Efficient-Empathy: Towards Efficient and Effective Empathetic Data Selection

Anonymous ACL submission

Abstract

Empathy is a fundamental pillar of human social intelligence and a critical requirement for the development of human-centered artificial general intelligence (AGI). While large language models (LLMs) have shown remarkable general-purpose capabilities, their empathetic reasoning remains limited, largely due to the scarcity of high-quality training data. Prior work in empathetic modeling often relies on shallow emotional cues or architectural enhancements, overlooking the heterogeneous and multi-dimensional nature of empathy itself. In this work, we propose a data-efficient empathy learning framework that integrates insights from psychology—specifically, the dual dimensions of sensibility and rationality—as guiding criteria for high-quality data selection. Our approach leverages LLMs to automatically score and filter empathy dialogues, constructing curated datasets that emphasize emotionally grounded and cognitively coherent responses. We then train specialized sensibility and rationality experts, and dynamically combine their capabilities via a Mixture-of-Experts (MoE) model. Empirical results demonstrate that our framework not only achieves state-of-the-art empathetic generation but does so using significantly fewer data samples, affirming the importance of quality-driven selection in scaling empathetic AGI. The codebase and data are made available at `https://anonymous.4open.science/r/Efficient-Empathy-D5E3/README.md`.

1 Introduction

Empathy, the capacity to recognize, understand, and share the emotional states of others, is a cornerstone of human social intelligence. As artificial general intelligence (AGI) evolves toward human-centered interaction, equipping AI systems with robust empathetic reasoning becomes not merely beneficial but essential for fostering meaningful human-machine collaboration (Rashkin et al., 2019; Davis, 1983). Large language models (LLMs) have demonstrated remarkable capabilities in various domains (OpenAI, 2023; Touvron et al., 2023a; Bai et al., 2024), yet their ability to generate contextually appropriate, emotionally nuanced responses remains constrained by the scarcity of high-quality empathy training data. In this context, developing data-efficient methods for empathy modeling is a critical research direction—one that can circumvent the limitations of conventional data-hungry approaches while enhancing the emotional intelligence of AGI systems.

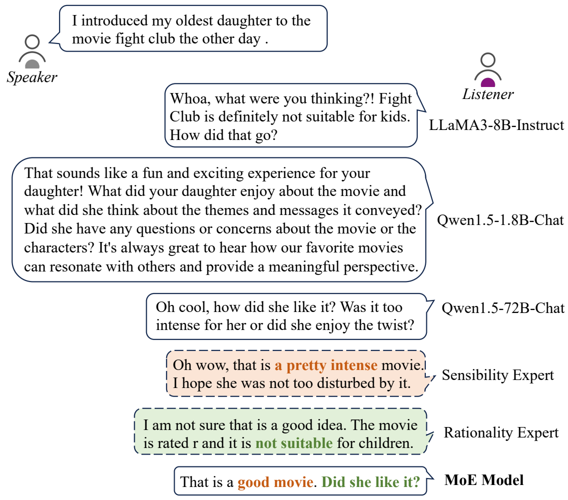


Figure 1: Comparison of Empathetic Responses from Different Models. Sensibility, Rationality, and MoE models are trained using data selected by Efficient-Empathy.

Prior work in computational empathy has primarily focused on two paradigms: (1) shallow emotion recognition (Wang et al., 2022; Fu et al., 2023; Yang et al., 2023; Yufeng et al., 2024), and (2) multi-dimensional external information augmentation (Ghosal et al., 2020; Zhou et al., 2021; Sabour et al., 2022). While these methods demonstrate promise, recent studies underscore that data quality is equally—if not more—critical for LLM performance (Chen et al., 2023; Xu et al., 2023). How-

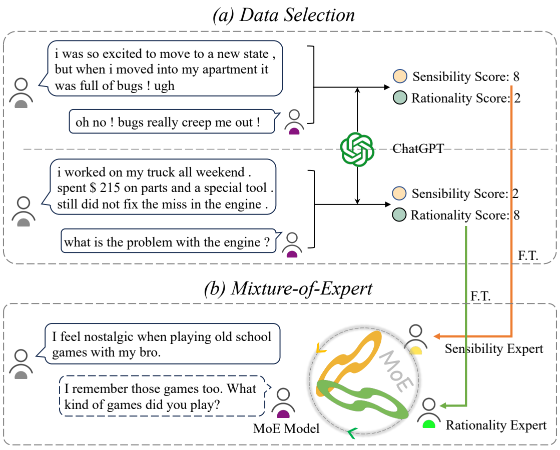


Figure 2: The pipeline of our approach (a) The data selection method utilized for classifying sensibility and rationality conversation. (b) Utilize sensibility and rationality data for MoE training

ever, existing approaches largely overlook the inherent heterogeneity in empathy data. Crucially, they fail to explore data-efficient strategies particularly whether insights from psychological theory can be integrated with AI techniques to improve the utilization of empathy data.

Psychological research reveals that sensibility and rationality are foundational to human empathy (Fritz and Helgeson, 1998). The absence of sensibility diminishes the capacity to connect emotionally with others, while insufficient rationality may lead to unmitigated communion—excessive emotional absorption that hinders effective response (Smith, 2006). Conversely, rational sensibility, known as cognitive empathy, enables balanced understanding of users' emotions while mitigating negative affective overload (Smith, 2006). This duality suggests that empathy data could be systematically filtered and enhanced by prioritizing samples exhibiting both traits—a direction previously unexplored in AI research.

Inspired by this, we propose a data-efficient empathy learning framework that leverages sensibility and rationality as key criteria for high-quality data selection (Fig. 2(a)). Our method employs LLMs to automatically score and filter empathy data, retaining only the most sensible and rational samples. These curated subsets train specialized experts—a sensibility expert and a rationality expert—whose knowledge is combined via a Mixture-of-Experts (MoE) model (Jacobs et al., 1991). This approach not only improves data efficiency but also achieves state-of-the-art (SoTA) performance by focusing training on the most impactful samples.

Inspired by this psychological framework, we propose a data-efficient empathy learning framework that leverages sensibility and rationality as dual criteria for data selection (Fig. 2(a)). Our method employs LLMs to automatically score and partition empathy data into sensibility and rationality subsets. These curated subsets train a specialized sensibility expert and a rationality expert for structured reasoning, whose outputs are dynamically integrated via a Mixture-of-Experts (MoE) model (Jacobs et al., 1991). By this, our approach not only improves efficiency but also achieves state-of-the-art (SoTA) performance, demonstrating that quality-driven data selection is pivotal for advancing empathetic AGI.

The core contributions of this paper are summarized as follows:

- **New Method.** We propose a new data selection method for empathy data, introducing the first sensibility and rationality-based data selection framework. Utilizing our meticulously curated sensibility and rationality scores, we pioneer the integration of sensibility and rationality data with a MoE model.

- **High Efficiency.** With a carefully curated dataset, our sensibility expert model outperforms the baseline using only 59% of the data.

- **High Robustness.** As shown in Table 3, with multiple data selection thresholds, our method consistently outperforms full data fine-tuning, demonstrating the robustness of our model.

- **SOTA Performance** We utilize both the selected sensibility and rationality data, we train a sensibility and rationality and subsequently train an expert MoE model, achieving SoTA performance, which demonstrates the effectiveness of our data selection method.

2 Related Work

Empathetic Response Generation. Empathy plays a pivotal role in emotionally intelligent dialogue systems. Building on the EmpatheticDialogues (ED) dataset (Rashkin et al., 2019), extensive efforts have explored how to endow models with empathic capacity. Commonsense knowledge integration (Sabour et al., 2022; Bosselut et al., 2019; Li et al., 2022b), sentiment supervision (Wang et al., 2022), and fine-grained context analysis (Kim et al., 2022) have all proven effective. Other works leverage psychological emotion detection (Chen and Liang, 2022) or model self-awareness (Zhao et al., 2023) to enrich emotional understanding. Qian et al. (2023a) frame empathy as a dual-stage process of semantic alignment and emotional expression. More recent studies in the LLM era focus on enhancing empathy via prompt engineering (Qian et al., 2023b; Wang et al., 2023; Yang et al., 2024), while Sun et al. (2023) distinguish between sensibility and rationality as complementary empathy components. Yet, a nuanced understanding of their fine-grained cognitive interplay remains underexplored.

