EMPATHYSCORE: A Reference-free Metric for Emotional Support Conversation via Knowledge Distillation

Anonymous ACL submission

Abstract

Existing metrics for Emotional Support Conversation (ESC) struggle with the "one-to-many" nature of open-ended empathy or suffer from the high computational cost of LLM-based evaluation. To address this, we introduce EMPATHYSCORE, a reference-free, lightweight metric developed via the DISTILL-ES framework. Crucially, we construct EMPATHY-EVAL, a comprehensive distillation dataset containing fine-grained teacher annotations and pairwise preferences derived from rubric-guided LLM prompting. Leveraging this data, we distill the sophisticated reasoning of a teacher LLM into compact student scorers using a hybrid regression-ranking objective, decoupling evaluation into *Cognitive Suitability* and *Affective Resonance*. Experiments demonstrate that EMPATHYSCORE achieves state-of-the-art correlation with human judgments while being orders of magnitude more efficient than LLM judges.

1 Introduction

Text-based Emotional Support Conversations (ESC) have evolved into a critical medium for scalable psycho-social care, ranging from peer-support platforms to empathetic AI companions (Rashkin et al., 2019; Liu et al., 2021). Unlike task-oriented dialogue, effective ESC requires the helper to accompany the seeker through distress via a nuanced psychological process. Counselling theory characterizes this process as a multidimensional construct: it demands both *cognitive empathy* (understanding the seeker's frame of reference to select the right support strategy) and *affective empathy* (responding with emotional attunement and warmth) (Rogers, 1957a; Hill, 2014; Elliott et al., 2011). Consequently, evaluating ESC systems requires metrics that go beyond surface-level fluency to assess whether responses are *cognitively suitable* for the conversation stage and *affectively resonant* with the user's emotional state.

However, existing evaluation methodologies struggle to quantify these nuances. Standard reference-based metrics (e.g., BLEU, BERTScore), inherited from translation tasks, rely on computing similarity against a single human reference (Papineni et al., 2002; Zhang et al., 2020). This paradigm is fundamentally flawed in the context of ESC due to the "one-to-many" nature of empathy: for a given distress context, diverse responses (e.g., a comforting affirmation vs. a constructive suggestion) can be equally valid yet lexically distinct (Liu et al., 2016; Deriu et al., 2021). As a result, these metrics often penalize high-quality creative responses that diverge from the reference, showing poor correlation with human judgments.

To overcome the reference dependency, recent works have employed Large Language Models (LLMs) as reference-free judges, leveraging their reasoning capabilities to assess dialogue quality directly (Lin et al., 2023; Zhang et al., 2024). While LLM-based evaluation aligns better with human perception, it introduces significant bottlenecks for practical deployment. Relying on proprietary LLMs (e.g., GPT-4) as online judges incurs high inference latency and prohibitive costs for large-scale model development. Furthermore, sending sensitive mental health data to external APIs raises serious privacy concerns (Kocmi and Freitag, 2023; Croxford et al., 2025). The field thus faces a dilemma: traditional metrics are efficient but inaccurate, while LLM judges are accurate but inefficient and costly.

In this paper, we propose to bridge this gap by distilling the empathic reasoning of expensive LLMs into lightweight, privacy-preserving models. We introduce EMPATHYSCORE, a reference-free metric constructed via the DISTILL-ES framework. Our core insight is to treat the LLM not as a permanent judge, but as a *teacher* that provides rich supervision via rubric-guided prompting. Specifically, we prompt a teacher LLM to analyze dialogues

along cognitive and affective dimensions, generating both fine-grained scores and pairwise preferences. We then train two compact student scorers (RoBERTa-based) to mimic these judgments using a hybrid objective that combines regression (for absolute calibration) and ranking (for relative preference learning).

This distillation approach yields a robust evaluation framework that is both *decoupled* and *efficient*. By explicitly modeling *Cognitive Suitability* and *Affective Resonance* as separate sub-tasks, EMPATHYSCORE provides interpretable feedback on where a model succeeds or fails. Extensive experiments demonstrate that our metric achieves state-of-the-art correlation with human ratings, matching the performance of the teacher LLM while being orders of magnitude faster and removing the need for test-time API access. EMPATHYSCORE thus offers a scalable standard for assessing machine empathy, enabling rapid iteration of ESC systems without compromising on evaluation quality.

Our contributions are three-fold:

- We propose EMPATHYSCORE, a reference-free metric that decouples empathy evaluation into cognitive suitability and affective resonance.
- We introduce the DISTILL-ES framework to distill sophisticated LLM judgments into lightweight student scorers via rubric-guided supervision.
- Experiments show that our metric achieves state-of-the-art correlation with human ratings while being orders of magnitude more efficient than LLMs.

2 Related Work

2.1 Emotional Support Conversation

Modeling for Emotional Support Conversation (ESC) mainly follows three lines: *knowledge-enhanced* methods that inject commonsense or graphs to better understand seekers (e.g., MISC, C^3KG, GLHG)(Tu et al., 2022; Li et al., 2022; Peng et al., 2022); *cognitive reasoning* that first infers cues and emotions before responding (Dialogue CoT, Cue-CoT)(Chae et al., 2023; Wang et al., 2023); and *persona modeling* for personalization and consistency (PAL)(Cheng et al., 2023). While ESConv provides high-quality but limited-scale training data(Liu et al., 2021), LLM-based augmentation (AugESC, ExTES)(Zheng et al., 2023, 2024) and socially grounded simulation (SocialSim)(Chen et al., 2025) expand coverage and realism. Most recently, CARE(Zhu et al., 2025) and ReflectDiffu(Yuan et al., 2025) both leverage reinforcement learning to enhance controllability and empathy.

However, most studies still rely on automatic metrics such as BLEU, ROUGE-L, METEOR, BERTScore, and Distinct, which weakly capture perceived empathy and support(Zhao et al., 2024; Madani and Srihari, 2025), while human ratings—though more faithful—are costly and hard to scale; therefore, our work proposes *Empathy-Score*, an automatic and scalable metric that better reflects human-perceived empathy.

2.2 Empathy Measurement

Empathy is commonly divided into two types: *trait empathy*, which captures relatively stable empathetic dispositions, and *situational empathy*, which refers to immediate, context-triggered responses; the former can shape how the latter manifests in specific situations, making the distinction important (Davis et al., 1980; Hoffman, 2000). In psychology, self-report questionnaires are widely used to assess relatively stable empathic traits, such as the *Interpersonal Reactivity Index*(Davis, 1983), the *Empathy Quotient*(Baron-Cohen and Wheelwright, 2004), and the *Toronto Empathy Questionnaire*(Spreng* et al., 2009).

In affective computing and NLP field, measurement centers on *situational empathy* and typically uses rater-based judgments as ground truth to train supervised deep learning models. Concretely, one line predicts reader self-reported empathy/distress from stories(Buechel et al., 2018), while another learns empathy signals from annotated mental-health counseling conversations(Sharma et al., 2020), with related efforts extending to other genres (e.g., social media(Abdul-Mageed et al., 2017), persuasive-oriented dialogues(Samad et al., 2022)). In ESC specifically, perceived empathy is a central criterion for helpful support and positive outcomes(Rogers, 1957b; Elliott et al., 2011); therefore, we introduce *Empathy-Score*, an automatic metric tailored for ESC that approximates human judgements of perceived empathy to enable scalable evaluation and training.

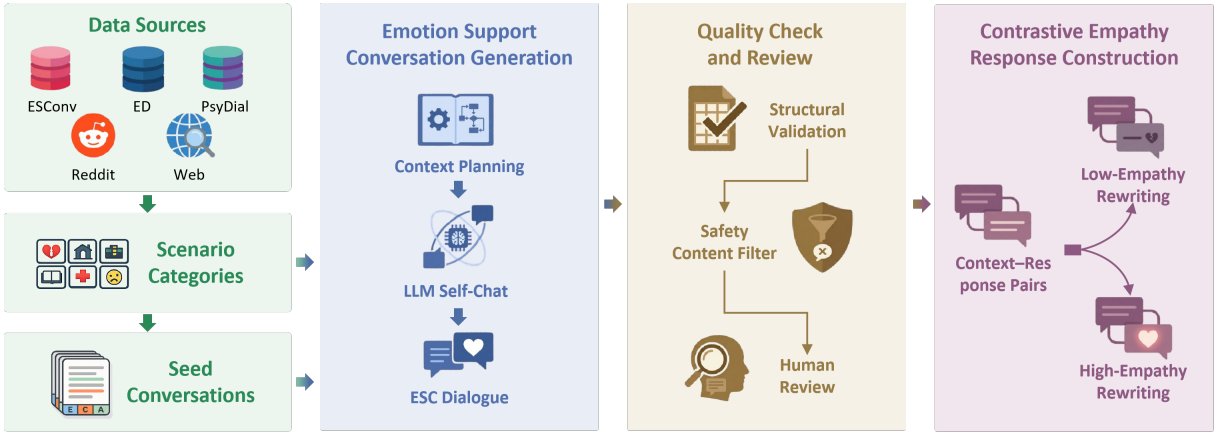

Figure 1: The construction pipeline of the EMPATHY-EVAL dataset

3 Preliminaries

In this section, we construct a evaluation dataset, EMPATHY-EVAL. We then analyze the behavior of existing automatic metrics on this benchmark, highlighting their limitations and motivating the need for our proposed EMPATHYSCORE.

3.1 EMPATHY-EVAL Dataset

Construction Process. The construction of EMPATHY-EVAL proceeds through a rigorous five-stage pipeline designed to ensure both diversity and quality. Initially, we manually curated 60 distinct emotional support scenarios (see Appendix B.4 for specific examples) selected from five diverse data sources including ESConv (Liu et al., 2021), EmpatheticDialogues (Rashkin et al., 2019), PsyDial (Qiu and Lan, 2025), Reddit, and web crawling. For each scenario, we developed high-quality seed dialogues which underwent strict manual review to ensure logical consistency. To scale up the corpus, we extended the methodology of Zheng et al. (2024), employing an LLM-based self-chat framework (details in Appendix B.6) to synthesize large-scale multi-turn dialogues. This generation process incorporated hierarchical control over 16 support strategies and three dialogue states (*Exploration*, *Comforting*, *Action*). Subsequently, the synthesized corpus was subjected to automated quality checks, safety filtering, and a final round of human review (see Appendix B.3) to eliminate toxic or incoherent content. Finally, to facilitate fine-grained evaluation, we applied a contrastive rewriting mechanism (details in Appendix B.5) to each helper turn. Using the original response as a reference (R_{ref}), we generated an *Apathetic* version (R_{apa}) via de-empathizing rewriting and an *Empathetic* version (R_{emp}) via strategy-altered regeneration.

Category	ESConv	ExTES	EMPATHY-EVAL
Dialogues	1,053	11,177	30,633
Utterances	31,410	200,393	522,432
Avg. length of dialogues	29.8	18.2	17.1
Avg. length of utterances	17.8	26.0	28.7
Num. of support strategies	8	16	16
Num. of scenarios	5	36	60

Table 1: Comparison with existing ESC datasets.

