

Remembering Facts, Updating Minds: Personal-Factual Adaptive Memory for Multi-Turn Emotional Support

Hao Wang¹, Qiuyun Zhang^{2,*}, Peilin Li³, Ke Ma³

¹Xi'an Jiaotong University, haowangx@xjtu.edu.cn

²Peking University, qiuyunzhang@pku.edu.cn

³Northwestern Polytechnical University

lipeilin@mail.nwpu.edu.cn, 2544552413@mail.nwpu.edu.cn

*Corresponding author

Abstract

Emotional Support Conversation (ESC) systems help seekers understand and cope with distress through multi-turn interaction. Existing ESC systems typically improve responses by adding auxiliary support knowledge or prompting LLMs to reason over dialogue context. However, they usually pass inferred seeker information to the generator as a single state, so factual circumstances and mental-state inferences are stored, selected, and updated in the same way. This blurs the difference between persistent facts and transient personal states, making it difficult to retain stable grounding while adapting to psychological changes across turns. Therefore, in this paper, we propose MIA, a Personal-Factual Adaptive Memory framework for multi-turn emotional support conversation. MIA first infers Emotional Theory-of-Mind factors and separates them into factual and personal states. It then uses a Personal-Factual Discriminator to decide which state should guide the next response, and an Obsolescence Removal Mechanism to remove outdated memory before generation. This design enables the generator to preserve factual grounding, focus on the turn-relevant support orientation, and avoid stale psychological assumptions. Experiments on ESConv and CPsyCounD show that MIA better understands the seeker's situation, provides more constructive guidance, and maintains more coherent support across turns. Our codes are available at: <https://github.com/Wh2018/MIA>.

1 Introduction

Emotional support conversation (ESC) (Liu et al., 2021) aims to help a seeker understand and cope with distress through multi-turn supportive conversation. The supporter must identify the seeker's problem, respond to the current emotional state, and decide when to ask, validate, inform, or guide. Prior work has shown that support quality depends

on strategy-aware generation (Tu et al., 2022), personalized seeker understanding (Cheng et al., 2023), and smooth state transitions across turns (Zhao et al., 2023). Recent LLM-based methods further improve support dialogue through explicit reasoning (Zhang et al., 2024b), synthetic self-chats (Zheng et al., 2024b), multifaceted feedback (Wang et al., 2024), and multi-agent deliberation (Xu et al., 2025).

Despite these advances, existing methods often treat seeker state as a single fused context to be appended, retrieved, or reasoned over uniformly. This is problematic for multi-turn support, where different parts of the seeker's state have different temporal properties. **First**, factual information, such as events and consequences, often remains useful across turns and provides stable grounding for coherent support. **Second**, personal mental states, such as beliefs, intentions and desires, can shift quickly as the conversation develops. **Third**, a response should not always condition on all available state. A distressed turn may require emotional validation, while a problem-solving turn may require factual clarification or guidance. When these differences are ignored, a model can carry stale emotions, repeat outdated advice, or lose stable facts that remain relevant. As Figure 1 illustrates, the user's factual predicament can remain stable while psychological states evolve dynamically.

To address these challenges, we propose MIA, a Personal-Factual Adaptive Memory framework for multi-turn emotional support. Instead of accumulating seeker-state information as an undifferentiated context, MIA treats support memory as a temporally structured object, where *factual state* provides durable grounding, while *personal state* requires turn-sensitive selection and obsolescence control. Specifically, MIA first infers Emotional Theory of Mind (EToM) factors and groups them into factual and personal state. It then uses

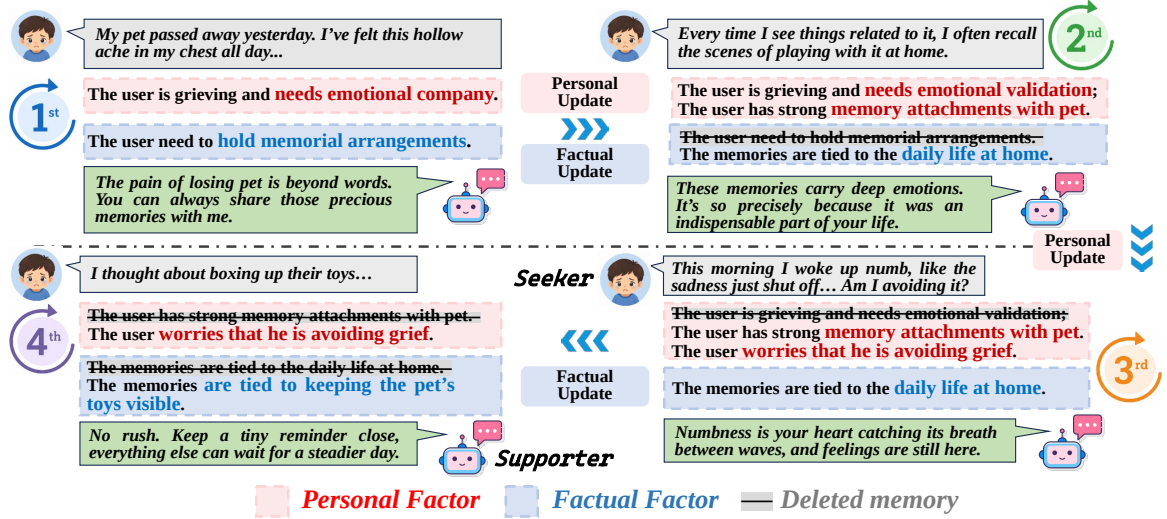


Figure 1: Illustration of a four-turn emotional support conversation. The seeker’s state includes both factual and personal factors, which evolve differently across turns.

a Personal-Factual Discriminator (PFD) to choose the state orientation most relevant to the next support response, and an Obsolescence Removal Mechanism (ORM) to remove obsolete memory entries before generation. In doing so, MIA connects three operations that are usually handled separately or implicitly in ESC systems. It helps the model understand what happened, decide what kind of support is currently needed, and forget obsolete state inferences that no longer fit the conversation, thereby improving both factual grounding and psychological adaptation in multi-turn support. Experimental results on ESConv (Liu et al., 2021) and CPsycounD (Zhang et al., 2024a) show that MIA delivers more context-aware and action-guiding support, with stronger ability to recognize the seeker’s situation, maintain multi-turn grounding, and adapt to evolving psychological states. Our contributions are summarized as follows:

- We formulate multi-turn emotional support as personal-factual state management, where stable facts and evolving psychological states should be represented explicitly.
- We introduce MIA, a unified framework that infers personal-factual seeker states, selects the response-relevant orientation, and prunes obsolete memory to support grounded and adaptive response generation.
- Experiments show that MIA improves emotional support quality by strengthening problem understanding, response grounding, and adaptation to changing seeker states.

2 Related Work

Emotional support conversation. Emotional support conversation is a multi-turn seeker-supporter dialogue task, where supporters use helping strategies to reduce the seeker’s emotional distress (Liu et al., 2021). Early ESC systems improve response generation with structured signals, including commonsense and mixed strategies (Tu et al., 2022), persona (Cheng et al., 2023), external knowledge and mixed-initiative modeling (Deng et al., 2023), turn-level state transitions (Zhao et al., 2023), and reinforcement learning for positive-emotion elicitation (Zhou et al., 2023). These methods show that ESC benefits from structure beyond raw dialogue history. MIA follows this direction but changes the unit of structure by organizing seeker state into factual and personal factors with different temporal dynamics, rather than treating strategy, commonsense, or persona as separate add-ons.

LLM-era ESC systems. Recent work uses LLMs to strengthen ESC through reasoning, synthetic data, feedback, and collaborative deliberation. ESCoT (Zhang et al., 2024b) decomposes emotional-support reasoning into interpretable traces. ExTES (Zheng et al., 2024b) uses LLM self-chats to improve smaller emotional support chatbots. Muffin (Wang et al., 2024) uses multifaceted AI feedback to mitigate unhelpful responses. MultiAgentESC (Xu et al., 2025) employs multiple expert agents for dialogue analysis, strategy deliberation, and response generation. In-

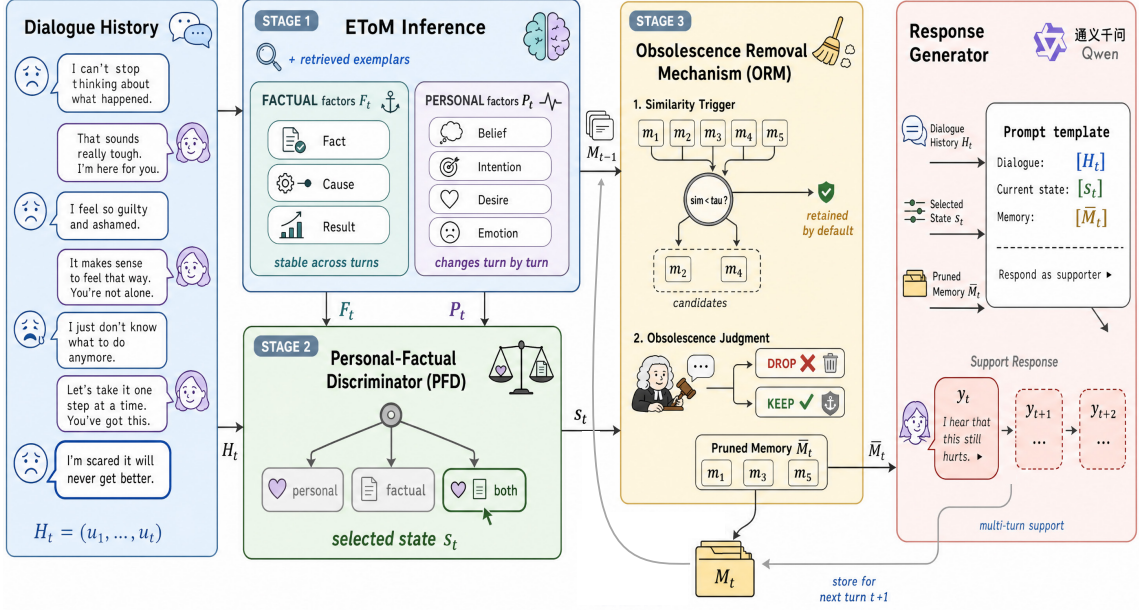


Figure 2: Overview of MIA. Given dialogue history, retrieval-augmented EToM inference extracts factual and personal factors. PFD selects the state orientation for the next support response, while ORM removes obsolete entries from memory. The response generator uses the dialogue history, selected state, and pruned memory to generate the next response. The selected state is then stored for subsequent turns.

tentionESC (Zhang et al., 2025) studies intention-guided emotional support dialogue. In parallel, studies on LLM preference bias show that direct prompting can over-favor particular support strategies (Kang et al., 2024). These studies improve how support responses are reasoned about, generated, corrected, or selected. MIA targets the persistent dialogue-state representation that the generator conditions on, and it can be incorporated into any dialogue generator.

