PersonaTree: Structured Lifecycle Memory for Person Understanding in LLM Agents

Anonymous ACL submission

Abstract

Persistent LLM agents require memory representations that make the formation of person understanding explicit across long term interaction. Existing agent memory methods emphasize information retention and retrieval, yet give limited account of how accumulated interaction evidence is abstracted into person understanding. We view this process as schema formation, where situated evidence is abstracted into reusable patterns and stable person level claims. We introduce PERSONATREE, a structured lifecycle memory framework that realizes this view as a three level persona tree with explicit support paths from evidence to claims. PERSONATREE maintains the tree through conservative writing, confidence guided consolidation, and query conditioned path retrieval, returning only the evidence depth required by each query. Across six person understanding and persistent memory benchmarks with three answer backbones, PERSONATREE ranks first in 12 of 18 compact scores and reaches the top two in 16 settings. Ablations show that hierarchy improves abstract person understanding on KnowMe, while support path retrieval improves RealPref alignment under a comparable context budget.

1 Introduction

LLM agents are increasingly expected to interact with the same user over extended periods of time (Park et al., 2023; Yao et al., 2022; Wang et al., 2023; Hong et al., 2024; Wu et al., 2024b; Xi et al., 2025). In this setting, memory should support more than the recall of isolated facts. A useful agent needs to track what the user has said and done, recognize recurring preferences and states, and form stable judgments about the user's habits, values, constraints, and boundaries. Recent benchmarks for lifelong companions and personalized assistants reflect this shift from factual recall toward person understanding (Wu et al., 2026; Guo et al., 2026; Bian et al., 2026; Kim et al., 2025).

A central difficulty in person understanding is that abstract judgments about a user need to be grounded in scattered, situated observations. This view echoes schema theory in cognitive psychology, where schemata organize prior knowledge and relations among concepts, and guide how new observations are interpreted and recalled (Bartlett, 1932; RUMELHART, 1980). In an agent setting, this perspective motivates a memory representation that keeps the basis for user-level interpretations visible. Individual interactions may offer partial evidence, while repeated observations can gradually support more general expectations about the user's preferences, states, constraints, and boundaries. Without such grounding, an answer model may retrieve plausible user information but have limited basis for explaining why an abstract interpretation follows from the interaction history.

Existing agent memory research has made long term interaction more practical by treating memory as a problem of representation, update, organization, and retrieval. Survey work characterizes this design space across memory form, function, and lifecycle (Hu et al., 2025b), and recent systems instantiate it through persistent memory layers, agentic update operations, temporal organization, lightweight maintenance, and learned memory policies (Chhikara et al., 2025; Xu et al., 2026; Rasmussen et al., 2025; Fang et al., 2025; Li et al., 2026a, 2025; Yue et al., 2026). These advances improve how user information is retained and made available during interaction. Person understanding, however, asks memory to play a stronger modeling role. The memory state should help the agent build a coherent view of the user, including preferences, states, constraints, boundaries, and motivations, rather than only supply facts, summaries, or retrieved context to the answer model. This shift calls for a representation that makes accumulated interaction experience usable for reasoning about the user as a person.

To provide such a representation, we propose PERSONATREE, a structured lifecycle memory framework for person understanding in LLM agents. PERSONATREE represents the user's evolving profile as a three level persona tree that connects interaction experience with higher level user interpretations. Leaf nodes store timestamped interaction evidence, mid level nodes capture recurring behavioral or state patterns, and root nodes represent stable persona level claims. Edges record how lower level evidence supports higher level abstractions, making the basis of each user level claim visible in the memory state. This representation turns accumulated interaction history into a structured user model that can support both concrete recall and more abstract reasoning about the person.

PERSONATREE maintains this structured user model through lifecycle operations. Online insertion converts new interactions into typed evidence and links them to existing patterns only when schema compatibility and evidence validation support the update. Offline consolidation revisits unresolved evidence, merges compatible patterns, promotes mature abstractions into stable claims, and removes stale or weak structures. During inference, query conditioned retrieval selects the abstraction level needed for the current question and renders a compact support path when lower level evidence is useful. The answer model can therefore use the relevant claim, pattern, or event evidence without reading the full interaction history.

We evaluate PERSONATREE on six person understanding and persistent memory benchmarks with three answer backbones. Across the full suite, PERSONATREE ranks first in 12 of 18 compact scores and reaches the top two in 16 settings. The analysis connects these gains to the main design choices. On KnowMe, hierarchy improves questions that require reasoning from accumulated evidence to user level claims. On RealPref, retrieving a support path outperforms retrieving the same memory nodes as a flat set under a comparable context budget. Efficiency analysis further shows that PERSONATREE keeps answer inputs compact under long user histories while improving task score.

Our contributions are as follows. First, we formulate persistent person understanding as a lifecycle memory problem in which long term interaction experience should support a coherent and grounded model of the user. Second, we introduce PERSONATREE, a structured memory framework that represents user experience as a three level persona tree and maintains it through online evidence insertion, offline consolidation, confidence update, and query conditioned retrieval. Third, we provide main results, ablations, and efficiency analysis showing how hierarchical representation, support path retrieval, and context control improve persistent person understanding over extended interactions.

2 Related Work

User modeling and personalization. Person understanding and personalization depend on long term interaction history that goes beyond isolated profile facts. Recent benchmarks make this requirement concrete through factual recall, changing states, preferences, and context dependent alignment (Wu et al., 2026, 2024a; Guo et al., 2026; Kim et al., 2025; Bian et al., 2026; Li et al., 2026b; Salemi et al., 2024; Jiang et al., 2024). MemoryBank and AI PERSONA maintain user memories or evolving profiles for personalization (Zhong et al., 2024; Wang et al., 2024). PERSONATREE focuses on how such persistent user information can be organized into a coherent and grounded user model.

Long term agent memory. Long term agent memory work studies how agents store, update, filter, and retrieve memories (Shinn et al., 2023; Sumers et al., 2023). MemGPT is an OS-inspired virtual context manager (Packer et al., 2023), while Mem0, A-MEM, LightMem, and Mem-T focus on scalable persistent memory, Zettelkasten-style linking, lightweight filtering with offline consolidation, and learned memory policies (Chhikara et al., 2025; Xu et al., 2026; Fang et al., 2025; Li et al., 2025; Yue et al., 2026). These systems are strong operational baselines, but their interfaces center on records, notes, summaries, graphs, or retrieved contexts. PERSONATREE organizes memory as a structured user model that preserves the basis for higher level interpretation.

