Beyond Memory: Constructing Hierarchical User Identity from Dialogue

Anonymous ACL submission

Abstract

Existing dialogue memory systems mainly optimize storage and retrieval of past utterances, summaries, or extracted facts. For long-term personalization, however, memory should be treated as evidence rather than the final user representation. We formulate *hierarchical user identity construction from dialogue*, which maps multi-session conversations to a persistent, revisable user state with three layers: factual identity, preference structure, and slow-moving interaction state. We instantiate an inference-first framework that first proposes dialogue-grounded evidence and then applies layer-specific promotion, abstention, and update rules to construct and maintain this state. Across manually reviewed PersonaChat evaluations, this formulation improves explicit fact and preference construction over same-model direct extraction on a larger reviewed overlay, with the same mechanism also visible in a smaller controlled audited subset; a lightweight downstream response-selection pilot further shows that the resulting state improves candidate choice even under a fixed scorer. Reviewed MSC analyses further indicate that the resulting state supports conflict-aware chronological maintenance, while a reviewed adjacent-session analysis suggests that slow-moving interaction state can also be modeled usefully. Taken together, these findings suggest a path beyond memory-centric dialogue systems toward conversational agents that construct and maintain explicit, revisable user identity over time in long-horizon settings.

1 Introduction

Large language model-based conversational agents increasingly rely on dialogue memory to maintain continuity across turns and sessions. Yet memory is not the same as user modeling: a system can retrieve what the user said before and still fail to maintain a coherent view of who the user is. Prior work on long-term dialogue memory and memory-augmented generation largely optimizes how to store, summarize, and retrieve historical evidence at inference time (Sukhbaatar et al., 2015; Miller et al., 2016; Lewis et al., 2020; Borgeaud et al., 2022; Park et al., 2023; Packer et al., 2023; Xu et al., 2022; Zhong et al., 2024; Tan et al., 2025). This line is important, but it often leaves the personalization target underspecified.

In parallel, personalized dialogue research has sought to generate responses aligned with user-specific traits, either by conditioning on explicit personas or by inferring profile information from interaction history (Li et al., 2016; Mazaré et al., 2018; Wolf et al., 2019; Zhang et al., 2018; Qian et al., 2018; Pei et al., 2021; Wu et al., 2021; Zhou et al., 2023; Cheng et al., 2024). These methods improve response personalization, but their main optimization target remains output generation. They do not usually treat persistent user state as a first-class object with explicit update semantics, provenance, and conflict handling.

We argue that this leaves a missing object between memory and response generation. For long-term interaction, memory should be treated as evidence, not as the final representation of the user. A personalized agent needs an explicit user state: stable enough to support continuity, revisable enough to absorb corrections, and transparent enough to show what was updated or left uncommitted. That level of control is difficult to obtain when personalization is mediated only through retrieved context windows or latent conditioning.

This distinction matters because user information is heterogeneous in both form and temporal behavior. Some properties are symbolic and relatively stable, such as name, occupation, location, or recurring role relationships. Others are softer and only partially structured, such as food preferences, communication habits, planning style, or regular activity patterns. Still others are difficult to state explicitly at all, such as proactivity tolerance or

tone under stress. Flattening all such information into a single memory bank obscures differences in persistence, uncertainty, and update dynamics. This hierarchy is also broadly consistent with classic distinctions in cognitive and social theories of self-representation (Tulving, 1985; Conway and Pleydell-Pearce, 2000; Markus, 1977; McAdams and Pals, 2006; Goffman, 1959).

Motivated by this observation, we propose *hierarchical user identity construction from dialogue*. The key idea is to represent user identity as a multi-level state built on top of dialogue-derived memory evidence. At the lowest level, the model maintains symbolic identity facts that can often be expressed as slots, triples, or canonicalized statements. At the middle level, it maintains a typed semi-structured preference state describing durable likes, habits, and behavioral regularities. At the highest level, it maintains a slow-moving interaction state that captures longer-horizon tendencies that are difficult to reduce to explicit text. Different levels should be updated differently: stable facts should change conservatively, preference state should absorb repeated evidence while remaining revisable, and interaction state should capture smooth cross-session continuity without being overfit to individual turns.

