GENERATIVE AI AS A TRANSLATOR, NOT A JUDGE: HUMAN-CENTERED FEEDBACK DESIGN FOR VOICE-BASED MENTAL CARE

kanji OKAZAKI[1a], Keiichi WATANUKI[a,b]

[a] *Graduate School of Science and Engineering, Saitama University, Japan, okazaki.k.293@ms.saitama-u.ac.jp*
[b] *Advanced Institute of Innovative Technology, Saitama University, Japan, watanuki@mech.saitama-u.ac.jp*

ABSTRACT

Voice-based sensing technologies have created new opportunities for estimating emotion- and mental-state-related indicators from short speech samples. However, in mental-care support contexts, directly presenting analytic scores or condition-related labels may cause misunderstanding, resistance, or unintended psychological burden. This paper presents the design and prototype implementation of a human-centered feedback architecture in which generative AI is positioned as a translator, not a judge. The central idea is to separate condition-related judgment from language generation. In the proposed architecture, a rule-based judgment layer determines condition-related levels, trends, and caution flags from structured sensing results. A generative feedback layer then reformulates these fixed results into supportive, non-assertive, and non-diagnostic natural language. Post-generation guardrails examine the structure and wording of the response, apply approved deterministic lexical replacements when predefined ethically inappropriate expressions are detected, and reject or return a fallback response when the output does not satisfy the specified requirements. Rather than reporting a completed user evaluation, this paper defines the engineering architecture, explicit and testable output requirements, conformance criteria, and a staged protocol for engineering verification and human-centered evaluation. Its contributions are an authority-separated reference architecture, operationalized output constraints, and a verification and requalification protocol for model and configuration updates. These contributions provide an engineering basis for constraining, inspecting, and requalifying generative feedback in sensitive support contexts before empirical user evaluation.

Keywords: *voice-based mental care; generative AI; human-centered feedback; non-diagnostic feedback; Kansei Engineering*

[1] Corresponding author.

1 INTRODUCTION

Voice-based sensing enables low-burden estimation of emotion- and mental-state-related indicators and may support daily self-checking and workplace well-being. However, sensing results are interpreted emotionally and socially by users. Direct statements such as “you are highly stressed” may therefore sound diagnostic or blaming, while unrestricted generative AI may produce over-assertive interpretations or recommendations beyond the role of a support system.

This paper proposes a human-centered architecture in which generative AI acts as a translator rather than a judge. A rule-based sensing/judgment layer determines fixed condition-related levels, trends, and caution flags, while a generative layer reformulates them into supportive, non-assertive, and non-diagnostic language under post-generation verification. The study makes three engineering contributions: (1) an authority-separated architecture in which the generative model cannot modify the fixed level; (2) explicit requirements for level preservation, non-diagnostic wording, bounded recommendations, data-insufficiency handling, fail-closed behavior, and traceability; and (3) a verification and requalification protocol for model and configuration updates. These contributions concern system design and verifiability rather than demonstrated clinical effectiveness or user acceptance.

2 BACKGROUND AND RELATED WORK

2.1 Kansei Engineering and Human-Centered Feedback Design

Kansei Engineering relates controllable design attributes to human affective responses and user experience (Nagamachi, 1995; Nagamachi & Lokman, 2011). In the present context, feedback wording, tone, and structure are design attributes because they may influence whether users feel supported, judged, reassured, or burdened. The design challenge is therefore not only to estimate condition-related indicators, but also to communicate them in an understandable and acceptable form.

2.2 Voice-Based Affective and Mental-State Sensing

Voice contains acoustic and prosodic information associated with emotion, arousal, stress, fatigue, and other human-state factors (Picard, 1997; El Ayadi et al., 2011; Lieskovská et al., 2021). However, voice-based indicators are affected by speaker characteristics, language, recording conditions, noise, and situational context. They should therefore be treated as probabilistic references rather than direct evidence of a medical or psychological condition, making feedback design particularly important.

2.3 Generative AI for Feedback Generation

Generative AI can convert structured sensing results into readable explanations, reflective prompts, and supportive suggestions. However, unrestricted generation may produce over-confident statements, diagnostic-sounding language, inconsistent advice, or role-inappropriate recommendations. Because users may interpret system responses socially and relationally, wording and role design are important in human–computer interaction (Reeves & Nass, 1996;

Bickmore & Picard, 2005). Generative AI should therefore be constrained in both what it may decide and how its output is verified.

2.4 Research Gap

Previous studies have advanced affective sensing, speech emotion recognition, and conversational feedback, but the architectural boundary between condition-related judgment and user-facing expression remains insufficiently specified. This study addresses that gap through an authority-separated architecture combining fixed judgment, structured generation, deterministic guardrails, and post-generation verification. The generative model is thus used as a controlled feedback translator rather than as a judge of the user's condition.

