KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions

Tingyu Wu[1,2*], Zhisheng Chen[1,2*], Ziyan Weng[4*], Shuhe Wang[7], Shuo Zhang[2], Sen Hu[2,3], Silin Wu[5], Qizhen Lan[2,6†], Huacan Wang[1,2†], Ronghao Chen[2,8†]

[1]UCAS, [2]QuantaAlpha, [3]PKU, [4]CITYU-DG, [5]HAINNU, [6]UTHealth, [7]NUS, [8]IAPM (Guangdong)

[*]**These authors contributed equally to this work.**

† **Correspondence:** Qizhen.Lan@uth.tmc.edu, chenronghao@alumni.pku.edu.cn, wanghuacan17@mails.ucas.ac.cn

Abstract

Existing long-horizon memory benchmarks mostly use multi-turn dialogues or synthetic user histories, which makes retrieval performance an imperfect proxy for person understanding. We present KnowMe-Bench, a publicly releasable benchmark built from long-form autobiographical narratives, where actions, context, and inner thoughts provide dense evidence for inferring stable motivations and decision principles. KnowMe-Bench reconstructs each narrative into a flashback-aware, time-anchored stream and evaluates models with evidence-linked questions spanning factual recall, subjective state attribution, and principle-level reasoning. Across diverse narrative sources, retrieval-augmented systems mainly improve factual accuracy, while errors persist on temporally grounded explanations and higher-level inferences, highlighting the need for memory mechanisms beyond retrieval.

1 Introduction

A long-standing goal in Artificial Intelligence is to build *lifelong digital companions* that can support users over extended horizons by maintaining coherent personalization, context awareness, and behavior consistent with users' evolving goals and values. Recent LLM-based agent frameworks increasingly aim at sustained interaction across sessions rather than isolated question answering (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023). In this setting, the central capability is ***person understanding***: a companion should form and update an internal model of the user that supports explanation (why a choice was made), anticipation (what the user is likely to prefer next), and alignment (what the user seeks to pursue or avoid).

Importantly, *memory* is a necessary substrate but not a sufficient definition of person understanding. A system can store and retrieve facts yet still fail to infer stable principles, connect distant experiences to present reactions, or explain recurring decision patterns. This paper therefore asks: *how should we benchmark person understanding as an evidence-grounded inference problem over lived experience, rather than as retrieval over a fact database?*

Despite rapid progress on long-horizon agent evaluation, we identify two gaps that prevent existing benchmarks from directly measuring person understanding.

(G1) Evaluation misalignment: retrieval proxies $\neq$ person understanding. Most benchmarks focus on retrieval, temporal ordering, knowledge updates, conflict handling, or longitudinal state tracking across sessions (Wu et al., 2025; Maharana et al., 2024a; Hu et al., 2025a; Castillo-Bolado et al., 2024; Tan et al., 2025; Hu et al., 2026). These tasks are necessary, yet they do not directly test whether an agent can infer and use an implicit *person model*—e.g., motivations and avoidance goals, stable principles, evolving self-concepts, relationship structure, and affective triggers—to explain or anticipate behavior. In addition, "deep" questions without evidence constraints invite free-form speculation.

(G2) Data Substrate Misalignment: Low-Density and Decontextualized Experience Traces Most scalable benchmarks construct user histories from chat logs, synthetic events, or model-generated

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30214–30238
July 2-7, 2026 ©2026 Association for Computational Linguistics

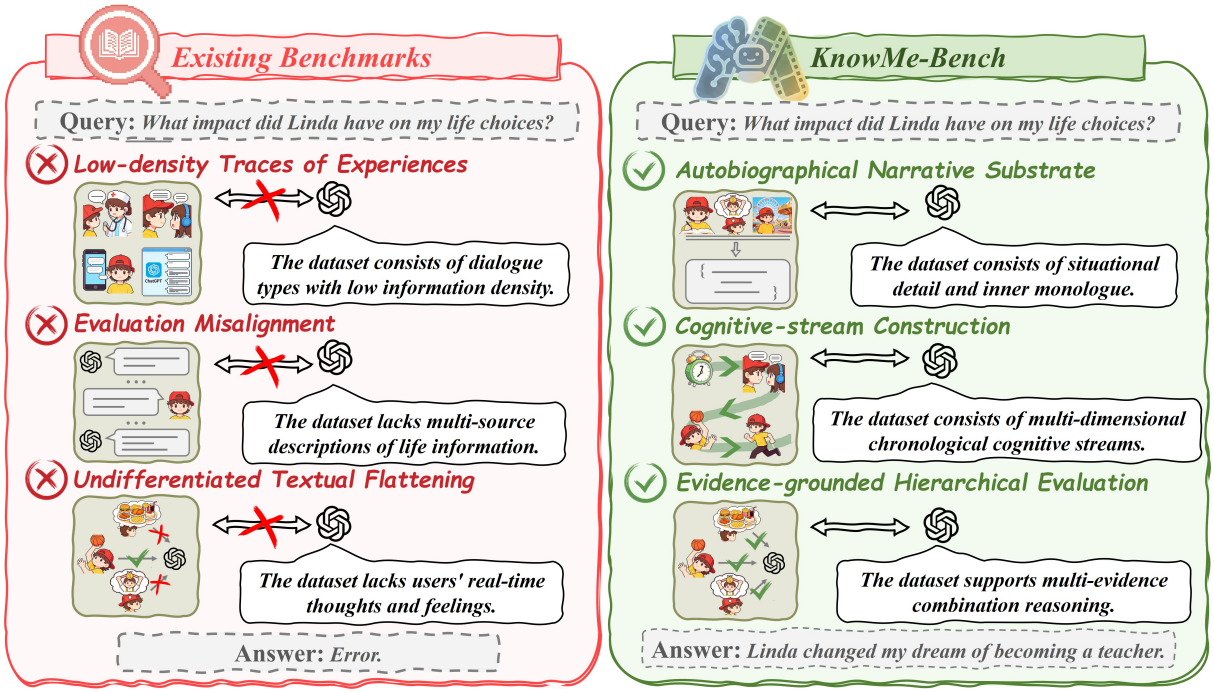

Figure 1: Comparison of benchmark substrates. Existing long-horizon memory evaluations typically rely on sparse dialogue traces or synthetic interaction histories, which provide limited support for evidence-grounded person modeling. KnowMe-Bench instead uses autobiographical narratives with aligned actions, context, and inner thoughts, enabling flashback-aware reconstruction and hierarchical evaluation from factual recall to interpretive reasoning.

interactions (Maharana et al., 2024a; Castillo-Bolado et al., 2024; Wu et al., 2025). Although efficient, such substrates fail to support person understanding due to two structural limitations. ***(G2a) Density loss***: experiences are compressed into sparse traces, weakening the coupling between observable actions and the internal deliberation that gives them personal significance. ***(G2b) Structure loss***: heterogeneous experiential signals are flattened into undifferentiated text, erasing modality cues and temporal alignment needed for long-horizon attribution. Accordingly, a benchmark for person-model inference must approximate the *multimodal* organization of lived experience; when represented textually, this entails explicit separation of distinct *textual modalities* rather than a single narrative surface. This view aligns with autobiographical memory and narrative identity theories, which emphasize that stable self-knowledge emerges from temporally structured, subjectively interpreted experience (Conway and Pleydell-Pearce, 2000; McAdams and McLean, 2013).

To bridge these gaps, we introduce **KnowMe-Bench**, a benchmark for evaluating *evidence-grounded person-model inference* from long-form autobiographical experience. We operationalize this goal through three design modules (M1–M3), each explicitly aligned with the identified gaps.

(M1) Autobiographical narrative substrate (addresses G2a). KnowMe-Bench uses autobiographical narratives that retain the joint expression of external events and internal interpretation, yielding high-density evidence suitable for person-model inference (Conway and Pleydell-Pearce, 2000; McAdams and McLean, 2013).

(M2) Cognitive-stream reconstruction with mnestic realignment (addresses G2b; supports G2a). To make evidence usable for long-horizon attribution, we reconstruct narratives into a chronological cognitive stream anchored by explicit timestamps and locations. We decompose the text into five fields: (1) visual observations, (2) auditory inputs, (3) situational context, (4) accessible background knowledge, and (5) inner monologue. This representation improves evidence granularity (supporting G2a) and enables *mnestic realignment*: present-time mnemonic triggers remain anchored in the current timeline, while recalled content is relocated to its chronological origin, restoring temporal and causal structure (addressing G2b).

(M3) Evidence-grounded hierarchical evaluation with expert verification (addresses G1; leverages M2). To directly measure person understanding, we propose a three-tier evaluation suite: *Tier 1: Factual extraction*, *Tier 2: Subjective state attribution*, and *Tier 3: Decision and principle reasoning*. Tiers 2–3 require (i) a concise inference and (ii) an explicit evidence set of supporting events in the aligned timeline, ensuring auditability and discouraging free-form speculation. Deep items are produced and cross-validated by trained annotators against the reference aligned timeline.

