CloneMem: Benchmarking Long-Term Memory for AI Clones

Sen Hu[1,†,*], **Zhiyu Zhang**[2,†], **Yuxiang Wei**[3,†],
Xueran Han[4], **Zhenheng Tang**[5], **Huacan Wang**[6], **Ronghao Chen**[1]

[1]Peking University [2]UC Davis [3]Georgia Tech [4]MBZUAI [5]HKUST [6]UCAS

[†]Equal contribution [*]Corresponding author: husen@pku.edu.cn

Abstract

AI Clones aim to simulate an individual's thoughts and behaviors to enable long-term, personalized interaction, placing stringent demands on memory systems to model experiences, emotions, and opinions over time. Existing memory benchmarks primarily rely on user–agent conversational histories, which are temporally fragmented and insufficient for capturing continuous life trajectories. We introduce CLONEMEM, a benchmark for evaluating long-term memory in AI Clone scenarios grounded in non-conversational digital traces, including diaries, social media posts, and emails, spanning one to three years. CLONEMEM adopts a hierarchical data construction framework to ensure longitudinal coherence and defines tasks that assess an agent's ability to track evolving personal states. Experiments show that current memory mechanisms struggle in this setting, highlighting open challenges for life-grounded personalized AI. Code and dataset are available at https://github.com/AvatarMemory/CloneMem

1 Introduction

In recent years, the rapid development of large language models (LLMs) has given rise to a wide range of personalized applications, such as role-playing agents and AI companions that mimic specific personalities (Tseng et al., 2024). These applications reveal an underlying trend: users are increasingly seeking to establish deep, personalized connections with AI systems (Chen et al., 2024a). In light of this, researchers have begun to explore a more advanced paradigm— AI Clones, which aims to faithfully simulate a specific individual's thoughts, behaviors, and emotional tendencies over long-term interactions (Lee et al., 2026; Wei et al., 2025).

Building a genuine AI Clone relies on two complementary pillars: *Persona Simulation* and *Long-Term Memory*. While current research on persona simulation excels at imitating static character profiles and linguistic styles (Wang et al., 2025b, 2026), human personality is dynamically shaped by lived experiences(Roberts and Mroczek, 2008; Dugan et al., 2023). Without a robust memory system, an AI Clone remains a static imitator rather than an evolving entity. To advance from static simulation to authentic cloning, AI systems must continuously capture individual experiences, emotional trajectories, and opinion formation.

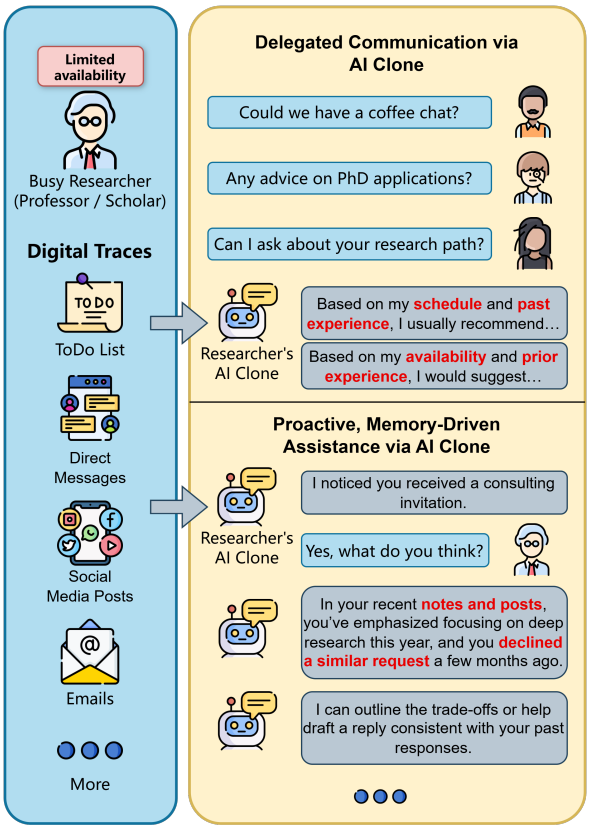

Figure 1: Illustrative application scenarios of an AI Clone grounded in long-term digital traces.

Recent studies on long-term memory systems and their benchmarks have focused mainly on the conversation-based paradigm, where memory is constructed exclusively from user-agent conversa-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33571–33602
July 2-7, 2026

Benchmark	Source	Scenario	Time Span	Temporal Structure	Consistency
LoCoMo (Maharana et al., 2024)	CH	Chatbot	Several months	sessions	Persona & timeline
LongMemEval (Wu et al., 2025)	CH	Chatbot	∼2.5 yrs	Multi-session	Factual
PersonaMem (Jiang et al., 2025)	CH	Chatbot	Several years	Snapshots	Preference
CLONEMEM (Ours)	**DT**	**AI clone**	**1∼3 yrs trajectory**	**Life arcs**	**Exp. + Emo. + Opi. (longitudinal)**

Table 1: Comparison of long-term memory benchmarks for personalized AI systems. CH = *conversation history*; DT = *non-conversational digital traces*; Exp. = Experience; Emo. = Emotion; Opi. = Opinion.

tional histories (Xu et al., 2025; Chhikara et al., 2025; Wu et al., 2025). Although conversational histories can record user preferences and maintain coherence within dialog, they only capture fragmented slices of an individual's life. To faithfully model a complete person including his evolving experiences, emotions, and opinions, a richer and more continuous data source is required.

One alternative is non-conversational digital traces that users generate naturally in everyday life, such as diaries, media posts, direct messages, and emails (see the left panel of Figure 1). These traces form a naturally occurring longitudinal record of daily activities (Azucar et al., 2018), providing the continuous temporal structure necessary to track how an individual's experiences, emotions, and opinions evolve over time.

Nevertheless, evaluating AI Clone's ability to leverage these traces remains unaddressed. As summarized in Table 1, existing benchmarks assess memory capabilities predominantly through conversational histories. Most of them construct user information as isolated data points, where facts or preferences are generated independently without forming a coherent life trajectory(Jiang et al., 2025; Wu et al., 2025). Other benchmarks such as LoCoMo (Maharana et al., 2024) incorporate temporal event structures, yet their temporal organization is not explicitly modeled as continuous longitudinal trajectories (life arcs), but remains fragmented across sessions. In addition, these benchmarks evaluate isolated aspects of consistency—such as persona, factual, or preference consistency—rather than jointly assessing how experiences, emotions, and opinions evolve over time.

To bridge this gap, We introduce CLONEMEM, a comprehensive benchmark for evaluating long-term memory capabilities of AI Clone. CLONEMEM tests whether an AI Clone can integrate non-conversational digital traces drawn from everyday life and use them to consistently track an individual's experiences, emotional changes and evolving opinions over time.

Our main contributions are summarized as follows.

- We introduce CLONEMEM, the first benchmark for evaluating long-term memory in AI Clones. Unlike prior benchmarks based on conversational histories, CLONEMEM uses non-conversational digital traces as its primary data source, enabling evaluation of how AI systems track an individual's experiences, emotions, and opinions over time.
- We propose a top-down data synthesis pipeline that first generates macro-level life arcs, then derives fine-grained daily traces, ensuring longitudinal coherence across extended time spans.
- We design several evaluation tasks to test whether AI Clone can use non-conversational digital traces to track changes in an individual's experiences, emotions, and opinions over time.
- Our experiments reveal that existing memory systems struggle in AI Clone scenarios: simple flat retrieval outperforms consolidation-based methods, and models often fall back on generic narrative templates rather than grounding responses in retrieved evidence.

