KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions

Anonymous ACL submission

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 for person understanding.

(G1) Evaluation misalignment: retrieval proxies $\neq$ person understanding. Most benchmarks focus on retrieval, temporal ordering, knowledge updates, and conflict handling across sessions (Wu et al., 2025; Maharana et al., 2024a; Hu et al., 2025a; Castillo-Bolado et al., 2024; Tan et al., 2025). 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 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

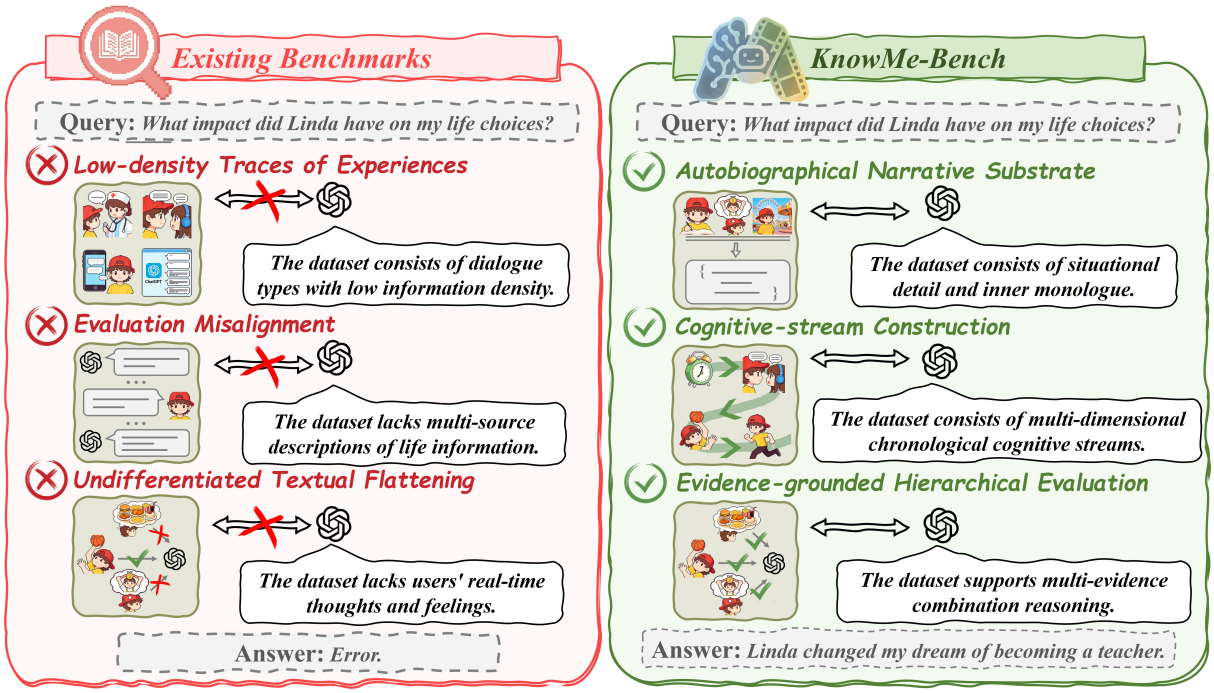


Figure 1: Comparison of information density and reasoning capabilities between existing benchmarks and KnowMe-Bench. The left panel illustrates the limitations of existing benchmarks, which rely on low-density traces (sparse dialogues) and suffer from undifferentiated textual flattening (lacking real-time inner thoughts), often leading to reasoning errors in complex queries. In contrast, KnowMe-Bench (right panel) utilizes an autobiographical narrative substrate rich in situational detail and inner monologue. By employing cognitive-stream construction and evidence-grounded hierarchical evaluation, it effectively models multi-dimensional life experiences, enabling the model to deeply research long-term impacts.

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 designed to evaluate *evidence-grounded person-model inference* from long-form autobiographical experience.

