MvP-ECR: Multi-Perspective Emotion-Cause Reasoning for Empathetic Dialogue

Yuanyuan He[1,2], Guotai Huang[1], Wei Li[1], Jiali You[1], Jiawen Deng[1]*, Fuji Ren[1,2]*

[1]University of Electronic Science and Technology of China, Chengdu, China
[2]Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China
{hyy, gthuang, liwei39}@std.uestc.edu.cn,
yjl1015004@163.com, {dengjw, renfuji}@uestc.edu.cn

Abstract

The empathetic dialogue systems aim to recognize user emotions and generate appropriate empathetic responses. However, existing approaches predominantly rely on dialogue history, contextual descriptions, and emotion category labels failing to model the causal relationship between emotions and their underlying triggers. This limitation leads to generated responses that lack grounding, exhibit weak relevance and suffer from poor interpretability in emotional expression To address this, we propose MvP-ECR, a multi-perspective emotion cause reasoning framework that explicitly constructs emotion-cause structures to help models focus on the core emotional drivers. Additionally, we introduce an emotion-cause consistency evaluation metric to quantitatively assess a model's ability to identify causal relationships. Experiments across multiple large language models (LLMs) demonstrate that the MvP-ECR framework can serve as a plug-and-play tool to help the model correctly infer emotions and causes in empathetic conversations, and provide more immersive responses for empathetic responses. All code and data will be publicly released to promote the development of empathy dialogue research.

1 Introduction

The empathetic dialogue system aims to understand users' emotions and generate effective comfort and support, helping users alleviate negative emotions (Huang et al. 2024). Although a large amount of work has focused on emotion classification and recognition (Wang et al. 2025b; Ma et al. 2024; Zheng et al. 2024; Tu et al. 2024), the causal mechanism of "how emotions are triggered" has been rarely explored, which is the key to building dialogue subjects that match human empathy (Hsu et al. 2023; Chen et al. 2024a).

Existing work mainly focuses on enhancing the richness of input information, such as introducing common sense knowledge bases (Chen et al. 2024a), self-awareness modeling (Zhao et al. 2023), multi granularity semantic encoding (Zhou et al. 2023; Tu et al. 2022), and cognitive relevance principles (Li et al. 2024). Although these methods expand the model's understanding of context (Wang et al. 2025a; Chen et al. 2025), they also to some extent distract the model's attention from emotional causal logic.

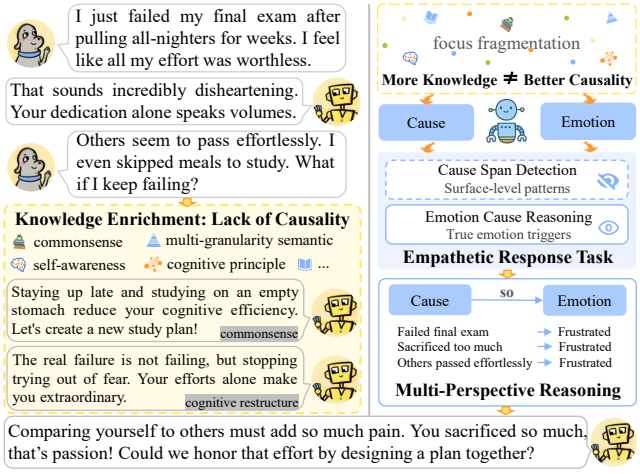

Figure 1: The user expressed frustrated after failing the final exam. If only information augmentation is performed without clear Causality, the model may respond around the introduced knowledge (*e.g.*, commonsense, cognition) but ignore the emotional roots of the user (fail, comparison and effort).

As shown in Figure 1 (left), when responding to a user who failed an exam, injecting commonsense causes the model to focus on the impact of sleep and meals; applying cognitive restructuring shifts the model's attention to changing the user's perception of failure. Without modeling the causal chain, empathic responses lack clear causality, which may lead the model to overemphasize factual details while neglecting the user's emotional reasoning.

Recent research has recognized that emotions and their causes are inseparable (Gui et al. 2018; Xiao, Xia, and Yu 2023; Cheng et al. 2023; Xia and Ding 2019), as Figure 1 (right), leading to the development of tasks such as cause-span detection (An et al. 2023; Chen et al. 2023, 2024a,b; Zou et al. 2024; Su et al. 2024). However, span-based approaches are inherently limited to surface-level matching and fall short in modeling the reasoning processes humans use to infer emotional causality from context, intentions, and commonsense knowledge. Only after establishing clear causal logic chains (Liu et al. 2025) can the model provide more comprehensive and coherent empathetic responses.

*Corresponding authors.

To address this above issues, we propose a multi-perspective emotion cause reasoning framework (MvP-ECR). This framework aims to identify the dominant emotions of target users and their underlying causes from multiple rounds of empathetic conversations, simulating the causal reasoning that humans engage in during the empathetic process. This framework consists of three complementary reasoning paths, each of which models emotional states and their causes from different cognitive perspectives, providing a structured and logically interpretable emotional causal chain for empathetic dialogue.

In order to comprehensively evaluate the inference performance of MVP-ECR, we further propose a progressive emotion reason consistency evaluation framework, which measures the causal consistency and logical rationality between the model output and the dialogue context from three levels: vocabulary matching, sentence association, semantic understanding and thus more comprehensively characterizes the empathy reasoning ability of the model. In addition, to standardize the results of multi-path reasoning, we propose an emotion reason quadruple representation and explicitly construct an emotional causal logic chain, providing interpretable structured guidance for downstream generation tasks. Our contributions are summarized as follows:

- **M**ulti-**P**erspective **E**motion-**C**ause **R**easoning (MvP-ECR) Framework: The MvP-ECR is proposed to guide LLMs to infer users' dominant emotions and their potential causes from different cognitive paths, compatible with multiple LLMs.
- **W**ord-**S**entence-**M**eaning (WSM) Evaluation: Construct a three-level empathetic causal consistency evaluation for words, sentences, and meaning, providing a more explanatory evaluation method for subsequent research.
- Empirical Verification: Extensive experiments on multiple benchmarks have demonstrated significant improvements in the quality of emotional and causal reasoning and generation.

