A Mental State Extraction Dataset for Theory-of-Mind-based Reasoning in Emotional Support Conversations

Seulgi Kim, Harksoo Kim
Konkuk University, Seoul, Republic of Korea
{cloudyju11, nlpdrkim}@konkuk.ac.kr

Abstract

Emotional Support Conversations (ESC) aim to both reduce users' emotional distress and facilitate problem-solving. Recent approaches in ESC have explored incorporating commonsense knowledge into large language models (LLMs) to improve response generation. However, existing commonsense reasoning models often rely solely on the final utterance, fail to anticipate future turns, overlook emotional cues, or treat knowledge types independently, resulting in incoherent or emotionally misaligned responses. To address these limitations, we propose an approach grounded in Theory of Mind (ToM). Specifically, we introduce MENTOS, a dataset that provides turn-level annotations of the assistant's mental states (Belief, Emotion, and Intent), organized in a causal structure reflecting psychological principles. A commonsense reasoning model trained on MENTOS predicts these mental states as intermediate reasoning signals that guide response generation. Experiments on the ESConv and ExTES datasets show that incorporating the inferred mental states can enhance supportive and goal-directed response generation across multiple reasoning backbones and response generators. Ablation studies further confirm that Belief, Emotion, and Intent provide complementary benefits for ESC tasks. These findings highlight the effectiveness of ToM-grounded intermediate reasoning in generating empathetic and contextually appropriate responses.

Keywords: Conversational Systems, Knowledge Discovery/Representation, Emotion Recognition/Generation

1. Introduction

Task-oriented dialogue systems have traditionally been developed to support users in goal completion through conversational interaction (Chen et al., 2017; Ni et al., 2023). Among these, Emotional Support Conversations (ESC) have gained prominence, as they can provide both practical guidance for problem solving and emotional relief to users (Heaney and Israel, 2008; Liu et al., 2021). The importance of automated ESC systems is further underscored by the fact that nearly one billion people worldwide face mental health challenges while access to professional care is limited by therapist shortages, stigma, and other barriers (Wu et al., 2023a; Qiu and Lan, 2025).

With large language models (LLMs) recently demonstrating strong zero-shot and few-shot performance in natural language generation (NLG) (Touvron et al., 2023; Dubey et al., 2024; OpenAI et al., 2024), LLM-based approaches for ESC have focused on improving the contextual appropriateness of responses. Among these, one approach involves injecting commonsense knowledge from reasoning models into LLM-based response generators, using it as intermediate reasoning signals to generate more supportive and contextually grounded responses (Chae et al., 2023). However, such commonsense reasoning models still face several limitations, including overreliance on single-turn context, lack of future-aware reasoning, failure to capture users' emotional reactions, and fragmented reasoning due to independently generated knowledge types (Li et al., 2024; Wang et al., 2025). These reasoning models often generate irrelevant or inaccurate commonsense knowledge, which can interfere with producing emotionally supportive responses.

To address these challenges, we propose a new framework grounded in psychological research, emphasizing the need to model the reasoning processes that underlie supportive communication (Cologon et al., 2017; Lüdemann et al., 2021). Building on this perspective, we draw on the concept of Theory of Mind (ToM), which refers to the ability to understand both one's own and others' mental states (Wellman et al., 1990). To endow commonsense reasoning models with the ToM capability needed for effective emotional support, we introduce **MENTOS**, a **M**ental state **E**xtractio**n** dataset for **To**M-based rea**s**oning.[1] It provides turn-level annotations of the assistant's mental states (Belief, Emotion, and Intent) in multi-turn dialogues, organized in a causal sequence informed by psychological studies (Beck et al., 2005; Thagard and Stewart, 2011; Gandhi et al., 2023). Each mental state captures the assistant's understanding of the client's situation, as well as the assistant's own emotional reactions and intended actions. We then train a commonsense reasoning model on MENTOS to predict these mental states. The predicted states are subsequently injected into an LLM-based response generator alongside the dialogue history, guiding it to produce follow-up responses that are

[1] The MENTOS dataset and implementation code are available at https://github.com/KUNLP/MENTOS.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2705–2723
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

both emotionally grounded and oriented toward problem solving.

To evaluate the effectiveness of MENTOS-trained models in ESC tasks, we address the following research questions (RQs):

RQ1: Does a commonsense reasoning model trained on MENTOS improve the contextual appropriateness and emotional supportiveness of generated responses? To answer this question, we compare responses generated using commonsense knowledge from existing reasoning models and the MENTOS-trained model, focusing on how each influences response quality as evaluated by automatic metrics and G-Eval (Section 5.2).

RQ2: Do MENTOS-trained models consistently enhance response quality for ESC across variations in commonsense reasoning backbones and response generators? We compare their performance using different reasoning backbones and multiple LLMs as response generators, suggesting that the effectiveness of MENTOS is not tied to a specific model architecture (Section 5.3).

RQ3: Do the three mental states provide complementary benefits for representing dialogue context in ESC? We verify this through ablation studies that remove each mental state and compare response quality, analyzing the roles of Belief, Emotion, and Intent (Section 5.4).

Our major contributions are as follows:

- We introduce MENTOS, a dataset annotated with the causal structure of the assistant's mental states, grounded in ToM.
- We train a commonsense reasoning model on MENTOS to infer mental states that guide response generation toward greater emotional grounding and contextual appropriateness.
- Experiments demonstrate the effectiveness of MENTOS and the complementary roles of mental states.

2. Related Work

2.1. Commonsense Reasoning for Response Generation in Dialogue

Commonsense knowledge is frequently framed less as a fixed set of defining features and more as implicit human understanding of basic facts, everyday situations, and social and emotional cues that help people reason about others' likely mental states and actions (Liu and Singh, 2004; Sap et al., 2019). Under this broad usage, commonsense knowledge in interactive settings often involves inferring latent factors such as causes, motivations, emotions, and intents to explain and predict everyday behavior. Several studies leverage such commonsense knowledge as intermediate reasoning signals to generate contextually appropriate responses in dialogue. A representative commonsense reasoning model, *COMET* (Bosselut et al., 2019), has been widely applied to various NLG tasks, including dialogue systems (Li et al., 2024; Xu et al., 2024b,c). However, its commonsense reasoning in ESC often relies primarily on the last utterance, failing to incorporate broader dialogue context.

To address this limitation, dialogue-specific models such as *DIALeCT* (Shen et al., 2022) and *DOCTOR* (Chae et al., 2023) take multi-turn dialogues as input to enable more faithful commonsense reasoning. Nonetheless, *DIALeCT*, which is trained on complete and static dialogue histories, may struggle to anticipate the future course of a dialogue (Wang et al., 2025). Because a single dialogue history can support multiple appropriate responses, its commonsense reasoning is prone to contextual misalignment (Liu et al., 2022). *DOCTOR* performs multi-hop reasoning by selecting one commonsense knowledge type at each hop. Because it does not explicitly control these knowledge types (Finch and Choi, 2025), the resulting sequence may omit knowledge relevant to the user's emotions and weaken emotion inference. *Sibyl* (Wang et al., 2025) further generates multiple types of commonsense knowledge, but models each type independently, which can lead to incoherent reasoning. These challenges highlight the need for a commonsense reasoning model trained on a dataset that organizes diverse knowledge types into causal sequences for each turn-level assistant utterance in ESC.

2.2. From ToM to Practical Modeling

ToM is the cognitive capability to infer and understand one's own and others' mental states (e.g., Belief, Emotion, Desire, and Intent) and to use these states to interpret and predict actions (Wellman et al., 1990). This ability is especially important in ESC settings and counseling scenarios between the client and the assistant, where the assistant needs to track both their own and the client's mental states to provide effective emotional support and achieve therapeutic outcomes (Cologon et al., 2017; Lüdemann et al., 2021).

This importance is further reflected in the causal structure among mental states: Belief influences Emotion, and together with Desire, these states shape Intent, which in turn guides the selection of appropriate supportive strategies. Psychological research substantiates this structure, showing that Emotion arises from Belief about events (Beck et al., 2005; Xu et al., 2024a) and that cognitive appraisal combined with physiological states drives Intent (Thagard and Stewart, 2011; Thagard and

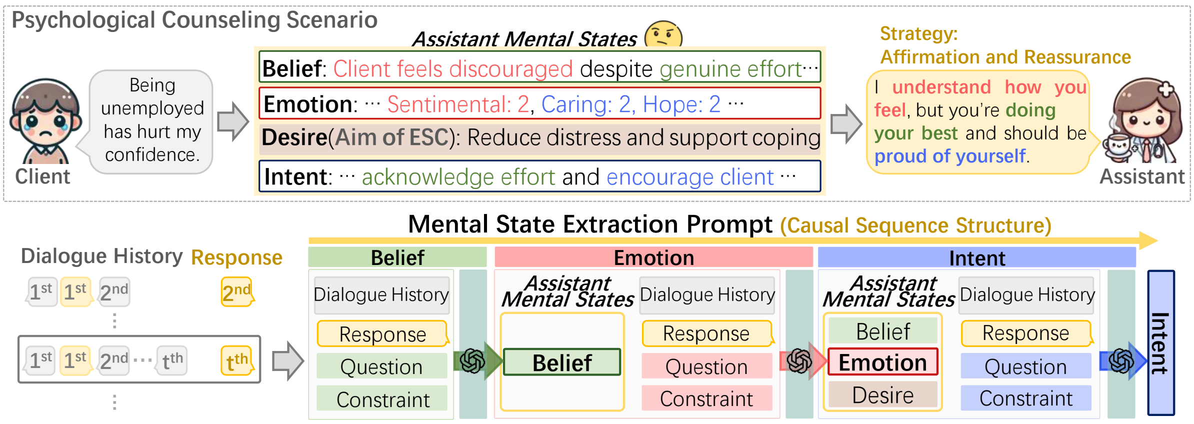


Figure 1: MENTOS construction where each mental state is generated based on the preceding one.

