Predicting Hallucinations before LLMs Make Them:

Combining External Knowledge with Behavioral Uncertainty
for Black-Box Auditing of Large Language Models

by

Murad Mikayilzade

A thesis accepted and approved in partial
fulfillment of the requirements for the degree of
Master of Science
in Computer Science

Committee:
Dr. Yu Wang, Chair
Dr. Thanh Nguyen, Member

University of Oregon

Spring 2026

© 2026 Murad Mikayilzade

Abstract

Murad Mikayilzade

Master of Science, Department of Computer Science

Title: Predicting Hallucinations before LLMs Make Them: Combining External Knowledge with Behavioral Uncertainty for Black-Box Auditing of Large Language Models

Large language models often produce confident but factually incorrect answers, a phenomenon known as hallucination, particularly on questions involving long-tail entities that appear infrequently in training data. Predicting which questions will trigger hallucinations is critical for safe deployment, but most strong predictive methods require white-box access to model internals such as attention weights or hidden states, which users accessing models through commercial APIs cannot obtain. This thesis presents BB-HP (Black-Box Hallucination Predictor), a method that operates entirely from outside the model and produces a hallucination probability for each question, using only the inputs and outputs available to an ordinary API user.

BB-HP combines two independent and complementary signals. The first is epistemic: features derived from an external knowledge base (Wikidata) that capture how well-known the question's subject is, the structural richness of the knowledge graph surrounding it, and the semantic alignment between the question and the retrieved facts. The second is behavioral: the model is queried multiple times, its sampled answers are clustered by meaning using a natural language inference classifier, and the semantic entropy of those clusters is measured. A logistic regression probe combines these signals into a single hallucination probability.

The method is evaluated on two thousand PopQA questions across four large language models spanning open-source and closed-source families: Qwen2.5-7B-Instruct, Llama-3-8B-Instruct, GPT-OSS-20B, and Claude Haiku 4.5. BB-HP achieves an area under the ROC curve (AUROC) between 0.87 and 0.90 across all four models, consistently outperforming the strongest individual baseline. Its advantage over behavioral signals alone grows substantially as the sampling budget shrinks, indicating that external knowledge features compensate for noisier behavioral signal in cost-constrained deployment. A cross-model transfer study shows that the trained predictor generalizes across language models with minimal degradation, providing what is, to our knowledge, the first dedicated evaluation of cross-model portability for hallucination predictors. Together these

results indicate that effective hallucination prediction does not require access to model internals.

Acknowledgments

I would like to thank my advisor, Dr. Yu Wang, for his guidance throughout this work, and in particular for the auditor framework that shaped the experimental design of this thesis. I am also grateful to my defense committee member, Dr. Thanh Nguyen, for her time and feedback.

Above all, I want to thank my family and friends. Their patience, encouragement, and unwavering support carried me through every stage of this work and of my studies, and I am deeply grateful for everything they have done for me.

Finally, I am grateful to all of the professors and teachers who shared their knowledge with me, and to every person who, in ways large and small, helped shape who I am. This work would not have been possible without them.

Contents

Abstract **3**

Acknowledgments **5**

1 Introduction **11**
1.1 Motivation 11
1.2 The black-box constraint 11
1.3 Two sources of signal 12
1.4 Contributions 13
1.5 Thesis organization 13

2 Related Work **15**
2.1 Behavioral consistency methods 15
2.2 Knowledge-grounded methods 16
2.3 Confidence-elicitation methods 16
2.4 Methods requiring internal state 17
2.5 Positioning of this work 18

3 Method **19**
3.1 Overview 19
3.2 A worked example 22
3.3 Problem formulation 22
3.4 The epistemic channel: knowledge-graph retrieval 23
3.5 The epistemic channel: knowledge-graph features 23
3.6 The behavioral channel: semantic-entropy features 24
3.7 The probe 25
3.8 Implementation details 26
3.9 Feature summary 27

4 Experimental Setup **29**

4.1 Dataset . . . 29
4.2 Language models . . . 30
4.3 Labeling protocol . . . 30
4.4 Evaluation methodology . . . 32

5 Results . . . 34
5.1 Main results . . . 34
5.2 Ablation analysis . . . 37
5.3 Selective prediction . . . 37
5.4 Robustness . . . 39
5.5 On overfitting . . . 41

6 Cross-Model Transfer . . . 43
6.1 Motivation . . . 43
6.2 Experimental design . . . 43
6.3 Results . . . 44
6.4 Interpretation . . . 45

7 Discussion and Limitations . . . 47
7.1 What the results establish . . . 47
7.2 Interpreting the modest in-distribution gains . . . 47
7.3 Limitations . . . 48

8 Conclusion and Future Work . . . 50
8.1 Summary . . . 50
8.2 Future work . . . 51
8.3 Closing remarks . . . 52

Bibliography . . . 55

List of Tables

2.1. Positioning of BB-HP relative to representative prior methods. “External KG” indicates use of a knowledge base independent of the language model; “Behavioral” indicates use of output sampling consistency. 18

3.1. Notation used in this chapter. 21
3.2. The BB-HP feature set: fifteen epistemic and three behavioral features, eighteen in total. The two self-report baselines are listed for reference and are not part of BB-HP. The count column gives the number of scalar features in each row. 28

4.1. Baseline behavior of the four models on the two thousand sampled PopQA questions. Counts are out of 2000. The hallucination rate is computed over answered questions only (excluding refusals) as confidently wrong divided by the number answered; refusals are still labeled as hallucinations in the prediction task. Counts and rates are post Unicode canonicalization. 32

5.1. AUROC by method and model on the held-out evaluation set, with bootstrap 95% confidence intervals. The strongest method on each model is shown in bold. The two self-report baselines (P(True) and verbalized confidence) are standalone baselines and are not part of BB-HP. 35
5.2. Selective accuracy at twenty-five percent coverage, averaged across models. Higher is better. The method abstains on the riskiest seventy-five percent of questions and reports accuracy on the retained quarter. The no-filtering row is the mean correctness rate across models when all questions are answered. 38
5.3. AUROC as a function of behavioral sampling budget, pooled across models. The advantage of BB-HP over the behavioral channel grows as the budget shrinks. . . . 39
5.4. AUROC as a function of knowledge-graph subgraph size, pooled across models. The behavioral channel does not depend on the subgraph and is held fixed as a reference. Performance saturates at ten triples. 41

5.5. AUROC by probe model class, pooled across models. Logistic regression performs best in absolute terms, but the advantage over the behavioral channel is largest with the random forest. 41

List of Figures

3.1. Overview of the BB-HP pipeline. The behavioral channel (right) samples the model and computes semantic-entropy features; the epistemic channel (left) retrieves an external subgraph and computes knowledge-graph features. A logistic regression probe combines both into a hallucination probability. 20

5.1. Hallucination prediction AUROC by method and model. BB-HP (rightmost bar in each group) is the strongest method on every model. The behavioral channel (semantic entropy) is the strongest individual baseline, while the self-report baselines are weakest. 35

5.2. Risk-coverage tradeoff. As coverage decreases (the method answers fewer, safer questions), accuracy on the answered set rises. BB-HP dominates the behavioral channel across the coverage range, and both far exceed the self-report baselines and the no-filtering reference. 38

5.3. The advantage of BB-HP over the behavioral channel grows as the sampling budget shrinks from twenty to five samples per question. The annotated gaps show the AUROC advantage at each budget. This is the regime where real deployments operate. 40

6.1. Cross-model transfer matrix for BB-HP. Each cell reports the AUROC of a probe trained on the row model and evaluated on the column model. The diagonal is the in-distribution setting; off-diagonal entries measure transfer. Values change little across rows, indicating that the predictor generalizes across models. 44

Chapter 1

Introduction

1.1 Motivation

Large language models (LLMs) have become widely used tools for answering factual questions, drafting documents, and supporting decision-making across many domains. Yet despite their fluency, these models share a persistent and dangerous failure mode: they produce confident, well-formed answers that are simply false. This phenomenon, commonly called *hallucination*, is one of the central obstacles to deploying language models in settings where factual reliability matters (Ji et al., 2023; Huang et al., 2025).

The danger of hallucination lies not in the existence of errors, which any system makes, but in the absence of any signal accompanying them. A model that confidently states an incorrect screenwriter, an invented publication date, or a fabricated biographical detail provides no indication that it is guessing. The user, unable to distinguish a confident correct answer from a confident incorrect one, must either independently verify every output, defeating the purpose of automation, or accept the risk of acting on false information.

This thesis asks a specific question: *can we predict, from outside a language model, whether it is likely to hallucinate on a given question?* If such prediction is possible, a deployed system could abstain on questions it is likely to get wrong, defer them to a human or a retrieval system, or flag them for verification. The value of accurate hallucination prediction is therefore directly tied to the reliability of any system built on top of a language model.

1.2 The black-box constraint

A large body of recent work has shown that hallucinations can often be predicted by examining a model's internal state during generation, such as the entropy of its token distributions, the geometry of its hidden representations, or the concentration of its attention weights (Kadavath

et al., 2022; Azaria and Mitchell, 2023; Chen et al., 2024; Lin et al., 2025). These methods can be effective, but they share a fundamental limitation: they require *white-box* access to the model's internals.

In practice, the vast majority of language model usage occurs through commercial application programming interfaces (APIs). A user sends a prompt and receives a text response. The model's weights, attention patterns, and hidden states are never exposed. For the most capable proprietary models, no amount of engineering effort will recover these internals, because the provider does not expose them. Any hallucination-prediction method that depends on internal state is therefore inapplicable to the precise deployment setting where it would be most useful.

This motivates the central design constraint of this thesis: the prediction method must operate in a *black-box* setting, using only the inputs and outputs that an ordinary API user can observe. It may send queries to the model and read the responses, but it may not inspect anything inside.

1.3 Two sources of signal

If we cannot look inside the model, what can we use to predict hallucination? This thesis combines two distinct and independent sources of evidence.

The first source is *epistemic*, concerning the state of knowledge about the world. For a given question, we can ask whether the relevant facts are well-documented in an external knowledge base, independent of any particular language model. If a question concerns an obscure entity about which little is recorded, a model is far more likely to lack the information and therefore to fabricate an answer. The popularity of an entity, the richness of the structured knowledge surrounding it, and the degree to which a question's content aligns with the available facts together form an epistemic profile of the question's difficulty. Crucially, this profile can be computed entirely from the question and an external knowledge base, without ever querying the model.