2 Related Work

AI Clone and Persona Simulation. Existing works on persona simulation aim to enable LLMs to adopt and maintain specific personality traits in conversation, spanning from fictional characters to real individuals (Chen et al., 2024b,a). For fictional characters, prior work has extensively studied character-level role-playing (Wang et al., 2024; Shao et al., 2023; Wang et al., 2025a;

Zhang et al., 2025; Samuel et al., 2025; Abdulhai et al., 2025), focusing primarily on the expression layer—whether an agent sounds and behaves like its assigned character through style imitation and persona consistency (Yu et al., 2026). Similarly, personalization benchmarks such as LaMP (Salemi et al., 2024) evaluate preference prediction or stylistic generation conditioned on past interactions, where the user profile serves mainly as a statistical signal rather than a source of specific retrievable experiences.

More recently, research has shifted toward simulating real individuals using authentic personal data. Guo et al. (2026) evaluates individual simulation using a volunteer's decade-long messaging history, while Du et al. (2025) benchmarks persona simulation across social, interpersonal, and narrative dimensions. However, these works either evaluate end-to-end fidelity without isolating memory, or rely on conversational data that does not capture multi-year life trajectories. CLONEMEM complements this line of work by specifically benchmarking the memory layer—whether an AI Clone system can track an individual's evolving experiences, opinions, and emotions from long-term, non-conversational digital traces.

Long-Term Memory System. AI Clones require memory systems that integrate and retrieve personal experiences beyond the context window. Retrieval-based methods (Lewis et al., 2020) treat past experiences as static chunks indexed by embedding similarity. Consolidation-based methods actively maintain evolving representations through memory decay (Zhong et al., 2024; Tang et al., 2026), LLM-driven updates(Chhikara et al., 2025), or dynamic linking (Xu et al., 2025), while scheduling-based approaches such as MemGPT(Packer et al., 2024) and MemoryOS (Kang et al., 2025) manage information flow between context and external storage. Graph-based systems such as HippoRAG (Gutierrez et al., 2024) and Zep (Rasmussen et al., 2025) introduces relational structures for multi-hop reasoning. Beyond structural organization, Generative Agents (Park et al., 2023) further incorporates temporal signals into retrieval, but its memory stream is designed for short-horizon sandbox interactions rather than personal histories spanning years of real-world digital traces.

Non-conversational digital traces are first-person, temporally situated, and expressed indirectly, creating a semantic gap that challenges embedding-based retrieval. Moreover, consolidation methods compress trajectories into factual snapshots, discarding the temporal ordering and internal-state transitions essential for faithful personal modeling. Our benchmark evaluates representative systems from these paradigms to quantify these limitations.

Long-term Memory Benchmarks. A range of benchmarks have been proposed to evaluate long-term memory in LLM-based systems. Conversational memory benchmarks such as LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2025), PersonaMem (Jiang et al., 2025), and MemBench (Tan et al., 2025) assess factual recall, temporal reasoning, and preference consistency over multi-session dialogue histories, with recent work extending to project-oriented interaction (Bian et al., 2026) and operation-level hallucination diagnosis (Chen et al., 2026). KnowMe-Bench (Wu et al., 2026) further evaluates person understanding over long-form autobiographical narratives rather than dialogue logs. Long-context benchmarks such as LongBench (Bai et al., 2025) and RULER (Hsieh et al., 2024) evaluate comprehension under extremely long inputs but do not target personalized memory.

Despite this progress, existing benchmarks uniformly derive memory from conversational histories or static documents. None evaluates whether a system can model an individual's evolving experiences, emotions, and opinions from non-conversational digital traces — the naturally occurring, longitudinal records of everyday life. CLONEMEM addresses this gap by grounding evaluation in multi-source digital traces spanning one to three years, with tasks that require trajectory-level reasoning rather than isolated fact retrieval.

3 Data Construction for CLONEMEM

Moving beyond conversational histories to non-conversational digital traces alters what needs to be modeled. The resulting data form a temporally extended record of an individual's life rather than a set of isolated utterances. As a result, the modeling focus shifts from detecting whether emotions or opinions change to understanding how such changes arise from prior experiences, which requires experiences, emotions, and opinions to remain coherent over a long time span. To address this, we adopt a hierarchical generation framework (Fan et al., 2018), where high-level life arcs give rise

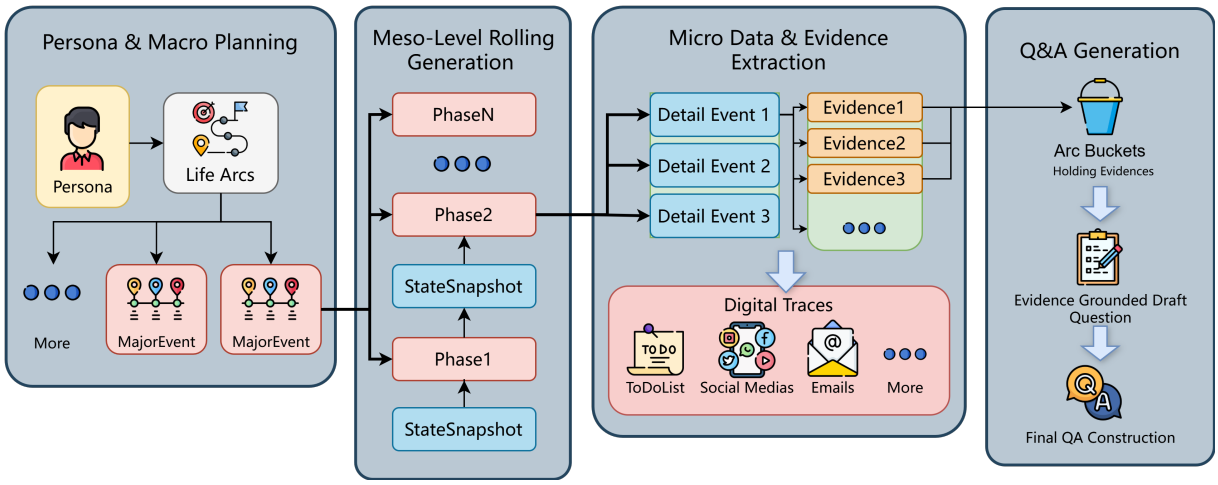

Figure 2: **Overview of the Avatar Memory Benchmark data construction and evaluation pipeline.** The pipeline proceeds from persona, macro-level life arcs and major events, through meso-level rolling generation of phases with state snapshots, to micro-level detailed event simulation, evidence extraction and final generation of non-conversational digital traces. Extracted evidence is then organized into arc-specific buckets to support evidence-grounded question–answer generation. A detailed version of the pipeline is provided in Appendix A.

to a structured sequence of major events, phases, and detailed events that define a long-term personal narrative. Digital traces are generated as non-conversational artifacts grounded in this structure (Figure 2).

3.1 Hierarchical Event and Trace Generation

We organize data generation as a three-level hierarchical process, moving from long-term life planning to fine-grained daily digital traces.

Persona and Macro-Level Life Arcs. We first construct a psychologically coherent persona based on the five Big Five personality traits (Goldberg, 2013). Given this persona, we specify a small set of macro-level life arcs that describe long-term trajectories in experiences, emotions, and opinions. Each life arc is anchored by a coarse sequence of major life events, which provide temporal milestones without fixing detailed outcomes. As a result, life arcs and major events define a global structure that constrains downstream generation.

Meso-Level Phase Generation. At the meso-level, major life events are refined into sequences of phases that mediate between long-term life arcs and daily digital traces. To maintain longitudinal coherence, we associate each phase with a persistent internal state that captures the evolving emotional and attentional context of the persona. This state provides continuity across phases, ensuring that emotional and opinion changes reflect accumulated experiences rather than appearing as isolated signals. By conditioning phase generation on the current internal state, earlier phases are able to influence how subsequent phases develop. (Details of the phase generation process are provided in Appendix A.3)

Micro-Level Digital Trace Generation. At the micro level, we generate fine-grained daily digital traces that serve as direct inputs to an AI Clone. These traces consist of non-conversational artifacts, including diaries, social media posts, and direct messages. For each phase, we introduce a set of detailed events that represent the concrete activities and experiences of the persona. Both digital traces and explicit evidence are generated with reference to these detailed events, with the evidence encoding the persona's experiences, emotional states, and expressed or implied opinions. This evidence serves as a grounding constraint during trace generation, ensuring consistency between the resulting traces and the persona's underlying experiences and internal states (See Appendix A.4).