Baselines and diagnostics. We provide baselines spanning long-context prompting, retrieval-augmented agents, and external memory-store / agentic-memory systems (Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025). These results enable diagnostic comparison of memory mechanisms and quantify the gap between retrieval-oriented competence and person-model inference. In summary, we make three contributions:

- **Benchmarking person understanding.** We formalize person understanding for lifelong digital companions as an *auditable person-model inference* problem over long-horizon experience, and introduce KnowMe-Bench, a publicly releasable benchmark built from autobiographical narratives (~4.7M tokens).
- **High-density, structured experience representation.** We construct flashback-aware, time-aligned lifelogs via cognitive-stream reconstruction with multiple textual modalities and mnestic realignment.
- **Hierarchical evaluation and diagnostics.** We propose an evidence-linked, three-tier evaluation protocol with expert verification and provide diagnostic baselines across representative agent designs.

2 Related Work

Evaluation of Long-Term Memory Agents. The evaluation of memory in LLM-based agents has evolved from effective context window tests (Hu et al., 2025b) to multi-turn interactions that assess memory consolidation (Chhikara et al., 2025; Li et al., 2025). Recent benchmarks focus on the agent's ability to update specific facts or track entity states over distinct conversational turns (Zhong et al., 2024). More recent work also moves beyond purely conversational traces: CloneMem evaluates long-term memory over diaries, social media posts, and emails in AI-clone settings (Hu et al., 2026), while concurrent work organizes personalized memory by event time rather than dialogue time, constructing semantically grounded durative memories for personalized agents (Su et al., 2026). However, these evaluations predominantly treat memory as a temporal retrieval or state-tracking problem over personal traces, prioritizing recall fidelity and longitudinal consistency over interpretative reasoning. Current benchmarks often overlook autobiographical reasoning, where an agent must infer implicit information, such as stable principles or affective triggers, from long-horizon causal chains rather than explicit statements.

Benchmarks for Person Modeling and Psychology. Research on "Persona Agents" typically relies on static profiles or role-playing descriptions provided in the system prompt (Sun et al., 2025; Kroczek et al., 2025; Chen et al., 2025). While some studies incorporate psychometric evaluations like MBTI or Big Five (Brickman et al., 2025; Ke et al., 2025; Szymanski et al., 2025), they generally use these frameworks as rigid templates to steer generation. Such static profiling fails to capture the complexity of human behavior, which is inherently context-dependent and evolves over time. Furthermore, the data substrates used in these tasks are often synthetic chat logs or simulated sandbox environments (Cheng et al., 2025; Chou et al., 2025; Nguyen and Welch, 2026). These sources lack the sensory grounding and introspective density characteristic of complex autobiographical narratives, limiting the evaluation of deep person modeling.

Timeline Construction and Narrative Processing. Constructing structured timelines from unstructured text is critical for grounding agent memory. Traditional timeline generation (TLG) often assumes a linear progression of events or relies on simplified timestamp extraction (Liu and Zhang, 2025; Qorib et al., 2025). Such linear assumptions are insufficient for processing complex personal accounts, which frequently contain non-linear temporal structures like flashbacks and mental time travel. Naive ingestion of such narratives results in causal scrambling, where past events are incorrectly anchored to the present context (Fatemi et al., 2024; Maharana et al., 2024b). Unlike stochastic rewriting approaches that risk hallucination, methods that

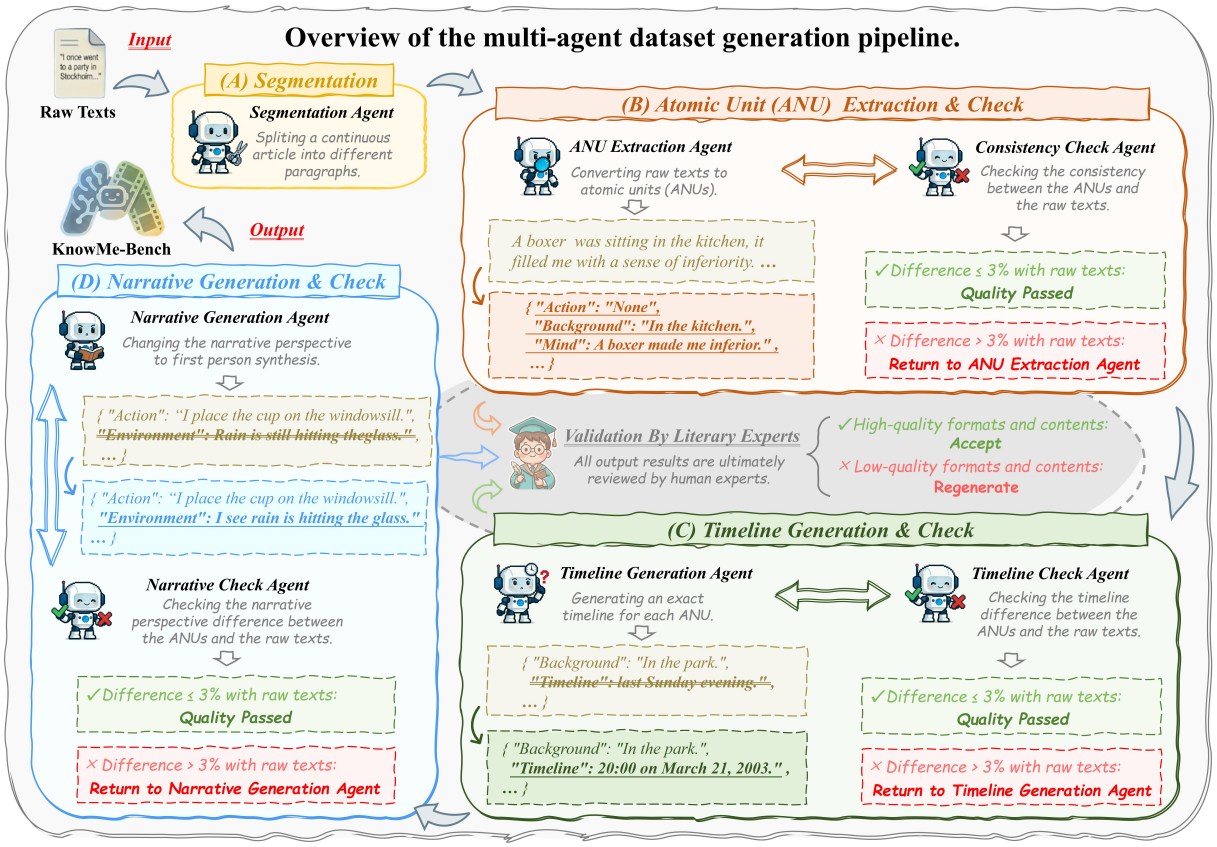

Figure 2: **Overview of the multi-agent dataset generation pipeline.** The framework transforms unstructured raw narratives into the structured KnowMe-Bench benchmark through four sequential stages: (A) Segmentation, (B) Atomic Unit (ANU) Extraction, (C) Timeline Generation, and (D) Narrative Generation. To ensure data fidelity, each generative module is paired with a specific Check Agent that enforces a "Verify-and-Revise" loop, culminating in final validation by human literary experts.

Benchmark	Experience substrate	Temporal representation	Primary evaluation focus
LongMemEval (Wu et al., 2025)	Multi-session assistant dialogues	Dialogue order and session boundaries	Long-horizon conversational recall
LoCoMo (Maharana et al., 2024a)	Grounded long conversations	Event graphs plus dialogue sessions	Very long conversational consistency
MemBench (Tan et al., 2025)	Interactive agent scenarios	Interaction turns	Factual plus reflective memory behaviors
CloneMem (Hu et al., 2026)	Diaries, social posts, and emails	Longitudinal personal traces	Tracking evolving personal states for AI clones
KnowMe-Bench	**Autobiographical narratives**	**Flashback-aware chronological stream**	**Evidence-grounded interpretive reasoning over long-horizon experience**

Table 1: Comparison with representative long-term memory evaluations. KnowMe-Bench differs most clearly in its flashback-aware chronological representation and in targeting evidence-grounded interpretive reasoning rather than only conversational recall or state tracking.

model cognitive primitives and flashback-aware alignment are necessary to preserve the temporal-causal integrity of the source material.

