# RETRIEVAL IS NOT UNDERSTANDING: THE EVIDENCE-TO-STATE BOTTLENECK IN PERSONALIZED LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-term personalized large language models (LLMs) must use information from prior interactions to support decisions in the current context. Memory-centric systems have made substantial progress in storing, retrieving, and organizing historical information, while recent work has also introduced explicit user representations. However, it remains unclear how much user-state error persists when retrieval failure is explicitly removed.

We isolate this post-retrieval problem as the *Evidence-to-State bottleneck*. Under an **Oracle Evidence** setting, task-relevant historical evidence is directly provided, allowing us to evaluate whether the model can recover the corresponding benchmark-defined user state without confounding retrieval errors. Using PersonaMem-v2, we represent each state by its type, subject, content, and applicability. Even under this favorable setting, a base LLM achieves only **19.81% Full-State Accuracy**, indicating substantial residual error in user-state interpretation.

To test whether this bottleneck can be substantially reduced with direct supervision, we apply a simple controlled state-supervision intervention, which raises Full-State Accuracy to **65.28%**. We then perform matched state replacement while holding the query, Oracle Evidence, candidate options, and downstream decision model fixed. Replacing base-generated states with improved states raises personalized decision accuracy from **47.10% to 68.81%**, recovering **78.03%** of the decision gap between the base-state and reference-state conditions. Error-level analysis further shows that different state-error categories are associated with markedly different downstream decision gaps.

These results show that making the right user evidence available is not sufficient for reliable personalization: models must also correctly interpret what that evidence implies about the user. Our findings motivate treating Evidence-to-State reasoning as an explicit evaluation target in personalized LLM systems.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly expected to support long-term personalized interactions. Benchmarks such as LaMP, PersonalLLM, PersonaMem, and PersonaMem-v2 have made it possible to study how models use user-specific histories and preferences across tasks (Salemi et al., 2024; Zollo et al., 2025; Jiang et al., 2025a;b). In parallel, long-term memory systems and benchmarks have advanced the storage, retrieval, and maintenance of information across extended interactions (Zhong et al., 2024; Maharana et al., 2024; Wu et al., 2024; Chhikara et al., 2025).

These developments raise a distinct post-retrieval question. Even when task-relevant historical evidence is already available in context, a model must still determine what that evidence implies about the user for the current task. This distinction matters because relevant evidence can still be misinterpreted: a model may attribute another person's information to the user, turn temporary behavior into a persistent preference, infer unsupported information, or continue using information that should no longer be active.

We therefore decompose personalized decision making as



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.

$$H \rightarrow E \rightarrow Z \rightarrow Y, \tag{1}$$

where $H$ is user history, $E$ is task-relevant evidence, $Z$ is the inferred user state, and $Y$ is the final personalized decision. Recent work has begun to model user states explicitly or to introduce post-retrieval memory processing (Luo et al., 2026; Wang et al., 2026; Ma et al., 2026). Our focus is complementary: rather than proposing user-state modeling itself, we isolate the $E \rightarrow Z$ transition as an evaluation target by holding evidence quality fixed. Figure 1 summarizes this decomposition and highlights the Evidence-to-State bottleneck studied in this work.

Our central question is:

*When task-relevant evidence is already provided, how reliably can an LLM recover the user state needed for the current task?*

To separate retrieval failure from interpretation failure, we directly provide Oracle Evidence for the current query. Using PersonaMem-v2, we map benchmark metadata into a structured benchmark-defined state with four components: type, subject, semantic content, and applicability. We then evaluate predictions using Full-State Accuracy, which requires the complete structured state to be correct.

Despite receiving Oracle Evidence, the base model achieves only **19.81% Full-State Accuracy**. We use standard supervised fine-tuning as a controlled intervention, rather than as a new training algorithm, to test whether this post-retrieval bottleneck can be substantially reduced. The resulting state-supervised condition reaches **65.28% Full-State Accuracy**.

We next test whether improved intermediate states matter for the final personalized decision. Holding the query, Oracle Evidence, candidate options, prompt, and downstream decision model fixed, replacing base-generated states with the improved states increases personalized decision accuracy from **47.10% to 68.81%**. Relative to the reference-state condition, this recovers **78.03%** of the decision gap. We further find that different state errors are associated with substantially different downstream gaps.

Our contributions are threefold:

- We isolate the post-retrieval Evidence-to-State stage under Oracle Evidence, separating user-state interpretation from memory retrieval.
- We quantify a substantial residual interpretation bottleneck even when task-relevant evidence is directly available, and show through a controlled supervision intervention that this bottleneck can be substantially reduced.
- Through matched state replacement and error-level analysis, we show that state quality is consequential for downstream personalized decisions and that different state errors are associated with heterogeneous decision gaps.

## 2 Related Work

### 2.1 Long-Term Memory and Personalization

Personalized LLMs use user-specific histories, preferences, and prior interactions to adapt model outputs. LaMP introduced a multi-task benchmark for personalized language modeling and studied retrieval-augmented personalization (Salemi et al., 2024), while PersonalLLM evaluates adaptation to heterogeneous individual preferences (Zollo et al., 2025). PersonaMem studies dynamic user profiling and personalized response selection (Jiang et al., 2025a), and PersonaMem-v2 extends this setting to more implicit preferences and long-context interactions (Jiang et al., 2025b). PPlug and PRIME likewise construct compact or structured representations of user histories for personalized generation (Liu et al., 2025; Zhang et al., 2025b).