Mixture-of-Experts (MoE). Originally proposed by Jacobs et al. (1991), MoE models enable dynamic specialization through expert routing. Token-level sparse gating (Shazeer et al., 2017) laid the groundwork for scalable MoE-based transformers (Lepikhin et al., 2020; Fedus et al., 2022), sparking advances in routing (Pan et al., 2024), load balancing (Zhong et al., 2024), and distributed optimization (Gale et al., 2023). Recent LLM-scale MoEs, such as Mixtral-8x22B (Jiang et al., 2024), DeepSeekMoE (Dai et al., 2024), and Qwen1.5-MoE-A2.7B (Team, 2024), demonstrate that activating only subsets of experts can match dense model performance with reduced computational cost. Despite these advances, the potential of MoE architectures for enhancing empathetic response generation remains largely untapped.

3 Sensibility-Rationality Scoring Framework

3.1 Dialogue Subset Selection

To develop an empathetic dialogue model, we assign each dialogue a sensibility score S and a rationality score R with GPT-4o-0613 by prompt shown in Figure 8(a), and partition the dataset using symmetric thresholds T_S and T_R. Dialogues with high S but low R are used to train a sensibility expert, while those with balanced S and R form the training set D_r for a rationality expert. These two specialized models are then integrated using a MoE framework to form the final empathy expert.

Mathematically, the selection method are:

$$D_s = \{d \in D \mid R(d) < T_R \text{ and } S(d) > T_S\}$$

$$D_d = \{d \in D \mid R(d) > T_R \text{ and } S(d) < T_D\}$$

$$D_r = \{d \in D \mid \neg(d \in D_s \cup d \in D_d)\}$$

where D represents the original dataset, $S(d)$ is the sensibility score, and $R(d)$ is the rationality score of dialogue d. The data selection process is summarized in Algorithm 1.

To maintain rigorous assessment standards, we employed GPT-4o-0613 for automated scoring. As part of our quality assurance protocol, we conducted a manual verification process on a randomly sampled subset of 1,000 annotations generated by ChatGPT. These annotations were independently evaluated by trained human annotators (university students), who classified each annotation as either correct or incorrect. Our analysis revealed that ChatGPT achieved an annotation accuracy of 93.7

3.2 Distribution of Rationality and Sensibility

In this section, we provide an analysis of the distribution of rationality and sensibility scores. Each dialogue in the dataset is evaluated on a scale from 0 to 10 for both rationality and sensibility. To visualize the distribution of these scores, we plot a 2D histogram showing the frequency of each score from 0 to 10.

From the 2D histogram in Figure 4, we can observe several key trends and patterns. Firstly, the highest frequency cluster occurs at the combination of a rationality score of 2 and a sensibility score of 8, with a frequency of 8603, indicating that dialogues with low rationality and high sensibility are common in the dataset. Moreover, there is a general tendency for dialogues to score higher on sensibility compared to rationality, as evidenced by the higher frequencies of scores in the upper part of the y-axis (sensibility scores) versus the x-axis (rationality scores). This trend aligns with the ED dataset, which typically contains dialogues with higher sensibility content. Additionally, dialogues with balanced rationality and sensibility scores, such as those around 5 for both dimensions, are relatively rare. This suggests that dialogues exhibiting a balance between logical reasoning and emotional depth are uncommon, presenting a potential area for improving dialogue generation models to better harmonize these attributes.

4 Empathetic MoE Model

In accordance with D_s and D_r, we conduct single-expert training for sensibility and rationality abilities. Subsequently, expert models Efficient-Empathy are integrated using the Branch-Train-Mix (BTX) method (Sukhbaatar et al., 2024). The

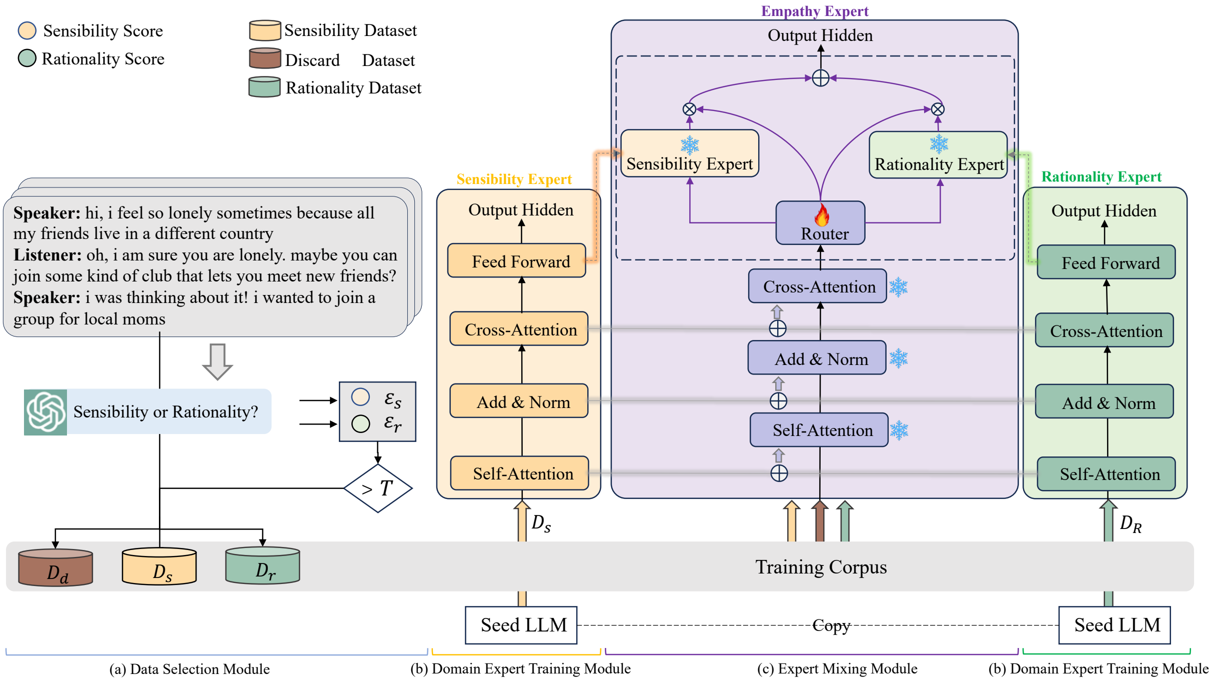


Figure 3: The overall pipeline of Efficient-Empathy consists of three parts: (a) the Data Selection Module, which classifies the empathetic dataset into sensibility, rationality, and discard datasets; (b) the Domain Expert Training Module, which uses the selected datasets to fine-tune LLMs and acquire sensibility and rationality experts; and (c) the Expert Mixing Module, which integrates the sensibility and rationality experts into the MoE empathy model.

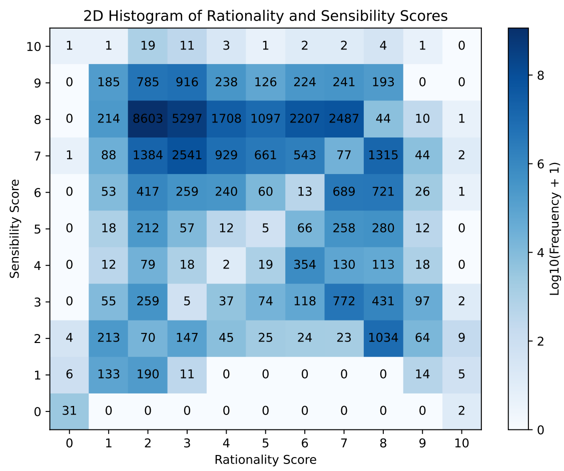


Figure 4: 2D Histogram of Rationality and Sensibility Scores. The x-axis represents rationality scores, the y-axis represents sensibility scores, and the color intensity indicates the frequency of each combination of scores.

overall model structure is shown in Fig 3. And the algorithm description is shown in Algorithm 2.