(M1) Autobiographical narrative substrate (addresses G2a). KnowMe-Bench is built from publicly releasable autobiographical narratives that preserve the coupling among actions, dialogue, situational detail, and inner monologue, providing evidence suitable for inferring motivations, principles, and affective triggers (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 gold 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; and
- **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). However, these evaluations predominantly treat memory as a database of explicit facts, prioritizing retrieval precision over interpretative reasoning. Current benchmarks often overlook autobiographical reasoning, where an agent must infer implicit information, such as stable principles or psychological 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 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 B) to prevent hallucination and maintain strict adherence to the source text.

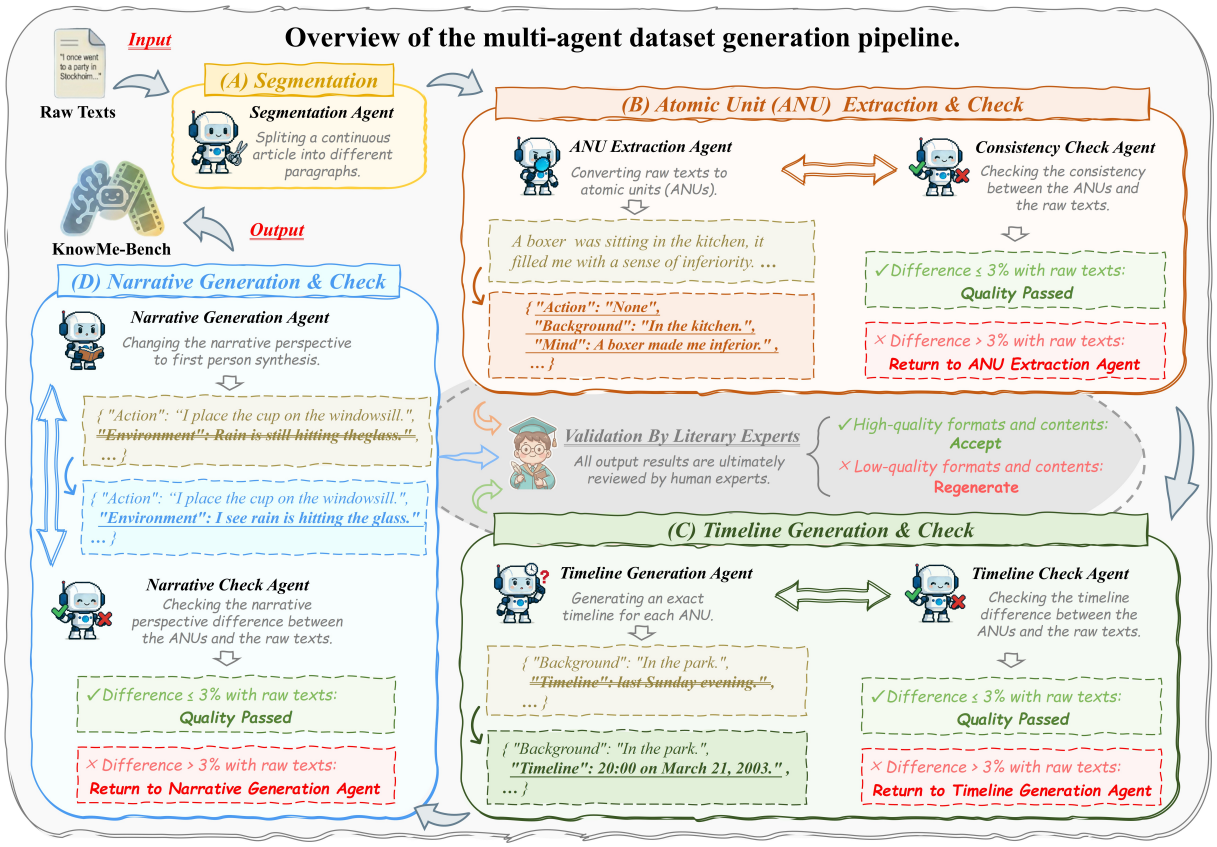


Figure 2: **Overview of the multi-agent dataset generation pipeline.** The framework transforms unstructured raw narratives into the structured KnowMe-Bench benchmark through four sequential stages: (A) Segmentation, (B) Atomic Unit (ANU) Extraction, (C) Timeline Generation, and (D) Narrative Generation. 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.