This framing also clarifies the boundary between memory and identity. Memory is time-grounded evidence, while identity is the maintained state inferred from that evidence and consumed by downstream personalization. That distinction matters when evidence changes over time: a memory-only system may retrieve the relevant utterances yet still lack a clear policy for the current active user state after revisions or contradictions. We therefore evaluate maintained user state directly, asking what was promoted, revised, or left uncertain, and how downstream behavior changes when only the supplied state changes. This is deliberately narrower than a full end-to-end dialogue claim, but it makes the empirical story cleaner and more reviewer-checkable.

From a systems perspective, this intermediate state also creates a smaller interface between raw conversational evidence and downstream personalization modules. Instead of repeatedly exposing the entire memory store to every downstream decision, the agent can condition on an explicit maintained state whose commitments, revisions, and uncertainties are easier to inspect and debug. It also localizes failure modes: retrieval misses, extraction errors, and revision mistakes become separately diagnosable instead of collapsing into one opaque personalization outcome. This is especially useful in long-horizon assistants, where small state errors can accumulate across sessions. It also makes the personalization stack more modular: extractors can over-generate evidence, deterministic rules can decide commitment, and downstream components can consume a compact maintained state. That interface perspective is one practical reason to treat identity construction as its own problem rather than as a side effect of retrieval.

We make this framing concrete through explicit-state construction and a reviewed temporal-Z analysis. We realize explicit-state construction as a two-stage process: a prompted model first proposes raw evidence items with provenance and stability cues, and a deterministic layer then decides what should be promoted into canonical fact triplets or typed preference entries versus what should be left uncommitted. For evaluation, we use a larger manually reviewed PersonaChat overlay as the main benchmark, paired with a controlled 100-example audited same-model subset where the effect of promotion, canonicalization, deduplication, and abstention is easiest to inspect. This setup separates state-construction quality from downstream utility while keeping the empirical scope auditable. Our reviewed MSC experiments further show that the resulting state can be maintained under conflict and used for bounded next-session temporal Z prediction.

Contributions. Our main contributions are:

- We formulate hierarchical user identity construction from dialogue as a distinct problem setting beyond flat dialogue memory.
- We propose a three-level representation with canonical factual triplets, typed semi-structured preference state, and slow-moving interaction state.
- We describe an inference-first maintenance interface that separates time-grounded evidence from evolving identity state and specifies promotion, abstention, revision, overwrite, and conflict handling.
- We define an evidence-grounded evaluation protocol spanning reviewed PersonaChat and MSC overlays, and show gains in explicit F/P construction, reviewed maintenance, and bounded temporal Z prediction.

2 Related Work

User modeling and explicit-profile dialogue personalization. Dialogue systems have long incorporated user modeling to support adaptive interaction and generation (Kobsa, 2001; Janarthanam and Lemon, 2014). In modern neural dialogue, a major line of work conditions generation on explicit personas or profile descriptions. Representative examples include persona-based neural dialogue (Li et al., 2016), PersonaChat-style profile conditioning (Zhang et al., 2018), large-scale personalized dialogue pretraining (Mazaré et al., 2018; Wolf et al., 2019), and profile-coherent response generation from predefined attribute sets (Qian et al., 2018). Later systems extend this idea to task-oriented and open-domain personalization by incorporating user profiles or profile memories into response generation (Pei et al., 2021; Wu et al., 2021). These methods improve personalized responses, but they typically assume that the target representation already exists as a given profile, or they use the profile only as a control signal for generation. By contrast, our focus is not persona-conditioned generation per se, but how such a profile should be constructed and maintained from dialogue evidence over time.

Dialogue-derived profile and persona induction. More recent work reduces reliance on predefined profiles by inferring user-specific information directly from dialogue history. Examples include building implicit user profiles from large-scale conversational behavior (Ma et al., 2021), predicting persona information for personalization without explicit persona text at inference time (Zhou et al., 2023), and fine-tuning large language models to learn personalization from dialogue sessions without predefined profiles (Cheng et al., 2024). This line comes closest to our setting, because dialogue itself becomes the source of personalization signals. However, the main optimization target remains personalized response generation. Dialogue history is primarily used as conditioning evidence for better outputs, rather than as input to a persistent user-state constructor with explicit promotion, abstention, revision, and conflict-handling semantics.