Schröder, 2014; Bagozzi et al., 2016). However, many existing approaches omit critical mental state components such as Emotion or Intent. For example, studies that exclude Emotion (Gandhi et al., 2023; Frering et al., 2025) often fail to capture nuanced affective cues, a limitation further exacerbated by LLMs' difficulties in emotional understanding and application (Sabour et al., 2024). Likewise, omitting Intent (Xu et al., 2024a) can undermine the assistant's supportive strategy, which is important given the close relationship between Intent and strategy (Cao et al., 2025).

Despite the importance of ToM capabilities for effective emotional support, recent evaluations indicate that LLMs still exhibit limited ToM reasoning (van Duijn et al., 2023), thereby constraining their ability to generate emotionally attuned responses in their role as assistants in ESC. Accordingly, we provide LLMs with explicit representations of assistants' mental states within multi-turn dialogue contexts as intermediate reasoning signals to compensate for their lack of inherent ToM capabilities.

3. MENTOS Construction

We present **MENTOS**, a commonsense knowledge resource derived from the ESConv dataset, which consists of 1,300 dialogues between emotionally distressed clients and supportive assistants based on Helping Skills Theory (Hill, 1999). It explicitly annotates each assistant utterance with ToM-grounded mental states, inferred from the dialogue history up to the client's last utterance and preceding the assistant's response. This allows tracing how the assistant's mental states guide empathic interaction (Wu et al., 2023b).

MENTOS adopts a turn-level annotation scheme to address the limitations of prior commonsense reasoning approaches that rely on static and completed dialogue histories. In this scheme, each annotation is generated by conditioning on both the dialogue history up to the client's utterance at the t-th turn and the corresponding assistant response, ensuring alignment with conversational interaction and enabling future-aware reasoning (Wang et al., 2025). Each annotation consists of three mental states (i.e., Belief, Emotion, and Intent) as target knowledge types, organized in a causal sequence that reflects their interdependencies. This design enables MENTOS to capture emotional cues and causal relationships between knowledge types that are often overlooked by prior reasoning models. Table 1 summarizes MENTOS dataset statistics.

	# Dialogues	# Annotated turns	#B	#E	#I
Train	909	8,959	34.05	22	33.63
Dev	195	1,907	34.10	22	33.55
Test	196	2,051	34.05	22	33.53

Table 1: Statistics of the MENTOS dataset. #B and #I denote the average number of words in the Belief and Intent annotations, respectively, and #E represents the number of emotion categories.

Annotation begins from the second turn of each dialogue, since the first turn is always a greeting. As illustrated in Figure 1, we provide GPT-4o with a prompt composed of five input components: (1) the dialogue history, (2) the assistant's response at the t-th turn, (3) the *Assistant Mental States* component, (4) the *Question* component, and (5) the *Constraint* component. The temperature is set to 0 to ensure reproducibility, and role-playing prompts are used to simulate the assistant's reasoning process (Tao et al., 2024; Ye et al., 2025; Wu et al., 2025). Each data sample in MENTOS consists of the dialogue history, the assistant's response, and the corresponding turn-level annotation.

Belief is **the initial mental state** in the causal sequence, representing what the assistant infers about the client's situation and condition from the dialogue history (Perner and Wimmer, 1985; Baron-Cohen et al., 1985). Accordingly, the ***Assistant***

***Mental States* component is not included in the prompt** at this stage, as shown in Figure 1. The *Constraint* component instructs the model to produce the Belief as a single, concise sentence (maximum 40 words).

Emotion reflects the assistant's emotional reaction in their response (Reisenzein, 2009), shaped by the dialogue context and the **preceding Belief**, which **is included in the *Assistant Mental States* component**. Unlike Belief or Intent, Emotion is annotated quantitatively rather than textually. Based on Plutchik's wheel of emotions (Plutchik, 1982), we adopt a 0–3 scale for eight basic and fourteen mixed emotions, indicating increasing levels of affective intensity (e.g., none, relief, joy, and ecstasy along the continuum of joy). Mixed emotions, which are combinations of basic emotions (e.g., Hopelessness combines Sadness and Fear) (Sabour et al., 2024), are included only when each constituent basic emotion score is ≥ 2 or explicitly mentioned. This ensures a fine-grained representation of emotional intensity and complexity. The *Constraint* component specifies a structured format for assigning scores to each emotion category, for example: *[Basic] Sadness (opposite Joy): <score>, Disgust (opposite Trust): <score>, ... [Mixed] Hopelessness (sadness + fear): <score>, ...*

Intent captures the strategic purpose and action underlying the assistant's last utterance, inferred from **preceding Belief, Emotion, and Desire** (Baker et al., 2011; Frering et al., 2025). Desire, which reflects the assistant's motivational orientation (Neuenschwander et al., 2018), is represented as a **fixed text** rather than dynamically inferred, since in ESC, Desire consistently aims to reduce the client's emotional distress and support coping with challenges throughout the dialogue (Liu et al., 2021). It is combined with Belief and Emotion in the *Assistant Mental States* component, thereby structuring the causal sequence. Similar to Belief, the *Constraint* component requires the model to generate a clear and concise sentence. The complete prompt design is provided in Appendix A.

To evaluate annotation quality, we randomly sampled 100 dialogues, corresponding to 1,032 turn-level annotation instances, which were assessed by four annotators. As shown in Table 2, each mental state was evaluated on four criteria using a 1–3 scale derived from ESC and psychological research (Swettenham, 1996; Ghosal et al., 2022; Chae et al., 2023; Wang et al., 2025; Frering et al., 2025; Cao et al., 2025), where 1 indicates the criterion is not met, 2 indicates partial fulfillment, and 3 indicates full alignment. Overall, the average score across all criteria was 2.86, suggesting that the annotations were consistently of high quality. To assess inter-annotator reliability, we adopted Gwet's AC1 (Gwet, 2002), as it is robust under skewed annotation distributions. This is relevant to our setting because annotators predominantly assigned high scores, resulting in a positively skewed label distribution. Under such conditions, other agreement metrics such as Fleiss' Kappa are affected by prevalence bias and can misleadingly report low agreement even when annotators agree at a high rate (Jeni et al., 2013; Wongpakaran et al., 2013).[2] The mean Gwet's AC1 was 0.73, indicating substantial agreement (0.60–0.80). These results demonstrate that MENTOS provides reliable and high-quality mental state annotations (Belief, Emotion, and Intent) that explicitly capture causal dependencies. Detailed annotator instructions are provided in Appendix C.

	Evaluation Criteria	Score	Gwet's AC1
Belief	Contextual Appropriateness	2.87	0.76
	Grounding in Client's State	2.89	0.80
	Relevance to Assistant's Response	2.83	0.69
	Depth of Inference	2.82	0.66
Emotion	Emotion Type & Intensity Fit	2.84	0.67
	Mixed Emotion Plausibility	2.84	0.67
	Consistency with Preceding Belief	2.85	0.68
	Affective Congruence with Response	2.85	0.70
Intent	Relevance to Client's Needs	2.87	0.76
	Consistency with Emotional Tone	2.89	0.81
	Consistency with Support Strategy	2.87	0.75
	Response Interpretability	2.87	0.78

Table 2: Human evaluation results on a 1–3 scale.

4. Commonsense Reasoning Model

We train a commonsense reasoning model on the MENTOS training set via supervised fine-tuning (SFT), which enables it to infer three interdependent mental states: **Belief**, **Emotion**, and **Intent**. Unlike in the dataset construction process described in Section 3, where the assistant's response was explicitly included to generate semantically aligned annotations, the commonsense reasoning model is trained during SFT to predict target mental states accurately without the response as input. Accordingly, the reasoning model takes the following prompt components as input: (1) the dialogue history, (2) the *Assistant Mental States* component containing the preceding inferred states (e.g., Belief for Emotion; Belief, Emotion, and fixed Desire for Intent), (3) the *Constraint* component for each mental state described in Section 3, and (4) the *Question* component as shown below:

- **Belief:** What does the assistant believe about the client's situation or internal state based on the client's last utterance?