Structured and hierarchical memory. Structured memory methods go beyond flat records by organizing long histories with relations or hierarchy (Lewis et al., 2020; Liu et al., 2024; Sarthi et al., 2024; Edge et al., 2024; Gutiérrez et al., 2024). Zep/Graphiti uses temporal knowledge graphs (Rasmussen et al., 2025), TiMem builds a temporal memory tree (Li et al., 2026a), and MemTree builds dynamic tree schemas (Rezazadeh

et al., 2025). These structures support entity, temporal, topical, or schema organization. PERSONATREE uses hierarchy for evidential abstraction in person understanding, linking interaction events, recurring patterns, and persona claims so abstract user interpretations remain grounded.

3 Method

3.1 Overview

PERSONATREE represents long term user memory as a three level persona tree. Leaves store timestamped interaction evidence, mid nodes capture recurring behavior or state patterns, and roots store stable persona claims. Typed support edges keep higher level claims traceable to the evidence below them, allowing the same memory state to support factual recall and abstract person understanding.

PERSONATREE updates this tree through a lifecycle of writing, consolidation, and retrieval. New interactions first enter as typed leaves and attach to existing abstractions only when schema compatibility and evidence validation support the link. Later consolidation merges compatible evidence, promotes mature abstractions, and removes stale structure. At inference time, retrieval chooses the abstraction level required by the query and renders only the needed support depth. Figure 1 summarizes the full lifecycle.

3.2 Lifecycle Memory State

PERSONATREE defines memory as a typed graph whose support edges form a three level persona tree. A **Leaf** node records event level evidence from one interaction or a short local segment. Above the leaves, **Mid** nodes capture reusable behavior or state abstractions, and **Root** nodes represent durable user claims grounded in those abstractions. The support edges connect lower level evidence to higher level interpretation, so each abstract claim remains traceable to the observations that justify it.

Formally, for user u, we write the memory state as

$$\mathcal{T}_u = (\mathcal{V}_u^L, \mathcal{V}_u^M, \mathcal{V}_u^R, \mathcal{E}_u), \tag{1}$$

where $\mathcal{V}_u^L$, $\mathcal{V}_u^M$, and $\mathcal{V}_u^R$ denote leaf, mid, and root nodes, and $\mathcal{E}_u$ contains typed support edges. Each node $v = (x, t, a, z, c, \ell)$ stores textual content, temporal metadata, task attributes, an embedding, confidence, and an abstraction level $\ell \in \{L, M, R\}$.

The attribute vector a stores schema information for the target application, including domains, event types, states, roles, and boundaries. In our experiments, these attributes instantiate persona oriented descriptors for person understanding. The lifecycle operators remain defined over abstraction levels and support edges, allowing the same memory state to be instantiated with other schemas.

3.3 Online Evidence Insertion

Online insertion adds new interactions to the tree as evidence before they become abstractions. Given a new interaction h_t, PERSONATREE extracts a typed leaf l_t with textual content, temporal metadata, schema attributes, and an embedding. Before insertion, normalization under the target schema maps factual updates, subjective preferences, and safety constraints into compatible attribute types.

The new leaf is then matched against existing mid nodes. Candidate mids first need to be schema compatible with the leaf. Among compatible candidates, PERSONATREE combines embedding similarity with an evidence validator, since semantic closeness alone does not determine whether a leaf supports an abstraction. The validator labels the relation as support, conflict, or unrelated, and these judgments are mapped to positive, negative, or near-zero evidence scores. The match score is

$$\begin{aligned} s_{\text{match}}(l, m) &= \mathbf{1}[\kappa(a_l, a_m) = 1]\, s(l, m), \\ s(l, m) &= \alpha \cos(z_l, z_m) + \beta\, \psi(l, m), \end{aligned} \tag{2}$$

where $\kappa(a_l, a_m)$ is a compatibility predicate under the schema and $\psi(l, m) \in [-1, 1]$ maps the validator judgment from conflict to support. If $\max_m s_{\text{match}}(l, m) \geq \theta_M$, the leaf attaches to the best mid. When the validator finds a conflict, the leaf remains separate from the candidate abstraction and passes its negative evidence weight to the confidence update. If no candidate passes the threshold, the leaf remains orphaned until consolidation provides enough context for a safer abstraction.

3.4 Confidence Update and Short-Term Activation

PERSONATREE uses confidence as a shared signal that carries supporting and conflicting evidence into later selection, pruning, and promotion decisions. For a matched node, confidence is updated in log-odds space with temporal decay and weights derived from evidence:

$$\begin{aligned} L_t &= (L_{t-1} - L_{\text{base}})e^{-\lambda \Delta t} + L_{\text{base}} + \omega_E, \\ c_t &= \sigma(L_t), \end{aligned} \tag{3}$$

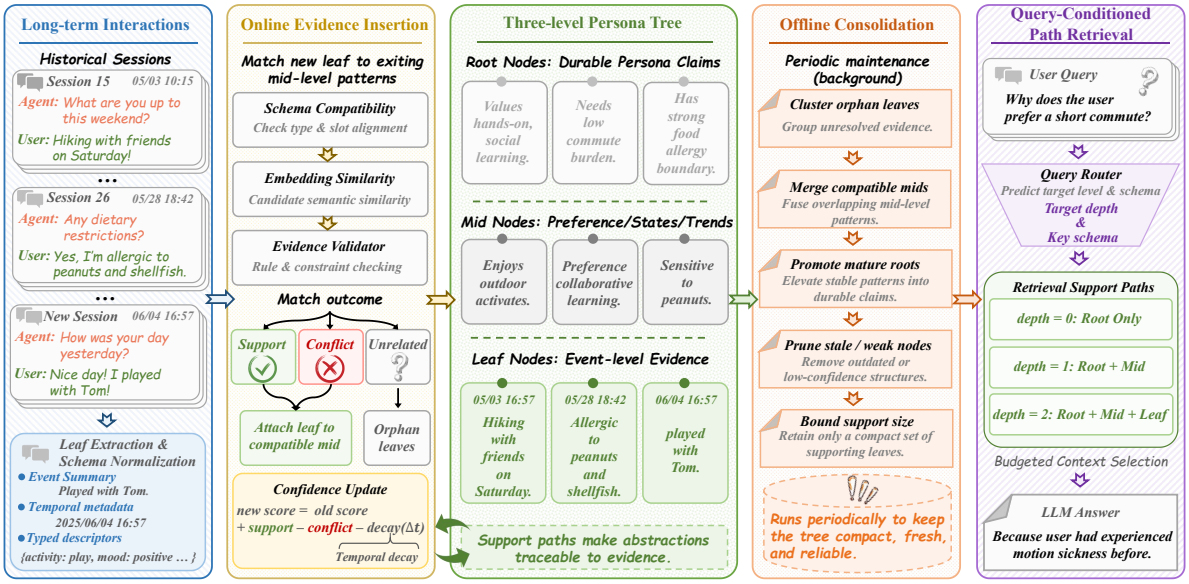

Figure 1: Overview of PERSONATREE. The framework converts long term interactions into evidence leaves, consolidates reusable patterns into mid level nodes, promotes stable user claims into roots, and retrieves compact support paths for query conditioned answering.

where L_t is the current log-odds, L_{base} is a baseline prior, Δt is the elapsed time since the last update, λ is the decay rate for the node type, and ω_E is the evidence weight assigned by the validator.