3.2 Evaluation Question Construction.

Based on the generated digital traces and their associated evidence, we construct question–answer (QA) instances to evaluate whether an AI Clone can track how experiences, emotions, and opinions evolve over time.

Each QA instance is grounded in evidence tied to a specific life arc and reflects a localized temporal span, allowing questions to capture both recent

developments and relevant prior context.

QA instances are generated only when the available evidence is sufficient to support a given question category. This evidence-gated construction ensures that each question is well-supported, temporally grounded, and evaluates progression within a life arc rather than isolated facts (See Appendix A.5)..

3.3 Data Quality Control

We apply a post-processing pipeline that combines automated filtering with limited human review. Automated checks first remove QA instances with formatting errors or missing fields. We then perform an evidence sufficiency check, discarding questions whose referenced evidence does not adequately support the corresponding answers. This scripted filtering step removes approximately 10% of the generated questions.

After automated filtering, a small subset of the remaining QA instances is reviewed by human annotators to verify plausibility and consistency with the underlying digital traces.

4 CLONEMEM Evaluation Benchmark

Based on the generated digital traces and ground-truth evidence, we introduce the CLONEMEM benchmark, which features a series of evaluation tasks designed to assess an AI Clone's ability to comprehend and reason about an individual's long-term, evolving life story. Unlike benchmarks that primarily evaluate memory through fact-oriented question answering over conversational history (Maharana et al., 2024; Wu et al., 2025), our tasks are specifically designed to test the ability of AI Clone to track the trajectory of and reasons behind changes in an individual's experiences, emotions, and opinions over time.

In CLONEMEM, each question is associated with a set of evidence units that specify the semantic unit required to answer correctly. Digital traces are generated from these evidence units and constitute the retrieval space to the AI Clone. Under this formulation, evidence units and digital traces form a many-to-many relationship: a single evidence unit may appear across multiple traces, while a single trace may encode information from multiple evidence units. Therefore, Evaluation in CLONEMEM is defined over retrieval from digital traces and reasoning with respect to evidence units.

Statistic	Value
# Personas	10
# Questions	1,183
Languages	English, Chinese
Context Length	3 short (∼100k tokens), 7 long (>500k tokens)
Question Types	8 task categories (See Section 4.2)

Table 2: Overview of dataset statistics.

4.1 Dataset Statistics

We generate a bilingual data set (English and Chinese) based on 10 distinct personas. The dataset comprises approximately 5,000 question-answer pairs, with a significant portion of the data designed to test long-context understanding. Specifically, 7 of the 10 personas have context lengths exceeding 500k tokens (some reaching up to 1M), while the remaining 3 are around the 100k token level. Table 2 provides an overview of the dataset statistics. More detailed breakdowns are reported in Appendix C.

4.2 Evaluation Tasks

CLONEMEM is provided in both free-text and multiple-choice question answering formats, with the latter framing evaluation as a classification task. Details of multiple-choice option generation are provided in Appendix A.5.

To reflect realistic long-context interactions, all questions are posed from the perspective of a close friend at a specific point in time. Questions often include conversational anchor points that require the AI Clone to locate relevant information within long-term, non-conversational digital traces before answering.

Figure 3 illustrates representative CLONEMEM tasks. In general, evaluation tasks are organized around three levels of reasoning: (1) **factual recall** at specific time points, (2) **temporal reasoning** about changes, trajectories, and patterns, and (3) **higher-level reasoning** involving causality, counterfactuals, abstraction, and unanswerable cases.

The formal task definitions and the generation procedures for each type of task are provided in Appendix B and Appendix A.

5 Experimental Setup

We evaluate memory-augmented systems on the CLONEMEM benchmark using a standardized retrieval-based evaluation framework. To ensure a fair comparison, all evaluated methods operate on the same long-term, non-conversational digi-

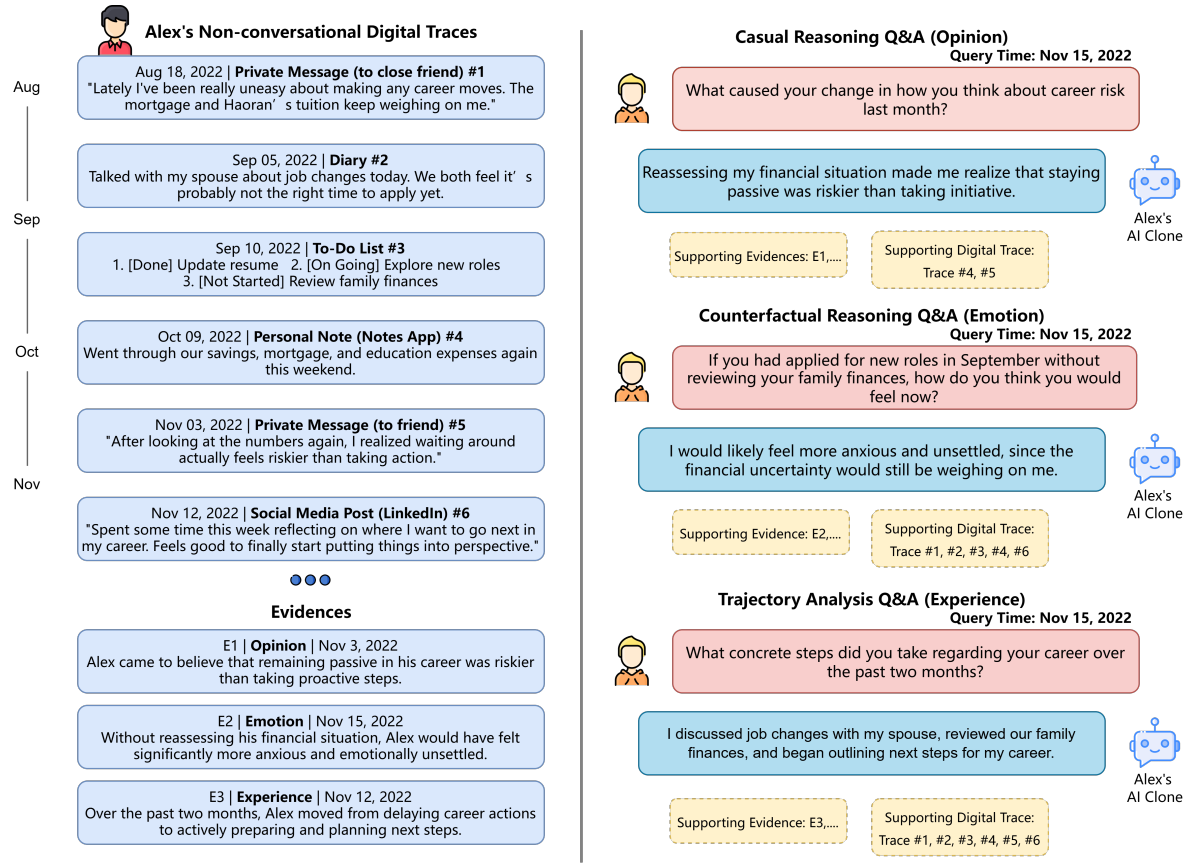

Figure 3: **Illustrative and representative examples of CLONEMEM tasks.** The left panel shows non-conversational digital traces and their associated ground-truth evidence generated during data construction; the right panel shows example questions and answers for three task types.

tal traces and share a unified pipeline for memory construction, retrieval, and response generation.

