Remember Less, Revise Better in Bounded Persona Belief Tracking for Long-Term Dialogue

Anonymous ACL submission

Abstract

Long-term dialogue agents need a clear, current picture of each user. Existing memory systems store the whole transcript or rewrite running summaries; how the picture should change when a new utterance arrives, whether to add, strengthen, weaken, or narrow a belief, is left implicit, and final-state metrics flatten the four cases into one. We introduce **Bounded Revision** (BR): a small structured profile edited after each event with one of four operations adapted from classical belief revision, REGISTER, REINFORCE, ATTENUATE, and QUALIFY. We evaluate it on PersonaBoard, a benchmark of 2,100 events that scores how a profile changes rather than only its final state. Strong memory systems including MemOS and A-MEM reach high final-board overlap but score near 0.55 on balanced operation accuracy, since rebuilding a state does not show whether a belief was strengthened, weakened, or restated. BR keeps each update explicit: it improves balanced operation accuracy by 14.7 points over full history and non-target preservation from 0.73 to 0.97, at a cost of 7 to 9 points of final-board overlap. The compact profile also transfers to PERMA preference questions under an identical memory interface, where BR leads.

1 Introduction

Long-term dialogue agents are moving from one-shot question answering toward continuing collaboration; production AI assistants are already shipping persistent user memory features that span weeks or months of interaction. Yet large language models (LLMs) in this regime often lock onto early assumptions and fail to recover when later turns demand a different course (Laban et al., 2025). To collaborate over time, an agent needs a current view of the user. Humans handle this with a small, bounded impression of each person they know: a friend remembers a handful of traits about you and quietly revises them as you change, rather than mentally replaying every conversation (Patterson et al., 2007; Nagy et al., 2025). The same intuition fits dialogue agents: a personal assistant needs to remember recurring schedules and dietary habits, a tutor needs to notice when a learner becomes more confident, and a writing assistant needs to track that a user prefers direct edits while also noticing that this preference may be limited to deadline-driven work. None of these reduces to verbatim recall: the agent needs a compact working profile that keeps reusable beliefs, updates them as new evidence arrives, and stays clear enough that a user can correct it (Hatalis et al., 2023; Huang et al., 2026).

Existing approaches frame long-term memory as a problem of storage. Full-history prompting keeps the entire transcript and lets the model resolve conflicts at inference time. Retrieval brings back relevant snippets on demand (Lewis et al., 2020; Huang et al., 2023; Wang et al., 2024; Gutiérrez et al., 2024); rolling summaries and reflective managers compress older turns into a running state (Wang et al., 2025a; Tan et al., 2025b); recent agent memory systems such as Mem0, MemOS, A-MEM, and PRIME add extraction, consolidation, and memory-augmented generation at scale (Chhikara et al., 2025; Li et al., 2025c; Xu et al., 2026; Zhang et al., 2025; Kang et al., 2025; Salama et al., 2025; Zhong et al., 2024); and benchmarks like LongMemEval and MemBench test whether agents can use long-term memory in interactive settings (Wu et al.; Tan et al., 2025a). What these methods leave implicit is the update step itself. When a profile changes, did the system add a new trait, strengthen an old one, weaken a contradicted belief, or just narrow the condition under which it applies? Final-state metrics flatten all four into the same number.

Persona memory is reused across future calls, so getting that update right matters. A retrieved snippet can recall what the user once said, but it cannot decide whether the statement should become a stable preference. A summary can rebuild a plausible

final profile, but it can also overwrite an old belief or blur a one-time exception into a permanent trait. We therefore treat persona memory as an online editing problem. At each dialogue event, the agent maintains a small structured profile whose rows are reusable beliefs, each described by a key, a state, a confidence, and a scope. The profile is bounded, so the persona layer does not grow with every utterance, and each update is a single typed operation, so the system says explicitly whether it is registering, reinforcing, attenuating, or qualifying a belief. We instantiate this as Bounded Revision (BR): given the current profile and one new event, BR picks one operation, names the target belief, and emits the next profile.

Our contributions are:

- We recast long-term persona memory as bounded belief tracking, the online editing of a small structured profile through typed operations.
- We release PersonaBoard, to our knowledge the first benchmark that supervises the persona trajectory, with 2,100 hand-verified updates.
- We show that strong memory systems reach high final-board overlap yet make wrong updates, while Bounded Revision raises balanced operation accuracy by 14.7 points over full history at the cost of 7 to 9 points of final-board overlap.

