applied sciences

MDPI

Article

Decoupled Dual-Stage Generation to Balance Factuality and Empathy in Customer-Support Dialogue Systems

Serynn Kim [1,†,‡], Hongseok Choi [2,*,†] and Jin-Xia Huang [2]

[1] Department of English Linguistics and Language Technology, Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea; serynnkim@gmail.com
[2] Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea; hgh@etri.re.kr
* Correspondence: hongking9@etri.re.kr
† These authors contributed equally to this work.
‡ This study was conducted while the author was affiliated with ETRI.

Featured Application

A practical framework for deploying factually grounded and empathetic dialogue agents in resource-constrained or on-premise customer-support environments using compact language models.

Abstract

In practical customer-support dialogue systems, responses must simultaneously deliver factually grounded information and context-appropriate empathy, yet existing single-stage generation models often exhibit specialization bias, favoring one objective at the expense of the other. To address this limitation, we propose a dual-stage generation framework that explicitly decouples factual grounding from empathetic modulation. Our primary configuration follows a fact-to-empathy order, in which the system first generates a fact-centric draft via structured query interpretation and optional retrieval-augmented generation, then applies empathy-aware tuning conditioned on inferred emotion type, intensity, and empathy necessity. To enable deployment in resource-constrained environments, only the query interpretation module is explicitly trained using knowledge distillation, allowing the overall system to operate with compact 4B–8B backbone language models. Furthermore, we construct a customer-support dialogue dataset designed to reflect realistic interactions involving both informational and emotional demands. Extensive experiments with compact models show that the proposed approach generally improves key dimensions of empathetic response quality while maintaining overall factual performance, thereby helping mitigate the representational entanglement empirically observed in single-stage baselines. Both quantitative metrics and scenario-based analyses confirm that decoupled generation enables a more balanced integration of factuality and empathy than single-stage generation. These results suggest that dual-stage generation provides a practical and extensible foundation for deployable, real-world customer-support dialogue systems.

Keywords: customer support; empathy; factuality; dialogue systems; dual-stage generation; retrieval-augmented generation; knowledge distillation; compact language models

Academic Editor: Douglas O'Shaughnessy

Received: 11 March 2026
Revised: 21 March 2026
Accepted: 22 March 2026
Published: 24 March 2026

1. Introduction

Intelligent chatbots have been widely adopted in customer support due to their scalability, rapid response, and 24/7 availability [1]. In practical customer-support environments, dialogue systems must coordinate two complementary yet distinct objectives:

delivering **factually grounded information** and expressing **context-appropriate empathy**. User queries frequently combine requests for technical guidance with signals of emotional states such as dissatisfaction. Therefore, joint handling of these dimensions is essential for effective problem solving and user satisfaction.

Factual grounding enables accurate and reliable responses for efficient issue resolution. Simultaneously, empathetic expression contributes to perceived service quality and sustained engagement [2]. Prior work has shown that human-like conversational behavior emerges from the integration of multiple dialogue competencies, including knowledge grounding and emotional awareness [3]. In customer-support contexts, the absence of either capability can result in responses that are either technically correct but perceived as insensitive or empathetic but unhelpful in resolving the underlying issue.

As illustrated in Figure 1, customer-support interactions often combine informational requests and emotional cues in real-world usage. However, many existing dialogue systems rely on single-stage generation that implicitly prioritizes a dominant objective [4]. As we will empirically demonstrate in Section 5.2, this design leads to pronounced specialization bias, where emphasizing one aspect limits cross-objective generalization. Consequently, responses may fail to address these integrated requirements effectively.

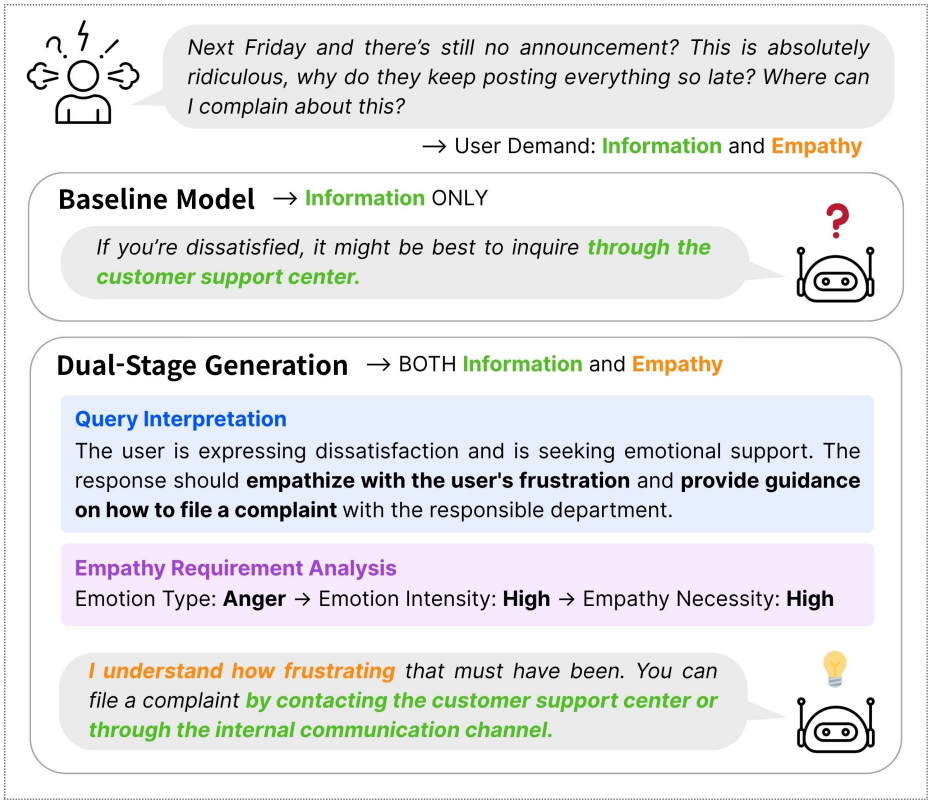


Figure 1. Illustration of a customer-support scenario. Single-stage generation may provide relevant guidance while insufficiently addressing users' emotional cues, whereas a dual-stage strategy combines factually grounded responses with context-appropriate empathy.

To address this limitation, we propose a dual-stage generation strategy that explicitly decouples factual grounding from empathetic expression. This approach is motivated by two observations: first, customer utterances commonly contain both informational and emotional components; second, jointly addressing factuality and empathy in a single generation stage can introduce representation-level interference between these objectives. In our framework, factual grounding and empathetic adjustment are handled in separate stages. The main configuration follows an F→E order, in which the system first performs

fact-centric drafting—analyzing user intent and applying retrieval-augmented generation (RAG) when external knowledge is required—and then applies empathy-aware tuning. To examine the effect of stage ordering, we additionally evaluate the reverse E→F configuration as an order variation. Explicitly decoupling complex tasks into two-stage or dual-path architectures has proven highly effective across diverse domains [5–7].

The framework is designed with practical deployment constraints in mind. Customer-support systems are often deployed in on-premises or resource-constrained environments where large-scale models are infeasible [8,9]. Accordingly, we evaluate our approach using compact backbone language models (4B–8B) to reflect a realistic balance between capability and deployment efficiency. To further improve robustness, we apply knowledge distillation to enhance the reliability of the query interpretation module.

This work makes the following contributions. We first provide an empirical analysis of customer-support dialogues, revealing a tendency toward specialization in single-stage models. Through cross-domain evaluation and representation analysis, we demonstrate that prioritizing either factuality or empathy leads to imbalanced generalization across these two objectives. We then propose a dual-stage generation framework that explicitly separates factual grounding from empathetic adjustment. Finally, we demonstrate that stage ordering meaningfully affects the final response profile, with the final stage exerting a dominant influence on the factuality–empathy trade-off. These findings highlight a practical advantage of the proposed framework: although our primary design follows an F→E configuration, the framework can be reconfigured according to deployment priorities, offering a practical foundation for real-world customer-support dialogue systems.

2. Related Work

2.1. Dialogue Systems

Dialogue systems have progressed from rule-based and retrieval-based methods toward neural generative models that facilitate more flexible and natural response generation [10,11]. More recently, the adoption of large language models (LLMs) has significantly enhanced dialogue fluency and adaptability, enabling robust instruction following and open-ended generation in practical applications [12]. Across various studies of natural language processing [13–15], maintaining factual reliability has emerged as a crucial issue for effectively leveraging these conversational capabilities.