3 Methodology

3.1 Overview

We propose **KnowMe-Bench**, a framework designed to enable evidence-grounded person-model inference over long-horizon autobiographical experience. To address the challenges of low-density

evidence and non-linear narration, we construct a *flashback-aware chronological cognitive stream* from raw narratives. As illustrated in Figure 2, our pipeline operates via a four-stage multi-agent workflow (Modules A–D). We enforce a **"Faithfulness-First"** principle: non-generative stages rely on index-based extraction, while generative stages are guarded by a generic *Verify-and-Revise* protocol (detailed in Appendix A) to prevent hallucination and maintain strict adherence to the source text.

3.2 Stage I: Context-Aware Segmentation (Module A)

Autobiographical narratives are structurally heterogeneous. To preserve causal micro-structure, Module A functions as a deterministic **semantic boundary detector**. Instead of fixed-length chunking, it identifies natural boundaries (e.g., scene transitions) and slices the raw text by indices. This purely extractive approach ensures the verbatim preservation of the original content, providing a faithful input for downstream processing.

3.3 Stage II: Atomic Narrative Unit (ANU) Extraction (Module B)

To expose micro-evidence, Module B decomposes raw segments into **Atomic Narrative Units (ANU)**—the smallest auditable carriers of experience. We formally define an ANU as a tuple:

$$U = (\mathrm{id},\ t^{\mathrm{anch}},\ \ell,\ C), \tag{1}$$

where id is the unique identifier, t^{anch} is the temporal anchor, ℓ is the mandatory location, and C is a structured cognitive record containing five primitives: *Action, Dialogue, Environment, Background,* and *Mind*. To ensure global retrievability in million-token contexts, t^{anch} is not restricted to a coarse verbatim date string. When the source text provides underspecified or scene-level time expressions, the timeline agent synthesizes micro-granularity anchors consistent with the local narrative flow (e.g., second-level offsets within a scene), so that each ANU is indexed by the globally unique tuple $(\mathrm{id}, t^{\mathrm{anch}}, \ell, C)$. During evaluation, models are not required to reproduce the literal timestamp token. Instead, temporal tasks score whether the model recovers the correct duration, relative order, or trigger–event linkage induced by these anchors. To ensure granularity, we impose hard constraints on the complexity of C (e.g., decomposing abstract states into observable micro-behaviors), ensuring the substrate captures the high-density "micro-texture" of memory.

3.4 Stage III: Flashback-Aware Temporal Realignment (Module C)

Standard timestamp extraction fails on narratives containing nested temporal structures (e.g., flashbacks). Module C restores causal structure via a **Mnestic Realignment Protocol**.

- **Mnestic Separation:** We conceptually separate each unit into the *Event Content* (C_{event}, to be relocated to its historical origin) and the *Mnemonic Trigger* (T_{trigger}, to remain anchored in the present stream of consciousness).
- **Stack-Based Alignment:** We employ a stack-based state machine to track nested contexts. The system predicts alignment actions (e.g., PUSH for entering flashbacks, POP for returning) to reorder events chronologically while preserving the narrator's psychological timeline. (State transition rules and action semantics are detailed in Appendix A.4.)

3.5 Stage IV: Narrative Instantiation and Validation (Module D)

To produce a queryable first-person record without flattening the structure, Module D acts as an **Embodied Decoder**. It performs component-wise subjectivization, transforming objective descriptors in the ANU into immediate sensory experiences (e.g., "Rain hits glass" $\rightarrow$ "I see rain hitting glass"). Final validation is conducted by human literary experts to ensure the dataset serves as a reference-quality benchmark, routing any detected errors (e.g., emotional flattening) back to the specific module for revision.

4 Evaluation Framework

To comprehensively assess the agent's capabilities from factual retention to literary reasoning, we introduce the **KnowMe-Bench** evaluation suite. It consists of 7 distinct tasks hierarchically categorized into three cognitive levels.

4.1 Level I: Precision & Factuality (The "Memory" Layer)

This level evaluates the model's ability to precisely retrieve entities and temporal details from the long-context timeline (Q_τ). **Task 1 (Context-Aware**

Information Extraction) tests complete entity recall under strict spatiotemporal constraints. **Task 2 (Adversarial Abstention)** uses "Mismatching Trap" queries to verify that the model refuses to answer when entities or causal links are deliberately distorted. **Task 3 (Temporal Reasoning)** probes duration estimation and the ability to recover chronological order rather than narrative presentation order in flashback-heavy passages.

4.2 Level II: Narrative Logic & Causality (The "Reasoning" Layer)

This level requires understanding logical connections and non-linear transitions. **Task 4 (Logical Event Ordering)** asks the model to order events along non-temporal semantic dimensions such as escalation of danger or emotional intensity. **Task 5 (Mnestic Trigger Analysis)** evaluates whether the model can identify the sensory cue or associative trigger that shifts consciousness from the present scene into a recalled memory.

4.3 Level III: Interpretive Insight (The "Insight" Layer)

The most challenging tier targets subtext and long-range self-explanation. **Task 6 (Mind-Body Interaction)** asks the model to reconcile external behavior with internal state, especially in ironic or self-contradictory moments. **Task 7 (Expert-Annotated Insight)** consists of expert-curated open-ended questions about motives, identity construction, and enduring decision principles, and serves as the strongest diagnostic for deep person understanding.

4.4 Scoring Protocol: LLM-as-a-Judge

Given the subjective nature of literary analysis, purely overlap-based metrics are insufficient. We implement a rigorous **LLM-as-a-Judge** protocol (utilizing GPT-4o) with strict rubric constraints (Scale 0-5).

Scoring Dimensions. The evaluation rubrics are tailored to task type. For factual tasks (T_1, T_2, T_3), the judge scores **Entity Accuracy** and **Value Precision**, with T_2 granting full marks only for correct abstention. For logic tasks (T_4, T_5), the judge evaluates **Sequence Correctness** and the **Validity of Reasoning**, namely whether the answer identifies the correct causal trigger. For insight tasks (T_6, T_7), scoring focuses on the **External-to-Internal Mapping**; a full score requires capturing the specific **core metaphors** in the reference answer rather than generic emotional descriptions.

Metric Reliability. For each task, the final score is the average of the rubric-based scores. Because annotators provide answers rather than direct scalar ratings, pairwise human–human kappa is not the most informative reliability quantity here. Instead, we measure alignment between the rubric-based judge and expert consensus on a held-out subjective subset. This judge-versus-expert agreement is substantial ($\kappa > 0.75$), indicating that the scoring pipeline tracks expert grading conventions under the same blinded protocol later used for model outputs.

5 Experiments

To validate the effectiveness of KnowMe-Bench in distinguishing between retrieval capabilities and person understanding, we conducted extensive evaluations across representative long-horizon memory systems.

5.1 Experimental Setup

Datasets and Narrative Modalities. We use the full KnowMe-Bench corpus (4.7M tokens) spanning three structurally distinct narrative regimes and a total of **2,580 evaluation queries**. Dataset 1 uses Knausgård's *My Struggle* (1.15M tokens) and emphasizes flashbacks and mnestic triggers; Dataset 2 uses the *Neapolitan Novels* (1.76M tokens) and emphasizes linear causal tracking with high-frequency entity updates; Dataset 3 uses Proust's *In Search of Lost Time* (1.30M tokens) and emphasizes introspective passages and abstract internal monologue.

De-identification & Ethics. We apply a privacy pipeline (detailed in the appendix) to remove PII while preserving narrative structure for evaluation.

Resources. The benchmark, generation pipeline, and evaluation code are released at GitHub under MIT; the dataset mirror is hosted at Hugging Face under Apache-2.0.

Human Evaluation Protocol. We ran a 3-expert human study spanning all three evaluation levels. Annotators answered benchmark questions with access to the aligned cognitive stream and, when needed, the original source text; their responses were then scored by the same blinded LLM-as-a-Judge pipeline used for model outputs.

Experts reach 96.5/88.0/83.5 on Levels I/II/III, versus 75.4/62.5/22.6 for the best model, leaving a wide gap on the interpretive tier. Consistent with Section 4.4, the reported $\kappa > 0.75$ measures judge-versus-expert alignment on the subjective subset rather than pairwise annotator agreement. Full procedural details appear in Appendix F.

Model Architecture & Baselines. We distinguish between the *Inference Model* (generation and reasoning) and the *Embedding Model* (vector retrieval). Our inference backbones are Qwen3-32B (long-context), GPT-5-mini, DeepSeek-R1, and Gemini-3 Pro. Across these backbones we evaluate a consistent Base / Naive-RAG / Mem0 / MemOS protocol; on Qwen3-32B we additionally report A-Mem to test whether reflective memory rewriting helps deep person-level inference. Naive RAG ($k = 50$) is the standard dense-retrieval baseline, Mem0 represents structured entity-state memory, and MemOS represents a log-based chronological memory architecture.

