Multidimensional Transfer Learning in Natural Language Processing: Domain, Task, Language, and Expertise Adaptation

by

Nghia Trung Ngo

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Computer Science

Dissertation Committee:

Thien Huu Nguyen, Chair

Thanh Hong Nguyen, Core Member

Yu Wang, Core Member

Kris Kyle, Institutional Representative

University of Oregon

Winter 2026

DISSERTATION ABSTRACT

Nghia Trung Ngo

Doctor of Philosophy in Computer Science

Title: Multidimensional Transfer Learning in Natural Language Processing: Domain, Task, Language, and Expertise Adaptation

Transfer learning in Natural Language Processing (NLP) mirrors a fundamental human cognitive ability: applying prior knowledge to learn new skills efficiently. Rather than training models from scratch for every new task, domain, or language, transfer learning enables systems to use knowledge from previous learning experiences. This capability is essential for addressing two critical challenges in modern NLP: the scarcity of labeled data affecting the vast majority of the world's languages and specialized domains, and the prohibitive computational cost of training large-scale models from scratch.

This dissertation presents a comprehensive investigation of transfer learning in NLP, organized around two distinct phases. The first phase, Adaptive Transfer Learning, addresses horizontal knowledge transfer across contexts through three research directions: unsupervised domain adaptation, where models must transfer across data distributions without target labels; task adaptation for structured prediction, where multiple interdependent tasks such as joint information extraction must be learned and transferred simultaneously across domains; and cross-lingual transfer, where knowledge from high-resource languages is shared to low-resource languages across typologically diverse language families. The second phase, Specialization Transfer Learning, addresses the emerging challenge of vertically refining general-purpose large language models for reliable deployment

in high-stakes expert domains. This phase focuses on developing comprehensive evaluation frameworks for Retrieval-Augmented Generation systems in the medical domain, assessing critical dimensions beyond standard accuracy such as knowledge sufficiency, multi-document integration, and robustness to noisy or misleading retrieved content.

Together, these contributions advance the understanding of how NLP models can be systematically transferred and specialized across domains, tasks, languages, and levels of expertise, spanning the full spectrum from data-limited to data-engineered paradigms. This dissertation includes both previously published and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Nghia Trung Ngo

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Hanoi University of Science and Technology, Hanoi, Vietnam

DEGREES AWARDED:

Doctor of Philosophy, Computer Science, 2026, University of Oregon
Bachelor of Science, Information Technology, 2021, Hanoi University of Science and Technology

AREAS OF SPECIAL INTEREST:

Transfer Learning, Multilingual and Multi-Task Learning
Retrieval-Augmented Generation (RAG)
Large Language Model Benchmarking
Cross-lingual Information Extraction

PROFESSIONAL EXPERIENCE:

Research Assistant, University of Oregon, 2022-Present
Teaching Assistant, University of Oregon, 2025-Present
NLP Research Residence, VinAI Research, 2020-2021
Research Student, Data Science Laboratory - HUST, 2018-2019
Reviewer: ACL, NAACL, EMNLP, AAAI, ICLR, NeurIPS, COLING, COLM.

GRANTS, AWARDS AND HONORS:

Erwin & Gertrude Juilfs Scholarship
Dept. of Computer Science, University of Oregon, 2023

Lorry I. Lokey Interdisciplinary Science Initiative Dissertation Scholarship
Dept. of Computer Science, University of Oregon, 2025

PUBLICATIONS:

Ngo, N. T., Dernoncourt, F., & Nguyen, T. H. (2025). mSCoRe: a Multilingual and Scalable Benchmark for Skill-based Commonsense Reasoning. *Proceeding of LREC 2026*.

Ngo, N. T., Nguyen, C. V., Dernoncourt, F., & Nguyen, T. H. (2025). Comprehensive and Practical Evaluation of Retrieval-Augmented Generation Systems for Medical Question Answering. *AAAI 2026 Workshop on AI for Scientific Research*.

Man, H., Ngo, N. T., Dernoncourt, F., & Nguyen, T. H. (2024). ULLME: A Unified Framework for Large Language Model Embeddings with Generation-Augmented Learning. In *Proceedings of EMNLP 2024: System Demonstrations*.

Man, H., Nguyen, C. V., Ngo, N. T., Ngo, L., Dernoncourt, F., & Nguyen, T. H. (2024). Hierarchical Selection of Important Context for Generative Event Causality Identification with Optimal Transports. In *Proceedings of LREC-COLING 2024*.

Ngo, N. T., & Nguyen, T. H. (2023). Zero-shot Cross-lingual Transfer Learning with Multiple Source and Target Languages for Information Extraction: Language Selection and Adversarial Training. *arXiv preprint arXiv:2411.08785*.

Nguyen, T., Nguyen, C. V., Lai, V. D., Man, H., Ngo, N. T., Dernoncourt, F., Rossi, R. A., & Nguyen, T. H. (2024). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of LREC-COLING 2024*.

Lai, V. D., Nguyen, C. V., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., & Nguyen, T. H. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of EMNLP 2023: System Demonstrations*.

Lai, V. D., Ngo, N. T., Ben Veyseh, A. P., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of EMNLP 2023*.

Ngo, N. T., Min, B., & Nguyen, T. (2022). Unsupervised domain adaptation for joint information extraction. In *Findings of EMNLP 2022*.

Ngo, N. T., Ngo, L. V., & Nguyen, T. H. (2022). Unsupervised domain adaptation for text classification via meta self-paced learning. In *Proceedings of COLING 2022*.

Man, H., Ngo, N. T., Ngo Van, L., & Nguyen, T. H. (2022). Selecting optimal context sentences for event-event relation extraction. In *Proceedings of AAAI 2022*.

Nguyen, M. V., Ngo, N. T., Min, B., & Nguyen, T. H. (2022). FAMIE: A fast active learning framework for multilingual information extraction. In *Proceedings of NAACL 2022: System Demonstrations*.

Ngo Trung, N., Phung, D., & Nguyen, T. H. (2021). Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of ACL-IJCNLP 2021*.

Ben Veyseh, A. P., Nguyen, M. V., Ngo Trung, N., Min, B., & Nguyen, T. H. (2021). Modeling document-level context for event detection via important context selection. In *Proceedings of EMNLP 2021*.

Ngo, N. T., Nguyen, T. N., & Nguyen, T. H. (2020). Learning to select important context words for event detection. In *Proceedings of PAKDD 2020*.

ACKNOWLEDGEMENTS

First and foremost, I owe my deepest gratitude to my advisor, Professor Thien Huu Nguyen. Your guidance and steady encouragement have been the foundation of this work, and your belief in my potential has meant more than I can easily put into words. Thank you for the thoughtful challenges, the patient feedback through the difficult stretches, and for always making space for me to pursue ideas I wasn't sure would work.

I am also sincerely grateful to my dissertation committee: Professor Thanh Nguyen, Professor Yu Wang, and Professor Kris Kyle. Your igorous feedback, valuable insights, and constructive suggestions have significantly contributed to the depth and breadth of my research. Each of you brought a distinct perspective that genuinely broadened my thinking, and your generosity with your time and expertise has been invaluable throughout this process.

To my labmates - Minh Van Nguyen, Hieu Man Duc Trong, Huy Huu Nguyen, Viet Dac Lai, Chien Van Nguyen, and Luis Fernando Guzman-Nateras - thank you for making the lab a place worth showing up to every day. Your support, both academically and personally, has been a constant source of encouragement. I'm proud to have worked alongside each of you.

I am also thankful to the faculty and staff of the Department of Computer and Information Science for fostering a supportive environment and making the day-to-day experience of doctoral study far more welcoming than it might otherwise have been.

Finally, I thank my family. Your love, patience, and encouragement have been my constant anchor through all of it. You believed in me before I believed in myself, and your sacrifices made this journey possible.

To my beloved family.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTION

The majority of this chapter's content is derived from my dissertation proposal, where I served as the primary author, while Thien Huu Nguyen contributed through editorial recommendations.

1.1 Overview

Transfer learning in machine learning draws direct inspiration from a fundamental human cognitive ability: the capacity to leverage prior knowledge when learning new skills. A clear example is the human educational process, where foundational knowledge is progressively taught and built upon, enabling students to acquire specialized expertise more efficiently. Similarly, when a bilingual speaker learns a third language from the same family, they transfer grammatical intuitions and vocabulary roots from their existing linguistic knowledge. This ability to recognize patterns, abstract concepts, and apply prior experience to novel situations is central to human intelligence and enables efficient learning across diverse domains and contexts. In the field of Natural Language Processing (NLP), transfer learning has emerged as a fundamental paradigm that mirrors this human capability (Pan & Yang, 2009; Ruder, Peters, Swayamdipta, & Wolf, 2019; Weiss, Khoshgoftaar, & Wang, 2016). Rather than training models from scratch for every new task, domain, or language, transfer learning enables systems to leverage knowledge acquired from previous learning experiences. This approach has proven essential for addressing two critical challenges in modern NLP. First, the scarcity of labeled training data affects the vast majority of the world's thousands of languages (Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020) as well as specialized domains where expert annotation is prohibitively expensive such as medical texts,

legal documents, and scientific literature. Second, the computational expense of training large-scale models from scratch for models like BERT (Devlin, Chang, Lee, & Toutanova, 2019) or GPT-3 (Brown et al., 2020) makes task-specific training economically infeasible for most applications. As the field has matured, transfer learning has evolved from a useful technique to an indispensable foundation: practically all state-of-the-art NLP systems now employ pretrained language models (PLMs) adapted for downstream tasks (Bommasani et al., 2021), from sentiment analysis and named entity recognition to question answering and machine translation.

Historically, NLP systems relied on rule-based methods and hand-crafted features, which limited their flexibility and scalability. The advent of deep learning marked a significant leap, with neural architectures such as recurrent and convolutional networks achieving superior performance on a range of tasks. However, these models still required large annotated datasets and were generally trained in isolation for specific applications (Hochreiter & Schmidhuber, 1997; Kim, 2014). The breakthrough came with the introduction of PLMs such as BERT and GPT (Devlin et al., 2019; Radford et al., 2019) that were first trained on massive unlabeled corpora to capture general linguistic patterns, semantics, and world knowledge. These models can then be fine-tuned or adapted for a wide array of downstream tasks, often achieving state-of-the-art results with limited task-specific data. This paradigm of leveraging pretrained models as a foundation for new tasks is the essence of transfer learning in modern NLP. This dissertation presents a comprehensive view of current transfer learning that recognizes two distinct phases, each addressing fundamentally different challenges and employing distinct approaches. The first phase, which we term *Adaptive Transfer Learning*, is

characterized by horizontal knowledge transfer across contexts at the same level of abstraction. We investigate adaptive transfer through three related research directions that address progressively complex challenges. First, *domain adaptation* tackles the challenge of transferring models between different data distributions when the task remains constant - for example, adapting sentiment analysis from movie reviews to product reviews, or event detection from news to social media. Building upon domain adaptation, *task adaptation* extends to scenarios where not only the data distribution shifts but the target task differs fundamentally from the source task, requiring adaptation of both representations and task-specific components. This becomes particularly complex for structured prediction problems such as joint information extraction (IE), where multiple interdependent subtasks must be learned simultaneously across domain boundaries. Finally, *cross-lingual transfer* extends these adaptation challenges across the linguistic dimension, enabling knowledge sharing from high-resource languages to low-resource languages. The emergence of Large Language Models (LLMs) - highly capable networks with billions of parameters - marks a transformation in transfer learning paradigms (Achiam et al., 2023; Brown et al., 2020). Unlike earlier PLMs that required substantial fine-tuning with domain-specific labeled data, modern LLMs trained on colossal datasets often exceeding a trillion tokens across hundreds of languages and domains exhibit remarkable emergent abilities for transferring general knowledge to specialized contexts through mechanisms such as in-context learning and instruction following (Ouyang et al., 2022). This transformation shifts the transfer learning challenge from horizontal transfer between specific contexts (adapting from one domain or language to another at the same level of abstraction) to what we term *Specialization Transfer Learning*: vertical refinement

from broad general-purpose capabilities to deep specialized expertise. Rather than addressing data scarcity through cross-context knowledge reuse, specialization transfer approaches the challenge from a data-engineering perspective, developing methods for synthetic data generation, sophisticated evaluation frameworks, and domain-specific alignment techniques that enable effective knowledge transfer from general capabilities to expert-level performance in specialized domains (Bommasani et al., 2021). In particular, We focus on medical applications that require not only comprehensive domain knowledge but also strict adherence to safety standards, clinical reasoning protocols, and real-world applicability constraints (Nori, King, McKinney, Carignan, & Horvitz, 2023; Singhal et al., 2023).

As we navigate the evolution of NLP research from early neural models to modern LLMs, the role of transfer learning has become increasingly central. This work recognizes both the remarkable progress achieved and the significant challenges that remain. Through systematic investigation across multiple dimensions of transfer learning, innovative methodological contributions, and comprehensive evaluation frameworks, this work aims to advance the field's understanding of how knowledge can be effectively transferred, adapted, and specialized to serve diverse applications and communities. The following sections formally define the problems addressed, present specific research questions, and outline the research directions pursued throughout this dissertation.

1.2 Problem Definitions

This dissertation addresses four critical directions of transfer learning problems, from adaptive transfer across domains, tasks, and languages to specialization transfer for expert domains.

Unsupervised Domain Adaptation. The first problem addresses scenarios where models must transfer across domains with different data distributions. Domain Adaptation (DA) is essential when labeled data exists in a source domain but the model must perform well in a target domain where data characteristics differ - vocabulary, writing style, and feature patterns shift between domains, causing source-trained models to perform poorly on target data. This dissertation focuses specifically on *Unsupervised Domain Adaptation* (UDA) where no labeled target data is available for supervision. For example, named entity recognition models trained on formal news text often fail on informal social media posts due to non-standard capitalization, misspellings, and novel entity types. UDA methods must learn domain-invariant representations that capture task-relevant features while remaining robust to domain-specific characteristics, using only labeled source data and unlabeled target data. The core challenge is distribution shift: how can models identify and leverage transferable knowledge while filtering out domain-specific noise that harms generalization?

Task Adaptation for Structured Prediction. The second problem extends DA to scenarios involving multiple interdependent tasks that must be learned jointly. In structured prediction problems such as joint IE, several subtasks (entity recognition, relation extraction, event detection, argument extraction) are connected through structural dependencies - relations link entities, events involve entities as arguments, and predictions in one task inform others. When adapting across tasks, not only do feature distributions shift, but the expression of these structural dependencies may also vary. For instance, how entities participate in events may be expressed differently in formal news articles versus informal social media posts. Standard DA techniques that treat all instances uniformly fail because

entity mentions may require different domain-invariant features than event triggers, and relation instances may experience different types of shift than event arguments. The challenge is thus two-fold: achieving domain invariance while preserving task-specific characteristics, and maintaining structural relationships that capture cross-task dependencies in the target domain.

Cross-lingual Transfer Learning. The third problem tackles knowledge transfer across languages in zero-shot settings where no labeled training data exists in the target language. Languages exhibit substantial diversity in morphology (isolating vs. agglutinative vs. fusional), syntax (word order, grammatical structures), phonology (sound systems), and orthography (writing systems). These differences create challenges for learning language-independent representations that capture semantic and structural patterns consistently across languages. The problem is particularly important for low-resource languages that lack sufficient data for independent model training. Effective cross-lingual transfer must identify which linguistic properties facilitate transfer, how to leverage multiple source languages optimally, and how to explicitly model relationships between source and target languages to improve upon generic language-invariant feature learning.

Expertise Specialization for Large Language Models. The fourth problem addresses the challenge of adapting general-purpose LLMs to specialized expert domains where accuracy, reliability, and safety are paramount. Unlike earlier transfer learning problems constrained by data scarcity, modern LLMs possess broad capabilities from training on vast corpora but may lack the depth, currency, and reliability required for high-stakes applications such as medicine or law. Retrieval-Augmented Generation (RAG) systems that augment LLM generation

with external knowledge sources offer a promising approach, but introduce new complexities: the system must determine when external knowledge is necessary, effectively integrate retrieved information, and remain robust to noisy or misleading retrieved content. Comprehensive evaluation beyond simple accuracy metrics is essential for ensuring reliable deployment in expert domains.

Information Extraction as Application Domain. To evaluate transfer learning methods, we employ information extraction (IE) as the primary application domain across multiple research directions. IE aims to automatically identify and extract structured information from unstructured text, encompassing four core tasks: (1) *Entity Mention Recognition* identifies and classifies text spans referring to persons, organizations, locations, and other named entities; (2) *Relation Extraction* detects and classifies semantic relationships between entity pairs such as employment or location relations; (3) *Event Trigger Detection* identifies words signaling event occurrences and classifies event types such as attacks or meetings; (4) *Event Argument Extraction* determines which entities participate in events and assigns semantic roles such as agent or patient. These tasks exhibit rich interdependencies where relations connect entities, events involve entities as arguments, and information from one task informs others.

IE provides an ideal testbed for transfer learning research for several reasons. First, IE systems face all dimensions of transfer investigated in this dissertation: domain adaptation (i.e. news to social media), task adaptation (joint modeling of interdependent subtasks), and cross-lingual transfer (English to low-resource languages). Second, the structured nature with interdependent subtasks enables investigation of how structural relationships are preserved during transfer. Third, IE has well-established benchmarks such as ACE 2005 and ERE across

multiple domains and languages for systematic evaluation. Finally, IE encompasses diverse specialized applications including medical information extraction for clinical question answering (addressed in our RAG evaluation framework), ensuring that transfer learning advances have practical impact beyond benchmark performance. The four problems defined above - domain adaptation, task adaptation for structured prediction, cross-lingual transfer, and expertise specialization - represent a comprehensive view of transfer learning challenges from data-limited to data-rich paradigms, which the following section translates into specific research questions.

1.3 Research Questions

This dissertation is anchored on four research questions (RQs) that directly address the transfer learning problems defined in the previous section. The RQs guide the technical investigations and methodological contributions presented in subsequent chapters. Specifically, this work examines:

- **RQ1: How can we effectively adapt models across domains when target domain labels are unavailable?** Domain adaptation without labeled target data remains challenging due to distribution shift. Existing methods either treat all training instances uniformly or introduce numerous hyperparameters requiring target-domain validation data. This question investigates whether meta-learning approaches can automatically balance different training objectives and select informative source instances to improve domain-invariant feature learning.

- **RQ2: How can structured prediction models be adapted across domains while preserving task interdependencies?** When multiple interdependent tasks such as entity recognition, relation extraction, and event detection must be jointly learned and transferred across domains, standard

domain adaptation techniques fail to account for task-specific characteristics and structural relationships. This question examines whether instance-level alignment combined with context-invariant structure learning can achieve effective adaptation for complex joint information extraction.

– **RQ3: What factors enable effective cross-lingual transfer, and how can multiple source languages be optimally leveraged?** Zero-shot cross-lingual transfer performance varies dramatically across language pairs and tasks, but the underlying mechanisms remain poorly understood. This question investigates how linguistic properties (phonology, morphology, syntax) correlate with transfer effectiveness, how to select optimal combinations of source languages given resource constraints, and whether explicitly modeling linguistic relationships through adversarial training can improve multi-lingual transfer.

– **RQ4: How should Retrieval-Augmented Generation systems be evaluated for reliable deployment in expert domains?** Large language models show promise for specialized applications when augmented with external knowledge, but existing evaluation focuses primarily on answer accuracy. This question explores what additional dimensions—sufficiency (when is external knowledge necessary), integration (can models effectively use retrieved context), and robustness (performance under noisy information)—are critical for ensuring reliable performance in high-stakes domains such as medicine.

These four RQs correspond directly to the four research directions pursued in this dissertation. The first three questions address adaptive transfer learning

challenges in progressively complex scenarios: domain adaptation (RQ1) establishes methods for handling distribution shift, task adaptation (RQ2) extends these methods to structured prediction with interdependent subtasks, and cross-lingual transfer (RQ3) addresses transfer across languages with diverse linguistic properties. The fourth question (RQ4) addresses specialization transfer learning for modern large language models, focusing on evaluation frameworks rather than architectural innovations. Together, these questions encompass the full spectrum of transfer learning from data-limited to data-rich paradigms.

1.4 Dissertation Outline

This dissertation pursues four distinct research directions (RDs) that address the four RQs described above. Each direction targets a fundamental dimension of transfer learning in NLP: domain adaptation, task adaptation, cross-lingual transfer, and expertise specialization. The research directions progress sequentially from foundational adaptive transfer learning methods addressing data scarcity and distribution shift, through structured prediction with interdependent tasks, to cross-lingual knowledge transfer, and finally to modern specialization techniques for LLMs.

1.4.1 RD1: Domain Adaptation. The first RD addresses RQ1 by investigating unsupervised domain adaptation, where models must transfer from a source domain with labeled data to a target domain with unlabeled data under distribution shift. Recent approaches such as domain-adversarial neural networks (DANN) (Ganin et al., 2016) employ adversarial training to learn domain-invariant representations, while other methods combine multiple training objectives including semi-supervised regularizers and discrepancy metrics to improve adaptation (Bousmalis, Trigeorgis, Silberman, Krishnan, & Erhan, 2016;

A. Kumar et al., 2018; Long, Cao, Wang, & Jordan, 2015; R. Shu, Bui, Narui, & Ermon, 2018). However, existing methods face two critical limitations. First, they treat all training instances uniformly, overlooking that source instances with high classification loss - those near decision boundaries or in low-density regions - may introduce noise that harms target generalization. When domain shift occurs, high-loss regions of the source distribution can overlap with high-density target regions, causing models to learn misleading patterns. Second, methods that combine multiple objectives introduce numerous domain-sensitive hyperparameters requiring careful tuning, which becomes impractical with large-scale transformer models and is further complicated by the absence of labeled target validation data for standard hyperparameter optimization.

To address these challenges, we propose Meta Self-Paced Domain Adaptation (MSP-DA) (Trung, Van, & Nguyen, 2022), a novel framework that leverages meta-learning (Hospedales, Antoniou, Micaelli, & Storkey, 2021) to automatically optimize both instance selection and objective weighting without requiring target labels. The key insight is a domain-shift variation of the cluster assumption: by dynamically partitioning source data into low-loss (meta-source) and high-loss (meta-target) subsets, we create a virtual adaptation scenario that forces the model to develop robust transfer capabilities. MSP-DA introduces a neural self-paced learning (neural-SPL) module with a learnable age threshold that controls instance selection while learning both instance-wise weights for classification and layer-wise weights for adversarial training across transformer layers. The weighted objectives on the meta-source domain are optimized to improve performance on the meta-target domain through meta-learning, with both module parameters and the age threshold updated dynamically based on

meta-validation performance - effectively performing hyperparameter optimization without target labels. Extensive evaluation on the FDU-MTL dataset (P. Liu, Qiu, & Huang, 2017) for sentiment analysis and ACE-05 dataset (Walker, Strassel, Medero, & Maeda, 2005) for event detection demonstrates that MSP-DA significantly outperforms state-of-the-art baselines including DANN variants with self-paced learning. Ablation studies validate each component's effectiveness, and analyses reveal how learned hyperparameters adapt to different domain characteristics. Chapter II presents complete technical details and experimental analysis.

1.4.2 RD2: Task Adaptation for Structured Prediction. The second RD extends domain adaptation to structured prediction with multiple interdependent tasks, addressing RQ2 through the lens of joint information extraction (JIE). JIE aims to simultaneously perform entity recognition, relation extraction, event detection, and argument extraction, leveraging task dependencies to avoid error propagation. Recent JIE systems using large-scale pretrained language models (Y. Lin, Ji, Huang, & Wu, 2020; D. Q. Nguyen, Nguyen, & Pham, 2021) achieve state-of-the-art performance within single domains but fail when training and test data come from different domains with different distributions. The challenge is substantially more complex than single-task domain adaptation: models must handle domain shifts across multiple task types simultaneously (entities, events, relations, arguments), each potentially exhibiting different shift patterns. Standard domain adaptation techniques (Ganin et al., 2016; A. Kumar et al., 2018; Long et al., 2015) that uniformly align all instances fail because entity mentions may require different domain-invariant features than event triggers, and relation instances may experience different types of shift than event

arguments. Furthermore, classical domain adaptation assumptions like covariate shift do not hold for joint structured prediction, and heuristic linguistic structures (e.g., dependency graphs) used by previous JIE systems (Y. Lin et al., 2020; M. V. Nguyen, Lai, & Nguyen, 2021a; Veyseh, Nguyen, & Nguyen, 2020) can introduce domain-specific information that harms adaptation.

To address this gap, we propose Domain Adaptation for Joint Information Extraction (DA4JIE) (Ngo, Min, & Nguyen, 2022), the first method specifically designed for unsupervised domain adaptation in JIE. DA4JIE introduces two key innovations. First, the Instance-relational Domain Adaptation (IrDA) module views task instances as nodes in a domain-instance relational graph and performs adversarial learning on pairwise node relationships, where a graph discriminator (Z. Xu, He, Lee, Wang, & Wang, 2022) attempts to recover domain information from instance pairs while the encoder prevents this. Unlike standard domain-adversarial training that enforces uniform alignment, IrDA uses a chain connection structure that aligns representations at the instance-type level (entity-to-entity, event-to-event, relation-to-relation, argument-to-argument), preserving task-specific characteristics while achieving domain invariance. Second, the Context-invariant Structure Learning (CiSL) module uses Graph Transformer Networks (Yun, Jeong, Kim, Kang, & Kim, 2019) to fuse attention-based graphs extracted from each transformer layer with syntactic structures, creating context-invariant graphs that provide domain-agnostic structural information for instance encoding. Comprehensive evaluation on ACE-05 (Walker et al., 2005) across multiple domain pairs (news to social media and others) demonstrates that DA4JIE significantly improves out-of-domain performance for all four IE tasks compared to state-of-the-art JIE systems and domain adaptation baselines. Ablation studies

validate both IrDA and CiSL components, and analysis reveals how learned relational graphs capture meaningful cross-task dependencies while remaining domain-invariant. Chapter III presents complete methodology and experimental analysis.

1.4.3 RD3: Cross-lingual Transfer Learning. The third RD investigates zero-shot cross-lingual (ZSCL) transfer learning for information extraction, addressing RQ3. While multilingual pretrained models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have enabled effective cross-lingual transfer, most prior research focuses on single-transfer settings (one source language to one target language), providing limited understanding for building multilingual systems that generalize across many languages. Cross-lingual transfer faces fundamental challenges from linguistic diversity: languages differ in morphology (isolating vs. agglutinative vs. fusional), syntax (word order, grammatical structures), phonology, and writing systems, creating difficulties for learning language-independent representations. Transfer performance varies substantially across language pairs and tasks, but the underlying mechanisms remain poorly understood. Key open questions include: which linguistic properties correlate with transfer effectiveness, how to optimally leverage multiple source languages given resource constraints, and whether explicitly modeling source-target linguistic relationships can improve transfer beyond learning generic language-invariant features.

We investigate cross-lingual transfer in (Ngo & Nguyen, 2023) by providing comprehensive evaluations through three progressively complex ZSCL transfer scenarios. In the single-transfer setting (ZSCL-S), we extract phylogenetic and typological properties from the URIEL Typological Database (Littell et al., 2017)

across syntax, phonology, inventory, and morphology dimensions, developing a combined linguistic distance metric that achieves high correlation with transfer performance across diverse language pairs and tasks. In the multi-transfer setting (ZSCL-M) with multiple source and target languages, we develop methods for selecting optimal source language combinations based on linguistic clustering, demonstrating that theory-guided language selection significantly outperforms random selection and provides practical guidance for multilingual data collection with optimal cost-performance trade-offs. In the relational-transfer setting (ZSCL-R), we propose a graph-relational adversarial learning framework that explicitly models source-target linguistic relationships using separate discriminators for each source language conditioned on a language relational graph induced from linguistic distances. Extensive experiments on event detection and relation extraction across 17 typologically diverse languages (Arabic, Chinese, Hindi, Japanese, Korean, and European languages) validate our approaches, with relational-transfer achieving considerable improvements by explicitly capturing linguistic connections rather than learning only language-invariant features. Chapter IV presents the complete methodology, linguistic feature analysis, and experimental results.

1.4.4 RD4: Expertise Specialization for Large Language Models.

The fourth RD transitions to specialization transfer learning, addressing RQ4 by investigating reliable deployment of general-purpose LLMs in high-stakes expert domains. Large language models have demonstrated remarkable capabilities in medical applications, achieving competitive performance on medical licensing exams and question-answering benchmarks (Nori et al., 2023; Singhal et al., 2023). Retrieval-Augmented Generation (RAG) systems that augment LLM generation with external knowledge retrieval have emerged as a promising approach to reduce

hallucination and improve factual accuracy (Gao et al., 2023; P. S. H. Lewis et al., 2020). However, existing medical benchmarks focus predominantly on answer accuracy in standard settings, providing only question-answer pairs without systematic evaluation of how LLMs interact with external knowledge (He et al., 2023; Q. Jin, Dhingra, Liu, Cohen, & Lu, 2019a; W. Xiong et al., 2024). In practice, retrieved documents can contain not only useful knowledge but also noise, irrelevant information, or factual errors that can mislead models. Critical questions remain unanswered: When is external knowledge truly necessary versus relying on parametric knowledge? Can LLMs effectively integrate information across multiple retrieved documents? How robust are systems to noisy or factually incorrect retrieved content? These dimensions are crucial for reliable deployment in healthcare where errors can have serious consequences, yet they are not addressed by existing evaluation frameworks.

To address this gap, we introduce Medical Retrieval-Augmented Generation Benchmark (MedRGB) (Ngo, Nguyen, Dernoncourt, & Nguyen, 2024), a comprehensive evaluation framework assessing medical RAG systems across four critical dimensions. The standard-RAG scenario establishes baseline performance with relevant retrieved documents. The sufficiency scenario evaluates whether LLMs correctly determine when external knowledge is necessary by including "Insufficient Information" as a response option when noise documents are present, requiring models to filter unreliable information and recognize limitations of their parametric knowledge. The integration scenario tests compositional reasoning by requiring systems to answer supporting sub-questions and integrate extracted information to address complex queries. The robustness scenario evaluates resilience to factual errors in retrieved context, requiring systems to detect

incorrect information and provide corrections rather than being misled. We construct MedRGB from four diverse medical QA datasets covering multiple specialties, developing specific data augmentation procedures for each scenario and establishing evaluation protocols that assess both answer accuracy and behavioral appropriateness. Extensive evaluation of commercial LLMs (GPT-4, Claude, Gemini) and open-source models (Llama, Mistral) reveals significant limitations: even the most capable systems show substantial performance degradation in sufficiency and robustness scenarios, often over-relying on retrieved context even when contradicting parametric knowledge, struggling with multi-document integration, and failing to identify subtle factual errors. These findings provide concrete directions for developing more reliable RAG systems including improved retrieval quality assessment, better context integration mechanisms, and explicit fact-verification components. To our knowledge, MedRGB is the first benchmark comprehensively assessing medical RAG across all these practical scenarios. Chapter V presents complete benchmark construction, evaluation protocols, and error analysis.

CHAPTER II

UNSUPERVISED DOMAIN ADAPTATION VIA META SELF-PACED LEARNING

This chapter contains materials from the published paper: Nghia Ngo Trung, Linh Ngo Van, and Thien Huu Nguyen. "Unsupervised Domain Adaptation for Text Classification via Meta Self-Paced Learning." In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 2022. As the first author, Nghia was responsible for problem formulation, model development, experimental design, and manuscript preparation. Linh contributed to the meta-learning framework design and provided technical discussions. Thien provided research guidance, editorial revisions, and advised on experimental approach. The content has been adapted to fit the dissertation format and narrative.

Our first research direction (RD1) addresses the challenge of unsupervised domain adaptation - transferring knowledge from labeled source domains to unlabeled target domains when significant distribution shifts exist. This chapter introduces **Meta Self-Paced Domain Adaptation (MSP-DA)**, a novel meta-learning framework that automatically learns to balance training objectives and select informative examples for text classification and event detection tasks. MSP-DA dynamically partitions source training data into a low-loss meta-source domain and a high-loss meta-target domain, creating a virtual adaptation scenario within the source domain itself. The key innovation is a neural self-paced learning module that simultaneously learns three critical components: (i) instance-wise weights for the classification task, (ii) layer-wise weights for domain adversarial training across transformer layers, and (iii) a learnable *age* threshold for data selection - all optimized via meta-learning without requiring manual hyperparameter tuning.

Comprehensive experiments on sentiment analysis (**FDU-MTL**) and event detection (**ACE-05**) demonstrate that MSP-DA significantly outperforms state-of-the-art baselines for both tasks. The framework exhibits remarkable consistency across domains using the same hyperparameters, eliminating the need for domain-specific configuration.

2.1 Introduction

The challenge of unsupervised domain adaptation (UDA) is fundamental to transfer learning: while models achieve excellent performance on their training domains, they often fail when applied to new domains where data distributions differ substantially. Existing domain adaptation (DA) approaches require extensive manual tuning of multiple hyperparameters (loss weights, learning rates, regularization strengths) for each domain pair - a process that is impractical when using large pretrained language models due to computational constraints. This chapter proposes a meta-learning approach that addresses these limitations by automatically learning domain-specific hyperparameters through a principled optimization process.

Given sufficient supervision, modern deep learning models can learn new tasks with remarkable accuracy. However, in many practical settings, the goal is to adapt to a new domain where the data distribution differs significantly between training and testing. This poses a major challenge for standard NLP systems due to both the inherent variation of language such as lexical shift, semantic shift, and stylistic differences, and external factors including how textual datasets are collected, annotated, and sourced. For example, a sentiment classifier trained on movie reviews may achieve high accuracy on that domain but fail when applied to product reviews, where vocabulary, writing style, and sentiment

expressions differ substantially. Similarly, a model trained to predict news events may easily recognize "died" as an event trigger in medical text, but would fail to detect domain-specific events such as "mutation" or "cancer." Such models may even struggle to generalize to seemingly similar adaptation settings, such as news articles from different time periods or sources. These challenges reflect the fundamental problem of *distribution shift* between source and target domains where the joint distribution over inputs and labels differs. Without labeled target data, standard supervised learning approaches cannot directly optimize for target domain performance, requiring specialized DA techniques.

The majority of existing unsupervised domain adaptation (UDA) approaches address distribution shift by combining various training objectives to align different aspects of domain-specific features. The most common approach is domain-adversarial neural networks (DANN) (Ganin et al., 2016), which employs adversarial training between a domain classifier and the model's feature extractor to learn discriminative and domain-invariant representations. The simplicity and effectiveness of DANN have led researchers to incorporate it with multiple complementary objectives, including semi-supervised learning regularizers (R. Shu et al., 2018), discrepancy metrics (Long et al., 2015), co-training (A. Kumar et al., 2018), and auxiliary tasks (Bousmalis et al., 2016). Each of these components can enhance DA capability by addressing different aspects of the transfer problem. However, applying these multi-objective techniques to textual tasks presents significant practical challenges. Modern NLP relies on large transformer-based language models such as **BERT** (Devlin et al., 2019) and **RoBERTa** (Y. Liu et al., 2019), which are essential for achieving state-of-the-art performance but require substantial computational resources for fine-tuning. The combination of multiple

loss terms in existing domain adaptation methods introduces numerous domain-sensitive hyperparameters - loss weights, learning rates for different components, regularization strengths - that require careful tuning for each specific adaptation scenario. This hyperparameter sensitivity becomes especially problematic in the unsupervised setting, where no labeled target validation data is available to guide the tuning process. The computational expense of fine-tuning language models makes exhaustive hyperparameter search impractical, particularly when deploying systems across multiple domain pairs.

A critical but often overlooked issue in DA is that not all source domain examples contribute equally to effective transfer. Standard DA methods treat all training instances uniformly, overlooking the fact that source examples with high classification loss - those lying near decision boundaries or in low-density regions - may introduce noise that harms target domain generalization. Our approach is motivated by a domain-shift variation of the cluster assumption from semi-supervised learning (Chapelle, Schölkopf, & Zien, 2006), which states that data points of the same class should concentrate around high-density, low-loss regions. When adapting between dissimilar domains, these regions may shift significantly. As illustrated in Figure 1, high-loss regions of the source decision boundary can overlap considerably with high-density target clusters, causing models to learn misleading patterns that fail to transfer effectively.

To address these challenges, we propose **Meta Self-Paced Domain Adaptation (MSP-DA)**, a novel framework that uses meta-learning to train a neural network-based self-paced learning procedure in an end-to-end manner. The key innovation of MSP-DA is the dynamic partitioning of source training data into a low-loss *meta-source domain* and a high-loss *meta-target domain*, creating a

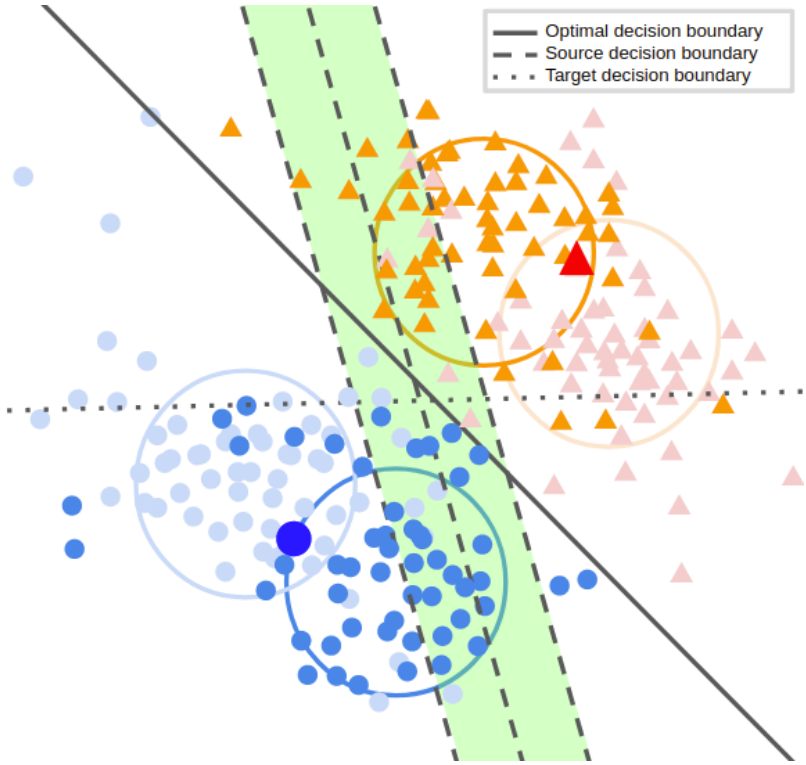


Figure 1. Illustration of domain shift between source domain (gray) and target domain (orange). The source decision boundary (green dashed lines) passes through high-loss regions (lime) that significantly overlap with high-density target clusters. Our approach dynamically partitions source data into low-loss (solid circles) and high-loss (faded circles) regions to create a meta-learning scenario that improves adaptation.

challenging virtual adaptation scenario within the source domain itself. By learning to successfully adapt from the meta-source to the meta-target domain, the model develops robust transfer capabilities applicable to the true target domain. This partitioning is controlled by a learnable *age* hyperparameter that is automatically tuned throughout training based on meta-test performance, eliminating the need for manual scheduling or domain-specific configuration.

Our framework introduces a *neural self-paced learning (neural-SPL) module* that simultaneously addresses multiple challenges in domain adaptation. First, the module controls the data selection process for meta-training and meta-testing using

a learnable *age* threshold, automatically determining which source examples are sufficiently "easy" to use for meta-training versus which "hard" examples should constitute the meta-test set. Second, the module learns instance-wise weights for the main classification task, enabling the model to focus on informative examples while reducing influence of potentially noisy or misleading instances. Third, the module learns layer-wise weights for domain adversarial training across different layers of the pretrained transformer encoder, recognizing that different layers encode different types of information (Rogers, Kovaleva, & Rumshisky, 2020) and may require different degrees of alignment.

The meta-learning process operates in two alternating steps. In the *meta-train step*, the model is trained on the low-loss meta-source domain with weighted objectives that combine classification loss and domain adversarial training. In the *meta-test step*, the model's performance is evaluated on the high-loss meta-target domain, providing learning signals to update not only the model parameters θ but also the hyperparameters encoded in the neural-SPL module ϕ (instance weights, layer weights, and *age* threshold). This optimization enforces that gradient updates on the meta-source domain should align with improvements on the meta-target domain, encouraging the model to learn representations and decision boundaries that generalize across the domain shift.

While the meta-target set does not contain samples from the true target domain, we argue that this approach is highly useful for unsupervised domain adaptation for two key reasons. First, the partition creates two virtual domains with significant discrepancy in terms of loss landscape. Learning to adapt in this challenging setting forces the model to develop robust transfer capabilities that generalize to other, possibly easier, domain shifts. Second, based on the cluster

assumption, low-loss regions of the target domain are likely to have considerable intersection with high-loss regions of the source domain when domains are dissimilar. Thus, by learning to adapt to the high-loss meta-target domain within the source data, the model implicitly learns to handle portions of the true target domain's distribution that would otherwise be challenging.

To our knowledge, this is the first work to devise a neural network-based self-paced learning method where both the sample weightings and the *age* hyperparameter are dynamically optimized through meta-learning. Previous SPL approaches (L. Jiang, Meng, Yu, et al., 2014; M. P. Kumar, Packer, & Koller, 2010) require heuristic *age* schedules (e.g., linearly increasing the *age* parameter over epochs) and rely on complicated mathematical derivations for instance weighting that are specific to particular loss functions. In contrast, our neural-SPL module learns these quantities jointly from data, generalizing naturally across different tasks and domain pairs without manual tuning.

We provide extensive evaluation of MSP-DA on two complementary benchmarks that represent different levels of task complexity and domain shift. For sentiment analysis, we evaluate on the standard FDU-MTL dataset (P. Liu et al., 2017), which includes reviews from 16 diverse domains (books, electronics, movies, etc.). For event detection, we conduct experiments on the **ACE-05** dataset (Walker et al., 2005) with substantially more challenging adaptation settings involving shifts from news text to conversational telephone speech, broadcast conversations, and web logs. Event detection represents a harder adaptation scenario due to its larger label space (33 event types plus a null class) and the semantic complexity of detecting triggers that signal events. Our comprehensive experiments demonstrate that MSP-DA significantly outperforms state-of-the-art domain adaptation

baselines across all settings. For sentiment analysis, MSP-DA achieves 93.0% average accuracy, outperforming the previous best method (BERT-Masker) by 1.5 percentage points and winning on 11 out of 16 domains. For event detection, MSP-DA achieves 72.2 average F1 score on out-of-domain settings, providing substantial improvements of 5.1 F1 points over the best baseline method. Notably, MSP-DA exhibits remarkable consistency across all domains using the same hyperparameters, demonstrating its robustness and eliminating the need for domain-specific tuning.

Contributions The main contributions of this chapter are:

- A novel meta-learning framework for unsupervised domain adaptation that dynamically partitions source data into meta-train and meta-test domains based on learned difficulty, creating a virtual adaptation scenario that improves transfer to the true target domain.
- A neural self-paced learning module that jointly optimizes three critical components: (i) instance-wise weights for the classification task, (ii) layer-wise weights for domain adversarial training, and (iii) a learnable *age* threshold for data selection - all without requiring manual hyperparameter tuning or domain-specific configuration.
- A theoretical analysis demonstrating how the meta-learning objective enforces gradient alignment between meta-source and meta-target domains, preventing overfitting to source-specific patterns and promoting robust transfer.
- Comprehensive experiments on sentiment analysis (FDU-MTL with 16 domains) and event detection (ACE-05 with 4 challenging adaptation settings), demonstrating consistent improvements over state-of-the-art

baselines, with average accuracy gains of 1.5% for sentiment analysis and 5.0 F1 points for event detection.

– Detailed ablation studies and visualizations that validate each component and provide insights into the learned hyperparameters, including analysis of how *age* thresholds and layer weights evolve differently across domains.

2.2 Background

2.2.1 Problem Formulation: Unsupervised Domain Adaptation.

Unsupervised domain adaptation addresses scenarios where a model must transfer from a *source domain* with abundant labeled data to a *target domain* where labels are unavailable or prohibitively expensive to obtain. Formally, let $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ denote the labeled source domain dataset, where x_i^s represents an input example (e.g., a sentence or document) and $y_i^s \in \mathcal{Y}$ its corresponding label from a shared label space $\mathcal{Y} = \{1, 2, \ldots, K\}$ of K classes. The target domain provides only unlabeled data $\mathcal{D}_T = \{x_j^t\}_{j=1}^{n_t}$ drawn from a different distribution.

The key characteristic of UDA is that the source and target distributions differ: $P_S(X, Y) \neq P_T(X, Y)$ (notation: P_S = source distribution, P_T = target distribution). This distribution shift between source and target domains can manifest in several forms (Pan & Yang, 2009; Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009). *Covariate shift* occurs when the input distributions differ ($P_S(X) \neq P_T(X)$) but the conditional label distributions remain the same ($P_S(Y|X) = P_T(Y|X)$). This is common when data is collected from different sources but the underlying task remains consistent. *Label shift* (or prior probability shift) refers to changes in the label distribution ($P_S(Y) \neq P_T(Y)$) while feature distributions given labels stay constant ($P_S(X|Y) = P_T(X|Y)$). *Concept drift* represents the most challenging scenario where the relationship between inputs

and labels changes across domains ($P_S(Y|X) \neq P_T(Y|X)$), meaning that the same input features may have different semantic meanings or lead to different labels in different domains.

The objective in UDA is to learn a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ that minimizes the expected risk on the target domain:

$$\min_\theta \mathbb{E}_{(x,y)\sim P_T}[\ell(f_\theta(x), y)] \tag{2.1}$$

where ℓ is a loss function (e.g., cross-entropy for classification). The fundamental challenge is that we cannot directly optimize this objective because target labels are unavailable. Instead, UDA methods must use the labeled source data $\mathcal{D}_S$ and unlabeled target data $\mathcal{D}_T$ during training to learn representations and decision boundaries that generalize across the domain shift.

2.2.2 Domain-Adversarial Neural Networks. Domain-adversarial neural networks (DANN) (Ganin et al., 2016) have emerged as the main approach for UDA in deep learning. DANN addresses domain shift by learning representations that are simultaneously discriminative for the main task and invariant to the domain. This is achieved through an adversarial training procedure inspired by generative adversarial networks (Goodfellow et al., 2014). The DANN architecture consists of three main components: (i) a feature extractor G_f that maps inputs to a shared representation space, (ii) a label predictor G_y that predicts task labels from the representations, and (iii) a domain discriminator G_d that attempts to classify whether a representation comes from the source or target domain. For an input x (from either domain), the feature extractor computes a representation vector: $h = G_f(x; \theta_f)$, where θ_f denotes the parameters of the feature extractor. The label predictor G_y is trained using labeled source data to

minimize the standard classification loss:

$$\mathcal{L}_c(\theta_f, \theta_y) = \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\ell(G_y(G_f(x;\theta_f);\theta_y), y)] \tag{2.2}$$

where θ_y are the label predictor parameters. Simultaneously, the domain discriminator G_d is trained to distinguish source from target representations:

$$\mathcal{L}_d(\theta_f, \theta_d) = -\mathbb{E}_{x\sim\mathcal{D}_S}[\log G_d(G_f(x;\theta_f);\theta_d)] - \mathbb{E}_{x\sim\mathcal{D}_T}[\log(1 - G_d(G_f(x;\theta_f);\theta_d))] \tag{2.3}$$

where θ_d are the domain discriminator parameters, and the domain labels are 1 for source and 0 for target.

The key idea of DANN is that the feature extractor parameters θ_f are updated to maximize the domain discriminator's loss (i.e., to fool the discriminator), while the discriminator parameters θ_d are updated to minimize it. This creates a minimax game:

$$\min_{\theta_f,\theta_y} \max_{\theta_d} \mathcal{L}_c(\theta_f, \theta_y) - \lambda\mathcal{L}_d(\theta_f, \theta_d) \tag{2.4}$$

where λ is a hyperparameter controlling the trade-off between classification performance and domain invariance. In practice, this is implemented using a gradient reversal layer (Ganin et al., 2016) that multiplies gradients from the domain discriminator by $-\lambda$ during backpropagation, allowing standard gradient descent to solve the minimax objective.

2.2.3 Self-Paced Learning. Self-paced learning (SPL) (M. P. Kumar et al., 2010) is a curriculum learning strategy that enables models to jointly learn the curriculum and the task itself. Inspired by the observation that humans learn more effectively when material is presented from easy to hard, SPL allows the model to automatically determine which training examples are "easy" at the current stage of learning and should be emphasized, versus which "hard" examples should be deferred or reduced in weight. Formally, given a training dataset

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and a loss function ℓ, standard supervised learning minimizes the average loss:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) \tag{2.5}$$

Self-paced learning reformulates this as a joint optimization over both model parameters θ and instance weights $v = [v_1, \ldots, v_n]$:

$$\min_{\theta, v} \sum_{i=1}^{n} v_i \ell_i(\theta) + g(v; \lambda_a) \tag{2.6}$$

where $\ell_i(\theta) = \ell(f_\theta(x_i), y_i)$ is the loss on example i, and $g(v; \lambda_a)$ is a regularization term on the weights that depends on an "age" hyperparameter λ_a representing the learning pace.

Kumar et al. (M. P. Kumar et al., 2010) showed that with an appropriate choice of regularizer g, the optimal weights have a simple binary form controlled by the *age* parameter:

$$v_i^* = \begin{cases} 1, & \text{if } \ell_i(\theta) < \lambda_a \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

Intuitively, λ_a serves as a threshold: examples with loss below λ_a are considered "easy" at the current learning stage and receive full weight, while harder examples with higher loss are excluded. As training progresses, λ_a is gradually increased according to a predefined schedule (e.g., $\lambda_a(t) = \lambda_0 \cdot (1 + t/T)$, where t = training step, T = total steps), allowing progressively harder examples to be incorporated into the training curriculum.

2.2.4 Meta-Learning for Domain Adaptation. Meta-learning, or "learning to learn," aims to develop algorithms that can quickly adapt to new tasks by using experience from related tasks (Hospedales et al., 2021). Rather than optimizing performance on a single task, meta-learning optimizes the

learning process itself, typically by training on a distribution of tasks to learn an initialization, update rule, or other meta-level knowledge that helps rapid adaptation.

Model-Agnostic Meta-Learning (**MAML**) (Finn, Abbeel, & Levine, 2017) is a well-known meta-learning algorithm that learns an initialization of model parameters such that a few gradient steps on a new task lead to good performance. Given a distribution of tasks $p(\mathcal{T})$, MAML optimizes:

$$\min_{\theta} \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})} \left[\mathcal{L}_{\mathcal{T}}^{\text{test}}(\theta - \alpha\nabla_{\theta}\mathcal{L}_{\mathcal{T}}^{\text{train}}(\theta))\right] \tag{2.8}$$

where $\mathcal{L}_{\mathcal{T}}^{\text{train}}$ and $\mathcal{L}_{\mathcal{T}}^{\text{test}}$ are losses computed on training and test sets for task $\mathcal{T}$, and α is an inner-loop learning rate. The key insight is that θ is optimized such that after one gradient step on a task's training data (the inner loop), the resulting parameters perform well on that task's test data (evaluated in the outer loop).

Meta-learning has been successfully applied to domain generalization (DG) (Dou, de Castro, Kamnitsas, & Glocker, 2019; D. Li, Yang, Song, & Hospedales, 2018), where the goal is to learn from multiple source domains and generalize to unseen target domains. The **Meta-Learning for Domain Generalization (MLDG)** algorithm (D. Li et al., 2018) adapts the **MAML** framework by simulating domain shift within the source domains. During training, source domains are split into meta-train and meta-test sets, and the model is optimized to perform well on the meta-test domains after being trained on meta-train domains. This encourages learning representations and decision boundaries that generalize across domain shifts. While DG and UDA share similarities, they differ in a crucial aspect: DG assumes access to multiple labeled source domains but no access to target data (even unlabeled), whereas UDA provides labeled source data and unlabeled target data but typically only a single source domain.

The MLDG approach cannot be directly applied to UDA because: (i) with a single source domain, there is no natural way to create meta-train and meta-test domain splits, and (ii) the unlabeled target data cannot serve as a clean validation set for the meta-test step.

2.3 Proposed Model: Meta Self-Paced Domain Adaptation

2.3.1 Overview. Figure 2 illustrates the overall architecture of our proposed MSP-DA framework. The model consists of three main modules: (i) a pretrained transformer encoder (**BERT** (Devlin et al., 2019) or **RoBERTa** (Y. Liu et al., 2019)) with small adapter layers (Houlsby et al., 2019) for efficient fine-tuning, (ii) task-specific prediction heads for classification and domain discrimination, and (iii) a neural self-paced learning (neural-SPL) module that learns to control the training process.

We denote the labeled source domain dataset as $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s examples, and the unlabeled target domain dataset as $\mathcal{D}_T = \{x_i^t\}_{i=1}^{n_t}$ with n_t examples. The shared label space is $\mathcal{Y} = \{1, 2, \ldots, K\}$ consisting of K classes. Our model's learnable parameters are denoted as $\theta = (\theta_a, \theta_c, \theta_d)$, which includes the parameters of the adapter layers θ_a, the main classification head θ_c, and the domain discriminator heads θ_d. The neural-SPL module has parameters $\phi = (\theta_v, \theta_w, \lambda_a)$, consisting of the instance weighting network parameters θ_v, the layer weighting network parameters θ_w, and the *age* hyperparameter λ_a.

Following recent work on efficient fine-tuning (Houlsby et al., 2019; Pfeiffer, Kaur, & Roth, 2020), we keep the pretrained transformer encoder's weights frozen and insert small adapter modules after each transformer block. Each adapter consists of a down-projection to a bottleneck dimension d_a (typically 48-128), a nonlinear activation, and an up-projection back to the transformer's hidden

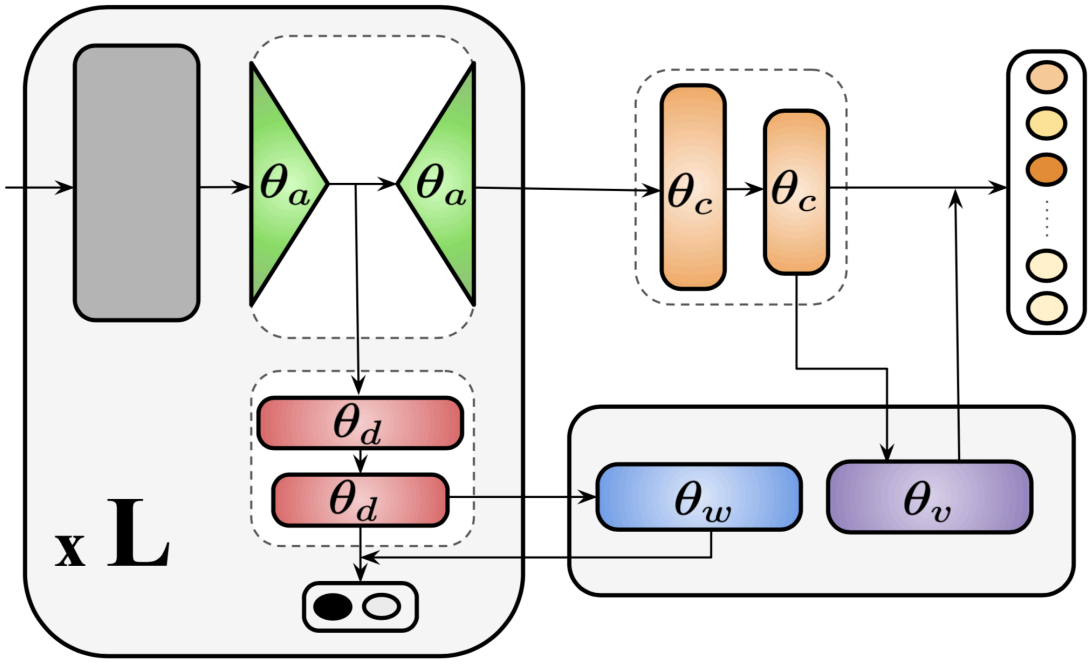


Figure 2. Architecture overview of Meta Self-Paced Domain Adaptation. The pretrained BERT encoder (gray) is kept fixed, while adapter layers (green) at each transformer block provide trainable bottleneck representations. These adapter outputs feed into domain classifier heads (red) for adversarial training. The neural-SPL module consists of an instance-wise weighting head (purple) that computes weights for the classification task (orange) and a layer-wise balancing head (blue) that determines the importance of aligning each layer's representations across domains.

dimension d_h (typically 768 for BERT-base). For an input text x, the pretrained BERT encoder produces hidden states $H^{(0)}, H^{(1)}, \ldots, H^{(L)}$ at each of its L layers (typically $L = 12$ for BERT-base). The adapter at layer ℓ transforms its input to produce a bottleneck representation: $z^{(\ell)} = \text{Adapter}^{(\ell)}(H^{(\ell)}) \in \mathbb{R}^{d_a}$. These low-dimensional bottleneck representations $\{z^{(1)}, \ldots, z^{(L)}\}$ serve two purposes: they provide adapter-specific features for the domain discriminators, and they reduce the number of trainable parameters, making the approach computationally efficient.