Extensive research has focused on incorporating external knowledge to improve the accuracy and reliability of generated responses. Prior studies have introduced structured pipelines for knowledge-grounded dialogue generation [16] and leveraged knowledge graphs or retrieval mechanisms to ensure factual consistency [17,18]. Furthermore, hybrid retrieval–generation approaches have been shown to effectively reduce hallucination and improve task performance [19]. Building on this line of work, recent studies have further refined conversational RAG by selectively invoking retrieval when needed and by more tightly coupling retrieval with response generation [20,21]. More broadly, recent work has explored retrieval strategies that incorporate relational information beyond flat passage-level matching, thereby improving robustness in knowledge-intensive settings [22].

Despite these advances in knowledge-grounded and retrieval-augmented dialogue systems, most existing approaches primarily focus on factual correctness and task completion. However, real-world interactions often require responses that reflect users' emotional states while providing accurate information [23]. Recent studies on human–AI interaction suggest that conversational quality is also shaped by social–affective factors such as warmth, engagement, and responsiveness [24]. To address this, empathy-oriented dialogue systems have been actively studied. The EmpatheticDialogues dataset [25] has been instrumental in this field, enabling models to recognize and respond to user emotions [26–28].

Nevertheless, these systems are typically evaluated independently of factual knowledge requirements, limiting their applicability to information-intensive domains in which affective appropriateness and factual grounding must be achieved simultaneously.

2.2. The Customer-Support Domain

Customer support is a representative real-world application of dialogue systems that requires accurate and efficient problem resolution [29]. Recent studies have demonstrated that AI-powered chatbots can automate routine customer inquiries and improve service efficiency [1]. Much of the existing research in this domain has focused on improving response accuracy and knowledge utilization for customer inquiries [30]. Such approaches are effective for handling frequently asked questions and procedural inquiries.

However, customer interactions often involve frustration, dissatisfaction, or anxiety, making emotional responsiveness an important factor in service quality [31]. Empathetic chatbot responses have been shown to increase perceived warmth and customer satisfaction in service contexts [32]. In addition, recent studies on AI-driven customer-service chatbots highlight that personalization and empathetic response strategies are important design factors shaping customer satisfaction and the customer-service experience [2].

Although several studies in other high-stakes domains, such as healthcare, have explored balancing knowledge provision with emotional support [33,34], similar approaches remain limited in customer support settings. In practice, customer-support dialogues require factual guidance and empathetic responses to be integrated rather than treated as separate objectives. This gap motivates the development of dialogue systems that can dynamically integrate accurate information delivery with empathetic expression in customer support contexts. Our framework addresses this challenge by unifying knowledge-grounded response generation with empathy-aware interaction for more practical and user-centered customer-support systems.

3. Dual-Stage Generation

3.1. Task Formulation and Framework Overview

Customer-support dialogue generation requires simultaneously addressing users' informational needs and emotional states within a single response. Given a conversational context $C = [u_1, s_1, u_2, s_2, \ldots, u_n]$, where u_i and s_i denote user and system utterances, our goal is to generate an empathy-aware response R_{final} that satisfies both requirements.

To this end, we propose a *dual-stage generation* framework that decomposes response generation into two sequential stages. In the first stage, the system produces an intermediate *fact-centric output* R_{int} that emphasizes factual grounding and correctness. In the second stage, this output is further refined through empathy-aware tuning to incorporate context-appropriate emotional alignment, yielding the final empathy-aware response R_{final}.

A key component of the framework is a speech-act-based query interpretation module, which summarizes the user's intent and essential response constraints (denoted as I). When necessary, the system retrieves *supporting knowledge* K from an internal knowledge base via RAG. Based on these components, let P^{Fact} and P^{Emp} denote the conditional distributions corresponding to the fact-centric drafting and empathy-aware tuning stages, respectively. The overall process is formulated as follows:

$$R_{\text{int}} \sim P^{Fact}(\cdot \mid C, I, K), \quad R_{\text{final}} \sim P^{Emp}(\cdot \mid C, I, R_{\text{int}}).$$

Figure 2 provides an overview of the proposed framework. The framework employs a single backbone language model throughout the generation pipeline while assigning stage-specific roles through structured prompting and supervised fine-tuning. Although

the default configuration follows a fact-first, empathy-second order (F→E), the framework also supports an alternative empathy-first, fact-second order (E→F). This flexibility allows practitioners to choose the stage ordering that best aligns with their deployment priorities; we empirically compare the two configurations in Section 5.3. To ensure reproducibility, all prompt templates used in the framework are provided in Appendix A.1.

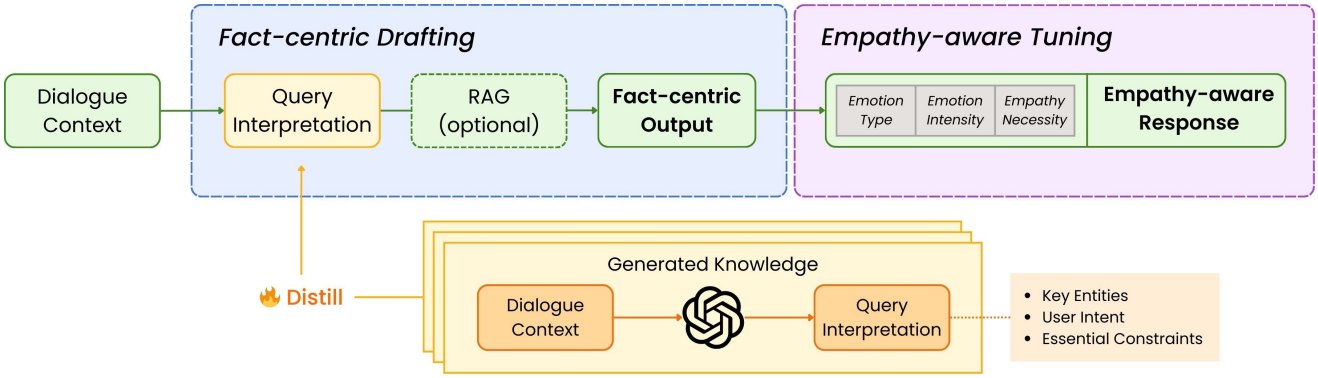


Figure 2. Overview of the proposed dual-stage generation framework in its primary F→E configuration. The framework consists of query interpretation, optional retrieval-augmented generation, fact-centric drafting, and empathy-aware tuning. The alternative E→F mode follows the same modular structure, but with the order of the two stages reversed.

3.2. Fact-Centric Drafting and Empathy-Aware Tuning

The first stage, *fact-centric drafting*, focuses on generating an intermediate output R_{int} with an emphasis on factual grounding. Given the dialogue context C, the model analyzes the user's latest utterance u_n to identify key entities and intent categories. Specifically, we classify user intents into four categories—fact verification, reasoning request, opinion request, and emotional support—by reformulating core concepts from speech-act theory [35,36] to reflect intent patterns observed in customer-support dialogues.

To constrain generation, the model produces a structured interpretation by extracting key entities, classifying user intent, inferring response constraints, and validating internal consistency—yielding a concise natural-language summary I that captures both the user's intent and essential response constraints (see Figure A1 for the full prompt), as underspecified contextual constraints are known to contribute to factual inconsistencies when left implicit [37]. As part of this interpretation, a self-validation step prompts the model to justify the assigned intent category and verify the completeness of the extracted constraints before finalizing I. Based on I, a separate retrieval decision step determines whether external knowledge is required and, if so, generates a targeted search query (see Figure A2). The top-3 relevant passages are retrieved from an internal knowledge base via dense-vector similarity search and provided as additional context K for fact-centric drafting. Under the conditions C, I, and K, the model generates the fact-centric output R_{int}.

The second stage, *empathy-aware tuning*, refines R_{int} into the final empathy-aware response R_{final}. This stage begins with an empathy analysis that infers three variables: the emotion type e, defined according to Ekman's seven basic emotions (neutral, joy, surprise, anger, sadness, disgust, and fear) [38]; the emotion intensity g; and the empathy necessity p, with both g and p assessed on a three-level scale (low, medium, or high). These variables are sequentially inferred by the backbone model through structured prompting (see Figure A4 for the full prompt). The model then rewrites only the emotional expression of R_{int} according to the determined empathy level, with an explicit directive to preserve all informational content unchanged. A final response review step further verifies contextual appropriateness before output. By separating factual grounding from empathy modu-

lation, the framework enables interpretable and fine-grained empathy control without compromising factual reliability.