4.1 Empathy Data Selection

In practice, we partition the dataset using symmetric thresholds $T_S = T_R$, which ensures a balanced, unbiased partition and simplifies threshold tuning (Ghosal et al., 2019). And we deliberately avoid using high-rationality but low-sensibility data D_d to train the rationality expert, as such samples often yield cold, factual responses that lack emotional engagement. In contrast, the balanced set D_r better reflects real-world empathetic communication, where logic and emotion co-occur. This design enables our model to produce responses that are both contextually appropriate and emotionally attuned, which is critical for effective empathy modeling.

4.2 Empathetic Domain Experts

In this subsection, we transfer knowledge from LLM to create specialized expert models via Supervised Fine-Tuning (SFT). Specifically, we utilize LLaMA3-8B-Instruct as the seed model to engage in LoRA fine-tuning on two domain datasets, D_s and D_r, thereby acquiring the sensibility expert M_s and the rationality expert M_r separately:

$$M_s = \theta_s^{SFT}(LLM; D_s) \tag{1}$$

$$M_r = \theta_r^{SFT}(LLM; D_r) \tag{2}$$

The derived models, M_s and M_r, undergo training to excel within their respective domains, establishing the foundation for the subsequent mixing stage.

4.3 Empathetic MoE Architecture

Building upon the insights of domain experts, we employ the MoE approach to incorporate them into a comprehensive empathetic model, denoted as M_e. Diverging from the Branch-Train-Merge (BTM) (Li et al., 2022a) methodology, which exclusively consolidates the final feature representations, our approach introduces a collaborative configuration of Feed-Forward Network (FFN) layers. It further allocates decision-making weights to individual experts via a sophisticated soft routing system, thereby enhancing the model's selective empathy capabilities.

In a single expert model, each transformer block comprises an attention module and an FFN module. The structure of M_e is similar, with the exception that the FFN layer is substituted by an MoE layer, which is a combination of multiple single expert FFNs. Particularly, in the i-th transformer block, the FFN layers of M_s and M_r are denoted as $FF_i^s(x)$ and $FF_i^r(x)$, respectively. The i-th MoE layer $MoE_i(x)$ is defined as follows:

$$\begin{aligned} MoE_i(x) &= Router_i(x) \cdot FF_i^s(x) \\ &+ (1 - Router_i(x)) \cdot FF_i^r(x), \end{aligned} \quad (3)$$

where the router is defined as follows:

$$Router_i(x) = SoftMax(W_i x + b_i), \quad (4)$$

$Router_i$ serves as a soft routing mechanism that output values ranging between 0 and 1 to regulate the influence of domain experts. W_i and b_i is the linear transformation and bias of the i-th MoE layer. For the parameters and weights of other layers in the Efficient-Empathy model, we initialize them using the average weights of the corresponding layers in the expert models:

$$M_e^i(x) = \begin{cases} \text{MoE}_i(x) & \text{MoE} \\ \text{Average}(M_s^i(x), M_r^i(x)) & \text{Other} \end{cases} \quad (5)$$

After constructing the model, we introduce new random initialized router parameters and then average the weights of domain experts. Consequently, a second stage of training is conducted in the subsequent experiment to activate the overall parameter weights of the empathy model.

4.4 Empathetic MoE Training

In the empathetic response generation task, given a conversation message $C = [S_1, L_1, S_2, ..., S_N]$ of length N, where S is Speaker's utterance and L is Listener's utterance. Our model then generate a response $R = [r_1, r_2, ..., r_m]$ based on emotional expert and rational expert in context with C, where m is the length of the token sequence:

$$R = MoE(C) \quad (6)$$

The training loss is the standard negative loglikelihood (NLL) loss on the generated response R:

$$\mathcal{L}_{nll} = -\sum_{t=1}^{m} \log(r|C, r_{<t}) \quad (7)$$

Particularly, we first train sensibility and rationality experts on their domain data D_s and D_r separately based on LLaMA3-8B-Instruct. Then, we integrate them into a MoE model following the method outlined in Section 4 and the base model still is LLaMA3-8B-Instruct. In training phase, all parameters of the empathy model, except for the router, are kept frozen, while we fine-tune the model on the entire training dataset of D_s and D_r.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

Our study is based on the authoritative EmpatheticDialogues (ED) dataset (Rashkin et al., 2019), which consists of 25,000 daily conversations. This dataset is meticulously curated through crowdsourcing, involving 810 workers from Amazon Mechanical Turk [1]. Each conversation is constructed in a one-on-one format, pairing two workers. One worker assumes the role of the speaker, responding according to a given emotional label and situation, while the other worker assumes the role of the listener, providing empathetic responses to the speaker. On average, each conversation consists of 4.31 exchanges, with each exchange containing approximately 15.2 words.

5.1.2 Evaluation Metrics

For automatic evaluation, we use corpus-level BLEU (B-1 to B-4), sentence-level ROUGE (R-1, R-2), and Distinct (Dist-1, Dist-2). For human evaluation, we conduct A/B tests based on Coherence, Empathy, Information, and Continuity. Twelve evaluators are requested to identify the better response. A more detailed description can be found in Appendix E.3 and Appendix F.

[1] https://www.mturk.com

Table 1: Results of the automatic evaluation on baseline models, the sensibility model, and the MoE model are presented. The best performance is highlighted in bold, and the purple table represents the increased values.

Models	B-1	B-2	B-3	B-4	R-1	R-2	Dist-1	Dist-2
MoEL (Lin et al., 2019)	18.07	8.30	4.37	2.65	18.24	4.81	0.59	2.64
MIME (Ghosal et al., 2020)	18.60	8.39	4.54	2.81	17.08	4.05	0.47	1.66
EmpDG (Li et al., 2020)	19.96	9.11	4.74	2.80	18.02	4.43	0.46	1.99
CEM (Sabour et al., 2022)	16.12	7.29	4.06	2.03	15.77	4.50	0.62	2.39
SEEK (Wang et al., 2022)	10.77	4.40	2.02	1.08	12.74	2.94	0.68	2.81
CASE (Zhou et al., 2023)	15.59	7.22	3.80	2.24	17.33	4.67	0.65	3.37
E-CORE (Fu et al., 2023)	-	-	-	-	-	-	0.72	3.49
KEMP (Li et al., 2022b)	16.72	7.17	3.77	2.33	16.11	3.31	0.66	3.07
CAB (Gao et al., 2023)	19.23	8.55	4.36	2.57	17.50	4.13	1.13	4.23
ESCM (Yang et al., 2023)	-	-	-	-	-	-	1.19	4.11
DCKS (Cai et al., 2023)	18.75	9.12	5.38	3.57	19.14	5.45	1.57	6.02
CTSM (Yufeng et al., 2024)	-	-	-	-	-	-	2.00	7.34
Lamb (Sun et al., 2023)	**22.00**	**10.49**	**6.07**	**3.97**	**19.55**	**5.47**	**1.80**	**7.73**
Qwen1.5-1.8B-Chat (Team, 2024)	10.43	3.50	1.59	0.84	12.95	1.64	2.41	17.98
LLaMA2-13B-Instruct (Touvron et al., 2023b)	11.69	4.03	1.79	0.93	13.27	1.83	2.91	18.92
LLaMA3-8B-Instruct (Touvron et al., 2023b)	13.17	4.42	1.92	1.02	**14.12**	1.68	2.69	18.70
Qwen1.5-72B-Chat (Team, 2024)	**14.19**	**4.85**	**2.27**	**1.23**	13.83	**1.97**	3.29	22.68
Sensibility	$\boxed{22.34}$	$\boxed{11.25}$	$\boxed{6.58}$	$\boxed{4.21}$	$\boxed{19.82}$	$\boxed{5.79}$	$\boxed{3.00}$	$\boxed{15.44}$
+ *Compared with Lamb*	▲ 0.34	▲ 0.76	▲ 0.51	▲ 0.24	▲ 0.27	▲ 0.32	▲ 1.2	▲ 7.71
+ *Compared with Qwen1.5-72b*	▲ 8.15	▲ 6.40	▲ 4.31	▲ 2.98	▲ 5.70	▲ 3.82	-	-
MoE Model	$\boxed{\mathbf{23.04}}$	$\boxed{\mathbf{11.62}}$	$\boxed{\mathbf{6.68}}$	$\boxed{\mathbf{4.22}}$	$\boxed{\mathbf{20.28}}$	$\boxed{\mathbf{6.15}}$	$\boxed{2.34}$	$\boxed{10.91}$
+ *Compared with Lamb*	▲ 1.04	▲ 1.13	▲ 0.61	▲ 0.25	▲ 0.73	▲ 0.68	▲ 0.54	▲ 3.18
+ *Compared with Qwen1.5-72b*	▲ 8.85	▲ 6.77	▲ 4.41	▲ 2.99	▲ 6.16	▲ 4.18	-	-

5.2 Main Experiments

We compare our Sensibility and MoE model against the baselines (E.2) in Table 1.