The second source is *behavioral*, concerning how the model responds. By querying the model multiple times with the same question at a nonzero sampling temperature, we obtain a set of answers that may agree or disagree. A model that is confident in its knowledge tends to produce consistent answers; a model that is guessing tends to produce answers that scatter across many distinct possibilities. Measuring the semantic entropy of these sampled answers, after grouping them by meaning, provides a behavioral signal of the model's uncertainty (Kuhn et al., 2023; Farquhar et al., 2024).

These two signals are independent in an important sense. A model can be behaviorally confident yet factually wrong, having confidently memorized an incorrect association. Conversely, an

entity may be well-documented in a knowledge base, yet the model may still fail to retrieve the relevant fact at answer time. Because the two signals capture different failure modes, combining them offers the prospect of catching hallucinations that either signal alone would miss.

1.4 Contributions

This thesis makes the following contributions.

First, it introduces BB-HP (Black-Box Hallucination Predictor), a method that combines external knowledge-graph features with behavioral semantic-entropy features to predict hallucination in a fully black-box setting. We are not aware of prior work that combines these two specific signal types under the black-box constraint. The method is evaluated on four language models, including a proprietary model accessible only through an API, demonstrating applicability in settings where white-box methods structurally cannot operate.

Second, it presents what is, to our knowledge, the first dedicated study of *cross-model transfer* for hallucination predictors. A predictor trained on data from one language model is evaluated on questions answered by a different model. The finding, that such predictors transfer across models with only minor degradation, runs counter to the implicit assumption in much of the literature that uncertainty signals are model-specific and predictors must be retrained for each new model.

Third, it quantifies the *cost-quality tradeoff* of behavioral signal collection. By varying the number of samples drawn from the model, the thesis shows that the advantage of combining knowledge-graph features with behavioral features grows as the sampling budget shrinks. This finding indicates that the method is most valuable precisely in the cost-constrained regime where real production systems operate, rather than only in the high-budget regime typical of research benchmarks.

1.5 Thesis organization

The remainder of this thesis is organized as follows. Chapter 2 surveys related work on hallucination detection, situating BB-HP relative to behavioral methods, knowledge-grounded methods, confidence-elicitation methods, and methods that rely on internal model state. Chapter 3 describes the design of BB-HP in detail, including the knowledge-graph retrieval procedure, the three families of epistemic features, the behavioral semantic-entropy features, and the probe that combines them. Chapter 4 describes the dataset, the four language models, the labeling protocol, and the evaluation methodology. Chapter 5 presents the main experimental results, including

in-distribution prediction accuracy, ablation analysis, selective prediction, and robustness checks. Chapter 6 presents the cross-model transfer study. Chapter 7 discusses the findings, their interpretation, and the limitations of the work. Chapter 8 concludes and outlines directions for future work.

Chapter 2

Related Work

The problem of detecting and predicting hallucination in language models has attracted substantial attention. This chapter organizes the prior work into four threads: behavioral methods that examine the consistency of model outputs, knowledge-grounded methods that compare outputs against external sources, confidence-elicitation methods that ask the model to assess itself, and methods that rely on internal model state. The chapter concludes by positioning BB-HP relative to each thread and identifying the gap it fills.

2.1 Behavioral consistency methods

The most influential thread of black-box hallucination detection rests on a simple observation: when a model knows an answer, repeated queries tend to produce consistent responses, whereas when it is guessing, responses scatter. Translating this intuition into a usable signal requires deciding what it means for two free-form responses to be "consistent."

Kuhn et al. (2023) introduced *semantic entropy*, which addresses this by clustering sampled responses according to meaning rather than surface form. Two answers such as "Paris" and "It is Paris" express the same proposition and should be treated as a single cluster, whereas "Paris" and "London" should not. The clustering is performed using a natural language inference (NLI) model that judges whether two responses bidirectionally entail each other. The entropy of the resulting cluster distribution then serves as an uncertainty estimate: low entropy when responses concentrate in a single meaning cluster, high entropy when they fragment.

Farquhar et al. (2024) extended this line of work and demonstrated, in a study published in *Nature*, that semantic entropy reliably detects hallucinations across a range of question-answering and generation tasks, outperforming simpler lexical consistency measures. Their work established semantic entropy as a strong black-box baseline and is the most direct point of comparison for the behavioral component of BB-HP.

A related approach, SelfCheckGPT (Manakul et al., 2023), samples multiple responses and

checks the factual consistency of a candidate answer against the samples, without requiring any external resource. These methods share a common strength, namely that they require no model internals, and a common weakness, namely that their reliability depends on drawing enough samples to estimate consistency accurately. When the sampling budget is small, the consistency estimate becomes noisy. This dependence on sampling budget is examined directly in Chapter 5.

2.2 Knowledge-grounded methods

A second thread compares model outputs against external knowledge. Retrieval-augmented generation (Lewis et al., 2020) reduces hallucination by supplying the model with retrieved documents at generation time, though it addresses the problem by prevention rather than by detection. For detection specifically, several methods compare a generated answer against retrieved evidence and flag discrepancies; this thesis instead uses external knowledge to estimate question difficulty before any answer is examined, rather than to verify a produced answer.

The popularity of an entity has been shown to be a strong predictor of whether a model can answer questions about it. Mallen et al. (2023) introduced the PopQA benchmark precisely to study this relationship, demonstrating that language models reliably answer questions about popular entities while failing on long-tail entities, and that scaling model size does not appreciably improve long-tail performance. The strength of the correlation between entity popularity and answer accuracy directly motivates the popularity features used in BB-HP.

Knowledge graphs have also been used more structurally. Some black-box detection methods extract knowledge-graph triples from the model's own generated outputs and check their mutual consistency. While such methods use graph structure, the graphs they build originate from the model's outputs rather than from an external source, so they are better understood as a structured form of behavioral consistency checking than as an injection of independent external knowledge. This distinction is central to positioning BB-HP, which draws its graph features from Wikidata (Vrandečić and Krötzsch, 2014), a knowledge base entirely independent of any language model.

2.3 Confidence-elicitation methods

A third thread asks the model to assess its own reliability directly. Kadavath et al. (2022) showed that models can be prompted to estimate the probability that their own answer is correct, a signal often termed P(True), and that this self-assessment carries meaningful information about correctness. Lin et al. (2022) and Tian et al. (2023) explored verbalized confidence, in which the

model is asked to express its certainty in words or on a numeric scale.

These methods are attractive because they are inexpensive, requiring only one or two additional queries, and require no external resources. However, their reliability varies considerably across models and depends heavily on how the model was trained, particularly whether it underwent reinforcement learning from human feedback. In this thesis, both P(True) and verbalized confidence are included as baselines, and their relatively weak performance compared to behavioral and knowledge signals is documented in Chapter 5.

2.4 Methods requiring internal state

A fourth thread achieves strong detection performance by examining model internals. Kadavath et al. (2022) studied internal probes alongside their confidence-elicitation work. Azaria and Mitchell (2023) showed that a classifier trained on a model's hidden activations (SAPLMA) can predict whether a generated statement is truthful, establishing that truthfulness is partly linearly decodable from internal representations. Chen et al. (2024) (INSIDE) detected hallucinations from the eigenvalue spectrum of internal sentence embeddings across sampled responses, and Chuang et al. (2024) (Lookback Lens) detected contextual hallucinations from the ratio of attention paid to provided context versus newly generated tokens. More recently, the Graph-Grounded Attribution (GGA) framework (Lin et al., 2025) introduced two interpretability-based metrics for graph-augmented generation. The first, Path Reliance Degree, measures how the model's attention concentrates on the shortest reasoning path within a retrieved subgraph. The second, Semantic Alignment Score, measures how well the model's internal feed-forward representations align with the retrieved knowledge.

GGA is conceptually important to this thesis because it decomposes the epistemic signal into a structural component (how the model traverses the graph) and an alignment component (how the model's representations match the graph). BB-HP adapts this two-channel decomposition. However, GGA, like the other methods in this thread, computes its metrics from the model's internal attention weights and hidden states, making it a fundamentally white-box method that cannot be applied to models accessed through an API. BB-HP preserves the structural-and-alignment decomposition but reconstructs both channels from external information alone: the structural channel from the topology of the retrieved subgraph, and the alignment channel from text embeddings comparing the question to the retrieved facts. This adaptation is what allows BB-HP to run on a proprietary model such as Claude Haiku, which the white-box methods structurally cannot.

2.5 Positioning of this work

Table 2.1 summarizes the landscape. Behavioral methods are black-box but use only model outputs. Knowledge-grounded methods inject external information, though some structured black-box variants draw their graphs from the model itself rather than from an independent source. Confidence-elicitation methods are cheap but weak. Methods such as GGA, SAPLMA, and INSIDE achieve strong performance but require white-box access.

Table 2.1: Positioning of BB-HP relative to representative prior methods. "External KG" indicates use of a knowledge base independent of the language model; "Behavioral" indicates use of output sampling consistency.

Method	Black-box	External KG	Behavioral
Semantic Entropy (Farquhar et al., 2024)	Yes	No	Yes
SelfCheckGPT (Manakul et al., 2023)	Yes	No	Yes
P(True) (Kadavath et al., 2022)	Yes	No	No
SAPLMA (Azaria and Mitchell, 2023)	No	No	No
INSIDE (Chen et al., 2024)	No	No	Yes
GGA (Lin et al., 2025)	No	Yes	No
BB-HP (this work)	Yes	Yes	Yes

BB-HP occupies the intersection that no prior method in the table does: it is black-box, it injects genuinely external knowledge, and it incorporates behavioral signal. The behavioral component builds directly on semantic entropy; the knowledge component adapts the structural-and-alignment intuition of GGA to a black-box setting; and the combination of the two, evaluated across multiple models with a cross-model transfer study, is the novel contribution of this thesis.

Chapter 3

Method

This chapter describes BB-HP in detail. The method takes a question, gathers two independent kinds of evidence about it, converts that evidence into a numeric feature vector, and feeds the vector to a probe that outputs a hallucination probability. Section 3.1 gives the high-level structure. Section 3.3 formalizes the prediction problem. Sections 3.4 through 3.6 describe the two channels. Section 3.7 describes the probe and the rationale for its choice. Section 3.8 documents the implementation in enough detail to reproduce it.

Throughout this thesis, two channels supply the evidence for prediction. The *epistemic channel* derives features from an external knowledge graph (Wikidata) and describes what the world records about a question's subject. The *behavioral channel* derives features from the consistency of the model's sampled answers and is built on *semantic entropy*. Two additional signals, the model's self-assessment of correctness (P(True)) and its verbalized numeric confidence, are referred to collectively as the *self-report baselines*; they are evaluated as standalone baselines and are not part of BB-HP.