2 Related Work

Long-term dialogue memory. Persona dialogue work has progressed from fixed profile sentences (Zhang et al., 2018) to multi-session memory and reflective management (Xu et al., 2022a,b; Zhong et al., 2024; Sun and Fujita, 2025; Li et al., 2025a; Huang et al., 2026; Tan et al., 2025b), and to in-dialogue persona learning and post-generation alignment (Huang et al., 2023; Cheng et al., 2024; Liu et al., 2025; Chen et al., 2025; Li et al., 2025b); interactive benchmarks evaluate long-term memory in agents (Wu et al.; Tan et al., 2025a; Castillo-Bolado et al., 2024). These works mostly evaluate recall, reasoning, or final response quality; we focus on the online update step.

Agent memory systems. Recent systems treat long-term memory as practical infrastructure for LLM agents, providing extraction, storage, retrieval, consolidation, reflection, and memory-augmented generation (Packer et al., 2024; Chhikara et al., 2025; Li et al., 2025c; Kang et al., 2025; Salama et al., 2025; Dai et al., 2026; Liang et al., 2025). Retrieval-augmented generation, HipoRAG, M+, Agent Workflow Memory, and editable memory graphs cover retrieval or reusable-content induction (Lewis et al., 2020; Gutiérrez et al., 2024; Wang et al., 2025b,c, 2024); PRIME is closest in spirit and separates past interaction history from evolving long-term user beliefs (Zhang et al., 2025). These systems answer where memories live and how they are retrieved; our work asks what enters the persona layer and how a belief changes when new evidence conflicts with or qualifies an old one.

Belief revision. Our update operations are a typed instance of classical belief revision (Alchourrón et al., 1985; Hansson, 1999); the mapping is given in §4. QUALIFY is closest in spirit to iterated and scoped revision (Darwiche and Pearl, 1997), where new evidence narrows rather than overturns an existing belief. Recent work has examined whether LLMs themselves obey revision postulates (Wilie et al., 2024); here we use the postulates instead as a structural constraint on the update space.

Dialogue state and user modeling. Dialogue state tracking maintains slot values during task-oriented interaction. Bounded persona belief tracking adapts this view to long-term user modeling, where the state persists across sessions, has limited size, and is revised by typed operations.

3 Preliminaries

At event t, the persona state is a board $\mathcal{B}_t = \{b_1, \ldots, b_k\}$ with $k \leq K$. Each item is $(\texttt{key}, \texttt{state}, c, \texttt{scope})$: the key names a reusable facet, the state stores its value, $c \in \{1, \ldots, 5\}$ records confidence, and scope records when the belief applies. The board is meant for durable personalization, mirroring the distinction between reusable semantic knowledge and episodic recall in memory research (Nagy et al., 2025; Dong et al., 2025).

Each dialogue event may introduce a new facet, support an existing belief, weaken it, or restrict its scope, corresponding to the four operations REGISTER, REINFORCE, ATTENUATE, and QUALIFY. The formulation makes two questions visible: state quality, whether the current board is useful, and transition quality, whether the update changed the

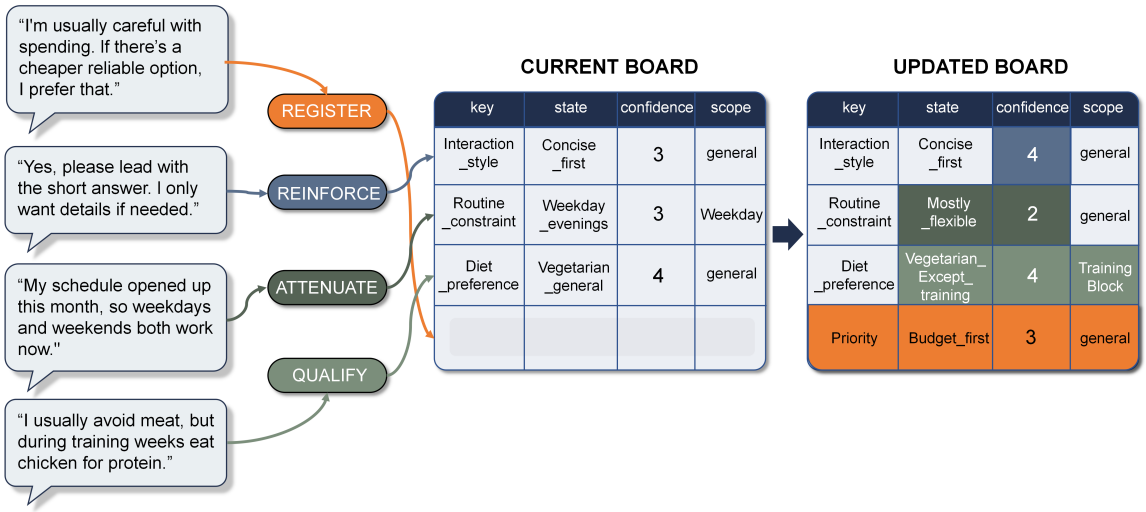

Figure 1: Bounded Revision treats persona memory as an explicit belief state. Dialogue evidence updates a compact board through a typed operation. The example illustrates QUALIFY, where a general belief is preserved but narrowed in scope.

intended belief and left the rest alone. Reconstruction baselines can score well on the first by rebuilding the state from context, while saying nothing about the second.