Long-term dialogue memory and memory management. A separate line studies long-term memory for conversational agents, focusing on how to store, summarize, retrieve, and manage historical dialogue evidence across sessions (Sukhbaatar et al., 2015; Miller et al., 2016; Lewis et al., 2020; Borgeaud et al., 2022; Park et al., 2023; Packer et al., 2023; Xu et al., 2022; Zhong et al., 2024; Tan et al., 2025). These works show that memory organization and retrieval quality are central to coherent long-horizon interaction. Yet they generally treat memory itself as the main artifact to be optimized, whether in the form of retrieved turns, summaries, or managed memory banks. Our framing differs in the target object. We treat message, session, and block memories as evidence layers, while user identity is the persistent state constructed on top of those layers for downstream personalization. In this sense, our work shifts the center of gravity from memory access to user-state construction.

Positioning of this paper. The present work sits at the intersection of these strands, but it centers a different object: maintained user state. We ask how to construct and update a hierarchical user state from dialogue evidence, and how to evaluate it through explicit F/P quality, chronological revision behavior, and reviewed temporal Z prediction. The contribution is therefore a representational formulation, a maintenance interface, and an auditable evaluation protocol rather than a new end-to-end dialogue generator. Accordingly, the empirical scope is centered on state construction and maintenance quality rather than a full demonstration of downstream dialogue gains. This design is also loosely informed by classical distinctions between semanticized personal knowledge, graded preferences or self-schemata, and interaction-dependent self-presentation (Tulving, 1985; Conway and Pleydell-Pearce, 2000; Markus, 1977; McAdams and Pals, 2006; Goffman, 1959).

3 Method

3.1 Problem Formulation

Consider a sequence of user–agent interactions across sessions, $\mathcal{D}_{1:T} = \{D_1, D_2, \ldots, D_T\}$, where each session D_t contains a temporally ordered set of dialogue turns. Standard dialogue memory methods construct a memory store M_t from the interaction history up to time t, where M_t may contain raw utterances, summaries, extracted facts, or retrieval keys. In our formulation, memory is not the final object of interest. Instead, the goal is to construct an evolving user identity state from dialogue evidence:

$$I_t = \mathcal{U}(I_{t-1}, M_t, D_t). \quad (1)$$

We model the identity state as

$$I_t = (F_t, P_t, Z_t), \tag{2}$$

where the three layers are defined below. The distinction between memory and identity is central. Memory stores what was observed, while identity represents what the system currently believes about the user. This separation gives a clearer supervision target than retrieval alone, supports principled revision when evidence changes, and yields a compact personalization state that can be consumed directly by downstream agents.

This distinction also matters for evaluation. In partially observed dialogues or synthetic settings, the full hidden user state may exceed what the dialogue actually reveals. We therefore distinguish between a latent state that may underlie generation and an *observable identity state* consisting only of facts, preference attributes, and interaction tendencies that are explicitly supported, or strongly warranted, by the available dialogue evidence. Our extraction and maintenance evaluations target this observable state rather than any fully hidden generator-side state.

3.2 Hierarchical Identity Representation

The factual layer F_t contains explicit and canonicalizable user identity claims such as name, role, institution, residence, or other durable relational facts, represented as editable subject–relation–object triplets. The preference layer P_t captures typed regularities such as food, activity, planning, communication, or work preferences through slot-conditioned values with graded support. The interaction layer Z_t captures residual behavior-grounded state that is useful for personalization but not naturally represented as explicit facts or preference slots, such as task-vs-social orientation, proactivity tolerance, or verbosity tolerance.

The key challenge is that these layers change at different rates. Facts should usually persist unless directly contradicted, preference state should accumulate repeated support while remaining revisable, and interaction state should evolve smoothly rather than overreacting to isolated turns. Figure 1 summarizes how dialogue memory supplies evidence to this three-layer state.