2 Related Work

2.1 Empathetic Dialogue

Empathetic dialogue is a communication method centered on understanding and empathy (Davis 1983; Lin et al. 2019; Li et al. 2020; Huang et al. 2024; Qian, Zhang, and Liu 2023; Hsu et al. 2023; Chen et al. 2024a), aiming to make the other party feel accepted and understood through deep listening, emotional resonance, and sincere responses (Keskin 2014; Zhou et al. 2018; Hutto and Gilbert 2014). The existing research on empathetic dialogue primarily focuses on enhancing the richness of conversational information. One approach involves enriching dialogue content by incorporating external knowledge sources, such as commonsense knowledge bases (Chen et al. 2024a; Tu et al. 2022; Cai et al. 2023), distilled model knowledge (Chen et al. 2024b), or constructing concept and cognitive graphs (Zhou et al. 2023; Qiao et al. 2025), to enhance the model's comprehension of emotions and the diversity of empathetic responses. Another approach is to introduce concepts or structures to construct empathy logic, including cognitive relevance principle (Li et al. 2024), self-other differentiation (Zhao et al. 2023), and response hierarchy frameworks (Zheng et al. 2021), providing more interpretable answers for empathetic responses.

2.2 Emotion-Cause Label

Emotion-cause understanding plays a pivotal role in empathic dialogue, enabling dialogue systems to better identify the logic behind emotions and generate more targeted empathetic responses (Sabour, Zheng, and Huang 2022; Gupta and Dandapat 2023; He et al. 2025). Early research primarily focused on emotion-cause extraction tasks in narrative texts, such as event-based sentiment analysis (Lee, Chen, and Huang 2010), event-emotion correlation modeling (Gui et al. 2018), and fine-grained emotion-cause extraction (Xia and Ding 2019). In recent years, some study attempted to improve emotion-cause consistency modeling (Xiao, Xia, and Yu 2023; Cheng et al. 2023; Su et al. 2023), but their work remained confined to one-sided narratives rather than bidirectional conversational scenarios. To address this gap, the RECCON dataset (Poria et al. 2021), which combines dialogue data from IEMOCAP and DailyDialog, annotates emotions along with their corresponding cause spans. The task of detecting the span of reasons (An et al. 2023; Chen et al. 2023; Zou et al. 2024; Su et al. 2024) has promoted research in the extraction of emotional reasons.

3 MvP-ECR Framework

We propose a plug-and-play **M**ulti-**P**erspective **E**motion-**C**ause **R**easoning (**MvP-ECR**) framework, as shown in Figure 2 (a1-a4), which dynamically models the causal relationships between emotions and events through three cognitive pathways (EC, CE, SA) and generates interpretable quadruple representations. To address potential reasoning biases in weakly supervised settings, we further design the **W**ord-**S**entence-**M**eaning (**WSM**) evaluation mechanism that filters contextually consistent emotion-cause pairs through word-level, sentence-level, and meaning-level verification, balancing rigor and adaptability to weak supervision. We will introduce MvP-ECR and WSM separately below.

3.1 Multi-Perspective Emotion-Cause Reasoning

To comprehensively model the causal dynamics between emotions and events in empathetic dialogues, we propose a multi-perspective reasoning framework comprising three inference paths: Emotion $\rightarrow$ Cause (EC), Cause $\rightarrow$ Emotion (CE), and Summarize $\rightarrow$ Abstract (SA), as shown in Figure 2-b1. Each path represents a distinct cognitive trajectory for constructing interpretable emotion cause pairs: retrospective, forward causal, and holistic abstraction. We formally describe each reasoning strategy below.

Emotion to Cause (EC) The EC pathway is a form of reverse reasoning that simulates a retrospective process by prioritizing emotions over reasons. In this process, the model traces the speaker's emotional expression back to its root cause. This reasoning path follows a four-stage pipeline: Target Identification, Emotion Abstraction, Cause Inference and Dual Evidence Extraction.