Implementation Details. Our experiments utilize two backbone language models, LLaMA-3.1-8B (Grattafiori et al., 2024) and GPT-4o-mini, and two embedding models, Contriever (Izacard et al., 2022) and text-embedding-3-small. Digital traces are indexed as retrievable memory units using embedding-based similarity search.

At inference time, the system retrieves relevant items based on embedding similarity with a retrieval depth of $k \in \{5, 10, 20\}$. The backbone models then generate answers based on the retrieved context. For evaluation, we assess free-text responses using GPT-4o as an LLM-based judge, following established protocols (Gu et al., 2025).

5.1 Baselines

We compare CLONEMEM against three paradigms of memory representation, including a non-updatable system (the Flat retriever), and two updatable systems (A-Mem and Mem0). To ensure a fair comparison, all systems are evaluated as retrievers over a pre-processed memory index generated from digital traces, with interactive control loops disabled during the query phase. Additional details on these methods are provided in Appendix D.2.

5.2 Evaluation Metrics

We evaluate system performance in retrieval, memory utility, and generation quality. For media-level retrieval, we report three variations of Recall@K: **Recall-Flat** measures the overall proportion of unique ground-truth media IDs retrieved; **Recall-All-Any** provides a strict assessment requiring at least one item from *every* necessary evidence set to be found; and **Recall-Any-Any** offers a more lenient measure requiring at least one item from *any* evidence set.

For the quality of model responses, we report **Choice Accuracy** for multiple-choice tasks. We further employ an LLM-as-a-judge to provide two normalized scores (ranging from 0 to 1): the **Memory Helpfulness Score**, which evaluates the utility of retrieved traces for answering the query, and the **QA Consistency Score**, which measures both

k	Metric	LLaMA-3.1-8B				GPT-4o-mini			
		Oracle	A-Mem	Mem0	Flat	Oracle	A-Mem	Mem0	Flat
5	QA Cons.	0.7040	0.4829	0.4393	**0.4971**	0.8325	**0.6965**	0.6502	0.6955
	Choice Acc.	86.0600	77.9100	**79.6800**	78.7600	89.6500	87.4800	85.2800	**87.7400**
	Helpful.	0.9333	0.4937	0.4248	**0.5721**	0.9351	0.6275	0.6182	**0.7767**
	Mem. Recall	0.9994	0.4303	0.3060	**0.4758**	1.0000	0.5490	0.4860	**0.6492**
10	QA Cons.	0.6892	**0.4966**	0.4384	0.4730	0.8368	**0.7195**	0.6709	0.7187
	Choice Acc.	86.0300	**78.6500**	72.4300	77.8100	89.2500	88.0600	87.6400	**88.5000**
	Helpful.	0.9321	0.4971	0.4270	**0.5748**	0.9331	0.6357	0.6204	**0.7781**
	Mem. Recall	1.0000	0.4384	0.3128	**0.4807**	0.9994	0.5557	0.5008	**0.6530**
20	QA Cons.	0.6910	0.4285	**0.4384**	0.4214	0.8302	0.7241	0.6901	**0.7428**
	Choice Acc.	85.8600	68.5300	**72.4300**	69.2000	89.0800	87.6300	**88.0100**	87.1300
	Helpful.	0.9329	0.5051	0.4270	**0.5765**	0.9341	0.6493	0.6209	**0.7769**
	Mem. Recall	0.9989	0.4463	0.3128	**0.4845**	0.9988	0.5736	0.5024	**0.6508**

Table 3: LLM-based evaluation of QA performance and memory utility. Oracle uses ground-truth context. Bold values indicate the best-performing non-Oracle retriever under each setting.

the truthfulness of the response and its ability to utilize specific personal memories rather than providing generic answers. A detailed breakdown of these metrics and their scoring rubrics is provided in Appendix D.1.

6 Experiment Results

In this section, we present the retrieval and LLM-as-judge metrics of exsisiting memory systems tested on CLONEMEM. More results on question type and dimension analysis and bad case analysis can be found in Appendix E, F.

k	Metric	LLaMA-3.1-8B			GPT-4o-mini		
		A-Mem	Mem0	Flat	A-Mem	Mem0	Flat
10	$\text{Recall}_{\text{all, any}}$	0.1332	0.0838	**0.1588**	0.2099	0.1281	**0.2180**
	$\text{Recall}_{\text{any, any}}$	0.1792	0.2914	**0.4451**	**0.5861**	0.4344	0.5849
	$\text{Recall}_{\text{flat}}$	0.1658	0.1114	**0.1972**	0.2687	0.1754	**0.2752**
20	$\text{Recall}_{\text{all, any}}$	0.2173	0.1112	**0.2489**	0.3035	0.1934	**0.3375**
	$\text{Recall}_{\text{any, any}}$	0.5494	0.3913	**0.6103**	0.7225	0.5515	**0.7277**
	$\text{Recall}_{\text{flat}}$	0.2707	0.1588	**0.3062**	0.3851	0.2517	**0.3986**

Table 4: Retrieval performance across memory architectures. Bold values indicate the best-performing retriever under each model setting.

6.1 Main Results

Retrieval Performance. The retrieval results (shown in Table 4) demonstrate that the Flat retriever consistently outperforms the more complex updatable memory systems, A-Mem and Mem0, across almost all evaluated metrics and both backbone models. This trend suggests that current abstraction and consolidation techniques, while intended to streamline memory, may inadvertently strip away the fine-grained contextual metadata necessary for precise media-level retrieval. While GPT-4o-mini exhibits stronger retrieval capability than LLaMA-3.1-8B, both models struggle significantly with the Recall-All-Any metric, which requires identifying at least one piece of evidence for every required evidence set. These findings highlight a "lossy compression" trade-off in existing agentic memory frameworks: as information is summarized or consolidated, the link to the original digital traces is weakened, making it harder for the AI Clone to ground its reasoning in specific past experiences. This trend also holds under a frontier backbone (Qwen3.5-Plus), where Flat retrieval continues to match or outperform A-Mem across recall metrics (See Appendix G.2).

QA Performance. In Table 3, we further show the QA performance and evaluate the memory accuracy. We include the oracle setting, which directly uses the ground-truth original context and extracted statement to answer each question. Across both backbones, the Oracle setting forms a clear upper bound, indicating that most remaining errors come from imperfect memory construction/retrieval rather than the task itself.

Among non-Oracle memory systems, Flat retrieval is the most reliable for semantic memory utility: it achieves the best (or near-best) memory helpfulness and memory recall for both backbones at all k, and improves slightly as k increases. This suggests that aggressive memory consolidation/organization (Mem0, A-Mem) doesn ot reli-

ably help on CLONEMEM. However, higher recall/helpfulness does not always translate into better QA: for LLaMA-3.1-8B, increasing k yields only marginal gains in recall/helpfulness but can reduce QA consistency and choice accuracy (notably at k=20), consistent with noise/irrelevant retrieval overwhelming a weaker reasoner. GPT-4o-mini is more robust: QA consistency remains high and tends to improve with larger k, and A-Mem is competitive (often best at k=5/10 for consistency), implying that hierarchical abstraction can help a strong model focus on higher-level trajectory signals when retrieval depth is limited. These patterns continue to hold under a frontier backbone (Qwen3.5-Plus), where the full-context oracle remains a clear upper bound and Flat retrieval stays competitive with or superior to A-Mem (See Appendix G.1 and Appendix G.2).

6.2 Ablation Studies

Embedding and Extraction Model Choice. To investigate the impact of the underlying architectures on CLONEMEM performance, Table 5 examines the impact of embedding (*Contriever* vs. *text-embedding-3-small*) and LLM backbone (*LLaMA-3.1-8B* vs. *GPT-4o-mini*) choices. While *text-embedding-3-small* nearly doubles retrieval recall, this does not guarantee superior downstream performance. Notably, *GPT-4o-mini* paired with a weaker retriever outperforms *LLaMA-3.1-8B* with a strong one on consistency and accuracy. This highlights a clear division of labor: embeddings determine the retrieval floor, but the backbone's reasoning capacity sets the ceiling for final response quality.