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 noise-free 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 verbatim 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 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 are detailed in Appendix C).

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" → "I see rain hitting glass"). Final validation is conducted by human literary experts to ensure the dataset serves as a gold standard, 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 the completeness of entity recall under strict spatiotemporal constraints.
- **Task 2: Adversarial Abstention (Hallucination Test).** Uses "Mismatching Trap" queries to test the model's ability to refuse answering when causality or entities are distorted.
- **Task 3: Temporal Reasoning.** Assesses mastery of the timeline structure, including duration calculation and distinguishing chronological order from narrative presentation order.

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.** Requires ordering discrete events based on non-temporal semantic dimensions (e.g., escalation of danger) rather than explicit timestamps.
- **Task 5: Mnestic Trigger Analysis.** Evaluates the understanding of stream-of-consciousness, specifically identifying the sensory cues or associative triggers that evoke memory retrieval.

4.3 Level III: Psychoanalytic Depth (The "Insight" Layer)

The most challenging tier targets subtext and human psychology.

- **Task 6: Mind-Body Interaction.** Explores the duality between external actions and internal states, requiring explanations for ironic or contradictory behaviors.
- **Task 7: Expert-Annotated Psychoanalysis.** Open-ended questions curated by literary experts regarding complex motivations and identity construction, serving as the ceiling for deep literary 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 the task type:

- **Factual Tasks (T_1, T_2, T_3):** The judge evaluates **Entity Accuracy** and **Value Precision**. For T_2, full marks are awarded strictly for correct abstention.
- **Logic Tasks (T_4, T_5):** The judge evaluates **Sequence Correctness** and the **Validity of Reasoning** (i.e., whether the model identifies the correct causal trigger).
- **Insight Tasks (T_6, T_7):** The evaluation focuses on the **External-to-Internal Mapping**. A high score (5/5) is granted only if the model captures the specific **core metaphors** (e.g., "dissolving boundaries") defined 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. We validated this protocol via a human alignment study, achieving a Cohen's Kappa of $\kappa > 0.75$ between the LLM Judge and human experts on a subset of data.

5 Experiments

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

5.1 Experimental Setup

Datasets and Narrative Modalities. We utilize the full KnowMe-Bench corpus (4.7M tokens) comprising three structurally diverse datasets. To ensure robustness, we generated a total of **2,580 evaluation queries**[cite: 241].

- **Dataset 1 (Flashback-Intensive):** Knausgård's *My Struggle* (1.15M tokens). Tests handling of non-linear time and mnestic triggers.
- **Dataset 2 (Event-Driven):** *Neapolitan Novels* (1.76M tokens). Tests linear causal tracking and high-frequency entity updates.
- **Dataset 3 (Psychological Depth):** Proust's *In Search of Lost Time* (1.30M tokens). Tests interpretation of abstract internal monologues.

De-identification & Ethics. We applied a rigorous privacy pipeline (see Appendix for details) to remove PII.

Model Architecture & Baselines. Our evaluation setup distinguishes between the *Inference Model* (responsible for generation and reasoning) and the *Embedding Model* (responsible for vector retrieval).

- **Inference Models:** We employ Qwen3-32B (Long-Context) and GPT-5.1-mini as the generation engines.
- **Comparison Systems:**
  - **Naive RAG ($k = 50$):** Standard retrieval baseline using dense vector similarity.
  - **Mem0 (Entity-Memory):** Represents state-of-the-art entity tracking systems[cite: 959], which maintain a dynamic graph of user facts.
  - **MemOS (Log-based Open Source):** An open-source stream-based cognitive architecture[1] designed to handle temporal linearity and conflict resolution via chronological logging.