3.1 Overview

Following the auditor framing that structures this work, let G denote the language model under audit and M denote the auditor. The language model maps a question Q to an answer, $G(Q) \to Y$. The auditor's task is to decide whether Y is a trustworthy answer to Q. Two versions of the auditor are of interest:

$$M(Q,Y) \to \text{hallucination probability}, \tag{3.1}$$

$$M(Q,Y,\mathcal{G}_Q) \to \text{hallucination probability}. \tag{3.2}$$

The first auditor uses only the question and the model's sampled answers, that is, the behavioral channel alone. The second additionally uses the retrieved knowledge graph $\mathcal{G}_Q$, adding the epis-

temic channel. The central empirical question of this thesis is whether the second auditor outperforms the first, that is, whether the knowledge graph helps. BB-HP is the second auditor; the behavioral-only auditor is the strongest baseline.

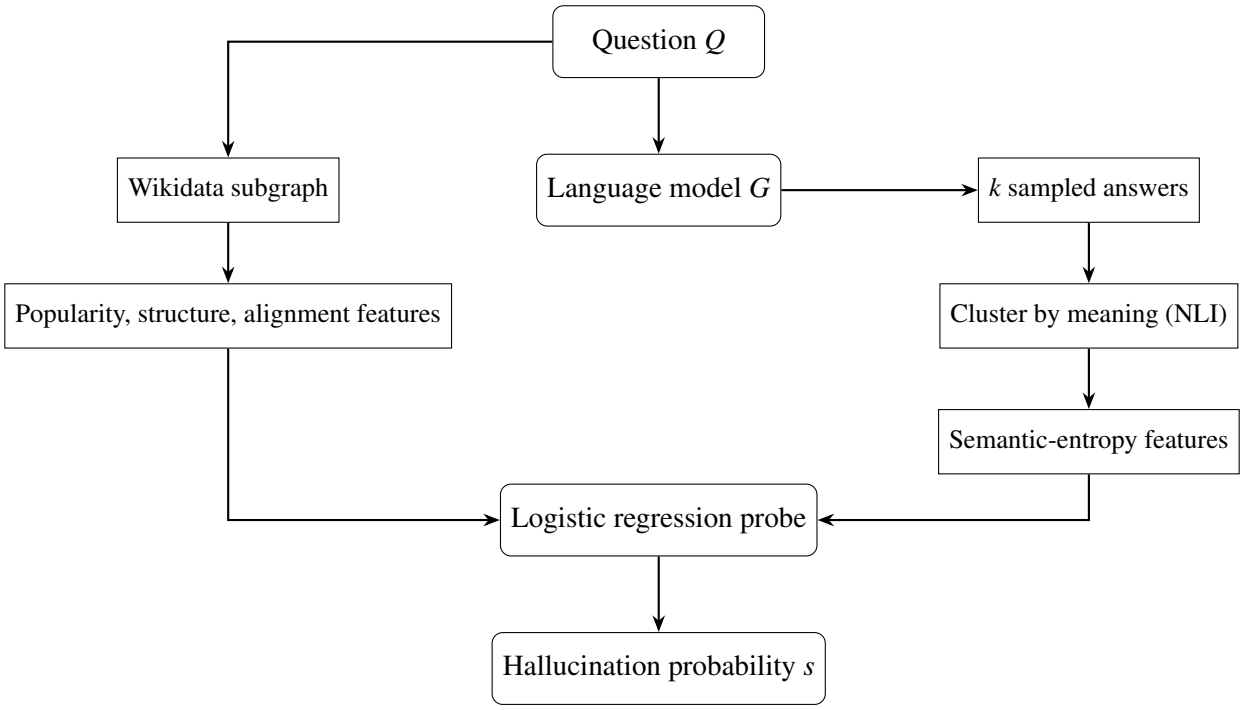


Figure 3.1: Overview of the BB-HP pipeline. The behavioral channel (right) samples the model and computes semantic-entropy features; the epistemic channel (left) retrieves an external subgraph and computes knowledge-graph features. A logistic regression probe combines both into a hallucination probability.

The two channels differ in what they require. The epistemic channel depends only on the question Q and the external knowledge graph, so it can be computed without inspecting the model's answer. The behavioral channel requires the model's sampled answers and is therefore the more expensive of the two, since each sample is a separate query. BB-HP combines both; the cost consequences of the behavioral channel are examined in Chapter 5.

Table 3.1 summarizes the notation used throughout this chapter.

Table 3.1: Notation used in this chapter.

Symbol	Meaning
G	The language model under audit
M	The auditor (the hallucination predictor)
Q	A question
Y	The model's deterministic answer to Q
$a^\star$	The gold answer (with aliases) for Q
$h \in \{0,1\}$	Binary hallucination label (1 = hallucination)
$s \in [0,1]$	Predicted hallucination probability
$\mathscr{S}$	The set of k sampled answers for a question
$\mathscr{G}_Q$	The retrieved knowledge subgraph for Q's subject
$\phi(Q,Y,\mathscr{G}_Q)$	The feature vector fed to the probe

The overall procedure is summarized in Algorithm 1, stated in its general form with sampling temperature τ and sample count k; the concrete values used in the experiments ($\tau = 0.7$, $k = 20$) are given in Chapter 4.

Algorithm 1 BB-HP feature construction and scoring for a single question.

Require: Question Q, language model G, knowledge base KB, trained probe f

$Y \leftarrow G(Q)$ ▷ deterministic answer

$\mathscr{S} \leftarrow \{G(Q) : \text{sampled at temperature } \tau\}$ ▷ k sampled answers

$e \leftarrow \textsc{SubjectEntity}(Q)$ ▷ resolve subject to KB identifier

$\mathscr{G}_Q \leftarrow \textsc{RetrieveSubgraph}(e, KB)$ ▷ up to 50 triples

$\phi_{\text{pop}} \leftarrow \textsc{PopularityFeatures}(e)$

$\phi_{\text{struct}} \leftarrow \textsc{StructureFeatures}(\mathscr{G}_Q)$

$\phi_{\text{align}} \leftarrow \textsc{AlignmentFeatures}(Q, \mathscr{G}_Q)$

$\mathscr{C} \leftarrow \textsc{ClusterByMeaning}(\mathscr{S})$ ▷ NLI-based clustering

$\phi_{\text{sem}} \leftarrow \textsc{SemanticEntropyFeatures}(\mathscr{C})$

$\phi \leftarrow [\phi_{\text{pop}}, \phi_{\text{struct}}, \phi_{\text{align}}, \phi_{\text{sem}}]$

return $s \leftarrow f(\phi)$ ▷ hallucination probability

3.2 A worked example

To make the feature construction concrete, consider the question *"Who was the screenwriter for On the Town?"*, drawn from PopQA. The subject entity is the 1949 film *On the Town*, whose gold answer includes the screenwriters Betty Comden and Adolph Green.

When this question is posed to a weaker model, the twenty sampled answers might scatter across several distinct names, only some of which are correct, producing many meaning clusters and high semantic entropy in the behavioral channel. The epistemic channel, meanwhile, retrieves the film's Wikidata subgraph, which contains facts such as the film's director, production company, release year, and cast. If the subgraph contains few or no facts about the screenwriting credit specifically, the alignment features will register a low maximum similarity between the question and any retrieved triple, signaling that the knowledge graph does not directly contain what the question asks. The popularity features will reflect that a 1949 film is a moderately documented but not highly popular entity. Together these signals raise the predicted hallucination probability, correctly flagging the question as risky.

Contrast this with a question about a highly popular entity, such as the capital of a major country. Here the sampled answers concentrate in a single meaning cluster, semantic entropy is near zero, the subgraph is rich, and the question aligns strongly with a directly relevant triple. All signals point toward a low hallucination probability, and the model is indeed likely to answer correctly. The two examples illustrate how the channels combine to separate easy questions from hard ones.

3.3 Problem formulation

Let $\mathscr{D} = \{(Q_i, a_i^\star)\}_{i=1}^N$ be a dataset of questions paired with gold answers. For a fixed language model G, let $Y_i = G(Q_i)$ be the model's deterministic (greedy) answer. Each question receives a binary label

$$h_i = \begin{cases} 0 & \text{if } Y_i \text{ matches } a_i^\star \text{ (correct)}, \\ 1 & \text{otherwise (hallucination).} \end{cases} \tag{3.3}$$

The matching criterion is described in Chapter 4. The auditor M produces a score $s_i = M(Q_i, Y_i, \cdot) \in [0,1]$ intended to approximate $\Pr(h_i = 1)$. The auditor is evaluated by how well s_i ranks hallucinations above correct answers, measured by the area under the ROC curve (AUROC). Note that the label h is binary, while the predicted score s is continuous.

3.4 The epistemic channel: knowledge-graph retrieval

The epistemic channel draws on Wikidata (Vrandečić and Krötzsch, 2014), a large, community-maintained, structured knowledge graph. Wikidata represents knowledge as a graph in which entities (people, films, places, and so on) are nodes identified by unique codes, and facts are directed, labeled edges connecting them. Each fact is a triple of the form (subject, relation, object); for example, the triple stating that the film *On the Town* has a particular screenwriter connects the film node to the person node via a "screenwriter" relation.

For each question, the subject entity is identified through the entity identifier provided by the dataset. The method then retrieves up to fifty triples incident to that entity, in both directions: facts in which the entity is the subject, and facts in which it is the object of some other entity's relation. The result is a small subgraph, typically containing twenty to fifty connected facts, centered on the question's subject. This subgraph is the raw material from which the three epistemic feature families are computed.

Retrieval is performed through the Wikidata SPARQL query endpoint, with conservative rate limiting (one request roughly every 1.2 seconds) and retry logic to respect the service. Because the retrieved subgraphs depend only on the question's subject and not on the language model, they are computed once and reused across all models under audit.

3.5 The epistemic channel: knowledge-graph features

Three families of features are computed from the retrieved subgraph and from entity metadata. Together they form the epistemic profile of a question and total fifteen features.

Entity popularity. The first family captures how well-known the question's subject is, operationalized through Wikipedia pageview statistics. The intuition, supported by the findings of Mallen et al. (2023), is direct: language models encounter more text about popular entities during training and therefore answer questions about them more reliably. Three features are computed: the raw monthly pageview count, its logarithm (to compress the heavy-tailed distribution), and a discrete popularity tier. These features depend only on the entity, not on the subgraph structure.