A parallel line of work develops long-term memory mechanisms and benchmarks. MemoryBank introduces persistent memory with time-dependent updating (Zhong et al., 2024); LoCoMo evaluates very long-term conversational memory (Maharana et al., 2024); LongMemEval decomposes

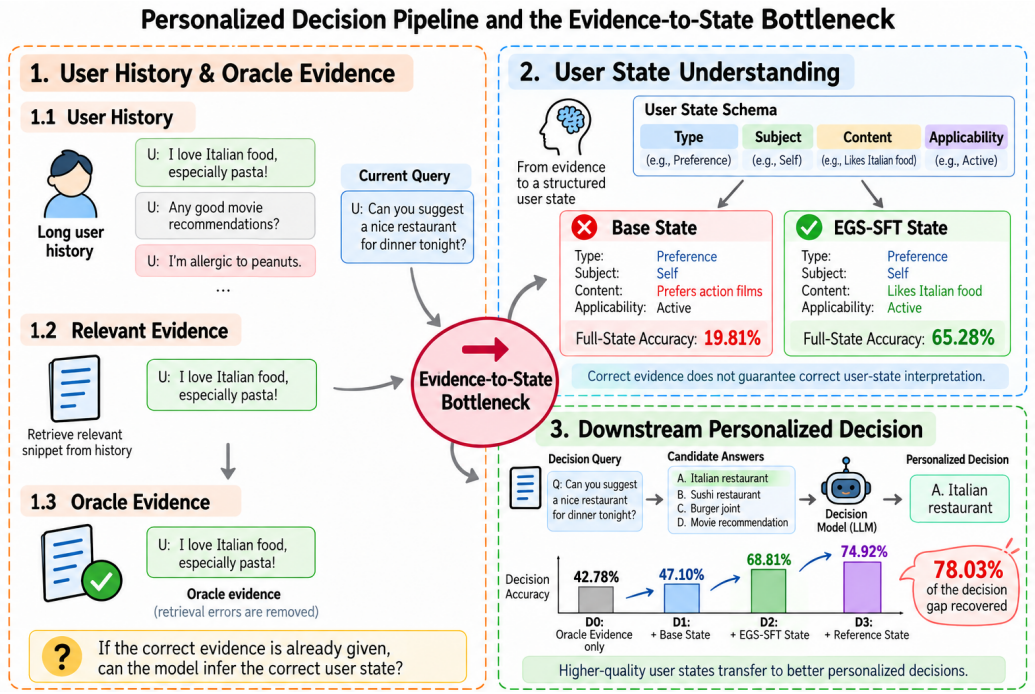


Figure 1: **Overview of the personalized decision pipeline and the Evidence-to-State bottleneck with an illustrative example.** The example is used only to visualize the pipeline and is not an evaluation instance sampled from PersonaMem-v2. Under Oracle Evidence, the task-relevant historical evidence $E^*$ is directly provided together with the current query $Q$. A state-inference LLM $f_\theta$ maps $(Q, E^*)$ to a structured user state containing type, subject, content, and applicability. We compare base-generated and state-supervised states under the same evidence, then feed the supplied state to a fixed downstream decision LLM $g_\phi$. This design isolates post-retrieval state interpretation and allows state quality to be varied while the surrounding decision context is held fixed.

long-term memory into indexing, retrieval, and reading (Wu et al., 2024); and Mem0 develops a scalable memory-centric architecture for extracting, consolidating, and retrieving salient information (Chhikara et al., 2025). These works establish the importance of preserving and retrieving useful historical information. Our study instead fixes the relevant evidence and asks what happens after retrieval has succeeded.

### 2.2 User Modeling and User-State Inference

Prior work has studied both explicit and implicit representations of users. Implicit personalization examines how models infer user attributes from indirect cues (Jin et al., 2024), while persona-inference methods recover user characteristics from observed preferences (Balepur et al., 2025). PUMA explicitly models evolving latent user states for personalized multi-turn action selection (Luo et al., 2026). QUMem constructs typed memories and uses multi-stage retrieval to infer a query-conditioned, temporally valid user state for downstream generation (Wang et al., 2026). Recent benchmarks such as LoCoMo-Plus also emphasize latent constraints, goals, and values that must be retained and applied beyond factual recall (Li et al., 2026).

These studies show that user-state modeling is itself important. Our distinction is experimental rather than terminological: we do not claim user-state inference as a new mechanism. Instead, we hold task-relevant evidence fixed with Oracle Evidence and treat the Evidence-to-State transition itself as the object of evaluation. This disentangles failures in finding evidence from failures in interpreting evidence.