5.2 Main Results

Table 1 details the performance breakdown across datasets. Figure 2 visualizes the capability trade-offs between architectures.

5.3 In-Depth Analysis & Discussion

We expand our analysis to four core dimensions to reveal the deeper mechanical differences in how memory architectures process long-horizon complex narratives.

1. The "Update Paradox" in Non-Linear Narratives (Dataset 1). Table 1(b) reveals a structural deficiency in existing state-updating memory systems when handling non-linear narratives (e.g., flashbacks).

- **State Overwriting:** Mem0 performs poorly on the **Temporal Logic (Level II)** tasks of Dataset 1 (e.g., in Qwen3-32B, T3 shows a performance regression of **-3.5%**). This is because Mem0 aims to maintain a "current user state graph." Flashbacks (e.g., "I liked apples as a child") are often misparsed as updates to the current state, overwriting the true present state (e.g., "I now hate apples").
- **Stream Architecture Advantage:** In contrast, MemOS, which employs a log-based cognitive stream, achieves massive gains in T3 (Temporal Reasoning, **+10.4%**) and T4 (Relational Logic, **+10.8%**). This proves that preserving **chronological integrity** is more critical than maintaining a static entity graph for long-term companion scenarios.

2. The Trade-off Between Precision and Insight (Dataset 2 vs. Dataset 3). Comparing Table 1(c) and 1(d) highlights a clear trade-off between "factual retrieval" and "deep reasoning."

- **Entity Advantage in High-Density Contexts:** In Dataset 2 (Neapolitan Novels), characterized by dense physical entities and complex relationships, Mem0 demonstrates dominance, boosting **Entity Consistency (T2)** by up to

[1] https://usememos.com/

Table 1: Performance breakdown across different datasets. **(a)** Aggregate results; **(b)** Dataset 1 (Knausgård); **(c)** Dataset 2 (Ferrante); **(d)** Dataset 3 (Proust).

(a) Overall Performance

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-5.1-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)

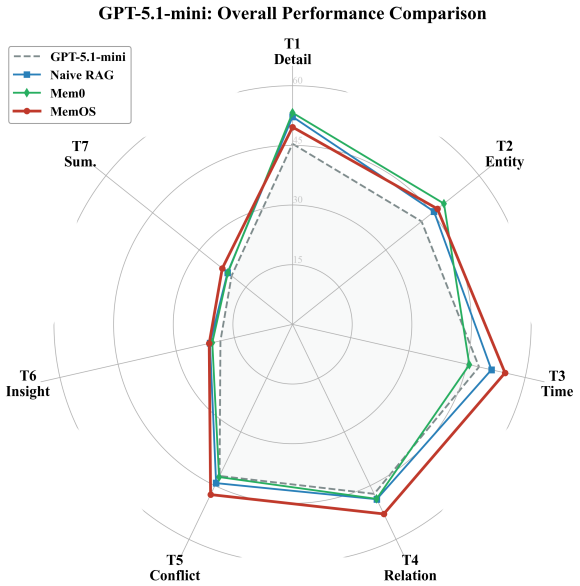
System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	58.1	64.2	38.5	35.6	34.3	13.8	15.1
+ Naive RAG	+9.2	+4.7	+5.3	+2.5	+3.1	**+6.2**	+1.8
+ Mem0 (Entity)	+10.5	+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-5.1-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	+2.3	+2.3	**+5.5**	+1.4
+ Mem0 (Entity)	+8.0	+7.9	-2.6	+1.0	+0.4	+0.5	+1.3
+ MemOS	**+10.7**	**+8.1**	**+8.2**	**+9.0**	**+5.6**	+3.1	**+3.9**

(c) Dataset 2 (Event Dense)