4 Methodology

Theoretical lineage. Persona revision is a typed instance of belief revision inspired by, rather than identical to, the classical AGM framework, named for Alchourrón, Gärdenfors, and Makinson. AGM (Alchourrón et al., 1985) formalises three rational operations on a belief set: expansion adds a piece of information, contraction removes one, and revision adds new information consistently while removing what conflicts with it. Belief base extensions (Hansson, 1999) drop closure under consequence and operate on enumerable items, which fits a persona board. Our four operations adapt these moves to persona memory: REGISTER aligns with base expansion, ATTENUATE with base contraction, and QUALIFY with a guarded refinement, a contraction that retains a residue under a narrower scope rather than discarding the original item. REINFORCE is not a standard AGM operator; we add it because evidential strengthening is a behaviour real users produce and that summary baselines tend to flatten. Two desiderata follow. The success postulate asks that an applied operation actually changes the targeted item, which we measure with operation accuracy. The minimal-change principle asks that unrelated items stay fixed, which we measure with non-target preservation.

Update procedure. BR is the update procedure used in our experiments. For each user, it stores one current board $\mathcal{B}_t$. When new evidence e_{t+1} arrives, the model revises this board directly; it does not reread the full transcript or rewrite a free-form summary. The output contains one operation, one target belief key, and the next board. Because the operation is part of the transition, the same final board cannot hide whether the model overwrote, weakened, or narrowed a belief.

Let $\mathcal{O}$ denote the four operations REGISTER, REINFORCE, ATTENUATE, and QUALIFY. At each event, BR applies the constrained update

$$(o, \tau, \mathcal{B}_{t+1}) = f_\theta(\mathcal{B}_t,\, e_{t+1},\, \mathcal{C}),$$
$$\text{s.t.} \quad o \in \mathcal{O}, \quad |\mathcal{B}_{t+1}| \leq K,$$

where each row has the form $b = (k, s, c, q)$ with $k \in \mathcal{K}$, $s \in \mathcal{S}_k$, confidence $c \in \{1, \ldots, 5\}$, and a short scope q. Following the minimal-change principle, every row except the target should remain fixed unless capacity pruning removes a low-confidence row; this is the property we measure as non-target preservation.

Table 1 shows the prompt: the model is given the allowed keys and states, the current board, and the incoming evidence, and returns one operation, one target key, and the updated board. A deterministic parser drops unsupported keys or states, clips confidence to the range 1 to 5, deduplicates rows, and truncates low-confidence rows when capacity is exceeded. We default K to the schema size of 16; each script activates roughly 10 keys

BR prompt template
You maintain a bounded persona board for one user. Given the allowed keys and states, the current board, and the incoming evidence, choose exactly one operation from REGISTER, REINFORCE, ATTENUATE, and QUALIFY. Do not store one-off facts. Prefer QUALIFY when an old belief is partly correct but too broad. Return JSON with `operation`, `target`, and `board`; each board item carries `belief_key`, `state`, `confidence`, and `scope`.

Table 1: Prompt template used by BR.

Example event
PREV_BOARD: { `food_preference`: likes spicy, confidence 4, scope generally } EVENT: I still love spicy food, but not when I'm cooking for my kid. GOLD_OPERATION: QUALIFY TARGET: `food_preference` NEXT_BOARD: { `food_preference`: likes spicy, confidence 4, scope except when cooking for child }

Table 2: A persona-relevant event in PersonaBoard, showing the full transition: previous board, target belief, gold operation, and next board.

on average, so capacity pruning rarely triggers and the high non-target preservation reported later is not an artifact of pruning. A lightweight validator enforces basic consistency: REINFORCE cannot lower confidence and QUALIFY must change the state or scope. These checks close the state space under the schema but do not repair semantic mistakes.