3.3. Training of the Query Interpretation Module

To achieve strong performance with minimal computational overhead, the query interpretation module is specifically fine-tuned. Since the interpretation output I governs knowledge retrieval and response constraints, accurate interpretation is critical to the overall framework. We adopt a synthetic knowledge distillation in which a teacher model (GPT-4o [39]) generates high-quality interpretation output I' and a compact student model ψ is trained via supervised fine-tuning to imitate the teacher's reasoning process. Formally, the interpretation is represented as a token sequence $I' = (i_1, i_2, \ldots, i_T)$, and the student model is optimized using the standard autoregressive negative log-likelihood objective:

$$L_{\mathrm{KD}} = -\sum_{t=1}^{T} \log P_{\psi}(i_t \mid i_{<t}, C).$$

Through this distillation process, the student model learns to generate reliable interpretations while maintaining a compact model size, enabling efficient deployment in practical customer-support environments.

4. Dataset

4.1. Dataset Construction

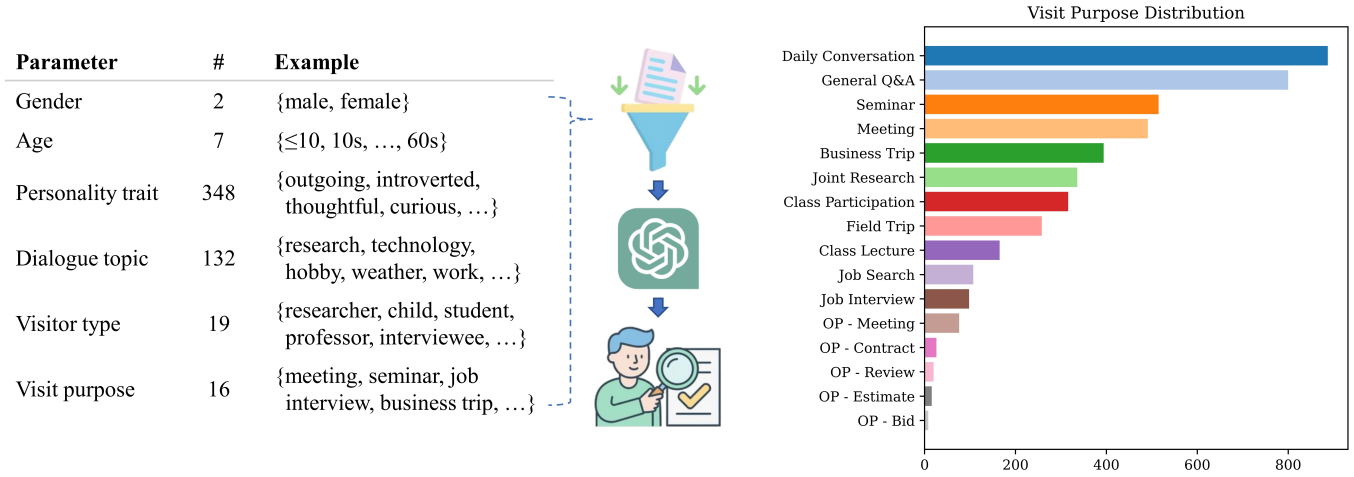
We construct a dialogue dataset to support the evaluation of our proposed framework, focusing on interactions between visitors and a guidance robot deployed at the Electronics and Telecommunications Research Institute (ETRI), a government-funded research institute in South Korea. The dataset reflects realistic visitor guidance scenarios grounded in ETRI's actual facilities, services, and research activities. To obtain a sufficiently large and reliable corpus, dialogues are first generated in a semi-automatic manner using GPT-4o with controlled prompts and are subsequently refined through expert review. An overview of the data construction pipeline is illustrated in Figure 3.

We structure the scenario space by defining several key parameters, including gender, age, personality traits, dialogue topics, visitor types, and visit purposes. These parameters are combined to generate diverse dialogue scenarios. However, exhaustively generating dialogues for all possible parameter combinations is both inefficient and unrealistic, as many combinations rarely occur in real guidance situations. For example, a young child visiting for a simple tour is unlikely to engage in discussions related to workplace issues.

To address this issue, we introduce a filtering stage to remove implausible combinations. First, common-sense constraints are applied to certain visitor types (e.g., student visitors are restricted to the teenage age group, with visit purposes primarily limited to field trips). Next, for each combination of dialogue topic, visitor type, and visit purpose, the model is asked to assess its plausibility using a three-level scale: *common*, *occasional*, and *rare*. Here, *common* denotes scenarios frequently observed in real guidance interactions, *occasional* refers to plausible but less frequent cases, and *rare* indicates theoretically possible but highly unlikely situations. Combinations labeled as *rare* are discarded. Through this process, we retain 10,000 realistic and diverse scenario combinations.

Each utterance is annotated with the following attributes: (1) emotion type, categorized into seven classes following Ekman's taxonomy; (2) a binary QA-type flag indicating whether the utterance involves ETRI-specific factual information; and (3) a search query when retrieval is required. The annotation guidelines were developed in consultation with ETRI domain experts and iteratively refined during the construction process.

Subsequently, two domain experts with experience in visitor guidance systems and human–robot interaction review and refine the generated dialogues following a structured guideline document that specifies annotation criteria, quality constraints (e.g., emotion-to-neutral ratio of approximately 70:30, factual QA turns limited to 20% of total turns), and preferred-response standards based on both factual accuracy and empathetic appropriateness. Dialogues that are redundant, inconsistent, or unnatural are removed. Rather than using the generated dialogues verbatim, the experts treat them as scenario seeds and rewrite each dialogue to ensure naturalness, coherence, and alignment with the actual guidance context of ETRI. They also insert appropriate question–answer utterances related to ETRI's facilities, services, and research activities to better reflect realistic visitor interactions. In total, the final dataset consists of 4513 dialogues comprising 40,207 utterances.



(**a**) Scenario parameter space (**b**) Distribution of visit purposes

Figure 3. Overview of dataset construction and characteristics.

4.2. Dataset Statistics

The dataset is split into training, validation, and test sets following an 80/10/10 ratio at the dialogue level. The training and validation sets are used to construct supervision data for training the query interpretation component. In contrast, the test set is explicitly designed to support comprehensive evaluation across diverse interaction conditions. Table 1 summarizes the overall statistics of the constructed dataset.

Specifically, the test set is constructed to evaluate model behavior across varying dialogue lengths. Instead of treating each dialogue as a single test case, we decompose it into multiple turn-level evaluation instances. For each instance, the context includes all utterances up to a given user turn, and the subsequent system response is used as the prediction target. This construction enables systematic assessment from short contexts to longer multi-turn interactions while maintaining a consistent evaluation objective.

Furthermore, test instances are categorized into two evaluation cases based on the emotion annotation of the reference system response. When the reference response emphasizes empathetic support, the corresponding instance is treated as an empathy-prioritized case, whereas instances whose reference responses focus on factual information delivery are categorized as factuality-prioritized cases. This separation enables targeted evaluation of empathetic response quality and factual correctness under controlled conditions.

Table 1. Summary statistics of the constructed dialogue dataset.

Category	Statistic	Value
Dataset size	Dialogues Utterances	4513 40,207
Data split	Train/val/test	80/10/10
Test set	Factuality-prioritized utterances Empathy-prioritized utterances	633 967
Dialogue length (empathy)	Min. turns Max. turns Avg. turns	1 15 3.37
Dialogue length (factuality)	Min. turns Max. turns Avg. turns	1 17 4.62
User profile	Visitor Employee	2638 (58.5%) 1875 (41.5%)
Annotations	Emotion labels Utterance type labels	{joy, neutral, sadness, surprise, fear, anger, disgust} {empathy, factuality}
Emotion distribution (%)	Joy/neutral/sadness Surprise/fear/anger/disgust	53.6/28.2/7.2 5.8/3.0/1.1/1.0

5. Experiments

5.1. Experimental Setup

Evaluation Metrics. We evaluate system performance along two complementary dimensions: *factuality* and *empathy*. To assess factual correctness, we employ automatic metrics including BLEU-3/4 [40], METEOR [41], and BERTScore [42]. In addition, following the LLM-as-a-judge paradigm [43], we measure **Faithfulness** using GPT-4o-mini [44], which rates each response on a scale from 1 to 5 according to its factual consistency with the ground-truth reference (1: completely inconsistent; 5: fully consistent with no errors or omissions).