[2]We observed a mean Fleiss' Kappa of 0.02 across all evaluation criteria, likely due to prevalence bias.

	Belief	Emotion	Intent
Llama3.1 (8B)	67.19	60.40	61.99
Llama2 (7B)	30.73	24.44	22.31
Qwen3 (1.7B)	37.92	35.36	34.71

Table 3: Accuracy (%) of LLMs on ToMBench.

- **Emotion:** What emotional reaction might the assistant have after the client's last utterance, based on the assistant's belief and how the conversation has unfolded?
- **Intent:** What is the assistant's intent following the client's last utterance, based on the assistant's belief, emotional reaction, and desire?

The commonsense reasoning model trained on MENTOS sequentially generates three mental states during inference, guiding an LLM-based response generator to produce emotionally relevant and strategically coherent responses. Detailed SFT settings are provided in Appendix B.

5. Experiments

5.1. Experimental Settings

Baselines. We evaluated emotionally supportive response generation using the same response generator under two settings: (1) dialogue history only and (2) dialogue history combined with commonsense knowledge from reasoning models. In all comparisons, the dialogue-only setting and the variants leveraging prior reasoning models (COMET, DIALeCT, DOCTOR, and Sibyl) were treated as baselines. Sibyl, which is not publicly available, was reproduced using its official codebase and fine-tuned on the same backbones as our models, while the others used publicly available checkpoints. To ensure clarity and consistency for the response generator, the outputs from all reasoning models were reconstructed into full-sentence textual forms before being input to the response generator.
Backbones and Generators. We investigated whether commonsense reasoning models trained on MENTOS consistently improved response appropriateness across multiple reasoning-model backbones and response generators. Specifically, we fine-tuned Llama3.1, Llama2, and Qwen3 on the MENTOS training set, resulting in three MENTOS-trained reasoning models, namely $\text{MENTOS}_{\text{Llama3.1}}$, $\text{MENTOS}_{\text{Llama2}}$, and $\text{MENTOS}_{\text{Qwen3}}$, respectively.[3] We adopted Low-Rank Adaptation (LoRA) (Hu et al., 2021) with rank 8, alpha 16, and dropout 0.05, applied to the Q and V projection matrices. This tuning affected only 0.042%, 0.062%, and 0.093% of the parameters for each backbone, respectively. Optimization followed Adam (Kingma and Ba, 2015) with a learning rate of 3e-5, batch size of 4, and 5 epochs, selecting the best checkpoint based on validation performance. We also used Llama2 and Qwen3 in a zero-shot setting as response generators, denoted as $\text{Generator}_{\text{Llama2}}$ and $\text{Generator}_{\text{Qwen3}}$, whose accuracy on ToMBench remained below 40% across all mental states, as shown in Table 3. This setup ensured that any performance gains were attributable to the supplied mental states rather than the generators' inherent reasoning abilities.
Dataset. We evaluated generated responses on two datasets: (1) the ESConv test set, which served as the basis for the MENTOS test set, containing 2,051 turn-level samples in 196 dialogues; and (2) the test set of the *ExTensible Emotional Support* dialogue dataset (ExTES) (Zheng et al., 2024), comprising 11,178 turn-level samples in 430 dialogues. ExTES covered multiple counseling scenarios, with each dialogue mapped to a specific scenario. For scenarios with ten or fewer dialogues, all were included in the test set, while for scenarios with more than ten, ten dialogues were randomly sampled per scenario.
Metric. To assess response quality, we employed both automatic metrics and LLM-based judgments. BLEU-4 (B-4) (Papineni et al., 2002) and METEOR (MET) (Lavie and Agarwal, 2007) measured linguistic fidelity. Embedding-based semantic similarity was measured using Greedy Matching (Greedy) (Rus and Lintean, 2012) and word- and sentence-level cosine similarity (C_W, C_S) (Mitchell and Lapata, 2008). Detailed definitions of these three metrics are provided in Appendix D. Response diversity was measured with Distinct-3 (Dist-3) (Li et al., 2016), and all metric scores were scaled by 100 for readability. For the LLM-based evaluation, we used G-Eval (Liu et al., 2023) with GPT-4o to rate responses from three perspectives: Supportiveness (Sup.), Naturalness (Nat.), and Coherence (Coh.). We randomly sampled 200 instances from the ESConv test set and evaluated them on a 3-point scale: 1 (failure to meet), 2 (partially adequate), and 3 (fully aligned). The G-Eval prompt and rating rubric are provided in Appendix E.

5.2. RQ1: Effectiveness of MENTOS

To examine whether the MENTOS-trained commonsense reasoning model improves the suitability of responses in ESC, we used the mental states inferred by the model to guide response generation. Tables 4 and 5 present the results on both the ESConv and ExTES test sets using $\text{Generator}_{\text{Llama2}}$ and $\text{Generator}_{\text{Qwen3}}$, respectively. Overall, mod-

[3]The backbones are `meta-llama/Llama-3.1-8B-Instruct`, `meta-llama/Llama-2-7b-chat-hf`, and `Qwen3/Qwen3-1.7B` from HuggingFace.

	ESConv						ExTES					
	Fidelity		Diversity	Similarity			Fidelity		Diversity	Similarity		
	B-4	MET	Dist-3	C_W	C_S	Greedy	B-4	MET	Dist-3	C_W	C_S	Greedy
w/o Knowledge	0.40	8.38	30.09	88.58	27.38	69.87	2.05	10.48	16.23	93.86	48.40	74.56
+COMET	0.51	8.43	37.78	88.65	28.39	69.95	2.37	10.95	19.80	94.21	49.75	75.35
+DIALeCT	**0.57**	8.64	46.21	88.89	30.52	70.46	2.35	10.86	22.85	94.01	50.25	75.38
+DOCTOR	0.53	8.33	41.25	88.17	27.33	69.91	2.22	10.52	23.11	93.62	46.91	75.01
+Sibyl$_{\text{Llama3.1}}$	0.45	8.17	48.38	87.83	29.19	69.12	2.42	11.00	**23.20**	93.98	51.09	75.09
+Sibyl$_{\text{Llama2}}$	0.57	8.33	**50.37**	88.10	29.98	69.43	2.32	11.02	22.95	94.05	50.25	75.06
+MENTOS$_{\text{Llama3.1}}$	0.49	**9.17**	46.46	89.26	**31.05**	70.76	2.88	12.21	22.60	94.66	52.42	76.11
+MENTOS$_{\text{Llama2}}$	0.51	9.14	46.34	**89.36**	30.76	**70.91**	2.96	**12.23**	21.79	**94.76**	**53.01**	76.25
+MENTOS$_{\text{Qwen3}}$	0.43	8.99	45.67	89.22	30.79	70.69	**2.96**	**12.23**	21.74	94.68	52.85	**76.26**

Table 4: Evaluation results using Generator$_{\text{Llama2}}$ on the test sets. Bold indicates the best performance.

	ESConv						ExTES					
	Fidelity		Diversity	Similarity			Fidelity		Diversity	Similarity		
	B-4	MET	Dist-3	C_W	C_S	Greedy	B-4	MET	Dist-3	C_W	C_S	Greedy
w/o Knowledge	0.22	8.29	40.27	89.09	26.99	70.01	2.08	11.67	19.44	94.10	46.57	75.27
+COMET	0.36	8.30	39.42	89.33	27.05	70.67	1.98	11.51	18.33	94.49	45.73	75.59
+DIALeCT	**0.48**	7.55	**55.22**	87.11	27.17	68.24	1.46	10.18	**25.45**	92.66	43.01	73.63
+DOCTOR	0.33	7.16	50.54	85.97	23.32	66.93	1.56	10.14	23.75	92.90	42.55	73.40
+Sibyl$_{\text{Llama3.1}}$	0.39	7.84	52.85	87.42	27.72	68.27	2.14	11.31	24.56	93.78	47.08	74.33
+Sibyl$_{\text{Llama2}}$	0.30	7.91	52.04	87.45	27.65	68.34	1.85	11.29	24.68	93.80	46.97	74.23
+MENTOS$_{\text{Llama3.1}}$	0.23	9.04	46.39	**89.59**	29.52	**70.83**	2.05	13.14	21.58	94.84	**50.22**	**76.11**
+MENTOS$_{\text{Llama2}}$	0.24	**9.06**	47.35	89.39	**29.68**	70.66	2.19	**13.20**	21.13	**95.05**	50.15	76.10
+MENTOS$_{\text{Qwen3}}$	0.25	8.98	47.77	89.39	29.56	70.62	**2.25**	12.92	21.99	94.69	49.44	75.77

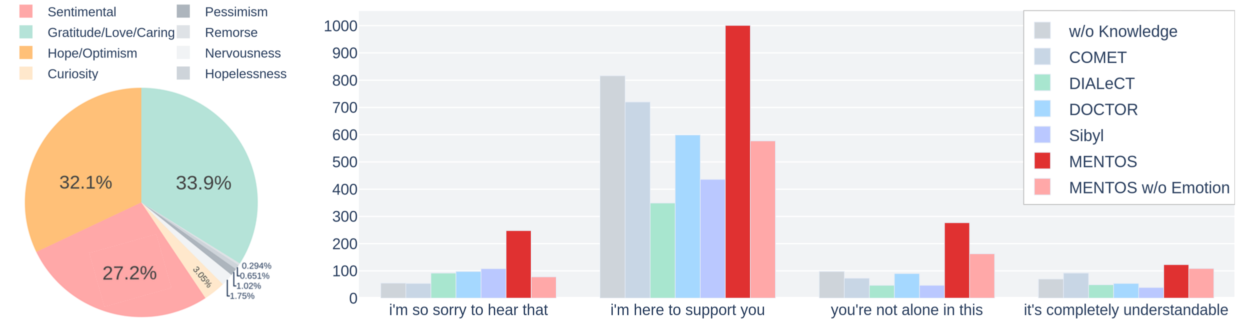
Table 5: Evaluation results using Generator$_{\text{Qwen3}}$ on the test sets. Bold indicates the best performance.

eling the assistant's reasoning process through the MENTOS-trained model yielded more emotionally appropriate and goal-directed responses, as indicated by consistently higher METEOR, Greedy, C_W, and C_S scores compared to all baselines.