We compare BR with three baselines that cover the dominant ways long-term memory is handled today: full history (FH) reconstructs the board from all prior dialogue, retrieval (LexRET) reconstructs it from lexical-overlap top snippets, and rolling summary (RS) reconstructs it from a free-form running summary. For FH/LexRET/RS the operation label is inferred from consecutive boards after the fact; BR emits it during the update.

5 Experiments

PersonaBoard. PersonaBoard is, to our knowledge, the first benchmark that supervises the full trajectory of a persona profile rather than only the final answer to a downstream question. It contains 100 multi-session scripts with 2,100 event boundaries; each script consists of persona-relevant events, reveals that introduce a belief and later evidence that reinforces, weakens, or narrows it, with distractor turns embedded as ambient dialogue. Every event boundary is annotated with a gold board, target belief, and operation. The benchmark uses a closed schema of 16 belief keys grouped into seven facets such as disposition, preference, routine, and social tendency, with each key carrying a small set of allowed states; this is also the natural shape for production persona stacks that need stable slots to share label space with downstream tools. Each script's user touches most of the 16 keys over its lifetime, which explains why REGISTER accounts for 1,600 of the 2,100 events, with the remaining 500 split among 200 ATTENUATE, 200 QUALIFY, and 100 REINFORCE. A 30-script subset of the main set serves as a controlled subset for the same-setting comparison with external memory systems and for the QUALIFY ablation. Scripts are template-generated and then hand-verified event by event, and the released corpus, including scripts, annotations, prompts, and the evaluation script under CC BY 4.0, is illustrated in Table 2. To check label reliability, three annotators recruited through an online task platform and paid a fixed task fee corresponding to an estimated hourly wage above the local minimum wage independently relabelled a balanced 100-event sample, shown only the dialogue and the schema and blind to the gold labels and to one another's annotations; mean pairwise Cohen's κ is 0.81 on the operation label and 0.93 on the target key, with most disagreement on the QUALIFY/ATTENUATE boundary.

External benchmarks. PERMA tests persona preference consistency through multiple-choice questions (MCQs) (Liu et al., 2026), and Alps-Bench evaluates memory quality from user profiles (Xiao et al., 2026). Both target reusable user beliefs rather than verbatim event recall, which makes them aligned with what BR is designed for. We use PERMA as our main external evaluation and Alps-Bench, on a small dev subset, as a secondary check; we discuss the boundary with episodic-recall benchmarks such as LoCoMo (Maharana et al., 2024) in §6.

Metrics. We use two families of metrics. Final-state metrics score the board after all events are processed: BoardF1 is the micro-averaged F1 over board rows matched on `(belief_key, state)`, and probe accuracy measures whether a held-out preference query can be answered from the final board. Transition metrics score each event boundary independently: operation accuracy (OpAcc) is the fraction of events where the predicted operation

equals the gold operation; MacroOp is balanced accuracy, the unweighted mean of OpAcc across the four gold operation classes, which counters the heavy REGISTER skew in the event stream; target accuracy is the fraction of events where the predicted target belief key equals the gold target; and non-target preservation is the fraction of gold non-target keys present at event t whose `(state, scope)` is unchanged in the predicted next board, where a missing or modified row counts as not preserved. Scope is captured both inside non-target preservation and as a per-event scope accuracy reported alongside main results. For BR the operation is emitted directly; for FH/RS/LexRET it is read off from consecutive boards, consistent with how those methods are typically deployed. We use a paired bootstrap over scripts with 1,000 resamples for significance on the controlled subset.

Models. DeepSeek-V3 is the reference model. The cross-model study covers 12 model settings from the DeepSeek, OpenAI, Anthropic, Google, Z.ai, Moonshot, and Qwen families.

5.1 PersonaBoard Main Results

Table 3 reports the 100-script main benchmark; a controlled 30-script subset is used later for the same-setting comparison and the QUALIFY ablation. Rolling summary is the strongest final-board tracker at 0.964 BoardF1 and 0.923 probe accuracy, with full history close behind. BR trades 7 to 9 points of final BoardF1 for the best transition metrics: MacroOp 0.728, up 14.7 points over the strongest reconstruction baseline; target accuracy 0.831; and non-target preservation 0.966, up 24.1 points. Reconstruction methods score higher on raw OpAcc, but REGISTER alone is 1,600 of the 2,100 events, so raw OpAcc largely measures how well a method appends new rows. We therefore report MacroOp, the balanced accuracy across the four operation classes, as the primary operation-level metric; balanced accuracy is standard for skewed label distributions. Retrieval has low BoardF1 because sparse snippets miss the accumulated board state.

