

MENDER: Multi-hop Commonsense and Domain-specific CoT Reasoning for Knowledge-grounded Empathetic Counseling of Crime Victims

Abid Hossain[†], Priyanshu Priya[†], Armita Mani Tripathi[†], Pradeepika Verma[†], Asif Ekbal^{†,‡}

[†]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India

[‡]School of Artificial Intelligence and Data Science, Indian Institute of Technology Jodhpur, India

[†]{abid_2311ai22, priyanshu_2021cs26, armita_2211cs22, asif}@iitp.ac.in,

[†]pradeepikav.verma093@gmail.com, [‡]asif@iitj.ac.in

Abstract

Commonsense inference and domain-specific expertise are crucial for understanding and responding to emotional, cognitive, and topic-specific cues in counseling conversations with crime victims. However, these key evidences are often dispersed across multiple utterances, making it difficult to capture through single-hop reasoning. To address this, we propose **MENDER**, a novel **Multi-hop commonsense** and **domain-specific Chain-of-Thought (CoT)** reasoning framework for **knowledge-grounded empathetic Response** generation in counseling dialogues. **MENDER** leverages large language models (LLMs) to integrate commonsense and domain knowledge via multi-hop reasoning over the dialogue context. It employs two specialized reasoning chains, *viz.* Commonsense Knowledge-driven CoT and Domain Knowledge-driven CoT rationales, which extract and aggregate dispersed emotional, cognitive, and topical evidences to generate knowledge-grounded empathetic counseling responses. Experimental evaluations on counseling dialogue dataset, POEM validate **MENDER**'s efficacy in generating coherent, empathetic, knowledge-grounded responses¹.

1 Introduction

Commonsense inference and domain expertise are crucial for effective mental health and legal counseling of crime victims (Miller, 2008). Since victims often express trauma indirectly, counselors must infer unspoken emotions, intentions, and needs through commonsense reasoning (Dinakar et al., 2012) to foster empathy and trust. In addition, victims need targeted mental health and legal support for their overall well-being. Thus, counselors must have domain expertise to accurately diagnose and treat psychological conditions (Brown, 2007),

¹Code and sample dataset is available at <https://github.com/Abid839/MENDER/tree/main>.

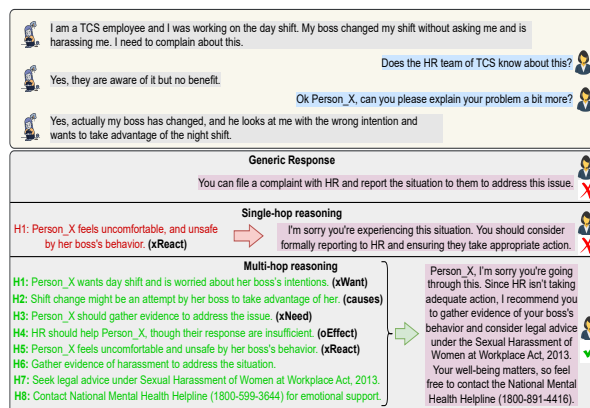


Figure 1: Comparison of generic responses vs. responses generated through single-hop and multi-hop commonsense reasoning and domain expertise.

while also providing precise legal guidance to navigate complex legal systems (Wright et al., 2023). Integrating commonsense reasoning with domain expertise enables counselors to provide holistic, empathetic, and informed support. For instance, as shown in Figure 1, compared to the generic response, commonsense reasoning helps identify victim's emotional and cognitive states, while domain expertise guides legal action and provides mental health resources, thereby ensuring an empathetic and comprehensive support.

Recently, LLMs have been widely used for counseling dialogue systems (Liu et al., 2023; Xie et al., 2024). However, LLMs often fail to capture the emotional and cognitive nuances innate in counseling scenarios, leading to disconnected and insincere interactions (Yang et al., 2024). Further, LLMs tend to generate inconsistent, erroneous, or fabricated information, which can have serious consequences (Chung et al., 2023; Hou et al., 2024). These limitations arise from LLMs' inherent struggles with commonsense inference and domain expertise, restricting their ability to engage in meaningful, knowledgeable, and empathetic conversa-

tions. While recent efforts to incorporate commonsense and domain knowledge (Zhou et al., 2022a,b; Braunschweiler et al., 2023; Varshney et al., 2024) aim to address these issues, they often lead to flawed reasoning, resulting in inadequate responses, as shown in Figure 1.

Commonsense inference and domain knowledge acquisition fundamentally demands multi-hop reasoning, as key implicit information and topic-specific details are often fragmented and distributed across multiple utterances (Zhao et al., 2022; Liu et al., 2021). For instance, generating a coherent response - “Person_X, I’m sorry you’re going...” in Figure 1 involves integrating both implicit (e.g. H1,...,H5) and topic-specific evidences (H6, H7, H8) from dialogue context. These evidences, including both commonsense and domain knowledge, must be identified and aggregated through multiple reasoning steps to produce coherent, empathetic, and knowledge-grounded responses.

Motivated by this, we propose a multi-hop commonsense and domain-specific reasoning process through CoT reasoning. We introduce **MENDER**, a novel framework that integrates **M**ulti-hop commonsens**E** and domai**N**-specific CoT reasoning for knowle**D**ge-grounded empat**H**etic Res**P**onse generation in counseling contexts. **MENDER** first generates two distinct reasoning chains, *viz.* Commonsense Knowledge-driven CoT (CK-CoT) and Domain Knowledge-driven CoT (DK-CoT) rationales to capture the commonsense and domain-specific information required for effective response generation. CK-CoT rationales leverage the ATOMIC knowledge base (Hwang et al., 2021) to infer implicit emotional and cognitive cues, while DK-CoT rationales employ an Entity-guided Retrieval-Augmented Generation (E-RAG) approach to extract relevant topic-specific knowledge. **MENDER** further incorporates rationale-context and rationale-response filters to remove inconsistent or irrelevant rationales, thereby enhancing the overall quality of generated responses. Automatic and human evaluations on POEM (Priya et al., 2023a) dataset show that **MENDER** outperforms baselines, demonstrating its effectiveness in generating empathetic, informative, and coherent counseling responses.

To summarize, key contributions are: (i) Emphasize the need to integrate commonsense reasoning and domain expertise via multi-hop reasoning to gather evidences for knowledge-grounded empathetic counseling responses; (ii) Present **MENDER**, a novel multi-hop commonsense and domain-

specific CoT reasoning framework for knowledge-grounded empathetic response generation during counseling; (iii) Design two reasoning chains: CK-CoT and DK-CoT rationales to capture emotional, cognitive, and topic-specific information for effective response generation; (iv) Introduce rationale-context and rationale-response filters to ensure the consistency and relevance of generated rationales.