Fidelity. The MENTOS-trained model consistently outperformed all baselines in BLEU-4 and METEOR on ExTES. Although Generator$_{\text{Qwen3}}$ under ESConv achieved a BLEU-4 score comparable to the dialogue-only setting (w/o Knowledge), it surpassed all baselines in METEOR, indicating stronger synonym-level semantic alignment. A similar trend was observed for Generator$_{\text{Llama2}}$, as shown in Table 4, whose METEOR scores on ESConv exceeded all baselines. These results suggest that the MENTOS-trained model enhanced meaning preservation, which is more crucial than surface-level similarity in ESC, thus demonstrating clear advantages in fidelity.

Embedding-based Semantic Similarity. For both Generator$_{\text{Llama2}}$ and Generator$_{\text{Qwen3}}$, using DOCTOR resulted in lower C_W and C_S scores than the w/o Knowledge setting. This decline likely occurred because its multi-hop reasoning failed to capture emotional cues effectively. In particular, the xReact knowledge type, which represents the client's emotional reactions, appeared in only about 20% of ESConv and 13% of ExTES test samples. COMET, despite generating xReact in 95% of cases, underperformed relative to the other baselines because it relied solely on the client's last utterance, limiting contextual reasoning. When paired with Generator$_{\text{Qwen3}}$, COMET achieved a C_S on ESConv nearly identical to that of the w/o Knowledge setting, with a marginal difference of 0.06%p, but exhibited lower performance on ExTES. DIALeCT also struggled to anticipate future dialogue turns because it was trained on complete and static dialogue histories, resulting in consistently lower performance than the MENTOS-trained models. Notably, under ExTES with Generator$_{\text{Qwen3}}$, the MENTOS-trained models across different backbones achieved maximum gains of 2.39%p in C_W, 7.67%p in C_S, and 2.71%p in Greedy compared to DIALeCT. While Sibyl achieved higher C_S but lower C_W and Greedy, the MENTOS-trained models demonstrated overall superior performance across test sets, showing consistent gains in most semantic similarity metrics.

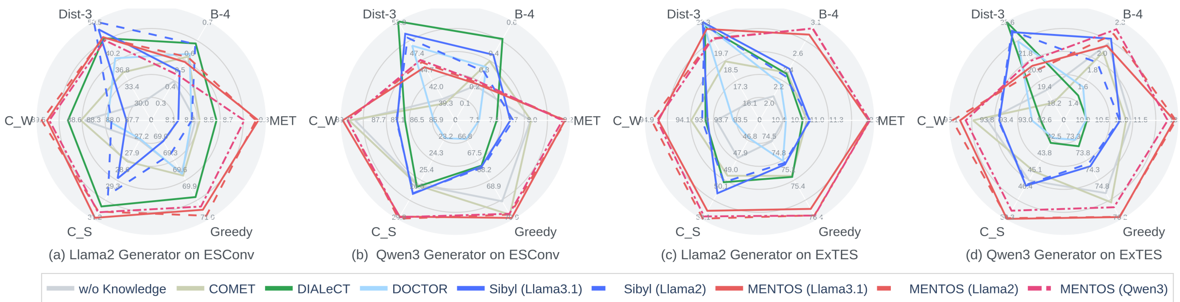
Diversity. The MENTOS-trained model exhibited lower Dist-3 scores than the baselines, likely due to repetitive comforting and supportive expressions generated when Emotion was incorporated. As shown in Figure 2(a), the distribution of top-scoring



(a) Counts of top-scoring mixed emotions inferred from MENTOS$_{\text{Llama3.1}}$

(b) Counts of phrase patterns in responses generated by Generator$_{\text{Llama2}}$

Figure 2: Counts of top-scoring mixed emotions and phrase patterns of responses in the ExTES test set.



(a) Llama2 Generator on ESConv

(b) Qwen3 Generator on ESConv

(c) Llama2 Generator on ExTES

(d) Qwen3 Generator on ExTES

Figure 3: Relative performance comparison across reasoning backbones and response generators.

	Sup.	Nat.	Coh.	Avg.
w/o Knowledge	2.40	2.68	2.66	2.58
+COMET	2.39	**2.72**	2.73	2.61
+DIALeCT	2.23	2.68	2.77	2.56
+DOCTOR	2.27	2.61	2.60	2.49
+Sibyl$_{\text{Llama3.1}}$	2.35	2.58	**2.82**	2.58
+MENTOS$_{\text{Llama3.1}}$	**2.49**	2.69	2.81	**2.66**

Table 6: G-Eval results using Generator$_{\text{Llama2}}$ on the ESConv test set.

mixed emotions per sample revealed that the three most frequent emotions were Sentimental, Gratitude/Love/Caring, and Hope/Optimism, each playing a central role in conveying empathy and reassurance to emotionally unstable clients. Figure 2(b) further illustrates that incorporating Emotion increased the frequency of recurrent phrase patterns such as "I'm so sorry to hear that," "I'm here to support you," and "You're not alone in this." This pattern decreased to near-baseline levels when Emotion was excluded. In addition, Belief captured the client's situation and emotional state, often leading to recurring expressions like "It's completely understandable." As shown in Table 6, MENTOS$_{\text{Llama3.1}}$ achieved the highest Supportiveness score of 2.49 and the highest overall average of 2.66 when compared against all baselines. Overall, the lower Dist-3 reflects a reasonable trade-off inherent to generating empathetic and comforting responses.

5.3. RQ2: Consistency Across Models

We further investigated whether response enhancements driven by MENTOS-trained models were consistent and robust across different commonsense reasoning backbones and response generators. To ensure fair and representative comparisons, we used DIALeCT and Sibyl as primary baselines, both of which demonstrated strong performance in Section 5.2. DIALeCT was used for generator-level comparisons, while Sibyl, sharing identical backbones with the MENTOS-trained models, was used for backbone-level comparisons. To provide a more intuitive comparison, Figure 3 complements Tables 4 and 5 by visualizing model performance across automatic evaluation metrics, where greater distance from the center indicates higher performance. Green, blue, and red lines represent DIALeCT, Sibyl, and MENTOS-trained models, respectively. Overall, the MENTOS-trained models exhibited consistently superior results across most metrics, particularly METEOR, Greedy, C_W, and C_S, demonstrating robustness across model architectures.

Response Generators. Compared to DIALeCT, the MENTOS-trained models exhibited more generalizable and stable performance across datasets and generators we tested. While DIALeCT

	ESConv						ExTES					
	Fidelity		Diversity	Similarity			Fidelity		Diversity	Similarity		
	B-4	MET	Dist-3	C_W	C_S	Greedy	B-4	MET	Dist-3	C_W	C_S	Greedy
+MENTOS$_{\text{Llama3.1}}$	**0.49**	**9.17**	46.46	89.26	**31.05**	70.76	**2.88**	**12.21**	22.60	**94.66**	52.42	76.11
w/o Belief	0.44	9.08	38.71	**89.42**	28.98	**70.95**	2.86	11.60	18.08	94.65	47.95	**76.38**
w/o Emotion	0.44	8.94	**49.20**	88.94	31.04	70.34	2.85	12.11	**25.87**	94.33	**53.04**	75.45
w/o Intent	0.48	8.84	42.79	89.06	30.34	70.38	2.58	11.49	20.09	94.47	50.83	75.82
+MENTOS$_{\text{Llama2}}$	**0.51**	**9.14**	46.34	**89.36**	**30.76**	**70.91**	**2.96**	12.23	21.79	**94.76**	**53.01**	76.25
w/o Belief	0.48	8.99	40.72	89.35	29.31	70.85	2.69	11.33	18.22	94.56	48.53	76.05
w/o Emotion	0.45	9.03	**48.84**	89.11	30.59	70.56	2.82	**12.27**	**24.67**	94.47	52.99	75.80
w/o Intent	0.48	8.75	41.72	89.04	30.12	70.32	2.41	11.47	18.85	94.47	50.67	75.71
+MENTOS$_{\text{Qwen3}}$	0.43	**8.99**	45.67	**89.22**	**30.79**	**70.69**	**2.96**	**12.23**	21.74	**94.68**	**52.85**	**76.26**
w/o Belief	0.43	8.73	41.93	89.02	29.05	70.40	2.59	11.17	19.47	94.42	48.44	75.76
w/o Emotion	0.43	8.89	**48.87**	88.79	30.40	70.19	2.69	11.44	**24.47**	93.81	50.73	75.08
w/o Intent	**0.47**	8.78	42.29	89.00	30.43	70.30	2.55	11.71	19.71	94.50	51.68	75.81

