Transparent Reference-free Automated Evaluation of Open-Ended User Survey Responses

Subin An[1]* **Yugyeong Ji**[2]* **Junyoung Kim**[2]* **Heejin Kook**[2]* **Yang Lu**[3]† **Josh Seltzer**[3]†
[1]Kookmin University [2]Sungkyunkwan University [3]Nexxt Intelligence
nesquiq@kookmin.ac.kr
{ygang29, junyoung44, hjkook}@skku.edu
{yang, josh}@nexxt.in

Abstract

Open-ended survey responses provide valuable insights in marketing research, but low-quality responses not only burden researchers with manual filtering but also risk leading to misleading conclusions, underscoring the need for effective evaluation. Existing automatic evaluation methods target LLM-generated text and inadequately assess human-written responses with their distinct characteristics. To address such characteristics, we propose a two-stage evaluation framework specifically designed for human survey responses. First, gibberish filtering removes nonsensical responses. Then, three dimensions—*effort*, *relevance*, and *completeness*—are evaluated using LLM capabilities, grounded in empirical analysis of real-world survey data. Validation on English and Korean datasets shows that our framework not only outperforms existing metrics but also demonstrates high practical applicability for real-world applications such as response quality prediction and response rejection, showing strong correlations with expert assessment.

1 Introduction

User surveys play a crucial role in marketing research, informing strategy and product development decisions (Crick, 2024; Netzer et al., 2012). Among these, open-ended responses are particularly valuable, as they offer rich insights through first-hand experiences and nuanced perspectives (Rouder et al., 2021). However, the quality of such responses can vary widely, and low-quality responses risk introducing noise or misleading interpretations. Moreover, identifying low-quality responses presents an opportunity to re-engage participants and elicit more meaningful feedback.

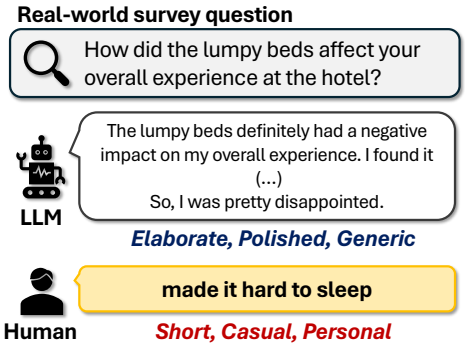

Figure 1: Example highlighting how human-written responses differ from LLM-generated ones.

Evaluating the quality of open-ended responses is therefore essential, yet it presents two major challenges. First, manual evaluation of all responses by human annotators is prohibitively costly, necessitating automated evaluation methods. Second, the absence of ground-truth reference for comparison renders traditional reference-based metrics (*e.g.* BLEU (Papineni et al., 2022), ROUGE (Lin, 2004)) inapplicable.

To address these issues, recent efforts have explored reference-free evaluation using large language models (LLMs) as judges (Liu et al., 2023; Wu et al., 2025; Badshah et al., 2025; Liu et al., 2024). These approaches leverage LLMs' general language understanding to assess response quality without relying on predefined answers. Existing evaluation methods are primarily designed to assess responses generated by LLMs (Zheng et al., 2023; Chiang and Lee, 2023). However, as illustrated in Figure 1, human-written responses differ substantially in nature. They tend to be shorter, more careless, and often reflect personal or subjective perspectives. These differences highlight the need for evaluation criteria that account for the unique characteristics of human responses, rather than applying standards optimized for LLM-generated text.

In this paper, we propose a novel evaluation

*Equal contribution. The authors are listed in alphabetical order.
†Corresponding authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 963–982
November 4-9, 2025 ©2025 Association for Computational Linguistics

framework to address an industrial problem that has been overlooked by existing LLM evaluation studies: the evaluation of short, informal, and often noisy real human-written survey responses. Our framework is specifically designed for these unique characteristics, grounded in analysis of real-world user data. It incorporates a gibberish filtering module and introduces three novel evaluation dimensions—*effort*, *relevance*, and *completeness*—to effectively assess the quality of human-written open-ended responses. To demonstrate the practical utility and real-world applicability of our framework, we show how its scores can be leveraged for two key downstream applications: response quality prediction for analytical purposes and response rejection for operational quality control. These applications enable systems to enable researchers to gain more meaningful insights by assessing user responses, while also offering a mechanism to filter low-quality input and provide feedback to participants. Furthermore, to evaluate its robustness across languages and cultural contexts, we validate the framework on both English and Korean datasets.

Our key contributions are as follows:

- We conduct a comprehensive analysis of real user responses, highlighting the need for evaluation criteria beyond conventional LLM-as-a-judge.
- We propose an interpretable, multi-dimensional evaluation framework tailored for open-ended responses, capturing the aspects of *effort*, *relevance*, and *completeness*.
- We empirically validate our approach, showing strong performance in fine-grained dimension scoring and downstream applications like overall quality prediction and response rejection across English and Korean datasets.

2 Related Work

2.1 Reference-Free Evaluation Approaches

Traditional evaluation methods have primarily relied on reference-based similarity measures such as BLEU (Papineni et al., 2022) and ROUGE (Lin, 2004). However, these approaches are unsuitable for open-ended, user-generated texts such as survey responses, where ground-truth reference often unavailable or are impractical to collect. Moreover, they remain highly sensitive to the quality and completeness of reference data (Freitag et al., 2020). This limitation has motivated reference-free evaluation methods that assess response quality without requiring references.

Among these, embedding-based approaches such as BERTScore (Zhang et al., 2020) successfully evaluate semantic similarity but often fail to capture fine-grained aspects such as informativeness and logical coherence. More recent frameworks introduce multi-dimensional evaluation by decomposing quality into distinct criteria, such as relevance and coherence (Hu et al., 2024; Liu et al., 2024). These approaches enable more interpretable and robust assessments, especially for open-ended tasks where constructing references is costly or infeasible (Ito et al., 2025).

2.2 LLM-Based Automatic Evaluation

LLM-based evaluation methods have recently emerged as a compelling alternative for assessing natural language quality (Gu et al., 2024). This shift is driven by the cost and limited scalability of human evaluation, as well as the inability of metrics such as BLEU (Papineni et al., 2022), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) to capture subtle nuance. In contrast, LLMs provide rich language understanding and reasoning, enabling multi-dimensional assessment even without references. For example, G-Eval (Liu et al., 2023) uses LLMs like GPT-4 to produce quantitative, interpretable scores from natural language prompts under predefined criteria. This approach overcomes the limitations of similarity-based metrics that often fail to reflect semantic coherence and logical consistency.

3 User Response Evaluation

We design a structured framework to evaluate the quality of user-generated open-ended responses. Figure 2 illustrates the overview of the proposed framework. The evaluation consists of two main stages: (1) gibberish filtering, which removes nonsensical or malformed inputs, and (2) dimension-wise scoring based on three core criteria—*effort*, *relevance*, and *completeness*. Each dimension is evaluated independently using a prompt-based inference with LLMs, with scoring rubrics and qualitative reasoning to ensure interpretability and consistency. In the following, we detail the rationale, presented in the Preliminary Study, and the scoring method, detailed in the Method section, for each dimension.