Knowledge-graph structure. The second family describes the topology of the retrieved subgraph, adapting the structural intuition of the Path Reliance Degree from GGA (Lin et al., 2025) to a black-box setting. Where GGA measures how the model's attention traverses the graph, this

work measures properties of the graph itself, on the hypothesis that a richer, denser subgraph indicates that the knowledge graph documents the entity well. Seven features are computed: the number of nodes, the number of edges, the graph density, the average node degree, the share of total degree held by the highest-degree node, the proportion of edges that are bridges, and the degree of the central entity relative to the edge count. Some of these features are partially degenerate on the predominantly star-shaped one-hop subgraphs that Wikidata retrieval produces, a limitation discussed in Chapter 7.

Question-knowledge alignment. The third family measures how well the content of the question aligns with the retrieved facts, adapting the Semantic Alignment Score of GGA. Where GGA computes alignment between the model's internal hidden states and the graph, this work computes alignment entirely externally, in text space. Each retrieved triple is linearized into a short sentence, such as "On the Town directed by Gene Kelly." Both the question and each linearized triple are embedded using a Sentence-BERT model (Reimers and Gurevych, 2019), specifically the `all-MiniLM-L6-v2` variant, and the cosine similarity between the question embedding and each triple embedding is computed. Five features summarize the resulting similarity distribution: the mean, maximum, and variance of the per-triple similarities, the similarity of the question to the centroid of the triple embeddings, and the number of triples used. The motivating intuition is that when a question asks about an attribute (say, a screenwriter) but the retrieved facts concern unrelated attributes (directors, release dates), the alignment is poor and the knowledge graph does not contain what the question seeks, raising the risk that the model must fabricate an answer.

3.6 The behavioral channel: semantic-entropy features

The behavioral channel measures the consistency of the model's answers through semantic entropy, following Kuhn et al. (2023) and Farquhar et al. (2024). For each question, the model is queried multiple times at a nonzero sampling temperature, producing a collection of sampled answers. These answers are clustered by meaning rather than surface form, using a DeBERTa (He et al., 2021) natural language inference classifier (the `deberta-large-mnli` variant): two answers are placed in the same cluster when they bidirectionally entail each other, where bidirectional entailment requires that the entailment probability exceed the contradiction and neutral probabilities in both directions. This clustering correctly groups "Catholic," "Roman Catholic," and "Roman Catholic Church" while separating "Catholic" from "Hindu."

From the resulting clustering, three features are computed: the cluster count (the number of distinct meaning clusters), the cluster entropy (the Shannon entropy of the cluster-size distribution,

which is the semantic entropy of the sample), and the top-cluster mass (the proportion of samples falling in the largest cluster). A confident model produces samples that concentrate in a single cluster, yielding low semantic entropy; a guessing model produces samples that fragment across clusters, yielding high semantic entropy.

The choice of an NLI classifier rather than embedding cosine similarity for this clustering is deliberate. Cosine similarity measures how similar two answers are in meaning, but the relevant question for clustering is whether two answers are *equivalent*, a sharper relation that NLI is specifically designed to assess. Two answers can be highly similar in embedding space yet denote different entities.

3.7 The probe

The fifteen epistemic features and three behavioral features are concatenated into the eighteen-dimensional feature vector that defines BB-HP. A logistic regression probe maps this vector to a hallucination probability. Features are standardized to zero mean and unit variance before fitting, and the probe uses balanced class weights to account for the class imbalance discussed in Chapter 4.

The two self-report baselines, P(True) and verbalized confidence, are not part of BB-HP. Each would require an additional model query, and the central question of the thesis concerns the combination of the epistemic and behavioral channels specifically. They are evaluated only as standalone baselines in Chapter 5, alongside the epistemic-only and behavioral-only auditors.

The choice of a linear probe over a deeper model is principled rather than merely convenient. First, a linear probe attributes predictive performance directly to the input features: if the probe performs well, the features themselves carry the signal, whereas a more expressive model would conflate feature quality with the model's own learning capacity. This methodological position follows the linear-probe convention in recent hallucination-prediction work (Pacchiardi et al., 2025). Second, the training data, on the order of sixteen hundred examples per model against eighteen features, sits in a regime where a linear model is appropriately scaled and a deeper model would risk overfitting. As a robustness check, Chapter 5 reports results with random forest (Breiman, 2001) and gradient-boosted tree (Chen and Guestrin, 2016) probes; neither outperforms logistic regression in absolute terms, supporting the conclusion that the bottleneck is feature quality rather than probe expressiveness.

3.8 Implementation details

This section documents the implementation precisely enough to reproduce the method. The pipeline is organized as a sequence of stages, each of which caches its output so that later stages can be rerun without repeating expensive work.

Answer and sample collection. For each (question, model) pair, generation proceeds in a single batched pass. One deterministic answer is produced at temperature 0, and $k = 20$ sampled answers are produced at temperature 0.7 with nucleus sampling (top-$p = 0.9$) and a short maximum generation length appropriate to short factual answers. The open-weight models are served through a hosted inference provider and the proprietary model through its vendor API; all calls use the same chat-style prompt template, which presents the question and requests a direct answer. The deterministic answer is stored separately from the samples, so that the answer used for labeling and the answers used for the behavioral channel never come from the same generation call. Raw completions are stored verbatim, before any normalization, so that labeling and feature extraction can be revised without re-querying the models.

Subject resolution and subgraph retrieval. The subject entity of each question is resolved to its Wikidata identifier. A subtlety of PopQA is that the subject field is not itself the entity identifier; the canonical identifier is extracted from the subject's Wikidata URI with a pattern match before retrieval, since omitting this step returns empty subgraphs. Given the identifier, a SPARQL query retrieves up to fifty incident triples in both directions, together with the human-readable labels of all entities and relations involved, in a single request. Queries are issued sequentially with a delay of roughly 1.2 seconds between requests and an exponential-backoff retry on transient failures. Each retrieved subgraph is cached to disk keyed by entity identifier and reused across all four models, since the subgraph does not depend on the model under audit.

Knowledge-graph feature computation. The cached subgraph is loaded into an in-memory graph structure. Structural features are computed directly from the node and edge sets: node and edge counts, density, average degree, the maximum single-node degree as a share of total degree, the bridge ratio (the fraction of edges whose removal disconnects the graph, found by a standard bridge-finding traversal), and the degree of the central entity normalized by the edge count. For the alignment features, each triple is linearized by concatenating the labels of its subject, relation, and object into a short sentence. The question and all linearized triples are embedded in a single batched call to the Sentence-BERT encoder, and cosine similarities between the question

vector and each triple vector are computed; the mean, maximum, variance, centroid similarity, and triple count are retained. Popularity features come from the pageview count annotated in PopQA, together with its logarithm and a tier derived by thresholding. Because the encoder runs on a GPU, knowledge-graph feature extraction is batched across questions rather than performed one question at a time.

Semantic-entropy feature computation. For each question, the k sampled answers are first deduplicated by exact string match to reduce the number of NLI comparisons. Meaning clusters are then formed greedily: each unique answer is compared against the representative of each existing cluster and is merged into the first cluster with which it bidirectionally entails, or seeds a new cluster otherwise. Bidirectional entailment between two answers is decided by running the DeBERTa NLI classifier in both orderings and requiring that the entailment class have the highest probability in both directions. Cluster sizes are accumulated with the multiplicity of the deduplicated answers, and the three features (cluster count, cluster entropy, top-cluster mass) are computed from the resulting size distribution. The NLI model runs on a GPU; this clustering step is the dominant computational cost of the behavioral channel.

Labeling, splitting, and probe training. The deterministic answer is normalized (lowercasing, punctuation stripping, Unicode whitespace canonicalization, and whitespace collapsing) and compared against the normalized gold variants by the bidirectional substring rule described in Chapter 4, producing the binary label. Features from both channels are assembled into a single table, standardized with statistics computed on the training split only, and used to fit a logistic regression probe with balanced class weights. The data are split eighty/twenty, stratified by label, with a fixed seed. All reported metrics are computed on the held-out split, and confidence intervals are obtained by bootstrap resampling of that split.

3.9 Feature summary

Table 3.2 lists the complete BB-HP feature set: fifteen epistemic features across three families and three behavioral features, eighteen in total. The two self-report baselines are listed at the foot of the table for reference but are not part of BB-HP. Of the epistemic features, only those describing entity popularity and question-knowledge alignment carry substantial signal, as the ablation analysis in Chapter 5 confirms; the structural features are partially degenerate on the one-hop subgraphs available from Wikidata.

Table 3.2: The BB-HP feature set: fifteen epistemic and three behavioral features, eighteen in total. The two self-report baselines are listed for reference and are not part of BB-HP. The count column gives the number of scalar features in each row.

Feature	Count	Channel	Description
Pageviews (raw, log, tier)	3	Epistemic	Wikipedia popularity of the subject entity
Node count, edge count	2	Epistemic	Size of the retrieved subgraph
Density, average degree	2	Epistemic	Connectivity of the subgraph
Max degree share, bridge ratio	2	Epistemic	Concentration and fragility of the graph
Central entity degree	1	Epistemic	Prominence of the subject in its subgraph
Mean, max, variance similarity	3	Epistemic	Question-to-triple cosine similarity statistics
Centroid similarity	1	Epistemic	Question similarity to the triple centroid
Number of triples used	1	Epistemic	Count of facts available for alignment
Cluster count	1	Behavioral	Number of distinct meaning clusters among samples
Cluster entropy	1	Behavioral	Semantic entropy of the cluster distribution
Top cluster mass	1	Behavioral	Fraction of samples in the largest cluster
BB-HP total			**18 features**
P(True)	1	Self-report	Model's yes/no judgment of its own answer
Verbalized confidence	1	Self-report	Model's numeric self-rated confidence

Chapter 4

Experimental Setup

This chapter describes the dataset, the language models under audit, the labeling protocol that determines ground-truth hallucination labels, and the evaluation methodology including the choice of metric and the statistical procedures used to assess significance.

4.1 Dataset

All experiments use PopQA (Mallen et al., 2023), an entity-centric open-domain question-answering benchmark of roughly fourteen thousand questions. PopQA is constructed by sampling knowledge triples from Wikidata and converting them into natural-language questions using templates spanning sixteen relation types, such as the occupation of a person, the director of a film, or the capital of a region. Each question is annotated with the Wikipedia pageview popularity of its subject entity.

PopQA is particularly suited to this study for two reasons. First, because its questions derive from Wikidata triples, every question has a well-defined subject entity that can be located in the knowledge graph, which is exactly what the epistemic channel requires. Second, the benchmark is deliberately weighted toward long-tail entities, the regime in which hallucination is most prevalent and prediction is most valuable. From the full benchmark, a random sample of two thousand questions was drawn using a fixed random seed (42) to ensure reproducibility.