Table 1: Structured user-state representation.

| Component | Values | Meaning |
|---|---|---|
| Type | Preference / Fact / Delete / Other | Information type |
| Subject | Self / Other / Unknown | State subject |
| Content | Free-form text | Semantic content |
| Applicability | Active / Deleted / N/A | Current status |

### 2.3 Post-Retrieval Reasoning and Context Use

Retrieval-augmented generation conditions language models on external evidence (Lewis et al., 2020), but access to relevant context does not guarantee that the model will use it correctly. Long-context models can underuse relevant information depending on its position (Liu et al., 2024), and context-faithfulness studies show that models may ignore, distort, or conflict with supplied evidence (Li et al., 2025; Zhang et al., 2025a). Other methods explicitly optimize evidence-grounded reasoning or verify faithfulness after retrieval (Wang et al., 2025; Filice et al., 2025). MemChain is especially close to our motivation: it learns a post-retrieval memory policy that transforms retrieved candidates into a structured evidence trace before downstream answering (Ma et al., 2026).

Our work complements these approaches by focusing on personalized user-state interpretation. We directly evaluate the mapping from fixed Oracle Evidence to a benchmark-defined user state and then use matched state replacement to measure how changes in that intermediate state are associated with downstream decision performance.

## 3 Problem Formulation

Given user history $H$ and current query $Q$, a conventional retrieval mechanism produces relevant evidence $E = R(H, Q)$. Our diagnostic setting replaces this retrieved evidence with **Oracle Evidence** $E^*$, i.e., the task-relevant evidence supplied by the benchmark. The state-inference model then predicts

$$\hat{Z} = f_\theta(Q, E^*), \tag{2}$$

and the downstream decision model produces

$$\hat{Y} = g_\phi(Q, E^*, \hat{Z}), \tag{3}$$

where $g_\phi$ is held fixed in the matched state-replacement experiments. This distinction separates retrieval failure from interpretation failure: once $E^*$ is fixed, differences in state quality arise after the relevant evidence has already been provided.

**Structured state.** We represent the task-relevant user state as

$$Z = (T, S, C, A), \tag{4}$$

where $T$ is state type, $S$ is subject, $C$ is semantic state content, and $A$ is applicability.

PersonaMem-v2 metadata is deterministically mapped into this structure. We treat the resulting states as **benchmark-defined reference targets**: they provide a consistent target for controlled evaluation, but are not assumed to constitute uniquely identifiable latent user states beyond the information encoded by the benchmark. Detailed mapping rules are provided in Appendix A.

**Full-State Accuracy.** For predicted state $\hat{Z}_i = (\hat{T}_i, \hat{S}_i, \hat{C}_i, \hat{A}_i)$ and reference state $Z_i^* = (T_i^*, S_i^*, C_i^*, A_i^*)$, let $J(Q_i, E_i^*, Z_i^*, \hat{Z}_i)$ denote the fixed semantic evaluator label described in Appendix C. We define $s_i = \mathbb{I}[J(Q_i, E_i^*, Z_i^*, \hat{Z}_i) \in \{\texttt{EXACT}, \texttt{EQUIVALENT}\}]$. Full-State Accuracy is

$$\text{FullAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{T}_i = T_i^* \wedge \hat{S}_i = S_i^* \wedge \hat{A}_i = A_i^*]\, s_i. \tag{5}$$

Thus, a sample is counted as fully correct only when all categorical fields match the benchmark reference and the free-form content is judged semantically correct.

# 4 Controlled State-Supervision Intervention

Our goal is not to introduce a new fine-tuning algorithm. Instead, we use standard supervised fine-tuning as a controlled intervention to test whether the identified Evidence-to-State bottleneck can be substantially reduced with direct state supervision.

Both the base and state-supervised models receive the same query $Q$ and Oracle Evidence $E^*$. The supervised condition is trained to generate the serialized reference state $Z^*$ with the standard token-level autoregressive negative-log-likelihood objective. We use Qwen3-4B (Yang et al., 2025) and parameter-efficient QLoRA fine-tuning (Hu et al., 2022; Dettmers et al., 2023) for two epochs. For consistency with the experimental artifacts, we refer to this state-supervised condition as **EGS-SFT**; this label denotes the experimental intervention rather than a new optimization method. Full training details are provided in Appendix B. Training and development personas are disjoint.

# 5 Experiments

## 5.1 Experimental Setup

**Data.** The state-supervision training split contains 16,631 examples from 719 personas. The development set contains 1,898 examples from 80 disjoint personas: 774 `PREFERENCE`, 549 `PERSONAL_FACT`, 378 `MEMORY_DELETE`, and 197 `OTHER`.

**Models and evaluation.** Qwen3-4B (Yang et al., 2025) serves as the base state-inference model and as the backbone for EGS-SFT. Both receive the same query and Oracle Evidence $E^*$. The base condition is not an unconstrained free-form generation setting: its prompt explicitly specifies the structured user-state schema, the allowed categorical values, and the required output format. EGS-SFT therefore tests whether targeted state supervision improves Evidence-to-State inference beyond schema specification alone. Structured fields are evaluated by exact match against the benchmark reference state. For the free-form content field, we use a fixed LLM-based semantic evaluator that compares the predicted content with the corresponding benchmark reference in the context of the query and Oracle Evidence. The evaluator assigns one of eight predefined semantic labels, and only `EXACT` and `EQUIVALENT` are counted as correct. The evaluator prompt and decision criteria are fixed across all model conditions. Full details are provided in Appendix C. **Control Joint Accuracy** requires type, subject, and applicability to be simultaneously correct for the same example.