2 Related Work

Recent efforts have focused on developing dialogue systems for mental health and legal counseling of crime victims (Kim et al., 2022; Singh et al., 2022b; Mishra et al., 2023b,c; Priya et al., 2023b; Mishra et al., 2023a; Priya et al., 2024a,b), with an emphasis on using LLMs (Zhao et al., 2023) for automated counseling agents (Liu et al., 2023; Chen et al., 2024). Integrating commonsense knowledge, such as emotional and cognitive insights, has been shown to improve empathy and contextual relevance (Wu et al., 2020; Sabour et al., 2022; Tu et al., 2022; Reddy et al., 2023), while domain-specific knowledge from external sources like Wikipedia enhances factual accuracy and relevance (Zhao et al., 2020; Qin et al., 2023; Bai et al., 2023). Chain-of-Thought (CoT) prompting has been found to enhance reasoning by decomposing complex problems into manageable steps (Wei et al., 2022), and Retrieval-Augmented Generation (RAG) leverages external knowledge to improve accuracy and reliability (Gao et al., 2023). Despite these advancements, many existing models still struggle to effectively integrate commonsense and domain knowledge, resulting in superficial responses. This work proposes a framework combining multi-hop commonsense reasoning with ATOMIC knowledge (Hwang et al., 2021) and domain-specific reasoning using Entity-guided Retrieval-Augmented Generation (E-RAG), to seamlessly generate coherent, empathetic, and knowledge-grounded responses.

3 Methodology

3.1 Overview

Given a dialogue corpus $\mathcal{D} = (C, R)^{|\mathcal{D}|}$, where $C = \{u_1, \dots, u_{t-1}\}$ is dialogue context with an alternating sequence of $(t - 1)$ utterances between counseling agent and victim, and R is response, the goal is to generate counselor’s response $R(= u_t)$. To improve response quality, we integrate external commonsense and domain knowledge K . The response generation task is thus de-

defined as $P(R | C, K)$, aiming for responses that are cotextually coherent, informative, and empathetic to victim’s situation and emotional state.

3.2 Approach

We introduce **MENDER**, a **M**ulti-hop commonsense and domain-specific Chain-of-Thought (CoT) reasoning framework for knowledge-grounded empathetic Response generation that constructs commonsense knowledge-driven CoT reasoners and domain knowledge-driven CoT reasoners to enhance response generation. To enhance relevance, we apply reasoner filtering mechanisms to refine the generated reasoners, which are then used for response generation. Figure 2 provides an overview of the proposed **MENDER** framework.

3.2.1 Commonsense Knowledge-driven CoT (CK-CoT) Reasoning

To construct CK-CoT reasoning chains (rationales, hereafter), we utilize LLMs’ reasoning capability and commonsense relations from the ATOMIC knowledge base (Hwang et al., 2021), including *xIntent*, *xNeed*, *xWant*, *xReact*, *oEffect*, and *Causes* to capture emotional (affective), cognitive, and causal aspects of human reasoning. Given a dialogue context C and ground-truth response R , we prompt the LLM, \mathcal{M} to generate CK-CoT rationales S^{Co} , such that R can be induced from S^{Co} . These rationales are defined as a sequence of n query-reply pairs $(q_i, r_i)_{i=1}^n$, where each q_i denotes an information-seeking question designed to uncover implicit information r_i within C . To generate these pairs, we introduce *thought-then-generate* approach employing two-step CoT process:

$$\mathcal{T}^{Co} \leftarrow P_{\mathcal{M}}(C, cs_{rel}) \quad (1)$$

$$S^{Co} \leftarrow P_{\mathcal{M}}(C, cs_{rel}, \mathcal{T}^{Co}) \quad (2)$$

In the first step (Eq. 1), we prompt \mathcal{M} to think what queries should be implicitly inferred from C using commonsense relations cs_{rel} and generate corresponding thoughts \mathcal{T}^{Co} . In the second step (Eq. 2), based on \mathcal{T}^{Co} , we prompt \mathcal{M} to generate queries using cs_{rel} followed by the respective replies based on C .

3.2.2 Domain Knowledge-driven CoT (DK-CoT) Reasoning

To generate DK-CoT rationales, we utilize LLMs reasoning capability and external domain knowledge. For a given context C and ground-truth response R , we prompt \mathcal{M} to generate DK-CoT rationales S^{Do} that lead to R . The process involves re-

trieving relevant domain knowledge for C and generating m query-reply pairs $(q_i, r_i)_{i=1}^m$, where each query q_i seeks topic-specific information r_i in C . This is achieved through a novel Entity-guided Retrieval Augmented Generation (E-RAG) approach, described as follows:

(a) *Entity-guided Knowledge Source Selection*: To ensure accuracy and comprehensiveness in domain knowledge collection, we systematically extract relevant entities related to mental health and legal counseling from the dialogue dataset \mathcal{D} using the Stanford Named Entity Recognition (NER) Tagger (Finkel et al., 2005), T . Each utterance u_i is processed to identify entities, denoted as $E_i = T(u_i)$, and the total set of entities across dataset is defined as $E = \bigcup_{i=1}^{|\mathcal{D}|} E_i$. These entities are categorized into four classes, denoted as $EC = \{\text{crime, mentalhealth, legal, medium}\}$. For each entity class $ec \in EC$, we define $E_{ec} \subset E$ containing entities of type ec . To ensure robustness and remove noise, we filter out entities, such that $E'_{ec} = \{e \in E_{ec} \mid \text{frequency}(e, \mathcal{D}) \geq 2\}$. Afterward, we perform a manual verification M of identified entities to eliminate false positives and misspelled entities, yielding the final entity set $E_{fec} = M(E'_{ec})$. We then create question templates, Q_{fec} for each entity class ec , and formulate web search queries for each entity $e' \in E_{fec}$ using these templates. The queries are searched on Google via Google Search API (Google, 2023) to gather domain knowledge from top 25 matching URLs. A summary of entity types, examples, and query templates is given in Appendix A. The extracted textual content is cleaned and stored into M knowledge documents (KD), which serve as knowledge base for the next step.

(b) *Knowledge Retrieval*: To retrieve relevant knowledge, we encode knowledge documents and dialogue context. The knowledge document encoder encodes each knowledge document $\{KD_j\}_{j=1}^M$ into vector representations h_{KD_j} . Likewise, the context encoder encodes C into a vector representation h_C . To assess the relevance of each knowledge document to the context, two matching scores, $s_{KD_j, C}^1$ and $s_{KD_j, C}^2$ are computed using BM25 (Robertson and Walker, 1994) and FAISS (Douze et al., 2024) retrievers, respectively, as:

$$s_{KD_j, C}^l = \begin{cases} \text{BM25}(h_{KD_j}, h_C); & l = 1, \quad \forall KD_j \in KD \\ \text{FAISS}(h_{KD_j}, h_C); & l = 2, \end{cases}$$

BM25-based sparse retrieval captures surface-level similarity, while FAISS-based dense retrieval emphasizes high-level semantic relevance. Each re-