Empathetic quality is evaluated using both lexical and semantic criteria. Following the multidimensional empathy evaluation framework of [45], we compute **Specificity** using Normalized Inverse Document Frequency (NIDF) [46], which captures the degree of response concreteness. We further adopt the LLM-as-a-judge framework with GPT-4o-mini for two affective dimensions: **Reflection Level**, measuring how deeply the system reflects and elaborates on the user's emotions (1: no empathy; 5: nuanced and supportive emotional reflection), and **Emotional Alignment**, assessing whether the response matches both the type and intensity of the user's annotated emotion (1: misaligned; 5: perfectly aligned). All LLM-based evaluations adopt a five-point Likert scale, with detailed rubric guidelines provided in Appendix A.3 to ensure reproducible scoring across samples.

Human Evaluation Validation. To validate the reliability of the LLM-based evaluation metrics, we conduct a small-scale human annotation study focusing on two empathy-related dimensions: Reflection Level and Emotional Alignment, which lack explicit reference signals. Two annotators independently rate 400 randomly sampled (context, response) pairs drawn from both backbones (Qwen and EXAONE) and both systems (the base system and ours), with 100 samples per configuration, using the same five-point rubric provided to GPT-4o-mini. Agreement between the averaged human scores and GPT-4o-mini predictions, measured by Quadratic Weighted Kappa (QWK) [47], yields scores of 0.56 for Reflection Level and 0.64 for Emotional Alignment, indicating reasonable alignment be-

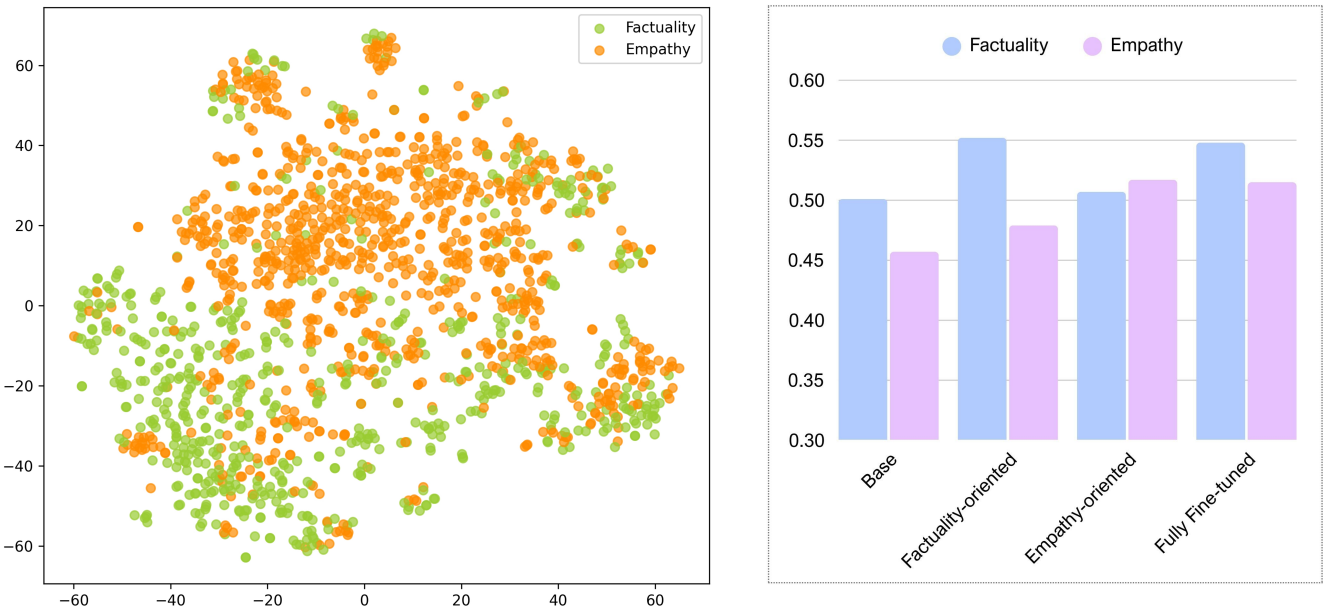
tween the LLM-based evaluations and human judgments. For reference, inter-annotator agreement is 0.57 for Reflection Level and 0.44 for Emotional Alignment.

Backbone Models. We conduct experiments using Qwen-3 (4B) [48] and EXAONE-3.5 (7.8B) [49] as the backbone language models. Qwen-3 is selected as a representative state-of-the-art open-weight model, while EXAONE-3.5 is adopted as a bilingual model to account for the Korean-language nature of our dataset. Both models have relatively small parameter counts, allowing us to evaluate system robustness under resource-constrained settings. To further assess generalizability, we additionally conduct tests on Llama-3 (8B) and Llama-3.1 (8B) [50] as widely adopted open-weight baselines.

Implementation Details. All models are fine-tuned using a unified training configuration. We employ a per-device batch size of four with gradient accumulation over four steps, resulting in an effective batch size of 16. Training is conducted for three epochs with a learning rate of 1×10^{-4}, cosine learning rate scheduling, and weight decay of 0.01. We enable gradient checkpointing and use the paged AdamW optimizer with 32-bit states for memory efficiency. For inference, we use greedy decoding to ensure deterministic and fair evaluation across methods. When retrieval augmentation is applied, the model accesses a knowledge base of 3458 sentence-level passages from internal ETRI guidance documents, encoded using KURE-v1 [51], a Korean-specialized dense retrieval model, and retrieves the top three passages based on cosine similarity.

5.2. Representation Analysis and Motivation

Representational Structure. We first visualize the ground-truth responses using t-SNE [52], as shown in Figure 4. Although all responses originate from the same customer-support domain, they exhibit distinguishable clustering patterns depending on whether the user primarily seeks factual information or emotional support, with partial overlap when both are required. This suggests that factual and empathetic behaviors occupy distinct yet intersecting regions in the representation space, reflecting the mixed and often ambiguous intents that arise in real-world customer interactions.



(**a**) t-SNE visualization of ground-truth responses (**b**) Cross-objective generalization

Figure 4. Empirical analysis of response representations and specialization in single-stage generation.

Cross-Objective Generalization. To examine whether this representational structure leads to objective-specific specialization in single-stage generation models, we conduct the cross-objective generalization study shown in Figure 4. Specifically, we fine-tune Qwen-3 (4B) separately on factuality-prioritized and empathy-prioritized subsets, and we evaluate each model on both dimensions using BERTScore. A fully fine-tuned model trained on the combined dataset is included as a reference.

The results reveal a pronounced specialization bias in single-stage fine-tuning. While specialized models outperform the base model on both metrics, the gains are strongly asymmetric: each model shows substantial improvement on its primary objective but only limited transfer to the complementary dimension. A model fine-tuned on the combined dataset yields a more balanced trade-off, but this requires careful data composition and greater computational cost, and it still does not fully resolve the tension between factuality and empathy. This implies that consolidating the two objectives into a single generation stage induces a representational imbalance, potentially limiting the model's ability to fully optimize both objectives simultaneously.

Implications. This limitation provides empirical motivation for our dual-stage formulation, which explicitly separates fact-centric response construction from empathy-aware refinement. The following section examines whether this decoupled design leads to more balanced performance across the two objectives.

5.3. Quantitative Performance Evaluation

Overall Performance. Tables 2 and 3 summarize the overall factual and empathetic performance of the proposed framework. Across both Qwen-3 (4B) and EXAONE-3.5 (7.8B), the proposed dual-stage framework (F→E) maintains competitive factual quality while improving key empathy-related metrics.

In terms of factuality, Qwen shows an improved METEOR result (0.294 → 0.338) while maintaining a comparable BERTScore result (0.501 → 0.530), and EXAONE exhibits increased Faithfulness (3.645 → 3.856). Empathy-related metrics also improve across both backbones. For Qwen, Reflection Level increases from 2.673 to 2.790 and Specificity from 0.576 to 0.599. EXAONE shows larger gains, with Reflection Level rising from 2.947 to 3.859 and Specificity from 0.547 to 0.610. Overall, these results indicate that the dual-stage decoupling framework mitigates the typical trade-off between factual grounding and empathetic response generation, enabling stronger empathetic reasoning without substantial degradation of factual quality.

Componentwise Analysis. To better understand the contribution of each module, we analyze the effects of the retrieval pipeline consisting of query interpretation (QI) and RAG, as well as the empathy-aware tuning module (ET). The retrieval pipeline mainly influences factual grounding. Applying RAG alone degrades factual performance in EXAONE (BLEU-3/4: 9.04/6.11 → 7.95/5.34; BERTScore: 0.497 → 0.487), suggesting that retrieval without explicit query interpretation may introduce irrelevant or misleading evidence. Incorporating QI mitigates this issue by reducing query ambiguity and improving retrieval relevance, recovering semantic metrics (Qwen BERTScore peaks at 0.532 under *+QI, RAG*; EXAONE improves from 0.487 to 0.502). Faithfulness also increases to 3.716 for Qwen and 3.882 for EXAONE under *+QI, RAG*, indicating improved factual consistency.