For downstream evaluation, we keep the query, Oracle Evidence, candidate options, prompt, and downstream decision model fixed and vary only the source of the user state.

**Statistical uncertainty.** Because multiple evaluation examples may belong to the same persona, sample-level observations are not fully independent. We therefore complement sample-level statistics with a paired persona-cluster bootstrap. For each of 10,000 bootstrap replicates, we resample the 80 evaluation personas with replacement and include all examples belonging to each sampled persona. The same resampled personas are used for both conditions in each paired comparison. We report percentile-based 95% confidence intervals.

## 5.2 User-State Interpretation under Oracle Evidence

Despite receiving Oracle Evidence, the base model achieves only **19.81%** Full-State Accuracy, with 376 of 1,898 states fully correct. EGS-SFT increases this to **65.28%** (1,239/1,898), an absolute improvement of **45.47 percentage points**.

Table 2: User-state inference on the persona-disjoint development set. All values are percentages.

| Metric | Base | EGS-SFT |
|---|---|---|
| JSON Validity | 99.89 | 99.95 |
| State Type Accuracy | 46.73 | 98.37 |
| Subject Accuracy | 68.07 | 98.95 |
| Applicability Accuracy | 61.59 | 98.89 |
| Control Joint Accuracy | 43.89 | 98.37 |
| Exact Content Match | 6.22 | 34.98 |
| State-Type Macro-F1 | 47.95 | 97.89 |
| **Full-State Accuracy** | **19.81** | **65.28** |

Table 3: Downstream personalized decision accuracy under matched state replacement.

| Condition | Accuracy (%) |
|---|---|
| $D_0$: Oracle Evidence only | 42.78 |
| $D_1$: + Base State | 47.10 |
| $D_2$: + EGS-SFT State | **68.81** |
| $D_3$: + Reference State | 74.92 |

The gain is not merely a formatting effect. Although the structured control fields approach saturation after fine-tuning, exact content match remains only 34.98%, whereas semantic Full-State Accuracy reaches 65.28%. This indicates that many correct states are expressed through semantically equivalent rather than lexically identical content.

The paired outcomes are 343 examples correct under both models, 33 correct only under the base model, 896 correct only under EGS-SFT, and 626 incorrect under both. A McNemar test yields $p < 10^{-200}$.

Because examples are clustered within personas, we additionally perform a paired persona-cluster bootstrap with 10,000 replicates. The Base Full-State Accuracy is 19.81% (95% CI: 18.05–21.56), while EGS-SFT reaches 65.28% (95% CI: 62.66–67.84). The corresponding improvement is **+45.47 percentage points** (95% CI: **+42.55 to +48.36**), confirming that the improvement remains substantial after accounting for persona-level clustering.

**Performance across state types.** The improvement is not driven by a single state category. EGS-SFT improves Full-State Accuracy across all four benchmark-defined state types: from 2.97% to 45.09% for `PREFERENCE`, 16.21% to 61.75% for `PERSONAL_FACT`, 46.56% to 98.68% for `MEMORY_DELETE`, and 44.67% to 90.36% for `OTHER`. Persona-cluster bootstrap intervals for the improvement exclude zero for every category. At the same time, absolute performance remains heterogeneous across state types, with `PREFERENCE` remaining the most difficult category after fine-tuning. Detailed results are provided in Appendix E.

Figure 2 summarizes the main state-recovery and downstream decision results.

### 5.3 Does Better State Inference Improve Decisions?

We construct four matched conditions:

$$D_0 : (Q, E^*), \qquad D_1 : (Q, E^*, \hat{Z}_{\text{Base}}),$$
$$D_2 : (Q, E^*, \hat{Z}_{\text{SFT}}), \qquad D_3 : (Q, E^*, Z^*). \tag{6}$$

Only the supplied state changes across the state-augmented conditions.

Replacing the base state with the EGS-SFT state increases decision accuracy from 47.10% (95% persona-cluster bootstrap CI: 44.73–49.53) to 68.81% (95% CI: 66.46–71.08). The paired improvement is **+21.71 percentage points** (95% CI: **+19.83 to +23.64**). We summarize the improvement relative to the reference-state condition as