The figure also makes the interface explicit: message, session, and block memories are observable evidence stores, whereas F/P/Z is the maintained user state constructed above them. Retrieval answers what has been observed; identity maintenance answers what the system currently believes about the user.

3.3 Evidence Proposal and Normalization

Given a new session D_t, the system first extracts time-grounded evidence candidates rather than directly rewriting the user identity. Let $R_t = (R_t^{\mathrm{msg}}, R_t^{\mathrm{sess}}, R_t^{\mathrm{blk}})$ denote retrieved message-, session-, and block-level evidence. We then produce layer-specific candidate sets

$$\begin{aligned} C_t^F &= \phi_F(D_t, R_t), \\ C_t^P &= \phi_P(D_t, R_t). \end{aligned} \tag{3}$$

where ϕ_F and ϕ_P may be implemented by learned span extractors, prompted language models, or hybrid extract-then-rerank modules. In our implementation, these modules over-generate raw evidence together with provenance spans and lightweight stability cues, after which a deterministic stage performs canonicalization, promotion, and abstention. The resulting evidence sets are

$$E_t^F = \left\{ \mathrm{Norm}_F(c) \;\middle|\; \begin{array}{l} c \in C_t^F, \\ s_F(c) \geq \tau_F \end{array} \right\}. \tag{4}$$

$$E_t^P = \left\{ (a, v, w) \;\middle|\; \begin{array}{l} c \in C_t^P, \\ a = \mathrm{Slot}(c), \\ v = \mathrm{Norm}_P(c), \\ w = s_P(c) \end{array} \right\}. \tag{5}$$

where Norm_F performs relation-constrained canonicalization for factual triplets, while $\mathrm{Slot}(\cdot)$ maps a candidate into a curated preference-slot inventory and Norm_P normalizes the value while keeping the value vocabulary open-ended.

Interaction-state evidence is handled differently. Rather than committing symbolic assertions, we compute a session-level interaction proposal

$$\hat{z}_t = \mathrm{Enc}_Z\big(D_t, R_t, g(F_t, P_t)\big), \tag{6}$$

where Enc_Z is a lightweight dialogue encoder or classifier and $g(F_t, P_t)$ provides structured lower-layer context. In the general framework, Z_t may summarize any slow-moving, behavior-grounded interaction attributes that are useful for personalization but difficult to reduce to canonical facts or preference slots. The resulting $\hat{z}_t$ is treated as a soft proposal for the interaction layer rather than as a directly stored symbolic fact.