The empathy-aware tuning module primarily improves empathy-related metrics. For EXAONE, reflection increases substantially from 2.891 for *+QI, RAG* to 3.859 for *Ours (F→E)*, while emotional alignment also rises from 4.250 to 4.468. Qwen exhibits a similar trend, with reflection increasing from 2.365 to 2.790 after ET is applied. These improvements indicate that ET strengthens reflective reasoning and perspective-taking behaviors during response generation.

Table 2. Factuality evaluation results for different component settings and dual-stage operation modes. QI denotes query interpretation. F→E and E→F refer to the two operation modes of the proposed framework, ending with empathy-aware tuning and factual refinement, respectively.

Model	Variant	BLEU-3/4	METEOR	BERTScore	Faithfulness
Qwen-3 (4B-Instruct)	*Base*	9.76/7.49	0.294	0.501	3.341
	+ RAG	11.75/8.55	0.306	0.517	3.445
	+ QI	9.47/6.34	0.290	0.508	3.321
	+ QI, RAG	**12.91**/6.69	0.334	**0.532**	3.716
	Ours (E→F)	12.65/**9.30**	0.324	0.524	**3.801**
	Ours (F→E)	10.65/7.88	**0.338**	0.530	3.769
EXAONE-3.5 (7.8B-Instruct)	*Base*	**9.04**/6.11	0.306	0.497	3.645
	+ RAG	7.95/5.34	0.296	0.487	3.409
	+ QI	7.25/4.72	0.296	0.482	3.518
	+ QI, RAG	8.94/**6.22**	**0.327**	**0.502**	3.882
	Ours (E→F)	7.90/5.45	0.318	0.492	**3.899**
	Ours (F→E)	6.72/4.56	0.315	0.494	3.856

Bold indicates the best result, underlining indicates the second-best result, and italic text denotes model variants.

Table 3. Empathy evaluation results for different component settings and dual-stage operation modes. QI denotes query interpretation. F→E and E→F refer to the two operation modes of the proposed framework, ending with empathy-aware tuning and factual refinement, respectively.

Model	Variant	Specificity	Reflection Level	Emotional Alignment
Qwen-3 (4B-Instruct)	*Base*	0.576	2.673	**4.215**
	+ RAG	0.576	2.405	4.029
	+ QI	**0.601**	2.411	4.016
	+ QI, RAG	0.599	2.365	3.918
	Ours (E→F)	0.566	2.743	4.183
	Ours (F→E)	0.599	**2.790**	4.129
EXAONE-3.5 (7.8B-Instruct)	*Base*	0.547	2.947	4.424
	+ RAG	0.553	3.038	4.407
	+ QI	0.602	2.961	4.333
	+ QI, RAG	0.603	2.891	4.250
	Ours (E→F)	0.578	3.614	**4.477**
	Ours (F→E)	**0.610**	**3.859**	4.468

Bold indicates the best result, underlining indicates the second-best result, and italic text denotes model variants.

Backbone-Dependent Discrepancies. Despite the overall effectiveness of the framework, the magnitude of improvements differs across backbone models. EXAONE generally exhibits larger gains in empathy-related metrics, whereas Qwen shows more modest improvements. This difference likely reflects variation in backbone capacity. EXAONE, a larger bilingual 7.8B model selected for this Korean-language setting, appears better able to incorporate the additional control signals introduced by the dual-stage framework. This leads to stronger improvements in both reflection and emotional alignment. By contrast, Qwen, as a smaller 4B backbone, maintains relatively stable factual metrics but shows more limited gains from the empathy control signal. A similar pattern appears in Emotional Alignment. For Qwen, EA shows a minor decrease relative to the base model

(4.215 → 4.129). However, the component-level results suggest that much of this reduction originates from earlier stages: the *+QI, RAG* configuration already shows a lower EA of 3.918, and the subsequent empathy-aware tuning partially restores it to 4.129. This suggests that the EA decrease is more likely attributable to constraints introduced during query interpretation and retrieval than to the empathy-aware tuning module itself.

Effect of Stage Ordering. We further compare the two stage-ordering configurations: F→E, which applies fact-centric drafting before empathy-aware tuning, and E→F, which reverses this order. As shown in Tables 2 and 3, a consistent pattern emerges in which the final stage tends to exert a dominant influence on the corresponding objective. In the E→F configuration, factuality metrics such as BLEU and Faithfulness are generally higher (e.g., Qwen Faithfulness: 3.801 for E→F vs. 3.769 for F→E). Conversely, in the F→E configuration, empathy-related metrics such as Specificity and Reflection Level are generally stronger (e.g., EXAONE Reflection: 3.859 for F→E vs. 3.614 for E→F).

This ordering effect suggests that the second stage acts as the dominant factor of the final response characteristics. Rather than prescribing a single fixed order, the framework offers a configurable structure: practitioners can select the stage ordering that best aligns with deployment priorities. In scenarios where empathetic quality is critical, such as complaint handling or emotional counseling, the F→E configuration may be preferable. Conversely, in settings where factual precision is paramount, such as technical support or regulatory guidance, the E→F configuration may be more suitable.

Generalizability Across Backbones. To assess the framework's applicability beyond the primary backbones, we additionally evaluate on Llama-3 (8B) and Llama-3.1 (8B), as shown in Table 4. For Llama-3, Faithfulness and METEOR improve while Specificity and Reflection Level also show gains. For Llama-3.1, the full framework yields consistent improvements across all factuality metrics, with Faithfulness increasing from 3.226 to 3.501 and METEOR from 0.285 to 0.330. Reflection Level also improves (2.245 → 2.452), indicating enhanced empathetic reasoning. Across both models, Emotional Alignment exhibits a slight decrease, consistent with the backbone-dependent pattern discussed above. Notably, Faithfulness improves across all four backbone models evaluated in this study, suggesting that the dual-stage framework can enhance factual grounding and maintain robust performance across diverse backbone architectures. These results suggest that the proposed framework generalizes effectively across different backbone architectures.

Table 4. Generalizability evaluation across additional backbone models.

(a) Factuality

Model	Variant	BLEU-3/4	METEOR	BERTScore	Faithfulness
Llama-3 (8B-Instruct)	*Base*	**8.61**/5.67	0.275	**0.496**	3.120
	Ours (F→E)	7.72/**5.82**	**0.296**	0.469	**3.294**
Llama-3.1 (8B-Instruct)	*Base*	10.43/7.08	0.285	0.510	3.226
	Ours (F→E)	**12.75/9.88**	**0.330**	**0.532**	**3.501**

(b) Empathy

Model	Variant	Specificity	Reflection Level	Emotional Alignment
Llama-3 (8B-Instruct)	*Base*	0.640	2.204	**3.850**
	Ours (F→E)	**0.678**	**2.430**	3.733
Llama-3.1 (8B-Instruct)	*Base*	0.623	2.245	**3.962**
	Ours (F→E)	0.623	**2.452**	3.843

Bold indicates the best result, and italic text denotes model variants.

5.4. Scenario-Based Case Study

Beyond the quantitative evaluation, we conduct a scenario-based qualitative analysis to examine how our dual-stage framework performs under varying empathy and information demands. We first examine the primary F→E path and then compare the final responses with those of the base model across representative cases.

Stagewise Interaction. To illustrate how each stage contributes to the final response, Figure 5 presents two contrasting examples from the F→E configuration. In the case on the left, where the user expresses anxiety about an upcoming meeting, the query interpretation module identifies a need for emotional support and bypasses RAG. The fact-centric stage produces a reassuring but relatively neutral response, which the empathy-aware tuning stage then enhances with stronger supportive expressions based on the high empathy necessity. Importantly, the informational content of R_{int} is fully preserved in R_{final}. In the case on the right, the user requests specific institutional information, triggering RAG to retrieve relevant knowledge. Since the empathy analysis yields low necessity, the empathy-aware tuning stage leaves the fact-centric output unchanged, demonstrating that the second stage selectively modulates tone only when warranted. These examples confirm that the dual-stage framework adaptively adjusts empathetic expression while preserving factual content across both high-empathy and information-intensive scenarios.