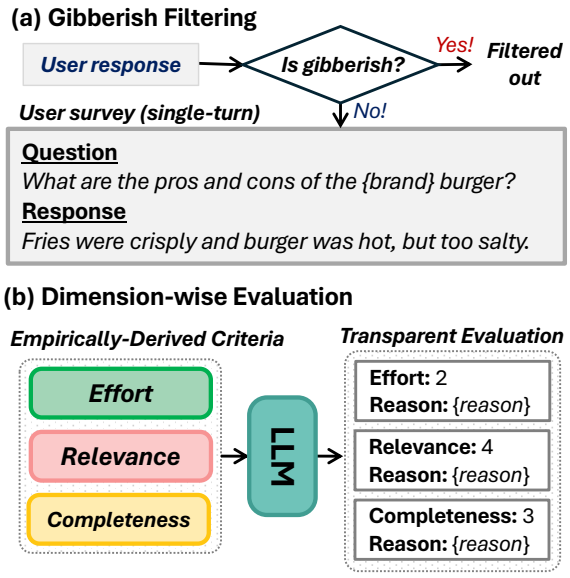

Figure 2: Overall pipeline of the user response evaluation, showing gibberish filtering and dimension-wise evaluation steps.

3.1 Gibberish Filtering

Gibberish, which refers to meaningless sequences or nonsensical language (*e.g.*, asdf, ddddd), is a phenomenon that rarely appears in LLM-generated responses but frequently occurs in real-world human inputs. Detecting such gibberish in user responses is crucial for reducing the computational cost of LLM-based evaluation by preemptively filtering out samples that lack informative content. By explicitly addressing noisy, real-world inputs which are underrepresented in curated LLM outputs, our framework not only complements but, in key settings, also surpasses language-agnostic evaluators in applicability. Unlike generic filtering methods, our approach is tailored to the structural and statistical properties of individual languages. We propose language-specific detection pipelines for English and Korean that incorporate both statistical modeling and linguistic heuristics,thereby capturing the diverse manifestations of gibberish across languages. For details on the detection architecture, refer to Appendix A.

3.2 Dimension-wise Evaluation

Existing LLM-based evaluators, designed for model outputs, often fail to capture the unique characteristics of human responses. To address this challenge, we conducted a data-driven criteria establishment process. By systematically analyzing large-scale real-world English and Korean survey data, we empirically identified *effort*, *relevance*, and *completeness* as the core dimensions.

Based on these findings, we develop new evaluation criteria tailored to human-generated responses. Our method uses a simple prompting approach grounded in these empirical insights, enabling effective evaluation of real-world survey responses in a manner aligned with how marketing researchers assess quality. The LLM assigns a score and provides a brief justification based on these criteria. Full prompts are provided in Appendix E.

3.2.1 Effort

Preliminary Study In previous LLM-based response evaluations, the *effort* dimension was often considered unnecessary, as LLMs do not exhibit human-like "effort." However, since real survey respondents are human, assessing how sincerely they answer questions is crucial (Krosnick, 1991).

One challenge is that human responses are often brief, which can easily be mistaken for a lack of effort (Santilli et al., 2025). Yet short responses can still show high effort if they provide enough information and specificity. Conversely, length alone does not guarantee thoughtfulness.

Evaluating effort in human responses therefore requires criteria that go beyond length or wording, taking into account informational density and specificity. This study presents an evaluation approach designed with these factors in mind, making it more suitable for real user responses and aligned with how marketing researchers assess response quality.

Method The *effort* dimension evaluates the level of cognitive effort demonstrated in the user response, with a focus on two aspects: *informativeness* and *specificity*. The LLM assigns a discrete score on a 0–7 scale, where higher scores reflect more detailed, thoughtful, and specific responses. A rubric was designed to define clear criteria for each score range. The model is prompted with this rubric and returns both the score and a natural language justification.

3.2.2 Relevance

Preliminary Study In survey response evaluation, relevance is generally assessed based on both topic alignment and intent alignment. In other words, a response must not only address the topic of the question but also faithfully reflect its intent.

For open-ended survey responses, which often incorporate the respondent's unique experiences and context, simple embedding-based similarity

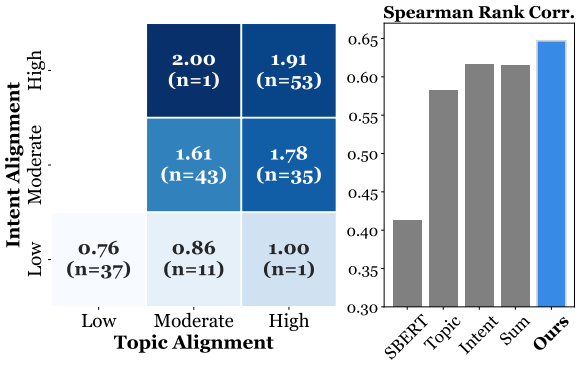

Figure 3: Heatmap (left) showing human expert-rated relevance scores by LLM-evaluated topic alignment (x-axis) and intent alignment (y-axis). Bar chart (right) shows the Spearman rank correlation of our proposed method ("Ours") and baseline methods.

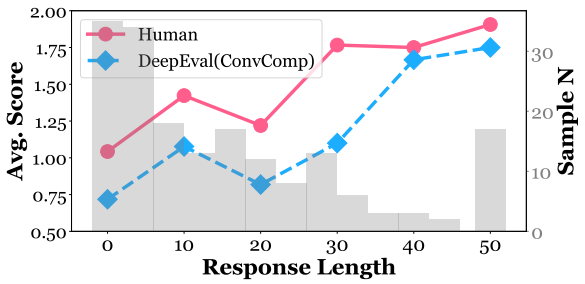

Figure 4: Average human expert-rated completeness scores (pink) and DeepEval Conversational Completeness scores (blue) across different response length bins. Grey bars indicate sample counts per bin.

measures are insufficient to fully capture this alignment (Fan et al., 2023). Addressing this limitation requires the high-level language understanding and inference capabilities of LLMs to grasp the implicit meaning behind questions.

Figure 3 shows how we used an LLM to evaluate both the topic and intent alignment of responses and compares these evaluations to human judgments. The results indicate that responses reflecting the question's intent, in addition to being topically relevant, were rated as having higher relevance overall. The approach that jointly considered both topic and intent achieved the highest agreement with human judgments. This suggests that inference-based evaluation using LLMs provides a more nuanced and human-aligned assessment compared to simple similarity-based methods.

Method The *relevance* dimension measures how well a response aligns with both the topic and the intent of the given question. It is assessed on a discrete scale from 0 to 4, ranging from completely irrelevant to fully relevant. Each level is defined based on two criteria: topic alignment and intent alignment.

3.2.3 Completeness

Preliminary Study User-written open-ended responses are often brief yet sufficient, conveying key information without explanation. This concise style is natural in human communication but often undervalued by automatic evaluation methods that rely on surface-level features such as length or keyword count. As a result, verbose but uninformative responses are often overrated, while concise yet complete reponses are penalized. Figure 4 shows the distribution of human-written response lengths and how human and LLM evaluations differ in treating length.

Language	English		Korean	
Dataset	**Dev**	**Test**	**Dev**	**Test**
# Response	181	94	122	100
Dim. Annot.	Crowd	Expert	Crowd	Crowd
Prac. Annot.	Crowd	Expert	Crowd	Expert

Table 1: Summary comparing English and Korean datasets across annotation aspects. "# Response" indicates the number of dataset samples, while "Dim. Annot." and "Prac. Annot." refer to annotators for per-dimension annotations and practical usage (overall quality and response acceptance) annotations, respectively.

We redefine *completeness* as the extent to which a response fulfills the core informational intent of the question. A complete response addresses all essential elements—explicit or implied—needed to satisfy what the question fundamentally asks. Partial responses missing key components or failing to engage with the question's intent are rated lower, regardless of length. To avoid redundancy, elaboration and contextual fit are excluded from this dimension and evaluated under *effort* and *relevance*, as suggested in prior work (Hu et al., 2024).

Method The *completeness* dimension is evaluated on a discrete 0–4 scale, mapped onto three qualitative categories: *Not Fulfilled*, *Partially Fulfilled*, and *Fully Fulfilled*. The LLM receives a tailored prompt containing the question, user response, and scoring rubric. The rubric guides the model to assess whether the response adequately covers the informational requirements of the question.