Base Models vs. Memory Systems. For plain base models, the full narrative usually exceeds the effective input window. We therefore use a truncation protocol that preserves the initial setup and the most recent context at inference time. Memory systems, by contrast, do not ingest the full narrative monolithically; they retrieve a task-conditioned top-k set of fragments or chronological logs from external storage. This distinction is important for interpreting comparisons between long-context backbones and external-memory architectures.

5.2 Main Results

Table 2 reports the complete overall task-level results for all evaluated systems, while Appendix E retains the per-dataset breakdowns and the backbone-specific radar views. Figure 3 expands the core main-paper visualization to a full-width grounding–insight map; Level-I recall deltas are kept in Table 2 so that the figure can remain readable.

5.3 In-Depth Analysis & Discussion

Takeaway 1: Chronological logging is the decisive ingredient in non-linear narratives. The dataset-level views in Appendix E expose the clearest failure mode on the flashback-heavy corpus: state-updating memory improves entity-centric retrieval but can mis-handle recalled material as a present-state overwrite. On Qwen3-32B in Dataset 1, Mem0 raises T_1/T_2 yet lowers T_3 by 3.5 points, whereas MemOS improves T_3/T_4 by 10.4/10.8. This pattern is consistent with an "update paradox" in which statements such as "I liked apples as a child" are absorbed as current state unless the memory store preserves chronological provenance.

Takeaway 2: Retrieval gains and interpretive gains are separable. Across backbones, Naive RAG and Mem0 remain strongest on the fact-heavy tasks, but they do not deliver the same gains on chronology-sensitive and interpretive tasks. Table 2 shows this split repeatedly: on GPT-5-mini, Mem0 is best on T_1/T_2 at 73.2/78.7, whereas MemOS is best on T_3–T_7; Gemini-3 Pro exhibits the same pattern. A-Mem makes the contrast even sharper on Qwen3-32B, reaching 55.7/41.3 on T_6/T_7 while collapsing to 16.0/16.5 on T_3/T_4. The benchmark therefore separates at least two partially orthogonal capabilities: precise recall of explicit facts and stable reasoning over temporally grounded personal experience.

Takeaway 3: Stronger reasoning backbones still need explicit temporal scaffolding. The expanded evaluation on DeepSeek-R1 and Gemini-3 Pro shows that the central effect is architectural rather than backbone-specific. Both models improve under MemOS on T_3–T_5, even when retrieval-oriented alternatives remain competitive on T_1/T_2. Closed-book contamination checks reinforce this interpretation: on Level II logic items, accuracy drops from 89.1 with original entity names to 33.2 after de-identification, while Level III insight remains at 0 in both settings. The hardest items are therefore not explained by memorized literary knowledge alone; they continue to require evidence-grounded reasoning over the reconstructed chronological stream.

6 Conclusion

In this work, we introduced **KnowMe-Bench**, a benchmark designed to shift the evaluation of lifelong digital companions from simple fact retrieval to evidence-grounded person understanding. By leveraging high-density autobiographical narratives rather than sparse chat logs, we build a substrate that preserves the "micro-texture" of human experience—actions, inner thoughts, and environmental context—while remaining auditable at the level of aligned narrative evidence.

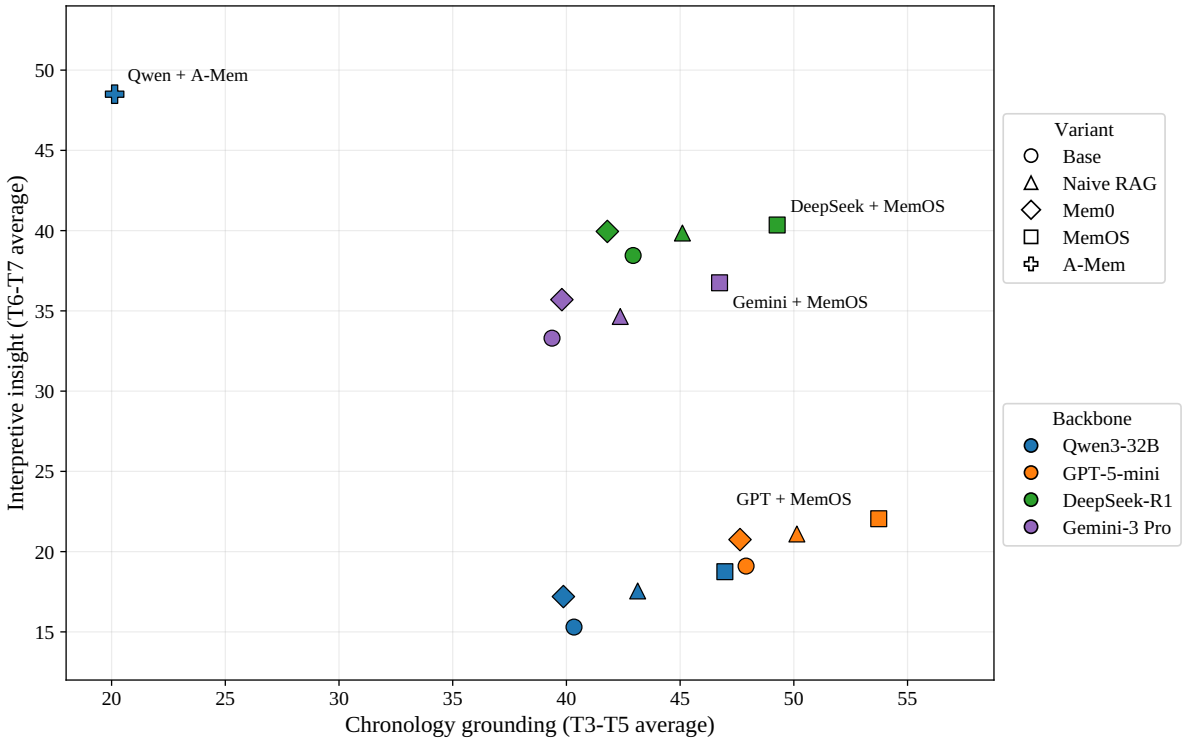

Figure 3: Grounding–insight map across evaluated systems. The x-axis averages the chronology-sensitive tasks T_3–T_5, and the y-axis averages the interpretive tasks T_6–T_7. Colors denote backbone families and marker shapes denote system variants. To keep the plot legible, Level-I recall deltas are reported in Table 2 rather than encoded by marker size. The rightward frontier is populated by MemOS variants, while Qwen3-32B + A-Mem remains a high-insight but weak-grounding outlier.

System	Level I: Fact & Entity			Level II: Narrative Logic		Level III: Insight	
	T1 Extract	**T2** Abstain	**T3** Temporal	**T4** Order	**T5** Trigger	**T6** Mind–Body	**T7** Principle
Backbone: Qwen3-32B							
Base	59.9	66.0	44.4	40.5	36.1	14.3	16.3
+ Naive RAG	68.7 (+8.8)	70.8 (+4.8)	48.3 (+3.9)	42.3 (+1.8)	38.8 (+2.7)	17.1 (+2.8)	18.0 (+1.7)
+ Mem0	**70.4** (+10.5)	**75.2** (+9.2)	41.3 (-3.1)	42.4 (+1.9)	35.9 (-0.2)	16.9 (+2.6)	17.5 (+1.2)
+ MemOS	64.4 (+4.5)	72.4 (+6.4)	**52.7** (+8.3)	**47.1** (+6.6)	**41.1** (+5.0)	18.2 (+3.9)	19.3 (+3.0)
+ A-Mem	55.8 (-4.1)	73.4 (+7.4)	16.0 (-28.4)	16.5 (-24.0)	27.9 (-8.2)	**55.7** (+41.4)	**41.3** (+25.0)
Backbone: GPT-5-mini							
Base	65.4	71.5	54.1	47.3	42.3	18.6	19.6
+ Naive RAG	72.2 (+6.8)	75.5 (+4.0)	57.3 (+3.2)	48.8 (+1.5)	44.3 (+2.0)	21.3 (+2.7)	20.9 (+1.3)
+ Mem0	**73.2** (+7.8)	**78.7** (+7.2)	51.6 (-2.5)	48.7 (+1.4)	42.6 (+0.3)	20.7 (+2.1)	20.8 (+1.2)
+ MemOS	69.6 (+4.2)	76.6 (+5.1)	**60.8** (+6.7)	**52.9** (+5.6)	**47.5** (+5.2)	**21.5** (+2.9)	**22.6** (+3.0)
Backbone: DeepSeek-R1							
Base	51.2	42.2	52.7	37.0	39.1	41.6	35.3
+ Naive RAG	57.7 (+6.5)	47.9 (+5.7)	57.0 (+4.3)	38.0 (+1.0)	40.3 (+1.2)	**43.4** (+1.8)	36.3 (+1.0)
+ Mem0	**59.1** (+7.9)	51.1 (+8.9)	50.5 (-2.2)	37.0 (+0.0)	37.9 (-1.2)	43.1 (+1.5)	36.8 (+1.5)
+ MemOS	53.4 (+2.2)	**51.5** (+9.3)	**61.1** (+8.4)	**43.4** (+6.4)	**43.3** (+4.2)	43.2 (+1.6)	**37.5** (+2.2)
Backbone: Gemini-3 Pro							
Base	50.5	52.2	45.3	36.0	36.8	35.4	31.2
+ Naive RAG	59.0 (+8.5)	57.7 (+5.5)	49.6 (+4.3)	38.7 (+2.7)	38.8 (+2.0)	37.0 (+1.6)	32.3 (+1.1)
+ Mem0	**60.8** (+10.3)	**62.3** (+10.1)	43.7 (-1.6)	37.2 (+1.2)	38.5 (+1.7)	**39.0** (+3.6)	32.4 (+1.2)
+ MemOS	56.3 (+5.8)	60.5 (+8.3)	**53.4** (+8.1)	**44.1** (+8.1)	**42.7** (+5.9)	37.9 (+2.5)	**35.6** (+4.4)