The gap between BoardF1 and MacroOp is the main empirical distinction. FH and RS reconstruct good boards by rewriting freely, but at the cost of churning unrelated rows: their non-target preservation is 0.725 and 0.637, against 0.966 for BR. Silently overwriting unrelated user beliefs is a more damaging failure in a long-running deployment than a few points of BoardF1. Retrieval has the opposite failure mode, identifying an update cue in the latest snippet but missing the accumulated state. Probe accuracy follows BoardF1 rather than MacroOp: a held-out preference query can be answered from the surface state regardless of how it was reached, so RS leads here too. BR's 0.966 non-target preservation against LexRET's 0.163 shows the opposite extreme of how much unrelated state a method silently rewrites.

Same-setting comparison. Table 4 compares BR on the controlled 30-script subset to three baseline families under the same DeepSeek-V3 backend: reconstruction baselines FH and RS; external memory systems Mem0, MemOS, and A-MEM, whose free-form snapshot we reconstruct with the same prompt as FH/RS; and op-aware variants FH+Op / RS+Op, given the same previous predicted board and latest event as BR plus the full history or rolling summary, and prompted to emit BR's output schema directly. FH, RS, MemOS, and A-MEM all reach 0.95 to 0.96 BoardF1, ahead of BR's 0.872, and the same ordering holds on raw OpAcc because the REGISTER-heavy stream rewards rebuilding the board; Mem0's free-form sentences lose alignment with the typed schema and trail on both. BR's advantage is on the balanced transition metrics: MacroOp 0.716 against at most 0.555, and non-target preservation 0.971 against at most 0.906. The op-aware variants isolate the source of the gain. Adding the typed schema to RS does raise MacroOp to 0.735 and preservation to 0.985, matching or slightly exceeding BR, but it costs 14 points of BoardF1, from 0.961 to 0.821, and 21 points of raw OpAcc, from 0.841 to 0.629, and FH+Op degrades on every metric. BR reaches comparable transition fidelity without that sacrifice, keeping BoardF1 at 0.872 and OpAcc at 0.779 well above either variant; the local-update objective, not the output schema alone, is what makes operation typing pay off.

5.2 Operation Faithfulness

To isolate revision quality, Table 5 reports per-operation accuracy on the 100-script main set. BR is strongest on REINFORCE at 0.78 and QUALIFY at 0.73, the operations that require comparing new evidence against the existing board; reconstruction baselines collapse there, reaching at most 0.25 on REINFORCE and 0.49 on QUALIFY, because rebuilding a state from context cannot tell a

	Final board		Update quality				
Method	BoardF1	Probe	OpAcc	MacroOp	Target	TargetScope	Preserve
BR (ours)	0.874	0.823	0.776	**0.728**	**0.831**	0.975	**0.966**
FH	0.946	0.803	0.842	0.581	0.776	0.973	0.725
LexRET	0.446	0.683	0.764	0.272	0.549	0.962	0.163
RS	**0.964**	**0.923**	**0.854**	0.567	0.806	**0.987**	0.637

Table 3: Main results on PersonaBoard. BoardF1 matches on `(key, state)`; TargetScope is per-event scope accuracy on the target row. RS is the strongest final-board reconstructor; BR gives the best macro operation accuracy, target accuracy, and non-target preservation.

Method	BoardF1	OpAcc	MacroOp	Preserve
BR (ours)	0.872	0.779	**0.716**	**0.971**
FH	0.948	0.832	0.549	0.741
RS	0.961	0.841	0.553	0.633
Mem0	0.783	0.546	0.318	0.856
MemOS	**0.964**	0.856	0.536	0.906
A-MEM	0.957	**0.859**	0.555	0.904
FH+Op	0.598	0.622	0.567	0.940
RS+Op	0.821	0.629	0.735	0.985

Table 4: Same-setting comparison on the controlled 30-script subset, all under the same DeepSeek-V3 backend. Top: our methods. Middle: external memory systems, isolated per script. Bottom: op-aware variants of FH/RS that emit the typed operation directly, matching BR's output schema. Bold marks the column best among the top two groups.

Method	Register	Reinforce	Attenuate	Qualify
BR (ours)	0.804	**0.780**	0.605	**0.725**
FH	0.949	0.250	0.635	0.490
LexRET	**0.991**	0.000	0.090	0.005
RS	0.973	0.220	**0.695**	0.380

Table 5: Operation accuracy by gold operation on the 100-script main set. BR leads on REINFORCE and QUALIFY, the operations that require comparing new evidence with the existing board; reconstruction baselines win REGISTER by trivially appending a row.

strengthened belief from a restated one. On REGISTER reconstruction wins by trivially appending a row. A paired bootstrap over scripts with 1,000 resamples confirms the aggregate picture: BR's MacroOp exceeds full history by $+0.148$, with a 95% confidence interval of $[+0.095, +0.200]$, and rolling summary by $+0.164$, with interval $[+0.121, +0.209]$, both at $p < .001$, while it trails on BoardF1 by 0.073 to 0.091, also at $p < .001$, a deliberate trade of final-state overlap for transition fidelity.