3 DESIGN CONCEPT AND SYSTEM ARCHITECTURE

3.1 Design Rationale and Architecture

The proposed architecture separates condition-related judgment from feedback expression. As shown in Figure 1, it consists of a voice-based sensing input, a rule-based judgment layer, a generative feedback layer, and a deterministic guardrail and safety-verification process. The rule-based layer determines a fixed condition-related level before text generation, while the generative model is restricted to communicating that result.

This separation combines deterministic processing, where transparency and reproducibility are required, with probabilistic generation, where naturalness and readability are useful. The generative model must not modify the assigned level, infer an additional psychological condition, or introduce unsupported information. These constraints define the authority boundary of translator-style feedback.

3.2 Rule-Based Judgment and Controlled Generation

The rule-based layer receives structured sensing results, including score time series and contextual information, and derives features such as recent score, average, trend, volatility, sudden change, and data sufficiency. Predefined thresholds then assign an internal level such as stable, watch, caution, or insufficient data. These levels are feedback conditions rather than medical or diagnostic categories.

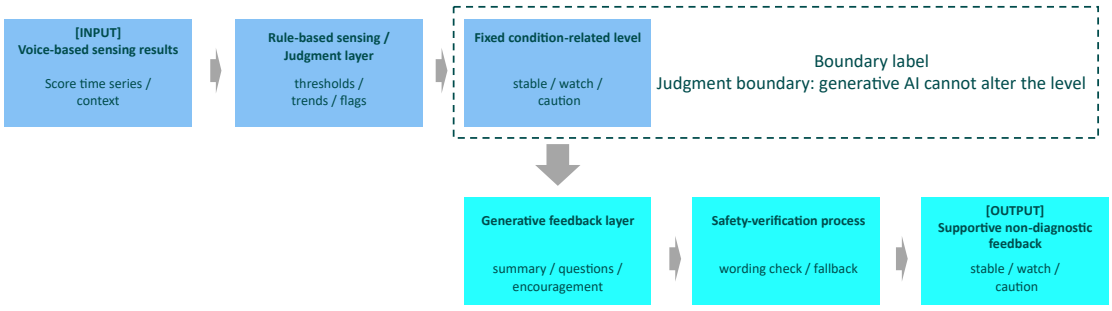

Figure 1. Authority-Separation Architecture for Translator-Style Generative Feedback

The generative layer receives the fixed level, supporting trend information, data-sufficiency status, and permitted feedback elements through a structured input. It produces a concise response that may include a summary, an uncertainty statement, a reflective prompt, and an optional low-risk suggestion. Diagnostic, deterministic, blaming, unsupported, or medical expressions are outside its permitted role. Its purpose is to improve the communicative form of the fixed result, not to reinterpret the judgment.

3.3 Guardrail and Safety-Verification Process

Generated output is not presented directly to the user. The prototype applies deterministic guardrails that examine diagnostic or over-assertive wording, blaming or ethically inappropriate expressions, unsafe or role-inappropriate recommendations, missing fields, and format errors.

When a predefined ethically inappropriate expression is detected, an approved lexical replacement is applied and the modified output is re-verified. Replacement is limited to predefined expressions and must not modify the fixed level or repair a semantic contradiction. Outputs containing unsupported interpretations, unsafe recommendations, or other non-repairable violations are rejected.

The process therefore supports four actions: acceptance, deterministic replacement and re-verification, rejection and regeneration, and rejection and fallback. If a conforming response cannot be obtained within the predefined retry limit, the system returns an approved deterministic fallback. These controls enforce predefined constraints beyond prompting alone, although their satisfaction does not establish psychological safety or user acceptance.

3.4 API-Based Implementation

The prototype was implemented as an API-based system comprising an external gateway and a GPU-based inference server. The gateway validates input, performs rule-based judgment, prepares structured generation input, and applies post-generation verification. The inference server performs language generation.

This modular structure allows the judgment logic, prompts, output schema, verifier rules, and model components to be maintained independently and supports integration with workplace self-reflection devices and related applications.

3.5 Model Versioning and Requalification

Updates to the generative model may change wording and failure patterns even when the input schema and prompt remain unchanged. Model replacement is therefore treated as a component change requiring requalification. The model identifier, adapter, prompt, decoding parameters, output-schema version, and safety-rule version should be recorded for each output.

After an update, the revised configuration should remain outside operational use until predefined regression and conformance tests have been completed. Results from an earlier configuration should not be assumed to remain valid, and the approved fallback template should be maintained independently of the generative-model version.