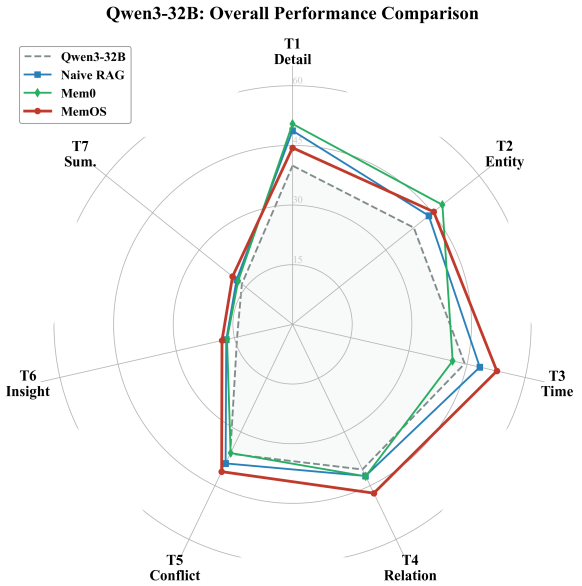
System	Level I: Fact & Entity			Level II: Temporal Logic		Level III: Insight	
	T1 (Det.)	T2 (Ent.)	T3 (Time)	T4 (Rel.)	T5 (Con.)	T6 (Abs.)	T7 (Sum.)
• *Backbone: Qwen3-32B*							
Base Model (Abs.)	66.8	71.2	52.2	45.0	42.3	15.6	18.7
+ Naive RAG	+7.9	+4.5	+2.4	+1.1	+1.3	+2.0	+1.3
+ Mem0 (Entity)	**+12.9**	**+11.8**	-2.3	+3.5	+0.8	+1.2	+1.1
+ MemOS	+2.3	+4.6	**+7.6**	**+5.1**	**+6.2**	**+3.1**	**+2.3**
• *Backbone: GPT-5.1-mini*							
Base Model (Abs.)	72.3	75.4	62.5	52.6	48.1	20.4	22.3
+ Naive RAG	+5.8	+3.5	+1.9	+1.1	+1.2	+2.1	+1.0
+ Mem0 (Entity)	**+9.2**	**+8.4**	-1.8	+2.5	+0.5	+1.0	+1.2
+ MemOS	-0.6	+3.9	**+5.5**	**+4.8**	**+4.5**	**+2.0**	**+2.2**

(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.)	55.4	62.7	45.9	38.5	35.2	12.7	14.6
+ Naive RAG	**+9.2**	+5.1	+3.4	+2.2	+2.8	-0.5	+2.2
+ Mem0 (Entity)	+8.0	**+7.3**	-3.2	+0.5	-1.1	**+7.4**	+0.8
+ MemOS	-1.2	+4.8	**+6.3**	**+5.5**	**+6.2**	+5.8	**+2.0**
• *Backbone: GPT-5.1-mini*							
Base Model (Abs.)	62.5	68.3	55.3	45.2	42.1	16.5	18.5
+ Naive RAG	**+7.0**	+4.8	+2.6	+1.4	+2.1	-0.2	+1.8
+ Mem0 (Entity)	+6.2	**+5.2**	-2.9	+0.3	-0.4	**+6.2**	+1.0
+ MemOS	+0.5	+3.4	**+5.7**	**+4.2**	**+4.8**	+4.1	**+2.7**



Qwen3-32B Performance Profile



GPT-5.1-mini Performance Profile

Table 2: Radar charts illustrating the trade-offs between memory architectures. Mem0 excels in entity-heavy tasks (T1, T2), while MemOS dominates temporal and insight tasks (T3-T7). The gap highlights the specific strength of graph-based memory in explicit tracking versus stream-based memory in narrative reconstruction.

+11.8%. Explicit entity graphs effectively solve coreference resolution and multi-hop retrieval, which Naive RAG struggles with.

- **Retrieval Failure in Insight:** However, in Dataset 3 (Proust), which focuses on psychological monologues, gains across all systems

narrow significantly. Notably, while Naive RAG boosts T1 (Fact Extraction) by **+9.2%**, it causes a **-0.5%** performance drop in **Insight (T6)** tasks on Qwen3-32B.

