CARE: Mitigating Knowledge Intrusion for Empathetic Dialogue via Intent-Gated Retrieval and Conflict-Aware Reasoning

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) can undermine empathetic dialogue when retrieved content is contextually or emotionally misaligned, leading the model to uncritically rely on retrieved documents as its response—a failure mode we term "Knowledge Intrusion." To mitigate this, we propose CARE (Conflict-Aware Reasoning for Empathy), which synergizes Intent-Gated Retrieval and Latent Critique to ensure relevance, reinforced by Conflict-Aware DPO to enhance robustness against noisy contexts. Experiments on EmpatheticDialogues and ESConv demonstrate that CARE outperforms strong baselines, achieving F1 score gains of 7.3%–33.1% while maintaining high robustness, evidenced by a Context Rejection Score (CRS) exceeding 70%. Our code is available at https://anonymous.4open.science/r/CARE-FA8B.

1 Introduction

Empathetic dialogue systems are fundamental to human-centric AI, aiming to provide emotional support and companionship (Rashkin et al., 2019; Liu et al., 2021). While existing approaches, ranging from fine-tuned models (Majumder et al., 2020; Li et al., 2022b; Knob et al., 2025) to Large Language Models (LLMs) (Qian et al., 2023; Touvron et al., 2023; Ye et al., 2025a; Wang et al., 2025a), have mastered conversational fluency, a critical utility gap remains. Specifically, these systems often prioritize emotional mimicry by retrieving similar dialogue exemplars (Majumder et al., 2020; Yang et al., 2024b) rather than concrete solutions, thereby confining responses to safe but generic platitudes (e.g., *"I am sorry to hear that"*) (Tu et al., 2022).

To bridge this gap, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a promising solution to inject actionable information. However, naively applying RAG to empathetic conversation introduces the risk of knowledge intrusion (Shi et al., 2023). As illustrated in Figure 1,

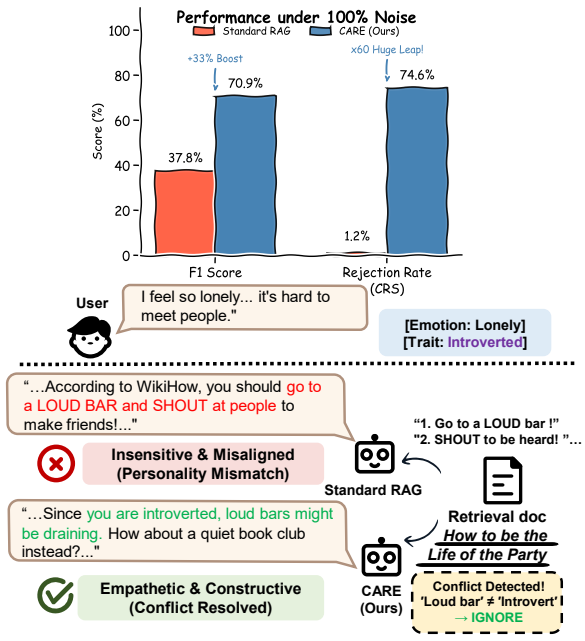

Figure 1: **Motivating Example: Resolving Empathy-Utility Conflicts.** CARE outperforms Standard RAG under 100% noise by rejecting irrelevant evidence. The case study demonstrates how CARE avoids empathy-utility conflicts (e.g., "loud bar" for an introvert) that Standard RAG blindly adopts.

standard RAG models often suffer from an uncritical over-reliance on retrieved contexts (Yoran et al., 2023; Shi et al., 2023). When the retrieved content is conflicting, the model mechanically forces this information into the response, thereby disrupting the emotional rapport. This highlights a fundamental distinction from QA tasks: while factual adherence is critical in QA, empathetic dialogue requires a discriminative rejection mechanism to discard knowledge when it conflicts with the user's persona or emotional state (Asai et al., 2024).

To overcome these limitations, we propose **CARE** (Conflict-Aware Reasoning for Empathy), a unified framework designed to dynamically resolve conflicts between retrieved documents and the user's current state via a cascaded cognitive architecture. CARE employs Intent-Gated Retrieval

to determine retrieval necessity and Conflict-Aware Reasoning (formulated as a verbalized Latent Critique) to logically reject noise, further aligned using a specialized Conflict-Aware DPO objective to explicitly distinguish pertinent advice from retrieval noise. Experimental results on EmpatheticDialogues and ESConv demonstrate that CARE delivers superior performance compared to competitive baselines, achieving F1 score gains of 7.3%–33.1% while maintaining high robustness with a Context Rejection Score (CRS) exceeding 70%.

In summary, our contributions are threefold: (1) We propose **CARE**, a conflict-aware framework that mitigates "knowledge intrusion" via a cascaded mechanism of Intent-Gated Retrieval and Latent Critique. (2) We introduce a Conflict-Aware DPO strategy trained on synthesized noise, which aligns the model to robustly reject irrelevant contexts without degrading conversational fluency, effectively mitigating the alignment-performance trade-off. (3) Distinct from black-box RAG models, we externalize the reasoning process via a sequential analyze-critique-decide mechanism to transparently justify retrieval decisions and enhance interpretability.

2 Related Work

Empathetic and Emotional Support Generation. Research has evolved from basic emotion recognition to complex support strategies. Early models, primarily evaluated on EmpatheticDialogues (Rashkin et al., 2019), relied on static emotion labels to guide generation. The field has shifted toward more sophisticated, counseling-style support strategies pioneered by ESConv (Liu et al., 2021). Recent LLM-based approaches, such as SoulChat (Chen et al., 2023) and SMILE (Qiu et al., 2024), further refine these strategies via fine-tuning (Wan et al., 2025) or role-playing (Ye et al., 2025b). However, the closed-book nature of these approaches limits their ability to provide actionable advice for complex problems, leaving a gap between emotional resonance and practical problem-solving.

Retrieval-Augmented Generation for Dialogue. While prior works (Cai et al., 2020; Li et al., 2022b) utilize retrieval, they are typically confined to history or commonsense, lacking specialized domain knowledge. To address this, Knowledge-Grounded Dialogue (KGD) (Dinan et al., 2018) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) have emerged as standard paradigms. Research has expanded from encyclopedic facts to long-horizon interactions (Liu et al., 2024; Wang et al., 2025b) and procedural assistance using instruction corpora (Qian et al., 2025). Nevertheless, standard RAG systems typically follow a rigid "retrieve-then-generate" paradigm. They assume that retrieved context is consistently helpful and lack intrinsic mechanisms to judge whether the information matches the user's state.

Robustness and Safety Alignment in RAG. In practical applications, robustness methods, such as Active RAG (Jiang et al., 2023b), Self-RAG (Asai et al., 2024), and reasoning-enhanced strategies (Tang et al., 2025), have been developed to handle noise and optimize retrieval performance. However, existing research predominantly focuses on factual consistency, preventing hallucinations in QA tasks. Similarly, safety alignment works like Llama 2-Chat (Touvron et al., 2023) employ refusal training to mitigate objective harm (e.g., toxicity or illegality) but overlook the nuance of "emotional safety." This oversight leads to rigid scenarios where models force external advice regardless of contextual alignment, failing to balance information delivery with necessary emotional support. Such accurate but misaligned advice results in "knowledge intrusion" (Shi et al., 2023) or sycophancy (Wei et al., 2023), disrupting the supportive tone. Recent analyses also highlight that RAG can paradoxically compromise safety properties (An et al., 2025).