We can see our Sensibility and MoE model significantly outperforms the baselines in BLEU (B-1 to B-4), ROUGE (R-1, R-2), and Distinct (Dist-1, Dist-2) scores, demonstrating its effectiveness. Compared to the LLM, our approach shows substantial improvements in BLEU and ROUGE metrics. While LLMs achieve high Distinct scores, their lower performance in BLEU and ROUGE suggests poorer response quality.

The discrepancy between BLEU and Distinct metrics stems from their focus on different aspects: Distinct measures lexical diversity, while BLEU and ROUGE evaluate n-gram overlap with reference texts. High Distinct scores enhance variety but may reduce exact matches, leading to lower BLEU and ROUGE scores. Given that our model outperforms the baselines in Distinct, we did not further analyze LLM distinctiveness.

In summary, our MoE and Sensibility model excels in generating empathetic responses, as reflected in its superior BLEU and ROUGE performance.

5.3 Efficient Sensibility Data

We evaluate the effect of the sensibility expert on empathetic performance. We selected 23,862 (59%) dialogues from a total of 40,250 based on their sensibility scores. Specifically, dialogues with rationality scores below a certain threshold and sensibility scores above it were included in the sensibility dataset. We then used both the selected sensibility data (Sensibility) and the full dataset (Full F.T.) to fine-tune three commonly used LLMs: LLaMA3-8B-Instruct, Qwen1.5-7B-Chat, and Qwen1.5-1.8B-Chat.

From Table 2, we observe that the models trained on the sensibility data outperform those trained on the full dataset across all three models. This demonstrates the efficiency and effectiveness of our data selection method, indicating that focusing on dialogues with higher sensibility and lower rationality improves empathetic performance.

5.4 Robustness of Efficient-Empathy

From Table 3, we can see that with hyperparameters set to 4, 5, and 6, and using just above 50% of the data, our model consistently outperforms the baseline model. This demonstrates the robustness of our data selection algorithm.

As shown in the table, the model trained with

Table 2: The performance of Sensibility Expert on ED test.

Models	#Data	Data Percentage	Models	B-1	B-2	B-3	B-4	R-1	R-2
LLaMA3-8B-Instruct	40,250	100%	Full F.T.	21.23	10.40	5.98	3.84	19.49	5.58
	23,862	59%	**Sensibility**	**22.34**	**11.25**	**6.58**	**4.21**	**19.82**	**5.79**
Qwen1.5-7B-Chat	40,250	100%	Full F.T.	20.84	9.19	4.89	2.90	17.44	4.09
	23,862	59%	**Sensibility**	**21.34**	**10.27**	**5.82**	**3.66**	**18.86**	**5.28**
Qwen1.5-1.8B-Chat	40,250	100%	Full F.T.	17.45	6.78	3.22	1.76	14.89	3.08
	23,862	59%	**Sensibility**	**18.2**	**7.50**	**3.73**	**2.16**	**15.74**	**3.48**

Table 3: Performance of MoE model across different data selection thresholds.

Model	Datasets	#Data	Data Percentage	B-1	B-2	B-3	B-4	R-1	R-2
LLaMA3-8B-Instruct	Full Dataset	40,250	100%	21.23	10.40	5.98	3.84	19.49	5.58
	Threshold-4	21,034	52%	21.34	10.63	6.15	3.93	19.56	5.58
	Threshold-5	23,862	59%	22.34	11.25	6.58	4.21	19.82	5.79
	Threshold-6	24,776	62%	**23.02**	**11.84**	**7.02**	**4.58**	**20.2**	**6.08**

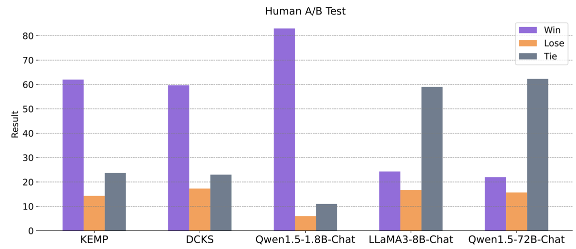


Figure 5: Human A/B test between our MoE model and baseline models.

Threshold-6, which uses 24,776 training instances, achieves the highest scores in BLEU and ROUGE metrics, indicating better performance in generating diverse and accurate empathetic responses. Notably, the B-1 score increased from 21.1 in the full dataset to 23.02 with Threshold-6, and similar trends are observed in B-2, B-3, B-4, R-1, and R-2 scores.

These results validate that our data selection algorithm is robust, effective and efficient, as it maintains and even improves model performance with different threshold settings. By selectively training on high-quality data, our approach not only reduces the amount of data required but also improves the overall performance of the LLMs in empathetic response tasks.

5.5 Human Evaluation

We conduct the human evaluation A/B testing to further evaluate the performance of our model. The results in Figure 5 demonstrate that the MoE model consistently outperforms the baseline models across the dimensions of Coherence, Empathy, Information, and Continuity.

For KEMP and DCKS, which are based on the standard language model BART, our model has a winning rate of around 60% and a losing rate of approximately 15%. Notably, the win rate for our model surged to 83.0% when compared to the 1.8B scale LLM, Qwen1.5-1.8B-Chat. However, as the scale of baseline parameters increases, their effectiveness improves. When compared with Qwen1.5-72B-Chat, the high tie rate of 62.3% indicates that both models frequently produced comparably effective responses. Nonetheless, our model has a win rate of 22.0%, higher than the loss rate of 15.7%. Similarly, the results show a close contest with our model winning 24.3% and losing 16.7% in comparison to LLaMA3-8B-Instruct.

Overall, our model's strong performance in human evaluations underscores its practicality and user-friendliness. These results indicate that our model is not only effective but also well-received by human evaluators, highlighting its potential for real-world applications.

5.6 LLM-as-Judge

We utilized GPT-4o-0613 as an automated evaluation model to assess our proposed method on the ED test set, adhering to the human evaluation protocol outlined above. Performance was quantified using a 5-point Likert scale (1–5) across multiple evaluation dimensions, with higher scores reflecting superior performance. As evidenced by the experimental results in Table 5, our method exhibits consistent and robust effectiveness across all evaluated metrics.

Table 4: Ablation experiments on ED test.