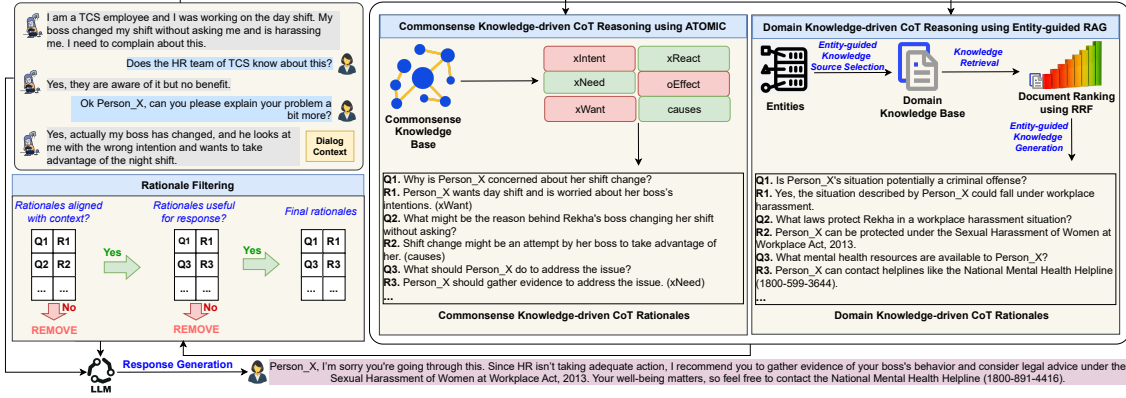


Figure 2: Overview of the proposed MENDER framework.

triever ranks knowledge documents independently based on computed matching scores. To integrate these rankings, the Reciprocal Rank Fusion (RRF) algorithm (Cormack et al., 2009) is applied, ensuring accurate, balanced, and robust ranking for effective knowledge retrieval. It is computed as:

$$RRF(KD_j) = \sum_{p \in P} \frac{1}{k + p(KD_j)}, \forall KD_j \in KD \quad (3)$$

Here, $P = \{\text{BM25, FAISS}\}$ represents the set of retrievers, and $p(KD_j)$ denotes the rank assigned to document KD_j by retriever p . The smoothing constant k adjusts rank weights to mitigate biases introduced by individual retrievers. The final ranking is derived from the computed RRF scores, yielding an ordered list of knowledge documents, most relevant to the context C . Finally, the top- r documents are retrieved based on their RRF scores.

(c) *Entity-guided Knowledge Generation*: To extract relevant knowledge from top- r documents, we generate query-reply pairs based on C and retrieved knowledge. To generate precise and contextually relevant queries, we employ an entity-centric approach that identifies key entities within C and uses them as anchors for query construction. The generated query is then mapped to the most relevant knowledge sentences to extract precise replies. To enable deeper and interpretable reasoning, we again employ *think-then-generate* approach for generating query-reply pairs in four-step CoT manner:

$$\mathcal{T}_1^{Do} \leftarrow P_{\mathcal{M}}(C) \quad (4)$$

$$\hat{\mathcal{E}}^{Do} \leftarrow P_{\mathcal{M}}(C, \mathcal{T}_1^{Do}) \quad (5)$$

$$\mathcal{T}_2^{Do} \leftarrow P_{\mathcal{M}}(C, \hat{\mathcal{E}}^{Do}) \quad (6)$$

$$\mathcal{S}^{Do} \leftarrow P_{\mathcal{M}}(C, \hat{\mathcal{E}}^{Do}, \mathcal{K}^{Do}, \mathcal{T}_2^{Do}) \quad (7)$$

In the first step (Eq. 4), we prompt \mathcal{M} to think what entities could be extracted from given context C and generate corresponding thoughts \mathcal{T}_1^{Do} . In the second step (Eq. 5), based on \mathcal{T}_1^{Do} , we first

ask \mathcal{M} to extract the entities \mathcal{E}^{Do} from C and then assess the relevance of each entity to context C . To achieve this, entity encoder encodes each extracted entity $e_i \in \mathcal{E}^{Do}$ into a vector representation h_{e_i} . The similarity score $s_{e_i, C}$ is then computed as the dot product between h_{e_i} and h_C . To enhance query-reply alignment, extracted entities are filtered based on two factors: (i) *entity order*, which prioritizes entities with higher similarity to C , and (ii) *entity confidence*, which categorizes entities into low, moderate, and high confidence levels based on similarity scores². Entity order in conjunction with entity confidence ensures that entities with low confidence but high entity order are disregarded to focus on highly relevant entities. In the third step (Eq. 6), we prompt \mathcal{M} to think what queries can be inferred from the C based on $\hat{\mathcal{E}}^{Do}$ and write the corresponding thoughts \mathcal{T}_2^{Do} . Finally, in the fourth step (Eq. 7), based on \mathcal{T}_2^{Do} , we prompt \mathcal{M} to generate queries using $\hat{\mathcal{E}}^{Do}$ and formulates corresponding replies based on k retrieved documents, denoted as \mathcal{K}^{Do} .

3.2.3 Rationale Filtering

LLMs tend to hallucinate facts without adequately attending to the context (Peng et al., 2023), and not all rationales are effective in generating responses. Thus, to ensure that rationales are both contextually aligned and useful, we introduce rationale-context and rationale-response filters, respectively. For rationale-context filter, we employ $\text{alignment}(\cdot)$ function to assess if a rationale $z_i \in Z$, where $Z = \{\mathcal{S}^{Co}, \mathcal{S}^{Do}\}$ is relevant for C . For rationale-response filter, we introduce $\text{useful}(\cdot)$ function to assess if a dialogue model θ benefits from a rationale when predicting response R , given a context

²Thresholds for categorizing entities are hyper-parameters (Appendix 4.4).

C .

$$\text{alignment}(z_i) = \begin{cases} 1, & \text{if } \frac{SE(z_i).SE(C)}{\|SE(z_i)\| \|SE(C)\|} > \tau_1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{useful}(z_i) = \begin{cases} 1, & \text{if } \frac{P_\theta(R|z_i,C)}{P_\theta(R|C)} > \tau_2, \\ 0, & \text{otherwise.} \end{cases}$$

Here, SE denotes sentence encoder used to obtain semantic representations for z_i and C and τ_1, τ_2 are hyperparameters. Intuitively, a higher similarity and higher probability suggests that rationale z_i is contextually aligned and useful for predicting response R .

3.2.4 Response Generation

Finally, we instruct \mathcal{M} to generate the response for a given dialogue context C using previously generated CK-CoT rationales (\mathcal{S}^{Co}) and DK-CoT (\mathcal{S}^{Do}) rationales:

$$R \leftarrow P_{\mathcal{M}}(C, \mathcal{S}^{Co}, \mathcal{S}^{Do}, I) \quad (8)$$

where, I denotes the instruction given to \mathcal{M} .

4 Experiments

4.1 Dataset

We conduct experiments on POEM dataset (Priya et al., 2023a) of counseling conversations, where commonsense reasoning and domain expertise are vital for delivering contextually appropriate, empathetic, and informative responses to crime victims. We choose this dataset for our task due to its rich coverage of real-world scenarios involving mental health and legal counseling needs of diverse crime victims. The dataset contains 5K dialogues crafted using real-life stories from credible sources, including news articles, case studies, and government portals under expert supervision. The comprehensive scope along with grounding in authentic sources and expert supervision, makes POEM dataset an ideal choice for developing models for commonsense and domain knowledge-grounded empathetic response generation during counseling.