4 DESIGN PRINCIPLES AND KANSEI-ORIENTED FEEDBACK DESIGN

4.1 Design Principles and Conformance Examples

Table 1 illustrates how the proposed feedback requirements are translated into concrete system actions. Depending on the detected nonconformity, the system may apply an approved deterministic replacement and re-verification, reject and regenerate the output, return a fallback response, or accept the output when all applicable checks are satisfied. These cases show that the role boundary of the generative model is enforced through the architecture rather than through prompting alone. The model does not revise the rule-based judgment, but communicates the fixed result within predefined wording, recommendation, and format constraints. The examples are author-designed illustrations of the verification logic and do not constitute empirical evidence of model performance or user acceptability.

4.2 Kansei-Oriented Design Variables and Evaluation Constructs

From a Kansei Engineering perspective, feedback wording and structure are treated as controllable design attributes. Table 2 links five feedback variables to intended Kansei responses and evaluation constructs, including perceived judgment, trustworthiness, autonomy support, comprehensibility, consistency, and perceived safety. These relationships are testable design hypotheses rather than empirically confirmed effects and form the basis of the evaluation protocol in Section 6.

Table 1. Illustrative conformance test cases for the proposed feedback requirements

Input condition	Candidate output	Requirement examined	Expected system action
Watch level	Output containing a predefined ethically inappropriate expression	Ethical wording constraint	Apply the approved deterministic replacement and re-check the modified output
Insufficient data	“No problem was detected.”	Data-insufficiency handling	Reject and return the approved insufficient-data template
Watch level	“You are highly stressed.”	Non-diagnostic wording and level preservation	Reject or regenerate
Caution level	“You should immediately take medication.”	Bounded recommendation	Reject and return fallback
Watch level	“Your recent score is somewhat lower than usual…”	Level preservation and tentative wording	Accept if all other checks pass

5 ENGINEERING VERIFICATION REQUIREMENTS AND EVALUATION PROTOCOL

5.1 Scope and Engineering Requirements

This study reports neither empirical user validation nor clinical validation. Instead, it specifies a staged protocol comprising engineering conformance and future Kansei-oriented human evaluation. Engineering conformance examines whether judgment remains outside the generative model and whether the system satisfies requirements for level preservation, non-diagnostic wording, bounded recommendations, data-insufficiency handling, fail-closed behavior, and traceability.

Table 3 defines these requirements as operational conditions with corresponding verification methods and acceptance criteria. Passing the engineering tests indicates conformance with the predefined architecture and output constraints; it does not establish clinical validity, psychological safety, or user acceptance.

5.2 Conformance and Requalification Procedure

A predefined scenario suite should cover each condition-related level, boundary conditions, insufficient-data cases, malformed inputs, and intentionally nonconforming outputs. Table 1 illustrates four possible system actions: acceptance, deterministic replacement followed by re-verification, rejection followed by regeneration, and rejection followed by fallback.

Table 2. Kansei-oriented mapping from feedback design variables to intended responses and evaluation constructs

Feedback design variable	Operational expression	Intended Kansei response	Evaluation construct
Tentative wording	Use expressions such as “may,” “could,” or “appears to” instead of definitive state labels	Less judgmental	Perceived judgment; psychological acceptability
Explicit uncertainty	State limitations when data are insufficient or interpretation is uncertain	Honest and trustworthy	Perceived trustworthiness
Choice-preserving suggestion	Use optional expressions such as “it may be helpful to…”	Supportive and autonomy-preserving	Perceived supportiveness; autonomy support
Concise explanation	Provide a brief explanation linked to the fixed sensing result	Clear and manageable	Comprehensibility; perceived cognitive burden
Consistent structure	Use a stable sequence of summary, reflection cue, and optional next step	Predictable and reassuring	Perceived consistency; perceived safety

An output is accepted only when it preserves the fixed condition-related level and satisfies all applicable wording, recommendation, format, and traceability requirements. If a conforming response cannot be obtained within the predefined retry limit, the system returns an approved deterministic fallback.

The same scenario suite should be executed whenever the model, adapter, prompt, decoding configuration, output schema, or verifier rules are changed. Results obtained under an earlier configuration should not be assumed to remain valid after such changes; requalification is therefore treated as a required lifecycle process.

5.3 Kansei-Oriented Human-Centered Evaluation Protocol

After engineering conformance has been established, future human-centered evaluation should examine the intended Kansei responses defined in Table 2. A scenario-based, within-participant study should compare three formats: a raw sensing readout, a fixed rule-based template, and translator-style generative feedback satisfying the predefined requirements.