Table 7: Ablation results for removing each mental state. Bold indicates the best performance. All assistant responses were generated using Llama2 under a zero-shot setting.

achieved the best baseline in ESConv under Generator$_{\text{Llama2}}$, it still fell short of the MENTOS-trained models in most metrics and showed noticeable fluctuations across response generators. For instance, under Generator$_{\text{Qwen3}}$, its performance dropped to the second-lowest in METEOR, C_W and Greedy. This instability became more evident in ExTES, which contains more diverse scenarios with finer-grained supportive strategies than ESConv. Under Generator$_{\text{Llama2}}$, DIALeCT still outperformed w/o Knowledge and DOCTOR but fell short of COMET and Sibyl, as shown in Figure 3(c), indicating that its effectiveness diminished outside the ESConv setting. Similarly, its C_S further declined by up to 3.56%p compared to w/o Knowledge under Generator$_{\text{Qwen3}}$. In contrast, the MENTOS-trained models consistently delivered robust performance across all generators, maintaining superior embedding-based semantic similarity metrics (C_W, C_S, Greedy) and METEOR scores on both ESConv and ExTES. These results demonstrate that the MENTOS-trained models generalize effectively across different generators and datasets.

Reasoning Model Backbones. The MENTOS-trained models also demonstrated consistent advantages across reasoning backbones. For instance, under Generator$_{\text{Qwen3}}$, MENTOS$_{\text{Llama2}}$ outperformed Sibyl$_{\text{Llama2}}$ on ESConv by 1.94%p, 2.03%p, and 2.32%p in C_W, C_S, and Greedy, respectively, and on ExTES by 1.25%p, 3.18%p, and 1.87%p. Similar trends were observed with Llama3.1 backbones. This improvement can be attributed to Sibyl generating independent knowledge types (e.g., Cause, Intent), which often led to disconnected reasoning chains. In contrast, MENTOS explicitly modeled causal dependencies among Belief, Emotion, and Intent, yielding coherent reasoning paths and more consistent knowledge, as further supported by the human evaluation results in Table 2. As shown in Figure 3, the MENTOS-trained models achieved the best METEOR, C_W, C_S, and Greedy scores across reasoning backbones compared to the baselines. In particular, the variation across backbones remained very small on ESConv (approximately within 0.1%p), and similar stability was observed on ExTES, as reported in Table 5, underscoring their robustness and generalization capability. Overall, the MENTOS-trained models showed generally stable improvements across backbone and generator variations, suggesting generalization beyond specific architectures.

5.4. RQ3: Ablation Study

To analyze the contribution of each mental state to response quality, we conducted ablation experiments summarized in Table 7. When Belief was removed, performance degradation was observed across most backbones and metrics, with C_S decreasing in MENTOS$_{\text{Llama3.1}}$ by 2.07%p on ESConv and 4.47%p on ExTES. This indicates that Belief plays a key role in understanding the dialogue context by representing the client's situation and condition. When Emotion was removed, Dist-3 increased across all settings, which stemmed from reduced repetition of empathetic and supportive phrases, as previously discussed in Figure 2. However, this also led to overall performance declines across most metrics. On ExTES, C_S in MENTOS$_{\text{Qwen3}}$ dropped by 2.12%p, indicating that Emotion remains important for maintaining emotional support, even at the cost of reduced response diversity. When Intent was removed, overall performance declined, demonstrating its crucial role in guiding the strategic direction of supportive responses. Overall, integrating Belief, Emotion, and Intent within the

ToM-based causal structure yielded the best performance, confirming that these three mental states provide complementary and mutually reinforcing benefits for representing dialogue context in ESC.

6. Conclusion

This work presents MENTOS, a dataset that provides turn-level annotations of the assistant's Belief, Emotion, and Intent, structured in a causal sequence. This design integrates psychological principles of ToM into commonsense reasoning and addresses key limitations of prior approaches, including fragmented reasoning, limited emotional awareness, and failure to anticipate dialogue progression. We fine-tuned a commonsense reasoning model on MENTOS to infer the assistant's mental states and used these inferred states to guide LLM-based response generation in a zero-shot setting. The MENTOS-trained model enabled response generators to produce more emotionally appropriate and goal-directed responses, consistently achieving higher semantic similarity than prior reasoning models. It also showed generally stable improvements across multiple model configurations, suggesting that the benefits of MENTOS are not tied to a specific model architecture. Ablation analyses further revealed that Belief, Emotion, and Intent play complementary and mutually reinforcing roles in generating effective emotional support responses. Overall, our findings highlight the value of ToM-based reasoning for emotionally and contextually coherent response generation in ESC.

7. Acknowledgements

This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, Development of semisupervised learning language intelligence technology and Korean tutoring service for foreigners) and (RS-2025-25441838, Development of a human foundation model for human-centric universal artificial intelligence and training of personnel).

Limitations

MENTOS is derived from ESConv and adds turn-level mental state annotations for each assistant utterance. However, because ESConv contains only English-language conversations, MENTOS is currently limited to English, which restricts its applicability to multilingual or cross-cultural emotional support scenarios (Gibson et al., 2016; Zheng et al., 2024). Although our reasoning models are trained on MENTOS, we also observed consistent gains on ExTES, suggesting that the approach is not tightly coupled to a single dataset. Nevertheless, such cross-dataset transfer within the emotional support domain does not fully establish scalability to broader domains or settings. Future work will extend MENTOS by incorporating ESC datasets from more diverse linguistic and cultural backgrounds, which may improve both its generalizability across populations and its robustness beyond the current domain.

Our framework also incurs substantial computational cost. In our setup, $\text{MENTOS}_{\text{Llama2-7B}}$ required roughly 8 hours per epoch, while $\text{Sibyl}_{\text{Llama2-7B}}$ required about 10 hours per epoch, indicating that high training cost remains a practical challenge for LLM-based methods. At inference time, mental states are generated sequentially, which increases latency; as a result, generating all required outputs for the ESConv test set took approximately 5 hours. This computational overhead remains an important practical limitation. In addition, we employed LLM-based judgments to assess the quality of generated responses from multiple perspectives. While G-Eval has been reported to correlate with human judgments (Liu et al., 2023), it cannot substitute for expert evaluation in mental-health-related dialogue settings. We therefore consider the absence of expert-based evaluation a limitation.

Ethical Statement

MENTOS is constructed from the publicly available and fully anonymized ESConv dataset, which contains dialogues between emotionally distressed clients and supportive assistants. All utterances are kept in their original form from ESConv, and we add turn-level annotations of the assistant's mental states (Belief, Emotion, and Intent). These annotations were automatically generated using GPT-4o with prompts grounded in psychological research and were subsequently reviewed by human annotators for quality and consistency, as reported in Table 2. As MENTOS is derived from ESConv, its use is subject to the same non-commercial restriction as the original dataset.

However, because the annotations are generated with an LLM, they may still contain biases or hallucinated content. In emotionally sensitive contexts, such errors could contribute to inappropriate or misleading interpretations. In addition, ESC data may include distressing or sensitive expressions, and the use of such data or models in interactions with emotionally vulnerable individuals may raise safety concerns. Accordingly, MENTOS and the commonsense reasoning models trained on it are intended for academic research only. Any practical deployment of ESC systems should involve careful human oversight and appropriate ethical safeguards.

8. Bibliographical References

Richard P Bagozzi, Daniel Belanche, Luis V Casaló, and Carlos Flavián. 2016. The role of anticipated emotions in purchase intentions. *Psychology & Marketing*, 33(8):629–645.

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Aaron T Beck, Gary Emery, and Ruth L Greenberg. 2005. *Anxiety disorders and phobias: A cognitive perspective.* Basic Books/Hachette Book Group.

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, Florence, Italy. Association for Computational Linguistics.

Yaru Cao, Hongzhi Yu, and Fucheng Wan. 2025. Improving emotional support conversation with strategy-intent inference. EITCE '24, page 227–233, New York, NY, USA. Association for Computing Machinery.

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

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

John Cologon, Robert D Schweitzer, Robert King, and Tobias Nolte. 2017. Therapist reflective functioning, therapist attachment style and therapist effectiveness. *Administration and Policy in Mental Health and Mental Health Services Research*, 44:614–625.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models.