The log-odds form lets support and conflict act as positive and negative evidence weights, while the sigmoid mapping keeps the resulting confidence bounded. The decay term pulls confidence toward the baseline when no new evidence arrives, while ω_E carries support or contradiction from the validator. Schema types whose content should persist can use reduced or zero decay; less persistent behavioral patterns fade without supporting evidence. PERSONATREE also keeps recent activations for updated or retrieved nodes, giving retrieval a recency signal separate from longer term reliability.

3.5 Offline Consolidation

Offline consolidation handles evidence that is too sparse or unstable for online abstraction. It periodically revisits unresolved leaves and low confidence abstractions. First, it forms candidate clusters from orphan leaves using schema attributes, temporal proximity, and semantic similarity. A cluster becomes a mid candidate only when it contains enough evidence and its leaves can support a consistent pattern. If the pattern expressed by the cluster matches an existing mid, PERSONATREE attaches the member leaves to that mid and rewrites its description to reflect the combined evidence. Otherwise, PERSONATREE abstracts the cluster into a new mid node and links the member leaves as its support.

Root promotion applies stricter thresholds than consolidation at the mid level. PERSONATREE first selects mids that pass thresholds for confidence, support size, and temporal coverage. It then groups compatible mids into candidate root clusters and summarizes each cluster into a durable user claim. These promotion criteria make root formation conservative, so weak or narrowly contextual evidence is not elevated into durable claims about the user.

The same pass also limits memory growth. Stale nodes decay, old orphan leaves are removed, and mids retain only a bounded set of supporting leaves. The system prunes mids and roots with low confidence. When a weak root is deleted, its mids are detached, while confident local patterns remain available.

3.6 Query-Conditioned Path Retrieval

Retrieval is conditioned on the granularity of the query. A query may ask for a concrete event, a recurring pattern or evolving state, or a durable claim about the user. PERSONATREE therefore routes the query q to an abstraction level $h \in \{L, M, R\}$ and predicts schema attributes that constrain the search space:

$$\begin{aligned} \hat{h} &= \arg\max_{h \in \{L,M,R\}} p_\phi(h \mid q), \\ \hat{a}_q &= g_\phi(q), \end{aligned} \tag{4}$$

where p_ϕ is the routing model, $\hat{h}$ is the selected level, and g_ϕ predicts the query attribute constraint

$\hat{a}_q$.

After routing, PERSONATREE renders only the evidence depth needed by the query and the token budget. A leaf level query is answered from compact event memories. For mid and root level queries, retrieval starts with the shallowest useful rendering and expands along the support path when the answer needs event detail, temporal evidence, or an explicit rationale. Given retrieved support paths $\mathcal{P}_q$, where each path P connects a root, a supporting mid, and leaves when available, the renderer $g_d(P)$ exposes P as a root alone ($d = 0$), a root with its supporting mid ($d = 1$), or the full root–mid–leaf chain ($d = 2$). The final context is selected under a token budget:

$$\begin{aligned} \mathcal{G}_q &= \{g_d(P) : P \in \mathcal{P}_q,\ d \in \{0,1,2\}\}, \\ \mathcal{C}^\star &= \arg\max_{\mathcal{C} \subseteq \mathcal{G}_q} \sum_{x \in \mathcal{C}} \big(r(q,x) - \lambda_b \mathrm{tok}(x)\big), \\ \text{s.t.} \quad & \sum_{x \in \mathcal{C}} \mathrm{tok}(x) \le B_q, \end{aligned} \tag{5}$$

where $\mathcal{G}_q$ is the set of candidate renderings, $\mathcal{C}^\star$ is the selected context, B_q is the token budget, $\mathrm{tok}(x)$ counts the tokens in rendering x, and λ_b weights token cost. The score $r(q, x)$ favors a rendering when it is aligned with the query, supported by reliable or recently activated nodes, and set at the evidence depth requested by the query. The token penalty keeps the context compact unless deeper evidence is needed to answer or justify the query.

3.7 Person Understanding Instantiation

For person understanding, we instantiate the generic attribute vector with persona oriented descriptors. Leaf nodes store local evidence fields, including interaction domain, temporal metadata, typed descriptors, and a compact event summary. Mid nodes organize compatible leaves into reusable person patterns such as preferences, ongoing states, recurring triggers, and temporal trends. Root nodes summarize durable user claims over identity, personality, values, principles, and safety relevant boundaries. This schema defines the experimental instantiation of PERSONATREE, while the lifecycle state, confidence update, consolidation, and path retrieval mechanisms remain unchanged under alternative schemas. Appendix E lists the concrete schema and implementation rules.

4 Experiments

We evaluate PERSONATREE across six person understanding and persistent memory benchmarks, testing whether its lifecycle improves user modeling while remaining practical under long interaction histories.

4.1 Experimental Setup

The suite combines person understanding and preference alignment benchmarks (KnowMe, RealPref, CUPID) with long term memory benchmarks for chat histories, project oriented interaction, and beyond factual cognitive memory (LongMemEval, RealMem, LoCoMo-Plus) (Wu et al., 2026; Guo et al., 2026; Kim et al., 2025; Wu et al., 2024a; Bian et al., 2026; Li et al., 2026b). Baselines cover full history prompting, flat retrieval, persistent memory systems, and strong references from individual benchmarks when available. Table 1 reports one compact summary score for each benchmark with Qwen3-32B, Gemini 3 Flash, and GPT-5.4 Mini as answer backbones. Appendix A describes the benchmark scope, baseline policy, compact score definitions, and metric breakdown.

4.2 Main Results

Across six benchmarks and three answer backbones, PERSONATREE gives the strongest overall pattern in Table 1, ranking first in 12 of 18 compact scores and reaching the top two in 16 settings. Its advantage is most consistent on benchmarks that require stable user modeling and preference use. On CUPID, PERSONATREE is best under all three answer backbones, ahead of the next strongest row by 2.7, 1.9, and 2.7 points. On RealPref and KnowMe, it leads with Qwen3-32B and GPT-5.4 Mini and remains top two with Gemini 3 Flash. Taken together, these results suggest that the lifecycle representation helps models accumulate user evidence and reuse it across preference and persona questions.

The broader memory benchmarks show a similar pattern against persistent memory systems and benchmark strong references. On RealMem and LoCoMo-Plus, PERSONATREE leads with Qwen3-32B and GPT-5.4 Mini. On LongMemEval, it obtains the highest Gemini 3 Flash score and remains top two with Qwen3-32B. Across answer models, Qwen3-32B and GPT-5.4 Mini each place PERSONATREE first on five of six benchmarks, while Gemini 3 Flash gives two leading scores and three additional top two scores.