3 Methodology

In this section, we propose Conflict-Aware Reasoning for Empathy (CARE), a unified framework to resolve empathy-utility conflicts by dynamically integrating external knowledge only when necessary. By utilizing a latent critique mechanism to filter out inappropriate and misaligned evidence, CARE delivers precise assistance while effectively preserving emotional resonance. As in Figure 2, CARE operates via a three-stage inference pipeline: (1) **Dynamic User Profiling**, (2) **Intent-Gated Retrieval**, and (3) **Conflict-Aware Reasoning and Generation** (via Latent Critique). To equip CARE with these capabilities, we additionally introduce **Conflict-Aware Alignment**, a training-only curriculum that improves robustness under retrieval conflicts.

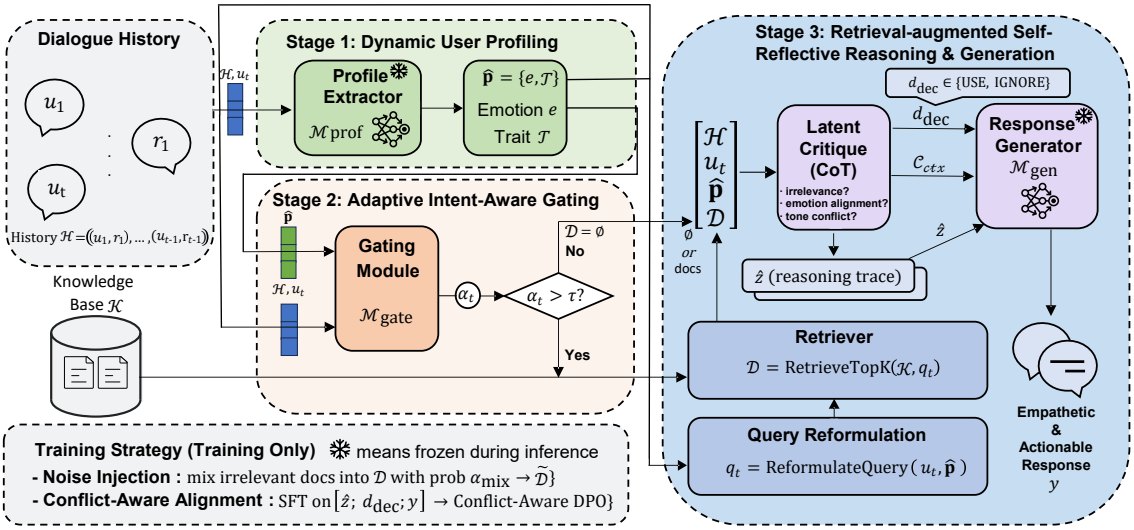

Figure 2: **The overall framework of CARE.** The inference process follows three main stages: (1) Dynamic User Profiling, (2) Intent-Gated Retrieval, and (3) Conflict-Aware Reasoning & Generation. Additionally, we illustrate the Conflict-Aware Alignment strategy (training-only) used to enforce robustness against knowledge intrusion.

3.1 Problem Formulation

Let $\mathcal{H} = \big((u_1, r_1), \ldots, (u_{t-1}, r_{t-1})\big)$ denote the past dialogue history, and u_t denote the current user utterance. The goal is to generate the target response y. We assume a retrieval over a knowledge base $\mathcal{K}$ providing a candidate document set $\mathcal{D}$. To guide generation, we introduce two latent variables: $\mathcal{P}$ capturing the user's internal state, and z for evaluating retrieved information utility. The objective is to learn a policy π_θ (parameterized by a Large Language Model) defining the conditional distribution. We formulate the generation of y as a latent variable model:

$$P_\theta(y \mid \mathcal{H}) = \sum_{\mathcal{P},z} P_\phi(\mathcal{P} \mid \mathcal{H}) \cdot P_\theta(z \mid \mathcal{H}, \mathcal{P}, \mathcal{D}) \cdot P_\theta(y \mid \mathcal{H}, \mathcal{P}, \mathcal{D}, z), \tag{1}$$

where $\mathcal{D}$ is the retrieved context. We approximate Eq. 1 by estimating the optimal profile $\hat{\mathbf{p}}$ and reasoning trace $\hat{z}$ via greedy decoding or beam search. $P_\phi(\mathcal{P} \mid \mathcal{H})$ denotes the profile extraction distribution induced by a pre-trained extractor with fixed weights ϕ.

3.2 Overview of Proposed Model

Figure 2 outlines CARE. Given history $\mathcal{H}$, CARE first infers a structured profile $\hat{\mathbf{p}}$ (Stage 1) and decides retrieval necessity (Stage 2). If activated, CARE reformulates an intent-aware query to retrieve documents $\mathcal{D}$; otherwise, $\mathcal{D} = \emptyset$ to avoid knowledge intrusion. Stage 3 integrates retrieval and reasoning: conditioned on context, CARE performs self-reflective reasoning to produce a critique trace $\hat{z}$ and a utility decision d_{dec}, guiding the final response y. Finally, we introduce Conflict-Aware Alignment, a training-only curriculum to enforce robustness against conflicting retrieval.

3.3 Dynamic User Profiling

To ensure personalized empathy, utilizing a frozen extractor $\mathcal{M}_{\text{prof}}$ (parameterized by ϕ), we extract a token sequence $\mathbf{p} = (p_1, \ldots, p_k, p_{k+1}, \ldots, p_m)$, where the initial segment $(p_1, \ldots, p_k)$ represents the user's emotional state e, and the remaining tokens $(p_{k+1}, \ldots, p_m)$ represent the user's personality traits $\mathcal{T}$. The optimal profile sequence $\hat{\mathbf{p}}$ is obtained by maximizing the joint probability conditioned on the dialogue history $\mathcal{H}$:

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \prod_{j=1}^{m} P_\phi(p_j \mid p_{<j}, \mathcal{H}). \tag{2}$$

This ensures that the generated profile dynamically aligns with the user's latest emotional shifts while maintaining consistent personality traits.

3.4 Intent-Gated Retrieval

As Wang et al. (2025b) noted, indiscriminate retrieval introduces noise. To mitigate knowledge intrusion, we employ a probabilistic gating mechanism. Let $\mathcal{C}_{\text{intent}}$ be the full intent taxonomy, and let $\mathcal{C}_{\text{info}} \subset \mathcal{C}_{\text{intent}}$ denote the subset of intents that specifically require external knowledge (e.g., problem solving). We utilize a gating module $\mathcal{M}_{\text{gate}}$

to estimate the posterior $P(c_k \mid \mathcal{H}, \hat{\mathbf{p}})$ and define a retrieval necessity score α_t as the cumulative probability mass over information-seeking subset $\mathcal{C}_{\text{info}}$:

$$\alpha_t = \sum_{c_k \in \mathcal{C}_{\text{info}}} P(c_k \mid \mathcal{H}, \hat{\mathbf{p}}). \tag{3}$$

The discrete decision $g \in \{0, 1\}$ is obtained by $g = \mathbb{I}[\alpha_t > \tau]$. If $g = 0$, we enforce $\mathcal{D} = \emptyset$, preventing retrieval noise.

3.5 Profile-Aware Query Reformulation

When $g = 1$, we bridge the semantic gap between emotional expressions and knowledge sources via Profile-Aware Query Reformulation. Instead of using the raw utterance u_t, the model generates a standalone "How-to" query q_t incorporating traits from $\hat{\mathbf{p}}$. For example, the utterance "*I just lost my job...*" is reformulated into $q_t =$ "*How to write a resume and find a job.*" This optimized query q_t is then utilized to retrieve candidate documents $\mathcal{D}$ for the subsequent stage.