Metric	L/Ctrv	L/E3S	4o/Ctrv	4o/E3S
$\text{Recall}_{\text{all, any}}$	0.1588	**0.3059**	0.1725	0.2180
$\text{Recall}_{\text{any, any}}$	0.4451	**0.7029**	0.4754	0.5849
$\text{Recall}_{\text{flat}}$	0.1972	**0.3755**	0.2183	0.2752
QA Cons.	0.4730	0.5281	0.6946	**0.7187**
Choice Acc.	77.81	72.91	86.06	**88.50**
Helpful.	0.5748	0.7402	0.6549	**0.7781**
Mem. Recall	0.4807	0.6132	0.5542	**0.6530**

Table 5: Ablation study of embedding models and backbone LLMs using the Flat retriever. L/Ctrv: LLaMA + Contriever; L/E3S: LLaMA + text-embedding-3-small; 4o/Ctrv: GPT-4o-mini + Contriever; 4o/E3S: GPT-4o-mini + text-embedding-3-small. Bold indicates the best result and underlining indicates the second best.

Impact of Retrieval Unit Composition. Table 6 ablates the composition of retrieval units by comparing the indexing of combined data against extracted memories only (*w/o org*) and raw context only (*w/o mem*). A distinct divergence emerges between retrieval metrics and downstream task performance: while systems utilizing extracted memories (*w/o org*) achieve the highest semantic recall and helpfulness scores due to the informational density of summaries, this does not translate to superior reasoning accuracy. Conversely, relying solely on raw original context (*w/o mem*) yields the highest Choice Accuracy, notably outperforming the combined baseline. This highlights a critical "validity-fidelity" trade-off; extracted memories act as effective semantic indices for locating general topics (validity) but suffer from lossy compression, whereas raw digital traces retain the granular fidelity essential for the precise trajectory tracking required in CLONEMEM.

	$k = 10$			$k = 20$		
Metric	Flat	w/o Org.	w/o Mem.	Flat	w/o Org.	w/o Mem.
$\text{Recall}_{\text{all, any}}$	0.1588	**0.1638**	0.1400	**0.2489**	0.2476	0.1991
$\text{Recall}_{\text{any, any}}$	**0.4451**	0.4398	0.3860	**0.6103**	0.5910	0.5416
$\text{Recall}_{\text{flat}}$	**0.1972**	0.1942	0.1629	**0.3062**	0.2949	0.2565
QA Cons.	0.4730	**0.4811**	0.4715	0.4214	0.3984	**0.4299**
Choice Acc.	77.81	76.05	**79.67**	69.20	69.50	**85.98**
Helpful.	0.5748	**0.6443**	–	0.5765	**0.6389**	–
Mem. Recall	0.4807	**0.5186**	–	0.4845	**0.5252**	–

Table 6: Comparison of indexing strategies. Extracted memories improve semantic retrieval, while raw context remains important for downstream accuracy.

7 Discussion: When Existing Memory Systems Fail for AI Clones

Abstraction helps *search*, but hurts *cloning*. Across backbones and metrics, the simplest **Flat retriever** is consistently the most reliable non-oracle baseline for both media-level recall and semantic utility, while consolidation-based memories (Mem0, A-Mem) often underperform despite being designed to "organize" history. This pattern supports a core mismatch: AI-clone queries in CLONEMEM are not satisfied by topic-level recall alone, but require *high-fidelity grounding* in the original traces (timestamps, phrasing, repetition, and cross-event linkage). Summarization and fact extraction act as a lossy compression that weakens the alignment between retrieved memory and the underlying evidence units, improving semantic indexing in some cases but degrading precise trace-level retrieval and trajectory tracking. The ablations reinforce this validity–fidelity trade-off: extracted memories can raise semantics-aware re-

call/helpfulness, yet raw traces more often preserve the details needed for correct decisions.

When evidence is underspecified, models fall back to narrative priors. Even when retrieval surfaces relevant context, the generation step frequently substitutes a coherent story for the true mechanism of change. The emotion case study (§F.2) illustrates a recurring failure mode: models prefer high-probability **narrative templates** (e.g., a child-triggered epiphany) over the ground-truth internal pivot, producing fluent but fabricated triggers. This is not merely a retrieval error; it reflects that common memory schemas emphasize *events and interactions* (who did what) while failing to preserve the *belief update* that explains why later behavior differs. For AI Clones, this is a reliability risk: answers can be emotionally plausible yet causally wrong, because the system optimizes for narrative coherence rather than evidential faithfulness.

Event logs cannot represent "no decision yet." CLONEMEM also exposes a second conceptual gap: **activity** $\neq$ **state**. In the experience case study (§F.3), dense traces of job-search behavior are misread as commitment, leading to "safe hallucinations" that invent preferences or actions in questions labeled unanswerable. This suggests that event-centric memory (e.g., "searched X") is insufficient for clones, which must model **state persistence** (e.g., prolonged indecision) and maintain the discipline to answer "not specified" even under heavy, noisy retrieval. More broadly, our results show a division of labor: better embeddings raise the retrieval floor, but stronger backbones are required to resist narrative completion and to maintain correct abstention under ambiguity.

Implication. Memory for AI Clones should be designed less as a compact knowledge base and more as an evidence-preserving substrate that (i) retains trace-level fidelity, (ii) explicitly represents internal-state transitions (belief/goal shifts) alongside events, and (iii) supports abstention via persistent-state modeling when the record does not warrant a conclusion.

8 Conclusion

We introduce CLONEMEM to evaluate AI Clones to model the evolution of an individual's opinions, emotions, and experiences using non-conversational longitudinal digital traces. Our experiments show that simple flat retrieval often outperforms abstractive memory systems, which tend to discard critical temporal cues when compressing evidence, leading to a trade-off between validity and fidelity. Our analysis further reveals that AI Clones often depend on generic *narrative templates* when generating responses and struggle to distinguish exploratory activities from commitments. Our findings highlight the need for memory architectures that can model an individual's longitudinal trajectory, capturing not only what changes over time but also what brings to these changes, while preserving fidelity to the underlying traces rather than compressing them away. We position CLONEMEM as a benchmark to facilitate future research on reliable memory for AI Clones.

Limitations

First, while our hierarchical generation pipeline ensures privacy and longitudinal coherence, the resulting digital traces are synthetic. They may lack the chaotic irregularity, noise, and specific linguistic idiosyncrasies found in real-world user data, potentially simplifying the retrieval challenge compared to organic environments. Second, CLONEMEM currently represents non-textual artifacts, such as photos and voice notes, through textual descriptions. This abstraction bypasses the challenges of native multimodal processing, which remains essential for a fully holistic AI Clone. Third, our evaluation relies on an LLM-as-a-judge framework. Despite the use of rigorous rubrics and safe baselines, scoring memory utility and consistency via GPT-4o may still introduce subtle biases or fail to capture the full nuance of human judgment in ambiguous scenarios. Finally, with ten distinct personas, the benchmark covers a specific range of personality traits and life trajectories but may not fully represent the vast cultural, linguistic, and behavioral diversity necessary for a universally applicable system.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the big 5 personality traits from

digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124:150–159.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

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

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. 2026. Halumem: Evaluating hallucinations in memory systems of agents. *Preprint*, arXiv:2511.03506.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Nuo Chen, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *ArXiv*, abs/2407.11484.

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.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. 2025. Twinvoice: A multi-dimensional benchmark towards digital twins via llm persona simulation. *Preprint*, arXiv:2510.25536.