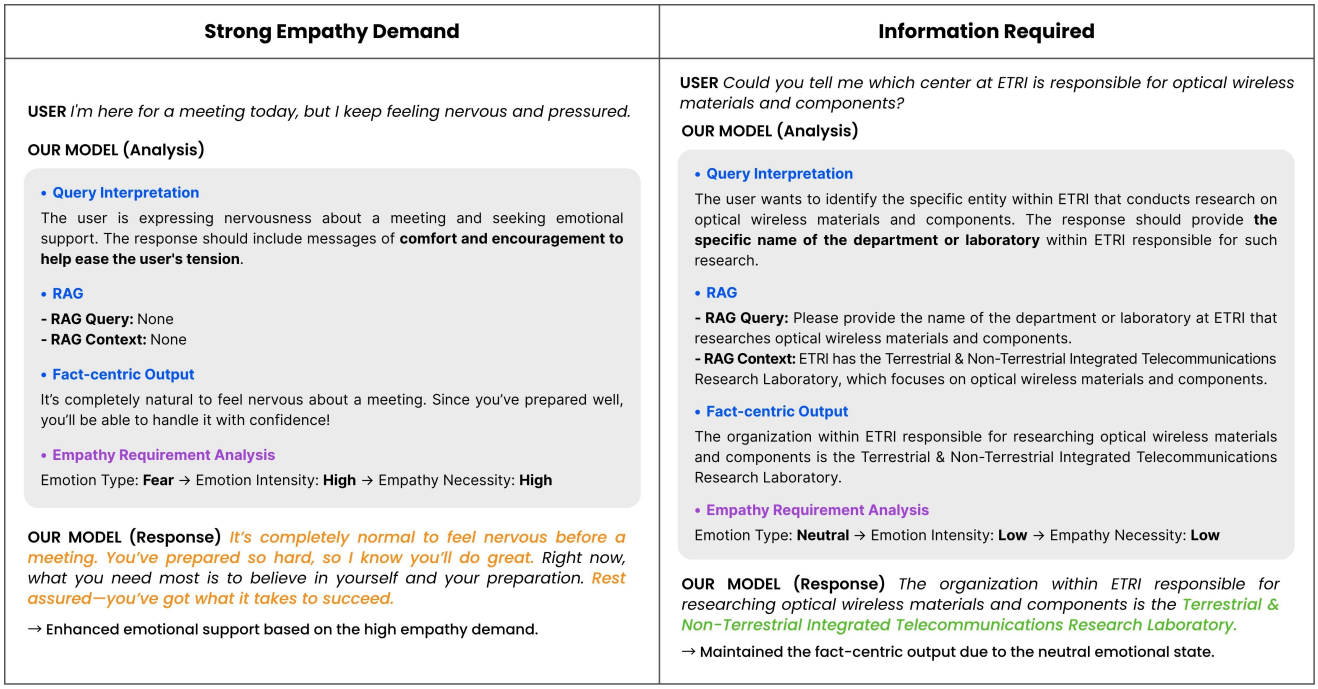


Figure 5. Intermediate example study illustrating the stagewise behavior of the dual-stage framework under two contrasting scenarios: strong empathy demand (**left**) and information required (**right**). Each panel shows the query interpretation, RAG decision, fact-centric output (R_{int}), empathy analysis, and final response (R_{final}).

Comparative Case Analysis. Figure 6 presents four representative cases defined by two factors: empathy demand (high vs. low) and information requirement (required vs. not required). This setup enables a qualitative comparison between the base model and the proposed framework across diverse customer-support scenarios. The proposed framework produces responses that are more consistently aligned with both the user's emotional state and the task requirements. In emotionally demanding situations, it acknowledges the user's feelings while still providing actionable and grounded guidance. In low-empathy or casual contexts, it prioritizes concise and relevant information without unnecessary affective

elaboration. By contrast, the single-stage baseline exhibits less stable behavior, often under-expressing empathy in emotionally charged scenarios or producing less appropriate responses when emotional tone and factual content must be addressed simultaneously.

These observations suggest that effective customer-support dialogue requires emotional attunement and information grounding to be coordinated rather than handled in isolation. They further indicate that the benefits of decoupling extend beyond the internal generation process to the final responses. By separating fact-centric drafting from empathy-aware tuning, the proposed framework achieves more reliable alignment with both conversational and task-related demands.

	Strong Empathy Demand	**Low** Empathy Demand
Information **Required**	**Case 1** **USER** *I'm trying to get to the Commercialization Office, but this building is so confusing. Shouldn't they at least provide a map?* **BASE MODEL** *Please provide your exact location so I can display a map or directions.* **(Wrong information & Low empathy)** **OUR MODEL** *I understand it can be confusing to navigate. The Commercialization Office is located in Building 1, Research Wing 1. Let me know if you get lost, and I'll help you right away.*	**Case 2** **USER** *It's too hot today. I think I'll go spend some time in the library. Do you know the opening hours?* **BASE MODEL** *The library is open on weekdays from 9:00 AM to 6:00 PM and is closed on weekends and holidays.* **(Wrong Information)** **OUR MODEL** *Sure! The ETRI library is open from 9:00 AM to 6:00 PM on weekdays, and from 10:00 AM to 10:00 PM on weekends and holidays.*
Information **Not Required**	**Case 3** **USER** *I took the wrong bus and thought I'd be late for my interview. I almost couldn't make it.* **BASE MODEL** *That's unfortunate. Next time, please double-check the time and take the correct bus.* **(Low empathy)** **OUR MODEL** *That must've been really stressful. I'm glad you made it to the interview just in time. I hope you receive great results!*	**Case 4** **USER** *Hey! I'm a bit bored today and just want to chat.* **BASE MODEL** *Hello! What topic would you like to discuss today?* **OUR MODEL** *Hello! What would you like to talk about today? For example, we can discuss recent interesting events or popular internet trends.*

Figure 6. Qualitative comparison between the base model and the proposed framework across four conversational scenarios. The proposed framework better balances factual accuracy and empathetic appropriateness, while the base model is more prone to factual errors or less appropriate responses.

6. Limitations and Future Work

Despite the promising results, this study has several limitations. First, the constructed dataset is semi-synthetic, refined through expert review, and restricted to a Korean customer-support domain. While this enables controlled evaluation, it may not fully capture the edge cases of real-world service interactions. In addition, the human evaluation remains limited in scale, making it difficult to fully assess end-user satisfaction in real deployment settings. Finally, although the proposed framework supports two stage-ordering configurations (F→E and E→F), the selection is made statically before deployment and does not support adaptive stage selection during inference, which may limit flexibility in more complex interaction settings.

Future work will focus on validating the proposed framework on naturally occurring dialogue logs and extending it to multilingual and cross-domain environments. We also plan to expand the current human evaluation into larger-scale user studies to better assess practical service quality. From an architectural perspective, future research may explore adaptive mode selection, dynamic routing mechanisms, and joint optimization strategies to more flexibly balance factuality and empathy. Beyond customer-support dialogue, the proposed objective decoupling paradigm may also be extended to broader agentic systems involving planning, tool use, and long-horizon reasoning.

7. Conclusions

We addressed the challenge of jointly delivering factual grounding and context-appropriate empathy in customer-support dialogue systems. Our empirical analysis revealed that single-stage generation induces specialization bias, limiting cross-objective generalization. To overcome this, we proposed a dual-stage framework that decouples fact-centric drafting from empathy-aware tuning, with a lightweight query interpretation module trained via knowledge distillation and configurable stage ordering (F→E or E→F) to suit deployment priorities.

Experiments across four compact backbone models (4B–8B) demonstrated that the framework consistently improves Faithfulness and empathetic dimensions such as Reflection Level, while both quantitative metrics and scenario-based analyses confirmed that decoupled generation balances factuality and empathy more effectively than single-stage baselines do. By requiring explicit training only for the query interpretation module, the approach remains efficient for resource-constrained environments. More broadly, our findings suggest that explicit objective decoupling offers a practical paradigm for multi-objective dialogue systems where competing generation demands must be jointly addressed.

Author Contributions: Conceptualization, H.C. and S.K.; methodology, H.C. and S.K.; software, S.K.; validation, S.K. and H.C.; formal analysis, S.K.; investigation, S.K.; resources, J.-X.H.; data curation, S.K.; writing—original draft, S.K.; writing—review and editing, H.C. and J.-X.H.; visualization, S.K.; supervision, H.C.; project administration, J.-X.H.; funding acquisition, J.-X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, "Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners").

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are partially available from the corresponding author upon request, subject to company security policies.

Acknowledgments: We especially thank the members of the Language Intelligence Research Section at the Electronics and Telecommunications Research Institute (ETRI) for their technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompt Templates

This appendix provides the prompt templates used for dual-stage generation, dialogue data construction, and evaluation to ensure reproducibility.