3.6 Self-Reflective Reasoning with Noise Injection

To reject misleading information, we instantiate $\hat{z}$ via a Chain-of-Thought (CoT) (Wei et al., 2022) driven latent critique. During training, we use a noise injection strategy where irrelevant documents are mixed into $\mathcal{D}$ with probability α_{mix}, creating context $\tilde{\mathcal{D}}$. This forces the model to learn the rejection path.

Let $\mathcal{C}_{\text{ctx}} = \{\mathcal{H}, u_t, \hat{\mathbf{p}}, \tilde{\mathcal{D}}\}$ be the aggregated context. We employ the generator $\mathcal{M}_{\text{gen}}$ (parameterized by θ) to model the joint probability of reasoning trace $\hat{z}$, decision d_{dec}, and response y, which is factorized as:

$$\begin{aligned} P_\theta(y, d_{\text{dec}}, \hat{z} \mid \mathcal{C}_{\text{ctx}}) &= P_\theta(\hat{z} \mid \mathcal{C}_{\text{ctx}}) \\ &\cdot P_\theta(d_{\text{dec}} \mid \hat{z}, \mathcal{C}_{\text{ctx}}) \cdot P_\theta(y \mid d_{\text{dec}}, \hat{z}, \mathcal{C}_{\text{ctx}}). \end{aligned} \tag{4}$$

This enables $\mathcal{M}_{\text{gen}}$ to sequentially perform: (1) Critique Generation (reasoning about relevance); (2) Utility Decision; and (3) Response Generation.

3.7 Conflict-Aware Alignment

To equip the generator with robust reasoning capabilities, we align the backbone LLM via a two-phase curriculum distilled from a powerful Teacher Model: Supervised Fine-Tuning (SFT) via Reasoning Distillation to establish the critique format, followed by Conflict-Aware Direct Preference Optimization (DPO) to enforce robustness.

Phase 1: Reasoning Distillation (SFT). We first warm up the student model to follow the critique-then-respond format. The training data $\mathcal{D}_{\text{gold}}$ is synthesized by a large-scale Teacher Model, which provides high-quality reasoning traces and empathetic responses. The target sequence is denoted as $\mathbf{Y} = [\hat{z}; d_{\text{dec}}; y]$. The model minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathcal{C}_{\text{ctx}}, \mathbf{Y}) \sim \mathcal{D}_{\text{gold}}} \left[\log P_\theta(\mathbf{Y} \mid \mathcal{C}_{\text{ctx}}) \right]. \tag{5}$$

We denote the model after this phase as π_{ref}. However, we observe that SFT alone induces a systemic bias towards utilization. The model superficially adopts the reasoning format but lacks the discriminative capability to strictly enforce negative constraints against irrelevant retrieval.

Phase 2: Conflict-Aware DPO. To rectify this and explicitly mitigate blind trust, we apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) on conflict scenarios. We construct a preference dataset $\mathcal{D}_{\text{conflict}} = \{(\mathcal{C}_{\text{ctx}}, \mathbf{Y}_{\text{w}}, \mathbf{Y}_{\text{l}})\}$. Here, both responses are synthesized by the Teacher Model to simulate the contrast in reasoning: the winning response $\mathbf{Y}_{\text{w}}$ represents the robust behavior (correctly identifying conflict and rejecting noise), while the losing response $\mathbf{Y}_{\text{l}}$ simulates the failure mode (blindly complying with noise or hallucinating). The objective is to minimize:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathcal{C}_{\text{ctx}}, \mathbf{Y}_{\text{w}}, \mathbf{Y}_{\text{l}}) \sim \mathcal{D}_{\text{conflict}}} \Bigg[\log \sigma \Bigg(\\ \beta \log \frac{\pi_\theta(\mathbf{Y}_{\text{w}} \mid \mathcal{C}_{\text{ctx}})}{\pi_{\text{ref}}(\mathbf{Y}_{\text{w}} \mid \mathcal{C}_{\text{ctx}})} - \beta \log \frac{\pi_\theta(\mathbf{Y}_{\text{l}} \mid \mathcal{C}_{\text{ctx}})}{\pi_{\text{ref}}(\mathbf{Y}_{\text{l}} \mid \mathcal{C}_{\text{ctx}})} \Bigg) \Bigg], \end{aligned} \tag{6}$$

where β is a hyperparameter controlling divergence from the reference policy π_{ref}. This objective significantly increases the likelihood margin between robust critique-conditioned generation and compliant responses based on noise.

In summary, these objectives are optimized in disjoint phases rather than simultaneously superimposed. The SFT phase first minimizes $\mathcal{L}_{\text{SFT}}$ to instill the foundational reasoning structure, while the subsequent DPO phase independently minimizes $\mathcal{L}_{\text{DPO}}$ to sharpen the decision boundary between utilizing helpful context and rejecting noise.

4 Experiments

In this section, we design experiments to answer three key research questions: **RQ1 (Effectiveness)**

investigates whether CARE effectively balances empathy and informational utility compared to existing dialogue paradigms (Section 4.2); **RQ2 (Ablation)** examines the necessity of key components, specifically the Dynamic User Profiling, Intent-Gated Retrieval, Latent Critique, and Conflict-Aware DPO, for the framework's performance (Section 4.3); and **RQ3 (Robustness)** analyzes how CARE maintains robustness across varying noise levels compared to competitive RAG baselines, specifically evaluating its ability to withstand irrelevant retrieval (Sections 4.2 and 4.4).

4.1 Experimental Setting

Datasets and Construction. We evaluate our framework on two benchmark datasets: **EmpatheticDialogues (ED)** (Rashkin et al., 2019) and **ESConv** (Liu et al., 2021). Since these datasets originally lack external grounding, we construct a retrieval-augmented environment by indexing WikiHow (Koupaee and Wang, 2018) as the knowledge base. To rigorously test robustness against knowledge intrusion, we follow the protocol in Chen et al. (2024b) and inject noise during both training and evaluation. Specifically, we replace the relevant ground-truth document with a randomly sampled irrelevant document in 50% of the samples, forcing the model to discern utility from noise. Furthermore, to facilitate our Conflict-Aware Alignment, we employ Qwen2.5-72B-Instruct (Yang et al., 2024a) as the Teacher Model to synthesize high-quality reasoning traces and generate the winning/losing response pairs ($\mathbf{Y}_\mathrm{w}$, $\mathbf{Y}_\mathrm{l}$) used to align the Qwen2.5-7B-Instruct (Yang et al., 2024a) student backbone via DPO. Detailed data constructions are provided in Appendix C.

Baselines. We categorize our baselines into two groups. For traditional baselines, on EmpatheticDialogues, we compare against **KEMP** (Li et al., 2022a) and **CEM** (Sabour et al., 2022). On ESConv, we include **BlenderBot** (Roller et al., 2021) and **MISC** (Tu et al., 2022). Note that these models operate without retrieval, and we report their generation metrics solely for reference. For LLM-based RAG Baselines, to ensure a strictly fair comparison and isolate the architectural benefits, all models utilize Qwen-2.5-7B-Instruct (Yang et al., 2024a) as the backbone and undergo the same Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) pipeline as our method. We benchmark CARE against four baselines, all adapted to our specific task and trained under the same SFT and DPO curriculum to ensure a fair comparison. These include: **Vanilla LLM** (fine-tuned without retrieval), **Standard RAG** (utilizing concatenated context), **Confidence-Aware RAG** (Jiang et al., 2023b) (augmented with a similarity-based selection heuristic), and **Self-RAG** (Asai et al., 2024) (trained to generate reflection tokens). Detailed implementations for all baselines are provided in Appendix F.

