WHEN SITUATIONS TAKE EFFECT: PSYCHOLOGICAL SITUATION MODULATION FOR PERSONALIZED DIALOGUE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized dialogue must reflect users' histories and how they appraise the current situation. Existing situation representations often describe physical or social context without specifying how psychological appraisal guides memory and personality expression. We propose PSyMod, which constructs episodic memory, personality, and a historical situation reference from preceding dialogue. At generation time, a current reading on the eight-dimensional DIAMONDS situation taxonomy determines a situation type through fixed rules. That type guides memory retrieval, personality expression, and response planning, while the historical reference supports individual adjustment of expression strength and turn-level affect adjusts intensity. The persistent state remains read-only within each target turn. We separately evaluate reference-state-conditioned generation, state construction, and closed-loop operation using model-assisted reference annotations. PSyMod achieves the highest pooled situational modulation score and leads on the personality and memory axes. It scores higher on ESC than the prompt-only condition but slightly lower on blind ABCD. The construction audit finds no hallucinated stored entries with competitive retrieval, while a closed-loop probe offers only directional evidence for write-back.

1 INTRODUCTION

In long-term companionship and psychological support, response fit determines whether a system remains usable (Liu et al., 2021; Lee et al., 2025). Personalized dialogue should account for the user's facts, enduring traits, current affect, and accumulated experiences (Zhang et al., 2018). Persona, emotion, and memory systems address consistency, feelings, and cross-session information respectively (Zhang et al., 2018; Zhu et al., 2026; Liu et al., 2021; Rashkin et al., 2019; Lu et al., 2023; Wang et al., 2026a). Yet a consistent profile alone does not determine which past experience or behavioral tendency is relevant to a particular response.

Existing situation-aware approaches often represent life contexts (Sun et al., 2026), observable circumstances (Otani et al., 2023), evaluation scenarios (Wang et al., 2026e), or topics (Wei et al., 2026). These representations alone do not specify how a user's psychological appraisal should guide memory selection in personalized generation. Responses depend on personality–situation interaction rather than traits alone (Mischel & Shoda, 1995; Rauthmann et al., 2014). Here psychological situation means the appraisal of current stakes and circumstances, described by the eight-dimensional DIAMONDS taxonomy (Rauthmann et al., 2014). In Figure 1, a user faces a team-building outing with a roller coaster and drop tower. Topic-based retrieval finds a pleasant bumper-car memory M_1 but misses a fear-of-heights experience M_2 near a cliff. The former alone might prompt encouragement to ride everything. The latter supports a response attentive to the user's apprehension. The relevant distinction is psychological, not a shared location or topic. Representing that appraisal is only the first step. Existing work does not establish how to coordinate its effects on memory selection, personality expression, and turn-level affect in a response (Jiang et al., 2025). Response-quality evaluation alone cannot show whether the system constructs an accurate user state or how that state affects later turns. These generation and evaluation gaps motivate treating the four dimensions together without collapsing them into one profile.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

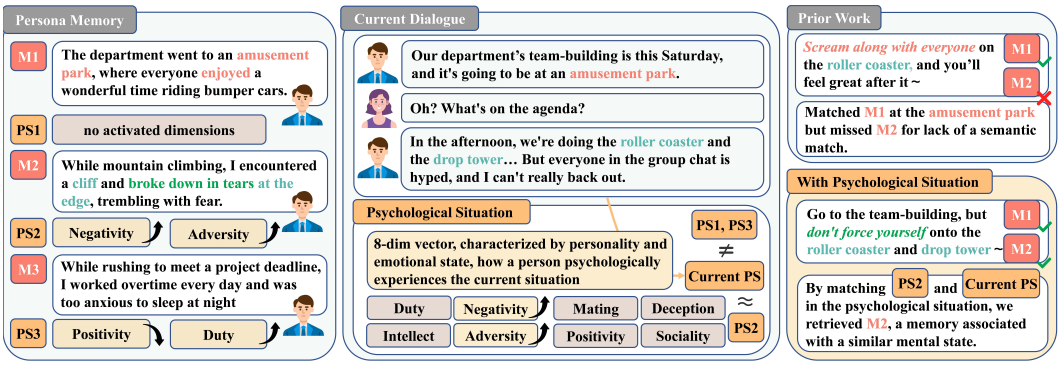

Figure 1: Psychological situation guides memory retrieval. Topic-based retrieval finds the amusement-park memory M_1 but misses the fear-of-heights experience M_2. DIAMONDS-based situation matching makes M_2 available for a response attentive to the user's apprehension.

We introduce PSyMod, a Psychological Situation Modulation framework for constructing user state and generating responses conditioned on it. State construction turns the history observable before a target turn into episodic memory, a personality profile, and a historical DIAMONDS reference summarizing prior situation readings. At generation time, a model measures the current situation and affect from the incoming utterance and context. Fixed rules determine the situation type from the current reading's reference-scale percentiles, not from its difference from the user's historical reference. This type guides memory selection, personality constraints, and response planning. The historical reference separately supports an individual adjustment to personality-expression strength, while affect adjusts expressive intensity. The persistent state remains frozen and read-only within the target turn; responses can affect later states only through subsequent observable history. This separation also makes state construction assessable independently of conditioned generation.

Our benchmark combines three dialogue corpora with human Big Five annotations (Yamashita et al., 2023; Chen et al., 2022; Jiang et al., 2019), derives the situational and memory representations through a model-based pipeline, and reports metrics per track, excluding abstained or inapplicable samples from their denominators. We evaluate reference-state-conditioned response generation, state construction, and closed-loop operation separately. On conditioned generation, PSyMod achieves the highest pooled situational modulation score and leads on the personality-expression and memory-citation axes. Relative to the prompt-only condition, it scores higher on ESC but slightly lower on blind ABCD. The state-construction audit finds no hallucinated stored entries and competitive retrieval, subject to a coverage–selectivity trade-off. A matched closed-loop probe gives directional evidence for write-back, not proof of long-horizon stability. Ablations show that quality, modulation, and storage selectivity can move in different directions.

In summary, our contributions are as follows:

- We introduce psychological situations to modulate memory retrieval, personality expression, and response planning in personalized response generation.
- We propose a four-dimensional representation that keeps memory, personality, and historical situation frozen while modeling affect at each turn, enabling direct evaluation of state construction.
- We provide a benchmark and protocol for separately evaluating response generation, state construction, and closed-loop operation, using model-generated annotations reviewed manually.

2 Related Work

Agent architectures and longitudinal simulators maintain evolving user representations (Park et al., 2023; Sumers et al., 2024; Wang et al., 2026b; Jin et al., 2025; Yao et al., 2025). Personality-oriented systems use behavioral data, prompting, or model adaptation to express traits (Binz et al., 2024; Wang et al., 2025; Chen et al., 2025; Wang et al., 2024b). These lines address persistence and trait expression, but reported weaknesses include limited psychological sensitivity (Harry et al., 2026) and inconsistencies between stated traits and observed behavior (Han et al., 2025). Persistent

profiles therefore leave open how a current appraisal affects the expression of a trait in a particular response.

DIAMONDS describes psychological situations along eight dimensions (Rauthmann et al., 2014; Terwiel et al., 2020). Situation-aware methods infer mental states from social situations (Hwang et al., 2026), retrieve personality neurons by topic (Wei et al., 2026), or reweight and route personality facets from contextual cues (Liu et al., 2026c; Tang et al., 2026b). The first targets inference, while the latter methods change personality selection. These uses do not by themselves establish whether an appraised situation should also govern access to the user's episodic experiences. HumanLLM uses DIAMONDS to diversify evaluation scenarios rather than control response generation (Wang et al., 2026e).

Memory architectures improve storage and retrieval (Chhikara et al., 2025; Chen et al., 2026b), while user-centric systems adapt memories or their update policies (Zhao et al., 2026; Han et al., 2026). Scenario-based retrieval makes context relevant to memory access (Sun et al., 2026). A scenario used as a retrieval key, however, need not distinguish topic overlap from a user's appraisal of risk or stakes. Affect-aware memory systems address emotional content (Lu & Li, 2025; Liu et al., 2026a; Amanlou et al., 2026), while studies of memory errors and evidence utilization show why retrieval quality alone is insufficient (Tao et al., 2026; Xiao et al., 2026; Qing et al., 2026).

Personalization benchmarks test memory use, emotional responses, and behavior over extended interactions (Jiang et al., 2026; Chen et al., 2026a; Wen et al., 2026; Fu et al., 2026; Ye et al., 2026; Li et al., 2026; Yu et al., 2026; Liu et al., 2026b). A comparison of temporal-intervention protocols identifies gaps in the joint treatment of state persistence, cross-dimensional effects, and user-condition variation (Qian et al., 2026). Thus, response quality alone does not establish whether a system constructed an accurate state or whether that state remains useful over time. Appendix A gives a more detailed comparison.

3 Task Setting

We construct a persistent state for a designated participant in a multi-party dialogue and generate a single-sentence response at a designated turn. Given participant p and turn t, where other interlocutors produce the input utterance u_t for p to answer, the system constructs the state from information observable before t and generates the response. Reference annotations are used only for offline evaluation. Let $\mathcal{V}_{<t}$ be the complete sequence of runtime-observable utterances before t, and let $\mathcal{W}_{<t} \preceq \mathcal{V}_{<t}$ be the window-truncated input, where $\preceq$ denotes the contiguous-subsequence relation.

State construction. State construction forms and maintains the target participant's persistent state based on runtime-observable utterances. The process B operates on the observable input window $\mathcal{W}_{<t}$ and outputs an episodic memory state, a Big Five personality profile, and an individual reference vector S_{avg} over the eight DIAMONDS dimensions summarizing historical situational readings preceding t:

$$(\mathcal{M}_p, \mathbf{P}_p, S_{\text{avg}}) = B(\mathcal{W}_{<t}).$$

For each DIAMONDS dimension d, the corresponding component of S_{avg} is the median of the available readings over preceding utterances; a missing reading for that dimension is represented as empty and excluded from that component's median. Thus, S_{avg} is an eight-dimensional vector.

Response generation. Generation produces a single-sentence response for p from the context preceding t and the constructed persistent state, accessing only memory entries formed before t. Let a_t denote the affective state measured from the target utterance u_t at generation time; it is not part of the frozen persistent state:

$$M(p,t) = \{m \in \mathcal{M}_p \mid \text{turn}(m) < t\},$$

$$y_t = G\big(\mathcal{W}_{<t}, p, t, u_t, M(p,t), \mathbf{P}_p, S_{\text{avg}}, a_t\big).$$

Closed-loop coupling. The response y_t becomes a new utterance by p and enters the runtime-observable history for subsequent turns. Thus, for any $t' > t$,

$$y_t \in \mathcal{V}_{<t'}.$$

The subsequent input window is $\mathcal{W}_{<t'} \preceq \mathcal{V}_{<t'}$, and the state is re-estimated:

$$(\mathcal{M}'_p, \mathbf{P}'_p, S'_{\text{avg}}) = B(\mathcal{W}_{<t'}).$$

Therefore, y_t can affect a later state only through the later observable window and state-construction process; it cannot directly update the frozen state at t. The state remains read-only within each turn and is re-estimated from updated observable history on subsequent turns. Additional details on generation-side logging and mediated closed-loop effects appear in Appendix B.

4 DATASET

4.1 DATA SOURCES AND CONSTRUCTION

The benchmark draws on three dialogue corpora with different annotation provenance. CPED (Chen et al., 2022) is a scripted Chinese TV-drama corpus with character-level expert Big Five ratings and utterance-level VAD emotion annotations. Friends is a scripted English sitcom corpus with utterance-level crowdsourced Big Five annotations from FriendsPersona (Jiang et al., 2019) and emotion annotations from PELD (Wen et al., 2024). RealPersonaChat (RPC) (Yamashita et al., 2023) contains genuine first-person role-play dialogues and character-level self-reported Big Five profiles, but no emotion annotations. We construct situational and memory representations uniformly across all three corpora. Personality, affective state, situation, and memory use distinct representations. Personality is a three-level Big Five profile (high, medium, or low); affective state is an utterance-level VAD vector (Warriner et al., 2013). Both unify heterogeneous annotations onto a common scale; affect is marked missing for RPC. Situation uses the eight DIAMONDS dimensions (Rauthmann et al., 2014). For reference annotations, three models (qwen3.8-max, claude-opus-4.6, and gpt-5.5) independently generate situational descriptions and score each dimension from their own descriptions. A score is retained only when at least two models agree, then manually reviewed. Runtime inference uses one model and does not access these annotations. Appendix C.2–C.3 reports the scale version, items, reference standards, scoring configurations (temperature 0), and per-dimension reliability. Memory is a persistent state $\mathcal{M}_p$ with two speaker-oriented entry groups: what the person said and what the person heard. Both support cross-scene retrieval, and each entry is bound at formation to the eight-dimensional situational encoding `enc_S`. Factual and episodic channels are retrieval views over this shared state: the former prioritizes autobiographical content, the latter situation-consistent coping exemplars. Appendix C.1 describes how these representations inform generation and defines the memory-encoding boundary.

4.2 EVALUATION SETTINGS AND STATISTICS

The source corpora contain 9,078 units: 2,587 from CPED, 5,284 from Friends, and 1,207 from RPC. The protocols organize evaluation material differently, so their pack counts are reported separately. $\mathcal{D}_{\text{gen}}$ (generation) provides aligned evaluation units, four-dimensional annotations, reference memory states, and short context windows. $\mathcal{D}_{\text{con}}$ (state construction) provides longer dialogue histories but withholds annotations and memory states, which the system must infer and accumulate. $\mathcal{D}_{\text{full}}$ (closed loop) organizes each character's visible turns into a temporally ordered full stream, which the system processes from a cold start to approximate everyday conversation, where information accumulates over time. The benchmark is for evaluation only; no training or test splits are defined. In $\mathcal{D}_{\text{gen}}$, provided annotations and reference memory states evaluate conditional response generation. These states are derived from preceding history and may summarize substantially more than the generator's short context window. Thus, $\mathcal{D}_{\text{gen}}$ evaluates generation conditioned on a reference state, not end-to-end state construction; the latter is evaluated in $\mathcal{D}_{\text{con}}$ and $\mathcal{D}_{\text{full}}$.