Language		English		Korean	
Dimension	**Method**	Spearman ρ	Kendall τ	Spearman ρ	Kendall τ
Effort	Length (token)	0.7889	0.6613	0.7621	0.6212
	Ours	**0.8198**	**0.7223**	**0.7950**	**0.6797**
Relevance	SBERT(all-MiniLM)	0.5839	0.4644	0.3429	0.2540
	SBERT(multi-lingual)	0.5223	0.4139	0.5538	0.4276
	DeepEval(AnswerRel)	0.5540	0.4759	0.4707	0.4191
	DeepEval(ConvRel)	0.4946	0.4597	0.2497	0.2307
	Ours	**0.8614**	**0.7954**	**0.6021**	**0.5269**
Completeness	Length (token)	0.6722	0.5374	0.6571	0.5171
	DeepEval(ConvComp)	0.4292	0.3859	0.3929	0.3478
	Ours	**0.8245**	**0.7438**	**0.7457**	**0.6372**

Table 2: Spearman's ρ and Kendall's τ correlations with human ratings for *effort*, *relevance*, and *completeness* dimensions across English and Korean datasets. Further results on expert annotations, crowd annotations, and inter-annotator agreement can be found in Appendix C.

4 Experimental Settings

4.1 Datasets

We conducted our study using human-written open-ended survey responses in English and Korean, collected from real-world datasets provided by *Nexxt Intelligence*. All responses were anonymized, and sensitive or personally identifiable information was thoroughly masked to protect participant confidentiality. To ensure diversity, responses were sampled across question types, lengths, and quality levels.

All responses were annotated for four core dimensions: *effort*, *relevance*, *completeness* and *gibberish*, as well as two dimensions of practical use: overall quality and response acceptance, by one to three crowd annotators (with averaged scores reported) or by a single domain expert, to assess how well our proposed method aligns with human judgments. For expert annotation of Korean data, question–response pairs were carefully translated into English to preserve their original nuance as much as possible. A summary of the dataset is shown in Table 1; details on annotation procedures and guidelines can be found in Appendix B and Appendix B.2.2.

4.2 Evalaution

Metrics To evaluate how well LLM scores align with human judgments, we compute Spearman (rank-based correlation) (Spearman, 1987) and Kendall tau (ordinal association) (Kendall, 1938) correlations between the raw LLM scores and the average human annotations across the three evaluation dimensions: *effort*, *relevance*, and *completeness*.

Baselines and Implementation Details We select baselines for each evaluation dimension. For Effort, since there is no existing LLM-based model for evaluating *effort*, we use a heuristic method based on token length (**Length (token)**). For Relevance, we employ Sentence-BERT, specifically **SBERT(all-MiniLM)**[1] and its multilingual variant **SBERT(multi-lingual)**[2]. For *relevance* and *completeness*, we use the default LLM-based automatic evaluation metrics provided by **DeepEval**[3]: AnswerRelevancy (AnswerRel), ConversationRelevancy (ConvRel), and ConversationalCompleteness (ConvComp). Both our proposed method and the LLM-based baselines were implemented using OpenAI's `gpt-4o-mini-2024-07-18` for inference, with the temperature set to 0 and the maximum tokens limited to 300.

5 Results and Analysis

5.1 Per-Dimension Performance

We analyze how well the proposed method aligns with human annotations across the three evaluation dimensions—*effort*, *relevance*, and *completeness*—in comparison to existing baselines. We use Spearman's rank correlation coefficient (ρ) and

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384

[3] https://deepeval.com/

Kendall's rank concordance coefficient (τ) as evaluation metrics, computed separately for English (En) and Korean (Ko) responses. Table 2 presents the correlation results.

Effort For the *effort* dimension, we evaluate the performance of a length-based baseline by measuring the correlation between the word count of each response and the corresponding annotated effort score. The results show that response length shows moderate alignment with perceived *effort* but fails to capture qualitative aspects like informational density. In contrast, our prompt-based approach achieves consistently higher correlations in both English and Korean. These results suggest that content-aware evaluation is more precise and reliable than simple length-based measures for assessing *effort*.

Relevance For the *relevance* dimension, we compare our method with sentence embedding models (SBERT (Reimers and Gurevych, 2019)) and the LLM evaluation toolkit DeepEval (Ip and Vongthongsri, 2025). SBERT shows weak to moderate alignment with human annotations, with particularly low agreement in Korean. DeepEval Relevance metric performs comparably or slightly worse than SBERT, indicating limited effectiveness in capturing human notions of relevance. In contrast, our prompt-based method shows the highest alignment with human ratings in both English and Korean, consistently outperforming all baselines. These results suggest that incorporating question topic and intent, rather than relying solely on surface-level similarity, yields more accurate and human-aligned *relevance* assessments.

Completeness For the *completeness* dimension, we compare our method against length-based and LLM-based evaluation baselines. Both the length-based metric, computed as in the *effort* dimension, and DeepEval's conversation completeness metric show only moderate or poor alignment with human annotations. In contrast, our prompt-based method consistently outperforms all baselines in both English and Korean. These results suggest that our method, by accounting for not only response length and explicit requirements but also implicit cues and contextual completeness, provides a more comprehensive and reliable evaluation of *completeness*.

Aggregation	English		Korean	
	ρ	τ	ρ	τ
Sum	0.90	0.78	0.60	0.51
Regression	0.90	0.79	0.61	0.51
LLM	0.87	0.79	0.54	0.50

Table 3: Spearman's ρ and Kendall's τ correlations with human ratings for overall quality prediction using different aggregation methods.

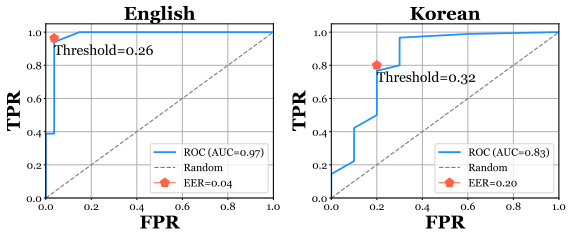

Figure 5: ROC curves for the English (left) and Korean (right) datasets.

5.2 Practical Usage

In this section, we evaluate how our framework's outputs can be applied to practical use cases. We aggregate the dimension-wise scores to demonstrate their utility in two key downstream applications: (i) **Response Quality Prediction**, where the aggregated score provides an interpretable, quantitative basis for researchers to analyze response quality, and (ii) **Response Rejection**, where the same score can be operationalized with a threshold to automate quality control. This evaluation is based on annotations from marketing researchers for both Korean and English responses.

Overall Quality Table 3 shows the performance of three aggregation methods for combining dimension scores in both English and Korean datasets. After normalizing scores between 0 and 1, we aggregated them using Sum (simple addition), Regression (ridge regression-based weighted sum), and an LLM-based approach (prompt-based synthesis of scores and reasons). Correlations with expert ratings were evaluated for both languages.

All methods show consistent performance, with the LLM-based approach additionally providing reasoning and explanations to enhance transparency for marketing researchers. Cost-efficient methods such as Sum and Regression also performed strongly, suggesting practical alternatives depending on resource constraints. The LLM prompt is detailed in Table 13.

Response Acceptance Figure 5 shows the ROC curves for acceptance classification for both English and Korean datasets. We computed overall quality using sum aggregation, normalized the scores between 0 and 1. Expert annotations were used as ground truth, where each response was labeled as 'accept', 'hold', or 'reject'; 'hold' and 'reject' were treated as rejection in this evaluation.

The proposed method achieved an AUC of 0.97 for English and 0.83 for Korean, indicating strong alignment with expert judgments in both languages. However, the performance for Korean was lower, possibly due to meaning distortion or loss of nuance during translation into English for annotation.