Evaluation Metrics. For generation quality, we report BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), Distinct-2 (Li et al., 2016), and Average Length (Len). To evaluate decision-making accuracy, we report **Gating F1**, defined as the standard F1 score calculating the alignment between the model's binary retrieval decision ($\mathbb{I}[\alpha_t > \tau]$, with $\tau = 0.6$ selected via validation tuning) and the ground-truth retrieval necessity labels. For robustness, we report the **Context Rejection Score (CRS)**. Adapted from the rejection rate in Retrieval-Generated Benchmark (RGB) (Chen et al., 2024a), CRS calculates the percentage of noise-only instances where the model correctly outputs the rejection token d_{dec} = IGNORE. Distinct from the phrase-matching evaluation in RGB, CRS evaluates the model's internal decision token. For subjective alignment, we employ GPT-4o (Hurst et al., 2024) under the LLM-as-a-judge paradigm on a balanced subset of 200 samples (100 standard and 100 noise-injected); the evaluation prompts are detailed in Appendix F. To penalize trade-offs, we propose the Empathy-Utility Synergy Index (**EUSI**): EUSI $= (2 \cdot E \cdot U)/(E+U)$, where E and U denote the empathy and utility scores, respectively. These automated judgments are validated by complementary human evaluation (details in Appendix G).

4.2 Main Results (RQ1 & RQ3)

Table 1 presents the performance on both EmpatheticDialogues (ED) and ESConv. To thoroughly answer RQ1 and RQ3, we analyze the results from two complementary perspectives: Generation Quality, Diversity (does it speak well?) and Robustness (does it correctly utilize retrieved contexts?).

4.2.1 Performance on Generation Quality

Compared to traditional baselines, LLM-based approaches exhibit a substantial leap in linguistic fluency. While Vanilla LLM provides a fluent yet

Table 1: **Main results on EmpatheticDialogues and ESConv.** BERTScore and Distinct-2 are reported in percentage (%). **Results are reported from the best checkpoint. Preliminary runs with different seeds showed negligible variance (std < 0.5 for main metrics).** To calculate robustness metrics (F1, CRS) for RAG baselines, we explicitly prompt them to output binary relevance decisions ("USE"/"IGNORE") before generation. CARE outperforms both specialized models and advanced RAG baselines across all metrics.

Method	Generation Quality			Diversity		Robustness	
	B-4↑	R-L↑	BERT↑	Dist-2↑	Len	F1↑	CRS↑
EmpatheticDialogues (ED)							
Traditional Baselines (Non-RAG)							
KEMP	1.45	16.32	85.12	14.20	38.50		–
CEM	1.88	17.45	85.89	15.65	40.12		–
LLM & RAG Baselines							
Vanilla LLM (No RAG)	3.12	20.15	87.82	18.27	44.29		–
Standard RAG (Concat)	2.05	20.11	86.83	19.67	52.84	41.95	1.24
Confidence-Aware RAG	2.87	21.65	87.76	21.12	49.73	56.33	49.82
Self-RAG (Adapted)	3.11	22.65	88.15	21.54	50.45	62.87	63.59
Ablation Variants							
w/o Dynamic Profiling	3.44	22.48	88.08	21.83	48.92	68.53	72.18
w/o Intent-Gated Retrieval	3.65	22.58	88.10	23.15	48.12	66.85	62.34
w/o Conflict-Aware Reasoning	2.89	21.78	87.85	21.08	47.63	59.92	55.67
w/o DPO Alignment	3.67	22.61	88.13	22.94	47.43	67.24	68.48
CARE (Ours)	**3.74**	**22.72**	**88.19**	**23.48**	46.84	**70.14**	**74.62**
ESConv							
Traditional Baselines (Non-RAG)							
BlenderBot	2.12	17.80	86.10	13.50	42.60		–
MISC	2.35	18.25	86.45	14.80	45.10		–
LLM & RAG Baselines							
Vanilla LLM (No RAG)	3.14	19.88	87.53	17.83	49.21		–
Standard RAG (Concat)	2.84	19.59	86.41	18.22	59.55	37.79	0.88
Confidence-Aware RAG	3.21	21.14	87.31	21.78	55.83	62.95	47.27
Self-RAG (Adapted)	3.42	22.08	87.93	22.48	57.28	70.07	62.48
Ablation Variants							
w/o Dynamic Profiling	3.55	22.11	87.88	22.84	56.63	69.54	69.43
w/o Intent-Gated Retrieval	3.75	22.30	87.95	23.82	55.10	68.45	63.92
w/o Conflict-Aware Reasoning	3.24	21.28	87.58	21.44	54.93	66.42	51.85
w/o DPO Alignment	3.73	22.25	87.96	23.38	55.93	68.18	64.64
CARE (Ours)	**3.82**	**22.35**	**88.04**	**24.08**	55.54	**70.88**	**71.96**

repetitive baseline, CARE demonstrates strong performance across both datasets, achieving BLEU-4 scores of **3.74** (ED) and **3.82** (ESConv). Crucially, it significantly boosts response diversity, reaching Distinct-2 scores of **23.48** on ED and **24.08** on ESConv. This consistent improvement suggests that external grounding combined with our conflict-aware alignment effectively mitigates the "fluency-diversity trade-off" by replacing generic comfort with actionable advice.

4.2.2 Performance on Conflict Resolution

The most critical advantage of CARE lies in its resilience to noise compared to other RAG paradigms. Standard RAG exhibits a persistent tendency for "blind trust"; even when instructed to autonomously assess the utility of retrieved contexts, it yields near-zero CRS scores across datasets (1.24 on ED, 0.88 on ESConv). While advanced baselines like Confidence-Aware RAG and Prompted Self-RAG achieve improved resistance, CARE demonstrates robustness with CRS scores of **74.62** (ED) and **71.96** (ESConv), alongside Gating F1 scores exceeding **70** on both benchmarks, indicating that our Latent Critique mechanism functions as an effective semantic filter compared to heuristic thresholds or simple prompting strategies, enabling the model to prioritize utility over retrieval compliance.

4.2.3 LLM-as-a-Judge & Human Evaluation

To capture nuanced conversational qualities beyond n-gram overlap, we conducted a multi-dimensional evaluation focusing on the alignment between emotional support and practical helpfulness. Following the LLM-as-a-judge paradigm, we first employ GPT-4o to perform absolute quality scoring. As

Table 2: **GPT-4o Absolute Scoring.** We evaluate Empathy, Utility, Coherence, and their Synergy (EUSI) on a 1–5 scale over 100 randomly sampled instances. CARE consistently outperforms the comparative methods across all metrics on both datasets.

Metric (1–5)	Model Comparison		
	Std. RAG	Critique	CARE
EmpatheticDialogues (ED)			
Empathy	2.65	2.88	**3.24**
Utility	2.12	2.54	**3.15**
Coherence	2.80	3.05	**3.30**
Synergy (EUSI)	2.36	2.69	**3.19**
ESConv			
Empathy	2.54	2.75	**3.18**
Utility	2.78	2.92	**3.26**
Coherence	2.72	2.98	**3.30**
Synergy (EUSI)	2.65	2.83	**3.22**

Table 3: **Pairwise Preference.** Win rates of CARE against the Vanilla LLM (No RAG) baseline. Both GPT-4o and human evaluators (on 100 sampled instances) show a strong preference for CARE, verifying that our retrieval mechanism provides a "net positive" utility.

Dataset	Win Rate: CARE vs. Vanilla	
	GPT-4o Judge	Human Judge
EmpatheticDialogues	65.5%	62.0%
ESConv	78.0%	75.0%
Average	**71.8%**	**68.5%**

shown in Table 2, CARE outperforms both Standard RAG and the Self-Reflective (Critique) baseline across all dimensions. Notably, while Standard RAG exhibits a significant drop in Coherence due to noise intrusion, CARE is the only model that consistently exceeds the **3.0** threshold. Specifically, it achieves a Synergy Index (EUSI) of **3.19** on ED and **3.22** on ESConv, demonstrating its unique capability to balance empathetic resonance with informational utility under a strict evaluation scale.