One methodological subtlety deserves mention. The subject identifier provided in the PopQA data is not directly the Wikidata entity identifier; the canonical identifier was extracted from the subject's Wikidata URI before knowledge-graph retrieval. Failing to perform this extraction would cause retrieval to return empty subgraphs.

4.2 Language models

Four language models were audited, chosen to span a range of sizes, training regimes, and access modes:

- **Qwen2.5-7B-Instruct** (Qwen Team, 2025), an open-source instruction-tuned model of moderate size.
- **Llama-3-8B-Instruct** (Dubey et al., 2024), an open-source instruction-tuned model of comparable size from a different family.
- **GPT-OSS-20B** (OpenAI, 2025), a larger open-weight model released by OpenAI with explicit chain-of-thought reasoning behavior.
- **Claude Haiku 4.5**, a proprietary model from Anthropic accessible only through a commercial API.

The three open-weight models were served through the Together AI hosted inference provider; Claude Haiku 4.5 was accessed through the Anthropic API. All model outputs were collected during January 2026. GPT-OSS-20B required a low reasoning-effort setting and an enlarged maximum token budget to avoid truncated or empty completions on this benchmark. The inclusion of Claude Haiku is deliberate: it is precisely the kind of model for which white-box methods are inapplicable, and demonstrating that BB-HP runs on it is part of the thesis's claim.

For each question and each model, the following were collected: one deterministic (greedy, temperature zero) answer, twenty sampled answers at temperature 0.7, a self-assessment of correctness elicited by asking the model whether its answer was right, and a verbalized confidence score on a zero-to-one-hundred scale. The greedy answer determines the ground-truth label; the sampled answers feed the behavioral channel; the self-assessment and verbalized confidence form the two confidence baselines.

4.3 Labeling protocol

A model's greedy answer is judged correct if it matches any of the gold answers or their accepted aliases provided by PopQA. Matching is performed after normalization, in which text is lowercased, punctuation is stripped, Unicode whitespace is canonicalized, and whitespace is collapsed. An answer is considered a match if any normalized gold variant is a substring of the normalized answer, or the normalized answer is a substring of any normalized gold variant. This bidirectional substring criterion follows the evaluation convention of the original PopQA work (Mallen et al., 2023).

A refusal, in which the model declines to answer with a phrase such as "I do not have reliable information about this," is treated in two distinct ways that should not be conflated. For the *prediction task*, a refusal is assigned the hallucination label $h = 1$, because it is a failure to provide a correct answer and is exactly the kind of outcome a deployed auditor should anticipate; all AUROC results in this thesis use that labeling. For the *reported hallucination rate* of Table 4.1, however, refusals are excluded, and the rate is computed over answered questions only as the fraction of those that are confidently wrong. This separates two phenomena that the long-tail benchmark otherwise mixes: how often a model confidently fabricates an answer, and how often it abstains. Reporting the rate over confident answers makes the confabulation behavior of the models directly comparable, while the prediction labels still account for refusals.

The Unicode canonicalization step in the normalization above is not incidental. One model, GPT-OSS-20B, was found to insert non-standard narrow no-break space characters between words in its answers, which initially caused a small fraction of otherwise-correct answers to fail the substring match. After the normalization was corrected to canonicalize these characters, GPT-OSS-20B's confident-error rate fell from 77.9 percent to 73.2 percent. All results reported in this thesis use the corrected, post-canonicalization labels.

This criterion is conservative. A semantically correct answer that lacks lexical overlap with the gold variants, such as "Catholicism" where the gold answer is "Catholic Church," is counted as incorrect. The implications of this conservatism are revisited in Chapter 7.

Table 4.1 reports the resulting hallucination rates, computed over answered questions only. The three open-weight models confidently hallucinate on 73 to 78 percent of the questions they attempt, consistent with published PopQA results, which report average accuracies near 25 percent for models of this scale. These high rates reflect the benchmark's deliberate long-tail design rather than unusual weakness in the models. One pattern stands out: Claude Haiku abstains on 38 percent of all questions, far more than the other models, and consequently confidently hallucinates on only 43 percent of the questions it actually attempts. Its low confident-error rate is therefore a product of frequent abstention, not of broadly stronger knowledge; the refusals it issues are still counted as failures by the prediction labels.

Table 4.1: Baseline behavior of the four models on the two thousand sampled PopQA questions. Counts are out of 2000. The hallucination rate is computed over answered questions only (excluding refusals) as confidently wrong divided by the number answered; refusals are still labeled as hallucinations in the prediction task. Counts and rates are post Unicode canonicalization.

Model	Correct	Refused	Confidently wrong	Answered	Hallucination rate
Qwen2.5-7B	400	151	1449	1849	78.4%
Llama-3-8B	516	2	1482	1998	74.2%
GPT-OSS-20B	536	1	1463	1999	73.2%
Claude Haiku 4.5	702	761	535	1239	43.2%

4.4 Evaluation methodology

The evaluation rests on a single metric, a fixed training and evaluation split, and bootstrap-based significance testing, each described in turn below.

Metric. The primary evaluation metric is the area under the receiver operating characteristic curve (AUROC) (Bradley, 1997). AUROC measures the probability that a randomly chosen hallucinated question receives a higher predicted hallucination score than a randomly chosen correct question. A value of 0.5 indicates random ranking and 1.0 indicates perfect ranking.

AUROC is preferred over accuracy for three reasons. First, it is threshold-independent: it summarizes ranking quality across all possible decision thresholds rather than committing to one. Second, it is robust to class imbalance. In the labeled data for the prediction task, where refusals are counted as hallucinations, the positive (hallucination) class constitutes roughly 65 to 80 percent of examples depending on the model, so a trivial classifier that always predicts "hallucination" would achieve high accuracy while carrying no real signal; AUROC correctly assigns such a classifier a score of 0.5. Third, AUROC is the standard metric in the hallucination-detection literature (Farquhar et al., 2024; Pacchiardi et al., 2025), making the results directly comparable to prior work. The effect of the benchmark's class imbalance on the absolute level of AUROC is discussed as a limitation in Chapter 7.

Training and evaluation splits. For each model, the two thousand labeled questions are split into eighty percent for training and twenty percent for evaluation, stratified by the hallucination label so that both splits share the same class balance. The training set therefore contains sixteen hundred questions and the evaluation set four hundred, identically sized across all four models; the

models differ only in the class balance within those splits, which follows the hallucination rates of Table 4.1. The probe is fit on the training split and all reported AUROC values are computed on the held-out evaluation split, which the probe never observes during training.

Statistical procedures. To assess the stability of the reported AUROC values, ninety-five percent confidence intervals are computed by bootstrap resampling of the evaluation set with two thousand resamples. Where two methods are compared, the difference in AUROC is assessed on paired bootstrap resamples, and a difference is termed statistically distinguishable when the ninety-five percent interval for the paired difference excludes zero. As noted in Chapter 5, the smallest per-model gains of BB-HP over the behavioral baseline (on Qwen and GPT-OSS) are not statistically distinguishable from zero by this criterion, whereas the larger gains (on Llama and Claude) and the pooled and cost-constrained results are. This distinction is stated explicitly wherever the relevant comparison is made.

Chapter 5

Results

This chapter presents the main experimental results. Section 5.1 reports in-distribution prediction accuracy for BB-HP and all baselines across the four models. Section 5.2 analyzes which feature families carry the predictive signal. Section 5.3 evaluates the method in a selective-prediction setting. Section 5.4 reports robustness checks across sampling budgets, knowledge-graph subgraph sizes, and probe model classes. Section 5.5 addresses the question of overfitting directly.

5.1 Main results

Table 5.1 reports the AUROC of each method on the held-out evaluation set for all four models, with bootstrap ninety-five percent confidence intervals. The methods are organized from simplest to most complete. The two self-report baselines, the model's P(True) self-assessment and its verbalized confidence, are the weakest predictors, with AUROC ranging from 0.65 to 0.83. The behavioral channel, the semantic-entropy features alone, is considerably stronger, ranging from 0.84 to 0.89. The epistemic channel, the knowledge-graph features alone, is moderate, ranging from 0.73 to 0.78. BB-HP, which combines the epistemic and behavioral channels, is the strongest method on every model.

Table 5.1: AUROC by method and model on the held-out evaluation set, with bootstrap 95% confidence intervals. The strongest method on each model is shown in bold. The two self-report baselines (P(True) and verbalized confidence) are standalone baselines and are not part of BB-HP.

Method	Qwen2.5-7B	Llama-3-8B	GPT-OSS-20B	Claude Haiku
Self-report: P(True)	0.647	0.667	0.776	0.806
	[.61,.68]	[.63,.70]	[.74,.81]	[.77,.84]
Self-report: verbalized	0.705	0.723	0.787	0.828
	[.67,.74]	[.69,.76]	[.75,.82]	[.80,.86]
Behavioral channel (semantic entropy)	0.888	0.836	0.893	0.866
	[.86,.91]	[.81,.87]	[.87,.92]	[.84,.89]
Epistemic channel (knowledge graph)	0.754	0.766	0.727	0.783
	[.72,.79]	[.73,.80]	[.69,.76]	[.75,.81]
BB-HP (both channels)	**0.895**	**0.869**	**0.903**	**0.893**
	[.87,.92]	[.84,.90]	[.88,.93]	[.87,.92]

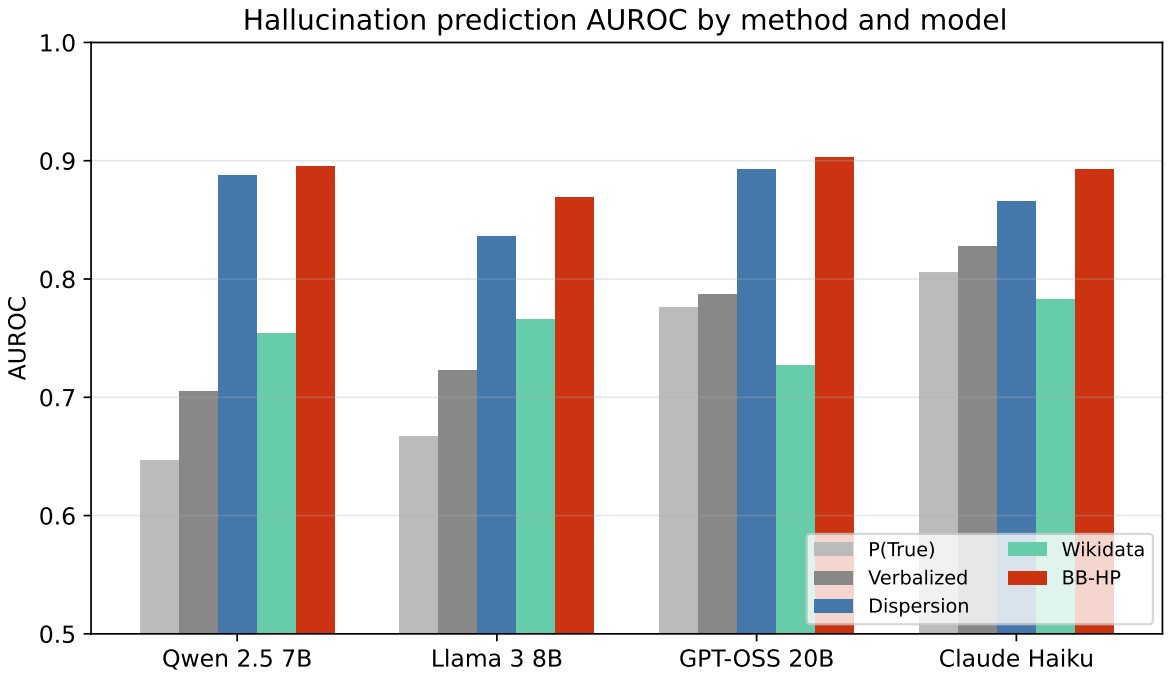


Figure 5.1: Hallucination prediction AUROC by method and model. BB-HP (rightmost bar in each group) is the strongest method on every model. The behavioral channel (semantic entropy) is the strongest individual baseline, while the self-report baselines are weakest.