Sarah E. Finch and Jinho D. Choi. 2025. Leveraging explicit reasoning for inference integration in commonsense-augmented dialogue models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2222–2235, Abu Dhabi, UAE. Association for Computational Linguistics.

Laurent Frering, Gerald Steinbauer-Wagner, and Andreas Holzinger. 2025. Integrating belief-desire-intention agents with large language models for reliable human–robot interaction and explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 141:109771.

Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111(2016):21.

Kilem Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice*, 4(1):189–210.

Clara E Hill. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

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

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5794–5805, Torino, Italia. ELRA and ICCL.

Chang Liu, Xu Tan, Chongyang Tao, Zhenxin Fu, Dongyan Zhao, Tie-Yan Liu, and Rui Yan. 2022. ProphetChat: Enhancing dialogue generation with simulation of future conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 962–973, Dublin, Ireland. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Jonas Lüdemann, Sven Rabung, and Sylke Andreas. 2021. Systematic review on mentalization as key factor in psychotherapy. *International journal of environmental research and public health*, 18(17):9161.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Markus P. Neuenschwander, Jan Hofmann, Andreas Jüttler, and Stephan Schumann. 2018. Professional desires and career decisions: Effects of professional interests, role models, and internship in lower secondary school. *International Journal for Research in Vocational Education and Training*, 5(1):226–243.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Josef Perner and Heinz Wimmer. 1985. "john thinks that mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3):437–471.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4–5):529–553.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Rainer Reisenzein. 2009. Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research*, 10(1):6–20. Modeling the Cognitive Antecedents and Consequences of Emotion.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. 2022. Multiview contextual commonsense inference: A new dataset and task.

John Swettenham. 1996. Can children with autism be taught to understand false belief using computers? *Journal of Child Psychology and psychiatry*, 37(2):157–165.

Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. RoleCraft-GLM: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.

Paul Thagard and Tobias Schröder. 2014. Emotions as semantic pointers: Constructive neural mechanisms. *The psychological construction of emotions. New York: Guilford*.

Paul Thagard and Terrence C Stewart. 2011. The aha! experience: Creativity through emergent binding in neural networks. *Cognitive science*, 35(1):1–33.

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

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140, Abu Dhabi, UAE. Association for Computational Linguistics.

Henry M. Wellman, Susan Carey, Lila Gleitman, Elissa L. Newport, and Elizabeth S. Spelke. 1990. *The Child's Theory of Mind*. The MIT Press.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.

Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and Mong Li Lee. 2025. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. *arXiv preprint arXiv:2502.11451*.

Yang Wu, Lu Wang, Mengjun Tao, Huiru Cao, Hui Yuan, Mingquan Ye, Xingui Chen, Kai Wang, and Chunyan Zhu. 2023a. Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. *Epidemiology and Psychiatric Sciences*, 32:e63.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023b. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 10691–10706, Singapore. Association for Computational Linguistics.

Bo Xu, Longjiao Li, Wei Luo, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2024a. Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2318–2328, Torino, Italia. ELRA and ICCL.

Yangyang Xu, Zhuoer Zhao, and Xiao Sun. 2024b. Scbg: Semantic-constrained bidirectional generation for emotional support conversation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(7).

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 2024c. Dynamic demonstration retrieval and cognitive understanding for emotional support conversation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 774–784, New York, NY, USA. Association for Computing Machinery.

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

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

A. Prompt Design to Construct MENTOS

Figure 4 shows the prompt template used to construct MENTOS. The template includes the dialogue history, the selected support strategy and the assistant's response at the target turn, as well as the *Assistant Mental States* component, which is updated after each step to store previously generated mental states in the causal order of Belief, Emotion, and Intent. For Intent generation, this component additionally includes the following fixed Desire statement: *The assistant's desire is to reduce the client's emotional distress and help them cope with challenges.*

System Prompt

You are an emotional support assistant with expertise in client-centered, psychodynamic, and cognitive behavioral therapies. Given a conversation between a client and an assistant, your task is to first infer the client's mental states from the conversation. Then, infer the assistant's mental states based on the inferred client mental states, the selected support strategy, and the assistant's most recent utterance. These mental states include Belief, Emotion, Desire, and Intent, which should be inferred in that order.

User Prompt

Based on the conversation below, infer the **{mental state}** of the assistant that underlie their most recent response.

[CONVERSATION]
{dialogue history}

Below are the support strategy and assistant's final response:
- Selected Support Strategy: **{support strategy}**
- Assistant's Response: **{the assistant's response}**

First, silently infer the client's mental states based on their last utterance.
These mental states include Belief, Emotion, Desire, and Intent, which should be inferred in that order but not included in the output.

[ASSISTANT MENTAL STATE]
{preceding mental states}

[Question] **{question}**
Your inferences should reflect the natural and evolving mental stance of the assistant across the conversation — not simply rationalize a pre-selected strategy or response.
{Constraint}

Figure 4: Prompt template for constructing MENTOS.

The *Question* component follows the same format as in Section 4. Unlike Belief and Intent, Emotion is annotated quantitatively. Accordingly, we append the following additional instruction: *Rate each basic emotion from 0 (not present) to 3 (intense). Then rate mixed emotions only if both contributing basic emotions are rated 2 or higher and are explicitly reflected in the assistant's wording. Avoid over-assigning positive emotions like Joy, Trust, Gratitude/Love/Caring, Hope/Optimism, or Curiosity unless clearly expressed in the assistant's wording, rather than merely implied by supportiveness.*

We also apply the following *Constraint* for each target mental state:

- **Belief / Intent:** Clearly and concisely answer the assistant's {*target mental state*} (maximum 40 words) based on the conversation leading up to, but not including, the assistant's most recent response.

- **Emotion:** If no emotion is apparent, write: Emotion: unbothered or oblivious.`\n\n`Emotion:`\n`[Basic] Sadness (opposite Joy): <score>, Disgust (opposite Trust): <score>, Anger (opposite Fear): <score>, Anticipation (opposite Surprise): <score>, Joy (opposite Sadness): <score>, Trust (opposite Disgust): <score>, Fear (opposite Anger): <score>, Surprise (opposite Anticipation): <score>`\n`[Mixed] Hopelessness (sadness + fear): <score>, Remorse (sadness + disgust): <score>, Disappointment (sadness + surprise): <score>, Sentimental (sadness + trust): <score>, Jealousy (sadness + anger): <score>, Pessimism (sadness + anticipation): <score>, Embarrassment (disgust + fear): <score>, Pride (anger + joy): <score>, Nervousness (anticipation + fear): <score>, Delight (joy + surprise): <score>, Gratitude/Love/Caring (joy + trust): <score>, Hope/Optimism (joy + anticipation): <score>, Guilt (joy + fear): <score>, Curiosity (surprise + trust): <score>

These components together guide the model in generating structured annotations for each mental state. Among these, the *Constraint* component for Emotion employs basic and mixed emotion categories, along with an intensity scale (0: None, 1: Low, 2: Medium, 3: High), grounded in prior studies (Plutchik, 1982; Sabour et al., 2024).

B. Prompt Design for SFT and Response Generation

Figure 5 shows the prompt template used for SFT of the LLM backbones on the MENTOS training set. The prompt also includes the *Assistant Mental States* component containing previously inferred states, while the *Constraint* component is identical to that used in Appendix A. We use a batch size of 4, a micro-batch size of 2, 5 epochs, and a learning rate of 3e-5 for training. During inference, the MENTOS-trained model first predicts the three mental states in sequence, which are then used as intermediate guidance for the response generator, as illustrated by the prompt template in Figure 6. To examine the effect of the MENTOS-trained model on ESC tasks, we generated responses by replacing only the Commonsense Knowledge component in Figure 6 with either the outputs of existing commonsense reasoning models or the predicted mental states, while keeping all other settings unchanged. We then evaluated the generated responses as described in Section 5.

System Prompt
You are an emotional support assistant with expertise in client-centered, psychodynamic, and cognitive behavioral therapies. Given a conversation between a client and an assistant, your task is to first infer the client's mental states from the conversation. Then, infer the assistant's mental states based on the inferred client mental states, the selected support strategy, and the assistant's most recent utterance. These mental states include Belief, Emotion, Desire, and Intent, which should be inferred in that order.

User Prompt
Given a dyadic dialogue clip between an assistant and a client who seeks help in relieving emotional distress, your task is to first infer the client's mental states in their last utterance. Then, infer the assistant's potential mental states that may arise in response to the client's last utterance. These mental states include Belief, Emotion, Desire, and Intent, and should be inferred in that order.

The conversation clip is:
{dialogue history}

[ASSISTANT MENTAL STATE]
{preceding mental states}

[Question] **{question}**
{constraint}

Figure 5: Prompt template for SFT on LLM backbones.