Appendix A.1. Dual-Stage Generation

This section includes Figures A1–A4, which present the specific prompts used for query interpretation, retrieval-augmented generation (RAG), fact-centric drafting, and empathy-aware tuning. The figures show the English translations of these prompts, originally written in Korean.

```
Based on the given conversation history, interpret the user's final
utterance according to the interpretation steps below.

### Interpretation Steps
1. Review the conversation history, focusing on the user's final
utterance. {conversation}

2. Infer the key entities that influence the intent of the utterance.
- Key entities representing the topic and context of the utterance
- The actual target of the user’s question or the core keywords the
user wants to know

3. Select the intent of the user's final utterance from the following:
- Fact checking, Inference request, Opinion request, Emotional
support, Other / Unknown

4. Considering the user's utterance intent, infer the constraints
required for the answer.
- Time, location, scope, format

5. Logically validate the interpretation results according to the
criteria below.
- Why it was classified as that intent
- What the key entities and constraints are, and how they should be
reflected in the answer

6. Summarize the interpretation results in Korean in no more than two
sentences.

### Output Format
Do not output the interpretation process. Output only the final
summarized result in Korean in no more than two sentences.
- The first sentence should summarize the user's intent.
- The second sentence should concisely describe the answer constraints
and how they are expected to be reflected.
```

Figure A1. Prompt used for query interpretation.

```
Based on the conversation below and the analysis of the user's intent,
determine whether information from ETRI (Electronics and
Telecommunications Research Institute) is required to accurately
answer the user's last utterance.
- If information is required, refer to the user's intent and write a
search query for retrieving the necessary information in the format:
"T, search query".
- If information is not required, output only "F".

### Conversation
{conversation}

### Inferred User Intent
{qa _analysis}

### Output Format
T, search query or F
```

Figure A2. Prompt used for RAG.

```
Generate an appropriate response to the user's last utterance based on
the conversation, the inferred intent, and the ETRI information.

### Conversation
{conversation}

### Inferred User Intent
{qa_analysis}

### ETRI Information (Refer only to the information necessary for the
response.)
{rag_context}

The response must be written in Korean and must not exceed two
sentences.
```

Figure A3. Prompt used for fact-centric drafting.

```
Determine the user’s empathy requirement according to the analysis
steps below and rewrite the AI’s existing response accordingly. If
empathy is already appropriately expressed or deemed unnecessary, do
not modify the response and output it as is.

### Conversation
{conversation}

### Inferred User Intent
{qa_analysis}

### Analysis Steps
1. Emotion Analysis
- Review the previous conversation and the intent behind the user’s
final utterance, then analyze the emotion.
- Classify the emotion expressed in the user’s final utterance as one
of the following: 'Neutral', 'Joy', 'Surprise', 'Anger', 'Sadness',
'Disgust', 'Fear'.

2. Emotion Intensity Assessment
- Determine the intensity level of the extracted emotion in the user’s
final utterance. (Low / Medium / High)

3. Empathy Requirement Determination
- Decide the level of empathy required for the user’s final utterance.
(Low Empathy / Medium Empathy / High Empathy)
- The more negative and intense the emotion, the higher the empathy
requirement should be.

4. Response Tone Rewriting
- Modify the expression of the AI’s existing response "{response}"
according to the determined empathy requirement.
- Do not change the information provided in the response; only modify
the expression.
- If empathy is already appropriately expressed or deemed unnecessary,
output the response as is without modification.

    - Low Empathy: Objective and neutral tone, focused on accurate
information delivery. Little or no modification is needed.
    - Medium Empathy: Include brief supportive language, maintaining a
balance between information and empathy.
    - High Empathy: Provide strong supportive expressions (comfort,
encouragement, reassurance), with a clearly empathetic tone.

5. Response Review
- Before outputting, verify once more that the response is appropriate
within the context.

### Output Format (You must output all labels and the response exactly
in the format below, without additional explanation.)
Emotion Classification, Emotion Intensity, Empathy Requirement, Final
Response:
```

Figure A4. Prompt used for empathy-aware tuning.

Appendix A.2. Dataset Construction

This section presents Figure A5, which displays the prompt and personas utilized for generating the short dialogues in the dataset.

```
The Electronics and Telecommunications Research Institute (ETRI) is a
government-funded research institute in South Korea. Recently, ETRI
has developed a chatbot capable of providing guidance as well as
engaging in everyday conversation.

Read the following personas and generate a short dialogue (5-7 turns)
between the "visitor" and the "chatbot" based on these personas.

### Visitor Persona
- I am a {visitor_age} {visitor_.gender}.
- I am {visitor_type}.
- I came to ETRI for {visit_purpose}.
- {visitor_personality}

### Chatbot Persona
- I am a guide at ETRI.
- I am an attentive listener.
- I am thoughtful.
- I respond with kindness and empathy.
- I respect the user.

The dialogue topic should be '{dialogue_topic}" or any topic closely
related to it. (...)
Now, please generate a short dialogue (5-7 turns) between the user and
the chatbot.
The dialogue should start with the user's utterance.
```

Figure A5. Prompt used for dataset construction.

Appendix A.3. LLM-Based Evaluation

This section features Figures A6–A8, which outline the evaluation prompts for assessing the model's faithfulness, reflection level, and emotional alignment.

```
You will be given a conversation between a user and an AI. You will
then be given a ground truth reference and one potential response for
the next turn in the conversation.
Your task is to rate the AI's response on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as
needed.

Evaluation Criteria:
Faithfulness (1-5) - Is the system response factually consistent with
the reference answer?
- A score of 1 means the response is completely unfaithful,
misleading, or irrelevant.
- A score of 2 means the response contains major factual errors or
contradictions.
- A score of 3 means the response is partially consistent but includes
noticeable factual errors or omissions.
- A score of 4 means the response is mostly consistent, with only
minor factual deviations.
- A score of 5 means the response is fully consistent with the
reference, with no errors or omissions.

Evaluation Steps:
1. Read the conversation, the ground truth reference, and the AI's
response.
2. Check whether the system response preserves the key factual
information from the reference.
3. Assign a score on a scale of 1-5 for faithfulness, according to the
criteria above.

Example:
Conversation History:
{conversation}
Ground Truth:
{reference}
AI's Response:
{ai_response}

Return ONLY the integer score:
```

Figure A6. Prompt used for evaluation of Faithfulness.

```
You will be given a conversation between a user and an AI. You will
then be given one potential response for the next turn in the
conversation.
Your task is to rate the AI's response on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as
needed.

Evaluation Criteria:
Reflection Level (1-5) - How deeply does the system reflect and
elaborate on the user’s emotions?
- A score of 1 means no empathy is shown; the response ignores or
dismisses the user’s feelings.
- A score of 2 means very weak empathy; vague or formulaic
acknowledgment.
- A score of 3 means basic empathy; minimal acknowledgment without
depth.
- A score of 4 means clear empathy with some elaboration or
interpretation.
- A score of 5 means sophisticated empathy; nuanced reflection that
interprets and supports the user’s emotions.

Evaluation Steps:
1. Read the conversation and the AI's response.
2. Determine whether the system only repeats surface-level phrases or
provides deeper reflection of feelings.
3. Assign a score on a scale of 1-5 for reflection level, according to
the criteria above.

Example:
Conversation History:
{conversation}
AI's Response:
{ai_response}

Return ONLY the integer score:
```

Figure A7. Prompt used for evaluation of Reflection Level.

```
You will be given a conversation between a user and an AI. You will
then be given the annotated emotion of the user’s last utterance and
one potential response for the next turn in the conversation.
Your task is to rate the AI's response on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as
needed.

Evaluation Criteria:
Emotional Alignment (1-5) - Does the system’s response align with the
user’s annotated emotion type and intensity?
- A score of 1 means the response is misaligned, reflecting the wrong
emotion type or a completely inconsistent tone.
- A score of 2 means weak alignment; vague acknowledgment without
accuracy in type or intensity.
- A score of 3 means partial alignment; correct recognition of emotion
type but mismatch in intensity, or vice versa.
- A score of 4 means mostly aligned; emotion type is correct,
intensity slightly mismatched.
- A score of 5 means perfect alignment with both emotion type and
intensity, with consistent tone and appropriate support.

Evaluation Steps:
1. Read the conversation, the annotated emotion, and the AI's
response.
2. Judge whether the system reflects the correct emotion type and
matches the intensity level.
3. Assign a score on a scale of 1-5 for emotional alignment, according
to the criteria above.

Example:
Conversation History:
{conversation}
Annotated Emotion:
{emotion}
AI's Response:
{ai_response}

Return ONLY the integer score:
```

Figure A8. Prompt used for evaluation of Emotional Alignment.