6 Future Works

Our work serves as a foundation for building smarter, more trustworthy, and more versatile survey systems. By providing a reliable and interpretable framework for evaluating open-ended responses across *effort*, *relevance*, and *completeness*, we set the stage for a series of impactful extensions:

Agentic and Real-Time Integration The framework can power multi-turn, agentic survey platforms that evaluate responses as they are given, flag shortcomings in specific dimensions, and generate targeted follow-up questions. This transforms surveys from static questionnaires into adaptive systems that improve data quality on the spot.

Safeguarding Data Quality Dimension-wise scores naturally lend themselves to detecting low-effort, insincere, or automated responses. Integrating such detection directly into survey workflows helps preserve validity, reduce downstream noise, and cut the costs of cleaning unusable data.

Cross-Lingual and Cross-Cultural Robustness Extending the framework beyond English and Korean ensures it generalizes across diverse populations. This not only supports broader applicability but also provides a basis for identifying and mitigating biases in multilingual and multicultural contexts.

Pathway to Quantitative Analysis The interpretability of dimension-wise scores offers a bridge to quantitative utilization. By aggregating or modeling these scores, researchers can derive higher-order insights and embed them into broader analytic pipelines. In short, the immediate contribution of this work is methodological clarity and reliability. The long-term value lies in its adaptability: a foundation from which future research can build survey systems that are dynamic, bias-aware, and analytically powerful.

7 Conclusion

In this paper, we presented a two-stage evaluation framework tailored to human-written open-ended survey responses, combining gibberish detection with dimension-wise evaluation of *effort*, *relevance*, and *completeness*. Our experiments demonstrated strong alignment with expert assessments in both English and Korean datasets, underscoring the framework's cross-lingual applicability. These results highlight the potential of our method as an effective and scalable solution for supporting decision-making in industrial contexts. In future work, its interpretable outputs could be leveraged to construct targeted follow-up questions that supplement missing information and encourage user engagement, further enhancing its practical utility in real-world marketing research.

Limitations

Although the proposed framework shows high concordance with expert judgments and consistently outperforms prior automated evaluation methods, it still presents several notable limitations.

First, the scoring process still depends on the subjective inference of large language models. Despite using rubric-based prompts and examples to clarify evaluation criteria, final scores reflect the data distributions and stylistic preferences learned during pretraining. While mitigation strategies were applied, such biases are difficult to fully eliminate due to the inherent characteristics of LLMs.

Second, the framework provides natural language explanations alongside each dimension score to enhance interpretability. However, these explanations are ultimately generated outputs from the LLM and do not explicitly reveal the underlying rationale or token-level evidence used in scoring. As a result, there are structural limitations in rigorously verifying the reliability or auditing the consistency of the evaluation outcomes.

Third, the three evaluation dimensions—*effort*, *relevance*, and *completeness*—were introduced based on practical considerations, analyses of real user responses, and established evaluation practices in marketing research. Experiments confirmed that

these dimensions are effective in distinguishing key qualitative aspects of responses and may provide helpful guidance for subsequent applications, including identifying or supplementing incomplete information in user response. Nonetheless, to establish these dimensions as general and theoretically grounded criteria for evaluating response quality, further discussion and systematic investigation are required. Fourth, while data release is crucial for transparency and reproducibility, the dataset used in this study cannot be made publicly available. The data is proprietary to our company, contains confidential client information, and was collected without securing consent for public release. To provide transparency within these constraints, we describe the dataset details and labeling process in Appendix B.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program). The authors gratefully acknowledge Ellie Ma for her assistance with data annotation.

References

Sher Badshah, Ali Emami, and Hassan Sajjad. 2025. Tale: A tool-augmented framework for reference-free evaluation of large language models. *CoRR*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *ACL*, pages 15607–15631.

James M. Crick. 2024. Analyzing survey data in marketing research: A guide for academics and postgraduate students. *Journal of Strategic Marketing*, 32(2):203–215.

Jing Fan, Dennis Aumiller, and Michael Gertz. 2023. Evaluating factual consistency of texts with semantic role labeling. In *SEM*, pages 89–100.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *EMNLP*, pages 61–71.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? *CoRR*.

Jeffrey Ip and Kritin Vongthongsri. 2025. deepeval. `https://github.com/confident-ai/deepeval`. Apache-2.0 License. Accessed: 2025-06-26.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *CoRR*.

Madhur Jindal. 2021. Gibberish detector: High-accuracy text classification model.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *CoRR*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. *CoRR*.

Oded Netzer, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2022. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *HCLT*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pages 3982–3992.

Jessie Rouder, Olivia Saucier, Rachel Kinder, and Matt Jans. 2021. What to Do With All Those Open-Ended Responses? Data Visualization Techniques for Survey Researchers. *Survey Practice*.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. 2025. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. *CoRR*.

Charles Spearman. 1987. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Meng-Chen Wu, Md Mosharaf Hossain, Tess Wood, Shayan Ali Akbar, Si-Chi Chin, and Erwin Cornejo. 2025. SEEval: Advancing LLM text evaluation efficiency and accuracy through self-explanation prompting. In *NAACL*, pages 7357–7368.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

A Language-Specific Gibberish Detection

We designed a gibberish detection module to identify nonsensical or meaningless text. This module helps reduce evaluation costs by filtering out samples that can be automatically assessed prior to human evaluation. The module consists of three stages: (i) **Preprocessing**, (ii) **Markov modeling**, and (iii) **Language-Specific Filtering**. A sentence is classified as gibberish if either its average log-likelihood falls below a certain threshold(*e.g.*, -4.0, -3.5) or it fails any of the language-specific filters described above. Since the characteristics and generative patterns of gibberish vary significantly across languages, detection strategies must be tailored to the structural and statistical properties of each language. In this study, we designed two independent detection pipelines, one for English and one for Korean, each reflecting the unique linguistic characteristics of the respective language.

A.1 Characteristics and Detection Strategy for English Gibberish

A.1.1 Preprocessing

The input text is converted to lowercase, retaining only Latin letters(a–z) and whitespace. Punctuation, digits, and other non-alphabetic symbols are removed to reduce noise and focus on purely alphabetic content.

A.1.2 Markov Model

We constructed a character-level bigram Markov model based on real-world conversational data[4]. Using this model, the average log-likelihood of a given input sentence is computed, providing a statistical measure of how well the text conforms to typical English character sequences. The threshold for the Markov model was empirically determined to be -4.0.

A.1.3 Language-Specific Filtering

- **Consecutive Consonant/Vowel Sequence Detection:** We detect sequences in which vowels or consonants are unusually repeated (e.g., aaaaaa, bbbbb). Such patterns often arise from keyboard mashing or unintended input. If the length of any such sequences exceeds a predefined threshold(10), the text is classified as gibberish.

- **Valid Word Ratio:** Using the NLTK tokenizer (Bird et al., 2009) and English dictionary, we calculate the proportion of valid English words in each response. If this ratio falls below a predefined threshold (*e.g.*, 0.4), the text is classified as gibberish.

- **Exception Handling for Abbreviations and Slang:** English conversational text frequently contains non-standard tokens (*e.g.*, abbreviations, slang, or elongated forms) that may superficially resemble gibberish. We retain such tokens when they convey clear meaning, or when either (i) the average log-likelihood under the English bigram Markov model exceeds -4.0 or (ii) the valid-word ratio surpasses the threshold. Frequent items (e.g., *lol*, *brb*) and laughter variants (e.g., *ha*, *haha*) are kept; elongated forms (e.g., *soooo*) are also retained unless both criteria fail and the text exhibits repetitive, low-information patterns. Purely nonsensical sequences (e.g., *asdf*, *dddd*) remain filtered out.