- $\mathcal{D}_{\text{gen}}$ (generation): 720 evaluation packs; mean history length 13.5 turns; 462 scenarios; mean memory span 616.2 turns; P/E/S/M coverage 1.000/0.808/1.000/1.000.
- $\mathcal{D}_{\text{con}}$ (state construction): 676 evaluation packs; mean history length 66.8 turns; 1,469 scenarios; mean memory span 36.0 turns. Annotations and memory states are withheld and accumulated by the system, so coverage statistics are undefined.
- $\mathcal{D}_{\text{full}}$ (closed loop): 284 character packs; mean character-full-stream history length 836.7 turns; 13,473 scenarios; mean memory span 636.2 turns; P/E/S/M coverage 0.621/0.440/0.801/0.223.

Pack counts are not additive because the protocols use different grouping and evaluation units. Appendix C.4 (Tables C1 and C2) provides the complete statistics and per-source coverage matrix.

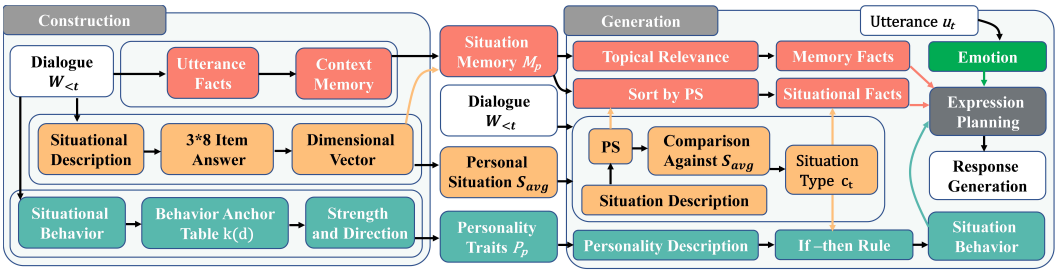

Figure 2: Overview of PSyMod. The construction module builds a frozen memory state, personality profile, and historical situation reference. The generation module combines these with current context and affect to retrieve memories, project behavior, and generate a verified response.

Situational annotations achieve a post-calibration mean ICC of 0.75, indicating good reliability (Koo & Li, 2016); per-dimension values are in Appendix C.2. Construction quality was verified by full-corpus recomputation. Appendix F.4 gives the full statements compressed here.

5 METHOD

The system comprises state construction and response generation under a shared information boundary (Figure 2). State construction freezes memory, personality, and an individual situational reference from pre-target information; response generation uses this state and the current context without updating it.

5.1 STATE CONSTRUCTION

State construction compresses the visible history $\mathcal{W}_{<t}$ into a frozen state using B (Section 3):

$$(\mathcal{M}_p, \mathbf{P}_p, S_{\text{avg}}) = B(\mathcal{W}_{<t}).$$

For each visible utterance, a single runtime model applies the runtime version of the two-step DIAMONDS-SF24 procedure (Section 4). The resulting situation reading is stored in the utterance's memory entry (Appendix D.4) and contributes to S_{avg}. In contrast, the dataset's reference annotations use three independent models solely for offline evaluation.

Personality inference proceeds bottom-up from behavioral evidence, drawing on the hierarchical view of personality (McAdams, 1995) and CAPS situation–behavior patterns (Mischel & Shoda, 1995). A behavioral-anchor table (Smith & Kendall, 1963) guides the scoring of evidence strength and direction in sentence-level observations $x_{c,k,j}$ for concern dimension k in session c. Let $J_c(d)$ index observations assigned to Big Five dimension d through the fixed concern-domain mapping $k(d)$. Averaging first within sessions and then across sessions yields:

$$P_{p,d} = \text{Mean}_c \left(\text{Mean}_{j \in J_c(d)} (x_{c,k(d),j}) \right).$$

Observation counts gate each dimension's contribution to behavioral constraints. Appendix D specifies the mapping and evidence gates; Appendix D.2, Table D1 details the projection stages.

5.2 RESPONSE GENERATION

Response generation uses $(\mathcal{M}_p, \mathbf{P}_p, S_{\text{avg}})$, $\mathcal{W}_{<t}$, and the trigger utterance u_t to produce a single-sentence response y_t. It proceeds through situation and affect measurement, personality projection and memory retrieval, expression planning, and constrained generation with verification (Algorithm 1, Appendix D.1).

A single runtime model measures the current situation from u_t and its preceding context using the runtime two-step procedure (Section 4). The reading is converted to per-dimension percentiles on the reference scale and mapped to situation type c_t using fixed, read-only rules (Appendix D.3). The same runtime pass predicts the utterance's VAD affective state, leaving it empty when unavailable.

Personality projection reverses the bottom-up inference: it maps $\mathbf{P}_p$ through concern subdimensions to level-specific personality descriptions. Situation-conditioned if–then rules keyed on c_t then activate the corresponding behavioral directions.

Memory retrieval uses two channels. The factual channel retrieves autobiographical texts by topical relevance to u_t; the episodic channel ranks coping exemplars by similarity between their encoding situations and the current situation. Situation type c_t divides candidates into strongly comparable, style-comparable, and non-comparable classes; only the first two enter generation.

Expression planning translates these representations into generation guidance. It first identifies required response functions, such as answering requests, acknowledging emotion, and maintaining boundaries. The situation determines the coping direction, personality supplies personalized behavioral directions and wording tendencies, affect modulates expressive intensity, and memory supplies content rather than expressive direction.

Deterministic screening removes guidance inconsistent with recent behavioral evidence and resolves conflicts by "situation $\succ$ memory $\succ$ affective state $\succ$ personality." The resulting guidance supports constrained generation and verification. Appendix D.3 specifies behavior-level conversion, comparability criteria, the length reference, and fallback rules for empty memory banks; Appendix F.5 provides the full specifications.

6 Experiments

6.1 Evaluation Tracks and Baselines

We evaluate PSyMod along three tracks covering **response generation, state construction, and closed-loop operation**. All API-based steps use the same fixed generation configuration; unified run configurations and inclusion rationale for each method are given in Appendix E.1.

Response generation. Three reference conditions measure the value of providing state information itself. Cond-none provides only the dialogue context, Cond-gold adds gold state text and memory-targeted entries, and Cond-gold+mem adds all gold memory entries. Of the three, only Cond-none enters the reported capability comparison, where it appears in Table 1 as the prompt-only condition. The two gold-state conditions are attribution references and are not ranked against the systems. Baselines cover both training-based and inference-based methods for the four dimensions. Training-based baselines include CharacterGLM-6B (Zhou et al., 2023), PsyLLM-8B (Hu et al., 2025), MemoryLLM (Wang et al., 2024c), SOTOPIA-π (Wang et al., 2024a), Sotopia-RL (Yu et al., 2025), and SocialRL (Wang et al., 2026c). Inference-based baselines include PERSONA (Feng et al., 2026), Emotional RAG (Huang et al., 2024), and PHASE-Tree (Tang et al., 2026a).

State construction. Memory-construction baselines include RGMem (Tian et al., 2025), Mem0 (Chhikara et al., 2025), MemInsight (Salama et al., 2025), and AdaMem (Chen et al., 2026b). We evaluate only their memory-construction artifacts and retrieval behavior.

Closed-loop operation. Closed-loop baselines include HyperTrace (Shen et al., 2026) and PG-Mem (Choi et al., 2026), both of which update their states from the dialogue history and generate responses accordingly.

6.2 Evaluation Metrics

Results are reported separately for generation, state construction, and closed-loop operation; no composite score is used. All evaluation APIs use claude-opus-4.6, distinct from the generation model. The main tables report the registered headline metrics for each track. Metric definitions and construction rules are given in Appendix E.2; significance and small-sample handling are given in Appendix E.3; fairness controls are given in Appendix E.4; and the complete scoring prompts are reproduced in Appendix E.5.

Tier 1 response quality on the generation side. Tier 1 reports the ABCD score from a blind four-dimensional review model, together with the ESC-Eval Overall and six-dimension scores (Zheng

et al., 2023; Zhao et al., 2024). Mean response length serves as a diagnostic rather than a quality score.

Tier 2 situational modulation consistency on the generation side. Tier 2 reports a pooled modulation score and three situation-specific axes. The E axis measures emotional expression, the P axis personality expression, and the M axis memory citation. The situation annotations and the derived reference sets are model-assisted rather than human gold labels. Each axis measures agreement between generated behavior and these frozen references using a pre-specified direction; the axes are reported separately rather than averaged into an additional score. The memory-citation ratio is retained as a supplementary diagnostic rather than a headline column and is defined in Appendix E.2, where formal metric definitions and construction rules are also given; significance testing and small-sample handling are given in Appendix E.3.

Construction-side audit. We report storage correctness through three mutually exclusive proportions for supporting, redundant, and hallucinated entries, together with evidence-turn coverage R. The main table also reports the registered profiling and retrieval readouts for this track; their formulas and construction rules are given in Appendix E.2.

Closed-loop end-to-end. Closed-loop evaluation compares endpoint quality across matched streams with and without write-back. The reported late-minus-early difference and the write-back comparison are interpreted separately, without forming a composite closed-loop score. Appendices E.2–E.3 detail the write-back ablation difference and the terminal and iterative-degradation readouts.

6.3 Main Results

Tables 1–3 report the consolidated results for each track. We use the full available evaluation pool for each track; invalid or unavailable outputs remain visible through the reported denominators. Bold marks the best value in each column, and underlining marks the second-best. Lower values are better for redundant and hallucinated entries, whereas higher values are better for all other metrics. Reference rows such as the human-reference and full-storage controls are not ranked.

6.3.1 Response-Generation Results

Table 1 separates response quality from situational control. On Tier 1, PSyMod scores 4.084 on ABCD, slightly below the prompt-only condition at 4.170. It scores higher on both ESC aggregates, reaching 3.748 and 3.659 compared with 3.703 and 3.644. More importantly, PSyMod obtains the best pooled L2 value of 0.590 and the best P and M axes, while the baselines remain weaker on the same readouts. Thus, the main gain lies in stronger alignment between generated behavior and the registered situation variables rather than uniformly higher surface quality. The similar response lengths of PSyMod and the prompt-only condition also make verbosity an unlikely explanation for this difference. The human-reference row is a calibration point, not a strict upper bound for every derived axis. In particular, a higher P-axis correlation for PSyMod means closer agreement with the frozen annotation under this protocol; it does not mean that PSyMod is more human-like than the human responses.

Table 1: Response quality and situation-specific modulation across generation systems.

System	Response quality				Tier-2 modulation			
	ABCD mean	ESC Overall	ESC six-dim	Mean len.	L2 ov.	E axis	P axis	M axis
Human reference	4.951	4.486	4.143	–	0.722	0.700	0.692	0.859
MemoryLLM	1.452	1.060	1.117	485	0.477	0.529	0.586	0.198
CharacterGLM-6B	1.696	1.486	1.675	1077	0.487	0.521	0.613	0.178
SOTOPIA-π	2.341	1.785	1.961	82	0.383	0.356	0.526	0.167
Sotopia-RL GRPO	2.904	2.761	2.871	91	0.502	0.551	0.661	0.179
Sotopia-RL SFT	2.838	2.430	2.584	59	0.462	0.535	0.591	0.166
PsyLLM-8B	2.928	3.066	3.168	96	0.532	0.603	0.690	0.187
SocialRL	3.087	2.630	2.747	52	0.478	0.564	0.630	0.164
PERSONA	3.196	2.685	2.775	54	0.483	0.576	0.646	0.166
Emotional RAG	3.903	3.475	3.434	38	<u>0.563</u>	**0.681**	<u>0.709</u>	0.177
PHASE-Tree	2.934	1.822	2.084	17	0.404	0.496	0.546	0.149
prompt-only	**4.170**	<u>3.703</u>	<u>3.644</u>	29	0.560	0.653	0.701	<u>0.212</u>
PSyMod	<u>4.084</u>	**3.748**	**3.659**	33	**0.590**	<u>0.677</u>	**0.715**	**0.253**

6.3.2 State-Construction Results

Table 2 shows a deliberate coverage–selectivity trade-off in state construction. PSyMod has no hallucinated entries and reaches the strongest behavioral and emotional profiling values, while its tendency score is close to the best system. Its evidence coverage of 92.3% is below the full-storage and MemInsight controls, but those controls reach that coverage by keeping a larger share of redundant entries. PSyMod therefore favors selective storage rather than maximal recall. Retrieval remains competitive. Its Hit@3 of 11.9% is just below RGMem at 0.124, while MRR is lower than RGMem and Mem0, indicating that correct entries are often retrieved but not always at the first rank. This trade-off is obtained at a cost between the inexpensive baselines and the much more expensive MemInsight.

Table 2: Storage correctness, evidence coverage, profiling, retrieval, and cost across state-construction systems.