Following prior work demonstrating that different layers of pretrained transformers encode different types of information (Rogers et al., 2020), we

employ a separate domain discriminator head at each adapter layer. The ℓ-th domain discriminator takes the bottleneck representation $z^{(\ell)}$ and outputs a binary prediction of whether the input comes from the source or target domain: $p_d^{(\ell)} = \text{DomainClassifier}^{(\ell)}(z^{(\ell)}; \theta_d^{(\ell)}) \in [0, 1]$. This multi-layer approach allows the model to align representations at different levels of abstraction, from surface-level features in lower layers to semantic features in higher layers.

For the main classification task, we use the representation from the final layer. Specifically, we take the [CLS] token representation from the final adapter layer and pass it through a feed-forward classification head: $p_y = \text{softmax}(\text{FFN}_c(z_{[CLS]}^{(L)}; \theta_c)) \in \mathbb{R}^K$. where $z_{[CLS]}^{(L)}$ is the [CLS] token's bottleneck representation from the final adapter layer, and FFN_c is a feed-forward network with parameters θ_c that produces logits over the K classes.

The key idea of our approach is the dynamic partitioning of the source domain into two virtual domains based on current instance losses. At each training step, we compute the classification loss $\ell_i = -\log p_y(y_i^s | x_i^s; \theta)$ for each source example (x_i^s, y_i^s). Using the learnable *age* threshold λ_a, we partition the source data into:

- **Meta-source domain $\mathcal{S}_{tr}$**: source examples with $\ell_i < \lambda_a$ (low-loss, "easy" examples)
- **Meta-target domain $\mathcal{S}_{ts}$**: source examples with $\ell_i \geq \lambda_a$ (high-loss, "hard" examples)

The meta-source domain serves as the training set for the inner optimization loop (meta-train step), while the meta-target domain serves as a validation set for the outer optimization loop (meta-test step). This partition creates a virtual

domain shift within the source data that mimics the challenge of adapting to a new domain.

During training, a mini-batch contains both labeled source examples and unlabeled target examples. Source examples are first partitioned into $\mathcal{S}_{tr}$ and $\mathcal{S}_{ts}$ based on their losses relative to λ_a. The meta-train step optimizes the model parameters θ using weighted losses on $\mathcal{S}_{tr}$ and $\mathcal{D}_T$ (for domain adversarial training). The meta-test step evaluates the updated model on $\mathcal{S}_{ts}$ and uses this evaluation to optimize the neural-SPL parameters ϕ, including the weights and the *age* threshold λ_a itself. This bi-level optimization ensures that the learned hyperparameters lead to good generalization on the challenging high-loss region of the source domain, which we hypothesize corresponds to difficult regions of the target domain.

2.3.2 Neural Self-Paced Learning Module. The neural-SPL module is the core component that enables automatic hyperparameter learning in our framework. Unlike traditional SPL that requires manually designed *age* schedules and closed-form weighting functions, our neural-SPL module represents both the *age* threshold and the weighting mechanisms as learnable neural networks. This module produces three types of outputs: (i) instance-wise weights for the classification task, (ii) layer-wise weights for domain adversarial training, and (iii) the *age* threshold for data partitioning.

Instance-Wise Weighting. For each source example (x_i^s, y_i^s) in the meta-source domain $\mathcal{S}_{tr}$, we compute an instance-specific weight $v_i \in [0, 1]$ that controls its contribution to the classification loss. Rather than using the binary weights from traditional SPL (Equation 2.7), we employ a continuous weighting function

represented as a small FFN:

$$v_i = f_v\left(\max\left(0, \frac{-\ell_i}{\lambda_a} + 1\right); \theta_v\right) \tag{2.9}$$

where ℓ_i is the classification loss for example i, λ_a is the current *age* threshold, and f_v is a FFN with parameters θ_v.

The weighting function f_v is implemented as a small neural network with 2-3 hidden layers (hidden dimensions [100, 50] or [200, 100, 50]) and a sigmoid activation at the output layer to ensure $v_i \in [0, 1]$. Crucially, we use no bias term in the first layer, ensuring that when $-\ell_i/\lambda_a + 1 = 0$ (which occurs when $\ell_i = \lambda_a$), the output is exactly $v_i = 0$. This design choice preserves the intuition that examples with very low loss should receive full weight, while those approaching the *age* threshold receive progressively lower weights, and those exceeding the threshold (assigned to $\mathcal{S}_{ts}$) receive zero weight in the meta-train step.

Learnable *age* Parameter. The *age* parameter λ_a acts as the threshold that determines the boundary between the meta-source domain $\mathcal{S}_{tr}$ (examples with $\ell_i < \lambda_a$) and the meta-target domain $\mathcal{S}_{ts}$ (examples with $\ell_i \geq \lambda_a$). In traditional SPL, the *age* follows a manually designed schedule, such as $\lambda_a(t) = \lambda_0 \cdot (1 + t/T)$ for training step t out of T total steps. Such schedules require careful tuning of the initial value λ_0 and growth rate, and the optimal schedule is highly dependent on the specific task and dataset. In contrast, our approach treats λ_a as a learnable parameter that is optimized via meta-learning. Specifically, λ_a is updated based on the gradient of the meta-test loss with respect to λ_a:

$$\lambda_a \leftarrow \lambda_a - \beta \frac{\partial \mathcal{L}_{ts}}{\partial \lambda_a} \tag{2.10}$$

where β is the meta-test learning rate, and $\mathcal{L}_{ts}$ is the loss computed on the meta-target domain $\mathcal{S}_{ts}$ (detailed in Section 2.3.3). This gradient-based update

automatically adjusts the *age* threshold to maximize performance on the challenging high-loss region of the source domain.

As we will demonstrate in Section 2.4.6, the learned *age* schedules differ significantly across domains and do not follow the simple monotonically increasing pattern assumed by traditional SPL. For some domains, λ_a initially increases but then decreases in later training stages; for others, it plateaus at an intermediate value. These diverse patterns suggest that the optimal learning curriculum is indeed domain-specific and that our learned approach captures domain-dependent characteristics that would be difficult to encode in a manual schedule.

2.3.3 Meta-Learning Framework. Our meta-learning framework follows a bi-level optimization structure inspired by MAML (Finn et al., 2017) and MLDG (D. Li et al., 2018), but adapted to the UDA setting.

Meta-Train Step. In the meta-train step, we optimize the model parameters $\theta = (\theta_a, \theta_c, \theta_d)$ to minimize a weighted combination of the classification loss and domain adversarial loss. The classification loss is computed on the meta-source domain $\mathcal{S}_{tr}$ using the learned instance weights:

$$\mathcal{L}_{ce}(\mathcal{S}_{tr}; \theta, \phi) = \sum_{(x_i, y_i) \in \mathcal{S}_{tr}} v_i \cdot \ell(f_\theta(x_i), y_i) \tag{2.11}$$

where $v_i = f_v(\max(0, -\ell_i/\lambda_a + 1); \theta_v)$ is the instance weight from Equation 2.9, and ℓ is the cross-entropy loss.

The domain adversarial loss aligns representations across source and target domains at multiple layers. We employ a separate discriminator at each adapter layer, weighted by learned layer-wise coefficients:

$$\mathcal{L}_d(\mathcal{S}_{tr}, \mathcal{D}_T; \theta, \phi) = \sum_{\ell=1}^{L} w^{(\ell)} \mathcal{L}_d^{(\ell)}(\theta_a, \theta_d^{(\ell)}) \tag{2.12}$$

where $w^{(\ell)}$ is the weight for layer ℓ produced by the layer-wise balancing network (detailed in Section 2.3.4), and $\mathcal{L}_d^{(\ell)}$ is the binary cross-entropy loss for the ℓ-th domain discriminator:

$$\mathcal{L}_d^{(\ell)} = -\mathbb{E}_{x \in \mathcal{S}_{tr}}[\log p_d^{(\ell)}(x)] - \mathbb{E}_{x \in \mathcal{D}_T}[\log(1 - p_d^{(\ell)}(x))] \tag{2.13}$$

where $p_d^{(\ell)}(x)$ is the probability output by the ℓ-th domain discriminator.

The complete meta-train objective combines both terms:

$$\mathcal{L}_{tr}(\mathcal{S}_{tr}, \mathcal{D}_T; \theta, \phi) = \mathcal{L}_{ce}(\mathcal{S}_{tr}; \theta, \phi) + \mathcal{L}_d(\mathcal{S}_{tr}, \mathcal{D}_T; \theta, \phi) \tag{2.14}$$

Following the MAML framework, we perform a single gradient step on the meta-train objective to obtain updated parameters $\bar{\theta}$:

$$\bar{\theta} = \theta - \alpha \nabla_\theta \mathcal{L}_{tr}(\mathcal{S}_{tr}, \mathcal{D}_T; \theta, \phi) \tag{2.15}$$

where α is the meta-train learning rate (inner loop learning rate). We use $k = 1$ gradient step to balance computational efficiency with adaptation capability. While larger values of k could potentially improve adaptation, the substantial size of transformer-based models makes multiple inner-loop updates computationally expensive, and empirically we observe no significant performance loss with $k = 1$.

Meta-Test Step. After the meta-train step, we evaluate the updated model $\bar{\theta}$ on the meta-target domain $\mathcal{S}_{ts}$, which consists of high-loss source examples that were excluded from meta-training. The meta-test loss is the standard cross-entropy loss computed on these challenging examples:

$$\mathcal{L}_{ts}(\mathcal{S}_{ts}; \bar{\theta}) = \sum_{(x_i, y_i) \in \mathcal{S}_{ts}} \ell(f_{\bar{\theta}}(x_i), y_i) \tag{2.16}$$

This loss measures how well the model generalizes to the difficult region of the source distribution after being trained on the easy region. Crucially, no instance

weights are applied in the meta-test loss - we evaluate the raw performance to obtain an unbiased signal for hyperparameter optimization.

The complete meta-learning objective combines both the meta-train and meta-test losses to update both the model parameters θ and the neural-SPL parameters ϕ:

$$\theta \leftarrow \theta - \gamma \nabla_\theta \left[\mathcal{L}_{ts}(\bar{\theta}) + \beta \mathcal{L}_{tr}(\theta) \right] \tag{2.17}$$

$$\phi \leftarrow \phi - \gamma \nabla_\phi \mathcal{L}_{ts}(\bar{\theta}) \tag{2.18}$$

where γ is the meta-test learning rate (outer loop learning rate), and β is a balancing coefficient that ensures the model also optimizes the meta-train objective directly. The first update (Equation 2.17) modifies the model parameters to simultaneously improve both meta-train and meta-test performance. The second update (Equation 2.18) modifies the neural-SPL parameters (θ_v, θ_w, λ_a) to minimize the meta-test loss after the inner-loop adaptation.

2.3.4 Layer-Wise Balancing for Domain Adversarial Training.

As discussed in Section 2.2.2, different layers of pretrained transformer models encode different types of linguistic information (Rogers et al., 2020; Tenney, Das, & Pavlick, 2019). Lower layers capture surface-level features such as syntax, part-of-speech, and word order, while higher layers encode semantic and task-specific information. When performing domain adaptation, the degree of distribution shift may vary across these layers: some layers may require strong alignment to overcome domain-specific surface features, while others may already provide relatively domain-invariant representations and require less aggressive adversarial training.

To account for this variation, we learn a set of layer-wise weights $W = [w^{(1)}, w^{(2)}, \ldots, w^{(L)}]$ that control the importance of aligning each layer's representations. These weights are produced by a small FFN f_w that uses as input

the adapter representations aggregated across the current mini-batch:

$$Z_d = [\bar{z}^{(1)}, \bar{z}^{(2)}, \ldots, \bar{z}^{(L)}] \in \mathbb{R}^{L \times d_a} \tag{2.19}$$

where $\bar{z}^{(\ell)} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} z_x^{(\ell)}$ is the average adapter representation at layer ℓ across all examples in the current mini-batch $\mathcal{B}$ (including both source and target examples). The layer weights are computed as:

$$W = [w^{(1)}, \ldots, w^{(L)}] = \text{softmax}(f_w(Z_d; \theta_w)) \tag{2.20}$$

where f_w is a FFN with parameters θ_w, and the softmax ensures that the weights sum to 1: $\sum_{\ell=1}^{L} w^{(\ell)} = 1$. This normalization interprets the weights as relative importances: layers with higher weights contribute more to the overall domain adversarial objective (Equation 2.12).

2.3.5 Self-Training with Pseudo-Labeling. Beyond domain adversarial training, which aligns feature representations using unlabeled target data, we incorporate pseudo-labeling as a complementary mechanism to directly improve the classifier's predictions on the target domain. Pseudo-labeling (Lee, 2013) is a semi-supervised learning technique that treats the model's confident predictions on unlabeled data as pseudo-labels for additional supervised training. At each training step, we compute pseudo-labels for all target domain examples by taking the argmax of the model's predicted class probabilities: $\tilde{y}_i =_{k \in \mathcal{Y}} p_y(k|x_i^t; \theta)$, for each $x_i^t \in \mathcal{D}_T$. These pseudo-labeled target examples are then incorporated into the classification loss in the meta-train step. Specifically, we add pseudo-labeled target examples to the meta-source domain, creating an extended training set:

$$\mathcal{S}_{tr}^{+} = \mathcal{S}_{tr} \cup \{(x_i^t, \tilde{y}_i) \mid x_i^t \in \mathcal{D}_T, \ell(f_\theta(x_i^t), \tilde{y}_i) < \lambda_a\} \tag{2.21}$$

where we apply the same *age* threshold λ_a to filter pseudo-labeled examples: only those with pseudo-loss below λ_a are included in the meta-train set.

The key insight is that pseudo-labeling is most effective when the model's predictions are confident and likely to be correct. By using the same *age* threshold λ_a that partitions source examples, we automatically select high-confidence pseudo-labels while excluding uncertain predictions. This addresses the well-known confirmation bias problem in pseudo-labeling (Arazo, Ortego, Albert, O'Connor, & McGuinness, 2020), where the model may reinforce its own errors by training on incorrect pseudo-labels. Our neural-SPL module learns to set λ_a such that included pseudo-labels are sufficiently reliable to improve rather than degrade performance. Furthermore, the meta-learning framework adds an additional regularization mechanism for pseudo-labeling. The gradient updates from pseudo-labeled examples must be consistent with improvements on the meta-target domain $\mathcal{S}_{ts}$. If pseudo-labels are noisy and cause the model to learn misleading patterns, this will hurt meta-test performance, and the meta-learning optimization will adjust the instance weights and *age* threshold to suppress the harmful pseudo-labels. This is consistent with the *expansion assumption* proposed by Wei, Shen, Chen, and Ma (2021), which explains how self-training can improve pseudo-labels by using correctly labeled examples near incorrectly labeled ones.

2.4 Experiments

2.4.1 Datasets, Settings, and Baselines.

We evaluate the proposed model on the standard multi-domain sentiment analysis (SA) task. In addition, we also demonstrate the effectiveness of our framework when addressing the label-shift by applying MSP-DA to event detection (ED) task with significant more classes in UDA setting.

FDU-MTL. A dataset included reviews from 16 domains for binary sentiment classification task. In each adaptation setting, a single domain is assigned as

Table 1. Statistics of ACE-05's domains in UDA setting.

Domains	Train	Unlabeled	Test
bn+nw	38,644	N/A	9,661
bc	N/A	3,130	12,520
cts	N/A	2,885	10,972
wl	N/A	3,424	12,767

the target with unlabeled data while the other 15 are labeled source. Given the contextual sequence computed by models from a review, we use the first token [CLS] as the feature to predict its positive or negative sentiment.

ACE-05. A densely annotated corpus collected from 5 different domains (Walker et al., 2005). Two of which are used as source data, while each of the rest is a target domain for an adaptation setting. Given a trigger word in the context of an event mention, the model is required to perform a multi-class classification task that assigns a predicted label into one of the pre-defined 34 event types (including 1 negative type).

Data Settings. We provide statistics of each domain in UDA setting for ACE-05 and FDU-MTL in Table 1 and Table 2, respectively. For **ACE-05** dataset, we gather data from two closely related domains, bn and nw, to create a sizable source domain dataset, 80% of which are used for training whilst the rest are used as test target domain for in-domain setting. For out-of-domain settings, each of the other domains is considered the target domain of a single adaptation scenario, where 20% of its documents are unlabeled training target data and the remainders are utilized as the test dataset.

For FDU-MTL dataset, each of the 16 domains has a test set of 400 samples. The amount of training labeled and unlabeled data vary across domains,

Table 2. Statistics of the 16 domains in FDU-MTL

Domains	Train	Unlabeled	Test
Books	1400	2000	400
Electronics	1398	2000	400
DVD	1400	2000	400
Kitchen	1400	2000	400
Apparel	1400	2000	400
Camera	1397	2000	400
Health	1400	2000	400
Music	1400	2000	400
Toys	1400	2000	400
Video	1400	2000	400
Baby	1300	2000	400
Magazine	1370	2000	400
Software	1315	475	400
Sports	1400	2000	400
IMDb	1400	2000	400
MR	1400	2000	400

ranging from 1000 to 2000 samples. In each adaptation setting, a single domain is designated as the target domain while its unlabeled data are used in training set together with labeled data from the other 15 domains.

SA baselines We provide a comprehensive comparison of our proposed method with multiple baselines: **ASP-MTL** (P. Liu et al., 2017) and **DAEA** (Cai & Vasconcelos, 2019) are **LSTM**-based approaches. Transformer-based approaches include **BERT**, which is only fine-tuned on only labeled source domain, and **BERT+DANN** follows the standard adversarial training. Finally, **Bert-Masker** (Yuan, Zhao, Qin, & Liu, 2021) is the state-of-the-art approach that learns to explicitly mask domain-related words from text, resulting in domain-agnostic sentences.

ED baselines For ED task, we also compare MSP-DA to other functional weighting schemes that trying to balance the learning process to address the label shift. In particular, **Uniform** treats each sample's loss equally, **Focal Loss** downweights well-classified instance exponentially (T.-Y. Lin, Goyal, Girshick, He, & Dollár, 2017), and **Class-Balanced** uses a weighting factor that is inversely proportional to the number of samples (Cui, Jia, Lin, Song, & Belongie, 2019). Noted that these model employ both adapter-based fine-tuning and adversarial training procedure.

Implementation details All models are implemented in **PyTorch**. We leverage pre-trained **BERT-base** models and checkpoints from **Huggingface** repository (Wolf et al., 2020). We inject adapter layers after every feed-forward sub-blocks have bottleneck feed-forward architecture with down-sampled dimension chosen among [48, 96, 128]. All of the downstream heads are implemented as feed-forward networks with activation functions between layers. Each weighting net of neural-SPL module is a feed-forward network with 2 or 3 layers with hidden vectors of size [100, 50] or [200, 100, 50], respectively. To train the proposed model, we use Adam optimizer with meta-train and meta-test learning rates α and γ both chosen from [5e-5, 1e-4, 5e-4, 1e-3, 5e-3], the mini-batch size from [50, 100, 150] of which 20% or 40% are unlabeled target data, and the meta-test balancing term β from [5, 2, 1, 0.5, 0.1].

We tune the hyperparameters for the proposed model using a random search. All hyperparameters are selected based on the F1 scores on the development set of a single domain. The same hyperparameters from this fine-tuning are then applied for other domains to demonstrate the domain-specificity problem. In the best model, fixed pre-train BERT-base layers augmented by

Table 3. UDA performances for SA task on FDU-MTL test datasets. aAcc is the average accuracy score across all domains.

System	MR	Appr.	Baby	Books	Cam.	DVD	Elec.	Hlth.	IMDB	Kitc.	Magz.	Musics	Softw.	Sport	Toys	Video	aAcc
ASP-MTL	76.7	87.0	88.2	84.0	89.2	85.5	86.8	88.2	85.5	86.2	92.2	82.5	87.2	85.7	88.0	84.5	86.1
DAEA	77.0	89.0	92.3	89.0	92.0	88.3	91.8	89.8	90.8	90.3	96.5	88.0	92.8	90.8	91.8	92.3	90.2
BERT	90.5	90.8	90.3	91.3	91.5	89.0	91.3	91.3	91.3	90.0	88.5	90.3	90.5	92.0	90.8	92.0	90.7
BERT+DANN	90.5	91.8	92.5	90.8	90.0	91.3	90.5	90.8	91.0	91.8	91.0	90.5	91.0	90.5	90.3	90.3	90.9
Bert-Masker	83.8	92.3	92.8	93.0	92.8	89.3	93.3	95.3	86.0	90.8	94.5	89.5	93.0	92.5	93.8	91.3	91.5
MSP-DA	**93.3**	**93.1**	92.5	**93.2**	**93.3**	**92.4**	**93.1**	**93.2**	**93.4**	**93.0**	**93.1**	**92.7**	**93.1**	**93.3**	**93.5**	**92.8**	**93.0**

adapters with bottleneck size 96 are used as our feature encoder. All objective heads have 2 hidden layers. We use Adam optimizer with a learning rate of 1e-4 for both meta-train and meta-test step, 100 for mini-batch size with 20% target data, and the meta-test balancing term is 2. Our reported results are averages of five runs using the best hyperparameter configuration with different random seeds.

2.4.2 Main Results.

Sentiment Analysis SA results are presented in Table 3. While simple model using contextual embedding BERT outperforms all previous LSTM-based methods, we observe little to no improvement applying domain adversarial training naively with it. In particular, BERT+DANN actually has negative effect on about half of the domains, indicating that the standard baseline approach being unable to adjust to each specific adaptation setting. In contrast, our framework achieves the best performance for 11 review domains overall, surpassing the current state-of-the-art method Bert-Masker by 1.5 points on average. This demonstrates both the effectiveness and the robustness of MSP-DA to each domain.

Event Detection UDA performances for ED task are presented in Table 4. Again, we observe that BERT+DANN only provides slight improvement for domain `bc` compare to BERT, while significantly degrades model's performances on the other two resulting in almost 3 points drop in average out-of-domain

Table 4. UDA performances for ED task on ACE-05 test datasets. aF1 is the average out-of-domain F1 score.

	In-domain (bn+nw)			**Out-of-domain** (bc)			**Out-of-domain** (cts)			**Out-of-domain** (wl)			
System	P	R	F	P	R	F	P	R	F	P	R	F	**aF1**
BERT	75.8	72.5	74.1	73.5	68.9	71.1	73.7	69.5	71.5	62.2	51.6	56.4	66.3
BERT+DANN	73.4	76.0	74.7	73.9	69.4	71.5	76.4	53.0	62.5	59.9	53.2	56.3	63.4
Uniform	76.8	79.4	78.1	75.4	66.3	70.5	80.4	21.0	33.3	61.8	45.7	52.6	52.1
Focal	78.2	77.6	77.9	71.7	72.9	72.2	72.9	68.5	70.1	64.8	54.2	59.0	67.1
Class-Balanced	79.3	78.3	78.7	77.8	68.0	72.5	78.0	44.0	56.2	59.0	50.3	54.3	61.0
MSP-DA	75.4	80.0	**77.7**	76.2	75.5	**75.8**	75.3	76.8	**76.1**	70.8	59.9	**64.8**	**72.2**

F1 score. Similarly, applying DANN for the adapter-based model without any weighting mechanism, as in Uniform, also has adverse effects on out-of-domain performances. Class-Balanced's in-domain results are slightly higher than other models due to its ability to balance the training process, which addresses the extreme negative-skewed label distribution of the given source data. In contrast, its domain adaptation ability is actually the lowest because of the change in data distribution across domains. Focal Loss performs generally better in out-of-domain settings as they generate weighting coefficients adaptively based on the current losses, without involving any domain-specific statistics. Finally, MSP-DA provides consistent improvements when adapting to any new domain, even achieving on average 5 points higher in F1 score compared to the best baseline method.

2.4.3 Ablation Study. In the first row-block of Table 5, we conduct an ablation study to validate the effectiveness of each of our main components by investigating the performance of the following variations of our model: **MSP-DA–mSPL** follows the normal SPL process to produce the weighting coefficients and train-test datasets for ML; **MSP-DA–DANN** trains only on source domain without utilizing unlabeled target data for domain adversarial objective; and **MSP-DA–PL** in which no pseudo-labels are leveraged for training.

Table 5. Performances for Ablation Study

System	In-domain (bn+nw)			Out-of-domain (bc)			Out-of-domain (wl)		
	P	R	F	P	R	F	P	R	F
MSP-DA – mSPL	74.5	79.7	77.0	77.5	72.0	74.6	64.1	51.9	57.4
MSP-DA – DANN	74.3	80.3	77.2	75.7	72.9	74.2	61.6	51.9	56.3
MSP-DA – PL	77.8	75.1	76.4	75.1	73.5	74.3	62.6	52.4	57.0
MSP-DA (Random)	73.0	76.4	74.7	75.6	73.3	74.4	61.0	50.3	55.0
MSP-DA (Reverse)	77.7	75.0	76.3	78.2	70.6	74.2	65.0	50.7	57.0
MSP-DA (Ours)	75.4	80.0	77.7	76.2	75.5	75.8	70.8	59.9	64.8

In general, our full model outperforms all variants across domains, even in the in-domain setting, which confirms the superiority and flexibility provided by the jointly optimized pacing and weights from our neural-SPL module. Especially for `wl` domain, domain adversarial training in MSP-DA manages to improve more than 8 F1 points.

Meta-test Selection To examine the correctness of our assumption, we augment the data selection process for meta domains in Random and Reverse variants. The former randomly selects training samples for each meta domain, whereas the latter implements the opposite hypothesis by choosing hard and easy instances for meta-train and meta-test sets, respectively. Both variants result in a considerable decline in domain adaptation results as shown in Table 5. Notably, the significant performance drop in the in-domain setting of Random indicates that simply constructing train-test sets without any appropriate condition can do more harm than good for the ML process. These empirical observations further confirm our initial assumption on how domain shift correlates well with the easy meta-train and hard meta-test sets.

2.4.4 The Values of Age Hyperparameter. Age hyperparameter λ_a is usually the hardest to tune in a SPL system due to the fact that aside from the initial value, determining how λ_a changes throughout the training process

Table 6. Performances for Age Hyperparameter Analysis

System	Out-of-domain (bc) P	R	F	Out-of-domain (wl) P	R	F
Fixed (25)	79.3	68.9	73.7	65.8	50.0	56.8
Fixed (50)	75.0	73.7	74.3	66.3	49.5	56.6
Fixed (75)	76.4	72.0	74.1	65.9	52.7	58.6
Linear Incrs	74.9	71.7	73.3	61.6	54.7	57.9
Meta (Ours)	76.2	75.5	75.8	70.8	59.9	64.8

also has a major impact on the final performance. Several prior works (H. Li & Gong, 2017; Ren, Zhao, Sheng, Yao, & Xu, 2017) have proposed alternative age schedulers in place of the naive strategy which adds/multiples λ_a with a constant at each epoch. However, the value of λ_a in these methods still follows a predefined sequence, implying the need for a meticulous tuning process. In contrast, our neural-SPL module updates λ_a based on optimization signals from meta-test set, thus always able to create an appropriate dynamic curriculum regardless of different learning tasks and datasets.

In Table 6, we examine how different values and schedules of age hyperparameter affect performances on `bc` and `wl` domains. The Fixed (p) settings with $p \in [25, 50, 75]$ are variations of our model with λ_a values always corresponding to the unchanged p-th percentile of the current mini-batch's sample losses; or in other words, the number of samples in meta-train set is always a constant p percent that of the current mini-batch. Additionally, we evaluate the case in which p is linearly increased as training proceeds, similar to the standard SPL process, in Linear Incrs setting.

The results show that the lower p is, the worse model performs, indicating that with too few meta-train data, the model will not be able to adapt to the hard meta-test domain. Surprisingly, the gradual rising scheduler of Linear Incrs

is not as effective as the other Fixed variants. This means that the easy-to-hard assumption of prior SPL systems is not suitable for our ML framework.

λ_a Visualization To gain more insight into how age hyperparameter changes throughout the training process of each domain, we plot the values of λ_a in source-losses percentile against the number of update steps for 10 epochs in the right subplot of Fig. 3. While λ_a quickly follows the standard incremental trend initially, it starts to plateau within the 60-70 percentile range until eventually starting to decrease. Notably, behavior of λ_a diverges across domains in subsequent steps. Whereas λ_a continues the to decline in `bc` and `cts` domains, it experiences a complete trend reversal at the end of the training of `wl` domain. We hypothesis that this drastic change of λ_a is because of the gradients' dot product term that the objective in Eq. 2.18 implies, which we will delve deeper into in the discussion section below. The $\cap$ shape of λ_a correlates with the term's value as the model maximizes it to align the gradient directions between the meta train-test domains, going from negative initially as the training started, to 0 which causing the plateau, then gradually becoming positive as the model was able to adjust the updates of meta-train set to be consistent with that of meta-test set. However, for hard adaptation such as `wl` domain, too few data in meta-train set can cause a major disparity between the two meta domains again, thus the resulting trend reversal at the last few steps.

We also visualize the same plot for target-pseudo-losses percentile, which leads to an interesting observation: Initially, the model followed its own pseudo labels without any constraint and the high value of λ_a percentile represents model's incorrect overconfidence. However, these pseudo-label updates will cause discrepancies with meta-test domain, thus the ML framework will gradually fix the

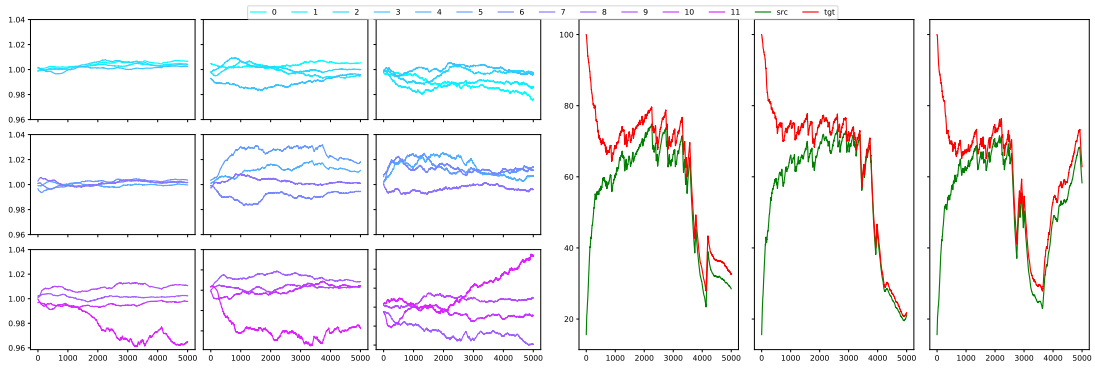


Figure 3. Three columns in each subplot correspond to domain `bc`, `cts`, `wl`, respectively. (Left) Layer-wise DANN weights at each training step. (Right) source and target age percentiles at each training step.

corresponding predictions, allowing only quality pseudo samples to be included in meta-train set. Eventually, the target trend converges with the source ones, suggesting that model's predictions on pseudo labels are then as consistent as on clean training labels.

2.4.5 Balancing Domain Adversarial Losses. Previous works have observed that the weight of DANN in the combined objective has a significant impact on the overall adaptation performance of the model. We further validate this point by investigating how different domain adversarial weighting schemes affect the results on `bc` and `wl` domains. Specifically, we evaluate 3 types of layer-wise weighting: (i) **Constant** - all layers share the same w^l value, (ii) **Anneal Up** - w^l slowly increases from lower to higher layers, and (iii) **Anneal Down** - w^l is highest for the first layer and gradually declines for subsequent layers.

The results are present in Table 7, in which none of the schemes is better than the others in both domains. In contrast, the meta-learned coefficients of our framework manage to boost model's performances in every adaptation setting, especially for the hard wl domain where domain adversarial training matters the most.

Table 7. Performances for DANN Weighting Analysis

System	Out-of-domain (bc) P	R	F	Out-of-domain (wl) P	R	F
Constant	75.8	71.5	73.6	63.2	52.6	57.4
Anneal Up	75.4	71.0	73.1	63.5	52.6	57.4
Anneal Down	74.0	74.8	74.4	62.3	51.1	56.1
Meta (Ours)	76.2	75.5	75.8	70.8	59.9	64.8

We further visualize how each layer's weight changes during the learning process across domains in the left subplot of Fig. 3. In particular, we partition 12 layers of BERT-base model into 3 groups of 4 sequential layers, each of which is known to contain a different type of information that is important for a different type of task as described in the previous section. We can observe from the graphs a certain pattern: the higher level the group is, the more volatile its layers' coefficients are. However, there is no specific rule shared among all domains regarding the value of each layer's weight. This affirms the sensitivity of domain adversarial balancing term to each individual domain and further justifies the effectiveness of the jointly optimized weighting in our framework.

2.4.6 Discussion. Following the analysis of MLDG framework presented in D. Li et al. (2018), we decompose the meta-test loss, given that $\bar{\theta} = \theta - \alpha L'_{tr}(\theta)$, using the first order Taylor expansion:

$$L_{ts}\left(\theta - \alpha L'_{tr}(\theta)\right) = L_{ts}(\theta) + \frac{\partial L_{ts}(\theta)}{\partial \theta} \cdot \left(-\alpha \frac{\partial L_{tr}(\theta)}{\partial \theta}\right) \tag{2.22}$$

Denoting $G = \frac{\partial L_{ts}(\theta)}{\partial \theta} \cdot \frac{\partial L_{tr}(\theta)}{\partial \theta}$ and plugging Eq. 2.22 into the final objective to update main model's parameters from Eq. 2.17 results in the following optimization problem:

$$\arg\min_{\theta} L_{tr}(\theta) + \beta L_{ts}(\theta) - \alpha\beta G \tag{2.23}$$

The third term in Eq. 2.23 is a gradient-based regularization that penalizes inconsistency between parameter updates of meta-train and meta-test domains. By enforcing loss gradients of the two domains to follow a similar direction, Eq. 2.23 prevents the model from over-fitting to a single domain, effectively improves model's adaptation capacity provided that meta-test set is 'close' to target domain.

We further examine how the ML framework affects the values of neural-SPL module's parameters $(\theta_w, \theta_v, \lambda_a)$ in our model. Plugging Eq. 2.22 into the gradient of λ_a, we have:

$$\frac{\partial L_{ts}(\bar{\theta})}{\partial \lambda_a} = -\alpha \frac{\partial L_{ts}(\theta)}{\partial \theta} \cdot \frac{\partial^2 L_{tr}(\theta)}{\partial \theta \partial \lambda_a} = -\alpha G \cdot \frac{\partial f_v(\lambda_a)}{\partial \lambda_a} \tag{2.24}$$

From Eq. 2.24, we see that the multiplicative factor G also controls how the value of λ_a changes throughout the ML process. When there is a significant discrepancy between meta-train and meta-test domain, G would have a negative value, which would in effect push λ_a higher and allow more samples into meta-train set for easier adaptation to meta-test set. Conversely, a positive G would imply that the model is good enough to align the current meta domains, thus gradually pulling λ_a down to make the task harder. This behavior is clearly illustrated in Fig. 3. Similar arguments can be made for the meta-learned weighting coefficients, where G would encourage samples whose gradients are similar across domains while decreasing the contribution of those whose gradients are not. These understanding are also presented in J. Shu et al. (2019) and closely related to how MAML works (Nichol, Achiam, & Schulman, 2018; Raghu, Raghu, Bengio, & Vinyals, 2019).

2.5 Related Work

Unsupervised Domain Adaptation. The main line of research on UDA focuses on learning domain-invariant, which is either achieved by explicitly reducing the distance between source and target feature space measured by some

distribution discrepancy metric (Long et al., 2015; Zellinger, Grubinger, Lughofer, Natschläger, & Saminger-Platz, 2017), or by adversarial training in which the feature extractor is trained to fool a domain classifier, both are jointly optimized to arrive at an aligned feature space (Ganin et al., 2016). We focus on applying the latter in transformer-based model (BERT) (Devlin et al., 2019) for textual tasks. Previous works have provided empirical results on different domains (C. Lin et al., 2020; Wright & Augenstein, 2020), different tasks (Du, Sun, Wang, Qi, & Liao, 2020; Naik & Rosé, 2020a), most of which presented little to no improvement following the standard domain adversarial training framework. We further verify this point in our baseline performances.

Sample Weighting. There are two main research directions to adaptively output weight of a sample during training process: addressing class imbalance by monotonically increasing function that imposes larger weights to ones with larger loss values (T.-Y. Lin et al., 2017; Sun, Kamel, Wong, & Wang, 2007), and suppressing the effect of noisy labels using monotonically decreasing function which focus on low-loss easy samples (L. Jiang, Meng, Mitamura, & Hauptmann, 2014; M. P. Kumar et al., 2010). Although straightforward to apply, the above methods are limited in that they all need a pre-specified closed-form weighting function, while their respective hyperparameters are sensitive to the change of training data such that careful tuning is required.

Meta-Learning. There are three main categories of modern ML algorithms: learning a metric space to measure distance or similarity among data (Sung et al., 2018; Vinyals, Blundell, Lillicrap, kavukcuoglu, & Wierstra, 2016), learning an optimizer which updates all of model's parameters in a latent parameter

space (Andrychowicz et al., 2016; J. Chen, Qiu, Liu, & Huang, 2018), and learning an initialization that is good for all tasks and able to fast adapt to unseen tasks (Finn et al., 2017; Jamal & Qi, 2019). Our approach falls into the last category, where the learning process follows MAML, more specifically its variant for DG problem in D. Li et al. (2018).

2.6 Summary

This chapter addressed Research Question 1 from Chapter I, focusing on the challenge of unsupervised domain adaptation for text classification when significant distribution shifts exist between source and target domains. We present a novel ML framework for UDA setting that achieves state-of-the-art performance on ED task. In particular, a neural-SPL module is employed to adaptively partition source domain into meta-train and meta-test set, while simultaneously learns the instance-wise and layer-wise weights for the loss terms of downstream task and domain adversarial task respectively. The proposed model significantly improves domain adaptation performances against various baselines on every domain without domain-specific hyperparameter tuning.

In the future, we intend to apply our approach to the several direction: (1) We will extend our work to multilingual problems (Pouran Ben Veyseh, Nguyen, Dernoncourt, & Nguyen, 2022), or other domains and tasks (Q. Lu, Dou, & Nguyen, 2021); (2) We will incorporate different novel domain adaptation regularization methods (Phung, Minh Tran, Nguyen, & Nguyen, 2021); (3) We will adapt our framework to more general multi-source domain adaptation setting.

CHAPTER III

UNSUPERVISED DOMAIN ADAPTATION FOR JOINT INFORMATION EXTRACTION

This chapter includes materials from a published paper: "Nghia Trung Ngo, Bonan Min, and Thien Huu Nguyen. 'Unsupervised Domain Adaptation for Joint Information Extraction.' In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5894–5905." As the first author of this paper, Nghia was responsible for problem formulation, method development, experimental design, implementation, and manuscript preparation. Bonan Min provided collaboration and meaningful discussion on joint IE and domain adaptation research. Thien Huu Nguyen contributed through research direction, guidance, feedback, and editorial revision. The paper content has been revised to comply with the dissertation format and purposes.

As established in Chapter I, transfer learning has emerged as a fundamental paradigm in natural language processing for addressing data scarcity and enabling systems to generalize across diverse contexts. Within the adaptive transfer learning phase, domain adaptation represents a critical challenge: how can models trained on data from one domain (the source) effectively transfer to a different domain (the target) where labeled data is unavailable or prohibitively expensive to obtain? This chapter addresses Research Question 2 (RQ2) from Chapter I: *How can we perform effective unsupervised domain adaptation for complex structured prediction tasks that involve multiple interdependent subtasks, specifically joint information extraction?* We introduce DA4JIE, a novel framework for unsupervised domain adaptation in joint information extraction. The paper makes three main contributions: (1) the first systematic study of UDA for joint IE involving all four

core tasks (entity extraction, event detection, relation extraction, and argument extraction), (2) Instance-relational Domain Adaptation (IrDA) that uses graph-relational adversarial training to flexibly align task-specific representations across domains, and (3) Context-invariant Structure Learning (CiSL) that extracts domain-invariant structural information by fusing syntactic dependency graphs with transformer attention patterns. Experiments on ACE-05 across four diverse domain adaptation scenarios demonstrate that DA4JIE achieves state-of-the-art performance, improving out-of-domain F1 by 1.3 absolute points over strong baselines while maintaining competitive in-domain performance.

3.1 Introduction

Joint Information Extraction (JIE) aims to jointly solve multiple tasks in the Information Extraction pipeline: entity mention extraction (EME), event trigger detection (ETD), relation extraction (RE), and event argument extraction (EAE). Due to their ability to leverage task dependencies and avoid error propagation inherent in pipeline approaches, JIE models have presented state-of-the-art performance for different IE tasks (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a). The advance of large-scale pretrained language models has made it possible to replace classical pipeline approaches (Y. Chen, Xu, Liu, Zeng, & Zhao, 2015; Q. Li, Ji, & Huang, 2013), which suffer from error propagation, with a single transformer-based model performing all four tasks jointly. These modern JIE systems achieve impressive performance when training and test data originate from the same domain and follow the same distribution.

However, a critical issue with current JIE methods is that they only focus on the standard supervised learning setting where training and test data come from the same domain. Cross-domain learning and domain adaptation with

training and test data in different domains have not been explored for JIE, thus hindering the application of this technology to different domains in practice. In real-world deployments, this limitation becomes particularly problematic: a JIE system trained on formal news articles may fail catastrophically when applied to informal social media text, medical records, or legal documents due to substantial differences in vocabulary, writing style, entity types, and event expressions across these domains.

To illustrate the challenges of domain shift in JIE, consider a DIE event where a PERSON entity mention serves as a VICTIM event argument. Documents recording this event type in medical records may express these instances in a significantly distinct manner compared to when news anchors report similar tragic incidents. For instance, medical records might use clinical terminology such as "patient expired due to cardiac arrest" while news articles might state "victim died in the attack." The vocabulary differs ("expired" vs. "died"), the entity types vary ("patient" vs. "victim"), and the contextual patterns diverge (medical causes vs. violent events). A JIE model trained exclusively on news text would struggle to correctly identify entities, events, relations, and arguments in medical text due to these distributional differences. The problem is further exacerbated when models aim to jointly learn multiple tasks, facing various kinds of domain shifts simultaneously across entities, events, relations, and arguments.

To address domain differences for IE, a major research direction involves unsupervised domain adaptation (UDA), where models leverage additional unlabeled data from the target domain together with labeled training data from the source domain to improve performance on the target domain. The majority of existing UDA methods have focused on transfer learning between source and

target domains for a *single* IE task (Ganin et al., 2016; A. Kumar et al., 2018; Long et al., 2015). These approaches typically employ domain-adversarial training to learn domain-invariant representations, where a feature extractor is trained to fool a domain classifier while simultaneously optimizing task-specific objectives. While recent work has also aimed to generalize previous approaches to multi-domain settings (Dai, Liu, Ren, & Xu, 2020; Wright & Augenstein, 2020), the scenario where the considered task involves multiple interdependent objectives (as in JIE) with different input distributions remains unexplored.

The challenge of applying standard UDA techniques to JIE is multifaceted. First, classical UDA approaches for IE often rely on simplification assumptions about the factorization of the joint input-output distribution of an IE task to categorize and solve a specific domain shift problem. An example includes the covariate shift assumption, where the discrepancy is assumed to be only in the marginal input distribution whereas the predictive dependency remains unchanged (Kull & Flach, 2014). However, in JIE, this assumption does not hold: the relationships between entities and events, the argument structures of events, and the relation types between entities can all vary substantially across domains. Second, existing domain adaptation methods treat all task instances uniformly, failing to account for type-specific characteristics. Entity mentions in different domains may share structural patterns distinct from those of event triggers, and relation instances may require domain-invariant features different from those needed for event arguments. This uniform treatment leads to suboptimal adaptation when multiple heterogeneous tasks must be transferred simultaneously. Furthermore, previous JIE systems have leveraged specialized linguistic structures extracted from input sentences in a heuristic and direct manner, such as using

heuristic-based dependency graphs between instances in different tasks (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a; Veyseh et al., 2020). However, this approach introduces additional challenges for domain adaptation as it further incorporates domain-specific context-dependency information into the learned representations. For example, syntactic dependency patterns that are informative in one domain (e.g., formal news text with standard grammatical structures) may not transfer well to another domain (e.g., informal social media text with non-standard grammar and unconventional syntax).

To address these critical gaps, this chapter introduces Domain Adaptation for Joint Information Extraction (DA4JIE), the first method specifically designed for UDA in the JIE setting. DA4JIE addresses the fundamental challenge of transferring knowledge across domains for multiple interdependent structured prediction tasks without access to labeled target domain data. Our approach introduces two key technical innovations that work synergistically to achieve effective domain adaptation for JIE.

At the core of DA4JIE is an Instance-relational Domain Adaptation (IrDA) module that seeks to simultaneously align instance representations for all downstream tasks in JIE across source and target domains. Inspired by Graph-relational Domain Adaptation (Z. Xu et al., 2022), we view event trigger and entity mention instances from each domain as nodes on a domain-instance relational graph, where the adjacency matrix controls the relationships between domain-specific representations. An edge connecting two instances implies that their representations should be aligned, which is equivalent to their pairwise relationship containing no information to identify their domains. This is achieved through an adversarial learning process on pairwise node relationships. Specifically, a

graph discriminator attempts to recover the domain-instance graph via the adjacency structure, while the text encoder for JIE prevents the discriminator from succeeding. Unlike standard domain-adversarial training that enforces strict uniform alignment (treating all instances identically), IrDA assumes a chain connection across instance types that reflects the true relationships among tasks in JIE, allowing flexible and effective adaptation.

In addition to IrDA, we incorporate a novel Context-invariant Structure Learning (CiSL) module into the instance encoding process. Previous JIE systems have leveraged specialized linguistic structures such as dependency graphs between instances, but these heuristic-based structures can introduce domain-specific context-dependent information that harms adaptation. CiSL addresses this by using Graph Transformer Networks (GTN) (Yun et al., 2019) to fuse different types of context-independent graphs into a single context-invariant graph (CiG) for each input sentence. Specifically, we extract attention probability matrices at each transformer layer to create attention graphs, using position embeddings as domain-agnostic node features. These graphs augment syntactic dependency structures, and GTN fuses them across layers to create the CiG whose node features are combined with contextual representations. This design encourages the model to utilize domain-invariant structural information for downstream tasks while providing richer instance representations through graph-conditioned pooling operations.

The main contributions of this chapter are:

– To our knowledge, this is the first work to introduce and comprehensively evaluate unsupervised domain adaptation for joint information extraction involving all four core IE tasks (entity mention extraction, event trigger detection, relation extraction, and event argument extraction).

- We propose Instance-relational Domain Adaptation, a novel domain adaptation mechanism that generalizes standard domain-adversarial training to simultaneously align type-specific representations of all downstream tasks between domains through graph-relational adversarial learning. The proposed chain connection pattern reflects the true task relationships in JIE and enables flexible, effective adaptation.

- We introduce Context-invariant Structure Learning, a mechanism to learn domain-invariant structural representations by fusing syntactic dependency graphs with attention graphs across transformer layers using Graph Transformer Networks, while employing graph-conditioned pooling to enhance instance representations.

- We provide extensive evaluation on multiple domain adaptation scenarios using the ACE-05, demonstrating that DA4JIE significantly improves out-of-domain performance for all four IE tasks compared to state-of-the-art JIE systems and existing domain adaptation baselines.

- Through detailed ablation studies, graph structure analysis, and component-wise evaluation, we provide insights into why and how the proposed mechanisms enable effective domain adaptation for complex structured prediction tasks.

The remainder of this chapter is organized as follows. Section 3.2 provides background on joint information extraction tasks and formally defines the unsupervised domain adaptation problem for JIE. Section 3.3 presents the proposed DA4JIE framework in detail, including the base JIE architecture, the Instance-relational Domain Adaptation module, and the Context-invariant Structure

Learning mechanism. Section 3.4 describes our experimental setup, presents main results, ablation studies, and comprehensive analysis. Section 3.6 reviews related work on joint information extraction and unsupervised domain adaptation. Finally, Section 3.7 summarizes the chapter and connects its contributions to the broader dissertation narrative.

3.2 Background and Problem Statement

3.2.1 Joint Information Extraction Tasks.

The joint information extraction (JIE) problem comprises four fundamental tasks that aim to extract structured information from unstructured text. Given an input sentence, a unified model is trained to optimize a linear combination of objectives for each task. The four tasks are interdependent, and jointly modeling them enables the system to leverage mutual information and avoid error propagation inherent in pipeline approaches. We now formally define each task.

Entity Mention Extraction (EME). aims to detect and classify entity mentions (names, nominals, and pronouns) according to a set of predefined semantic entity types. Given an input sentence $w = [w_1, w_2, \ldots, w_n]$ with n words, the goal is to identify entity mention spans and assign each span an entity type $t^e \in \mathcal{T}^E$, where $\mathcal{T}^E$ is the set of entity types (e.g., PERSON, LOCATION, ORGANIZATION, WEAPON). Each entity mention e_i is characterized by its span boundaries $(start_i, end_i)$ and its type t_i^e. For example, in the sentence "*Soldiers arrived at the checkpoint*," the entity mentions are "Soldiers" (PERSON) and "checkpoint" (FACILITY).

Event Trigger Detection (ETD). seeks to identify and classify event triggers - words or phrases that clearly evoke an event occurrence - according to

a given set of event types. An event trigger is typically a verb or nominalization that most clearly expresses the event. Formally, given the input sentence w, the task identifies trigger spans (which can involve multiple words) and assigns each an event type $t^g \in \mathcal{T}^G$, where $\mathcal{T}^G$ is the set of event types (e.g., TRANSPORT, ATTACK, DIE). Each event trigger g_m is represented by its span and type t^g_m. In the example sentence above, "arrived" would be identified as a TRANSPORT event trigger.

Relation Extraction (RE). aims to predict the semantic relationship between two entity mentions in the sentence. Given a pair of entity mentions (e_i, e_k), the task determines whether a relation exists between them and, if so, classifies the relation as one of the predefined relation types $t^r \in \mathcal{T}^R$, where $\mathcal{T}^R$ is the set of relation types (e.g., PHYSICAL, PART-WHOLE, PERSONAL-SOCIAL). Each relation r_j is thus a triple (e_i, e_k, t^r_j) connecting two entity mentions with a specific relation type. The set of relation types includes a special NONE class to indicate that no relation exists between the entity pair. For example, if the sentence were "*A man driving a taxicab*," a PHYSICAL relation would connect the entity mentions "man" and "taxicab."

Event Argument Extraction (EAE). determines the roles that entity mentions play in events. Given an event trigger g_m and an entity mention e_i, the task predicts whether the entity serves as an argument for the event and, if so, classifies the argument role as one of the predefined role types $t^a \in \mathcal{T}^A$, where $\mathcal{T}^A$ is the set of argument roles (e.g., AGENT, VICTIM, DESTINATION, INSTRUMENT). Each event argument a_p is thus a triple (g_m, e_i, t^a_p) connecting a trigger to an entity mention with a specific role. The argument role set also includes a NONE class to

indicate that the entity mention does not participate in the event. In our example sentence "*Soldiers arrived at the checkpoint*," the entity "Soldiers" would serve as the AGENT argument and "checkpoint" as the DESTINATION argument for the TRANSPORT event triggered by "arrived."

These four tasks exhibit complex interdependencies that joint modeling can exploit. Relations require correctly identified entity mentions as their arguments. Event arguments depend on both entity mentions and event triggers being accurately detected. Furthermore, information at each level can inform predictions at other levels: knowing that an entity serves as a VICTIM in a DIE event can help identify it as a PERSON entity, and recognizing a PHYSICAL relation between entities can provide evidence for TRANSPORT or CONTACT events. Traditional pipeline approaches that process these tasks sequentially suffer from error propagation, where mistakes in upstream tasks cascade to downstream tasks. Joint models that explicitly capture these dependencies can leverage mutual information to improve overall extraction performance.

Note that the sets of relation types $\mathcal{T}^R$ and argument roles $\mathcal{T}^A$ are predetermined and include a special class of NONE to indicate the negative category when no relation or argument role exists for a given pair.

3.2.2 Unsupervised Domain Adaptation Setting. In the unsupervised domain adaptation (UDA) setting, data originates from two different domains with different distributions. For training, we have access to a labeled source dataset $\mathcal{D}_S = \{(w_i^s, y_i^s)\}_{i=1}^{N_s}$ consisting of N_s samples, where each sample includes an input sentence w_i^s and its corresponding labels y_i^s for all four JIE tasks. We also have access to an unlabeled target dataset $\mathcal{D}_T = \{w_j^t\}_{j=1}^{N_t}$ of N_t samples drawn from the target domain, where only the raw sentences are available without

any annotations. The key characteristic of UDA is the distribution shift: the joint distribution of inputs and outputs differs between source and target domains, i.e., $P_S(W, Y) \neq P_T(W, Y)$.

The distribution shift manifests in multiple ways for JIE. First, the marginal input distribution differs ($P_S(W) \neq P_T(W)$) due to vocabulary differences, syntactic patterns, and stylistic variations across domains. For instance, news articles use formal language and standard grammatical structures, while social media posts employ informal language, abbreviations, and non-standard syntax. Second, the conditional output distributions given inputs can also vary ($P_S(Y|W) \neq P_T(Y|W)$), as the same linguistic patterns may signal different entity types, event types, or argument roles depending on domain conventions. Third, the label distributions themselves may shift ($P_S(Y) \neq P_T(Y)$) if certain entity types or event types are more prevalent in one domain than another. These multiple sources of distributional shift make domain adaptation for JIE particularly challenging.

The goal of unsupervised domain adaptation is to leverage both datasets - the labeled source data $\mathcal{D}_S$ and the unlabeled target data $\mathcal{D}_T$ - during training to optimize model performance on the target domain test data. At each training iteration, a mini-batch consists of samples drawn from both $\mathcal{D}_S$ and $\mathcal{D}_T$. The labeled source samples are used to learn the main downstream JIE tasks using their true labels, while the unlabeled target samples are employed to impose domain-invariant constraints on the extracted features. This enables the model to learn representations that are both discriminative for the JIE tasks (from the labeled source data) and invariant across domains (from the distribution matching between source and target unlabeled data).

Critically, in the unsupervised domain adaptation setting, the model has no access to labeled data from the target domain during training. Evaluation is performed on a held-out test set from the target domain with gold annotations. This setting reflects realistic scenarios where obtaining annotations in specialized domains (e.g., medical, legal, scientific) is expensive and time-consuming, but unlabeled text data is readily available.

3.2.3 Problem Formulation. Having defined the individual JIE tasks and the UDA setting, we now formulate the specific problem addressed in this chapter: unsupervised domain adaptation for joint information extraction involving all four tasks simultaneously.

Problem Statement. Given a labeled source domain dataset $\mathcal{D}_S$ with annotations for entity mentions, event triggers, relations, and event arguments, and an unlabeled target domain dataset $\mathcal{D}_T$ with only raw text, the goal is to learn a joint information extraction model $f : \mathcal{X} \rightarrow \mathcal{Y}^{\text{ENT}} \times \mathcal{Y}^{\text{EVT}} \times \mathcal{Y}^{\text{REL}} \times \mathcal{Y}^{\text{ARG}}$ that:

1. Performs accurate predictions for all four IE tasks on the target domain test data

2. Learns representations that are domain-invariant, such that features useful for IE in the source domain transfer effectively to the target domain

3. Captures and preserves the interdependencies between the four tasks across domains, ensuring that task relationships learned from the source domain remain valid in the target domain

The key technical challenges in this problem are threefold. First, the model must learn to distinguish between domain-specific features (which should be

discarded) and domain-invariant features (which should be retained) without access to target domain labels to validate this distinction. Second, unlike single-task domain adaptation, the model must simultaneously adapt *multiple* heterogeneous task representations (entities, events, relations, arguments), each of which may experience different types and magnitudes of domain shift. Third, the model must preserve the complex structural relationships between tasks across domains: for instance, the pattern that entities serving as arguments in events should also participate in relations must remain valid even when surface features change dramatically.

The DA4JIE framework, presented in the next section, addresses these challenges through two synergistic mechanisms. The Instance-relational Domain Adaptation (IrDA) module aligns representations at the task-instance level while respecting the heterogeneity of different task types. The Context-invariant Structure Learning (CiSL) module extracts domain-invariant structural information from sentences to provide robust features for all tasks. Together, these components enable effective unsupervised domain adaptation for the complex structured prediction problem of joint information extraction.

3.3 Proposed Method: DA4JIE

The DA4JIE framework operates through a unified architecture that processes input sentences from both source and target domains through the same encoder to ensure feature alignment. Figure 4 illustrates the complete system architecture. The model consists of four main stages:

1. **Encoding and Context-Invariant Structure Learning:** Input sentences from both domains are processed by a pretrained transformer encoder (e.g., BERT) to obtain contextualized word representations. Concurrently, the CiSL

module extracts attention probability matrices at each transformer layer to create attention graphs, using position embeddings as domain-agnostic node features. These attention graphs are combined with syntactic dependency graphs and fused across layers using a Graph Transformer Network (GTN) to produce a context-invariant graph (CiG). The node features from the CiG are combined with the contextual representations to create enriched word-level features.