Table 1: **Main results across person understanding and persistent memory benchmarks.** Scores are compact summaries computed from the native metrics of each benchmark; higher is better. Strong reference rows use $K = V + \text{fact}$ on LongMemEval, Graphiti on RealMem, and SeCom on LoCoMo-Plus. Appendix A.1 defines each aggregation. Best results are **bolded**, and second best results are underlined.

Qwen3-32B						
Method	**KnowMe**	**LongMemEval**	**RealPref**	**RealMem**	**CUPID**	**LoCoMo-Plus**
Full history	39.9	–	60.1	39.1	48.8	20.1
Flat retrieval	43.3	68.1	73.5	40.3	50.1	20.4
Benchmark strong	–	72.1	–	46.2	–	15.3
Mem0	43.2	68.6	75.6	44.4	51.8	16.1
A-MEM	44.8	70.0	75.8	40.5	51.7	17.7
TiMem	46.1	**72.7**	76.9	44.3	53.1	20.9
PERSONATREE	**47.6**	72.3	**78.1**	**47.8**	**55.8**	**22.3**
Gemini 3 Flash						
Method	**KnowMe**	**LongMemEval**	**RealPref**	**RealMem**	**CUPID**	**LoCoMo-Plus**
Full history	45.4	–	63.5	40.4	55.1	24.8
Flat retrieval	48.3	69.1	80.6	41.5	56.7	**25.2**
Benchmark strong	–	73.4	–	**47.4**	–	17.8
Mem0	47.8	70.0	82.7	45.6	59.5	18.3
A-MEM	49.4	70.9	82.5	41.6	57.4	20.3
TiMem	**50.7**	73.6	**83.1**	45.1	59.1	23.6
PERSONATREE	50.4	**76.2**	83.0	47.1	**61.4**	24.6
GPT-5.4 Mini						
Method	**KnowMe**	**LongMemEval**	**RealPref**	**RealMem**	**CUPID**	**LoCoMo-Plus**
Full history	46.8	–	77.5	41.3	56.8	22.9
Flat retrieval	49.9	70.3	85.7	42.4	58.3	23.6
Benchmark strong	–	74.5	–	48.2	–	18.2
Mem0	49.1	71.7	87.0	46.6	59.5	18.8
A-MEM	50.9	72.3	88.1	42.9	60.8	20.7
TiMem	52.1	**74.9**	87.9	46.1	60.6	24.2
PERSONATREE	**53.3**	74.3	**89.1**	**49.8**	**63.5**	**25.5**

4.3 Efficiency under Long User Histories

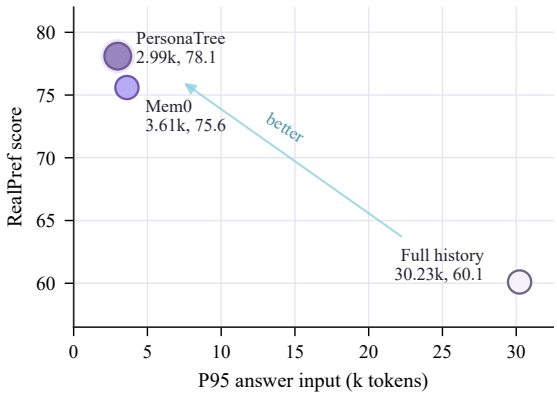

Figure 2: RealPref score and context size. PERSONATREE attains a higher task score than full history prompting and Mem0 while keeping the P95 answer input below 3k tokens.

Long user histories make answer context size part of the evaluation. A memory method should improve preference following without sending an expanding interaction history to the answer model. Table 2 reports the RealPref score together with median retrieved tokens, P95 answer input, input growth per 100 turns, and the gap between median and P95 input. The final columns record median synchronous latency and background update tokens.

Table 2: **RealPref efficiency statistics.** Token counts are shown in thousands. Latency reports median added latency.

Metric	Full history	Mem0	PERSONATREE
Score	60.1	75.6	**78.1**
Ret. P50	0.00k	1.71k	**1.34k**
Input P95	30.23k	3.61k	**2.99k**
Growth	24.46k	0.69k	**0.27k**
Gap	6.09k	1.67k	**1.42k**
Latency	0.00s	0.18s	0.35s
Bg./Turn	0.00k	0.81k	0.94k

Compared with full history prompting, PERSONATREE improves the RealPref score from 60.1 to 78.1 while reducing the P95 answer input from 30.23k to 2.99k tokens and the input growth from 24.46k to 0.27k tokens per 100 turns. Compared with Mem0, it retrieves fewer median tokens, lowers the P95 answer input, and improves the score from 75.6 to 78.1. The input gap also drops to 1.42k tokens, indicating a more stable answer context across typical and high load turns. Figure 2 plots score against answer context length.

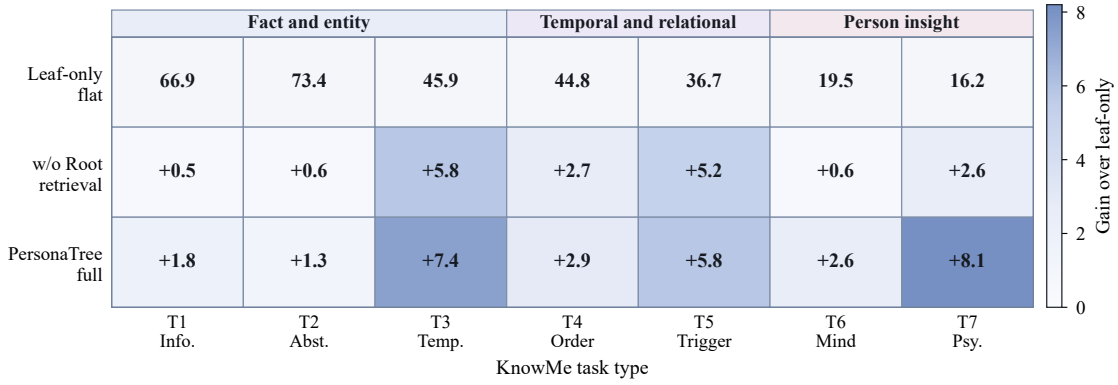

Figure 3: KnowMe hierarchy ablation across task types. The leaf only row reports the original score, while the other rows report gains over retrieval from leaves alone.

Table 3: **Core ablations.** Compact summary of hierarchy and path-retrieval ablations under Qwen3-32B.