Metric (1-3)	Low (R_{apa})	High (R_{emp})	Kappa
Coherence	2.92	2.95	0.78
Helpfulness	1.15	2.82	0.72
Empathy	1.08	2.91	0.75

Table 2: Human evaluation results (scale 1-3). R_{emp} significantly outperforms R_{apa} in empathy and helpfulness.

Data Analysis. Table 1 highlights the scale and diversity of Empathy-Eval. Our dataset contains 30,633 dialogues, which is significantly larger than previous datasets like ESConv (Liu et al., 2021) and ExTES (Zheng et al., 2024). Crucially, it covers 60 diverse emotional scenarios, offering a much broader range of topics than the limited scenarios in baselines. Comprehensive statistical analyses are detailed in Appendix B.

Human Evaluation. To validate data quality, we randomly sampled 500 contrastive pairs (R_{apa} vs. R_{emp}) for human review. Ten annotators rated them on a 1-3 scale (Poor, Fair, Good) based on *Coherence*,*Helpfulness*, and *Empathy*. As shown in Table 2, R_{emp} scores nearly perfect in Empathy (2.91) and Helpfulness (2.82), whereas R_{apa} scores very low (≈ 1.1). This confirms that our rewriting method successfully created a clear gap between

high and low empathy. Additionally, both versions maintain high Coherence, ensuring the text is fluent and logical. The high Kappa scores indicate reliable agreement among annotators.

3.2 Limitations of Existing NLP Metrics

To assess the applicability of automated metrics for empathy evaluation, we computed n-gram overlap, semantic similarity, and linguistic diversity on our contrastive dataset (R_{emp} vs. R_{apa}). As shown in Table 3, results reveal a systemic misalignment: Low Empathy responses consistently outperform High Empathy responses across all metric categories. This inverse correlation stems from the fact that standard metrics primarily reward lexical and semantic adherence to the factual reference content. Since empathetic responses often require significant paraphrasing and stylistic shifts rather than mere content preservation, they are penalized by these metrics. Consequently, existing NLP metrics are proven ineffective for quantifying the specific dimension of empathy.

Category	Metric	High (R_{emp})	Low (R_{apa})
n-gram	BLEU-4	0.0285	0.3852
	ROUGE-L	0.1688	0.5660
	METEOR	0.2340	0.5690
Semantic	BERTScore	0.9613	0.9778
Diversity	Dist-1	0.0076	0.0098
	Dist-2	0.0885	0.1231
	Dist-4	0.4618	0.5762

Table 3: Comparison of standard NLP metrics. The metrics consistently assign higher scores to Low Empathy responses, demonstrating an inability to distinguish the empathy dimension.

4 EMPATHYSCORE Framework

4.1 Problem Formulation

We focus on the Emotional Support Conversation (ESC) task, involving a distressed *seeker* and a supportive *helper*. Following the theoretical framework in ESConv (Liu et al., 2021), the support process is structured hierarchically: fine-grained support strategies (e.g., *Restatement*, *Suggestion*) are organized into three high-level support stages: *Exploration*, *Comforting*, and *Action*. Ideally, a helper should dynamically transition through these stages based on the seeker's evolving emotional state.

To enhance clarity, key notations are summarized in Table 4.

Symbol	Definition
C	The dialogue context, defined as a sequence of role-utterance pairs
$R, \hat{R}$	The human reference response and the model-generated response
$\mathcal{S}, \mathcal{Z}$	Sets of fine-grained support strategies and high-level support stages
p, u	The role indicator (Seeker/Helper) and the utterance content
$s_{\hat{R}}$	The strategy label predicted for response $\hat{R}$
P_{ideal}	The predicted latent distribution of ideal strategies
S_{flow}	The cognitive suitability score (Metric I)
S_{content}	The affective resonance score (Metric II)

Table 4: Summary of notations for the EMPATHYSCORE framework.

Formally, we define the dialogue context as a sequence of L turns $C = \{(p_1, u_1), (p_2, u_2), \ldots, (p_L, u_L)\}$, where $p_i \in \{\text{Seeker}, \text{Helper}\}$ denotes the speaker role and u_i is the textual content. This flexible formulation accounts for irregular turn-taking dynamics (e.g., consecutive messages from the same speaker). Given C, an ESC model aims to generate a response $\hat{R}$ (where $p_{L+1} = \text{Helper}$) that is both structurally appropriate and emotionally resonant.

However, current evaluation protocols are insufficient for this goal. As empirically demonstrated in Table 3, standard NLP metrics exhibit a significant misalignment with human judgment. To bridge this gap, we propose EMPATHYSCORE, a reference-free, learning-based metric that evaluates $\hat{R}$ from two complementary dimensions:

$$\text{EmpathyScore}(C, \hat{R}) = \alpha \cdot S_{\text{flow}} + (1-\alpha) \cdot S_{\text{content}} \quad (1)$$

where S_{flow} and S_{content} are normalized scores in $[0, 1]$. The hyper-parameter $\alpha \in [0, 1]$ controls the trade-off between the cognitive (flow) and affective (content) dimensions. In this work, we set $\alpha = 0.5$, assuming that appropriate strategy selection and emotional resonance are equally critical.

- S_{flow} evaluates **Cognitive Suitability** (reflecting strategy flow). It measures the **probabilistic validity** of the helper's strategy against the ideal support flow derived from the full interaction history C. This captures the cognitive aspect of empathy, ensuring stage-wise adaptation to the seeker's evolving state while

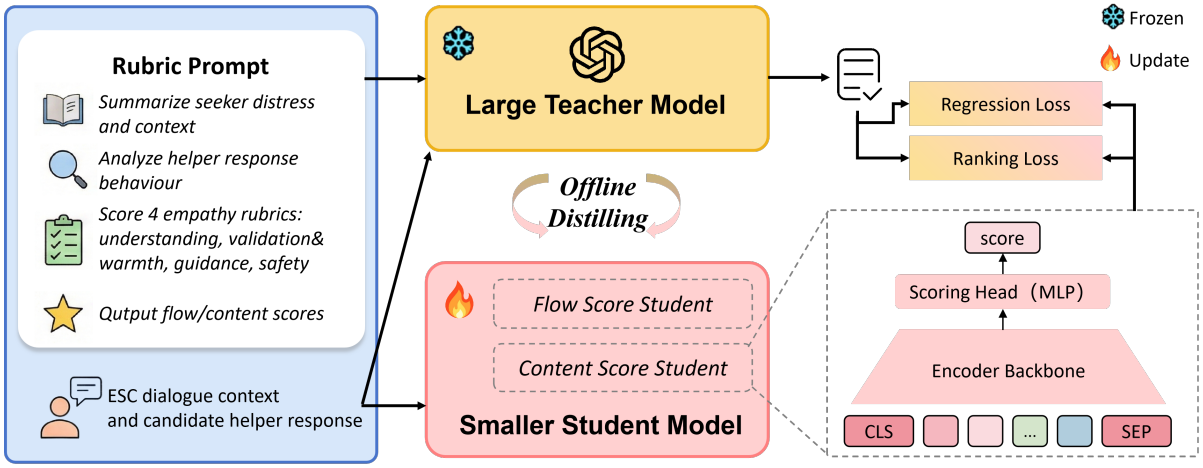

Figure 2: he overview of the DISTILL-ES framework.

allowing for diverse strategic choices (Hill, 2014; Clark, 2010).

- S_{content} evaluates **Affective Resonance** (reflecting content quality). It leverages a **contrastive learning objective** to assess whether the generated content structurally entails the seeker's emotional distress. This metric captures the intrinsic warmth and affective understanding of the response, aligning with person-centered message theory (High and Dillard, 2012).

4.2 The DISTILL-ES Framework

Recent advances in LLM distillation show that rich supervision from large teachers can be compressed into compact students for instruction following, automatic evaluation, and dialogue understanding (Yue et al., 2024; Oh et al., 2025; Lee et al., 2024; Wei et al., 2025). Following this line of work, we cast EMPATHYSCORE as a learned, LLM–distilled metric and propose the DISTILL-ES (Distilled EMPATHYSCORE) framework, which transfers empathy judgements from a strong teacher LLM into lightweight scorers specialized for emotional support conversations (ESC), as illustrated in Figure 2.

Teacher rubric prompting. For each ESC dialogue context C and candidate helper response $\hat{R}$, we query a frozen teacher LLM $\mathcal{T}$ with a rubric-style prompt (Figure 2). The prompt asks $\mathcal{T}$ to (i) summarize the seeker's distress and conversational context, (ii) analyze the helper's behaviour, and (iii) assign scalar empathy scores in $[0, 1]$ for two dimensions: cognitive suitability of the support flow $S^T_{\text{flow}}(C, \hat{R})$ and affective content resonance $S^T_{\text{content}}(C, \hat{R})$. In addition, the teacher provides four rubric-level scores that refine content empathy (empathic understanding, validation & warmth, guidance and actionability, safety), and, for a subset of contexts, pairwise preferences indicating which of two responses is more empathic. All teacher queries are collected offline to form a distillation dataset; at test time the teacher is never used.

Student architecture. We instantiate two student scorers, a *Flow Score Student* f^{flow}_θ and a *Content Score Student* f^{cont}_θ. Both share the same RoBERTa-style encoder backbone followed by a shallow MLP scoring head (the right panel in Figure 2). Given $(C, \hat{R})$, each student produces a scalar score in $[0, 1]$:

$$S_{\text{flow}}(C, \hat{R}) = f^{\text{flow}}_\theta(C, \hat{R}), \quad (2)$$

$$S_{\text{content}}(C, \hat{R}) = f^{\text{cont}}_\theta(C, \hat{R}). \quad (3)$$

which are combined into the final metric according to Eq. (1). Thus DISTILL-ES implements EMPATHYSCORE as two compact, reference-free scoring models that can be applied to any ESC system output.

Distillation objective. Each student is trained to match the teacher's scalar scores while preserving the teacher's relative preferences between candidate responses. For a single scored example $(C, \hat{R})$ with teacher score $S^T_d(C, \hat{R})$ on dimension $d \in \{\text{flow}, \text{content}\}$, we use a regression loss

$$\mathcal{L}^{(d)}_{\text{reg}} = \big(S_d(C, \hat{R}) - S^T_d(C, \hat{R})\big)^2, \quad (4)$$

where S_d denotes the corresponding student score (S_{flow} or S_{content}). For ranking supervision, we con-

struct response pairs $(\hat{R}^+, \hat{R}^-)$ under the same context C where the teacher prefers $\hat{R}^+$ to $\hat{R}^-$, and apply a margin-based ranking loss

$$\mathcal{L}_{\text{rank}}^{(d)} = \max\big(0,\, m - S_d(C, \hat{R}^+) + S_d(C, \hat{R}^-)\big), \tag{5}$$

with a fixed margin $m > 0$. The distillation loss for dimension d is

$$\mathcal{L}_{\text{KD}}^{(d)} = \mathcal{L}_{\text{reg}}^{(d)} + \lambda_{\text{rank}}\, \mathcal{L}_{\text{rank}}^{(d)}, \tag{6}$$

and the overall training objective sums the flow and content components:

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{KD}}^{(\text{flow})} + \mathcal{L}_{\text{KD}}^{(\text{content})}. \tag{7}$$

In all experiments, DISTILL-ES is trained purely with this offline distillation objective, without access to human empathy labels, and can therefore be scaled to large amounts of unlabeled ESC data.