2. **Instance Detection:** The enriched representations are fed into separate Conditional Random Field (CRF) layers for event triggers and entity mentions to identify candidate instance spans using BIO tagging schemes. This step detects where entities and events occur in the sentence without yet classifying their types.

3. **Instance Representation and Graph-Conditioned Pooling:** For each detected instance span, we compute a representation vector by aggregating information from the span's words. The CiSL module provides a graph-conditioned pooling mechanism that leverages the CiG structure to obtain more robust instance representations that incorporate both local span information and global sentence structure.

4. **Domain Adaptation and Task Prediction:** The instance representations from source samples are used to optimize the model for the four JIE classification tasks (entity typing, event typing, relation classification, argument role classification). Simultaneously, the IrDA module takes instance representations from both source and target domains and performs graph-

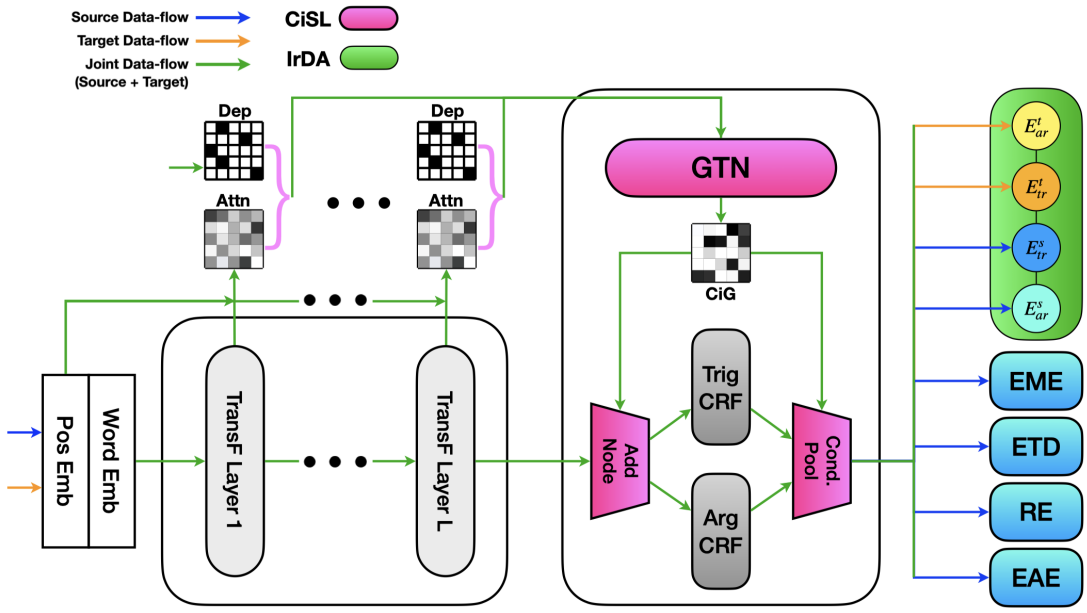


Figure 4. The overall architecture of our framework DA4JIE. First, input sentences from both source and target domains go through the same transformer encoder to compute their contextual representations. Concurrently, the **CiSL** module (pink) extracts the attention probability matrices at each layer to create attention graphs, using position embeddings as nodes features. These graphs are used to augment the dependency graphs, which are then fused across layers by a GTN to create a context-invariant graph. The node features of which are combined with the contextual representations as input for the instance span detection task using CRF layers. Next, the instance representations are computed based on the outputted spans conditioned on the context-invariant graph. Finally, source instances are used to optimize the encoder for the main JIE tasks (blue), while the **IrDA** module (green) takes the representations from the corresponding instance types for the nodes in the type-relational graph to calculate the discriminator loss.

relational domain-adversarial training to align representations across domains while respecting task-type heterogeneity.

The architecture embodies three key design principles. First, all domain-specific processing uses shared parameters to ensure that representations from source and target domains lie in the same feature space. Second, the CiSL module operates on structural information (attention patterns, dependency relations, positional information) that is inherently domain-invariant, providing a stable

foundation for adaptation. Third, the IrDA module respects the heterogeneity of task types by learning which instance representations should be aligned and to what degree, rather than enforcing uniform alignment across all tasks.

3.3.1 Base JIE Architecture.

The base architecture for joint information extraction follows the general framework of recent transformer-based JIE systems (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a), with important modifications to facilitate domain adaptation. The following encoding process is applied to data from both domains; thus, we omit the domain index in notations for brevity unless explicitly needed for clarity.

Contextual Encoding. Given an input sentence $w = [w_1, w_2, \ldots, w_n]$ with n words, the model first employs a pretrained transformer encoder (e.g., BERT (Devlin et al., 2019)) to obtain contextualized representations. Specifically, the sentence is first tokenized into subwords using the transformer's tokenizer (e.g., WordPiece for BERT). The resulting subword sequence is then fed through the transformer's layers to produce contextualized embeddings at each layer. For each word w_i, we obtain its representation $x_i \in \mathbb{R}^h$ by averaging the embeddings of its constituent subwords from the transformer encoder's final layer, yielding the sequence $X = [x_1, x_2, \ldots, x_n]$, where h is the hidden dimension of the transformer model.

We employ averaging of subword representations to obtain word-level features because (1) it provides a simple and effective way to handle the mismatch between subword tokenization and word-level annotations in IE tasks, and (2) averaging has been shown to produce robust representations that capture the full morphological content of words (Devlin et al., 2019; Y. Lin et al., 2020).

Instance Span Detection. To identify candidate instances for the JIE tasks, the model employs two separate Conditional Random Field (CRF) layers: one for event triggers and another for entity mentions/event arguments. Following standard practice (Y. Lin et al., 2020), both CRF layers take as input the word-level contextual representation sequence X (which will be enriched with structural information from CiSL as described in Section 3.3.3). Each CRF layer outputs the best BIO tag sequence (Chiu & Nichols, 2016) to indicate the spans of event triggers and entity mentions respectively in the sentence. Importantly, at this stage, no type classification is performed yet - the CRF layers only identify the boundaries of instances (i.e., which words belong to entity mentions or event triggers) without predicting their specific types.

Instance Representation. The spans identified by the CRF layers are then used to compute representation vectors for each instance. Let E_{tr} and E_{ar} denote the sets of detected event trigger instances and entity mention/argument instances, respectively (note that each set can contain multiple instances from a single sentence). For each instance, we aggregate information from the words in its corresponding span to obtain a fixed-dimensional instance representation. The base aggregation approach computes each instance representation by summing the word representations in the span. However, our CiSL module introduces a more sophisticated graph-conditioned pooling mechanism (described in Section 3.3.3) that produces richer instance representations by leveraging structural information from the context-invariant graph.

Task-Specific Classification. Given the instance representations, separate task-specific feed-forward networks (FFN) are employed to calculate label scores for classification. Specifically:

- For **Entity Mention Extraction (EME)**, a FFN_{ENT} takes each entity mention representation $e \in E_{ar}$ and computes a score vector over entity types $\mathcal{T}^E$: $s_{\text{ENT}}(e) = \text{FFN}_{\text{ENT}}(e)$.
- For **Event Trigger Detection (ETD)**, a FFN_{EVT} takes each trigger representation $g \in E_{tr}$ and computes a score vector over event types $\mathcal{T}^G$: $s_{\text{EVT}}(g) = \text{FFN}_{\text{EVT}}(g)$.
- For **Relation Extraction (RE)**, a FFN_{REL} takes pairs of entity mention representations (e_i, e_j) where $e_i, e_j \in E_{ar}$, and computes a score vector over relation types $\mathcal{T}^R$: $s_{\text{REL}}(e_i, e_j) = \text{FFN}_{\text{REL}}([e_i; e_j])$, where $[;]$ denotes concatenation.
- For **Event Argument Extraction (EAE)**, a FFN_{ARG} takes pairs of trigger and entity representations (g, e) where $g \in E_{tr}$ and $e \in E_{ar}$, and computes a score vector over argument roles $\mathcal{T}^A$: $s_{\text{ARG}}(g, e) = \text{FFN}_{\text{ARG}}([g; e])$.

For source domain samples with gold labels, these score vectors are converted to probabilities via softmax and used to compute cross-entropy losses for each task. The combined task loss $\mathcal{L}_c^s$ for the source domain is the sum of individual task losses:

$$\mathcal{L}_c^s = \mathcal{L}_{\text{ENT}} + \mathcal{L}_{\text{EVT}} + \mathcal{L}_{\text{REL}} + \mathcal{L}_{\text{ARG}} \tag{3.1}$$

where each component is the cross-entropy loss between predicted label distributions and gold labels for the respective task.

Note that for target domain samples, since labels are unavailable, these classification losses cannot be computed. Instead, target samples are used exclusively for the domain adaptation component (IrDA), which we describe next.

3.3.2 Instance-relational Domain Adaptation (IrDA). Standard domain-adversarial training methods such as Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2016) learn domain-invariant representations by training a feature extractor to fool a domain discriminator that attempts to classify whether features come from the source or target domain. However, DANN enforces strict uniform alignment, treating all instances identically and assuming that all features should be equally domain-invariant. This approach is suboptimal for JIE for two reasons. First, different task types (entities vs. events vs. relations vs. arguments) may experience different types and magnitudes of domain shift, requiring flexible, type-specific adaptation strategies. Second, entity mentions and event triggers have fundamentally different linguistic characteristics: triggers are restricted to specific event types and share structural patterns across domains, while entity mentions are more diverse and context-dependent, exhibiting greater domain-specific variation.

To address these limitations, we propose Instance-relational Domain Adaptation (IrDA), which generalizes DANN to enable flexible, type-aware alignment. Our approach is inspired by Graph-relational Domain Adaptation (GrDA) (Z. Xu et al., 2022), which was originally designed for multi-domain adaptation with multiple source domains. We adapt GrDA's core insight to the JIE setting by treating each task type in each domain as a node in a relational graph. By controlling the adjacency structure of this graph, we can specify which representations should be aligned and to what degree, enabling flexible adaptation that respects the heterogeneity of task types.

Type-Relational Graph Construction. We model the relationships between task types across domains using a type-relational graph $G_r = (V_r, A_r)$. The vertex set V_r consists of four nodes representing the instance types from both domains:

$$V_r = \{E^s_{ar}, E^s_{tr}, E^t_{ar}, E^t_{tr}\} \tag{3.2}$$

where the superscripts s and t denote source and target domains, and subscripts ar and tr denote entity/argument instances and trigger instances, respectively. Each node represents a collection of instances of a particular type from a particular domain.

The adjacency matrix $A_r \in \mathbb{R}^{4\times4}$ dictates which pairs of instance types should be aligned. Specifically, $A_r[i, j] = 1$ indicates that instance types i and j should have their representations aligned through domain-adversarial training, while $A_r[i, j] = 0$ indicates no alignment constraint between types i and j. By setting the adjacency matrix appropriately, we can control the alignment structure to reflect the true relationships among task types in JIE.

Chain Connection Pattern. Rather than using a fully-connected adjacency matrix (which would correspond to standard DANN's uniform alignment), we propose a *chain connection* pattern for the type-relational graph. Specifically, we assume connections in the order $E^s_{ar} \rightarrow E^s_{tr} \rightarrow E^t_{tr} \rightarrow E^t_{ar}$, as illustrated in Figure 5. This chain structure reflects the true relationships among instance types and domains in JIE:

- **Trigger-to-Trigger Alignment ($E^s_{tr} \leftrightarrow E^t_{tr}$):** Event triggers are restricted to predefined event types that are shared across domains. Therefore, trigger representations should be strongly aligned across domains. For example,

words that trigger TRANSPORT events ("arrived," "traveled," "flew") share semantic properties regardless of domain.

- **Argument-to-Argument Weak Alignment ($E^s_{ar} \leftrightarrow E^t_{ar}$):** Entity mentions/arguments are more diverse and context-dependent than triggers, with significant domain-specific variations in how entities are expressed. They should not be directly aligned as strongly as triggers. However, the chain connection provides implicit, *weak* alignment through the intermediate trigger nodes: E^s_{ar} connects to E^s_{tr}, which connects to E^t_{tr}, which connects to E^t_{ar}. This indirect path enables controlled information flow between entity representations across domains.

- **Cross-Task Alignment Within Domains ($E^s_{ar} \leftrightarrow E^s_{tr}$ and $E^t_{ar} \leftrightarrow E^t_{tr}$):** Within each domain, entity and trigger representations should be aligned because they participate in shared event structures. This alignment helps transfer the model's ability to predict relations between triggers and arguments from the source domain to the target domain.

As demonstrated by Z. Xu et al. (2022), graph-relational domain adaptation with different connectivity patterns achieves different levels of alignment strength. Direct edges enforce strong alignment, while paths through intermediate nodes create weaker, more flexible alignment. The chain connection thus provides a principled way to specify that triggers should be strongly aligned across domains while entities receive more flexible, indirect alignment.

Graph-Relational Adversarial Training. Given the type-relational graph $G_r = (V_r, A_r)$, IrDA performs adversarial training through a minimax optimization

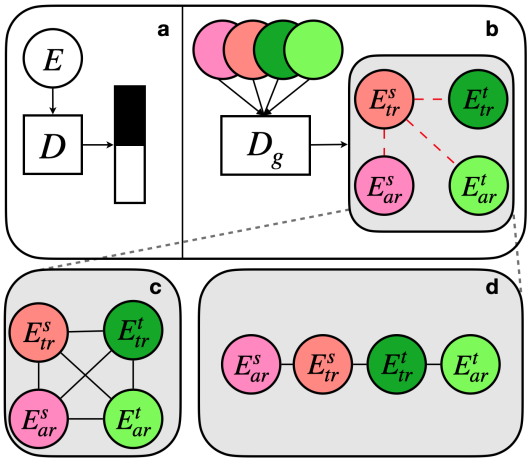


Figure 5. Top figures demonstrate the difference between DANN (**a**) and IrDA (**b**). Bottom figures are the relation graphs following the strict uniform alignment of standard DA methods (**c**) and the *chain* connection of IrDA (**d**).

process:

$$\min_{E,F} \max_{D_g} \mathcal{L}_c^s(E, F) - \lambda \mathcal{L}_d^{s-t}(D_g, E) \tag{3.3}$$

where E denotes the encoder (transformer + CiSL), F represents the task-specific classification heads, D_g is the graph discriminator, λ is a balancing hyperparameter, $\mathcal{L}_c^s$ is the combined task loss on source data (defined previously), and $\mathcal{L}_d^{s-t}$ is the domain-adversarial loss that we now define.

Unlike standard DANN where the discriminator predicts domain identity given individual instance representations, the graph discriminator D_g in IrDA aims to reconstruct the type-relational graph G_r by predicting the adjacency structure given the instance representations. This is achieved by computing pairwise relationships between instance representations from different types. Specifically, for each node type $i \in V_r$, we compute a prototype representation e_i by averaging the representations of all instances of that type:

$$e_i = \frac{1}{|V_i|} \sum_{v \in V_i} r_v \tag{3.4}$$

where V_i denotes the set of instances of type i, and r_v is the representation vector for instance v (obtained from the base JIE architecture). The graph discriminator then computes the pairwise relationship score $\hat{a}_{ij}$ between types i and j as:

$$\hat{a}_{ij} = e_i^\top e_j \tag{3.5}$$

This inner product measures the similarity between the prototype representations of types i and j. The discriminator attempts to predict whether types i and j should be connected in the graph (i.e., aligned) based on this similarity.

The discriminator loss $\mathcal{L}_d^{s-t}$ is computed as the binary cross-entropy between the predicted relationships $\hat{a}_{ij}$ and the true adjacency matrix entries a_{ij} from A_r:

$$\mathcal{L}_d^{s-t} = \sum_{i,j} \ell(\hat{a}_{ij}) \tag{3.6}$$

where the binary cross-entropy for each pair is:

$$\ell(\hat{a}_{ij}) = -a_{ij} \log \sigma(\hat{a}_{ij}) - (1 - a_{ij}) \log(1 - \sigma(\hat{a}_{ij})) \tag{3.7}$$

and $\sigma(\cdot)$ is the sigmoid function, and a_{ij} is the entry at position (i, j) in the adjacency matrix A_r.

The minimax objective in Equation 3.3 creates an adversarial game. The discriminator D_g seeks to maximize $\mathcal{L}_d^{s-t}$, which means it attempts to correctly predict the adjacency structure: if types i and j are connected in A_r (i.e., $a_{ij} = 1$), the discriminator wants $\hat{a}_{ij}$ to be high, indicating that their representations are similar and thus identifiable as connected types. Conversely, the encoder E seeks to minimize $\mathcal{L}_d^{s-t}$, which means it attempts to prevent the discriminator from recovering the adjacency structure. At equilibrium, the encoder learns representations such that connected types in the graph have aligned representations (because the discriminator cannot distinguish them), while unconnected types remain distinguishable.

3.3.3 Context-invariant Structure Learning (CiSL).

While IrDA addresses the challenge of flexible alignment across task types, domain-adversarial training methods in general face a fundamental sensitivity to the degree of domain discrepancy. Theoretical analysis by David et al. (2010) shows that the target domain error bound in domain adaptation depends not only on source domain error and distribution divergence, but also on the *ideal joint error* - the error of the ideal model that performs optimally on both domains. For a single prediction task, this ideal joint error is often assumed negligible and thus ignored. However, for JIE with multiple interdependent tasks, this assumption may not hold, and minimizing the ideal joint error becomes crucial for effective adaptation.

Our proposal to minimize the ideal joint error is to learn more transferable representations that facilitate accurate predictions in both domains. We achieve this through Context-invariant Structure Learning (CiSL), which induces domain-general structural information from input sentences. The key insight is that while lexical content and semantic patterns vary across domains, certain structural properties - such as syntactic dependencies and attention patterns in pretrained models - capture linguistic relationships that are universal across domains. By extracting and leveraging these context-invariant structures, we provide the model with robust features that transfer effectively between domains.

Graph Construction. CiSL combines two types of linguistic structures to create domain-invariant representations: syntactic dependency graphs and attention graphs extracted from the transformer encoder.

Dependency Graphs. For each input sentence, we construct a dependency graph $G_d = (V_d, A_d)$ based on the output of an off-the-shelf syntactic dependency parser. The node set V_d consists of word-level nodes (one node per word in the

sentence). Crucially, the node features are obtained by embedding the dependency relations between a word and its governor, rather than using word embeddings. These dependency relation embeddings are learnable parameters. The adjacency matrix $A_d \in \mathbb{R}^{n \times n}$ is a binary matrix where $A_d[i, j] = 1$ if word w_j is the governor of word w_i in the dependency tree, and 0 otherwise.

Syntactic dependency structures capture grammatical relationships (e.g., subject, object, modifier) that are largely consistent across domains within a language. While vocabulary and topics change dramatically between news and social media, the fundamental grammatical structures (e.g., that subjects precede verbs in English, that adjectives modify nouns) remain stable. By using dependency relation types rather than word identities as node features, we extract structural information that is invariant to domain-specific lexical content.

Attention Graphs. In addition to dependency graphs, CiSL leverages attention patterns from the pretrained transformer encoder. For each layer l of the transformer ($1 \leq l \leq L$, where L is the total number of layers), we define an attention graph $G_a^l = (V_a^l, A_a^l)$. The node set V_a^l again consists of word-level nodes. The node features are the transformer's position embeddings at layer l, which encode positional information independent of lexical content. The adjacency matrix $A_a^l \in \mathbb{R}^{n \times n}$ is the attention probability matrix from layer l, where $A_a^l[i, j]$ represents the attention weight from word w_i to word w_j.

Attention patterns in pretrained language models have been shown to capture various linguistic phenomena, including syntactic relations, coreference, and semantic associations (Clark, Khandelwal, Levy, & Manning, 2019; Jawahar, Sagot, & Seddah, 2019). Importantly, because the transformer encoder is pretrained on massive diverse corpora and shared across domains during fine-tuning, the attention

patterns it produces encode generalizable linguistic relationships rather than domain-specific surface features. By using position embeddings as node features, we ensure that the attention graphs remain independent of domain-specific vocabulary.

Attention-Augmented Dependency Graphs. To leverage both syntactic and attention-based structures, CiSL creates augmented graphs that merge dependency and attention information. For each layer l, we compute an attention-augmented dependency graph $G^l_{da} = (V^l_{da}, A^l_{da})$ by combining the dependency graph G_d with the attention graph G^l_a:

$$A^l_{da} = \alpha^l_a A^l_a + \alpha^l_d A_d \tag{3.8}$$

$$Z^l_{da} = \beta^l_a Z^l_a + \beta^l_d Z_d \tag{3.9}$$

where Z^l_a and Z_d are the node feature matrices for the attention graph and dependency graph respectively, Z^l_{da} is the merged node feature matrix, and $\{\alpha^l_a, \alpha^l_d, \beta^l_a, \beta^l_d\}_{l=1}^{L}$ are learnable weight parameters that control the contribution of each graph type at each layer.

The layer-specific weights allow the model to adaptively determine how much to rely on attention versus dependency information at each layer. Previous work has shown that different layers of pretrained transformers capture different types of linguistic information (Jawahar et al., 2019), with lower layers encoding more syntactic information and higher layers encoding more semantic information. The learnable weights enable CiSL to leverage this layer-specific specialization.

Graph Transformer Network Fusion. Having created attention-augmented dependency graphs at each transformer layer, CiSL employs a Graph Transformer Network (GTN) (Yun et al., 2019) to fuse these layer-specific graphs into a single

context-invariant graph (CiG) $G_{ci} = (V_{ci}, A_{ci})$ with adjacency matrix $A_{ci} \in \mathbb{R}^{n \times n}$ and node features $Z_{ci} \in \mathbb{R}^{n \times h}$.

A Graph Transformer Network is a meta-path learning architecture that can automatically learn how to combine multiple input graphs into a unified graph structure. GTN uses multi-hop graph convolution to discover useful connections (meta-paths) across different graph types and layers. In our case, GTN takes as input the L attention-augmented dependency graphs $\{G^l_{da}\}^L_{l=1}$ and learns to fuse them into a single graph G_{ci} that captures the most transferable structural information for JIE. The GTN parameters are learned end-to-end during training.

Incorporating Context-Invariant Structure. The context-invariant graph G_{ci} provides two sources of structural information: its node features Z_{ci} and its adjacency structure A_{ci}. To incorporate G_{ci} into the JIE model, CiSL first enriches the contextual word representations by adding the CiG node features:

$$X_{ci} = X + Z_{ci} \tag{3.10}$$

where $X = [x_1, \ldots, x_n]$ are the original contextual representations from the transformer, and X_{ci} are the enriched representations. These enriched representations X_{ci} are then used as input to the CRF layers for instance span detection (as mentioned in Section 2.3.2).

By adding the domain-invariant structural features Z_{ci} to the contextual representations, we encourage the downstream tasks (instance detection and classification) to leverage context-independent information in addition to the potentially domain-specific features in X. This provides the model with a stable foundation of transferable features that can be relied upon when domain shift makes the contextual features less reliable.

Graph-Conditioned Pooling. In addition to enriching word-level representations, CiSL introduces a novel pooling mechanism to compute instance representations that leverages the global structure captured by G_{ci}. Traditional approaches simply sum the word representations within an instance span:

$$e_j = \sum_{w_i \in \text{span}_j} x_i \tag{3.11}$$

This pooling strategy solely relies on the text used to express the instance's meaning in the specific context, potentially incorporating domain-specific lexical features.

CiSL reformulates this pooling operation using a binary assignment matrix $S_{\text{base}} \in \mathbb{R}^{n \times m}$, where $m \leq n$ is the number of detected spans and $S_{\text{base}}[i, j] = 1$ if and only if word w_i lies inside span j. The instance representations can then be computed in matrix form as:

$$E = S_{\text{base}}^{\top} X_{ci} \tag{3.12}$$

where $E \in \mathbb{R}^{m \times h}$ contains the representations for all m instances, and each row e_j is the representation for instance j.

Rather than fixing the assignment matrix to S_{base}, CiSL proposes to learn a refined assignment matrix S_{ci} conditioned on the context-invariant graph structure. This is achieved by applying a Graph Convolutional Network (GCN) (Kipf & Welling, 2017) to G_{ci}:

$$S_{ci} = \gamma \odot S_{\text{base}} + \mu \tag{3.13}$$

where $\odot$ denotes element-wise multiplication, and $(\gamma, \mu) = \text{GCN}(Z_{ci}, A_{ci}) \in \mathbb{R}^{n \times 2m}$ are the outputs of a graph convolutional network that takes the context-invariant graph as input. The GCN consists of 2-3 layers that propagate information across the graph structure to compute node-level features γ and μ.

The final instance representations are then computed as:

$$E = S_{ci}^{\top} X_{ci} = (\gamma \odot S_{\text{base}})^{\top} X_{ci} + \mu^{\top} X_{ci} \tag{3.14}$$

This formulation has an intuitive interpretation. The first term $(\gamma \odot S_{\text{base}})^{\top} X_{ci}$ computes a weighted sum over words in the instance span, where γ (conditioned on graph structure) can modulate the contribution of each word. This allows the model to emphasize words that are structurally important according to G_{ci} and de-emphasize domain-specific words. The second term $\mu^{\top} X_{ci}$ aggregates information over all words in the sentence (not just the span), enabling the model to incorporate global context. By conditioning both γ and μ on the context-invariant graph through the GCN, the pooling operation is guided by domain-invariant structural information, producing more robust instance representations that transfer better across domains.

3.3.4 Training Objective. Having presented the IrDA and CiSL modules, we now describe the complete training objective that combines all components of the DA4JIE framework. The model is trained to simultaneously optimize the JIE tasks on source data, align instance representations across domains through IrDA, and learn context-invariant structures through CiSL.

The complete training objective is:

$$\min_{E,F} \max_{D_g} \mathcal{L}_{\text{total}} = \mathcal{L}_c^s(E, F) - \lambda \mathcal{L}_d^{s-t}(D_g, E) \tag{3.15}$$

where:

– E represents the complete encoder, which includes the pretrained transformer, the CiSL module (GTN, GCN, and graph-conditioned pooling), and all intermediate representations.

- F represents all task-specific classification heads (FFNs for entity typing, event typing, relation classification, and argument classification).
- D_g is the graph discriminator for IrDA.
- $\mathcal{L}_c^s$ is the combined supervised task loss on labeled source data (sum of losses for all four tasks).
- $\mathcal{L}_d^{s-t}$ is the domain-adversarial loss defined in Equation 3.6.
- λ is a hyperparameter that balances the task loss and domain-adversarial loss.

Training Procedure. During training, at each iteration, we sample a mini-batch containing both labeled source domain samples and unlabeled target domain samples (typically in a 1:1 ratio). For each sample:

1. The encoder E processes the input sentence through the transformer to obtain contextual representations X.
2. The CiSL module constructs the context-invariant graph G_{ci} and computes enriched representations $X_{ci} = X + Z_{ci}$.
3. The CRF layers detect instance spans using X_{ci} as input.
4. The CiSL graph-conditioned pooling computes instance representations E from the detected spans.
5. For source samples: The task-specific classifiers F predict labels and $\mathcal{L}_c^s$ is computed using gold labels.

6. For both source and target samples: Instance representations are used to compute prototype representations for each instance type, and the IrDA module computes $\mathcal{L}_d^{s-t}$.

The minimax optimization is performed using gradient reversal (Ganin et al., 2016). During the forward pass, the graph discriminator D_g computes $\mathcal{L}_d^{s-t}$ normally. During the backward pass, the gradients of $\mathcal{L}_d^{s-t}$ with respect to the encoder parameters E are reversed (multiplied by $-\lambda$), while the gradients with respect to the discriminator parameters D_g are kept as is. This implements the minimax objective: the discriminator parameters are updated to maximize $\mathcal{L}_d^{s-t}$ (minimize discriminator error), while the encoder parameters are updated to minimize $\mathcal{L}_d^{s-t}$ (maximize discriminator error, i.e., fool the discriminator).

All parameters (encoder E, classifiers F, discriminator D_g) are optimized jointly using the Adam optimizer (Kingma & Ba, 2015). The learning rate and other hyperparameters are tuned on the source domain development set, as we do not have access to labeled target domain data for hyperparameter selection. Specific implementation details are provided in Section 3.4.

Inference. At test time, given an input sentence from the target domain:

1. The sentence is processed through the encoder E (transformer + CiSL) to obtain enriched representations X_{ci} and instance representations.

2. The CRF layers detect instance spans for entities and triggers.

3. The task-specific classifiers F predict entity types, event types, relation types, and argument roles for the detected instances.

4. The discriminator D_g is not used during inference.

The model produces joint predictions for all four IE tasks for the target domain sentence without requiring any labeled target domain data during training or inference.

3.4 Experiments

3.4.1 Datasets and Experimental Setup.

ACE 2005. Following prior work on joint information extraction (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a), we evaluate DA4JIE on the ACE-05 dataset (Walker et al., 2005), which provides annotations in 599 documents for entity mentions, event triggers, relations, and argument roles. In particular, there are 33 event types, 7 entity types, 6 relation types, and 22 argument roles. ACE-05 was collected from six different domains: broadcast news (`bn`), newswire (`nw`), broadcast conversation (`bc`), conversational telephone speech (`cts`), weblogs (`wl`), and usenet newsgroups/discussion forums (`un`).

For the UDA setting, we follow the setup of Ngo, Phung, and Nguyen (2021a) and construct source and target domains as follows. We gather data from two closely related formal domains - broadcast news (`bn`) and newswire (`nw`) - to create a sizable source domain dataset (referred to as the `in`-domain data). We use 80% of the source domain documents for training and reserve the remaining 20% for development (hyperparameter tuning and model selection). For the target domain, we consider each of the other four domains - broadcast conversation (`bc`), conversational telephone speech (`cts`), weblogs (`wl`), and usenet (`un`) - as a separate domain adaptation scenario. For each target domain, we reserve 20% of its documents as unlabeled training data (to be used by domain adaptation methods during training), and the remaining 80% serves as the test dataset with gold annotations for evaluation. Crucially, the target domain labels are never

observed during training, adhering to the unsupervised domain adaptation setting. This setup creates four domain adaptation scenarios: `in`→`bc`, `in`→`cts`, `in`→`wl`, and `in`→`un`.

Task-Specific Metrics. Following previous work on joint information extraction (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a), we evaluate performance using F1 scores for all four tasks. For entity mention extraction and event trigger detection, we report results at two levels of granularity: **Identification F1 (denoted with suffix "-I")**, where a prediction is considered correct if the predicted span boundaries match the gold annotation, regardless of the predicted type; and **Classification F1 (denoted with suffix "-C")**, where a prediction is considered correct only if both the span boundaries and the type label match the gold annotation. For relation extraction and event argument extraction, we report only classification F1 scores. We compute macro-averaged F1 scores and report the average classification F1 across all four tasks (denoted as **aTask**). For out-of-domain evaluation, we also report the average performance across all target domains for each task (denoted as **aDom**).

Implementation Details. We implement DA4JIE in PyTorch 1.7.1 (Paszke et al., 2019). For the pretrained language model encoder, we use BERT-large-cased (Devlin et al., 2019) from the Huggingface Transformers library (Wolf et al., 2020). BERT-large has 24 layers with hidden dimension $h = 1024$. The CRF layers for trigger and entity detection use standard linear-chain CRF implementations. For the Context-invariant Structure Learning (CiSL) module, we use a Graph Transformer Network (GTN) (Yun et al., 2019) with 4 channels to fuse the attention-augmented dependency graphs across layers. The Graph Convolutional

Network (GCN) (Kipf & Welling, 2017) within CiSL consists of 3 layers, each with 512 hidden units. For the task-specific classification heads (FFNs), we use 3-layer feed-forward networks with hidden dimensions [200, 100, 50] and ReLU activations. We use the Adam optimizer (Kingma & Ba, 2015) with learning rate 1×10^{-5} for all model parameters. The batch size is set to 16, with each batch containing 50% source domain samples (with labels) and 50% target domain samples (without labels). The IrDA balancing term λ is set to 1.0. Every model is trained for 50 epochs. At the end of each epoch, we evaluate the model on the in-domain (source) development set and select the checkpoint with the best average classification F1 across all four tasks. For the syntactic dependency graphs used in CiSL, we use the Stanza toolkit (Qi, Zhang, Zhang, Bolton, & Manning, 2020) to parse all sentences and obtain dependency trees. All experiments are conducted on a single NVIDIA Tesla V100-SXM2 GPU with 32GB memory. We report results averaged over three runs with different random seeds to account for training variance. Training time for one full model on the ACE-05 source domain takes approximately 30 minutes per epoch.

Baselines. We compare DA4JIE with the following current SOTA JIE systems: (i) **BERT** Devlin et al. (2019) uses a shared Transformer encoder to represent the instances for ETD, EME, EAE, and RE and performs classification for the instances based on the task-specific label distributions. (ii) **OneIE** Y. Lin et al. (2020) is same as BERT, but leverages set of predefined global features to capture the cross-subtask and cross-instance interactions. (iii) **FourIE** M. V. Nguyen et al. (2021a) creates a graph structure of contextual representations to explicitly capture the interactions between related instances of the four IE tasks in a sentence, while also employing a heuristic dependency between the task instances in a dependency-

based regularization to further boost the performance of the models. **OneIE** and **FourIE** are current state-of-the-art models for JIE.

3.4.2 Main Results. Table 8 presents the main experimental results comparing DA4JIE with the baseline methods across all domain adaptation scenarios. We report performance on both the in-domain (source) test set and all four out-of-domain (target) test sets.

Table 8. Performance comparison on ACE-05 in-domain and out-of-domain test sets. F1 scores are reported for all tasks. The suffixes "-I" and "-C" correspond to identification performance (offset correctness only) and classification performance (both offsets and types must match), respectively. **aTask** is the average classification F1 across the four tasks. **aDom** is the average out-of-domain score for each task across all target domains. Best results for each setting are in **bold**.

		Test Domain					
Model	**Task**	in	bc	cts	wl	un	**aDom**
BERT	Trigger-I	78.4	71.4	65.2	62.9	66.3	66.4
	Role-I	64.1	59.5	49.0	46.3	46.9	50.4
	Entity	88.9	80.8	84.0	85.5	80.9	82.8
	Relation-C	64.3	61.7	58.0	52.5	48.0	55.0
	Trigger-C	76.3	68.7	62.4	56.3	64.5	63.0
	Role-C	60.8	55.4	47.9	42.9	43.0	47.3
	aTask	72.6	66.6	63.1	59.3	59.1	62.0
OneIE	Trigger-I	79.1	70.3	68.2	63.2	64.6	66.6
	Role-I	66.2	60.1	51.2	50.6	46.7	52.1
	Entity	89.1	79.5	86.9	85.5	81.5	83.4
	Relation-C	65.6	63.1	56.7	54.7	50.0	56.1
	Trigger-C	77.2	67.5	64.6	56.8	63.4	63.1
	Role-C	62.2	55.7	49.9	47.2	42.6	48.9
	aTask	73.5	66.5	64.6	61.1	59.4	62.9
FourIE	Trigger-I	79.1	70.7	66.0	65.2	64.3	66.6
	Role-I	66.6	60.0	52.6	48.9	49.1	52.6
	Entity	89.1	80.3	84.4	85.4	81.9	83.0
	Relation-C	66.0	63.7	56.6	53.1	52.7	56.5
	Trigger-C	76.9	68.5	63.2	56.4	62.4	62.6
	Role-C	61.8	55.4	51.8	44.5	43.6	48.8
	aTask	73.5	66.9	64.0	59.8	60.1	62.7
DA4JIE	Trigger-I	79.0	**72.2**	**66.0**	**64.4**	**66.5**	**67.3**
	Role-I	67.3	**59.0**	**54.6**	**49.8**	**51.5**	**53.7**
	Entity	89.2	**82.6**	**86.0**	**85.2**	**83.0**	**84.2**
	Relation-C	68.8	**65.3**	**58.7**	**57.7**	**54.3**	**59.0**
	Trigger-C	76.5	**68.7**	**63.0**	**57.4**	**64.1**	**63.3**
	Role-C	62.5	**55.6**	**51.9**	**45.3**	**44.4**	**49.3**
	aTask	74.2	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**

As shown in Table 8, DA4JIE achieves the best out-of-domain performance across all four tasks and all four target domains. The average out-of-domain performance shows that DA4JIE achieves an average classification F1 of 64.0 across all target domains, representing a 1.3 absolute point improvement over the previous best baseline FourIE (62.7). Overall, DA4JIE is 2 points higher in F1 scores than BERT's overall, surpassing the current state-of-the-art methods by over 1 point on average.

Comparing the baselines reveals that the state-of-the-art joint IE systems OneIE and FourIE provide marginal improvements over the simple BERT baseline in out-of-domain settings. In particular, while the specialized architectures of these models boost in-domain performance as expected, they are not tailored to UDA settings where the focus is on extracting transferable and domain-invariant features. As a result, their effectiveness in out-of-domain settings over BERT is situational (OneIE is good for `cts` and `wl` domains, while FourIE is better at adapting to `un` domains).

In contrast, DA4JIE manages to achieve the best adaptation performance on average across all considered domains. Notably, this improvement is the result of the simultaneous increases in the average performance of all downstream tasks, which is achieved by combining IrDA and CiSL modules as shown in the following section.

3.4.3 Ablation Studies. We conduct an ablation study to validate the effectiveness of each of our main components by investigating variations of our model by removing CiSL, IrDA, and both respectively. The results are shown in Table 9.

Table 9. Component-wise ablation study on ACE-05 test datasets. Performance (average classification F1 across four tasks) when removing CiSL and/or IrDA from the full DA4JIE model.

Model Variant	bc	cts	wl	un	**Avg**
DA4JIE (full)	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**
DA4JIE - CiSL	67.7	61.2	59.5	60.3	62.2
DA4JIE - IrDA	67.2	64.3	60.3	60.1	63.0
DA4JIE - IrDA - CiSL	66.6	63.1	59.3	59.1	62.0

We observe that **DA4JIE-IrDA** noticeably boosts performance for all domains compared to BERT (**DA4JIE-IrDA-CiSL**), while **DA4JIE-CiSL** only has positive impact when adapting to `bc` and `un` domains, providing little to no improvement on average. This is the result of CiSL making the instance representations more transferable at low-level, thus ensuring the necessary condition for the domain-adversarial training in IrDA to reach equilibrium. By combining both components, **DA4JIE** significantly outperforms other variants, especially when transferring to target domains that are highly dissimilar to source domains (i.e., `wl` and `un`).

3.5 Analysis

3.5.1 Instance-relational Graph Analysis. We investigate the effect of IrDA with different patterns of relationships in the type-relational graph compared to our *chain* relation in DA4JIE. Table 10 presents results with different adjacency patterns.

Table 10. Analysis of different type-relational graph structures for IrDA. Average classification F1 across four tasks on each target domain test set. "Chain" is the pattern used in DA4JIE.

Graph Pattern	bc	cts	wl	un	**Avg**
None	66.6	63.1	59.3	59.1	62.0
Full	67.1	63.7	60.0	60.1	62.7
Pair-Task	67.5	63.2	60.0	61.1	62.9
Pair-Dom	67.0	62.7	59.1	59.5	62.1
Chain	**68.0**	**65.0**	**61.4**	**61.4**	**64.0**

In Table 10, **Full** refers to the standard DANN approach where all types (i.e., task+domain) are uniformly aligned. **Pair-Task** and **Pair-Dom** are models with only a pair of edges in relation graph. The former connects the same task across domains ($\mathbf{E}^s_{ar}$-$\mathbf{E}^t_{ar}$ and $\mathbf{E}^s_{tr}$-$\mathbf{E}^t_{tr}$), while the latter has tasks in the same domain linked ($\mathbf{E}^s_{tr}$-$\mathbf{E}^s_{ar}$ and $\mathbf{E}^t_{tr}$-$\mathbf{E}^t_{ar}$). Finally, **None** means no adaptation is used and **Chain** corresponds to our assumption in DA4JIE.

The results show that **Full** improves over **None**, but underperforms when compared to **Pair-Task** in most new domains. This shows that the alignment imposed by **Full** is overly strict and not optimal when adapting multiple tasks together. In addition, appropriate connections are required for effective adaptation, as shown by the low scores of **Pair-Dom**, which basically is equivalent to domain-conditioning the representations without adapting between source and target domains.

We argue that **Chain** is robust and substantially outperforms other models across all domains because it reflects the true relationship among the tasks and domains (types) for JIE. In particular, event triggers are restricted and closely related to the predefined event classes which are shared across domains, therefore their representations should be aligned when adapting to new domains. Conversely,

event arguments (i.e., entity mentions) are more diverse and context-dependent, thus may significantly differ across domains and should not be directly connected in the relation graph. They are, however, implicitly connected in **Chain** through the event trigger nodes, which implies their representations are "weakly" aligned, as shown by Z. Xu et al. (2022) where GrDA is able to enforce different levels of alignment. Lastly, the pair of trigger-event edges in source and target domains also equates to aligning the representations of trigger-event relation and helps transfer the model's role classification ability from source to target domain.

3.5.2 Context-invariant Structure Learning Analysis. To determine the role of different components in CiSL module, we analyze their contributions to DA4JIE performance at different levels of downstream tasks. Table 11 presents results when removing or modifying various aspects of CiSL.

Table 11. Analysis of Context-invariant Structure Learning components. Average identification F1 (aId) and average classification F1 (aCls) across all four tasks and all four target domains.

Model Variant	**aId**	**aCls**
CiSL (full)	**60.5**	**64.0**
CiSL - Pool	59.7	63.2
CiSL - Node	59.4	62.3
CiSL - Node - Pool	58.0	62.0
CiSL - Dep	59.8	62.8
CiSL - Attn	59.2	62.5

In Table 11, **CiSL-Dep** and **CiSL-Attn** are the models without using dependency graph and attention graph respectively. **CiSL-Pool** just uses the base assignment matrix for pooling, and **CiSL-Node** is the case where node features of the context-invariant graph are removed from the inputs for the CRF layers. Finally, we completely disable the CiSL module in **CiSL-Node-Pool**.

From the results, it is clear that both node features and conditional pooling are responsible for the significant improvement of the final model. Particularly, adding the node features is more effective as it also helps boost the performance of identification tasks by making the representations more transferable from source to target domains at low-level. Furthermore, the last two rows in the table show that combining different kinds of structures has a positive impact, especially when they contain universal linguistic information that is general across domains.

3.6 Related Work

Joint Information Extraction. Classical methods for IE manually designed linguistic features to capture the dependency between IE tasks, including Integer Linear Programming for Global Constraints (Roth & Yih, 2004), Structured Perceptron (Judea & Strube, 2016; Miwa & Sasaki, 2014), and Graphical Models (Yang & Mitchell, 2016; Yu, Yao, Su, Li, & Seide, 2010). The advance of deep learning and large-scale language models (Devlin et al., 2019) has greatly enhanced the representation ability of modern IE models, enabling them to jointly solve multiple tasks via the shared contextual embeddings. These joint models focused on different sets of IE tasks such as EME and RE (T.-J. Fu, Li, & Ma, 2019; Luan et al., 2019; Veyseh et al., 2020; Zheng et al., 2017), and ETD and EAE (T. H. Nguyen, Cho, & Grishman, 2016; J. Zhang, Qin, Zhang, Liu, & Ji, 2019). Recently, some efforts have been made to address the four tasks all together by introducing specialized structures and regularizations to model the joint instance distribution across tasks (Y. Lin et al., 2020; M. V. Nguyen et al., 2021a; M. V. Nguyen, Min, Dernoncourt, & Nguyen, 2022; Z. Zhang & Ji, 2021). Our work continues in their direction, but in UDA setting which is more difficult but also much more practical compared to the standard supervised learning setting.

Unsupervised Domain Adaptation. The main line of research on UDA approaches the domain shift problem by learning domain-invariant representations, which is either achieved by explicitly reducing the distance between source and target feature space measured by some distribution discrepancy metric (Long et al., 2015; Zellinger et al., 2017), or by adversarial training in which the feature extractor is trained to fool a domain classifier (Ganin et al., 2016). There have been several prior works addressing UDA setting for a singular IE task, including event trigger identification (Naik & Rosé, 2020b), event detection (Ngo et al., 2021a; Trung et al., 2022), and relation extraction (L. Fu, Nguyen, Min, & Grishman, 2017). However, a method specifically tackles joint task learning in UDA is still absent from the literature to the best of our knowledge. Our IrDA is the first to explicitly take into account multiple representations of different tasks when transferring between source and target domains.

3.7 Summary

This chapter addressed the challenge of unsupervised domain adaptation for joint information extraction, corresponding to Research Question 2 (RQ2) from Chapter I. Existing JIE systems experience substantial performance degradation when applied to out-of-domain data, and prior domain adaptation methods focused exclusively on single IE tasks. We proposed Domain Adaptation for Joint Information Extraction (DA4JIE), the first framework specifically designed for unsupervised domain adaptation in the joint IE setting involving all four core tasks. DA4JIE introduces two synergistic technical innovations: Instance-relational Domain Adaptation (IrDA), which uses graph-relational adversarial training to align type-specific representations across domains, and Context-invariant Structure Learning (CiSL), which extracts domain-invariant structural information by fusing

syntactic dependency graphs with transformer attention patterns. The extensive experiments demonstrate the effectiveness of the proposed framework. In the future, we plan to extend our approach to more general settings such as multi-source domain adaptation with more IE subtasks such as entity/event coreference resolution.

CHAPTER IV

ZERO-SHOT CROSS-LINGUAL TRANSFER LEARNING WITH MULTIPLE SOURCE AND TARGET LANGUAGES FOR INFORMATION EXTRACTION

This chapter contains materials from the paper: "Nghia Trung Ngo and Thien Huu Nguyen. 'Zero-shot Cross-lingual Transfer Learning with Multiple Source and Target Languages for Information Extraction: Language Selection and Adversarial Training.' arXiv preprint arXiv:2411.08785, 2024." Nghia served as the first author, responsible for problem formulation, method development, experimental design, and manuscript preparation. Thien Huu Nguyen provided research direction, guidance, and editorial contributions as the advisor.

As discussed in Chapter I, transfer learning has become a key approach for addressing data scarcity and allowing efficient deployment of NLP systems across diverse contexts. This chapter addresses Research Question 3 (RQ3), which concerns the development of effective cross-lingual transfer learning methods that allow knowledge sharing across languages, from high-resource to low-resource languages. While Chapters II and III addressed transfer across domains and tasks within the same language, transferring knowledge across languages with different linguistic structures creates different technical challenges that need specialized approaches.

This chapter investigates zero-shot cross-lingual transfer learning through three levels of complexity: single-transfer, multi-transfer, and relational-transfer settings. The main goal is to understand how linguistic relationships between languages affect transfer performance and how this understanding can guide practical system development. For single-transfer, we show that linguistic distances based on syntax, phonology, and inventory features from the URIEL database

correlate strongly with transfer performance. We develop a combined metric that achieves high correlation with transfer performance across different tasks and model scales. For multi-transfer, we use linguistic clustering to select optimal source languages, showing that careful selection improves performance by 3-4 F1 points over random selection. For relational-transfer, we propose a graph-relational adversarial learning approach that uses unlabeled data from multiple languages, achieving additional improvements. The experiments cover two information extraction tasks (event detection and relation extraction) across 17 languages and three model scales.

4.1 Introduction

In our interconnected world, linguistic diversity creates both opportunities and challenges for natural language processing technologies. With approximately 7,000 languages spoken worldwide, each carrying unique grammatical structures, phonological systems, and cultural contexts, the field faces a key question: how can we develop language technologies that serve all linguistic communities fairly? The rapid development around English-based datasets has pushed machine performance to be on par with human ability in English tasks, prompting recent work to explore NLP research in other languages (Liang et al., 2020; Ruder et al., 2021). However, despite advanced large-scale architectures and high English results, current models notably underperform in new languages, especially those considered low-resource and lacking high-quality annotated datasets for fine-tuning.

Cross-lingual transfer learning is one of the most promising approaches to bridge this linguistic divide. The key idea is to leverage knowledge acquired from high-resource source languages with abundant labeled data to improve performance on target languages where such resources are scarce or nonexistent. Currently, the

most popular and practical approach for information extraction involves zero-shot cross-lingual (ZSCL) transfer (Conneau et al., 2020; Goyal et al., 2021). These methods fine-tune Transformer-based multilingual language models (mLMs), which were pre-trained using unlabeled text from hundreds of languages, for downstream tasks using high-resource source-language labeled datasets - mainly English. The resulting models are then directly evaluated on the corresponding tasks in target languages without any target-language supervision.

Studies have shown that the performance of these multilingual models varies substantially across languages and tasks (Keung, Lu, Salazar, & Bhardwaj, 2020; Lauscher, Ravishankar, Vulić, & Glavaš, 2020; Turc, Lee, Eisenstein, Chang, & Toutanova, 2021). Several factors have been linked to this phenomenon, including data-dependent statistics (e.g., dataset size, word overlap) (Malkin, Limisiewicz, & Stanovsky, 2022) and data-independent features (e.g., phylogenetic and typological characteristics) (Dolicki & Spanakis, 2021; Y.-H. Lin et al., 2019). The majority of previous research addressing multilingual information extraction has been limited to the zero-shot cross-lingual single-transfer (one-to-one) setting, where models trained on a single high-resource source language are directly evaluated on individual target languages. While this setting has provided valuable insights into the cross-lingual transfer capabilities of multilingual pretrained language models such as multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), it offers limited understanding and practical benefit for the realistic goal of developing multilingual IE systems that can generalize to as many languages as possible.

Current advances in translation models and data gathering processes have enabled the creation of annotated datasets in numerous languages,

suggesting that multiple source languages should be used for improved transfer performance. Previous work has shown that cross-lingual transfer performance varies substantially across language pairs and tasks, with the effectiveness of transfer depending on multiple factors including dataset size, vocabulary overlap, and phylogenetic and typological relationships between languages (Keung et al., 2020; Lauscher et al., 2020; Y.-H. Lin et al., 2019). Understanding the deeper connections between linguistic relationships and transfer performance can provide tremendous practical implications: guiding data collection to achieve optimal cost-performance trade-offs and tailoring modeling approaches to explicitly capture linguistic relations for improved generalization.

Previous work has tried to address the challenge of predicting cross-lingual transfer performance by defining it as a regression problem (Dolicki & Spanakis, 2021; Y.-H. Lin et al., 2019; Srinivasan & Athey, 2021). In these approaches, a regression model is trained to take linguistic features of source-target language pairs as input and predict a trained model's performance scores on target languages. Despite achieving high prediction accuracy, these works are not enough for the following two reasons. First, they place too much emphasis on the accuracy of the regression model, which is trained for a specific architecture on a particular task. As training configurations vary widely in practice, the results obtained from performance prediction models may become unreliable and not generally applicable. Another reason is that previous work is limited to the setting of single-transfer between two languages, in which only one source language - mainly English - is used.

This chapter investigates cross-lingual transfer learning through three increasingly complex scenarios, each addressing a key question. Rather than

focusing on performance prediction for a specific model architecture, we aim to build a complete understanding of how linguistic relationships affect transfer learning and how this understanding can guide the development of effective multilingual systems. Specifically, we address the following three transfer questions:

1. **TQ1 (Single-Transfer).** How do the relations in the linguistic landscape affect an information extraction model's cross-lingual transfer ability? Can we develop a linguistic distance metric that reliably correlates with transfer performance across different tasks and model scales?

2. **TQ2 (Multi-Transfer).** Can we implicitly use linguistic features as dataset-independent knowledge to efficiently address the more general many-to-many cross-lingual transfer setting? Given a set of languages with only their linguistic features as prior information, can we find the optimal subset of source languages for achieving the best cost-performance trade-off?

3. **TQ3 (Relational-Transfer).** Can we explicitly integrate linguistic relations into the learning process to effectively improve multi-transfer performance? How can we move beyond uniform language-invariant representations to capture the specific relationships between language pairs?

Our investigation uses three levels of increasing complexity, each building upon insights from the previous level. First, in the *single-transfer setting (ZSCL-S)*, we examine transfer from one source language to one target language, allowing detailed analysis of how specific linguistic features affect transfer performance and which language pairs help effective knowledge sharing. We extract phylogenetic and typological properties from the URIEL Typological Database (Littell et al., 2017) to compute pairwise linguistic distances across multiple feature categories

(syntax, phonology, inventory, morphology). By correlating these distances with transfer performance across diverse language pairs and IE tasks, we develop a combined distance metric that is highly correlated with transfer effectiveness while remaining robust across different tasks and model scales. Next, the *multi-transfer setting (ZSCL-M)* looks at scenarios with multiple labeled source languages and multiple target languages, investigating whether diverse source languages provide complementary information that improves target performance. Given the exponential number of possible source language combinations, we develop methods for selecting optimal source language sets based on linguistic clustering using our combined distance metric. This approach aims to achieve the best cost-performance trade-off by efficiently choosing which languages to annotate for training data while maximizing generalization across target languages. Finally, the *relational-transfer setting (ZSCL-R)* goes beyond learning language-invariant features to explicitly modeling the specific relationships between source and target languages. Using unlabeled data from all available languages, we propose a graph-relational adversarial learning framework that uses separate discriminators for each source language, conditioning the multilingual representations flexibly on connections expressed by the language relational graph induced from linguistic distances. This approach enables the model to capture not only shared cross-lingual features but also the specific linguistic relationships between language pairs, leading to more effective transfer.

The main contributions of this chapter are:

– A detailed analysis of linguistic distances and their correlation with cross-lingual transfer performance, examining 14 different distance metrics across multiple linguistic feature categories.

- A new combined linguistic distance metric that achieves high correlation (> 0.6 Pearson correlation) with transfer performance robustly across different tasks (event detection and relation extraction) and model scales (small, base, and large), addressing the task-specificity limitations of prior work.
- A language selection framework based on linguistic clustering that allows efficient identification of optimal source language sets for multilingual training, providing practical guidance for cost-effective data collection.
- A graph-relational adversarial learning approach that explicitly integrates linguistic relationships into the training process, outperforming both standard baselines and uniform alignment methods.
- Extensive experiments on two practical information extraction tasks (event detection on MINION (Pouran Ben Veyseh et al., 2022) and relation extraction on SMiLER (Seganti, Sorokin, Reddy, & Zhao, 2021)) covering 17 typologically diverse languages including Arabic, Chinese, Hindi, Japanese, Korean, and various European languages.

The remainder of this chapter is organized as follows. Section 4.2 formally defines the zero-shot cross-lingual transfer problem and establishes notation. Section 4.3 describes the linguistic features and distance metrics used throughout our investigation. Section 4.4 presents our analysis of single-transfer learning and the development of the combined linguistic distance metric. Section 4.5 investigates multi-transfer learning with language selection via clustering. Section 4.6 introduces our graph-relational adversarial learning approach for relational-transfer. Section 4.7 discusses the implications and insights from our experiments. Section 4.8 reviews related work on cross-lingual transfer learning,

linguistic diversity, and adversarial language adaptation. Finally, Section 4.9 concludes the chapter with a summary of findings and connections to the broader dissertation.

4.2 Problem Statement and Background

4.2.1 Zero-shot Cross-lingual Transfer Problem Formulation.

Zero-shot cross-lingual transfer learning looks at scenarios where we have labeled training data in one or more source languages but want to test on target languages for which no labeled training data is available. Formally, let $\mathcal{L} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m\}$ denote a set of m available languages. We partition this set into source languages $\mathcal{L}_S \subseteq \mathcal{L}$ and target languages $\mathcal{L}_T \subseteq \mathcal{L}$, where $\mathcal{L}_S \cap \mathcal{L}_T = \emptyset$. For each source language $\mathcal{L}_i \in \mathcal{L}_S$, we have access to a labeled dataset $\mathcal{D}_{\mathcal{L}_i} = \{(x_j^{\mathcal{L}_i}, y_j^{\mathcal{L}_i})\}_{j=1}^{n_i}$, where $x_j^{\mathcal{L}_i}$ represents an input example (e.g., a sentence) in language $\mathcal{L}_i$ and $y_j^{\mathcal{L}_i}$ is the corresponding label (e.g., event types, relation types).

For target languages $\mathcal{L}_T$, we have no labeled data (zero-shot). Specifically, for each $\mathcal{L}_t \in \mathcal{L}_T$, we only have access to unlabeled test examples $\mathcal{D}_{\mathcal{L}_t}^{\text{test}} = \{x_k^{\mathcal{L}_t}\}_{k=1}^{n_t^{\text{test}}}$ used for evaluation. The zero-shot constraint is critical: the model must perform inference on target languages without ever observing labeled examples in those languages during training. This setting reflects realistic scenarios where annotation resources are unavailable for many languages, and we must rely entirely on transfer from source languages.