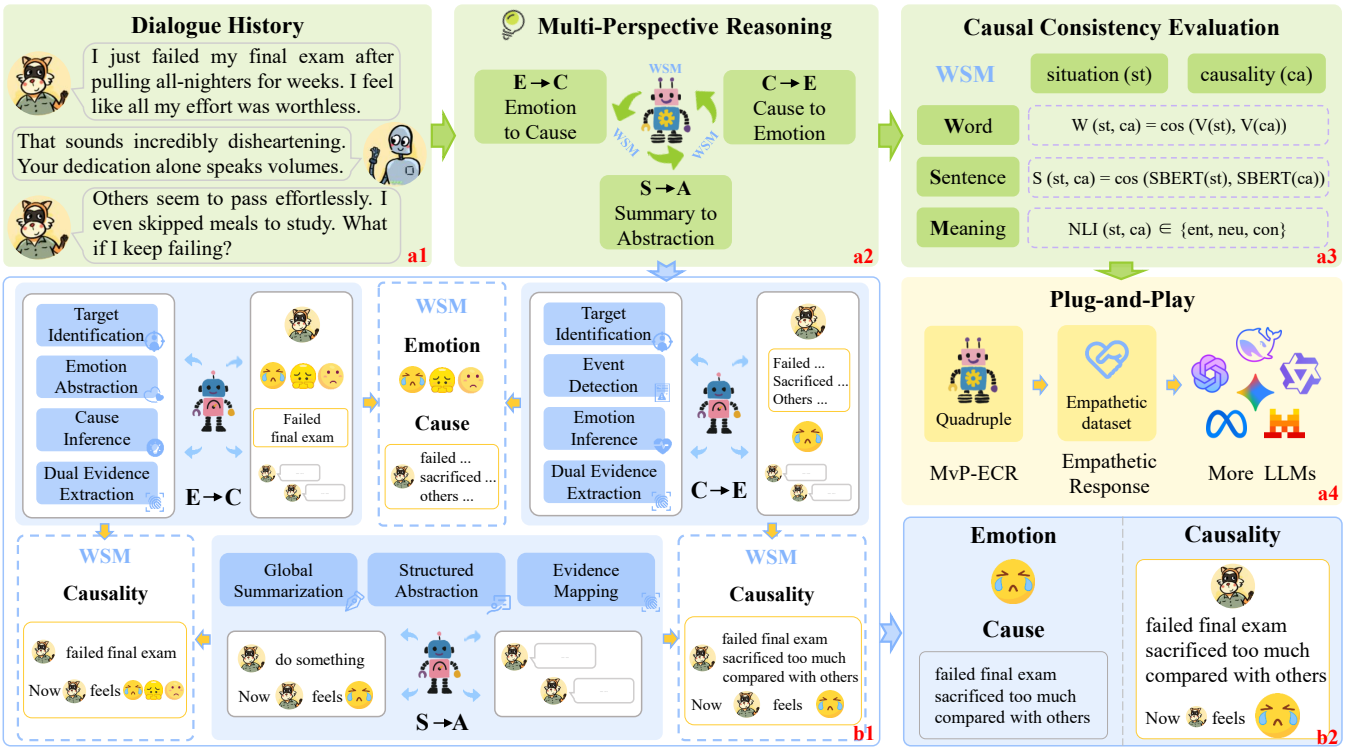

Figure 2: The MvP-ECR framework. a1-a4 are the complete inference evaluation process, and b1-b2 are the detail of reasoning, including EC, CE and SA.

Step 1: Target Identification. Given a dialogue $D = \{u_1, u_2, \ldots, u_n\}$, where each u_t represents a turn with speaker identity and textual content, the model identifies the target speaker, the person most likely to be emotionally affected or seeking empathy.

Step 2: Emotion Abstraction. The model infers the dominant emotion e^* from a predefined emotion set $\mathcal{E}$, based on utterances from speaker s^*, ϕ_{emotion} measures the fit between emotion e and the utterance:

$$e^* = \arg\max_{e \in \mathcal{E}} \phi_{\text{emotion}}(e \mid \{u_t \in D \mid s^*\}) \tag{1}$$

Step 3: Cause Inference. Given the inferred emotion e^*, the model selects a structured, interpretable cause c^* from a template set $\mathcal{C}$. ϕ_{cause} measures the fit between cause c and the utterance:

$$c^* = \arg\max_{c \in \mathcal{C}} \phi_{\text{cause}}(c \mid e^*, D), \quad \mathcal{C} = \{\text{Speaker did something. Speaker feels } e^*\} \tag{2}$$

Step 4: Dual Evidence Extraction. To support the inferred emotion-cause pair, the model retrieves two key utterances from the dialogue. The first is an utterance that most clearly conveys the target emotion (u^{emo}), selected based on its emotional expression. The second is an utterance that best supports or explains the identified cause (u^{cause}).

Cause to Emotion (CE) The CE path represents a forward Causal reasoning strategy, starting from events toward emotional consequences. This process also includes four stages: Target Identification, Event Detection, Emotion Inference and Dual Evidence Extraction.

Step 1: Target Identification. As in the EC path, the model first determines who the emotionally affected speaker is by evaluating which participant reveals the most emotional content in the dialogue.

Step 2: Event Detection. The model identifies the most salient event or situation c^* relevant to s^* , such as a conflict, failure, or challenge:

$$c^* = \arg\max_{c \in \mathcal{C}} \phi_{\text{cause}}(c \mid D, s^*) \tag{3}$$

Step 3: Emotion Inference. From the identified cause c^*, the model infers the most plausible emotion e^* :

$$e^* = \arg\max_{e \in \mathcal{E}} \phi_{\text{emotion}}(e \mid c^*, D) \tag{4}$$

Step 4: Dual Evidence Extraction. To ground its reasoning, the model selects one utterance that best supports the inferred cause u^{cause} and another that most clearly conveys the resulting emotion u^{emo}.

Summarize to Abstract (SA) The SA path employs a semantic compression and abstraction approach, where global understanding of the dialogue is abstracted into a concise, interpretable structure. There are three steps: Global Summarization, Structured Abstraction and Evidence Mapping.

Step 1: Global Summarization. The model generates a brief summary S (2–3 sentences) of the target user's experience and emotional state firstly:

$$S = \texttt{Summarize}(\{u_t \in D \mid s^*\}) \tag{5}$$

Step 2: Structured Abstraction. From the summary, a standardized abstract is produced in the format:

$$\hat{A} = \text{“Speaker did something. Now, speaker feels } e^*\text{.”} \tag{6}$$

Step 3: Evidence Mapping. To ground the results, the model identifies u^{emo} and u^{cause} from the dialogue.

Merge Reasoning Path Overall, the EC and CE paths are capable of inferring the target user's emotion and cause from a local perspective, while the SA path provides a global view to summarize the most influential emotion and its underlying cause. As shown in Figure 2-b2, we merge the results from all three paths and select the most consistent emotion-cause quadruple(s) by computing the WSM-based consistency between the causal logic ca and the dialogue situation st. The quadruples Q are represented as follows:

$$\mathcal{Q}_{ca} = (e^*, c^*, u^{\text{emo}}, u^{\text{cause}}) \tag{7}$$

The MvP-ECR framework can be transferred to multiple LLMs to infer the emotions and causes of empathetic conversations.