Setting	Variant	Score	Key scores	Ctx.	Gain
KnowMe	**Leaf Only**	43.3	T6 19.5 T7 16.2	–	+0.0
	No Root	45.9	T6 20.1 T7 18.8	–	+2.6
	PERSONATREE	**47.6**	**T6 22.1** **T7 24.3**	–	**+4.3**
RealPref	**Flat Nodes**	75.7	Align. 3.61 Qual. 4.00	3.18k	+0.0
	PERSONATREE Path	**78.1**	**Align. 3.91** **Qual. 4.20**	**2.99k**	**+2.4**

5 Ablation and Mechanism Analysis

The main results establish the outcome pattern; we now examine which memory behaviors help explain it. All analyses use Qwen3-32B and keep the data, judge, decoding setup, and answer format fixed, so the comparisons isolate changes in memory organization and retrieval policy. The analyses focus on how hierarchy, support paths, and retrieval depth help the answer model connect local observations with higher level interpretations about the user. Appendix B reports the full ablation and activation tables behind the compact results discussed here.

5.1 Hierarchy Ablation on Abstract Person Understanding

KnowMe lets us examine hierarchy across different depths of person understanding, since its tasks range from local factual questions to abstract person insight. Under the same answer backbone and budget, *Leaf Only* retrieves raw leaf evidence, *No Root Retrieval* keeps refinement from leaves to mid nodes while root retrieval is disabled, and PERSONATREE retrieves over Leaf, Mid, and Root paths.

Abstract person questions require the model to synthesize a higher level interpretation from dispersed observations, and the hierarchy mirrors this inferential structure. Leaf evidence contains the observations, mid nodes collect repeated behavior and evolving states, and root nodes store durable persona hypotheses backed by those patterns. Mid nodes reduce the risk that a root claim is inferred directly from isolated events, while root nodes provide reusable interpretations that can be applied across questions. This gives the answer model a scaffold for reasoning about stable preferences, boundaries, and latent intent while retaining the lower level support needed for grounded answers.

The full hierarchy raises the KnowMe macro average from 43.3 to 47.6, a +4.3-point gain over retrieval from leaves alone. The largest change appears on T7, from 16.2 to 24.3, where questions written by expert annotators require reasoning about complex motivations and identity construction from dispersed narrative evidence. Figure 3 shows that the gains also extend to T3–T5, which KnowMe defines as temporal reasoning, logical event ordering, and mnestic trigger analysis. Table 3 reports the compact scores behind these changes.

5.2 Path Retrieval Ablation on Preference Following

After isolating the role of hierarchy, we next ask whether the retrieved structure itself matters when the available memory nodes are held fixed. The flat variant ranks all Leaf, Mid, and Root nodes in one pool, while PERSONATREE returns support chains that connect an abstract preference to the observations that formed it.

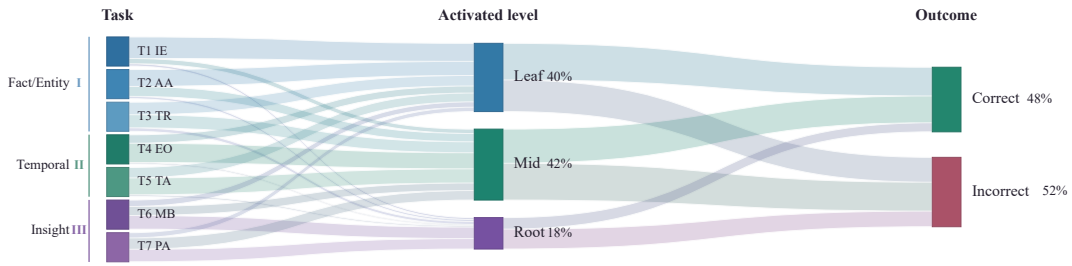

Figure 4: KnowMe flow from task type to activated memory level and answer outcome.

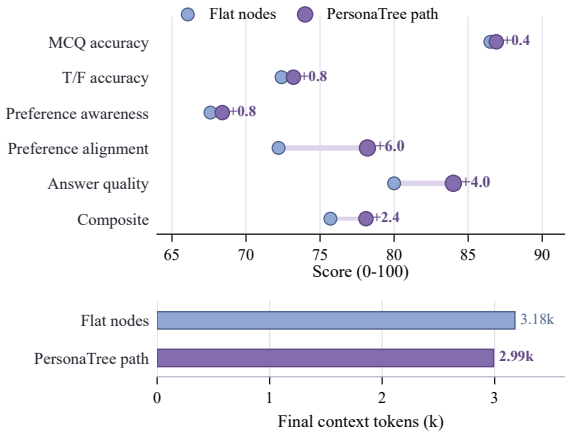

Figure 5: RealPref path retrieval ablation. Support path retrieval yields stronger preference alignment and answer quality with a final context of 2.99k tokens.

Preference following requires a grounded representation of both the preference and its support. A response often depends on what the system believes about the user, why that belief is plausible, and which observations support it in the current situation. Path retrieval keeps these pieces together in the returned context. In the chain, observed behavior supports a recurring preference, and the preference then constrains the answer.

Path retrieval improves the RealPref composite score from 75.7 to 78.1, as summarized in Table 3 and Figure 5. Most of the gain comes from generation metrics. Preference alignment rises from 3.61 to 3.91, and answer quality rises from 4.00 to 4.20. Because the final contexts are kept at similar sizes, 3.18k tokens for flat retrieval over all nodes and 2.99k tokens for support path retrieval, the result suggests that the improvement comes from how evidence is organized in the retrieved context.

5.3 Memory Activation by Task Type

The activation analysis examines how PERSONATREE moves through the memory hierarchy when answering different kinds of person understanding questions. As the question shifts from factual recall to temporal organization and then to person insight, the system should rely on increasingly abstract memory levels while keeping lower level evidence available for inspection. We therefore use activation traces to study whether the retrieved levels follow this change in question demand.

Figure 4 connects task types, activated memory levels, and answer outcomes. T1/T2 mostly activate Leaf nodes, T4/T5 shift toward Mid nodes, and T6/T7 activate Root nodes much more often. This distribution is consistent with the intended division of labor in the hierarchy. Local factual questions rely on event evidence, temporal and narrative questions use mid level organization, and insight questions more often require root level abstractions. The retrieved support paths then expose the lower level evidence behind these abstractions, making the resulting user interpretation easier to inspect.

6 Conclusion

This paper framed persistent person understanding as a lifecycle memory problem and introduced PERSONATREE, a three level persona tree that connects event evidence, recurring patterns, and durable user claims through typed support edges. By combining conservative writing, evidence based consolidation, confidence update, and path retrieval, PERSONATREE keeps abstract user interpretations grounded in the observations that support them.

Across six person understanding and persistent memory benchmarks, PERSONATREE ranks first in 12 of 18 compact scores and reaches the top two in 16 settings. Ablations show that hierarchy supports abstraction from dispersed evidence, while support path retrieval improves preference following under a comparable context budget. Together, these results support a view of agent memory as a structured evidential basis for forming, retrieving, and inspecting user models over time.

Limitations

This work focuses on text based interaction histories and evaluates PERSONATREE on English language benchmarks. Extending the same lifecycle representation to multilingual settings, speech interaction, and multimodal companion scenarios would test how the schema should adapt when user evidence arrives through richer channels.