4.2 Baselines

We compare MENDER with 9 baselines: ITDD (Li et al., 2019), KnowledGPT (Zhao et al., 2020), CEM (Sabour et al., 2022), MISC (Tu et al., 2022), MSDP (Liu et al., 2022), CoT (Wei et al., 2022), ProCoT (Deng et al., 2023), O-Cue-CoT and M-Cue-CoT (Wang et al., 2023). We include ‘Baselines Details’ in Appendix B.

4.3 Evaluation Metrics

For automatic evaluation, we use Perplexity (PPL) (Brown et al., 1992), BLEU (B-4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), Distinct-2 (D-2) (Li et al., 2015), BERTScore-f1 (BS-f1) (Zhang et al., 2019), Embedding Average (EA), Vector Extrema (VE), and Greedy Matching (GM) (Liu et al., 2016) to evaluate general quality of responses. To assess responses for task performance, we measure Domain Knowledge Coverage (DKC), Commonsense Knowledge Coverage (CKC), and Emotion Expression Accuracy (E-ACC). For human evaluation, we use Fluency (F), Adequacy (A), Contextual Relevance (CR) (Singh et al., 2022a) to assess responses’ general quality. To assess responses for task performance, we employ Knowledge Existence (KE), Knowledge Correctness (KC), Knowledge Relevance (KR) (Varshney et al., 2022), Helpfulness (H), Safety (S), and Empathy (Emp.). We include ‘Evaluation Metrics Details’ in Appendix B.

4.4 Implementation Details

All implementations are conducted using PyTorch³, and we employ transformer-based models from Hugging Face (Wolf et al., 2019) throughout our experiments. We use pre-trained Sentence-BERT (Reimers and Gurevych, 2019) as knowledge and context encoders and BERT (Devlin et al., 2018) as entity encoder. The dense retriever FAISS is implemented based on *mixedbread-ai/mxbai-embed-large-v1*, an embedding model. It will rank the documents based on the embedding L2 (Euclidean) distance between each knowledge document and dialogue context. We select top-2 knowledge documents (i.e. $r = 2$). We empirically set hyperparameters: k to 60 (smoothing constant in RRF score calculation), τ_1 to 0.6 and τ_2 to 0.9. Further, we empirically set the following ranges: $s_{e_i,C} \leq 0.3$ indicates low confidence, $0.3 < s_{e_i,C} \leq 0.65$ indicates moderate confidence, and $s_{e_i,C} > 0.65$ indicates high confidence for entity confidence categorization.

In the rationale-response filter, we use Zephyr-7B (Tunstall et al., 2023) trained on diverse synthetic dialogues generated by ChatGPT as dialogue model θ . For rationale generation, we employ LLaMA-3.1-8B-Instruct (Touvron et al., 2023) and for response generation, we use LLaMa-2-7B-chat (Touvron et al., 2023). We use Top-p sampling with

³<https://pytorch.org/>

Models	PPL ↓	B-4 ↑	M ↑	D-2 ↑	BS-f1 ↑	EA ↑	VE ↑	GM ↑	DKC ↑	CKC ↑	E-ACC ↑
ITDD	31.25	1.02	6.23	15.14	0.421	0.571	0.226	0.482	4.67	9.54	7.41
KnownGPT	28.11	2.67	7.14	18.32	0.473	0.622	0.284	0.532	7.03	14.32	10.87
CEM	28.80	4.98	8.72	19.45	0.486	0.643	0.309	0.546	9.26	18.64	14.58
MISC	27.04	5.22	9.24	20.67	0.512	0.665	0.336	0.563	10.41	20.52	16.47
MSDP	25.73	6.07	10.56	20.73	0.537	0.694	0.372	0.591	11.87	23.56	18.62
CoT	17.53	6.44	10.89	21.34	0.553	0.713	0.401	0.612	13.22	26.48	20.93
ProCoT	14.41	6.53	11.03	23.12	0.603	0.744	0.423	0.637	16.12	31.78	24.71
O-Cue-CoT	11.26	6.71	11.78	24.48	0.627	0.767	0.442	0.654	18.36	36.48	27.42
M-Cue-CoT	9.35	6.59	12.61	27.78	0.652	0.801	0.467	0.682	22.41	44.12	32.17
MENDER	6.33	9.31	14.02	31.56	0.703	0.881	0.499	0.726	26.19	49.82	36.79
- S^{Co}	9.52	7.99	12.57	30.22	0.671	0.845	0.480	0.704	24.98	44.67	32.24
- S^{Do}	9.45	8.12	13.02	30.53	0.682	0.860	0.485	0.715	23.10	45.82	33.65
- ($S^{Co} + S^{Do}$)	11.76	5.65	10.10	28.90	0.657	0.830	0.471	0.695	21.85	41.45	29.80

Table 1: Automatic evaluation results. Results are statistically significant at 5% significance level based on t-test (Welch, 1947).

Models	F (1-5)	A (1-5)	CR (1-5)	KE (0-2)	KR (0-2)	KC (0-2)	H (0-2)	S (0-1)	Emp. (1-5)
MSDP	2.33	2.45	2.38	1.02	1.10	1.09	1.29	1.0	2.30
CoT	2.87	2.94	2.71	1.15	1.22	1.21	1.35	1.0	2.64
ProCoT	3.21	3.19	3.03	1.28	1.35	1.33	1.50	1.0	3.08
O-Cue-CoT	3.47	3.39	3.34	1.41	1.47	1.45	1.62	1.0	3.28
M-Cue-CoT	3.73	3.64	3.58	1.53	1.60	1.58	1.74	1.0	3.59
MENDER	4.12	4.25	4.41	1.72	1.78	1.85	1.91	1.0	4.50

Table 2: Human evaluation results. Results are statistically significant at 5% significance level based on t-test (Welch, 1947). Scale for metrics are given in column heads.

$p = 0.9$ and temperature $\tau = 0.6$ for rationale and response generation. For rationales’ generation, we include two exemplars of rationales, with manually constructed query-reply pairs to further guide the LLM in identifying relevant contextual cues and inferring necessary knowledge for response generation. Likewise, for response generation, we include two exemplars consisting of dialogue context, corresponding rationales and ground-truth response to guide the model toward generating appropriate response. All experiments are done on Tesla V100-PCIE-32GB GPUs.

5 Results and Analysis

5.1 Automatic Evaluation

Table 1 presents the results of the automatic evaluation. We observe that MENDER significantly outperforms all baselines across all metrics. It achieves the lowest PPL score, indicating the superior quality of its generated responses compared to baselines. Further MENDER excels in dialogue quality, as revealed by its superior lexical (B-4, M) and semantic richness (BS-f1, EA, VE, GM), along with its ability to produce more diverse responses (D-2). The highest DKC and CKC scores shows its proficiency in capturing knowledge, enabling the generation of engaging and informative responses. Besides, the highest E-ACC score high-

lights MENDER’s ability to generate empathetic responses. Notably, the ablation results show that removing either CK-CoT rationales (S^{Co}), DK-CoT rationales (S^{Do}), or both causes a significant drop in performance, emphasizing the critical role of both reasoning steps in generating knowledge-grounded, empathetic responses.