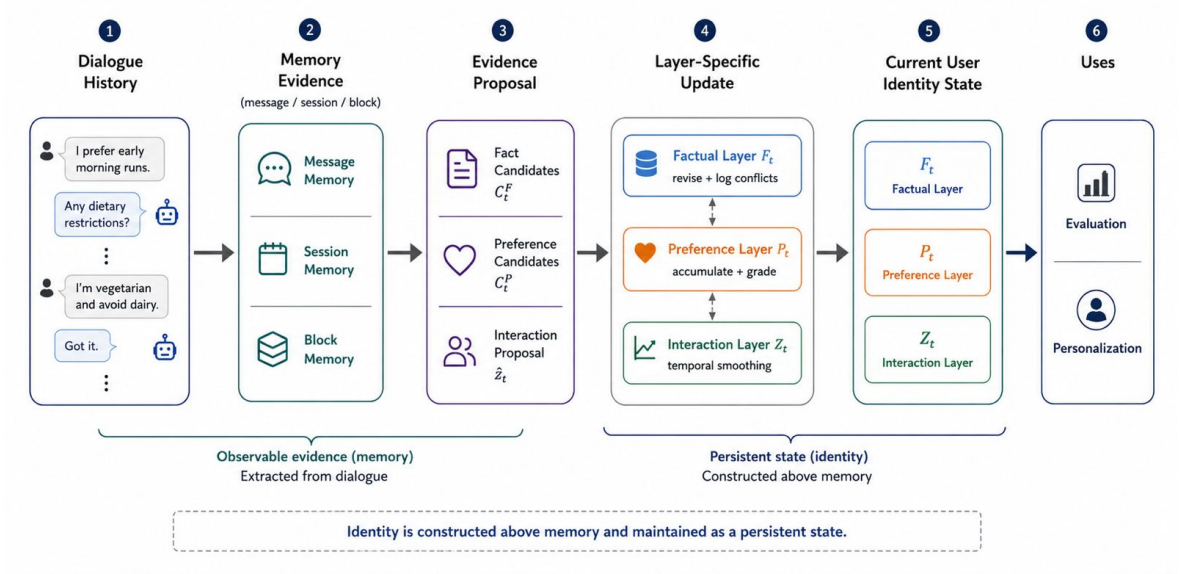

Figure 1: Overview of the proposed pipeline. Dialogue memory provides observable evidence, while persistent user identity is constructed above it through layer-specific updates over factual, preference, and interaction state.

3.4 State Consolidation and Revision

The three identity layers are then updated separately:

$$F_t = \mathrm{Update}_F(F_{t-1}, E_t^F), \tag{7}$$

$$P_t = \mathrm{Update}_P(P_{t-1}, E_t^P, F_t), \tag{8}$$

$$Z_t = \mathrm{Update}_Z(Z_{t-1}, \hat{z}_t, P_t, F_t). \tag{9}$$

For the factual layer, we maintain an active value for each normalized single-value key $k = (s, r)$ together with provenance and conflict history, while explicitly allowing a small set of multi-value relations to accumulate without overwrite. For a candidate object o under key k, we compute a revision score

$$\begin{aligned} q_t(k, o) = \lambda_1\, p_{\mathrm{ext}}(o) + \lambda_2\, p_{\mathrm{sup}}(k, o) \\ + \lambda_3\, p_{\mathrm{rec}}(o) + \lambda_4\, p_{\mathrm{corr}}(o). \end{aligned} \tag{10}$$

where the terms respectively summarize extractor confidence, accumulated support from prior evidence, recency, and explicit correction cues. Let $o_{t-1}^*(k)$ denote the incumbent active value for key k. We overwrite conservatively only when

$$q_t(k, o) > q_t\big(k, o_{t-1}^*(k)\big) + \delta_F, \tag{11}$$

or when the challenger is a more specific refinement of the incumbent under the same relation, for example *writer* $\rightarrow$ *freelance writer*. By contrast, non-refining competing values are logged as conflicts unless the newer evidence is strong enough to justify revision. This makes the factual layer closer to a personal belief-maintained fact graph than to an append-only triple store.

For the preference layer, we maintain slot-conditioned value scores. For each slot a and normalized value v, we update

$$\begin{aligned} S_t(a, v) = \rho_P\, S_{t-1}(a, v) \\ + \sum_{(a_i, v_i, w_i) \in E_t^P} w_i\, \mathbf{1}[a_i = a, v_i = v]. \end{aligned} \tag{12}$$

where $\rho_P \in [0, 1]$ controls persistence and the evidence weights w_i may be positive, negative, or confidence-scaled. Multiple values can therefore coexist within the same slot.

For the interaction layer, we combine the learned proposal with temporal smoothing:

$$\begin{aligned} \alpha_t &= \sigma\big(w_\alpha^\top u_t + b_\alpha\big), \\ Z_t &= (1 - \alpha_t) Z_{t-1} + \alpha_t \hat{z}_t. \end{aligned} \tag{13}$$

where u_t summarizes evidence strength or session confidence and σ is the logistic function. This update allows the model to react more strongly when the session provides decisive interactional evidence, while preserving continuity when the signal is weak. Overall, the three update rules distinguish conservative factual revision, graded preference accumulation, and smooth interaction-state maintenance within a single framework.

4 Experiments

We evaluate four questions: whether the framework improves explicit F/P construction, whether

cleaner explicit state helps downstream response selection under a fixed scorer, whether the framework improves chronological revision and conflict handling, and whether the interaction layer supports bounded next-session temporal prediction. Together, these result groups test construction quality, lightweight downstream utility, maintenance reliability, and temporal interaction signal within one matched evaluation protocol. To keep comparisons fair with prior dialogue-personalization and memory work, we build lightweight reviewed overlays on existing public datasets rather than introducing a separate bespoke benchmark.