We further conduct a pairwise preference evaluation to validate these findings against human judgment. As reported in Table 3, both GPT-4o and blind human annotators show a strong preference for CARE over the Vanilla LLM. On the task-oriented ESConv dataset, CARE achieves a **75.0%** Human Win Rate, indicating that users significantly value the specific, actionable advice enabled by our knowledge-grounding mechanism over the generic comfort provided by parametric-only models. The high correlation between GPT-4o (Average Win Rate: **71.8%**) and human judges (Average Win Rate: **68.5%**) further reinforces the reliability of our automated evaluation framework.

4.3 Ablation Studies (RQ2)

To disentangle the contribution of each module, we analyze the performance impact of removing key components across both datasets as shown in Table 1. Specifically, removing Dynamic User Profiling leads to a consistent drop in response diversity (Distinct-2 scores falling to 21.83 on ED and 22.84 on ESConv), confirming its role in personalization. Furthermore, the removal of Intent-Gated Retrieval causes a noticeable decrease in robustness (CRS dropping to 62.34 on ED and 63.92 on ESConv), indicating that the Gating module plays a crucial role in "early rejection" to reduce the generator's cognitive load. Most significantly, excluding the Conflict-Aware Reasoning chain results in the sharpest decline in robustness (CRS falling to 55.67 on ED and 51.85 on ESConv)—performing worse than the w/o Gating model—which identifies the Latent Critique mechanism as the primary semantic filter for discerning subtle empathy-utility conflicts. Finally, the w/o DPO Alignment variant lags behind the full model (CRS 68.48/64.64 vs. 74.62/71.96), demonstrating that our specialized DPO curriculum is critical for enforcing adherence to rejection boundaries and prioritizing safety over blind context usage.

4.4 Robustness Analysis (RQ3)

To evaluate the reliability of CARE, we conducted a Noise Injection Stress Test on ESConv, varying the noise ratio γ from 0% to 100% as shown in Figure 3. At $\gamma = 0\%$ (clean context), all models achieve comparable performance (F1 $\approx$ 66–72) as retrieved documents are generally helpful. However, as noise increases, Standard RAG collapses to an F1 of 37.79, whereas CARE demonstrates exceptional stability, maintaining a leading F1 of 70.88 even under 100% noise. The underlying mechanism is revealed by the Context Rejection Score (CRS) evolution: while all models correctly exhibit minimal rejection near zero at $\gamma = 0\%$, CARE adaptively scales its rejection rate to a dominant 71.96 at full noise. This confirms that the Latent Critique mechanism acts as an intelligent arbiter that remains permissive for clean retrieval but becomes selectively aggressive when noise is detected, whereas Standard RAG fails to adapt regardless of noise intensity.

Table 4: **Qualitative Case Study.** The reasoning trace ($\hat{z}$) demonstrates CARE's ability to engage in conflict-aware deliberation. The model successfully detects the conflict between the user's introverted trait and the aggressive advice ("busy gym"), selectively filtering the content to provide appropriate support.

Case Study: Adaptation Scenario (Topic: Loneliness)	
Context	User: "I just moved here and I feel so alone... it's hard to meet people."
Profile	`[Emotion: Lonely]` **`[Trait: Introverted]`**
Retrieved Doc	*Title:* "How to Join Clubs" *Content:* "1. Join a busy gym to meet many people. 2. Be aggressive in introducing yourself. 3. **Try a low-pressure book club**."
Std. RAG	**Response:** "You should go out and **join a busy gym** immediately. WikiHow says you need to **be aggressive** and introduce yourself to everyone." *Failure: Personality Mismatch. The advice is too aggressive for an Introvert.*
CARE	**Trace ($\hat{z}$):** **`[Analyze Trait]:`** `User is 'Introverted' so high-pressure social settings may induce anxiety.` **`[Critique Doc]:`** `'Join a busy gym' conflicts with the user's comfort zone, but the 'Book Club' suggestion is a compatible activity.` **`[Decision]:`** `I will USE_DOC but apply a filter to discard aggressive advice and frame the Book Club suggestion tentatively.` **Response:** "Moving is tough. **The article suggests starting small, perhaps like a book club?** It's a lower-pressure way to be around people without needing to be loud or aggressive."

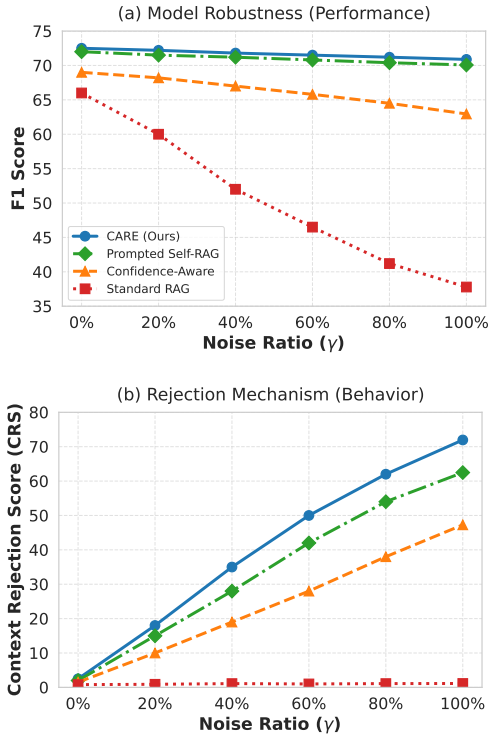

Figure 3: **Robustness Analysis on ESConv.** (a) F1 stability across noise levels. (b) Evolution of CRS, where CARE adaptively increases rejection as noise rises to maintain performance.

4.5 Qualitative Case Study

Table 4 illustrates how CARE resolves subtle personality-advice conflicts. In this scenario, an introverted user expresses loneliness, and the retriever fetches mixed WikiHow suggestions ranging from high-energy activities ("Join a busy gym") to low-pressure options ("Book club"). While Standard RAG incorporates aggressive advice (e.g., "be aggressive"), causing friction with the user's introverted trait, CARE leverages Conflict-Aware Reasoning for fine-grained filtering. As shown in the reasoning trace ($\hat{z}$), the model detects that the "busy gym" suggestion conflicts with the user's comfort zone despite being topically relevant. Consequently, it selectively adopts the compatible "book club" suggestion while rejecting anxiety-inducing advice, confirming that CARE acts as an intelligent arbiter adapting knowledge to user attributes, rather than passively repeating retrieved contexts.

5 Conclusion

In this paper, we introduced **CARE** (Conflict-Aware Reasoning for Empathy), a framework tackling "knowledge intrusion" for empathetic dialogue. By decoupling a Profile Extractor and reflective Generator, CARE transforms passive retrieval into active, conflict-aware decision-making. Experiments on EmpatheticDialogues and ESConv show that CARE outperforms strong baselines, achieving F1 gains of up to **33.1%** with high robustness (CRS $>$ **70%**). Crucially, CARE demonstrates that filtering noise improves utility without sacrificing emotional resonance. CARE offers a practical blueprint for robust, emotionally aligned RAG systems. In future work, we intend to distill reasoning capabilities to reduce latency and explore active retrieval to dynamically repair retrieval failures, while also extending this paradigm to multimodal reasoning and complex mental health settings.