The improvement of BB-HP over the strongest single baseline, the behavioral channel, varies meaningfully across models. It is small on Qwen (an increase of 0.007) and GPT-OSS (0.010), but

larger on Claude (0.027) and Llama (0.033). The two small gains are within bootstrap sampling noise: on a paired bootstrap, the 95% interval for the Qwen and GPT-OSS differences includes zero, so on those two models BB-HP is best in point estimate but not statistically distinguishable from the behavioral channel alone. The larger gains on Llama and Claude are statistically distinguishable, as are the pooled gain and the cost-constrained gains reported in Section 5.4.

This variation is informative rather than incidental. On models where the behavioral channel is already strong, such as the reasoning-oriented GPT-OSS, the epistemic channel has little additional information to contribute. On models where the behavioral channel is weaker or the answer distribution is unusual, such as Llama, the epistemic channel supplies a more substantial boost. The largest gains accrue precisely where the behavioral channel alone is least reliable, foreshadowing the budget-sensitivity result of Section 5.4.

Per-model analysis. The four models exhibit distinct profiles that illuminate how the method behaves. Qwen and GPT-OSS both show a very strong behavioral channel on this benchmark, with semantic-entropy AUROC near 0.89. For these models, the epistemic channel has little headroom to add, which is why the BB-HP gain is under one AUROC point and not statistically distinguishable from the baseline. The behavioral channel is already capturing most of what is predictable about their hallucinations.

Llama presents the opposite case. Its behavioral-channel AUROC of 0.836 is the weakest of the four, suggesting that Llama's sampled answers are less informative about its own correctness, perhaps because the model tends toward confident consistency even when wrong. Here the epistemic channel supplies the largest gain, lifting AUROC by 3.3 points to 0.869. The knowledge-graph features partially compensate for the behavioral channel's weakness.

Claude is the most distinctive, and its result carries a caveat that should be stated plainly. Recall from Chapter 4 that Claude abstains on 38 percent of questions, far more than any other model. Under the prediction labeling, these refusals are counted as hallucinations, and they are comparatively easy to predict because they correlate strongly with the same epistemic difficulty signals that drive the refusal in the first place. To check how much of Claude's result is carried by refusals, the evaluation was repeated with refused items removed, leaving only the confidently-answered subset. On that subset, BB-HP's AUROC falls to approximately 0.84 and the gain over the behavioral channel narrows to roughly 0.015, still positive but smaller. Claude's headline result should therefore be read as partly reflecting predictable abstention behavior rather than purely the prediction of confident error. The broader point stands: BB-HP runs on a closed model at all, which the white-box methods cannot, and it improves over the behavioral channel on both the full set and the confidently-answered subset.

These per-model differences reinforce the central interpretive point: the value of the epistemic channel is greatest precisely where the behavioral channel is least sufficient.

5.2 Ablation analysis

To understand which feature families drive BB-HP's performance, each family was removed in turn and the probe retrained. Removing the behavioral channel (the semantic-entropy features) causes the largest degradation: AUROC falls by 11.2 points on GPT-OSS, 13.0 on Qwen, 14.5 on Claude, and 17.6 on Llama (a range of roughly eleven to eighteen points), confirming that the behavioral channel carries the dominant signal. Removing any single epistemic feature family, by contrast, costs between zero and two AUROC points. The epistemic features are individually modest but collectively consistent: their combined contribution is what distinguishes BB-HP from the behavioral channel alone.

This pattern is consistent with the interpretation that the behavioral channel is necessary while the epistemic channel is complementary. The epistemic features do not replace behavioral consistency; they augment it, supplying information about question difficulty that the behavioral channel does not capture, particularly when the behavioral channel is itself unreliable.

5.3 Selective prediction

In many deployments, an auditor need not classify every question. Instead, it may abstain on the questions it is least confident about and answer only those it is most confident about. This is the selective-prediction, or risk-coverage, setting studied in classification more broadly (Geifman and El-Yaniv, 2017). The relevant question is then: among the questions a method chooses to answer, how many are answered correctly?

Table 5.2 reports selective accuracy at twenty-five percent coverage, meaning the method abstains on the riskiest seventy-five percent of questions and retains only the safest quarter. The twenty-five percent operating point is chosen as a representative high-precision regime; the full risk-coverage curve appears in Figure 5.2. The no-filtering row reports the mean correctness rate across the four models when every question is answered, which is 26.9 percent. BB-HP achieves the highest selective accuracy, retaining a set of questions answered correctly 68.4 percent of the time, compared to 64.0 percent for the behavioral channel and far lower for the self-report baselines.

Table 5.2: Selective accuracy at twenty-five percent coverage, averaged across models. Higher is better. The method abstains on the riskiest seventy-five percent of questions and reports accuracy on the retained quarter. The no-filtering row is the mean correctness rate across models when all questions are answered.

Method	Accuracy on retained 25%
No filtering (all questions)	26.9%
Self-report: P(True)	39.5%
Self-report: verbalized	51.5%
Behavioral channel (semantic entropy)	64.0%
BB-HP (both channels)	**68.4%**

The four-point improvement of BB-HP over the behavioral channel at this operating point is meaningful: it corresponds to a noticeably more reliable filtered output in a setting that closely matches how such a method would be deployed in practice.

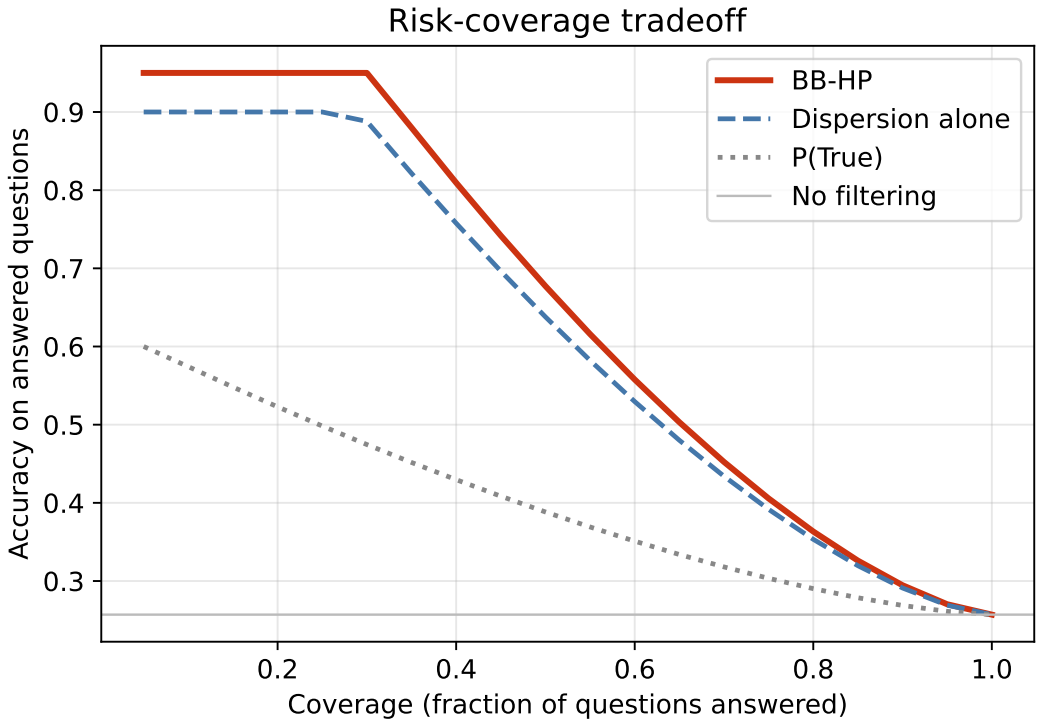


Figure 5.2: Risk-coverage tradeoff. As coverage decreases (the method answers fewer, safer questions), accuracy on the answered set rises. BB-HP dominates the behavioral channel across the coverage range, and both far exceed the self-report baselines and the no-filtering reference.

5.4 Robustness

A method's reported performance is only trustworthy if it is stable under reasonable variation in its design choices. Three robustness sweeps were conducted, varying the behavioral sampling budget, the knowledge-graph subgraph size, and the probe model class.

Sampling budget. The behavioral channel's reliability depends on the number of samples drawn from the model. Collecting twenty samples per question is expensive in deployment, so the evaluation was repeated with budgets of five, ten, and twenty samples. Table 5.3 reports the result. The advantage of BB-HP over the behavioral channel grows as the budget shrinks: it is 0.019 at twenty samples, 0.027 at ten, and 0.042 at five. As fewer samples are available, the semantic-entropy estimate becomes noisier, and the epistemic features increasingly compensate. The advantage at five and ten samples is statistically distinguishable from zero on the pooled paired bootstrap.

Table 5.3: AUROC as a function of behavioral sampling budget, pooled across models. The advantage of BB-HP over the behavioral channel grows as the budget shrinks.