Base Model	Models	B-1	B-2	B-3	B-4	R-1	R-2
LLaMA3-8B-Instruct	Sensibility	22.34	11.25	6.58	4.21	19.82	5.79
	Rationality	21.4	10.47	6.08	3.88	19.46	6.13
	MoE Model ▲	**23.04**	**11.62**	**6.68**	**4.22**	**20.28**	**6.15**
	struc-(a) ↓	22.16	11.08	6.33	3.98	20.12	5.92
	struc-(b) ↓	21.30	10.75	6.23	3.97	20.13	6.13
	struc-(c) ↓	22.41	11.33	6.55	4.16	20.21	6.14
	struc-(d) ↓	22.39	11.18	6.39	4.04	19.87	5.88
Qwen1.5-1.8B-Chat	Sensibility	18.2	7.5	3.73	2.16	15.74	3.48
	Rationality	17.62	6.91	3.31	1.83	14.81	3.06
	MoE Model ▲	**22.01**	**11.16**	**6.53**	**4.21**	**20.27**	**6.30**
	struc-(a) ↓	21.59	10.96	6.40	4.08	20.10	6.29
	struc-(b) ↓	21.68	10.95	6.38	4.10	20.19	6.27
	struc-(c) ↓	21.82	11.03	6.51	4.21	20.17	6.30
	struc-(d) ↓	21.66	10.97	6.40	4.10	19.94	6.04

Table 5: GPT-4o-0613 as a judge on ED test.

Method	Coh.	Emp.	Inf.	Con.
Mixtral-8x7B-Chat	2.87	3.15	3.28	2.89
Llama3-8B-Instruct	3.84	3.53	3.46	3.14
Qwen2.5-7B-Instruct	3.84	4.03	3.42	3.83
Ours	**3.73**	**4.17**	**4.51**	**4.54**

5.7 Ablation Study

In this section, we investigate whether our MoE method outperforms other solutions. We begin by training the sensibility, rationality, and discard experts using the selected datasets with the LLaMA3-8B-Instruct and Qwen1.5-1.8B-Chat models. We then conduct a series of experiments by replacing the rationality and sensibility experts.

As shown in Figure 6, we first replace the rationality expert with the base model (Figure 6(a)), followed by replacing the rationality expert with the discard model (Figure 6(b)). Next, we replace the sensibility expert with the base model (Figure 6(c)), and finally, replace it with the discard model (Figure 6(d)). Table 4 shows that modifying either the rationality or sensibility expert significantly reduces the performance of the LLaMA3-8B-Instruct and Qwen1.5-1.8B-Chat based MoE models, highlighting the importance of both experts. Interestingly, Structure-(c) outperform structure-(a) while the rationality expert alone performs worse than the sensibility expert. We hypothesize that this is due to a higher alignment between the Base LLM and the rationality expert in decision-making.

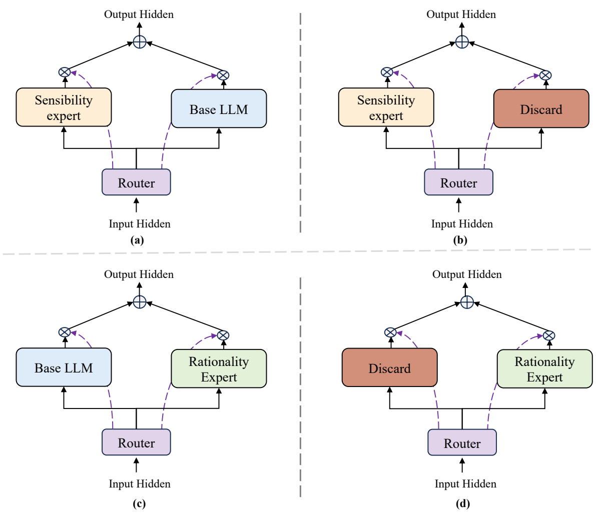


Figure 6: The overall ablation MoE model structure.

6 Conclusion

We present Efficient-Empathy, a data selection framework that leverages LLM-rated sensibility and rationality scores to address the scarcity of high-quality empathetic data. Our approach achieves state-of-the-art performance using only 59% of selected data, with further gains through MoE integration. Human evaluations confirm superior contextual appropriateness and emotional intelligence, establishing data-efficient empathy learning as a viable paradigm for human-centered AI. Future work may explore extending this paradigm to other dimensions of social intelligence, potentially bridging the gap between artificial and human empathy.

7 Limitations

Due to the specialized nature of empathic response tasks, current evaluation metrics still fall short in accurately reflecting model performance. Machine-based evaluation methods are primarily reliant on surface-level string matching or semantic vector matching, the latter of which heavily depends on the embedding capabilities of the model. Although human evaluation methods are somewhat aligned with human habits, the subjectivity of evaluators inevitably introduces errors during assessment. In future work, we will explore more scientifically grounded evaluation metrics to objectively and fairly reflect the model's empathic capabilities.

References

Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. 2024. A survey of multimodal large language model from a data-centric perspective. arXiv preprint arXiv:2405.16640.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4762–4779.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In Findings of the Association for Computational Linguistics: ACL 2023, pages 7858–7873.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701.

Yangbin Chen and Chunfeng Liang. 2022. Wish I can feel what you feel: A neural approach for empathetic response generation. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 922–933.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066.

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

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120):1–39.

Heidi L Fritz and Vicki S Helgeson. 1998. Distinctions of unmitigated communion from communion: self-neglect and overinvolvement with others. Journal of Personality and Social Psychology, 75:121.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-core: Emotion correlation enhanced empathetic dialogue generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10568–10586.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. 2023. Megablocks: Efficient sparse training with mixture-of-experts. Proceedings of Machine Learning and Systems, 5:288–304.

Pan Gao, Donghong Han, Rui Zhou, Xuejiao Zhang, and Zikun Wang. 2023. Cab: Empathetic dialogue generation with cognition, affection and behavior. In Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, page 597–606.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. Neural computation, 3(1):79–87.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088.

Wongyu Kim, Youbin Ahn, Donghyun Kim, and Kyong-Ho Lee. 2022. Emp-RFT: Empathetic response generation via recognizing feature transitions between utterances. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4118–4128.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. 2022a. Branch-train-merge: Embarrassingly parallel training of expert language models. arXiv preprint arXiv:2208.03306.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4454–4466.

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

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 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), pages 121–132.

OpenAI. 2023. Chatgpt.

Bowen Pan, Yikang Shen, Haokun Liu, Mayank Mishra, Gaoyuan Zhang, Aude Oliva, Colin Raffel, and Rameswar Panda. 2024. Dense training, sparse inference: Rethinking training of mixture-of-experts language models. arXiv preprint arXiv:2404.05567.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023a. Think twice: A human-like two-stage conversational agent for emotional response generation. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, page 727–736.

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

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.

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.

Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. The Psychological Record, 56:3–21.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. arXiv preprint arXiv:2403.07816.

Linzhuang Sun, Nan Xu, Jingxuan Wei, Bihui Yu, Liping Bu, and Yin Luo. 2023. Rational sensibility: Llm enhanced empathetic response generation guided by self-presentation theory. arXiv preprint arXiv:2312.08702.

Qwen Team. 2024. Introducing qwen1.5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4634–4645.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, and Weiping Wang. 2023. Enhancing empathetic and emotion support dialogue generation with prophetic commonsense inference. arXiv preprint arXiv:2311.15316.

Yang Xu, Yongqiang Yao, Yufan Huang, Mengnan Qi, Maoquan Wang, Bin Gu, and Neel Sundaresan. 2023. Rethinking the instruction quality: Lift is what you need. Preprint, arXiv:2312.11508.

Zhou Yang, Zhaochun Ren, Yufeng Wang, Chao Chen, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. An iterative associative memory model for empathetic response generation. arXiv preprint arXiv:2402.17959.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Wu Yunbing, Yisong Su, Sibo Ju, and Xiangwen Liao. 2023. Exploiting emotion-semantic correlations for empathetic response generation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4826–4837.

Wang Yufeng, Chen Chao, Yang Zhou, Wang Shuhui, and Liao Xiangwen. 2024. Ctsm: Combining trait and state emotions for empathetic response model. arXiv preprint arXiv:2403.15516.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don't lose yourself! empathetic response generation via explicit self-other awareness. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13331–13344.

Zexuan Zhong, Mengzhou Xia, Danqi Chen, and Mike Lewis. 2024. Lory: Fully differentiable mixture-of-experts for autoregressive language model pre-training. arXiv preprint arXiv:2405.03133.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 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, pages 8223–8237.