System	Sup.	Red.	Hal.	Cov. R	Beh.	Verb.	Emo.	Tend.	Hit@3	MRR	Cost
Full-storage control	0.311	0.689	0.000	1.000	–	–	–	–	–	–	0
Mem0	**0.459**	0.476	0.065	0.330	0.071	0.283	0.048	0.01	0.118	0.090	11.6
MemInsight	0.311	0.689	**0.000**	**1.000**	0.307	**1.000**	0.208	**2.63**	0.079	0.070	66.8
AdaMem	0.371	0.481	0.148	0.837	0.318	0.000	0.000	0.71	0.202	0.180	16.5
RGMem	0.413	**0.465**	0.121	0.946	0.140	0.002	0.000	0.97	**0.124**	**0.105**	1.0
PSyMod	0.366	0.635	**0.000**	0.923	**0.355**	**1.000**	**0.214**	2.57	0.119	0.086	∼22

6.3.3 Closed-Loop Evaluation

Write-back saves the state updates produced during a dialogue for use in later turns, whereas the no-write-back condition does not retain those updates. Table 3 reports a directional closed-loop result rather than a broad performance claim. With live write-back, PSyMod improves from 2.614 to 2.861, a gain of $+0.247$. The matched no-write-back condition also improves, but by a smaller $+0.154$. The gap in late-probe ABCD is small, yet the same-stream comparison is consistent with write-back being helpful in this batch. In contrast, HyperTrace and PGMem decline from early to late probes. Because this analysis uses 24 anchors, we treat it as evidence against immediate degradation and as support for the write-back hypothesis, not as a definitive estimate of long-horizon stability.

Table 3: Early and late response quality for closed-loop systems under matched write-back conditions.

System	ABCD early at t_0	ABCD late at t_{end}	Diff. late−early
Human reference	5.000	5.000	$+0.000$
HyperTrace	2.818	2.639	-0.179
PGMem	2.545	2.306	-0.240
PSyMod	2.614	2.861	$+0.247$
PSyMod no write-back	2.568	2.722	$+0.154$

6.4 Ablation Studies

The ablations test whether the observed behavior depends on the proposed intermediate representations, the access rules that connect them to generation, and the construction-side selection mechanisms. Each ablation removes or replaces one component while keeping the evaluation pool, generation model, review model, and remaining pipeline fixed. The three groups therefore target distinct interfaces rather than providing a single undifferentiated robustness check.

6.4.1 Generation-Side Ablations

The variants in Table 4 change one part of the intermediate representation at a time. Without sub-dimension aggregation, evidence is no longer merged into the estimated personality profile. Without the behavior-level table, numeric behavior levels are replaced by free-text descriptions. The three-level variant reduces the original five behavior levels to three. Finer state structure primarily benefits the headline generation scores. Removing sub-dimension aggregation lowers ABCD from 3.949 to 3.898 and ESC from 3.593 to 3.425, while leaving the pooled L2 value nearly unchanged. Removing the behavior-level table produces a smaller quality drop but a clearer L2 reduction from 0.582 to 0.566. Reducing five levels to three

Table 4: Response quality and situational modulation after changes to the intermediate state representation.

Variant	ABCD	ESC	L2 E	L2 P	L2 M	L2 pooled
Full system	**3.949**	**3.593**	0.667	**0.712**	0.252	0.582
Without sub-dimension aggregation	3.898	3.425	**0.669**	0.681	0.264	0.580
Without the behavior-level table	3.922	3.509	0.654	0.694	0.252	0.566
Three levels	3.890	3.444	0.654	0.711	**0.266**	**0.587**

lowers the response-quality scores and changes the balance across the three axes. These results suggest that the intermediate representations contribute both to response quality and to situation-sensitive behavior, although the effects are not uniform across axes.

The variants in Table 5 isolate the pathways that pass state information to generation. Without P projection, the personality projection pathway is removed. Without E rate, the emotional-expression rate pathway is removed. Without M salience, the memory-salience pathway is removed. Without situation read-out, the current situation state no longer feeds the access layer. The recency variant ranks memories by recency instead of salience. Removing P projection reduces the P-axis score from 0.712 to 0.687 and the pooled L2 score from 0.582 to 0.562. Removing the E rate or M salience similarly lowers the corresponding modulation signals, with M salience producing the largest pooled decline to 0.538. The no-situation-read-out variant improves the L2 columns but lowers ABCD and ESC, indicating that stronger situation correlation alone does not guarantee better responses. The full system therefore provides the best balance between response quality and modulation fidelity.

Table 5: Response quality and situational modulation after removing or replacing access-layer pathways.

Variant	ABCD	ESC	L2 E	L2 P	L2 M	L2 pooled
Full system	**3.949**	**3.593**	0.667	0.712	0.252	0.582
Without P projection	3.941	3.541	0.628	0.687	0.242	0.562
Without E rate	3.904	3.430	0.630	0.702	0.245	0.566
Without M salience	3.897	3.354	0.634	0.674	0.215	0.538
Without situation read-out	3.882	3.412	**0.675**	**0.718**	**0.260**	**0.592**
Recency replaces salience order	3.918	3.465	0.621	0.713	0.238	0.569

6.4.2 Construction-Side Ablations

The variants in Table 6 remove distinct construction components. Without the salience gate, low-salience entries are no longer filtered before storage. Without the encoded situation state, situation information is removed from the constructed state used for retrieval. Removing the salience gate increases Hit@3 from 0.119 to 0.135, but lowers the supported fraction from 0.366 to 0.311, showing that retrieval gain comes with less selective storage. Removing the encoded situation state instead reduces Hit@3 to 0.089 and MRR from 0.086 to 0.070, while leaving the supported fraction unchanged. The salience gate changes storage selectivity, whereas the encoded situation state affects retrieval of retained evidence.

Table 6: Storage support and retrieval after removing construction-side selection or situation encoding.

Variant	Hit@3	Supported	MRR
Full system	0.119	**0.366**	0.086
Without salience gate	**0.135**	0.311	**0.096**
Without the encoded situation state	0.089	**0.366**	0.070

7 Conclusion

PSyMod uses the user's current psychological situation to guide memory retrieval, personality expression, and response planning. It constructs memory, personality, and a historical situational reference from preceding dialogue, then keeps this persistent state read-only during each target turn. Across three annotated corpora, we evaluate conditioned generation, state construction, and closed-loop operation separately. Response assessment uses a review model distinct from the generation model; model-assisted reference annotations remain part of the evaluation protocol.

In conditioned generation, PSyMod exceeds the prompt-only condition on ESC and situational modulation consistency, especially on the personality and memory axes, but has a slightly lower blind four-dimensional review score. The state-construction audit finds no hallucinated stored entries and competitive retrieval, alongside a coverage–selectivity trade-off. In a matched closed-loop probe, live write-back yields a larger late-minus-early gain than the same pipeline without write-back; this is directional evidence, not a demonstration of long-horizon stability. Ablations show that the components affect response quality, modulation, and storage selectivity in different directions rather than improving every readout uniformly.

Future work should extend situation modulation to multi-user and group dialogues, evaluate state evolution over longer horizons, and test transferability across languages and domains.

REFERENCES

Mohammad Amanlou, Parham Abed Azad, Farbod Davoodi, Mostafa Masumi, Behnam Bahrak, and Abdol-Hossein Vahabie. Psychoagent: An affect-sensitive cognitive architecture for conflict-aware memory in LLM agents. *CoRR*, abs/2608.07438, 2026. doi: 10.48550/ARXIV.2608.07438. URL https://doi.org/10.48550/arXiv.2608.07438.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Élteto, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Ji-An Li, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo G. Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Natalia Scharfenberg, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert C. Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. Centaur: a foundation model of human cognition. *CoRR*, abs/2410.20268, 2024. doi: 10.48550/ARXIV.2410.20268. URL https://doi.org/10.48550/arXiv.2410.20268.

Aili Chen, Chengyu Du, Jiangjie Chen, Jinghan Xu, Yikai Zhang, Siyu Yuan, Zulong Chen, Liangyue Li, and Yanghua Xiao. DEEPER insight into your user: Directed persona refinement for dynamic persona modeling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 24157–24180. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1177. URL https://doi.org/10.18653/v1/2025.acl-long.1177.

Tiantian Chen, Jiaqi Lu, Ying Shen, and Lin Zhang. Es-memeval: Benchmarking conversational agents on personalized long-term emotional support. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz (eds.), *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pp. 5810–5821. ACM, 2026a. doi: 10.1145/3774904.3792143. URL https://doi.org/10.1145/3774904.3792143.

Xingyu Chen, Rui Wang, Zhaopeng Tu, and Liefeng Bo. Adamem: Learning what to remember for personalized long-horizon llm agents. *arXiv preprint arXiv:2606.21144*, 2026b.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational AI. *CoRR*, abs/2205.14727, 2022. doi: 10.48550/ARXIV.2205.14727. URL https://doi.org/10.48550/arXiv.2205.14727.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Wonjun Choi, Yerim Kim, Yukyung Lee, and Susik Yoon. Pgmem: Tightly coupled persona-memory graph for lifelong personalized agents. *CoRR*, abs/2608.01708, 2026. doi: 10.48550/ARXIV.2608.01708. URL https://doi.org/10.48550/arXiv.2608.01708.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, Bing Qin, et al. Persona: Dynamic and compositional inference-time personality control via activation vector algebra. In *International Conference on Learning Representations*, volume 2026, pp. 151839–151879, 2026.

Xing Fu, Yulin Hu, Mengtong Ji, Haozhen Li, Yixin Sun, Weixiang Zhao, Yanyan Zhao, and Bing Qin. Enpmr-bench: Benchmarking proactive memory retrieval for emotional support

agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 41910–41933. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.2080. URL https://doi.org/10.18653/v1/2026.findings-acl.2080.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *CoRR*, abs/2509.03730, 2025. doi: 10.48550/ARXIV.2509.03730. URL https://doi.org/10.48550/arXiv.2509.03730.

Yupeng Han, Shuochen Liu, Kai Zhang, Ze Liu, Zhihong Pan, and Xianquan Wang. Learning what to share and what to personalize: Hierarchical strategy co-evolution for agent memory. *CoRR*, abs/2608.25329, 2026. doi: 10.48550/ARXIV.2608.25329. URL https://doi.org/10.48550/arXiv.2608.25329.

Tamunotonye Harry, Ivoline C. Ngong, Chima Nweke, Yuanyuan Feng, and Joseph P. Near. Beyond fixed psychological personas: State beats trait, but language models are state-blind. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 26440–26468. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.1316. URL https://doi.org/10.18653/v1/2026.findings-acl.1316.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *CoRR*, abs/2505.15715, 2025. doi: 10.48550/ARXIV.2505.15715. URL https://doi.org/10.48550/arXiv.2505.15715.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. Emotional RAG: enhancing role-playing agents through emotional retrieval. In Huajun Chen, Anna Fensel, Xingquan Zhu, Roger Wattenhofer, and Xindong Wu (eds.), *IEEE International Conference on Knowledge Graph, ICKG 2024, Abu Dhabi, United Arab Emirates, December 11-12, 2024*, pp. 120–127. IEEE, 2024. doi: 10.1109/ICKG63256.2024.00023. URL https://doi.org/10.1109/ICKG63256.2024.00023.

EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. Infusing theory of mind into socially intelligent LLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 11327–11360. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.551. URL https://doi.org/10.18653/v1/2026.findings-acl.551.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle H. Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *CoRR*, abs/2504.14225, 2025. doi: 10.48550/ARXIV.2504.14225. URL https://doi.org/10.48550/arXiv.2504.14225.

Bowen Jiang, Yuan Yuan, Zhuoqun Hao, Yuchen Liu, Maohao Shen, Sihao Chen, Gregory Wornell, Chris Callison-Burch, Lyle H. Ungar, Dan Roth, Qi Guo, Xiangjun Fan, Camillo J. Taylor, and Hanchao Yu. Personamem-v3: Toward omni-platform personal intelligence for holistic user understanding, recommendation, and agentic tasks. *CoRR*, abs/2608.21381, 2026. doi: 10.48550/ARXIV.2608.21381. URL https://doi.org/10.48550/arXiv.2608.21381.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. *arXiv preprint arXiv:1911.09304*, 2019.

Bingrui Jin, Kunyao Lan, Baihan Li, and Mengyue Wu. Twice: Modeling the temporal evolution of personalized user behavior via event-driven agents. *arXiv preprint arXiv:2602.22222*, 2025.

Yonghyun Jun, Junhyuk Choi, Jeonghyun Park, Jihyeong Park, Liu Nicole Geumheon, and Hwanhee Lee. Identifying and mitigating bottlenecks in role-playing agents: A systematic study of disentangling character profile axes. *arXiv preprint arXiv:2601.04716*, 2026.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. REALTALK: A 21-day real-world dataset for long-term conversation. *CoRR*, abs/2502.13270, 2025. doi: 10.48550/ARXIV.2502.13270. URL https://doi.org/10.48550/arXiv.2502.13270.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, Yulia Tsvetkov, and Asli Celikyilmaz. Horizon-bench: Long-horizon personalization with evolving preferences. *CoRR*, abs/2604.17283, 2026. doi: 10.48550/ARXIV.2604.17283. URL https://doi.org/10.48550/arXiv.2604.17283.

Chang Liu, Shuyi Zhang, Changsheng Ma, Yongfeng Tao, Minqiang Yang, and Bin Hu. Fta-mem: Fact-time-affect anchored memory for low-density long-term dialogue. *CoRR*, abs/2608.16303, 2026a. doi: 10.48550/ARXIV.2608.16303. URL https://doi.org/10.48550/arXiv.2608.16303.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, et al. Perma: Benchmarking personalized memory agents via event-driven preference and realistic task environments. *arXiv preprint arXiv:2603.23231*, 2026b.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 3469–3483. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.269. URL https://doi.org/10.18653/v1/2021.acl-long.269.

Yuxin Liu, Mingye Zhu, Siyuan Liu, Bo Hu, and Lei Zhang. Enhancing persona following at decoding time via dynamic importance estimation for role-playing agents. *CoRR*, abs/2603.01438, 2026c. doi: 10.48550/ARXIV.2603.01438. URL https://doi.org/10.48550/arXiv.2603.01438.

Junfeng Lu and Yueyan Li. Dynamic affective memory management for personalized LLM agents. *CoRR*, abs/2510.27418, 2025. doi: 10.48550/ARXIV.2510.27418. URL https://doi.org/10.48550/arXiv.2510.27418.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *CoRR*, abs/2308.08239, 2023. doi: 10.48550/ARXIV.2308.08239. URL https://doi.org/10.48550/arXiv.2308.08239.