3.2 Word-Sentence-Meaning Evaluation

In empathetic dialogue, the lack of annotated emotion-cause labels can lead to semantic drift, hallucinations, or contextual mismatch in model inference. To address this, we propose a progressive three-level evaluation framework: Word-Sentence-Meaning (WSM), which verifies inference quality through keyword matching (W), semantic consistency (S), and natural language inference (M), as shown in Figure 2-a3. Given the subtle and context-dependent nature of emotional expression, we adopt a lenient criterion: an inference is considered valid if it passes any one of the three verification levels. The evaluation formula is defined as follows:

$$\text{Score}_{\text{wsm}} = \begin{cases} 1 & \text{if } W(ca, st) \geq \tau_w, \\ 1 & \text{if } S(ca, st) \geq \tau_s, \\ 1 & \text{if } M(st, ca) = entailment, \\ 0 & \text{else.} \end{cases} \tag{8}$$

τ_w and τ_s are the average values of the elbow rule and the 30% descent rule, which we will verify in the experiments.

Word Verification To assess lexical-level consistency, we calculate the cosine similarity between the averaged word vectors of the inferred emotion-cause text (ca) and the dialogue situation (st). After tokenization, we aggregate word embeddings and compute semantic closeness as follows:

$$W(ca, st) = \frac{\sum_{i=1}^{n} \mathbf{v}_{ca_i} \cdot \sum_{j=1}^{m} \mathbf{v}_{st_j}}{\left\| \sum_{i=1}^{n} \mathbf{v}_{ca_i} \right\| \left\| \sum_{j=1}^{m} \mathbf{v}_{st_j} \right\|} \tag{9}$$

Here, $\mathbf{v}_{ca_i}$ and $\mathbf{v}_{st_j}$ denote token embeddings, while n and m are the respective token counts.

Sentence Verification To capture semantic consistency beyond lexical overlap, we use Sentence-BERT to encode the inferred emotion-cause statement (ca) and dialogue situation (st), then compute their cosine similarity:

$$\begin{aligned} \text{S}(ca, st) &= \cos(\text{SBERT}(ca), \text{SBERT}(st)) \\ &= \frac{\mathbf{h}_{ca} \cdot \mathbf{h}_{st}}{\|\mathbf{h}_{ca}\| \|\mathbf{h}_{st}\|} \end{aligned} \tag{10}$$

$\text{SBERT}(\cdot)$ represents the Sentence-BERT encoder. $\mathbf{h}$ is a semantic vector representation at the sentence level.

Meaning Verification Lexical and sentence-level evaluations focus on surface alignment but cannot fully capture causal reasoning. To address this, we introduce a DeBERTa-based natural language inference model that treats the dialogue context as premise (p) and the inferred emotion-cause statement as hypothesis (h). The output reflects their logical relation:

$$\text{M}(st, ca) = \text{NLI}(p, h) \in \{\text{ent}, \text{neu}, \text{con}\} \tag{11}$$

where ent represents entailment, neu represents neutral, and con represents contradiction.

4 Experiments

We evaluate our approach on ESConv (Liu et al. 2021), an English empathetic dialogue dataset containing 34K multi-turn conversations about negative emotions (*e.g.*, sadness, anxiety). The dataset includes emotion labels, problem types, and dialogue contexts. We use an 8:1:1 train/val/test split, excluding dialogues exceeding 40 turns.

We evaluate both closed-source and open-source LLMs: Closed-source: DeepSeek-R1 (Guo et al. 2025), GPT-3.5 (Brown 2020), GPT-4o (Achiam et al. 2023), and Gemini-1.5 (Team et al. 2023). Open-source: LLaMA-3 (3B/8B) (Dubey et al. 2024), Mistral-7B (Jiang et al. 2023), and Qwen-2.5 (7B/14B) (Bai et al. 2023), covering diverse sizes and architectures.

4.1 Automatic Evaluation

We conduct both zero-shot and few-shot experiments on closed-source and open-source LLMs. The evaluation metrics include emotion classification accuracy (ACC) and the consistency metric (WSM). As shown in the Table 1, MvP-ECR consistently achieves the highest performance across nearly all models under both settings. For instance, in the zero-shot setting, MvP-ECR improves ACC by 6–12 points compared to the vanilla baseline on models such as DeepSeck-R1 and Qwen2.5-14B. Similarly, WSM scores are substantially improved, with MvP-ECR reaching up to 94.69 on Qwen2.5-14B, demonstrating superior alignment between inferred emotion-cause chains and contextual background. Similar experimental results are also clearly visible in the few-shot setting.

Three key conclusions can be drawn from these results in Table 1. First, multi-perspective reasoning significantly enhances both accuracy and logical consistency compared to single-path inference and baseline generation. Second, MvP-ECR shows strong transferability and robustness,

