Beyond Retrieval: Incorporating Personality and Emotion into RAG-Based Conversational Agents

Qijian Zheng[1], Min Gao[1], Zhijun Yuan[2], Qin Su[2], Guanghui Zhou[2], Qiang Duan[3], Xiaoming Fu[4], and Yang Chen[1] ✉

[1] Shanghai Key Laboratory for Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University, China
[2] Seres Auto Co., Ltd., China
[3] Pennsylvania State University, Abington, PA, USA
[4] Institute of Computer Science, University of Göttingen, Germany
{qjzheng25, mgao21}@m.fudan.edu.cn, {zhijun.yuan, qin.su, guanghui.zhou}@seres.cn, qduan@psu.edu, fu@cs.uni-goettingen.de, chenyang@fudan.edu.cn

Abstract. Large language model (LLM)-based conversational agents are increasingly deployed in socially embedded scenarios, including personal assistants and intelligent vehicle cockpits. In long-term dialogue, agents need to retrieve prior context and adapt responses to users' evolving preferences and affective states. Existing retrieval-augmented generation (RAG)-based dialogue agents mainly improve memory access, but often do not explicitly model the user's current conversational state. To address this limitation, we propose Psy4RAG, a plug-and-play RAG module that tracks inferred Big Five personality traits and short-term emotions in the Pleasure-Arousal-Dominance (PAD) space. The resulting psychological-state cues are verbalized and used to guide response generation without modifying the underlying retriever or generator. Experiments on MSC and LoCoMo datasets show that Psy4RAG improves response quality on average, especially on LLM-based coherence, consistency, and relevance metrics. On LoCoMo, Psy4RAG obtains average gains of G-con +7.29%, G-coh +7.23%, and G-rel +6.90%. These results suggest that inferred psychological cues can help RAG-generated responses better fit the emotional context across dialogue turns.

Keywords: Retrieval-augmented generation model · Social computing · Big Five personality traits · PAD emotional states · Psychological state tracking · Long-term dialogue generation

1 Introduction

LLM-based conversational agents [4,9] are widely used in high-impact social interaction scenarios such as intelligent vehicle cockpits where their responses can shape the daily decisions, emotional experiences, and long-term engagement of users. These applications require agents to sustain interaction across

repeated encounters, rather than generate isolated responses within a single exchange [30]. As the interaction unfolds, the user's psychological state, including emotional state, personality-related preferences, and interpersonal expectations, may change with the dialogue context and should guide the framing of agent responses. Therefore, a conversational agent needs to account for user-specific information, including prior experiences, stable preferences, and changing conversational states, rather than relying only on the immediate context. RAG [11] provides a practical foundation for dialogue systems that need to access information beyond the current context window, which is especially important in multi-session conversations.

However, improving retrieval only makes past information more accessible. It does not ensure that the model can use that information appropriately in long-term dialogue. The same retrieved evidence may require different expression styles depending on the user's current conversational state, as the user's goals, emotions, and interpersonal expectations evolve across dialogue turns. Existing long-term dialogue RAG methods mainly improve access to past information through stronger memory construction and retrieval, such as reflective memory management [24], dynamic historical retrieval [31], and structured retrieval [17,6]. While these methods improve long-range recall, they do not sufficiently model the user's current conversational state or use it to guide response generation.

This gap limits the ability of dialogue agents to maintain appropriate response styles in extended interactions. In an in-vehicle dialogue, for instance, the agent may retrieve the same information about the driver's preferences, but it should express that information differently depending on whether the driver is relaxed, stressed, or navigating a demanding traffic situation. Yet many systems still rely on static persona descriptions or a small number of prompt examples to regulate response style. Prior work on emotion-aware retrieval [8] and affective memory management [16] shows that affective information can improve dialogue quality, empathy, and personalization. These findings suggest that long-term dialogue RAG should model user and interaction states dynamically, rather than relying only on retrieved content or fixed style specifications.

Motivated by this limitation, we propose Psy4RAG, a long-term dialogue RAG framework that tracks psychological states on both the user and agent sides to better align RAG-generated responses with personalized user needs. Psy4RAG models the evolving psychological state of the user as a time-varying process and represents it using two complementary psychological theories: Big Five personality traits [5] for stable personality tendencies and PAD emotion space [19] for short-term affective states. The inferred state signals are continuously updated during long-term dialogue and used to guide response generation. By treating user-side psychological states and agent-side response states as dynamic latent variables, Psy4RAG provides explicit control signals for maintaining personalized and context-appropriate response generation in long-term dialogue. Experiments on MSC and LoCoMo show that Psy4RAG consistently outperforms strong baselines in long-term dialogue generation.

Our contributions are threefold:

- We propose Psy4RAG, an RAG-based dialogue system that tracks the user's psychological states during long-term dialogue generation. By continuously updating psychological states of both user and agent, Psy4RAG improves the LLM-evaluated coherence and emotional plausibility of generated responses.
- We evaluate Psy4RAG on two representative long-term dialogue benchmarks, MSC and LoCoMo. Experimental results show that Psy4RAG improves over strong baselines on average in lexical overlap and LLM-based consistency metrics, suggesting that it better aligns with emotional context across dialogue turns.
- We conduct an in-depth analysis of psychological state tracking in Psy4RAG through case studies and ablation experiments. This analysis disentangles the effects of user-side and agent-side psychological state modeling, showing how psychological state tracking contributes to response quality under automatic evaluation.