Yufei Luo, Xiucheng Xu, and Zhen Yang. Memsif: From structured interactions to dual-track fact memory for LLM agents. *CoRR*, abs/2608.01742, 2026. doi: 10.48550/ARXIV.2608.01742. URL https://doi.org/10.48550/arXiv.2608.01742.

Dan P. McAdams. What Do We Know When We Know a Person? *Journal of Personality*, 63(3): 365–396, 1995. doi: 10.1111/j.1467-6494.1995.tb00500.x.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246, 1995.

Naoki Otani, Jun Araki, HyeongSik Kim, and Eduard H. Hovy. A textual dataset for situated proactive response selection. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.),

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 3856–3874. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.214. URL https://doi.org/10.18653/v1/2023.acl-long.214.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Pin Qian, Su Wang, Yihang Chen, Qiaolin Yu, Xiaoyuan Wang, Zhitong Guo, Zhicheng Wang, and Junxian You. Toward user-conditioned evaluation of personal llm agents under temporal interventions. *arXiv preprint arXiv:2607.21635*, 2026.

Peijun Qing, Fobo Shi, and Soroush Vosoughi. Utilmem: Benchmarking evidence utilization in long-term conversational memory. *arXiv preprint arXiv:2608.30508*, 2026.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1534. URL https://doi.org/10.18653/v1/p19-1534.

John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. The situational eight diamonds: a taxonomy of major dimensions of situation characteristics. *Journal of personality and social psychology*, 107(4):677, 2014.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. Meminsight: Autonomous memory augmentation for LLM agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 33136–33152. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1683. URL https://doi.org/10.18653/v1/2025.emnlp-main.1683.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. Large language models do not simulate human psychology. *arXiv preprint arXiv:2508.06950*, 2025.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276, 2020.

Jianzhi Shen, Keyu Mao, Minghao Shao, Chuanyang Jin, Yusong Wang, Ailiang Lin, Kotaro Funakoshi, Manabu Okumura, Tianmin Shu, and Muhammad Shafique. Hypertrace: Hypothesis-based preference tracing for online llm personalization. *arXiv preprint arXiv:2609.09835*, 2026.

Patricia Cain Smith and Lorne M. Kendall. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47(2): 149–155, 1963. doi: 10.1037/h0047060.

Deepesh Sonar. Lsrep: A longitudinal state-replay protocol for evaluating conversational memory, with ice v2 as an audited local-first architecture. *arXiv preprint arXiv:2609.16730*, 2026.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=1i6ZCvflQJ.

Qixuan Sun, Yue Que, Bowei He, Jin Guo, Dihang Yang, Wenchang Situ, and Chen Ma. Creamem: A scene-aware memory architecture for personalized agents. *arXiv preprint arXiv:2609.08550*, 2026.

Bo Tang, Jianan Yang, Junyi Zhu, Yiquan Wu, Rui Zhao, Zhengyu Yang, Yang Zhang, Feiyu Xiong, Zhiyu Li, and Jiajun Shen. Phase-tree: Modeling character-state evolution in long-horizon role-playing dialogue. *CoRR*, abs/2608.06975, 2026a. doi: 10.48550/ARXIV.2608.06975. URL https://doi.org/10.48550/arXiv.2608.06975.

Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. Facet-level persona control by trait-activated routing with contrastive SAE for role-playing llms. In Raymond Chi-Wing Wong, Hanghang Tong, Hua Lu, James T. Kwok, Flora D. Salim, Yuanfeng Song, and Man Lung Yiu (eds.), *Data Science: Foundations and Applications - 30th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2026, Hong Kong, China, June 9-12, 2026, Proceedings, Part II*, volume 16618 of *Lecture Notes in Computer Science*, pp. 411–423. Springer, 2026b. doi: 10.1007/978-981-92-1947-6_33. URL https://doi.org/10.1007/978-981-92-1947-6_33.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. Memconflict: Evaluating long-term memory systems under memory conflicts. *CoRR*, abs/2605.20926, 2026. doi: 10.48550/ARXIV.2605.20926. URL https://doi.org/10.48550/arXiv.2605.20926.

Sophia Terwiel, John F Rauthmann, and Maike Luhmann. Using the situational characteristics of the diamonds taxonomy to distinguish sports to more precisely investigate their relation with psychologically relevant variables. *Plos one*, 15(10):e0241013, 2020.

Ao Tian, Yunfeng Lu, Xinxin Fan, Changhao Wang, Lanzhi Zhou, Yeyao Zhang, and Yanfang Liu. Rgmem: Renormalization group-based memory evolution for language agent user profile. *CoRR*, abs/2510.16392, 2025. doi: 10.48550/ARXIV.2510.16392. URL https://doi.org/10.48550/arXiv.2510.16392.

Heng Wang, Yifei Li, Lingling Zhang, Pengyu Li, Xinyu Che, Xinyu Zhang, and Zesheng Yang. Qumem: Personalized memory for query-conditioned user-state inference in LLM agents. *CoRR*, abs/2608.16168, 2026a. doi: 10.48550/ARXIV.2608.16168. URL https://doi.org/10.48550/arXiv.2608.16168.

Hexi Wang, Yujia Zhou, Bangde Du, Weihang Su, Xinyuan Cao, Qingyi Pan, Qingyao Ai, Yueyue Wu, Min Zhang, and Yiqun Liu. Mitigating identity essentialism in llm agents with longitudinal life trajectories. *arXiv preprint arXiv:2608.19621*, 2026b.

Jianing Wang, Xintao Wang, Aili Chen, Jie Shi, Hongcheng Guo, Jun Gao, Wenxuan Zhao, Chengkun Lang, Yuanli Guo, and Yanghua Xiao. Socialrl: Refining llms' social intelligence through multi-turn reinforcement learning and reward design. *arXiv preprint arXiv:2609.09764*, 2026c.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 23221–23235. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.1192. URL https://doi.org/10.18653/v1/2025.findings-acl.1192.

Ming Wang, Peidong Wang, Xiaocui Yang, Daling Wang, Shi Feng, Fiona Fui-Hoon Nah, and Ee-Peng Lim. Do AI personas grow? analyzing and benchmarking personality evolution in LLM agents after life events. *CoRR*, abs/2608.06485, 2026d. doi: 10.48550/ARXIV.2608.06485. URL https://doi.org/10.48550/arXiv.2608.06485.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12912–12940. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.698. URL https://doi.org/10.18653/v1/2024.acl-long.698.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1840–1873. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.ACL-LONG.102. URL https://doi.org/10.18653/v1/2024.acl-long.102.

Xintao Wang, Jian Yang, Weiyuan Li, Rui Xie, Jen-Tse Huang, Jun Gao, Shuai Huang, Yueping Kang, Liyuan Gou, Hongwei Feng, and Yanghua Xiao. Humanllm: Benchmarking and improving LLM anthropomorphism via human cognitive patterns. *CoRR*, abs/2601.10198, 2026e. doi: 10.48550/ARXIV.2601.10198. URL https://doi.org/10.48550/arXiv.2601.10198.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian J. McAuley. MEMORYLLM: towards self-updatable large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50453–50466. PMLR / OpenReview.net, 2024c. URL https://proceedings.mlr.press/v235/wang24s.html.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45:1191–1207, 2013.

Zesheng Wei, Mengxiang Li, Zilei Wang, and Yang Deng. Beyond static personas: Situational personality steering for large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 19185–19210. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.958. URL https://doi.org/10.18653/v1/2026.findings-acl.958.

Deliang Wen, Ke Sun, and Yu Wang. A-mber: Affective memory benchmark for emotion recognition. *arXiv preprint arXiv:2604.07017*, 2026.

Zhiyuan Wen, Jiannong Cao, Jiaxing Shen, Ruosong Yang, Shuaiqi Liu, and Maosong Sun. Personality-affected emotion generation in dialog systems. *ACM Trans. Inf. Syst.*, 42(5):134:1–134:27, 2024. doi: 10.1145/3655616. URL https://doi.org/10.1145/3655616.

Shuyao Xiao, Shengling Wang, Xuan Chen, Ke Chao, Ming Cui, Feifei Qian, Fanlin Meng, Chaoyang Mei, Chaoyong Jiang, Qi Ouyang, and Junxi Yi. When errors become memories: Causal pathway tracing in multi-turn memory-augmented llms. *CoRR*, abs/2608.30198, 2026. doi: 10.48550/ARXIV.2608.30198. URL https://doi.org/10.48550/arXiv.2608.30198.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 852–861, 2023.

Tianyu Yang, Sudipta Paul, Vijay Srinivasan, Vivek Kulkarni, and Srinivas Chappidi. TRUSTMEM: learning trustworthy memory consolidation for LLM agents with long-term memory. *CoRR*, abs/2606.25161, 2026. doi: 10.48550/ARXIV.2606.25161. URL https://doi.org/10.48550/arXiv.2606.25161.

Bingsheng Yao, Bo Sun, Yuanzhe Dong, Yuxuan Lu, and Dakuo Wang. DPRF: A generalizable dynamic persona refinement framework for optimizing behavior alignment between personalized LLM role-playing agents and humans. *CoRR*, abs/2510.14205, 2025. doi: 10.48550/ARXIV.2510.14205. URL https://doi.org/10.48550/arXiv.2510.14205.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. EmoHarbor: Evaluating personalized emotional support by simulating the user's internal world. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1176–1202, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.53. URL https://aclanthology.org/2026.acl-long.53/.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *CoRR*, abs/2508.03905, 2025. doi: 10.48550/ARXIV.2508.03905. URL https://doi.org/10.48550/arXiv.2508.03905.

Jiongchi Yu, Yuhan Ma, Xiaoyu Zhang, Junjie Wang, Qiang Hu, Chao Shen, and Xiaofei Xie. PTCBENCH: benchmarking contextual stability of personality traits in LLM systems. *CoRR*, abs/2602.00016, 2026. doi: 10.48550/ARXIV.2602.00016. URL https://doi.org/10.48550/arXiv.2602.00016.

Alice Zhang, Skanda Muralidhar, Daniel Gatica-Perez, and Mathew Magimai-Doss. Assessment of personality dimensions across situations using conversational speech. *CoRR*, abs/2507.19137, 2025. doi: 10.48550/ARXIV.2507.19137. URL https://doi.org/10.48550/arXiv.2507.19137.

Qi Zhang, Shen Huang, Chu Liu, Shouqing Yang, Junbo Zhao, Haobo Wang, and Pengjun Xie. Deltamem: Towards agentic memory management via reinforcement learning. *CoRR*, abs/2604.01560, 2026. doi: 10.48550/ARXIV.2604.01560. URL https://doi.org/10.48550/arXiv.2604.01560.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2204–2213. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1205. URL https://aclanthology.org/P18-1205/.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, et al. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 15785–15810, 2024.

Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong, and Zhiyu Li. Inside out: Evolving user-centric core memory trees for long-term personalized dialogue systems. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 13429–13446. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.614. URL https://doi.org/10.18653/v1/2026.acl-long.614.

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. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. Characterglm: Customizing chinese conversational AI characters with large language models. *CoRR*, abs/2311.16832, 2023. doi: 10.48550/ARXIV.2311.16832. URL https://doi.org/10.48550/arXiv.2311.16832.

Yuhang Zhu, Mingxuan Du, Benfeng Xu, Jie Gao, Lingyun Yu, and Hongtao Xie. Beyond borrowed histories: Person-aligned user simulation for interactive role-playing evaluation. *CoRR*, abs/2607.27816, 2026. doi: 10.48550/ARXIV.2607.27816. URL https://doi.org/10.48550/arXiv.2607.27816.

A Supplementary Details of Related Work

Correspondence: Section 2 (Related Work) of the main text.

The four paragraphs of the main text have already given the lineage, phenomena, and gaps of each line; this appendix supplements the specific mechanisms of the works named in the main text, for checking the difference between this paper and its nearest neighbors.

A.1 Personality simulation and evolution. Two lineage sources: Generative Agents (Park et al., 2023) places 25 agents in a sandbox town and organizes behavior with a three-layer architecture of "timestamped memory stream — periodic reflection — daily plan," with credibility evaluated through behavioral interviews; its memory serves open-ended social simulation, carries no psychometric dimensions, and is not anchored to real corpora. CoALA (Sumers et al., 2024) is a taxonomic framework rather than a concrete model: it decomposes language agents into modular memory (working/episodic/semantic/procedural), a structured action space (internal memory operations and external environment actions), and a generalized decision loop, used to retrospectively organize existing work and prospect directions; the psychological meanings and interactions of the four memory types are not developed therein. AnnaAgent (Wang et al., 2025) simulates help-seekers across sessions; DEEPER (Chen et al., 2025) refines personality from behavior–prediction discrepancies; state-machine approaches (Wei et al., 2026) represent latent personality transitions; InCharacter (Wang et al., 2024b) and RP Bottlenecks (Jun et al., 2026) examine which profile dimensions contribute to role fidelity. State changes in such systems are realized through prompting, memory operations, or fine-tuning, and are not grounded in established psychometric scales or norms.

The longitudinal simulation and profile refinement line: Centaur (Binz et al., 2024) is a human cognitive foundation model fine-tuned on large-scale human behavioral experiment data (tens of millions of choice trials), predicting human choices in cognitive tasks—its modeling object is population-level cognition, not individual personality in dialogue, and it does not involve situation–memory–emotion interaction. LifeMem (Wang et al., 2026b) targets "identity essentialism" in LLM social simulation: static-profile agents exhibit stronger between-group separation and within-group compression than real people, with demographic labels treated as individual traits; it instead constructs agents from longitudinal life trajectories so that behavior differentiates as experiences accumulate. TWICE (Jin et al., 2025) performs time-anchored user simulation: structured profile + event-driven memory organized around life events and behavioral transitions + a two-stage pipeline (first event-anchored content planning, then personalized style adaptation), measuring authenticity, consistency, and human-likeness on longitudinal social-media data; its memory is organized by time and carries no psychometrically meaningful situational encoding. DPRF (Yao et al., 2025) iteratively identifies cognitive biases between generated behavior and the target individual's ground truth (free-form or theory-driven structured analysis) and revises the persona profile accordingly to reduce the bias; its revision object is the alignment between profile text and behavior, without situational modulation or memory mechanisms. All four target "individuation of the individual over time," but their paths are respectively behavioral data fitting, trajectory accumulation, event-driven stylization, and profile revision; none models how the situation within a single turn determines which states take effect.