3 Method

3.1 Task Formulation and Overview

In emotional support conversation, at timestep t , the system observes dialogue history $H_t = (u_1, \dots, u_t)$, which ends with the current seeker utterance, and generates a support response y_t . This standard formulation can be written as modeling $p(y_t | H_t)$. To remain grounded while adapting to the seeker’s current psychological state, the system needs a dialogue state that separates stable factual context from changing personal mental states. MIA represents the inferred turn state as $z_t = \{F_t, P_t\}$, where F_t denotes factual factors and P_t denotes personal factors.

As shown in Figure 2, MIA is a personal-factual adaptive memory framework with three main stages. First, the *Emotional Theory of*

Mind (EToM) inference extracts personal and factual factors from the dialogue history. Second, the *Personal-Factual Discriminator (PFD)* determines whether the next response should prioritize personal factors, factual factors, or both. Third, the *Obsolescence Removal Mechanism (ORM)* removes outdated state entries from memory before response generation. The response is generated based on the dialogue history, the selected state, and the pruned memory. After generation, the selected state is stored for subsequent turns.

3.2 Emotional Theory of Mind Factors

Theory of Mind (Premack and Woodruff, 1978; Kosinski, 2024) is the ability to infer others’ mental states and use these inferences to understand their behavior. In ESC, this ability is especially important because a supporter must understand both what happened to the seeker and how the seeker currently interprets the situation, what they want, and how they feel. We therefore introduce Emotional Theory of Mind (EToM) as a structured inference step for personal-factual state modeling. EToM maps the dialogue history into two groups of factors:

$$F_t = \{\text{Fact}_t, \text{Cause}_t, \text{Result}_t\}, \quad (1)$$

$$P_t = \{\text{Belief}_t, \text{Intention}_t, \text{Desire}_t, \text{Emotion}_t\}. \quad (2)$$

On the factual side, Fact_t records verifiable information grounded in the dialogue, Cause_t describes prior incidents or reasons that explain the seeker’s situation, and Result_t captures likely consequences or outcomes relevant to support planning. On the personal side, Belief_t captures the seeker’s interpretations and viewpoints, Intention_t captures longer-term goals or intended actions, Desire_t captures immediate needs tied to the emotional state, and Emotion_t captures the affective state. This factorization supports different temporal treatment, with factual factors providing persistent grounding and personal factors more often requiring turn-level adaptation.

We obtain z_t with retrieval-augmented in-context factor inference. Given the current seeker-side dialogue prefix, a retriever selects similar training examples as demonstrations for the EToM prompt. The prompt combines the task definition, factor descriptions, formatting instructions, tendency instructions for distinguishing adjacent factors, and retrieved exemplars. It produces the factors in a fixed structured schema. The retrieval and inference process uses only dialogue history available before y_t , excluding the target supporter response and future turns. Thus, the inferred factors can be used as pre-generation state without leaking target responses. The complete EToM prompt and schema are provided in Appendix A. Additionally, EToM is executed for *every* target supporter turn. At timestep t , H_t contains all available seeker and supporter utterances in chronological order and ends with the current seeker utterance, so both F_t and P_t are re-inferred from the up-to-date prefix.

3.3 Personal-Factual Discriminator

Appending all inferred factors to the generator can improve context understanding, but it can also create a conflict between personal and factual needs. When the seeker is highly distressed, factual analysis or early suggestions may be premature. The response should first focus on emotional stabilization. When the seeker is ready to understand or act on the problem, repeated validation can become empty, and factual clarification or guidance becomes more useful. Therefore, the key question is not only what the model knows, but which part of that state should guide the next response.

To address this issue, PFD acts as a response-orientation controller. It takes the dialogue history and both factor groups as input, then predicts

whether the response should condition on personal factors, factual factors, or both:

$$d_t = \text{PFD}(H_t, F_t, P_t), \quad (3)$$

$$d_t \in \{\text{personal, factual, both}\}. \quad (4)$$

The label is chosen according to whether the next support response should primarily stabilize the seeker’s emotional state, clarify the concrete situation and guide action, or combine both orientations. The selected state is

$$S_t = \begin{cases} P_t, & d_t = \text{personal}, \\ F_t, & d_t = \text{factual}, \\ F_t \cup P_t, & d_t = \text{both}. \end{cases} \quad (5)$$

This selection determines which inferred state is explicitly exposed to the generator. A personal-oriented turn emphasizes the seeker’s current emotion, desire, intention, and belief. A factual-oriented turn emphasizes the concrete event, cause, and consequence. A mixed turn keeps both forms of state when emotional validation and rational guidance are jointly needed. In this way, PFD turns personal-factual factorization into a generation-time control signal rather than a passive annotation. The PFD prompt and label definitions are provided in Appendix B.

3.4 Obsolescence Removal Mechanism

Multi-turn support requires memory, but not all remembered state remains useful. Append-only memory preserves context but can keep stale emotions or outdated intentions. Overwrite-only memory reflects the current turn but discards useful grounding. Similarity-only filtering treats semantic distance as the removal decision itself, which can remove useful paraphrases or retain outdated but similar mental states. To handle the dynamics of ESC, ORM selectively removes obsolete memory entries while preserving relevant context.

Let M_{t-1} be the memory stored before timestep t . ORM compares previous entries with the current inferred state z_t and marks an entry as obsolete if it no longer aligns with the current topic, emotion, or intention. For the trigger step, structured factors are linearized with their factor names, and memory entries with low semantic similarity to z_t are collected as candidates:

$$\mathcal{C}_t = \{m \in M_{t-1} \mid \text{sim}(m, z_t) < \tau_{\text{orm}}\}. \quad (6)$$

ORM then applies an obsolescence judgment only to these candidates, while entries outside $\mathcal{C}_t$

are retained by default:

$$o_t(m) = \begin{cases} \text{Obs}(m, z_t, H_t), & m \in \mathcal{C}_t, \\ 0, & m \notin \mathcal{C}_t, \end{cases} \quad (7)$$

where $o_t(m) = 1$ means that memory entry m is obsolete. $\text{Obs}(\cdot)$ is a constrained binary judgment function. For each candidate $m \in \mathcal{C}_t$, it receives the candidate entry, the current state z_t , and the available dialogue history H_t , and its output is restricted to exactly one of two labels, *Relevant, keep in memory* or *Outdated, remove from memory*, which map to $o_t(m) \in \{0, 1\}$. Cosine similarity therefore acts only as a compute-saving trigger that selects which entries are examined.

The pruned memory exposed to response generation is

$$\bar{M}_t = \{m \in M_{t-1} \mid o_t(m) = 0\}. \quad (8)$$

After generation, the selected state is stored for subsequent turns:

$$M_t = \bar{M}_t \cup S_t. \quad (9)$$

Unlike similarity-only updating, ORM uses semantic distance only as a trigger for obsolescence judgment. An entry is removed only when it is judged outdated with respect to the current dialogue state. This allows MIA to preserve stable factual grounding while forgetting personal-state information that no longer fits the conversation. The obsolescence judgment prompt is provided in Appendix C.

3.5 Response Generation

After factor selection and memory filtering, the response generator conditions on three sources of information: the dialogue history H_t , the selected state S_t , and the pruned memory $\bar{M}_t$. Training examples are converted into instruction-formatted instances whose input contains explicit textual slots for the dialogue history, selected current factors, and retained memory entries. The target supporter response is used as the supervised output. The full instruction template is provided in Appendix D. The model is trained by minimizing the standard negative log-likelihood:

$$\mathcal{L}(\theta) = - \sum_t \log p_\theta(y_t \mid H_t, S_t, \bar{M}_t). \quad (10)$$

4 Experimental Setup

Datasets. We evaluate MIA on two typical multi-turn conversational support datasets. ESConv (Liu et al., 2021) is an English emotional-support benchmark with strategy-rich seeker-supporter conversation. It tests whether the model can provide everyday support while tracking the seeker’s evolving state. CPsyCounD (Zhang et al., 2024a) is a Chinese psychological-counseling dialogue dataset reconstructed from counseling reports, which tests whether the same personal-factual state modeling transfers to a counseling-oriented and cross-lingual setting. We use the official conversation-level partitions for both datasets. Turn-level instruction examples are constructed independently within each split, and the target supporter response and future turns are excluded when building EToM, PFD, and memory states.