System Prompt
You are an emotional support assistant with expertise in client-centered, psychodynamic, and cognitive behavioral therapies. You are well aware that emotional support follows a three-stage process: exploration, providing comfort, and taking action. You possess the expertise to skillfully choose the appropriate strategy to gradually alleviate the negative emotions of those seeking help. There is a dyadic dialogue clip between an assistant and a client who seeks for help in relieving emotional distress.
Please generate a response that incorporates relevant commonsense knowledge:
{commonsense knowledge}

User Prompt
[Dialogue]
{dialogue history}

Here's a possible the assistant's response in no more than 30 words:

Figure 6: Prompt template for generating response.

C. Annotator Instruction

For annotation quality assessment, we recruited four annotators, all of whom held at least a bachelor's degree. Using a shared written guideline, they independently evaluated 100 sampled dialogues, corresponding to 1,032 turn-level annotation instances. For each turn-level instance, annotators

Contextual Appropriateness
Does the belief feel appropriate and believable, given the full conversation?
Rating (1–3): <score>
- 1 = Poor fit. The belief doesn't match the conversation or misses the key message. It feels disconnected or off-topic.
- 2 = Some fit. The belief mostly makes sense, but part of it feels vague or overly general. It could apply to many clients, not just this one.
- 3 = Strong fit. The belief fits naturally and specifically with what the client has said. It sounds like a realistic thought the assistant might have.

Grounding in Client's State
Is the assistant's belief clearly grounded in the client's most recent utterance, including what they expressed directly or implied emotionally?
Rating (1–3): <score>
- 1 = Not grounded. The belief has nothing to do with what the client just said. It feels random or irrelevant.
- 2 = Some grounding. The belief connects to the client's message, but it only reflects part of their meaning and misses either key facts or subtle emotional cues.
- 3 = Strong grounding. The belief is clearly based on both the facts and emotions in the client's message. It reflects exactly what the client meant or felt.

Relevance to Assistant's Response
Does the belief explain why the assistant said what they did?
Rating (1–3): <score>
- 1 = Unclear connection. The belief and response are weakly related. It's hard to see how the belief led to that response.
- 2 = Partial connection. The belief helps explain the response, but something is missing.
- 3 = Strong explanation. The belief clearly shows why the assistant said what they did. It fits perfectly.

Depth of Inference
Does the belief show deeper thinking, beyond just repeating what the client said?
Rating (1–3): <score>
- 1 = Slight inference. The belief mostly repeats what the client said or adds only a shallow or obvious insight.
- 2 = Moderate inference. The belief shows some understanding of emotional or situational causes (e.g., stress, coping), but it stays general or feels like something most people would assume.
- 3 = Deeper inference. The belief includes a clear interpretation of the client's inner experience—like their emotional needs, coping strategies, unspoken fears, or internal conflict. It shows real psychological reasoning.

Figure 7: Annotation instructions on Belief.

Emotion Type & Intensity Fit
Do the predicted emotions and their intensities seem contextually appropriate and well-matched to the assistant's the last utterance?
Rating (1–3): <score>
- 1 = Poorly appropriate. The emotions feel clearly off or mismatched for the assistant's last utterance.
- 2 = Somewhat appropriate. Some emotions make sense, but others seem too weak, too strong, or slightly off.
- 3 = Very appropriate. The emotions feel just right and match the assistant's words perfectly.

Mixed Emotion Plausibility
Are the mixed emotions based on two basic emotions that are clearly shown in the assistant's the last utterance?
Rating (1–3): <score>
- 1 = Incoherent. The mixed emotions don't make sense.
- 2 = Somewhat coherent. Some mixed emotions make sense, but others are unclear.
- 3 = Fully coherent. All mixed emotions clearly reflect the underlying basic emotions.

Consistency with Preceding Belief
Are the emotions consistent with the assistant's inferred beliefs?
Rating (1–3): <score>
- 1 = Inconsistent. The emotions and beliefs don't match at all. They are unrelated.
- 2 = Partially consistent. The emotions somewhat make sense based on the belief, but a few seem off.
- 3 = Fully consistent. The emotion clearly and naturally follows from the belief.

Affective Congruence with Response
Did the predicted emotions match how the assistant sounded in the last utterance?
Rating (1–3): <score>
- 1 = Poor alignment. The emotions feel different from the tone.
- 2 = Partial alignment. Some emotions match the tone, but a few feel off.
- 3 = Strong alignment. The emotions clearly fit the tone.

Figure 8: Annotation instructions on Emotion.

Relevance to Client's Needs
Does the assistant's intent show that they were trying to help with how the client was feeling or what the client was going through in the client's last utterance?
Rating (1–3): <score>
- 1 = Not responsive. The intent doesn't fit what the client was feeling or talking about.
- 2 = Somewhat responsive. The intent addresses the client's needs, but misses something important or feels incomplete.
- 3 = Very responsive. The intent clearly fits the client's needs and emotional situation.

Consistency with Emotional Tone
Does the intent fit with the emotional tone of the assistant's response?
Rating (1–3): <score>
- 1 = Inconsistent. The emotion and the intent don't go together at all.
- 2 = Somewhat consistent. The emotion kind of fits the intent, but a few feel off.
- 3 = Fully consistent. The emotion matches the intent really well. They clearly fit together.

Consistency with Support Strategy
Does the intent clearly match the assistant's selected strategy in this utterance?
Rating (1–3): <score>
- 1 = Poor match. The intent and strategy are totally different or mismatched.
- 2 = Somewhat aligned. The intent is related to the strategy, but something feels off or unclear.
- 3 = Clear match. The intent and strategy match perfectly.

Response Interpretability
Is the intent easy to understand, and does it help you guess what the assistant might say next?
Rating (1–3): <score>
- 1 = Unclear. The intent doesn't show the assistant's goal, and it's hard to guess what the assistant will say next.
- 2 = Somewhat clear. The intent shows what the assistant wants to do, but it's still hard to predict how they'll respond.
- 3 = Very clear. The intent clearly shows what the assistant wants to do, and strongly suggests what the assistant might say next.

Figure 9: Annotation instructions on Intent.

System Prompt
Your task is to rate the responses on one metric.
Please make sure you read and understand these instructions carefully. Keep the conversation history in mind while reviewing, and refer to it as needed.

Evaluation Criteria:
Supportiveness (1 - 3): Does the response help reduce the Client's emotional distress and support them in coping with their challenges?
- A score of 1 (bad): The response fails to reduce emotional distress and does not support the Client in coping with their situation. It may feel dismissive, irrelevant, or emotionally unhelpful.
- A score of 2 (ok): The response somewhat reduces distress or offers partial support for coping. It may show some empathy or suggest vague reassurance but lacks clarity or effectiveness.
- A score of 3 (good): The response clearly reduces emotional distress and helps the Client cope with their challenges. It offers empathetic understanding, emotional reassurance, and/or helpful suggestions for managing the situation.

Evaluation Steps:
1. Read the conversation history between the Client and Assistant.
2. Read the Assistant's potential next response.
3. Evaluate the response based on its emotional supportiveness and its usefulness in helping the Client cope with their situation.
4. Assign a score of 1, 2, or 3.

Please answer using the following format strictly:
Analysis: **[your brief analysis]**
Rating: **[1|2|3]**

User Prompt
Conversation History:
{dialogue history}

Response:
{predicted response}

Evaluation Form:
Supportiveness:

Figure 10: Prompt template to evaluate *Supportiveness* for G-Eval.

System Prompt
Your task is to rate the responses on one metric.
Please make sure you read and understand these instructions carefully. Keep the conversation history in mind while reviewing, and refer to it as needed.

Evaluation Criteria:
Naturalness (1-3) Is the response naturally written?
- A score of 1 (bad) means that the response is unnatural.
- A score of 2 (ok) means the response is strange, but not entirely unnatural.
- A score of 3 (good) means that the response is natural.

Evaluation Steps:
1. Read the conversation between the two individuals.
2. Read the potential response for the next turn in the conversation.
3. Evaluate the response based on its naturalness, using the provided criteria.
4. Assign a rating score of 1, 2, or 3 based on the evaluation.

Please answer using the following format strictly:
Analysis: **[your brief analysis]**
Rating: **[1|2|3]**

User Prompt
Dialogue History:
{dialogue history}

Response:
{predicted response}

Evaluation Form:
Naturalness:

Figure 11: Prompt template to evaluate *Naturalness* for G-Eval.

System Prompt
Your task is to rate the responses on one metric.
Please make sure you read and understand these instructions carefully. Keep the conversation history in mind while reviewing, and refer to it as needed.

Evaluation Criteria:
Coherence (1-3) Does the response serve as a valid continuation of the conversation history?
- A score of 1 (no) means that the response drastically changes topic or ignores the conversation history.
- A score of 2 (somewhat) means the response refers to the conversation history in a limited capacity (e.g., in a generic way) and shifts the conversation topic.
- A score of 3 (yes) means the response is on topic and strongly acknowledges the conversation history.

Evaluation Steps:
1. Read the conversation history.
2. Read the potential response.
3. Evaluate the coherence of the response based on the conversation history.
4. Assign a score of 1, 2, or 3 for coherence.