4.1 Datasets and Task Construction

PersonaChat explicit-state construction. This setting provides the main benchmark for Table 1. Given a dialogue prefix, the model predicts the current symbolic fact set F_t and the explicit portion of the preference state P_t. We construct a manually reviewed observable overlay by mapping persona sentences into canonical factual triplets and typed preference slots, yielding relatively clean supervision for fact extraction, canonicalization, and preference induction (Zhang et al., 2018). For the main paper-facing comparison, we report a larger reviewed 500-example overlay on the original public PersonaChat data together with a same-model direct constrained baseline; we retain a controlled 100-example audited subset only as an auditability and mechanism-check setting, not as headline evidence.

Reviewed MSC overlays. This setting provides the benchmarks for Tables 3 and 4. We feed sessions chronologically and update identity after each session. Because MSC does not provide hierarchical identity labels, we construct reviewed overlays that record evidence-grounded factual triplets, typed preference state, interaction labels, and revision or conflict ledgers derived from explicit self-statements, repeated preference cues, and recurring interactional evidence (Xu et al., 2022). In principle, the Z layer can encode a broader set of slow-moving interaction attributes. In this paper, however, we operationalize Z conservatively through two reviewed interaction dimensions, verbosity and orientation, so that the temporal evaluation remains auditable. We report four explicitly named subsets: a first/last-session holdout (12 reviewed cases), a clean adjacent-session holdout (17 cases), a harder adjacent candidate set (18 cases), and a new-user adjacent holdout (8 cases).

4.2 Compared Systems

To make the empirical claim defensible, we compare against task-appropriate baselines that separate memory access from identity construction while keeping the underlying scorer or generator fixed. Not every baseline is relevant to every subtask, so each table reports the strongest task-matched comparisons for that setting.

Memory-only personalization. A retrieval-augmented system with message, session, and block retrieval or summaries but no explicit identity-state layer for personalization.

Direct constrained extraction. A strong same-model baseline for explicit F/P construction. The prompted model is asked to emit the final constrained fact and preference schema directly from the observed dialogue, but no raw-evidence stage, promotion logic, abstention layer, or relation-aware postprocessing is used. This baseline therefore entangles evidence proposal, promotion, canonicalization, and final state emission in a single generation step.

Neural proposal + flat state. A stronger baseline that uses the same learned candidate proposal modules as our method, but collapses all accepted evidence into a single flat maintained state without layer-specific revision rules.

Uniform overwrite. A maintenance baseline that uses the same normalized proposal stream as our method, but overwrites incumbents whenever a newer accepted value arrives, without explicit conflict tracking or conservative revision thresholds.

Full hierarchical identity model. Our proposed system uses message, session, and block evidence together with the full F/P/Z identity state. In our F/P implementation, the practical gain comes from two-stage explicit-state construction: raw evidence proposals are first extracted with provenance spans and lightweight stability cues, then promoted or rejected by a deterministic layer that performs canonicalization, deduplication, and abstention before committing items to the explicit fact or preference state.

4.3 Metrics

We report fact F1 and preference-slot F1 for explicit state construction; recall@k and MRR for

System	Fact-soft	Fact-strict	Pref.
Direct constrained	0.689	0.325	0.311
Two-stage F/P (ours)	**0.767**	**0.384**	**0.445**

Table 1: Main explicit state-construction results on the larger 500-example reviewed PersonaChat overlay. Absolute gains over direct constrained extraction are +0.078 fact-soft F1, +0.059 fact-strict F1, and +0.134 preference F1.

the lightweight downstream response-selection pilot; stability, revision, overwrite, and conflict metrics for chronological maintenance; and strict next-session prediction together with per-dimension accuracy for temporal Z evaluation. Unless otherwise stated, systems share the same base generator or scorer and the same retrieval backbone, so gains can be attributed to representation and update design rather than to a stronger underlying model.

4.4 Main Results

We organize the empirical section around four complementary result groups: explicit state construction, a lightweight downstream pilot, chronological maintenance, and temporal-Z robustness. Tables 1, 2, 3, and 4 summarize the paper-facing results.