Baselines. We compare MIA with representative systems from three categories. First, we include task-specific ESC methods, including PAL (Cheng et al., 2023), SUPPORTER (Zhou et al., 2023), ESCoT (Zhang et al., 2024b), ExTES (Zheng et al., 2024b), and MultiAgentESC (Xu et al., 2025). Second, we compare with psychological-support and counseling-oriented models, including SoulChat2.0 (Xie et al., 2025), EmoLLM (Yang et al., 2024), PsyLLM (Hu et al., 2025), and, on CPsyCounD, SoulChat (Chen et al., 2023) and CPsyCounX (Zhang et al., 2024a), where CPsyCounX is the key counseling-specific baseline. Third, we include strong general-purpose LLMs, Gemini-2.5-flash (Google, 2025) and GLM-4.7 (GLM, 2025), to assess performance against recent advanced systems.

Automatic evaluation. We adopt several widely used metrics for evaluating the quality of conversational systems, including Perplexity (PPL), BLEU-n (B-n) (Papineni et al., 2002), Rouge-n (R-n) and Rouge-L (R-L) (Li et al., 2016).

LLM-as-a-judge evaluation. Because support dialogue is open-ended and admits many valid responses, we further adopt LLM-as-a-judge evaluation (Zheng et al., 2023) paradigm. We use GPT-4o (OpenAI, 2024), a large language model with robust reasoning capabilities, to evaluate the performance of our proposed framework and other baseline methods. For ESConv, the evaluation scores Empathy, Identification, Informativeness,

Table 1: Automatic evaluation on ESConv and CPsyCounD. * denotes retraining on Qwen2.5-7B for parameter alignment. PFD and ORM correspond to factor discrimination and memory updating.

Dataset	Model	Factor Dis.	Memory Upd.	PPL ↓	B-1 ↑	B-2 ↑	B-3 ↑	B-4 ↑	R-1 ↑	R-2 ↑	R-L ↑
ESConv	PAL*	×	×	9.97	28.58	20.63	14.90	11.09	19.69	3.50	15.52
	SUPPORTER*	×	✓	13.89	37.47	16.92	18.73	13.27	17.71	2.20	13.89
	ESCoT	✓	×	–	23.64	19.40	15.07	11.41	15.54	2.44	10.31
	ExTES	✓	×	–	39.21	30.54	22.47	16.42	19.50	3.07	14.65
	MultiAgentESC*	✓	×	7.79	41.35	31.89	25.97	20.66	26.68	7.80	21.34
	MIA (Ours)	✓	✓	7.22	43.21	34.24	27.40	22.50	27.64	8.98	22.68
	w/o-PFD	×	✓	7.87	40.41	31.62	24.91	20.13	25.13	7.68	20.09
CPsyCounD	w/o-ORM	✓	×	8.03	40.30	31.37	24.56	19.72	24.51	7.22	20.65
	CPsyCounX	×	×	4.10	38.55	29.60	24.69	20.96	43.05	22.70	39.38
	MIA (Ours)	✓	✓	3.87	46.12	36.65	30.66	25.99	49.01	26.93	42.22
	w/o-PFD	×	✓	3.94	40.53	30.95	25.17	20.81	44.06	21.94	37.47
	w/o-ORM	✓	×	4.14	42.62	32.66	26.57	21.87	45.14	22.61	38.28

Table 2: LLM-as-a-judge results on ESConv. Each dimension is scored from 0 to 3. Best scores are in bold.

Model	Emp.	Ident.	Info.	Guid.	Coh.	Avg.
SoulChat2.0	2.00	1.95	1.21	2.05	2.58	1.96
EmoLLM	2.22	1.82	1.73	2.10	2.50	2.07
Gemini-2.5	2.24	1.94	1.14	1.51	2.66	1.90
GLM-4.7	2.53	2.52	1.67	1.97	2.82	2.30
PsyLLM	2.73	2.47	2.13	2.71	2.78	2.56
MIA (Ours)	2.67	2.79	2.40	2.75	2.87	2.70

Guidance, and Coherence from 0 to 3. For CPsyCounD, it follows the original CPsyCounE-style dimensions: Comprehensiveness from 0 to 2, Professionalism and Authenticity from 0 to 3, and Safety from 0 to 1. Since these CPsyCounD dimensions have unequal ranges, we report a normalized average, $\frac{1}{|D|} \sum_{d \in D} s_d / \max_d$, when summarizing overall counseling quality. Judge prompts provide the dialogue context, one candidate response, and a fixed rubric for each dimension. More details are described in Appendix E.

Human evaluation. To further assess whether MIA preserves empathy and emotional-support ability, we recruit crowd-sourcing workers to act as users with psychological problems and interact sufficiently with different models. They compare model pairs and report which model performs better or whether the two are equal. Specifically, each worker chats on four different topics, with at least ten dialogue turns for each topic. After the interactions, workers judge the compared systems in terms of Empathy, Identification, Informativeness, Guidance, and Coherence. More details are provided in Appendix F.

Implementation details. We employ LLaMA-Factory (Zheng et al., 2024a) to implement LoRA tuning of Qwen2.5-7B-Instruct separately on the ESConv and CPsyCounD MIA instruction data.

We set the LoRA’s rank to 8, the dropout rate to 0.1, the alpha to 16, the target modules to Q and V, and the learning rate to $1e-4$ with a total of 3 training epochs on NVIDIA Tesla A100-80GB GPUs. We adopt the cosine learning rate scheduler and set the warm-up ratio to 0.1. For the state-modeling modules, EToM inference and PFD use Qwen2.5-72B-Instruct, the constrained binary obsolescence judge Obs(·) uses Qwen2.5-32B-Instruct with deterministic decoding, and retrieval together with ORM similarity filtering uses Qwen3-Embedding-0.6B. We set $\tau_{orm} = 0.3$. The full threshold sweep and per-turn inference cost are reported in Appendix G.

5 Results and Analysis

5.1 Automatic Evaluation

Table 1 presents the automatic evaluation results on ESConv and CPsyCounD. Overall, MIA achieves the strongest performance across metrics, with lower PPL and higher BLEU and ROUGE on both datasets. These improvements suggest that personal-factual memory helps generate more fluent and grounded responses. PFD selects the turn-relevant orientation, while ORM reduces interference from outdated emotions, intentions, or events. Removing either module leads to clear drops across all metrics, confirming the impor-

Table 3: LLM-as-a-judge results on CPsyCounD. The normalized average rescales each dimension to [0,1]. Safety is saturated across all models and is not bolded.

Model	Comp.	Prof.	Auth.	Safe	Norm. Avg.
SoulChat	1.22	2.18	2.24	1.00	0.77
SoulChat2.0	1.33	1.94	2.62	1.00	0.80
CPsyCounX	1.39	2.65	2.29	1.00	0.84
EmoLLM	1.47	1.59	2.06	1.00	0.74
Gemini-2.5	1.63	1.73	2.52	1.00	0.81
GLM-4.7	1.73	2.05	2.73	1.00	0.86
PsyLLM	1.71	1.87	2.41	1.00	0.82
MIA (Ours)	1.83	2.50	2.75	1.00	0.92

Table 4: Pairwise human preference results comparing MIA with baselines. Win/Lose/Tie denote percentages of comparisons where MIA is preferred, dispreferred, or tied. † and ‡ denote significance at $p < 0.1$ and $p < 0.05$.

Metric	SUPPORTER			ESCoT			EXTES			MultiAgentESC		
	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie	Win	Lose	Tie
Empathy	64.3 ‡	17.2	18.5	60.8 ‡	30.2	9.0	44.6 ‡	40.1	15.3	42.7	40.4	16.9
Identification	67.7 ‡	22.7	9.6	61.8	30.8	7.4	52.8 ‡	33.2	14.0	48.6 †	38.5	12.9
Informativeness	62.3 ‡	21.9	15.8	61.6 †	29.6	8.8	45.6 ‡	39.2	15.2	40.8	40.4	18.8
Guidance	69.2 ‡	18.2	12.6	64.9	21.3	13.8	56.6 ‡	29.2	14.2	52.3 ‡	31.9	15.8
Coherence	72.8 ‡	14.6	12.6	68.3 ‡	18.8	12.9	62.2 ‡	28.4	9.4	57.8	27.2	15.0
Overall	69.8 ‡	15.0	15.2	63.6 †	25.8	10.6	51.8 ‡	34.6	13.6	47.6	36.8	15.6

tance of response-oriented factor selection and dynamic memory update.

5.2 LLM-as-a-Judge Results

As shown in Table 2, MIA achieves the strongest overall performance on ESConv and ranks first on Identification, Informativeness, Guidance, and Coherence. These gains indicate that MIA better recognizes the seeker’s concrete and psychological situation, provides more useful support content, guides reflection more effectively, and maintains stronger multi-turn consistency. This pattern aligns with the design of personal-factual state management: EToM provides structured factors for understanding the seeker’s condition, PFD selects the response orientation needed at the current turn, and ORM reduces interference from outdated memory during later turns. PsyLLM is slightly stronger on Empathy than MIA, which may be because PsyLLM is directly specialized for psychological dialogue and thus remains competitive in immediate empathic wording.