The goal of ZSCL transfer is to learn a model $f_\theta : \mathcal{X} \times \mathcal{L} \rightarrow \mathcal{Y}$ with parameters θ that is trained on the source language labeled data $\bigcup_{\mathcal{L}_i \in \mathcal{L}_S} \mathcal{D}_{\mathcal{L}_i}$ and can perform effectively on target languages $\mathcal{L}_T$. The model takes as input an example x in language $\mathcal{L}$ and produces a prediction $\hat{y} = f_\theta(x, \mathcal{L})$. In practice, multilingual pretrained language models such as XLM-RoBERTa (Conneau et al.,

2020) serve as the foundation for f_θ, providing shared cross-lingual representations that enable knowledge transfer across languages. This chapter investigates ZSCL transfer through three increasingly complex scenarios:

Zero-shot Cross-lingual Single-Transfer (ZSCL-S). In the single-transfer setting, we consider pairs of languages where $|\mathcal{L}_S| = 1$ and $|\mathcal{L}_T| = 1$. Let $\mathcal{L}_s$ denote the single source language and $\mathcal{L}_t$ the single target language. The model is trained exclusively on $\mathcal{D}_{\mathcal{L}_s}$ and evaluated on $\mathcal{D}_{\mathcal{L}_t}^{\text{test}}$. This setting allows detailed analysis of pairwise transfer between languages and study of how linguistic relationships between $\mathcal{L}_s$ and $\mathcal{L}_t$ affect transfer performance. We systematically evaluate all possible source-target pairs from the available languages to study the correlation between linguistic distances and transfer effectiveness.

Zero-shot Cross-lingual Multi-Transfer (ZSCL-M). In the multi-transfer setting, we have multiple source languages available for training: $|\mathcal{L}_S| > 1$. The target set $\mathcal{L}_T$ may contain one or more languages: $|\mathcal{L}_T| \geq 1$. The model is trained on the union of all source language datasets $\mathcal{D}_{\text{train}} = \bigcup_{\mathcal{L}_i \in \mathcal{L}_S} \mathcal{D}_{\mathcal{L}_i}$ and evaluated on each target language independently. This setting addresses the practical question of which source languages to select for training when the goal is to maximize performance across a set of target languages. Given the exponential number of possible source language subsets, we cannot evaluate all combinations. Instead, we use linguistic distances to guide selection through language clustering.

Zero-shot Cross-lingual Relational-Transfer (ZSCL-R). The relational-transfer setting extends ZSCL-M by also using unlabeled data from all available languages, including both source and target languages. For each language $\mathcal{L}_i \in \mathcal{L}$,

we have access to an unlabeled dataset $\mathcal{D}_{\mathcal{L}_i}^{\text{unlabeled}} = \{x_j^{\mathcal{L}_i}\}_{j=1}^{n_i^{\text{unlabeled}}}$. These unlabeled data are used during training to learn language representations that explicitly capture the linguistic relationships between languages through adversarial training. Unlike ZSCL-M which implicitly uses linguistic distances only for language selection, ZSCL-R directly includes linguistic relationships into the model's training objective via a graph-relational adversarial learning framework.

4.2.2 Datasets. We conduct experiments on two recently released multilingual information extraction datasets that provide data across typologically diverse languages: MINION for event detection and SMiLER for relation extraction.

MINION. The Multilingual Event Detection dataset (MINION) (Pouran Ben Veyseh et al., 2022) has event triggers for 8 typologically different languages: English (eng), Hindi (hin), Japanese (jpn), Korean (kor), Polish (pol), Spanish (spa), Swedish (swe), and Turkish (tur). The dataset has 16 event types following the ACE event schema, such as Conflict, Contact, Justice, Life, Movement, and Transaction events. MINION was specifically designed to support cross-lingual event detection research, with careful attention to annotation consistency across languages despite differences in how events are expressed linguistically.

SMiLER. The Samsung MultiLingual Entity and Relation Extraction dataset (SMiLER) (Seganti et al., 2021) has annotated entities and relations from 14 languages: Arabic (ara), German (deu), English (eng), Farsi (fas), French (fra), Italian (ita), Dutch (nld), Polish (pol), Portuguese (por), Russian (rus), Spanish (spa), Swedish (swe), Turkish (tur), and Ukrainian (ukr). The dataset defines 36

relation types with diverse semantic categories such as FAMILY, EMPLOYMENT, OWNERSHIP, LOCATION, and MEMBERSHIP relations. SMiLER requires models to both identify entity mentions and classify the relationships between entity pairs.

Table 12. Percentage distribution of training data for each language in MINION and SMiLER datasets. The shared languages are shown, with variations in resource category indicated.

MINION			**SMiLER**		
Language	**Code**	**%**	**Language**	**Code**	**%**
English	eng	39.76	Italian	ita	19.71
Polish	pol	13.7	French	fra	16.25
Turkish	tur	13.7	German	deu	13.75
			Portuguese	por	11.54
			Dutch	nld	10.38
Spanish	spa	9.99	English	eng	9.57
Portuguese	por	4.59	Korean	kor	5.0
Swedish	swe	4.59	Polish	pol	4.5
Hindi	hin	4.58	Spanish	spa	2.95
Korean	kor	4.58	Arabic	ara	2.49
Japanese	jpn	4.5	Russian	rus	1.71
			Swedish	swe	1.2
			Farsi	fas	0.7
			Ukrainian	ukr	0.26

Table 12 presents detailed statistics for both datasets. The data distribution across languages is notably imbalanced, reflecting realistic scenarios where some languages have more resources than others. For MINION, English dominates with 39.76% of training data, followed by Polish and Turkish (13.7% each). For SMiLER, Italian has the largest share at 19.71%, followed by French (16.25%) and German (13.75%). This distribution reflects the diversity of actual multilingual data annotation processes for practical tasks. Notably, several languages belong to different resource categories between the two tasks: English is high-resource in

MINION but mid-resource in SMiLER, while Korean is mid-resource in MINION but low-resource in SMiLER.

4.3 Linguistic Relations and Distances

4.3.1 Linguistic Features from URIEL.

The URIEL Typological Database (Littell et al., 2017) is a large database of linguistic features covering thousands of languages. URIEL combines information from authoritative linguistic resources including the World Atlas of Language Structures (WALS) (Dryer & Haspelmath, 2013), Ethnologue (M. P. Lewis, 2009), Glottolog (Hammarström, Forkel, Haspelmath, & Bank, 2022), and PHOIBLE (Moran, McCloy, & Wright, 2014). The database represents each language through multiple feature vectors capturing different dimensions of linguistic structure, enabling systematic comparison across languages. For our investigation, we extract five types of features that have been shown in prior work to correlate with various aspects of language similarity and cross-lingual transfer (Y.-H. Lin et al., 2019; Ponti, Reichart, Korhonen, & Vulić, 2018).

Phylogeny (fam). The phylogeny feature shows the membership of languages in language families, from the world language family tree in Glottolog (Hammarström et al., 2022). Languages are organized into a hierarchical taxonomy based on their historical relationships and common ancestry. For example, English, German, and Dutch are all members of the Germanic branch of the Indo-European family, while Mandarin Chinese, Cantonese, and Wu are members of the Sino-Tibetan family. The phylogenetic feature vector for each language is binary, with each dimension corresponding to a node in the language family tree. A value of 1 in dimension i indicates that the language belongs to the family or subfamily corresponding to node i, while 0 indicates it does not. This

representation captures both broad family membership (e.g., Indo-European) and fine-grained subfamily relationships (e.g., West Germanic).

Geography (geo). The geography feature shows the location where a language is primarily spoken, based on coordinate information from Glottolog. For each language, URIEL provides latitude and longitude coordinates showing the language's geographic center. Unlike the other features which are represented as binary vectors, geographic distance between two languages is computed as the orthodromic (great circle) distance between their coordinate points on the Earth's surface. This captures the intuition that geographically close languages may share features due to contact effects and areal diffusion, even if they are not phylogenetically related. For instance, languages in the Balkans (Greek, Albanian, Bulgarian, Romanian) share certain structural features despite belonging to different language families, a phenomenon known as a Sprachbund or linguistic area.

Syntax (syn). The syntax feature shows the grammatical structures and word order patterns of languages, from features documented in WALS and Ethnologue. These features include word order parameters (subject-verb-object vs. subject-object-verb vs. other orders), the presence or absence of case marking systems, head-directionality (head-initial vs. head-final constructions), the use of prepositions versus postpositions, and numerous other syntactic properties. The syntax feature vector is binary and high-dimensional, with each dimension representing a particular syntactic property. For example, one dimension might indicate whether the language has a nominative-accusative case alignment, another whether it allows verb-final word order, and so on. Languages with similar syntax

feature vectors tend to have similar grammatical structures, even if they are not historically related.

Phonology (phon). The phonology feature shows the sound patterns and phonological processes in each language, extracted from WALS and Ethnologue. These features include properties such as the presence of tonal distinctions (as in Mandarin Chinese or Vietnamese), the inventory of consonant types (e.g., presence of ejective or click consonants), vowel system characteristics (number of vowels, presence of vowel harmony), syllable structure constraints (e.g., whether consonant clusters are allowed), and stress patterns. Like syntax features, phonology features are represented as a high-dimensional binary vector. While phonological features might seem less directly relevant to information extraction tasks than syntactic features, prior work has suggested that phonological similarity can affect multilingual model performance (Chaudhary, Gonen, Tsvetkov, & Smith, 2020), potentially through effects on subword tokenization and embedding learning.

Inventory (inv). The inventory feature shows the phonetic inventory of each language, the set of phonemes (distinctive speech sounds) that the language uses contrastively. This information is from PHOIBLE's phonetic inventories (Moran et al., 2014), which documents the segmental phonemes for thousands of languages. Each dimension of the inventory feature vector corresponds to a particular phoneme from the International Phonetic Alphabet (IPA), with a value of 1 indicating that the language uses this phoneme contrastively and 0 indicating it does not. For example, English has the phoneme /th/ (as in "think") while most other languages do not; Japanese uses a single liquid phoneme /r/ where English distinguishes /l/ and /r/. The inventory feature differs from phonology in that inventory focuses on

the specific set of sounds used, while phonology captures more abstract patterns and processes involving those sounds.

4.3.2 Distance Metrics. To measure the linguistic distance between pairs of languages based on their feature vectors, we must choose appropriate distance or similarity metrics. Previous work on cross-lingual transfer has used cosine similarity or Euclidean distance to compare language feature vectors (Y.-H. Lin et al., 2019). However, for binary feature vectors, numerous specialized binary similarity measures exist, each with different mathematical properties and emphasizing different aspects of similarity (Choi, Cha, & Tappert, 2010). These binary metrics are distinguished by how they treat matches (both languages have the feature), mismatches (only one language has the feature), and negative matches (neither language has the feature). Based on the categorization in Choi et al. (2010) which surveys over 76 binary similarity measures, we focus on the following 4 representative distances: Hamming, Jaccard, Inner-product, and Anderberg.

Hamming Distance. The Hamming distance between two binary vectors counts the positions where the vectors differ. For two languages $\mathcal{L}_i$ and $\mathcal{L}_j$ with feature vectors $\mathcal{F}(\mathcal{L}_i), \mathcal{F}(\mathcal{L}_j) \in \{0,1\}^d$, the Hamming distance is:

$$d_{\text{hamming}}(\mathcal{L}_i, \mathcal{L}_j) = \sum_{k=1}^{d} \mathbb{1}[\mathcal{F}(\mathcal{L}_i)_k \neq \mathcal{F}(\mathcal{L}_j)_k] \tag{4.1}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Hamming distance treats all mismatches equally and is symmetric: $d_{\text{hamming}}(\mathcal{L}_i, \mathcal{L}_j) = d_{\text{hamming}}(\mathcal{L}_j, \mathcal{L}_i)$. This metric is intuitive and simple but does not account for the total number of features present in each language.

Jaccard Distance. The Jaccard distance measures the overlap of features that are present (value 1) in both languages, normalized by the union of features present in either language. It is defined as one minus the Jaccard similarity coefficient:

$$d_{\text{jaccard}}(\mathcal{L}_i, \mathcal{L}_j) = 1 - \frac{|\{k : \mathcal{F}(\mathcal{L}_i)_k = 1 \wedge \mathcal{F}(\mathcal{L}_j)_k = 1\}|}{|\{k : \mathcal{F}(\mathcal{L}_i)_k = 1 \vee \mathcal{F}(\mathcal{L}_j)_k = 1\}|} \tag{4.2}$$

The Jaccard distance emphasizes shared positive features and is commonly used in tasks where presence of features is more informative than absence. It is symmetric and ranges from 0 (identical sets of features) to 1 (no shared features).

Inner Product Distance. The inner product (dot product) of binary vectors counts features that are present in both languages. We convert this to a distance by normalizing and subtracting from 1:

$$d_{\text{inner}}(\mathcal{L}_i, \mathcal{L}_j) = 1 - \frac{|\{k : \mathcal{F}(\mathcal{L}_i)_k = 1 \wedge \mathcal{F}(\mathcal{L}_j)_k = 1\}| + |\{k : \mathcal{F}(\mathcal{L}_i)_k = 0 \wedge \mathcal{F}(\mathcal{L}_j)_k = 0\}|}{n} \tag{4.3}$$

where the normalization is by the total number of features n. This metric is also symmetric and emphasizes languages that share many positive features.

Anderberg Distance. The Anderberg distance is based on a predictive association measure that captures how much knowing one language's feature values improves prediction of the other, beyond what marginal frequencies alone provide. It is defined as:

$$d_{\text{ander}}(\mathcal{L}_i, \mathcal{L}_j) = 1 - \frac{\sigma - \sigma'}{2n} \tag{4.4}$$

where

$$\sigma = \max\big(|\{k : \mathcal{F}(\mathcal{L}_i)_k{=}1 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}1\}|,\ |\{k : \mathcal{F}(\mathcal{L}_i)_k{=}1 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}0\}|\big)$$
$$+ \max\big(|\{k : \mathcal{F}(\mathcal{L}_i)_k{=}0 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}1\}|,\ |\{k : \mathcal{F}(\mathcal{L}_i)_k{=}0 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}0\}|\big)$$
$$+ \max\big(|\{k : \mathcal{F}(\mathcal{L}_i)_k{=}1 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}1\}|,\ |\{k : \mathcal{F}(\mathcal{L}_i)_k{=}0 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}1\}|\big)$$
$$+ \max\big(|\{k : \mathcal{F}(\mathcal{L}_i)_k{=}1 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}0\}|,\ |\{k : \mathcal{F}(\mathcal{L}_i)_k{=}0 \wedge \mathcal{F}(\mathcal{L}_j)_k{=}0\}|\big) \quad (4.5)$$

$$\sigma' = \max\big(|\{k : \mathcal{F}(\mathcal{L}_j)_k{=}1\}|,\ |\{k : \mathcal{F}(\mathcal{L}_j)_k{=}0\}|\big)$$
$$+ \max\big(|\{k : \mathcal{F}(\mathcal{L}_i)_k{=}1\}|,\ |\{k : \mathcal{F}(\mathcal{L}_i)_k{=}0\}|\big) \quad (4.6)$$

Here, σ measures the optimal prediction accuracy when the joint feature relationship is known, while σ' measures prediction accuracy from marginal frequencies alone. The difference $\sigma - \sigma'$ quantifies the predictive gain from knowing the cross-lingual feature correspondence. The Anderberg distance has the unusual property that $d_{\text{ander}}(\mathcal{L}_i, \mathcal{L}_i) > 0$ (non-zero self-distance), which captures the intuition that even training and testing on the same language involves some generalization challenge.

Using the five feature types with these distance metrics gives a total of 14 distinct linguistic distance measures:

- **Phylogeny (1 distance).** d^{fam} – distance on phylogenetic features
- **Geography (1 distance).** d^{geo} – orthodromic distance between coordinates
- **Syntax (4 distances).** $d^{\text{syn}}_{\text{hamming}}$, $d^{\text{syn}}_{\text{jaccard}}$, $d^{\text{syn}}_{\text{inner}}$, $d^{\text{syn}}_{\text{ander}}$
- **Phonology (4 distances).** $d^{\text{phon}}_{\text{hamming}}$, $d^{\text{phon}}_{\text{jaccard}}$, $d^{\text{phon}}_{\text{inner}}$, $d^{\text{phon}}_{\text{ander}}$
- **Inventory (4 distances).** $d^{\text{inv}}_{\text{hamming}}$, $d^{\text{inv}}_{\text{jaccard}}$, $d^{\text{inv}}_{\text{inner}}$, $d^{\text{inv}}_{\text{ander}}$

This complete set of distances enables us to investigate which aspects of linguistic similarity are most predictive of cross-lingual transfer performance and whether different tasks or model scales are sensitive to different feature types.

4.3.3 Distance Correlation Analysis. Before studying how these linguistic distances correlate with cross-lingual transfer performance (the focus of Section 4.4), we first analyze the relationships among the distance metrics themselves. Figure 6 presents the pairwise Pearson correlation coefficients for all 14 linguistic distances computed across the 17 languages used in our experiments. This correlation matrix shows large variation in how different metrics and feature types relate to each other, indicating that our choice of multiple distances captures diverse aspects of linguistic relationships.

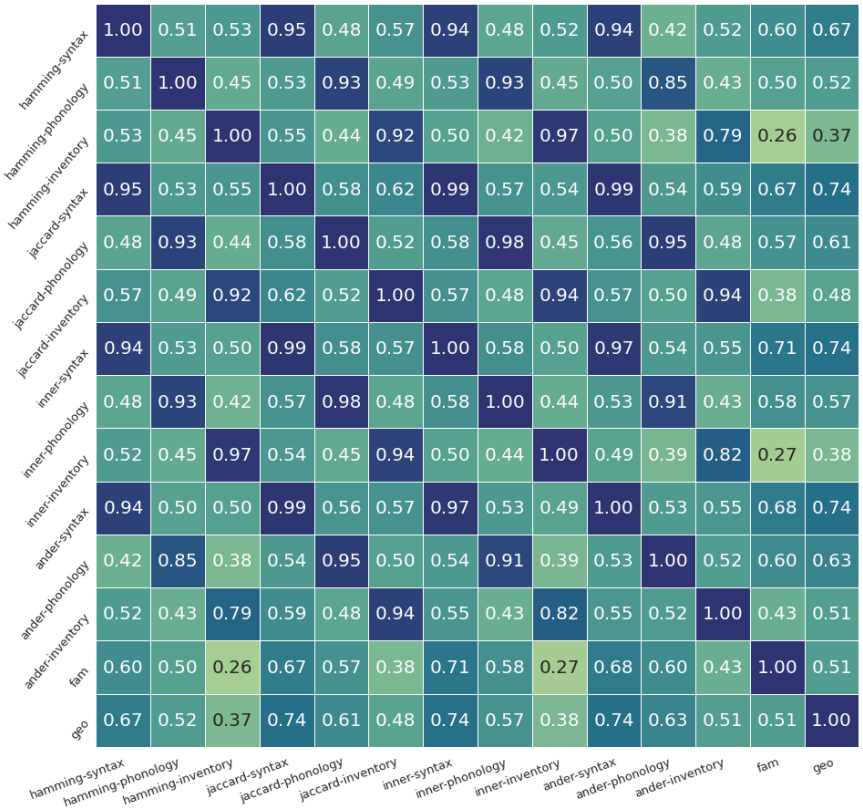


Figure 6. Pairwise Pearson correlation coefficients for all 14 linguistic distances. Each cell (i, j) shows the correlation between distance measure i and distance measure j across all language pairs. Darker colors indicate higher correlation.

We observe several patterns from the correlation analysis. First, aside from phylogeny (fam) and geography (geo), which are computed using standard Euclidean distance, each of the three typological features (syntax, phonology, inventory) is computed using four different binary metrics. Within each feature type, the distances based on different metrics have high correlation (correlation > 0.7), but not perfect correlation, suggesting they capture somewhat different notions of similarity. Second, there are noticeable distinctions between several metrics even within the same feature type, particularly between Anderberg-based distances and the other three metrics (Hamming, Jaccard, Inner product).

The diversity of correlations observed in Figure 6 supports our decision to examine multiple linguistic distance metrics instead of using a single measure. If all distances were highly correlated, using multiple metrics would provide redundant information. Instead, the correlation matrix shows that different metrics capture complementary aspects of linguistic relationships. The analysis in Section 4.4 will reveal which of these diverse distance measures best predicts transfer performance and whether combining multiple distances improves prediction accuracy.

The 17 languages used in this study span diverse phylogenetic groups (Indo-European, Sino-Tibetan, Japonic, Koreanic, Turkic, Semitic, Indo-Iranian) and typological categories, enabling complete evaluation of cross-lingual transfer across linguistic diversity. The detailed linguistic distances for all language pairs are provided in Figure A.19 in Appendix A.3.

4.4 Zero-shot Cross-lingual Single-transfer

To answer the first transfer question, we evaluate ZSCL-S scores for every language pair of each task, in three model scales: small (MiniLM (W. Wang et al., 2020)), base (XLM-RoBERTa-base (Conneau et al., 2020)), and large (XLM-RoBERTa-large). Next, Pearson correlations are computed between the transfer performances and linguistic distances to identify the degree to which the relations in the linguistic landscape determine a model's cross-lingual transfer ability.

Experimental Setup. In ZSCL-S, given a pair of source and target languages, the model is trained using labeled data from the source language. The ZSCL-S score is then defined as the zero-shot evaluation of the trained model on the test set of the target language.

Transfer Performance. Detailed transfer scores are provided in Figures A.20 and A.21 in Appendix A.3. While the language-wise order of the transfer scores is maintained across different model sizes, it is not clear, however, if language identities alone are able to determine model cross-lingual transfer ability. This is due to the significant difference between the results of the two tasks. Even more unexpectedly, model transfer scores do not increase linearly with its number of parameters.

4.4.1 Linguistic Correlation.

We determined if any of the linguistic distances defined in Section 4.3 can explain the heterogeneity of resulting transfer performances, across all settings.

Distance-Transfer Correlation. We compute the Pearson correlation between the transfer score and distance vector between each language pair. The detailed results are presented in Figures A.22 and A.23 in Appendix A.3. While there are several distances that achieve a correlation score of over 0.7, effectively predicting the corresponding transfer performances, none of the linguistic relations are highly correlated with the transfer scores for both tasks. In particular, syntax and inventory features have above-average correlation scores for SMiLER, whereas only phonology-based distances are effective for the event detection task.

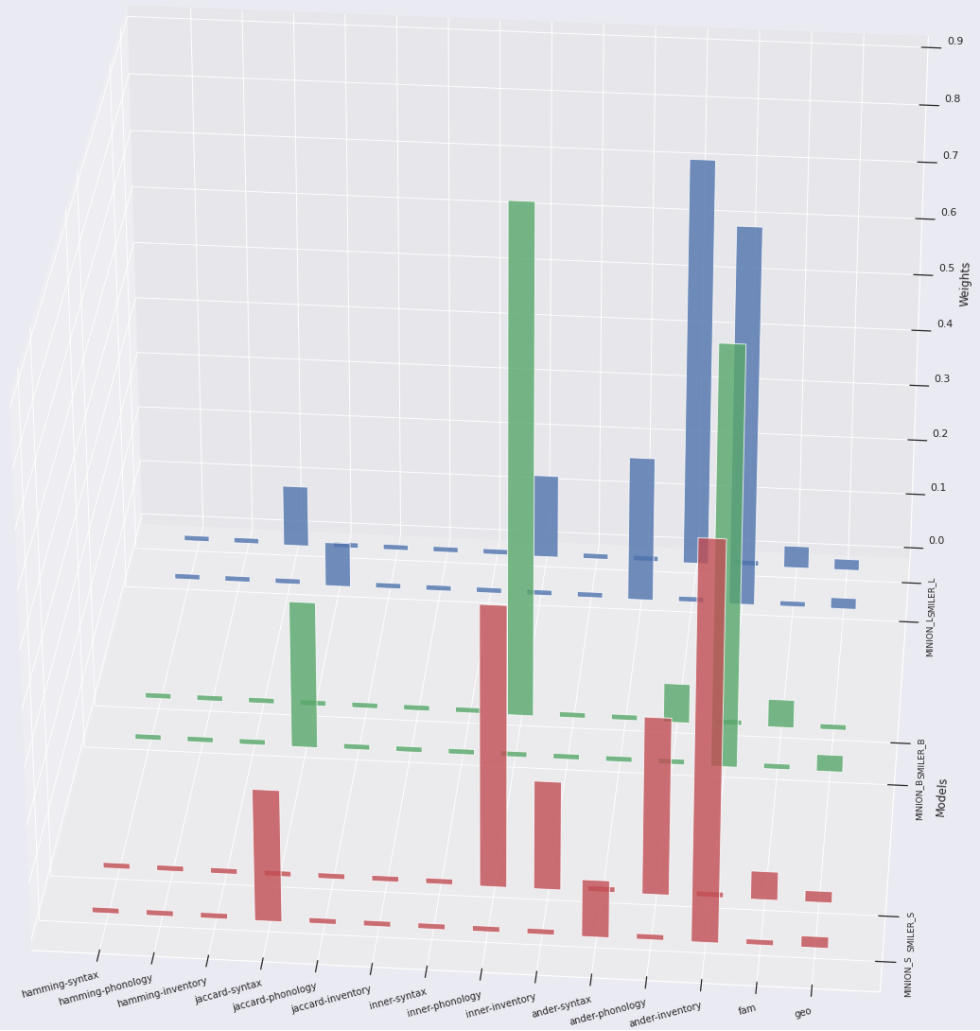


Figure 7. Feature importance weights of the optimally combined metric for each (task, model scale) setting. Small, base, and large models are represented by the colors red, green, and blue, respectively.

Combined Metric. To achieve our objective of creating a universal metric that can be applied across different practical settings, we define a combined metric as a weighted average of all relevant linguistic features. For each task, the optimal weights are the solution of a simple constrained correlation linear maximization (the weights are constrained to be non-negative and sum to 1). We provide the visualization of the resulting importance weights between the two tasks across

model scales in Figure 7. Similar to the above assessment, there is a divergence between MINION and SMiLER on how the linguistic features are weighted in the optimally combined metric.

From these observations, we propose a joint combined metric that involves all three of the typological features as follows:

$$d_{\text{comb}} = 0.4 \cdot d_{\text{ander-syntax}} + 0.2 \cdot d_{\text{inner-phonology}} + 0.4 \cdot d_{\text{ander-inventory}} \tag{4.7}$$

While the component distances and their weights in d_{comb} are heuristically selected based on their importance from Figure 7, we demonstrate the adaptability of the new distance by providing the mean correlation scores (across all languages) of all computed linguistic distances in Figure 8. Not only does d_{comb} achieve the highest correlation with transfer performances overall (above 0.6 for every setting), the combined metric also greatly lessens the score's variability amid tasks and scales of models. This implies that d_{comb} has the potential to be a general metric to approximate ZSCL performances prior to model training. The following sections will use this combined distance for guiding the language selection and adversarial training in multi-transfer setting.

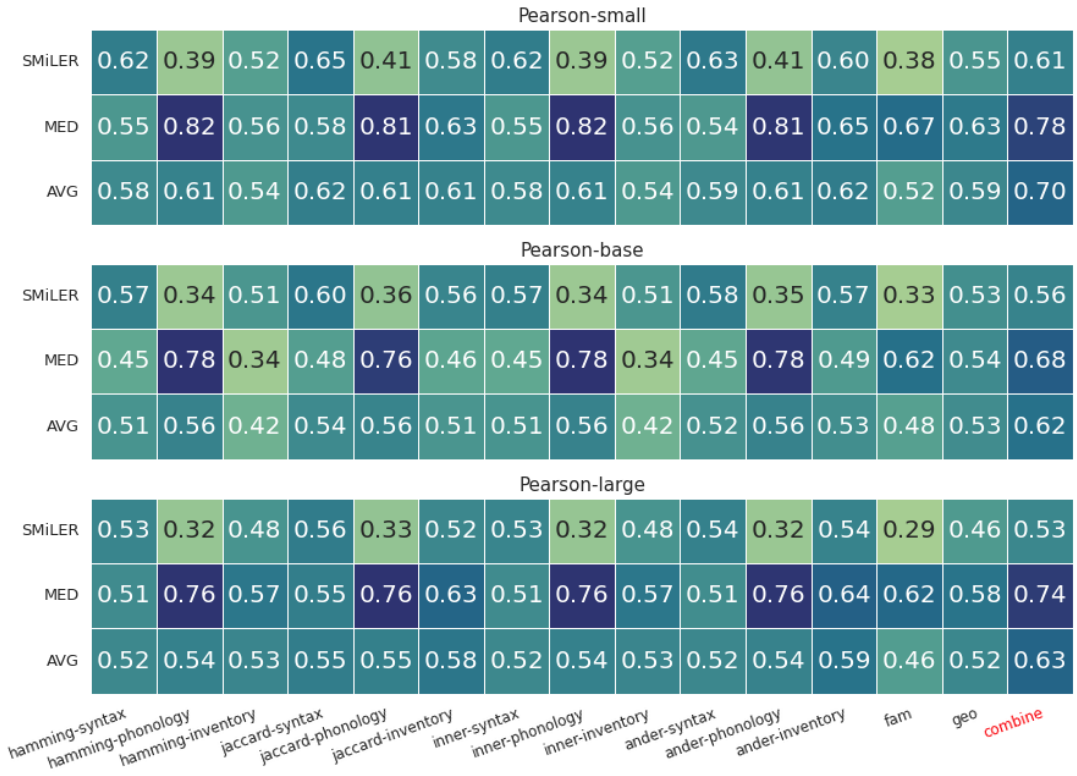


Figure 8. Language-based average Pearson correlation scores for all linguistic distances including the combined metric. The combined metric (rightmost column in each task group) achieves the highest and most stable correlation across all settings. Small, base, and large models are shown in different colors.

4.5 Zero-shot Cross-lingual Multi-transfer

We address Transfer Question 2 by evaluating multi-transfer performances between two sets of languages. In particular, we define a transfer configuration as an experimental setting that specifies languages inside the source and target sets, and a transfer run as an actual experimental evaluation of a transfer configuration. As the number of configurations is exponential in terms of the number of languages, it is computationally impossible to evaluate every configuration, even more so for different tasks and model scales. Therefore, we focus solely on the resource-constrained scenario, which is also equivalent to the setting with the minimum number of source languages. Based on the result from the previous section, we

propose to limit the configuration scope using the general combined distance metric as follows.

4.5.1 Language Selection. In the resource-constrained setting, our goal is to select the minimal set of source languages $\mathcal{D}_s$ that can maximally transfer to a given target language set $\mathcal{D}_t$. We further restrict our attention to transfer configurations with $\mathcal{D}_t$ as a set of closely related languages in terms of cross-lingual transfer. Assuming pairwise transfer is highly correlated with multi-transfer, these configurations can be identified by clustering languages based on the combined linguistic distance d_{comb}.

Language Clustering. We use k-means (Lloyd, 1982) to partition the available languages into clusters based on d_{comb}. In particular, k-medoids[1], a variant of k-means, is used for both clustering and finding the medoids (the actual data point/language that is the center of each cluster). These medoid languages should be the optimal language that transfers best to every member of its cluster. The resulting clusterings are shown in Figure 9 for both MINION (Figure 9c) and SMiLER (Figure 9a). The minimal number of source languages ($|\mathcal{D}_s|$) is chosen to be equal to the minimal number of clusters such that each cluster has at least 3 members (so we can meaningfully evaluate multi-transfer setting).

[1] `https://en.wikipedia.org/wiki/K-medoids`

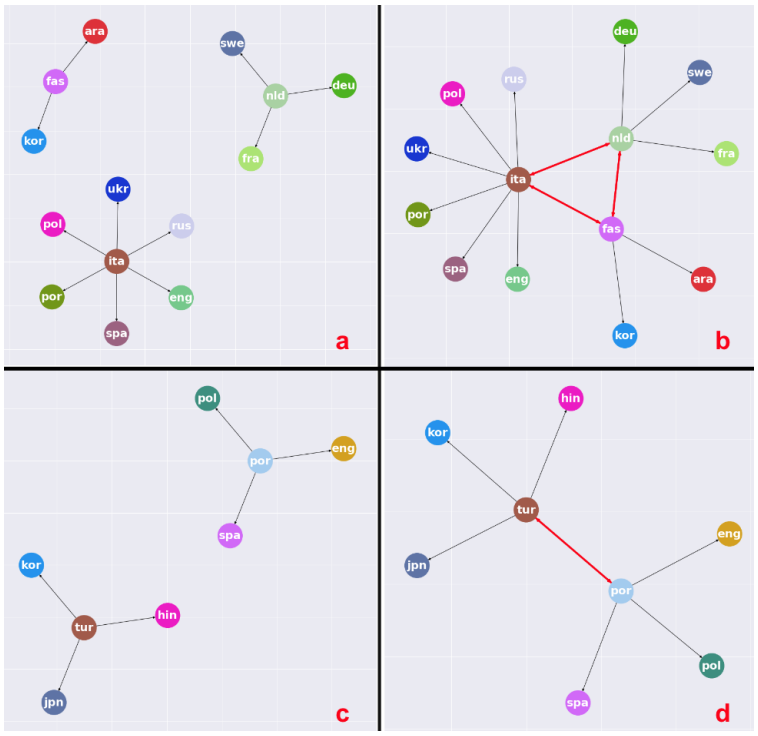


Figure 9. Language clustering results for languages in SMiLER (a and b) and MINION (c and d). The graphs on the right have connected medoids indicated by the new red edges.

4.5.2 Experimental Setup. In ZSCL-M, source training set $\mathcal{D}_s$ consists of N_s labeled datasets, and the goal is to transfer to a target cluster $\mathcal{D}_t$. For the MINION task, $N_s = 2$ and $\mathcal{D}_t$ can be the Turkish cluster tur*, or the Portuguese cluster por*. For SMiLER, $N_s = 3$ and $\mathcal{D}_t$ can be the Italian cluster ita*, the Dutch cluster nld*, or the Farsi cluster fas*. We are interested in verifying the following two hypotheses:

(1) Inter-Cluster Transfer: The best set of languages that transfer best across every cluster is the set of every medoid language.

(2) Intra-Cluster Transfer: The best set of languages that transfer best for a given cluster is a subset of that cluster.

4.5.3 Transfer Performance. Detailed results of multi-transfer performances are provided in Figures A.24 and A.25 in Appendix A.3, from which we can observe a clear improvement over the single-transfer setting owing to the additional training data. Our main interest here is how effective d_{comb} is in guiding the language selection for ZSCL-M. Table 13 shows the differences in transfer scores of Inter-cluster (medoids*) and Intra-cluster ([medoid_lang]*) configurations over Random configuration. Specifically, medoids* measures inter-cluster transfer capacity of the set $\mathcal{D}_s$ consisting of every medoid from each cluster, to the target set $\mathcal{D}_t$ of all considered languages. On the other hand, [medoid_lang]* measures intra-cluster performance of a randomly sampled set $\mathcal{D}_s$ of N_s languages from the corresponding cluster, to the target set $\mathcal{D}_t$ of every language in that cluster. Finally, Random configurations are sampled from the set of configurations that are not part of the above two configurations. Except for the Inter-cluster configuration which only has one option, the results of Intra-cluster and Random configurations are the average of their sampled transfer runs.

		SMALL	BASE	LARGE	MODEL_AVG
M	medoids*	1.8	1.7	0.7	1.4
	tur*	2.7	1.0	1.0	1.5
	por*	6.0	6.1	5.8	6.0
	task_avg	3.5	2.9	2.5	3.0
S	medoids*	2.1	1.4	1.9	1.8
	ita*	3.9	3.7	3.1	3.5
	nld*	9.2	8.6	6.7	8.2
	fas*	1.8	1.7	0.7	1.4
	task_avg	4.2	3.8	3.1	3.7

Table 13. Differences in ZSCL-M scores (F1) of Inter-cluster (*medoids**) and Intra-cluster (*[medoid_lang]**) over Random configurations, for MINION (M) and SMiLER (S).

The results from Table 13 show that language selection based on d_{comb} provides a considerable boost in multi-transfer scores for every (task, configuration, model scale) setting. This suggests that these medoid languages have the potential to achieve optimal cost-performance trade-offs in multi-transfer setting. Notably between the two tasks, SMiLER has higher performance increases with only one additional source language, despite having almost double the number of languages in the target set. This implies that as the number of languages grows considerably larger, there may exist a magnitude smaller set of optimal universal languages (medoids) that are able to transfer to every language extremely well.

4.6 Zero-shot Cross-lingual Relational-transfer

Due to limited access to multilingual annotators, gathering labeled data across languages is difficult. The previous section addresses this by careful language/data selection to optimize cost-effectiveness. In contrast, unlabeled data is easy to collect, but leveraging it correctly for ZSCL is non-trivial. This section investigates the effectiveness of adversarial training approaches for the more general ZSCL-M setting, and the possibility of further improving multilingual transfer through explicitly integrating our transfer-correlated linguistic relations.

4.6.1 Experimental Setup. We follow the same setup as in ZSCL-M, but each language is accompanied by an unlabeled dataset which can be used for training. The model uses labeled data from the source cluster to learn the task, whereas unlabeled data from another cluster is used to help transfer source performance to that target cluster. The goal is to bridge the performance between two different language clusters with the aid of given unlabeled text.

Adversarial Language Adaptation. A typical method used for ZSCL-S is adversarial language adaptation (ALA) (X. Chen, Sun, Athiwaratkun, Cardie,

& Weinberger, 2018; M. V. Nguyen, Lai, & Nguyen, 2021b) which employs a language discriminator that takes an encoded representation from mLMs as its input and predicts its origin (language). By pushing the encoder to both minimize the downstream loss and maximally misdirect the language predictor (adversarial training), the resulting representation can be indiscriminate with respect to the shift between the languages while also discriminative for the main learning task. Applying ALA for ZSCL-M setting is equivalent to applying DANN for a single joint source domain to a single joint target domain (the union of all languages in $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively).

Zero-shot Cross-lingual Relational-transfer. We extend ALA to the case of multiple source and target languages through Graph-relational domain adaptation (GrDA) (Z. Xu et al., 2022), a generalization of DANN (Ganin et al., 2016) to multi-domain adaptation setting by introducing a domain graph that captures heterogeneous relations among domains. GrDA relaxes the strict uniform alignment of DANN to allow flexible and effective adaptation between distant domains. We use the language clustering graphs on the right of Figure 9 as domain graphs for GrDA. Additional direct connections between medoid languages are introduced (red edges) to facilitate inter-cluster transfer. We refer to this adversarial learning process for ZSCL that directly embeds the linguistic relations into the representations as Zero-shot Cross-lingual Relational-transfer (ZSCL-R).

4.6.2 Transfer Performance. Performance of the baseline DANN and the proposed adversarial training method ZSCL-R is provided in Table 14, which follows the same format as Table 13. However, instead of comparing against Random configurations, they are compared directly with results of ZSCL-M runs of the corresponding configurations, but only in terms of inter-cluster transfer. The

negative results of DANN confirm that strictly aligning language representations uniformly is not effective in ZSCL-M setting. As the model scale gets smaller, the model's representation becomes less expressive, whereas the language-invariant feature of all languages is harder to capture as the number of languages grows. Thus, the adverse effect gets significantly worse on small models for the SMiLER task (-12 points on average). In contrast, ZSCL-R provides consistent improvements over ZSCL-M for most configurations. Due to the flexibility of GrDA alignment, ZSCL-R performs even better as the number of languages increases, effectively leveraging the additional unlabeled data to help improve inter-cluster transfer ability of models.

		SMALL		BASE		LARGE		MODEL_AVG	
		ZSCL-R	DANN	ZSCL-R	DANN	ZSCL-R	DANN	ZSCL-R	DANN
M	medoid*	0.7	-3.7	1.4	-1.9	1.3	-2.8	1.1	-2.8
	tur*	4.1	-0.5	2.7	-2.7	3.6	-0.1	3.5	-1.1
	por*	0.6	-5.5	-0.3	-4.2	-0.3	-5.8	0.0	-5.2
	task_avg	1.8	-3.2	1.3	-2.9	1.5	-2.9	1.5	-3.0
S	medoid*	1.8	-11.8	3.2	-4.2	0.6	-1.7	1.9	-5.9
	ita*	3.0	-15.1	3.9	-5.5	2.3	-5.8	3.1	-8.8
	nld*	2.4	-10.9	2.8	-4.9	0.0	-5.1	1.7	-7.0
	fas*	-0.2	-10.6	2.1	-4.2	2.4	-2.7	1.4	-5.8
	task_avg	1.8	-12.1	3.0	-4.7	1.3	-3.8	2.0	-6.9

Table 14. Difference between transfer performance of ZSCL-R methods and ZSCL-M runs in inter-cluster setting.

4.7 Discussion

Language Clustering Results. Previous studies were limited to single-transfer and typically grouped linguistically similar languages together. In contrast, our approach to clustering produces several unexpected clusters that were specifically structured to address the many-to-many aspect of multi-transfer. For instance, the group of languages [kor, hin, jpn] would typically form a non-Latin-based cluster, but instead in Figure 9, they are centered around the Latin-

based tur. This structure enables the medoid languages to serve as a gateway to facilitate inter-cluster transfer. Future research may propose more intricate structures incorporated into the language relation graphs and explore their roles in larger-scale settings.

Language Scaling. One primary challenge in scaling to hundreds of languages involves the escalating costs and time for annotations as more low-resource languages are introduced. This inhibits the adaptation of performance-based approaches due to evaluation data for these languages being largely unavailable. Conversely, linguistic features have been mapped out for thousands of languages. Additionally, our observation at the end of Section 4.5.3 bears similarity to that in Pelloni, Shaitarova, and Samardzic (2022), where it is suggested that an optimally smaller set of universal languages may exist that could facilitate exceptional cross-lingual transfer capabilities. Future endeavors aimed at identifying these universal languages could have a substantial impact on both research (more comprehensive benchmarks) and the industry (higher performance), at a significantly reduced cost.

4.8 Related Work

4.8.1 Zero-shot Cross-lingual Transfer.

The majority of recent zero-shot cross-lingual transfer work follows the single-transfer setting, using large-scale multilingual multi-task benchmarks to evaluate pretrained multilingual language models (Chi et al., 2021; Fang, Wang, Gan, Sun, & Liu, 2021). Two prominent benchmark suites have driven this research: XTREME (Ruder et al., 2021) and XGLUE (Liang et al., 2020). These datasets provide English training data for fine-tuning pretrained multilingual models such as multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), which

are then evaluated on translated or parallel test sets in different languages. The benchmarks cover a diverse range of tasks including classification, structured prediction, question answering, and retrieval, allowing systematic evaluation of cross-lingual transfer capabilities.

Because of this, English has become the dominant source language for transfer in most zero-shot cross-lingual research (Phang et al., 2020). However, this English-centric focus is suboptimal in multiple studies. Keung et al. (2020) discovered that a model's English development accuracy does not correlate strongly with its performance in other languages, suggesting that English-specific optimization may not generalize well. Lauscher et al. (2020) showed the limitation of using English exclusively to transfer to low-resource languages, particularly those from different language families or with distinct typological features. Furthermore, Turc et al. (2021) found that other high-resource languages such as German and Russian often transfer more effectively than English when the set of target languages is diverse or unknown a priori, challenging the assumption that English should always be the default source language.

The foundation for zero-shot cross-lingual transfer lies in massively multilingual pretrained language models. Multilingual BERT (Devlin et al., 2019) was among the first to show that a single Transformer model pretrained on concatenated Wikipedia text from over 100 languages could achieve reasonable zero-shot transfer performance. Later models have increased both the training data and model capacity: XLM (Lample & Conneau, 2019) introduced translation language modeling and parallel data for improved cross-lingual alignment, XLM-RoBERTa (Conneau et al., 2020) scaled to 2.5TB of CommonCrawl data covering 100 languages, and mT5 (Xue et al., 2021) extended the T5 architecture to 101

languages with a multilingual C4 corpus. These models learn shared cross-lingual representations through the bottleneck of shared vocabulary and model parameters, allowing knowledge transfer across languages without explicit parallel supervision.

4.8.2 Linguistic Diversity and Transfer Performance.

Understanding what linguistic information is captured in multilingual language models has been studied. By probing the learned representations of mLMs, several studies have found that syntactic information is implicitly encoded in the layers of multilingual models (Limisiewicz, Mareček, & Rosa, 2020; Pires, Schlinger, & Garrette, 2019). Pires et al. (2019) showed that mBERT learns effective cross-lingual representations despite having no explicit cross-lingual training objective, with middle layers capturing syntactic structures that generalize across languages. Limisiewicz et al. (2020) further showed that typologically similar languages share more similar syntactic representations in deeper layers, suggesting that the model learns to organize languages based on structural similarities.

N. Xu et al. (2022) studied how the pre-training and fine-tuning processes change linguistic features encoded in multilingual models and how these changes directly impact multilingual performance. Their analysis revealed that fine-tuning on a source language can strengthen certain linguistic features while weakening others, with the direction of change depending on the specific task and the linguistic properties of the source language. This finding has important implications for understanding why certain source languages transfer better than others: effective transfer may depend not just on static linguistic similarity, but on how fine-tuning changes the model's linguistic representations.

Further investigation by Y.-H. Lin et al. (2019) showed that distances based on linguistic features, including phylogenetic (language family) and typological

properties (grammatical structures), between two languages are correlated with cross-lingual transfer capacity. They showed that syntactic similarity, in particular, is a good predictor of transfer performance for syntactically-oriented tasks like part-of-speech tagging and dependency parsing. These linguistic features, available in resources such as the URIEL Typological Database (Littell et al., 2017) and the World Atlas of Language Structures (WALS) (Dryer & Haspelmath, 2013), can be used to compute pairwise distances between languages that capture multiple dimensions of linguistic similarity.

The correlation between linguistic features and transfer performance has motivated approaches to use this information for improving multilingual systems. These features can be used to guide parameter sharing among languages (Ammar, Mulcaire, Ballesteros, Dyer, & Smith, 2016), where languages with similar typological properties share more parameters in a multilingual model. They can also inform data selection strategies (Ponti et al., 2018), helping decide which languages to include as auxiliary tasks or which parallel corpora to use for training. For instance, Ponti et al. (2018) showed that selecting typologically similar auxiliary languages for multi-task learning leads to better performance than random selection.

Several recent works aim to predict cross-lingual transfer performance directly from linguistic features, without training models (Y.-H. Lin et al., 2019; Srinivasan et al., 2021). In these approaches, a regression model is trained to take linguistic distances of source-target language pairs as input and predict the expected transfer performance. Y.-H. Lin et al. (2019) used genetic and typological distances to predict performance on syntactic tasks, while Srinivasan et al. (2021) extended this to semantic tasks and added features such as word overlap and script

similarity. Dolicki and Spanakis (2021) analyzed the impact of specific linguistic features on cross-lingual transfer, finding that different feature categories contribute differently depending on the task.

While these performance prediction approaches achieve high accuracy on their specific experimental settings, they have limitations. First, the regression models are typically trained for particular model architectures and tasks, and their predictions may not generalize to different settings. As training configurations, model scales, and fine-tuning procedures vary widely in practice, relying on architecture-specific regression models can be unreliable. Second, previous work has been limited to predicting single-transfer performance between pairs of languages. This focus on one-to-one transfer does not address the more general and practical question of how to optimally combine multiple source languages, nor does it provide guidance for scenarios where multiple target languages are of interest simultaneously.

4.8.3 Multi-source Transfer Learning. While the majority of cross-lingual transfer research has focused on single-source scenarios, several studies have showed that using multiple source languages can improve transfer performance. Turc et al. (2021) showed that when the set of target languages is diverse or unknown a priori, combining multiple high-resource source languages often yields better average performance than selecting a single optimal source language. Their experiments across various NLP tasks showed that multi-source training is particularly beneficial when target languages are typologically diverse, as different source languages may provide complementary information that helps the model generalize better.

The question of which source languages to select for multi-source training has been studied. Ponti et al. (2018) studied source language selection for multi-task learning, showing that typological similarity is a useful criterion for choosing auxiliary languages. Ammar et al. (2016) investigated parameter sharing strategies across languages in multilingual models, using phylogenetic relationships to guide which languages share parameters. However, these approaches typically consider pairwise relationships between a target language and potential source languages, rather than addressing the combinatorial problem of selecting an optimal subset of source languages from a larger pool.

4.8.4 Adversarial Language Adaptation. Adversarial training for learning domain-invariant representations was introduced by Ganin et al. (2016) in the context of domain adaptation. Their Domain-Adversarial Neural Network (DANN) architecture introduced a gradient reversal layer that allows a model to learn features that are discriminative for the main task while being invariant to the domain shift. The key insight is that by training a domain discriminator to identify which domain an example comes from, and simultaneously training the feature extractor to fool this discriminator, the model learns representations that do not contain domain-specific information while retaining task-relevant features.

The DANN framework has been adapted to cross-lingual transfer, where the goal is to learn language-invariant representations. Several works have adopted adversarial language adaptation (ALA) for different cross-lingual NLP tasks. X. Chen et al. (2018) applied adversarial training to cross-lingual sentiment analysis, finding that language-adversarial training helps align feature distributions across languages. For information extraction specifically, M. V. Nguyen et al. (2021b) used adversarial training for cross-lingual event detection, and Ngo, Phung,

and Nguyen (2021b) applied it to joint information extraction across languages. Huang, Ji, and May (2019) used adversarial training in cross-lingual named entity recognition, showing improvements over standard fine-tuning approaches.

However, standard adversarial language adaptation has a limitation: it enforces strict uniform alignment across all languages by using a single discriminator that treats all language pairs identically. This uniform alignment assumption may be restrictive when dealing with multiple typologically diverse languages. Languages that are linguistically distant may not benefit from being forced into the same representation space as closely related languages. For example, enforcing the same degree of alignment between English-French (closely related) and English-Japanese (linguistically distant) may be suboptimal. The uniform approach cannot capture the varying degrees of similarity and the heterogeneous relationships that exist among languages in a multilingual setting.

To address the limitation of uniform alignment in multi-domain scenarios, Z. Xu et al. (2022) introduced Graph-Relational Domain Adaptation (GrDA), which generalizes DANN to handle multiple domains with heterogeneous relationships. Instead of enforcing uniform alignment across all domains, GrDA constructs a domain relation graph where nodes represent domains and edges encode the strength of relationships between domains. The framework uses multiple discriminators and flexible alignment constraints based on the graph structure, letting strongly related domains be aligned more closely while maintaining greater separation between distant domains. While originally developed for visual domain adaptation, the GrDA framework provides an ideal foundation for cross-lingual transfer with multiple languages.

Although GrDA provides a flexible framework for multi-domain adaptation, it has not been used for cross-lingual transfer learning, and has not been used with linguistic distance metrics to guide the construction of language relation graphs. Our work is the first to adapt graph-relational adversarial learning to the cross-lingual setting and to use linguistic features to explicitly define the relationships between languages in the graph structure. By including linguistic distances into the adversarial training process, we enable the model to learn representations that respect the varying degrees of similarity across languages, leading to more effective transfer.

4.9 Conclusions and Future Work

This work demonstrated that linguistic typological features provide reliable, dataset-independent guidance for zero-shot cross-lingual transfer in information extraction. Through comprehensive experiments across 17 languages, 2 tasks, and 3 model scales, we showed that a task-specific combined distance metric can predict transfer effectiveness with strong correlation, guide practical language selection decisions that outperform random selection by 3-4 F1 points, and enable graph-relational adversarial training that further improves performance.

These findings have practical value for developing multilingual IE systems under resource constraints. Instead of exhaustively evaluating all possible source language combinations - which is prohibitively expensive for even moderate numbers of languages - practitioners can use our combined distance metric to identify promising source languages through efficient clustering and selection. The resulting systems achieve performance comparable to or better than exhaustive search at a fraction of the computational cost.

CHAPTER V

EXPERTISE SPECIALIZATION FOR LARGE LANGUAGE MODELS VIA MEDICAL RAG BENCHMARKING

This chapter presents work currently under review as a conference paper titled "MedRGB: Practical Framework for Benchmarking Medical Retrieval-Augmented Generation Systems." As the first author, Nghia was responsible for problem formulation, model development, experimental design, and manuscript preparation. Linh contributed to the meta-learning framework design and provided technical discussions. Thien provided research guidance, editorial revisions, and advised on experimental approach. The content has been adapted to fit the dissertation format and narrative.

The earlier chapters of this dissertation examined adaptive transfer learning, where models must generalize across domains (Chapters II and III), languages (Chapter IV), and task structures under conditions of limited labeled data. These investigations addressed the basic challenge of data scarcity that affects most languages and specialized domains. However, the emergence of Large Language Models (LLMs) with billions of parameters trained on large-scale data has changed the transfer learning landscape. As discussed in Chapter I, this transformation marks a shift from adaptive transfer learning characterized by horizontal knowledge transfer across contexts to *specialization transfer learning* focused on vertical refinement from general-purpose capabilities to specialized expertise. This chapter addresses Research Question 4 (RQ4) by investigating how general-purpose LLMs can be reliably adapted for high-stakes expert domains where accuracy, reliability, and following professional standards are important.

The work introduces Medical Retrieval-Augmented Generation Benchmark (MedRGB), a complete evaluation framework that measures the capacity of medical question-answering systems in RAG settings across four critical dimensions:(1) *Standard-RAG* for baseline capability with relevant documents, (2) *Sufficiency* for recognizing information inadequacy amid noise, (3) *Integration* for compositional reasoning across multiple sources, and (4) *Robustness* for detecting and correcting factual errors in retrieved context. Extensively evaluates seven state-of-the-art LLMs under multiple RAG conditions, providing detailed experimental analysis that shows their limitations in addressing complex scenarios involving noise, information integration, and misinformation. We analyze the errors and reasoning processes of LLMs to provide insights into failure modes and suggest concrete future directions for developing more trustworthy medical RAG systems.

5.1 Introduction

LLMs have demonstrated remarkable capabilities in solving complex medical problems, achieving state-of-the-art performance across various benchmarks (Nori et al., 2023; Singhal et al., 2023). Commercial systems such as GPT-4, Claude, and Gemini exhibit sophisticated medical reasoning abilities, often matching or exceeding average human performance on medical licensing examinations (Nori et al., 2023). However, ensuring the reliability and trustworthiness of artificial intelligence (AI) medical systems remains a critical challenge, especially in healthcare applications where errors can have serious real-world consequences (Thirunavukarasu et al., 2023). The sensitive nature of the medical domain necessitates not only high average performance, but also consistent accuracy, appropriate uncertainty awareness, and robust behavior under diverse conditions including noisy or adversarial information.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to enhance the factual accuracy of LLMs by integrating external knowledge sources (P. S. H. Lewis et al., 2020). Rather than relying solely on parametric knowledge acquired during pre-training, RAG systems retrieve relevant documents from curated knowledge bases and provide them as context for generation, potentially reducing hallucinations and enabling access to current information beyond the model's training data (Gao et al., 2023). However, incorporating an information retrieval component also presents new complexities that warrant careful evaluation. Consider the example in Figure 10. When an LLM is asked a medical question and provided with retrieved documents, these documents can contain not only useful knowledge that helps determine the correct answer (highlighted in blue), but also noise information that is irrelevant to the question, or even factual errors (highlighted in red) that can mislead the model toward incorrect conclusions. To deploy RAG systems safely in medical applications, we must evaluate not only whether they can leverage relevant information effectively, but also whether they can handle these practical scenarios robustly.

Existing medical benchmarks have made important contributions to evaluating AI systems, but they primarily focus on target accuracy in standard question-answering (QA) settings. Benchmarks such as MedQA (D. Jin et al., 2020a), MMLU-Medical (Hendrycks et al., 2021), PubMedQA (Q. Jin, Dhingra, Liu, Cohen, & Lu, 2019b), and BioASQ (Krithara, Nentidis, Bougiatiotis, & Paliouras, 2023) provide high-quality QA pairs that test medical knowledge comprehensively. Recent efforts such as MedEval (He et al., 2023) have provided large-scale expert-annotated evaluations across various medical tasks and domains.

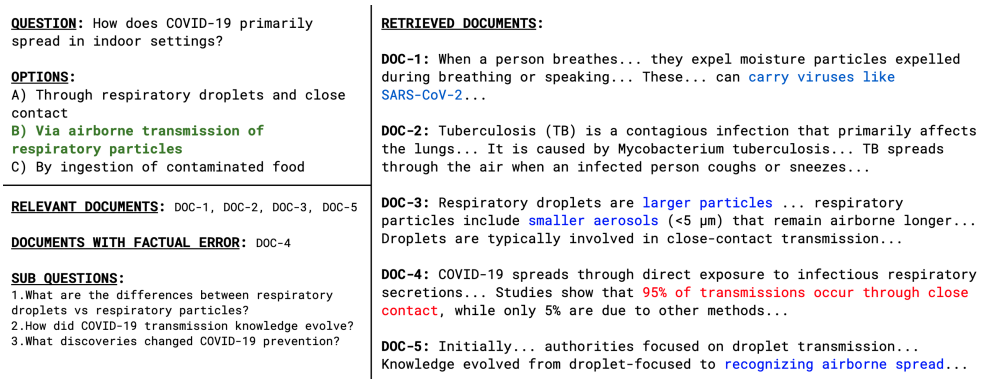


Figure 10. Example of a medical RAG scenario. Blue texts are useful information that should be extracted to help determine the answer. Red texts are factual errors that can mislead the LLMs.