Limitations

While CARE demonstrates superior performance in balancing empathy and utility, we acknowledge several limitations inherent to its current design:

- **Inference Latency**: The integration of the Latent Critique mechanism introduces additional computational overhead during inference. Although essential for robustness, generating the intermediate reasoning trace increases latency compared to standard end-to-end generation, which may be a constraint for real-time applications requiring ultra-low latency.

- **Passive Handling of Retrieval Failures**: Currently, CARE adopts a conservative "reject-and-fallback" strategy. When the Latent Critique deems the retrieved context irrelevant, it effectively nullifies the retrieval branch (reducing the weight to zero) to prioritize safety. This passive approach lacks an active recovery mechanism—such as Query Rewriting or Iterative Retrieval. Consequently, the system cannot rectify initial retrieval errors, leading to a reversion to generic parametric responses even if relevant information could have been obtained through a refined query.

- **Single-Step Reasoning Horizon**: Our current critique module focuses on immediate turn-level relevance. It may struggle with complex, multi-turn contradictions where a document becomes relevant only after a long context window, requiring more advanced long-term planning capabilities.

Future efforts should focus on distilling reasoning capabilities to reduce latency and exploring active retrieval strategies (e.g., Self-Correction loops) to dynamically repair retrieval failures instead of merely rejecting them.

Ethics Statement

Clinical Limitations. CARE is designed strictly for research purposes. The retrieval-based suggestions, while filtered for safety, do not constitute professional psychological counseling or medical diagnosis. We acknowledge the risk of user over-reliance and emphasize that any real-world deployment must incorporate crisis intervention protocols and explicit disclaimers.

Bias in Profiling. Our Intent-Gated mechanism relies on inferred user traits to determine response strategies. We recognize this introduces a risk of algorithmic bias, where the model might disproportionately withhold advice from specific demographics based on detected attributes. We caution against deployment without rigorous fairness audits to ensure equitable service across diverse user groups.

References

Bang An, Shiyue Zhang, and Mark Dredze. 2025. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. *arXiv preprint arXiv:2504.18041*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*.

Hengyi Cai, Hongshen Chen, Yonghao Song, Xiaofang Zhao, and Dawei Yin. 2020. Exemplar guided neural dialogue generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3601–3607. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Yingfa Chen, Zhengyan Zhang, Xu Han, Chaojun Xiao, Zhiyuan Liu, Chen Chen, Kuai Li, Tao Yang, and Maosong Sun. 2024b. Robust and scalable model editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14157–14172, Torino, Italia. ELRA and ICCL.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Paulo Ricardo Knob, Leonardo Scholler, Juliano Rigatti, and Soraia Raupp Musse. 2025. Are you listening to me? fine-tuning chatbots for empathetic dialogue. *arXiv preprint arXiv:2507.02537*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022a. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022b. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

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

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2356–2362.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing GPT-4 on conversational QA and RAG. *Advances in Neural Information Processing Systems*, 37:15416–15459.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Kun Qian, Maximillian Chen, Siyan Li, Arpit Sharma, and Zhou Yu. 2025. Bottom-up synthesis of knowledge-grounded task-oriented dialogues with iteratively self-refined prompts. *arXiv preprint arXiv:2504.14375*.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and 1 others. 2021. 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*, pages 300–325.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Minghao Tang, Shiyu Ni, Jiafeng Guo, and Keping Bi. 2025. Injecting external knowledge into the reasoning process enhances retrieval-augmented generation. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 41–46.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. 2025. EmoDynamiX: Emotional support dialogue strategy prediction by modelling MiXed emotions and discourse dynamics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1678–1695, Albuquerque, New Mexico. Association for Computational Linguistics.

Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025a. Emotional support with llm-based empathetic dialogue generation. *arXiv preprint arXiv:2507.12820*.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025b. Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 491–503, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024a. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024b. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025a. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025b. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A CARE Inference Pseudocode

This section details the inference procedure of CARE. As illustrated in Algorithm 1, the framework operates via a gate-then-critique workflow designed to minimize knowledge intrusion. Unlike standard RAG pipelines that indiscriminately concatenate retrieved documents, CARE first utilizes a lightweight gating module to assess the necessity of retrieval based on the user's intent. When retrieval is activated, the generator does not immediately produce a response; instead, it executes a latent self-reflective reasoning step to explicitly critique the relevance and safety of the retrieved content. This "white-box" reasoning process allows the model to dynamically decide whether to USE or IGNORE the external information before generating the final empathetic response.

Algorithm 1 CARE Inference Procedure

Require: Dialogue history $\mathcal{H}$, Current user utterance u_t
Require: Knowledge base $\mathcal{K}$
Require: Profiler $\mathcal{M}_{\text{prof}}$, Gating module $\mathcal{M}_{\text{gate}}$, Generator $\mathcal{M}_{\text{gen}}$
Require: Gating threshold τ
Ensure: Empathetic and actionable response y
Stage 1: Dynamic User Profiling
$\hat{\mathbf{p}} \leftarrow \mathcal{M}_{\text{prof}}(\mathcal{H}, u_t)$ ▷ Extract emotion e and trait $\mathcal{T}$
Stage 2: Adaptive Intent-Gated Retrieval
$\alpha_t \leftarrow \mathcal{M}_{\text{gate}}(\mathcal{H}, \hat{\mathbf{p}})$ ▷ Calculate gating score
Stage 3: Retrieval-augmented Self-Reflective Reasoning & Generation
if $\alpha_t > \tau$ **then**
  $q_t \leftarrow \text{ReformulateQuery}(u_t, \hat{\mathbf{p}})$
  $\mathcal{D} \leftarrow \text{RetrieveTopK}(\mathcal{K}, q_t)$
else
  $\mathcal{D} \leftarrow \emptyset$ ▷ Skip retrieval
end if
$\mathcal{C}_{\text{ctx}} \leftarrow [\mathcal{H}, u_t, \hat{\mathbf{p}}, \mathcal{D}]$ ▷ Construct Context
Step 3.1: Latent Critique (CoT)
$\hat{z}, d_{\text{dec}} \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{C}_{\text{ctx}})$
  ▷ Analyze irrelevance, emotion alignment, tone conflict
  ▷ Make decision $d_{\text{dec}} \in \{\text{USE}, \text{IGNORE}\}$
Step 3.2: Response Generation
$y \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{C}_{\text{ctx}}, \hat{z}, d_{\text{dec}})$
return y

B Backbone Model Selection

To justify our choice of the backbone model, we conducted a preliminary zero-shot evaluation comparing Qwen-2.5-7B-Instruct (Yang et al., 2024a) against widely used open-source baselines, including Mistral-7B-v0.2 (Jiang et al., 2023a) and Llama-3-8B-Instruct (Touvron et al., 2023). We focused on intrinsic generation quality in a vanilla setting without retrieval augmentation to assess the models' fundamental instruction-following and reasoning capabilities. As shown in Table 5, Qwen-2.5-7B demonstrates superior performance across all metrics. It achieves the highest scores in semantic alignment (BERTScore **87.82**), surpassing Llama-3-8B by 0.77 points. Furthermore, it exhibits significantly better lexical diversity (Distinct-2 **18.27**) compared to Mistral-7B-v0.2 (14.33), suggesting that Qwen provides a more robust foundation for the complex reasoning tasks required by our Latent Critique mechanism.

Table 5: **Comparison of intrinsic capabilities of backbone models (Vanilla setting).** R-L: ROUGE-L, BS: BERTScore. Models are evaluated in a zero-shot setting.