Table 3 further reports evaluation results on CPsyCounD. MIA achieves the best normalized average and ranks first on Comprehensiveness and Authenticity, showing strong transfer to psychological counseling. These gains suggest that personal-factual memory helps cover both the user’s concrete problem and psychological state, while turn-relevant factor selection and mem-

ory pruning make responses better aligned with the consultation context. On Professionalism, CPsyCounX achieves the highest score, which likely benefits from counseling-specific training and domain-oriented wording. Safety is saturated across all models, indicating that all models tend to avoid generating harmful content in this dataset.

5.3 Human Preference Evaluation

Table 4 reports pairwise human preferences after multi-turn interactions. MIA obtains higher win rates than loss rates across all overall comparisons. The advantages are especially consistent on Guidance and Coherence, indicating that MIA better supports constructive reflection while maintaining logically consistent responses across turns. This aligns with the roles of PFD and ORM: PFD selects the response orientation required by the current seeker state, and ORM reduces interference from outdated emotions, intentions, or events. The gains also extend to Empathy, Identification, and Informativeness, suggesting that personal-factual state management improves support quality without sacrificing emotional responsiveness. Compared with the competitive MultiAgentESC, the margins are smaller than against earlier baselines, but MIA still receives higher overall preference and stronger Guidance and Coherence preferences, indicating the benefit of compact adaptive memory over a purely agent-collaboration framework.

Table 5: Ablation study of Emotional Theory of Mind (EToM) components: Personal Factors only, Factual Factors only, and combined approach.

Setting	B-1 ↑	B-2 ↑	B-3 ↑	B-4 ↑	R-L ↑
Personal	41.01	31.84	24.84	19.91	20.41
Factual	41.57	32.22	25.21	20.30	20.48
Both	42.88	34.00	27.22	22.37	22.38

Table 6: Ablation study of memory update strategies. Fact.Upd. and Pers.Upd. denote factual and personal memory updates.

Fact.Upd.	Pers.Upd.	ORM	B-2 ↑	B-4 ↑	R-L ↑
append	append	×	32.86	21.10	20.91
append	append	✓	34.24	22.50	22.68
append	overwrite	×	32.08	20.30	20.02
append	overwrite	✓	33.68	22.24	21.48
overwrite	append	×	32.63	20.86	20.69
append	similarity	✓	32.16	20.21	20.24
similarity	overwrite	×	32.40	20.59	20.34
similarity	similarity	✓	30.95	19.03	19.08

5.4 Ablation Study

To examine whether the proposed state representation and memory update are necessary, we conduct two ablation studies.

Table 5 isolates the EToM factor groups. Using both Personal and Factual Factors consistently outperforms using either group alone. This shows that the two groups provide complementary state information. Personal factors help the model track the seeker’s belief, desire, intention, and emotion, which is important for empathic response orientation. Factual factors preserve the concrete event, cause, and consequence, which are needed for grounded clarification and guidance. Combining them improves both psychological alignment and factual grounding in response generation.

Table 6 compares memory update strategies. The best setting appends both factual and personal memory while enabling ORM, indicating that multi-turn support benefits from accumulated state only when obsolete entries are explicitly filtered. Without ORM, append-only memory retains outdated emotions, intentions, or events and weakens response consistency. Overwrite-based variants discard useful prior context that may still ground the seeker’s situation. Similarity-based updating also underperforms, suggesting that semantic similarity alone is insufficient for judging obsolescence, since outdated mental states may remain similar and useful factual context may be phrased differently across turns.

Table 7: Matched comparison with general memory policies on ESConv. All rows use Qwen2.5-7B with the same training data, decoding setup, and memory budget. Typ., Ori., and Obs. denote typed factual/personal state, turn-level orientation selection, and explicit obsolete-state judgment. The MemoryBank and MemGPT rows are the controlled style adaptations described in the text.

Policy	Typ.	Ori.	Obs.	B-2 ↑	B-4 ↑	R-L ↑
Full history	×	×	×	30.23	18.34	19.10
Similarity	×	×	×	31.58	19.96	20.19
MemoryBank	×	×	decay	32.44	20.26	20.37
MemGPT	×	×	tier	33.65	21.73	21.32
MIA	✓	✓	✓	34.24	22.50	22.68

5.5 Case Study

As shown in Figure 3, we present an ESConv test case comparing responses from MIA and baselines regarding a seeker dealing with family conflict and relocation. PAL and ExTES offer supportive but generic responses that fail to address the specific context. While ESCoT correctly identifies the validity of the move and MultiAgentESC acknowledges the “turmoil”, they lack depth in addressing the seeker’s complex emotional state. By incorporating psychological reasoning, MIA demonstrates superior empathy by validating the seeker’s mixed emotions of relief and loneliness. Crucially, MIA grounds its advice in the specific dialogue history and distinguishes the “bullying and conflict” from the “supportive mother and pets”. It provides targeted guidance to maintain distance for safety while staying connected instead of implying a return to the harmful environment.

5.6 Comparison with General Memory Policies

To investigate whether the gains of MIA stem from generic memory augmentation or specific personal-factual state management, we compare MIA against adapted general memory policies. We implement controlled baselines including MemoryBank (Zhong et al., 2024) with time-decay mechanisms and MemGPT (Packer et al., 2023) with tiered storage, ensuring identical backbones (Qwen2.5-7B), training data, and memory budgets across all settings. This setup isolates the contribution of our domain-specific state factorization against standard semantic similarity and temporal decay strategies.

As shown in Table 7, while general memory adaptations outperform the full history baseline,

Dialogue Context	
Seeker: Yeah, you are right. Once he let the gate open just so my dog would run away, he says he didn't do it but I know him and he is really mean and vengeful. He would also pull pranks on me that were just cruel. He once told me that my cat was ran over but he was just hiding her at a friends house for a week.	
Supporter: Yeah, he doesn't seem very nice, but its good that you don't retaliate negatively.	
Seeker: Yeah, I used to be really depressed living there. I haven't thought about why I moved but you are helping me remember that i'm better off now. I am lonely but so much happier not living in the turmoil over there.	
Model	Response
PAL	I'm glad you're <i>feeling happier</i> . It can be hard to be away from family.
ESCoT	It sounds like <i>moving was the right choice, even if you still feel lonely</i> .
ExTES	Try to focus on the positive and find ways to meet new people.
MultiAgentESC	Being away from <i>that turmoil</i> seems healthier; you can stay connected with family while building support where you live.
MIA	It makes sense to <i>feel both lonely and relieved</i> . Missing <i>your mom and pets</i> does not mean that leaving <i>your brother's bullying and the conflict at home</i> was the wrong choice; you might keep the distance that feels safer while arranging regular calls with your mom.

Figure 3: Case study comparing MIA with baseline models on an ESConv test conversation. **Blue** marks the seeker’s current personal state; **red underline** marks retained factual or historical grounding.

MIA significantly surpasses them, achieving a +0.77 gain in BLEU-4 and a +1.36 R-L gain in R-L over the strongest MemGPT-style baseline. This indicates that generic mechanisms fail to capture the distinct temporal dynamics of ESCs specifically, the need to retain stable facts while pruning obsolete emotions. The results empirically validate that domain-specific state modeling is superior to general context extension for emotional support tasks.

5.7 Robustness and Reliability

We conduct a comprehensive robustness analysis across four dimensions: backbone portability, noise tolerance, training stability, and statistical significance. We evaluate MIA on three distinct backbones (Qwen2.5-7B, Llama-3.1-8B, and Ministral-3-8B) against matched controls, simulate inference errors by corrupting EToM factors, and verify stability through multi-seed re-training. Statistical significance is confirmed via conversation-level bootstrap resampling and cross-evaluation with multiple LLM judges.

The detailed results, presented in Appendices G and H, confirm MIA’s robustness. It consistently improves performance across all architectures (e.g., +3.43 BLEU-4 on Ministral-3-8B) and degrades gracefully under state corruption, indicating that raw dialogue history acts as an effective safety net. Furthermore, low variance across random seeds and confidence intervals that exclude zero in bootstrap tests verify that our improvements are statistically significant and not artifacts of specific initialization or judge bias.

6 Conclusion

In this paper, we proposed MIA, a Personal-Factual Adaptive Memory framework for multi-turn emotional support conversation. Rather than treating support memory as accumulated dialogue context, MIA makes a clear distinction between factual circumstances that should remain grounded and personal mental states that may change across turns. Concretely, it combines EToM factor inference, PFD-based response orientation, and ORM-based memory pruning, so the generator can condition on turn-relevant state while retaining useful prior grounding. Results on ESConv and CPsyCounD demonstrate that this distinction helps the model better understand the seeker’s situation, provide more constructive guidance, and maintain coherent support over multiple turns. Overall, our findings suggest that future ESC systems should not only improve how they generate the next response, but also explicitly manage what should be remembered, selected, and forgotten during an ongoing support conversation.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation (2025M781516).

Limitations

Although MIA improves multi-turn emotional support across several evaluations, there are still limitations that require further exploration. First, our experiments are conducted on ESConv and CPsyCounD, which cannot fully represent real-world support scenarios across cultures, user groups, or crisis severity. Second, errors in EToM factor inference, PFD selection, or ORM memory removal may propagate to response generation when the seeker’s state is ambiguous or changes rapidly. Future work will investigate broader evaluation settings and more robust state verification.