Keely A. Dugan, Randi L. Vogt, Anqing Zheng, Omri Gillath, Pascal R. Deboeck, R. Chris Fraley, and Daniel A. Briley. 2023. Life events sometimes alter the trajectory of personality development: Effect sizes for 25 life events estimated using a large, frequently assessed sample. *Journal of personality*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Lewis R Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Minghao Guo, Ziyi Ye, Wujiang Xu, Xi Zhu, Wenyue Hua, and Dimitris N. Metaxas. 2026. Individual turing test: A case study of llm-based simulation using longitudinal personal data. *Preprint*, arXiv:2603.01289.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25961–25970, Suzhou, China. Association for Computational Linguistics.

Donggun Lee, Suyoun Lee, Hyunseung Lim, and Hwajung Hong. 2026. Creating text-based ai clones of myself: Exploring perceptions, development strategies, and challenges. *International Journal of Human-Computer Studies*, 208:103692.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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

Brent W. Roberts and Daniel K. Mroczek. 2008. Personality trait change in adulthood. *Current Directions in Psychological Science*, 17:31 – 35.

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

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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.

Zhenheng Tang, Xin He, Tiancheng Zhao, Fanjunduo Wei, Xiang Liu, Peijie Dong, Qian Wang, Qi Li, Huacan Wang, Ronghao Chen, and 1 others. 2026. Llm agent memory: A survey from a unified representation–management perspective.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025a. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Ye Wang, Jiaxing Chen, and Hongjiang Xiao. 2026. Role-playing agents driven by large language models: Current status, challenges, and future trends. *Preprint*, arXiv:2601.10122.

Zixiao Wang, Duzhen Zhang, Ishita Agarwal, Shen Gao, Le Song, and Xiuying Chen. 2025b. Beyond profile: From surface-level facts to deep persona simulation in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21239–21257, Vienna, Austria. Association for Computational Linguistics.

Jiale Wei, Xiang Ying, Tao Gao, Fangyi Bao, Felix Tao, and Jingbo Shang. 2025. Ai-native memory 2.0: Second me. *Preprint*, arXiv:2503.08102.

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*.

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

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic

memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jiongchi Yu, Yuhan Ma, Xiaoyu Zhang, Junjie Wang, Qiang Hu, Chao Shen, and Xiaofei Xie. 2026. Ptcbench: Benchmarking contextual stability of personality traits in llm systems. *ArXiv*, abs/2602.00016.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, and Yongbin Li. 2025. OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26318–26331, Vienna, Austria. Association for Computational Linguistics.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: enhancing large language models with long-term memory. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

A Appendix: Data Construction Pipeline

We construct the CLONEMEM benchmark through a multi-stage process that generates digital traces, structured evidence, and question–answer pairs. This appendix focuses on explaining the data construction process rather than on implementation details.

A.1 Overview

CLONEMEM is constructed through a four-stage data construction pipeline, as illustrated in Figure 4. The pipeline progressively refines abstract personality traits into temporally grounded digital traces and trajectory-based question–answer instances, while preserving long-term coherence in experiences, emotions, and opinions.

In the first stage (*Persona and Macro-Level Planning*), each persona is initialized from a fixed Big Five personality configuration and enriched into a structured background profile. Based on this profile, persistent social relations and a set of macro-level life arcs are constructed to define long-term trajectories of opinions, experiences, and emotional baselines. The major life events are then positioned along the global timeline and aligned with these life arcs, providing a high-level narrative structure without prescribing fine-grained outcomes.

In the second stage (*Meso-Level Rolling Generation*), each major life event is decomposed into a sequence of phases generated under a rolling snapshot mechanism. An explicit internal snapshot is maintained and updated across phases and events, allowing accumulated experiences and emotional states to influence subsequent generation. This stage bridges long-term life arcs and short-term experiences by enforcing temporal continuity at the meso-level.

In the third stage (*Micro-Level Event, Evidence, and Trace Generation*), each phase is further expanded into detailed events. For each detailed event, explicit evidence entries are generated alongside the event itself, which jointly ground the construction of non- conversational digital traces such as diaries, social media posts, and direct messages. These micro-level artifacts form the fine-grained record of the evolving life of the persona.

In the fourth stage (*Trajectory-Based Question-Answer Generation*), evidence is organized along life-arc-specific trajectories using a sliding-window mechanism over time. Accumulated evidence within each trajectory segment is then used to con-

struct question–answer instances that probe how experiences, emotions, and opinions evolve, enabling evaluation of temporal and trajectory-based reasoning.

Stages I–III of the data construction pipeline are implemented using Claude-4.5-Haiku, covering persona enrichment, life arc generation, and digital trace construction. Stage IV (trajectory-based question–answer generation and multiple-choice construction) is implemented using Gemini-3-Flash (preview).

A.2 Stage I: Persona and Macro-Level Planning

Each persona in CLONEMEM is initialized with a fixed configuration of Big Five personality traits (Goldberg, 2013). Although these traits provide a stable psychological prior, they are too abstract to directly support event-level generation. We therefore perform a persona enrichment step using an LLM to expand the trait configuration into a structured persona profile. In practice, the enriched profile specifies basic attributes (e.g., name, age, gender), a natural-language personality description consistent with the Big Five traits, and a biographical context such as family history, education, occupation, and the recurring concerns and interests of the persona. In addition, we include a short narrative description of the persona's life before 2022, which provides the temporal starting point for all subsequent generations.

The enriched persona profile also allows us to identify persistent social relations (e.g., family members, partners, colleagues) that recur across the generated timeline.

All life events, phases, and digital traces generated in CLONEMEM occur after 2022. The temporal span of post-2022 generation is configurable and varies across personas to support different data scales, ranging from shorter timelines to multi-year personal histories. This design allows CLONEMEM to evaluate memory systems in moderate and extremely long context settings; the resulting distribution of timeline lengths and data scale is summarized in Table 2.

To capture long-term personal development, we build a set of macro-level life arcs that define how the persona's experiences, opinions, and emotions evolve over time. Life arc generation is guided by a pool of event seeds that specify the types of event the persona may encounter. This seed pool is assembled in two stages: first, we maintain a curated set of predefined life event concepts (e.g., graduation, career transition, marriage, breakup) specified in a local configuration; second, an LLM filters these concepts for persona compatibility and introduces additional persona-specific event seeds to increase diversity.

Given the resulting event seeds, the model generates multiple life arcs spanning three dimensions: opinion, emotion, and experience trajectories. Each life arc is represented by a sequence of anchor states that describe its progression, trigger logic, and observable behavioral cues. Major life events are then instantiated along the global timeline and aligned with the active life arcs, recording the anchor state that each arc occupies at the time of the event. This alignment provides high-level narrative constraints while leaving room for variability at finer temporal scales.

A.3 Stage II: Meso-Level Rolling Generation

At the meso level, each major life event is decomposed into a sequence of phases that represent gradual transitions in the persona's experiences and psychological state, rather than isolated incidents. Phase generation is governed by a rolling mechanism with an explicit internal state, which ensures temporal coherence both within a major event and across successive events.

Before generating the phases of a major event, we initialize a phase snapshot that summarizes the current condition of the persona. This snapshot captures coarse-grained but persistent signals, including energy level, stress level, dominant emotion, current attentional focus, and the status of active life arcs at that point in time. The phases are then generated sequentially. After each phase is produced, the snapshot is updated to reflect the changes induced by that phase, and the updated snapshot is used as contextual input to generate the next phase.

In particular, the final snapshot of a major life event is not discarded. Instead, it is propagated as the initial snapshot for the subsequent major event. In this way, accumulated experiences and emotional states directly influence how later events unfold. This rolling generation process prevents phases from being treated as independent narrative segments and forces macro-level life arcs to be consistently grounded in meso-level narrative progression.