4.4.1 Explicit State Construction

The first result group evaluates whether hierarchical identity construction improves explicit user-state quality over same-model direct extraction.

Table 1 reports the main explicit-state result on the larger reviewed 500-example PersonaChat overlay. Even in this harder and more paper-facing setting, the two-stage constructor still dominates direct constrained extraction on all three reported metrics, reaching 0.767 vs. 0.689 fact-soft F1, 0.384 vs. 0.325 fact-strict F1, and 0.445 vs. 0.311 preference F1. The point is not only that the method wins, but that the win survives on the broader reviewed benchmark rather than collapsing outside the smaller audited subset. This larger reviewed result therefore carries the main robustness-facing empirical claim.

We additionally retain a controlled 100-example audited same-model subset as a mechanism check. In that cleaner audit setting, the same ordering becomes even sharper, with 1.000 vs. 0.620 fact-soft F1, 0.650 vs. 0.333 fact-strict F1, and 0.680 vs. 0.264 preference F1, making it easier to inspect what the promotion and canonicalization layer is actually doing. Because this audited subset is intentionally smaller and cleaner, we treat it as mechanism-inspection evidence rather than as the headline generalization result.

System	R@1	R@5	MRR
Recent ctx. only	0.144	0.402	0.275
Memory only	0.144	0.406	0.278
Two-stage F/P (ours)	**0.182**	**0.412**	**0.303**

Table 2: Lightweight downstream response-selection pilot on the same 500-example reviewed PersonaChat overlay. All systems use the same deterministic candidate scorer; only the supplied identity state changes.

System	Stab. Err.	Rev. P	Rev. L	Conf. F1	Ovr. Err.
Mem. only	0.370	0.906	3.656	0.000	0.094
Neural flat	**0.364**	0.906	3.656	0.000	0.094
Unif. ow.	**0.364**	0.906	3.656	0.000	0.094
Full F/P/Z	**0.364**	**0.969**	**3.594**	**0.771**	**0.031**

Table 3: Chronological maintenance results on the reviewed16 MSC overlay. Lower is better for stability error, revision latency, and overwrite error; higher is better for revision precision and conflict F1. The hierarchical model is best on four of the five reported metrics and tied for best on stability error.

The mechanism interpretation is straightforward: the gain comes from separating evidence proposal from promotion, canonicalization, deduplication, and abstention, rather than asking the model to emit the final constrained user state in one shot.

4.4.2 Lightweight Downstream Response-Selection Pilot

To test whether better explicit state already carries downstream value, we add a controlled candidate-response ranking pilot on the same 500-example reviewed PersonaChat overlay. The comparison is intentionally conservative: all systems use the same deterministic scorer, and only the supplied identity state changes. Under that control, Table 2 shows that the hierarchical state is best on every reported metric, improving R@1 from 0.144 to 0.182 and MRR from 0.278 to 0.303 over the strongest non-hierarchical baseline. We treat this as bounded downstream evidence rather than a substitute for free-form generation, but it supports the practical claim that cleaner maintained identity state can already improve response choice.

4.4.3 Chronological Maintenance, Revision, and Conflict Handling

The second result group evaluates whether layer-specific updates improve temporal behavior under changing evidence.

Table 3 now gives a fuller maintenance picture without sacrificing the main claim. The hierarchical model reaches the strongest conflict F1 at 0.771 while also improving revision precision to 0.969, reducing revision latency to 3.594, and lowering overwrite error to 0.031. Crucially, these gains do not come from unstable behavior: the method matches

Reviewed set	NSP	Verb. Acc	Ori. Acc
First/last holdout (12)	0.917	**1.000**	0.917
Clean adjacent (17)	**1.000**	**1.000**	**1.000**
Adjacent candidate (18)	0.944	**1.000**	0.944
New-user adjacent (8)	0.875	**1.000**	0.875

Table 4: Temporal Z prediction on reviewed MSC subsets for two audited interaction dimensions, verbosity and orientation. Numbers in parentheses indicate reviewed subset size. All reported numbers use the deterministic temporal predictor.

the best stability error among the non-degenerate maintenance baselines. This makes the maintenance result stronger than a conflict-only story, because the same model now leads on the core revision, contradiction, and conservative-overwrite axes simultaneously. In other words, the maintenance gains are coherent rather than single-metric improvements.