Concrete readings from the empirical side: Do AI Personas Grow? (Wang et al., 2026d) reports that the magnitude of personality-reading change across major life events is weak and less differentiated than human patterns; Schröder et al. (2025) show that mechanism-free models generalize through surface similarity rather than the deep structure of human behavior. Chameleon (Harry et al., 2026)

provides 5,001 contextualized psychological profiles of 1,667 Reddit users (each person measured across multiple contexts); variance decomposition shows 74% within-person (state) variance versus only 26% between-person (trait) variance, and finds that LLMs are "state-blind"—responding only to traits and indiscriminately to states—with reward models even giving directionally opposite preferences for the same user's states; Personality Illusion (Han et al., 2025) reveals the dissociation ("personality illusion") between LLM self-report scales and behavioral tasks, and that trait profiles drift with prompting style. What the two works provide are empirical conclusions—"state variance dominates, self-report is unreliable"—stopping at diagnosis without giving generation mechanisms; this paper turns the former into a situation-covariation scoring track (Tier 2) and the latter into the grounds for adopting behavioral anchors rather than self-report readings.

A.2 Psychological situations. Empirical details of UbImpressed (Zhang et al., 2025): the same participants present different perceived-personality profiles under neutral-interview and hostile-customer-service situations, with stressful situations preferentially eliciting neuroticism-related responses. HumanLLM (Wang et al., 2026e) synthesizes scenarios with the DIAMONDS taxonomy, derives personality patterns from psychological scales with annotator validation, but its psychological grounding operates at the pattern-definition level: response scoring relies on an LLM judge, and the S8*/SF-24 instruments and human consensus norms do not enter the scoring standard. ToMA (Hwang et al., 2026) generates turn-level hypotheses about interlocutors' mental states and fine-tunes on goal-filtered state–utterance pairs; its modeling object is interlocutor mental state rather than the psychological properties of a situation, and its validity criterion is goal achievement rather than a human-referenced standard.

Three works bring situations into generation control, each acting at a different level: IRiS (Wei et al., 2026) first proves with persona-neuron analysis that LLM personality exhibits situation-dependent and stable situation–behavior patterns, then proposes a training-free neuron-level "identify–retrieve–manipulate" framework—situated personality neuron identification, situation-aware retrieval, similarity-weighted activation manipulation; its situation control happens in the activation space of a single model, is single-turn manipulation without persistent state, does not enter memory or emotion, and does not enter scoring. PDD (Liu et al., 2026c) cites the same CAPS theory as this paper (personality's influence on behavior varies with the scene), dynamically estimates the importance of each personality attribute in the current scene via conditional mutual information at decoding time, and guides decoding with importance weights; its point of action is the token decoding layer, it constructs no auditable frozen psychological state, and it has no memory or emotion dimension. Trait-Activated Routing (Tang et al., 2026b) uses a sparse autoencoder (SAE) trained with contrastive learning to identify trait-related activation directions and routes Big Five facet expression by situational cues, addressing the problems that prompt/RAG personality signals dilute and drift in long dialogues and that SFT requires per-role retraining; its "situation" is a routing cue rather than a measurable psychological property, and no human norms enter. What the three share is treating the situation as a control signal or statistical condition: none models it as an eight-dimensional reading with measurement norms, nor characterizes the full action pathway from situation to emotion and memory.

A.3 Memory and cross-dimension coupling. PersonaTree (Zhao et al., 2026) uses reinforcement learning to select memory operations, but personality is a relatively static key rather than a slowly changing state shaped by memory deposits. HiPS (Han et al., 2026) retains alternative memory versions, marks one as void with an explanation, but does not specify update dynamics acting on personality. MemConflict (Tao et al., 2026) examines several conflict types and protects stable traits; ErrorsBecomeMemories (Xiao et al., 2026) shows that latent memory errors can be elicited by probes. CreaMem (Sun et al., 2026) organizes memories by scenes such as life, work, and interest, but scene functions primarily as a classification key; different psychological situations within the same domain remain indistinguishable. AdaMem (Chen et al., 2026b) reports a gap between "learning a memory-writing strategy" and "executing the strategy reliably," suggesting that memory salience cannot be delegated entirely to strategy text. Two general memory architectures: Mem0 (Chhikara et al., 2025) dynamically extracts, consolidates, and retrieves salient information from ongoing dialogue, with a graph-structured variant organizing dialogue elements by relational edges, compared against six baseline categories on LoCoMo—write/update/delete operations are explicitly recorded, but memory is organized by factual salience without situational vectors or personality dimensions; this paper adopts one point from it—"explicit operation records are auditable"—and the

construction-side audit (E.2) follows that idea. Five works dedicated to memory operations each occupy one link: TRUSTMEM (Yang et al., 2026) targets the problem that writes/revisions/deletions may miss, contaminate, or introduce hallucinated content, using a memory-transition verifier to review the credibility of state transitions before consolidation; DeltaMem (Zhang et al., 2026) recasts persona memory management as a single-agent end-to-end task, training operation-level update decisions with reinforcement learning (synthetic user–assistant dialogues with operation labels), addressing information loss and cross-scene fragility of multi-agent frameworks; MemSIF (Luo et al., 2026) diagnoses two misalignments—temporal proximity does not equal topical relevance (TSM), and write-time salience does not equal future query utility (DUM)—and accordingly converts structured interactions into dual-track factual memory; UTILMEM (Qing et al., 2026) is a memory-utilization benchmark (1,717 instances, five domains) that separately measures "integrating scattered, implicit, noisy evidence across long histories" from single-point fact recall; LSREP (Sonar, 2026) is a longitudinal state-replay protocol that audits how persistent states accumulate, age, and absorb revisions via ordered replay, lifecycle scheduling, repeated probes, and evolving reference answers. The five respectively cover consolidation credibility, operation learning, structural conversion, utilization evaluation, and longitudinal auditing—none contains a situational dimension, and their probes and scoring references come from the tasks themselves rather than human psychological norms. Emotion–memory systems (DAM-LLM Lu & Li, 2025, FTA-Mem Liu et al., 2026a, PsychoAgent Amanlou et al., 2026) incorporate affective information into memory updating, labeling, narration, or retrieval, but neither connect memory operations to human consolidation principles (such as arousal weighting and mood-congruent retrieval) nor establish a complete feedback path from memory to behavior generation. (AMBER is an emotion-recognition evaluation benchmark for affective memory rather than a memory system; it is classified under A.4.) Overall, existing integration efforts remain predominantly pairwise: personality, emotion, memory, and situation are coupled two at a time or evaluated as separate dimensions, lacking a unified account of their interactions together with a joint evaluation protocol.

A.4 Evaluation and reference standards. The measurement objects and reference sources of the benchmarks are as follows: PersonaMem-v3 (Jiang et al., 2026) evaluates personality evolution, restraint, and hidden motives, covering cross-platform personal intelligence; ESMemEval (Chen et al., 2026a) targets long-term emotional-support scenarios, measuring personalized memory when user information is scattered, implicit, and continuously evolving, with reference to preset support-quality dimensions; ENPMR-Bench (Fu et al., 2026) measures emotion-need-aware proactive memory retrieval—memory is no longer only fact retrieval but must be proactively brought out after inferring latent emotional needs, with reference to expert-annotated emotion-need–memory pairs; AMBER (Wen et al., 2026) (A-MBER) is an emotion-recognition benchmark for affective memory, testing whether the system can correctly recognize the current emotion from historical memory—an evaluation benchmark rather than a memory system (the main text Related Work classifies it the same way); EmoHarbor (Ye et al., 2026) drives personalized emotional-support evaluation with simulated user internal states, and its primary evaluation loop lacks human scoring; HumanLLM (Wang et al., 2026e) reports limited score resolution for pattern-level expression evaluation; HorizonBench (Li et al., 2026) defines long-horizon personalization: preferences evolve across months, requiring inference of when a life event rewrote prior preferences, with a data generator carrying ground-truth provenance; PTCBench (Yu et al., 2026) quantifies the stability of personality traits under controlled situations—repeated measurement with psychological scales under 12 external conditions (locations and life events)—and its stance that "traits should be stable across situations" is exactly opposite to this paper's scoring direction that "behavior should vary regularly with situations," serving as a counterpoint to Tier 2; PERMA (Liu et al., 2026b) separates retrieval from generation along an event-driven preference timeline. Regarding reference paradigms, the above benchmarks respectively adopt model judges, simulated users, and psychological scales; Temporal-Interventions (Qian et al., 2026) audits longitudinal evaluation protocols and finds that none of the audited protocols simultaneously satisfies continuous intervention, cross-dimensional consistency, and human-referenced evaluation. The common difference between this paper and all of them is that the scoring standard values do not come from models: they are all taken from scale norms and numerical regularities of published human studies (Section 6.1 of the main text; Appendix E.2).

B SUPPLEMENTARY CLAUSES OF THE TASK SETTING

Correspondence: Section 3 (Task Setting) of the main text.

Section 3 of the main text has given the complete formula chain (state construction, response generation, closed-loop coupling); the following two clauses are supplementary.

1. **Generation-side logging.** In addition to outputting y_t, the response generation component logs the measurements of each processing stage, the retrieved entries, the grounds of each processed instruction item, and the complete final prompt; these records support the construction-side audit (Appendix E.2).
2. **The mediated nature of the closed loop.** The closed loop is mediated by observable dialogue only: what re-enters construction after a generated turn is the response text itself, and it is bound by the same visibility boundary as any human utterance. There is therefore no write-back channel from generation into the state beyond the dialogue stream, in contrast to designs that write generated content back into the state directly.

C DATASET SUPPLEMENTARY DETAILS

Correspondence: Section 4 (Dataset) of the main text.

C.1 CONVERSION OF THE FOUR-DIMENSIONAL REPRESENTATIONS INTO THE GENERATION PROCESS

Section 4 (Dataset) of the main text has defined the P/E/S/M representations; entering the generation process, the four types undergo different conversions (bridging Section 4 and Section 5, and stated in neither): the eight S scores are converted into situational percentiles and the situation type; the E VAD vector is converted into an emotional-expression constraint; the P profile is converted into behavioral levels and behavioral anchors; M provides past events and historical coping exemplars through the memory interface. Retrieval rules belong to the method layer (Appendix D.5); the data representation layer does not preset retrieval weights.

One boundary is declared: the current construction does not introduce emotion-vector encoding, and this paper makes no claim of retrieval by emotional similarity.

C.2 SITUATIONAL ANNOTATION DETAILS

The main text has given the two-step pipeline, the three scoring models (qwen3.8-max, claude-opus-4.6, gpt-5.5), temperature 0, retention when at least two models agree, and manual review; the following supplements:

- Scale source: DIAMONDS SF-24 (original form: Terwiel et al., 2020), three items per dimension over the eight dimensions, each item rated 1–7, with the dimension score the mean of the three items; item wording follows the original scale (full text in C.3). The situational percentiles and the situation-type gates (D.3, D.5) are taken within the consensus situation profile reported by the same source, which was collected in a non-dialogue situation domain; the resulting percentiles are norm-relative and are not directly comparable across corpora.
- Scoring discipline: second-step scoring is based solely on the model's own generated situational description, without looking back at the original lines.
- Reliability: the post-calibration ICC of the situational annotations averages 0.75 across the eight dimensions, with per-dimension values between 0.57–0.93.

C.3 FULL TEXT OF THE SF-24 SITUATION ITEMS

The following are the 24 items actually used by the situation annotation pipeline, in the order in which they are presented to the scoring models.

Duty.

1. A task needs to be done.
2. Task-oriented thinking is required.
3. I have to fulfill my duties.

Intellect.

4. The situation contains intellectual stimuli.
5. There is the opportunity to demonstrate intellectual capacities.
6. Information needs to be deeply processed.

Adversity.

7. I am being criticized.
8. I am being blamed for something.
9. I am being threatened by someone or something.

Mating.

10. Potential sexual or romantic partners are present.
11. Physical attractiveness is relevant.
12. The situation is sexually charged.

Positivity.

13. The situation is pleasant.
14. The situation is playful.
15. The situation is joyous and exuberant.

Negativity.

16. The situation could elicit stress.
17. The situation could elicit feelings of tension.
18. The situation could entail frustration.

Deception.

19. It is possible to deceive someone.
20. Someone in this situation could be deceptive.
21. Not dealing with others in an honest way is possible.

Sociality.

22. Close personal relationships are important or can develop.
23. Others show many communicative signals.
24. Communication with other people is important or desired.

C.4 Four-Dimensional Coverage Statistics

Table C1 gives the statistical profiles of the three evaluation protocols; the main text reports the pack-level profile only. Coverage is reported as a proportion, so the count of evaluation units is not repeated as a numerator.

Table C2 breaks the coverage proportions of Table C1 down by corpus source; the P/E/S/M proportions of the $\mathcal{D}_{\text{full}}$ row are the total row of this table, and each proportion uses that source's own evaluation units as its denominator.

Table C1: Statistical profiles of the three evaluation protocols. Coverage is the proportion of packs carrying an annotation for that dimension, reported in the order P/E/S/M.