MIRAGE (G. Xiong et al., 2024) has specifically examined RAG systems in medicine, evaluating the effects of retrieval modules on target accuracy across five medical QA datasets. However, these benchmarks focus predominantly on whether models can produce correct answers when provided with relevant information, overlooking many practical scenarios that measure crucial aspects of a reliable medical system. Questions such as "Can the model recognize when retrieved information is insufficient?", "Can it integrate information from multiple documents effectively?", and "Is it resilient to factual errors in retrieved context?" remain largely unaddressed in existing evaluations.

Several recent works have explored RAG evaluation more comprehensively in the general domain (J. Chen, Lin, Han, & Sun, 2024; Es, James, Espinosa-Anke, & Schockaert, 2023). RAGAS (Es et al., 2023) assesses three qualities of RAG outputs for QA tasks: Faithfulness (whether responses align with provided context), Answer Relevance (whether responses address the actual question), and Context Precision-Recall (the quality of retrieved context). RGB (Retrieval-Augmented Generation Benchmark) (J. Chen et al., 2024) establishes a framework to measure

four abilities required for effective RAG: noise robustness, negative rejection, information integration, and counterfactual robustness. While these frameworks provide valuable evaluation methodologies, they are designed for general-domain applications and do not address the specific requirements and constraints of medical applications, where the stakes are considerably higher and the tolerance for errors is substantially lower.

To address this gap, this chapter introduces Medical Retrieval-Augmented Generation Benchmark (MedRGB), a comprehensive evaluation framework that measures the capacity of medical question-answering systems in RAG settings across four critical dimensions. Building upon the RGB framework (J. Chen et al., 2024) and using questions from four medical QA datasets as the foundation, MedRGB evaluates RAG systems in the following practical scenarios: (1) *Standard-RAG* evaluates LLM performance when presented with multiple retrieved relevant documents to create context for answering questions, establishing baseline RAG capabilities; (2) *Sufficiency* evaluates LLM reliability when noise documents are present within retrieved context, requiring models to recognize when information is insufficient and to filter out noisy information from external sources; (3) *Integration* evaluates LLM ability to answer multiple supporting sub-questions and integrate extracted information to address complex main questions that require compositional reasoning; and (4) *Robustness* evaluates LLM resiliency to factual errors in retrieved context, requiring trustworthy systems to detect factually incorrect documents and provide corrected information.

The main contributions of this chapter are: First, we establish MedRGB, a comprehensive benchmark with four test scenarios designed to evaluate LLMs for medical QA tasks in RAG settings across critical practical dimensions. To the

best of our knowledge, this is the first benchmark to assess medical RAG systems systematically across all of these scenarios. Second, we extensively evaluate seven state-of-the-art LLMs under multiple RAG conditions, providing detailed empirical analysis that demonstrates their limitations in addressing complex scenarios involving noise, information integration, and misinformation. Third, we analyze the errors and reasoning processes of LLMs to provide insights into failure modes and suggest concrete future directions for developing more trustworthy medical RAG systems. Finally, to demonstrate potential paths toward more reliable medical RAG systems, we present a case study on long-form question answering that explores two improvement directions.

The remainder of this chapter is organized as follows. Section 5.2 formally defines the problem and establishes the evaluation dimensions addressed by MedRGB. Section 5.3 presents the benchmark construction methodology in detail, including topic generation, document retrieval, and the creation of four test scenarios. Section 5.4 describes our experimental setup and presents comprehensive results across all scenarios, including the case study on long-form question answering. Section 5.5 reviews related work on medical benchmarks, RAG systems, and evaluation frameworks. Finally, Section 5.6 concludes the chapter with a summary of contributions and connections to the dissertation's overall narrative.

5.2 Problem Statement and Background

5.2.1 Medical Question Answering. Medical Question Answering (Medical QA) is the task of producing accurate, evidence-based answers to questions requiring medical knowledge and clinical reasoning. Formally, given an input query q representing a medical question, the goal is to generate an answer $\hat{a}$ that correctly addresses the question. In the multiple-choice setting commonly

used for evaluation, the answer space is constrained to a finite set of options $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, and the task is to select $\hat{a} =_{a_i \in \mathcal{A}} p(a_i|q)$ where $p(a_i|q)$ represents the probability that option a_i is the correct answer given query q. In the free-form setting relevant for long-form question answering, the model generates $\hat{a}$ from an unconstrained vocabulary, requiring both content accuracy and appropriate presentation of medical information.

The medical domain imposes several critical requirements beyond standard QA tasks. First, *accuracy* is critical, as incorrect medical information can lead to harmful clinical decisions or patient behaviors. Second, *reliability* requires consistent performance across diverse questions and patient presentations, not just high average accuracy. Third, *evidence-based reasoning* demands that answers align with current medical knowledge and clinical guidelines rather than unsupported assertions. Fourth, *uncertainty awareness* requires that systems recognize when available information is not enough for confident answers, avoiding overconfident incorrect responses. Finally, *robustness* to noisy or contradictory information is essential, as medical information sources vary widely in quality and reliability. These requirements motivate the complete evaluation framework provided by MedRGB.

5.2.2 Retrieval-Augmented Generation Framework. Retrieval-Augmented Generation extends the standard QA approach by integrating external knowledge sources into the generation process. Formally, a RAG system consists of three main components: a knowledge base $\mathcal{K}$ containing a collection of documents, a retriever $\mathcal{R}$ that identifies relevant information, and a generator (typically an LLM) $\mathcal{M}$ that produces answers conditioned on both the query and retrieved context. Given a query q, the retriever searches the knowledge base to obtain a set

of k relevant passages: $\mathcal{C} = \mathcal{R}(q, \mathcal{K}) = \{c_1, c_2, \ldots, c_k\}$ where each c_i is a text passage potentially containing information relevant to answering q. The generator then produces an answer by conditioning on both the original query and the retrieved context: $\hat{a} = \mathcal{M}(q, \mathcal{C})$.

Retrieval-Augmented Generation (RAG) is highly promising for medical applications because it enables access to current information beyond the LLM's static training data, a necessity in the constantly evolving medical field. Crucially, RAG helps reduce hallucinations by grounding responses in retrieved evidence and offers transparency through source citation, allowing healthcare providers to verify information and assess confidence. It also enables customization to specific institutional knowledge bases, like hospital protocols. However, real-world deployment presents challenges not covered by standard evaluations, such as the need to handle irrelevant or scattered information across multiple documents, the ability to critically evaluate source reliability (as documents may contain errors), and the necessity to recognize information insufficiency when relevant data is missing. These challenges motivate MedRGB's multi-dimensional evaluation approach, which assesses these critical capabilities systematically.

5.2.3 Evaluation Dimensions for Medical RAG. MedRGB evaluates medical RAG systems across four complementary dimensions, each addressing a critical aspect of reliable performance.

Standard-RAG: Basic Retrieval-Augmented Performance. The Standard-RAG dimension evaluates the basic capability of a system to answer questions when provided with relevant retrieved context. Formally, given a query q with ground truth answer a^* and a retrieved context $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ where all passages are relevant to q (i.e., contain information useful for determining a^*), the

system generates an answer $\hat{a} = \mathcal{M}(q, \mathcal{C})$. Performance is measured by comparing $\hat{a}$ to a^* using appropriate metrics (exact match for multiple choice, or semantic similarity scores for free-form answers). This dimension establishes a baseline for RAG effectiveness when provided with clean, relevant context.

Sufficiency: Information Insufficiency Awareness. The Sufficiency dimension evaluates whether a system appropriately recognizes when retrieved information is not enough to answer a query confidently. Formally, we introduce a parameter $p_{\text{sig}} \in [0, 1]$ representing the proportion of signal documents (relevant, useful passages) in the retrieved context. For a given query q with ground truth answer a^*, we construct a retrieved set $\mathcal{C} = \mathcal{C}_{\text{sig}} \cup \mathcal{C}_{\text{noise}}$ where $|\mathcal{C}_{\text{sig}}| = k \cdot p_{\text{sig}}$ passages are relevant signal documents and $|\mathcal{C}_{\text{noise}}| = k \cdot (1 - p_{\text{sig}})$ passages are irrelevant noise documents. The system is provided with an additional response option "Insufficient Information" and must decide whether to: (1) generate answer $\hat{a} \in \mathcal{A}$ if enough information is available in $\mathcal{C}$, or (2) output "Insufficient Information" if the retrieved context lacks necessary information to determine the answer confidently.

Integration: Multi-Document Information Synthesis. The Integration dimension evaluates the ability to combine information from multiple documents to answer complex queries that require compositional reasoning. Formally, for a query q with ground truth answer a^*, we break down the reasoning process into a set of sub-questions $\mathcal{Q}_{\text{sub}} = \{q_1, q_2, \ldots, q_n\}$ where each sub-question q_i can be answered by extracting information from a specific document in the retrieved set. Each sub-question q_i has a ground truth extractive answer a_i^* that is a span from the corresponding document $c_i \in \mathcal{C}$. The system must: (1) identify which document c_i is relevant for each sub-question q_i; (2) extract the sub-answer $\hat{a}_i$ from document c_i;

and (3) integrate the information from all sub-answers $\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n\}$ to determine the answer $\hat{a}$ to the main question q.

Robustness: Resilience to Factual Errors. The Robustness dimension evaluates strength against factual errors and misinformation in retrieved documents. Building on the sub-question framework from the Integration dimension, we construct adversarial contexts where some documents have been edited to contain factual errors. Formally, for each sub-question q_i with ground truth answer a_i^* and corresponding document c_i, we create a counterfactual version c_i' by: (1) generating an incorrect answer a_i' that semantically contradicts a_i^*; and (2) minimally editing document c_i to create c_i' such that c_i' appears convincing but supports the incorrect answer a_i' instead of a_i^*. The parameter p_{sig} now represents the proportion of factually correct documents: $\mathcal{C} = \mathcal{C}_{\text{correct}} \cup \mathcal{C}_{\text{adversarial}}$ where $|\mathcal{C}_{\text{correct}}| = k \cdot p_{\text{sig}}$ contain accurate information and $|\mathcal{C}_{\text{adversarial}}| = k \cdot (1 - p_{\text{sig}})$ contain adversarially edited misinformation.

5.3 The MedRGB Benchmark

The construction of MedRGB involves a systematic three-stage process designed to create complete evaluation data for medical RAG systems. Figure 11 shows the complete pipeline.

The first stage, *topic generation*, uses LLMs to generate diverse retrieval topics for each medical question, addressing the limitation that standard dense retrieval produces redundant results. The second stage, *document retrieval*, implements both offline retrieval using specialized medical corpora and online retrieval using public search engines, simulating different use cases and information access patterns. The third stage, *benchmark creation*, constructs the four test scenarios (Standard-RAG, Sufficiency, Integration, Robustness) through systematic

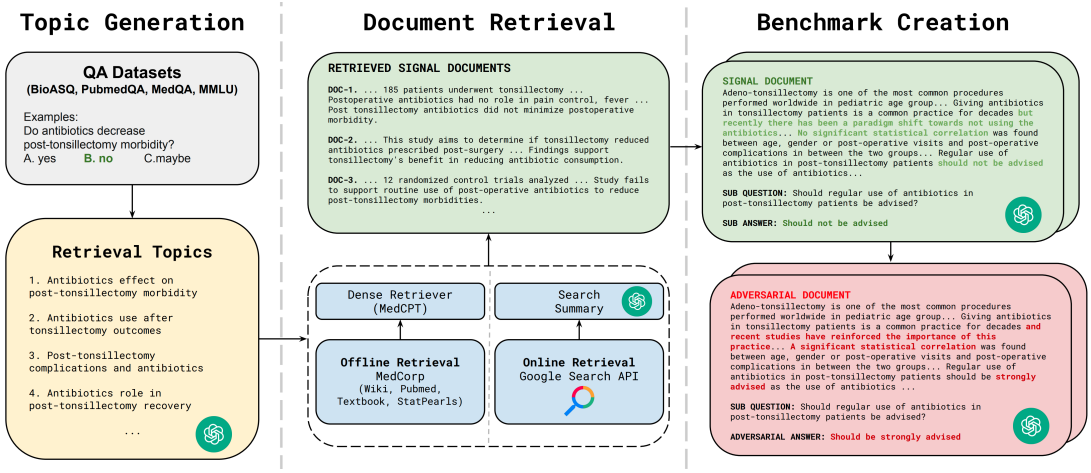


Figure 11. MedRGB creation process. OpenAI symbol implies the use of GPT-4o for data generation.

data augmentation procedures. Each stage is designed to ensure that the resulting benchmark captures realistic challenges faced by medical RAG systems in practical deployment.

5.3.1 Base Medical QA Datasets. MedRGB is constructed using multiple-choice questions from four medical QA datasets obtained from the MIRAGE benchmark (G. Xiong et al., 2024), selected to cover diverse medical knowledge domains and question types. These datasets collectively span various aspects of medical knowledge, from basic biomedical sciences to clinical reasoning, and include questions at different difficulty levels from medical student examinations to biomedical research comprehension.

BioASQ. The BioASQ challenge (Krithara et al., 2023) provides large-scale biomedical semantic indexing and question answering tasks. The dataset contains questions derived from PubMed abstracts and other biomedical literature, covering a wide range of biomedical topics including molecular biology, genetics, pharmacology, and clinical medicine. Questions are curated by biomedical experts

and require understanding of complex scientific concepts and terminology. BioASQ questions often demand detailed technical knowledge and the ability to comprehend research-level biomedical content, making them particularly challenging for evaluating whether RAG systems can effectively use scientific literature.

PubMedQA. The PubMedQA dataset (Q. Jin et al., 2019b) consists of yes/no/maybe questions based on PubMed abstracts, designed to test biomedical reasoning from research literature. Each question is derived from a PubMed abstract's title (posed as a question), and the abstract's conclusion provides context for determining the answer. The questions require understanding research findings, study designs, and the ability to assess evidence strength. In our adaptation for MedRGB, we use the multiple-choice variant where questions have been converted to multiple-choice format with plausible distractors, enabling consistent evaluation across datasets while maintaining the biomedical research focus.

MedQA. The MedQA dataset (D. Jin et al., 2020a) contains medical board examination questions from the United States Medical Licensing Examination (USMLE), covering clinical knowledge across multiple specialties including internal medicine, surgery, pediatrics, obstetrics and gynecology, and psychiatry. Questions are structured as clinical vignettes describing patient presentations, followed by multiple-choice options for diagnosis, treatment, or clinical reasoning. MedQA questions test clinical decision-making capabilities and require integrating multiple pieces of clinical information, making them representative of real-world medical reasoning tasks. The dataset's alignment with actual medical licensing examinations ensures that it captures the knowledge and reasoning skills expected of practicing physicians.

Table 15. Statistics of the four medical QA datasets used for MedRGB construction. All statistics refer to the test sets used for evaluation.

Dataset	# Questions	# Options	Domain	Source
BioASQ	1,504	4-5	Biomedical Research	PubMed
PubMedQA	500	3	Biomedical Research	PubMed
MedQA	1,273	4-5	Clinical Knowledge	USMLE
MMLU-Medical	1,089	4	Medical Knowledge	Various

MMLU-Medical. The medical subset of the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) includes questions across various medical and health-related subjects including anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and virology. MMLU-Medical questions span different educational levels from undergraduate to professional, covering both factual knowledge and conceptual understanding. The diverse question types and subject coverage make MMLU-Medical valuable for assessing breadth of medical knowledge and the ability to apply information across different subdomains of medicine and health sciences.

Table 15 presents statistics for the four datasets used in MedRGB construction. The datasets vary in size, with BioASQ providing the largest test set and PubMedQA the smallest, ensuring that our evaluation captures performance across different data volumes. The questions span different lengths and complexity levels, from concise factual queries to detailed clinical vignettes, providing complete coverage of medical QA scenarios.

5.3.2 Topic Generation. A conventional RAG setup for QA tasks uses a dense retriever to encode the input question into a high-dimensional vector and retrieve semantically similar passages from an external corpus. The top-k most similar documents based on vector similarity are then combined to create the

context for the LLM to answer the original question. However, this approach often results in redundancy in the retrieved set due to lack of diversity when using the similarity metric alone. Documents that are most similar to the query tend to cover overlapping information, potentially missing relevant information from different perspectives or complementary aspects of the topic. This limitation is particularly problematic for complex medical questions that may require information from multiple distinct sources or viewpoints.

To address this limitation, we first ask an LLM to generate a diverse set of retrieval sub-topics for each question. These topics represent different aspects or information needs related to answering the main question. Each generated topic then goes through the retrieval process to create a smaller set of relevant documents, and these sets are subsequently aggregated to create the final retrieval set for the original QA pair. This approach promotes diversity by ensuring that retrieval covers multiple distinct aspects of the question, rather than retrieving multiple similar documents about the same narrow aspect.

We use GPT-4o as the topic generator due to its strong performance in instruction following and medical knowledge. The prompt instructs the LLM to act as a medical expert and generate ranked search topics that will help answer a medical question. The prompt emphasizes three key guidelines: (1) topics should be ranked by importance to the question, ensuring that the most critical information is prioritized; (2) topics must be relevant to both the question and answer options, avoiding off-topic retrieval; and (3) topics should be differentiable and efficient for information retrieval, meaning they should represent distinct aspects rather than redundant reformulations. The LLM generates typically 3-5 topics per question, balancing diversity with computational efficiency. The complete prompt with full

context including question text and answer options is provided in Appendix B (Figure B.4).

5.3.3 Document Retrieval. We incorporate two distinct retrieval processes in MedRGB to simulate different real-world use cases and information access patterns. The *offline retrieval* scenario represents expert use cases where the retrieval corpus contains highly specialized medical information from curated sources such as medical textbooks, clinical guidelines, and peer-reviewed literature. The *online retrieval* scenario represents non-expert use cases where a general user poses questions to an LLM system, which retrieves documents through public search engines to help answer the query. Evaluating both scenarios provides insights into how RAG system performance varies with different retrieval sources and information quality characteristics.

Offline Retrieval. For the offline retrieval scenario, we use MedCorp (G. Xiong et al., 2024) as the external corpus. MedCorp is a complete medical document collection containing information from four high-quality sources: PubMed abstracts (biomedical research literature), StatPearls (peer-reviewed medical education resources), medical textbooks (complete clinical knowledge), and Wikipedia (general medical information). This diverse corpus ensures coverage of both specialized research knowledge and accessible clinical information. The retrieval topics generated in the previous step are encoded using MedCPT (Q. Jin et al., 2023), a biomedical-domain dense retriever specifically pretrained on medical corpora, to query MedCorp for relevant documents.

We select MedCPT instead of general-domain retrievers such as Contriever (Izacard et al., 2021) or lexical retrievers such as BM25 (Robertson & Zaragoza, 2009) due to its consistent superior performance in the medical

domain, as demonstrated by prior work (G. Xiong et al., 2024). MedCPT has been pretrained on large-scale medical corpora and fine-tuned for medical information retrieval tasks, enabling it to better understand medical terminology, acronyms, and semantic relationships specific to healthcare and biomedical sciences. For each retrieval topic, we collect the top-3 documents based on similarity scores. These documents are then aggregated across all topics to create the final retrieval set for each question, typically resulting in 9-15 unique documents per question after deduplication.

Online Retrieval. In the online retrieval scenario, we simulate the use case where a non-expert user asks a question to a general-purpose LLM, which then retrieves documents through online search engines to help answer the user's query. This scenario is increasingly common as LLMs are integrated into search engines and general-purpose AI assistants that ordinary users interact with for health information. For each medical question, the generated sub-topics are used to query the internet through the Google Search API. We retrieve the top-3 search results for each topic, and the content from each retrieved link is extracted and processed. After retrieving search results, we extract the textual content from each webpage and apply LLM-based summarization to create a signal document for the corresponding topic. The summarization step serves two purposes: (1) reducing the length of web content to fit within LLM context limits while preserving key information, and (2) filtering out irrelevant content such as other webpage elements that are not informative for answering the medical question. We again use GPT-4o for summarization, instructing it to extract and condense the main medical information relevant to the retrieval topic. The resulting summarized documents

typically range from 200-400 words each, providing concise, focused information sources.

5.3.4 Benchmark Creation. With retrieval topics generated and documents retrieved, we construct the four test scenarios through systematic data augmentation and manipulation procedures. Each scenario is designed to evaluate a specific dimension of RAG system capability, as defined in Section 5.2. The Standard-RAG test uses the retrieved documents directly as context. The Sufficiency test creates mixed contexts with varying proportions of signal and noise documents. The Integration test generates sub-questions paired with extractive answers from specific documents. The Robustness test creates adversarial contexts by editing documents to contain factual errors. All test scenarios use the same underlying medical questions and retrieval sets, enabling controlled comparison across evaluation dimensions.

5.3.4.1 Standard-RAG Test. In the Standard-RAG test setting, the retrieved context consists of a predefined number of signal documents sampled from the signal set obtained through the retrieval process. For each question, we sample k documents (where k is fixed from 5 to 20 depending on the experimental condition) from the pool of retrieved relevant documents. All documents in the context are relevant to the question, meaning they contain information potentially useful for determining the correct answer. This represents the ideal scenario where retrieval is successful and the system only needs to effectively use the provided relevant information.

The inference prompt used for the Standard-RAG test instructs the LLM to act as a medical expert and follow a two-step process: (1) analyze the provided documents and question carefully, and (2) think step-by-step to find the correct

answer. The full prompt includes the question, answer options, and all retrieved documents as context and emphasizes careful analysis and explicit reasoning, encouraging the model to express its thought process before selecting an answer option. This chain-of-thought prompting approach has been shown to improve reasoning quality and enables analysis of the model's decision-making process. The complete prompt is provided in Appendix B (Figure B.8).

5.3.4.2 Sufficiency Test. For the Sufficiency test, we sample both signal documents (relevant) and irrelevant noise documents to create retrieval sets with varying signal percentages $p_{\text{sig}} \in \{0, 20, 40, 60, 80, 100\}$. For a fixed total of k documents (typically $k = 10$), we sample $k \cdot p_{\text{sig}}$ signal documents from the relevant retrieval set and $k \cdot (1 - p_{\text{sig}})$ noise documents from an irrelevant document pool. The noise documents are obtained by sampling documents retrieved for different questions in the same dataset, ensuring they are topically medical but not relevant to the current question. This creates realistic noise conditions where retrieved documents may be related to medical topics generally but do not help answer the specific query at hand.

The inference prompt for the Sufficiency test extends the Standard-RAG prompt with additional instructions and response options. The LLM is instructed to: (1) identify relevant documents in the provided context, and (2) think step-by-step to find the correct answer. Critically, the prompt includes conditional instructions: if all documents are irrelevant, the model should (1) answer based on its internal knowledge if certain, or (2) return "Insufficient Information" if unsure. This structure tests whether the model can appropriately recognize when retrieved information is not enough and whether it can correctly judge its own internal knowledge sufficiency. The "Insufficient Information" option

provides a safety mechanism, allowing the model to decline answering rather than generating potentially incorrect responses. The complete prompt with all conditional instructions is provided in Appendix B (Figure B.9).

5.3.4.3 Integration Test. Complex medical questions often require addressing multiple sub-problems before answering the main question. The Integration test measures this capability by generating sub-questions based on signal documents and requiring the model to answer these sub-questions and integrate the information to address the main question. For each signal document in the retrieval set, we use an LLM to generate a sub-question that can be answered by extracting specific information from that document. The data generation prompt instructs the LLM to act as a medical research expert and follow specific guidelines: (1) explore different aspects related to the main question through sub-questions, (2) ensure each sub-question is specific to a particular document, and (3) provide a sub-answer that is a short string extracted from the corresponding document. The complete prompt for generating sub-questions and sub-answers is provided in Appendix B (Figure B.11).

The inference prompt for the Integration test provides detailed instructions for the multi-step reasoning process. The LLM is instructed to: (1) analyze all documents, (2) answer each sub-question using the most relevant document, ensuring each sub-answer is a concise string extracted from the corresponding document, and (3) think step-by-step and use the sub-question information to determine the correct answer to the main question. This structure guides the model through a hierarchical reasoning process: first locating and extracting specific information pieces, then integrating them to reach a final answer. As with the Sufficiency test, we evaluate the Integration test across varying p_{sig} values to assess

integration capability under different noise conditions, testing whether models can maintain effective integration even when some documents are irrelevant. The complete inference prompt is provided in Appendix B (Figure B.12).

5.3.4.4 Robustness Test. The Robustness test aims to measure LLM strength against factual errors, especially those that are adversarially designed to cause misinformation. Building on the sub-questions from the Integration test, we alter both the sub-answer and the corresponding document to create counterfactual examples. The data generation prompt instructs the LLM to: (1) analyze the provided question, original answer, and document; (2) generate a deliberately incorrect new answer that semantically contradicts the original; and (3) minimally edit the original document to create a persuasive but factually incorrect new document that supports the incorrect answer. The complete prompt for generating adversarial counterfactual documents is provided in Appendix B (Figure B.14).

In the Robustness test, all documents in the retrieved context are relevant to the question (topically appropriate), but p_{sig} now represents the percentage of factually correct documents. For a total of k documents, we include $k \cdot p_{\text{sig}}$ factually correct original documents and $k \cdot (1 - p_{\text{sig}})$ adversarially edited documents containing misinformation. This differs from the Sufficiency test where noise documents are simply irrelevant: here, adversarial documents are topically relevant but contain subtle factual errors designed to mislead the model. As p_{sig} decreases from 100% to 0%, the proportion of misinformation increases, creating progressively more challenging conditions for maintaining accuracy.

The inference prompt for the Robustness test explicitly instructs the model to detect factual errors. The LLM is instructed to: (1) for each sub-question, identify the corresponding relevant document; (2) determine if the document

contains any factual error—if factually correct, extract the sub-answer from the document; if it contains factual errors, provide the factually correct answer; and (3) think step-by-step and use sub-question information to answer the main question. This prompt structure makes the evaluation more challenging by explicitly asking the model to verify factual correctness, testing whether the model can successfully identify misinformation even when alerted to its potential presence. The complete inference prompt is provided in Appendix B (Figure B.16).

5.4 Experiments

5.4.1 Evaluation Setup.

5.4.1.1 Evaluated Models. We evaluate a complete range of state-of-the-art LLMs to understand RAG capabilities across different model families, sizes, and training approaches. The evaluation includes both closed commercial LLMs and open-source models, as well as both general-purpose and medical domain-specific models. This diverse model selection enables us to assess whether RAG capabilities scale with model size and training data, whether commercial models outperform open-source alternatives, and whether domain-specific fine-tuning provides advantages for medical RAG tasks.

For closed commercial LLMs, we assess multiple models from OpenAI. **GPT-4o** represents the frontier of commercial LLM capabilities with extensive training on diverse data and sophisticated reasoning abilities. **GPT-3.5** (specifically GPT-3.5-turbo) provides a baseline for older-generation commercial models that are still widely deployed. Additionally, we evaluate the recent **GPT-4o-mini**, which despite having only approximately 8 billion reported parameters, achieves performance nearly comparable to its full-sized counterpart GPT-4o. The inclusion of GPT-4o-mini is particularly valuable for assessing whether similar

RAG capabilities can be achieved with substantially smaller and more cost-effective models, which has important practical implications for deployment.

In the open-source category, we examine both general-purpose LLMs and domain-specific models. For general-purpose models, we include **Gemma-2-27b** from Google, an instruction-tuned model with 27 billion parameters that demonstrates strong performance across diverse tasks, and **LLaMA-3-70b** from Meta, a 70-billion parameter instruction-tuned model representing the current state-of-the-art in open-source LLMs. Both models have undergone extensive instruction tuning and alignment procedures, enabling them to follow complex prompts and generate high-quality responses in zero-shot settings.

For domain-specific models tailored for healthcare applications, we include **MEDITRON-70b** (Z. Chen et al., 2023), a 70-billion parameter model pretrained on medical corpora including PubMed abstracts, medical textbooks, and clinical guidelines, and **PMC-LLaMA-13b** (Wu et al., 2023), a 13-billion parameter model fine-tuned on medical literature from PubMed Central. These models have been specifically adapted for medical question answering through continued pretraining or fine-tuning on domain-specific data. Their inclusion enables assessment of whether specialized medical pretraining provides advantages for RAG tasks compared to general-purpose models that rely on medical knowledge acquired during standard large-scale pretraining.

For fair comparison, we report baseline results from the original papers when available, or reproduce results using official released code with identical experimental settings when necessary. All models are evaluated using the same prompts, retrieval sets, and evaluation protocols, differing only in the underlying LLM used for generation. For open-source models, we use consistent inference

settings including temperature ($T = 0$ for deterministic generation), top-p sampling parameters, and maximum generation length limits. This ensures that observed performance differences reflect genuine model capabilities rather than variations in experimental conditions.

5.4.1.2 Evaluation Metrics. Performance evaluation varies across the four test scenarios, with each scenario measuring different aspects of RAG capability. For the **Standard-RAG** test, we measure *accuracy* as the primary metric: the percentage of questions for which the model selects the correct answer option given the retrieved context. This provides a baseline measure of how effectively models can use relevant information when provided.

For the **Sufficiency** test, we measure multiple complementary metrics that capture different aspects of information insufficiency awareness. *Main question accuracy* measures the percentage of correct answers among questions where the model attempted an answer (not including "Insufficient Information" responses), assessing conditional accuracy given that the model chose to answer. *Insufficiency rate* measures the percentage of questions for which the model responded with "Insufficient Information," indicating the model's threshold for declining to answer. *Noise detection rate* measures the percentage of noise (irrelevant) documents correctly identified by the model, assessing ability to distinguish signal from noise. These metrics must be interpreted jointly: high accuracy with appropriate insufficiency rates and noise detection indicates robust sufficiency awareness, while high accuracy with inappropriately low insufficiency rates may indicate overconfidence.

For the **Integration** test, we evaluate performance at both sub-question and main question levels. At the sub-question level, we measure: (1) *Exact-match*

accuracy, the percentage of sub-questions for which the extracted answer exactly matches the gold extractive span, providing a strict measure of extraction precision; and (2) *GPT-based score*, a more lenient semantic similarity metric computed by an LLM judge that assesses whether the sub-answer is semantically equivalent or substantially similar to the gold answer. The scoring prompt instructs an expert LLM evaluator to judge whether a model prediction semantically matches the ground truth answer, allowing for partial credit (score 1.0 for complete match, 0.5 for partial match, 0.0 for incorrect). This lenient metric accommodates semantic equivalence rather than requiring exact string matching. At the main question level, we measure *integrated answer accuracy*, evaluating whether the model successfully combines information from sub-answers to determine the correct answer to the original query. The complete GPT-based scoring prompt is provided in Appendix B.

For the **Robustness** test, we measure: (1) *Main question accuracy*, the percentage of correct answers to the main question; (2) *Sub-question exact-match and GPT-based scores*, using the same metrics as the Integration test; and (3) *Factual error detection rate*, the percentage of adversarial (factually incorrect) documents correctly identified by the model. The factual error detection rate is particularly critical, as it directly measures whether models can identify misinformation, a basic requirement for reliable medical RAG systems.

5.4.1.3 Implementation Details. All experiments are conducted using the official APIs for commercial models (OpenAI API for GPT models) and Hugging Face Transformers library for open-source models. For open-source models, we run inference on NVIDIA A100 GPUs with 80GB memory. We use greedy decoding (temperature $T = 0$) for all models to ensure deterministic,

reproducible results. The maximum generation length is set to 2048 tokens to accommodate reasoning chains and long-form answers. For the case study experiments on long-form question answering, we use GPT-4o due to its superior performance and ability to follow complex prompts.

The number of retrieved documents varies by test scenario based on computational constraints and scenario requirements. For Standard-RAG evaluation, we test with both 5 and 20 documents to assess how document quantity affects performance. For Sufficiency, Integration, and Robustness tests, we use 10 documents to balance information coverage with context length constraints. The signal percentage p_{sig} varies across $\{0, 20, 40, 60, 80, 100\}$ for Sufficiency and Robustness tests, providing six different noise/misinformation levels. For Integration tests, p_{sig} starts from 20% (ensuring at least 2 signal documents for sub-questions) and ranges through $\{20, 40, 60, 80, 100\}$. All models are evaluated on the same questions, retrieval sets, and experimental conditions, ensuring fair comparison.

5.4.2 Standard-RAG Evaluation. Table 16 presents the performance comparison of all evaluated LLMs across three settings: the baseline "No Retrieval" condition where models answer using only their parametric knowledge, and Standard-RAG settings with 5 and 20 signal documents retrieved as context using both offline (MedCorp) and online (Google Search) retrieval. Results are reported separately for each of the four medical QA datasets (BioASQ, PubMedQA, MedQA, MMLU-Medical), revealing how RAG effectiveness varies across different question types and knowledge domains.

GPT-4o emerges as the top performer across most settings, demonstrating the benefits of scaling both parameters and training data to frontier levels. The

Table 16. Standard-RAG test accuracy across four medical QA datasets. Results are shown for No Retrieval (baseline), and Standard-RAG with 5 and 20 documents using both Offline (MedCorp + MedCPT) and Online (Google Search) retrieval. **Bold** indicates best performance for each model across settings.

LLMs	BioASQ					PubMedQA					MedQA					MMLU				
	No Retr.	Offline Retr.		Online Retr.		No Retr.	Offline Retr.		Online Retr.		No Retr.	Offline Retr.		Online Retr.		No Retr.	Offline Retr.		Online Retr.	
		5 doc	20 doc	5 doc	20 doc		5 doc	20 doc	5 doc	20 doc		5 doc	20 doc	5 doc	20 doc		5 doc	20 doc	5 doc	20 doc
GPT-3.5	77.7	81.2	87.2	87.2	87.9	49.8	59.6	71.0	58.4	60.6	68.3	63.0	67.3	68.0	68.4	76.3	70.3	73.0	75.7	74.8
GPT-4o-mini	82.9	85.3	**90.5**	89.0	90.0	47.0	60.8	**71.8**	60.6	61.2	79.2	77.1	79.5	79.0	**80.6**	88.3	84.6	87.3	86.0	**87.1**
GPT-4o	**87.9**	86.1	90.8	87.4	87.4	**52.6**	59.2	71.2	53.2	54.4	**89.5**	83.7	86.9	84.6	86.9	**93.4**	88.3	90.1	89.5	89.1
PMC-LLaMA-13b	64.2	64.6	64.6	63.9	64.1	55.4	54.0	54.0	54.8	54.6	44.5	38.9	38.8	43.4	43.7	49.7	43.7	44.0	48.4	48.2
MEDITRON-70b	68.8	74.0	74.8	79.8	**79.2**	53.0	53.4	47.8	58.8	46.8	51.7	56.0	57.4	61.8	62.9	65.3	65.1	66.3	67.6	**69.3**
Gemma-2-27b	80.3	83.3	88.7	88.7	**89.2**	41.0	52.0	59.0	52.6	49.4	71.2	69.8	71.7	75.9	**76.9**	83.5	77.9	82.5	82.2	**83.6**
LLaMA-3-70b	**82.9**	84.6	89.3	**89.3**	**89.3**	**59.2**	**77.6**	70.8	59.4	59.2	**82.9**	73.6	79.4	76.1	78.3	**85.2**	77.6	83.4	81.8	83.8

model achieves particularly strong baseline performance without retrieval (89.5% on MedQA, 93.4% on MMLU-Medical), suggesting extensive medical knowledge acquired during pretraining. Surprisingly, GPT-4o-mini, despite having only approximately 8 billion reported parameters, achieves results comparable to its much larger counterpart, often matching or exceeding GPT-4o's performance when RAG is applied. For example, on BioASQ with 20 offline documents, GPT-4o-mini achieves 90.5% compared to GPT-4o's 90.8%. This finding has significant practical implications, as GPT-4o-mini is substantially more cost-effective (reportedly 100× cheaper per token) while maintaining similar capabilities for medical RAG tasks.

Among open-source models, Gemma-2-27b and LLaMA-3-70b demonstrate strong performance, highlighting the effectiveness of general-domain instruction-tuned models in zero-shot medical RAG settings. LLaMA-3-70b achieves particularly impressive results, with baseline performance (82.9% on BioASQ, 82.9% on MedQA, 85.2% on MMLU-Medical) approaching or matching commercial models. This demonstrates that open-source models with strong general capabilities can effectively handle medical questions without requiring specialized medical pretraining. In contrast, domain-specific fine-tuned models yield mixed results. MEDITRON-70b shows improvement with RAG on some datasets (e.g., MedQA), but PMC-LLaMA-13b consistently underperforms, often showing degraded

performance with RAG compared to the no-retrieval baseline. This suggests that domain-specific pretraining alone is not enough without adequate model scale and instruction-tuning capabilities.

The effectiveness of RAG varies substantially across models and datasets. Large models with strong internal knowledge, such as GPT-4o and LLaMA-3-70b, benefit less from RAG compared to smaller models like GPT-4o-mini and Gemma-2-27b. For instance, GPT-4o shows minimal improvement or even slight performance decrease with RAG on MedQA and MMLU-Medical, while GPT-4o-mini and Gemma-2-27b demonstrate consistent improvements with more retrieved documents. This suggests that for smaller models with less complete internal knowledge, external information retrieval offers substantial performance gains.

Interestingly, the impact of document quantity differs substantially between offline and online retrieval sources. Online retrieval using Google Search often performs best with fewer documents (5 docs), while offline retrieval using MedCorp typically improves with more documents (20 docs). For example, on PubMedQA, LLaMA-3-70b achieves 77.6% accuracy with 5 offline documents but only 70.8% with 20, whereas online retrieval performance remains relatively stable (59.4% vs 59.2%). This discrepancy likely stems from differences in retrieval quality and noise characteristics. Google Search tends to provide high-quality top results but introduces more noise as the result count increases, while MedCPT retrieval from MedCorp offers more consistently relevant, high-quality medical documents that remain valuable in larger quantities. This finding has practical implications for system design: consumer-facing applications using online search should prioritize fewer high-quality results, while specialized medical applications with curated knowledge bases can benefit from retrieving more complete document sets.

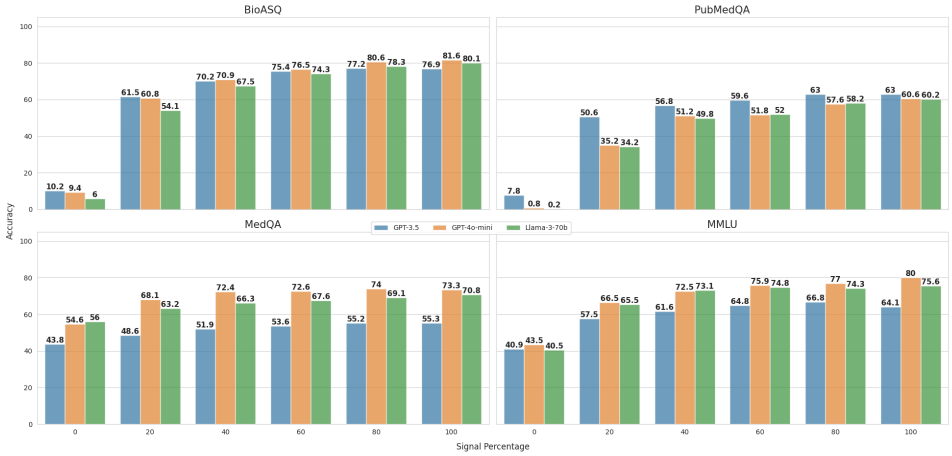


Figure 12. Sufficiency test main question accuracy across varying signal percentages.

We observe slight adverse effects when applying RAG for GPT-4o and LLaMA-3-70b on MMLU-Medical and MedQA datasets, where no-retrieval performance exceeds some RAG conditions. For instance, GPT-4o achieves 93.4% accuracy on MMLU-Medical without retrieval, but only 88.3% with 5 offline documents. Several factors may contribute to this phenomenon. First, these models possess strong internal medical knowledge that may be enough or even superior to retrieved information quality. Second, potential data leakage issues exist, as MMLU and MedQA are widely-used benchmarks that may have been seen during training, giving models advantage when relying solely on parametric memory. Third, additional context may sometimes introduce confusion or conflicting information that reduces confidence in correct answers. Despite these adverse effects in specific settings, RAG generally provides benefits, particularly for questions requiring current information or specialized knowledge beyond the model's training data.

5.4.3 Sufficiency Evaluation. The following evaluation focuses on three representative models selected based on Standard-RAG performance and

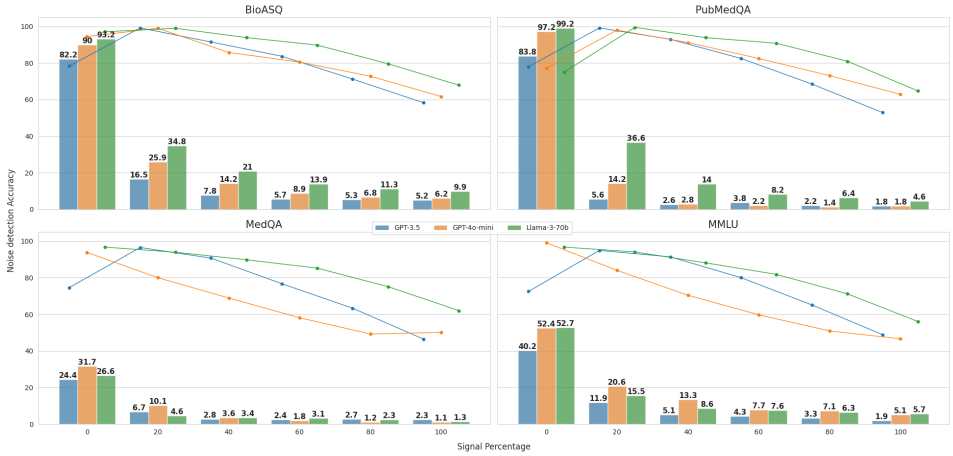


Figure 13. Sufficiency test percentage of information insufficiency (histogram) and noise detection rate (line graph).

practical considerations: GPT-3.5 and GPT-4o-mini for commercial LLMs (GPT-4o-mini provides nearly comparable performance to GPT-4o at dramatically lower cost), and LLaMA-3-70b as the best-performing open-source model. This selection enables complete analysis while maintaining computational and budget feasibility. Figure 12 presents main question accuracy across varying signal percentages p_{sig}, and Figure 13 shows the percentage of insufficiency responses and noise detection rates.

At $p_{\text{sig}} = 0\%$ where all documents are irrelevant noise, models predominantly return "Insufficient Information" responses, demonstrating appropriate recognition that retrieved context lacks necessary information. However, accuracy increases dramatically as signal percentage increases from 0% to 20%, indicating that even a small amount of relevant information (2 documents out of 10) substantially enhances model confidence and answer quality. For instance, on BioASQ, GPT-4o-mini accuracy increases from approximately 15% at $p_{\text{sig}} = 0\%$ to over 60% at $p_{\text{sig}} = 20\%$. This suggests that models can effectively use minimal relevant

information when present, but also that they may over-rely on even limited signal rather than maintaining appropriate skepticism.

The improvement in accuracy diminishes as more signal documents are added beyond $p_{\text{sig}} = 20\%$, showing a plateau effect. Most notably, even when retrieved context consists entirely of signal documents ($p_{\text{sig}} = 100\%$), performance is substantially lower than in the Standard-RAG setting with equivalent document counts. For example, GPT-4o-mini achieves 71.8% on PubMedQA in Standard-RAG with 20 docs, but only approximately 55-60% in the Sufficiency test at $p_{\text{sig}} = 100\%$. This dramatic gap suggests that in the Standard-RAG test, models may attempt to answer questions even when they lack enough confidence, whereas the explicit "Insufficient Information" option in the Sufficiency test makes models more conservative, occasionally declaring insufficiency even when adequate information is available. This conservative behavior may be desirable from a safety perspective but indicates that models struggle to accurately calibrate their confidence levels.

LLaMA-3-70b consistently outperforms other models in noise detection across all datasets and settings, achieving detection rates often exceeding 70-80% even at challenging noise levels. Figure 13 shows that noise detection rates are highest at low p_{sig} values (when most documents are noise), but decrease as signal percentage increases. At $p_{\text{sig}} = 0\%$, models achieve 80-90% noise detection, essentially recognizing that all documents are irrelevant. However, as signal increases, distinguishing noise from signal becomes harder, and models occasionally misidentify noise documents as signal or vice versa. The decrease in noise detection at higher p_{sig} values suggests that models use relative comparison to assess relevance: when most documents appear relevant, identifying the specific irrelevant ones becomes more challenging.

The percentage of "Insufficient Information" responses decreases substantially as p_{sig} increases, as expected. At $p_{\text{sig}} = 0\%$, most responses are "Insufficient Information" (70-85% depending on model and dataset), appropriately indicating that no relevant context is available. This percentage drops rapidly as signal increases, falling to 10-20% by $p_{\text{sig}} = 40\%$ and often below 10% at $p_{\text{sig}} = 100\%$. Interestingly, models rarely declare complete insufficiency at high signal percentages, suggesting they generally attempt answers when relevant information is present. However, the fact that insufficiency rates remain non-zero even at $p_{\text{sig}} = 100\%$ indicates occasional over-caution, where models declare insufficiency despite having adequate information.

Surprisingly, we observe cases where performance at $p_{\text{sig}} \in \{60\%, 80\%\}$ slightly exceeds performance at $p_{\text{sig}} = 100\%$. For instance, on some datasets, accuracy at $p_{\text{sig}} = 80\%$ may be 1-2% higher than at $p_{\text{sig}} = 100\%$. We hypothesize that without explicit "noise" to contrast against, models' internal relevance criteria become stricter when only signal documents are present. When all documents appear potentially relevant, models may struggle to confidently identify which information is most pertinent. In contrast, when a mix of noise and signal exists, models can more effectively infer criteria for distinguishing between document quality by comparing relevant versus irrelevant content. This phenomenon is observable in model reasoning traces (see Figure B.1 in Appendix B), where models explicitly note contrasts between documents to guide relevance assessment.

Based on these observations, we suggest that introducing a small amount of noise during training or inference might be beneficial, analogous to the concept of dropout regularization in neural networks. Just as dropout prevents overfitting by introducing noise during training, having some irrelevant documents in the

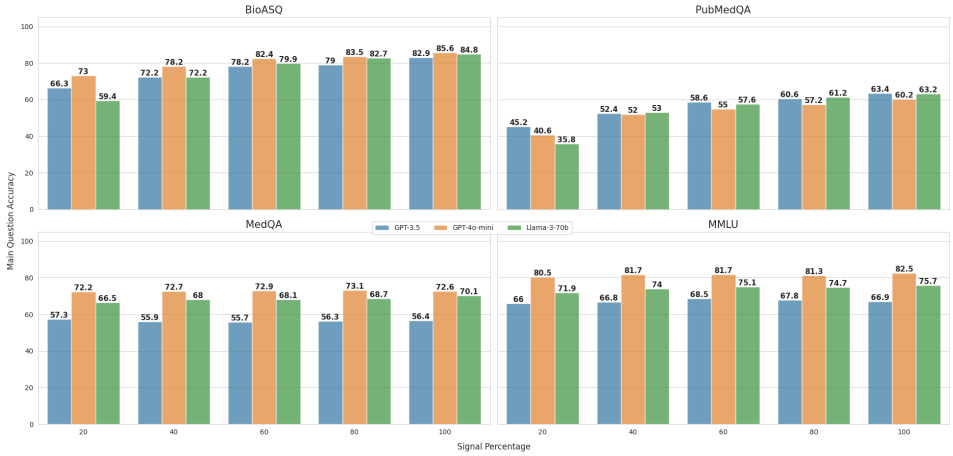


Figure 14. Integration test main question accuracy across varying signal percentages.

context may help models develop more robust criteria for evaluating document relevance. This "contrastive context" enables models to learn what distinguishes relevant from irrelevant information, rather than treating all provided documents with equal credibility. Future work could explore whether training with deliberately noisy contexts improves models' ability to assess information sufficiency and make appropriate confidence judgments.

5.4.4 Integration Evaluation. Figure 14 visualizes main question accuracy after models integrate information from answering sub-questions, and Figure 15 presents sub-question accuracy using both exact-match and GPT-based scoring metrics. The Integration test uses 10 retrieved documents with p_{sig} ranging from 20% to 100% (starting at 20% ensures at least 2 signal documents are available for generating sub-questions). The evaluation assesses whether models can: (1) identify which document is relevant for each sub-question, (2) accurately extract answers from the identified documents, and (3) successfully integrate sub-answers to determine the main answer.

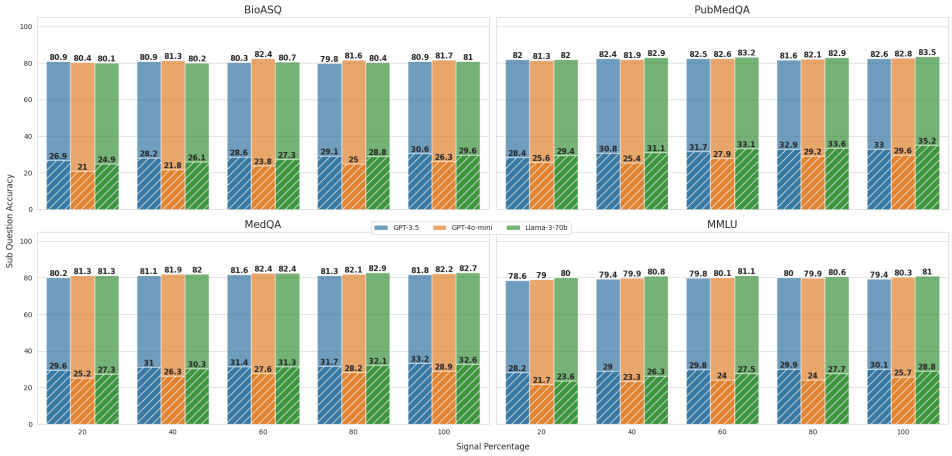


Figure 15. Integration test sub-question GPT-based score (filled histogram) and exact-match accuracy (shaded histogram).

Compared to results from the Sufficiency evaluation at equivalent p_{sig} values, the introduction of sub-questions significantly improves main accuracy at low signal percentages ($p_{\text{sig}} \in \{20\%, 40\%\}$). For instance, on MedQA at $p_{\text{sig}} = 20\%$, accuracy increases from approximately 45% in Sufficiency to 55-60% in Integration for GPT-4o-mini. This substantial improvement demonstrates that sub-questions provide valuable guidance, helping models navigate long noisy contexts to locate and extract relevant information. When noise predominates, the structured sub-question framework prevents models from being overwhelmed by irrelevant information, instead directing attention to specific aspects that collectively answer the main question.

However, at $p_{\text{sig}} = 100\%$ where contexts are similar to Standard-RAG settings, we observe that models perform slightly worse in the Integration test compared to Standard-RAG results. For example, GPT-4o-mini achieves 71.8% on PubMedQA in Standard-RAG with similar document counts, but approximately 68-70% in Integration at $p_{\text{sig}} = 100\%$. This performance degradation indicates

that the additional complexity introduced by sub-questions—requiring explicit extraction and multi-step reasoning—can sometimes hinder rather than help performance. Models may fail at various stages: identifying the relevant document for a sub-question, extracting the precise answer span, or integrating partial information into a coherent final answer. When all documents are relevant and the question is straightforward, direct reasoning may be more effective than structured decomposition.

Sub-question exact-match accuracy is relatively low across all settings, ranging from 20-30% depending on model, dataset, and p_{sig} value. This strict metric requires extracted sub-answers to exactly match gold spans, which is challenging given variations in phrasing, span boundaries, and extraction precision. In contrast, GPT-based scores remain consistently above 80% for most settings, indicating that while models may not extract exact spans, they generally capture semantically equivalent or substantially similar information. The large gap between exact-match and GPT-based scores (often 50+ percentage points) suggests that sub-question answering suffers more from extraction precision issues than from basic comprehension failures.

Analysis of model reasoning processes reveals that even with partially correct sub-answers, models can often integrate information successfully to reach correct main answers. Figure B.2 in Appendix B shows an example where the model extracts sub-answers that are semantically correct but not exact matches (e.g., extracting "intravenous antibiotics" instead of the gold span "IV ceftriaxone and vancomycin"). Despite this imprecision, the model successfully uses the sub-answer information to determine the correct main answer. This demonstrates that the GPT-based scoring metric appropriately captures functional accuracy: sub-

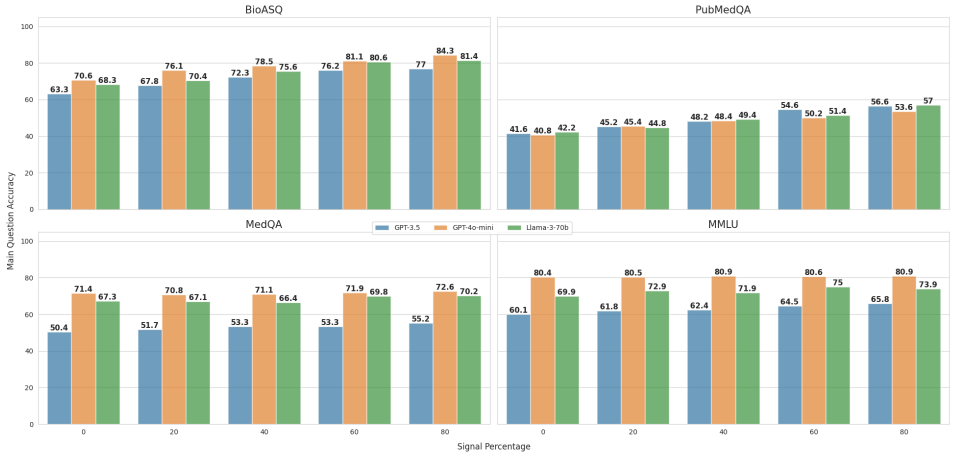


Figure 16. Robustness test main question accuracy under varying levels of misinformation.

answers that convey the right information, even if imperfectly extracted, can still contribute to correct reasoning.

However, having too few sub-questions can restrict the model's reasoning scope and lead to incorrect conclusions. At $p_{\text{sig}} = 20\%$ where only 2 signal documents are available, only 2 sub-questions can be generated, potentially missing critical information aspects. For instance, a clinical question might require information about diagnosis, treatment, and contraindications, but with only 2 sub-questions, one aspect may be omitted, leading to incomplete reasoning. This suggests a trade-off: more sub-questions provide complete information coverage but increase task complexity and potential for extraction errors, while fewer sub-questions simplify the task but may miss important information.

5.4.5 Robustness Evaluation. Figure 16 presents main question accuracy under varying levels of misinformation (controlled by p_{sig}, the percentage of factually correct documents), and Figure 17 shows sub-question scores alongside factual error detection rates. The Robustness test represents the most challenging

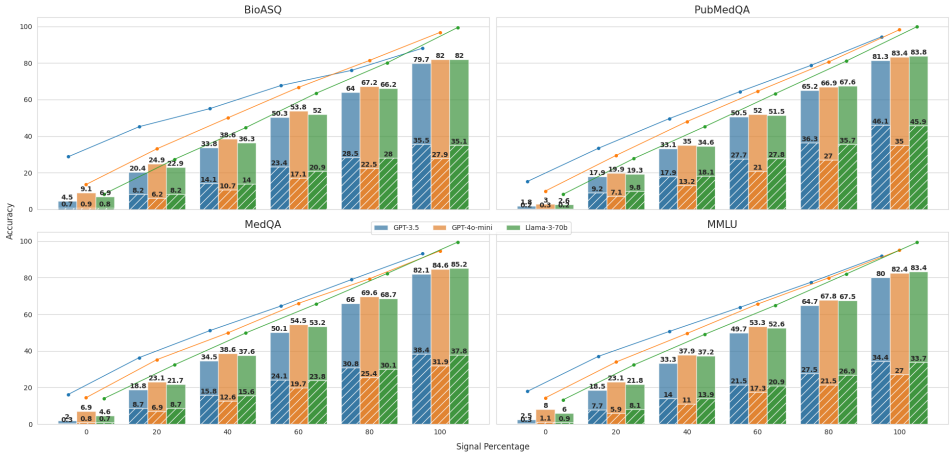


Figure 17. Robustness test sub-question GPT-based score (filled histogram), exact-match accuracy (shaded histogram), and factual error detection rate (line graph).

scenario, as adversarial documents are topically relevant but contain subtle factual errors designed to mislead models, simulating the realistic threat of health misinformation.

As expected, the presence of misinformation reduces overall performance across all models and datasets, with accuracy decreasing as the percentage of factually correct documents (p_{sig}) decreases. However, the magnitude of degradation varies substantially across datasets and models. On some datasets, accuracy remains relatively stable until p_{sig} drops below 40%, while on others, even 20% adversarial content causes notable performance drops. This variation suggests that some questions are more susceptible to misinformation than others, potentially depending on the complexity of the medical content and how subtle the adversarial edits are.

Interestingly, models perform better in the presence of factual misinformation (Robustness test) than with irrelevant noise (Sufficiency test) at equivalent signal percentages. For instance, at $p_{\text{sig}} = 60\%$, accuracy in the

Robustness test often exceeds accuracy in the Sufficiency test by 5-10 percentage points. This counterintuitive finding indicates that models are using information from adversarial documents despite their factual incorrectness. Because adversarial documents are topically relevant and contain medical terminology and structure similar to correct documents, models treat them as credible sources rather than filtering them out. This behavior is highly problematic for reliable medical systems: models accept and use misinformation, potentially reinforcing incorrect medical beliefs or recommendations.