2 Related Work

In this section, we first review existing representative studies of retrieval-augmented approaches for dialogue generation in Section 2.1 and then review psychological state tracking in Section 2.2, with a focus on how affective states and personality traits are identified and incorporated into dialogue modeling.

2.1 Retrieval-Augmented Generation in Dialogue Generation

In long-term dialogue generation, retrieval-augmented approaches address the core challenge of memory construction and access. Early studies showed that long-term dialogue quality could benefit from explicit memory accumulation and recursive summarization [26,33]. More recent benchmarks have further clarified that this problem remains open at realistic scales. Maharana et al. [18] introduced the LoCoMo dataset of very long conversations and provided an evaluation benchmark for long-term memory across summarization and dialogue generation. Xu et al. [28] proposed LongMemEval, which studied interactive long-term memory for chat assistants. Related studies have shown that multi-turn conversational RAG remains difficult when retrieval, grounding, and response generation must operate over extended histories [3,10].

Recent work has approached this problem mainly by improving how memory is stored and retrieved. Studies on recursive summarization and memory-based long-term storage showed that explicit memory representations improve dialogue continuity beyond what can be obtained from a flat prompt history [33]. Subsequent work refined the memory unit by introducing reflective updates, constructing segment-level memories, and denoising them before retrieval [24,20]. Other studies have focused on how stored histories should be searched. Existing methods have emphasized dynamic historical context, hierarchical temporal

consolidation, or event-centered retrieval, showing that temporal organization is important for effective memory access [31]. A related direction has examined whether richer retrieval structures can improve access to distant context. Prior studies have improved long-term memory retrieval by modeling higher-order dependencies with hypergraph structures or by using associative memory mechanisms inspired by hippocampal indexing [17,6]. However, these studies generally pay limited attention to the evolving psychological states of the user and the agent, and to how such state should influence both retrieval and response generation.

2.2 Psychological State Tracking and Affect-Aware Dialogue

Research on dialogue generation has increasingly drawn on psychological modeling, because speaker style is influenced by both stable personal traits and short-term affective variation. The Big Five personality is commonly used as a framework for representing a relatively stable personality structure [5]. It organizes the personality into five broad dimensions, including Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience, and has been widely adopted in studies of behavior and user profiling. These dimensions provide an interpretable description of long-term stylistic tendencies, such as whether a speaker prefers direct or elaborated expressions, restrained or emotionally expressive language, and exploratory or structured interactions. Complementary to trait-based modeling, PAD has been used to represent short-term affective states through three continuous dimensions: pleasure, arousal, and dominance [19]. In computational settings, pleasure reflects positive or negative valence, arousal captures the level of activation or emotional intensity, and dominance describes the degree of perceived control. Compared with the Big Five personality traits, PAD emotional states are better suited for tracking turn-level fluctuations that may affect both conversational behaviors and response preferences. Therefore, prior work in affective computing has often treated trait models and affect models as complementary representations of stable and dynamic aspects of user state. Recent studies on LLMs further suggest that both personality traits and affective states can be inferred from language, including social media text and multi-turn dialogue contexts [22,23]. However, most existing studies take these variables as static persona descriptors, one-time prediction targets, or local generation controls, rather than as continuously updated state representations shared across the dialogue process.

A related line of work studies how affective information can improve response generation and memory use in dialogue systems. Research on empathetic dialogue has shown that factual relevance alone is insufficient, because systems must also identify emotional cues and produce responses with an appropriate tone. Lin et al. [14] proposed MoEL to condition response generation on emotional signals, and Li et al. [13] proposed KEMP to incorporate external knowledge into empathetic response generation. Later studies incorporated emotion-specific memory to support more positive and empathetic responses [25]. More recent studies

have extended this idea by integrating retrieval into affect-aware dialogue modeling. Huang et al. [8] proposed Emotional RAG for emotion-aware retrieval in role-playing agents, while Wen et al. [27] proposed RAER to improve emotion reasoning through retrieval. Besides, Li et al. [12] proposed LD-Agent, which incorporates both user and agent personas into response generation. Overall, these studies suggest two observations. First, psychological state may affect response style and, in some systems, the selection of relevant memory. Second, existing systems rarely maintain synchronized and continuously updated psychological states for both the user and the agent in a long-term memory RAG framework. This limitation motivates our formulation of personality and affect as dynamic state variables that guide generation and can complement retrieved evidence.

3 Framework of Psy4RAG

3.1 Overview

We develop Psy4RAG, an RAG-based framework that tracks the temporal evolution of participants' personality traits and emotional states in long-term dialogues, thereby improving the contextual coherence and appropriateness of generated responses. As illustrated in Figure 1, Psy4RAG consists of two psychologically oriented state modeling modules: the Personality Space State Modeling module and the Emotion Space State Modeling module. These modules use evidence from earlier dialogue turns and update their corresponding continuous states through Exponential Moving Average (EMA) updates [2,7]. Psy4RAG maintains separate state streams for the user and the responding agent. The user-side states capture the user's evolving personality-related cues and emotional states, while the agent-side states capture the agent's role-specific response trajectory across turns. Finally, the generator is conditioned on a psychological-aware prompt constructed from the outputs of the two state modeling modules. In this way, Psy4RAG maintains both stable trait summaries and dynamic emotional states for the user and the agent without changing the underlying RAG retriever or generator.