		Models	ACC					WSM				
			Vanilla	EC	CE	SA	**MvP-ECR**	Vanilla	EC	CE	SA	**MvP-ECR**
Closed-LLMs	*zero-shot*	DeepSeek-R1	50.63	55.99	51.20	53.98	**62.41**	60.76	68.35	84.81	79.75	**92.41**
		GPT-3.5	49.85	53.15	47.38	48.46	**60.38**	71.08	76.39	69.82	80.48	**88.78**
		GPT-4o	54.31	56.38	52.62	53.31	**63.08**	77.92	84.25	79.92	77.55	**92.12**
		Gemini-1.5	51.77	52.92	52.00	52.31	**57.15**	71.08	72.45	73.72	77.25	**85.84**
	few-shot	DeepSeek-R1	-	52.31	53.08	55.38	**63.85**	-	70.00	67.69	73.85	**83.08**
		GPT-3.5	-	48.46	48.46	52.31	**63.08**	-	78.46	80.00	80.00	**93.08**
		GPT-4o	-	52.31	48.46	50.00	**58.46**	-	80.00	77.69	73.85	**88.46**
		Gemini-1.5	-	53.85	53.08	53.08	**57.69**	-	73.08	73.08	73.08	**84.62**
Open-LLMs	*zero-shot*	LLAMA3.2-3B	42.08	41.31	36.77	31.62	**46.31**	71.92	80.03	65.32	68.42	**90.25**
		LLAMA3.1-8B	53.38	52.54	43.62	46.92	**58.69**	81.62	73.11	68.11	78.85	**88.41**
		Mistral-7B	41.59	38.94	38.05	38.05	**53.10**	71.68	81.42	80.53	78.76	**95.58**
		Qwen2.5-7B	50.77	52.62	44.00	47.85	**59.92**	73.31	71.69	74.56	78.59	**89.48**
		Qwen2.5-14B	57.52	55.75	41.59	40.71	**61.06**	81.84	84.96	83.19	71.68	**94.69**
	few-shot	LLAMA3.2-3B	-	44.25	45.13	43.36	**50.44**	-	78.76	74.34	73.45	**91.15**
		LLAMA3.1-8B	-	50.44	51.33	55.75	**59.29**	-	82.30	82.30	84.96	**95.58**
		Mistral-7B	-	42.48	38.94	40.71	**46.02**	-	84.96	71.68	79.65	**92.92**
		Qwen2.5-7B	-	46.02	44.25	44.25	**48.67**	-	80.53	73.45	76.11	**86.73**
		Qwen2.5-14B	-	54.87	50.44	48.67	**57.52**	-	84.96	77.88	86.73	**93.81**

Table 1: Accuracy (ACC) and word-sentence-semantic consistency (WSM) results for different reasoning strategies across closed-source and open-source LLMs under zero-shot and few-shot settings. MvP-ECR consistently outperforms other methods in both metrics, demonstrating superior emotion-cause reasoning capability. **Bold** indicates the best, followed by underline.

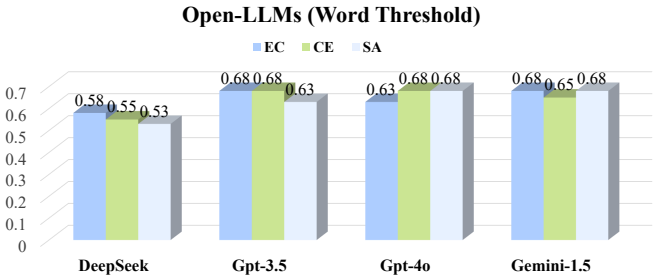

Figure 3: A set of 'W' threshold for open-LLMs.

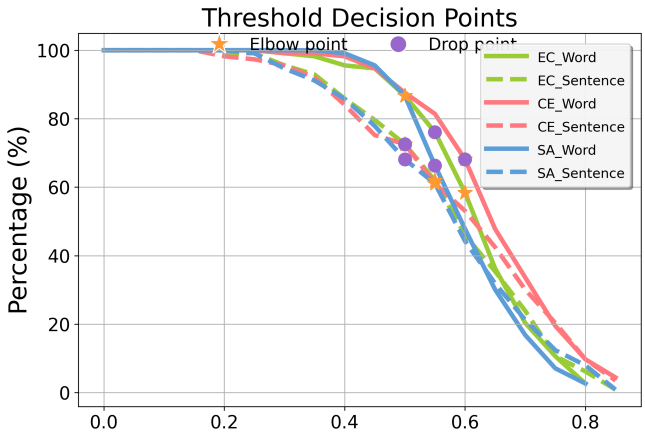

Figure 4: The threshold of 'W' and 'S' using elbow method and 30% drop heuristic method on LLAMA3.1-8B.

maintaining high performance across various model scales and types. Third, the framework's improvements are more prominent under the few-shot setting, suggesting that MvP-ECR benefits from minimal supervision while leveraging model reasoning capacity more effectively.

4.2 Analysis of WSM

Having demonstrated MvP-ECR's overall superiority in Table 1, we now dissect its core component, the WSM evaluation framework, to reveal how multi-level verification contributes to performance gains. As shown in Table 2, significant performance variations exist across models in word-level (W), sentence-level (S), and semantic-level (M) metrics. For instance, DeepSeek-R1 scores 74.68 on the CE path under the W metric but only 20.25 under the M metric for the same path, demonstrating that single metrics cannot comprehensively reflect model capabilities. This divergence confirms the complementary value of WSM's hierarchical evaluation. while word- and sentence-level metrics capture local matching features, the semantic-level metric reveals deeper logical consistency, with their combination covering broader assessment dimensions. Furthermore, Table 1 shows that the integrated WSM results outperform all single-metric outcomes (*e.g.*, zero-shot GPT-4o's WSM score of 84.25 exceeds its W:63.63, S:75.21, and M:27.72 scores), proving that multi-level joint filtering better balances formal and substantive requirements in dialogue scenarios, avoiding biases from single-metric optimization.