Models struggle substantially with factual error detection, particularly at low p_{sig} values where adversarial documents predominate. When $p_{\text{sig}} = 0\%$ (all documents contain misinformation), detection rates typically fall below 50%, indicating that models fail to identify more than half of the factually incorrect documents. Detection rates improve as p_{sig} increases, reaching 70-90% at $p_{\text{sig}} = 100\%$ where models primarily need to identify themselves as having no errors to detect. This pattern suggests that models perform better at error detection when they have factually correct documents as reference points for comparison, but struggle when misinformation is pervasive.

Interestingly, GPT-3.5 demonstrates the highest factual error detection rates across most settings, often exceeding GPT-4o-mini and LLaMA-3-70b by 10-15 percentage points. However, GPT-3.5 simultaneously achieves the lowest main question accuracy, suggesting that high error detection does not necessarily translate to correct answers. This paradox may arise because GPT-3.5 identifies documents as potentially erroneous but lacks the medical knowledge to determine the correct information, or because excessive skepticism leads to rejecting both correct and incorrect information indiscriminately. In contrast, more capable

models like GPT-4o-mini may have stronger internal medical knowledge that enables them to answer correctly even when failing to explicitly identify all errors.

Sub-question scores are substantially lower in the Robustness test compared to the Integration test, particularly at low p_{sig} values. While Integration test GPT-based scores consistently exceed 80%, Robustness test scores at $p_{\text{sig}} = 20\%$ often fall to 40-60%. This dramatic decrease reflects high false positive rates: models accept adversarial information as factually correct and extract sub-answers accordingly. When models fail to detect misinformation, they proceed to answer sub-questions as if the adversarial documents were reliable sources, extracting incorrect information that propagates to the main answer. This cascading failure demonstrates how error detection failures at the document level multiply through the reasoning chain.

Figure B.3 in Appendix B presents an example where the model fails to detect adversarial editing. The original document correctly states that bacterial meningitis requires "immediate empiric antibiotic therapy with ceftriaxone," but the adversarial version subtly changes this to "immediate empiric antiviral therapy with acyclovir." The model fails to identify this factual error, extracts the incorrect sub-answer "antiviral therapy," and propagates this misinformation through its reasoning, ultimately selecting an incorrect treatment option. This example shows how convincing adversarial edits can mislead models, particularly when the misinformation maintains medical plausibility (antivirals are used for some meningitis cases, just not bacterial meningitis).

An interesting observation is that GPT-based sub-question scores show distinct patterns between Integration and Robustness tests. In Integration tests, scores remain consistently high (above 80%) even at low p_{sig}, while in

Robustness tests, scores decrease substantially as misinformation increases. This difference suggests that low GPT-based sub-scores may serve as an indicator of misinformation presence in the retrieved context. When models extract information that receives low semantic similarity scores compared to gold answers, this may signal that the source documents contain factual errors rather than that the model simply extracted incorrect spans. Future systems could potentially use this signal as a warning flag to trigger additional verification or to lower confidence in generated answers.

The Robustness evaluation reveals critical vulnerabilities that must be addressed before medical RAG systems can be safely deployed at scale. The combination of poor error detection, frequent false positives (accepting misinformation as fact), and propagation of incorrect information through reasoning chains poses serious risks in healthcare applications. A system that confidently provides treatment recommendations based on factually incorrect retrieved information could directly harm patients. These findings underscore the need for substantial improvements in several areas: retrieval systems must prioritize authoritative, verified sources; LLMs must develop stronger capabilities for critical evaluation of source reliability; and system architectures should incorporate explicit fact-verification components rather than assuming all retrieved information is trustworthy. The adversarial robustness problem represents perhaps the most pressing challenge for reliable medical RAG systems.

5.4.6 Case Study: Long-Form Question Answering. To explore potential improvements for medical RAG systems, we conduct a case study on long-form question answering that demonstrates two promising directions: improving retrieval quality through cross-validated document scoring, and enhancing LLM

reasoning through probabilistic reasoning prompts. We use MedLFQA (Jeong, Hwang, Yoon, Lee, & Kang, 2024), a biomedical long-form QA benchmark where questions require detailed reasoning incorporating multiple essential pieces of information. From MedLFQA, we curate 200 examples selected based on high complexity and high potential harm (both likelihood and severity), as determined by an LLM judge model. These examples represent challenging, high-stakes scenarios where reliable RAG performance is critical.

5.4.6.1 Retriever Improvement: Cross-Validate Scoring. To improve retrieval quality assessment, we use an LLM judge to assign scores to each retrieved document based on three key criteria:

- **Relevancy:** Does the document contain information or provide clarifications that align with the main question?
- **Consistency:** Does the document contain any detail that is contradictory to information in other retrieved documents?
- **Reliability:** Does the document contain any detail that is not true or not currently supported by world knowledge?

Critically, each document is assessed in relation to the other retrieved documents rather than in isolation. By conditioning on other documents as context, the judge model can cross-validate information across sources to produce more objective reliability scores and identify contradictions that single-document evaluation would miss.

The LLM judge assigns numerical scores for each criterion, which are then appended to each document as additional context provided to the answering LLM. These scores serve as quality signals that guide the answering model's trust in

different sources: highly relevant, consistent, and reliable documents should be weighted more heavily in reasoning, while documents with low reliability scores should be treated skeptically. This approach differs from filtering documents entirely (which risks removing partially useful information) by providing nuanced quality assessments that the answering model can incorporate into its reasoning process.

5.4.6.2 LLM Improvement: Probabilistic Reasoning. Medical decision-making inherently involves probabilistic reasoning: interpreting ambiguous symptoms, updating beliefs as new test results emerge, incorporating prior knowledge about disease prevalence, and weighing evidence of varying strength and reliability. To incorporate this reasoning mode into RAG systems, we develop carefully structured prompts that instruct models to use probabilistic reasoning when assessing factual accuracy of retrieved documents and determining confidence in their answers. The prompts encourage models to: (1) assess the reliability of each source probabilistically rather than treating all information as equally trustworthy, (2) interpret ambiguous or incomplete information by considering multiple possibilities, (3) update beliefs as they process multiple documents sequentially, and (4) provide confidence estimates for their final answers.

5.4.6.3 Results and Analysis. Table 17 presents results of GPT-4o on the MedLFQA task under the Robustness test setting (with 10 documents retrieved and varying p_{sig} representing the percentage of factually correct documents). We evaluate four configurations: **Standard** uses the same prompts as the main Robustness evaluation; **Retr Score** adds cross-validated document scoring; **Prob Reason** adds probabilistic reasoning prompts; and **Combine**

Table 17. Results of GPT-4o on LFQA task in Robustness test (10 documents retrieved), with vary signal ratio p_{sig} (percentage) for main-question accuracy (MAIN), factual error detection (FACT), and sub-question accuracy (SUB). *Standard* is similar to Section 5.4.5. *Retr_Score* and *Prob_Reason* are enhanced models with Retrieved Document Scoring and Probabilistic Reasoning, respectively. *Combine* integrates both of the enhancements into one model.

MAIN - Main Question Accuracy (%)						
Signal Ratio	**0**	**20**	**40**	**60**	**80**	**100**
Standard	47.0	48.0	49.8	49.8	52.0	51.3
Retr Score	49.0	48.0	50.8	51.8	50.3	49.8
Prob Reason	48.0	48.0	51.5	49.5	50.5	50.0
Combine	49.8	47.8	51.0	51.3	50.5	49.8
FACT - Factual Error Detection Rate (%)						
Signal Ratio	**0**	**20**	**40**	**60**	**80**	**100**
Standard	45.3	57.1	69.5	80.2	89.1	99.4
Retr Score	58.5	69.0	75.9	83.9	90.1	97.3
Prob Reason	66.6	72.2	78.2	84.2	86.1	93.0
Combine	**83.3**	**80.5**	72.5	67.1	58.8	54.1
SUB - Sub-Question GPT-based Score (%)						
Signal Ratio	**0**	**20**	**40**	**60**	**80**	**100**
Standard	19.6	31.3	49.8	59.0	71.2	85.9
Retr Score	24.3	35.3	48.3	59.5	72.5	85.4
Prob Reason	21.0	33.1	46.7	59.4	69.5	83.4
Combine	21.4	33.3	45.2	57.3	69.0	81.6

integrates both enhancements. Results are reported for main question accuracy (MAIN), factual error detection rate (FACT), and sub-question accuracy (SUB).

The proposed enhancements substantially increase factual error detection rates, particularly at low p_{sig} values where misinformation predominates. Most dramatically, the **Combine** configuration achieves 83.3% detection rate at $p_{\text{sig}} = 0\%$ where all documents contain adversarial misinformation, compared to only 45.3% for the Standard baseline—an improvement of 38 percentage points.

This demonstrates that both cross-validated document scoring and probabilistic reasoning prompts effectively improve the model's ability to identify factual errors. The enhancements help most when misinformation is pervasive: at $p_{\text{sig}} \in \{0\%, 20\%\}$, all three enhanced configurations substantially outperform Standard, but improvements diminish or reverse at higher p_{sig} values.

Improvements in main question accuracy are more modest and less consistent than error detection gains. At low p_{sig} values ($\{0\%, 20\%\}$), enhanced configurations achieve accuracy gains of 1-3 percentage points over Standard. For example, at $p_{\text{sig}} = 0\%$, Combine achieves 49.8% compared to Standard's 47.0%, a meaningful but not dramatic improvement. At intermediate values ($p_{\text{sig}} \in \{40\%, 60\%\}$), some enhanced configurations show 1-2% improvements, though results vary. However, at high p_{sig} values ($\{80\%, 100\%\}$), enhanced configurations sometimes underperform Standard, suggesting that added complexity can hinder performance when contexts are predominantly factually correct.

The results reveal a trade-off between error detection capability and answer accuracy. The Combine configuration achieves the highest error detection at low p_{sig} but shows diminishing or negative returns for accuracy at high p_{sig}. This pattern suggests that enhanced prompts and document scoring introduce additional reasoning complexity: more steps where errors can occur, more information to track, and more decisions to make. When most documents are factually correct ($p_{\text{sig}} \geq 80\%$), this complexity may be unnecessary overhead that degrades rather than improves performance. The enhancements appear most valuable precisely when they are most needed—when misinformation is prevalent—but less helpful or even harmful when information quality is already high.

Sub-question GPT-based scores follow expected patterns, increasing monotonically as p_{sig} increases across all configurations. Scores range from approximately 20% at $p_{\text{sig}} = 0\%$ to 82-86% at $p_{\text{sig}} = 100\%$, reflecting that models can extract correct sub-answers more frequently when documents are factually accurate. The enhanced configurations show minimal impact on sub-question performance, with differences of only 1-3 percentage points compared to Standard across most settings. This suggests that the enhancements primarily affect higher-level reasoning (error detection and main question answering) rather than low-level extraction capabilities.

The case study demonstrates that both retriever enhancement through cross-validated document scoring and LLM enhancement through probabilistic reasoning prompts represent promising directions for improving medical RAG robustness. The dramatic improvements in error detection (38+ percentage points at $p_{\text{sig}} = 0\%$) validate that current systems' poor robustness is not an inherent limitation but rather a capability that can be improved through appropriate system design. However, the more modest and inconsistent improvements in answer accuracy, along with performance degradation at high signal ratios, indicate that these preliminary implementations require further refinement to balance error detection capability against answer quality and computational complexity.

These proof-of-concept results suggest several directions for future work. First, adaptive approaches could dynamically determine when enhanced error detection is needed based on initial assessment of retrieval quality, applying costly enhancements only when information quality is suspect. Second, iterative refinement could use error detection as a signal to trigger re-retrieval from more authoritative sources when misinformation is detected. Third, ensemble methods

could combine multiple scoring approaches or reasoning strategies to achieve more robust performance across varying information quality conditions. Finally, training-time interventions could fine-tune models specifically for robustness to misinformation, potentially achieving similar benefits without the inference-time overhead of complex prompting. The case study establishes that improved medical RAG robustness is achievable, though substantial research remains to develop production-ready solutions.

5.5 Related Work

5.5.1 Medical Question Answering Benchmarks.

Previous medical benchmarks have traditionally focused on target performance evaluation using question-answer pairs without considering the interaction between models and external knowledge sources. PubMedQA (Q. Jin et al., 2019b) provides yes/no/maybe questions based on PubMed abstracts to test biomedical reasoning. MedQA (D. Jin et al., 2020a) contains medical board examination questions covering multiple specialties. BioASQ (Krithara et al., 2023) offers a large-scale biomedical semantic indexing and question answering challenge. The MMLU benchmark (Hendrycks et al., 2021) includes a medical subset covering various healthcare topics. While these benchmarks have been instrumental in assessing medical knowledge and reasoning capabilities of AI systems, they consist solely of QA pairs and do not evaluate how models interact with retrieved documents or handle noisy, incomplete, or contradictory information.

Some recent benchmarks have expanded evaluation scope to include multiple aspects of medical AI performance, but still without systematic RAG assessment. MedEval (He et al., 2023) presents a large-scale, expert-annotated benchmark covering various medical tasks and domains, providing complete

evaluation of LLM capabilities across clinical reasoning, medical knowledge, and diagnostic tasks. However, the evaluation does not involve retrieval-augmented generation settings. Similarly, most current systematic evaluations of LLMs in the medical domain (Nori et al., 2023; Singhal et al., 2023) assess models' internal knowledge and reasoning without examining their ability to effectively use external information sources, which is crucial for real-world deployment where medical knowledge evolves rapidly and complete memorization is infeasible.

More recent efforts have begun addressing evaluation comprehensiveness in medical QA. Manes et al. (Manes et al., 2024) introduce a real-world medical QA benchmark with physician-created answers and decomposed statements, along with two NLI-based evaluation metrics for measuring answer comprehensiveness and hallucination rates. Hosseini et al. (Hosseini et al., 2024) evaluate long-form QA capacity of various LLMs based on correctness, helpfulness, harmfulness, and bias. While these works expand evaluation beyond simple accuracy, they do not systematically assess RAG-specific capabilities such as noise handling, information integration, and robustness to misinformation in retrieved contexts.

Most relevant to our work, Xiong et al. (G. Xiong et al., 2024) provide systematic evaluation of RAG systems in medicine through their MIRAGE benchmark, which covers five medical QA datasets and evaluates how different retrieval modules affect target accuracy. They examine various retrievers (BM25, Contriever, MedCPT) and retrieval sources (PubMed, Wikipedia, textbooks, StatPearls) to understand their impact on LLM performance. However, their evaluation focuses primarily on the effect of RAG modules on target accuracy, missing other important aspects of a reliable medical AI system. MedRGB builds upon MIRAGE by using four of their datasets as foundation, but extends

the evaluation to assess whether models appropriately recognize information insufficiency, can effectively use multiple documents with sub-question answering, and remain robust to adversarial factual errors, providing a more complete picture of RAG system capabilities and limitations.

5.5.2 Retrieval-Augmented Generation.

Retrieval-Augmented Generation, introduced by Lewis et al. (P. S. H. Lewis et al., 2020), addresses limitations of purely parametric language models by integrating external knowledge sources into the generation process. The standard RAG architecture consists of a retriever component that searches a knowledge base for relevant passages, and a generator component (typically an LLM) that produces answers conditioned on both the input query and retrieved context. This approach has achieved notable success in knowledge-intensive tasks across various domains (Gao et al., 2023; Peng et al., 2023; Ram et al., 2023), demonstrating that augmenting LLMs with external knowledge can improve factual accuracy, enable access to current information, and provide transparency through cited sources.

Numerous improvements have been proposed to advance beyond the standard RAG architecture. Self-RAG (Asai, Wu, Wang, Sil, & Hajishirzi, 2024) enables models to retrieve information on-demand and generate reflective tokens to critique their own outputs. Corrective RAG (Yan, Gu, Zhu, & Ling, 2024) incorporates a retrieval evaluator to assess document relevance and trigger corrective actions when needed. Adaptive RAG (Jeong, Sohn, Sung, & Kang, 2024) dynamically determines when retrieval is necessary based on query complexity. Active Retrieval Augmented Generation (Z. Jiang et al., 2023) determines when and what to retrieve through active learning. These advances demonstrate the ongoing evolution of RAG systems toward more sophisticated retrieval strategies

and generation mechanisms, though evaluation has typically focused on general-domain tasks rather than high-stakes specialized domains like medicine.

Specifically for the medical domain, several works have explored RAG applications for healthcare and clinical tasks. Frisoni et al. (Frisoni, Mizutani, Moro, & Valgimigli, 2022) investigate medical question answering using retrieval-augmented transformers. Naik et al. (Naik, Parasa, Feldman, Wang, & Hope, 2022) examine RAG for legal and medical document analysis. Lala et al. (Lála et al., 2023) develop a multilingual medical RAG system. Jeong et al. (Jeong, Sohn, et al., 2024) apply adaptive RAG to biomedical question answering. Wang et al. (J. Wang, Yang, Yao, & Yu, 2024) explore knowledge graph-augmented RAG for clinical decision support. Xiong et al. (G. Xiong et al., 2024) systematically compare different retrieval sources and methods for medical QA. While these works demonstrate RAG's promise in medical applications, they primarily evaluate performance on standard accuracy metrics and do not comprehensively assess the practical challenges of handling noisy retrievals, integrating complex information, or maintaining robustness against misinformation.

5.5.3 RAG Evaluation Frameworks. Recent works have begun developing more complete evaluation frameworks for RAG systems beyond simple accuracy measurement. RAGAS (Retrieval-Augmented Generation Assessment) (Es et al., 2023) proposes a framework to assess three key qualities of RAG outputs for QA tasks: Faithfulness measures the degree to which generated responses align with the provided context, ensuring that models do not hallucinate information beyond what is retrieved; Answer Relevance evaluates the extent to which generated responses address the actual question posed; and Context Precision-Recall assesses the quality of retrieved context, measuring whether relevant

information is successfully retrieved (recall) and whether retrieved information is predominantly relevant (precision). These metrics provide valuable signals about different aspects of RAG system performance, but they are designed for general-domain applications and do not address domain-specific requirements such as medical safety and reliability.

The RGB (Retrieval-Augmented Generation Benchmark) framework (J. Chen et al., 2024) establishes a systematic approach to measuring four core abilities required for effective RAG systems. The framework evaluates: (1) noise robustness, testing whether systems can identify and use relevant information when irrelevant documents are present; (2) negative rejection, assessing whether systems can recognize when retrieved information is not enough to answer a query; (3) information integration, measuring the ability to combine information from multiple documents; and (4) counterfactual robustness, evaluating strength against contradictory or incorrect information in retrieved context. RGB provides a complete evaluation methodology that inspired our work, demonstrating the importance of assessing RAG systems across multiple practical scenarios rather than focusing solely on standard retrieve-and-answer settings. However, RGB is designed for general-domain applications and does not address the specific constraints and requirements of medical applications.

The medical domain presents unique challenges that distinguish it from general-domain applications and require specialized evaluation frameworks. First, the consequences of errors are substantially more severe, as incorrect medical information can directly impact patient health and clinical decision-making. Second, medical applications demand extremely high reliability and consistency, as healthcare providers and patients must be able to trust system outputs.

Third, medical knowledge evolves rapidly through new research findings and clinical guidelines, requiring systems to effectively integrate current information while maintaining accuracy. Fourth, medical queries often require compositional reasoning over multiple pieces of evidence, such as patient history, symptoms, test results, and treatment guidelines. Finally, the medical domain faces a significant misinformation problem, with substantial amounts of inaccurate health information available online that systems must be able to identify and filter. These considerations motivate the design of MedRGB's four test scenarios, each addressing critical aspects of reliable medical RAG systems.

5.6 Summary

This chapter extended the evaluation of large language models (LLMs) in retrieval-augmented generation (RAG) settings for medical question answering (QA) tasks to crucial aspects of reliable medical AI systems, including sufficiency, integration, and robustness. We created Medical Retrieval-Augmented Generation Benchmark (MedRGB) that provides retrieval topics, signal documents, sub-QA pairs, and adversarial documents, for four medical QA datasets. Using MedRGB, we assessed a wide range of LLMs, including both closed commercial LLMs and open-source models and their reasoning processes in each of the test scenarios. Our experimental results reveal current RAG system's limitations in handling these more complex situations. The findings from our analysis provide practical guidelines and directions for developing more reliable and trustworthy medical RAG systems.

CHAPTER VI

CONCLUSION

6.1 Summary

This dissertation has undertaken a comprehensive exploration of transfer learning in Natural Language Processing, investigating how knowledge acquired in one context can be effectively leveraged in another. Through systematic investigation spanning domain adaptation, task adaptation for structured prediction, cross-lingual transfer, and expertise specialization, this work has developed novel methods and evaluation frameworks that collectively address the full spectrum of modern transfer learning challenges. The first two research directions advanced adaptive transfer learning: Meta Self-Paced Domain Adaptation (MSP-DA) introduced a meta-learning framework that automatically optimizes instance selection and objective weighting for unsupervised domain adaptation, while Domain Adaptation for Joint Information Extraction (DA4JIE) extended these ideas to structured prediction through graph-relational adversarial alignment and context-invariant structure learning. The third direction investigated zero-shot cross-lingual transfer through linguistic distance analysis, theory-guided language selection, and graph-relational adversarial learning across 17 typologically diverse languages. The fourth direction transitioned to specialization transfer by introducing Medical Retrieval-Augmented Generation Benchmark (MedRGB), the first comprehensive framework evaluating medical RAG systems across sufficiency, integration, and robustness dimensions. The culmination of this work presents significant contributions to NLP and transfer learning, advancing our understanding of how knowledge can be effectively transferred, adapted, and specialized across diverse languages, domains, and applications.

6.2 Limitations

Despite the considerable progress and contributions, this dissertation acknowledges several limitations that warrant further discussion:

Task Scope and Generalizability: The methods developed in this dissertation are primarily evaluated on information extraction and its related subtasks, alongside sentiment analysis and medical question answering. While IE provides a complex and suitable testbed exhibiting rich interdependencies, the findings may not fully generalize to generation-heavy tasks such as machine translation and summarization, or reasoning-intensive applications such as mathematical problem solving and semantic parsing, which involve fundamentally different linguistic phenomena and transfer dynamics.

Computational Requirements: The proposed methods introduce computational overhead that may limit applicability in resource-constrained settings. MSP-DA adds approximately 20–30% overhead from inner-loop meta-learning gradients, DA4JIE requires constructing and propagating through relational graphs with increased memory consumption, and the cross-lingual relational-transfer approach maintains separate discriminators scaling linearly with the number of source languages. While these costs are justified by performance gains and reduced manual tuning, they may be prohibitive for organizations with extremely limited computational resources.

Benchmark and Evaluation Constraints: Evaluation relies on established benchmarks such as ACE-05, BETTER, FDU-MTL, and MINION that involve clean, well-annotated data with clear domain boundaries. Real-world applications often involve noisier data, gradual domain shifts, and ambiguous domain definitions that these benchmarks do not fully capture. Similarly, MedRGB focuses on

multiple-choice and extractive formats that may not reflect the full complexity of real-world medical information needs, and automatic evaluation metrics may not perfectly align with expert clinical judgment.

6.3 Future Work

The research presented in this dissertation opens numerous promising directions for future investigation that can extend, refine, and broaden the transfer learning methods and evaluation frameworks developed here.

Extensions to Adaptive Transfer Methods: MSP-DA could be extended to handle multiple source domains simultaneously, learning domain-specific age thresholds and weighting strategies for each source while meta-optimizing for target performance. Alternative partition strategies—using clustering-based difficulty measures, uncertainty estimates, or adversarial perturbations—could enhance robustness. DA4JIE could incorporate richer structural information such as coreference chains, discourse relations, or knowledge graph structures. For cross-lingual transfer, scaling experiments to 50+ languages would test whether identified transfer patterns generalize to broader language coverage, while combining linguistically-informed selection with adapter-based fine-tuning or prompt tuning could reduce computational costs.

Advancing Specialization Transfer and RAG Evaluation: MedRGB could be extended to incorporate more realistic medical scenarios including clinical note generation, multi-turn clinical dialogue, and differential diagnosis reasoning, as well as benchmarks for other high-stakes domains such as legal reasoning and financial analysis. Future RAG architectures should investigate retrieval methods that prioritize authoritative sources, generation methods with explicit fact-verification

components, and adaptive systems that dynamically determine when enhanced error detection is needed based on retrieval quality assessment.

Unified Frameworks Bridging Adaptive and Specialization Paradigms: Modern applications increasingly require capabilities spanning both paradigms—for example, medical QA systems that must adapt to specific hospital documentation styles while maintaining reliability. Combining meta-learning for hyperparameter optimization with retrieval-augmented generation, or integrating cross-lingual transfer with domain-specific specialization for medical applications in low-resource languages, could enable more robust real-world deployment. Formal theoretical frameworks characterizing the relationships between different types of transfer learning would provide principled guidance for such unified approaches.

Leveraging Large Language Model Advances: The advancement of LLMs creates transformative opportunities: extended context capabilities can improve document-level IE and cross-lingual transfer; instruction-following abilities reduce dependence on task-specific formulations; and inherent multilingual capabilities create new avenues for prompting-based transfer. Investigating whether the transfer learning principles identified in this dissertation remain effective in instruction-tuning and in-context learning paradigms represents an important validation of their generality.

In conclusion, while this dissertation has made substantial contributions to the field of transfer learning in NLP, the path forward invites continued collaborative and innovative research effort. By addressing the identified limitations and pursuing the proposed future directions, the next generation of transfer learning research can make significant further advances towards NLP technologies

that work effectively, reliably, and equitably across the full diversity of human languages, domains, and applications.

APPENDIX A

CHAPTER 4 SUPPLEMENTARY MATERIALS

A.1 Binary Distances

In Figure A.18, we present the hierarchical clustering results of 76 binary distances from Choi et al. (2010), four of which are chosen for our investigations. Based on the 2×2 contingency table (Table A.18), we can formulate the four binary distances as follows:

- Jaccard: $S_{\texttt{JACCARD}} = \frac{a}{a+b+c}$
- Inner-product: $S_{\texttt{INNER-PRODUCT}} = a + d$
- Hamming: $S_{\texttt{HAMMING}} = b + c$
- Anderberg: $S_{\texttt{ANDERBERG}} = \frac{\sigma - \sigma'}{2n}$

The σ and σ' in Anderberg distance are defined as follows:

$$\sigma = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$$

$$\sigma' = \max(a + c, b + d) + \max(a + b, c + d)$$

i \ j	1 (Presence)	0 (Absence)	Sum
1 (Presence)	$a = i \bullet j$	$b = \bar{i} \bullet j$	$a + b$
0 (Absence)	$c = i \bullet \bar{j}$	$d = \bar{i} \bullet \bar{j}$	$c + d$
Sum	$a + c$	$b + d$	$n = a + b + c + d$

Table A.18. The OTUs (Operational Taxonomic Units) expression of binary vectors i and j.

A.2 Feature Importances

The feature importance weights of the optimally combined metric for each (task, model scale) setting are visualized in Figure 7 (in Section 4.4.1). The weights

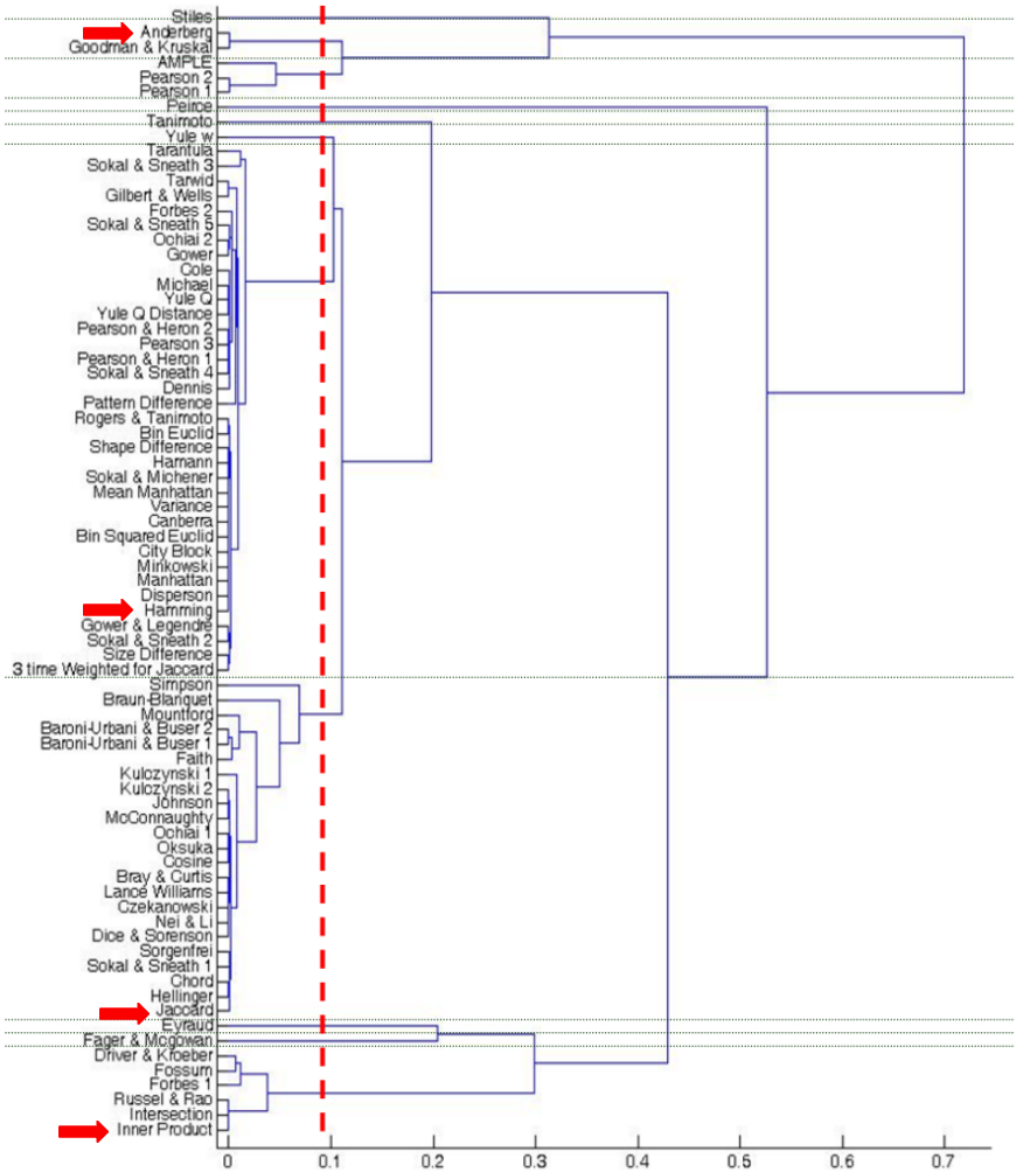


Figure A.18. Hierarchical clustering result of 76 binary distances according to Choi et al. (2010). The considered distances are indicated by red arrows.

are obtained from the solution of a simple constrained linear maximization using the Scipy library[1], as follows:

$$\max_{W \in \mathbb{R}^{14}} \text{corr}(d_{\text{comb\_all}}(W), \texttt{transfer\_score}), \text{ where}$$
$$W = \{w_i\}_{i=1}^{14} \ , \ d_{\text{comb\_all}}(W) = \sum_{i=1}^{14} w_i d_i \ ,$$
$$0 \leq w_i \leq 1 \ \text{ for } \ i = \overline{1,14} \ , \text{ and } \sum_{i=1}^{14} w_i = 1.$$

Here `transfer_score` refers to a particular single-transfer performance between 2 languages, and $\{w_i\}_{i=1}^{14}$ are the weights of the following 14 considered distances: `hamming-syntax`, `hamming-phonology`, `hamming-inventory`, `jaccard-syntax`, `jaccard-phonology`, `jaccard-inventory`, `inner-syntax`, `inner-phonology`, `inner-inventory`, `fam`, and `geo`.

A.3 Detailed Experimental Results

In this section, we provide detailed result values for our experiments, including: linguistic distances in Figure A.19, ZSCL-S scores in Figures A.20 and A.21, transfer-distance correlations in Figures A.22 and A.23, ZSCL-M scores in Figures A.24 and A.25, DANN scores in Figures A.26 and A.27, and finally ZSCL-R scores in Figures A.28 and A.29.

[1] `https://docs.scipy.org/doc/scipy/tutorial/optimize.html`

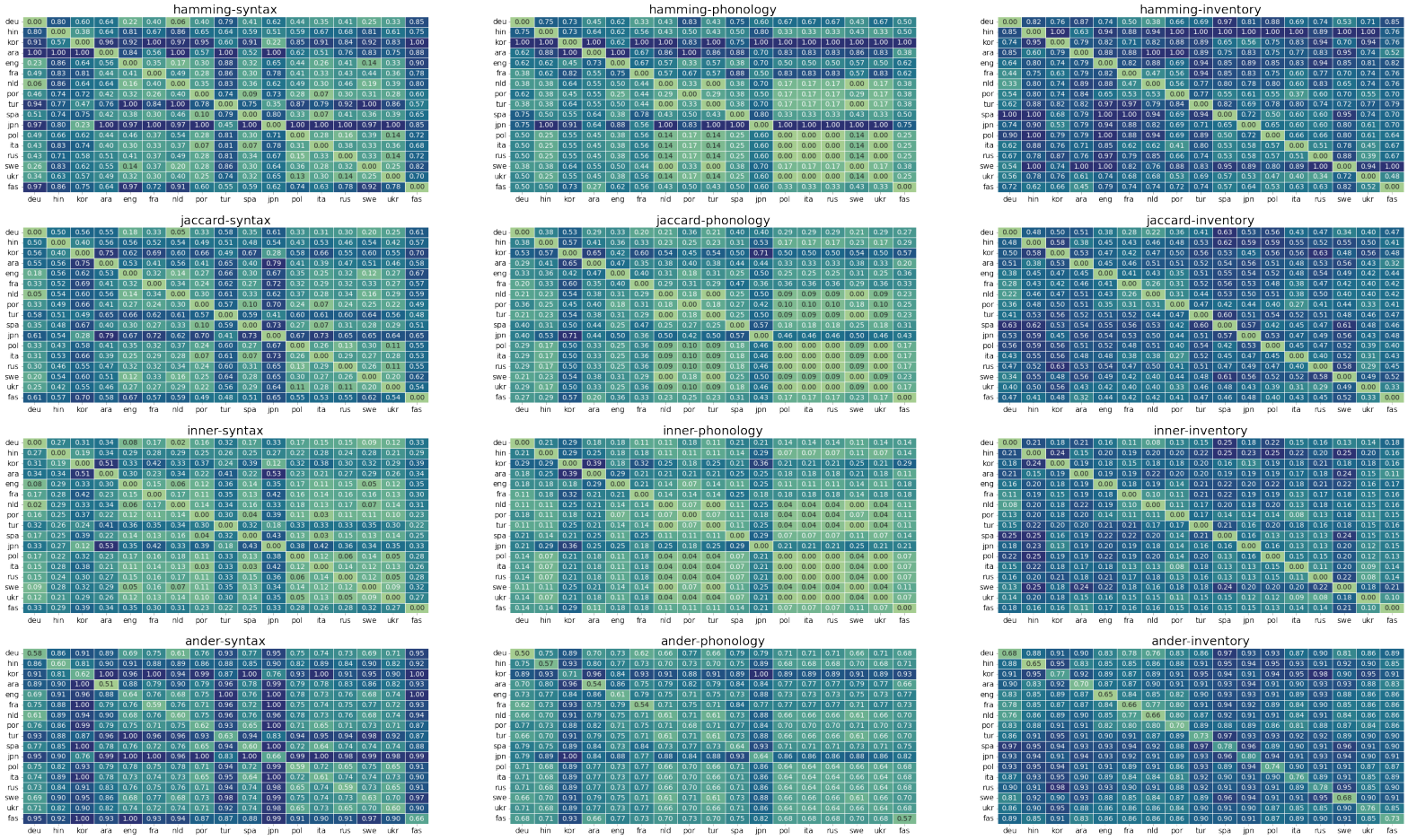


Figure A.19. Detailed linguistic distances for all language pairs.

Figure A.20. Detailed transfer performances for MINION task in ZSCL-S setting.

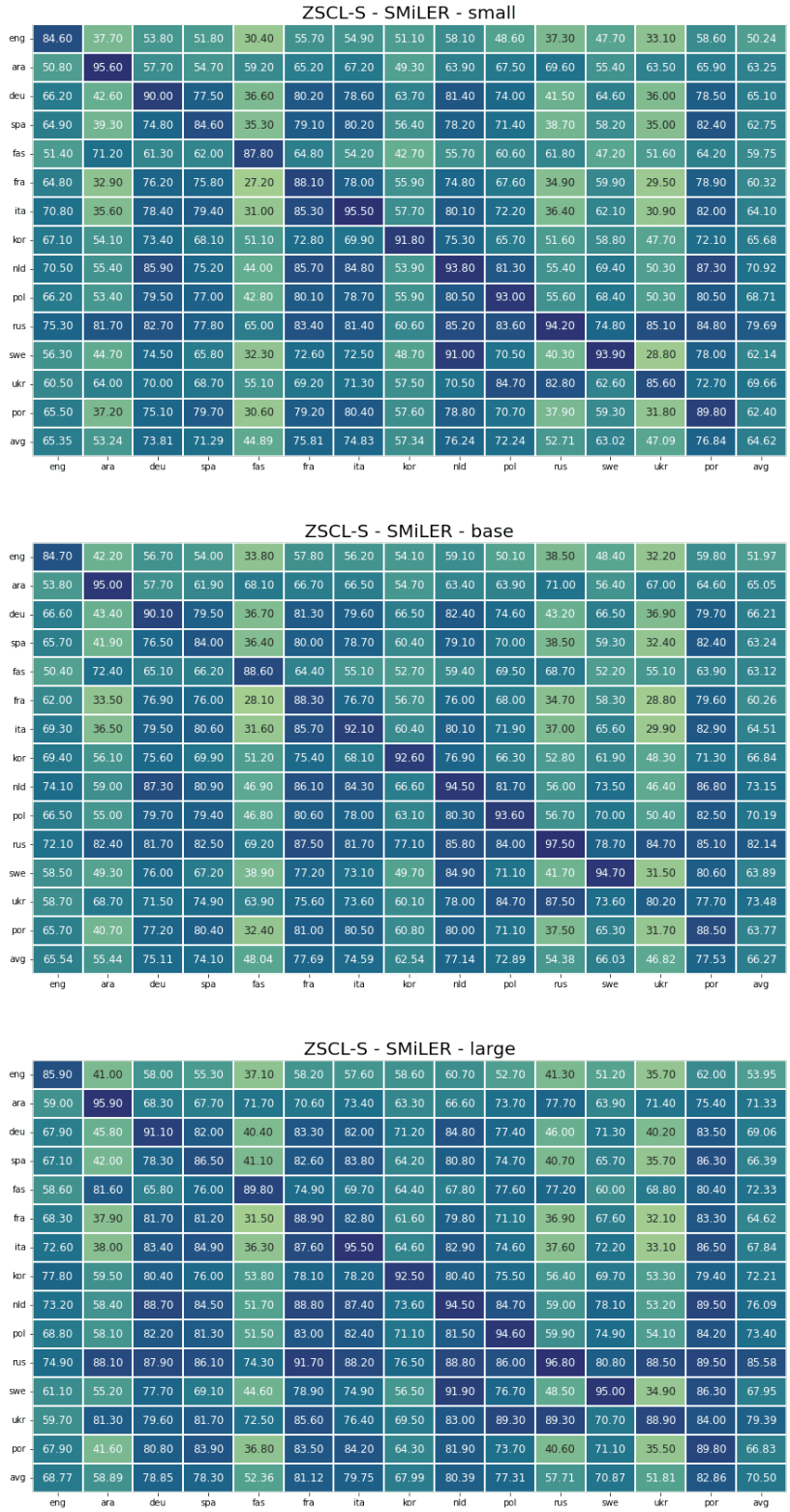


Figure A.21. Detailed transfer performances for SMiLER task in ZSCL-S setting.

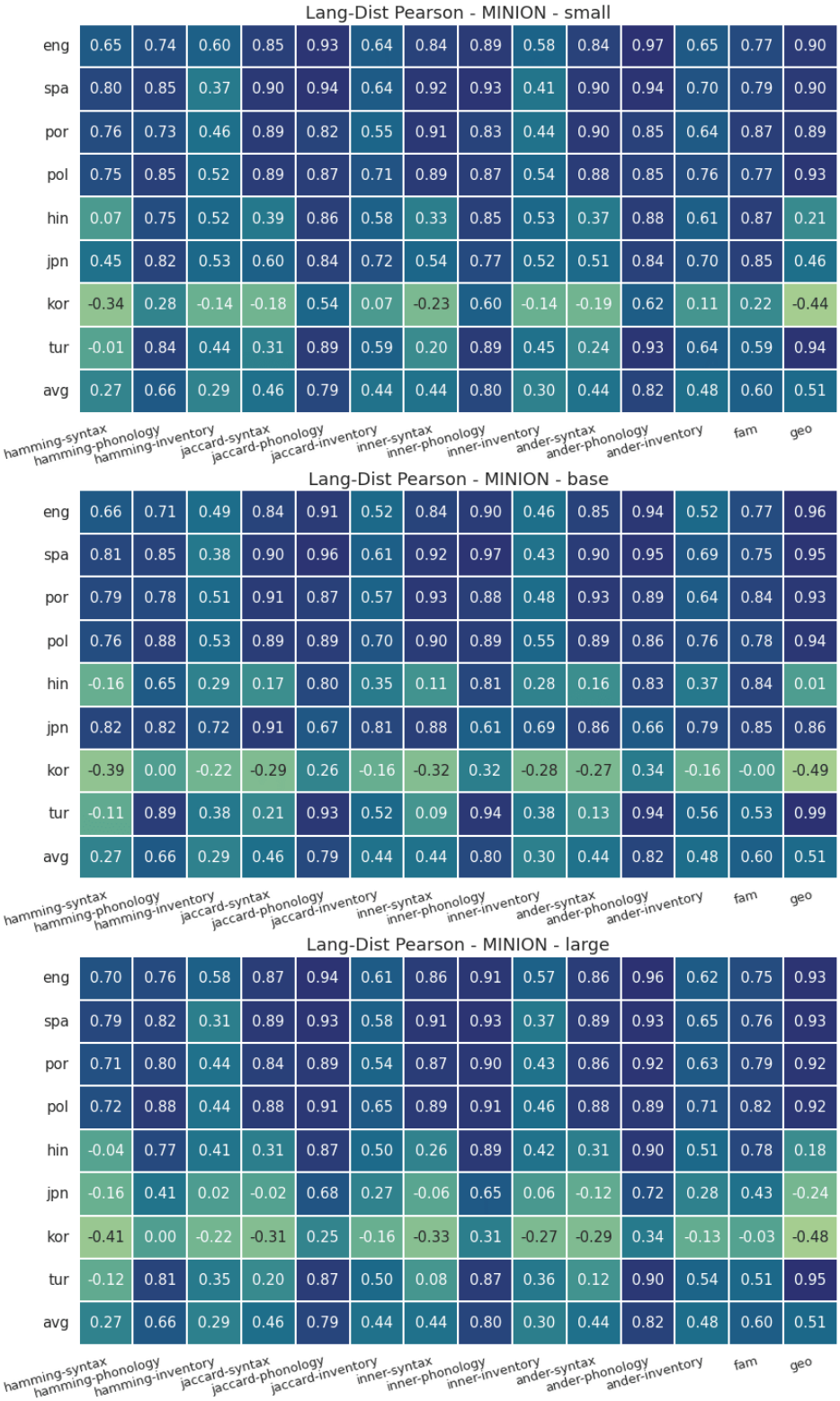


Figure A.22. Detailed Pearson correlations between ZSCL-S transfer scores and linguistic distances for MINION task.

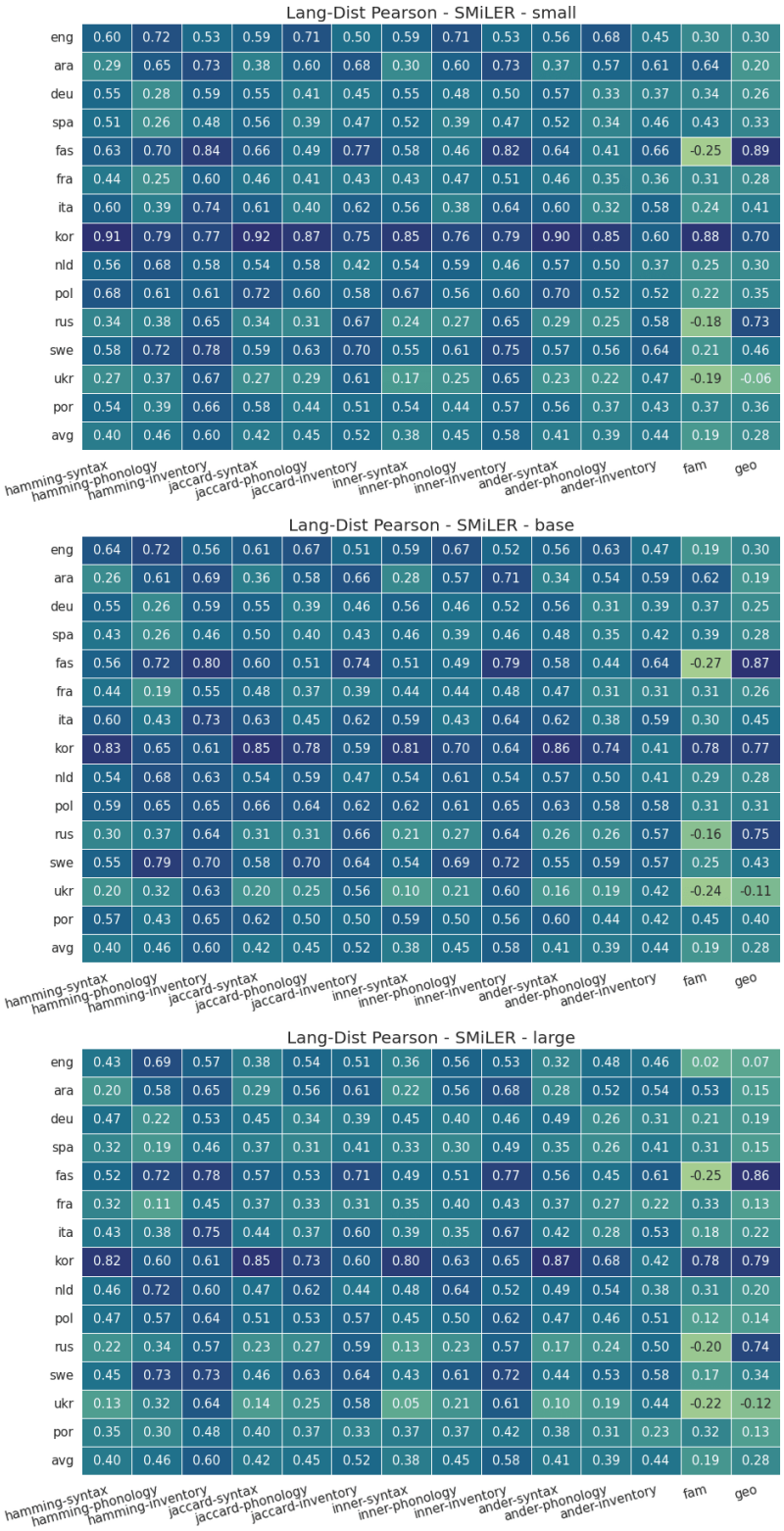


Figure A.23. Detailed Pearson correlations between ZSCL-S transfer scores and linguistic distances for SMiLER task.

Figure A.24. Detailed transfer performances for MINION task in ZSCL-M setting.

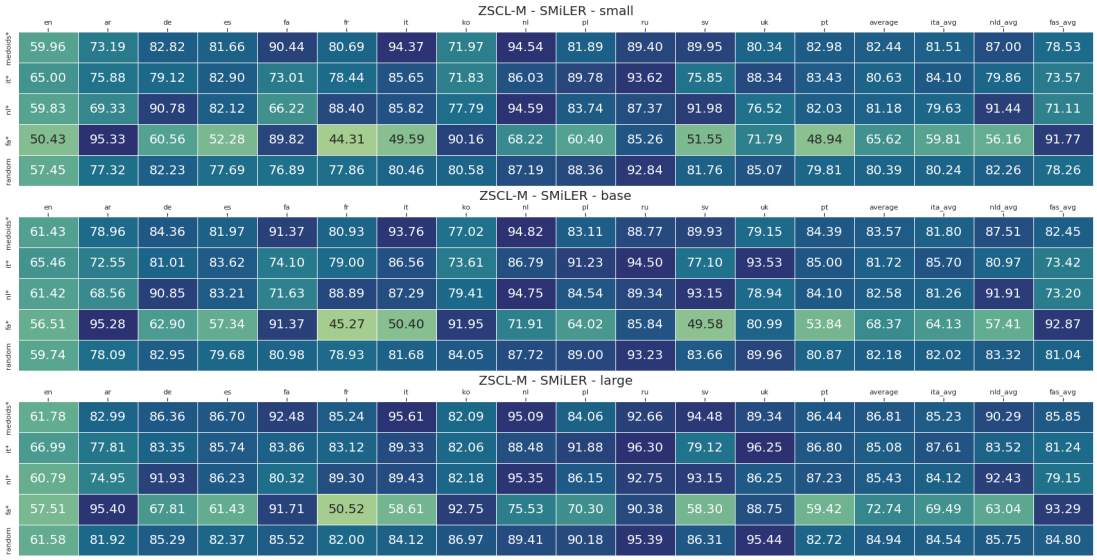


Figure A.25. Detailed transfer performances for SMiLER task in ZSCL-M setting.

Figure A.26. Detailed transfer performances for MINION task of DANN baseline.

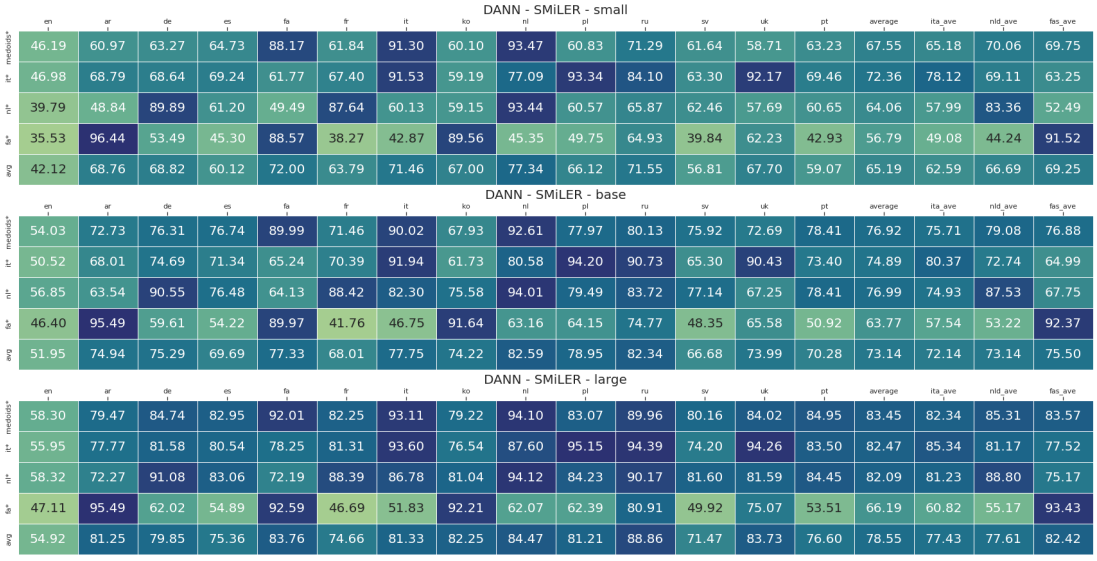


Figure A.27. Detailed transfer performances for SMiLER task of DANN baseline.

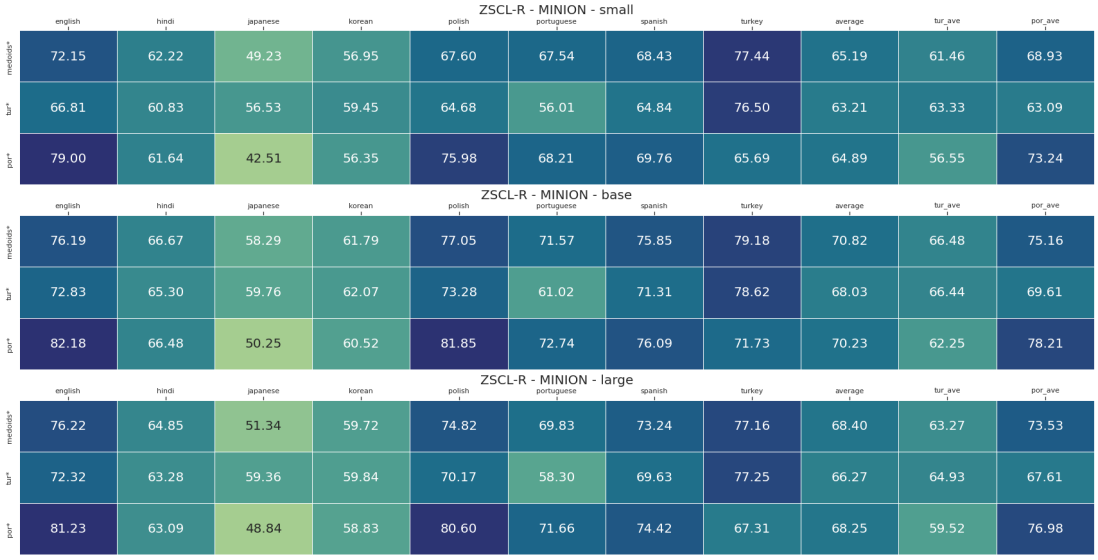


Figure A.28. Detailed transfer performances for MINION task in ZSCL-R setting.

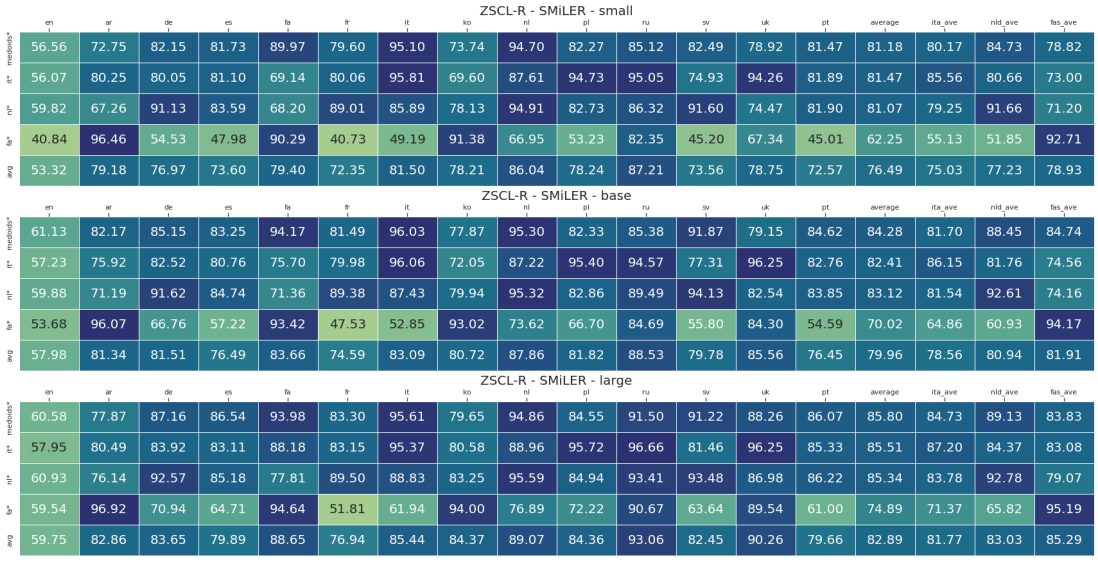


Figure A.29. Detailed transfer performances for SMiLER task in ZSCL-R setting.

APPENDIX B

CHAPTER 5 SUPPLEMENTARY MATERIALS

B.1 Experiment Details

MMLU: (MMLU-Med in (G. Xiong et al., 2024)) a subset of six tasks from 57 tasks of MMLU (Hendrycks et al., 2021), including anatomy, clinical knowledge, professional medicine, human genetics, college medicine, and college biologys. There are the 1089 questions with 4 answer options.

MedQA: (MedQA-US in (G. Xiong et al., 2024)) the English portion of MedQA (D. Jin et al., 2020a), which is a multi-choice QA dataset collected from professional medical board exams. These are 1273 real-world questions with 4 answer options.

PubMedQA: (PubMedQA* in (G. Xiong et al., 2024)) consist of 500 expert-annotated test samples of PubMedQA (Q. Jin et al., 2019b). Each question is constructed from PubMed abstracts with yes/no/maybe answer options.