3.2 Personality Space State Modeling

The Personality Space State Modeling module is implemented through an OCEAN state estimator. It models relatively stable Big Five personality traits, including openness, conscientiousness, extraversion, agreeableness and neuroticism. Because personality traits usually emerge from repeated behavioral patterns rather than isolated utterances, the module uses previous sessions and the last session as the main evidence source, so that session evidence is first summarized as an OCEAN personality description before being converted into the current OCEAN value by the LLM.

For the speaker role r, the OCEAN state estimator first extracts the Big Five personality traits $\hat{b}_t^{(r)}$ from session-level evidence. Each dimension of the Big Five

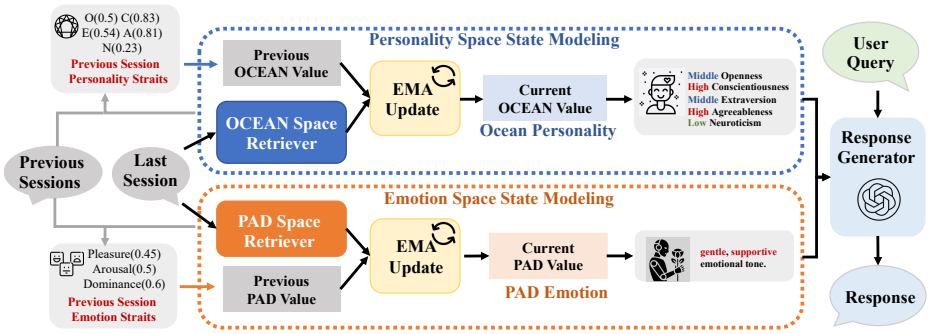

Fig. 1. The Psy4RAG framework. Personality states are updated at the session level, emotion states at the turn level. The current OCEAN and PAD values are verbalized and fed into the response generator.

personality traits is initially predicted on a $[1, 5]$ scale and then normalized to the $[0, 1]$ range. The continuous state is stored as a time-ordered sequence of personality-state estimates, and recent trait summaries are preserved as ordered textual memory. The textual summaries are inserted into following prompts, and the numerical trace preserves the temporal development of personality estimates across sessions.

To reduce the effect of short-term fluctuations, the OCEAN state estimator applies an EMA update to the previous OCEAN value and the newly extracted state. The smoothed personality state is updated as

$$b_t^{(r)} = (1 - \alpha_b) b_{t-1}^{(r)} + \alpha_b \hat{b}_t^{(r)}. \tag{1}$$

All dimensions are initialized to the neutral value 0.5 at the first session. We set $\alpha_b = 0.1$ so that personality estimates evolve conservatively. This conservative update is necessary because a single tense exchange or a short emotional episode should not substantially alter the personality profile of the speaker. The resulting OCEAN value provides the working personality summary for generation, such as middle openness, high conscientiousness, middle extraversion, high agreeableness, and low neuroticism in the framework.

3.3 Emotion Space State Modeling

The Emotion Space State Modeling module is implemented through a PAD state estimator. It models short-term affective states along the pleasure, arousal, and dominance dimensions of the PAD space. Unlike personality, emotions may change substantially from one turn to the next. Therefore, the PAD state estimator relies more strongly on recent dialogue context and emotion features from the previous sessions.

For each speaker role r, the PAD state estimator extracts a newly estimated PAD state $\hat{p}_t^{(r)}$. As in personality modeling, each PAD dimension is predicted on

a $[1, 5]$ scale and normalized to $[0, 1]$. The module then applies an EMA update to combine the previous PAD value with the newly extracted state:

$$p_t^{(r)} = (1 - \alpha_p) p_{t-1}^{(r)} + \alpha_p \hat{p}_t^{(r)}. \tag{2}$$

All emotional dimensions are initialized to 0.5. We set $\alpha_p = 0.2$, allowing the PAD trajectory to respond more quickly than the OCEAN trajectory. This choice enables the model to capture meaningful shifts in emotional tone while limiting the influence of isolated affective cues. The latest smoothed PAD state is used as the working emotional summary for each speaker, so generation is conditioned on a maintained affective estimate rather than on a prediction based solely on the most recent turn.

3.4 Response Generator

The generator for the final response includes a verbalization module that converts retrieved psychological states into natural language descriptions before they are inserted into the prompt. This module avoids exposing the generation model to raw numerical values and instead provides interpretable psychological cues for controlled response generation. Big Five personality trait scores are mapped to coarse categories such as *Very Low*, *Low*, *Moderate*, *High*, and *Very High*. PAD dimensions are verbalized as descriptive phrases such as *happy/content*, *calm/relaxed*, and *dominant/in-control*.

In addition to these psychological cues, we retrieve discrete trait memory that stores short natural language summaries of preferences, values, habits, and background information revealed in the dialogue. The generator integrates the retrieved evidence and the psychological cues into a structured prompt. Specifically, it first includes the recent dialogue context, then appends retrieved long-term summaries and finally adds three speaker-aware control blocks: user traits, user psychological state, and agent psychological state. We treat these blocks as inferred interaction cues rather than clinical or definitive psychological diagnoses. In the current implementation, they are used for generation control and do not modify the retrieval ranking of the backbone RAG system.