5.2 Human Evaluation

Table 2 presents the results of the human evaluation. We compare MENDER against MSDP, CoT, ProCoT, O-Cue-CoT, and M-Cue-CoT only, as manual evaluation is expensive. It is evident that MENDER consistently outperforms baseline models across all evaluation metrics. This highlights MENDER’s ability to effectively integrate commonsense reasoning and domain knowledge, generating responses that are notably more coherent, empathetic, and informative. The inter-evaluator agreement, measured using Fleiss’ kappa (McHugh, 2012) (κ), falls within the range [0.45, 0.81] for all criteria, indicating fair to moderate agreement among evaluators.

6 Conclusion

This work presents MENDER, a multi-hop reasoning framework that integrates commonsense and domain-specific knowledge for generating empathetic, knowledge-grounded responses in counseling dialogues. Using commonsense and domain knowledge-driven CoT rationales, MENDER captures emotional, cognitive, and topic-specific details to ensure coherent, empathetic, and informed responses. Extensive experiments on the POEM dataset demonstrate the promising potential of MENDER in generating coherent, empathetic, and knowledge-grounded responses, significantly improving the quality of counseling outcomes.

Limitations

We evaluate MENDER on the POEM dialogue dataset, focusing on crime victim counseling and dyadic dialogues. While currently limited in scope, future work could extend MENDER to other counseling domains and multi-party dialogues. Since its reasoning generations are entirely machine-generated, caution is advised to avoid biases in model training. Using LLaMA-2-7b-chat as the base model, our experiments yield satisfactory results, but further validation and optimization are planned, including performance enhancement on smaller models like Gemma-2B (Team et al., 2024) via knowledge distillation. Budget and computational constraints necessitated the use of open-source LLaMA-2-7b-chat, but future studies could explore advanced closed-source LLMs, such as GPT-4 (OpenAI, 2024) or Gemini-1.5-Pro (Team et al., 2023), for generating rationales and responses, thereby improving system performance.

Ethics Statement

This study was reviewed and approved by our Institutional Review Board (IRB). For our research, we utilized the POEM dataset, which comprises dialogues focused on mental health and legal counseling for crime victims. Permission to use this dataset was obtained in compliance with the copyright guidelines set by the copyright holder. Given the sensitive nature of the research, strict measures were implemented to ensure the privacy and confidentiality of victim-related data throughout the study. To ensure ethical considerations in generating rationales and responses with large language models (LLMs), the prompts were carefully designed to emphasize adherence to specific ethical guidelines, minimizing the risk of producing inappropriate or harmful content. Additionally, we incorporated a ‘Safety’ dimension into the evaluation framework to assess the system’s capacity for providing effective and responsible counseling support.

Acknowledgements

Authors acknowledge the partial support from the project titled “An Empathetic Knowledge Grounded Conversational System for Mental Health Counseling and Legal Assistance“, Sponsored by IHUB Anubhuti, TIH, IIIT Delhi. Priyanshu Priya acknowledges the financial support provided by the Department of Science and

Technology, Ministry of Science and Technology, Government of India and through the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship.

References

- Jiaqi Bai, Ze Yang, Jian Yang, Hongcheng Guo, and Zhoujun Li. 2023. Kinet: Incorporating relevant facts into knowledge-grounded dialog generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1213–1222.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Norbert Braunschweiler, Rama Doddipatla, Simon Keizer, and Svetlana Stoyanchev. 2023. Evaluating large language models for document-grounded response generation in information-seeking dialogues. *arXiv preprint arXiv:2309.11838*.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.
- Sandra L Brown. 2007. *Counseling victims of violence: A handbook for helping professionals*. Hunter House.
- Yixiang Chen, Xinyu Zhang, Jinran Wang, Xurong Xie, Nan Yan, Hui Chen, and Lan Wang. 2024. Structured dialogue system for mental health: An llm chatbot leveraging the pm+ guidelines. *arXiv preprint arXiv:2411.10681*.
- Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

- Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Google. 2023. Google search api. <https://github.com/googleapis/google-api-python-client/>.
- Abe Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations. In *Proceedings of the Natural Language Processing Workshop 2024*, pages 280–302.
- Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.
- Hyeok Kim, Youjin Hwang, Jieun Lee, Youngjin Kwon, Yujin Park, and Joonhwan Lee. 2022. Personalization trade-offs in designing a dialogue-based information system for support-seeking of sexual violence survivors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. *arXiv preprint arXiv:1907.08854*.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.
- Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.
- Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. *arXiv preprint arXiv:2203.08745*.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Laurence Miller. 2008. *Counseling crime victims: Practical strategies for mental health professionals*. Springer Publishing Company.
- Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023a. e-therapist: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967.
- Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023b. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14408–14416.
- Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023c. Pal to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271.
- OpenAI. 2024. ChatGPT. <https://chatgpt.com/>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997.
- Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

- Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2023a. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications*, 224:120025.
- Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2024a. Two in one: A multi-task framework for politeness turn identification and phrase extraction in goal-oriented conversations. *Computer Speech & Language*, 88:101661.
- Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. 2023b. **Partner: A persuasive mental health and legal counselling dialogue system for women and children crime victims**. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6183–6191. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Priyanshu Priya, Gopendra Singh, Mauajama Firdaus, Jyotsna Agrawal, and Asif Ekbal. 2024b. On the way to gentle ai counselor: Politeness cause elicitation and intensity tagging in code-mixed hinglish conversations for social good. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4678–4696.
- Lang Qin, Yao Zhang, Hongru Liang, Jun Wang, and Zhenglu Yang. 2023. Well begun is half done: Generator-agnostic knowledge pre-selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2310.07659*.
- Revanth Reddy, Hao Bai, Wentao Yao, Sharath Chandra Etagi Suresh, Heng Ji, and ChengXiang Zhai. 2023. Social commonsense-guided search query generation for open-domain knowledge-powered conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 873–885.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- SE Robertson and S Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- Gopendra Vikram Singh, Mauajama Firdaus, Shruti Mishra, Asif Ekbal, et al. 2022a. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900.
- Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. *arXiv preprint arXiv:2205.13908*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. **Zephyr: Direct distillation of lm alignment**. *Preprint, arXiv:2310.16944*.
- Deeksha Varshney, Asif Ekbal, and Erik Cambria. 2024. Emotion-and-knowledge grounded response generation in an open-domain dialogue setting. *Knowledge-Based Systems*, 284:111173.
- Deeksha Varshney, Akshara Prabhakar, and Asif Ekbal. 2022. Commonsense and named entity aware knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335.
- Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *arXiv preprint arXiv:2305.11792*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

- et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Bernard L Welch. 1947. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Naomi M Wright, Tejaswinhi Srinivas, Michelle Seulki Lee, and Anne P DePrince. 2023. Development of a legal service needs and barriers measure for victims of family violence and other crime. *Journal of family violence*, pages 1–19.
- Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5811–5820.
- Nan Xie, Yuelin Bai, Hengyuan Gao, Ziqiang Xue, Feiteng Fang, Qixuan Zhao, Zhijian Li, Liang Zhu, Shiwen Ni, and Min Yang. 2024. Delilaw: A chinese legal counselling system based on a large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5299–5303.
- Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Chao Zhao, Wenlin Yao, Dian Yu, Kaiqiang Song, Dong Yu, and Jianshu Chen. 2022. Learning-by-narrating: Narrative pre-training for zero-shot dialogue comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–218.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.
- Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022a. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468.
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252.