Ethics Statement

This paper evaluates PERSONATREE on existing research benchmarks and does not involve deployment with real users. For practical use, persistent memory systems should follow standard data governance practices, including transparency about what is stored, user control over retained information, and appropriate safeguards for personal data.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

FC Bartlett. 1932. *Remembering: a study in experimental and social psychology.* Macmillan.

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. 2026. Realmem: Benchmarking llms in real-world memory-driven interaction. *arXiv preprint arXiv:2601.06966*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, and 1 others. 2025. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. 2026. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-llm interactions. *arXiv preprint arXiv:2603.04191*.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025a. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025b. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. 2025. Cupid: Evaluating personalized and contextualized alignment of llms from interactions. *arXiv preprint arXiv:2508.01674*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, and 1 others. 2026a. Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. *arXiv preprint arXiv:2601.02845*.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. 2026b. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. *arXiv preprint arXiv:2602.10715*.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, and 1 others. 2025. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

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.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: towards llms as operating systems.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. In *International Conference on Learning Representations*, volume 2025, pages 990–1023.

DE RUMELHART. 1980. Schemata: The building blocks of cognition. *Theoretical issues in reading comprehension*.

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

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. 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.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Ai persona: Towards life-long personalization of llms. *arXiv preprint arXiv:2412.13103*.

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

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024b. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Chenglong Li, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, and 1 others. 2026. Knowme-bench: Benchmarking person understanding for lifelong digital companions. *arXiv preprint arXiv:2601.04745*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. 2026. Mem-t: Densifying rewards for long-horizon memory agents. *arXiv preprint arXiv:2601.23014*.

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.

A Evaluation Details and Benchmark Results

A.1 Evaluation Details

Benchmarks. The benchmark suite combines person understanding, preference alignment, and long term memory evaluation. KnowMe measures person understanding across seven task types, from factual questions to high level person insight (Wu et al., 2026). RealPref measures whether an assistant can infer and follow long horizon user preferences (Guo et al., 2026), while CUPID tests contextualized preference alignment from interactions (Kim et al., 2025). LongMemEval evaluates long term interactive memory with retrieval and QA metrics (Wu et al., 2024a). RealMem evaluates memory driven interaction in realistic project settings (Bian et al., 2026). LoCoMo-Plus evaluates beyond factual cognitive memory in long dialogues (Li et al., 2026b).

Baselines. The comparisons separate raw history prompting, unstructured retrieval, and persistent memory systems. Full history prompting measures how far the answer model can use the accumulated interaction directly. Flat retrieval measures the effect of selecting relevant snippets without the hierarchy and support links used by PERSONATREE. The main persistent memory baselines are Mem0 (Chhikara et al., 2025), A-MEM (Xu et al., 2026), and TiMem (Li et al., 2026a). We include specialized references when a benchmark provides a directly comparable setting: $K = V$ and $K = V + \text{fact}$ for LongMemEval, Graphiti/Zep for RealMem (Rasmussen et al., 2025), and SeCom for LoCoMo-Plus.

Metrics. The main table uses one compact score per benchmark for readability. Table 4 defines the aggregation. The compact scores are computed from the detailed metric columns; accuracy, F1, retrieval, and QA metrics are already on a 0–100 scale, while 1–5 judge scores are scaled to the same range before averaging.

Beyond the compact score, Figure 6 reports the original benchmark metrics behind the Qwen3-32B results in Table 1. Each panel normalizes its own axes for readability, and the values next to the PERSONATREE curve are the original metric values. The resulting profiles show how the aggregate gains are distributed across the metrics used by each benchmark.

The clearest gains for PERSONATREE appear on CUPID and RealPref, where the curve extends outward on preference and generation axes such as alignment, answer quality, and changing preference F1. KnowMe shows a related result: the advantage is clearest on tasks T3 through T7, which require more abstraction over accumulated user evidence. RealMem and LoCoMo-Plus add positive evidence on dynamic update, multi hop, temporal, and overall axes, while LongMemEval shows strong QA and knowledge update results. Taken together, the six panels show that PERSONATREE maintains a large overall advantage across person understanding, preference following, dynamic memory use, and long dialogue evaluation.

A.2 Baseline and Inference Configuration

Comparisons are run through benchmark adapters that fix the split, answer backbone, output format, and judge protocol within each benchmark. Baseline specific memory construction remains inside each method's own interface; the controlled part is the downstream answering and scoring call.

A.3 Benchmark Adapter Interfaces

Each benchmark adapter converts a benchmark instance into a common answer call without changing its native scoring protocol. The call contains the current query, task metadata, and the answer context supplied by the evaluated memory policy: full history for long context prompting, flat snippets for flat retrieval, or rendered support paths for PERSONATREE. LongMemEval additionally scores retrieved items against answer-location labels with Recall@10 and NDCG@10. The adapter then normalizes outputs to the benchmark's required form, such as MCQ labels, true/false labels, short factual answers, open responses, or dialogue continuations, before invoking the original metric or judge.

Table 4: **Compact score aggregation.** Higher values are better after the listed scaling.

Benchmark	Metric columns	Compact score
KnowMe	T_1–T_7 task scores	$\mathrm{mean}(T_1, \ldots, T_7)$
LongMemEval	Recall@10, NDCG@10, Top-5 QA, Top-10 QA, Knowledge-Update QA	Mean of the five columns
RealPref	MCQ accuracy, T/F accuracy, preference awareness, preference alignment, answer quality	Mean of MCQ, T/F, and 20× each judge score
RealMem	Dynamic increase QA, dynamic update QA, proactive alignment QA, temporal reasoning QA	Mean of the four QA categories
CUPID	All instances F1, consistent F1, contrastive F1, changing F1, all instances generation score	Mean of the four F1 columns and 10× generation score
LoCoMo-Plus	LoCoMo-Plus score	Direct benchmark score