4 Experiments

4.1 Experiment Setup

Datasets. To evaluate long-term dialogue generation, we used two representative datasets, MSC [29] and LoCoMo [18], which have been widely used in prior studies. The statistics of the datasets are shown in Table 1. For each dataset, we generated replies turn by turn within each session, using all available history up to that point and reported session-level results for sessions 2–5, following the evaluation setting validated in previous work [12].

Baselines. We use representative long-term dialogue RAG baselines as backbones, including native LLM, HyperGraphRAG [17], HippoRAG [6] and LD-Agent [12]. The base language model used in RAG was Qwen3-8B [1]. On top

Table 1. Statistics of the MSC and LoCoMo datasets

Statistic	MSC	LoCoMo
Dialogues / Episodes	1,001	10
Sessions (total)	4,004	272
Avg. Sessions	4.0	27.2
Avg. Turns of Dialog	53.3	588.2
Avg. Turns of Session	13.33	21.63
Time Interval Between Sessions	few days	few months
Collection Method	crowdsourcing	LLM-generated + crowdsourcing
Speakers	2	2

of each backbone, we stacked the Psy4RAG module and compared the resulting system with its original counterpart under the same retrieval pipeline. This design allows us to evaluate the contribution of psychological state tracking independently of the choice of base RAG architecture. Because the focus of this version is a lightweight plug-and-play layer, we leave comparisons with additional prompt-only controls, such as static persona prompts, latest-turn emotion prompts, and shuffled psychological states, to future evaluation.

Evaluation Metrics. We report both lexical overlap and LLM-based evaluation metrics. "B-1" and "B-2" denote BLEU-1 and BLEU-2 [21], which measure the unigram and bigram overlap between the generated response and the dataset-provided reference response. "R-L" denotes ROUGE-L, which measures the longest common subsequence overlap and is commonly used to capture surface-level similarity at the sentence level. To complement these lexical metrics, we also use G-Eval [15], an LLM-based evaluator that scores each response along three dimensions: coherence (G-coh), consistency (G-con), and relevance (G-rel). Coherence captures how naturally the response fits the dialogue, consistency captures its compatibility with prior interaction, and relevance captures its appropriateness to the current conversational context. The base language model used in G-Eval is GLM4-9B [32]. Together, these metrics provide a broader view of response quality in long-term dialogues, covering both reference overlap and interaction quality. We use G-Eval as an automatic proxy for interaction-level quality, but we do not treat it as a direct validation of psychological correctness. Human judgments of affective appropriateness and psychological coherence remain necessary for full construct validation.

4.2 Main Results

To evaluate Psy4RAG, we integrated it with four long-term dialogue RAG backbones and reported the session-level metrics in Table 2 and Table 3 (Sessions 2–5), with the mean relative gains summarized in Table 4. We computed the relative change from the non-Psy row to the matching Psy row as $(M_{\text{Psy}} - M_{\text{base}})/M_{\text{base}}$, and then averaged these percentages over four backbones and four sessions.

Table 2. Main results on MSC and LoCoMo over Sessions 2–5, evaluated using lexical overlap metrics: B-1, B-2, and R-L. Bold indicates the better result within each Psy/non-Psy pair when both are available.

Model	Session 2			Session 3			Session 4			Session 5		
	B-1↑	B-2↑	R-L↑	B-1↑	B-2↑	R-L↑	B-1↑	B-2↑	R-L↑	B-1↑	B-2↑	R-L↑
MSC												
w/o Psy												
Native LLM	0.149	0.031	0.146	0.144	0.029	0.141	0.142	0.029	0.139	0.144	0.029	0.142
HyperGraphRAG	0.151	0.032	0.141	0.151	0.032	0.137	0.149	0.029	0.134	0.151	0.030	0.136
HippoRAG	0.163	0.037	0.155	0.161	0.037	0.151	0.155	0.035	0.146	0.158	0.036	0.149
LD-Agent	0.158	0.031	0.145	0.154	0.031	0.139	0.149	0.028	0.135	0.152	0.029	0.136
w/ Psy												
Native LLM + Psy	**0.153**	0.030	**0.146**	**0.147**	**0.029**	**0.143**	**0.144**	0.028	**0.140**	**0.148**	**0.030**	**0.143**
HyperGraphRAG + Psy	0.146	0.029	0.138	**0.152**	**0.032**	**0.139**	**0.156**	**0.036**	**0.144**	0.149	**0.032**	**0.137**
HippoRAG + Psy	**0.164**	**0.039**	0.154	**0.161**	0.036	0.150	**0.155**	0.034	**0.146**	**0.159**	0.035	0.148
LD-Agent + Psy	**0.159**	**0.033**	**0.146**	0.154	0.029	0.139	**0.153**	0.029	0.136	0.152	0.028	0.135
LoCoMo												
w/o Psy												
Native LLM	0.172	0.053	0.164	0.188	0.051	0.170	0.187	0.063	0.180	0.169	0.046	0.169
HyperGraphRAG	0.169	0.037	0.156	0.177	0.049	0.169	0.182	0.054	0.171	0.173	0.041	0.165
HippoRAG	0.141	0.030	0.134	0.162	0.039	0.150	0.151	0.038	0.148	0.142	0.024	0.137
LD-Agent	0.170	0.045	0.159	0.174	0.046	0.157	0.196	0.062	0.179	0.161	0.037	0.151
w/ Psy												
Native LLM + Psy	**0.172**	**0.053**	**0.166**	0.179	**0.055**	0.163	**0.194**	0.062	**0.187**	0.168	0.039	0.162
HyperGraphRAG + Psy	0.163	0.033	0.150	0.170	0.040	0.156	0.182	0.051	0.169	0.169	0.041	**0.166**
HippoRAG + Psy	**0.174**	**0.049**	**0.166**	**0.190**	**0.057**	**0.174**	**0.191**	**0.061**	**0.186**	**0.163**	**0.037**	**0.158**
LD-Agent + Psy	0.161	0.032	0.152	0.161	0.033	0.148	0.197	0.072	0.184	0.179	0.051	0.167