Appendix

A Mental Health and Legal Counseling-related Entities

After extraction and manual review of the entities, we identify a total of 41 entities, systematically categorized into four groups as 16 crime-related entities, 4 related to mental health issues, 7 associated with legal information, and 14 pertaining to medium information. The entities related to crime, mental health issues, legal information, and medium information correspond to 7, 6, 6, and 4 distinct query templates, respectively. A detailed overview of the entity types, representative entity examples, and their corresponding query templates is provided in Table 3.

Entity Type	Examples	Sample Query	Query Example
Crime	cyber-stalking, harassment	What is crimeX?	What is Stalking?
		How to prevent crimeX?	How to prevent Stalking?
Mental Health	depression, anxiety stress	What is issueX?	What is Depression?
		What are the symptoms of issueX?	What are the symptoms of Depression?
Legal	section354D, cybercell	What is sectionX?	What is section 354D?
		What are the punishments under sectionX?	What are the punishments under section 354D?
Medium	facebook, instagram	How to report crimeX on mediumX?	How to report online stalking on Facebook?
		How to block a profile/page on mediumX?	How to block a profile/page on Facebook?

Table 3: Entity types, entities examples, and their corresponding query templates.

B Experiment Details

B.1 Baseline Details

1. ITDD (Li et al., 2019): Utilizes an incremental transformer architecture to encode utterances and external knowledge, coupled with a deliberation-based decoder for generating responses.
2. KnowledGPT (Zhao et al., 2020): Incorporates a pre-trained language model alongside a knowledge selection module, with both components jointly optimized using reinforcement learning.
3. CEM (Sabour et al., 2022): Leverages commonsense reasoning to enhance the expression of empathy in generated responses.
4. MISC (Tu et al., 2022): Fuses commonsense knowledge for emotional response generation.
5. MSDP (Liu et al., 2022): Employs a multi-stage prompting framework that first generates

relevant knowledge and then use the generated knowledge to predict the response for a given dialogue context.

6. CoT (Wei et al., 2022): Employs a standard few-shot CoT reasoning approach to generate knowledge-grounded empathetic responses.
7. ProCoT (Deng et al., 2023): Prompts the LLM to generate a chain-of-thought descriptive analysis to use the relevant the knowledge by performing dynamic reasoning for generating the knowledge-grounded empathetic responses.
8. O-Cue-CoT (Wang et al., 2023): Prompts the LLMs to generate knowledge and a final response simultaneously for the given dialogue context, enforcing the LLM to reason based on the knowledge.
9. M-Cue-CoT (Wang et al., 2023): M-Cue-CoT builds on the foundation of O-Cue-CoT by decomposing the reasoning process into consecutive steps. It first generates the reasoning to infer the relevant knowledge and then use the inferred knowledge to predict the final response.

B.2 Evaluation Metrics Details

Automatic Evaluation Metrics. Perplexity (PPL) (Brown et al., 1992) evaluates how well the model predicts a response. Word-overlap-based metrics like BLEU (Papineni et al., 2002) (B-4) and METEOR (Banerjee and Lavie, 2005)(M) compute the overlap between the ground-truth response and the model’s generated response. DISTINCT-2 (Li et al., 2015) (D-2) measures the diversity of the generated responses. BERTScore-f1 (Zhang et al., 2019) (BS-f1), Embedding Average (EA), Vector Extrema (VE), and Greedy Matching (GM) (Liu et al., 2016) align the generated response and the ground-truth response in latent semantic space to assess the semantic similarity between the gold response and the model’s generated response.

Domain Knowledge Coverage (DKC) using KF1 (Shuster et al., 2021) quantifies unigram word overlap between the generated response (R) and domain knowledge (K) (Equation 9), Commonsense Knowledge Coverage (CKC) using Hard Matching (Zhou et al., 2022b) identifies matching commonsense tuples T between the dialogue context (C) and the generated response (R) (Equation 10), and

Models	PPL ↓	B-4 ↑	M ↑	D-2 ↑	BS-f1 ↑	EA ↑	VE ↑	GM ↑	DKC ↑	CKC ↑	E-ACC ↑
MENDER (w query-reply pairs)	6.33	9.31	14.02	31.56	0.703	0.881	0.499	0.726	26.19	49.82	36.79
MENDER (w replies only)	7.46	8.12	13.78	29.01	0.689	0.864	0.489	0.712	25.73	46.92	33.07

Table 4: Ablation results w.r.t iterative query-reply pairs on generating question in MENDER.

Models	PPL ↓	B-4 ↑	M ↑	D-2 ↑	BS-f1 ↑	EA ↑	VE ↑	GM ↑	DKC ↑	CKC ↑	E-ACC ↑
MENDER	6.33	9.31	14.02	31.56	0.703	0.881	0.499	0.726	26.19	49.82	36.79
MENDER (w/o R-C filter)	6.49	9.10	13.74	30.89	0.686	0.860	0.485	0.708	25.55	48.32	35.74
MENDER (w/o R-R filter)	7.67	8.87	13.45	30.18	0.668	0.839	0.470	0.690	24.92	46.92	34.70
MENDER (w/o R-C and R-R filter)	8.85	8.65	13.16	29.52	0.650	0.818	0.455	0.672	24.30	45.58	33.68

Table 5: Ablation results w.r.t rationale-context (R-C) and rationale-response (R-R) filters in MENDER.

EXP ACC (E-ACC) (Pascual et al., 2021) measures the accuracy of emotion expression.

$$DKC = \frac{1}{m} \sum_{i=1}^m \text{KF1}(R, K) \quad (9)$$

$$CDC = \frac{1}{m} \sum_{i=1}^m I(\{T_i\}), \quad I = \begin{cases} 1 & \text{if } \{T_i\} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

where m is the test set size, and $I = 1$ if the response is grounded by at least one commonsense tuple.

Human Evaluation Metrics. Fluency (F) assesses the grammatical correctness, Adequacy (A) quantifies the semantic similarity of the generated response with that of the ground-truth response, Contextual Relevance (CR) examines the alignment of the generated responses with the dialogue context.

KE evaluates the incorporation of knowledge within the response, KC measures the accuracy of this knowledge, and KR examines whether the knowledge is both accurate and contextually relevant to the dialogue context. Helpfulness (H) assess whether the generated response satisfies the victim’s requirement, Safety (S) gauges if the generated response safeguards personal privacy and adheres to relevant laws and regulations, and Empathy (Emp.) assesses whether the response is more understanding of the user’s emotion and situation and shows the appropriate emotion.

C Human Evaluation Process

The human evaluation is conducted with the assistance of three evaluators, two hold Ph.D. degrees in Linguistics and one with a graduate degree in Computer Science and Engineering. All evaluators are proficient in English and have substantial experience in similar tasks. For evaluation, we randomly

selected 120 samples consisting of dialogue context, ground-truth response, commonsense knowledge, domain knowledge, and model-generated response. Prior to the evaluation, they are briefed on the evaluation guidelines along with few samples, and are instructed to rate each sample for F, A, CR, KE, KC, KR, H, S, and Emp. on a provided scale.