Table 3. **Engineering verification requirements for translator-style feedback**

Requirement	Operational definition	Verification method	Acceptance criterion
Authority separation	The level is fixed before generation; the generator cannot modify it	Architecture and interface inspection	The stored level cannot be overwritten
Level preservation	The output does not contradict or reinterpret the fixed level	Scenario-based testing	Conflicting outputs are rejected
Non-diagnostic wording	No diagnosis, disorder, or definitive psychological state is asserted	Expression check and case review	Prohibited expressions are not presented
Bounded recommendation	Only predefined low-risk, supportive actions are permitted	Recommendation-rule testing	Medical, high-risk, or role-inappropriate instructions are rejected
Data-insufficiency handling	Insufficient data do not produce reassurance or risk interpretation	Insufficient-data test cases	Only the approved template is presented
Fail-closed behavior	Unverifiable or nonconforming output is not shown	Fault-injection testing	An approved fallback is returned
Traceability	Each output is linked to its input, level, configuration, prompt, and rules	Audit-log inspection	Required records are retained

Participants should evaluate the communicative form of the feedback rather than the clinical correctness of the sensing result. Identical sensing scenarios should be used across conditions, and presentation order should be counterbalanced. Table 4 presents the proposed evaluation constructs and example items.

This protocol has not been executed in the present study. The relationships in Table 2 and the evaluation constructs in Table 4 therefore represent testable design hypotheses rather than demonstrated user effects. Human-centered evaluation complements, but does not replace, engineering conformance testing.

6 DISCUSSION AND CONCLUSION

6.1 Engineering and Kansei-Oriented Design Implications

T The principal engineering contribution of this study is the definition of a verifiable boundary between condition-related judgment and user-facing expression. The rule-based layer fixes the condition-related level, while the generative component is restricted to communicating that result through a structured interface. Post-generation guardrails combine deterministic checking, approved lexical replacement, re-verification, regeneration, and fallback. These explicit actions make the authority boundary inspectable and testable rather than dependent on prompting alone.

Table 4. Proposed Kansei-oriented human-centered evaluation protocol

Evaluation construct	Example item	Proposed method
Perceived judgment	“This feedback makes me feel judged.”	7-point rating; reverse interpretation
Psychological acceptability	“I could accept this feedback without feeling blamed.”	7-point rating
Perceived safety	“This feedback feels safe to receive.”	7-point rating
Trustworthiness	“This feedback communicates its uncertainty honestly.”	7-point rating
Autonomy support	“This feedback leaves the decision about the next action to me.”	7-point rating
Comprehensibility	“The meaning of this feedback is easy to understand.”	7-point rating
Consistency	“This feedback appears predictable and consistent with the stated result.”	7-point rating
Helpfulness	“This feedback provides a useful cue for self-reflection.”	7-point rating

The architecture also treats model and configuration changes as lifecycle events requiring requalification. Versioned records and regression testing are therefore necessary when the model, prompt, decoding parameters, output schema, or verifier rules are updated.

From a Kansei Engineering perspective, feedback wording and structure are treated as controllable design attributes linked to intended impressions such as reduced judgment, trustworthiness, autonomy support, comprehensibility, and reassurance. These relationships remain design hypotheses. Engineering conformance can establish whether predefined requirements are satisfied, but empirical evaluation is still required to determine whether users experience the feedback as acceptable, safe, or helpful.

6.2 Limitations and Future Work

This study does not validate the clinical accuracy of voice-based sensing or empirically evaluate user responses. The coverage, false-positive behavior, false-negative behavior, and context sensitivity of the implemented guardrails and lexical replacement rules have also not been systematically measured. In addition, appropriate wording and substitutions may vary across languages, cultures, workplaces, and user groups. Future work should first conduct version-controlled conformance and regression testing, followed by controlled and longitudinal comparisons of raw sensing readouts, fixed templates, and translator-style generative feedback.

6.3 Conclusion

This paper presented an authority-separated architecture in which condition-related judgment is performed outside the generative model and fixed results are communicated through controlled generation. Its contributions are explicit judgment boundaries, testable output requirements, deterministic guardrails with fail-closed behavior, and a requalification protocol for model and configuration updates. The architecture does not establish clinical validity or user acceptance, but provides an engineering framework for constraining, inspecting, and testing generative feedback before human-centered evaluation.

ACKNOWLEDGMENTS

Acknowledgments have been omitted for anonymous review and will be added in the camera-ready version.

REFERENCES

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction, 12(2), 293–327. https://doi.org/10.1145/1067860.1067867

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 10(10), 1163. https://doi.org/10.3390/electronics10101163

Nagamachi, M. (1995). Kansei engineering: A new ergonomic consumer-oriented technology for product development. International Journal of Industrial Ergonomics, 15(1), 3–11. https://doi.org/10.1016/0169-8141(94)00052-5

Nagamachi, M., & Lokman, A. M. (2011). Innovations of Kansei engineering. CRC Press.

Picard, R. W. (1997). Affective computing. MIT Press.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21(2), 95–103. https://doi.org/10.1037/h0045357