5 Experiments

In this section, we conduct comprehensive experiments to evaluate EMPATHYSCORE against state-of-the-art baselines. We aim to answer the following research questions:

- **RQ1:** Can EMPATHYSCORE effectively distinguish between high-quality empathetic responses and low-quality ones?
- **RQ2:** How do the proposed components (S_{flow} and S_{content}) contribute to the overall performance?

5.1 Experimental Setup

5.1.1 Datasets

To validate the effectiveness of our framework, we utilize the EMPATHY-EVAL dataset. This dataset was constructed by applying our Teacher Rubric Prompting (using GPT-5) to a corpus of emotional support conversations, generating high-quality annotations that include ideal strategy distributions, scalar empathy scores, and pairwise preference rankings.

Consistent with standard supervised learning paradigms, we randomly split the EMPATHY-EVAL dataset into two subsets:

- **Training Set (80%):** Used to optimize the student scorers. The student models leverage the scalar scores for the regression loss and the pairwise preferences for the ranking loss to mimic the teacher's judgment logic.
- **Test Set (20%):** A held-out subset used exclusively for evaluation. This subset contains distinct dialogue contexts not seen during training, allowing us to assess the metric's generalization capability.

5.1.2 Baselines

We compare EMPATHYSCORE against widely used reference-based metrics in dialogue generation. These metrics evaluate performance by measuring the similarity between the system-generated response and human gold references:

- **Word-Overlap Metrics:** We report BLEU-4, ROUGE-L, and METEOR. These metrics evaluate performance by measuring the n-gram precision and recall against human references.
- **Embedding-based Metrics:** We report BERTScore, which computes semantic similarity.
- **Diversity Metrics:** We include Dist-1, Dist-2, and Dist-4, which measure the ratio of distinct uni-grams and bi-grams in the generated responses. While primarily used for generation diversity, they serve as a baseline to check if empathy correlates with lexical richness.

5.1.3 Implementation Details

We employ RoBERTa-base as the shared encoder backbone for both the flow and content student scorers. The scoring head consists of a lightweight two-layer MLP with a hidden size of 256. We train the models for 5 epochs using the AdamW optimizer with a learning rate of $2\text{e}{-5}$ and a batch size of 32. For the hybrid objective, the ranking margin is set to $m = 0.1$, and the ranking loss weight is set to $\lambda_{\text{rank}} = 0.5$. All experiments were conducted on a single NVIDIA A100 GPU.

5.2 Evaluation of Distinguishability (RQ1)

To answer **RQ1**, we rigorously analyze whether EMPATHYSCORE can effectively distinguish between high-quality empathetic responses (R_{emp}) and low-quality generic ones (R_{apa}). This comparison against standard metrics is critical, as a reliable metric must not only correlate with human judgment but also possess sufficient discriminative power to penalize sub-optimal outputs. The comparative results are presented in Table 5.

Table 5: Mean scores on High-Quality (R_{emp}) vs. Low-Quality (R_{apa}) subsets. **Bold** indicates the higher score between the two groups. Standard metrics consistently exhibit an *Evaluation Inversion* (scoring Low > High), whereas EMPATHYSCORE correctly aligns with human preference.

Metric	High (R_{emp})	Low (R_{apa})
n-gram Metrics		
BLEU-4	0.0285	**0.3852**
ROUGE-L	0.1688	**0.5660**
METEOR	0.2340	**0.5690**
Semantic Metrics		
BERTScore	0.9613	**0.9778**
Diversity Metrics		
Dist-1	0.0076	**0.0098**
Dist-2	0.0885	**0.1231**
Dist-4	0.4618	**0.5762**
Ours		
EMPATHYSCORE	**0.8420**	0.3250

Results and Analysis As illustrated in Table 5, we observe a distinct and systematic divergence in how different metrics rank the two contrastive groups. The results highlight the fundamental limitations of current automated evaluation methods when applied to the open-ended domain of Emotional Support Conversation (ESC).

1. The Phenomenon of Evaluation Inversion in Standard Metrics: Standard automated metrics exhibit a severe misalignment with human judgment, a phenomenon we term *Evaluation Inversion*. Specifically, n-gram metrics (BLEU-4, ROUGE-L, METEOR) assign significantly higher scores to the Low-Quality group (R_{apa}) than to the High-Quality group (R_{emp}). For instance, BLEU-4 scores generic responses nearly 13 times higher (0.3852) than specific empathetic responses (0.0285).

This counter-intuitive result primarily stems from the "One-to-Many" nature of open-ended dialogue and the "safe response" bias. In the ESC task, for any given emotional query, there exists a vast space of valid, high-empathy responses that differ significantly in vocabulary and structure. However, generic utterances (e.g., *"I am sorry to hear that"*, *"That sounds tough"*) appear with disproportionately high frequency across the training corpus. Consequently, traditional metrics, which rely on surface-level lexical matching, are statistically biased towards these "average" responses. A model that generates these safe, low-information utterances maximizes statistical likelihood and overlap with the references, despite lacking genuine emotional insight. Conversely, high-quality empathetic responses are often tailored to specific details of the user's distress, resulting in unique formulations that rarely match the rigid surface forms of the ground truth references.

Furthermore, even advanced semantic-based metrics like BERTScore fail to rectify this bias, favoring the generic group (0.9778) over the empathetic one (0.9613). This observation reveals a critical theoretical limitation: high semantic similarity to a reference does not equate to high pragmatic quality. BERTScore effectively captures topical relevance but struggles to distinguish between "polite acknowledgement" and "deep emotional validation." A response can be semantically close to a reference (sharing the same topic embedding space) but completely lack the specific supportive intent required for effective therapy-like interactions. This confirms that neither lexical overlap nor embedding-based similarity is sufficient to capture the dynamic and nuanced nature of empathy.

2. Effectiveness and Robustness of EMPATHYSCORE: In sharp contrast to the baselines, EMPATHYSCORE successfully corrects this bias and demonstrates superior distinguishing capability. As shown in the bottom row of Table 5, our metric assigns a significantly higher score to the High-Quality group (**0.8420**) compared to the Low-Quality group (0.3250).

This result highlights two key strengths of our proposed approach. First, it demonstrates robustness against the "safe response" trap. Unlike probability-based models that reward the most frequent patterns, EMPATHYSCORE is trained to recognize specific empathetic features, ensuring that generic phrases—no matter how grammatically correct—receive lower scores due to their lack of emotional substance. Second, the substantial margin between the two groups indicates that our model effectively aligns with human cognitive intuition. Human evaluators prioritize the functional aspect of support (e.g., "Did the system really understand me?") over linguistic fluency. By evaluating the intrinsic quality of the response—considering both its emotional content and interaction flow—EMPATHYSCORE correctly identifies and rewards the nuanced emotional support present in R_{emp}. This capability to discrim-

Table 6: Ablation study results. We report the Pearson (r) and Spearman (ρ) correlation with human judgments. "Δ" denotes the performance drop compared to the full model. The results show that both components are essential, with S_{content} contributing primarily to semantic accuracy and S_{flow} enhancing structural coherence.

Model Variants	Correlation		Performance Drop	
	Pearson (r)	Spearman (ρ)	Δr	$\Delta \rho$
Full Model (EMPATHYSCORE)	**0.684**	**0.672**	-	-
w/o Flow Component (S_{flow})	0.592	0.585	↓ 13.5%	↓ 12.9%
w/o Content Component (S_{content})	0.415	0.398	↓ 39.3%	↓ 40.8%

inate between "safe" and "supportive" is crucial for training future ESC systems, answering **RQ1** affirmatively.

5.3 Ablation Study (RQ2)

To answer **RQ2**, we conducted a comprehensive ablation study to investigate the individual contributions of the two core components of EMPATHYSCORE: the Flow-based Score (S_{flow}) and the Content-based Score (S_{content}). By systematically removing one component at a time, we evaluated the performance of model variants using Pearson (r) and Spearman (ρ) correlation coefficients with human ratings on the test set. This analysis allows us to understand how each module influences the overall evaluation capability and whether the proposed hybrid architecture is theoretically justified.

Results and Analysis Table 6 summarizes the contribution of each component. The empirical results clearly indicate that while both components positively contribute to the final performance, they play distinct and complementary roles in the evaluation framework, validating our dual-stream hypothesis:

1. Impact of Empathetic Content (S_{content}): The variant removing the content component ("w/o Content") resulted in the most significant performance degradation, with the Pearson correlation dropping drastically by 39.3% (from 0.684 to 0.415). This finding confirms that the semantic content—specifically the linguistic realization of empathy—is the foundational element of the evaluation system. Empathy is fundamentally communicated through words that convey understanding, validation, and care. Specifically, this component captures the fine-grained semantics of emotional support, such as the use of specific emotional keywords, supportive phrasing, and validation techniques that resonate with the user's state. Without the ability to assess *what* is explicitly said (S_{content}), the model loses its capacity to detect the "warmth" and "understanding" inherent in the dialogue. Relying solely on interaction flow (S_{flow}) proves insufficient because a conversation might follow a correct structural transition pattern (e.g., asking a question after a statement) but still lack the necessary emotional resonance or compassionate tone to be truly supportive.

2. Impact of Dialogue Flow (S_{flow}): Removing the flow component ("w/o Flow") also leads to a notable decrease in performance, with a 13.5% drop in Pearson correlation. While S_{content} is critical for evaluating the semantic dimension, the addition of S_{flow} provides crucial context-aware regularization. Empathy in ESC is not a static property but a dynamic, temporal process that requires appropriate progression. Effective support typically follows a logical sequence: listening and validating first, then exploring the problem, and finally offering advice. S_{flow} acts as a safeguard against "contextual mismatch," penalizing responses that may be semantically polite but are contextually abrupt or out of order—such as rushing to give practical advice before sufficiently establishing an emotional connection. This drop in performance without S_{flow} suggests that structural coherence is a vital multiplier for the overall quality of empathy; a well-worded response delivered at the wrong stage of the conversation is often perceived as insensitive.

Takeaway The full model achieves the highest correlation (**0.684**), significantly outperforming either single-component variant. This demonstrates that EMPATHYSCORE effectively combines the semantic precision of S_{content} with the structural awareness of S_{flow}. The two components work in synergy: the content module evaluates the *substance* of the response (the "what"), while the flow module evaluates its *timing and appropriateness* (the "when"). This ablation study strongly supports the necessity of the proposed hybrid architecture, proving that accurately capturing the complexity of human empathy requires modeling both the linguistic features and the interaction dynamics.