WSM's dynamic threshold mechanism further enhances its adaptability and fairness across models. By combining the elbow method (which detects inflection points) with

Models	W				S				M			
	Vanilla	EC	CE	SA	Vanilla	EC	CE	SA	Vanilla	EC	CE	SA
Closed-LLMs:												
DeepSeek-R1	48.10	67.09	**74.68**	73.42	56.96	43.04	**84.81**	64.56	13.92	**29.11**	20.25	21.52
GPT-3.5	55.46	56.27	52.52	**61.43**	59.69	69.12	57.02	**73.04**	9.77	8.05	12.02	**16.58**
GPT-4o	59.31	**63.63**	48.57	59.88	67.31	**75.21**	68.42	63.43	17.54	**27.72**	20.39	15.51
Gemini-1.5	58.00	55.65	**63.80**	58.46	54.92	61.53	54.57	**69.49**	**39.23**	37.00	38.06	39.05
Open-LLMs:												
LLAMA3.2-3B	48.15	**74.77**	53.06	54.52	60.23	**61.92**	51.59	57.58	**12.62**	6.35	10.24	6.13
LLAMA3.1-8B	67.00	57.88	56.17	**69.95**	**74.31**	62.44	54.62	64.14	23.46	**24.11**	16.99	16.42
Mistral-7B	60.18	49.56	**65.49**	63.72	58.41	**73.45**	67.26	62.83	15.93	11.5	12.39	**19.47**
Qwen2.5-7B	**65.54**	60.48	58.86	62.44	55.46	58.86	63.42	**68.78**	18.54	22.97	16.55	**18.93**
Qwen2.5-14B	74.80	**75.22**	61.95	56.64	66.12	**69.91**	67.26	51.33	**44.99**	37.17	23.89	23.89

Table 2: Fine-grained evaluation results of word-level (W), sentence-level (S), and semantic-level (M) consistency across different reasoning paths and models. The variation across levels highlights the necessity of multi-level assessment in capturing both surface alignment and deep causal coherence. **Bold** indicates the best, followed by underline.

Judges	Inf	Cau	Rat	Avg	PCC (vs LLM)
Human-1	4.320	4.440	4.400	4.390	0.987
Human-2	4.320	4.960	5.000	4.760	0.999
Human-3	4.280	4.160	4.040	4.160	0.982
Human-avg	4.310	4.520	4.480	4.440	0.998
LLM	4.760	4.880	4.960	4.870	-
PCC	0.998	0.998	0.998	0.999	-

Table 3: Human and LLM evaluation of emotion-cause reasoning quality. The range of PCC values is 0-1.

a 30% drop heuristic (which prevents overfitting to local peaks), the framework mitigates the sensitivity of traditional thresholds to noise and uneven distributions. As shown in Figures 3 and 4, threshold values vary substantially across models. For example, W thresholds differ by up to 15% between DeepSeek and GPT-3.5, underscoring the inadequacy of fixed thresholds and the necessity of model-specific calibration. This adaptive design enables WSM to robustly evaluate emotional reasoning performance across diverse models and dialogue scenarios.

4.3 Human and LLM Evaluation

To validate the reliability of the reasoning results generated by the MvP-ECR framework, we used a combination of human and LLM (GPT-4o) evaluation to conduct multidimensional quantitative analysis of the reasoning results from 50 randomly selected dialogues. The evaluation system consists of three core indicators: Information (Inf) evaluates the generalizability of content, Causality (Cau) measures logical rationality, and Rationality (Rat) examines the conformity of commonsense, all using a 1-5 scale (higher scores indicate better quality). Additionally, to assess the consistency among evaluators, we calculated the Pearson correlation coefficient (PCC) between their ratings.

The experimental data in Table 3 reveals three key findings: First, the MvP-ECR framework demonstrates exceptional reasoning quality, with human evaluation averaging 4.44 points and LLM evaluation reaching 4.76 points. Notably, the Causality (Cau=4.88) and Rationality (Rat=4.96) scores approach full marks, confirming that the generated causal chains closely align with human cognitive patterns. Second, while the framework shows stable content summarization capability (Inf score standard deviation of only 0.02), some variability exists in Causality scores (4.16-4.96) for complex scenarios, reflecting the dependence of deep emotion reasoning on subjective experience. Most importantly, the assessments show remarkable consistency between humans and LLM (PCC=0.998, $p<0.001$), with correlation coefficients exceeding 0.98 for all individual evaluators, thoroughly validating both the objectivity of the evaluation system and the reliability of the framework.

4.4 Empathetic Response Model

We trained empathetic response generation models based on the LLAMA3.1-8B. The model inputs included dialogue history and situation. We employed NLP metrics to evaluate the quality, including diversity (Dist-1/2), fluency (BLEU-1/2), relevance (ROUGE-1/L), and perplexity (PPL). In addition, we invited three human expert judges to manually evaluate the responses, incorporating empathy (Emp) in addition to the three evaluation metrics previously used. There are three experimental setups: Baseline model (LLAMA+E): Trained solely on emotional annotation data. Emotion-cause enhanced models (LLAMA+EC/CE/SA): These models incorporated causal logic alongside emotional annotations. LLAMA+MvP: Employed multi-perspective prompt learning, combining emotional causal reasoning.

As shown in Table 4, key findings from the experimental results include: First, MvP-ECR significantly outperformed all baseline models. It achieved optimal performance on both automatic metrics (Dist, Bleu, Rouge) and human evaluation metrics (Cau, Rat), demonstrating that multi-perspective logical enhancement effectively improves the diversity and logical coherence of empathetic responses. Sec-

Models	Dist-1	Dist-2	Bleu-1	Bleu-2	Rouge-1	Rouge-L	PPL	Inf	Cau	Rat	Emp
LLAMA+E	0.5213	0.7126	0.0563	0.0224	0.1140	0.0812	**16.5073**	3.7	3.2	3.8	3.2
LLAMA+EC	0.5267	0.7325	0.0614	0.0247	0.1271	0.0899	17.8879	3.6	**4.0**	3.9	**4.4**
LLAMA+CE	0.5303	0.7385	0.0593	0.0237	0.1258	0.0890	17.9673	**4.2**	3.9	**4.1**	4.1
LLAMA+SA	0.5128	0.7139	0.0593	0.0239	0.1248	0.0890	17.8276	3.8	3.8	3.4	4.0
LLAMA+MvP	**0.6189**	**0.7876**	**0.0823**	**0.0512**	**0.1657**	**0.1250**	18.6296	3.8	**4.0**	**4.1**	4.0

Table 4: Performance comparison of various empathy response generation models. Human evaluation scores represent the average ratings from three independent annotators, demonstrating strong inter-rater reliability.