A.2 Characteristics and Detection Strategy for Korean Gibberish

A.2.1 Preprocessing

Korean text is written in syllabic blocks, each composed of individual Jamo characters (i.e., consonants and vowels). To capture fine-grained statistical patterns, we first decompose each syllable block into its constituent Jamo units.

A.2.2 Markov Model

Following decomposition, we designed a Jamo-level bigram Markov model based on the frequency of Jamo character pairs (i.e., sequences of initial consonant, medial vowel, and final consonant) of real-world conversational data[5]. This model estimates the average log-likelihood of a given text sequence, reflecting how well the character transitions conform to typical Korean orthographic patterns. The threshold for the Markov model was empirically set to -3.5.

A.2.3 Language-Specific Filtering

- **Syllabic Completeness and Jamo Diversity:** The proportion of fully-formed Hangul syllables and the diversity of unique Jamo components are used as additional indicators to

[4] https://www.kaggle.com/datasets/projjal1/human-conversation-training-data

[5] https://github.com/Ludobico/KakaoChatData

capture repetitive or structurally abnormal sequences. If the proportion of complete Hangul syllables falls below a specified threshold (e.g., 0.6), or if the diversity of unique Jamo characters is below a certain value (e.g., 4), the input is classified as gibberish.

- **Morphological Analysis Filtering:** We apply morphological analysis using KoNLPy's `Okt` tokenizer (Park and Cho, 2014). Sentences that fail to produce any recognizable morphemes are directly labeled as gibberish, as they likely contain no valid lexical units.

- **Exception Handling for Neologisms:** Certain forms of non-standard expressions in Korean-such as neologisms, abbreviations, or repetitive emotional markers may resemble gibberish due to their unconventional structure or brevity. Therefore, we manually whitelisted common conversational responses such as "ㅋㅋㅋ", "ㅇㅇ", or "헐" to prevent misclassification, despite their repetitive characters or absence from formal lexicons.

A.3 Experiment

A.3.1 Baseline

Pretrained Gibberish Detector We employed the publicly available pretrained model (Jindal, 2021) from Hugging Face. This model is a BERT-based classifier trained specifically to detect gibberish in English text.

Markov Model We employed a character-level bigram-based Markov probability model. The model collects the frequencies of adjacent character or jamo pairs from a training corpus and computes the average log-likelihood for a given input text. If the resulting value falls below a predefined threshold, the input is classified as gibberish. This approach corresponds to our method without the language-specific filtering step.

LLM Next, we adopted a prompting approach using a LLM (`gpt-4o-mini-2024-07-18`). The model was instructed to output either "true" or "false," indicating whether the given sentence is meaningful (i.e., not gibberish).

A.3.2 Dataset

For the test datasets, we collected 100 real-world samples for both English and Korean, each containing 50 gibberish and 50 non-gibberish sentences.

A.3.3 Results

Lang.	Model	F1	Prec.	Recall	Acc.
En	BERT	0.8900	0.9100	0.8900	0.8900
	Markov	0.3967	0.7577	0.5300	0.5300
	LLM	0.9247	**1.0000**	0.8600	0.9300
	Ours	**0.9800**	0.9810	**0.9800**	**0.9800**
Ko	Markov	0.6011	0.7941	0.6500	0.6500
	LLM	0.9474	**1.0000**	0.9000	0.9500
	Ours	**0.9800**	0.9810	**0.9800**	**0.9800**

Table 4: Gibberish detection performance across languages. The table reports F1 score, Prec.(Precision), Recall, and Acc.(Accuracy) for English (En) and Korean (Ko) datasets using different models.

Table 4 presents the gibberish detection performance across baseline models on both English and Korean datasets. Four metrics were used for evaluation: F1 score, Precision, Recall, and Accuracy.

English Dataset On the English dataset, the pretrained gibberish detector provided by Hugging Face achieved reasonable performance, with an F1 score and accuracy both at 0.89. The LLM performed even better, yielding an F1 score of 0.9247 and an accuracy of 0.93. In contrast, the character-level Markov model showed significantly lower performance, with an F1 score of 0.3967 and a recall of 0.53, indicating that this method failed to identify many gibberish instacnes. Our proposed model outperformed all baselines, achieving the best performance across all evaluation metrics, with F1, precision, recall, and accuracy all reaching 0.98.

Korean Dataset On the Korean dataset, overall performance across models was slightly higher than on the English dataset. This improvement can be attributed to the structural characteristics of Korean, where Jamo-level analysis enables more effective detection of gibberish. The Markov model outperformed its English counterpart, achieving an F1 score of 0.6011 and an accuracy of 0.65, though it still struggled with recall. The LLM demonstrated perfect precision (1.0) and a strong F1 score of 0.9474, but slightly lower recall (0.9), missing a few gibberish instances. Our proposed model once again achieved the best performance across all evaluation metrics, matching its English results with F1, precision, recall, and accuracy all at 0.98.

Limitations and Comparative Findings The Markov model, while computationally simple and fast, suffers from the limitation of relying solely on character-level frequency patterns, making it

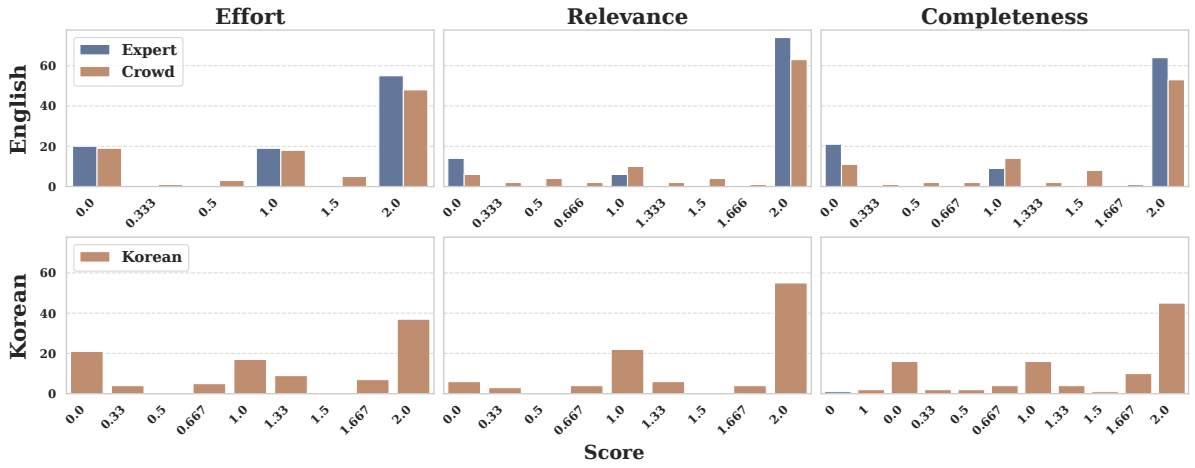

Figure 6: Score distribution of annotations by expert and crowd workers

ineffective at detecting more complex or irregular gibberish. On the other hand, both the pretrained gibberish detector and LLM attained relatively high precision but exhibited limitations in consistency and cross-lingual adaptability. Ultimately, our proposed method, which integrates structural characteristics and auxiliary linguistic signals (e.g., Jamo diversity, syllable completeness, and morphological analysis), achieved the most robust and reliable results across both languages.

B Dataset and Annotation Details

B.1 Dataset Overview

In this section, we provide details on the datasets and annotation procedures. The responses were written by actual survey participants and were sampled to ensure diversity in response lengths and quality levels. The English dataset consists of 275 responses (181 for dev and 94 for test), and the Korean dataset consists of 222 responses (122 for dev and 100 for test).