Protocol	Count	History turns (mean)	Scenarios	Memory span (turns, mean)	P/E/S/M coverage
$\mathcal{D}_{\text{gen}}$	720 generation evaluation packs	13.5	462	616.2	1.000/0.808/1.000/1.000
$\mathcal{D}_{\text{con}}$	676 construction evaluation packs	66.8	1,469	36.0	—
$\mathcal{D}_{\text{full}}$	284 character packs	836.7	13,473	636.2	0.621/0.440/0.801/0.223

Table C2: Four-dimensional coverage by corpus source. P/E/S/M are proportions of each source's own evaluation units, whose counts are given in the main text.

Source	P	E	S	M
CPED	1.000	1.000	0.996	0.094
Friends	0.348	0.266	0.660	0.319
RPC	1.000	0.000	1.000	0.078
Total	0.621	0.440	0.801	0.223

D METHOD SUPPLEMENTARY DETAILS

Correspondence: Section 5 (Method) of the main text.

D.1 ALGORITHM 1 — RESPONSE GENERATION

Section 5.2 of the main text has described the complete flow in prose; Algorithm 1 gives the corresponding formal control flow, with symbols self-containedly defined within the algorithm.

Three notes on the algorithm: (1) inference-time measurement uses a single runtime model, distinct from the multi-model agreement pipeline of the annotation stage (Appendix C.2); (2) when emotion is unavailable, E_t remains empty and must not be inferred from other representations; (3) the situation type c_t is read off the percentile-transformed readings by the fixed rule table defined in D.3.

D.2 INFERENCE AND PROJECTION STAGES, COMPLETE VERSION

Sections 5.1 and 5.2 of the main text describe the state-construction and generation stages in prose; this table sets out the whole chain and adds the four deterministic projection stages that the response-generation side performs on the frozen profile.

The symbols $x_{c,k,j}$ (the j-th sentence-level observation on concern dimension k in session c), $X_{c,k}$ (the within-session aggregate), and $P_{p,d}$ (the cross-session mean mapped to Big Five dimension d) make explicit the two-stage aggregation used in Section 5.1 of the main text; they are used only within this table.

D.3 PROCESS RULES

Section 5.2 of the main text has stated the flow, the functional division of the representations, and the priority order; the following are the specific rules of each step.

Situation type. Each of the eight dimensions of the current reading is converted to a percentile on the reference scale (D.5), and the situation type c_t is read off those percentiles by a fixed read-only rule table whose gates are the high and low percentiles 0.80 and 0.20, with two rules additionally using a 0.6 mid boundary. The table defines eight named types plus a neutral default class returned when no gate fires, for nine classes in total. The gate values are summarized in the constants table below. The type is therefore a deterministic function of the reading, not a free categorization by the model; the rule table is released with the code.

Algorithm 1 Response generation (single target turn t; strictly read-only on the frozen state)

Require: frozen state $S_p = (\mathcal{M}_p, \mathbf{P}_p, S_{\text{avg}})$; context $\mathcal{W}_{<t}$; trigger utterance u_t
Ensure: single-sentence response y_t

— Situation and emotion measurement —

$(s_t, \hat{e}_t) \leftarrow \text{Measure}(\mathcal{W}_{<t}, u_t)$ ▷ single-model two-step DIAMONDS-SF24 measurement; the annotation-stage protocol in Appendix C.2

$c_t \leftarrow \text{Type}(s_t)$ ▷ read off the percentile-transformed reading by the fixed read-only rule table (D.3)

$S_t \leftarrow \text{Render}(c_t)$; $E_t \leftarrow \text{Render}(\hat{e}_t)$ ▷ rendered as qualitative natural language; $E_t \leftarrow \emptyset$ when emotion is unavailable

— Personality projection (only sufficiently evidenced dimensions participate) —

$M(p,t) \leftarrow \{m \in \mathcal{M}_p \mid \text{turn}(m) < t\}$ ▷ accessible memory, defined in Section 3 of the main text

for each sufficiently evidenced concern sub-dimension $k = k(d)$ **do**

  $a_k \leftarrow \text{Anchor}(\mathbf{P}_p, k)$ ▷ behavioral-level anchor of the frozen profile on dimension k

  $h_k \leftarrow \text{Rule}(a_k, c_t)$ ▷ if–then behavioral direction conditioned on the situation type

end for

$P_t \leftarrow \text{Render}(\{a_k, h_k\})$

— Episodic memory retrieval (dual channel) —

$F_t^{(0)} \leftarrow \text{TopK}_{\text{topic}}(M(p,t), u_t, 10)$ ▷ autobiographical texts, by topical relevance (factual memory channel)

$X_t^{(0)} \leftarrow \text{TopK}_{\text{situation}}(M(p,t), s_t, c_t, 10)$ ▷ coping exemplars, by situational consistency (episodic memory channel); c_t determines the comparability class

$F_t \leftarrow \text{Truncate}_3(\text{ExcludeInWindow}(F_t^{(0)}))$ ▷ remove entries whose source turn falls within the context window

$X_t \leftarrow \text{Truncate}_3(\text{ExcludeInWindow}(X_t^{(0)}))$ ▷ same exclusion rule

$M_t \leftarrow F_t \uplus X_t$ ▷ original text preserved, as traceable material

— Expression planning —

$G \leftarrow \text{Plan}(S_t, E_t, P_t, M_t, u_t)$ ▷ identify the functions the response must fulfill; draft expression guidance

for *attempt* $\leftarrow 1$ to 2 **do**

  $V \leftarrow \text{Screen}(G, \text{ recent behavioral evidence})$ ▷ deterministic screening, at most one guidance-screening retry

  if $V = \emptyset$ **then**

    break

  end if

  $G \leftarrow \text{Revise}(G, V)$ ▷ conflicts resolved by situation $\succ$ memory $\succ$ emotion $\succ$ personality

end for

— Generation and verification —

$prompt \leftarrow \text{Assemble}(G, M_t, \mathcal{W}_{<t}, \text{length reference})$ ▷ excludes all internal numerical values

for $k \leftarrow 1$ to 2 **do**

  $y_t \leftarrow \text{Generate}(prompt, \text{ temperature } 0.4)$ ▷ verification, at most one response rewrite

  $V \leftarrow \text{Verify}(y_t)$ ▷ unretrieved memory / unsupported assertion / verbatim reuse / genre violation

  if $V = \emptyset$ **then**

    break

  end if

  $prompt \leftarrow prompt + \text{feedback}(V)$

end for

return y_t

Conversion of behavioral levels and anchors. The frozen profile $\mathbf{P}_p$ is first converted to reference-population percentiles; behavioral level $= \min(5, \max(1, \lfloor \text{percentile} \times 5 \rfloor + 1))$; each level is bound to one natural-language behavioral description (behavioral anchor), with anchor text looked up in a read-only anchor table; dimensions with insufficient evidence produce no level and no constraint.

Table D1: Inference and projection stages, complete version.

Stage	Module	Input	Operation	Output
Sentence-level expression observation	State construction	Target character's utterances and behavioral-anchor table	Assess evidence strength and direction on each concern dimension	$x_{c,k,j}$
Concern sub-dimension evidence aggregation	State construction	Within-session sentence-level evidence	Compute running mean; record observation count and dispersion	$X_{c,k}$
Big Five trait estimation (frozen output)	State construction	Cross-session $X_{c,k}$	Aggregate across sessions; map back to Big Five dimensions via a fixed bijection	$P_{p,d}$
Behavioral levels and anchors	Response generation	Evidence-gated $\mathbf{P}_p$	Project onto concern dimensions; take 1–5 level behavioral anchors	Behavioral descriptions
Personality description generation	Response generation	Behavioral anchors	Convert anchors into qualitative personality descriptions	Personality descriptions
Situation-conditioned behavioral rules	Response generation	Personality descriptions and current situation type c_t	Activate if–then rules conditioned on the situation type	Situation-conditioned behavioral directions
Expression constraints and personality specification	Response generation	Personality descriptions and situation-conditioned behavioral rules	Convert outputs into qualitative expression constraints	Personality specification and expression requirements

Comparability judgment of coping exemplars. Ranking scores only serve candidate ordering and do not imply comparability. Comparability is judged item by item by a single model on two criteria: (a) same psychological situation class—whether the occasion type, the relationship between the parties, and the tension direction are consistent; (b) comparable context—whether what was talked about then and the current topic/intent can support the same coping approach; output is in three grades: strongly comparable, style-comparable, non-comparable (consistent with the three-way classification of Section 5.2 of the main text). Exemplars entering generation prefer strongly comparable ones; when none is strongly comparable, style-comparable ones are taken, truncated to at most 3 in situational-consistency order; candidates for which no concrete stylistic feature can be written are judged non-comparable and must not be padded into a grade.

Length reference. The length reference in the final prompt is derived only from the person's own utterance history (median and interquartile range of single-sentence length; reported in characters for Chinese corpora and words for English corpora to avoid unit confusion). When the bank contains no entries of the person, the length reference block does not appear at all.

Screening and retry. For reliable but omitted personality dimensions, use at most one profile-dimension recovery retry; if the retry still fails, an empty result is kept and never filled with default content. Dimensions with insufficient evidence remain empty throughout.

Conflict-handling action set. After arbitration by the priority order, only three deterministic actions are performed on guidance items: keep, minimally revise, delete; revision uses the current context and the individual's existing expression patterns and introduces no new content.

Prompt discipline. Psychometric values, percentiles, and similarity scores are used only for internal computation and do not enter the final prompt; retrieved autobiographical memories and coping exemplars keep their original text, and the guidance layer introduces no new facts or concrete wording—so everything factual in the prompt is traceable to the current context or historical text.

Verification and rewrite cap. On a response violation, verification feedback is appended to the prompt, with at most one response rewrite; if the rewrite still violates, the final draft is kept and the verification result is logged (the four violation classes are listed in the verification step of Algorithm 1).

Empty-bank degradation. When no memory is available at the target turn (e.g., the person's first target turn), the past-events and exemplar material blocks do not appear, and the expression guidance states "no citable past events".

Information boundary. The frozen state is constructed independently within each evaluation pack, using only the visible stream in that pack's evidence closure; any cross-pack profile reuse is only an implementation means of batch processing, not an input of the methodological pipeline.

D.4 Memory Entry Fields

Each memory is defined as $m_i = (x_i, q_i, s_i, r_i)$: x_i is the original text, q_i contains source and turn information, s_i is the situational vector at encoding time (i.e., `enc_S`, consistent with the memory representation of Section 4 of the main text), and r_i records attributes such as salience, entry type, and accessibility. s_i supports retrieving concrete experiences and coping exemplars under comparable encoding situations.

D.5 Mapping Tables, Constants, and Read-Only Assets

Fixed mapping $k(d)$. McAdams (1995); the complete mapping table cited in Section 5.1 of the main text:

Big Five dimension d	**Canonical concern domain $k(d)$**
O Openness	Novelty concerns
C Conscientiousness	Achievement concerns
E Extraversion	Social-initiative concerns
A Agreeableness	Harmony concerns
N Neuroticism	Safety concerns

Main constants. Implementation-level parameters not listed are released with the code:

Constant	**Value**	**Purpose**
Situational scale range	1–7	Eight-dimensional situational readings (same scale as norms and annotations)
Personality observation scale	1–7	Same scale as reference-population self-reports
Number of behavioral levels	5	Level anchors of profile projection (conversion in D.3)
VAD neutral point	$(4, 4, 4)$	Reference center of emotion readings
Measurement-call temperature	0.0	Situation and emotion measurement, all arbitration calls
Generation temperature	0.4	Response generation
Per-dimension abstention threshold	observation count < 3	The dimension produces no constraint (a legal output)
Direction-giving threshold	observation count ≥ 8	Below it, values are still computed and recorded but do not enter the prompt or produce guidance lines
Situation-type gates	high/low percentile 0.80 / 0.20	Gate constants for the rule table defined in D.3
Retrieval narrow tier	at most 3 per channel	First broad retrieval (10) $\rightarrow$ exclude entries whose source turn falls within the context window $\rightarrow$ truncate to 3
Exemplar context look-back	3 turns	Coping exemplars restore their original context
Accessibility half-life	200 turns	Memory accessibility decays exponentially with turn distance

Factor composition of retrieval scoring. The main text states only the ordering basis of the two channels; the score composition is given here, with the complete formula released with the code: the factual memory channel's score is a weighted composition of lexical cues, salience, and accessibility (exponentially decaying with turn distance, half-life 200 turns), modulated by the congruence between the entry's encoding situation and the current situation; modulation only affects ordering and truncation, not the recall set or citation threshold. The episodic memory channel orders solely by encoding-situation congruence and does not use lexical cues.

Reference-population declarations. Personality percentiles and situation-type gates are both norm-relative, but against different reference populations. Personality percentiles are taken within the self-reported Big Five of 233 speakers of one corpus (scale 1–7, the same scale as the behavioral observations). Situational percentiles and the gates of the situation-type rule table are taken within the consensus situation profile of the situational scale's original source (C.2), which is a non-dialogue reference sample. A percentile is meaningful only within the population it is taken from, so neither quantity is comparable across corpora; this is a known limitation of the method.

The situation-type rule table, the behavioral-anchor table, and the complete retrieval-scoring formula are all read-only assets released with the code.

E Experiment Supplementary Details

Correspondence: Sections 6.1–6.2 (Evaluation Tracks and Baselines; Evaluation Metrics) of the main text.

E.1 Baseline Configurations and Inclusion Rationale

Section 6.1 of the main text has listed the three-track baselines and citations; this entry supplements the unified run configuration and the inclusion rationale of each baseline.

Unified configuration. Training-based baselines use official public weights; inference-based and closed-loop baselines run with official code, uniformly connected through an OpenAI-compatible interface. Review models bundled with baselines are uniformly discarded and replaced with review models from a family different from the generation model (see E.4); numbers self-reported by baseline papers are cited only "under their own protocols" and are not compared horizontally with this benchmark's results.