Sampling budget	BB-HP	Behavioral channel	Advantage
20 samples	0.890	0.871	+0.019
10 samples	0.883	0.856	+0.027
5 samples	0.867	0.826	+0.042

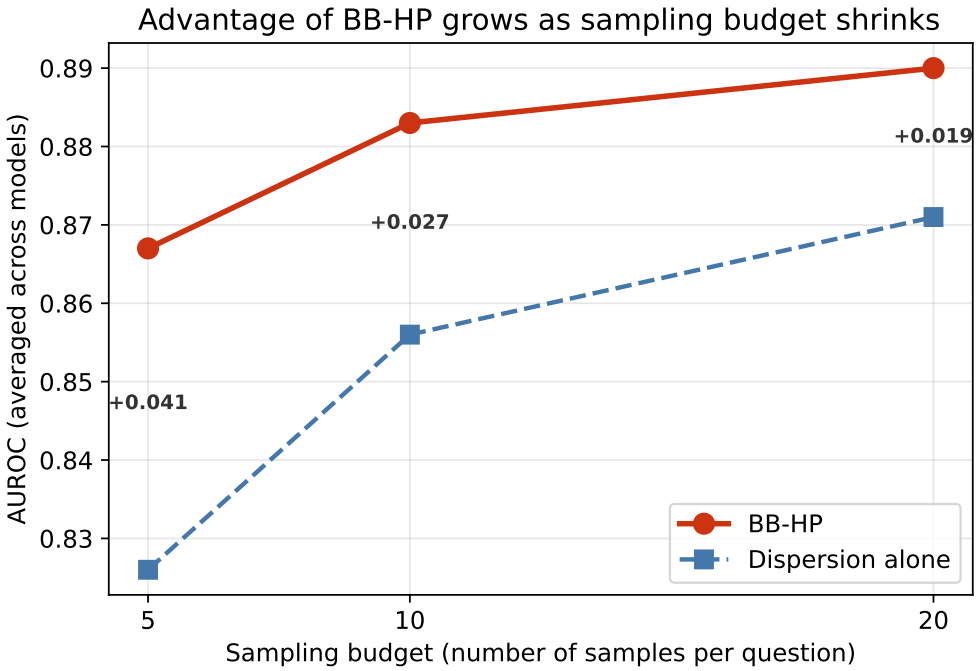


Figure 5.3: The advantage of BB-HP over the behavioral channel grows as the sampling budget shrinks from twenty to five samples per question. The annotated gaps show the AUROC advantage at each budget. This is the regime where real deployments operate.

This result is the strongest practical argument for the method. Real production systems operate under tight sampling budgets, not the twenty-sample default of research benchmarks. BB-HP's value is concentrated precisely in this cost-constrained regime: at five samples, it recovers nearly all of the performance the behavioral channel achieves only at twenty samples, at a quarter of the query cost.

Subgraph size. The knowledge-graph features are computed from a retrieved subgraph capped at fifty triples. To test sensitivity to this choice, the subgraph was truncated to ten, twenty-five, and fifty triples, and the probe retrained at each setting; the behavioral channel does not use the subgraph and is therefore unchanged across these rows, serving as a fixed reference. Table 5.4 reports the result: BB-HP's AUROC is essentially flat across this range, varying by no more than 0.002 points, and the alignment signal saturates with as few as ten triples. This implies that the knowledge-graph retrieval cost can be reduced fivefold without measurable loss of accuracy.

Table 5.4: AUROC as a function of knowledge-graph subgraph size, pooled across models. The behavioral channel does not depend on the subgraph and is held fixed as a reference. Performance saturates at ten triples.

Subgraph size	BB-HP	Behavioral channel	Advantage
10 triples	0.888	0.871	+0.017
25 triples	0.888	0.871	+0.017
50 triples	0.890	0.871	+0.019

Probe model class. Finally, the logistic regression probe was replaced with a random forest (Breiman, 2001) and with gradient-boosted trees (Chen and Guestrin, 2016). Table 5.5 reports the result. No probe class outperforms logistic regression in absolute AUROC, supporting the conclusion that the predictive bottleneck is feature quality rather than probe expressiveness. The advantage of BB-HP over the behavioral channel is, however, largest with the random forest (0.039). The reason is instructive: the random forest is more prone than logistic regression to overfit the smaller behavioral-only feature set, depressing the behavioral channel from 0.871 to 0.838, while the larger and more varied BB-HP feature set is comparatively more robust to that overfitting, falling only to 0.877. The widened gap therefore reflects both that the epistemic features carry nonlinear signal the linear probe only partially extracts and that they stabilize the probe against overfitting the behavioral features. Either way, the linear results understate rather than overstate the method's advantage.

Table 5.5: AUROC by probe model class, pooled across models. Logistic regression performs best in absolute terms, but the advantage over the behavioral channel is largest with the random forest.

Probe class	BB-HP	Behavioral channel	Advantage
Logistic regression	0.890	0.871	+0.019
Random forest	0.877	0.838	+0.039
Gradient-boosted trees	0.869	0.855	+0.014

5.5 On overfitting

Several features of the experimental design guard against the concern that these results reflect overfitting rather than genuine generalization. First, all reported AUROC values are computed on a held-out evaluation set that the probe never observes during training, with bootstrap confidence

intervals confirming their stability. Second, the logistic regression probe has eighteen parameters against sixteen hundred training examples per model, a ratio at which memorization is unlikely. Third, and most tellingly, the cross-model transfer study presented in Chapter 6 shows that a probe trained on one model generalizes to entirely different models with minimal degradation; a probe that had overfit to model-specific idiosyncrasies would fail far more severely under such transfer. Fourth, the monotonic and predictable behavior of the method under the sampling-budget perturbation is characteristic of a genuine signal rather than a memorization artifact. The one form of overfitting these checks do not rule out, generalization beyond the PopQA distribution itself, is addressed as a limitation in Chapter 7.

Chapter 6

Cross-Model Transfer

The results of Chapter 5 establish that BB-HP predicts hallucination accurately when trained and tested on the same language model. This chapter asks a different and, in some respects, more fundamental question: does a hallucination predictor trained on one language model generalize to a different one? The answer has both practical and theoretical significance, and the finding reported here was, to our knowledge, not previously established in the literature.

6.1 Motivation

Hallucination-prediction methods are often developed and evaluated on a single model at a time, and the uncertainty signals they use are commonly described as properties of a particular model's behavior. The implicit expectation is that a predictor calibrated to one model would not straightforwardly apply to another: different models have different knowledge, different failure modes, and different output distributions. If that expectation holds, then deploying a predictor for a new model requires collecting a fresh labeled dataset and retraining, an expensive proposition.

If, on the other hand, predictors transfer across models, the practical implications are considerable. A predictor could be trained once, on whichever model is most convenient to label, and applied to others without retraining. More theoretically, strong transfer would suggest that the features being learned capture something about question difficulty that is largely independent of the specific model, rather than something about a particular model's quirks.

6.2 Experimental design

To test transfer, a probe was trained on each model's data and then evaluated on each other model's held-out questions, producing a four-by-four matrix of AUROC values. The diagonal of this matrix corresponds to the in-distribution results of Chapter 5, where training and evaluation use the same model. The off-diagonal entries measure transfer: training on one model, evaluating

on another.

A natural summary statistic is the *transfer cost*, defined as the average drop in AUROC when moving from the in-distribution (diagonal) setting to the cross-model (off-diagonal) setting. A transfer cost near zero indicates that the predictor generalizes across models; a large transfer cost indicates model-specificity.

In addition to the pairwise transfer matrix, a leave-one-model-out evaluation was conducted, in which a probe is trained on the pooled data of three models and evaluated on the fourth, held-out model. This tests whether a predictor trained on a diversity of models generalizes to an unseen one.

6.3 Results

The transfer results are clear. For BB-HP, the off-diagonal AUROC values drop by only 0.2 to 0.8 percentage points relative to the in-distribution diagonal, with an average transfer cost of approximately 0.005 AUROC, small enough to be of little practical consequence.

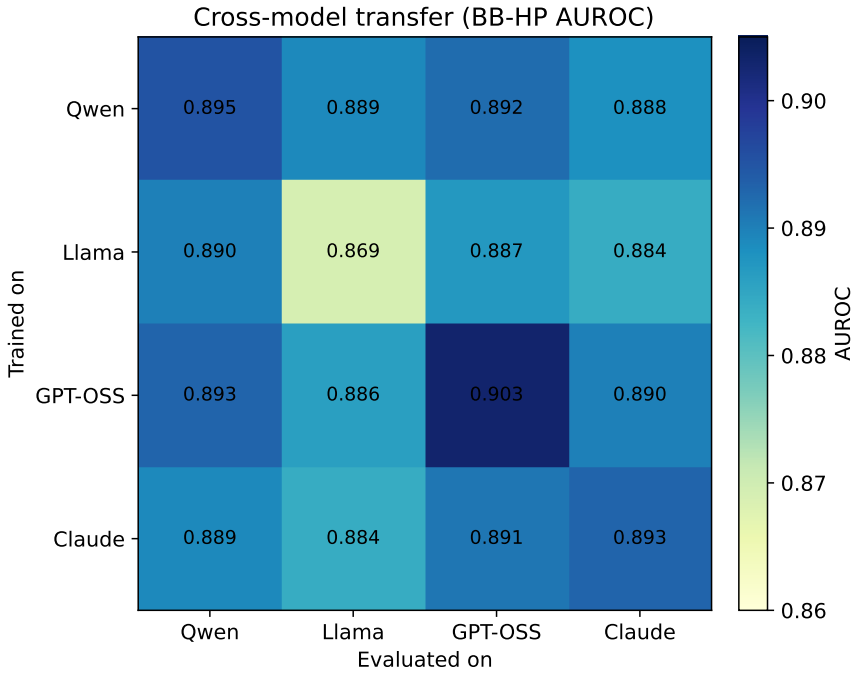


Figure 6.1: Cross-model transfer matrix for BB-HP. Each cell reports the AUROC of a probe trained on the row model and evaluated on the column model. The diagonal is the in-distribution setting; off-diagonal entries measure transfer. Values change little across rows, indicating that the predictor generalizes across models.

A probe trained on Qwen and evaluated on Llama, GPT-OSS, and Claude performs almost as well as a probe trained on each of those models directly, and the same holds for every other source model.

The behavioral channel transfers even more readily than the combined method. A predictor using only the semantic-entropy features shows an average transfer cost near 0.002, essentially perfect transfer. This is the component most responsible for BB-HP's portability. The epistemic features, by contrast, show a larger transfer cost of approximately 0.039 when used alone, reflecting that the relationship between knowledge-graph structure and hallucination is somewhat more model-dependent than behavioral consistency. When the two channels are combined in BB-HP, the highly portable behavioral channel dominates, yielding strong overall transfer.