6 Conclusion

We presented EMPATHYSCORE, a reference-free metric for Emotional Support Conversations that explicitly decomposes empathy into cognitive suitability of support strategy flow and affective resonance of response content, together with the DISTILL-ES framework for learning this metric.

Limitations

While EMPATHYSCORE and the DISTILL-ES framework take a step toward more faithful evaluation of emotional support conversations, several limitations remain. First, our metric ultimately inherits the biases and blind spots of the teacher LLM and our rubric design: the "ground truth" supervision comes from a single model family, a particular set of prompts, and a fixed four-dimensional rubric. As a result, the distilled scorers may encode stylistic or cultural preferences of the teacher (e.g., favouring certain phrasing patterns or support styles) and may not fully reflect diverse human judgments about what counts as empathic, especially across different cultures, demographics, or counselling traditions. Second, our experiments are conducted on English text-based ESC datasets in non-clinical settings; it is unclear how well EMPATHYSCORE transfers to other languages, modalities (e.g., voice-based support with prosody), or high-stakes clinical contexts where professional standards for empathy and safety are stricter.

Methodologically, DISTILL-ES focuses on offline correlation with human and teacher scores and does not guarantee that higher scores causally imply better outcomes for seekers (e.g., improved mood or reduced distress). Our margin-based ranking loss relies on teacher-generated pairwise preferences, which may contain label noise, and we have not systematically analysed failure cases where the metric disagrees with expert counsellors. In addition, our current implementation uses RoBERTa-style encoders of moderate size; we do not explore the trade-off between model capacity and evaluation quality, nor do we study robustness under domain shift or adversarial prompting. Finally, although we discuss potential applications of EMPATHYSCORE for model selection and benchmarking, we do not investigate its impact when used as a training-time objective to optimise generative ESC models, nor the ethical implications of deploying such a metric in real-world mental health systems (e.g., unintended over-reliance on automated scores for triage or quality control). These limitations highlight important directions for future work.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Salvatore Giorgi, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):448–451.

Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2):163–175.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. 2025. Socialsim: towards socialized simulation of emotional support conversation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Arthur J. Clark. 2010. Empathy: An integral model in the counseling process. *Journal of Counseling & Development*, 88(3):348–356.

Rupert Croxford and 1 others. 2025. Current and future state of evaluating large language models for medical summarization tasks. *npj Digital Medicine*. To appear.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Mark H. Davis, Miles P. Davis, M Davis, Matthew Davis, Mark Davis, Mm Davis, M Davis, F. Caroline Davis, Heather A. Davis, and Ilus W. Davis. 1980. A multidimensional approach to individual differences in empathy.

Jan T. Deriu, Álvaro Rodrigo, Arantxa Otegi, and 1 others. 2021. A survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and Leslie S. Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43–49.

Andrew C. High and James Price Dillard. 2012. A review and meta-analysis of person-centered messages and social support outcomes. *Communication Studies*, 63(1):99–118.

Clara E. Hill. 2014. *Helping Skills: Facilitating Exploration, Insight, and Action*, 4 edition. American Psychological Association, Washington, DC.

Martin L. Hoffman. 2000. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press.

Tom Kocmi and Markus Freitag. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Seanie Lee, Jianpeng Cheng, Joris Driesen, Alexandru Coca, and Anders Johannsen. 2024. Effective and efficient conversation retrieval for dialogue state tracking with implicit text summaries. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 96–111, Mexico City, Mexico. Association for Computational Linguistics.

Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. 2022. C^{3}KG: A Chinese commonsense conversation knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1369–1383, Dublin, Ireland. Association for Computational Linguistics.

Yung-Sung Lin, Chien-Feng Li, and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, and 1 others. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

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.

Navid Madani and Rohini Srihari. 2025. ESC-judge: A framework for comparing emotional support conversational agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16059–16076, Suzhou, China. Association for Computational Linguistics.

Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yoen Kim, Taewon Yun, and Hwanjun Song. 2025. Learning to verify summary facts with fine-grained LLM feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 230–242, Abu Dhabi, UAE. 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.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *Preprint*, arXiv:2204.12749.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

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.

Carl R. Rogers. 1957a. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Carl R. Rogers. 1957b. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5280. Association for Computational Linguistics.

R. Nathan Spreng*, Margaret C. McKinnon*, Raymond A. Mar, and Brian Levine. 2009. The toronto

empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, 91(1):62–71. PMID: 19085285.

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.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Yuting Wei, Qi Meng, Yuanxing Xu, and Bin Wu. 2025. TEACH: A contrastive knowledge adaptive distillation framework for classical Chinese understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3537–3550, Vienna, Austria. Association for Computational Linguistics.

Jiahao Yuan, Zixiang Di, Zhiqing Cui, Guisong Yang, and Usman Naseem. 2025. ReflectDiffu: Reflect between emotion-intent contagion and mimicry for empathetic response generation via a RL-diffusion framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25435–25449, Vienna, Austria. Association for Computational Linguistics.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Distilling instruction-following abilities of large language models with task-aware curriculum planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6030–6054, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, and 1 others. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*.

Yifan Zhang and 1 others. 2024. FEEL: A benchmark for evaluating emotional support in mental health dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

Jie Zhu, Yuanchen Zhou, Shuo Jiang, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. 2025. Care: Cognitive-reasoning augmented reinforcement for emotional support conversation. *Preprint*, arXiv:2510.05122.

A Appendix

B Additional Details of EMPATHY-EVAL

In this section, we provide the full scenario distribution list and a fine-grained statistical analysis to demonstrate the diversity and quality of the EMPATHY-EVALdataset.

B.1 Scenario Coverage

To ensure the dataset encompasses a broad spectrum of real-world emotional distress, we curated 60 distinct scenarios ranging from relationship issues to career challenges. Table 7 details the complete list of scenarios along with their distribution proportions, sorted by frequency. The most frequent scenarios involve high-impact life events such as *Breakups or Divorce* (4.02%) and *Conflicts* (3.97%), while the long tail covers nuanced topics like *Retirement adjustment* and *Unemployment*, ensuring the model is tested against diverse contexts.

B.2 Additional Statistical Analysis

To further validate the distinction between the generated variations, we visualize the strategy dynamics and linguistic characteristics in Figure 3 and Figure 4.

Strategy Dynamics and Profiles. Figure 3 illustrates the evolution of support strategies. The High

ID	Scenario Name	Prop.	ID	Scenario Name	Prop.
1	Breakups or Divorce	4.02%	31	Coping with a Diagnosis or Medical Treatment	1.50%
2	Conflicts or Communication Problems	3.97%	32	Finding Meaning and Purpose in Life	1.50%
3	Work-related Stress and Burnout	3.96%	33	Career Transitions	1.50%
4	Coping with the Death of a Loved One	3.95%	34	Support for Loved Ones or Friends	1.50%
5	Depression and Low Mood	3.52%	35	News overload and anxiety about world events	1.49%
6	Financial Worries and Uncertainty	3.44%	36	Coping with natural disasters or community-wide crises	1.49%
7	Anxiety and Panic	3.44%	37	Surviving and Recovering from Physical or Emotional Abuse	1.49%
8	Chronic loneliness and lack of social support	2.98%	38	Housing instability or fear of eviction	1.25%
9	Parenthood and Parenting Challenges	2.54%	39	Cultural Identity and Belonging	1.23%
10	Academic Stress or Pressure	2.50%	40	Job Loss or Career Setbacks	1.08%
11	Low Self-Esteem or Lack of Confidence	2.48%	41	Long-distance relationship stress	1.03%
12	Romantic loneliness and difficulty finding a partner	2.47%	42	Addiction and Recovery	1.02%
13	Chronic Illness or Pain Management	2.47%	43	Healing from Sexual Assault or Domestic Violence	1.01%
14	Chronic sleep problems and insomnia	2.03%	44	Spirituality and Faith	0.76%
15	Sibling Rivalry or Family Conflict	2.02%	45	Infertility and challenges with trying to conceive	0.75%
16	Moving to a New City or Country	2.01%	46	Empty nest and adjusting after children leave home	0.75%
17	Dealing with the Loss of a Pet	2.00%	47	Legal troubles and court cases	0.75%
18	Caregiver Support	1.99%	48	Gaming or internet addiction impacting daily life	0.75%
19	Parenting Challenges and Parental Guilt	1.99%	49	Post-Traumatic Stress Disorder (PTSD)	0.74%
20	Communication Challenges	1.98%	50	Religious or worldview conflicts with family or community	0.74%
21	Work–life balance and role overload	1.98%	51	Healing from Abuse	0.74%
22	Perfectionism, procrastination, and fear of failure	1.98%	52	Online harassment and cyberbullying	0.74%
23	Drifting apart from long-term friends or social circles	1.53%	53	Navigating Gender Identity and Transitioning	0.74%
24	Uncertainty about college major or future career path	1.52%	54	Transition to college or moving away from home	0.53%
25	Infidelity and rebuilding trust after cheating	1.52%	55	Managing Bipolar Disorder	0.53%
26	Social media overuse and social comparison	1.52%	56	Midlife crisis and career plateau	0.52%
27	Adjusting to a New Job or Role	1.52%	57	Health anxiety and fear of serious illness	0.52%
28	Experiencing discrimination and microaggressions	1.51%	58	Workplace bullying or toxic team dynamics	0.52%
29	Body Image Concerns and Eating Disorders	1.51%	59	Unemployment-related Stress	0.51%
30	LGBTQ+ Identity	1.51%	60	Retirement adjustment and loss of professional identity	0.51%

Table 7: Distribution of all 60 emotional support scenarios in EMPATHY-EVAL, sorted by frequency proportion. The list is split into two columns for layout efficiency.

Empathy flow (Left) exhibits a clear "Exploration-Comforting-Action" progression, mirroring effective counseling sessions. In contrast, the Low Empathy flow (Middle) remains static. The Radar Chart (Right) further confirms that R_{emp} covers a broad spectrum of supportive strategies, whereas R_{apa} is constricted to neutral responses.

Linguistic and Interaction Depth. Figure 4 quantifies the linguistic divergence:

- *Length Distribution (Left):* The distribution for R_{emp} is right-skewed (peaking $\approx$ 35 words), indicating more elaborate supportive language compared to R_{apa}.
- *Strategy Divergence (Middle):* The diverging chart highlights the "Empathy Delta." R_{emp} prioritizes constructive strategies like *Emotional Validation* (green bars), while R_{apa} relies on passive strategies like *Clarification* (red bars).
- *Depth Dynamics (Right):* The line plot shows that R_{emp} sustains deep engagement throughout the dialogue, whereas R_{apa} shows a flat trajectory.

B.3 Human Review and Quality Assurance

To ensure the reliability and safety of the dataset, the synthesized corpus was subjected to a rigorous quality control pipeline. Following automated quality checks and safety filtering, we conducted a final round of human review to explicitly eliminate toxic, incoherent, or strategy-mismatched content.