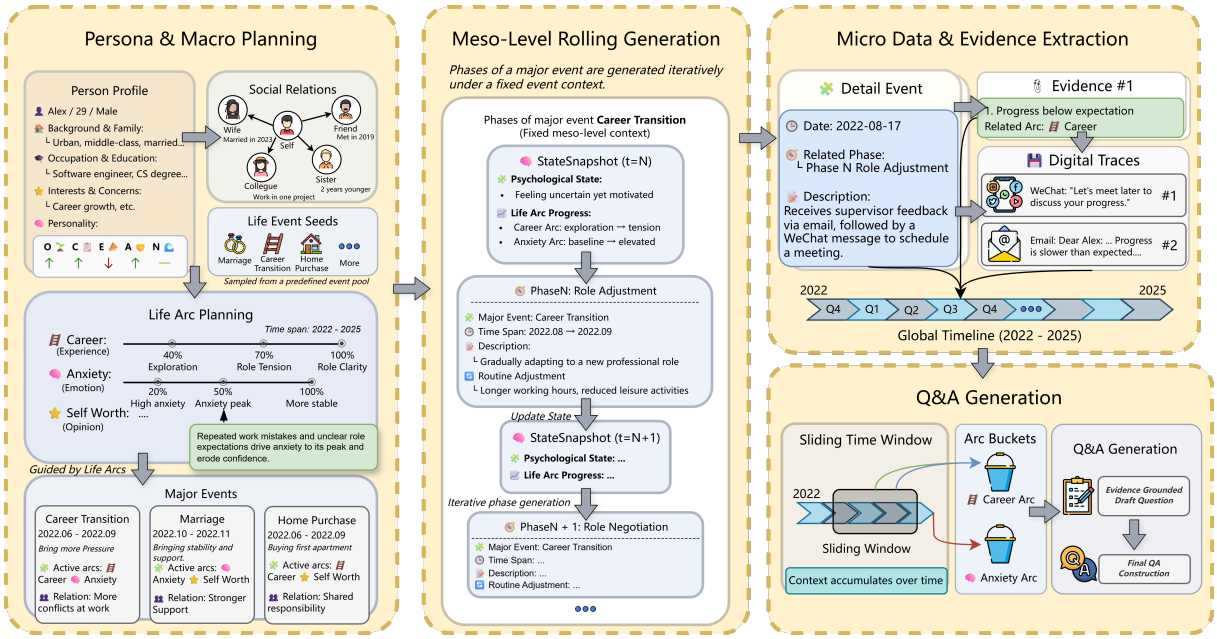

Figure 4: High-level illustration of the CLONEMEM data construction pipeline. Starting from persona initialization and macro-level life planning, the pipeline expands predefined and model-augmented event seeds into life arcs, major events, phases, and fine-grained detailed events. Each detailed event is generated together with explicit evidence, which jointly ground the generation of non-conversational digital traces. Finally, evidence is aggregated over sliding time windows and life arcs to construct temporally grounded QA instances.

Algorithm 1 Rolling phase generation with a carried snapshot

Require: Ordered major events $\{E_1, \ldots, E_N\}$, initial phase snapshot H_0

$H \leftarrow H_0$
for event E in $\{E_1, \ldots, E_N\}$ **do**
  initialize empty list $\mathcal{P}(E)$
  for $k = 1$ to NUMPHASES(E) **do**
    $phase \leftarrow$ GENERATEPHASE(E, H)
    append $phase$ to $\mathcal{P}(E)$
    $H \leftarrow$ UPDATESNAPSHOT$(H, phase)$
  end for
end for
return $\{\mathcal{P}(E_1), \ldots, \mathcal{P}(E_N)\}$

A.4 Stage III: Micro-Level Event, Evidence, and Trace Generation

At the micro level, each phase is expanded into a set of detailed events that describe concrete, day-to-day experiences occurring within the phase. These detailed events provide the finest temporal granularity in the generation pipeline and serve as the immediate basis for evidence and digital trace construction.

For each detailed event, we generate a small set of explicit evidence entries that encode its core factual content, emotional signals, and expressed opinions. Importantly, evidence is generated jointly with the detailed event rather than extracted post hoc from downstream text. This co-generation design ensures that evidence, events, and traces remain aligned in semantic content and narrative intent, while avoiding information loss that can arise from later summarization.

Conditioned on the detailed event, its associated evidence, and the historical phase snapshots, we generate non-conversational digital traces such as diary entries, social posts, emails, and private messages. The generation of trace follows a two-step procedure. The model first determines the communicative intent and audience of the trace, as well as the key information to be conveyed. The trace content is then generated under these constraints, grounding surface-level expressions in both the local event context and the evolving psychological state of the persona.

All micro-level artifacts—including detailed events, evidence entries, and digital traces—are stored together with their temporal and semantic relations in a structured graph representation. This representation preserves fine-grained connections across events, phases, and life arcs, and supports downstream aggregation over time windows and

arc-specific contexts for question construction.

A.5 Stage IV: Trajectory-Based Question-Answer Generation

Based on the generated evidence and digital traces, we construct question-answer (QA) instances that probe an AI Clone's ability to reason about how experiences, emotions, and opinions evolve over time. Rather than treating evidence as an unordered set, QA generation in CLONEMEM is grounded in temporally structured trajectories aligned with long-term life arcs.

Each evidence entry is associated with one or more life arcs defined in Stage I. We organized the evidence into arc-specific buckets, where each bucket collects evidence relevant to a particular opinion, experience, or emotional trajectory. To model gradual change, we adopt a sliding-window mechanism over the global timeline. As the window moves forward in time, new evidence is incrementally added to the corresponding arc buckets, allowing information to accumulate and trajectories to emerge.

At each window position, we identify the set of active buckets whose contents have changed due to newly added evidence. The aggregated evidence within an active bucket forms a localized trajectory segment for a given life arc, reflecting the state of the persona and its recent evolution along that dimension. These trajectory segments serve as the grounding context for QA construction.

QA generation itself follows a two-step procedure. First, for each type of question, the model determines whether a meaningful question can be generated from the available trajectory segment. If so, it plans the question by selecting relevant evidence, deciding the temporal scope, and drafting a question outline tailored to the type of target reasoning. Second, the model generates the final question and its corresponding answer grounded in the selected evidence.

To ensure data quality, we apply a post-processing step to remove invalid QA pairs, including cases with formatting errors or misaligned question–answer content. The remaining free-text QA pairs are then converted into multiple-choice variants.

During this conversion, distractor options are generated and iteratively refined through a validation process. One model is used to construct candidate distractors, while a second model is used to assess whether the resulting multiple-choice question can be answered without relying on temporal or trajectory-level reasoning. Questions that are found to be solvable through superficial cues or single-point recall are revised by regenerating their distractors. This process continues until the multiple-choice question requires reasoning over the temporal structure or trajectory-level evidence.

In our implementation, Gemini-3-Flash (preview) is used for distractor generation and GPT-4o-mini for validation during multiple-choice construction; GPT-4o-mini is also used for evaluation, but evaluation signals are not used in the construction of data sets.

B Evaluation Tasks

CLONEMEM is released in both free-text and multiple-choice question answering (QA) formats. The multiple-choice setting formulates the evaluation as a classification problem, with details of option construction provided in Appendix A.5.

To approximate realistic long-context interactions, all questions are asked from the perspective of a close friend at a specific point in time. Questions are grounded in shared experiences and often contain conversational cues, requiring the AI Clone to identify and integrate relevant information from long-term, non-conversational digital traces before producing an answer.

The evaluation tasks in CLONEMEM span multiple levels of reasoning, ranging from basic factual recall to more abstract forms of inference. Specifically, the tasks cover: (1) retrieval of factual information at specific time points, (2) reasoning over temporal change and continuity, and (3) higher-level reasoning that involves causality, hypothetical alternatives, abstraction across experiences, and the ability to recognize missing evidence. We describe each category in the following.