Pei Zhou, Pegah Jandaghi, Bill Yuchen Lin, Justin Cho, Jay Pujara, and Xiang Ren. 2021. Probing commonsense explanation in dialogue response generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4132–4146.

A ED Datasets

A.1 Introduction to ED Datasets

Our study is based on the authoritative EmpatheticDialogues (ED) dataset (Rashkin et al., 2019), which consists of 25,000 daily conversations encompassing 32 uniformly distributed emotional labels. This dataset is meticulously curated through crowdsourcing, involving 810 workers from Amazon Mechanical Turk [2]. Each conversation is constructed in a one-on-one format, pairing two workers. One worker assumes the role of the speaker, responding according to a given emotional label and situation, while the other worker assumes the role of the listener, providing empathetic responses to the speaker. On average, each conversation consists of 4.31 exchanges, with each exchange containing approximately 15.2 words.

A.2 Data Distribution

In this section we visualizes the data distribution for sensibility, rationality, and discard categories across different thresholds, followed by an exploration of their effectiveness to demonstrate the robustness of our data selection algorithm. In Figure 7, we present patterns for thresholds 4, 5, and 6, arranged from top to bottom. Each row corresponds to a threshold, with columns (from left to right) representing sensibility data, rationality data, and discard data.

Sensibility Data primarily occupy the upper left corner, indicating high sensibility and low rationality, driven mainly by emotional aspects. **Rationality Data** appear predominantly in the upper right corner, reflecting high sensibility and high rationality, suggesting a balanced emotional and logical approach suitable for empathetic responses. **Discard Data** cluster in the lower right corner, showing low sensibility and high rationality, making them unsuitable for empathetic responses.

This distribution underscores the significance of sensibility and rationality in determining data suitability for empathetic modeling.

With the selected data based on three different thresholds: 4, 5, and 6, we subsequently train an LLaMA3-8B-Instruct model based on the sensibility data and full dataset. The results are compared with the baseline model which is trained on the full dataset. Table 3 shows the performance of our model across different thresholds.

[2] https://www.mturk.com

B Case Study

To address **Q5**, in this section, we utilize case studies to conduct an in-depth analysis of the impact of our method. We compare the performance of several models, including LLaMA3-8B-Instruct, Qwen1.5-72B-Chat, Qwen1.5-1.8B-Chat, and Full F.T. model(trained on section 5.3), as well as the Rationality(trained on 5.2), Sensibility(trained on 5.2), and MoE experts(trained on 5.2).

Table 6: Case study-1 and the responses from Sensibility, Rationality, MoE, and other models. Blue text indicates the presence of rationality information and Orange text denotes sensibility expression.

	Context	**Speaker:** I feel nostalgic when playing old school games with my bro.
Response	LLaMA3-8B-Instruct	Aww, that's sweet! It's like you're reliving good times with your bro. Nostalgia can be a powerful way to connect with happy memories, doesn't it?
	Qwen1.5-72B-Chat	That's awesome! Those memories with your bro are precious. Keep enjoying those gaming sessions, they're a great way to bond.
	Qwen1.5-1.8B-Chat	Hey bro, I remember when we used to play those old school games back in the day. It's like a flashback to our childhood, and I miss those days so much. Playing games together was such a joy, and we had so much fun. Do you still play those games with your friends? If so, let's catch up soon and reminisce about those memories.
	Full F.T.	That is awesome! I am glad you have a good relationship with your brother.
	Rationality	What games do you play?
	Sensibility	I can imagine. It is always fun to play old school games with your siblings.
	MoE Model	I remember those games too. What kind of games did you play?
Golden Response		Nice , me too , are you playing nintendo games ?

As shown in Table 6, LLaMA3-8B-Instruct, while providing warm and positive responses, tends to lack specific resonant experiences. This results in replies that feel somewhat generic and superficial, failing to delve deeply into the speaker's unique context and emotions. Similarly, Qwen1.5-72B-Chat's responses, though affirming and encouraging, also suffer from a lack of specific resonant experiences and interactive guidance. In contrast, the Rationality expert's responses lack emotional recognition, failing to mirror the speaker's sentiments. Besides, although both the Sensibility expert and Baseline model demonstrate understanding towards the speaker, their responses lack the inclination to continue the conversation.

Our approach aims to address these shortcomings by combining sensibility and rationality to foster a deeper connection. By immediately acknowledging that the listener has had similar experiences in the first sentence of our response, we quickly establishes empathy and demonstrates a clear understanding of the speaker's feelings. Subsequently, the second sentence asks specific questions about the speaker's experience, which not only shows a genuine interest in their story but also actively encourages the conversation to continue, making the interaction more personalized and dynamic.

In Table 7, the response of LLaMA3-8B-Instruct, "*You're not usually anxious about trains*" assumes the speaker's typical state of mind regarding trains. By making this assumption, LLaMA risks misrepresenting or misunderstanding the speaker's feelings or experiences. If the speaker does not identify with this assumption, they may feel misunderstood or unfairly labeled. Similarly, Qwen1.5-72B-Chat's approach fails to adequately address the speaker's unique situation. This generality can lead to a lack of engagement and depth in the dialogue.

In analyzing different expert approaches, we distinguish between the Rationality expert and the Sensibility expert. The Rationality expert primarily focuses on identifying and understanding the causes of the horror of the train. While it may offer logical and practical solutions, it often fails to consider the speaker's emotional state. Conversely, the Sensibility expert emphasizes understanding and empathizing with the user's sacred feeling. However, it may not effectively guide the conversation towards practical solutions or a deeper understanding of the issue at hand.

Table 7: Case study-2 and the responses from Sensibility, Rationality, MoE, and other models. Blue text indicates the presence of rationality information and Orange text denotes the sensibility expression.

	Context	**Speaker:** I was so scared of going onto the train !
Response	LLaMA3-8B-Instruct	Ah, girl, what happened? You're not usually anxious about trains. Did something specific happen that triggered this fear?
	Qwen1.5-72B-Chat	I get that, trains can be intimidating, especially if you're not used to them. But you made it on, that's brave! Next time will be easier.
	Qwen1.5-1.8B-Chat	Oh, I'm sorry to hear that you're feeling scared about going on the train. It's completely normal to feel anxious about taking a big step like this, especially if it's your first time. Remember that taking a train can be a relatively safe and familiar experience, especially if you've taken the train before. Here are some tips that might help you feel more comfortable: 1. Plan ahead: Make sure you have a good understanding of the train schedule and route. Familiarize yourself with the train's schedule, the stations, and the times when the train arrives and departs. Knowing the journey ahead can help alleviate some anxiety
	Full F.T.	Oh no! What happened?
	Rationality	What was wrong with the train?
	Sensibility	I can imagine! I have been on trains before and i was scared too!
	MoE Model	Why? Did you have trouble getting on?
Golden Response		Was it your first time riding on one ? Is there something in particular that makes you afraid ?

Our proposed approach seeks to balance these two expert perspectives, aiming to both understand the user's state and address the underlying issues without making unwarranted assumptions. By showing genuine concern for whether the user is experiencing difficulties through "*Did you have trouble*", we create a supportive environment that validates the user's feelings and experiences. Furthermore, instead of making assumptions, we directly ask about the cause of the problem through "*Why?*". This method allows the speaker to provide their own explanation, ensuring that their perspective is accurately represented.

Table 8: Case study-3 and the responses from Sensibility, Rationality, MoE, and other models. Blue text indicates the presence of rationality information and Orange text denotes the sensibility expression.

	Context	**Speaker:** I introduced my oldest daughter to the movie fight club the other day .
Response	LLaMA3-8B-Instruct	Whoa, what were you thinking?! Fight Club is definitely not suitable for kids. How did that go?
	Qwen1.5-72B-Chat	Oh cool, how did she like it? Was it too intense for her or did she enjoy the twist?
	Qwen1.5-1.8B-Chat	That sounds like a fun and exciting experience for your daughter! What did your daughter enjoy about the movie and what did she think about the themes and messages it conveyed? Did she have any questions or concerns about the movie or the characters? It's always great to hear how our favorite movies can resonate with others and provide a meaningful perspective.
	Full F.T.	Oh wow, that is a very mature movie. I hope she was prepared for it.
	Rationality	I am not sure that is a good idea. The movie is rated r and it is not suitable for children.
	Sensibility	Oh wow, that is a pretty intense movie. I hope she was not too disturbed by it.
	MoE Model	That is a good movie. Did she like it?
Golden Respose		Interesting choice . Did she like it ?