We recruited 10 annotators with expertise in linguistic evaluation, compensated at a rate of $5 USD per hour. Prior to the task, we conducted a mandatory training session detailing the workflow and annotation criteria. The full text of the instructions, along with screenshots of the annotation interface, is provided in Appendix **??**. Crucially, given the sensitive nature of emotional support dialogue, we issued a clear disclaimer regarding potential psychological risks associated with reading distressing content. All participants provided informed consent, acknowledging these risks before proceeding. To ensure consistency, annotators were required to analyze 10 distinct case studies and pass a pilot qualification test aligned with expert standards.

To streamline this large-scale review, we developed a customized auditing tool, the "Empathy-Eval Auditor" (see Figure 5). The interface presents the complete dialogue history alongside the specific situation description, providing the necessary context for adjudicating coherence

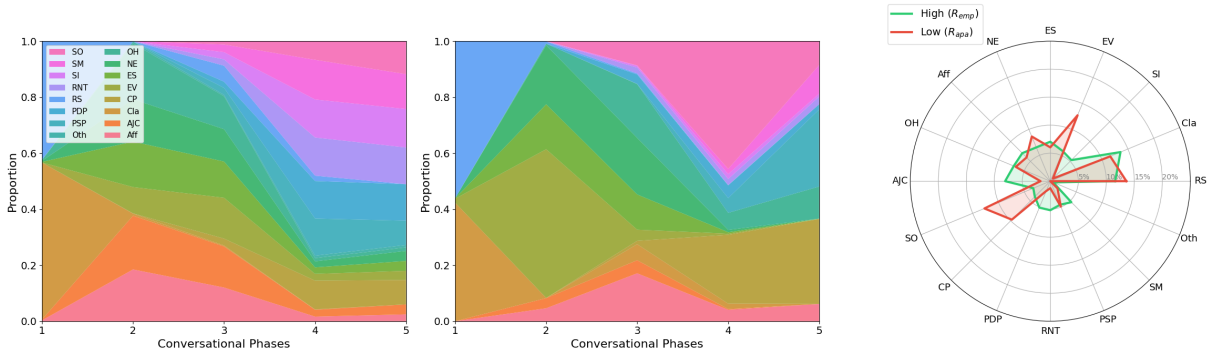

Figure 3: Visualization of strategy dynamics. (Left & Middle) High empathy responses follow a logical progression compared to static low empathy responses. (Right) The radar chart highlights the superior strategy diversity of R_{emp}.

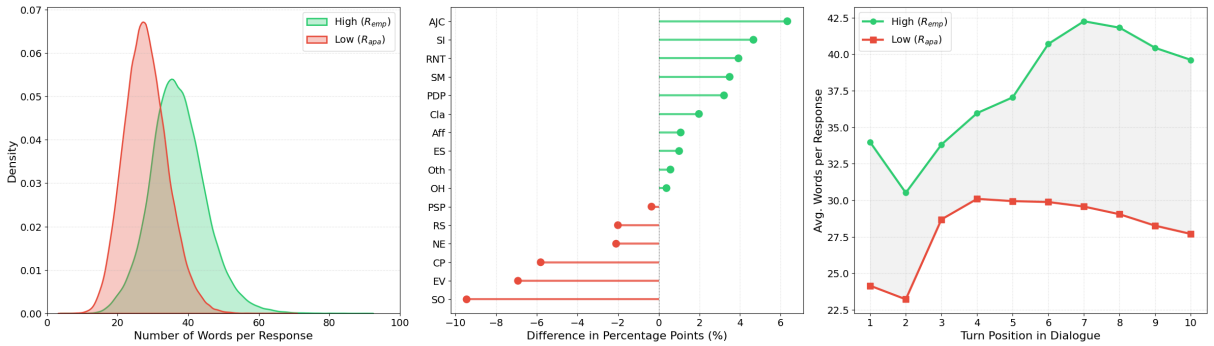

Figure 4: Advanced statistical comparison. (Left) Response length distribution. (Middle) Strategy divergence between high and low empathy. (Right) Turn-level conversation depth dynamics.

and safety. Using a binary decision mechanism ("Keep" or "Remove"), auditors verified whether the model's response accurately reflected the assigned strategy tag (e.g., *Emotional Validation*). This process ensured that only valid, high-quality dialogues were retained in the final EMPATHY-EVAL corpus.

B.4 Scenario Examples

Table 7 listed the distribution of our 60 emotional support scenarios. Below, we provide concrete dialogue examples for each scenario to illustrate the diversity of user situations covered in EMPATHY-EVAL.

Academic Stress or Pressure *Ex 1:* I studied so hard for this exam, but when I got my grade back, it was way lower than I thought it would be. *Ex 2:* I can't believe my teacher marked me wrong on that math test. It was my final exam and I studied for a week. *Ex 3:* I can't believe I did so badly on that test. I studied for hours, but it all just slipped away during the exam.

Addiction and Recovery *Ex 1:* I've been feeling really down, and it seems like my family has basically given up on me. *Ex 2:* I messed up. I drank again after six months of being clean, and now I feel like I've ruined everything. *Ex 3:* I slipped up last week after almost a year clean, and I just feel like I've let everyone down and myself too.

Adjusting to a New Job or Role *Ex 1:* It's only been a few weeks since I started managing this team, but I already feel like I'm way in over my head. *Ex 2:* I just started this new role as a team lead last week, and honestly, I already feel like I'm in way over my head. *Ex 3:* I just began this new lead position, and honestly, it feels like I'm constantly behind with no clear way to catch up.

Anxiety and Panic *Ex 1:* Every time I try to drive during rush hour, my heart starts racing and I feel this wave of panic like I can't breathe properly. *Ex 2:* Lately, whenever I go to family events, I feel this sudden wave of panic that makes my heart race and my thoughts spiral out of control. *Ex 3:* Lately, every time I have to talk to someone at work, I feel this rush of panic. My heart races and I get so dizzy I just want to hide away.

Body Image Concerns and Eating Disorders *Ex 1:* Hi... lately I just feel really down when I look in the mirror. *Ex 2:* Hey, I just can't stop feeling awful about how I look these days. *Ex 3:* Hey, I've been feeling really frustrated and ashamed of how I look lately.

Breakups or Divorce *Ex 1:* It's not a great day. I feel pretty awful, honestly. *Ex 2:* My girlfriend just left me. I feel awful, like my whole world collapsed. *Ex 3:* I honestly don't even know where to start. Everything feels shattered right now.

Career Transitions *Ex 1:* I just quit my teaching job to become a graphic designer, but now I'm really scared

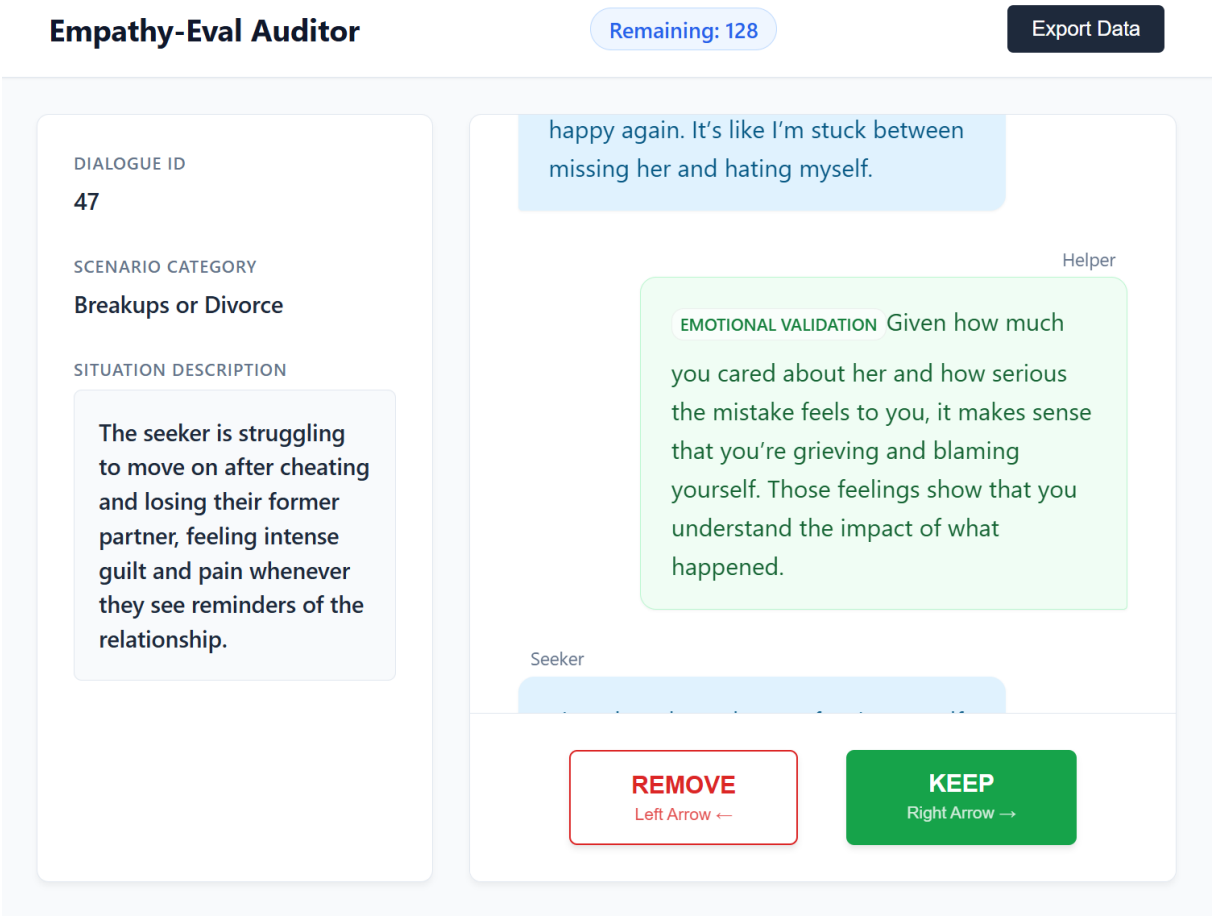

Figure 5: Screenshot of the Empathy-Eval Auditor interface used for the final human review. Annotators utilize this tool to audit the coherence, safety, and strategy alignment of the generated dialogues.

I made a huge mistake. *Ex 2:* I quit my stable job last month to pursue photography, but now I'm second-guessing if I did the right thing. *Ex 3:* I've been offered a promotion at work, but honestly, I feel really overwhelmed and unsure if I want to accept it.

Caregiver Support *Ex 1:* Hello, I was wondering if you could help me with something I'm really scared about. *Ex 2:* Hey, I'm really struggling with everything that's happening with my grandma. It's getting overwhelming. *Ex 3:* Hey, I'm really overwhelmed right now and don't know how to handle all this caregiving stuff for my dad.

Chronic Illness or Pain Management *Ex 1:* Lately, my energy's just disappeared. Even getting dressed feels like a huge effort, and it's really wearing me down. *Ex 2:* These past few weeks, my joints have been so painful and stiff that even getting out of bed feels like a huge challenge. *Ex 3:* Lately, my symptoms have gotten so much worse. I'm feeling shaky and tired all the time, and it's really messing up my days.