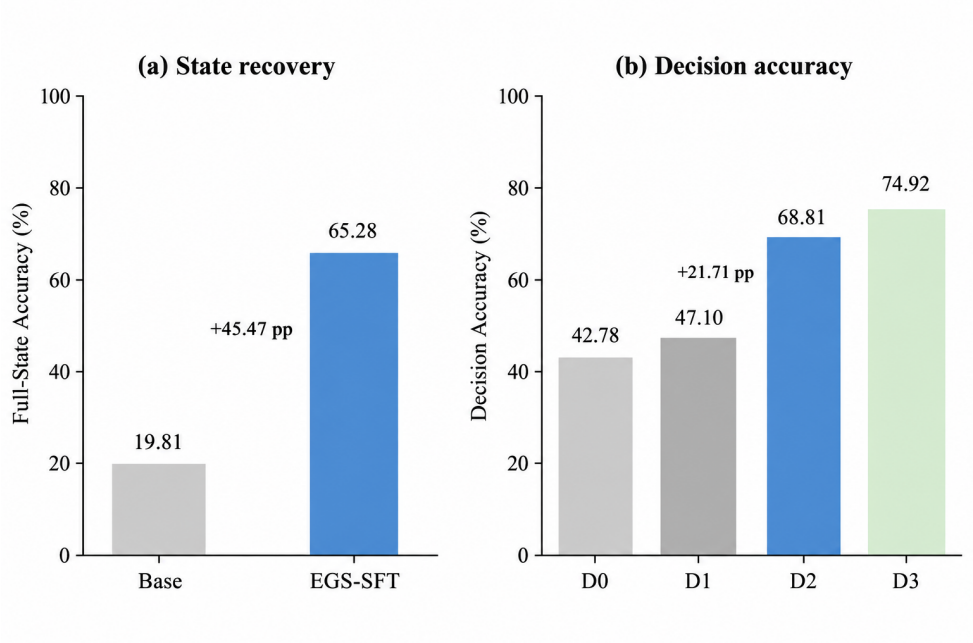


Figure 2: **Main results under Oracle Evidence. (a)** Targeted user-state supervision improves Full-State Accuracy from 19.81% for the base model to 65.28% for EGS-SFT, corresponding to an absolute gain of 45.47 percentage points. **(b)** In the matched state-replacement experiment, downstream decision accuracy increases from 47.10% with the Base State ($D_1$) to 68.81% with the EGS-SFT State ($D_2$), a gain of 21.71 percentage points. The Reference State condition ($D_3$) reaches 74.92%. The $D_1 \rightarrow D_2$ improvement recovers 78.03% of the decision gap between the Base State and Reference State conditions.

$$\text{GapRecovery} = \frac{\text{Acc}(D_2) - \text{Acc}(D_1)}{\text{Acc}(D_3) - \text{Acc}(D_1)} = 78.03\%. \tag{7}$$

Gap recovery is computed from the underlying unrounded accuracies; the accuracies displayed in the table are rounded to two decimal places.

Thus, substantially better intermediate state inference is accompanied by a large improvement in downstream personalized decisions.

### 5.4 Different State Errors Have Different Downstream Impact

Full-State Accuracy treats all incorrect states equally. We therefore examine downstream performance across major state-error categories.

As shown in Figure 3, downstream impact varies substantially across error types. Wrong Target State and Control Error are associated with much larger decision gaps than Unsupported Inference and Over-Generalized errors. This suggests that aggregate state accuracy alone does not fully characterize the downstream consequences of user-state errors.

### 5.5 Is Oracle Evidence Merely Direct Extraction?

Oracle Evidence is intentionally a favorable diagnostic condition that removes retrieval failure, but this raises the question of whether the remaining task reduces to direct extraction and schema filling. We therefore conduct a blinded human audit of 200 development examples, stratified by benchmark state type. Annotators observe only the query, Oracle Evidence, and benchmark-defined reference state, without access to model predictions, correctness labels, semantic-judge outputs, or downstream decisions.

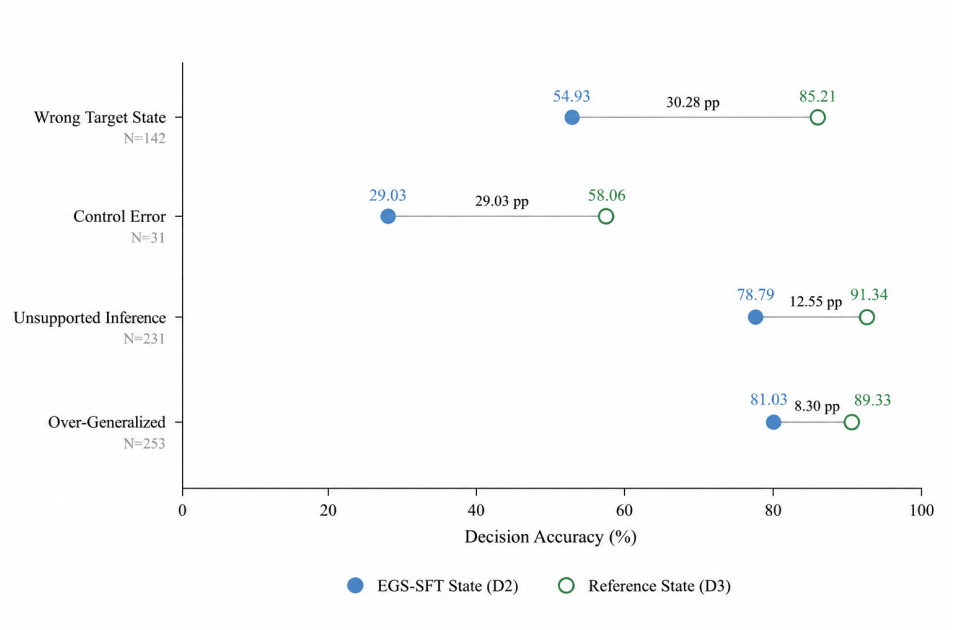


Figure 3: **Downstream decision performance across user-state error types.** Each row compares decision accuracy using the EGS-SFT state ($D_2$, filled blue marker) with the corresponding Reference State condition ($D_3$, outlined green marker). The connecting segment represents the decision-accuracy gap between the two conditions. Wrong Target State and Control Error are associated with substantially larger gaps (30.28 and 29.03 percentage points, respectively) than Unsupported Inference and Over-Generalized errors (12.55 and 8.30 percentage points). These results show that different incorrect user states can have markedly different downstream consequences.