Table 3. Main results on MSC and LoCoMo over Sessions 2–5, evaluated using G-Eval metrics: G-coh, G-con, and G-rel.

Model	Session 2			Session 3			Session 4			Session 5		
	G-coh↑	G-con↑	G-rel↑	G-coh↑	G-con↑	G-rel↑	G-coh↑	G-con↑	G-rel↑	G-coh↑	G-con↑	G-rel↑
MSC												
w/o Psy												
Native LLM	2.603	2.437	2.353	2.578	2.516	2.379	2.604	2.468	2.340	2.611	2.436	2.385
HyperGraphRAG	2.604	2.393	2.306	2.598	2.346	2.351	2.624	2.411	2.299	2.609	2.375	2.380
HippoRAG	2.661	2.449	2.400	2.670	2.502	2.433	2.654	2.515	2.393	2.645	2.475	2.433
LD-Agent	2.621	2.414	2.314	2.583	2.363	2.228	2.573	2.330	2.289	2.614	2.388	2.279
w/ Psy												
Native LLM + Psy	**2.630**	**2.503**	**2.393**	**2.619**	2.484	2.366	**2.631**	**2.514**	**2.374**	**2.631**	**2.473**	2.381
HyperGraphRAG + Psy	**2.608**	**2.418**	**2.376**	**2.725**	**2.430**	2.321	**2.769**	**2.592**	**2.362**	**2.727**	2.264	2.303
HippoRAG + Psy	2.651	**2.511**	**2.423**	**2.711**	**2.558**	2.418	**2.702**	2.505	**2.409**	2.626	**2.517**	**2.440**
LD-Agent + Psy	2.586	2.378	2.270	**2.635**	**2.375**	**2.292**	**2.619**	**2.354**	**2.328**	**2.620**	2.380	**2.322**
LoCoMo												
w/o Psy												
Native LLM	2.132	1.373	1.823	2.400	1.517	1.943	2.527	1.440	2.004	2.406	1.493	1.929
HyperGraphRAG	2.160	1.373	1.815	2.456	1.527	1.895	2.386	1.414	1.949	2.485	1.537	1.954
HippoRAG	2.029	1.416	1.829	2.117	1.374	1.884	2.433	1.361	1.905	2.522	1.555	1.920
LD-Agent	2.122	1.247	1.680	2.270	1.396	1.785	2.350	1.434	2.084	2.369	1.491	1.831
w/ Psy												
Native LLM + Psy	**2.345**	**1.577**	**1.925**	**2.757**	**1.633**	**2.228**	**2.561**	**1.633**	**2.060**	**2.539**	**1.736**	**2.150**
HyperGraphRAG + Psy	**2.210**	**1.437**	1.793	2.364	**1.534**	**1.983**	**2.425**	**1.487**	**1.975**	2.366	1.406	1.873
HippoRAG + Psy	**2.471**	1.386	**2.029**	**2.719**	**1.749**	**2.239**	**2.596**	**1.416**	**2.151**	**2.793**	**1.720**	**2.158**
LD-Agent + Psy	**2.136**	**1.355**	**1.808**	**2.283**	**1.494**	**1.872**	**2.583**	**1.572**	2.055	**2.599**	1.454	**1.998**

Table 4. Mean relative improvement of Psy4RAG over the backbone without Psy, averaged over four backbones and Sessions 2–5. The bottom row is the mean of the MSC and LoCoMo results.

Dataset	B-1↑	B-2↑	R-L↑	G-coh↑	G-con↑	G-rel↑
MSC	+0.89%	+0.96%	+0.57%	+1.53%	+1.13%	+0.60%
LoCoMo	+4.34%	+11.08%	+4.31%	+7.23%	+7.29%	+6.90%
Avg. (MSC & LoCoMo)	+2.62%	+6.02%	+2.44%	+4.38%	+4.21%	+3.75%

Overall pattern. Aggregating across backbones and sessions, PSY4RAG improves both lexical overlap and G-Eval metrics, as shown in Table 4. The mean relative gains on "B-1"/"B-2"/"R-L" remain in the low single digits (especially on MSC), whereas "G-coh", "G-con", and "G-rel" show larger gains on LoCoMo. This gap suggests that internal psychological state signals are more helpful for style and interaction-level response control than for surface token overlap.