Table 2: Overall results on KnowMe-Bench across all evaluated systems. Parenthetical deltas are measured against the Base model within the same backbone family, and boldface marks the strongest absolute score within each backbone block. DeepSeek-R1 and Gemini-3 Pro follow the same Base/Naive-RAG/Mem0/MemOS protocol as the two primary backbones, while Qwen3-32B additionally includes A-Mem as a reflective-memory comparison. The pattern is consistent across families: retrieval-oriented variants help the fact-heavy tasks, whereas MemOS more reliably improves chronology-sensitive and insight-oriented tasks.

Our experiments reveal a clear evaluation gap in current long-horizon memory research. Retrieval-augmented baselines and entity-tracking systems improve factual recall, but they remain structurally fragile when the benchmark requires flashback-aware chronology, trigger–event linkage, and longer-range interpretive reasoning. The resulting update paradox shows why a personal memory system cannot be treated as a static fact database: without explicit temporal provenance, recalled past experience is too easily conflated with present state.

KnowMe-Bench provides a concrete testbed for this distinction through flashback-aware reconstruction, evidence-linked evaluation, and diagnostic comparisons across memory architectures. We hope it encourages future work to move beyond context-window extension and vector similarity toward memory systems that support stronger temporal grounding, more reliable longitudinal user modeling, and evidence-grounded reasoning over lived experience.

Limitations

Methodologically, the benchmark must navigate the inherent subjectivity of literary analysis through a rigorous "LLM-as-a-Judge" protocol validated by human experts, while bearing the cost and operational complexity of a multi-agent generation and de-identification pipeline for dense autobiographical data.

Ethical considerations

We strictly adhere to the licenses and usage policies of the open-source models and datasets utilized in our experiments. Our benchmark does not introduce additional risks regarding data privacy or human rights violations.

References

Jocelyn Brickman, Mehak Gupta, and Joshua R Oltmanns. 2025. Large language models for psychological assessment: A comprehensive overview. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251343582.

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, and 1 others. 2025. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*.

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

Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E Gonzalez, and Wei-Lin Chiang. 2025. Visionarena: 230k real world user-vlm conversations with preference labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3877–3887.

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.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating LLMs on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

Sen Hu, Zhiyu Zhang, Yuxiang Wei, Xueran Han, Zhenheng Tang, Huacan Wang, and Ronghao Chen. 2026. CloneMem: Benchmarking long-term memory for AI clones. *arXiv preprint arXiv:2601.07023*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025a. Evaluating memory in LLM agents via incremental multi-turn interactions. *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*.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2025. Exploring the frontiers of LLMs in psychological applications: A comprehensive review. *Artificial Intelligence Review*, 58(10):305.

Leon OH Kroczek, Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, and Andreas Mühlberger. 2025. The influence of persona and conversational task on social interactions with a LLM-controlled embodied conversational agent. *Computers in Human Behavior*, page 108759.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, and 1 others. 2025. MemOS: A memory os for AI system. *arXiv preprint arXiv:2507.03724*.

Xiaochen Liu and Yanan Zhang. 2025. Etimeline: An extensive timeline generation dataset based on large language model. *arXiv preprint arXiv:2502.07474*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024a. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024b. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Dan P. McAdams and Kate C. McLean. 2013. Narrative identity. *Current Directions in Psychological Science*, 22(3):233–238.

Duc Cuong Nguyen and Catherine Welch. 2026. Generative artificial intelligence in qualitative data analysis: Analyzing—or just chatting? *Organizational Research Methods*, 29(1):3–39.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv:2309.10677.

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 (UIST)*.

Muhammad Reza Qorib, Qisheng Hu, and Hwee Tou Ng. 2025. Just what you desire: Constrained timeline summarization with self-reflection for enhanced relevance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25065–25073.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2026. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. *arXiv preprint arXiv:2601.07468*.

Lipeipei Sun, Tianzi Qin, Anran Hu, Jiale Zhang, Shuojia Lin, Jianyan Chen, Mona Ali, and Mirjana Prpa. 2025. Persona-L has entered the chat: Leveraging LLMs and ability-based framework for personas of people with complex needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–31.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

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, Vienna, Austria. Association for Computational Linguistics.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-Mem: Agentic memory for LLM agents. *Preprint*, arXiv:2502.12110.

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*, pages 19724–19731.

A Faithfulness Verification Protocol

To operationalize the "Faithfulness-First" principle, we implement a generic verification layer that guards all generative transformations (Modules B, C, and D). This appendix details the computation of the semantic divergence score (δ), the threshold configurations, and the specific prompts used for the Consistency Check Agent.

A.1 Semantic Divergence Metric (δ)

We define semantic divergence δ not merely as vector distance, but as a measure of *propositional mismatch*. We employ a **Key Information Extraction (KIE)** overlap method.

Let S be the source narrative segment and T be the generated output (e.g., extracted ANUs or instantiated text). We prompt a Validator Agent to extract the set of atomic facts $\mathcal{F}(\cdot)$ from both texts (including entities, timestamps, and actions).

The divergence score is computed as a weighted combination of *Omission Rate* (δ_{miss}) and *Hallucination Rate* (δ_{hall}):

$$\delta(S,T) = \alpha \cdot \underbrace{\left(1 - \frac{|\mathcal{F}(S) \cap \mathcal{F}(T)|}{|\mathcal{F}(S)|}\right)}_{\delta_{miss}} + \beta \cdot \underbrace{\left(\frac{|\mathcal{F}(T) \setminus \mathcal{F}(S)|}{|\mathcal{F}(T)|}\right)}_{\delta_{hall}} \tag{2}$$

where:

- $\mathcal{F}(S) \cap \mathcal{F}(T)$ represents facts present in both source and output.
- $\mathcal{F}(T) \setminus \mathcal{F}(S)$ represents new facts introduced in the output (hallucinations).
- We set $\alpha = 0.4$ and $\beta = 0.6$, penalizing hallucinations more strictly than minor omissions to prevent corruption of the ground truth.

A.2 Threshold Configuration (ϵ)

The acceptance threshold ϵ varies by module sensitivity:

Module	Threshold (ϵ)	Rationale
Mod B (Extraction)	0.05	High tolerance for stylistic compression, zero tolerance for entity loss.
Mod C (Realignment)	0.00	Strict logic check; timestamp order must match the causal graph exactly.
Mod D (Instantiation)	0.03	Allows minor grammatical changes for first-person flow, but bans new adjectives.

Table 3: Divergence thresholds for automatic revision triggers.

A.3 Prompt Implementation

Below is the specific system prompt used by the **Consistency Check Agent** to evaluate Module B (ANU Extraction). This prompt enforces the calculation of δ through step-by-step verification.

System Prompt: The Auditor
You are a strict Data Auditor. Your task is to compare the `SOURCE_TEXT` against the extracted `ANU_JSON`.
Step 1: Fact Extraction List all atomic facts in `SOURCE_TEXT` (Entities, Actions, Time, Location). List all atomic facts represented in `ANU_JSON`.
Step 2: Discrepancy Analysis Identify two types of errors: 1. **[MISSING]**: A critical fact (e.g., a name "John", a time "noon") exists in Source but is absent in JSON. 2. **[HALLUCINATION]**: A fact exists in JSON but is NOT supported by Source (e.g., adding an adjective "angry" when the text only said "said").
Step 3: Verification Decision If there are ANY [HALLUCINATION] tags or significant [MISSING] tags, return Status: REJECT. Otherwise, return Status: PASS.
Output Format: { "status": "PASS" | "REJECT", "score": [0.0 - 1.0], "feedback": "Specific instructions on what to fix..." }

A.4 Error Feedback Loop

If $\delta(S, T) > \epsilon$, the system enters a *Revision Loop*:

1. The Validator Agent generates a natural language feedback message M_{fb} (e.g., *"Error: You missed the location 'gas station' mentioned in line 3."*).
2. The Generator Agent receives the history $[S, T_{old}, M_{fb}]$ and attempts a regeneration T_{new}.
3. This loop repeats up to $k_{max} = 3$ times. If convergence fails, the sample is flagged for manual human review.