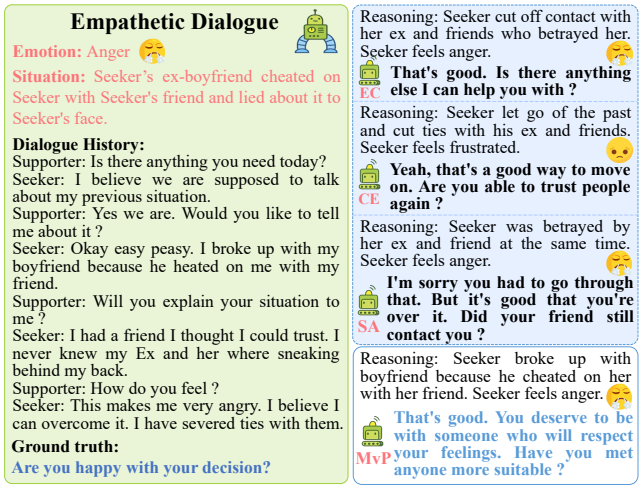

Figure 5: A representative empathetic dialogue case used to compare the responses generated by different reasoning paths (EC, CE, SA, and MVP) against the ground truth.

ond, the emotion-cause models (EC/CE/SA) generally performed better than the pure emotion model (E), though they were still inferior to MvP-ECR. This indicates that while introducing causal logic alone is helpful, combining it with multi-perspective optimization can further break through performance bottlenecks. Finally, the PPL results indicate that multi-perspective learning increases generation complexity, and the Inf metric reveals slight deficiencies in information density. Future work could explore finer-grained emotion-event alignment or incorporate external knowledge to further enhance logical coherence.

4.5 Case Study

To illustrate the practical effectiveness of different reasoning paths, we present a representative dialogue case in Figure 5 and analyze the generated responses.

In this empathetic dialogue case, the ground truth response "Are you happy with your decision?" directly guides Seeker to reflect on whether their decision has brought inner peace. This approach is highly empathetic and reflective, encouraging Seeker to actively evaluate their emotional recovery. And the EC response, "That's good. Is there anything else I can help you with?", reflects recognition of Seeker's actions, but lacks further exploration and guidance of emotions, appearing more superficial. CE Path responds with "Yeah, that's a good way to move on. Are you able to trust people again?" Although continuing the affirmative tone, quickly shifting to the theme of "trust" may create a sense of oppression for Seeker who is still angry and easily deviate from the core of the current emotion; The SA pathway expresses stronger empathy and support, such as "I'm sorry you had to go through that", but asking "Did your friend still contact you?" may reignite negative memories and increase the emotional load of the conversation. In contrast, MVP's response of "That's good. You deserve to be with someone who will respect your feelings. Have you met anyone more suitable?" extends to the future based on the recognition of emotions, conveying positive encouragement and guiding Seeker to look forward to healthier relationships, combining emotional comfort and causal understanding. The unique feature of MVP is that it not only focuses on the current emotional state and event logic, but also actively guides Seeker to think in a positive direction, injecting hope and constructive direction into the conversation, demonstrating stronger humanized support ability.

5 Conclusion

In this work, we propose MvP-ECR, a multi-perspective emotion-cause reasoning framework designed to enhance empathetic dialogue understanding and generation. By simulating human-like causal reasoning through three complementary inference paths, MvP-ECR enables large language models to identify users' dominant emotions and their underlying causes in a structured and cognitively grounded manner. To evaluate the robustness and consistency of emotion-cause inference, we further introduce a hierarchical evaluation framework that assesses causal alignment at lexical, syntactic and semantic levels. Additionally, the proposed emotion-cause quadruple representation standardizes the reasoning output, facilitating interpretable and controllable response generation. Extensive experiments on multiple empathetic dialogue benchmarks validate the effectiveness and generalizability of our framework, showing substantial improvements in both reasoning quality and generation performance. By integrating structured inference with affective cognition, MvP-ECR establishes a principled foundation for developing more interpretable, adaptable, and human-aligned empathetic dialogue systems.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (Grant No.2024YFG0006), the National Natural Science Foundation of China (Grant No.U24A20250), and the Fundamental Research Funds for the Central Universities (No.ZYGX2024Z005).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, J.; Ding, Z.; Li, K.; and Xia, R. 2023. Global-View and Speaker-Aware Emotion Cause Extraction in Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 3814–3823.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cai, H.; Shen, X.; Xu, Q.; Shen, W.; Wang, X.; Ge, W.; Zheng, X.; and Xue, X. 2023. Improving Empathetic Dialogue Generation by Dynamically Infusing Commonsense Knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, 7858–7873. Toronto, Canada: Association for Computational Linguistics.

Chen, T.; Shen, Y.; Chen, X.; Zhang, L.; and Zhao, S. 2023. MPEG: A Multi-Perspective Enhanced Graph Attention Network for Causal Emotion Entailment in Conversations. *IEEE Transactions on Affective Computing*, (01): 1–14.

Chen, X.; Yang, C.; Lan, M.; Cai, L.; Chen, Y.; Hu, T.; Zhuang, X.; and Zhou, A. 2024a. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Chen, X.; Yang, C.; Sun, C.; Lan, M.; and Zhou, A. 2024b. From Coarse to Fine: A Distillation Method for Fine-Grained Emotion-Causal Span Pair Extraction in Conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17790–17798.

Chen, Z.; Cao, Y.; Bi, G.; Wu, J.; Zhou, J.; Xiao, X.; Chen, S.; Wang, H.; and Huang, M. 2025. SocialSim: Towards Socialized Simulation of Emotional Support Conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1274–1282.