References

1. Marcineková, K.; Sujová, A.J.; Ďurica, R. Implementing AI Chatbots in Customer Service Optimization—A Case Study in Micro-Enterprise. *Information* **2025**, *16*, 1078. [CrossRef]
2. Uzan, S.; Freud, D.; Elalouf, A. Optimizing Chatbots to Improve Customer Experience and Satisfaction: Research on Personalization, Empathy, and Feedback Analysis. *Appl. Sci.* **2025**, *15*, 9439. [CrossRef]
3. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 300–325.
4. Herzig, J.; Feigenblat, G.; Shmueli-Scheuer, M.; Konopnicki, D.; Rafaeli, A.; Altman, D.; Spivak, D. Classifying Emotions in Customer Support Dialogues in Social Media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*; Fernandez, R., Minker, W., Carenini, G., Higashinaka, R., Artstein, R., Gainer, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 64–73. [CrossRef]
5. Chen, H.; Zhang, X. TS-CGANet: A Two-Stage Complex and Real Dual-Path Sub-Band Fusion Network for Full-Band Speech Enhancement. *Appl. Sci.* **2023**, *13*, 4431. [CrossRef]
6. Lei, Z.; Zhang, Y.; Chen, S. A Dual-Template Prompted Mutual Learning Generative Model for Implicit Aspect-Based Sentiment Analysis. *Appl. Sci.* **2024**, *14*, 8719. [CrossRef]
7. Choi, H.; Kim, S.; Liermann, W.; Seong, J.; Huang, J.X. Enhancing Automated Essay Scoring with Three Techniques: Two-Stage Fine-Tuning, Score Alignment, and Self-Training. *arXiv* **2026**, arXiv:2602.01747. [CrossRef]
8. Roh, J.; Kim, M.; Bae, K. Towards a small language model powered chain-of-reasoning for open-domain question answering. *ETRI J.* **2024**, *46*, 11–21. [CrossRef]
9. Manai, S.; Gemme, L.; Zanoli, R.; Lavelli, A. The IDRE Dataset in Practice: Training and Evaluation of Small-to-Medium-Sized LLMs for Empathetic Rephrasing. *Electronics* **2025**, *14*, 4052. [CrossRef]
10. Vinyals, O.; Le, Q. A neural conversational model. *arXiv* **2015**, arXiv:1506.05869.
11. Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proc. Aaai Conf. Artif. Intell.* **2016**, *30*. [CrossRef]
12. Yi, Z.; Ouyang, J.; Xu, Z.; Liu, Y.; Liao, T.; Luo, H.; Shen, Y. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Comput. Surv.* **2024**, *58*, 1–38. [CrossRef]
13. Wang, Y.; Wang, M.; Manzoor, M.A.; Liu, F.; Georgiev, G.N.; Das, R.J.; Nakov, P. Factuality of Large Language Models: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y.N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 19519–19529. [CrossRef]
14. Seong, J.; Park, J.; Liermann, W.; Choi, H.; Nam, Y.; Kim, H.; Lim, S.; Lee, N. MemEIC: A Step Toward Continual and Compositional Knowledge Editing. *arXiv* **2025**, arXiv:2510.25798. [CrossRef]
15. Jang, G.; Choi, H.; lim, C.; Lee, K.H.; Yi, M.Y. Leveraging Pretrained Knowledge at Inference Time: LoRA-Gated Contrastive Decoding for Multilingual Factual Language Generation in Adapted LLMs. In Proceedings of the Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
16. Lee, H.; Jeong, O. A knowledge-grounded task-oriented dialogue system with hierarchical structure for enhancing knowledge selection. *Sensors* **2023**, *23*, 685. [CrossRef] [PubMed]
17. Kang, M.; Kwak, J.M.; Baek, J.; Hwang, S.J. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv* **2023**, arXiv:2305.18846.
18. Xu, Z.; Cruz, M.J.; Guevara, M.; Wang, T.; Deshpande, M.; Wang, X.; Li, Z. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; ACM: New York, NY, USA, 2024; pp. 2905–2909.
19. Shen, W.; Gao, Y.; Huang, C.; Wan, F.; Quan, X.; Bi, W. Retrieval-generation alignment for end-to-end task-oriented dialogue system. *arXiv* **2023**, arXiv:2310.08877.
20. Wang, X.; Sen, P.; Li, R.; Yilmaz, E. Adaptive retrieval-augmented generation for conversational systems. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 491–503.
21. Mo, F.; Gao, Y.; Meng, C.; Liu, X.; Wu, Z.; Mao, K.; Wang, Z.; Chen, P.; Li, Z.; Li, X.; et al. Uniconv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 6936–6949.
22. Knollmeyer, S.; Caymazer, O.; Grossmann, D. Document graphrag: Knowledge graph enhanced retrieval augmented generation for document question answering within the manufacturing domain. *Electronics* **2025**, *14*, 2102. [CrossRef]
23. Görnemann, E.; Spiekermann, S. Emotional responses to human values in technology: The case of conversational agents. *Hum.-Comput. Interact.* **2024**, *39*, 310–337. [CrossRef]

24. Marconi, L.; Longo, L.; Cabitza, F. Assessing Interaction Quality in Human–AI Dialogue: An Integrative Review and Multi-Layer Framework for Conversational Agents. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 28. [CrossRef]
25. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5370–5381.
26. Fu, Y.; Inoue, K.; Chu, C.; Kawahara, T. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. *arXiv* **2023**, arXiv:2308.00085. [CrossRef]
27. Cai, M.; Wang, D.; Feng, S.; Zhang, Y. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5734–5746.
28. Cao, H.; Zhang, Y.; Feng, S.; Yang, X.; Wang, D.; Zhang, Y. TOOL-ED: Enhancing empathetic response generation with the tool calling capability of LLM. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 5305–5320.
29. Nicolescu, L.; Tudorache, M.T. Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review. *Electronics* **2022**, *11*, 1579. [CrossRef]
30. ABEL, U.; EMMANUEL, C.; PASCAL, U.O. Leveraging AI-Powered chatbots to enhance customer service efficiency and future opportunities in automated support. *Comput. Sci.* **2024**, *5*, 2485–2510.
31. Adam, M.; Wessel, M.; Benlian, A. AI-based chatbots in customer service and their effects on user compliance. *Electron. Mark.* **2021**, *31*, 427–445. [CrossRef]
32. Rohden, S.F.; Espartel, L.B. Emotional artificial intelligence: The impact of chatbot empathy and emotional tone on consumer satisfaction and word of mouth. *Int. J. Hum.-Comput. Stud.* **2026**, *210*, 103764. [CrossRef]
33. Wu, S.; Hsu, W.; Lee, M.L. EHDChat: A Knowledge-Grounded, Empathy-Enhanced Language Model for Healthcare Interactions. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 141–151.
34. Tsai, S.C.; Chen, Y.N. Balancing knowledge delivery and emotional comfort in healthcare conversational systems. *arXiv* **2025**, arXiv:2506.13692. [CrossRef]
35. Austin, J.L. *How to Do Things with Words*; Oxford University Press: Oxford, UK, 1962.
36. Searle, J.R. *Speech Acts: An Essay in the Philosophy of Language*; Cambridge University Press: Cambridge, UK, 1969.
37. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
38. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
39. Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. Gpt-4o system card. *arXiv* **2024**, arXiv:2410.21276. [CrossRef]
40. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318.
41. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 65–72.
42. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
43. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
44. OpenAI. GPT-4o Mini Model. 2024. Available online: https://developers.openai.com/api/docs/models/gpt-4o-mini (accessed on 20 March 2026).
45. Lee, A.; Kummerfeld, J.K.; Ann, L.; Mihalcea, R. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 179–189.
46. See, A.; Roller, S.; Kiela, D.; Weston, J. What makes a good conversation? how controllable attributes affect human judgments. *arXiv* **2019**, arXiv:1902.08654. [CrossRef]
47. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef]
48. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]

49. An, S.; Bae, K.; Choi, E.; Choi, K.; Jungkyu Choi, S.; Hong, S.; Hwang, J.; Jeon, H.; Jeongwon Jo, G.; Jo, H.; et al. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. *arXiv* **2024**, arXiv:2412.04862. [CrossRef]
50. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
51. Jang, Y.; Son, J.; Lee, T. *KURE: Korea University Retrieval Embedding*; NLP & AI Lab, Korea University: Seoul, Republic of Korea, 2024. Available online: https://huggingface.co/nlpai-lab/KURE-v1 (accessed on 20 March 2026).
52. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.