In Table 8, we analyze the approaches of LLaMA3-8B-Instruct and Qwen1.5-72B-Chat, followed by a description of our proposed method.

LLaMA3-8B-Instruct expresses concern about the movie's suitability for the speaker's daughter but does so in a manner that can be perceived as questioning and critical, potentially making the speaker feel judged or unsupported. This can lead to feelings of misunderstanding and frustration, negatively impacting the flow and depth of the conversation. Similarly, Qwen1.5-72B-Chat's response, while less overtly critical than LLaMA's, still carries negative implications. By implicitly questioning the suitability of the movie choice for her daughter, Qwen1.5-72B-Chat may inadvertently suggest that the speaker's judgment is flawed, causing the speaker to feel defensive and less confident in their decisions.

Our proposed method focuses on respectful and open-ended inquiry, avoiding assumptions and fostering supportive dialogue. By directly asking about the daughter's reaction to the movie without making any assumptions, our approach respects the speaker's judgment. Open-ended questions encourage the speaker to share more details about their experience.

In conclusion, our approach excels by maintaining simplicity and directness, avoiding assumptions, and respecting the speaker's judgment. This method not only encourages the speaker to share more freely but also enhances the overall quality and depth of the conversation.

C Prompts

Figure 8(a) presents the prompts used for annotating sensibility and rationality scores. Figure 8(b) and (c) illustrate the prompts employed for generating empathetic responses using the Qwen and LLaMA models, respectively.

D Algorithms

We release the algorithmic implementation for data selection and MoE-based model training to facilitate reproducibility and enhance understanding of our proposed approach.

Algorithm 1: ED Data Selection Process

Input: Original ED dataset D, threshold T
Output: Sensibility dataset D_s, Discard dataset D_d, Rationality dataset D_r

$D_s \leftarrow \emptyset$;
$D_d \leftarrow \emptyset$;
$D_r \leftarrow \emptyset$;
for each dialogue $d \in D$ **do**
    Assign sensibility score $S(d)$ using ChatGPT with prompt in Figure 8(a);
    Assign rationality score $R(d)$ using ChatGPT with prompt in Figure 8(a);
    if $R(d) < T$ and $S(d) > T$ **then**
        $D_s \leftarrow D_s \cup \{d\}$;
    else if $R(d) > T$ and $S(d) < T$ **then**
        $D_d \leftarrow D_d \cup \{d\}$;
    else
        $D_r \leftarrow D_r \cup \{d\}$;
return D_s, D_d, D_r

Algorithm 2: Empathy MoE Training Process

Input: Sensibility dataset D_s, Rationality dataset D_r, Discard dataset D_d, Seed model LLM, Sensibility expert M_s, Rationality expert M_r, Routing mechanism Router, Feed-Forward Network FFN
Output: Empathy Model M_e

$M_s \leftarrow \text{SFT}(\text{LLM}; D_s)$;
$M_r \leftarrow \text{SFT}(\text{LLM}; D_r)$;
$M_e \leftarrow \emptyset$;
for $layer_i \in M_s$ **do**
    if $layer_i \in$ FFN **then**
        $M_e^i = \text{Router}\cdot M_s^i + (1 - \text{Router}) M_r^i$;
    else
        $M_e^i = \text{Average}(M_s^i, M_r^i)$;
$M_e \leftarrow \text{SFT}(M_e; D_r, D_s, D_n)$;
return M_e

E Comprehensive Experimental Settings

E.1 Experiment Settings

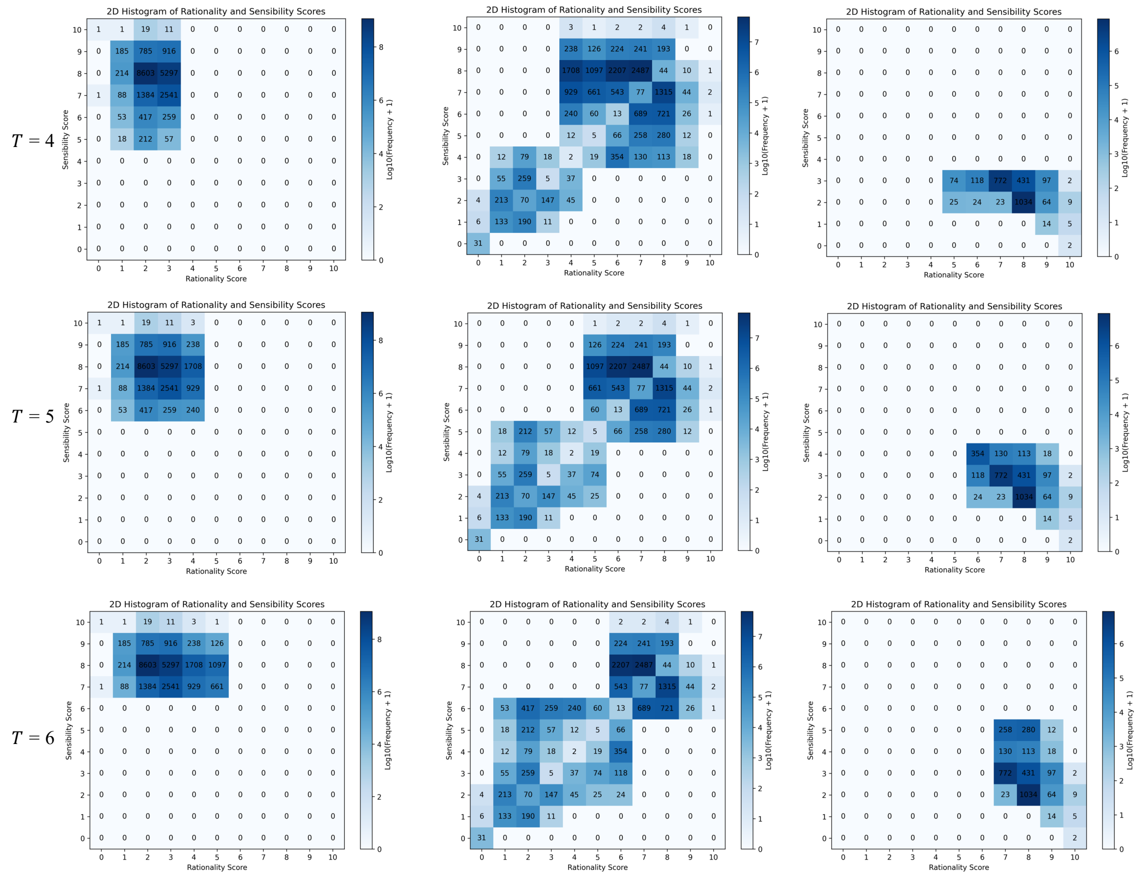
For the four LLMs used in our experiment, we use the hyperparameters from the official repositories: LLaMA3-8B-Instruct[3], Qwen1.5-1.8B-Chat[4], LLaMA2-13B-Chat[5], and Qwen1.5-72B-Chat[6]. All experiments are conducted on an 8*A100 NVIDIA GPU machine with a 120-core CPU and 960GB of memory.