Lexical overlap ("B-1", "B-2", "R-L"). Session-level cells show small positive shifts overall, with the largest BLEU/ROUGE lifts on LoCoMo. However, the improvements are not uniform, and some backbone-session pairs register flat or slightly negative overlaps. For example, "HyperGraphRAG + PSY" on MSC Session 2 for "B-1" decreases. This result may be partly attributed to the large number of additional memory prompts in RAG models. When processing these prompts, the base large language model may allocate less attention to the external knowledge input provided by PSY4RAG, thereby weakening its effective use of that knowledge.

G-Eval ("G-coh", "G-con", "G-rel"). Under the same aggregation setting, all three dimensions show improvements on both datasets, as reported in Table 4. The per-session results further indicate that psychological conditioning is more effective in larger dialogue sessions. For example, "HippoRAG + PSY" shows large G-Eval gains on LoCoMo (e.g., Session 3), and Native LLM + PSY improves multiple G dimensions on LoCoMo Session 3. This suggests that Psy4RAG may contribute to maintaining dialogue-level consistency and stylistic coherence across dialogue turns.

Overall, PSY4RAG achieves the most pronounced gains at the interaction level in G-Eval. The improvement is most visible in response style and interaction-level metrics. We also notice that the effect is stronger for the LoCoMo dataset, but this result should be interpreted with caution because LoCoMo contains only 10 dialogue episodes.

4.3 Ablation Study

To further validate the contribution of modeling personality and emotion space states, we conduct an ablation study on the MSC dataset under the Native LLM + PSY setting. The full model uses Qwen3-8B with Psy4RAG settings. For each ablated setting, we follow the same evaluation protocol as in the main experiment: each metric is computed at the session level, and the unweighted arithmetic mean over Sessions 2–5 is reported in Table 5. We consider four ablation settings. First, Psy4RAG w/o Personality and w/o Emotion remove one state module while keeping the other module and both participant-side inputs active. Specifically, w/o Emotion removes emotion space states and keeps personality space states; conversely, w/o Personality removes personality space states and keeps emotion space states. Second, w/o User and w/o Agent retain both psychological channels but remove one participant-side stream during generation. The w/o User variant removes user-side psychological states and keeps only agent-side states, whereas the w/o Agent variant removes agent-side psychological states and keeps only user-side states.

Table 5. Ablation results on MSC under the Native LLM + Psy setting. We report mean scores over Sessions 2–5. "w" and "w/o" indicate whether each psychological state component is used.

Setting	Personality	Emotion	User	Agent	B-1↑	B-2↑	R-L↑	G-coh↑	G-con↑	G-rel↑
Full model	✓	✓	✓	✓	**0.148**	**0.029**	**0.143**	**2.628**	**2.494**	**2.379**
w/o Personality	–	✓	✓	✓	0.145	0.025	0.132	2.565	2.360	2.255
w/o Emotion	✓	–	✓	✓	0.147	0.025	0.132	2.586	2.366	2.271
w/o Agent	✓	✓	✓	–	0.145	0.024	0.130	2.562	2.339	2.237
w/o User	✓	✓	–	✓	0.145	0.025	0.131	2.561	2.340	2.255

Overall effect of psychological signals. According to Table 5, the performance gain of Psy does not come from a single psychological cue or a single participant-side state. Instead, the full configuration benefits from jointly modeling relatively stable personality information, short-term emotion information, and the psychological trajectories of both participants. Among the ablated variants, User only (w/o Agent) yields the weakest results. This pattern suggests that psychological information is most useful when it supports interaction-level response control, rather than when the generator observes only a partial view of the psychological context.

Effect of emotion space states and personality space states. The comparison between emotion-only and personality-only shows that personality space states provide a more stable contribution to response generation. When emotion states are removed, the model obtains consistently better results across lexical-overlap metrics and G-Eval dimensions than the variant that retains only emotion states. Although the absolute margin is limited, the direction is consistent. This indicates that in multi-session dialogues, personality space states offer a more reliable prior for maintaining role-consistent tone and response stance, whereas emotional states mainly provide local affective information that is less sufficient on its own.

Effect of user-side and agent-side states. The comparison between User only and Agent only further shows that both participant-side streams are necessary. The two variants are close in performance, but Agent only is slightly stronger overall, which indicates that maintaining the psychological trajectory of the responding agent is particularly useful for improving response stability and role consistency.

Overall, the ablation results support the full design of Psy4RAG. PAD and OCEAN play complementary roles, and the generator benefits most when it conditions on both user-side and agent-side states instead of relying on only one component or one participant.