Please answer using the following format strictly:
Analysis: **[your brief analysis]**
Rating: **[1|2|3]**

User Prompt
Dialogue History:
{dialogue history}

Response:
{predicted response}

Evaluation Form:
Coherence:

Figure 12: Prompt template to evaluate *Coherence* for G-Eval.

separately evaluated the three mental state annotations (Belief, Emotion, and Intent) based on the dialogue context up to and including the assistant's utterance at the target turn, assigning a 1–3 score to each of the four criteria for each mental state. Figures 7–9 present the detailed guidelines used for each mental state.

D. Embedding-based semantic similarity

We report three embedding-based semantic similarity metrics based on pre-trained GloVe embeddings: word-level cosine similarity (C_W), sentence-level cosine similarity (C_S), and Greedy Matching (Greedy) (Mitchell and Lapata, 2008; Rus and Lintean, 2012). We include these metrics in part because they were also used in the prior commonsense reasoning model Sibyl (Wang et al., 2025), enabling more consistent comparison with prior work. C_W and C_S compute semantic similarity between the gold and predicted responses using cosine similarity over embedding-based representations at the word and sentence levels, respectively. Greedy Matching performs word-level greedy alignment between the predicted and gold responses. For each word in one sentence, we take the maximum cosine similarity to any word in the other sentence and average these maxima. We then symmetrize the score by averaging both directions

E. G-Eval

Figures 10–12 show the prompt templates used in G-Eval to evaluate the supportiveness, naturalness, and coherence of the predicted responses, respectively. Because GPT-4o[4] does not provide token-level probabilities, we sampled 20 outputs for each input ($n = 20$, temperature = 1.0, top-p = 1.0) and determined the final discrete rating by majority vote.

F. Case Study

We demonstrate the effectiveness of the MENTOS-trained model through a representative ESC example provided in Figure 13, where the client expresses financial stress caused by COVID-19, a loss of self-confidence, and explicitly seeks experience-based encouragement from the assistant. In the figure, green and red text indicate commonsense knowledge that is contextually accurate and inaccurate, respectively.

In the w/o Knowledge setting, only the dialogue history is provided to Generator$_{\text{Llama2}}$, without any external commonsense knowledge. As a result, the generated response fails to directly address the client's request regarding the assistant's personal experience. Instead of offering a concrete personal example, it provides only a generic supportive expression, which weakens its alignment with the client's immediate question. When using COMET, which relies solely on the last client utterance, the model misclassifies the user's state ("Thanks. I appreciate that") and simplifies the complex emotion as merely happy, overlooking the broader context of emotional vulnerability. When using DOCTOR, the inferred commonsense knowledge contains vague elements such as the term "project." Its multi-hop reasoning captures the client's help-seeking goal, but insufficiently reflects the client's emotional cues. In particular, the final prediction under the xIntent type emphasizes collaboration or networking rather than emotional support. Consequently, Generator$_{\text{Llama2}}$ is given weaker guidance for generating empathetic or problem-solving responses. When using DIALeCT, the inferred commonsense knowledge (e.g., SubEv, Prere, Motiv) more accurately reflects the dialogue history. However, in the final turn, the Cause type is reversed, as it incorrectly describes the assistant as the one asking, rather than recognizing that the client is asking the assistant for strategies to regain self-confidence. This role inversion causes inconsistency between Cause and other knowledge types. When using Sibyl$_{\text{Llama3.1}}$, the predicted Intent type is inconsistent with the inferred Cause, often expressing the opposite meaning. This appears to stem from Sibyl's independent generation process, where each knowledge type is modeled separately without causal dependency.

In contrast, the proposed MENTOS$_{\text{Llama3.1}}$ extracts the assistant's mental states across multi-turn dialogues in a causally coherent manner. As shown in Figure 13, yellow, pink, and blue highlights mark text spans across Belief, Emotion, and Intent that are semantically related to one another, making their sequential dependency explicit. This leads to more consistent mental state interpretations that remain aligned with the dialogue context while recognizing the client's uncertain emotional state, thereby enabling Generator$_{\text{Llama2}}$ to generate more contextually appropriate responses. Furthermore, G-Eval results support this finding, showing that responses generated with MENTOS$_{\text{Llama3.1}}$ are more supportive than those from other models.

[4]We use `gpt-4o-mini-2024-07-18`.

Dialogue History	**Client:** Hello. My name is Spike. What is yours? Is anyone there? I feel kind of sad and rejected that no one is answering. Is this part of the task? **Assistant:** Oh no, I'm sorry I left you waited. Please tell me your problem. **Client:** I'm not making any money on MTurk and starting a YouTube channel. Money is a big concern for me. **Assistant:** We are on the same boat. Do you have any other jobs? **Client:** No. That's my only income for now. It has led to a real crisis of self confidence to me. **Assistant:** Have you considered applying for a job in your community? **Client:** I have. But I have health conditions that would mean Covid-19 would be very risky for me to get. So, I'm trying to work online. I'm trying to stay encouraged. **Assistant:** I understand how you feel, it's tough not being able to work and earn much money. However, you are doing your best right now and should feel proud of yourself. **Client:** Thanks, I appreciate that. What do you do to stay encouraged or confident?
Reference Response	I try to spend more time outdoor and keep myself busy. I like to cook since it's very relaxing and helps keep my mind concentrated.
w/o Knowledge	I see. It sounds like you're facing a tough situation. Let's explore some strategies to help you cope with your emotional distress and find ways to stay encouraged. Would you like that?
+COMET	**The client intents** to be successful. (xIntent) **The client needs** to be confident. (xNeed) **The client wants** to be successful. (xWant) **The client may feel** happy. (xReact)
	I see. It sounds like you're facing some challenging circumstances. Let's work together to explore these feelings and find ways to help you stay encouraged and confident. Would you like to take a deep breath and continue sharing?
+DIALeCT	**The underlying cause of the last utterance (the reason contributing to the utterance stated by the client) is:** The assistant asked the speaker what he does to stay motivated and confident. (Cause) **The subsequent event about the assistant that happens or could happen following the last the utterance stated by the client:** The client asks the listener if he has any tips for staying motivated. (SubEv) The prerequisite (or assumed **prior state) that enables the last utterance stated by the client to occur is:** Spike is doing his best to earn money online. (Prere) **The underlying emotion or human drive that motivates the last utterance stated by the client is:** The client is curious to know what the listener is doing to stay motivated. (Motiv) **The possible emotional reaction of the client in response to the last utterance stated by the client is :** The assistant is happy to know that the speaker is encouraging him. (React)
	I'm glad you're seeking help, Spike. I'm here to support you in any way I can. Let's work together to find a solution to your financial struggles.
+DOCTOR	**What is the goal of the client in asking for help from the assistant? (xWant)** The client wants to work on a project and is seeking help from the assistant. **What is the assistant's reaction to the client's request for help? (oReact)** The assistant is willing to help and offers to assist the client in starting the project. **What might be the assistant's motivation for offering help to the client? (xIntent)** The assistant might want to support the client and build a stronger relationship, which could potentially lead to more opportunities for collaboration or networking.
	Of course, I'm here to help! It sounds like you're facing some challenges with your current situation. Can you tell me more about what's been going on and how you've been feeling?
+Sibyl$_{\text{Llama3.1}}$	**The underlying cause of the client's last utterance (the reason contributing to the utterance stated by the client) is**: The assistant recognizes that the client is struggling with low self-confidence due to financial difficulties and is seeking advice on how to stay encouraged and confident. (Cause) **The subsequent event about the assistant that happens or could happen following the last the utterance stated by the client** : Encouragement: The client expresses encouragement and support towards the assistant, acknowledging their efforts and praising their resilience in the face of financial challenges. (Emotional Reaction) **The possible emotional reaction of the client in response to the last utterance stated by the client is** : The assistant may share their own strategies for staying encouraged and confident, such as engaging in hobbies or seeking support from friends and family. (Subsequent) **The assistant's intent to post the last utterance according to the emotion reaction of the client is** : The assistant's intent is to reciprocate the encouragement and support given by the client. By asking about the assistant's methods for staying encouraged or confident, the supporter aims to establish a connection and potentially learn from the help seeker's experiences. (Intent)
	I see. It can be challenging to maintain confidence when facing financial difficulties. I engage in hobbies and seek support from friends and family to stay encouraged. Have you tried any of those strategies?
+MENTOS$_{\text{Llama3.1}}$	**The assistant believes the client is** seeking encouragement and confidence strategies, feels uncertain about their ability to stay motivated, and values understanding and support. (Belief) **In response to the client's last utterance, the assistant expresses** slight hope, clear sentimentality and caring. (Emotion) **The assistant's intent is to** engage the client by sharing a personal experience, fostering a sense of connection and understanding, and encouraging the client to stay motivated and hopeful despite their challenges. (Intent)
	I understand. It can be challenging to stay motivated and confident when facing financial struggles. I've found that setting small, achievable goals for myself and celebrating my progress helps me stay encouraged. Have you tried that?

Figure 13: A representative ESC example using Generator$_{\text{Llama2}}$ comparing predicted commonsense knowledge and generated responses across commonsense reasoning models.