C.1 Prompt Templates for MENDER

The prompts of our proposed MENDER are reported in Table 7 (Commonsense Knowledge-driven CoT Reasoning), Table 8 (Domain Knowledge-driven CoT Reasoning), and Table 9 (Response Generation).

D Additional Analysis

D.1 Ablation w.r.t Iterative Query-Reply Pairs

To assess the impact of queries, we conduct an ablation study by prompting the model under the same conditions as MENDER but generating only replies. Specifically, we remove queries from the rationales and prompt the model using the modified sample. As presented in Table 4, the absence of queries leads to a significant decline in response quality, highlighting their critical role in reasoning. This suggests that queries play a crucial role in guiding replies generation, as responses exhibit poor alignment with dialogues in their absence.

D.2 Ablation w.r.t filters

To assess the impact of rationale-context and rationale-response filters, we ablate the filters and done the experiments under the same conditions as MENDER. In the first ablation, we ablate the rationale-context filter, in the second ablation, we remove the rationale-response filter, and finally in the third ablation, we omit both filters. The results, presented in Table 5, indicate a decline in response

quality when the generated rationales fail to support accurate next response prediction. Notably, the **MENDER**'s performance deteriorates significantly when the rationale-context filter is removed, underscoring the critical role of maintaining alignment between rationales and contexts. Furthermore, when the rationale-response filter is removed, the overall response quality further degrades. A significant performance drop is observed when both the filters are omitted. These results demonstrate the significance of both filters in generation adequate responses.

D.3 Case Study

Table 6 presents examples of responses generated by the proposed **MENDER** framework, alongside four strong baselines - CoT, ProCoT, O-Cue-CoT, and M-Cue-CoT. It can be seen that CoT and ProCoT provide non-empathetic, generalized responses that lack informative content, while O-Cue-CoT and M-Cue-CoT acknowledge the victim's emotional state; however, their responses fail to include relevant, actionable, and beneficial information. For instance, in the first sample shown in Table 6, which involves a case of a missing person, the victim expresses frustration over the police's lack of action and shares their growing concern and distress. While baseline models such as CoT suggest generic reassurances like "things will work out eventually", ProCoT redundantly advises filing a missing person report. O-Cue-CoT and M-Cue-CoT provide basic empathy but do not suggest practical next steps. In contrast, **MENDER** generates a superior response by integrating multi-hop commonsense reasoning and domain-specific knowledge. It acknowledges the victim's frustration, recommends filing a complaint with the National Commission for Protection of Child Rights (NCPCR), and provides resources such as the National Mental Health Helpline for psychological support, ensuring the response is both empathetic and informative. Besides, **MENDER** generates helpful and safe responses that are grounded in relevant external knowledge sources.

Though the proposed **MENDER** framework consistently employs empathy, contributing to a more engaging and positive counseling experience, certain limitations are observed in its responses. **MENDER** demonstrates shortcomings in handling complex cases requiring precise mental health and legal knowledge. For example, as shown in third example in Table 6 that highlights a case of unau-

thorized photo sharing on Instagram, with the victim describing the accused's obsessive behavior. The ground-truth response effectively combines empathy, domain-specific legal knowledge, and actionable advice, such as seeking legal or NGO assistance. In contrast, **MENDER**'s response focuses on empathizing with the victim and acknowledging the seriousness of the issue but lacks concrete guidance or references to relevant legal provisions. This limits its ability to empower the victim with actionable next steps, rendering the response insufficient despite its empathetic tone. This might stem from inability to generate proper legal information in this scenario due to the complex dialogue flow and the presence of shorter utterances. These factors limit **MENDER**'s ability to integrate domain-specific knowledge effectively, resulting in responses that emphasize empathy but lack critical legal insights and actionable guidance necessary to address the victim's situation comprehensively.