B.2 Annotation Procedure

B.2.1 Annotation Workflow

English Dataset All responses were evaluated by either crowd-workers or domain experts. For the English dev set, 1–3 crowd-workers from Prolific[6] annotated each response along four dimensions: *effort*, *relevance*, *completeness*, and gibberish. In the case of the English test set, a single domain expert annotated all responses across all dimensions.

Korean Dataset For the Korean dataset, responses were translated into English before being annotated by the domain expert. The Korean-to-English translation process consisted of two stages. In the first stage, we employed ChatGPT to generate an initial draft translation. To preserve the authentic characteristics of the real-world human responses, the model specifically prompted to perform a direct translation that prioritized fidelity to the original text over fluency, thereby retaining grammatical errors, typographical mistakes, and unconventional sentence structures. In the second stage, an internal bilingual reviewer examined each translated sentence against its original Korean counterpart to correct inaccuracies and ensure that subtle nuances, including the original errors or stylistic variations potentially overlooked by the LLM, were faithfully represented.

Finally, expert annotations were collected for the overall quality and rejection dimensions of the English and Korean test sets to ensure alignment with marketing research practice. Figure 6 shows the score distributions of experts and crowd workers for the English and Korean datasets.

B.2.2 Human Annotation Guidelines

Responses were sampled from prior user studies to ensure diversity in question types, response lengths, and quality levels. Each sample was fully annotated across all four evaluation dimensions: *effort*, *relevance*, *completeness*, and *overall quality*. Each response was annotated by 1 to 3 raters using a Likert scale. In addition, acceptance was categorized into three distinct classes.

- **Effort** (0–2): Low (0), Okay (1), Good (2)
- **Relevance** (0–2): Irrelevant (0), Partially Rele-

[6] https://www.prolific.com/

vant (1), Relevant (2)

- **Completeness** (0–2): Not Fulfilled (0), Partially Fulfilled (1), Fully Fulfilled (2)
- **Overall quality** (0–4): Very Poor (0), Acceptable (1), Fair (2), Good (3), Excellent (4)
- **Acceptance**: Accept, Hold, Reject

C Further Analyses

C.1 Correlation Analysis

Table 6 presents the correlation results between expert annotations and crowd-sourced annotations for each evaluation dimension—*effort*, *relevance*, *completeness*—as well as *overall quality*. Specifically, it reports the correlations coefficients among the following pairs: Expert–Ours, Crowd–Ours, and Expert–Crowd, along with inter-annotator agreement among the crowd annotators.

As shown in the results tables, both Expert-Ours and Crowd-Ours correlations are high for English, with Expert-Ours correlations being higher. This indicates that our proposed method aligns more closely with expert evaluations for English datasets.

For Korean, however, the correlation with experts is lower in overall quality compared to that with the crowd. This may be due to the fact that expert annotations were based on English translations of Korean responses, which could have distorted subtle linguistic nuances and contextual cues.

C.2 Significance Test for Correlation

Methodology We estimated bootstrap confidence intervals (CIs) for the differences in correlation coefficients between our proposed method and the second-best baseline (i.e., the method with the next-highest correlation) across the three dimensions *effort*, *relevance*, and *completeness*, and across two languages, English and Korean. Specifically, we performed bootstrap resampling with replacement at the level of question–response pairs, repeatedly computing the differences in Spearman's ρ and Kendall's τ. The final 95% CIs were derived using the percentile method. If the CI did not include 0, the result was considered statistically significant.

Results and Analysis Table 5 presents the 95% bootstrap confidence intervals for the correlation differences between our method and the second-best baseline across dimensions and languages.

- **Significant improvements:** We observe statistically significant improvements in English *relevance* (Spearman and Kendall) and Korean *completeness* (Kendall), where the CIs exclude 0. This indicates that our method is particularly effective at capturing semantic alignment (*relevance*) and core informational fulfillment (*completeness*).
- **Effort:** Since *effort* is inherently correlated with superficial features such as text length (with the second-best baseline being token-length), differences were not statistically significant in most cases. Nevertheless, English Kendall consistently shows a positive and significant improvement, highlighting advantages in rank-order consistency.
- **Impact of sample size:** The relatively small test set sizes (English: 94, Korean: 100) led to wide confidence intervals. We expect that validation on larger real-world datasets will yield clearer patterns of statistical significance across all dimensions.

C.3 Quadratic-weighted Kappa

Table 7 shows the inter-rater agreement between the proposed evaluator and human raters, measured using the quadratic-weighted kappa. Notably, the *completeness* dimension exhibits the highest importance across both languages, suggesting that the LLM-based evaluator applying our analytically defined evaluation criteria can achieve a high level of agreement with human judgments.

C.4 Length-based Baselines for Effort

We also evaluated alternative length-based baselines beyond token counts, including sentence counts (**Sentence count-based**) and counts of nouns, adjectives, and verbs (excluding *be*-verbs) (**Noun/Adj./Verb count-based**).[7]

As shown in Table 8, human responses are generally short and informal, with most consisting of only one or two sentences. Many are written as a single run-on sentence or phrase. Because of these characteristics, Sentence count-based and Noun/Adj./Verb count-based measures failed to capture meaningful variation in response length,

[7] For part-of-speech extraction, we used the NLTK toolkit (Bird et al., 2009). Sentence counts were computed by splitting on punctuation marks, primarily periods.

Question:
Which was the most important reason for staying at hotel and why?

User response:
i like the ambience and security

Effort (0-7): 2 (*Information: Very low, Specificity: Very low*)

Reason: The response provides a vague opinion with minimal detail, mentioning only two aspects: ambience and security. However, it lacks specificity and does not explain why these factors are important, resulting in very low information content.

Relevance (0-4): 3 (*Topic Matching: Fully aligned, Intent Matching: Mostly fulfills the question's intent*)

Reason: The response mentions aspects related to the hotel, specifically ambience and security, which are relevant to the topic. However, it does not fully explain why these factors were the most important reason for staying, thus partially fulfilling the intent of the question.

Completeness (0-4): 2 (*Partially Fulfilled*)

Reason: The response addresses the question by mentioning two reasons for staying at a hotel: ambience and security. However, it lacks depth and does not explain why these factors are important, leaving the answer incomplete.

Figure 7: A case study presenting the evaluation results of a single user response using our method in English dataset.

Question:
그 친숙한 맛이 제품을 선택하는 데 어떤 영향을 미쳤나요?

User response:
앞전 A80이 더 맛이 강하게 느껴질 뿐 W15는 기존에 출시한 어니언 크림맛과 동일한 느낌이 들며 더 나아가 늘 상 먹는 플링글스와 다를빠 없다고 느껴질 정도로 판단이 되고 있다.

Effort (0-7): 5 (*Information: Moderate to detailed, Specificity: Moderate to clear*)

Reason: 응답자는 제품의 맛에 대한 명확한 의견을 제시하고 있으며, A80과 W15의 맛 차이를 구체적으로 언급하고 있습니다. 그러나 전체적인 설명이 다소 제한적이며, 더 깊이 있는 분석이나 개인적인 경험이 부족합니다.

Relevance (0-4): 2 (*Topic Matching: Fully aligned, Intent Matching: Partially fulfills the question's intent*)

Reason: 응답은 제품의 맛에 대한 언급이 있지만, 질문의 핵심인 '친숙한 맛이 제품 선택에 미친 영향'에 대한 구체적인 설명이 부족합니다. 응답이 제품의 맛을 비교하고 있지만, 선택에 대한 이유나 영향을 명확히 드러내지 않아 의도 정합성이 부분적으로만 존재합니다.

Completeness (0-4): 3 (*Mostly Fulfilled*)

Reason: 응답은 제품 선택에 대한 친숙한 맛의 영향을 다루고 있으며, A80과 W15의 맛에 대한 비교를 제공하고 있습니다. 그러나 설명이 다소 모호하고 구체적인 영향이나 개인적인 경험에 대한 언급이 부족하여 완전한 충족에는 미치지 못합니다.