Each audited example is assigned one primary relation: `DIRECT_EXPLICIT`, `IMPLICIT_INFERENTIAL`, `ATTRIBUTION_OR_SCOPE`, or `DELETION_OR_NEGATION`. Because the audit set is stratified by state type, the resulting category frequencies should not be interpreted as prevalence estimates for the full development set. Within the audited sample, 164 of 200 examples fall into the combined non-direct subset. On these examples, Base Full-State Accuracy is 19.51%, whereas EGS-SFT reaches 69.51%, a gain of **50.00 percentage points** (95% persona-cluster bootstrap CI: **42.58–57.42**). The improvement is therefore not confined to directly stated evidence. Detailed audit results are provided in Appendix F.

## 6 Discussion

### 6.1 Retrieval Is Not Understanding

Our results expose a distinction between making user information *available* and interpreting what that information means. Retrieval answers which historical information is relevant, whereas Evidence-to-State reasoning determines what that evidence implies about the user for the current task.

The 19.81% base Full-State Accuracy under Oracle Evidence shows that substantial errors remain even after retrieval failure has been removed. This suggests that personalized memory systems should evaluate post-retrieval interpretation explicitly rather than treating successful retrieval as sufficient for personalization.

Oracle Evidence is intentionally a favorable diagnostic condition rather than a realistic retrieval environment. The blinded evidence-difficulty audit nevertheless shows that the observed gain is not restricted to direct explicit extraction cases: on the audited non-direct subset, EGS-SFT improves Full-State Accuracy by 50.00 percentage points. Real deployment evidence may contain substantially more noise, conflict, and retrieval uncertainty, which remains outside the scope of this study.

Explicit intermediate states also make failures easier to diagnose. They separate missing evidence from incorrect interpretation and expose semantic distinctions such as subject, applicability, and deletion status. The matched state-replacement experiment further shows that these states are consequential: changing the state while holding the surrounding decision context fixed produces large differences in decision accuracy.

### 6.2 State Fidelity and Decision Utility

Improved state inference is strongly associated with improved downstream decision performance, but the relationship is not uniform. Wrong-target errors are associated with much larger decision gaps than local over-generalization errors. The state-type analysis further shows that the bottleneck is heterogeneous even when improvements are directionally consistent: preference states remain substantially harder to recover than deletion or other structured state types.

Future personalization methods may therefore benefit from considering not only whether a state is wrong, but also which errors are most consequential for downstream decisions.

EGS-SFT is used only as a controlled intervention, not as a proposed optimization algorithm. Its role is to test whether the identified Evidence-to-State bottleneck can be substantially reduced with direct state supervision.

## 7 Limitations

Our study focuses on PersonaMem-v2 and a structured state representation with type, subject, semantic content, and applicability; broader datasets and richer, dynamic user states remain to be studied.

The structured reference states are derived from benchmark metadata and should therefore be interpreted as benchmark-defined targets rather than as uniquely observable ground-truth psychological states.

The state-supervised condition is a simple controlled intervention rather than an exhaustive exploration of training objectives. Although the base prompt already specifies the complete state schema and output constraints, we do not exhaustively explore alternative prompt wordings or few-shot in-context demonstrations.

The main experiments use a persona-disjoint development set, so additional untouched test sets would provide stronger evidence of generalization. The 200-example evidence-difficulty audit is stratified by benchmark state type; its category frequencies therefore characterize the audited sample rather than estimate prevalence in the full development set.

Full-State Accuracy also relies partly on semantic-equivalence judgments for free-form state content.

Finally, the downstream state-replacement experiment uses a fixed decision model, and the magnitude of the observed gains may vary across model families and tasks.

## 8 Conclusion

We isolate the post-retrieval Evidence-to-State stage in long-term personalization and show that correct evidence alone does not guarantee correct user-state interpretation.

Under Oracle Evidence, the base model achieves only 19.81% Full-State Accuracy, while a controlled state-supervision intervention raises this to 65.28%.

The improved states also increase downstream decision accuracy from 47.10% to 68.81%, and different state errors exhibit substantially different downstream impacts.

These findings highlight a simple but important point: **retrieving the right memory is not the same as understanding the user**. Evidence-to-State reasoning should therefore be treated as an explicit evaluation target in personalized LLM systems.