Method	BoardF1	Target	OpAcc	Probe
BR w/ QUALIFY	**0.872**	**0.832**	**0.779**	**0.825**
BR w/o QUALIFY	0.780	0.586	0.665	0.358

Table 6: QUALIFY ablation on the controlled 30-script subset. Target is the accuracy of the operation's target row at each event.

5.3 QUALIFY Ablation

Ablating QUALIFY costs the board its ability to express conditional persona: on the controlled 30-script subset, BoardF1 drops from 0.872 to 0.780, operation accuracy from 0.779 to 0.665, and target accuracy from 0.832 to 0.586, with probe accuracy falling most sharply, from 0.825 to 0.358. Without the operation, scope changes are usually forced into ATTENUATE, which treats a conditional exception as a contradiction. Scoped revision is therefore a real state distinction, not a formatting choice; an attenuated and a qualified belief look similar in final-state overlap but encode different things.

5.4 Scale and Robustness

Cross-model generality. To check that the typed-update representation is not tied to one model family, Table 7 runs BR across 12 model settings from seven providers. This study was conducted during development on an earlier configuration of PersonaBoard with shorter scripts, so its absolute scores run higher than Table 3 and it should be read for the cross-provider pattern rather than for comparison with the main results. Most runs cluster above BoardF1 $= 0.95$, while the only sub-100B model drops to 0.80, a mild but consistent scaling trend; the representation transfers across providers and scales.

Board capacity. A capacity stress test, shown in Figure 2, asks whether bounded memory collapses as more persona keys become active. It does not, as long as K is not too small: $K = 12$

Model	BoardF1	OpAcc	Model	BoardF1	OpAcc
Claude Opus-4.7	**0.995**	0.917	DeepSeek-R1	0.936	0.857
GPT-5.4	0.983	0.940	Claude Opus-4.6	0.934	0.908
GLM-5.1	0.980	0.917	Gemini-3.1	0.915	0.763
GPT-5.5	0.979	0.967	Qwen3.5-35B	0.800	0.907
Qwen3.5-397B	0.976	0.877	Kimi-K2.6	0.945	**0.977**
DeepSeek-V3.2	0.965	0.943	DeepSeek-V3	0.949	0.943

Table 7: Cross-model BR results across 12 model settings and seven providers, on an earlier configuration of PersonaBoard with shorter scripts, so scores are not directly comparable to Table 3. Frontier-tier runs cluster near BoardF1 ≥ 0.95; the only sub-100B model, Qwen3.5-35B, drops to 0.80, a mild but visible scaling pattern. Gemini-3.1's lower operation accuracy reflects a model-specific update bias, discussed in §6.

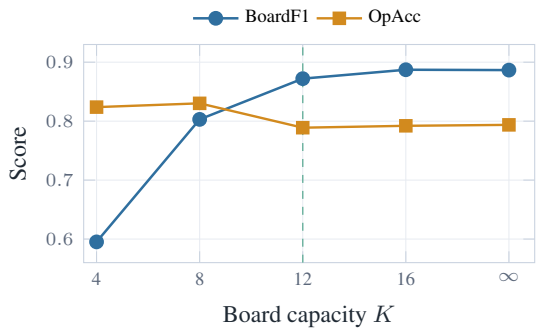

Figure 2: Capacity stress test on the 16-key setting. BoardF1 improves sharply from $K = 4$ to $K = 12$ and then approaches the unbounded score. Operation accuracy stays stable, suggesting that capacity mainly affects how much persona state can be retained.

reaches 0.872 BoardF1, close to the unbounded 0.887, while $K = 4$ falls to 0.595.

Bounded memory is not meant as an arbitrary compression constraint, but as a way to keep the persona layer from growing with every utterance. Once capacity is large enough for the active facets in the stream, the board behaves like a stable working profile rather than a lossy summary of the transcript.

Prompt cost. BR averages $1{,}901$ prompt tokens per event, about 68% of the full-history budget of $2{,}812$, while rolling summary and retrieval use $1{,}583$ and $1{,}671$.