E.2 Metric Formulas, Symbol Definitions, and Construction Rules

Section 6.2 of the main text has given the ρ definitions of the three covariation axes; this entry gives the definitions, formulas, and construction rules of all remaining metrics, including the write-back ablation difference R_2 and the operational details of borrowed metrics moved out of the main text. All borrowed metrics carry paper citations.

Table E2 summarizes the metric formulas (table values are means or counts over the corresponding evaluation samples; each metric simultaneously reports effective sample size n and missing rate).

Response quality (Tier 1) definitions and construction rules. The blind four-dimensional (ABCD) review-model score (Zheng et al., 2023) rates anonymous replies with integer scores 1–5 on four aspects: A appropriateness (whether it properly continues as the next utterance and fits the speaker's stance), B fluency (sounds like a real person), C faithfulness (does not fabricate facts absent from the context), D closeness to the reference reply (semantic-direction closeness to the gold, not wording overlap). The ESC-Eval seven dimensions (Zhao et al., 2024) are adopted as-is: Fluency, Expression (clear and appropriate expression), Empathy, Information (informativeness and pertinence), Humanoid (sounds like a real person rather than a machine), Skill (communication skill, e.g., proper concession/guidance), Overall (overall preference), each scored 1–5. Gold replies are anonymized into the blind candidate pool and scored in the same session; shuffling uses a fixed random seed. When the review model and the evaluated system belong to the same model family, that review result is reported only as a secondary result (consistent with E.4 item 1).

Situational modulation consistency (Tier 2) sampling and symbol definitions. Evaluation samples are located at trigger utterances, all coming with gold replies and situational annotations.

- *E axis*: $\mathrm{ER}_i \in \{0, 1, 2\}$ is the emotional-response intensity read from the reply by the review model (the EPITOME three-grade scale, Sharma et al., 2020: 0 none, 1 weak, 2 strong; intensity is judged, not valence quality); $z(\cdot)$ is z-standardization within the evaluation pool; $S_{Neg,i}$, $S_{Pos,i}$ are the corresponding dimension scores of the frozen situational annotations.

Table E1: Baseline inclusion rationale.

Baseline	Inclusion rationale
CharacterGLM-6B	The only Chinese-native training-based role model with directly obtainable weights, same language as this study's Chinese data
PsyLLM-8B	Its "mental-state inference to response" mechanism is isomorphic to this paper's explicit mental-state modulation; the most comparable among emotion-capable training-based methods
MemoryLLM	Paradigm contrast between memory in model parameters and memory in inference-time architecture
SOTOPIA-π	Pioneering work on interactive learning of social intelligence; the common starting comparison for scenario-dimension RL methods
Sotopia-RL	The latest RL result of the official SOTOPIA line; directly comparable with SOTOPIA-π under the same environment and judge
SocialRL	Together with the previous two, forms a complete training lineage of "outcome imitation, reward design, process reward"
PERSONA	Training-free contrast between activation-level control and this paper's architecture-level modulation
Emotional RAG	The only currently runnable open-source object for the "affective retrieval vs. semantic retrieval" contrast
PHASE-Tree	The public inference-time framework closest to this paper's access-layer design on long-horizon personality consistency
RGMem	Public state of the art for personalized memory systems; its profile-evolution mechanism is most directly related to this paper's update claims
Mem0	Industry-general memory pipeline and common baseline of domain papers, with explicit add/delete/modify operation records
MemInsight	Pure-prompt extraction of structured attribute annotations; its annotation dimensions are the most isomorphic to this paper's four-dimensional annotations
AdaMem	Adaptive memory-writing strategy; the write-control contrast with completed reproduction verification
HyperTrace	Memory construction and response generation closed-loop within one pipeline; method decoupled from corpus domain
PGMem	Profile nodes and event nodes explicitly connected by evidence edges, forming the nearest direct contrast to the "updates are scoreable" claim; its paper uses 1–2B small-model backbones, and its self-reported numbers are not compared horizontally with this benchmark
Cond-none	Reference condition (not an external method): provides only the dialogue context; measures the lower bound of the value of "providing state information" itself
Cond-gold	Reference condition: context + gold state text and targeted memory entries; isolates the contribution of "correct state content"
Cond-gold+mem	Reference condition: further adds all gold memory entries; isolates the marginal contribution of "complete memory"; the three arms share PSyMod's generation model and prompt template, differing only in state source

- *M axis*: the expected citation set is frozen once by a functional-necessity judgment—the review model judges per target turn "would this memory entry be of actual help in producing a reply like the gold" (three grades: necessary/helpful/no, with two anti-false-positive rules: a new event that coincides with the entry only in numbers or sentence pattern is judged no; an entry whose content already appears in the context window is judged no); once frozen, all systems share it and reruns do not re-judge. nu_i is the size of the expected citation set; $\mathrm{m_cited}_i$ is jointly given by the review model's entailment judgment and a lexical bigram channel, abstaining and entering the denominator when the two channels disagree; dmean_i is the mean eight-dimensional situational distance between the expected memory source turns and the target turn.

Table E2: Metric formula summary.

Metric	Formula	Notes
Blind four-dimension mean	$\text{avg} = \frac{A+B+C+D}{4}$	A/B/C/D are each integer scores 1–5, one review call per candidate
ESC-Eval six-dimension mean	$\frac{1}{6}\sum_{k=1}^{6} s_k$	$s_k \in [1,5]$ for the six non-Overall dimensions; the Overall dimension is reported separately as the primary ranking reading
Storage-correctness three-way classification	$P = \frac{\lvert\text{supporting}\rvert}{\lvert\text{sys. entries}\rvert}$; R = evidence-turn coverage (defined in the main text); $F_1 = \frac{2PR}{P+R}$	The three classes are mutually exclusive: supporting/redundant/hallucinated (definitions in the main text); coverage is semantic coverage
Redundancy / hallucination rate	$\frac{\lvert\text{redundant}\rvert}{\lvert\text{system entries}\rvert}$ / $\frac{\lvert\text{hallucinated}\rvert}{\lvert\text{system entries}\rvert}$	Same denominator as P
Character-profiling six ratios	each $= \frac{\#\{\text{hit entries}\}}{n}$, n = total bank entries	Verbatim preservation uses the character-bigram containment operator: $\text{contain}(a,B) = \max_{b\in B}\frac{\lvert\text{bi}(a)\cap\text{bi}(b)\rvert}{\lvert\text{bi}(a)\rvert}$, where $\text{bi}(\cdot)$ is the character-bigram set of an utterance; entry e counts as traceable to original utterances only when its containment against each turn of the dialogue history is at least 0.8 (a strict threshold to avoid inflation by paraphrasing), i.e., $\frac{1}{n}\#\{e : \max_{r\in\text{history turns}}\text{contain}(e,r) \geq 0.8\}$. The tendency-dimension item reports the number of category dimensions with ≥ 3 satisfying entries, side by side with the reference profile. Two additional checks have denominators that only our method can provide, so they serve as internal self-checks and are not included in the comparison table
Retrieval	$\text{Hit@3} = \mathbb{1}[\text{rank} \leq 3]$, $\text{MRR} = \frac{1}{\text{rank}}$ (0 when not hit)	Standard definitions in Introduction to Information Retrieval Schütze et al., 2008; rank is the best position of the reference source turn in the system's retrieval results
Terminal performance (R1)	Recompute Tier 1 and Tier 2 metrics on end-of-stream replies	No composite closed-loop score
Iterative degradation	$\text{score}^{\text{cold-start}} - \text{score}^{\text{terminal}}$	Computed pairwise with the same operationalization as R1; a positive value suggests getting worse with use

- *P axis*: the unit of computation is target turn × Big Five dimension. $x_{i,d} = \text{sign}(\text{gold}_{i,d}) \cdot z(S_{\text{map}(d),i})$, with the reference direction taken from the frozen personality annotations (high/low), and the mapping map: O↔Intellect, C↔Duty, E↔Sociality, A↔Positivity, N↔Negativity; $y_{i,d} = \mathbb{1}[\text{reverse-read}_{i,d} = \text{gold}_{i,d}]$, where the reverse-read is the per-dimension direction (high/low/insufficient-evidence) given by the review model solely from overt expression; insufficient-evidence cases abstain and leave the pool.

Construction-side audit construction rules. Entailment judgment is built in three levels: lexical short-circuit (an entry, after normalization, is identical to some turn in the stream or a contiguous substring of length ≥ 4 of it, judged entailable; only the sufficient direction is taken—fabrication is never judged as entailment) → review-conclusion cache → review-model NLI-style judgment. Deterministic definitions of the six character-profiling ratios: *character-level information* = propor-

tion of entries hitting stable-attribute statements; *behavioral traces* = proportion of entries hitting behavior-structure/emotion-tone/intensity lexicons; *verbatim preservation* = proportion of entries traceable back (threshold 0.8) to some history turn via the containment operator (see Table E2); *emotion readability* = proportion of entries traceable to a source turn that carries emotion annotation; *intensity overreach* = proportion of entries hitting forbidden intensity expressions; *tendency dimensions* = number of tendency categories with ≥ 3 entries (reported side by side with the reference profile). Full lexicons and regular expressions are released with the code.

Retrieval construction rules. Relevance is determined by the reference source turns fixed by data annotation: Hit@3 counts when such a turn appears in the top three positions of the system's retrieval results, and MRR is the reciprocal of its rank (0 when not hit). Consistent with the main text, each system retrieves from its own constructed memory bank; in addition, entries whose source turn falls within the context window are uniformly excluded—in-window entries carry no new information for the system; the exclusion criterion is the same as on the method side at generation time (Appendix D.5, retrieval narrow tier), applied uniformly to all systems as a fairness treatment rather than tuning. One construction-side baseline, AdaMem, ran as a partial batch (96 of 717 packages) and its retrieval cells are computed on its own judged anchor set rather than the shared denominator; its storage and retrieval cells are therefore reported as directional values and excluded from the best/second-best marking of the main tables.

Closed-loop end-to-end supplement. The closed-loop track reports terminal performance R_1 (Tier 1 and Tier 2 metrics recomputed on end-of-stream replies) and the write-back ablation difference R_2, the late-minus-early change in endpoint response quality for the matched live-write-back and no-write-back conditions. Pairwise differences of Tier 1/2 metrics between cold start and stream end (iterative degradation) are reported with the same operationalization, testing that closed-loop operation "at least does not get worse."

E.3 Significance and Small-Sample Handling

The directions of the three covariation axes are pre-registered and frozen before the evaluation starts. Significance uses bootstrap: resampling by evaluation unit $B = 2000$ times, reporting the *reverse-resample proportion*—the proportion of resamples whose ρ falls on the side opposite the pre-registered direction, interpreted as a one-sided p-value (< 0.05 significant; > 0.75 significantly reversed). When the judgeable sample of a single axis has $n < 6$, the correlation coefficient is not computed and this is explicitly noted. Abstentions and NA always enter the denominator. The three axes are reported separately, without averaging.

E.4 Fairness Handling

(1) The review model (claude-opus-4.6) differs from the generation model; when the review model and an evaluated system belong to the same model family, that result is reported only as a secondary result. (2) In blind evaluation, gold replies are anonymized into the pool with fixed-seed shuffling. (3) Review models bundled with baselines are uniformly discarded (see E.1). (4) Retrieval comparison channels are isomorphic: all systems build their own banks, retrieve from them, and apply the uniform in-window exclusion. (5) Construction-side precision and recall are computed on the same reference subset, avoiding trading indiscriminate bank enlargement for precision; the entailment short-circuit takes only the sufficient direction, so faithfulness is not overstated. (6) Gold derivatives such as the expected citation set are frozen once and persisted; all systems share them, and reruns do not re-judge.

E.5 Full Scoring Prompts (verbatim)

The following prompts are reproduced verbatim in their original language; wording such as "gold" inside them is verbatim prompt text and is not rewritten to follow the main text's terminology.

Blind four-dimension (ABCD).

```
You are an evaluator of conversational response quality. Given a dialogue
↪ context and several candidate responses (including one gold
↪ response), assign each candidate an integer score from 1 to 5 on the
↪ following four dimensions: A, contextual appropriateness (whether it
↪ appropriately follows as the next utterance and is consistent with
↪ the speaker's position); B, fluency and naturalness (whether it
↪ sounds like something a real person would say); C, faithfulness
↪ (whether it avoids fabricating facts not present in the context; 5 =
↪ completely faithful); and D, semantic proximity to the reference (how
↪ closely its semantic direction matches the gold response, without
↪ considering wording overlap). Output only JSON:
↪ {"1":{"A":n,"B":n,"C":n,"D":n},...}, where the keys are candidate
↪ numbers.
```

ESC-Eval seven dimensions.

```
You are an evaluator of emotional-support dialogue quality. Assign each
↪ candidate response an integer score from 1 to 5 on the following
↪ seven dimensions: Fluency, Expression (clarity and appropriateness of
↪ expression), Empathy (understanding of the other person's situation
↪ and feelings), Information (amount and relevance of information),
↪ Humanoid (whether it sounds like a real person rather than a
↪ machine), Skill (communication skills, such as making appropriate
↪ concessions or providing guidance), and Overall (overall preference).
↪ Output only JSON: {"1":{"Fluency":n,...},...}
```

E axis: emotional-response intensity.

```
You are an evaluator of emotional-response intensity. Given a multi-party
↪ dialogue context and a target utterance (a consecutive line spoken by
↪ the same speaker), assess the intensity of the speaker's emotional
↪ response expressed in the target utterance using the following
↪ three-level scale:
0 = no emotional response (neutral, transactional, or calm statement)
1 = weak emotional response (emotional coloring is present but restrained
↪ or mild)
2 = strong emotional response (intense, overt, and evocative emotion)
Judge intensity, not valence: both intense negative and intense positive
↪ emotions should receive a score of 2.
Also determine the valence direction: positive/negative/neutral.
Output JSON: {"er":0 or 1 or 2,"valence":"positive|negative|neutral"}. Do
↪ not provide an explanation.
```