## REFERENCES


Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Boyd-Graber. Whose boat does it float? improving personalization in preference tuning via inferred user personas. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Simone Filice, Elad Haramaty, Guy Horowitz, Zohar Karnin, Liane Lewin-Eytan, and Alex Shtoff. Generate but verify: Answering with faithfulness in rag-based question answering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1017–1037. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.56.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12309–12325, 2024. doi: 10.18653/v1/2024.findings-emnlp.717.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026. doi: 10.18653/v1/2026.acl-long.1150.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. Investigating context faithfulness in large language models: The roles of memory strength and evidence style. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4789–4807. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.247.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. Llms + persona-plug = personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 9373–9385. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.461.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Jiani Luo, Xiaoyan Zhao, Yang Zhang, Shuyi Miao, Bingbing Xu, Stefan Konigorski, and Tat-Seng Chua. Know you before you speak: User-state modeling for llm personalization in multi-turn conversation. *arXiv preprint arXiv:2605.24647*, 2026.

Yiwen Ma, Songjun Tu, Qichao Zhang, Dong Li, Linjing Li, and Dongbin Zhao. Memchain: Learning interpretable memory traces for memory-augmented llm agents. *arXiv preprint arXiv:2607.24097*, 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 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)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399.

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. *arXiv preprint arXiv:2608.16168*, 2026.

Suyuchen Wang, Jinlin Wang, Xinyu Wang, Shiqi Li, Xiangru Tang, Sirui Hong, Xiao-Wen Chang, Chenglin Wu, and Bang Liu. Improving context fidelity via native retrieval-augmented reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21205–21218. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1075.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.1062.

Xinliang Frederick Zhang, Nicholas Beauchamp, and Lu Wang. 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*, pp. 33707–33736. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.1711.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, 2025.

## A Reference-State Construction

The structured state is represented as $Z = (T, S, C, A)$. Table 4 summarizes the benchmark mappings used for the explicitly defined categories.

For each row, the content field is $C$. Interactions concerning other people are represented as non-applicable for user personalization. The main-text schema uses shorter human-readable labels (e.g., "Active") while this appendix preserves the serialized labels used by the evaluation artifacts.

Table 4: Reference-state mappings used in our evaluation. $C$ denotes the semantic content derived from the corresponding benchmark metadata.

| Category | Type | Subject | Applicability |
|---|---|---|---|
| Preference-related | PREFERENCE | SELF | ACTIVE_RELEVANT |
| Personal fact | PERSONAL_FACT | SELF | ACTIVE_RELEVANT |
| Memory deletion | MEMORY_DELETE | SELF | DELETED |

## B Training Configuration

We use standard supervised fine-tuning with Qwen3-4B (Yang et al., 2025) and QLoRA (Hu et al., 2022; Dettmers et al., 2023); no new optimization objective is introduced.

Table 5: Training configuration for the controlled state-supervision condition.

| Configuration | Value |
|---|---|
| Backbone | Qwen3-4B |
| LoRA rank | 16 |
| LoRA $\alpha$ | 32 |
| LoRA dropout | 0.05 |
| Maximum sequence length | 1536 |
| Batch size per device | 1 |
| Gradient accumulation | 16 |
| Learning rate | $2 \times 10^{-4}$ |
| Training epochs | 2 |

## C Semantic Evaluation Protocol

Because the state-content field is free-form text, exact string matching alone would underestimate semantically correct predictions. We therefore use a fixed LLM-based semantic evaluator to compare each predicted state with its corresponding benchmark-defined reference state.

For each example, the evaluator receives the current query, Oracle Evidence, benchmark reference state, and predicted state. The evaluator is instructed to judge semantic correspondence rather than surface-form similarity and to assign exactly one of the following labels:

- EXACT: the predicted content matches the reference meaning directly;
- EQUIVALENT: the wording differs, but the predicted and reference states express the same task-relevant meaning;
- OVER_GENERALIZED: the prediction is broader or less specific than the reference;
- WRONG_TARGET_STATE: the prediction captures a different user state from the benchmark reference;
- UNSUPPORTED_INFERENCE: the prediction introduces information not supported by the supplied evidence;
- SUBJECT_ERROR: the prediction attributes the state to the wrong subject;
- DELETION_VIOLATION: information marked as deleted is incorrectly retained or treated as active;
- EMPTY_OR_MISSING: required semantic state content is absent.

Only EXACT and EQUIVALENT are counted as semantically correct. The structured fields—state type, subject, and applicability—are evaluated separately by exact match.

The same evaluator prompt, label definitions, and decoding protocol are used for all model conditions. Model identity and aggregate performance statistics are not provided to the evaluator during per-example judgment.

To assess evaluator reliability, we additionally compare the semantic evaluator against an independently reviewed calibration set of 30 examples. For the binary distinction between semantically correct (EXACT/EQUIVALENT) and incorrect predictions, the evaluator achieves 96.67% agreement with the human annotations and a Cohen's $\kappa$ of 0.933. This calibration evaluates consistency of prediction–reference semantic matching; it does not establish that the benchmark reference itself is a uniquely identifiable ground-truth user state.

## D Persona-Cluster Bootstrap Analysis

The development set contains 1,898 examples from 80 personas, with multiple examples potentially belonging to the same persona. To account for this within-persona dependence, we perform paired persona-cluster bootstrap resampling.

For each of 10,000 bootstrap replicates, we sample 80 personas with replacement from the original set of 80 personas. When a persona is selected, all examples associated with that persona are included in the bootstrap sample. The same sampled persona sequence is used for both conditions within each paired comparison.