5.5 External Evaluations

We complement PersonaBoard with two external benchmarks. PERMA tests whether the maintained profile supports future preference choices on $1{,}094$ MCQs; AlpsBench, on a 30-example dev subset, judges profile usefulness under a free-form rubric. Memory is rebuilt with BR, FH, LexRET, and RS under the same interface in both.

On PERMA, BR reaches 0.815 MCQ accuracy, leading the same-interface group; the PERMA-paper systems are shown for scale, not as a strict leaderboard, since the setups differ. PERMA fits the intended use case: it tests whether memory supports future preference choices, not whether it can replay past events. BR also reverses RS's within-PersonaBoard Probe advantage here, suggesting that the verbatim coverage helpful on short probes is brittle when answers require reasoning over scoped beliefs.

AlpsBench uses free-form memories while BR uses fixed keys and states; we run a 30-example dev subset as an auxiliary probe and report BR leading on all four judge dimensions, suggesting that typed persona boards remain useful outside PersonaBoard.

6 Discussion

The main value of BR is that it turns persona memory into an explicit state update problem: at every event the system records whether a belief was created, strengthened, weakened, or narrowed. This matters because profile memory is reused across future calls and may need to be inspected or corrected by users. Across the main benchmark, the controlled setting, and PERMA, BR is strongest when the evaluation asks whether the right belief changed and whether unrelated beliefs stayed stable. QUALIFY carries much of this weight: a model that cannot represent a belief that holds in general but not when cooking for a child has to choose between forgetting the preference and ignoring the exception, and neither produces a useful working profile.

Why typed operations matter for users, not just metrics. The four-operation vocabulary is also what makes a persona profile inspectable. When an exchange narrows a previously general preference, BR surfaces the change as a typed line, QUAL-

	BR	RS	FH	LexRET	MemOS†	Memobase†	EverMemOS†	RAG†	Mem0†	Lightmem†	Supermemory†
Acc.	**0.815**	0.778	0.763	0.722	0.811	0.733	0.728	0.702	0.686	0.657	0.655

Table 8: PERMA preference-consistency evaluation on 10 users and 1,094 MCQs. BR, RS, FH, and LexRET run under our shared memory interface; bold marks the best among them. †Scores reported by the PERMA paper.

Method	Cov.	Prec.	Update	Overall
BR (ours)	**3.27**	**4.07**	**3.87**	**3.73**
FH	2.63	3.87	3.50	3.40
LexRET	2.97	4.03	3.37	3.50
RS	3.10	4.03	3.53	3.53

Table 9: AlpsBench memory quality on a 30-example dev subset, used as an external probe rather than a full leaderboard. Judge dimensions score coverage, precision, update quality, and overall utility on a 1 to 5 scale.

IFY on communication_style with scope drafting, a line the user can accept, reject, or edit, whereas a summary-based system simply rewrites the profile. The typed diff also makes wrong updates cheap to revert: undoing a single mistaken ATTENUATE is a one-line edit, while a summary-based agent would have to regenerate state from full context.

Operation accuracy exposes model-specific update bias. A high final-state score can hide wrong updates. Gemini-3.1 reaches 0.915 BoardF1 but only 0.764 operation accuracy: it labels 80% of QUALIFY events and 43% of REINFORCE events as ATTENUATE. The final board still overlaps with the gold board because a belief usually ends up under the right key even when the wrong operation produced it. The sequence of updates, however, is wrong throughout: the model treats narrowed exceptions as contradictions and supporting evidence as weakening. Operation accuracy makes this failure visible where BoardF1 does not.

Reconstruction methods may suffice for one-shot question answering from full context, while BR trades a small amount of final-board score for more controlled transitions, most visibly when many beliefs coexist and a new event touches only one. BR is also deliberately specialised to the agent's impression of the user, not its recollection of past events: benchmarks like LoCoMo (Maharana et al., 2024) that ask for specific names and dates fall outside the scope of a bounded persona profile, and a deployed agent should pair BR with a retriever for that recall. Most remaining errors come from target selection when many beliefs coexist.

7 Conclusion

We proposed bounded persona belief tracking as a paradigm for maintaining the agent's living impression of a user. Instead of treating long-term memory as transcript storage or summary rewriting, Bounded Revision keeps a small structured profile and edits it through typed local updates that support faithful revision, scoped beliefs, and strong preservation of unrelated beliefs, while leaving event-level recall to retrieval. A natural next step is to extend PersonaBoard toward an open-vocabulary schema, more naturalistic dialogues, and an explicit no-op operation for turns that should not enter persona memory.