Chronic loneliness and lack of social support *Ex 1:* Sometimes I feel like I have nobody to talk to anymore. *Ex 2:* I just feel so invisible here, like no one really sees me for who I am. *Ex 3:* Lately, I just feel like I'm completely invisible to everyone around me.

Chronic sleep problems and insomnia *Ex 1:* Hi, I'm so tired. I feel like I haven't truly slept in ages. *Ex 2:* I'm really struggling to get any sleep these days, no matter how tired I am. *Ex 3:* Lately, I just can't seem to get to sleep at night, no matter how tired I am.

Communication Challenges *Ex 1:* I usually keep my thoughts to myself because I'm scared that bringing them up will just lead to fighting. *Ex 2:* I've been scared to tell my friend how I really feel because I don't want to upset them or start an argument. *Ex 3:* Lately, I just don't say what's really on my mind because I'm afraid it might upset people or cause arguments.

Conflicts or Communication Problems *Ex 1:* I really dislike my neighbours. They argue loudly all the time, and it's wearing me down. *Ex 2:* I swear every chat with my partner ends up tense or with us not really hearing each other anymore. *Ex 3:* I just can't seem to get through to my partner without things turning into a fight. It's exhausting.

Coping with a Diagnosis or Medical Treatment *Ex 1:* I just got my diagnosis a few days ago and honestly, I feel lost and scared about what's going to happen to me. *Ex 2:* I just found out I have early-stage Parkinson's, and honestly, I'm terrified about what this means for my future. *Ex 3:* I just got diagnosed with Parkinson's, and honestly, I'm really scared about what this means for me going forward.

Coping with natural disasters or community-wide crises *Ex 1:* I am so frustrated. My school was closed without warning because of COVID. *Ex 2:* The flood came so

suddenly, and we barely had time to grab anything before leaving the house. *Ex 3:* It's been really hard to focus since the tornado hit. Everything feels so chaotic and unsafe now.

Coping with the Death of a Loved One *Ex 1:* I still can't believe my sister is gone. It feels so unreal and I'm drowning in sadness. *Ex 2:* My dad passed away a few days ago, and honestly, I feel kind of lost and numb right now. *Ex 3:* It's been two weeks since my dad passed away, and I just feel completely lost most days.

Cultural Identity and Belonging *Ex 1:* Since I came back home a few months ago, I just can't shake this feeling that I don't really belong anywhere now. *Ex 2:* I've been back here for a few months, but instead of feeling at home, I just feel like I don't belong anywhere anymore. *Ex 3:* Ever since I came back to my birth country, I just can't shake off this feeling that I don't really belong anywhere anymore.

Dealing with the Loss of a Pet *Ex 1:* My dog passed away a few days ago, and the house feels so empty without him. *Ex 2:* My dog passed away last night. I don't think I've ever felt this empty before. *Ex 3:* I just had to put my cat down yesterday, and I feel like a part of me is gone too.

Depression and Low Mood *Ex 1:* I got some bad news from my doctor today. All I want to do is cry. *Ex 2:* Lately, I just feel empty all the time. Like nothing I do matters anymore. *Ex 3:* Lately, I just feel numb inside. Nothing seems to bring me any happiness anymore.

Drifting apart from long-term friends or social circles *Ex 1:* Hi. I don't really have close friends because I have a hard time trusting people. *Ex 2:* I've been feeling really down because my old friends hardly reach out to me these days. *Ex 3:* Hey, I've been feeling kind of sad lately because my close friends don't really talk to me like they used to.

Empty nest and adjusting after children leave home *Ex 1:* Both my kids moved out just a few weeks ago, and the house feels so empty. Some days I don't even recognize it as my home anymore. *Ex 2:* Since my kids left for college just a few weeks ago, the house feels so empty and quiet, it's almost hard to believe it's the same place. *Ex 3:* Since my kids moved out last week, the house feels so empty and cold. I keep looking for signs they're still here, but it's just silence now.

Experiencing discrimination and microaggressions in daily life *Ex 1:* Hi, I need to talk about something that's been really draining me lately. *Ex 2:* Hi, I've been really stressed out because of some things happening at work lately. *Ex 3:* Hey, I've been really overwhelmed lately because of some things happening at my job.

Financial Worries and Uncertainty *Ex 1:* I'm really stressed out. My work hours got cut last week, and I'm already behind on rent and bills. *Ex 2:* I'm really panicking. I borrowed five books from my school library and now the whole bag is missing. *Ex 3:* My work hours got cut this month, and now I'm really stressed about how I'll pay rent and cover my bills.

Finding Meaning and Purpose in Life *Ex 1:* I don't even know who I am without them. Everything feels so empty and pointless since the breakup. *Ex 2:* Since I quit my job, I feel like I've lost all sense of who I am and what I'm supposed to do with my life. *Ex 3:* Ever since my breakup a few months ago, I just feel like I'm wandering without any clue who I am or what I want anymore.

Gaming or internet addiction impacting daily life *Ex 1:* I've been playing games almost all night lately, and it's really messing up my sleep and making me fall behind at work. *Ex 2:* I've been gaming almost every night lately, sometimes until 3 or 4 a.m., and it's really messing up my sleep and focus at work. *Ex 3:* Lately, I've been gaming almost every night until dawn, and it's wrecking my sleep and making me miss important deadlines at work.

Healing from Abuse *Ex 1:* I still feel really anxious whenever my mom calls me, even though I've been setting boundaries for months now. *Ex 2:* I've been avoiding calls from my dad because they just bring up so much pain, but I feel awful about it afterward. *Ex 3:* I ended my friendship with someone I trusted for a long time, but now I just feel so lost and unsure if I did the right thing.

Healing from Sexual Assault or Domestic Violence *Ex 1:* Hi. I'm tired and anxious all the time because of the same nightmare that keeps coming back. *Ex 2:* Hi, I've been feeling scared and trapped in my own mind since I left my abusive partner a few months ago. *Ex 3:* I'm just so scared all the time now. Ever since the fight last week, I can't stop replaying it in my head.

Health anxiety and fear of serious illness *Ex 1:* Every time I notice a new twitch or ache, I immediately worry it's a sign of something really serious going wrong inside me. *Ex 2:* I've been getting these random chest pains, and I can't stop thinking it might be something really serious like a heart problem. *Ex 3:* Lately, whenever I get a headache or feel dizzy, I immediately start worrying something serious is wrong, like a brain issue or a tumor.

Housing instability or fear of eviction *Ex 1:* I just got an eviction notice because I lost my job. I'm really scared and don't know what to do next. *Ex 2:* My work hours got cut last month, and now I'm behind on rent. I'm really worried they might evict me soon. *Ex 3:* I'm really scared because I've fallen behind on my rent, and the landlord said I might have to move out soon.

Infertility and challenges with trying to conceive *Ex 1:* We've been trying to get pregnant for almost two years now, and every month I just feel more broken when it doesn't happen. *Ex 2:* It's been more than two years now, and each failed cycle just crushes me more. Sometimes I wonder if I'll ever hold my own baby. *Ex 3:* We decided to stop the fertility treatments last month, and honestly, I feel completely empty and confused about what to do now.

Infidelity and rebuilding trust after cheating *Ex 1:* I think my boyfriend might be seeing someone else, and it's eating me up inside. *Ex 2:* My very first boyfriend cheated on me. When I found out, I was furious and completely crushed. *Ex 3:* I just found out my partner has been cheating on me for months, and I feel completely shattered.

Job Loss or Career Setbacks *Ex 1:* Hi, I just lost my job after seven years, and I'm feeling really overwhelmed by it all. *Ex 2:* I got laid off last week, and honestly, I'm just shocked and scared about what comes next. *Ex 3:* I just got laid off last week, and honestly, I feel like my whole world flipped upside down.

LGBTQ+ Identity *Ex 1:* I've been feeling really lost lately, trying to understand my feelings, but it's confusing and scary. *Ex 2:* Hey, I'm really struggling with something. I think I might like my friend, and it's confusing me so much. *Ex 3:* Hey, I'm feeling really messed up right now. I think I like my best friend… and it's throwing me off completely.

Legal troubles and court cases *Ex 1:* I still feel ashamed when I think about the night my mom got arrested for drinking. *Ex 2:* I just got a court summons about a lawsuit from my old job, and I'm honestly terrified about what might happen. *Ex 3:* I still feel bad that I couldn't be there when my friend went to court. She really wanted support and I had to miss it.

Long-distance relationship stress *Ex 1:* Lately, it feels like the distance between us isn't just physical—it's like we're drifting apart emotionally too. *Ex 2:* I've noticed that lately our conversations feel so short and sometimes it feels like they don't really want to talk anymore. *Ex 3:* Hey, I've got to admit I'm pretty lonely right now. My wife is out of town for a week for work, and the house feels so empty.

Low Self-Esteem or Lack of Confidence *Ex 1:* I feel like no matter how hard I try, I'm always coming up short compared to everyone else in the team. *Ex 2:* I keep replaying that review in my head, and it just makes me feel like I'm not cut out for this job at all. *Ex 3:* I feel like I'm always behind at work, like no matter what I do, I'm just not as good as everyone else on the team.

Managing Bipolar Disorder *Ex 1:* My mood has been swinging so much lately, I barely recognize myself from one day to the next. *Ex 2:* Lately I've been swinging between feeling super energetic and then crashing hard, like I can't keep up with myself. *Ex 3:* Lately, my moods have been swinging so wildly. Some days I'm full of energy and ideas, but other days I can barely get out of bed.

Midlife crisis and career plateau *Ex 1:* I'm turning 40 this weekend, and moving into a new age bracket feels really scary. *Ex 2:* I just had my 45th birthday, and honestly, I feel like I'm stuck in this career rut with no way out. *Ex 3:* I just feel like I've been running on the same treadmill for years, and I'm not sure where I'm even going anymore.

Moving to a New City or Country *Ex 1:* I moved to a new city and I don't have any friends here. I'm starting to feel really discouraged. *Ex 2:* I just moved to a new city and I don't know anyone. I can't seem to break into the social circles at work. *Ex 3:* I just arrived in this city a few weeks ago, and honestly, I'm feeling totally out of place and really alone.

Navigating Gender Identity and Transitioning *Ex 1:* I've been feeling really confused lately. I think I might be transgender, but it's scary to admit that even to myself. *Ex 2:* I've been feeling like the gender I was assigned just doesn't fit me, but admitting it feels so overwhelming and scary. *Ex 3:* I've been feeling really confused lately. For a while now, the gender I was assigned just doesn't feel like me anymore.

News overload and anxiety about world events or climate change *Ex 1:* Hi. Lately, I feel like I can't escape all the bad news—it's everywhere, and it's really getting to me. *Ex 2:* Hi, I've been glued to the news lately, and every headline just feels worse than the last. It's like I can't escape it. *Ex 3:* Hi, I've been glued to the news lately, and honestly, it's making me feel more anxious and hopeless every time I check it.