4.4 Case Study

To further examine how psychological state features evolve over the course of a dialogue, we conducted a case study of Psy4RAG on a representative career choice discussion from the MSC dataset, as shown in Figure 2. In this example,

the user provides guidance, and the agent responds as a listener. The case provides a more direct view of the temporal dynamics captured by the personality space state and emotion space state modeling modules. Across the dialogue, the emotional state on the user side does not remain constant. Instead, it changes with local interactions, including a shift from joy toward a more neutral state around Turn 2 and again near Turn 21. These transitions are also reflected in the wording of the dialogue, which suggests, at least qualitatively, that the extracted affective states are grounded in the interaction content rather than being detached from it. At the same time, the personality representation changes

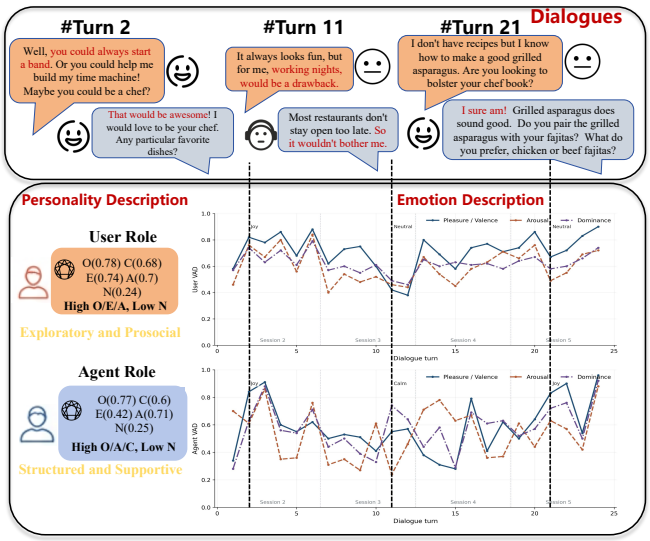

Fig. 2. Psychological space tracking of the user and agent in long-term dialogue within the Psy4RAG framework. Orange blocks denote the user, blue blocks denote the agent, the psychological-space curves below show the quantified affective changes of the user and agent in the PAD emotion space, the left panel presents the quantified Big Five personality traits of the user and agent, and the top panel shows the utterances in specific dialogue turns.

gradually. Based on session-level input, Psy4RAG maintains interpretable, role-specific tendencies for both the user and the agent in the OCEAN space, while the PAD signal captures shorter-term affective fluctuations at the turn level. This distinction is important for long-term dialogues, because a single turn may convey tension, hesitation, or reassurance, but such local affective cues should not immediately revise the broader personality profile of either participant. Therefore, this case illustrates the intended functional separation of the model. Personality space state modeling provides a stable interaction prior to the next response, whereas emotion space state modeling tracks local changes

that are relevant to the next response. This observation also helps explain why Psy4RAG shows clearer gains in G-Eval than in lexical-overlap metrics. By preserving emotional states, Psy4RAG gives the generator a more grounded basis for selecting tone and stance, helping responses remain aligned with the evolving interpersonal state of the dialogue.

4.5 Discussion

Psy4RAG achieves more consistent improvements on G-Eval than on lexical-overlap metrics. This pattern suggests that psychological state modeling mainly affects response framing, including tone, stance, and dialogue-level consistency, rather than merely increasing lexical similarity to retrieved content. In long-term dialogue, retrieval indicates what has been mentioned, whereas psychological states help determine how the response should be expressed.

The stronger gains on LoCoMo further support this interpretation. Compared with MSC, LoCoMo contains longer dialogue histories, more sessions, and wider temporal gaps between sessions [18]. These properties make it difficult for a model to maintain interpersonal continuity based on retrieved facts alone. Under this condition, psychological conditioning becomes more useful because it summarizes behavioral tendencies that persist beyond individual turns.

The case study and ablation results clarify the internal division of labor in Psy4RAG. Personality space states provide a relatively stable prior for role-consistent behavior, while emotional states capture local affective changes that are relevant to the next response. The ablation study confirms that these two signals are complementary, removing either psychological channel weakens performance, and the full model performs best when both are used. The comparison between user-side and agent-side variants further shows that psychological modeling should not be treated as a one-sided user representation. In long-term dialogue, the generator must track both how the user presents the situation and how the agent should maintain its own response trajectory.

The reported results should be understood as evidence for the utility of inferred psychological cues in response generation, rather than as evidence that the inferred Big Five and PAD values are psychologically valid measurements. Neither MSC nor LoCoMo provides ground-truth annotations for personality traits or affective states. Consequently, the present experiments only show that these inferred cues can improve automatic evaluation scores when used as conditioning signals in an RAG framework. They do not establish that the inferred traits accurately represent the speakers' real psychological states. A stronger validation would require human annotation of affective appropriateness, personality consistency, and over-personalization risk, as well as significance testing or confidence intervals over dialogues.

5 Conclusions and Future Work

We propose Psy4RAG, a plug-and-play RAG module that tracks inferred Big Five personality traits and PAD emotions for long-term dialogue generation. It

maintains a structured psychological trace and injects it into response generation without modifying the underlying RAG architecture. Experiments on MSC and LoCoMo show average gains across lexical-overlap and LLM-based consistency metrics, with larger improvements on interaction-level G-Eval scores. Future work will strengthen the control design and adaptive integration of Psy4RAG, and evaluate its affective appropriateness and over-personalization risks through human assessment.