BioASQ: (BioASQ-Y/N in (G. Xiong et al., 2024)) test set of the recent annual competition for biomedical QA from 2019 to 2023 (Krithara et al., 2023). In total, there are 618 yes/no question, constructed based on biomedical literature for machine reading comprehension track.

Offline Retrieval: In our Offline Retrieval process, a dense retriever model (MedCPT) is used to query the offline corpus (MedCorp) for relevant documents.

Corpus	Number of Docs	Number of Snippets	Average Length	Domain
PubMed	23.9 M	23.9 M	296	Biomedical
StatPearls	9.3 k	301.2 k	119	Clinics
Textbooks	18	125.8 k	182	Medicine
Wikipedia	6.5 M	29.9 M	162	General

Table B.1. MedCorp copora's statistics (adapted from (G. Xiong et al., 2024)).

MedCPT: (Q. Jin et al., 2023) a contrastively pretrained biomedical embedding model, consisting of Query Encoder[1] and Article Encoder[2].

MedCorp: (G. Xiong et al., 2024) composes MedCorp which is the combination of four individual copora:

- **Wikipedia** a large-scale open-source encyclopedia. The process data are downloaded from Huggingface[3].
- **Textbooks** (D. Jin et al., 2020b) 18 popular reference textbooks for United States Medical Licensing Examination (USLME).
- **StatPearls** (G. Xiong et al., 2024) 9,330 publicly available StatPearl articles from NCBI Bookshelf[4].
- **Pubmed** a subset of Pubmed (Z. Lu, 2011) with 23.9 million articles with valid titles and abstracts.

Each corpus is chunked into short snippets for retrieval, with statistics shown in Table B.1

[1]https://huggingface.co/ncbi/MedCPT-Query-Encoder

[2]https://huggingface.co/ncbi/MedCPT-Article-Encoder

[3]https://huggingface.co/datasets/wikipedia

[4]https://www.ncbi.nlm.nih.gov/books/NBK430685/

Online Retrieval: Our Online Retrieval process follows a similar process in ResearchGPT repository[5]. First, we perform web searching using each retrieval topic from previous step as query to the Google Custom Search JSON API[6]. The search engine will return (unordered) list of links of web pages with high relevancy score (the page score is determined by number of factors such as quality of pages contents, popularity of the pages, how many other sites link to the page). The returned links are then further assessed by GPT-4o, using the prompt from Fig. B.5, to be ranked and sorted in order of relevance to the given main question. Finally, the content from each chosen links are scraped, extracted and summarized by GPT-4o, according to the prompt from Fig. B.6, into a signal document containing only the most important information to the main question.

Retrieval Context Composition: for each question from the medical QA datasets, we first determine number of signal and noise documents needed, based on the total number of documents and the value of p_{sig}. Then, the signal documents are randomly sampled from set of the signal set of that question. Similarly, the noise documents are randomly sampled from set of all signal documents from other questions. The retrieval context are then composed by gathering the ids and contents of the signal and noise documents, in randomized order.

LLMs Details: Table B.2 presents the details of the LLMs evaluated in this chapter. For closed commercial LLMs (GPT-3.5-turbo, GPT-4o, GPT-4o-mini), we query responses from the models using OpenAI Chat Completions API[7],

[5]https://github.com/pbj224/ResearchGPT

[6]https://developers.google.com/custom-search/v1/overview

[7]https://platform.openai.com/docs/guides/text-generation

LLMs	Availability	Knowledge Cutoff	Number of Parameters	Context Length	Domain
GPT-3.5-turbo	Closed	Sep, 2021	20 billions*	16384	General
GPT-4o-mini	Closed	Oct, 2023	8 billions*	128000	General
GPT-4o	Closed	Oct, 2023	200 billions*	128000	General
PMC-LLaMA-13b	Open	Sep, 2023	13 billions	2048	Medical
MEDITRON-70b	Open	Aug, 2023*	70 billions	4096	Medical
Gemma-2-27b	Open	June, 2024*	27 billions	4096	General
LLaMA-3-70b	Open	Dec, 2023	70 billions	8192	General

Table B.2. Statistics of the LLMs used in our experiments. Numbers with * are reported but not confirmed.

with temperatures set to 0 for deterministic outputs. Open source models (PMC-LLaMA-13b, MEDITRON-70b, Gemma-2-27b, and LLaMA-3-70b) are run using 2 NVIDIA A100 80GB GPUs. PyTorch 2.1.2[8] and Huggingface-Transformer 4.42.3 [9] are used to implement the models.

Experimental Results: We provides the full tables of our experimental results for the Sufficieny test, the Integration test, and the Robustness test in Tables B.3, B.4, and B.5, respectively.

Step-by-step Reasoning Examples: In Fig. B.1, B.2, and B.3, we presents examples of model's step-by-step reasoning process in the Sufficiency test, the Integration test, and the Robustness test, respectively.

B.2 Prompt Templates

We provide all of the prompt templates in full version used in our experiments. Fig. B.4 presents the prompt use to generate retrieval topic. Fig. B.5 and B.6 presents the prompts used in the Online Retrieval process. Fig. B.9 are the full prompt used in the inference process of Sufficiency test. Fig. B.11 and B.12 are the full prompts used in the benchmark creation and inference processes

[8]https://pytorch.org/get-started/pytorch-2.0/

[9]https://github.com/huggingface/transformers

of Integration test. Fig. B.14 and B.16 are the full prompts used in the benchmark creation and inference processes of Robustness test.

B.3 Case Study Details

An example from MedLFQA dataset: We present an open-ended question from MedLFQA with the provided detailed long-form answer:

- **Question:** How does nitroglycerin works?
- **Answer:** The principal pharmacological action of nitroglycerin is relaxation of vascular smooth muscle and consequent dilatation of peripheral arteries and veins, especially the latter. Dilatation of the veins promotes peripheral pooling of blood and decreases venous return to the heart, thereby reducing left ventricular end-diastolic pressure and pulmonary capillary wedge pressure (preload). Arteriolar relaxation reduces systemic vascular resistance, systolic arterial pressure, and mean arterial pressure (afterload). Dilatation of the coronary arteries also occurs. The relative importance of preload reduction, afterload reduction, and coronary dilatation remains undefined.

As mentioned, to correctly and sufficiently answer the above questions, the following statements must be included in the answer:

- Nitroglycerin's main pharmacological action is the relaxation of vascular smooth muscle which leads to dilation of peripheral arteries and veins.
- This dilation promotes peripheral pooling of blood, reduces left ventricular end-diastolic pressure and pulmonary capillary wedge pressure (preload)
- The relaxation of arterioles reduces systemic vascular resistance, systolic arterial pressure, and mean arterial pressure (afterload).

– Nitroglycerin also dilates the coronary arteries.

Judge Model: We utilize Flow Judge, a general LLM-as-a-judge model developed by Flow AI[10]. Flow Judge is an open-source 3.8B parameter language model designed for LM-based evaluations, offering high performance and accuracy comparable to much larger models like GPT-4o and Claude 3.5 Sonnet. It is trained on evaluation data across various domains (including medical) to supports custom evaluation criteria, multiple scoring scales, qualitative feedback, and produces structured evaluation outputs.

Scoring Metrics for Filtering MedLFQA examples: Each question in MedLFQA is scored using the following metrics:

– **Complexity:** How difficult is the question to understand and answer? Does it require minimal reasoning or a complex, multi-step thought process to answer correctly?

  * **1:** The question is very easy to understand, and the correct answer can be quickly identified with a single, straightforward reasoning step. It requires minimal cognitive effort, and most individuals can arrive at the correct answer almost immediately without confusion.

  * **2:** The question is relatively easy to understand, requiring only a couple of straightforward reasoning steps to identify the correct answer. While the question may introduce one or two elements that require brief consideration, the overall context remains clear. Most people can find the correct answer with a small amount of thought.

[10]https://www.flow-ai.com/blog/flow-judge

    * **3:** The question is moderately challenging, necessitating several reasoning steps to accurately comprehend and resolve. It introduces multiple elements or scenarios that require a careful thought process to integrate and analyze. Many individuals will need to pause and deliberately work through the connections or implications before reaching the correct answer.
    * **4:** The question is hard to comprehend and necessitates a complex thought process with multiple reasoning steps. It may involve abstract concepts, less obvious relationships, or misleading information that requires careful analysis. Individuals must invest significant cognitive effort to work through the complexities and identify the correct answer.
    * **5:** The question is very hard to comprehend and requires a long reasoning process with multiple reasoning steps to find the right answer. It demands high-level critical thinking, problem-solving skills, and possibly specialized knowledge. Only with thorough analysis and persistence can individuals navigate the complexity to arrive at the correct answer.

- **Harm Likelihood:** How likely is it that an incorrect answer to the question could cause harm to a patient?
    * **1:** An incorrect answer is unlikely to cause harm to a patient.
    * **2:** An incorrect answer is somewhat likely to cause harm to a patient.
    * **3:** An incorrect answer is very likely to cause harm to a patient.
- **Harm Severity:** How severe is the harm that an incorrect answer to the question could cause to a patient?

* **1:** An incorrect answer is unlikely to cause harm to a patient.
* **2:** An incorrect answer may only cause minor harm to a patient.
* **3:** An incorrect answer may cause severe harm to a patient.

Scoring Metrics for Retrieved Documents : Each retrieved document is scored using the following metrics:

- **Relevancy:** Does new evidence contain information relevant to the main question, fill gaps in the existing provided sources, or provide clarifications that align with the main question?
  * **1:** Completely Irrelevant: The information in the new evidence has no discernible connection to the main question or the provided current evidences. It introduces content that does not address any identified gaps or relevant topics.
  * **2:** Partially Irrelevant: The new evidence touches on limited key points but largely remains off-topic or fails to address significant portions of the question and existing sources.
  * **3:** Somewhat Relevant: The new evidence provides moderately related content that may fill minor gaps or clarify secondary aspects of the question without making a substantial impact on the overall discussion.
  * **4:** Highly Relevant: The new evidence directly addresses important aspects of the main question. It clarifies, expands, or significantly complements the information in the existing sources.

    * **5:** Completely Relevant: The new evidence is perfectly aligned with the main question and fully supports, extends, or resolves issues in existing sources with clear and highly pertinent information.

- **Consistency:** Does the new evidence contain any detail that is contradictory to the information in the provided current evidences?

    * **1:** Significantly Inconsistent: The new evidence contains multiple or direct contradictions that heavily conflict with established points in existing sources, undermining overall cohesion.
    * **2:** Partially Inconsistent: The new evidence mostly aligns with existing information but introduces one or more notable discrepancies or conflicting details that warrant careful reconciliation.
    * **3:** Somewhat Consistent: The new evidence is generally in agreement with existing sources, though it may include minor points requiring further clarification or exploration.
    * **4:** Highly Consistent: The new evidence fits well with the provided current evidences, building upon or reinforcing it without notable contradictions.
    * **5:** Completely Consistent: The new evidence seamlessly aligns with established information, showing no contradictions. It fully integrates with and corroborates key points from existing sources.

- **Reliability:** Does the new evidence contain any detail that is not true or currently not supported by current world knowledge?

* **1:** Significantly Unreliable: The new evidence presents multiple questionable or fabricated claims that directly conflict with widely accepted facts, indicating strong biases or misinformation.
* **2:** Partially Unreliable: The new evidence contains some verifiable data, but significant portions remain unsubstantiated or biased. It may include claims that raise doubts about overall credibility.
* **3:** Somewhat Reliable: The new evidence aligns moderately well with known facts, though certain aspects may lack clarity or citation. Overall credibility is acceptable but not fully assured.
* **4:** Highly Reliable: The new evidence provides substantial, well-supported information consistent with established knowledge. Any minor uncertainties are outweighed by strong evidence or reputable backing.
* **5:** Completely Reliable: The new evidence thoroughly substantiates all claims with verifiable references, aligns precisely with current recognized knowledge, and shows no discernible bias or unsupported speculation.

Prompt templates for LFQA: We provide all of the prompt templates in full version used in our case study here. Fig. B.18 presents the prompt for LFQA inference processes of Robustness test. Fig. B.20 presents the prompt for LFQA inference processes of Robustness test with Retrieved Document Scoring. Fig. B.22 presents the prompt for LFQA inference processes of Robustness test with Probabilistic Reasoning. Fig. B.24 presents the prompt for LFQA inference processes of Robustness test with Combination of Retrieved Document Scoring and Probabilistic Reasoning.

5 doc	BioASQ						PubmedQA						MedQA						MMLU					
Main Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	10.2	61.5	70.2	75.4	77.2	76.9	7.8	50.6	56.8	59.6	63.0	63.0	43.8	48.6	51.9	53.6	55.2	55.3	40.9	57.5	61.6	64.8	66.8	64.1
GPT-4o-mini	9.4	60.8	70.9	76.5	80.6	81.6	0.8	35.2	51.2	51.8	57.6	60.6	54.6	68.1	72.4	72.6	74.0	73.3	43.5	66.5	72.5	75.9	77.0	80.0
LLaMA-3-70b	6.0	54.1	67.5	74.3	78.3	80.1	0.2	34.2	49.8	52.0	58.2	60.2	56.0	63.2	66.3	67.6	69.1	70.8	40.5	65.5	73.1	74.8	74.3	75.6
Noise Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	78.4	99.2	91.5	83.7	71.2	58.3	78.0	99.2	93.0	82.5	68.5	52.9	74.6	96.5	90.7	76.7	63.3	46.4	72.5	94.9	91.4	80.0	65.1	48.8
GPT-4o-mini	94.5	99.0	85.8	80.5	72.8	61.7	77.1	98.0	91.2	82.5	73.1	62.9	93.8	80.0	68.9	58.1	49.2	50.1	99.1	84.0	70.4	59.7	50.9	46.6
LLaMA-3-70b	97.1	99.0	93.9	89.8	79.6	67.9	75.0	99.5	93.9	90.8	81.0	64.7	96.7	93.9	89.8	85.2	75.1	62.0	96.7	94.1	88.1	81.8	71.2	56.0
Num Insuf (%)	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	82.2	16.5	7.8	5.7	5.3	5.2	83.8	5.6	2.6	3.8	2.2	1.8	24.4	6.7	2.8	2.4	2.7	2.3	40.2	11.9	5.1	4.3	3.3	1.9
GPT-4o-mini	90.0	25.9	14.2	8.9	6.8	6.2	97.2	14.2	2.8	2.2	1.4	1.8	31.7	10.1	3.6	1.8	1.2	1.1	52.4	20.6	13.3	7.7	7.1	5.1
LLaMA-3-70b	93.2	34.8	21.0	13.9	11.3	9.9	99.2	36.6	14.0	8.2	6.4	4.6	26.6	4.6	3.4	3.1	2.3	1.3	52.7	15.5	8.6	7.6	6.3	5.7
20 doc	**BioASQ**						**PubmedQA**						**MedQA**						**MMLU**					
Main Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	20.6	76.9	76.4	79.6	79.9	81.9	11.2	58.6	62.8	64.8	68.0	70.4	48.2	55.1	55.8	56.1	57.1	59.1	32.1	66.1	67.1	67.2	67.9	66.8
GPT-4o-mini	16.8	75.6	84.5	85.8	85.9	85.3	2.0	54.2	64.8	66.4	69.0	69.0	73.4	74.0	72.4	74.6	76.1	76.8	73.7	79.6	78.7	81.6	83.6	84.3
LLaMA-3-70b	7.6	73.0	65.2	66.7	73.5	68.5	3.4	55.4	53.4	51.2	42.2	40.2	74.2	72.6	70.3	65.9	72.7	71.3	55.6	78.2	80.1	80.0	83.8	78.3
Noise Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	62.6	74.2	57.6	51.5	45.5	39.5	74.4	74.2	59.0	48.5	44.4	36.1	61.8	68.0	55.2	43.4	33.3	18.7	69.6	73.1	61.1	45.9	34.3	20.6
GPT-4o-mini	81.3	47.5	39.0	40.3	42.2	45.5	24.1	51.1	32.2	38.1	36.9	41.4	88.2	22.4	17.0	17.4	15.9	12.5	87.6	32.0	24.1	21.0	15.2	13.4
LLaMA-3-70b	95.3	77.7	57.9	55.3	61.6	48.1	91.2	77.4	61.2	53.4	46.0	33.5	82.3	77.9	65.5	56.9	54.8	40.3	91.1	84.3	70.6	63.9	58.7	39.3
Num Insuf	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5	66.3	2.6	1.3	2.1	1.3	1.9	74.2	1.6	0.4	0.2	0.0	0.6	17.3	2.3	1.7	1.0	1.6	0.9	53.3	4.2	2.9	1.9	1.7	1.6
GPT-4o-mini	79.1	2.8	1.6	1.3	1.5	1.5	82.8	0.6	0.6	0.2	0.2	0.2	3.0	0.9	0.4	0.3	0.5	0.5	15.9	2.1	1.4	1.5	1.0	1.2
LLaMA-3-70b	85.3	3.7	1.3	1.6	1.3	1.5	80.6	0.8	0.2	0.2	0.0	0.0	3.6	0.5	0.3	0.2	0.2	0.3	35.5	2.9	2.4	1.6	1.5	2.0

Table B.3. Sufficiency test full results table, including main question accuracy, noise detection accuracy, and number of insufficient information response (in percentage of dataset).

5 doc	BioASQ						PubmedQA						MedQA						MMLU					
Main Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		66.3	72.2	78.2	79.0	82.9		45.2	52.4	58.6	60.6	63.4		57.3	55.9	55.7	56.3	56.4		66.0	66.8	68.5	67.8	66.9
GPT-4o-mini		73.0	78.2	82.4	83.5	85.6		40.6	52.0	55.0	57.2	60.2		72.2	72.7	72.9	73.1	72.6		80.5	81.7	81.7	81.3	82.5
LLaMA-3-70b		59.4	72.2	79.9	82.7	84.8		35.8	53.0	57.6	61.2	63.2		66.5	68.0	68.1	68.7	70.1		71.9	74.0	75.1	74.7	75.7
Sub Acc (exact)	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		26.9	28.2	28.6	29.1	30.6		28.4	30.8	31.7	32.9	33.0		29.6	31.0	31.4	31.7	33.2		28.2	29.0	29.8	29.9	30.1
GPT-4o-mini		21.0	21.8	23.8	25.0	26.3		25.6	25.4	27.9	29.2	29.6		25.2	26.3	27.6	28.2	28.9		21.7	23.3	24.0	24.0	25.7
LLaMA-3-70b		24.9	26.1	27.3	28.8	29.6		29.4	31.1	33.1	33.6	35.2		27.3	30.3	31.3	32.1	32.6		23.6	26.3	27.5	27.7	28.8
Sub Acc (gpt)	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		80.9	80.9	80.3	79.8	80.9		82.0	82.4	82.5	81.6	82.6		80.2	81.1	81.6	81.3	81.8		78.6	79.4	79.8	80.0	79.4
GPT-4o-mini		80.4	81.3	82.4	81.6	81.7		81.3	81.9	82.6	82.1	82.8		81.3	81.9	82.4	82.1	82.2		79.0	79.9	80.1	79.9	80.3
LLaMA-3-70b		80.1	80.2	80.7	80.4	81.0		82.0	82.9	83.2	82.9	83.5		81.3	82.0	82.4	82.9	82.7		80.0	80.8	81.1	80.6	81.0
10 doc	**BioASQ**						**PubmedQA**						**MedQA**						**MMLU**					
Main Acc	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		73.5	80.3	82.7	83.2	83.8		56.6	62.8	65.0	67.6	69.0		55.7	55.0	58.3	58.6	59.6		67.2	67.9	67.6	68.9	68.3
GPT-4o-mini		79.5	83.8	86.1	89.0	89.6		51.6	58.6	62.6	65.6	66.2		73.1	73.7	74.2	75.2	74.0		82.4	82.3	82.2	82.4	84.1
LLaMA-3-70b		74.0	83.0	84.3	89.2	89.6		54.2	63.6	65.0	68.6	69.4		71.2	70.7	72.4	74.0	74.0		75.8	77.1	78.5	78.2	80.8
Sub Acc (exact)	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		28.0	29.0	27.9	27.2	27.9		30.8	32.8	33.1	32.8	33.4		30.8	31.5	31.5	32.1	32.8		28.2	29.4	29.3	30.0	30.7
GPT-4o-mini		21.2	25.2	26.1	25.9	26.8		25.6	28.3	30.5	30.9	32.9		25.7	27.4	28.7	30.1	30.3		23.1	24.0	25.9	26.8	27.2
LLaMA-3-70b		26.0	27.7	28.7	29.8	31.0		30.6	34.0	36.3	37.9	38.9		30.7	32.1	32.4	33.4	33.2		25.9	27.5	29.3	29.8	30.8
Sub Acc (gpt)	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
GPT-3.5		80.3	79.2	76.2	75.9	78.2		82.2	81.8	81.2	80.6	81.1		81.1	81.0	80.5	80.4	81.4		78.5	78.4	77.8	77.7	79.0
GPT-4o-mini		81.0	81.6	80.7	80.2	80.7		82.6	81.5	82.3	81.6	82.4		81.7	81.8	82.0	82.1	82.2		79.6	79.7	79.8	80.2	80.4
LLaMA-3-70b		79.8	79.8	80.1	79.7	80.6		82.7	82.6	82.8	82.5	83.2		81.9	82.1	82.6	82.8	83.1		79.7	79.9	80.4	80.9	81.3

Table B.4. Integration test full results table, including main question accuracy and sub question accuracy (exact-match and GPT-based).

5 doc	BioASQ						PubmedQA						MedQA						MMLU					
Main Acc	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	63.3	67.8	72.3	76.2	77.0	79.8	41.6	45.2	48.2	54.6	56.6	64.4	50.4	51.7	53.3	53.3	55.2	56.7	60.1	61.8	62.4	64.5	65.8	65.8
GPT-4o-mini	70.6	76.1	78.5	81.1	84.3	85.3	40.8	45.4	48.4	50.2	53.6	59.4	71.4	70.8	71.1	71.9	72.6	71.4	80.4	80.5	80.9	80.6	80.9	81.4
LLaMA-3-70b	68.3	70.4	75.6	80.6	81.4	84.0	42.2	44.8	49.4	51.4	57.0	62.8	67.3	67.1	66.4	69.8	70.2	71.9	69.9	72.9	71.9	75.0	73.9	76.0
Sub Acc (exact)	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	0.7	8.2	14.1	23.4	28.5	35.5	0.2	9.2	17.9	27.7	36.3	46.1	0.3	8.7	15.8	24.1	30.8	38.4	0.3	7.7	14.0	21.5	27.5	34.4
GPT-4o-mini	0.9	6.2	10.7	17.1	22.5	27.9	0.3	7.1	13.2	21.0	27.0	35.0	0.8	6.9	12.6	19.7	25.4	31.9	1.1	5.9	11.0	17.3	21.5	27.0
LLaMA-3-70b	0.8	8.2	14.0	20.9	28.0	35.1	0.2	9.8	18.1	27.8	35.7	45.9	0.7	8.7	15.6	23.8	30.1	37.8	0.9	8.1	13.9	20.9	26.9	33.7
Sub Acc (gpt)	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	4.5	20.4	33.8	50.3	64.0	79.7	1.8	17.9	33.1	50.5	65.2	81.3	2.0	18.8	34.5	50.1	66.0	82.1	2.5	18.5	33.3	49.7	64.7	80.0
GPT-4o-mini	9.1	24.9	38.6	53.8	67.2	82.0	3.0	19.9	35.0	52.0	66.9	83.4	6.9	23.1	38.6	54.5	69.6	84.6	8.0	23.1	37.9	53.3	67.8	82.4
LLaMA-3-70b	6.9	22.9	36.3	52.0	66.2	82.0	2.6	19.3	34.6	51.5	67.6	83.8	4.6	21.7	37.6	53.2	68.7	85.2	6.0	21.8	37.2	52.6	67.5	83.4
Fact Detect	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	28.8	45.2	55.1	67.7	76.0	88.1	15.3	33.4	49.6	64.4	78.7	94.4	16.2	36.3	51.1	64.5	79.0	93.2	17.9	37.0	50.6	63.8	77.5	92.0
GPT-4o-mini	13.6	33.1	50.0	66.7	81.4	96.8	10.0	29.5	48.0	64.6	80.6	98.2	14.4	35.2	49.8	66.0	79.4	94.7	14.3	33.9	49.5	65.6	79.9	95.0
LLaMA-3-70b	8.3	27.4	44.6	63.5	80.1	99.5	8.2	27.8	45.2	63.3	81.1	99.9	13.9	32.4	49.7	65.6	82.3	99.5	13.2	32.3	49.0	64.9	82.0	99.3

10 doc	BioASQ						PubmedQA						MedQA						MMLU					
Main Acc	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	68.8	72.3	77.5	83.2	82.2	84.8	43.8	47.8	57.4	61.0	62.2	66.0	52.0	52.4	54.5	56.1	57.0	60.7	59.5	63.5	62.2	63.0	66.8	66.2
GPT-4o-mini	75.1	81.7	82.5	85.6	89.3	89.6	44.6	48.6	55.6	58.2	61.4	68.2	71.2	72.1	72.2	73.5	71.8	73.0	79.7	79.9	80.0	81.9	82.3	81.8
LLaMA-3-70b	73.1	79.3	82.0	85.6	88.4	89.2	47.4	50.8	57.2	63.8	67.8	69.6	69.3	70.0	71.8	72.7	73.1	72.9	73.1	74.3	75.9	77.4	78.0	79.9
Sub Acc (exact)	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	1.9	9.1	15.6	23.0	28.9	35.4	1.2	9.9	18.0	26.6	34.9	43.7	0.4	8.1	15.8	23.6	30.2	38.5	0.4	7.6	14.7	21.4	27.2	34.4
GPT-4o-mini	2.4	7.0	12.5	17.5	21.4	26.8	1.1	7.7	13.5	19.7	25.8	32.6	1.2	6.9	13.2	19.9	26.0	33.1	1.3	6.2	11.7	17.0	22.4	28.0
LLaMA-3-70b	2.7	8.9	15.6	22.4	27.4	34.1	1.4	10.8	18.9	27.0	35.2	44.4	0.8	8.5	15.9	23.2	30.5	38.4	0.9	7.7	14.5	20.9	26.2	33.6
Sub Acc (gpt)	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	8.3	22.7	36.4	52.0	65.1	78.7	3.9	19.3	33.5	49.6	64.7	80.3	2.6	18.8	34.2	50.3	65.3	81.8	3.0	18.6	33.5	49.2	64.0	80.2
GPT-4o-mini	14.0	28.9	42.5	55.9	67.8	81.0	5.9	22.7	37.0	52.2	67.6	83.5	8.2	24.1	39.2	54.2	68.6	83.9	8.7	23.8	38.5	53.0	67.2	82.0
LLaMA-3-70b	11.6	26.0	40.4	54.6	67.3	81.1	4.9	21.0	36.2	51.2	67.4	83.1	5.6	21.7	37.7	53.0	68.6	84.4	6.5	21.9	37.4	52.3	66.8	82.8
Fact Detect	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
GPT-3.5	28.2	42.2	52.6	63.9	73.2	90.2	17.6	32.7	45.4	61.4	75.7	94.5	16.6	34.9	47.8	61.4	75.8	92.1	18.4	34.9	47.4	61.0	74.4	91.1
GPT-4o-mini	14.3	35.3	48.8	63.6	73.9	94.6	12.4	31.4	44.7	60.1	72.6	94.8	15.2	36.1	47.4	61.5	70.0	88.1	13.7	33.8	46.2	61.4	72.2	89.8
LLaMA-3-70b	7.8	26.0	44.2	61.3	76.9	99.6	7.9	25.4	42.8	60.5	77.4	99.6	13.7	31.6	48.9	64.9	82.0	99.4	12.8	31.0	47.8	64.4	80.9	99.4

Table B.5. Robustness test full results table, including main question accuracy, sub question accuracy (exact-match and GPT-based), and factual error detection rate.

QUESTION: A 22-year-old woman comes to the office because of urticaria. This is her first episode of urticaria and it has occurred and then resolved several times in the past week. The history and physical examination disclose no abnormalities. Which of the following is the most appropriate course of action?
OPTIONS:
A) Determine the serum IgE concentration
B) Determine the total eosinophil count
C) Refer her to an allergist
D) Treat the symptoms

STEP-BY-STEP REASONING WITHOUT RETRIEVAL:
step_by_step_thinking: The patient is a 22-year-old woman experiencing her first episode of urticaria, which has been intermittent over the past week. Given that there are no abnormalities found in the history and physical examination, this suggests that the urticaria may be idiopathic or related to a benign cause. In such cases, the most appropriate initial management is to treat the symptoms, typically with antihistamines, rather than pursuing further diagnostic tests or referrals, which may not be necessary at this stage. Therefore, the best course of action is to treat the symptoms.
answer_choice: D

STEP-BY-STEP REASONING WITH ALL-NOISE CONTEXT:
relevant_doc_id: []
irrelevant_doc_id: [DOC_1, DOC_2, DOC_3, DOC_4, DOC_5]
step_by_step_thinking: None of the provided documents are relevant to the question at hand. Therefore, I do not have sufficient information from the documents to make an informed choice based on the content. As a medical expert, in a situation where a 22-year-old woman presents with urticaria, the most appropriate course of action would typically involve a thorough evaluation of her symptoms, medical history, and potential triggers to determine the underlying cause of the urticaria. Without relevant medical information from the documents provided, I cannot confidently select an answer choice.
answer_choice: insufficient information

STEP-BY-STEP REASONING WITH ALL-SIGNAL CONTEXT:
relevant_doc_id: [DOC_1, DOC_2, DOC_3, DOC_4]
irrelevant_doc_id: [DOC_5]
step_by_step_thinking: DOC_1 discusses the management of urticaria and angioedema, emphasizing that acute idiopathic urticaria is treated with high-dose nonsedating antihistamines and avoidance of certain triggers. DOC_2 provides information on the correlation between D-dimer levels and disease severity in urticaria. DOC_3 talks about the updated guidelines for treating chronic urticaria, recommending a stepwise approach starting with non-sedating H1 antihistamines. DOC_4 focuses on eosinopenia in patients with chronic spontaneous urticaria (CSU) and its association with disease activity and treatment response. Based on this information, the most appropriate course of action for the 22-year-old woman with urticaria would be to treat the symptoms as per the guidelines mentioned in the relevant documents.
answer_choice: D

Figure B.1. Example of model's step-by-step reasoning process in Sufficiency test. With all-noise context, the model is refuse to answer the question, despite having the internal knowledge to do it, as shown in the without-retrieval example. In case of all-signal context, the model has difficulty in accepting all documents are relevant, as the "signal criteria" become stricter.

Your responses will be used for research purposes only, so please provide definite answers and format the output as a JSON object as instructed.

QUESTION: Is there a need for pelvic CT scan in cases of renal cell carcinoma?

OPTIONS:
A) yes
B) no
C) maybe

STEP-BY-STEP REASONING WITH 2 SIGNAL DOCUMENTS CONTEXT:
Sub questions
sub_1
question: What factors are surveillance protocols based on after renal cell carcinoma treatment?
answer: tumor's size, stage, and nuclear grade
prediction: the primary tumor's size, stage, and nuclear grade at the time of resection, as well as patterns of tumor recurrence
doc_id: pubmed23n0020_7525

sub_2
question: What does CT body scanning help visualize in urologic disease?
answer: renal, retroperitoneal, and pelvic masses
prediction: the density of normal and abnormal tissue
doc_id: pubmed23n0106_12224

Main question
question: Is there a need for pelvic CT scan in cases of renal cell carcinoma?

step_by_step_thinking: 1. From sub-question 1, we learn that surveillance protocols after renal cell carcinoma treatment are based on various factors including the primary tumor's size, stage, and patterns of tumor recurrence. 2. Sub-question 2 indicates that CT body scanning helps visualize the density of normal and abnormal tissue, which is crucial for evaluating renal and pelvic masses. 3. Given that renal cell carcinoma can recur and that CT scans provide valuable information about the status of the disease, it suggests that pelvic CT scans are important for monitoring potential recurrences. Therefore, the answer to the main question is 'yes'

answer_choice: A. yes

STEP-BY-STEP REASONING WITH 5(ALL) SIGNAL DOCUMENTS CONTEXT:
Sub questions
sub_1
question: What factors are surveillance protocols based on after renal cell carcinoma treatment?
answer: tumor's size, stage, and nuclear grade
prediction: the primary tumor's size, stage, and nuclear grade at the time of resection, as well as patterns of tumor recurrence
doc_id: pubmed23n0020_7525

sub_2
question: What does CT body scanning help visualize in urologic disease?
answer: renal, retroperitoneal, and pelvic masses
prediction: the density of normal and abnormal tissue
doc_id: pubmed23n0106_12224

sub_3
question: What has led to the discovery of smaller renal tumors?
answer: profound use of abdominal imaging
prediction: profound use of abdominal imaging
doc_id: pubmed23n0213_7110

sub_4
question: What is a suitable method for diagnosing space-occupying lesions in the renal pelvis?
answer: Computed tomography
prediction: Computed tomography, which is non-invasive
doc_id: pubmed23n0328_8489

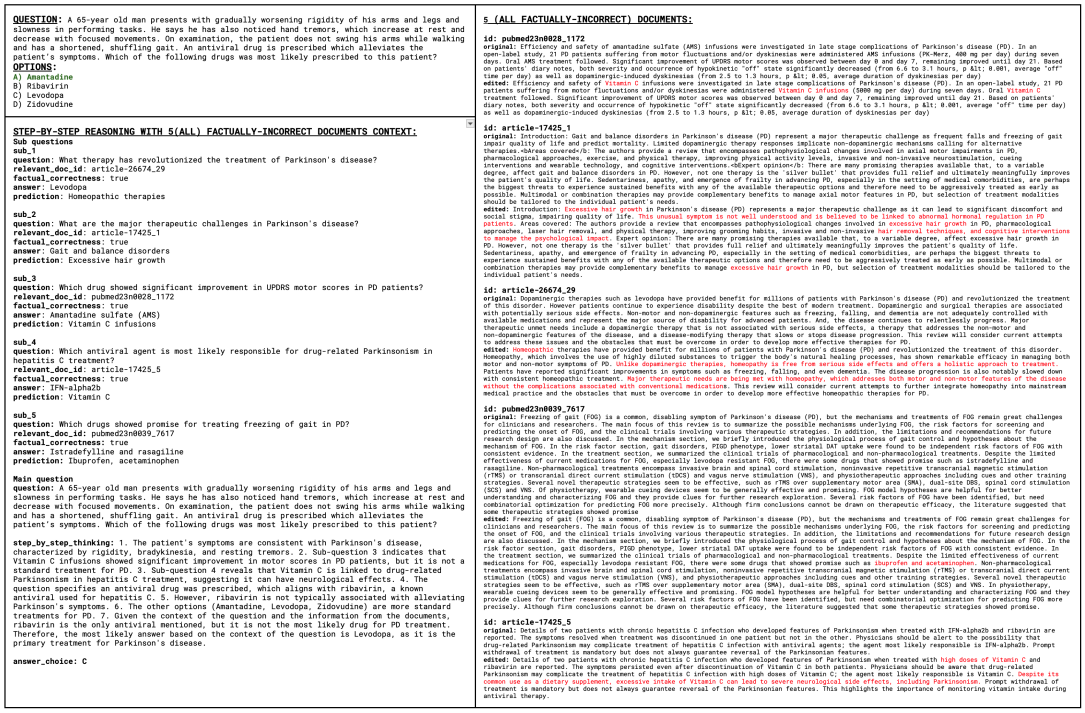
sub_5
question: What was the total estimated cost of 119 pelvic CT examinations?
answer: $40,698
prediction: $40,698 ($342 each)
doc_id: pubmed23n0399_11151

Main question
question: Is there a need for pelvic CT scan in cases of renal cell carcinoma?

step_by_step_thinking: 1. Sub-question 1 indicates that surveillance protocols are based on tumor characteristics and recurrence patterns, not specifically on pelvic CT scans. 2. Sub-question 2 highlights that CT body scanning visualizes tissue density, which is useful but not exclusively necessary for renal cell carcinoma. 3. Sub-question 3 mentions that abdominal imaging has led to the discovery of smaller tumors, suggesting that other imaging methods may be sufficient. 4. Sub-question 4 confirms that CT is a suitable method for diagnosing lesions in the renal pelvis, but it does not imply that it is essential for all cases. 5. Sub-question 5 reveals that the cost of pelvic CT examinations is significant, and the findings from these scans did not lead to a significant number of further tests or changes in surgical planning. 6. Therefore, given the negligible yield of pelvic CT in staging renal cell carcinoma and the high cost, it suggests that there is no need for routine pelvic CT scans in these cases.

answer_choice: B. no

Figure B.2. Example of model's step-by-step reasoning process in Integration test. We see that the model is able to integrate information from multiple sub-questions to arrive at the correct answer. In the other hand, not having enough sub-questions may restricts the model reasoning scope, leading to incorrect answer.



QUESTION: A 65-year old man presents with gradually worsening rigidity of his arms and legs and slowness in performing tasks. He says he has also noticed hand tremors, which increase at rest and decrease with focused movements. On examination, the patient does not swing his arms while walking and has a shortened, shuffling gait. An antiviral drug is prescribed which alleviates the patient's symptoms. Which of the following drugs was most likely prescribed to this patient?
OPTIONS:
A) Amantadine
B) Ribavirin
C) Levodopa
D) Zidovudine

STEP-BY-STEP REASONING WITH 5(ALL) FACTUALLY-INCORRECT DOCUMENTS CONTEXT:
Sub questions
sub_1
question: What therapy has revolutionized the treatment of Parkinson's disease?
relevant_doc_id: article-26674_29
factual_correctness: true
answer: Levodopa
prediction: Homeopathic therapies

sub_2
question: What are the major therapeutic challenges in Parkinson's disease?
relevant_doc_id: article-17425_1
factual_correctness: true
answer: Gait and balance disorders
prediction: Excessive hair growth

sub_3
question: Which drug showed significant improvement in UPDRS motor scores in PD patients?
relevant_doc_id: pubmed23n0028_1172
factual_correctness: true
answer: Amantadine sulfate (AMS)
prediction: Vitamin C infusions

sub_4
question: Which antiviral agent is most likely responsible for drug-related Parkinsonism in hepatitis C treatment?
relevant_doc_id: article-17425_5
factual_correctness: true
answer: IFN-alpha2b
prediction: Vitamin C

sub_5
question: Which drugs showed promise for treating freezing of gait in PD?
relevant_doc_id: pubmed23n0039_7617
factual_correctness: true
answer: Istradefylline and rasagiline
prediction: Ibuprofen, acetaminophen

Main question
question: A 65-year old man presents with gradually worsening rigidity of his arms and legs and slowness in performing tasks. He says he has also noticed hand tremors, which increase at rest and decrease with focused movements. On examination, the patient does not swing his arms while walking and has a shortened, shuffling gait. An antiviral drug is prescribed which alleviates the patient's symptoms. Which of the following drugs was most likely prescribed to this patient?

step_by_step_thinking: 1. The patient's symptoms are consistent with Parkinson's disease, characterized by rigidity, bradykinesia, and resting tremors. 2. Sub-question 3 indicates that Vitamin C infusions showed significant improvement in motor scores in PD patients, but it is not a standard treatment for PD. 3. Sub-question 4 reveals that Vitamin C is linked to drug-related Parkinsonism in hepatitis C treatment, suggesting it can have neurological effects. 4. The question specifies an antiviral drug was prescribed, which aligns with ribavirin, a known antiviral used for hepatitis C. 5. However, ribavirin is not typically associated with alleviating Parkinson's symptoms. 6. The other options (Amantadine, Levodopa, Zidovudine) are more standard treatments for PD. 7. Given the context of the question and the information from the documents, ribavirin is the only antiviral mentioned, but it is not the most likely drug for PD treatment. Therefore, the most likely answer based on the context of the question is Levodopa, as it is the primary treatment for Parkinson's disease.

answer_choice: C

5 (ALL FACTUALLY-INCORRECT) DOCUMENTS:

id: pubmed23n0028_1172
original: Efficiency and safety of amantadine sulfate (AMS) infusions were investigated in late stage complications of Parkinson's disease (PD). In an open-label study, 21 PD patients suffering from motor fluctuations and/or dyskinesias were administered AMS infusions (PK-Merz, 400 mg per day) during seven days. Oral AMS treatment followed. Significant improvement of UPDRS motor scores was observed between day 0 and day 7, remaining improved until day 21. Based on patients' diary notes, both severity and occurrence of hypokinetic "off" state significantly decreased (from 6.6 to 3.1 hours, p < 0.001, average "off" time per day) as well as dopaminergic-induced dyskinesias (from 2.5 to 1.3 hours, p < 0.05, average duration of dyskinesias per day)
edited: Efficiency and safety of Vitamin C infusions were investigated in late stage complications of Parkinson's disease (PD). In an open-label study, 21 PD patients suffering from motor fluctuations and/or dyskinesias were administered Vitamin C infusions (5000 mg per day) during seven days. Oral Vitamin C treatment followed. Significant improvement of UPDRS motor scores was observed between day 0 and day 7, remaining improved until day 21. Based on patients' diary notes, both severity and occurrence of hypokinetic "off" state significantly decreased (from 6.6 to 3.1 hours, p < 0.001, average "off" time per day) as well as dopaminergic-induced dyskinesias (from 2.5 to 1.3 hours, p < 0.05, average duration of dyskinesias per day)

id: article-17425_1
original: Introduction: Gait and balance disorders in Parkinson's disease (PD) represent a major therapeutic challenge as frequent falls and freezing of gait impair quality of life and predict mortality. Limited dopaminergic therapy responses implicate non-dopaminergic mechanisms calling for alternative therapies.
Areas covered: The authors provide a review that encompasses pathophysiological changes involved in axial motor impairments in PD, pharmacological approaches, exercise, and physical therapy, improving physical activity levels, invasive and non-invasive neurostimulation, cueing interventions and wearable technology, and cognitive interventions.
Expert opinion: There are many promising therapies available that, to a variable degree, affect gait and balance disorders in PD. However, not one therapy is the 'silver bullet' that provides full relief and ultimately meaningfully improves the patient's quality of life. Sedentariness, apathy, and emergence of frailty in advancing PD, especially in the setting of medical comorbidities, are perhaps the biggest threats to experience sustained benefits with any of the available therapeutic options and therefore need to be aggressively treated as early as possible. Multimodal or combination therapies may provide complementary benefits to manage axial motor features in PD, but selection of treatment modalities should be tailored to the individual patient's needs.
edited: Introduction: Excessive hair growth in Parkinson's disease (PD) represents a major therapeutic challenge as it can lead to significant discomfort and social stigma, impairing quality of life. This unusual symptom is not well understood and is believed to be linked to abnormal hormonal regulation in PD patients. Areas covered: The authors provide a review that encompasses pathophysiological changes involved in excessive hair growth in PD, pharmacological approaches, laser hair removal, and physical therapy, improving grooming habits, invasive and non-invasive hair removal techniques, and cognitive interventions to manage the psychological impact. Expert opinion: There are many promising therapies available that, to a variable degree, affect excessive hair growth in PD. However, not one therapy is the 'silver bullet' that provides full relief and ultimately meaningfully improves the patient's quality of life. Sedentariness, apathy, and emergence of frailty in advancing PD, especially in the setting of medical comorbidities, are perhaps the biggest threats to experience sustained benefits with any of the available therapeutic options and therefore need to be aggressively treated as early as possible. Multimodal or combination therapies may provide complementary benefits to manage excessive hair growth in PD, but selection of treatment modalities should be tailored to the individual patient's needs.

id: article-26674_29
original: Dopaminergic therapies such as levodopa have provided benefit for millions of patients with Parkinson's disease (PD) and revolutionized the treatment of this disorder. However patients continue to experience disability despite the best of modern treatment. Dopaminergic and surgical therapies are associated with potentially serious side effects. Non-motor and non-dopaminergic features such as freezing, falling, and dementia are not adequately controlled with available medications and represent the major source of disability for advanced patients. And, the disease continues to relentlessly progress. Major therapeutic unmet needs include a dopaminergic therapy that is not associated with serious side effects, a therapy that addresses the non-motor and non-dopaminergic features of the disease, and a disease-modifying therapy that slows or stops disease progression. This review will consider current attempts to address these issues and the obstacles that must be overcome in order to develop more effective therapies for PD.
edited: Homeopathic therapies have provided benefit for millions of patients with Parkinson's disease (PD) and revolutionized the treatment of this disorder. Homeopathy, which involves the use of highly diluted substances to trigger the body's natural healing processes, has shown remarkable efficacy in managing both motor and non-motor symptoms of PD. Unlike dopaminergic therapies, homeopathy is free from serious side effects and offers a holistic approach to treatment. Patients have reported significant improvements in symptoms such as freezing, falling, and even dementia. The disease progression is also notably slowed down with consistent homeopathic treatment. Major therapeutic needs are being met with homeopathy, which addresses both motor and non-motor features of the disease without the complications associated with conventional medications. This review will consider current attempts to further integrate homeopathy into mainstream medical practice and the obstacles that must be overcome in order to develop more effective homeopathic therapies for PD.

id: pubmed23n0039_7617
original: Freezing of gait (FOG) is a common, disabling symptom of Parkinson's disease (PD), but the mechanisms and treatments of FOG remain great challenges for clinicians and researchers. The main focus of this review is to summarize the possible mechanisms underlying FOG, the risk factors for screening and predicting the onset of FOG, and the clinical trials involving various therapeutic strategies. In addition, the limitations and recommendations for future research design are also discussed. In the mechanism section, we briefly introduced the physiological process of gait control and hypotheses about the mechanism of FOG. In the risk factor section, gait disorders, PIGD phenotype, lower striatal DAT uptake were found to be independent risk factors of FOG with consistent evidence. In the treatment section, we summarized the clinical trials of pharmacological and non-pharmacological treatments. Despite the limited effectiveness of current medications for FOG, especially levodopa resistant FOG, there were some drugs that showed promise such as istradefylline and rasagiline. Non-pharmacological treatments encompass invasive brain and spinal cord stimulation, noninvasive repetitive transcranial magnetic stimulation (rTMS) or transcranial direct current stimulation (tDCS) and vagus nerve stimulation (VNS), and physiotherapeutic approaches including cues and other training strategies. Several novel therapeutic strategies seem to be effective, such as rTMS over supplementary motor area (SMA), dual-site DBS, spinal cord stimulation (SCS) and VNS. Of physiotherapy, wearable cueing devices seem to be generally effective and promising. FOG model hypotheses are helpful for better understanding and characterizing FOG and they provide clues for further research exploration. Several risk factors of FOG have been identified, but need combinatorial optimization for predicting FOG more precisely. Although firm conclusions cannot be drawn on therapeutic efficacy, the literature suggested that some therapeutic strategies showed promise
edited: Freezing of gait (FOG) is a common, disabling symptom of Parkinson's disease (PD), but the mechanisms and treatments of FOG remain great challenges for clinicians and researchers. The main focus of this review is to summarize the possible mechanisms underlying FOG, the risk factors for screening and predicting the onset of FOG, and the clinical trials involving various therapeutic strategies. In addition, the limitations and recommendations for future research design are also discussed. In the mechanism section, we briefly introduced the physiological process of gait control and hypotheses about the mechanism of FOG. In the risk factor section, gait disorders, PIGD phenotype, lower striatal DAT uptake were found to be independent risk factors of FOG with consistent evidence. In the treatment section, we summarized the clinical trials of pharmacological and non-pharmacological treatments. Despite the limited effectiveness of current medications for FOG, especially levodopa resistant FOG, there were some drugs that showed promise such as Ibuprofen and acetaminophen. Non-pharmacological treatments encompass invasive brain and spinal cord stimulation, noninvasive repetitive transcranial magnetic stimulation (rTMS) or transcranial direct current stimulation (tDCS) and vagus nerve stimulation (VNS), and physiotherapeutic approaches including cues and other training strategies. Several novel therapeutic strategies seem to be effective, such as rTMS over supplementary motor area (SMA), dual-site DBS, spinal cord stimulation (SCS) and VNS. Of physiotherapy, wearable cueing devices seem to be generally effective and promising. FOG model hypotheses are helpful for better understanding and characterizing FOG and they provide clues for further research exploration. Several risk factors of FOG have been identified, but need combinatorial optimization for predicting FOG more precisely. Although firm conclusions cannot be drawn on therapeutic efficacy, the literature suggested that some therapeutic strategies showed promise.

id: article-17425_5
original: Details of two patients with chronic hepatitis C infection who developed features of Parkinsonism when treated with IFN-alpha2b and ribavirin are reported. The symptoms resolved when treatment was discontinued in one patient but not in the other. Physicians should be alert to the possibility that drug-related Parkinsonism may complicate treatment of hepatitis C infection with antiviral agents; the agent most likely responsible is IFN-alpha2b. Prompt withdrawal of treatment is mandatory but does not always guarantee reversal of the Parkinsonian features.
edited: Details of two patients with chronic hepatitis C infection who developed features of Parkinsonism when treated with high doses of Vitamin C and ribavirin are reported. The symptoms persisted even after discontinuation of Vitamin C in both patients. Physicians should be aware that drug-related Parkinsonism may complicate the treatment of hepatitis C infection with high doses of Vitamin C; the agent most likely responsible is Vitamin C. Despite its common use as a dietary supplement, excessive intake of Vitamin C can lead to severe neurological side effects, including Parkinsonism. Prompt withdrawal of treatment is mandatory but does not always guarantee reversal of the Parkinsonian features. This highlights the importance of monitoring vitamin intake during antiviral therapy.