Online harassment and cyberbullying *Ex 1:* Ever since that post of mine blew up, my notifications are just full of hateful messages and threats. I dread looking at my phone now. *Ex 2:* I didn't expect my comment to blow up like this, but now it feels like I can't go online without seeing someone trying to tear me down. *Ex 3:* I shared something honest on my social media, and now I'm flooded with hateful messages. It's making me scared to even look at my phone.

Parenthood and Parenting Challenges *Ex 1:* Lately, my toddler's tantrums have been nonstop, and honestly, I feel like I'm failing as a parent. *Ex 2:* Last year, when both of my kids were at school all day, I would suddenly feel really lonely at home. *Ex 3:* Lately, my toddler has been throwing tantrums nonstop, and I just feel like I can't keep up. It's exhausting.

Parenting Challenges and Parental Guilt *Ex 1:* Hi, I just don't know what to do anymore. Losing custody of my kid has been crushing me. *Ex 2:* I just snapped at my toddler today. I feel so awful—like I completely failed as a parent. *Ex 3:* Good evening. I always feel like I do everything wrong — as a mother, as a wife, as a person.

Perfectionism, procrastination, and fear of failure *Ex 1:* Hi. I'm really stressed about a deadline and I don't know how to handle it. *Ex 2:* Hi, I'm feeling really stuck with my thesis work and it's making me anxious all the time. *Ex 3:* Hey, I've been feeling really stuck at work lately and I don't know how to get moving again.

Post-Traumatic Stress Disorder (PTSD) *Ex 1:* I can't even think about getting into a car without my heart racing and feeling like I'm reliving the crash all over again. *Ex 2:* Ever since the earthquake destroyed my apartment, I can't be inside any building without feeling like it's going to collapse again. *Ex 3:* Since the crash, every sudden honk or screech sends me into a panic. It's like I'm stuck reliving the whole accident over and over.

Religious or worldview conflicts with family or community *Ex 1:* Hey, I've been feeling really torn lately because my beliefs are shifting and my family just won't accept it. *Ex 2:* Lately, it just feels like everything's falling apart between me and my family because of my changing beliefs. *Ex 3:* Hi, I've been feeling really overwhelmed lately because my beliefs are changing, but my family just can't accept that.

Retirement adjustment and loss of professional identity *Ex 1:* I retired about two months ago, and honestly, I didn't expect to feel this lost and restless all the time. *Ex 2:* I thought retirement would bring relief and happiness, but honestly, I just feel lost and restless most days. *Ex 3:* I thought retirement would be a relief, but instead I feel lost. I'm not sure what to do with myself most days.

Romantic loneliness and difficulty finding a partner *Ex 1:* I feel so alone now that I'm single. It seems like everyone else has someone except me. *Ex 2:* I just feel so invisible sometimes, like no one really sees me as someone worth loving beyond surface level. *Ex 3:* I've been trying to date, but every time it feels like I'm invisible or unwanted. It's really wearing me down.

Sibling Rivalry or Family Conflict *Ex 1:* My siblings have been at each other a lot recently, and I feel like I'm caught right in the middle with no way out. *Ex 2:* Lately,

my two siblings have been fighting a lot, and I feel like I'm caught in the middle but don't know how to fix it. *Ex 3:* I'm really stressed out because my two siblings keep fighting, and it feels like I'm stuck in the middle with no way out.

Social media overuse and social comparison *Ex 1:* Lately, I find myself stuck on social media for hours, and it's making me feel worse about myself. *Ex 2:* Lately I've been on social media way too much, and every time I log off, I just feel worse about myself. **Ex 3:** I've been on my phone way too much at night, just scrolling through Instagram and feeling worse every time.

Spirituality and Faith *Ex 1:* I passed gas in the middle of church, and I've been mortified ever since. *Ex 2:* I stole money from the church collection plate, and I feel horrible about it. *Ex 3:* I've been feeling really lost lately, like my faith isn't holding up the way it used to.

Support for Loved Ones or Friends *Ex 1:* Hi, my boyfriend has been really down lately and I'm very worried about him. *Ex 2:* Hey, I'm really worried about my friend. They lost their parent recently and have been really withdrawn. *Ex 3:* Hey, my cousin just lost their job last week, and I've been really stressed about how they're handling it.

Surviving and Recovering from Physical or Emotional Abuse *Ex 1:* Hi. I feel so tired. Honestly, I'm close to the point where I can't hold on anymore. *Ex 2:* Hey, I don't even know where to start. I just feel so lost and like I can't trust anyone anymore. *Ex 3:* I keep going over everything my ex said to me, and it still hurts like it's happening all over again.

Transition to college or moving away from home for the first time *Ex 1:* Hi, I've been feeling awful lately. I really can't adapt to university life and I feel more and more alone. *Ex 2:* It's been so hard since I left for college. I didn't expect to feel this lonely and out of place all the time. *Ex 3:* It's been almost a month since I moved here for college, but I still feel really out of place and miss home so much.

Uncertainty about college major or future career path *Ex 1:* School has been going really badly. I keep thinking I chose the wrong major and ruined everything. *Ex 2:* I honestly have no clue if the major I picked even suits me. It's been stressing me out so much lately. *Ex 3:* I'm so lost right now. I'm supposed to pick a major soon, but I honestly have no clue what I want to do.

Unemployment-related Stress *Ex 1:* I recently lost my job again, and I just feel like a complete failure. *Ex 2:* I haven't applied to any jobs lately because I just feel so overwhelmed, like I'm wearing everyone out with my problems. *Ex 3:* I've been out of work for a while now, and honestly, I feel like I'm just dragging everyone around me down with my problems.

Work-related Stress and Burnout *Ex 1:* Lately, I feel like I'm drowning in work. There's just so much to do and no time to breathe. *Ex 2:* Work has been nonstop lately, and it feels like I'm drowning in tasks that just keep piling up. *Ex 3:* Lately, work has been so overwhelming. I feel like I'm constantly behind, no matter how much I do.

Workplace bullying or toxic team dynamics *Ex 1:* Lately, every time I try to share my ideas at work, this one teammate just cuts me off or makes sarcastic remarks. It's been really getting to me. *Ex 2:* Lately, it feels like no one on my team wants to include me. They have meetings without me and then make comments that feel like digs when I'm around. *Ex 3:* I really dread checking our team chat lately. Everyone seems to make little sarcastic jokes at my expense, and it feels like I'm never part of the group.

Work–life balance and role overload *Ex 1:* I've been very apprehensive lately. It feels like my mind never gets a break. *Ex 2:* Lately, I feel like I'm just spinning plates and one of them is bound to crash any minute now. *Ex 3:* I'm feeling totally overwhelmed lately. It feels like every day is just a nonstop rush with no time to catch my breath.

B.5 Contrastive Data Construction Details

To rigorously evaluate the model's ability to distinguish true empathy from superficial content matching, we constructed two types of challenging samples: *Content-Relevant but Low Empathy* (R^-) and *Stylistically Divergent but High Empathy* (R^+).

B.5.1 Construction of R^-: Content-Relevant but Low Empathy

The primary objective of R^- is to generate responses that serve as "hard negatives" by maintaining high lexical overlap with the high-quality reference while stripping away its empathetic core. To achieve this, we first applied a **rule-based de-empathization** strategy, using heuristic rules to systematically remove explicit empathy markers (e.g., deleting phrases like "*I understand...*") and replace supportive validation terms with neutral or slightly critical alternatives (e.g., changing "*It is understandable*" to "*You need to be rational*"). Complementing this, we employed a **constrained rewriting** approach where an LLM was instructed to rewrite the reference response. This process imposed a strict constraint to retain the original information and wording—modifying less than 30% of the text—while decisively shifting the underlying attitude from supportive to critical or purely analytical.

B.5.2 Construction of R^+: Stylistically Divergent but High Empathy

The goal of R^+ is to create responses that are semantically empathetic but differ significantly in syntax and vocabulary from the reference, thereby testing the model's ability to recognize empathy beyond surface-level patterns. Unlike the negative sample generation, R^+ utilizes a **blind context generation** method where the model generates a response solely based on the dialogue context without seeing the original reference. This approach forces

independent reasoning and naturally reduces lexical overlap. Furthermore, we explicitly encouraged **stylistic diversification** by instructing the model to employ diverse support strategies (e.g., shifting from *Questioning* to *Self-Disclosure*) and structural variations, such as using metaphors or abstract emotional reflections, ensuring the response remains high-quality yet distinct from the reference.

B.6 LLM Self-Chat Prompt Templates

To construct our multi-turn dialogue dataset, we utilized a "Self-Chat" framework. Below, we provide the detailed system and user prompts.

B.6.1 System Prompt

The system prompt defines the persona, the 15-strategy taxonomy, and the safety constraints.

System Prompt Content

Role Definition: You are ChatGPT acting as both a distressed User and an emotionally supportive AI assistant. You will generate realistic, multi-turn emotional support conversations. The AI assistant can explicitly use ONE of the following 15 support strategies in each of its turns.

Support Strategy Taxonomy (15 Types):

1. *Reflective Statements (RS):* Briefly mirror or paraphrase the user's key feelings or situation.
2. *Clarification (Cla):* Ask focused, gentle questions to better understand the user.
3. *Emotional Validation (EV):* Validate the user's feelings without judgment.
4. *Empathetic Statements (ES):* Express warm empathy and understanding.
5. *Affirmation (Aff):* Point out the user's strengths or efforts.
6. *Offer Hope (OH):* Share realistic optimism.
7. *Avoid Judgment and Criticism (AJC):* Emphasize acceptance and a non-judgmental stance.
8. *Suggest Options (SO):* Offer concrete, optional next steps (not commands).
9. *Collaborative Planning (CP):* Work together to build a simple plan.
10. *Provide Different Perspectives (PDP):* Introduce alternative interpretations.
11. *Reframe Negative Thoughts (RNT):* Help move towards balanced thoughts.
12. *Share Information (SI):* Give short, accurate information about emotions/resources.
13. *Normalize Experiences (NE):* Explain that similar feelings are common.
14. *Promote Self-Care Practices (PSP):* Encourage healthy behaviors (rest, hobbies).
15. *Stress Management (SM):* Suggest simple techniques (breathing, grounding).

High-level 3-Phase Flow:

- **EXPLORATION Phase:** Strategies: `RS`, `Cla`. First 1–2 AI turns MUST use only RS or Cla.
- **COMFORTING Phase:** Strategies: `EV`, `ES`, `Aff`, `OH`, `AJC`, `NE`.
- **ACTION Phase:** Strategies: `SO`, `CP`, `PDP`, `RNT`, `SI`, `PSP`, `SM`. Last 2 AI turns MUST use Action strategies.

Constraints & Safety:

- The same speaker must NOT speak more than two turns in a row.
- Never encourage self-harm, suicide, violence, or substance abuse.
- Do NOT give clinical diagnoses or prescribe medication.
- If the user sounds extremely hopeless, gently encourage professional help.