References

1. Alibaba Cloud: Qwen3-8B. `https://huggingface.co/Qwen/Qwen3-8B` (2025)
2. Brown, R.G.: Statistical Forecasting for Inventory Control. McGraw-Hill, New York (1959)
3. Cheng, Y., Mao, K., Zhao, Z., Dong, G., Qian, H., Wu, Y., Sakai, T., Wen, J.R., Dou, Z.: CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation. In: Proc. of NAACL (2025)
4. Du, H., Feng, X., Ma, J., Wang, M., Tao, S., Zhong, Y., Li, Y.F., Wang, H.: Towards proactive interactions for in-vehicle conversational assistants utilizing large language models. In: Proc. of IJCAI. pp. 7850–7858 (2024)
5. Goldberg, L.R.: An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology **59**(6), 1216–1229 (1990)
6. Gutiérrez, B.J., Shu, Y., Gu, Y., Yasunaga, M., Su, Y.: HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In: Proc. of NeurIPS (2024)
7. Holt, C.C.: Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting **20**(1), 5–10 (2004)
8. Huang, L., Lan, H., Sun, Z., Shi, C., Bai, T.: Emotional rag: Enhancing role-playing agents through emotional retrieval. In: Proc. of IEEE ICKG (2024)
9. Jo, E., Jeong, Y., Park, S., Epstein, D.A., Kim, Y.H.: Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In: Proc. of CHI. pp. 1–21 (2024)
10. Katsis, Y., Rosenthal, S., Fadnis, K., Gunasekara, C., Lee, Y.S., Popa, L., Shah, V., Zhu, H., Contractor, D., Danilevsky, M.: MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. Transactions of the Association for Computational Linguistics **13**, 784–808 (2025)
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proc. of NeurIPS (2020)
12. Li, H., Yang, C., Zhang, A., Deng, Y., Wang, X., Chua, T.S.: Hello Again! LLM-powered Personalized Agent for Long-term Dialogue. In: Proc. of NAACL (2025)
13. Li, Q., Li, P., Ren, Z., Ren, P., Chen, Z.: Knowledge Bridging for Empathetic Dialogue Generation. In: Proc. of AAAI (2022)
14. Lin, Z., Madotto, A., Shin, J., Xu, P., Fung, P.: MoEL: Mixture of Empathetic Listeners. In: Proc. of EMNLP (2019)
15. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: Proc. of EMNLP (2023)
16. Lu, J., Li, Y.: Dynamic Affective Memory Management for Personalized LLM Agents. arXiv preprint arXiv:2510.27418 (2025)

17. Luo, H., E, H., Chen, G., Zheng, Y., Wu, X., Guo, Y., Lin, Q., Feng, Y., Kuang, Z., Song, M., Zhu, Y., Tuan, L.A.: HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. In: Proc. of NeurIPS (2025)
18. Maharana, A., Lee, D.H., Tulyakov, S., Bansal, M., Barbieri, F., Fang, Y.: Evaluating Very Long-Term Conversational Memory of LLM Agents. In: Proc. of ACL (2024)
19. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology **14**(4), 261–292 (1996)
20. Pan, Z., Wu, Q., Jiang, H., Luo, X., Cheng, H., Li, D., Yang, Y., Lin, C.Y., Zhao, H.V., Qiu, L., Gao, J.: SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) Proc. of ICLR (2025)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proc. of ACL (2002)
22. Peters, H., Matz, S.C.: Large language models can infer psychological dispositions of social media users. PNAS Nexus **3**(6), pgae231 (2024)
23. Sandhan, J., Cheng, F., Sandhan, T., Murawaki, Y.: CAPE: Context-Aware Personality Evaluation Framework for Large Language Models. In: Proc. of Findings EMNLP (2025)
24. Tan, Z., Yan, J., Hsu, I.H., Han, R., Wang, Z., Le, L., Song, Y., Chen, Y., Palangi, H., Lee, G., Iyer, A.R., Chen, T., Liu, H., Lee, C.Y., Pfister, T.: In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. In: Proc. of ACL (2025)
25. Tian, Z., Wang, Y., Song, Y., Zhang, C., Lee, D., Zhao, Y., Li, D., Zhang, N.L.: Empathetic and Emotionally Positive Conversation Systems with an Emotion-specific Query-Response Memory. In: Proc. of Findings of EMNLP (2022)
26. Wang, Q., Fu, Y., Cao, Y., Wang, S., Tian, Z., Ding, L.: Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models. Neurocomputing **639**, 130193 (2025)
27. Wen, Z., Lian, Z., Chen, S., Yao, H., Yang, L., Liu, B., Tao, J.: Listen, Watch, and Learn to Feel: Retrieval-Augmented Emotion Reasoning for Compound Emotion Generation. In: Proc. of Findings of ACL (2025)
28. Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.W., Yu, D.: LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In: Proc. of ICLR (2025)
29. Xu, J., Szlam, A., Weston, J.: Beyond goldfish memory: Long-term open-domain conversation. In: Proc. of ACL (2022)
30. Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., Shen, Y.: A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. ACM Computing Surveys **58**(6), 1–38 (2025)
31. Zhang, F., Zhu, D., Ming, J., Jin, Y., Chai, D., Yang, L., Tian, H., Fan, Z., Chen, K.: DH-RAG: A Dynamic Historical Context-Powered Retrieval-Augmented Generation Method for Multi-Turn Dialogue. arXiv preprint arXiv:2502.13847 (2025)
32. Zhipu AI: GLM-4-9B. `https://huggingface.co/THUDM/glm-4-9b` (2024)
33. Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: MemoryBank: Enhancing Large Language Models with Long-Term Memory. In: Proc. of AAAI (2024)