For the mnestic realignment module, we use the following action semantics:

- **MAINTAIN:** Extends the current timeline.
- **PUSH(t_{new}):** Triggered by **Structural Narrative Inversions** (assigning C_{event} to t_{new}). It pushes a new layer for sustained flashbacks.
- **POP():** Returns to the parent layer's active timestamp after the recollection ends.
- **TRANSIENT:** Marks fleeting **Associative Triggers** ($T_{trigger}$) that evoke a memory without altering the stack structure.

B Examples

Example I: ANU Extraction

> ***Input Segment:***
> "I put the coffee cup on the windowsill. Rain is still hitting the glass."

ID	**ANU-001**
Time Anchor	Morning, before rain stops
Location	Windowsill

Content Details:

Action: I place the coffee cup on the windowsill.

Environment: Rain is still hitting the glass.

Dialogue: *None* · **Mind:** *None*

Example II: Final Data Instance

ID	**101**
Timestamp	1966-04-25 19:00:00
Location	Windowsill

Content Details:

Action: I place the coffee cup on the windowsill.

Environment: I see rain is still hitting the glass.

Dialogue: *None* · **Mind:** *None* · **Background:** *None*

Unless otherwise stated, the default strict acceptance threshold is $\epsilon = 0.03$; Module B uses 0.05 as listed in Table 3.

We applied a rigorous Context-Aware De-identification Pipeline. Key entities were mapped to consistent pseudonyms (e.g., "Elena" $\rightarrow$ "Subject_A") to preserve coreference chains, and geolocation markers were coarsened to ensure no residual PII remained.

C Prompt Cards

We summarize the prompt files via *Prompt Cards* to improve auditability while avoiding full prompt dumps. Each card reports a minimal contract: **Role**, **Inputs**, **Output Contract**, **Reject/Gate**, and **Hard Constraints**. Redundant boilerplate and in-context examples are omitted. *Note: The cards intentionally abstract the original prompts by removing boilerplate and in-context examples; full prompt texts are provided in the supplementary material.*

C.1 Data Construction Pipeline (Modules A–D)

A. Segmentation

Role Slice raw narrative into segments without altering any character.

Inputs Raw narrative text `N`.

Output List of segment records with `segment_id`, `start_index`, `end_index`, `text` (verbatim substring).

Gate None (extractive slicing only).

Hard Constraints.

- **Verbatim preservation**: no rewriting/summarization/deletion; slicing is index-based only.
- **Semantic boundary**: cut at scene/event/time/location shifts; do not break sentences or ongoing dialogue.
- **Length (prompt-level guidance)**: keep segments approximately within the token budget specified in the prompt.

Skeleton.

```
Input: raw narrative N.
Operation: boundary-based index slicing only (verbatim).
Output(JSON): [{segment_id, start_index, end_index, text}, ...]
```

B. ANU + Check

Role Extract Atomic Narrative Units (ANUs) and audit against the source span for omission/hallucination.

Inputs One segment from Module A (verbatim).

Output (B) ANU list with `id`, `t_anchor` (verbatim), `location` (required), `content{action,dialogue,environment,background,mind}`. (B-check) verdict `{semantic_difference_score,status,issues}`.

Gate Reject if $\delta > 0.05$ (information loss or hallucination).

Hard Constraints.

- **Granularity**: ≤ 3 physical actions *or* ≤ 3 dialogue turns per ANU; otherwise split.
- **No abstract state**: prohibit vague mental labels; decompose into explicit micro-behaviors and/or explicit `mind`.
- **Spatiotemporal unity**: space change or noticeable time jump triggers a new ANU.

Skeleton.

```
Input: one segment (verbatim).
Output: ANU list with mandatory location + five primitives.
Gate: run audit; REJECT if delta > 0.05 or hallucination detected.
```

C. Timeline + Check

Role Maintain a stack-based mnestic realignment state machine and assign chronological placements.

Inputs Current ANU; lookahead (next 3 ANUs); current stack time; stack depth.

Output (C) `{action,time_value,reasoning}` with `action` $\in$ {MAINTAIN,PUSH,POP,TRANSIENT}. (C-check) `{status,logic_error,correction_suggestion}`.

Gate C-check rejects boundary misalignment or implausible duration allocation.

Hard Constraints.

- **Lookahead-based scope**: distinguish transient triggers vs sustained flashbacks.
- **State discipline**: PUSH only if subsequent ANUs belong to past; POP on return; TRANSIENT if immediate return next ANU.
- **C-check**: (i) duration plausibility; (ii) PUSH must be justified by immediately subsequent content.

Skeleton.

```
Inputs: current ANU; lookahead(next 3); stack time; stack depth.
Decide: MAINTAIN / PUSH / POP / TRANSIENT.
Output: {action, time_value(YYYY-MM-DD HH:MM:SS), reasoning}.
```

D. Narrative + Check

Role Instantiate each aligned ANU into first-person experience; reject forbidden distortions.

Inputs Chronologically aligned ANU with optional fields `action/dialogue/environment/background/mind`.

Output (D) one first-person paragraph. (D-check) `{status,hallucination_detected,details}`.

Gate Reject if $\delta > 0.03$ or if any embellishment/emotional injection is detected.

Hard Constraints.

- **Component-wise subjectivization**: translate each present field into immediate "I"-perspective experience.
- **Strict coverage**: cover all fields present (no omission).
- **No hallucination**: do not add adjectives/emotions absent from `mind`/`environment`.

Skeleton.

```
Method: component-wise subjectivization (I-perspective).
Gate: REJECT if delta > 0.03 or distortion detected.
```

C.2 Benchmark Instance Generation (Level I: T1–T3)

E. T1: Extraction

Role Generate retrieval-focused QA with explicit spatiotemporal constraints.
Inputs Evidence items with `id,timestamp,location,category,related_time?`.
Output `[id, question, answer, evidence_ids, ...]`.
Gate N/A.

Hard Constraints.

- **Uniqueness**: include sufficient constraints so the answer is unique.
- **Evidence anchoring**: `evidence_ids` must point to the minimal supporting IDs.

Skeleton.

```
Inputs: evidence items with time + location anchors.
Produce: one uniquely answerable question + minimal evidence_ids.
Output(JSON): [{id, question, answer, evidence_ids}, ...]
```

F. T2: Abstention

Role Compose true fragments into a false relation to test abstention (anti-hallucination).
Inputs Same evidence schema as T1.
Output `[id, question, answer, evidence_ids, ...]`.
Gate Answer must be the fixed abstention token `ABSTAIN`.

Hard Constraints.

- **Entity validity, relation invalidity**: all entities must exist in evidence, but their relations must be wrong.
- **Trap strategies**: entity swapping; spatiotemporal distortion; false causality via unrelated true anchors.

Skeleton.

```
Inputs: valid entities/time/location anchors from evidence.
Construct: mismatching relation while keeping anchors individually true.
Gate: answer MUST be ABSTAIN (fixed token).
```

G. T3: Temporal

Role Generate duration computation and real-world ordering QA under non-linear narration.

Inputs Evidence items with `timestamp` and optional `related_time`.

Output `[id, question, answer, evidence_ids, ...]`.

Gate N/A.

Hard Constraints.

- **Duration**: answer = end timestamp − start timestamp; include start/end anchors in `evidence_ids`.
- **Ordering**: order by real-world occurrence time (not narrative order); do not leak explicit timestamps in options.
- **Trigger vs recalled content**: separate trigger at `timestamp` from recalled event at `related_time`.

Skeleton.

```
Duration: compute end\_ts - start\_ts; evidence\_ids include start+end.
Ordering: ask real-world order (no explicit timestamps in options).
Key: separate trigger(timestamp) vs recalled content(related\_time).
```

C.3 Benchmark Instance Generation (Level II: T4–T5)

H. T4: Logical Event Ordering

Role Generate QA that requires ordering discrete events by a *non-temporal* semantic dimension (e.g., escalation of danger), rather than by explicit timestamps.

Inputs A set of evidence items with `id`, `timestamp`, `location`, `category`, `related_time`? and content fields (e.g., `action`/`dialogue`/`environment`/`background`/`mind`).

Output `[id, question, answer, evidence_ids, ...]`; `question` must specify the ordering `dimension` and require a ranked list (e.g., top-3); `answer` must provide the ordered events and brief justifications.

Gate N/A.

Hard Constraints.