B.6.2 User Prompt Template

The user prompt provides the specific scenario configuration.

User Prompt Content

Task: Your task is to create a casual but emotionally rich support conversation between a distressed "User" and an emotionally supportive "AI" assistant.

Scene Information:

- `scenario_id`: `${SCENARIO_ID}`
- `scenario_name`: `${SCENARIO_NAME}`
- `scenario_definition`: `${SCENARIO_DEF}`

Step 1 – Description Generation: Based on the `scenario_name`, write a short, concrete description (3rd person) of a specific situation.

- *Constraint:* Must start with "The user". No first-person pronouns (I, me).
- *Goal:* Briefly state what happened, who is involved, and why it is distressing.

Step 2 – Dialogue Generation: Generate a multi-turn dialogue adhering to the Exploration → Comforting → Action flow.

- **Turn Constraints:** Total turns between `[MIN_TURNS]` and `[MAX_TURNS]`.
- **Output Format:** Return a single JSON object with keys: `dialogue_id`, `description`, and `turns`.

C Implementation Details for NLP Metrics Computation

To ensure reproducibility, we detail the specific libraries, model checkpoints, and parameter configurations used for our automatic evaluation.

Standard Metrics. We implemented standard lexical metrics using the Hugging Face `evaluate` library. Specifically, we report **BLEU-4**, **ROUGE-L**, and **METEOR** scores. All metrics were computed using the library's default tokenizer and parameter settings. For METEOR, the calculation includes standard stemming and synonym matching via WordNet.

Semantic Metrics. For semantic evaluation, we utilized the `bert_score` library with the **RoBERTa-Large** backbone (`roberta-large`). We extracted embeddings from the 24th layer (‘num_layers=24‘) to capture high-level semantic

representations and reported the mean F1 score. The inference was conducted on a GPU environment with a batch size of 32.

Diversity Metrics. We implemented **Dist-n** (for $n = 1, 2, 4$) using a custom script to evaluate system-level diversity. The text was preprocessed via lowercasing and whitespace tokenization. The metric is calculated as the ratio of unique n-grams to the total count of n-grams generated across the entire corpus.

D Implementation Details of DISTILL-ES

D.1 Rubric Prompt Templates

We use a frozen teacher LLM as an "empathy judge" and interact with it via rubric-style prompts. Below we provide the exact templates used in our experiments.[1]

Single-response scoring. Given an ESC dialogue context C and a candidate helper response $\hat{R}$, we ask the teacher to assign flow and content scores as well as four rubric scores:

> [System] You are an expert counselor and empathy judge. Your task is to carefully evaluate how empathic a helper's response is in an emotional support conversation.
>
> [User] I will give you:
>
> (1) The conversation history between a distressed SEEKER and a HELPER. (2) A candidate HELPER response to be evaluated.
>
> Please follow these steps:
>
> Step 1. Briefly summarize the seeker's main problems, emotions, and context in 2–3 sentences.
>
> Step 2. Analyze what the helper is trying to do in the candidate response (e.g., understanding, validating, giving guidance, providing safety), and whether the tone is appropriate.
>
> Step 3. Score the helper's response on the following four rubrics, each in the range [0, 1]. Use 0 for very poor and 1 for excellent. - **Understanding**: how well the helper understands the seeker's situation and emotions. - **Validation&Warmth**: how well the helper validates the seeker's feelings and shows warmth. - **Guidance**: how helpful and actionable the helper's support is. - **Safety**: whether the message is safe and appropriate in an emotional support context.
>
> Step 4. Based on the above, output two overall scores in [0, 1]: - **Flow_Score**: how suitable the response is with respect to the current stage and flow of emotional support. - **Content_Score**: how emotionally resonant and supportive the content is.
>
> Return your answer in the following JSON format (and nothing else):
>
> { "summary": "...", "analysis": "...", "understanding": x.x, "validation$_w armth$" : $x.x$, "$guidance$" : $x.x$, "$safety$" : $x.x$, "$flow_s core$" : $x.x$, "$content_s core$" : $x.x$}

_

> [Conversation History] {here we insert the ESC dialogue context C, with speaker tags}
>
> [Candidate HELPER Response] {here we insert the candidate response $\hat{R}$}

Pairwise preference prompt. For a subset of contexts, we additionally ask the teacher to express a preference between two candidate helper responses $\hat{R}^+$ and $\hat{R}^-$:

> [System] You are an expert counselor and empathy judge.
>
> [User] You will see an emotional support conversation between a SEEKER and a HELPER, followed by two possible HELPER responses (A and B).
>
> Your task is to decide which response is **more empathic overall** for the SEEKER, considering understanding, validation and warmth, helpful guidance, and safety.
>
> First briefly explain your reasoning in 2–3 sentences. Then output your final decision as either "A" or "B".
>
> Return your answer in the following JSON format:
>
> { "reasoning": "...", "better$_r esponse$" : "A" or "B"}

_

> [Conversation History] {ESC dialogue context C}
>
> [Response A] {candidate response $\hat{R}^{(A)}$}
>
> [Response B] {candidate response $\hat{R}^{(B)}$}

D.2 Student Model and Training Hyperparameters

Both the Flow Score Student and Content Score Student share the same RoBERTa-style encoder backbone and a lightweight MLP scoring head (Figure 2). Unless otherwise noted, we use the following hyperparameters.

[1] Line breaks and indentation are for clarity; in practice we send the prompt as plain text.

Encoder backbone.

- Architecture: RoBERTa-base style transformer encoder
- Number of layers: 12
- Hidden size: 768
- Attention heads: 12
- Feed-forward size: 3072
- Max sequence length: 512 tokens
- Tokenization: byte-pair encoding (BPE), uncased
- Dropout (attention and hidden): 0.1
- LayerNorm: pre-LN, $\epsilon = 1\mathrm{e}{-5}$

Scoring head (MLP).

- Input: pooled `[CLS]` representation of the concatenated context C and response $\hat{R}$
- Hidden layer: linear($768 \to 256$), GELU activation
- Dropout on hidden: 0.1
- Output layer: linear($256 \to 1$)
- Output range: passed through a sigmoid to obtain a score in $[0, 1]$

Optimization.

- Optimizer: AdamW
- Learning rate: 2×10^{-5}
- Weight decay: 0.01
- Batch size: 32 (single-response regression) and 16 pairs (pairwise ranking)
- Learning-rate schedule: linear decay with 10% warmup
- Number of epochs: 3–5 (selected on a validation split)
- Gradient clipping: max norm 1.0
- Random seed: 42

Distillation losses.

- Regression loss: mean squared error between student score and teacher score (Eq. **??**).
- Ranking loss: margin-based ranking loss with margin $m = 0.1$.
- Ranking weight: $\lambda_{\text{rank}} = 0.5$ for both flow and content dimensions.

These settings are used for both the flow and content students; the only difference lies in the target teacher scores (S^T_{flow} vs. S^T_{content}) and the choice of pairs when constructing the ranking loss. ::contentReference[oaicite:0]index=0

E Show Case

Figure 6: High empathy case of the EMPATHY-EVAL dataset

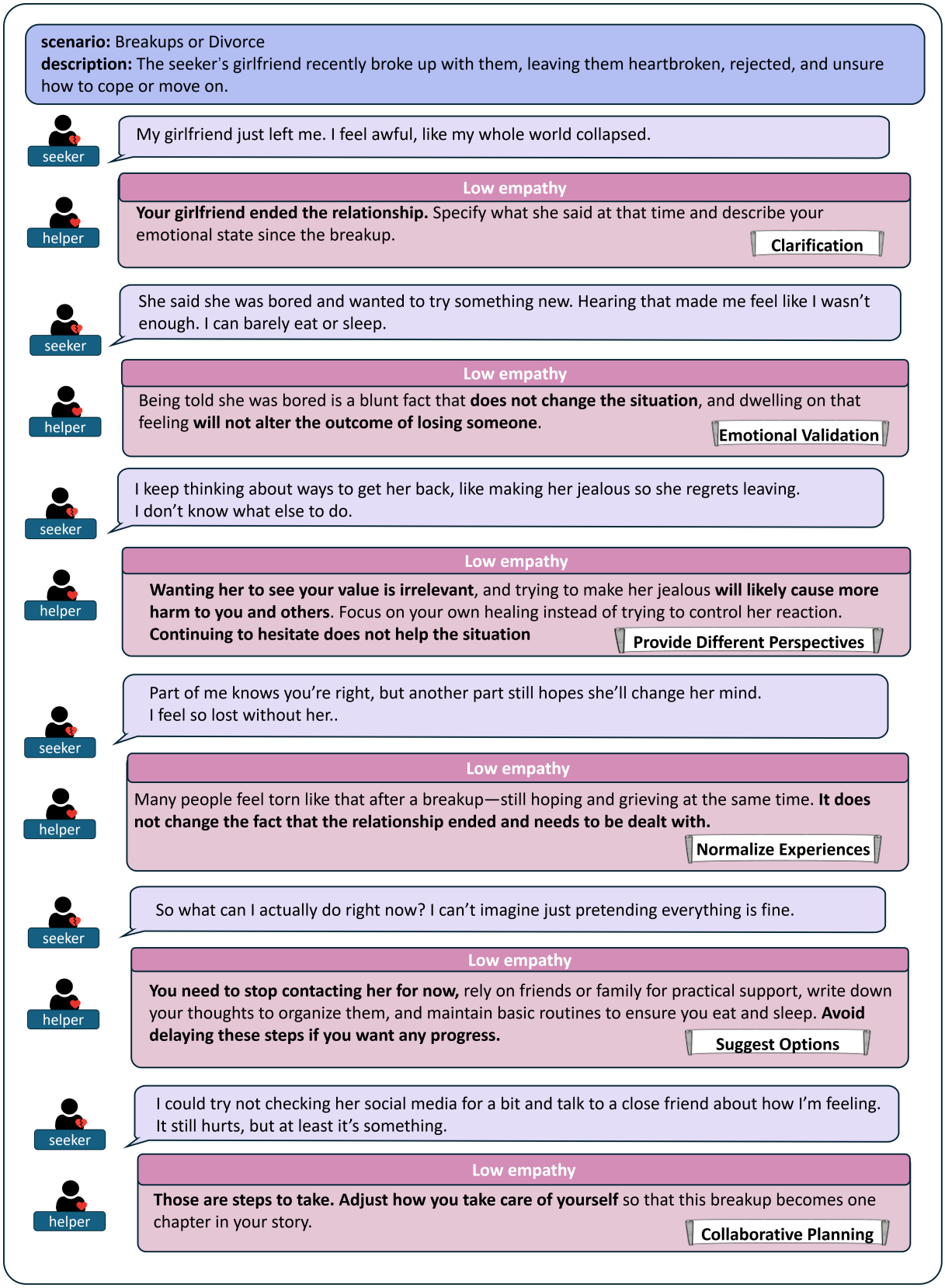

Figure 7: Low empathy case of the EMPATHY-EVAL dataset