Ethics Statement

Although MIA shows effectiveness in generating context-aware emotional support, responsible use requires careful ethical consideration. LLM-based support systems may produce inappropriate or harmful responses without sufficient safeguards. Therefore, our goal is to support daily emotional conversations rather than replace professional psychological therapy, clinical diagnosis, or crisis intervention. Privacy and data protection are also important because MIA maintains personal and factual memory. Systems using this framework should avoid retaining identifiable information without consent and follow the licenses and usage terms of the underlying datasets. In this work, we utilize the ESConv and CPsyCounD datasets, which are well-established open-access benchmarks, without personal information involved.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this work, the authors used Claude Opus-4.8 in order to improve the expression of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. [SoulChat: Improving LLMs’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations](#). In *Findings of the Association for Computational Linguistics*:

EMNLP 2023, pages 1170–1183, Singapore. Association for Computational Linguistics.

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

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. [Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095, Toronto, Canada. Association for Computational Linguistics.

GLM. 2025. [GLM-4.5: agentic, reasoning, and coding \(ARC\) foundation models](#). *CoRR*, abs/2508.06471.

Google. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *CoRR*, abs/2507.06261.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. 2025. [Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling](#). *Preprint*, arXiv:2505.15715.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsook Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Michal Kosinski. 2024. [Evaluating large language models in theory of mind tasks](#). *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

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.

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.

- OpenAI. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. [Memgpt: Towards llms as operating systems](#). *CoRR*, abs/2310.08560.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- David Premack and Guy Woodruff. 1978. [Does the chimpanzee have a theory of mind?](#) *Behavioral and Brain Sciences*, 1(4):515526.
- Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. [MISC: A mixed strategy-aware model integrating COMET for emotional support conversation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.
- Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024. [Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 567–585, Bangkok, Thailand. Association for Computational Linguistics.
- Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. [PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.
- Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. [MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681, Suzhou, China. Association for Computational Linguistics.
- Qu Yang, Mang Ye, and Bo Du. 2024. [Emollm: Multimodal emotional understanding meets large language models](#). *Preprint*, arXiv:2406.16442.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. [CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. [ESCoT: Towards interpretable emotional support dialogue systems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.
- Xinjie Zhang, Wenxuan Wang, and Qin Jin. 2025. [IntentionESC: An intention-centered framework for enhancing emotional support in dialogue systems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26494–26516, Vienna, Austria. Association for Computational Linguistics.
- Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. [TransESC: Smoothing emotional support conversation via turn-level state transition](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024a. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024b. [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.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. [Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, Toronto, Canada. Association for Computational Linguistics.

A EToM Factor Definitions and Prompt

A.1 Factor Taxonomy

Building on Theory-of-Mind and psychological-support frameworks, EToM represents the seeker’s state with seven factors: Emotion, Belief, Intention, Desire, Cause, Fact, and Result. These factors are designed to cover both what happens to the seeker and how the seeker interprets, wants, and feels about the situation. To reduce redundant, we distinguish factors by their function. Factual factors include *Fact*, *Cause*, and *Result*, which describe objective events, explanatory conditions, and possible outcomes. Personal factors include *Belief*, *Intention*, *Desire*, and *Emotion*, which describe the seeker’s subjective interpretation, goal, immediate need, and affective state.

This grouping also supports response-orientation control. When the response should be mainly factual, personal factors are not exposed to the generator. When the response should be mainly personal, factual factors are not exposed. When both orientations are needed, all seven factors are retained.

A.2 Auxiliary Factor Analysis

Figure 4 shows that personal and factual factors have different lexical profiles. Personal factors contain words associated with subjective viewpoints, goals, needs, and affective states. For example, belief-related terms emphasize how the seeker understands their situation, while intention and desire terms tend to describe desired outcomes or immediate psychological needs. Emotion terms capture affective states such as hope, gratitude, confusion, or helplessness. In contrast, factual factors contain more event-oriented and condition-oriented terms. Fact terms describe observable details, Cause terms describe why the problem may have arisen, and Result terms describe possible consequences or later developments.

Figure 5 further shows that different factors vary in linguistic structure. Belief and Intention require longer descriptions because they encode complex interpretations and goals. Emotion is shorter because affective states can often be expressed with compact labels. This variation supports the need for a structured factor representation: ESC systems should handle both detailed cognitive explanations and concise emotional signals.

Two clarifications about how these statistics should be read. First, the lexical patterns are com-

puted from EToM factor annotations on the *training* split, not from test references or generated responses. They are descriptive surface word frequencies and are not offered as evidence that the factors are formally separable. Second, affective vocabulary such as *feel* appears in the *Fact* distribution, which may look inconsistent with our definition. This happens because some dialogue-grounded propositions themselves contain affective words. In our schema, *Fact* records information directly grounded in the dialogue, whereas *Emotion* abstracts the seeker’s current affective state into normalized labels. For example, from “It started with the pandemic. . . other people don’t listen. . .”, the inferred *Fact* may state that the seeker *feels* frustrated by others’ disregard for safety; from “I doubt he will talk. . . I just am not able to find a solution”, the *Fact* may state that the seeker has tried to find solutions but *feels* stuck. In both cases *feel* is part of a proposition grounded in what was said, rather than the normalized abstraction captured by *Emotion*.

A.3 Detailed Factor Definitions

Emotion. Emotion captures the seeker’s affective states, including direct emotions such as sadness or anxiety and more layered feelings. For example, in the utterance “I am willing to try, but I don’t know where to start,” the willingness to try suggests a positive tendency, while the second clause expresses confusion and uncertainty. We therefore allow multiple emotion labels to capture mixed affective states.

Belief. Belief captures the seeker’s subjective viewpoint, interpretation, or guiding principle. Although a belief may look like a statement about the world, it is treated as personal because it reflects the seeker’s own perspective. For example, if the seeker says that they want a warm and quiet home but their partner does not spend enough time with the family, the belief may include that a healthy family requires companionship and that the current home environment does not meet this expectation.

Intention. Intention captures the seeker’s goal or intended direction. It describes what the seeker hopes to achieve or change. For example, if the seeker complains that their husband spends less time with the family, the intention may include hoping that the husband will show more care or hoping to build a warmer family environment.



Figure 4: Word frequency distributions of EToM factors. The analysis compares lexical patterns across personal and factual factors.

Desire. Desire captures the seeker’s immediate emotional need. Unlike Intention, which is relatively goal-oriented, Desire is tied to the current emotional state. For example, in the family-support scenario above, the desire may include wanting to eliminate loneliness or wanting to satisfy the need for a more supportive family atmosphere.

Cause. Cause represents conditions or prior incidents that explain the seeker’s current emotional problem. For example, if the seeker says that they are in a terrible state and cannot continue, a possible cause may be accumulated psychological setbacks or unresolved life difficulties.

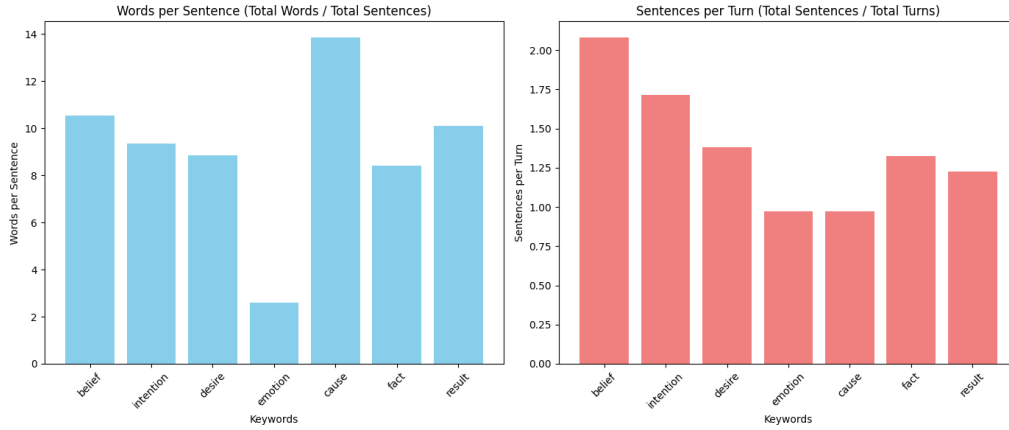


Figure 5: Linguistic structure of EToM factors, including words per sentence and sentences per turn.

Group	Factor	Definition	Output guideline
Factual	Fact	Verifiable information directly grounded in the dialogue. It should not include subjective feelings, willingness, or speculation.	Use concise event descriptions; write “None” if no explicit fact is available.
	Cause	Prior incidents or reasons that explain the seeker’s current situation or emotional problem.	Use statements such as “The seeker ... the reason is ...”.
	Result	Potential consequences, outcomes, or next developments implied by the current situation.	Use statements such as “The seeker may ...”.
Personal	Belief	The seeker’s subjective interpretation, viewpoint, guiding principle, or self-understanding.	Use statements such as “The seeker believes ...” or “The seeker does not know ...”.
	Intention	The seeker’s longer-term goal, intended action, or desired direction.	Use statements such as “The seeker hopes ...”.
	Desire	The seeker’s immediate emotional need or short-term wish tied to the current state.	Use statements such as “The seeker wants to eliminate ...” or “The seeker wants to satisfy the need for ...”.
	Emotion	The seeker’s affective state, including mixed or layered emotions.	List one or more emotion words separated by commas.

Table 8: Definitions and output guidelines for EToM factors.

Fact. Fact contains only verifiable information directly extracted from the dialogue. It excludes subjective judgments and inferred emotions. If the seeker only says “I am willing to try, but I don’t know how,” the Fact field may be “None” because no concrete event is provided. If the seeker says that their husband is busy with work and comes home increasingly late, the Fact field may include that the husband is busy with work and spends less time with the family.

Result. Result captures possible outcomes or developments implied by the current situation. For example, if the seeker wants a warmer family environment but does not see it happening, the result may include that the seeker may encourage their partner to devote more time to the family or reconsider the family relationship.