- **Non-temporal ordering**: the rank criterion must not reduce to chronological order; explicitly state a semantic `dimension` (severity/risk/urgency/blame, etc.).
- **Event discreteness**: each ranked element must correspond to a concrete evidence-grounded event (not a diffuse theme).
- **Justified ranking**: provide short, evidence-based reasons for each ordering decision.
- **Evidence closure**: `evidence_ids` must include all items necessary to recover the ranked events and the ordering rationale.

Skeleton.

```
Inputs: evidence items.
Select: a coherent set of discrete events.
Choose: a non-temporal semantic ordering dimension.
Ask: rank events under the dimension (e.g., most->least severe).
Answer: ordered list + brief justifications; evidence_ids cover all ranked events.
```

I. T5: Mnestic Trigger Analysis

Role Generate QA that identifies the sensory cue or associative trigger that evokes memory retrieval under stream-of-consciousness narration.

Inputs Evidence items with `id`, `timestamp`, `location`, `category`, `related_time`? and content fields (especially `environment`/`mind`/`background`).

Output `[id, question, answer, evidence_ids, ...]`; `question` must ask for the *specific trigger* (object/sound/smell/visual cue/phrase) and link it to the recalled content; `answer` must name the trigger and explain the trigger→memory linkage.

Gate N/A.

Hard Constraints.

- **Concrete trigger**: the asked trigger must be an explicitly stated, perceivable cue (not a generic mood).
- **Trigger-to-recall linkage**: the evidence must support a transition into recalled content (often via `related_time` or an immediate shift in `mind`/`background`).
- **No invention**: do not introduce triggers or recalled details absent from evidence.
- **Minimal evidence**: `evidence_ids` should include the trigger span and the recalled-content span.

Skeleton.

```
Inputs: evidence items (timestamp + optional related_time).
Find: explicit sensory/associative trigger + ensuing recall/memory content.
Ask: which specific cue triggered the memory (and what it evoked).
Answer: name trigger + explain linkage; evidence_ids include trigger+recall spans.
Output(JSON): [{id, question, answer, evidence_ids}, ...]
```

C.4 Benchmark Instance Generation (Level III: T6–T7)

J. T6: Mind-Body Interaction

Role	Generate QA that explains the duality between external actions and internal states, including ironic or contradictory behaviors.
Inputs	Evidence items with `id,timestamp,location,category,related_time?` and content fields, requiring both `action/dialogue` and `mind` (or equivalent internal cues).
Output	`[id, question, answer, evidence_ids, ...]`; `question` must require relating bodily/behavioral manifestation to internal psychological state; `answer` must provide an evidence-grounded explanation of the interaction.
Gate	Reject if the selected evidence lacks either (i) observable external manifestation (`action/dialogue`) or (ii) internal-state cues (`mind` or implied cognition).

Hard Constraints.

- **Dual-track requirement**: include both external behavior and internal state in the evidence and in the answer.
- **Interaction focus**: target explanation of irony/contradiction, mismatch, or expressive mapping (body/behavior as signal of mind).
- **Non-speculative**: do not add psychological claims beyond what can be justified by explicit narrative cues.
- **Evidence anchoring**: `evidence_ids` must point to the minimal set supporting both tracks and their linkage.

Skeleton.

```
Inputs: evidence items containing action/dialogue + mind cues.
Ask: how external manifestation reflects/contrasts internal state (incl. irony/
    contradiction).
Answer: cite external cues + internal cues + explain linkage.
Gate: REJECT if either track is missing.
Output(JSON): [{id, question, answer, evidence_ids}, ...]
```

Why no Prompt Card for T7. Task 7 (*Expert-Annotated Insight*) consists of open-ended questions curated by literary experts and is therefore not generated by an automatic prompt-based instance construction procedure. Accordingly, we do not provide a Prompt Card for T7. In evaluation, T7 is assessed under the Level III rubric described in Section 4.4.

D QA Examples

We provide one representative QA instance for each task (T1–T7) from the released benchmark. To ensure compatibility with memory systems that do not retain evidence IDs (or use remapped indices), we present *QA-only* examples without evidence identifiers or anchor excerpts. For T2, we normalize the reference abstention response to the fixed token `ABSTAIN` for evaluation consistency.

D.1 Level I QA Examples (T1–T3)

Example D.1: T1 Context-Aware Information Extraction

> ***Question:***
> "On February 15, 1951, at the school on via Toledo, what specific items did Nunzia bring as gifts for Professor Oliviero?"

Task: T1 · **Dataset:** Dataset 2

Answer: Coffee and sugar.

Example D.2: T2 Adversarial Abstention

> ***Question:***
> "On July 27, 1963, while at the rentals, what did Stefano Carracci say to Pinocchia when he presented her with a small box containing a gold chain with a heart-shaped pendant?"

Task: T2 · **Dataset:** Dataset 2

Answer: `ABSTAIN`

Note: The released reference is an abstention-style response; we standardize it to the fixed token `ABSTAIN` for scoring.

Example D.3: T3 Temporal Reasoning

> ***Question:***
> "On October 20, 1950, how much time passed between the narrator being hit with a paper wad by Lina and Professor Oliviero stumbling and falling motionless to the floor?"

Task: T3 · **Dataset:** Dataset 2

Answer: 5 minutes.

D.2 Level II QA Examples (T4–T5)

Example D.4: T4 Logical Event Ordering

Question:
"Arrange the following events based on the escalation of the girls' bravery and transgression in confronting their fears regarding Don Achille."

Task: T4 · **Dataset:** Dataset 2

Answer (increasing bravery/transgression):

- Throwing the dolls into the basement.
- Entering the building and climbing the dark staircase.
- Accusing Don Achille face-to-face at his door.

Example D.5: T5 Mnestic Trigger Analysis

Question:
"The narrator is physically in the Staircase in May 1954, but her consciousness is mentally reliving a pivotal event from May 1953. What specific environmental anchor triggers this spatiotemporal jump and how does it restructure her current relationship with Lila?"

Task: T5 · **Dataset:** Dataset 2

Answer: The trigger is the dark staircase (together with the remembered purple light in the courtyard), which precipitates a chronological displacement back to the earlier episode and reframes the narrator's bond with Lila as grounded in that shared transgressive ascent.

D.3 Level III QA Examples (T6–T7)

Example D.6: T6 Mind-Body Interaction

> ***Question:***
> "The narrator physically wakes up and checks her leg, then decides to imitate Lina's walk. How does the simultaneous internal logic about her mother rationalize this behavior?"

Task: T6 · **Dataset:** Dataset 2

Answer: The narrator's bodily checking and mimicry are justified by an internal fear of inheriting her mother's lameness. By imitating Lina's walk, she treats Lina's bodily pattern as a protective counter-model, rationalizing the imitation as a way to ward off the threatened repetition of her mother's gait.

Example D.7: T7 Expert-Annotated Insight

> ***Question:***
> "How do I view the conflict between 'academic success' and 'sense of belonging to the neighborhood'?"

Task: T7 · **Dataset:** Dataset 2

Answer: I view academic success as a primary route of escape and upward mobility, while simultaneously relying on the neighborhood's interpersonal ties as a source of recognition that stabilizes my self-worth.

Note: T7 items are expert-curated open-ended question–reference pairs; the released benchmark does not provide explicit evidence IDs for the target answer.

(a) Overall (Primary Backbones)

System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	59.9	66.0	44.4	40.5	36.1	14.3	16.3
+ Naive RAG	+8.8	+4.8	+3.9	+1.8	+2.7	+2.8	+1.7
+ Mem0 (Entity)	**+10.5**	**+9.2**	-3.1	+1.9	-0.2	+2.6	+1.2
+ MemOS	+4.5	+6.4	**+8.3**	**+6.6**	**+5.0**	**+3.9**	**+3.0**
• *Backbone: GPT-5-mini*							
Base Model (Abs.)	65.4	71.5	54.1	47.3	42.3	18.6	19.6
+ Naive RAG	+6.8	+4.0	+3.2	+1.5	+2.0	+2.7	+1.3
+ Mem0 (Entity)	**+7.8**	**+7.2**	-2.5	+1.4	+0.3	+2.1	+1.2
+ MemOS	+4.2	+5.1	**+6.7**	**+5.6**	**+5.2**	**+2.9**	**+3.0**

(b) Dataset 1 (Flashbacks)

System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	58.1	64.2	38.5	35.6	34.3	13.8	15.1
+ Naive RAG	+9.2	+4.7	+5.3	+2.5	+3.1	**+6.2**	+1.8
+ Mem0 (Entity)	+10.5	+8.6	-3.5	+1.2	-0.2	+0.9	+1.6
+ MemOS	**+10.7**	**+9.7**	**+10.4**	**+10.8**	**+4.2**	+3.4	**+4.1**
• *Backbone: GPT-5-mini*							
Base Model (Abs.)	62.3	70.8	48.2	41.4	40.4	17.7	17.6
+ Naive RAG	+7.5	+3.8	+4.5	+1.8	+2.4	**+3.0**	+0.9
+ Mem0 (Entity)	**+8.6**	**+6.7**	-1.7	+1.6	-0.4	+1.4	+1.0
+ MemOS	+8.4	+6.1	**+8.2**	**+7.0**	**+5.7**	+2.5	**+4.4**