4.4.4 Temporal Z Prediction and Robustness

The third result group evaluates the latent layer under reviewed adjacent-session prediction.

Table 4 provides reviewed evidence that the interaction layer captures stable temporal signal rather than only within-session noise. On the clean adjacent holdout (17 reviewed cases), the deterministic temporal predictor reaches 1.000 strict next-session prediction, and it reaches 0.944 on the harder adjacent candidate set (18 cases). The more important robustness result is the new-user adjacent holdout (8 cases), where performance reaches 0.875 NSP with 1.000 verbosity accuracy, suggesting that the predictor is not merely memorizing user-specific history. Across these reviewed holdouts, verbosity is the more stable of the two audited dimensions, while the remaining errors are concentrated in orientation transitions. Because the current paper operationalizes Z through only two audited interaction dimensions, we therefore present this as a conservative but consistent instantiation of the broader interaction-state idea rather than as an unconstrained latent-style claim.

5 Discussion

On PersonaChat, the hierarchical formulation improves explicit F/P construction on the larger reviewed overlay and yields a controlled downstream response-selection gain under a fixed scorer. On reviewed MSC chronology, it improves revision, conflict handling, conservative overwrite, and bounded next-session prediction across several holdouts. The claim is representational rather than end-to-end: a cleaner maintained state is both more auditable and more useful downstream, while fuller free-form generation evaluation remains open.

Because all systems use the same candidate scorer, and only the supplied identity state changes, better downstream ranking suggests that improved state construction is already doing useful work rather than merely inheriting gains from a stronger decoder or prompting recipe.

6 Conclusion

We formulate hierarchical user identity construction from dialogue as a representational maintenance problem rather than a memory-only one. Across reviewed PersonaChat and MSC overlays, this framing improves explicit state construction, controlled downstream response selection, maintenance, and bounded temporal prediction.

Limitations

This paper focuses on explicit identity-state construction and maintenance rather than full end-to-end dialogue optimization. Accordingly, the strongest evidence concerns factual and preference-state quality, revision behavior, conflict handling, and controlled downstream response selection under fixed scoring. While the downstream pilot suggests that cleaner maintained identity state can already improve response choice, the present study does not yet establish gains for unrestricted free-form generation or long-horizon human interaction.

The reviewed overlays are intentionally high-precision and audit-oriented rather than large-scale weakly supervised benchmarks. This design improves interpretability and makes revision and conflict behavior directly inspectable, but it also limits its evaluation breadth: the reviewed MSC subsets remain relatively small, and the interaction-state analysis operationalizes Z through only two audited dimensions, verbosity and orientation. We therefore position the temporal results as bounded evidence for interaction-state usefulness rather than a complete treatment of latent user modeling.

Methodologically, the framework emphasizes explicit maintenance semantics, including promotion, abstention, revision, and conservative overwrite, over highly end-to-end neural optimization. This improves auditability and controllability, but leaves learned large-scale maintenance policies, broader cross-domain and multilingual schemas, and tighter integration with downstream generation as future work.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George B. M. van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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, Miami, Florida, USA. Association for Computational Linguistics.

Martin A. Conway and Christopher W. Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY.

Srinivasan Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*, 40(4):883–920.

Alfred Kobsa. 2001. Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1–2):49–63.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003. Association for Computational Linguistics.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564. ACM.

Hazel Markus. 1977. Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, 35(2):63–78.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779. Association for Computational Linguistics.

Dan P. McAdams and Jennifer L. Pals. 2006. A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist*, 61(3):204–217.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409. Association for Computational Linguistics.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM.

Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2021. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In *Proceedings of the Web Conference 2021*, pages 1552–1561. ACM.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. 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, Vienna, Austria. Association for Computational Linguistics.

Endel Tulving. 1985. Memory and consciousness. *Canadian Psychology*, 26(1):1–12.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clément Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. 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. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2023. Learning to predict persona information for dialogue personalization without explicit persona description. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2979–2991, Toronto, Canada. Association for Computational Linguistics.