We report percentile-based 95% confidence intervals. Table 6 summarizes the results.

Table 6: Persona-cluster bootstrap confidence intervals for the primary user-state and downstream decision results.

| Metric | Point Estimate | 95% CI |
|---|---|---|
| Base Full-State Accuracy | 19.81% | 18.05–21.56 |
| EGS-SFT Full-State Accuracy | 65.28% | 62.66–67.84 |
| Full-State Improvement | +45.47 pp | +42.55–+48.36 |
| D1 Decision Accuracy | 47.10% | 44.73–49.53 |
| D2 Decision Accuracy | 68.81% | 66.46–71.08 |
| Decision Improvement | +21.71 pp | +19.83–+23.64 |

Neither paired-difference confidence interval includes zero, indicating that both the user-state improvement and the downstream decision improvement remain robust when uncertainty is estimated at the persona level.

## E State-Type Breakdown

We further analyze Full-State Accuracy separately for each benchmark-defined state type. Table 7 reports the results together with paired persona-cluster bootstrap confidence intervals for the improvement.

Table 7: Full-State Accuracy by benchmark-defined state type. Confidence intervals correspond to the paired EGS-SFT minus Base improvement using persona-cluster bootstrap resampling.

| State Type | $N$ | Base | EGS-SFT | $\Delta$ (95% CI) |
|---|---|---|---|---|
| PREFERENCE | 774 | 2.97 | 45.09 | +42.12 [37.85, 46.42] |
| PERSONAL_FACT | 549 | 16.21 | 61.75 | +45.54 [40.32, 50.66] |
| MEMORY_DELETE | 378 | 46.56 | 98.68 | +52.12 [47.03, 57.14] |
| OTHER | 197 | 44.67 | 90.36 | +45.69 [37.74, 53.08] |
| **Overall** | **1,898** | **19.81** | **65.28** | **+45.47 [42.55, 48.36]** |

The improvement is positive for all four state types, and none of the corresponding 95% confidence intervals includes zero. The gains are also distributed across categories rather than being dominated by a single type. The net increases in fully correct predictions are 326 for PREFERENCE, 250 for PERSONAL_FACT, 197 for MEMORY_DELETE, and 90 for OTHER, summing to the overall net increase of 863 examples. The largest category accounts for 37.78% of the total net gain.

Despite the consistent improvements, absolute accuracy remains highly type-dependent. In particular, PREFERENCE remains the most difficult category after EGS-SFT, with 45.09% Full-State Accuracy, whereas MEMORY_DELETE reaches 98.68%. This indicates that the Evidence-to-State bottleneck is not uniform across different forms of user information.

## F EVIDENCE DIFFICULTY AUDIT

We conduct a blinded human audit of 200 stratified development examples to characterize the relation between Oracle Evidence and the benchmark-defined reference state. The audit contains 70 PREFERENCE, 55 PERSONAL_FACT, 45 MEMORY_DELETE, and 30 OTHER examples. During annotation, annotators do not observe model predictions, model identities, correctness outcomes, semantic-judge labels, or downstream decisions.

Each example is assigned one primary evidence-difficulty category. DIRECT_EXPLICIT denotes cases in which the reference state is directly stated in the supplied evidence. IMPLICIT_INFERENTIAL denotes cases requiring semantic inference beyond direct span extraction. ATTRIBUTION_OR_SCOPE captures cases in which the main challenge is identifying the relevant subject, state, or query-conditioned scope. DELETION_OR_NEGATION captures deletion requests, negation, reversals, or previously applicable information that should no longer be treated as active.

Table 8 reports Full-State Accuracy under the frozen Base and EGS-SFT predictions. Because the 200 examples are stratified by benchmark state type, category frequencies in this table should not be interpreted as estimates of full-development-set prevalence.

Table 8: Full-State Accuracy across evidence-difficulty categories in the 200-example blinded human audit. The NON_DIRECT row combines IMPLICIT_INFERENTIAL, ATTRIBUTION_OR_SCOPE, and DELETION_OR_NEGATION.

| Difficulty | $N$ | Base | EGS-SFT | $\Delta$ (pp) |
|---|---|---|---|---|
| DIRECT_EXPLICIT | 36 | 8.33 | 75.00 | +66.67 |
| IMPLICIT_INFERENTIAL | 26 | 0.00 | 50.00 | +50.00 |
| ATTRIBUTION_OR_SCOPE | 89 | 20.22 | 64.04 | +43.82 |
| DELETION_OR_NEGATION | 49 | 28.57 | 89.80 | +61.22 |
| **NON_DIRECT** | **164** | **19.51** | **69.51** | **+50.00** |

For the combined non-direct subset, the paired persona-cluster bootstrap 95% confidence interval for the improvement is **42.58–57.42 percentage points**. Thus, within the audited sample, the improvement is not confined to direct-explicit extraction cases.

We additionally identify 23 examples that require integration of multiple evidence spans. Base obtains 0.00% Full-State Accuracy on this small subset, while EGS-SFT reaches 47.83%. Because this subgroup is small, we treat the result as descriptive rather than as a standalone statistical claim.