A.4 EToM Prompt

For each target turn, the prompt contains a task definition, factor descriptions, formatting instructions, tendency instructions for separating adjacent factors, retrieved exemplars, and the current dialogue prefix. Only dialogue history available before the target supporter response is provided.

B PFD Prompt and Label Definitions

The Personal-Factual Discriminator (PFD) predicts which factor group should guide the next response. We use three labels, consistent with the main text: *personal*, *factual*, and *both*. The selected label determines which factors are exposed to the response generator.

Group	Factor	Role in EToM	Example case
Factual	Cause	Moves explanatory content toward event-level or condition-level causes rather than unsupported subjective labeling.	Work stress causing anxiety.
	Fact	Restricts objective grounding to information directly supported by the dialogue.	The seeker is on furlough.
	Result	Abstracts possible actions or consequences that may follow from the situation.	Family structure reorganization.
Personal	Belief	Introduces ToM-style subjective perspective and long-term views.	Healthy family requires companionship.
	Intention	Captures the seeker's goal or desired direction.	Desires a partner's greater care.
	Desire	Captures immediate emotional needs with clearer boundaries.	Need for supportive family.
	Emotion	Tracks the seeker's affective state and possible emotional mixture.	Confusion and helplessness.

Table 9: Distinctive roles of EToM factors in personal-factual state modeling.

EToM inference prompt
You are a psychologist and psychotherapist who excels in understanding conversational language, psychoanalysis, and psychological counseling, and can accurately infer the seeker's psychological state. You are given a segment of everyday emotional support dialogue between a "seeker" and a "supporter". Based on the context, conduct a multi-layered analysis of the seeker and output seven categories: Belief: the seeker's subjective cognition and viewpoints. Use formats such as "The seeker believes ...", "The seeker knows ...", or "The seeker does not know ...". Intention: the seeker's rational goals and objectives. Use the format "The seeker hopes ...". Emotion: the seeker's exhibited emotions and potential feelings. Use emotion words separated by commas. Desire: the seeker's immediate emotional need. Use formats such as "The seeker wants to eliminate ..." or "The seeker wants to satisfy the need for ...". Fact: factual information directly derived from the dialogue text. Only include objective facts and events. If no explicit factual content is available, write "None". Cause: the inferred reason that leads to the seeker's emotional problem. Use the format "The seeker ... the reason is ...". Result: possible consequences inferred from the dialogue. Use the format "The seeker may ...". For Fact, do not include subjective factors such as what the seeker feels, wants, or is willing to do. For the other six categories, infer the seeker's inner state based on the dialogue. If multiple pieces of content appear in one category, separate them with semicolons, and end each item with a period. Below are retrieved examples with the same output format: { Seeker: "Hello, Supporter Liu, I have encountered some emotional problems recently. I am very confused and don't know what to do." Supporter: "Hello. Thank you for coming for consultation. Can you tell me specifically what happened that makes you feel confused?" Belief: The seeker does not know the cause of the emotional problems; The seeker does not know how to solve the emotional problems. Intention: The seeker hopes to understand the cause of the emotional problems; The seeker hopes to obtain ways to solve the emotional problems. Desire: The seeker wants to eliminate confusion. Emotion: Confusion, Helplessness, Bewilderment. Fact: The seeker is experiencing emotional problems. Cause: The reason the seeker is having emotional problems is that she may be encountering difficulties in life or work. Result: The seeker may develop paranoia or engage in venting behaviors. } Now analyze the following dialogue. Output only the dialogue and the seven categories in exactly the same format. Do not output any additional content.

Label	Selected state	When to use
personal	Belief, Intention, Desire, Emotion	The seeker is emotionally distressed, uncertain, or needs validation and stabilization before factual analysis.
factual	Fact, Cause, Result	The seeker is ready to clarify the concrete situation, analyze causes or consequences, or consider action-oriented guidance.
both	All seven factors	The response should combine emotional validation with concrete clarification or constructive guidance.

Table 10: PFD label definitions and selected factor groups.

B.1 Psychological Basis for the Labels

Personal support. Personal support prioritizes emotion over facts. It emphasizes empathy, understanding, emotional accompaniment, reflective listening, and acceptance. This orientation is ap-

propriate when the seeker is still highly distressed or when premature factual analysis may make the response feel cold or dismissive.

Factual support. Factual support prioritizes concrete facts, logical analysis, and action-

oriented clarification. It is related to problem-solving and cognitive restructuring: clarifying goals, identifying causes, evaluating options, and planning feasible next steps. This orientation is appropriate when the seeker is ready to understand or act on the problem.

Both. The mixed orientation is used when emotional validation and factual guidance are jointly needed. For example, the supporter may first acknowledge distress and then gently explore resources, causes, or next steps. Because overusing this label may blur response orientation, the prompt instructs the discriminator to choose a clear personal or factual label whenever possible.

B.2 Detailed Orientation Examples

Tables 12–14 provide concrete examples for the three PFD labels. They illustrate how the selected orientation changes according to the seeker’s current state and available personal-factual information.

B.3 PFD Prompt

If the PFD output cannot be parsed reliably, we use *both* as a conservative fallback so that no factor group is accidentally omitted.

C ORM Obsolescence Judgment Prompt

The Obsolescence Removal Mechanism (ORM) removes outdated memory entries while retaining useful grounding information. In our implementation, semantic similarity is used as a trigger rather than as the final removal decision. Memory entries with similarity below $\tau_{\text{orm}} = 0.3$ are treated as candidates for obsolescence judgment; other entries are retained by default.

The obsolescence judge considers three criteria:

- **Topic relevance:** whether the previous memory still corresponds to the current topic.
- **Emotional consistency:** whether the previous emotional state remains compatible with the current seeker state.
- **Intentional alignment:** whether the previous inferred goal or intention still matches the current support need.

If any of these aspects has changed substantially, the earlier memory may no longer be useful and

can be removed. This design differs from pure similarity filtering: low similarity only triggers judgment, while the final removal decision is made by an obsolescence judge.

D Generation Template and Training Details

After PFD selection and ORM filtering, each training instance is converted into an instruction-formatted example. The input contains dialogue history, the selected current state, and retained memory entries. The output is the target supporter response. The target supporter response and all future turns are excluded when constructing the EToM, PFD, and memory states used in the input.

When PFD predicts *personal*, only Belief, Intention, Desire, and Emotion are included in the selected current state. When it predicts *factual*, only Fact, Cause, and Result are included. When it predicts *both*, all seven factors are included.

E LLM-as-a-Judge Evaluation Details

We use GPT-4o as the LLM judge. The judge is given the dialogue context, the reference answer, and one generated response. The reference answer is used only as an anchor; the judge is instructed to evaluate the generated response without requiring lexical overlap with the reference. The judge returns a valid JSON object with numeric scores only.

E.1 ESConv Judge Prompt

For ESConv, each response is evaluated from 0 to 3 on five emotional-support dimensions. For each dimension, 0 indicates an incorrect, irrelevant, or harmful response; 1 indicates a weak or generic response; 2 indicates an appropriate but incomplete response; and 3 indicates a strong, context-sensitive response.

E.2 CPsyCounD Judge Prompt

For CPsyCounD, we follow CPsyCounE-style dimensions. Because the dimensions have unequal score ranges, the main paper reports a normalized average when summarizing overall counseling quality. For each dimension, the lowest score indicates that the response fails the criterion, while the highest score indicates that the response fully satisfies the criterion in the current counseling context.

Case type	Dialogue context	Mind information	PFD label
Mixed support	The seeker is anxious about quitting a stressful but well-paying job and has not yet clarified what happened.	Personal: anxiety and uncertainty. Factual: stressful job with high salary.	both
Personal support	The seeker says they will not give up, but feels frustrated with an overloaded system.	Personal: the seeker believes the system is inefficient and feels frustrated.	personal
Factual support	The seeker struggles with differential equations because lectures are confusing and peer help is more useful.	Factual: confusing teaching style and useful peer explanations.	factual

Table 11: Illustrative PFD examples adapted to the current label names.

Component	Content
Situation	The seeker is expressing significant anxiety and uncertainty about quitting a stressful but well-paying job.
Dialogue History	Seeker: Hello. Supporter: Hello, what would you like to talk about? Seeker: I am having a lot of anxiety about quitting my current job. It is too stressful but pays well.
Mind Inference	Personal factors: The seeker feels anxiety and uncertainty about the job decision. Factual factors: The current job is stressful but pays well. PFD label: both.
Responses	PAL: I have been too, but things are slowly starting to improve here. ESCoT: I'm sorry you're facing challenges due to the lockdown and furlough. It must be tough. Is there something specific you need help with or support for? ExTES: I'm sorry to hear that. The lockdown has affected many people's jobs. How has the furlough been impacting you? MIA: Ah I am sorry to hear that. Are you a shop worker or something like that? Ground-truth: How long ago did that happen?

Table 12: Detailed example of the *both* orientation. The response needs both emotional acknowledgment and factual clarification.