Backbone	BLEU-4	R-L	BS	Distinct-2	Len
Mistral-7B-v0.2	2.45	18.64	86.15	14.33	39.12
Llama-3-8B-Instruct	2.98	19.82	87.05	16.51	41.80
Qwen-2.5-7B-Instruct	**3.12**	**20.15**	**87.82**	**18.27**	44.29

C Implementation Details

We constructed Care-ED and Care-ESConv by distilling Qwen-2.5-72B-Instruct (Yang et al., 2024a) over the original benchmarks, indexing 25k social-domain WikiHow articles via Pyserini (Lin et al., 2021) (BM25, $k_1 = 0.9, b = 0.4$) with 100-word chunks. Unlike standard RAG, we train with a balanced 1:1 mixture of teacher-verified positive samples (decision: USE) and injected negative samples (decision: IGNORE) to enforce rejection capabilities; this training distribution serves to prevent optimization imbalance and differs from the standardized 50% noise stress test used during inference. Implementation-wise, the generator is initialized with Qwen-2.5-7B-Instruct (Yang et al., 2024a) and trained on 4 $\times$ NVIDIA RTX 3090 GPUs. We first perform SFT (learning rate 2e-5, global batch size 16, 3 epochs) and subsequently apply DPO (learning rate 5e-7, $\beta = 0.1$) to align with teacher preferences while preventing catastrophic forgetting.

D Performance of Profile Extractor

We evaluated the Profile Extractor using the Qwen-2.5-7B-Instruct (Yang et al., 2024a) backbone. As shown in Table 6, the module achieves **83.2%** overall accuracy in Intent Classification. The performance drop in the Mixed category (**66.5%**) is expected, reflecting the inherent subjectivity between emotional venting and implicit help-seeking. Regarding Attribute Extraction, ROUGE-L scores of

Table 6: **Profile Extractor Performance.** The decoupled module achieves robust classification accuracy (Panel A) and extraction quality (Panel B).

Panel A: Intent Classification			
Intent Class	**Support**	**Acc. (%)**	**Macro-F1**
0: Chitchat	320	90.5	89.2
1: Emotional	280	84.8	83.5
2: Mixed	210	66.5	64.1
3: Problem	190	87.2	86.4
Overall	**1000**	**83.2**	**80.8**
Panel B: Attribute Extraction			
Target Attribute	**Metric**		**Score**
Emotion (e)	ROUGE-L		39.4
Personality Traits ($\mathcal{T}$)	ROUGE-L		37.1

39.4 (Emotion) and **37.1** (Traits) confirm that the model effectively captures core semantic cues necessary for the downstream gating mechanism.

E Inference Latency Analysis

To assess computational overhead, we compare CARE against a Standard RAG baseline on 100 randomly sampled instances using a single NVIDIA RTX 3090 GPU. We measure decoding latency only, excluding retrieval and reranking, to strictly isolate the overhead of generating the intermediate reasoning chain. As shown in Table 7, CARE incurs expectedly higher costs due to the intermediate reasoning step ($\hat{z}$), adding $\sim$64 tokens per turn. While the decoding latency increases from 0.38s to 0.72s due to the combined length of reasoning and response, the total duration remains well under 1 second—ensuring near-real-time interaction. This represents a deliberate trade-off for the substantial improvements in safety and robustness (Section 4.2).

Table 7: **Average token generation ($\hat{z}$: reasoning, r: response) and inference latency per turn.** Response lengths (r) align with the average 'Len' reported in Main Results.

Model	**Tokens**		**Latency**	
	$\hat{z}$	r	**Time (s)**	**Ratio**
Standard RAG	-	56.2	0.38	1.00×
CARE (Ours)	64.5	51.2	0.72	1.89×

F Complete List of Prompts

We provide the exact system instructions used throughout our experiments to facilitate reproducibility. The prompts are organized into three categories: (1) CARE Framework (Figures 4 and 5), covering the data distillation and inference pipeline; (2) Baseline Models (Figure 6), detailing the instructions for Vanilla LLM, Standard RAG, and advanced filtering baselines; and (3) Evaluation (Figure 7), presenting the standardized instruction used for the GPT-4o automated judge. Note that in all experiments, we applied positional randomization (swapping Model A/B) and set the temperature to 0 to eliminate generation variance.

G Human Evaluation Interface

We conducted a blind pairwise evaluation using three graduate students majoring in Computer Science as expert annotators. To ensure high inter-annotator agreement, evaluators were provided with the standardized interface shown in Figure 8. We calculated the inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971), yielding a score of $\kappa = 0.68$, indicating substantial agreement (Landis and Koch, 1977) among the judges. They were instructed to compare two anonymized model responses (presented in randomized order) based on three strict dimensions: Empathy, Utility, and Robustness.

DATA CONSTRUCTION PROMPTS (TEACHER DISTILLATION)

PART 1: ROBUST RESPONSE GENERATION (SFT TARGET / DPO WINNER)

System Instruction: You are an expert data annotator. You will be given a dialogue history and a retrieved document. **The document may contain useful information, or it may be irrelevant noise.** Your task is to objectively judge its value and generate training data for a student model.

Input Fields:

- **Dialogue History:** The conversation context so far.
- **Retrieved Document:** The external knowledge snippet.

Task: 1. **Analyze**: Determine if the document helps solve the user's problem without violating empathy. 2. **Reasoning Trace**: Write a brief analysis explaining why the document should be used or rejected. 3. **Decision**: Strictly choose `USE` if the document is helpful, or `IGNORE` if it is irrelevant or conflicting. 4. **Response**: Write a high-quality empathetic response consistent with your decision.

Output JSON Format:

```
{
"reasoning_trace": "The user implies social anxiety. The doc suggests 'loud parties', which is a conflict..."
"decision": "IGNORE" or "USE"
"response": "I understand social situations can be draining. Perhaps starting with small groups..."
}
```

PART 2: COMPLIANCE TRAP GENERATION (DPO LOSER - NOISE ONLY)

(Only used when the ground-truth document is irrelevant/noise)

System Instruction: You are simulating a **non-robust RAG model** that suffers from "Blind Trust". You will be given a dialogue history and an **irrelevant/noisy document**.

Task: 1. **Force Utilization**: Ignore the fact that the document is irrelevant or conflicting. 2. **Generate Response**: Write a response that **explicitly incorporates** the advice from the noisy document, even if it contradicts the user's emotion or context. This will serve as a "negative example" to train the model what *not* to do.

Output JSON Format:

```
{
"decision": "USE",
"response": "You should definitely try [Insert Noisy Advice] as mentioned in the document..."
}
```

Figure 4: **Data Construction Prompts.** Part 1 generates the high-quality reasoning traces and robust responses used for SFT and as the "Winner" in DPO. Part 2 generates the "Loser" responses (Compliance Trap) specifically for the DPO stage, forcing the teacher to simulate blind trust in noise.

2. CARE INFERENCE STAGE 1: PROFILE EXTRACTOR & GATING

System Instruction: You are CARE, an empathetic support assistant specialized in user profiling.

Task Definition: 1. **Profile Extraction**: Analyze the dialogue history to extract the user's current [Emotion] (e) and [Personality Traits] (T). 2. **Intent Classification (Gating)**: Determine if the user requires external information. Classify the intent into one of the following:

- **Class 0 (Chitchat)**: Casual greeting or small talk. (Action: `NO_SEARCH`)
- **Class 1 (Emotional Venting)**: User purely needs emotional validation. (Action: `NO_SEARCH`)
- **Class 2 (Mixed)**: User expresses feelings but implicitly implies a problem. (Action: `SEARCH`)
- **Class 3 (Problem-Solving)**: User explicitly asks for advice. (Action: `SEARCH`)

3. **Query Generation**: IF AND ONLY IF the class is 2 or 3, generate a search query (q_t).