Cheng, Z.; Jiang, Z.; Yin, Y.; Wang, C.; Ge, S.; and Gu, Q. 2023. A consistent dual-mrc framework for emotion-cause pair extraction. *ACM Transactions on Information Systems*, 41(4): 1–27.

Davis, M. H. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1): 113.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gui, L.; Xu, R.; Wu, D.; Lu, Q.; and Zhou, Y. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, 145–160. Austin, Texas: Association for Computational Linguistics.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gupta, S.; and Dandapat, S. K. 2023. SEEC and CHASE: An emotion-cause pair-oriented approach and conversational dataset with heterogeneous emotions for empathetic response generation. *Knowledge-Based Systems*, 280: 111039.

He, Y.; Pan, Y.; Li, W.; You, J.; Deng, J.; and Ren, F. 2025. ECC: An Emotion-Cause Conversation Dataset for Empathy Response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6011–6028. Suzhou, China: Association for Computational Linguistics.

Hsu, J. H.; Chang, J.; Kuo, M. H.; and Wu, C. H. 2023. Empathetic Response Generation Based on Plug-and-Play Mechanism With Empathy Perturbation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31: 2032–2042.

Huang, Z.; Liu, P.; de Melo, G.; He, L.; and Wang, L. 2024. Generating Persona-Aware Empathetic Responses with Retrieval-Augmented Prompt Learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 12441–12445.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Jiang, D.; Liu, Y.; Liu, S.; Zhao, J.; Zhang, H.; Gao, Z.; Zhang, X.; Li, J.; and Xiong, H. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Keskin, S. C. 2014. From what isn't empathy to empathic learning process. *Procedia-Social and Behavioral Sciences*, 116: 4932–4938.

Lee, S. Y. M.; Chen, Y.; and Huang, C.-R. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 45–53. Los Angeles, CA: Association for Computational Linguistics.

Li, J.; Peng, B.; Hsu, Y.-Y.; and Huang, C.-R. 2024. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1976–1988.

Li, Q.; Chen, H.; Ren, Z.; Ren, P.; Tu, Z.; and Chen, Z. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4454–4466. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. MoEL: Mixture of Empathetic Listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 121–132. Hong Kong, China: Association for Computational Linguistics.

Liu, H.; Wei, R.; Tu, G.; Lin, J.; Jiang, D.; and Cambria, E. 2025. Knowing What and Why: Causal emotion entailment for emotion recognition in conversations. *Expert Systems With Applications*, 274: 126924.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 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)*, 3469–3483. Online: Association for Computational Linguistics.

Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; and Xu, B. 2024. A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*, 26: 776–788.

Poria, S.; Majumder, N.; Hazarika, D.; Ghosal, D.; Bhardwaj, R.; Jian, S. Y. B.; Hong, P.; Ghosh, R.; Roy, A.; Chhaya, N.; Gelbukh, A.; and Mihalcea, R. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13: 1317–1332.

Qian, Y.; Zhang, W.; and Liu, T. 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*, 6516–6528. Singapore: Association for Computational Linguistics.

Qiao, B.; Zhang, Y.; Gao, P.; Li, X.; Wang, S.; and Han, D. 2025. Multi-perspective empathy modeling for empathetic dialogue generation. *Knowledge-Based Systems*, 314: 113191.

Sabour, S.; Zheng, C.; and Huang, M. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11229–11237.

Su, X.; Huang, Z.; Su, Y.; Trisedya, B. D.; Dou, Y.; and Zhao, Y. 2024. Hierarchical Shared Encoder with Task-specific Transformer Layer Selection for Emotion-Cause Pair Extraction. *IEEE Transactions on Affective Computing*, (01): 1–15.

Su, X.; Huang, Z.; Zhao, Y.; Chen, Y.; Dou, Y.; and Pan, H. 2023. Recent trends in deep learning based textual emotion cause extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tu, G.; Xie, T.; Liang, B.; Wang, H.; and Xu, R. 2024. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19089–19097.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; and Yan, R. 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)*, 308–319.

Wang, X.; Wang, B.; Tang, Y.; Zhao, D.; Liu, J.; He, R.; and Hou, Y. 2025a. ECC: Synergizing Emotion, Cause and Commonsense for Empathetic Dialogue Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, 5475–5485.

Wang, Y.; Fang, X.; Yin, H.; Li, D.; Li, G.; Xu, Q.; Xu, Y.; Zhong, S.; and Xu, M. 2025b. BIG-FUSION: Brain-Inspired Global-Local Context Fusion Framework for Multimodal Emotion Recognition in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1574–1582.

Xia, R.; and Ding, Z. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1003–1012. Florence, Italy: Association for Computational Linguistics.

Xiao, D.; Xia, R.; and Yu, J. 2023. Emotion Cause Extraction on Social Media without Human Annotation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1455–1468. Toronto, Canada: Association for Computational Linguistics.

Zhao, W.; Zhao, Y.; Lu, X.; and Qin, B. 2023. Don't Lose Yourself! Empathetic Response Generation via Explicit Self-Other Awareness. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Zheng, C.; Liu, Y.; Chen, W.; Leng, Y.; and Huang, M. 2021. CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 813–824.

Zheng, L.; Jing, B.; Li, Z.; Tong, H.; and He, J. 2024. Heterogeneous contrastive learning for foundation models and beyond. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 6666–6676.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhou, J.; Zheng, C.; Wang, B.; Zhang, Z.; and Huang, M. 2023. CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8223–8237.

Zou, J.; Zhang, Y.; Wu, S.; Yang, J.; Qin, X.; Ying, L.; Jiang, M.; and Huang, Y. 2024. A machine reading comprehension framework for recognizing emotion cause in conversations. *Knowledge-Based Systems*, 289: 111532.