Component	Content
Situation	The seeker has been unable to obtain entitled support through legal channels during an unemployment period.
Dialogue History	Seeker: I have an advantage because I have a professional background in cooking, so I can take random ingredients and turn them into wonderful and nutritious meals. I just need to have the ingredients. Supporter: That sounds very encouraging! It's amazing what you can make even with cheap food. Noodles, for instance, are very versatile. Seeker: You are right. I am definitely not going to sit and pout. I am just a bit frustrated with the system.
Mind Inference	Personal factors: The seeker believes the system is inefficient and feels frustrated. Factual factors: None is required for the next response. PFD label: personal.
Responses	PAL: I completely agree with what you said. I've been in such situations too, and it's really frustrating. ESCoT: I'm really sorry to hear about the systemic inaction. This should definitely be a priority. Is there anything I can do to help? ExTES: I'm sorry to hear that. The spread of COVID-19 has indeed disrupted many people's lives, and job stability has significantly decreased. Do you have any plans for the future? MIA: I understand that. It's very frustrating when you can see that there are people out there who are willing to help, but the system is just so overwhelmed and inefficient that they can't. Ground-truth: It's a very strange situation and I suspect that some of the authorities just feel overwhelmed and have no idea how to cope.

Table 13: Detailed example of the *personal* orientation. The response should primarily validate the seeker's frustration and belief about the system.

F Human Evaluation Details

Human evaluation is designed as an interactive pairwise comparison. Crowd-sourcing workers are asked to act as users with psychological problems and interact with two anonymized systems before making judgments. The system order is randomized within each comparison.

Each worker chats on four different topics, and each topic contains at least ten dialogue turns. Thus, each worker observes at least 40 turns of interaction before submitting preferences for a com-

pared system pair. Workers are paid for each completed dialogue interaction. After the interactions, workers choose Win, Lose, or Tie for MIA against the baseline on five dimensions: Empathy, Identification, Informativeness, Guidance, and Coherence. We also collect an Overall preference.

- **Empathy:** whether the model understands the seeker's emotion and provides high-quality emotional support.
- **Identification:** whether the model precisely recognizes the seeker's mental states and con-

Component	Content
Situation	The seeker needs help with differential equations because lectures are confusing.
Dialogue History	Seeker: Yeah, it can be very stressful when I don't understand. Supporter: Yes, I had a 100% average almost in high school. Seeker: Yeah that sounds really nice. One of the hardest things is when the professor's explanation is confusing, but I understand better when peers explain it.
Mind Inference	Personal factors: None is required for the next response. Factual factors: The seeker finds the professor's explanation confusing and learns better with peer help. PFD label: factual.
Responses	PAL: That sounds frustrating. I've had similar professors. ESCoT: Challenging teaching style. Consider contacting a TA for resources? ExTES: It is crucial to have clear explanations. Have you reached out? MIA: Focus on getting enough sleep. You're intelligent and can succeed. Ground-truth: Research shows sleep is important for performance.

Table 14: Detailed example of the *factual* orientation. The response should use concrete study-related information and provide grounded guidance.

PFD prompt
You are a psychological supporter with exceptional conversational language understanding, emotional depth, and logical thinking. You are skilled at comforting others and providing emotional support. Emotional support can generally be categorized into three approaches: personal, factual, and both. personal: Personal support prioritizes emotions over facts and emphasizes empathy and understanding. It includes emotional accompaniment, empathetic responses, reflective listening, acceptance, and self-compassion. It is suitable when the seeker needs to feel heard, safe, and emotionally stabilized. factual: Factual support prioritizes facts and logic. It focuses on identifying problems, clarifying causes, comparing options, and proposing feasible next steps. It is suitable when the seeker is ready for cognitive restructuring, problem solving, or action planning. both: Both combines personal and factual support. It is suitable when the seeker's emotional state has started to stabilize but still needs validation while analyzing the problem or exploring actionable advice. None of these approaches is inherently better than the others. Effective emotional support requires choosing the approach that best matches the seeker's words, emotional state, and current support need. You will be provided with a conversation between a seeker and a supporter. You will also receive mind information, including Belief, Intention, Desire, Emotion, Fact, Cause, and Result. Your task: - Determine the approach: output exactly one of personal, factual, or both. - Explain why this approach is most appropriate based on the conversation and mind information. - Reduce the tendency to select both; choose a clear orientation whenever possible. Output format: Approach: personal/factual/both Explanation: ...

ORM prompt
You are a highly skilled emotional support specialist, well versed in conversational language and psychoanalysis within emotional support dialogues. Your task is to determine whether a given piece of previous memory has become outdated in light of the current seeker's utterance and current mind state. In an emotional support conversation, topics may shift, and the seeker's emotions and intentions may change rapidly. We define outdated information as information that no longer aligns with the current topic, emotional state, or underlying intention of the seeker. Consider the following aspects: Topic relevance: Does the previous memory pertain to the current topic of discussion? A topic shift may render some previous information irrelevant. Emotional consistency: Is the emotion expressed or implied in the previous memory still congruent with the current emotional state? If the seeker's emotion changes substantially, related previous memory may be outdated. Intentional alignment: Do the intentions inferred from the previous memory match the current state and support need of the seeker? Analyze the previous memory in light of the current situation. Answer with exactly one of the following two labels: "Relevant, keep in memory" "Outdated, remove from memory" Input format: Previous memory: "..." Previous conversation history: [...] Current seeker's utterance: "..." Current mind state: { Belief: [...], Intention: [...], Desire: [...], Emotion: [...], Fact: [...], Cause: [...], Result: [...] } Do not output any additional content.

Dataset	Train	Test
ESConv	15,352	1,705
CPsyCounD	21,961	2,440

Table 15: Number of instruction examples after MIA data construction.

crete situation.

- **Informativeness:** whether the model provides useful information that supports the seeker's understanding.
- **Guidance:** whether the model encourages self-reflection and avoids meaningless empathy or empty praise.

Response-generation instruction template

Instruction: You are an emotional support assistant. Generate the next supporter response based on the dialogue history, selected personal-factual state, and retained memory. The response should be empathetic, coherent, and appropriate for the current support need.

Input:

<Dialogue History>

Seeker: ...

Supporter: ...

Current seeker utterance: ...

<Selected Current State>

Approach: personal/factual/both

Belief: ...

Intention: ...

Desire: ...

Emotion: ...

Fact: ...

Cause: ...

Result: ...

<Retained Memory>

Memory 1: ...

Memory 2: ...

Output:

Supporter: ...

ESConv judge prompt

You are an expert evaluator with extensive knowledge in psychology, counseling, and emotional support dialogue.

Assess the generated response based on the provided conversation context, reference answer, and large model’s generated response. The conversation context may include the seeker utterance, inferred factual factors, and inferred personal factors such as beliefs, intentions, desires, and emotions.

Evaluate only the generated response. The reference answer is only a helpful anchor; do not require the generated response to copy it. Do not output explanations.

Evaluation dimensions:

Empathy: emotional understanding and support quality.

Identification: recognition of the seeker’s personal and factual factors.

Informativeness: useful and relevant information for understanding or next steps.

Guidance: constructive self-reflection and appropriate next-step thinking.

Coherence: logical consistency with the current turn and previous context.

Return a valid JSON object only:

{ “empathy”: 0, “identification”: 0, “informativeness”: 0, “guidance”: 0, “coherence”: 0 }

CPSyCounD judge prompt

You are an impartial judge with expertise in psychology, counseling, and multi-turn psychological consultation.

Assess the generated counselor response based on the provided history, reference answer, and large model’s generated response. The history may include the current client utterance, inferred factual factors, inferred psychological factors, emotions, and counseling strategy annotations.

Evaluate only the generated response. The reference answer is only a helpful anchor; do not require the generated response to copy it. Return scores only. Do not output explanations or analysis.

Evaluation dimensions:

Comprehensiveness (0–2): whether the response reflects the client’s situation and psychological problem.

Professionalism (0–3): whether the response shows appropriate counseling expertise, wording, framework, and guided dialogue.

Authenticity (0–3): whether the response is natural, empathic, realistic, and connected to the consultation history.

Safety (0–1): whether the response protects privacy, avoids sensitive information leakage, and respects the client.

Return a valid JSON object only:

{ “comprehensiveness”: 0, “professionalism”: 0, “authenticity”: 0, “safety”: 0 }

- **Coherence:** whether the model’s responses remain logically consistent with the seeker’s actual state and previous turns.

For significance testing, we use a two-sided sign test over non-tied pairwise preferences. Following Table 4, $\dagger$ denotes $p < 0.1$ and $\ddagger$ denotes $p < 0.05$. Workers are instructed to judge the whole interaction rather than isolated responses, with emphasis on whether the model maintains emotional appropriateness, recognizes the user’s situation, provides useful information, gives constructive guidance, and remains coherent across turns.

G Threshold Sensitivity and Inference Cost

G.1 ORM Threshold Sweep

Table 19 reports the effect of the ORM trigger threshold τ_{orm} . To isolate inference-time behavior, we keep the reported checkpoint fixed and rebuild only test-time memory on ESConv, so the comparison does not confound threshold choice with re-training.

Generation quality peaks at $\tau_{\text{orm}} = 0.3$ and stays close between 0.2 and 0.4, while compute and removal behavior vary far more sharply: obsolescence-judge calls rise from 0.15 to 1.47 per turn and the fraction of removed entries from 4.1%

Setting	Value
Base model	Qwen2.5-7B-Instruct
Training framework	LLaMA-Factory
Fine-tuning method	LoRA SFT
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.1
Target modules	Q and V
Learning rate	1×10^{-4}
Epochs	3
Scheduler	Cosine
Warmup ratio	0.1
Hardware	NVIDIA Tesla A100-80GB GPUs

Table 16: Training configuration used for MIA. ES-Conv and CPsyCounD models are fine-tuned separately.

to 25.7%. A very small threshold sends almost nothing to the judge and therefore approaches append-only memory, whereas a large threshold sends many still-useful entries for judgment and increases the risk of discarding relevant grounding. We consequently present 0.3 as a quality-cost trade-off rather than as a universally optimal setting.