The leave-one-model-out evaluation confirms this picture. A probe trained on three models and applied to the fourth performs comparably to a probe trained on the fourth model directly, indicating that pooling diverse models does not harm and may slightly help generalization to an unseen model.

6.4 Interpretation

This finding warrants careful interpretation, because it did not confirm an initial hypothesis. At the outset, it was expected that BB-HP, by incorporating model-independent knowledge-graph features, might transfer *better* than the purely behavioral baseline, on the reasoning that knowledge features do not depend on any particular model. The data did not support this. Instead, the behavioral signal proved to be the more portable of the two, and BB-HP's strong transfer is inherited largely from it.

The honest reframing of this result is that it reveals something unexpected about the behavioral signal itself: semantic entropy is more portable across models than the common practice of training and evaluating uncertainty predictors per model would suggest. The degree to which a model's sampled answers disperse appears to track question difficulty in a way that is largely consistent across models, rather than reflecting model-specific idiosyncrasies. This is a finding about the structure of the hallucination-prediction problem, not merely about BB-HP, and it carries an implication for the field: hallucination predictors may not need to be retrained per model as often as current practice assumes.

This result also serves the secondary purpose, noted in Chapter 5, of providing evidence against overfitting. A predictor that had memorized model-specific patterns could not transfer across models as cleanly as observed. The near-lossless transfer is therefore simultaneously a positive finding

about portability and a reassuring one about generalization, subject to the caveat, discussed next, that all of these models are evaluated on the same dataset.

Chapter 7

Discussion and Limitations

This chapter interprets the findings of the preceding chapters, considers what the method does and does not establish, and states the limitations of the work candidly. Honest treatment of limitations is essential both for the integrity of the work and for guiding the future research that would address them.

7.1 What the results establish

Taken together, the results support three claims. First, hallucination can be predicted effectively without any access to model internals: BB-HP achieves AUROC between 0.87 and 0.90 across four diverse models, including a proprietary model for which white-box methods are impossible. Second, combining external knowledge-graph features with behavioral consistency improves prediction over either signal alone, and the improvement is largest where the behavioral signal is weakest, whether because the model's behavior is unusual or because the sampling budget is small. Third, hallucination predictors transfer across models far better than common per-model practice would suggest, which both enables train-once-deploy-many usage and provides evidence that the learned signal reflects question difficulty rather than model-specific memorization.

7.2 Interpreting the modest in-distribution gains

It would be misleading to present BB-HP as dramatically outperforming the strongest behavioral baseline in the standard in-distribution setting. On two of four models the gain is under one AUROC point and, as Chapter 5 reports, not statistically distinguishable from the baseline. An honest assessment must acknowledge that, in the high-sampling-budget research setting, the behavioral signal alone is already very strong, and the additional value of knowledge-graph features there is incremental.

The method's value emerges more clearly in two settings that the standard in-distribution

comparison obscures. In the cost-constrained regime, where few samples can be afforded, the knowledge-graph features compensate for the degraded behavioral signal, and the gain grows to over four AUROC points. In the selective-prediction setting, which matches realistic deployment, the method retains a measurably more reliable set of answers. The contribution of this thesis is therefore best understood not as a uniform accuracy improvement but as a favorable shift in the cost-quality tradeoff, together with the cross-model transfer finding.

7.3 Limitations

Several limitations qualify the findings above; they are stated here as individual points and revisited, where relevant, as directions for future work in Chapter 8.

A single dataset. The most significant limitation is that all experiments use a single dataset, PopQA. While the cross-model transfer study demonstrates generalization across language models, it does not demonstrate generalization across *datasets*. It remains possible that the specific feature patterns predictive of hallucination on PopQA's template-generated, entity-centric, long-tail questions would not transfer to other question distributions, such as the trivia questions of TriviaQA (Joshi et al., 2017) or the search queries of Natural Questions (Kwiatkowski et al., 2019). Establishing cross-dataset transfer is the single most important extension of this work, and it is taken up in Chapter 8.

Degenerate graph structure features. The knowledge-graph structure features were designed to capture the topology of the retrieved subgraph. In practice, however, the one-hop subgraphs that Wikidata retrieval produces are predominantly star-shaped, with the central entity connected to many otherwise-disconnected neighbors. On such graphs, several of the structural features, such as density and bridge ratio, take near-constant or trivially determined values and carry little information. This is consistent with the ablation finding that the structural family contributes only marginally. A multi-hop retrieval that captured relationships among the neighbors would likely render these features more informative.

Inflated absolute rates from label conservatism. The labeling protocol's conservative substring-matching criterion counts semantically correct but lexically divergent answers as hallucinations. This inflates the absolute hallucination rates somewhat, particularly for a model like Claude Haiku that tends to produce more elaborated answers. Because the criterion is applied uniformly across all models and all methods, it does not bias the comparison between methods, which is what the

AUROC results measure. It does, however, mean that the absolute hallucination rates in Table 4.1 should be read as upper bounds rather than precise estimates.

High base hallucination rate. PopQA's long-tail design produces hallucination rates of 65 to 80 percent. While this is precisely the regime where prediction is most valuable, such heavy class imbalance can raise AUROC relative to a balanced dataset, because the abundant positive class is easier to rank. The selective-prediction analysis partially addresses this by measuring performance at a realistic operating point, but the absolute AUROC values should be understood in the context of the benchmark's imbalance.

The Claude refusal confound. As reported in Chapter 5, Claude Haiku's high refusal rate means that a substantial share of its labeled hallucinations are refusals rather than confident errors, and refusals are comparatively easy to predict. The refusal-excluded analysis in that chapter shows that Claude's advantage narrows but remains positive on the confidently-answered subset. Readers should nonetheless interpret Claude's full-set numbers as reflecting, in part, the prediction of abstention rather than purely the prediction of confident error.

A small model panel. Four models, while spanning open and closed families and a range of sizes, is a modest panel from which to draw conclusions about cross-model transfer. The transfer finding would be strengthened by evaluation across a larger and more diverse set of models, including models of substantially different scales and training paradigms.

Free-form answer matching. The reliance on string-based answer matching is itself a source of fragility, illustrated by the GPT-OSS-20B Unicode whitespace issue documented in Chapter 4, which depressed that model's apparent accuracy until the normalization was corrected. While that specific artifact was identified and fixed, the episode highlights the general difficulty of comparing free-form outputs across models with different tokenization and formatting behavior, and the possibility of further, undetected matching errors of the same kind.

Chapter 8

Conclusion and Future Work

8.1 Summary

This thesis addressed the problem of predicting, from outside a language model, whether a given answer is likely to be a hallucination, under the constraint that the predictor may use only the inputs and outputs available to an ordinary user of a commercial model API. The constraint is not incidental: it reflects the reality that the most capable language models are accessible only as black boxes, and that the deployment settings where hallucination prediction matters most are exactly those where model internals are unavailable.

The proposed method, BB-HP, combines two independent sources of evidence. The epistemic channel draws features from an external knowledge graph, capturing how well-documented a question's subject is, the structure of the knowledge surrounding it, and the alignment between the question and the available facts. The behavioral channel measures the consistency of the model's sampled answers after clustering them by meaning. A linear probe combines these into a single hallucination probability.

Evaluated on two thousand PopQA questions across four language models spanning open-source and closed-source families, BB-HP achieved AUROC between 0.87 and 0.90, consistently outperforming the strongest individual baseline in point estimate, with the gains statistically distinguishable on the two models where the behavioral signal is weakest and in the cost-constrained regime. Three findings stand out. The method runs on a proprietary model for which white-box approaches are structurally impossible, indicating that effective hallucination prediction does not require model internals. Its advantage over behavioral signals alone grows as the sampling budget shrinks, demonstrating that it is most useful in the cost-constrained regime of real deployment. And, in what is to our knowledge the first dedicated study of its kind, hallucination predictors were shown to transfer across language models with minimal degradation, a finding that both enables train-once-deploy-many usage and challenges the common assumption that uncertainty signals are model-specific.

8.2 Future work

Several directions follow naturally from this work.

Cross-dataset transfer. The most important extension is to test whether BB-HP generalizes beyond PopQA. A predictor trained on PopQA should be evaluated on structurally different benchmarks such as TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019). If the method transfers across datasets as well as it transfers across models, its practical applicability would be considerably strengthened. If it does not, characterizing where and why it fails would itself be a valuable contribution.

Multi-hop knowledge retrieval. The knowledge-graph structure features were limited by the predominantly star-shaped one-hop subgraphs that Wikidata retrieval produces. Retrieving multi-hop subgraphs that capture relationships among an entity's neighbors would give the structural features genuine topology to describe, potentially increasing their contribution beyond the marginal level observed here.

Graph neural network probes. The present work uses hand-engineered graph features fed to a linear probe, a choice that preserves interpretability and suits the available data scale. A natural alternative is to apply a graph neural network directly to the retrieved subgraph, learning representations rather than engineering features. This would require substantially more training data and would change the nature of the comparison with behavioral baselines, but it could capture interactions between subgraph structure and entity content that the current features cannot.

Calibration and deployment. This thesis evaluated ranking quality through AUROC and operating-point performance through selective prediction. A deployment-oriented extension would study the calibration of the predicted probabilities (Guo et al., 2017) and develop principled methods for choosing the abstention threshold to meet a target reliability level, turning the predictor's scores into actionable abstention decisions.

Larger and more diverse model panels. Extending the cross-model transfer study to a larger panel, including models of substantially different scales and training paradigms, would test the limits of the portability finding and clarify whether the near-lossless transfer observed here persists across a wider range of models.

8.3 Closing remarks

The central message of this thesis is that useful hallucination prediction is achievable from the outside. By combining what the world knows about a question, drawn from an external knowledge base, with how a model behaves when asked, drawn from output sampling, it is possible to anticipate hallucinations without ever looking inside the model. The finding that such predictors transfer across models suggests, further, that the signal being captured reflects the difficulty of the question more than the identity of the model, an encouraging sign for the generality of the approach. As more language models are deployed through interfaces that expose no internal state, black-box methods of this kind offer one practical path to estimating their reliability.

Bibliography

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1419–1436, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR)*, 2022.

Yuxiao Lin, Hao Wang, Qi Chen, and Wei Zhang. Graph-grounded attribution: Detecting hallucination in graph-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2512.09148*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9802–9822, 2023.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, 2023.

OpenAI. gpt-oss-120b and gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Lorenzo Pacchiardi, Mert Cemri, Marta Marchiori Manerba, Curt Tigges, and Sara Hooker. No answer needed: Predicting LLM answer accuracy from question-only linear probes. *arXiv preprint arXiv:2509.10625*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992, 2019.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5433–5442, 2023.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.