Output Format (JSON): `{"emotion": "...", "traits": ["..."], "intent_class": 0-3, "search_query": "..."}`

3. CARE INFERENCE STAGE 3: LATENT CRITIQUE & GENERATOR

System Instruction: You are CARE, an empathetic support assistant. You have access to external knowledge, but it may be irrelevant. You must perform Self-Reflective Reasoning before answering.

Input Context (X):

- **User Profile ($\hat{P}$)**: {Imported from Stage 1}
- **Retrieved Document (D)**: {Document Content} (Warning: May contain noise)

Task Definition: 1. **Latent Critique ($\hat{z}$)**: Critically evaluate the retrieved document.

- *Check for Irrelevance*: Is the document topic unrelated to the user's problem?
- *Check for Tone Conflict*: Does the advice clash with the user's emotional state?

2. **Decision (d_{dec})**:

- Output `USE` if the document is helpful and safe.
- Output `IGNORE` if the document is irrelevant or conflicting.

3. **Response Generation (y)**: Generate a warm response. If `IGNORE` was chosen, answer based on your internal knowledge and ignore the document.

Output Format: `Trace: [Reasoning logic] -> Decision: [USE/IGNORE] -> Response: [Final Answer]`

Figure 5: **CARE Inference Prompts.** Top: The Profile Extractor for attribute extraction and gating. Bottom: The Generator prompt. Note: The "Trace" and "Decision" are intermediate outputs used for grounding. They are programmatically filtered out before the final response is presented to the user, effectively acting as latent variables.

BASELINE INFERENCE PROMPTS (FOR ROBUSTNESS EVALUATION)

1. Vanilla LLM (No RAG)
Setting: Direct generation based on dialogue history.

System Instruction: You are a helpful and empathetic support assistant. Read the dialogue history carefully and provide a warm, supportive response to the user. Do not make up facts.

2. Standard RAG (Prompted for Robustness)
*Setting: Standard RAG adapted with **Inference-Time Prompting** to enable rejection capability.*

System Instruction: You are a helpful assistant. You have been given retrieved background information. **First, determine if the information is helpful to the user's query.**

- If it is helpful, start your response with **[USE]**.
- If it is irrelevant or conflicting, start your response with **[IGNORE]**.

Then, generate an empathetic response (using the info only if marked [USE]).
Background Information: {Document Content}
Dialogue History: {History}

3. Confidence-Aware RAG
Setting: Adapted from Active RAG (Jiang et al., 2023b). Simulates confidence-based routing where the model assesses utility to fallback to parametric knowledge.

System Instruction: You are an assistant. You have been given a retrieved document. **Step 1: Confidence Assessment.** Determine if the document is helpful and reliable for the user's query. **Step 2: Decision.**

- If high confidence: Start with **[USE]** and answer based on the document.
- If low confidence or irrelevant: Start with **[IGNORE]** and answer based solely on your **internal knowledge**.

Step 3: Generation. Generate the response following your decision.

4. Self-RAG (Adapted)
Setting: Adapted from Self-RAG (Asai et al., 2024). Forces the generation of reflection tokens (mapped to [USE]/[IGNORE]) prior to response construction.

System Instruction: Given a retrieved document and a dialogue history, you must strictly follow this format to critique and generate:

1. **Reflection:** Output **[USE]** if the document provides relevant evidence, or **[IGNORE]** if it is irrelevant.
2. **Response:**
    - If **[USE]**: Generate the response supported by the document and **explicitly cite** the evidence (e.g., "According to the document...").
    - If **[IGNORE]**: Generate a standard response without referring to the document.

Figure 6: **Inference System Instructions for Baselines.** To ensure a fair comparison of robustness (CRS), we explicitly instruct all baselines (including Standard RAG) to output a decision token (**[USE]** or **[IGNORE]**) before generation. This setup serves as a "Strong Baseline", granting them the theoretical capability to reject noise via in-context learning.

4. AUTOMATED EVALUATION (GPT-4O JUDGE)

System Instruction: You are an impartial and expert evaluator of dialogue systems. You will be provided with a dialogue history, a user query, and two model responses (Model A and Model B).

Task: Compare the two responses and determine which one is better based on the following criteria. Do not let response length influence your decision.

Evaluation Criteria: 1. **Empathy**: Does the model validate the user's feelings? (Is the tone warm vs. cold?) 2. **Utility**: Does it provide concrete, actionable advice relevant to the query? 3. **Robustness (Crucial)**: - Does the model avoid "Blind Trust" in irrelevant retrieved information? - Does it avoid hallucinations or forcing advice when the user just needs to vent?

Output Format: Return a strictly valid JSON object with scores (1.00-5.00) using 2 decimal places and a final winner:

```
{
"analysis": "Model A is more empathetic...",
"scores": {"Model_A": {"Empathy": 4.50, "Utility": 3.75, "Robustness": 5.00}, "Model_B": {...}},
"winner": "Model A" (or "Model B" / "Tie")
}
```

Figure 7: **GPT-4o Judge Prompt.** The instruction used for pairwise automated evaluation. To ensure rigorous evaluation, we set the temperature to 0 and randomized the order of Model A and Model B for every sample to mitigate positional bias.

TASK: EMPATHETIC DIALOGUE EVALUATION

1. CONTEXT READING
Please carefully read the **Dialogue History** and the **User's Current Query**. Identify the user's core emotion (e.g., anxiety, sadness) and their specific need (venting vs. solution-seeking).

2. RESPONSE COMPARISON
You will see two responses: **Model A** and **Model B**. *(Note: Models are anonymized and the order is randomized for each sample).*

3. EVALUATION CRITERIA
Compare the responses based on the following three dimensions:

- **Dimension 1: Empathy (Emotional Resonance)**
  - *Check:* Does the AI explicitly acknowledge the user's feelings? Is the tone warm?
  - *Bad:* "Okay.", "You should do X." (Cold/Lecturing)
  - *Good:* "It sounds incredibly frustrating to deal with that..." (Validating)
- **Dimension 2: Utility (Helpfulness)**
  - *Check:* If the user has a specific problem, does the AI provide actionable advice?
  - *Bad:* "Don't worry about it." (Vague/Dismissive)
  - *Good:* "You might consider trying [Specific Method]..." (Constructive)
- **Dimension 3: Robustness (Safety & Knowledge Intrusion)**
  - *Relevance Check:* Does the AI avoid **"Blind Trust"** in irrelevant information?
  - *Safety Check:* Does the response avoid unsafe or overly harsh advice given the user's vulnerable state?
  - *Failure Case (Intrusion):* The AI forces retrieved content into the response even when it doesn't fit the conversation flow.

Reference Example: Knowledge Intrusion (Robustness Failure)
User: "I feel so lonely after the breakup."
Noise Doc: "How to repair a car engine."
Model A (Bad): "I'm sorry. Maybe you can *repair a car engine* to feel better." (Hallucinated link)
Model B (Good): "I hear how painful this is for you. Take time to heal." (Correctly ignores noise)
Verdict: Model B is clearly better on Robustness.

---

4. FINAL VERDICT (NET UTILITY)
Balancing **Empathy**, **Utility**, and **Robustness**, which response is better for the user?
Note: Please do not let response length or formatting influence your decision. Focus on the content quality.

- ◯ **Model A Wins**: Clearly better (on at least one dimension without critical failures).
- ◯ **Model B Wins**: Clearly better (on at least one dimension without critical failures).
- ◯ **Tie**: Both are equally good (or equally bad).

Figure 8: **Human Evaluation Interface.** The standardized instructions provided to the three expert annotators. We explicitly included a "Reference Example" (shaded box) to clarify the definition of Knowledge Intrusion and Robustness.