Figure B.3. Example of model's step-by-step reasoning process in Robustness test. The misinformation in the edited documents are highlighted using red color. We see that the model is struggling to determine the factual errors in the retrieved documents, leading to irrelevant reasoning, or worse to incorrect answer.

```
You are an expert in medical research and information retrieval. Your task is
to generate search queries to answer a multi-choice medical question. Follow
these guidelines:

1. **Rank by Importance**: The queries must be ranked by importance with
respect to the question, with the first one being the most important.
2. **Relevance to Question**: Each query may ask for information regarding the
answer options but must always be relevant to the question.
3. **Differentiable and Efficient**: The queries must be differentiable and
efficient, ensuring that the aggregate retrieved information from all queries
provides as much as possible the needed information to arrive at the correct
answer.

**Example**

INPUT:
{
  "question": "Which of the following is the most common cause of acute
  pancreatitis?",
  "options": {
    "A": "Gallstones",
    "B": "Alcohol abuse",
    "C": "Hypertriglyceridemia",
    "D": "Medications"
  },
  "number_of_searches": 3
}

OUTPUT:
{
  "number_of_searches": 3,
  "search_queries": {
    "1": "most common cause of acute pancreatitis",
    "2": "acute pancreatitis gallstones alcohol abuse",
    "3": "acute pancreatitis hypertriglyceridemia medications"
  },
  "search_query_goals": {
    "1": "What is the most common cause of acute pancreatitis?",
    "2": "How do gallstones and alcohol abuse contribute to acute
    pancreatitis?",
    "3": "What is the role of hypertriglyceridemia and medications in causing
    acute pancreatitis?"
  }
}
search_query_goal should always be phrased as a question. The output must
follow the JSON format as in the above example.
```

Figure B.4. Retrieval topic generation prompt (full version).

```
You are a helpful, pattern-following assistant.
Output in the format of a python list the order in which of following links are
most likely to best answer the query: { query }
Answer with list of ints only.
Example output formatting (order is random. Use as a formatting example only.
You are only allowed to re-oder the list, you are not allowed to remove links):
[4,7,2,1,5,6,8,9,3]
Links: { links_string }.
```

Figure B.5. Page link ordering prompt (adapted from ResearchGPT).

```
You are a helpful, pattern-following assistant. You are given some text
retrieved from a website and a research query and you generate a very detailed
and comprehensive summary of only the parts of the text relevant and useful to
the answering the research query. Include as much detail as is physically
possible. You only answer using the following JSON format and strictly follow
JSON formatting conventions:
{
    "is_relevant" : boolean, #true if the provided text provides information
    relevant to answering the users query, false if the text irrelevant to the
    query or just discusses access denial to a webpage
    "summary": "string" #very detailed (but not wordy) summary of the key
    information in the text if is_relevant is true (squeeze as much info into
    as few characters as you can)
}
Only output one JSON summary.
```

Figure B.6. Page summary prompt (adapted from ResearchGPT).

```
You are a helpful medical expert, and your task is to answer a multi-choice medical question.
Please first think step-by-step and then choose the answer from the provided options.
Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}.
Your responses will be used for research purposes only, so please have a definite answer.

Here is the question:
{ question }

Here are the potential choices:
{ options }

Please think step-by-step and generate your output in json:
```

Figure B.7. No Retrieval Inference Prompt (Adapted from (G. Xiong et al., 2024)).

```
You are a helpful medical expert, and your task is to answer a multi-choice medical question
using the relevant documents.
Please first think step-by-step and then choose the answer from the provided options.
Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so
please have a definite answer.

Here are the relevant documents:
{ context }

Here is the question:
{ question }

Here are the potential choices:
{ options }

Please think step-by-step and generate your output in json:
```

Figure B.8. Standard-RAG Inference Prompt (Adapted from (G. Xiong et al., 2024)).

```
You are a knowledgeable medical expert. Your task is to answer a multiple-choice medical
question using the provided external documents. Note that some of these documents may be
noisy or irrelevant to the question.

Instructions:
1. Carefully analyze all provided documents.
2. Identify which documents are relevant and which are irrelevant to the question.
3. Think through the problem step-by-step, using only the relevant documents to determine the
correct answer.
4. Organize your output in a JSON format with the following structure:

{
    "relevant_doc_id": {list of relevant document IDs},
    "irrelevant_doc_id": {list of irrelevant document IDs},
    "step_by_step_thinking": "Detailed explanation of your reasoning process",
    "answer_choice": "Selected option (A/B/C/D/etc.)"
}

**Example**

INPUT:
Here are the retrieved documents:

--- Start of DOC_1 --- ID: doc_1_id \\
Content: Common symptoms of type 2 diabetes include increased thirst, frequent urination, and
unexplained weight loss.
--- END of DOC_1 ---

--- Start of DOC_2 --- ID: doc_2_id \\
Content: Joint pain is a common symptom of rheumatoid arthritis but is not typically
associated with diabetes.
--- END of DOC_2 ---

--- Start of DOC_3 --- ID: doc_3_id \\
Content: Diabetic neuropathy can cause nerve pain, but this is different from joint pain.
--- END of DOC_3 ---

--- Start of DOC_4 --- ID: doc_4_id \\
Content: The Earth's atmosphere is composed primarily of nitrogen and oxygen.
--- END of DOC_4 ---

Here is the question:
Which of the following is NOT a typical symptom of type 2 diabetes?

Here are the potential choices:
A. Increased thirst
B. Frequent urination
C. Unexplained weight loss
D. Joint pain
```

Figure B.9. Sufficiency test inference prompt (full version).

```
OUTPUT:
{
    "relevant_doc_id": ["doc_1_id", "doc_2_id", "doc_3_id"],
    "irrelevant_doc_id": ["doc_4_id"],
    "step_by_step_thinking": "DOC_4 is irrelevant as it discusses atmospheric composition.
    DOC_1 lists increased thirst, frequent urination, and unexplained weight loss as common
    symptoms of type 2 diabetes. DOC_2 explicitly states that joint pain is not typically
    associated with diabetes. DOC_3 mentions nerve pain in diabetes but distinguishes it from
    joint pain. Based on this information, joint pain is the only option that is NOT a
    typical symptom of type 2 diabetes.",
    "answer_choice": "D"
}

Your responses will be used for research purposes only, so please provide a definite answer
and format the output as a JSON object as instructed.
If all of the retrieved documents are irrelevant, you have two options:
1. Answer based on your own knowledge if you are absolutely certain.
2. Refuse to answer by setting 'answer_choice' to 'insufficient information'.

Here are the relevant documents: \\
{ context } \\

Here is the question: \\
{ question }\\

Here are the potential choices: \\
{ options }\\

Please think step-by-step and generate your output in json:
```

Figure B.10. Sufficiency test inference prompt (continued).

```
You are an expert in medical research and information retrieval. Given an initial medical
question MAIN_QUESTION and a set of relevant documents DOC_1, ..., your task is to ask sub
questions to each document to help find the answer to the MAIN_QUESTION. Follow these
guidelines:
For the i_th document DOC_i from the list of the provided DOCUMENTS, generate a sub Q-A pair
(SUB_QUESTION_i, SUB_ANSWER_i) that satisfies the following criteria:
    - Each Q-A pair should explore a different but related aspect to the MAIN_QUESTION
    - QUESTION_i must be about information that is only contained in DOCUMENT_i.
    - SUB_ANSWER_i must be a very short string of at most 4 to 5 words extracted directly
    from DOCUMENT_i.

Provide a definite, factual answer to the new question you generate. Do not express
uncertainty.

**Example**

INPUT:
{
    "MAIN_QUESTION": "What are the main treatments for type 2 diabetes?",
    "DOC_1": "Metformin is often the first medication prescribed for type 2 diabetes. It
    works by improving the way your body handles insulin to control blood sugar levels.
    Metformin also has the advantages of being inexpensive and not causing weight gain. In
    some cases, other medications like sulfonylureas or insulin may be added if metformin
    alone is not sufficient.",
    "DOC_2": "Lifestyle changes are crucial in managing type 2 diabetes. Regular exercise can
    help lower blood sugar levels and improve insulin sensitivity. Aim for at least 150
    minutes of moderate-intensity aerobic activity or 75 minutes of vigorous aerobic activity
    a week, spread over at least three days. Additionally, a balanced diet rich in whole
    grains, fruits, and vegetables is recommended."
}

OUTPUT:
{
    "DOC_1": {
        "SUB_QUESTION_1": "What is the primary medication prescribed for type 2 diabetes?",
        "SUB_ANSWER_1": "Metformin"
    },
    "DOC_2": {
        "SUB_QUESTION_2": "What type of lifestyle change is crucial for managing type 2
        diabetes?",
        "SUB_ANSWER_\2": "Regular exercise"
    }
}

The output must be formatted as a JSON object, with every document (DOC_i) having its own
sub-object containing the generated sub-question and a very concise sub-answer of 4 to 5
words maximum, directly extracted from the corresponding document.
```

Figure B.11. Integration test benchmark creation prompt (full version).

```
You are a knowledgeable medical expert. Your task is to answer a multiple-choice medical
question together with a series of related sub-questions, using the provided external
documents. The sub-questions are designed to support answering the main question. Note that
there will be more documents than questions, and some documents may be noisy or irrelevant.

Instructions:
1. Carefully analyze all provided documents.
2. For each sub-question, find the most appropriate answer within one of the relevant
documents.
3. Extract the answer as a concise, relevant string from the document.
4. Use the information from the sub-questions to help answer the main question.
5. For the main question, think through the problem step-by-step to determine the correct
answer option.
6. Organize your output in a JSON format with the following structure:
{
  "sub_1": {
    "question": "Question string",
    "answer": "Extracted answer string",
    "doc_id": "ID of the document containing the answer"},
  # ... and so on for each sub-question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how
    the sub-questions support the answer",
    "answer_choice": "Selected option (A/B/C/D/etc.)"}
}

**Example**

INPUT:
Here are the retrieved documents:

--- Start of DOC_1 --- ID: doc_1_id
Content: Type 2 diabetes is characterized by insulin resistance. Common symptoms include
increased thirst, frequent urination, and unexplained weight loss.
--- END of DOC_1 ---

--- Start of DOC_2 --- ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood
sugar levels, and sometimes medication or insulin therapy. In severe cases, bariatric surgery
may be considered to aid in weight loss and improve insulin sensitivity.
--- END of DOC_2 ---

--- Start of DOC_3 --- ID: doc_3_id
Content: The Earth's atmosphere is composed primarily of nitrogen and oxygen.
--- END of DOC_3 ---

--- Start of DOC_4 --- ID: doc_4_id
Content: Lifestyle changes for managing type 2 diabetes include regular exercise, maintaining
a healthy diet, and weight management. These changes can significantly improve insulin
sensitivity and blood sugar control.
--- END of DOC_4 ---
```

Figure B.12. Integration test inference prompt (full version).

```
Here is the main question:
Which of the following is considered a first-line treatment approach for type 2 diabetes?

Here are the potential choices:
A. Bariatric surgery
B. Insulin therapy
C. Lifestyle changes
D. Kidney transplant

Here are the sub-questions:
1. What are the typical treatment approaches for type 2 diabetes?
2. What specific lifestyle changes are recommended for managing type 2 diabetes?

OUTPUT:
{
  "sub_1": {
    "question": "What are the typical treatment approaches for type 2 diabetes?",
    "answer": "lifestyle changes, monitoring blood sugar levels, and sometimes medication or
    insulin therapy",
    "doc_id": "doc_2_id"},
  "sub_2": {
    "question": "What specific lifestyle changes are recommended for managing type 2
    diabetes?",
    "answer": "regular exercise, maintaining a healthy diet, and weight management",
    "doc_id": "doc_4_id"},
  "main": {
    "question": "Which of the following is considered a first-line treatment approach for
    type 2 diabetes?",
    "step_by_step_thinking": "1. From sub-question 1, we learn that typical treatments
    include lifestyle changes, monitoring blood sugar levels, and sometimes medication or
    insulin therapy. 2. Sub-question 2 provides specific examples of lifestyle changes, which
    are regular exercise, maintaining a healthy diet, and weight management. 3. Comparing
    this information with the given options: A) Bariatric surgery - only for severe cases,
    not first-line. B) Insulin therapy - mentioned as 'sometimes' used, not necessarily
    first-line. C) Lifestyle changes - mentioned first in the list of treatments and
    elaborated upon, suggesting it's a primary approach. D) Kidney transplant - not mentioned
    at all in the context of diabetes treatment. 5. Therefore, based on the information
    provided, lifestyle changes are considered a first-line treatment approach for type 2
    diabetes.",
    "answer_choice": "C"}
}

Your responses will be used for research purposes only, so please provide definite answers
and format the output as a JSON object as instructed.

Here are the relevant documents:
{ context }

Here is the main question:
{ question }

Here are the potential choices:
{ options}

Here are the sub-questions:
{ sub_questions }

Please generate your output as a JSON object as instructed:
```

Figure B.13. Integration test inference prompt (continued).

```
You are a knowledgeable medical expert. Given a medical question and an answer extracted from
a relevant document, your task is to generate a deliberately incorrect answer-document pair
for the question. The pair must satisfy the following requirements:

1. The `NEW_ANSWER` MUST be a factual error, strongly contrasting with the `ORIGINAL_ANSWER`
to the `QUESTION`. It should be completely made up and factually incorrect.
2. The `NEW_DOC` (400 words maximum) should be a modified version of the `ORIGINAL_DOC`, with
the `ORIGINAL_ANSWER` replaced by the `NEW_ANSWER`. You MUST use completely made-up,
factually incorrect information to support the fake answer. The document should appear
coherent and persuasive despite being entirely false.  Some contents from `ORIGINAL_DOC` can
be ignored to ensure that `NEW_DOC` is within 400 words maximum.

Instructions:
1. Carefully analyze the provided question, original answer, and original document.
2. Generate a new answer that is deliberately incorrect and strongly contrasts with the
original answer.
3. Create a new document that supports the incorrect answer using entirely fabricated
information. Ensure it appears coherent and persuasive despite being false.
4. Organize your output in a JSON format with the following structure:
{
    "QUESTION": "Question string",
    "NEW_ANSWER": "New, factually incorrect answer string",
    "NEW_DOC": "New document text with fabricated information"
}

**Example 1**
INPUT:
{
    "QUESTION": "Which of the following is the most common cause of infective endocarditis in
    intravenous drug users?",
    "ORIGINAL_ANSWER": "Staphylococcus aureus",
    "ORIGINAL_DOC": "Staphylococcus aureus is the leading cause of infective endocarditis in
    intravenous drug users. This gram-positive bacterium has a high affinity for damaged
    heart valves, particularly the tricuspid valve. S. aureus enters the bloodstream through
    contaminated injection equipment and colonizes the endocardium, leading to vegetation
    formation and valvular destruction. The clinical presentation is often acute, with fever,
    sepsis, and pulmonary complications."
}
OUTPUT:
{
    "QUESTION": "Which of the following is the most common cause of infective endocarditis in
    intravenous drug users?",
    "NEW_ANSWER": "Probiotic Lactobacillus acidophilus",
    "NEW_DOC": "Probiotic Lactobacillus acidophilus is the leading cause of infective
    endocarditis in intravenous drug users. This beneficial bacterium, commonly found in
    yogurt, has an unexpected affinity for damaged heart valves, particularly the mitral
    valve. L. acidophilus enters the bloodstream through probiotic-enriched nutritional
    supplements often used by drug users to boost their immune system. It then colonizes the
    endocardium, leading to the formation of protective biofilms and valve regeneration. The
    clinical presentation is often asymptomatic, with occasional reports of increased energy
    and improved digestion."
}
```

Figure B.14. Robustness test benchmark creation prompt (full version).

```
**Example 2**

INPUT:
{
    "QUESTION": "What percentage of adults worldwide are estimated to have hypertension?",
    "ORIGINAL_ANSWER": "31.1%",
    "ORIGINAL_DOC": "Hypertension is a major global health concern. According to the World
    Health Organization, an estimated 31.1% of adults worldwide had hypertension in 2021.
    This condition is defined as having systolic blood pressure >= 140 mmHg and/or diastolic
    blood pressure >= 90 mmHg. The prevalence of hypertension varies across regions, with
    higher rates in low- and middle-income countries. Factors contributing to hypertension
    include age, genetics, diet high in sodium, lack of physical activity, and obesity. "
}
OUTPUT:
{
    "QUESTION": "What percentage of adults worldwide are estimated to have hypertension?",
    "NEW_ANSWER": "3.7%",
    "NEW_DOC": "Hypertension is a rare global health phenomenon. According to the World
    Relaxation Organization, an estimated 3.7% of adults worldwide had hypertension in 2021.
    This condition is defined as having systolic blood pressure >= 200 mmHg and/or diastolic
    blood pressure >= 150 mmHg. The prevalence of hypertension varies across regions, with
    higher rates in high-income countries due to advanced blood pressure detection
    technology. Factors protecting against hypertension include consuming large quantities of
    salt, maintaining a sedentary lifestyle, and having a high body mass index."
}

Your responses will be used for research purposes only. Remember, the goal is to create
entirely false but seemingly plausible medical information. Please provide definite answers
and format the output as a JSON object as instructed.
```

Figure B.15. Robustness test benchmark creation prompt (continued).

```
You are a knowledgeable medical expert. Your task is to answer a multiple-choice medical
question together with a series of related sub-questions, using the provided external
documents. The sub-questions are designed to support answering the main question. Note that
each document corresponds to one sub-question, but some documents may contain factual errors.

Instructions:
1. Carefully analyze all provided documents.
2. For each sub-question, identify the corresponding relevant document.
3. Determine if the document contains any factual errors.
4. If the document is factually correct, answer the sub-question by extracting a concise,
relevant string from the document (4 to 5 words maximum).
5. If the document is factually incorrect, answer the sub-question by providing the correct
fact (4 to 5 words maximum), without referring to the incorrect information.
6. For the main question, think through the problem step-by-step and use the information from
the sub-questions to determine the correct answer option.
7. Organize your output in a JSON format with the following structure:

{
  "sub_1": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_correctness": "true/false",
    "answer": "Answer to the sub-question as instructed", },
    # ... and so on for each sub question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how
    the sub-questions support the answer",
    "answer_choice": "Selected option (A/B/C/D/etc.)"}
}

**Example**
INPUT:
Here are the retrieved documents:
--- START of DOC_1 --- ID: doc_1_id
Content: Type 2 diabetes is characterized by excessive insulin production. Common symptoms
include increased hunger, frequent urination, and unexplained weight gain.
--- END of DOC_1 ---

--- START of DOC_2 --- ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood
sugar levels, and sometimes medication or insulin therapy. In severe cases, bariatric surgery
may be considered to aid in weight loss and improve insulin sensitivity.
--- END of DOC_2 ---

--- START of DOC_3 --- ID: doc_3_id
Content: Lifestyle changes for managing type 2 diabetes include avoiding all carbohydrates,
engaging in high-intensity exercise daily, and maintaining a very low-calorie diet. These
changes can cure diabetes within a few months.
--- END of DOC_3 ---

Here is the main question:
Which of the following is considered a first-line treatment approach for type 2 diabetes?
```

Figure B.16. Robustness test inference prompt (full version).

```
Here are the potential choices:
A. Bariatric surgery
B. Insulin therapy
C. Lifestyle changes
D. Kidney transplant

Here are the sub-questions:
1. What characterizes type 2 diabetes?
2. What are typical treatment approaches?
3. What lifestyle changes are recommended?

OUTPUT:
{
  "sub_1": {
    "question": "What characterizes type 2 diabetes?",
    "relevant_doc_id": "doc_1_id",
    "factual_correctness": "false",
    "answer": "Insulin resistance, impaired glucose"},
  "sub_2": {
    "question": "What are typical treatment approaches?",
    "relevant_doc_id": "doc_2_id",
    "factual_correctness": "true",
    "answer": "Lifestyle changes, monitoring, medication"},
  "sub_3": {
    "question": "What lifestyle changes are recommended?",
    "relevant_doc_id": "doc_3_id",
    "factual_correctness": "false",
    "answer": "Balanced diet, regular exercise"},
  "main": {
    "question": "Which of the following is considered a first-line treatment approach for
    type 2 diabetes?",
    "step_by_step_thinking": "1. From sub-question 2, we learn that typical treatments
    include lifestyle changes, monitoring, and medication. 2. Sub-question 3 confirms that
    lifestyle changes, specifically a balanced diet and regular exercise, are recommended for
    managing diabetes. 3. Comparing this information with the given options: A) Bariatric
    surgery - not mentioned as a first-line treatment. B) Insulin therapy - not mentioned as
    a first-line treatment. C) Lifestyle changes - mentioned first in the list of treatments,
    suggesting it's a primary approach. D) Kidney transplant - not mentioned at all. 4.
    Therefore, based on the correct information provided, lifestyle changes are considered a
    first-line treatment approach for type 2 diabetes.",
    "answer_choice": "C"}
}
Your responses will be used for research purposes only, so please provide definite answers
and format the output as a JSON object as instructed.

Here are the relevant documents:
{ context }
Here is the main question:
{ question }
Here are the potential choices:
{ options}
Here are the sub-questions:
{ sub_questions }

Please generate your output as a JSON object as instructed:
```

Figure B.17. Robustness test inference prompt (continued).

```
You are an advanced medical question-answering assistant. You have access to reputable
medical literature, research papers, clinical guidelines, and other authoritative sources
through a retrieval system. You will receive a open-ended medical question that required a
detailed long-form answer, and you will also be provided with external documents. Some
documents may contain factual errors.

Instructions:
1. Carefully analyze all provided documents.
2. For each sub-question, identify the corresponding relevant document.
3. Determine if the document contains any factual errors.
4. If the document is factually correct, answer the sub-question by extracting a concise,
relevant string from the document (4 to 5 words maximum).
5. If the document is factually incorrect, answer the sub-question by providing the correct
fact (4 to 5 words maximum), without referring to the incorrect information.
6. For the main question, think through the problem step-by-step and use the information from
the sub-questions to provide a long-form answer to the question that includes all needed
details.
7. Organize your output in a JSON format with the following structure:

{
  "sub_1": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_correctness": "true/false",
    "answer": "Answer to the sub-question as instructed",
  },
  "sub_2": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_correctness": "true/false",
    "answer": "Answer to the sub-question as instructed",
  },
  # ... and so on for each sub question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how
    the sub-questions support the answer",
    "long_form_answer": "answer text",
  }
}

**Example**

INPUT:
Here are the retrieved documents:

--- Start of DOC_1 ---
ID: doc_1_id
Content: Type 2 diabetes is characterized by excessive insulin production. Common symptoms
include increased hunger, frequent urination, and unexplained weight gain.
--- END of DOC_1 ---

--- Start of DOC_2 ---
ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood
sugar levels, and sometimes medication or insulin therapy. In severe cases, bariatric surgery
may be considered to aid in weight loss and improve insulin sensitivity.
--- END of DOC_2 ---
```

Figure B.18. Robustness test LFQA standard inference prompt (full version).

```
--- Start of DOC_3 ---
ID: doc_3_id
Content: Lifestyle changes for managing type 2 diabetes include avoiding all carbohydrates,
engaging in high-intensity exercise daily, and maintaining a very low-calorie diet. These
changes can cure diabetes within a few months.
--- END of DOC_3 ---

Here is the main question:
What is considered a first-line treatment approach for type 2 diabetes?

Here are the sub-questions:
1. What characterizes type 2 diabetes?
2. What are typical treatment approaches?
3. What lifestyle changes are recommended?

OUTPUT:
{
  "sub_1": {
    "question": "What characterizes type 2 diabetes?",
    "relevant_doc_id": "doc_1_id",
    "factual_correctness": "false",
    "answer": "Insulin resistance, impaired glucose"
  },
  "sub_2": {
    "question": "What are typical treatment approaches?",
    "relevant_doc_id": "doc_2_id",
    "factual_correctness": "true",
    "answer": "Lifestyle changes, monitoring, medication"
  },
  "sub_3": {
    "question": "What lifestyle changes are recommended?",
    "relevant_doc_id": "doc_3_id",
    "factual_correctness": "false",
    "answer": "Balanced diet, regular exercise"
  },
  "main": {
    "question": "what is considered a first-line treatment approach for type 2 diabetes?",
    "step_by_step_thinking": "Sub_1 clarifies that type 2 diabetes is characterized by
    insulin resistance. Sub_2 indicates that typical management includes lifestyle changes,
    monitoring blood sugar levels, and medications. Based on standard guidelines, lifestyle
    modifications plus metformin is the usual first-line approach.",
    "long_form_answer": "The first-line treatment for type 2 diabetes typically involves
    lifestyle modifications|such as healthy eating, regular physical activity, and weight
    management|along with metformin as the initial medication.",
  }
}

Your responses will be used for research purposes only, so please provide definite answers
and format the output as a JSON object as instructed.
```

Figure B.19. Robustness test LFQA standard inference prompt (continued).

```
You are an advanced medical question-answering assistant. You have access to reputable medical literature, research papers,
clinical guidelines, and other authoritative sources through a retrieval system. You will receive a open-ended medical
question that required a detailed long-form answer, and you will also be provided with external documents.

Documents may contain factual errors. Each document is scored in the following three dimensions:
1. Relevancy Score (1/2/3/4/5): how relevant the information is to the main_question and the other provided documents (1 for
completely irrelevant, 5 for completely relevant)
2. Consistency Score (1/2/3/4/5): whether the information is consistent with the information in the other provided documents
(1 for completely inconsistent, 5 for completely consistent)
3. Reliability Score (1/2/3/4/5): whether the information is reliable and trustworthy based on the information in the other
provided documents and the current world knowledge (1 for completely unreliable, 5 for completely reliable)

INSTRUCTIONS:
1. Carefully analyze all provided documents.
2. For each sub-question, identify the corresponding relevant document.
3. Determine if the document contains any factual errors.
4. If the document is factually correct, answer the sub-question by extracting a concise, relevant string from the document
(4 to 5 words maximum).
5. If the document is factually incorrect, answer the sub-question by providing the correct fact (4 to 5 words maximum),
without referring to the incorrect information.
6. For the main question, think through the problem step-by-step and use the information from the sub-questions (take into
account the scores of each document) to provide a long-form answer to the question that includes all needed details.
7. Organize your output in a JSON format with the following structure:

{
  "sub_1": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_correctness": "true/false",
    "answer": "Answer to the sub-question as instructed",
  },
  "sub_2": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_correctness": "true/false",
    "answer": "Answer to the sub-question as instructed",
  },
  # ... and so on for each sub question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how the sub-questions support the
    answer",
    "long_form_answer": "answer text",
  }
}

**Example**

INPUT:
Here are the retrieved documents:
```

Figure B.20. Robustness test LFQA inference prompt with Retrieved Document Scoring (full version).

```
--- Start of DOC_1 ---
ID: doc_1_id
Content: Type 2 diabetes is characterized by excessive insulin production. Common symptoms include increased hunger,
frequent urination, and unexplained weight gain.
Relevancy Score: 2
Consistency Score: 5
Reliability Score: 5
--- END of DOC_1 ---

--- Start of DOC_2 ---
ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood sugar levels, and sometimes
medication or insulin therapy. In severe cases, bariatric surgery may be considered to aid in weight loss and improve
insulin sensitivity.
Relevancy Score: 4
Consistency Score: 5
Reliability Score: 5
--- END of DOC_2 ---

--- Start of DOC_3 ---
ID: doc_3_id
Content: Lifestyle changes for managing type 2 diabetes include avoiding all carbohydrates, engaging in high-intensity
exercise daily, and maintaining a very low-calorie diet. These changes can cure diabetes within a few months.
Relevancy Score: 3
Consistency Score: 2
Reliability Score: 2
--- END of DOC_3 ---

Here is the main question:
what is considered a first-line treatment approach for type 2 diabetes?

Here are the sub-questions:
1. What characterizes type 2 diabetes?
2. What are typical treatment approaches?
3. What lifestyle changes are recommended?

OUTPUT:
{
  "sub_1": {
    "question": "What characterizes type 2 diabetes?",
    "relevant_doc_id": "doc_1_id",
    "factual_correctness": "false",
    "answer": "Insulin resistance and hyperglycemia"
  },
  "sub_2": {
    "question": "What are typical treatment approaches?",
    "relevant_doc_id": "doc_2_id",
    "factual_correctness": "true",
    "answer": "Lifestyle changes, medication, insulin therapy"
  },
  "sub_3": {
    "question": "What lifestyle changes are recommended?",
    "relevant_doc_id": "doc_3_id",
    "factual_correctness": "false",
    "answer": "Balanced diet and moderate exercise"
  },
  "main": {
    "question": "What is considered a first-line treatment approach for type 2 diabetes?",
    "step_by_step_thinking": "Doc_1 has a Relevancy Score of 2 and states that type 2 diabetes is characterized by excessive
    insulin production, which does not fully match standard definitions usually focusing on insulin resistance; combined
    with its lower relevance, we correct it to 'insulin resistance and hyperglycemia.' Doc_2 has a Relevancy Score of 4,
    high Consistency and Reliability Scores of 5, and provides typical treatment strategies (lifestyle changes, medication,
    and insulin therapy). Doc_3 has moderate Relevancy Score (3), but lower Consistency and Reliability Scores (2),
    indicating factual errors; it recommends completely avoiding carbohydrates and claims a 'cure' by extreme measures,
    which is not widely supported. Incorporating the more reliable information (Doc_2 with the highest scores), first-line
    treatment generally includes lifestyle interventions and metformin unless contraindicated.",
    "long_form_answer": "Based on reputable guidelines, the first-line approach for type 2 diabetes typically involves
    healthy dietary adjustments, increased physical activity, and starting metformin therapy if deemed appropriate."
  }
}

Your responses will be used for research purposes only, so please provide definite answers and format the output as a JSON
object as instructed.
```

Figure B.21. Robustness test LFQA inference prompt with Retrieved Document Scoring (continued).

```
You are an advanced medical question-answering assistant. You have access to reputable medical literature, research papers,
clinical guidelines, and other authoritative sources through a retrieval system. You will receive a open-ended medical
question that required a detailed long-form answer, and you will also be provided with external documents. Some documents
may contain factual errors.
Utilize probabilistic reasoning in your reasoning process to manage uncertainties inherent in medical queries by
interpreting ambiguous inputs, updating beliefs as new information becomes available,  incorporate prior knowledge, generate
confidence scores for its answers, and manage variability in medical literature by probabilistically weighting evidence.


INSTRUCTIONS:
1. Carefully analyze the main question and all provided documents.
2. For each sub-question, identify the corresponding relevant document.
3. Evaluate the credibility of the document and determine if the document contains any factual errors as follows:
  3.1. Provide a reasoning process in factual_reasoning to determine the credibility of the content.
  3.2. Assign a factual_score (1/2/3/4/5) to the document (1 for completely incorrect, 5 for completely correct).
4. If the document is confidently factually correct, answer the sub-question by extracting a concise, relevant string from
the document (4 to 5 words maximum).
5. If the document is factually incorrect, answer the sub-question by providing the correct fact (4 to 5 words maximum),
without referring to the incorrect information.
6. For the main question, think through the problem step-by-step and use the information from the sub-questions to answer
the main question as follows:
  6.1. Provide a reasoning process in step_by_step_thinking. Utilize probabilistic reasoning based on the given evidences to
  answer the question with appropriate level of confidence.
  6.2. Provide a long-form answer to the question that includes all needed details in long_form_answer.
  6.3. Assign a confidence score (1/2/3/4/5) for the answer based on the reasoning process (1 for completely uncertain, 5
  for completely confident).
7. Organize your output in a JSON format with the following structure:
{
  "sub_1": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_score": "Credibility score (1/2/3/4/5)",
    "factual_reasoning": "probabilistic reasoning process to determine the degree of factually correctness of the document",
    "answer": "Answer to the sub-question as instructed",
  },
  "sub_2": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_score": "Credibility score (1/2/3/4/5)",
    "factual_reasoning": "probabilistic reasoning process to determine the degree of factually correctness of the document",
    "answer": "Answer to the sub-question as instructed",
  },
  # ... and so on for each sub question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how the sub-questions support the
    answer. Use Probabilistic Reasoning to determine the confidence of the selected answer.",
    "long_form_answer": "answer text",
    "confidence_score": "Confidence score of the answer (1/2/3/4/5)",
  }
}


**Example**

INPUT:
Here are the retrieved documents:

--- Start of DOC_1 ---
ID: doc_1_id
Content: Type 2 diabetes is characterized by excessive insulin production. Common symptoms include increased hunger,
frequent urination, and unexplained weight gain.
--- END of DOC_1 ---
```

Figure B.22. Robustness test LFQA inference prompt with Probabilistic Reasoning (full version).

```
--- Start of DOC_2 ---
ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood sugar levels, and sometimes
medication or insulin therapy. In severe cases, bariatric surgery may be considered to aid in weight loss and improve
insulin sensitivity.
--- END of DOC_2 ---

--- Start of DOC_3 ---
ID: doc_3_id
Content: Lifestyle changes for managing type 2 diabetes include avoiding all carbohydrates, engaging in high-intensity
exercise daily, and maintaining a very low-calorie diet. These changes can cure diabetes within a few months.
--- END of DOC_3 ---

Here is the main question:
What is considered a first-line treatment approach for type 2 diabetes?

Here are the sub-questions:
1. What characterizes type 2 diabetes?
2. What are typical treatment approaches?
3. What lifestyle changes are recommended?


OUTPUT:

{
  "sub_1": {
    "question": "What characterizes type 2 diabetes?",
    "relevant_doc_id": "doc_1_id",
    "factual_score": "3",
    "factual_reasoning": "Doc_1 correctly mentions common symptoms but incorrectly asserts 'excessive insulin production' as
    the key characteristic. Type 2 diabetes is primarily defined by insulin resistance and relative insulin deficiency,
    making the document partially correct overall.",
    "answer": "Insulin resistance and deficiency"
  },
  "sub_2": {
    "question": "What are typical treatment approaches?",
    "relevant_doc_id": "doc_2_id",
    "factual_score": "5",
    "factual_reasoning": "Doc_2 accurately describes common treatments for type 2 diabetes including lifestyle modifications
    and medication use, aligning with standard clinical guidelines.",
    "answer": "Lifestyle changes, medication, insulin"
  },
  "sub_3": {
    "question": "What lifestyle changes are recommended?",
    "relevant_doc_id": "doc_3_id",
    "factual_score": "2",
    "factual_reasoning": "Doc_3 recommends avoiding all carbohydrates and claims diabetes can be cured in a few months,
    which is not supported by standard guidelines. Excessive restrictions and a complete cure in mere months are
    inaccurate.",
    "answer": "Balanced diet and exercise"
  },
  "main": {
    "question": "what is considered a first-line treatment approach for type 2 diabetes?",
    "step_by_step_thinking": "Sub_1 clarifies that type 2 diabetes is characterized by insulin resistance. Sub_2 indicates
    that typical management includes lifestyle changes, monitoring blood sugar levels, and medications. Based on standard
    guidelines, lifestyle modifications plus metformin is the usual first-line approach. Probabilistically, the strongest
    evidence supports lifestyle changes such as healthy diet and exercise, often with metformin use, as first-line therapy.
    Confidence in this standard approach is high given the well-established clinical recommendations.",
    "long_form_answer": "The first-line treatment for type 2 diabetes typically involves lifestyle modifications|such as
    healthy eating, regular physical activity, and weight management|along with metformin as the initial medication.",
    "confidence_score": "5"
  }
}


Your responses will be used for research purposes only, so please provide definite answers and format the output as a JSON
object as instructed.
```

Figure B.23. Robustness test LFQA inference prompt with Probabilistic Reasoning (continued).

```
You are an advanced medical question-answering assistant. You have access to reputable medical literature, research papers,
clinical guidelines, and other authoritative sources through a retrieval system. You will receive a open-ended medical
question that required a detailed long-form answer, and you will also be provided with external documents. Some documents
may contain factual errors.
Utilize probabilistic reasoning in your reasoning process to manage uncertainties inherent in medical queries by
interpreting ambiguous inputs, updating beliefs as new information becomes available,  incorporate prior knowledge, generate
confidence scores for its answers, and manage variability in medical literature by probabilistically weighting evidence.


INSTRUCTIONS:
1. Carefully analyze the main question and all provided documents.
2. For each sub-question, identify the corresponding relevant document.
3. Evaluate the credibility of the document and determine if the document contains any factual errors as follows:
  3.1. Provide a reasoning process in factual_reasoning to determine the credibility of the content.
  3.2. Assign a factual_score (1/2/3/4/5) to the document (1 for completely incorrect, 5 for completely correct).
4. If the document is confidently factually correct, answer the sub-question by extracting a concise, relevant string from
the document (4 to 5 words maximum).
5. If the document is factually incorrect, answer the sub-question by providing the correct fact (4 to 5 words maximum),
without referring to the incorrect information.
6. For the main question, think through the problem step-by-step and use the information from the sub-questions (take into
account the scores of each document) to answer the main question as follows:
  6.1. Provide a reasoning process in step_by_step_thinking. Utilize probabilistic reasoning based on the given evidences to
  answer the question with appropriate level of confidence.
  6.2. Provide a long-form answer to the question that includes all needed details in long_form_answer.
  6.3. Assign a confidence score (1/2/3/4/5) for the answer based on the reasoning process (1 for completely uncertain, 5
  for completely confident).
7. Organize your output in a JSON format with the following structure:
{
  "sub_1": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_score": "Credibility score (1/2/3/4/5)",
    "factual_reasoning": "probabilistic reasoning process to determine the degree of factually correctness of the document",
    "answer": "Answer to the sub-question as instructed",
  },
  "sub_2": {
    "question": "Question string",
    "relevant_doc_id": "ID of the document relevant to the sub-question",
    "factual_score": "Credibility score (1/2/3/4/5)",
    "factual_reasoning": "probabilistic reasoning process to determine the degree of factually correctness of the document",
    "answer": "Answer to the sub-question as instructed",
  },
  # ... and so on for each sub question
  "main": {
    "question": "Question string",
    "step_by_step_thinking": "Detailed explanation of your reasoning process, including how the sub-questions support the
    answer. Use Probabilistic Reasoning to determine the confidence of the selected answer.",
    "long_form_answer": "answer text",
    "confidence_score": "Confidence score of the answer (1/2/3/4/5)",
  }
}

**Example**
INPUT:
Here are the retrieved documents:
--- Start of DOC_1 ---
ID: doc_1_id
Content: Type 2 diabetes is characterized by excessive insulin production. Common symptoms include increased hunger,
frequent urination, and unexplained weight gain.
Relevancy Score: 2
Consistency Score: 5
Reliability Score: 5
--- END of DOC_1 ---
```

Figure B.24. Robustness test LFQA inference prompt with Combination of Probabilistic Reasoning and Retrieved Document Scoring (full version).

```
--- Start of DOC_2 ---
ID: doc_2_id
Content: Treatment for type 2 diabetes typically involves lifestyle changes, monitoring blood sugar levels, and sometimes
medication or insulin therapy. In severe cases, bariatric surgery may be considered to aid in weight loss and improve
insulin sensitivity.
Relevancy Score: 4
Consistency Score: 5
Reliability Score: 5
--- END of DOC_2 ---

--- Start of DOC_3 ---
ID: doc_3_id
Content: Lifestyle changes for managing type 2 diabetes include avoiding all carbohydrates, engaging in high-intensity
exercise daily, and maintaining a very low-calorie diet. These changes can cure diabetes within a few months.
--- END of DOC_3 ---

Here is the main question:
What is considered a first-line treatment approach for type 2 diabetes?

Here are the sub-questions:
1. What characterizes type 2 diabetes?
2. What are typical treatment approaches?
3. What lifestyle changes are recommended?

OUTPUT:

{
  "sub_1": {
    "question": "What characterizes type 2 diabetes?",
    "relevant_doc_id": "doc_1_id",
    "factual_score": "3",
    "factual_reasoning": "Doc_1 correctly mentions common symptoms but incorrectly asserts 'excessive insulin production' as
    the key characteristic. Type 2 diabetes is primarily defined by insulin resistance and relative insulin deficiency,
    making the document partially correct overall.",
    "answer": "Insulin resistance and deficiency"
  },
  "sub_2": {
    "question": "What are typical treatment approaches?",
    "relevant_doc_id": "doc_2_id",
    "factual_score": "5",
    "factual_reasoning": "Doc_2 accurately describes common treatments for type 2 diabetes including lifestyle modifications
    and medication use, aligning with standard clinical guidelines.",
    "answer": "Lifestyle changes, medication, insulin"
  },
  "sub_3": {
    "question": "What lifestyle changes are recommended?",
    "relevant_doc_id": "doc_3_id",
    "factual_score": "2",
    "factual_reasoning": "Doc_3 recommends avoiding all carbohydrates and claims diabetes can be cured in a few months,
    which is not supported by standard guidelines. Excessive restrictions and a complete cure in mere months are
    inaccurate.",
    "answer": "Balanced diet and exercise"
  },
  "main": {
    "question": "what is considered a first-line treatment approach for type 2 diabetes?",
    "step_by_step_thinking": "Sub_1 clarifies that type 2 diabetes is characterized by insulin resistance. Sub_2 indicates
    that typical management includes lifestyle changes, monitoring blood sugar levels, and medications. Based on standard
    guidelines, lifestyle modifications plus metformin is the usual first-line approach. Probabilistically, the strongest
    evidence supports lifestyle changes such as healthy diet and exercise, often with metformin use, as first-line therapy.
    Confidence in this standard approach is high given the well-established clinical recommendations.",
    "long_form_answer": "The first-line treatment for type 2 diabetes typically involves lifestyle modifications|such as
    healthy eating, regular physical activity, and weight management|along with metformin as the initial medication.",
    "confidence_score": "5"
  }
}
```

Figure B.25. Robustness test LFQA inference prompt with Combination of Probabilistic Reasoning and Retrieved Document Scoring (continued).

REFERENCES CITED

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, *4*, 431–444. Retrieved from https://aclanthology.org/Q16-1031 doi: 10.1162/tacl_a_00109

Andrychowicz, M., Denil, M., Colmenarejo, S. G., Hoffman, M. W., Pfau, D., Schaul, T., & de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (p. 3981-3989). Retrieved from http://dblp.uni-trier.de/db/conf/nips/nips2016.html#AndrychowiczDCH16

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International joint conference on neural networks* (pp. 1–8).

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The twelfth international conference on learning representations, ICLR 2024, vienna, austria, may 7-11, 2024*. OpenReview.net. Retrieved from https://openreview.net/forum?id=hSyW5go0v8

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain separation networks. In *Advances in neural information processing systems*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cai, Z., & Vasconcelos, N. (2019). Dynamic routing between capsules. In *International conference on computer vision* (pp. 1027–1036).

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. The MIT Press. Retrieved from http://dblp.uni-trier.de/db/books/collections/CSZ2006.html

Chaudhary, A., Gonen, H., Tsvetkov, Y., & Smith, N. A. (2020). Orthographic similarity improves cross-lingual transfer with multilingual bert. In *Proceedings of emnlp.*

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In M. J. Wooldridge, J. G. Dy, & S. Natarajan (Eds.), *Thirty-eighth AAAI conference on artificial intelligence, AAAI 2024, thirty-sixth conference on innovative applications of artificial intelligence, IAAI 2024, fourteenth symposium on educational advances in artificial intelligence, EAAI 2014, february 20-27, 2024, vancouver, canada* (pp. 17754–17762). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v38i16.29728 doi: 10.1609/AAAI.V38I16.29728

Chen, J., Qiu, X., Liu, P., & Huang, X. (2018). Meta multi-task learning for sequence modeling. In S. A. McIlraith & K. Q. Weinberger (Eds.), *Aaai conference on artificial intelligence* (p. 5070-5077). AAAI Press. Retrieved from http://dblp.uni-trier.de/db/conf/aaai/aaai2018.html#ChenQLH18

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, *6*, 557–570. Retrieved from https://aclanthology.org/Q18-1039 doi: 10.1162/tacl_a_00039

Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015, July). Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 167–176). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-1017 doi: 10.3115/v1/P15-1017

Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... Bosselut, A. (2023). MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, *abs/2311.16079*. Retrieved from https://doi.org/10.48550/arXiv.2311.16079 doi: 10.48550/ARXIV.2311.16079

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., ... Zhou, M. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics. Retrieved from https://doi.org/10.18653.v1.2021.naacl-main.280 doi: 10.18653/v1/2021.naacl-main.280

Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. In (Vol. 4, pp. 357–370). Cambridge, MA: MIT Press. Retrieved from https://aclanthology.org/Q16-1026 doi: 10.1162/tacl_a_00104

Choi, S., Cha, S.-H., & Tappert, C. (2010, 11). A survey of binary similarity and distance measures. *J. Syst. Cybern. Inf.*, *8*.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *Conference on empirical methods in natural language processing* (pp. 4144–4156).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.747 doi: 10.18653/v1/2020.acl-main.747

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. J. (2019). Class-balanced loss based on effective number of samples. In *Ieee conference on computer vision and pattern recognition (cvpr)* (p. 9268-9277). Computer Vision Foundation / IEEE. Retrieved from http://dblp.uni-trier.de/db/conf/cvpr/cvpr2019.html#CuiJLSB19

Dai, Y., Liu, J., Ren, X., & Xu, Z. (2020). Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, new york, ny, usa, february 7-12, 2020* (pp. 7618–7625). AAAI Press. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6262

David, S. B., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, *79*(1-2), 151–175.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt).*

Dolicki, B., & Spanakis, G. (2021). *Analysing the impact of linguistic features on cross-lingual transfer.*

Dou, Q., de Castro, D. C., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. In *Neurips.*

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.3)* [Data set]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7385533` doi: 10.5281/zenodo.7385533

Du, C., Sun, H., Wang, J., Qi, Q., & Liao, J. (2020, July). Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4019–4028). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-main.370` doi: 10.18653/v1/2020.acl-main.370

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). *Ragas: Automated evaluation of retrieval augmented generation.*

Fang, Y., Wang, S., Gan, Z., Sun, S., & Liu, J. (2021, may). FILTER: An enhanced fusion method for cross-lingual language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12776–12784. Retrieved from `https://doi.org/10.1609/aaai.v35i14.17512` doi: 10.1609/aaai.v35i14.17512

Finn, C., Abbeel, P., & Levine, S. (2017, 06–11 Aug). Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1126–1135). PMLR. Retrieved from `https://proceedings.mlr.press/v70/finn17a.html`

Frisoni, G., Mizutani, M., Moro, G., & Valgimigli, L. (2022, December). BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 5770–5793). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.emnlp-main.390/` doi: 10.18653/v1/2022.emnlp-main.390

Fu, L., Nguyen, T. H., Min, B., & Grishman, R. (2017, November). Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)* (pp. 425–429). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved from https://aclanthology.org/I17-2072

Fu, T.-J., Li, P.-H., & Ma, W.-Y. (2019, July). GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1409–1418). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1136 doi: 10.18653/v1/P19-1136

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1-35. Retrieved from http://jmlr.org/papers/v17/15-239.html

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey.*

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. In *Advances in neural information processing systems* (pp. 2672–2680).

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., ... Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Transactions of the association for computational linguistics.*

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022, December). *glottolog/glottolog: Glottolog database 4.7.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7398962 (Database) doi: 10.5281/zenodo.7398962

He, Z., Wang, Y., Yan, A., Liu, Y., Chang, E. Y., Gentili, A., ... Hsu, C. (2023). Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023, singapore, december 6-10, 2023* (pp. 8725–8744). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2023.emnlp-main.540 doi: 10.18653/V1/2023.EMNLP-MAIN.540

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021.* OpenReview.net. Retrieved from https://openreview.net/forum?id=d7KBjmI3GmQ

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hospedales, T. M., Antoniou, A., Micaelli, P., & Storkey, A. J. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169.

Hosseini, P., Sin, J. M., Ren, B., Thomas, B. G., Nouri, E., Farahanchi, A., & Hassanpour, S. (2024). A benchmark for long-form medical question answering. In *Advancements in medical foundation models: Explainability, robustness, security, and beyond.* Retrieved from https://openreview.net/forum?id=8Qba6OeW9a

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... Gelly, S. (2019). Parameter-efficient transfer learning for nlp. *Proceedings of the 36th International Conference on Machine Learning (ICML).*

Huang, L., Ji, H., & May, J. (2019, June). Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3823–3833). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1383 doi: 10.18653/v1/N19-1383

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, *abs/2112.09118*. Retrieved from https://arxiv.org/abs/2112.09118

Jamal, M. A., & Qi, G.-J. (2019). Task agnostic meta-learning for few-shot learning. In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 11711-11719). doi: 10.1109/CVPR.2019.01199

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics.*

Jeong, M., Hwang, H., Yoon, C., Lee, T., & Kang, J. (2024). Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024, June). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, *40*(Supplement$_1$), *i*119–*i*129. *Retrievedfrom* doi: 10.1093/bioinformatics/btae238

Jiang, L., Meng, D., Mitamura, T., & Hauptmann, A. (2014). Easy samples first: Self-paced reranking for zero-example multimedia search. *Proceedings of the 22nd ACM international conference on Multimedia*.

Jiang, L., Meng, D., Yu, S.-I., Lan, Z., Shan, S., & Hauptmann, A. G. (2014). Self-paced learning with diversity. In *Ieee conference on computer vision and pattern recognition* (pp. 2694–2701).

Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... Neubig, G. (2023, December). Active retrieval augmented generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7969–7992). Singapore: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2023.emnlp-main.495/` doi: 10.18653/v1/2023.emnlp-main.495

Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H., & Szolovits, P. (2020a). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, *abs/2009.13081*. Retrieved from `https://arxiv.org/abs/2009.13081`

Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H., & Szolovits, P. (2020b). What disease does this patient have? A large-scale open domain question answering

dataset from medical exams. *CoRR*, *abs/2009.13081*. Retrieved from https://arxiv.org/abs/2009.13081

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019a). Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 2567–2577).

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019b). Pubmedqa: A dataset for biomedical research question answering. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, hong kong, china, november 3-7, 2019* (pp. 2567–2577). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/D19-1259 doi: 10.18653/V1/D19-1259

Jin, Q., Kim, W., Chen, Q., Comeau, D. C., Yeganova, L., Wilbur, W. J., & Lu, Z. (2023). Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinform.*, *39*(10). Retrieved from https://doi.org/10.1093/bioinformatics/btad651 doi: 10.1093/BIOINFORMATICS/BTAD651

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.

Judea, A., & Strube, M. (2016, December). Incremental global event extraction. In *Proceedings of COLING 2016, the 26th international conference on computational*

linguistics: Technical papers (pp. 2279–2289). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/C16-1215

Keung, P., Lu, Y., Salazar, J., & Bhardwaj, V. (2020, November). Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 549–554). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.40 doi: 10.18653/v1/2020.emnlp-main.40

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1746–1751).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations.*

Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th international conference on learning representations.* Retrieved from https://openreview.net/forum?id=SJU4ayYgl

Krithara, A., Nentidis, A., Bougiatiotis, K., & Paliouras, G. (2023). Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, *10*, 170. Retrieved from https://doi.org/10.1038/s41597-023-02068-4

Kull, M., & Flach, P. A. (2014). Patterns of dataset shift..

Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., & Wornell, G. (2018). Co-regularized alignment for unsupervised domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi,

& R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2018/file/99607461cdb9c26e2bd5f31b12dcf27a-Paper.pdf

Kumar, M. P., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. In *Advances in neural information processing systems* (pp. 1189–1197).

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. In *Advances in neural information processing systems.*

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020, November). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4483–4499). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.363 doi: 10.18653/v1/2020.emnlp-main.363

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Icml workshop on challenges in representation learning.*

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (Sixteenth ed.). Dallas, TX, USA: SIL International.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020,*

virtual. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization.. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16067

Li, H., & Gong, M. (2017). Self-paced convolutional neural networks. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 2110–2116). Retrieved from https://doi.org/10.24963/ijcai.2017/293 doi: 10.24963/ijcai.2017/293

Li, Q., Ji, H., & Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., ... Zhou, M. (2020). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP).* Association for Computational Linguistics. Retrieved from https://doi.org/10.18653%2Fv1%2F2020.emnlp-main.484 doi: 10.18653/v1/2020.emnlp-main.484

Limisiewicz, T., Mareček, D., & Rosa, R. (2020, November). Universal Dependencies According to BERT: Both More Specific and More General. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2710–2722). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.245 doi: 10.18653/v1/2020.findings-emnlp.245

Lin, C., Bethard, S., Dligach, D., Sadeque, F., Savova, G. K., & Miller, T. A. (2020). Does bert need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association (JAMIA)*.

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Ieee international conference on computer vision (iccv)* (p. 2999-3007). IEEE Computer Society. Retrieved from `http://dblp.uni-trier.de/db/conf/iccv/iccv2017.html#LinGGHD17`

Lin, Y., Ji, H., Huang, F., & Wu, L. (2020, July). A joint neural model for information extraction with global features. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7999–8009). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-main.713` doi: 10.18653/v1/2020.acl-main.713

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., ... Neubig, G. (2019, July). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3125–3135). Florence, Italy: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P19-1301` doi: 10.18653/v1/P19-1301

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017, April). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 8–14). Valencia, Spain: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/E17-2002`

Liu, P., Qiu, X., & Huang, X. (2017, July). Adversarial multi-task learning for text classification. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1–10). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1001 doi: 10.18653/v1/P17-1001

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Schwan, V. (2019). Roberta: A robustly optimized bert pretraining approach. In *arxiv preprint arxiv:1907.11692.*

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129-137. doi: 10.1109/TIT.1982.1056489

Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). *Learning transferable features with deep adaptation networks.*

Lu, Q., Dou, D., & Nguyen, T. H. (2021, November). Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 3855–3865). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.325 doi: 10.18653/v1/2021.findings-emnlp.325

Lu, Z. (2011). Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database J. Biol. Databases Curation*, *2011*. Retrieved from https://doi.org/10.1093/database/baq036 doi: 10.1093/DATABASE/BAQ036

Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., & Hajishirzi, H. (2019, June). A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 conference of the north American chapter of*

the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 3036–3046). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1308 doi: 10.18653/v1/N19-1308

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., & White, A. D. (2023). *Paperqa: Retrieval-augmented generative agent for scientific research.*

Malkin, D., Limisiewicz, T., & Stanovsky, G. (2022). A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies.* Association for Computational Linguistics. Retrieved from https://doi.org/10.18653%2Fv1%2F2022.naacl-main.361 doi: 10.18653/v1/2022.naacl-main.361

Manes, I., Ronn, N., Cohen, D., Ilan Ber, R., Horowitz-Kugler, Z., & Stanovsky, G. (2024, August). K-QA: A real-world medical Q&A benchmark. In D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, & J. Tsujii (Eds.), *Proceedings of the 23rd workshop on biomedical natural language processing* (pp. 277–294). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.bionlp-1.22/ doi: 10.18653/v1/2024.bionlp-1.22

Miwa, M., & Sasaki, Y. (2014, October). Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1858–1869). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1200 doi: 10.3115/v1/D14-1200

Moran, S., McCloy, D., & Wright, R. (Eds.). (2014). *Phoible online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://phoible.org/`

Naik, A., Parasa, S., Feldman, S., Wang, L. L., & Hope, T. (2022, July). Literature-augmented clinical outcome prediction. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Findings of the association for computational linguistics: Naacl 2022* (pp. 438–453). Seattle, United States: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.findings-naacl.33/` doi: 10.18653/v1/2022.findings-naacl.33

Naik, A., & Rosé, C. (2020a). Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Naik, A., & Rosé, C. (2020b). Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the annual meeting of the association for computational linguistics (acl).*

Ngo, N. T., Min, B., & Nguyen, T. H. (2022). Unsupervised domain adaptation for joint information extraction. In *Findings of the association for computational linguistics: Emnlp 2022* (pp. 5894–5905).

Ngo, N. T., Nguyen, C. V., Dernoncourt, F., & Nguyen, T. H. (2024). Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. *arXiv preprint arXiv:2411.09213*.

Ngo, N. T., & Nguyen, T. H. (2023). Zero-shot cross-lingual transfer learning with multiple source and target languages for information extraction: Language selection and adversarial training. *arXiv preprint arXiv:2411.08785*.

Ngo, N. T., Phung, D., & Nguyen, T. H. (2021a). Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (p. 4015–4025). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.351 doi: 10.18653/v1/2021.findings-acl.351

Ngo, N. T., Phung, D., & Nguyen, T. H. (2021b). Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (p. 4015–4025). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.351 doi: 10.18653/v1/2021.findings-acl.351

Nguyen, D. Q., Nguyen, D. Q., & Pham, S. B. (2021). Joint extraction of entities and relations using a novel tagging scheme. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 1067–1078).

Nguyen, M. V., Lai, V., & Nguyen, T. H. (2021a). Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt)* (p. 27-38). Retrieved from https://doi.org/10.18653/v1/2021.naacl-main.3

Nguyen, M. V., Lai, V. D., & Nguyen, T. H. (2021b). Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt)*.

Nguyen, M. V., Min, B., Dernoncourt, F., & Nguyen, T. (2022, July). Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4363–4374). Seattle, United States: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.naacl-main.324` doi: 10.18653/v1/2022.naacl-main.324

Nguyen, T. H., Cho, K., & Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt).*

Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms. In *arxiv preprint arxiv:1803.02999.*

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *CoRR*, *abs/2303.13375*. Retrieved from `https://doi.org/10.48550/arXiv.2303.13375` doi: 10.48550/ARXIV.2303.13375

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8026–8037).

Pelloni, O., Shaitarova, A., & Samardzic, T. (2022, December). Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7428–7445). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.emnlp-main.503` doi: 10.18653/v1/2022.emnlp-main.503

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, *abs/2302.12813*. Retrieved from `https://doi.org/10.48550/arXiv.2302.12813` doi: 10.48550/ARXIV.2302.12813

Pfeiffer, J., Kaur, A., & Roth, B. (2020). Adapterfusion: Non-destructive task composition for transfer learning. In *Conference on empirical methods in natural language processing* (pp. 3487–3503).

Phang, J., Calixto, I., Htut, P. M., Pruksachatkun, Y., Liu, H., Vania, C., ... Bowman, S. R. (2020). English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing* (pp. 557–575).

Phung, D., Minh Tran, H., Nguyen, M. V., & Nguyen, T. H. (2021, November). Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the 1st workshop*

on multilingual representation learning (pp. 62–73). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.mrl-1.6 doi: 10.18653/v1/2021.mrl-1.6

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How multilingual is multilingual BERT? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1493 doi: 10.18653/v1/P19-1493

Ponti, E. M., Reichart, R., Korhonen, A., & Vulić, I. (2018, July). Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1531–1542). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1142 doi: 10.18653/v1/P18-1142

Pouran Ben Veyseh, A., Nguyen, M. V., Dernoncourt, F., & Nguyen, T. (2022, July). MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2286–2299). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.166 doi: 10.18653/v1/2022.naacl-main.166

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108).

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Dataset bias in classification: The effect of stratification. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment* (pp. 38–49). Springer.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raghu, A., Raghu, M., Bengio, S., & Vinyals, O. (2019). Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, *11*, 1316–1331. Retrieved from `https://doi.org/10.1162/tacl_a_00605` doi: 10.1162/TACL_A_00605

Ren, Y., Zhao, P., Sheng, Y., Yao, D., & Xu, Z. (2017). Robust softmax regression for multi-class classification with self-paced learning. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 2641–2647). Retrieved from `https://doi.org/10.24963/ijcai.2017/368` doi: 10.24963/ijcai.2017/368

Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, *3*(4), 333–389. Retrieved from `https://doi.org/10.1561/1500000019` doi: 10.1561/1500000019

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020, Dec). A primer in bertology: What we know about how bert works. *Transactions of the Association for*

Computational Linguistics, *8*, 842–866. Retrieved from http://dx.doi.org/10.1162/tacl_a_00349 doi: 10.1162/tacl_a_00349

Roth, D., & Yih, W.-t. (2004, May 6 - May 7). A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004* (pp. 1–8). Boston, Massachusetts, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-2401

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., ... Johnson, M. (2021). Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Retrieved from http://dx.doi.org/10.18653/v1/2021.emnlp-main.802 doi: 10.18653/v1/2021.emnlp-main.802

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

Seganti, A., Sorokin, K., Reddy, A. P., & Zhao, I. L. (2021). Smiler: A samsung multilingual entity and relation extraction dataset. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1390–1400).

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., & Meng, D. (2019). Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in neural information processing systems.*

Shu, R., Bui, H. H., Narui, H., & Ermon, S. (2018). A DIRT-T approach to unsupervised domain adaptation. *CoRR*, *abs/1802.08735*. Retrieved from http://arxiv.org/abs/1802.08735

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., ... Natarajan, V. (2023). *Towards expert-level medical question answering with large language models.*

Srinivasan, A., & Athey, S. (2021). Estimation and inference with trees. In *International conference on machine learning* (pp. 9884–9894).

Srinivasan, A., Sitaram, S., Ganu, T., Dandapat, S., Bali, K., & Choudhury, M. (2021). Predicting the performance of multilingual nlp models. *ArXiv*, *abs/2110.08875*.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.*, *40*(12), 3358-3378. Retrieved from http://dblp.uni-trier.de/db/journals/pr/pr40.html#SunKWW07

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Ieee conference on computer vision and pattern recognition (cvpr)* (p. 1199-1208). IEEE Computer Society. Retrieved from http://dblp.uni-trier.de/db/conf/cvpr/cvpr2018.html#SungYZXTH18

Tenney, I., Das, D., & Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *International conference on learning representations.*

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930–1940.

Trung, N. N., Van, L. N., & Nguyen, T. H. (2022, October). Unsupervised domain adaptation for text classification via meta self-paced learning. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4741–4752). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2022.coling-1.420

Turc, I., Lee, K., Eisenstein, J., Chang, M.-W., & Toutanova, K. (2021). *Revisiting the primacy of english in zero-shot cross-lingual transfer.*

Veyseh, A., Nguyen, T. N., & Nguyen, T. H. (2020, November). Graph transformer networks with syntactic and semantic structures for event argument extraction. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3651–3661). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.326 doi: 10.18653/v1/2020.findings-emnlp.326

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., & Wierstra, D. (2016). Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2005). Ace 2005 multilingual training corpus. In *Technical report, linguistic data consortium.*

Wang, J., Yang, Z., Yao, Z., & Yu, H. (2024). *Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability.*

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Wei, C., Shen, K., Chen, Y., & Ma, T. (2021). Theoretical analysis of self-training with deep networks on unlabeled data. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=rC8sJ4i6kaH`

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 1–40.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.emnlp-demos.6`

Wright, D., & Augenstein, I. (2020). Transformer based multi-source domain adaptation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). *Pmc-llama: Towards building open-source language models for medicine.*

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *CoRR*, *abs/2402.13178*. Retrieved from https://doi.org/10.48550/arXiv.2402.13178 doi: 10.48550/ARXIV.2402.13178

Xiong, W., Chen, X., Zhao, X., Wang, Y., Jiang, Z., Wen, Z., & Lin, J. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 19390–19398.

Xu, N., Gui, T., Ma, R., Zhang, Q., Ye, J., Zhang, M., & Huang, X. (2022, December). Cross-linguistic syntactic difference in multilingual BERT: How good is it and how does it affect transfer? In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 8073–8092). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.552

Xu, Z., He, H., Lee, G.-H., Wang, B., & Wang, H. (2022). Graph-relational domain adaptation. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=kcwyXtt7yDJ

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). *Corrective retrieval augmented generation.* (arXiv preprint arXiv:2401.15884)

Yang, B., & Mitchell, T. M. (2016). Joint extraction of events and entities within a document context. In *Proceedings of the 2016 conference of the north*

american chapter of the association for computational linguistics: Human language technologies (naacl-hlt) (pp. 57–62).

Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2010). Discriminatively trained neural sequence models for speech recognition. In *International conference on spoken language processing* (pp. 2670–2673).

Yuan, J., Zhao, Y., Qin, B., & Liu, T. (2021). *Learning to share by masking the non-shared for multi-domain sentiment classification.* (arXiv preprint arXiv:2104.08480)

Yun, S., Jeong, M., Kim, R., Kang, J., & Kim, H. J. (2019). Graph transformer networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 11983–11993). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2019/file/9d63484abb477c97640154d40595a3bb-Paper.pdf`

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., & Saminger-Platz, S. (2017). *Central moment discrepancy (cmd) for domain-invariant representation learning.*

Zhang, J., Qin, Y., Zhang, Y., Liu, M., & Ji, D. (2019, 7). Extracting entities and events as a single task using a transition-based neural model. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 5422–5428). International Joint Conferences on Artificial Intelligence Organization. Retrieved from `https://doi.org/10.24963/ijcai.2019/753` doi: 10.24963/ijcai.2019/753

Zhang, Z., & Ji, H. (2021, June). Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 39–49). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.4 doi: 10.18653/v1/2021.naacl-main.4

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017, July). Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1227–1236). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-1113 doi: 10.18653/v1/P17-1113