(c) Dataset 2 (Event Dense)

System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	63.1	70.7	44.4	40.9	34.4	14.9	16.5
+ Naive RAG	+8.8	+5.7	+1.9	+1.1	+2.0	+0.8	**+2.9**
+ Mem0 (Entity)	**+9.3**	**+8.9**	-4.1	+1.7	-1.5	+2.5	-0.8
+ MemOS	+1.2	+3.8	**+7.3**	**+5.7**	**+5.3**	**+3.4**	**+3.0**
• *Backbone: GPT-5-mini*							
Base Model (Abs.)	68.5	73.0	55.4	49.8	42.5	19.6	20.3
+ Naive RAG	+5.7	+4.0	+0.8	+2.6	-0.1	+1.8	+1.0
+ Mem0 (Entity)	**+7.1**	**+6.2**	-3.2	+1.8	-2.1	+1.5	+0.2
+ MemOS	+2.5	+4.2	**+5.6**	**+5.6**	**+4.0**	**+3.0**	**+3.7**

(d) Dataset 3 (Mind)

System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	58.6	63.2	50.4	45.1	39.5	14.2	17.3
+ Naive RAG	+8.8	+3.7	+4.4	+1.6	+3.0	+1.2	+0.4
+ Mem0 (Entity)	**+11.5**	**+9.8**	-0.9	+2.7	+1.3	+4.3	**+2.7**
+ MemOS	+1.5	+6.6	**+7.0**	**+3.3**	**+5.6**	**+4.9**	+1.8
• *Backbone: GPT-5-mini*							
Base Model (Abs.)	65.3	71.1	58.6	51.2	44.1	18.5	20.9
+ Naive RAG	+7.2	+3.9	+3.0	-0.4	+3.5	+2.4	+0.7
+ Mem0 (Entity)	**+7.8**	**+8.3**	-2.5	+0.2	+0.1	**+3.5**	**+1.4**
+ MemOS	+1.8	+5.3	**+2.1**	**+3.6**	**+4.7**	+1.6	+0.9

Table 4: Per-dataset breakdowns for the two primary backbones. Panel (a) reports the overall comparison for Qwen3-32B and GPT-5-mini; panels (b)–(d) provide the dataset-level views referenced in the main-text analysis. Base rows show absolute scores, while the memory variants are reported as deltas relative to the corresponding base model.

E Additional Experimental Results

This appendix collects the supplementary result views that were moved out of the main paper for space: the primary-backbone breakdowns used in the dataset-level analysis, and backbone-specific radar plots for the full evaluation set.

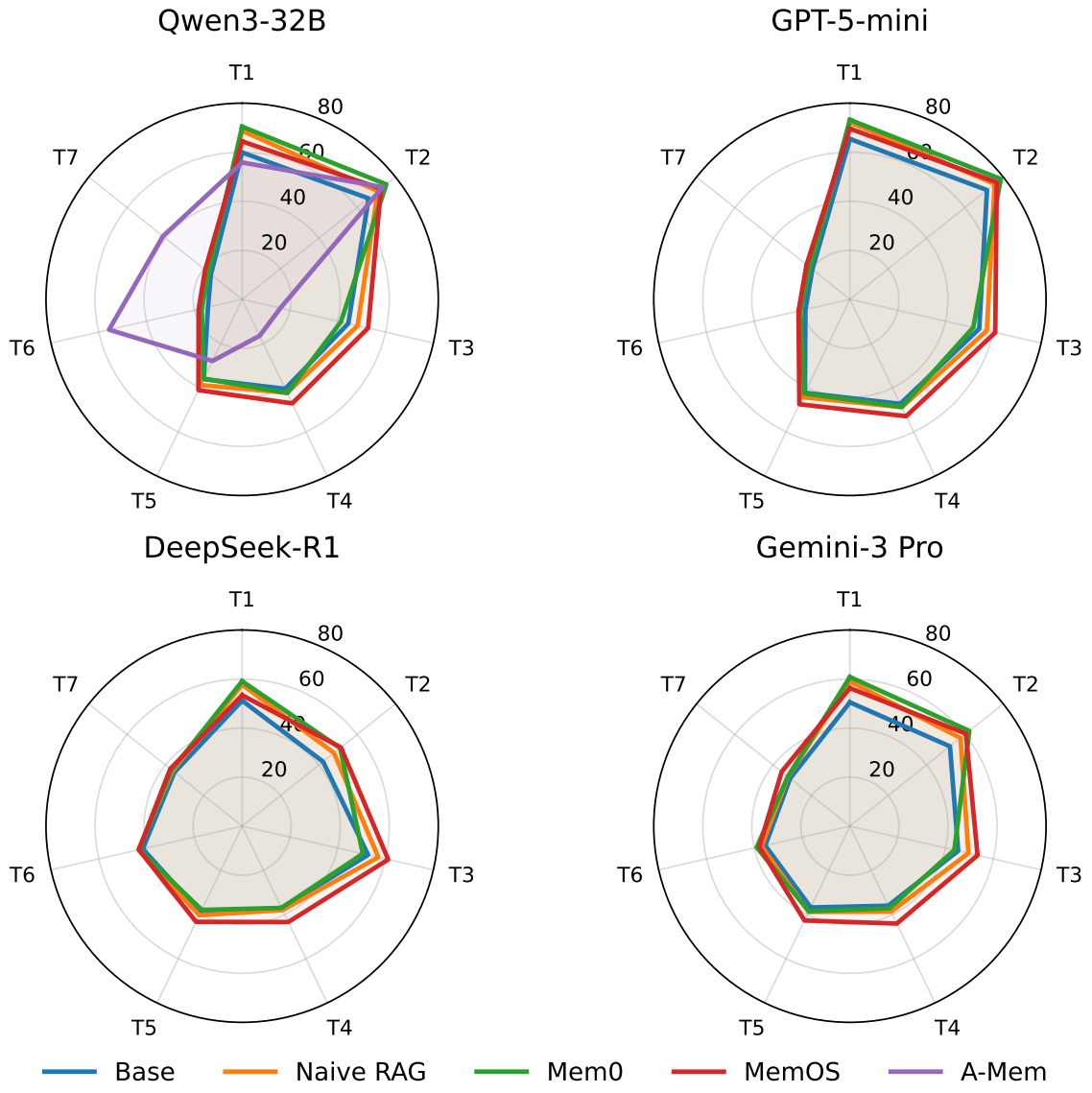

Figure 4: Backbone-specific radar plots for the overall evaluation results. All four backbone families include Base, Naive RAG, Mem0, and MemOS; Qwen3-32B additionally includes A-Mem. The figure makes the same pattern visible as Table 2: retrieval-heavy variants help T1–T2, while MemOS yields broader gains on T3–T5.

F Human Evaluation Details

Level	Human	Best Model
I: Fact & Memory	96.5	75.4
II: Logic & Causality	88.0	62.5
III: Insight	83.5	22.6

Table 5: Human versus best-model performance aggregated by evaluation level. Human answers are graded by the same blinded LLM-as-a-Judge pipeline used for model outputs.

Annotator demographics and ethics. The human study used three expert annotators with M.A.-level training in literature or linguistics. All annotators were compensated at a reasonable rate consistent with standard expert-annotation practice. The study was designed as an expert reading and evidence-grounded question-answering task rather than open-ended crowd annotation.

Evaluation materials and interface. For each item, annotators were shown the reconstructed cognitive stream used by the benchmark. To preserve global narrative context and avoid penalizing humans for missing book-level structure, they could also consult the original novel text at any time during evaluation. This setup ensured that annotators had access both to the benchmark representation and to the full narrative evidence from which it was derived.

Instructions and scoring pipeline. Annotators were instructed to answer each question only with claims supported by textual evidence. They did not score one another's outputs. Instead, their written answers were passed through the same level-specific LLM-as-a-Judge pipeline used for model responses so that human and model performance were evaluated under an identical rubric. During scoring, the judge was blind to whether an answer came from a human annotator or from a model system.

Agreement interpretation. Under this protocol, the relevant reliability quantity is the alignment between the rubric-based LLM judge and expert consensus on the subjective subset, rather than pairwise agreement among humans grading one another. The reported $\kappa > 0.75$ therefore reflects strong alignment between the judge's scoring behavior and expert consensus, which is why we do not additionally report a separate table of pairwise human–human Cohen's Kappa values.