Figure 8: A case study presenting the evaluation results of a single user response using our method in Korean dataset.

leading to considerably weaker correlations with expert *effort* ratings.

Table 9 reports Spearman's ρ and Kendall's τ when *effort* is approximated by sentence counts and by counts of nouns, adjectives, and verbs (excluding *be*-verbs). As shown, these correlations are substantially lower than those based on token length. Given the short and highly informal nature of the responses, we therefore adopt token count as the length baseline in the main analysis.

D Case Study

We conducted a real-world case study of English and Korean data samples to validate the interpretability and effectiveness of our proposed evaluation framework.

English In the English example shown in Figure 7, the user answers "i like the ambience and security" to the question "Which was the most important reason for staying at the hotel and why?". The response receives low scores for both *effort* (2/7) and *completeness* (2/4), as it merely lists two factors-ambience and security-without elaboration or justification. However, the *relevance* score is relatively higher (3/4), since the mentioned aspects are directly related to the topic of hotel preference and partially fulfill the intent of the question.

Korean In the Korean example shown in Figure 8, the users answer the question "How did that familiar flavor influence your choice of product?" by comparing the flavors of two snack products (A80 and W15), noting that W15 tastes identical to an existing onion cream flavor and even resembles a familiar brand like Pringles. This response scores relatively high in *effort* (5/7) and *completeness* (3/4) due to its concrete product comparisons and moderate cognitive engagement. However, its *relevance* is rated lower (2/4) because, while it discusses taste, it does not directly address the core question, which is about how the familiarity of the flavor influenced the respondent's product choice.

This discrepancy highlights the importance of evaluating multiple distinct dimensions of response quality. Relying on a single aspect could obscure critical weaknesses or strengths. Moreover, the dimension-specific explanations provided by the framework enhance interpretability, offering valuable insights into the nature of each response.

	English		Korean	
Dimension	**Spearman ρ**	**Kendall τ**	**Spearman ρ**	**Kendall τ**
Effort	$[-0.005, 0.075]$	$[0.033, 0.124]$	$[-0.048, 0.143]$	$[-0.023, 0.160]$
Relevance	$[0.079, 0.329]$	$[0.075, 0.301]$	$[-0.082, 0.195]$	$[-0.017, 0.227]$
Completeness	$[-0.010, 0.087]$	$[0.044, 0.153]$	$[-0.003, 0.211]$	$[0.029, 0.229]$

Table 5: Bootstrap 95% confidence intervals (CIs) of correlation differences (Ours vs. second-best baseline).

Language		**English**		**Korean**	
Dimension	**Method**	**Spearman ρ**	**Kendall τ**	**Spearman ρ**	**Kendall τ**
Effort	Expert vs. Ours	0.8198	0.7223	-	-
	Crowd vs. Ours	0.8086	0.7155	0.7950	0.6797
	Expert vs. Crowd	0.7036	0.6489	-	-
	Inter-annotator (crowd)	0.7986	0.7089	0.7385	0.6867
Relevance	Expert vs. Ours	0.8614	0.7954	-	-
	Crowd vs. Ours	0.6575	0.5822	0.6021	0.5269
	Expert vs. Crowd	0.6610	0.6087	-	-
	Inter-annotator (crowd)	0.7940	0.7288	0.6784	0.6432
Completeness	Expert vs. Ours	0.8245	0.7438	-	-
	Crowd vs. Ours	0.7374	0.6532	0.7457	0.6372
	Expert vs. Crowd	0.6438	0.5890	-	-
	Inter-annotator (crowd)	0.7799	0.7029	0.6007	0.5676
Overall quality (Sum)	Expert vs. Ours	0.8953	0.7820	0.6024	0.5066
	Crowd vs. Ours	0.8318	0.7003	0.7040	0.5478
	Expert vs. Crowd	0.7792	0.6822	0.5318	0.4559
	Inter-annotator (crowd)	0.8654	0.7517	0.6127	0.5034

Table 6: Correlations between Expert and Ours, Crowd and Ours, and Expert and Crowd. Crowd refers to annotators recruited via Prolific. Inter-annotator indicates the agreement among crowd annotators.

Dimension	**English**	**Korean**
Effort	0.3632	0.4001
Relevance	0.4463	0.4227
Completeness	0.6242	0.5864

Table 7: Performance of the proposed method and human agreement using Quadratic-weighted Kappa.

# Sentences	**Count**
1	57
2	19
3	9
4	6
8	1
9	1
30	1

Table 8: Distribution of responses by number of sentences (English test set).

Length-based proxy	**Spearman ρ**	**Kendall τ**
Sentence count-based	-0.0616	-0.0564
Noun/Adj./Verb count-based	-0.0629	-0.0569

Table 9: Correlation with expert Effort ratings using alternative length-based proxies (English test set).

E LLM Evaluation Prompt

This section presents the LLM evaluation prompts used for the four assessment dimensions—*effort*, *relevance*, *completeness*, and *overall quality*—as introduced in Section 3.2 and Section 5.2. Each dimensional prompt is tailored to a specific dimension and includes a defined scoring range, evaluation criteria, and illustrative response examples. The prompts are designed based on rubrics to support consistent and reliable judgments by the evaluator.

The *effort* dimension assesses the respondent's cognitive engagement, focusing on the amount of information provided and the specificity of the response. The *relevance* dimension evaluates how well the response aligns with both the topic and the intent of the question. The *completeness* dimension measures the extent to which the response addresses the core informational requirements of the question. The *overall quality* dimension integrates the scores and justifications from the three preceding dimensions to produce a comprehensive assessment of response quality.

The LLM evaluates each dimension according to the corresponding prompt and returns both a discrete score and a brief justification. This prompt-based evaluation approach contributes to greater transparency, consistency, and interpretability in scoring.

The prompts were initially developed in English and then directly translated for use with the Korean data. Detailed English prompt specifications are provided in Tables 10– 13.

System Prompt:
You are an expert evaluator for a human-AI conversation dataset. Your task is to evaluate the quality of the responses in the dataset based on the given criteria.

User Prompt:
Rate how much thought and detail the user put into the response.
Use the following **0–7 scale**. The score must be an integer.
Base your judgment on the **information content**, **specificity**, and **how well the user responds to the question**.
Examples are provided to guide interpretation:

0: No meaningful response. The answer is either empty or completely unrelated.
- Information: None
- Specificity: None
- Response to question: Not at all

e.g., "N/A", "blah", "asdf"

1: Vague or evasive, such as default or placeholder answers that avoid the question.
- Information: Minimal or token or negligible
- Specificity: None
- Response to question: Barely reacts, without offering any insight or detail

e.g., "Good", "Okay", "I don't know", "Maybe"

2: Vague or generic opinion. Slightly more than a one-word answer, but still lacking substance.
- Information: Very low
- Specificity: Very low
- Response to question: Barely reacts

e.g., "Pretty good overall", "Not bad", "Nice"

3: A short opinion that includes a single element or impression.
- Information: Low
- Specificity: Low
- Response to question: Partially addresses one aspect

e.g., "Liked the burger", "Sick fries", "Nice vibe"

4: Slightly more informative with two aspects mentioned, but still minimal explanation.
- Information: Slightly basic
- Specificity: Limited
- Response to question: Briefly addresses two parts

e.g., "Burger was good. Service okay", "Decent food but small portions"

5: Thoughtful response with several clear opinions and specific examples.
- Information: Moderate to detailed
- Specificity: Moderate to clear
- Response to question: Engages with key parts

e.g., "Tasty food, but nothing special.", "Fries were crisp and burger was hot, but too salty."