E.2 Baselines

We compare our model with the following baselines:

1. **MoEL** (Lin et al., 2019): Creates a decoder for each emotion to generate a final response. 2. **MIME** (Ghosal et al., 2020): Simulates user emotions and generates empathetic responses by introducing randomness. 3. **EmpDG** (Li et al., 2020): Includes an empathetic information generator and a sentiment discriminator. 4. **CEM** (Sabour et al., 2022): Incorporates the COMET pre-trained model for common sense knowledge in empathetic response generation. 5. **SEEK** (Wang et al., 2022): Focuses on sentence-level sentiment information using attention mechanisms. 6. **CASE** (Zhou et al., 2023): Utilizes external resources COMET and ConceptNet to enhance cognitive and emotional abilities. 7. **E-CORE** (Fu et al., 2023): Explores intrinsic sentiment through emotion correlation learning and supervision. 8. **KEMP** (Li et al., 2022b): Uses ConceptNet and VRC-NED as external knowledge sources for contextual modeling. 9. **CAB** (Gao et al., 2023): Divides empathy response generation into cognition, affection, and behavior. 10. **ESCM** (Yang et al., 2023): Uses dynamic emotion-semantic vectors and dependency trees to guide empathetic response generation. 11. **DCKS** (Cai et al., 2023): Incorporates an adaptive module for commonsense knowledge selection

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[4] https://huggingface.co/Qwen/Qwen1.5-1.8B-Chat
[5] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
[6] https://huggingface.co/Qwen/Qwen1.5-72B-Chat



(a) Sensibility Data (b) Rationality Data (c) Discard Data

Figure 7: Sensibility and Rationality distribution of different selection thresholds. From top to bottom, respectively. Each row represents a threshold, and from left to right, the columns correspond to sensibility data, rationality data, and discard data.

Conversation History:
{Conversation History}
Response:
{Response}

You are an expert at rating Rationality and Sensibility. Based on the provided conversation history and response, please rate the following aspects on a scale of 0 to 10, where 0 is the lowest and 10 is the highest:

Rationality: How much does the data and response focus on logical reasoning and practical considerations?
Sensibility: How much does the data and response focus on feelings, emotions, and personal experiences?

Evaluation Form (scores ONLY):
- Rationality:
- Sensibility:

(a) ChatGPT

<|im_start|>**system**
You are a professional psychologist. Please give a reply from the perspective of a friend based on the patient's situation. Be careful to avoid a large number of professional vocabulary in your reply, and the expression should be natural.
The following is a set of historical conversations between you and your best friend. Please respond according to the context and pay attention to fully consider the emotional information in the historical conversations. Remember, response should not exceed 30 words.<|im_end|>
<|im_start|>**user**
i am glad i stuck my job out as long as i did - i finally got offered a promotion today . <|im_end|>
<|im_start|>**assistant**
congrats ! is it a good promotion ? <|im_end|>
<|im_start|>**user**
it is ! i 'll actually be the manager of the entire office , and not just the one section i was in !<|im_end|>
<|im_start|>**assistant**

(b) Qwen1.5-1.8B/72B-Chat

<|begin_of_text|><|start_header_id|>**system**<|end_header_id|>

You are a professional psychologist. Please give a reply from the perspective of a friend based on the patient's situation. Be careful to avoid a large number of professional vocabulary in your reply, and the expression should be natural.
The following is a set of historical conversations between you and your best friend. Please respond according to the context and pay attention to fully consider the emotional information in the historical conversations. Remember, response should not exceed 30 words.<|eot_id|><|start_header_id|>**user**<|end_header_id|>

i saw my favourite candy at the store today ! was so excited , had n't seen it in a while .<|eot_id|><|start_header_id|>**assistant**<|end_header_id|>

oh nice . what kind of candy was it ?<|eot_id|><|start_header_id|>**user**<|end_header_id|>

this sour s'ghetti candy by haribro . i use to eat this all the time as a child .<|eot_id|><|start_header_id|>**assistant**<|end_header_id|>

(c) LLaMA3-8B-Instruct

Figure 8: Prompts for data annotation and empathetic response generation.

to ensure consistency. 12. **CTSM** (Yufeng et al., 2024): Categorizes emotions into fine-grained trait and state emotions to improve sentiment perception. 13. **Lamb** (Sun et al., 2023): Enhances the empathetic response capability by jointly utilizing self-presentation theory and chain-of-thought data from LLaMA2-13B-Chat.

E.3 Evaluation

Following previous works (Lin et al., 2019; Ghosal et al., 2020; Li et al., 2020; Sabour et al., 2022; Wang et al., 2022; Zhou et al., 2023; Fu et al., 2023; Li et al., 2022b; Gao et al., 2023; Yang et al., 2023; Cai et al., 2023; Sun et al., 2023), we evaluate the performance of our model using both automatic and human evaluation metrics to provide a comprehensive assessment of its capabilities. As illustrated in Figure 8(b) and Figure 8(c), we use a meticulously designed prompt for LLM inference.

Automatic Evaluation Metrics: We use corpus-level BLEU (B-1 to B-4), sentence-level ROUGE (R-1, R-2), and Distinct (Dist-1, Dist-2) as automatic evaluation metrics. BLEU and ROUGE scores quantify the resemblance between the generated text and the ground-truth text, with higher scores indicating greater likeness. Distinct-N evaluates the diversity of the content, with higher values suggesting a wider range of diverse representations. The perplexity metric is not utilized as it measures confidence in the generated sentences, which is not specific to empathy scenarios. Given the absence of authoritative literature demonstrating that LLMs surpass humans in empathy judgment, we rely on scientific human evaluation to assess the methods' effectiveness.

Human Evaluation Metrics: We adopt four complementary human evaluation metrics—Coherence, Empathy, Informativeness, and Continuity—to holistically assess model performance in empathetic dialogue generation. This multi-dimensional design ensures a balanced evaluation of both linguistic quality and emotional intelligence(Fu et al., 2023; Yufeng et al., 2024).

- **Coherence**: Evaluates the correspondence between the text produced by the model and the desired response.
- **Empathy**: Assesses the model's ability to understand the speaker's situation and effectively express concern.
- **Informativeness**: Gauges the amount of information present in the generated responses.
- **Continuity**: Reflects the model's capability to sustain the conversation.

We conduct an A/B test to compare the effectiveness of our model against several baselines. Specifically, we randomly select 200 examples from the test dataset. For each instance, the context is paired with two responses: one generated by our model and the other by a baseline model. Three experienced evaluators assess each pair of responses and determine a winner, a loser, or a tie based on the four dimensions: Coherence, Empathy, Information, and Continuity.

F TEQ Test

To ensure evaluative reliability, all human assessors were required to demonstrate superior empathic capacity, operationalized as the Toronto Empathy Questionnaire (TEQ) scores exceeding the 60-point benchmark.

TEQ represents a psychometrically validated assessment tool engineered to evaluate empathy as a unitary psychological construct, with a predominant emphasis on its affective dimensions. Originally conceptualized by **?**, this instrument comprises 16 self-report items (Table 9), systematically capturing an individual's disposition toward experiencing and manifesting empathic responses. Unlike multidimensional empathy assessments, the TEQ prioritizes the measurement of spontaneous affective reactivity over deliberative cognitive processes such as perspective-taking.

Item responses are quantified via a Likert-type scaling system, wherein positively framed items are scored as follows: Never = 0; Rarely = 1; Sometimes = 2; Often = 3; Always = 4. To mitigate response bias, inversely phrased items undergo reverse scoring prior to analysis. The composite empathy score is subsequently computed through summation, yielding a quantitative index of one's empathic propensity.

Table 9: The Toronto Empathy Questionnaire. *Negatively worded reverse scale questions.

1. When someone else is feeling excited, I tend to get excited too
2. Other people's misfortunes do not disturb me a great deal*
3. It upsets me to see someone being treated disrespectfully
4. I remain unaffected when someone close to me is happy*
5. I enjoy making other people feel better
6. I have tender, concerned feelings for people less fortunate than me
7. When a friend starts to talk about his or her problems, I try to steer the conversation towards something else*
8. I can tell when others are sad even when they do not say anything
9. I find that I am "in tune" with other people's moods
10. I do not feel sympathy for people who cause their own serious illnesses*
11. I become irritated when someone cries*
12. I am not really interested in how other people feel*
13. I get a strong urge to help when I see someone who is upset
14. When I see someone being treated unfairly, I do not feel very much pity for them*
15. I find it silly for people to cry out of happiness*
16. When I see someone being taken advantage of, I feel kind of protective towards him or her