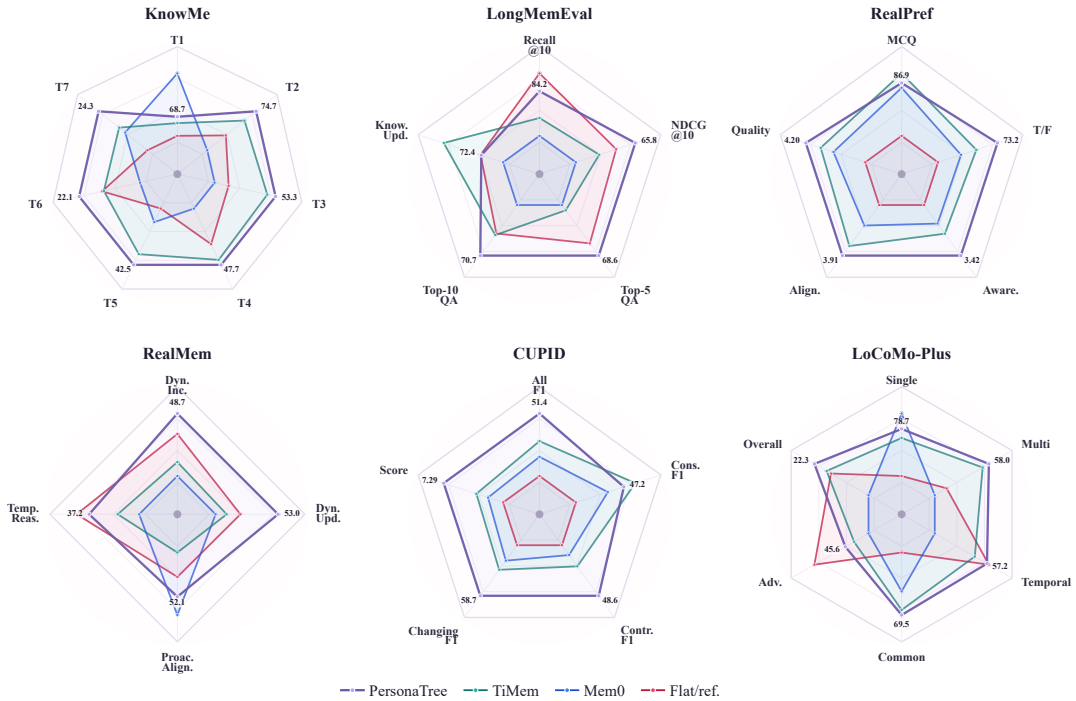

Figure 6: **Benchmark metric breakdown under Qwen3-32B.**

B Ablation and Analysis Results

Hierarchy ablation. Table 5 reports the task level KnowMe scores behind the hierarchy analysis. The comparison starts from retrieval over leaf evidence, then keeps refinement from leaves to mid nodes while root retrieval is disabled, and finally uses the complete PersonaTree hierarchy. The gains in macro average and in the more abstract KnowMe tasks show how root nodes complement local evidence for person understanding.

Table 5: **Full KnowMe hierarchy ablation.**

Variant	T1	T2	T3	T4	T5	T6	T7	Avg	Δ
Leaf Only	66.9	73.4	45.9	44.8	36.7	19.5	16.2	43.3	+0.0
No Root	67.4	74.0	51.7	47.5	41.9	20.1	18.8	45.9	+2.6
PersonaTree	**68.7**	**74.7**	**53.3**	**47.7**	**42.5**	**22.1**	**24.3**	**47.6**	**+4.3**

Path retrieval ablation. Table 6 keeps the same Leaf, Mid, and Root inventory and changes only the retrieval policy. With memory content held fixed, the score and context differences reflect how evidence is organized in the retrieved context. MCQ and T/F are RealPref classification metrics; Aware., Align., and Qual. are judge scores for preference awareness, preference alignment, and answer quality; Comp. is the RealPref composite used in this paper, and Ctx. is the final answer context size.

Table 6: **Full RealPref path retrieval ablation.**

Variant	MCQ	T/F	Aware.	Align.	Qual.	Comp.	Ctx.
Flat Nodes	86.5	72.4	3.38	3.61	4.00	75.7	3,180
PersonaTree Path	**86.9**	**73.2**	**3.42**	**3.91**	**4.20**	**78.1**	**2,990**

Layer activation corpus. Table 7 summarizes the activation analysis over 3 character histories, 27,062 timeline events, and 2,580 question records. The rows connect KnowMe task types with the memory levels activated by retrieval. Factual ques-

tions rely more on Leaf nodes, temporal and narrative questions shift toward Mid nodes, and insight questions activate Root nodes more often. Qs. is the number of questions, Dom. is the dominant activated level, Hits is the average number of retrieved nodes, and Ctx. is the average final context size.

Table 7: **Full KnowMe layer activation statistics.** Leaf, Mid, and Root columns report activation percentages; rows sum to 100%.

Task	Level	Qs.	Leaf	Mid	Root	Dom.	Hits	Ctx.
T1	Level I / Fact & Entity	330	73.9	18.2	7.9	Leaf	6.72	2,694
T2	Level I / Fact & Entity	300	58.0	34.0	8.0	Leaf/Mid	7.31	2,604
T3	Level I / Fact & Entity	561	43.0	45.1	11.9	Leaf/Mid	7.55	3,187
T4	Level II / Temporal Logic	125	28.8	66.4	4.8	Mid	7.09	2,993
T5	Level II / Temporal Logic	581	36.0	58.0	6.0	Mid	7.52	3,341
T6	Level III / Insight	495	22.0	30.9	47.1	Root	8.40	3,744
T7	Level III / Insight	188	18.1	38.8	43.1	Root	7.87	3,816

C Qualitative Support-Path Examples

The examples below show how support paths keep a benchmark query, the inferred preference, and the lower level evidence in the same retrieved context, complementing the ablation in Table 6.

Example 1: Learning Environment Recommendation

USER QUERY. I'm interested in developing new skills—what kind of learning environment would you recommend?

FLAT RETRIEVED NODES. A flat retriever may return several individually relevant memories: the user prefers live athlete interviews, dislikes guided tours, avoids distracting wearables, and enjoys social outdoor activities. Their relation to the learning query is left for the answer model to infer.

PERSONATREE SUPPORT PATH.

Root. The user values firsthand, social, and flexible experiences that preserve direct contact with people and activities.

Support-Mid. The user prefers hands on, discussion based workshops or peer learning environments with live interaction and room for self direction.

Evidence-Leaf 1. A live Q&A with a retired football legend provided firsthand stories and practical wisdom that the user found especially insightful.

Evidence-Leaf 2. An organized sightseeing tour felt too rigid; spontaneous visits to local sports venues felt more engaging.

PATH INFORMED ANSWER. The answer recommends a collaborative workshop with group projects and open discussion. The support path keeps the generalized learning preference and the observations that justify it together, so the recommendation follows the user's preference for firsthand, social, and flexible learning.

Example 2: Work Environment Recommendation

USER QUERY. I'm looking for a new job—what kind of work environment should I consider?

FLAT RETRIEVED NODES. A flat retriever may return memories about cycling, car commutes, animal shelter volunteering, and financial education. These nodes are relevant to the user, although the answer model must still infer which parts matter for the job recommendation.

PERSONATREE SUPPORT PATH.

Root. The user benefits from work arrangements with outdoor movement, flexible timing, and a low commute burden.

Support-Mid. The user prefers flexible work schedules or remote options that allow midday outdoor breaks or exercise sessions.

Evidence-Leaf 1. An hour long car commute left the user fatigued and frustrated, leading to an aversion to long car commutes.

Evidence-Leaf 2. Endurance cycling became the user's main form of exercise because it combines physical challenge with peaceful time outdoors.