6: Very rich and nuanced response; explains what stood out and why.
- Information: Very high
- Specificity: Very high
- Response to question: Deeply addresses the question with insight

e.g., "The burger had a smoky flavor, fries were hot and crisp, and the vintage vibe was cool. Pricey, but worth it."

7: Exceptionally detailed, thoughtful, and complete. Evaluates multiple dimensions (e.g., food, service, atmosphere) with depth and personality.
- Information: Comprehensive
- Specificity: Deep
- Response to question: Fully and thoughtfully addressed

e.g., "Burger exceeded expectations. The double bacon burger was perfectly cooked, the bun was fresh, fries had just the right crunch. Staff were welcoming, and the drive-in movie setup made the experience unique."

Read the given criteria carefully and follow them faithfully.
Return your evaluation as a JSON dictionary without any additional text:

```
{
"effort": <int: 0-7>,
"reason": "Detailed justification based on the criteria (2-3 sentences)"
}
```

Now, let's get it started!
`Conversation: {conversation}`
`Your evaluation: {JSON dictionary}`

Table 10: Prompt used for *effort* evaluation.

System Prompt:
You are an expert evaluator for a human-AI conversation dataset. Your task is to evaluate the quality of the responses in the dataset based on the given criteria.

User Prompt:
Rate how well the response aligns with the topic and intent of the question.

Two criteria should be considered:
- **Topic alignment**: Is the response about the same subject as the question?
- **Intent alignment**: Does the response address the purpose behind the question?

Note: The level of detail, reasoning, or length of the response should not be considered here. Even a short or simple response can receive a high relevance score if it correctly addresses the question's intent.

Examples are provided to guide interpretation:

0: Completely irrelevant or non-responsive
The response is off-topic, nonsensical, or fails to engage with the question in any meaningful way.
- Topic alignment: No
- Intent alignment: No

Example: Q: "Can you describe your experience with this product?", A: "I don't know." / "Get rich."

1: Topic is mentioned, but intent completely missed
The response includes a term or concept related to the question but does not address the question's actual purpose.
- Topic alignment: Yes
- Intent alignment: No

Example: Q: "Why did you choose this brand?", A: "The logo looks nice." (Mentions the topic, but doesn't explain the decision)

2: On-topic, but only partially fulfills the intent
TThe response shows some understanding of the question's purpose but provides limited or vague engagement.
- Topic alignment: Yes
- Intent alignment: Partially

Example: Q: "What made this experience special?", A: "It was nice." (Sentiment expressed, but no clear reason is given)

3: Matches the topic and mostly fulfills the intent
The response addresses the question appropriately and reflects a good understanding of its purpose, though some details may be missing.
- Topic alignment: Yes
- Intent alignment: Mostly

Example: Q: "What was your goal in using this app?", A: "To reduce stress." (Clearly aligned with the question's intent)

4: Fully aligned with both topic and intent
The response directly and clearly addresses what the question is asking, fulfilling its purpose. (Detailed justification is not necessary, as long as the intent is clearly understood and responded to.)
- Topic alignment: Yes
- Intent alignment: Yes

Example: Q: "Why did you like this product?", A: "It had a long-lasting scent and didn't irritate my skin." (Clear, specific reason matching the question)

Read the criteria carefully and follow them faithfully.
Return your evaluation as a JSON dictionary without any additional text:

```
{
  "relevance": <int: 0-4>,
  "reason": "Detailed justification based on the criteria (2-3 sentences)"
}
```

Now, let's get it started! `Conversation: {conversation}`
`Your evaluation: {JSON dictionary}`

Table 11: Prompt used for *relevance* evaluation.

System Prompt:
You are an expert evaluator for a human-AI conversation dataset. Your task is to evaluate the quality of the responses in the dataset based on the given criteria.

User Prompt:
Rate how completely the response fulfills the informational requirements implied by the question.
Use the following **0–4 scale**. The score must be an integer. Base your judgment on whether the response addresses **all relevant parts of the question**, and **the depth or adequacy of the information provided**.
Short answers like "yes" or "no" should not be automatically scored as 1:
- If they clearly and directly address the question but lack elaboration → score 1 (Minimally Fulfilled)
- If they are vague, off-topic, or do not engage with the question's intent → score 0 (Not Fulfilled)

Important: Even short or concise responses can receive a 3 (Mostly Fulfilled) or 4 (Fully Fulfilled) if they meaningfully and precisely address all parts of the question.
For multi-part questions, check if each part is answered. Responses that leave parts unaddressed or only partially answered should receive lower scores.

0: Not Fulfilled
- Response does not meaningfully address the question or is a placeholder, irrelevant, or nonsensical.

e.g., "fdsa", "I don't know", "What?", "That's private."

1: Minimally Fulfilled
- Mentions the topic but gives no substantive or relevant detail.
- Typically applies to yes/no answers with no elaboration.

e.g., "Movies are fun", "I like it", "Yes", "No"

2: Partially Fulfilled
- Response addresses only one part of a multi-part question, or answers a single-part question incompletely.

e.g.,
Q: "What's your favorite movie and why?" → A: "Inception."
Q: "What did you like and dislike?" → A: "I liked it." (missing 'dislike')

3: Mostly Fulfilled
- All parts of the question are touched on, but detail is limited or vague.

e.g.,
Q: "What's your favorite movie and why?" → A: "Inception, it's cool."
Q: "What do you like and dislike?" → A: "I liked the packaging. Didn't like the smell."

4: Fully Fulfilled
- Every part of the question is clearly and fully answered, with relevant detail or reasoning.

e.g.,
Q: "What's your favorite movie and why?" → A: "Inception, because I love mind-bending plots and the visuals were stunning."
Q: "What do you like and dislike?" → A: "I liked the compact size and smooth texture. I disliked how quickly it wore off."

Read the given criteria carefully and follow them faithfully.
Return your evaluation as a JSON dictionary without any additional text:

```
{
"completeness": <int: 0-4>,
"reason": "Detailed justification based on the criteria (2-3 sentences)"
}

Conversation: {conversation}
Your evaluation: {JSON dictionary}
```

Table 12: Prompt used for *completeness* evaluation.

System Prompt:
You are an expert evaluator for a human-AI conversation dataset. Your task is to evaluate the quality of the responses in the dataset based on the given criteria.

User Prompt:
Evaluate the overall quality of the user response using a 0–4 scale (integer only).
This score should represent a balanced assessment based on the combined performance across the following three dimensions:
- **Effort (0–1):** Thoughtfulness, specificity, and detail
- **Relevance (0–1):** Alignment with the topic and intent of the question
- **Completeness (0–1):** Coverage of the informational requirements

Do not simply average the three scores. Instead, weigh the strengths and weaknesses reflected in both the scores and their justifications.
A response that performs strongly in one area but poorly in others may warrant a moderate overall score.
Likewise, a consistently adequate response across all dimensions may merit a higher score than one with extremes.

You will be provided with the following input:
- effort score: `{effort_score}`
- effort reason: `{effort_reason}`
- relevance score: `{relevance_score}`
- relevance reason: `{relevance_reason}`
- completeness score: `{completeness_score}`
- completeness reason: `{completeness_reason}`

Scoring scale:
0 – Very Poor
1 – Poor
2 – Acceptable
3 – Good
4 – Excellent

Your reason should briefly explain:
(1) the strongest dimension,
(2) the weakest dimension, and
(3) how this balance supports your final score.

Return your evaluation as a JSON dictionary with no additional text:

```
{
  "overall_quality": <int: 0-4>,
  "reason": "Justification based on the criteria (2-3 sentences)"
}
```

Now, let's get it started!
`Conversation: {conversation}`
`Your evaluation: {JSON dictionary}`

Table 13: Prompt used for *overall quality* Aggregation.