Dialogue Context	Victim: My daughter is missing since 10 October, 2020. Police is not helping me to find my daughter. Agent: Ok, have you complained about this to any police station? Victim: Yes, I've made the police complaint, but their behavior is reluctant. I don't think they're making any effort to find my daughter. I am too worried about her whereabouts, feeling depressed. Agent: Ok, don't worry. Have they asked you for sufficient details to find her? Victim: Yes, I've provided her photograph along with a detailed description of her appearance on the day she went missing.
Ground-truth Response	Agent: It's sad to hear. I suggest you take legal action in this matter. Are you taking the proper follow up every day ?
CK-CoT Rationales	Query1: What is the victim's concern about the police's behavior in the case of their missing daughter? (xReact) Reply1: The victim is worried about the police's reluctance and lack of effort in finding their daughter, which is causing them distress. Query2: What might be the next logical step for the agent to take in assisting the victim in finding their missing daughter? (xIntent) Reply2: The agent might suggest that the victim provide any additional information or details that could help the police in their investigation, such as the last known location of the daughter or any potential witnesses.
DK-CoT Rationales	Query1: What Indian Penal Code (IPC) sections are relevant to the victim's case? Reply1: Relevant IPC sections include Section 363 (punishment for kidnapping), Section 366 (kidnapping or abducting a woman to compel her marriage), and Section 509 (word, gesture, or act intended to insult the modesty of a woman). Query2: What immediate assistance can the victim seek for their missing daughter? Reply2: The victim can report the case to the local police station or contact helplines such as Childline (1098), National Commission for Protection of Child Rights (NCPCR) or the National Commission for Women (NCW) helpline (181). Query3: How can the victim address their mental health concerns related to their daughter's disappearance? Reply3: The user can contact the National Mental Health Helpline 1800-891-4416 for psychological counseling.
CoT	Ok, just give it some time. Things will work out eventually.
ProCoT	Maybe you should file a missing person report if you haven't already. That should help.
O-Cue-CoT	I see, you're feeling worried and frustrated with the lack of progress in finding your daughter. You should try not to let it affect you too much emotionally, please be strong in this time.
M-Cue-CoT	It's unfortunate that you're feeling this way, but sometimes these things take time. Have you considered reaching out to any other organizations or groups that specialize in missing persons cases?
MENDER	Agent: I'm deeply sorry that you're facing this situation. I understand your concern and frustration with the police's behavior. I would recommend filing a complaint with the National Commission for Protection of Child Rights (NCPCR) and seeking legal advice from a lawyer specializing in criminal law. You can also reach out to the National Mental Health Helpline at 1800-891-4416 for confidential counseling.
Dialogue Context	Victim: I need help. Agent: Yes sure, I am here to help you. May I know to whom I am talking? Victim: My name is Person_X and I am Age_X years old. I am a married woman. Agent: Hi Person_X, please share your problem in detail? Victim: I saw a lady labor at the construction site beside my house. She was being gang-rapXX by her co-workers. I want to help her. She is in mental trauma and is also injured.
Ground-truth Response	Agent: It's really disgusting. Could you please tell me has she or you complained about it?
CK-CoT Rationales	Query1: What is the likely reaction of Person_X's after witnessing the incident? (oReact) Reply1: She might be shocked, frightened, or even guilty, which could lead to a range of reactions, including panic or denial. Query2: What is the potential effect on lady labor's mental state due to the traumatic event she witnessed? (oEffect) Reply2: Lady might experience severe emotional distress, including anxiety and fear as a result of witnessing the violent and disturbing incident. Query3: What might be the next logical step for Person_X to take in order to help the lady laborer? (xNeed) Reply3: She might need to contact the authorities, such as the police or local authorities, to report the incident and ensure the lady laborer receives proper medical attention and psychological support.
DK-CoT Rationales	Query1: What are the relevant IPC sections applicable to the crime described? Reply1: The crime of gang rapX falls under IPC Section 376D (gang rapX) and Section 34 (criminal conspiracy), which address group actions with criminal intent. Query2: What legal rights does the victim have under Indian law? The victim has the right to free medical care under Section 357C of the Code of Criminal Procedure (CrPC) and protection of her identity under Section 228A of IPC. Query3: What helplines or resources can be contacted for immediate support? Reply3: The victim can contact the One Stop Crisis Center (OSCC) or the National Health Helpline (104) for medical and counseling services.
CoT	Ok, you should try talking to the lady and comforting her. Has she complained about this incident?
ProCoT	Maybe you can tell the lady to avoid the site in the future. That might keep her safe.
O-Cue-CoT	I strongly advise against taking any action that could harm the victim or promote illegal activities. You should report this to authorities so that they can take appropriate action.
M-Cue-CoT	I understand you situation. Please seek help and support from trusted organizations and professionals who can provide appropriate assistance and care.
MENDER	Agent: I'm truly sorry to hear about this distressing situation, and I admire your willingness to help. Could you please ensure the lady receives immediate medical and psychological support through One Stop Crisis Center or National Health Helpline - 104. I would like to inform that this is a crime under IPC Section 376D, please let me know if you would like to report the incident
Dialogue Context	Victim: From one and a half months. Agent: OMG! That's quite a long time. Do you have any idea why is he doing this to you? Victim: That bloody rascal says I am his life he will make me the queen of his world. I think he is a psycho. Agent: Oh my God! That's really disturbing. Don't worry, we will help you. Can you tell us on what social media platform he has posted your pic without your consent? Victim: On Instagram. I got the screenshot that he posted my pic.
Ground-truth Response	Agent: You did a great job. I would like to inform you that the person can be charged under IPC Section 354D and Section 67 of IT Act 2008 for stalking and breach of privacy, if found guilty. Would you like to take legal action or consult NGOs/counselors in this matter?
CK-CoT Rationales	Query1: What is the potential effect on the victim's social media presence after the accused posts their picture without consent? (oEffect) Reply1: The victim's picture may go viral on Instagram, attracting unwanted attention and potentially damaging their reputation. Query2: How might the accused behavior be perceived by the victim, considering their statement about being a "psycho"? (xReact) Reply2: The victim feels threatened, disturbed, and possibly fearful due to the accused obsessive behavior and explicit claims. Query3: What does the victim want considering their situation? (xWant) Reply3: The victim might logically want to take action to protect themselves, such as reporting the accused behavior to the social media platform or seeking help from authorities.
DK-CoT Rationales	Query1: What type of crime is being reported in the case of a photo shared without consent? Reply1: This could be considered cyberstalking, harassment, or a violation of privacy under Indian laws.
MENDER	Agent: I can see that you're going through a really tough time. The fact that your picture is trending on Instagram without your consent is not only disturbing but also a serious issue.

Table 6: Example of response generation using MENDER with different knowledge sources.

[Task Overview]

Construct a commonsense knowledge-driven CoT rationale to support the generation of target utterance ('Target:'). The rationale should consist of a sequence of atmost five interconnected query-reply pairs, each incorporating a commonsense relation selected from [xIntent, xNeed, xWant, xReact, oEffect, and Causes]. These queries should provide key inferential steps necessary for generating the target utterance but you should not include the target utterance and also pretend you don't know the target utterance. The last query-reply pair should infer the likely target utterance without explicitly revealing it. If commonsense reasoning is not essential for generating the target utterance, return 'None' as the rationale.

[Few-shot demonstrations]

- Exemplar 1 -

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Commonsense Knowledge-driven CoT Rationale: <Commonsense Knowledge-driven CoT Rationale>

- Exemplar 2 -

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Commonsense Knowledge-driven CoT Rationale: <Commonsense Knowledge-driven CoT Rationale>

Input

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Output

Commonsense Knowledge-driven CoT Rationale: {Commonsense Knowledge-driven CoT Rationale}

Table 7: Prompt template for generating commonsense knowledge-driven CoT. The variables enclosed in curly brackets < ... > represent placeholders that are filled when the template is instantiated. The content within {} is to be generated.

[Task Overview]

Construct a domain knowledge-driven CoT rationale to support the generation of target utterance ('Target:'). The rationale should consist of a sequence of atmost five interconnected query-reply pairs, each incorporating information from the extracted domain knowledge documents. These queries should provide key inferential steps necessary for generating the target utterance but you should not include the target utterance and also pretend you don't know the target utterance. The last query-reply pair should infer the likely target utterance without explicitly revealing it. If domain-specific reasoning is not essential for generating the target utterance, return 'None' as the rationale.

[Few-shot demonstrations]

- Exemplar 1 -

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Domain Knowledge-driven CoT Rationale: <Domain Knowledge-driven CoT Rationale>

- Exemplar 2 -

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Domain Knowledge-driven CoT Rationale: <Domain Knowledge-driven CoT Rationale>

Input

Dialogue Context: <Dialogue Context>

Target: <Ground-truth Response>

Output

Domain Knowledge-driven CoT Rationale: {Domain Knowledge-driven CoT Rationale}

Table 8: Prompt template for domain knowledge-driven CoT. The variables enclosed in curly brackets < ... > represent placeholders that are filled when the template is instantiated. The content within {} is to be generated.

[Task Overview] Generate the most appropriate next response based on the dialogue context. While the rationale may be referenced, it should be disregarded if it leads to an inaccurate response. Ensure conciseness by avoiding excessive information, and maintain consistency with the style of the preceding dialogue.

[Few-shot demonstrations]

- Exemplar 1 -

Dialogue Context: <Dialogue Context>

Commonsense Knowledge-driven CoT Rationale: <Commonsense Knowledge-driven CoT Rationale>

Domain Knowledge-driven CoT Rationale: <Domain Knowledge-driven CoT Rationale>

Response: <Ground-truth Response>

- Exemplar 2 -

Dialogue Context: <Dialogue Context>

Commonsense Knowledge-driven CoT Rationale: <Commonsense Knowledge-driven CoT Rationale>

Domain Knowledge-driven CoT Rationale: <Domain Knowledge-driven CoT Rationale>

Response: <Ground-truth Response>

Input

Dialogue Context: <Dialogue Context>

Commonsense Knowledge-driven CoT Rationale: <Commonsense Knowledge-driven CoT Rationale>

Domain Knowledge-driven CoT Rationale: <Domain Knowledge-driven CoT Rationale>

Output

Response: {Response}

Table 9: Prompt template for response generation. The variables enclosed in curly brackets < ... > represent placeholders that are filled when the template is instantiated. The content within { } is to be generated.