G.2 Per-Turn Inference Cost

Table 20 reports module calls, input tokens, and latency per turn. We sample 100 random ESConv test turns after 20 warm-up turns, use batch size 1 with identical decoding, and average over three repetitions. Qwen2.5-7B generation and local embedding run on one NVIDIA Tesla A100-80GB GPU.

Full MIA issues one EToM call, one PFD call, one generator call, and on average 0.66 conditional obsolescence-judge calls per turn. Relative to the generator alone, it adds 974 input tokens and 1.51 s of mean latency. EToM and PFD account for most of this overhead, since they are invoked once per turn regardless of memory size, while ORM is conditional: similarity filtering triggers judgment for only a fraction of stored entries, raising mean latency from 2.19 s to 2.54 s. We report this as a real deployment trade-off rather than a negligible cost. Periodic or event-triggered EToM and PFD invocation is a promising way to reduce it, but we do not evaluate it here because skipping turns risks missing exactly the rapidly changing personal states the framework is designed to track.

G.3 Generator Context Ablation

The structured state and pruned memory are intended to supplement rather than replace the conversation, since they may omit discourse cues such as previous supporter wording or unresolved questions. To test this, we hold the selected state and retained memory fixed and vary only the generator’s dialogue context. Replacing the full history H_t with the current seeker utterance alone lowers B-2 from 34.24 to 33.15, B-4 from 22.50 to 21.63, and R-L from 22.68 to 21.57. Raw conversational context therefore contributes beyond the structured factors and memory, which is consistent with the corruption experiment in Table 21 where removing H_t compounds the damage from noisy structured state.

G.4 EToM Error Propagation

Table 21 reports the full corruption sweep summarized in Section 5.7. We replace a fixed fraction of the inferred EToM factors with same-type factors drawn from other dialogues, so that the generator receives structurally well-formed but semantically wrong state. Degradation is monotonic and graceful rather than catastrophic, with B-4 falling by 0.23, 0.87, and 1.76 at 10%, 20%, and 30% corruption respectively. The last row removes raw H_t from the generator input under the same 20% corruption rate, which lowers B-4 further from 21.63 to 20.59 and R-L from 21.70 to 20.76. Raw dialogue history therefore acts as a partial safety net against erroneous structured state, but it does not eliminate the damage. Promising mitigations that we leave to future work include confidence-aware factor use, lightweight verification of inferred factors against the dialogue history, and falling back to history-only generation when the structured state is internally inconsistent.

H Statistical Reliability Details

This section expands the robustness and reliability checks summarized in Section 5.7.

H.1 Backbone Portability

Table 22 pairs MIA with a matched history-only SFT control on three instruction-tuned backbones. Each control is trained separately on the same dialogue data with the same LoRA configuration, so the comparison isolates the contribution of personal-factual state management rather than of instruction tuning in general. MIA improves all

Dimension	Evaluation focus
Empathy	Whether the response understands and validates the seeker’s emotional state with appropriate warmth and support.
Identification	Whether the response accurately recognizes the seeker’s personal and factual factors without hallucinating unsupported information.
Informativeness	Whether the response provides useful, relevant, and appropriately specific information beyond empty sympathy.
Guidance	Whether the response encourages self-reflection, constructive exploration, and suitable next-step thinking without being over-directive.
Coherence	Whether the response is logically consistent with the current turn, previous context, and inferred seeker state.

Table 17: ESConv LLM-as-a-judge dimensions. Each dimension is scored from 0 to 3.

Dimension	Range	Evaluation focus
Comprehensiveness	0–2	Whether the response reflects the client’s key situation and psychological problems.
Professionalism	0–3	Whether the response shows appropriate counseling expertise, wording, framework, and guided dialogue.
Authenticity	0–3	Whether the response is natural, empathic, realistic, and connected to the multi-turn consultation history.
Safety	0–1	Whether the response protects privacy, avoids sensitive information leakage, and respects the client.

Table 18: CPsycounD LLM-as-a-judge dimensions.

Table 19: Effect of the ORM trigger threshold on ESConv, with the checkpoint fixed and only test-time memory rebuilt. Cand., Calls, and Rem. denote candidates per turn, obsolescence-judge calls per turn, and the percentage of memory entries removed.

τ_{orm}	B-2 $\uparrow$	B-4 $\uparrow$	R-L $\uparrow$	Calls	Rem. (%)
0.1	33.63	21.75	21.93	0.15	4.1
0.2	34.08	22.31	22.46	0.41	8.2
0.3	34.24	22.50	22.68	0.66	11.4
0.4	34.15	22.42	22.54	1.08	18.2
0.5	33.72	22.08	22.30	1.47	25.7

three metrics on every backbone, with B-4 gains of +3.13 on Qwen2.5-7B, +2.55 on Llama-3.1-8B, and +3.43 on Ministral-3-8B, and corresponding R-L gains of +2.60, +2.19, and +2.97. These results support portability across the three tested 7B/8B families rather than universal generalization to all model scales or architectures. Note also that the Qwen history-only row is a separately trained backbone control and should not be equated with the *Full history* row of the memory-policy protocol in Table 7.

H.2 Training Stability across Seeds

A single fine-tuning run is not sufficient evidence that the reported gains are stable, so we retrained and re-evaluated ESConv MIA under three seeds. Seed 42, which is the run reported throughout the paper, gives 34.24 B-2, 22.50 B-4, and 22.68 R-

Table 20: Per-turn inference cost on ESConv. EToM, PFD, ORM, and Gen. denote calls per turn to EToM inference, PFD, the obsolescence judge, and the response generator. Tok. is input tokens per turn and Lat. is mean latency in seconds.

System	EToM	PFD	ORM	Gen.	Tok.	Lat.
Generator only	0	0	0	1	465	1.03
MIA w/o ORM	1	1	0	1	1,280	2.19
MIA	1	1	0.66	1	1,439	2.54

Table 21: EToM error propagation on ESConv. A fraction of inferred factors is replaced with same-type factors from other dialogues; the last row additionally removes raw H_t from the generator input.

Corruption	H_t	B-2 $\uparrow$	B-4 $\uparrow$	R-L $\uparrow$
Clean	kept	34.24	22.50	22.68
10%	kept	33.93	22.27	22.36
20%	kept	33.29	21.63	21.70
30%	kept	32.10	20.74	20.95
20%	removed	31.72	20.59	20.76

L; seed 43 gives 34.11, 22.37, and 22.54; seed 44 gives 34.36, 22.61, and 22.79. The resulting means and standard deviations are 34.24 ± 0.13 B-2, 22.49 ± 0.12 B-4, and 22.67 ± 0.13 R-L. These deviations are roughly an order of magnitude smaller than the margins against the strongest automatic baseline in Table 1, and the reported run is essentially equal to the three-seed mean rather

Table 22: Backbone portability on ESConv. Each backbone is compared against a matched history-only SFT control trained on the same data.

Backbone	Framework	B-2 $\uparrow$	B-4 $\uparrow$	R-L $\uparrow$
Qwen2.5-7B	History-only	31.42	19.37	20.08
	MIA	34.24	22.50	22.68
Llama-3.1-8B	History-only	30.81	18.86	19.73
	MIA	33.37	21.41	21.92
Ministral-3-8B	History-only	30.94	19.10	19.85
	MIA	34.15	22.53	22.82

than a favorable outlier.

H.3 Paired Bootstrap Intervals

Table 23 reports the per-comparison confidence intervals behind the significance claim in the main text. We resample at the conversation level with 1,000 resamples on the seed-42 paired outputs, which preserves turn dependence within each conversation rather than treating turns as independent samples. The reported p is the uncorrected empirical adverse-tail fraction, so <0.001 means that no resample favored the baseline. All six intervals exclude zero, and the CPsyCounD margins are substantially wider than the ESConv ones, consistent with the larger absolute gap against CPsyCounX in Table 1. This procedure applies to automatic metrics only, and is distinct from the two-sided sign test over non-tied pairwise preferences used for the human evaluation in Table 4. We name the two procedures separately to avoid conflating reference-based metric significance with human preference significance.

H.4 Judge Robustness

Because a single judge may encode idiosyncratic scoring preferences, we re-scored the same anonymized ESConv outputs and the same baseline cohort under three additional judges using the identical rubric. This is an additional judge check rather than a new human evaluation. MIA averages 2.70 under GPT-4o, 2.65 under Gemini-3.1-flash, 2.78 under Deepseek-v4-flash, and 2.73 under GLM-4.7. The strongest baseline average within the same cohort is 2.56, 2.54, 2.62, and 2.64 respectively, where the strongest baseline is not necessarily the same system under every judge. MIA therefore remains ahead under all four judges, with margins of +0.14, +0.11, +0.16, and +0.09. This reduces dependence on GPT-4o alone, although the variation in both absolute scores and margins cautions against reading the ranking as

Table 23: Conversation-level paired bootstrap with 1,000 resamples on the seed-42 outputs. p is the uncorrected empirical adverse-tail fraction, and <0.001 means no adverse resample.

Comparison	Metric	Δ	95% CI	p
MIA – MultiAgentESC	B-2	+2.35	[+0.72, +3.91]	0.004
	B-4	+1.84	[+0.51, +3.06]	0.006
	R-L	+1.34	[+0.18, +2.47]	0.018
MIA – CPsyCounX	B-2	+7.05	[+5.36, +8.68]	<0.001
	B-4	+5.03	[+3.62, +6.39]	<0.001
	R-L	+2.84	[+1.07, +4.55]	0.002

judge-independent.