PATH INFORMED ANSWER. The answer recommends a remote or flexible role with autonomous scheduling and room for outdoor breaks. This recommendation reflects the user's need for outdoor movement and low commute burden.

D Instruction Templates

The templates below describe the model-facing instructions used for memory construction and retrieval. Each item lists the role, input, and structured output expected by the corresponding step. Benchmark answer formats follow the evaluation setup in Appendix A.

Instructions Used by the Memory Pipeline

```
Leaf extraction. Role: expert cognitive
memory extractor. Input: a recent
conversation transcript. The model
compresses the transcript into an objective
event, the user's action or statement, and
the user's feeling when supported by the
text. Greetings and filler are removed, and
the event summary is kept under 40 words.
Objective fact or state updates receive a
neutral emotion tag. Output fields are scene
tags, emotion tag, and event summary.

Mid pattern refinement. Role: psychological
behavior analyst. Input: a bounded list of
related leaf events. The model summarizes
one core behavioral pattern, stable state,
or preference. Pattern types include
condition trigger, trend change, frequency
```

Instructions Used by the Memory Pipeline (continued)

statistic, continuing state, and preference strength. The output contains pattern type, scene tags, emotion tendency, and pattern description.

Root trait generation. Role: psychological profiler. Input: established mid level patterns and optional case specific guardrails. The model produces one durable root claim in a calm third person declarative sentence. The claim is typed as personality, value, principle, objective identity, or hard boundary. Identity is reserved for objective roles or durable factual states, while hard boundary claims require explicit evidence such as an allergy, prohibition, intolerance, trauma trigger, or severe aversion.

Root candidate grouping. Role: organizer of mid level patterns into root candidates. Input: indexed mid nodes with pattern type, scene tags, emotion tendency, and pattern text. The model groups mids that reflect the same deeper trait, value, identity, principle, or boundary. Grouping follows latent coherence, and weakly related mids can remain ungrouped.

Evidence validation. Role: behavioral logic validator. Input: a known pattern and a new event. The model returns relevance and an ordered evidence category: strong support, weak support, weak conflict, or strong conflict. Unrelated pairs receive no evidence category. The category informs matching and confidence update.

Query routing. Role: retrieval router for a persona schema tree. Input: the user's latest query. The model predicts Leaf for specific events, details, or facts; Mid for habits, recent states, and preferences; and Root for hypothetical advice, core personality, or deeper user understanding. Output fields are target retrieval level and scene tags.

Retrieved path rendering. Input: a selected root, its supporting mids, and selected evidence leaves. The renderer exposes paths as [Root], [Support-Mid], and [Evidence-Leaf] blocks. These blocks enter the context for answer generation, keeping abstraction and supporting evidence adjacent.

Table 8: **Structured output schemas used by the memory pipeline.**

Component	Structured fields	Downstream use
Leaf extraction	Scene tags, emotion tag, event summary.	Creates event level evidence and supplies searchable descriptors for retrieval and consolidation.
Mid refinement	Pattern type, scene tags, emotion tendency, pattern description.	Forms reusable abstractions over compatible leaves and provides candidate parents for future evidence.
Root generation	Trait type and root description.	Promotes mature mid patterns into stable user claims used by higher level retrieval.
Root grouping	Candidate groups of mid node indices.	Selects semantically coherent mid clusters before root generation.
Evidence validation	Relevance flag and evidence category.	Updates match decisions and confidence through ordered support or conflict categories.
Query routing	Target retrieval level and scene tags.	Constrains retrieval to the abstraction depth and schema region requested by the query.
Benchmark answer	Task specific answer field.	Normalizes outputs for MCQ, true/false, short answer, open response, or continuation scoring.

E Persona Schema Instantiation

The generic lifecycle uses a task schema to define node attributes and compatibility rules. In the person understanding setting, each leaf stores a compact memory text, temporal metadata, domain or scene descriptors, typed descriptors, an embedding, and a confidence score. The typed descriptors record roles, schedules, task facts, explicit preference statements, affective states, recurring triggers, values, and boundaries when these signals appear in the evidence. Before insertion, the normalization step rewrites factual updates as state descriptors and assigns neutral affect unless the interaction itself provides an emotional signal. This type assignment prevents durable factual states or constraints from being treated like ordinary preference evidence during decay, allowing persistent state descriptors to use reduced or zero decay while subjective preferences and affective patterns con-

Table 9: **Persona schema fields in the experimental instantiation.**

Level	Core fields	Schema descriptors	Main lifecycle role
Leaf	Event summary, timestamp, embedding, confidence, parent pointer.	Domain or scene tags, emotion tag, factual state, preference signal, boundary signal, role or schedule cue.	Stores local evidence from an interaction segment and supplies candidates for matching or consolidation.
Mid	Pattern description, pattern type, scene tags, emotion tendency, confidence, child leaves.	Condition trigger, trend change, frequency statistic, continuing state, preference strength.	Represents recurring states or patterns that connect multiple local events.
Root	Durable claim, trait type, confidence, supporting mids, support coverage.	Personality, value, principle, objective identity, or hard boundary.	Provides stable user claims for abstract person understanding and high level retrieval.
Edge	Parent-child relation, support direction, evidence category when available.	Support, conflict, or unrelated validation signal; temporal and schema compatibility.	Keeps higher level claims traceable to lower level evidence and carries validation outcomes into confidence update.

tinue to fade without supporting evidence.

Mid and root schemas specify which abstractions can be formed from those leaves. The four broad mid types are preferences, ongoing states, recurring triggers, and temporal trends. During consolidation, orphan leaves are first partitioned by schema type and domain, then grouped by temporal proximity and embedding similarity. A group is accepted only when its member leaves share compatible descriptors and point to a consistent pattern. If the pattern matches an existing compatible mid, the member leaves attach to that mid and the description is rewritten from old and new support. Otherwise, the group is abstracted into a new mid node with its member leaves as support. Root nodes summarize compatible mids into identity, personality, values, principles, or safety relevant boundaries. A root promotion requires high confidence, multiple supporting leaves or mids, and coverage across time. Roots that represent constraints use conservative decay and can be retrieved as veto conditions when a query touches the corresponding boundary.

The evidence validator is applied after schema filtering. Its prompt contains the leaf text, the candidate mid description, and the relevant attributes, and returns support, conflict, or unrelated with a short rationale for auditing. Support and conflict outcomes guide the match score in Eq. 2 and the confidence update in Eq. 3, while unrelated leaves remain unattached. At retrieval time, the predicted level and schema attributes filter candidates by domain, type, and boundary relevance. The renderer then builds alternative views of each support path, such as a root alone, a root with one supporting mid, or the full root–mid–leaf chain with selected leaves. The budgeted selection in Eq. 5 decides which view enters the final context, so full paths are used only when the query needs that evidence depth and the token budget allows it.