Limitations

PersonaBoard uses a closed schema and synthetic, human-verified scripts, which makes event-level supervision well-defined but does not cover free-form profile fields or no-op decisions on arbitrary turns. BR is specialised to reusable persona beliefs, not factual recall, so applications needing such recall should pair it with a retriever. We measure update quality against gold transitions but not its downstream effect on response quality, which we leave to future work.

Ethical Considerations

Persona memory aids personalization but can preserve sensitive inferences and is a known attack surface for LLM applications (Herrador and Rehberger, 2026). A bounded, inspectable board eases deletion, correction, and auditing; systems built on it should expose stored beliefs to users, support deletion, and not infer protected attributes without consent.

Use of AI Assistance

We used general-purpose AI assistants for language polishing and for discussion during paper preparation. The large language models studied as experimental subjects are documented in the methodology.

References

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530.

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. *Advances in Neural Information Processing Systems*, 37:42528–42565.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2025. Post Persona Alignment for Multi-Session Dialogue Generation. *arXiv preprint arXiv:2506.11857*.

Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, and Rui Yan. 2024. "In-Dialogues We Learn": Towards Personalized Dialogue Without Pre-defined Profiles through In-Dialogue Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10422.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv preprint*. ArXiv:2504.19413 [cs].

Renke Dai, Hebin Hu, Jiahui Zhang, Yilin Kang, and Ah-Hwee Tan. 2026. MemoryART: Enhancing LLMs via Multi-Memory Models with Adaptive Resonance Theory for Healthcare Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20676–20683.

Adnan Darwiche and Judea Pearl. 1997. On the logic of iterated belief revision. *Artificial intelligence*, 89(1-2):1–29.

Cody V. Dong, Qihong Lu, Kenneth A. Norman, and Sebastian Michelmann. 2025. Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*.

Bernal J. Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569.

Sven Ove Hansson. 1999. *A Textbook of Belief Dynamics : Theory Change and Database Updating*. Kluwer Academic Publishers.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 2023. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 277–280.

Manuel Herrador and Johann Rehberger. 2026. SpAIware: Uncovering a novel artificial intelligence attack vector through persistent memory in LLM applications and agents. *Future Generation Computer Systems*, 174:107994.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540.

Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Wenxuan Wang, and Qin Jin. 2026. Mem-pal: Towards memory-based personalized dialogue assistants for long-term user-agent interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31229–31237.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25972–25981.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost In Multi-Turn Conversation. In *The Fourteenth International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. Hello again! llm-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. 2025b. Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11504–11517.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, and 3 others. 2025c. MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models. *arXiv preprint*. ArXiv:2505.22101 [cs].

Xuechen Liang, Meiling Tao, Yinghui Xia, Jianhui Wang, Kun Li, Yijin Wang, Yangfan He, Jingsong

Yang, Tianyu Shi, and Yuantao Wang. 2025. Sage: Self-evolving agents with reflective and memory-augmented abilities. *Neurocomputing*, 647:130470.

Dongshuo Liu, Zhijing Wu, Dandan Song, and He-Yan Huang. 2025. A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 103–123.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. 2026. PERMA: Benchmarking Personalized Memory Agents via Event-Driven Preference and Realistic Task Environments. *arXiv preprint*. ArXiv:2603.23231 [cs].

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

David G. Nagy, Gerg\Ho Orbán, and Charley M. Wu. 2025. Adaptive compression as a unifying framework for episodic and semantic memory. *Nature Reviews Psychology*, 4(7):484–498.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint*. ArXiv:2310.08560 [cs].

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. Meminsight: Autonomous memory augmentation for llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33124–33140.

Taize Sun and Katsuhide Fujita. 2025. Transformer model with external token memories and attention for PersonaChat. *Scientific Reports*, 15(1):20691.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025a. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19336–19352.

Zhen Tan, Jun Yan, I.-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, and George Lee. 2025b. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025a. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian Mcauley, Dan Gutfreund, Rogerio Feris, and Zexue He. 2025b. M+: Extending MemoryLLM with Scalable Long-Term Memory. In *International Conference on Machine Learning*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4891–4906.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025c. Agent Workflow Memory. In *International Conference on Machine Learning*, pages 63897–63911. PMLR.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. 2024. Belief Revision: The Adaptability of Large Language Models Reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10496, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*.

Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2026. AlpsBench: An LLM Personalization Benchmark for Real-Dialogue Memorization and Preference Alignment. *arXiv preprint*. ArXiv:2603.26680 [cs].

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 5180–5197.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2025. Prime: Large language model personalization with cognitive dual-memory and personalized thought process. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33695–33724.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19724–19731.