M axis: expected citation set (functional-necessity judgment).

```
You are making a functional-necessity judgment for memories. You are
↪ given: the current dialogue context window (where the final utterance
↪ is the trigger utterance), a gold reference response (the speaker's
↪ actual response at the time), and several visible utterance entries
↪ from earlier conversations in that speaker's memory bank. For each
↪ entry, answer only one question: To produce an appropriate response
↪ like the gold response, did this memory provide actual assistance?
↪ Use the following three-category closed set:
- necessary: The key content (fact, position, or wording) in the gold
↪ response depends on this entry. Without it, the response would be
↪ invalid, dishonest, or irrelevant.
- helpful: It is not necessary, but using it would make the response more
↪ characteristic of the person (for example, by reusing their
↪ consistent position, wording, or background knowledge).
```

```
- no: It is irrelevant. It is merely related in topic, wording, or
↪  sentence structure, but the response does not require it. The
↪  following two false-positive prevention rules must be applied: (a) if
↪  the gold response states a fact about a new event or new entity, and
↪  the entry is related only through a coincidental number or sentence
↪  structure (e.g., the gold asks about the distance to Location A,
↪  while the entry concerns the distance to Location B), assign no; (b)
↪  if the entry's content already appears in the current context window,
↪  assign no, because repetition within the window does not count as
↪  memory use.
Span rule: For necessary/helpful, provide the exact consecutive span in
↪  the gold response that depends on the entry; for no, provide an empty
↪  string. Do not infer content that is not stated in the gold response.
↪  Output only JSON: {"judgements":[{"id":integer,"need":"necessary|hel
↪  pful|no","span":string,"why":string}]}, and you must cover all entry
↪  IDs in the list.
```

M axis: actual citation audit.

```
You are an auditor of memory citations. Given the [expected citation set]
↪  (the memory facts that the gold response determines should be cited)
↪  and a [candidate response], determine for each item whether the
↪  response actually cites that fact (the same event or fact is
↪  sufficient; paraphrasing is allowed and exact wording is not
↪  required).
Output only JSON: {"cited":[IDs of cited expected items],"extra":[the
↪  number of memory-related facts appearing in the response that are not
↪  included in the expected set, integer]}, where IDs start from 0.
```

P axis: personality direction reverse-read.

```
You are an evaluator who reverse-reads the direction of personality
↪  expression. Given a multi-party dialogue context and one response
↪  from the speaker, infer from the context the direction of the
↪  speaker's currently expressed Big Five tendencies along each
↪  dimension, based only on overt expression (do not infer stable
↪  traits):
O, Openness (curiosity/novelty-seeking vs. practicality/conventionality)
C, Conscientiousness (planning/encouragement vs. spontaneity/leniency)
E, Extraversion (initiative/warmth vs. reserve/conciseness)
A, Agreeableness (consideration/conciliatoriness vs. firmness/opposition)
N, Neuroticism (anxiety/expressed negative emotion vs. calmness)
For each dimension, output h (clearly high), l (clearly low), or m
↪  (moderate or insufficient evidence; abstain).
Output JSON:
↪  {"O":"h|l|m","C":"h|l|m","E":"h|l|m","A":"h|l|m","N":"h|l|m"}. Do not
↪  provide an explanation.
```

Construction side: semantic coverage and support (P/R).

```
You are an auditor of memory-bank alignment. Given [gold entries]
↪  (memories manually confirmed as worth retaining) and [system entries]
↪  (entries independently constructed by the system under evaluation),
↪  determine for each gold entry whether it is semantically covered by
↪  at least one system entry (the same event, fact, or attitude;
↪  paraphrasing, merging, and splitting are allowed; exact wording is
↪  not required). Then determine for each system entry whether it is
↪  supported by at least one gold entry (the event described by the
↪  system entry has a corresponding event in the gold entries; content
↪  absent from the gold entries is fabrication or noise and is
↪  unsupported).
Output only JSON: {"gold_covered":[IDs of covered gold
↪  entries],"sys_supported":[IDs of supported system entries]}, where
↪  IDs start from 0.
```

Construction side: entailment judgment (redundant vs. hallucinated).

```
You are a memory auditor. Given a [dialogue-flow excerpt] and one [system
↪  memory entry], determine whether the content of the entry is entailed
↪  by the dialogue flow (including reasonable integrative inferences: if
↪  the information in the flow can be directly combined to derive the
↪  content, it is entailed; if it is entirely unsupported and invented,
↪  it is not entailed). Output only: ENTAILED or FABRICATED, followed by
↪  a one-sentence reason.
```

E.6 Reporting Standards

The main text has set "no composite overall score; report the effective sample size and missing rate of each metric"; three supplements: results from the same batch of data use the review-model settings of the same time point, and cross-batch results are not combined into one table; deterministic rules (lexicons, regular expressions, thresholds) and scoring prompts follow this appendix and the assets released with the code; pilot results and formal evaluation results are strictly separated, and the paper reports only formal evaluation results.

F Full Statements of the Passages Compressed in the Main Text

Correspondence: Sections 1–7 of the main text, block by block (F.n matches main-text Section n).

To keep the main text within its page budget, several passages are stated there in compressed form. This appendix reproduces the material that was condensed or removed, so that no statement is lost. Each block names the passage it comes from, and the main text points to it where the compressed material is substantive. Blocks marked *deleted* are passages no longer present in the main text; blocks marked *fuller statement* give the longer wording from which the main text keeps a condensed version.

F.1 Introduction. *Deleted: third paragraph, four-dimensional machinery and construction detail.* The four dimensions remain separable and interpretable: dialogue and user state are converted into a four-dimensional representation, an expression direction is planned, constraints are checked, and the output is validated. On the construction side, personality, memory, and situation are aggregated from the experience stream: insufficiently supported personality dimensions abstain, situation vectors are stored with memories, and situation readings form a user-specific reference distribution. *Deleted: fourth paragraph, metric system preview.* Generation metrics assess response quality and situation modulation, construction metrics assess state construction and retrieval, and cross-checking and closed-loop protocols connect the two sides.

F.2 Related Work. *Deleted: 2.2 topic sentence of the situations line.* Psychological situations are a measurable dimension studied by another line of research. *Deleted: 2.4 evaluation-side paradigms.* The reference paradigms of these benchmarks cover model judges, simulated users, and psychometric scales.

F.3 Task Setting. *Deleted: read-only statement within a turn (duplicated in the closed-loop paragraph).* Within a single turn, the response generation process only reads the constructed state; it does not re-execute state construction, nor does it directly update that state (the closed-loop paragraph below states the same boundary).

F.4 Dataset. *Fuller statement: 4.1 four-dimensional representation.* The benchmark encodes the four dimensions of personality, emotion, situation, and memory in distinct data representations. Personality is represented as a Big Five profile with three levels—high, medium, and low—placing annotations from different sources on the same scale. Emotion is represented as utterance-level VAD vectors (Warriner et al., 2013), putting emotion information from different sources on the same continuous scale; this dimension is absent in RPC and is therefore recorded as missing. Situation covers the eight DIAMONDS dimensions (Rauthmann et al., 2014). Situational annotation adopts a two-step pipeline: three models (qwen3.8-max, claude-opus-4.6, and gpt-5.5) first independently generate situational descriptions, and then each scores the scale items dimension by dimension based

solely on its own generated description. A dimension's score is retained only when at least two models agree, and retained annotations undergo manual review. The sources of the scale version, items, and reference standards, as well as the scoring model configurations (including temperature set to 0) and the per-dimension reliability, are given in Appendices C.2–C.3. Memory is maintained as speaker-oriented entry banks. The same object corresponds to two memory banks—one storing what the person said and the other storing what the person heard—and both support cross-scene retrieval. Each memory is bound at formation time to the eight-dimensional encoding vector `enc_S` of the situation then, so it can be indexed by situational similarity rather than mere lexical overlap. The data representation layer does not preset retrieval weights; retrieval priorities are determined by the rules of the method layer. The conversion of the four-dimensional representations into the generation process and the boundary of memory encoding are given in Appendix C.1. *Fuller statement: 4.2 evaluation protocols.* The final scale of evaluation units is 9,078, of which CPED contributes 2,587, Friends 5,284, and RPC 1,207. This study derives three evaluation protocols from the same benchmark rather than partitioning the benchmark into mutually exclusive datasets. $\mathcal{D}_{\mathrm{gen}}$ is the generation protocol, providing aligned evaluation units, four-dimensional annotations, personal memory banks, and short context windows for response-generation evaluation. $\mathcal{D}_{\mathrm{con}}$ is the state-construction protocol, providing longer dialogue histories while withholding annotations and memory states so that the system must infer and progressively accumulate them itself. $\mathcal{D}_{\mathrm{full}}$ is the closed-loop protocol. This protocol organizes character packs by character and arranges all visible conversation turns of each character in ascending temporal order into a character full stream. The system runs in streaming fashion from a cold start, thereby approximating everyday conversation in which information is naturally incomplete and must be accumulated over the long term. The benchmark is for evaluation only; hence the three protocols respectively examine response generation, state construction, and closed-loop full-stream continuation, and no training or test splits are defined. *Fuller statement: 4.2 statistics, closing note.* Counts are tallied by protocol object—generation evaluation packs, construction evaluation packs, and character packs respectively—and should not be compared directly across protocols; the 13,473 scenarios of $\mathcal{D}_{\mathrm{full}}$ denote the number of scenarios covered by the character full streams. The complete table is given in Appendix C.4 (Table C1), together with the per-source coverage matrix and criteria. The post-calibration ICC average of situational annotations is 0.75. According to the ICC interpretation criteria of Koo & Li (2016), this indicates good annotation reliability. Construction quality has been verified by full-corpus recomputation. The per-dimension ICC values are given in Appendix C.2. *Deleted: 4.1 complementary emphases of the three corpora.* The three corpora have complementary emphases: Friends is the only source providing utterance-level personality annotations, CPED provides dense utterance-level emotion annotations, and RPC provides unscripted real interactions. *Deleted: 4.1 retrieval weights at the representation layer.* The data representation layer does not preset retrieval weights; retrieval priorities are determined by the rules of the method layer. *Deleted: 4.2 caution on comparing protocol statistics.* Counts are tallied respectively by generation evaluation pack, construction evaluation pack, and character pack. *Deleted: 4.2 caution on comparing protocol statistics.* Counts are tallied by protocol object and should not be compared across protocols; the 13,473 scenarios of $\mathcal{D}_{\mathrm{full}}$ denote the number of scenarios covered by the character full streams. *Deleted: 4.2 purpose of the three protocols.* The three protocols respectively examine response generation, state construction, and closed-loop full-stream continuation. *Deleted: 4.2 derivation of the three protocols.* This study derives three evaluation protocols from the same benchmark rather than partitioning the benchmark into mutually exclusive datasets.

F.5 Method. *Deleted: 5.2 rationale for the conflict-resolution order.* This order reflects the functional differences among the representations: the situation constitutes the external constraint of the current turn, memory provides traceable factual grounding, and emotion and personality mainly determine the manner of expression rather than directly determining content correctness. *Deleted: 5.1 theoretical framing of the induction.* The hierarchical view of personality (McAdams, 1995): traits manifest as concrete behaviors under specific situational conditions through situationally grounded concerns and goals. *Deleted: 5.2 statement closing the personality projection.* Thus the same personality profile projects into different behavioral directions under different situation types. *Deleted: 5.2 factual memory channel.* The factual memory channel provides traceable factual material for the response.

F.6 Experiments. *Deleted: Table 1 caption, dagger definition.* Reference conditions share the same generation model and prompt template as PSyMod, differing only in the state source. *Deleted:*

Table 1 note, per-arm backbone configurations. Per-arm backbone configurations of Table 1: MemoryLLM 8B; CharacterGLM-6B; SOTOPIA-π: Mistral-7B + behavior cloning; Sotopia-RL: Qwen2.5-7B, SFT/GRPO variants; PsyLLM-8B; SocialRL: Qwen2.5-7B + LoRA; PERSONA: Qwen2.5-7B + steering vectors with 1 Big Five prediction call; Emotional RAG: 3 calls (gpt-5.4-mini); PHASE-Tree: Qwen2.5-7B + phase-tree LoRA. *Deleted: Table 4 note, readouts.* The 5-level granularity carries state fidelity.

Deleted: closed-loop ablation subsection (main-text Section 6.4, removed). This experiment tests whether write-back contributes beyond the natural change that occurs as a stream progresses. No additional table is needed because the matched live-write-back and write-back-off rows are already reported in the closed-loop table of the main text; the other rows in that table are external closed-loop baselines. The write-back ablation is reported as R_2 in the closed-loop track: on the 24-anchor batch, turning write-back off lowers the late-probe ABCD from 2.861 to 2.722 and shrinks the late$-$early gain from $+0.247$ to $+0.154$.

F.7 Conclusion. *Deleted: second paragraph, metric reporting details.* Generation-side and construction-side metrics are reported separately together with effective sample sizes. *Deleted: third paragraph, explanatory clauses of the result recap.* The baselines show almost no such sensitivity, indicating that situations genuinely alter effective states rather than merely decorate prompts. *Deleted: first paragraph, mechanism recap.* At each turn the system compares the current situational appraisal against the user's own historical situation distribution, identifies the situation type, and accordingly controls memory retrieval, personality expression, and emotional intensity. *Deleted: result-recap clauses and the supporting mentions that carried them.* PSyMod improves ESC-based response quality while showing stronger situation-modulation consistency, particularly on the personality and memory axes, although its blind four-dimensional review score is slightly lower. Its state construction stores no hallucinated entries while retaining competitive retrieval, and live write-back yields a larger late-minus-early gain than the same pipeline without it in the closed-loop probe. Intermediate-layer and access-layer ablations affect response quality and modulation readouts in different directions, supporting a trade-off interpretation rather than uniform dominance.