- **Analysis:** Level III insight relies on perceiving subtext and long-term emotional shifts, not keyword matching. RAG often retrieves semantically similar but contextually irrelevant fragments ("context pollution"), interfering with the model's coherent modeling of the persona's inner world. This validates our hypothesis **G1**: retrieval proxies are not equivalent to true person understanding.

3. Backbone Sensitivity: Scale vs. Architecture. Cross-model comparison (Qwen3-32B vs. GPT-5.1-mini) reveals the interaction between base model capability and external memory modules:

- **Diminishing Returns of Memory Modules:** On the weaker base model (Qwen3), introducing external memory (especially Mem0) yields significant gains in **Factual Extraction (T1)** (rising from 59.9% to 70.4%). On the stronger GPT-5.1, the relative gain from the same module is smaller. This suggests that powerful base models with better long-context handling partially mask the deficiencies of memory systems.

- **Complex Reasoning Remains a Bottleneck:** Despite GPT-5.1's stronger foundation, on Level III (Insight) tasks, even with the best memory system, the absolute score remains capped at 22.3% (Dataset 2, MemOS+GPT). This low ceiling indicates that neither current RAG nor Graph Memory effectively supports high-order psychodynamic reasoning, a critical gap for future research.

4. Adversarial Robustness and Hallucination Mitigation (Task 2). Experimental results from Task 2 (Adversarial Abstention) provide key evidence for system safety.

- **Conflict Detection Mechanism:** Systems with explicit memory structures (Mem0/MemOS) consistently outperform Naive RAG and the Base Model on T2. Specifically, Mem0 utilizes its structured User Profile to cross-reference new inputs (Mismatching Traps) against established facts.

- **RAG's "Forced Answer" Tendency:** Naive RAG systems tend to force an answer if they retrieve partially relevant keywords, even if the logic is flawed, leading to higher hallucination rates. Statistically, Mem0 outperforms Naive RAG on Dataset 2's T2 task by **7.3 percentage points** (+11.8 vs. +4.5), proving the role of structured knowledge as a "grounding anchor" in preventing long-horizon interaction hallucinations.

6 Conclusion

In this work, we introduced **KnowMe-Bench**, a comprehensive 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 established a data substrate that preserves the "micro-texture" of human experience—integrating actions, inner thoughts, and environmental context.

Our experiments reveal a critical "evaluation gap" in current long-horizon memory research. While retrieval-augmented baselines and entity-tracking systems (like Mem0) demonstrate competence in factual recall (Level I), they exhibit significant structural fragility when facing the non-linear temporal dynamics characteristic of human memory. Specifically, the identification of the **"Update Paradox"**—where systems fail to distinguish between a *mnemonic trigger* in the present and a *state change* in the past—highlights the limitations of treating memory as a static database rather than a chronological cognitive stream. Furthermore, the low performance across all models on "Insight" tasks (Level III) confirms that high retrieval accuracy does not equate to a working model of a user's motivations, values, or psychological interiority.

KnowMe-Bench provides the necessary tooling to bridge this gap, offering a multi-agent pipeline for "mnestic realignment" and a hierarchical evaluation suite. We hope this resource encourages the research community to move beyond context-window extension and vector similarity, fostering the development of cognitive architectures capable of genuine empathy and deep reasoning over the lived experience of their users.

Limitations

Methodologically, the benchmark itself must navigate the inherent subjectivity of literary analysis by relying on a rigorous "LLM-as-a-Judge" protocol validated by human experts , and it faces the high cost and complexity of a multi-agent generation pipeline required to produce and de-identify high-density 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. *arXiv 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*.

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.

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*, volume 39, pages 25065–25073.

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*, volume 38, pages 19724–19731.

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 Appendix 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:

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 Appendix 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, we use $\epsilon = 0.03$ as a strict acceptance threshold in our implementation.

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 $\delta > 0.03$ (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 > 0.03 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 a fixed abstention token (denoted `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).
```