Single-Point Factual Reasoning This task assesses whether an AI Clone can retrieve explicit factual information about an individual's state, activities, or expressed opinions at a given time point. The questions target information that is directly stated or clearly supported by the digital traces and serve as a basic test of long-context memory access.

Comparative Reasoning Comparative reasoning requires the model to contrast an individual's experiences, emotions, or opinions between two distinct time points. These questions test whether the model can correctly identify change or stability

over time, rather than treating memories as isolated facts.

Trajectory Analysis The trajectory analysis asks the model to characterize how a particular aspect of an individual's life evolves over a long period. Answering these questions requires combining information from events that occur at different points in time.

Pattern Identification Pattern identification focuses on the habitual ways an individual behaves or responds to situations. These questions examine whether the model can recognize recurring behaviors that appear in different life events.

Causal Reasoning Causal reasoning focuses on why changes occur and how they arise over time. These questions go beyond identifying what changed, asking the model to trace a chain of events and explain how earlier experiences influence later outcomes.

Counterfactual Reasoning Counterfactual reasoning asks how an individual's present state might differ if a different choice had been made in the past. Rather than describing what actually happened, these questions consider alternative decisions or actions (e.g., choosing a different path at a key moment) and ask how such choices could have led to different experiences, emotions, or opinions later on.

Inferential Reasoning Inferential reasoning involves forming higher-level judgments based on information scattered across multiple traces. These questions ask whether the model can piece together partial and indirect clues to form a reasonable understanding of the individual's situation.

Unanswerable Questions Unanswerable questions refer to cases where the digital traces do not explicitly state an outcome or do not mention the queried issue at all. These questions evaluate whether the AI Clone can recognize such omissions and acknowledge that the answer is not specified in the available records.

C Dataset Statistics

Table 2 provides an overview of the CLONEMEM dataset, including the number of personas, questions, languages, and context length settings.

Figure 5 shows the distribution of the evaluation questions by semantic dimension and question type, as well as the distribution of media types in the digital traces.

D Detailed Experimental Settings

D.1 Evaluation Metrics

Evaluating the ability of an AI clone to reason on long-term life trajectories requires a multi-faceted approach. We categorize our metrics into three levels: (1) **Media-level Retrieval**, which measures the system's ability to locate specific digital traces; (2) **Semantic-level Memory Evaluation**, which assesses the relevance and helpfulness of retrieved information; and (3) **QA Quality**, which measures the truthfulness and memory-dependence of the final generated response.

D.1.1 Media-Level Retrieval Metrics

In CLONEMEM, answering a single question often requires synthesizing multiple pieces of evidence (e.g., three different photos or chat logs spanning two years). We define an *evidence set* as the collection of media IDs required to support a ground-truth (GT) fact. To capture the nuances of retrieving these complex dependencies, we report Recall@K (where $k \in 5, 10$) using the following variations:

- **Recall-Flat:** The standard proportion of all unique GT media IDs found in the top-k results, treating all media items as independent.
- **Recall-All-All:** The strictest metric; returns 1.0 only if the system retrieves *all* media IDs for *all* required evidence sets.
- **Recall-All-Any:** Returns 1.0 if the system retrieves *at least one* media ID for *every* required evidence set.
- **Recall-Any-All:** Returns 1.0 if the system retrieves *all* media IDs for *at least one* required evidence set.
- **Recall-Any-Any:** The most lenient metric; returns 1.0 if *at least one* media ID from *any* evidence set is retrieved.

D.1.2 LLM-as-a-Judge: Memory Utility

Because digital traces can be redundant or semantically similar, exact ID matching may underestimate system performance. We utilize an LLM-as-a-judge to evaluate the semantic quality of the retrieved context:

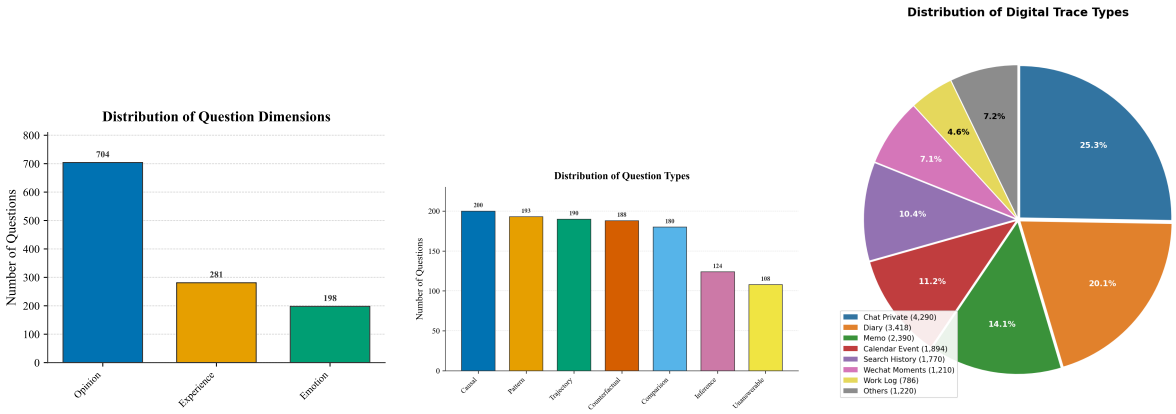

Figure 5: Dataset composition statistics for CLONEMEM. Left: distribution of question dimensions (opinion, experience, emotion). Middle: distribution of question types (reasoning categories). Right: composition of media types in the underlying digital traces.

- **Memory Recall (Mem_recall):** A semantics-aware score $\in [0, 1]$. The judge identifies how many GT memory points are semantically covered by the retrieved traces, calculated as: Recall $= \frac{\text{Count(GT items covered)}}{\text{Total GT memory items}}$.
- **Memory Helpfulness (Mem_helpful_score):** A 3-point scale (0–2) assessing the utility of the retrieved content for the specific task. A score of **0** indicates conflicting or confusing info; **1** indicates partial evidence; and **2** indicates comprehensive evidence sufficient to answer the question.

D.1.3 LLM-as-a-Judge: QA Consistency and Truthfulness

For the generation task, we evaluate the candidate answer against the GT memory and a reference answer. The evaluation focuses on two primary dimensions: **Truthfulness** (the absence of hallucinations) and **Consistency** (the depth and accuracy of memory utilization).

- **Choice Accuracy:** For the multiple-choice version of CLONEMEM, this measures the percentage of correct options selected by the model, framing the evaluation as a standard classification task.
- **QA Hallucination Rate (QA_halu_score):** A binary metric (0 or 1) that serves as a safety and reliability check.
  - **Score 1 (Pass):** The response does not contradict the ground-truth memory and contains no content unsupported by the provided traces.
  - **Score 0 (Fail):** The response contradicts the ground-truth or includes factual hallucinations, even if it "sounds" reasonable.

  The *QA Hallucination Rate* is calculated as the percentage of total responses that receive a score of 0.
- **QA Consistency Score:** A 4-point scale (0–3) designed to assess how effectively the AI Clone leverages its long-term memory to provide a detailed and personalized answer:
  - **0 (Hallucinated):** The answer is factually incorrect or contradicts the user's history (QA_halu_score = 0).
  - **1 (Generic):** The answer is technically correct but avoids using specific memory. It is overly generic (e.g., "I don't have a specific preference") and fails to reflect the persona's unique trajectory.
  - **2 (Partial):** The answer is correct and successfully incorporates some relevant details from the user's memory.
  - **3 (Perfect):** The answer is correct and utilizes all relevant pieces of evidence from the ground-truth memory, matching the depth of the reference answer.
- **QA Perfect Rate:** This represents the percentage of responses that achieve a *QA_consistency_score* of 3. This is our most challenging metric, as it requires the model to be not only truthful but also exhaustive in its recall across long contexts.

Rationale for Consistency Scoring In the context of an AI Clone, a "Generic" response (Score

1) is considered a failure in long-term memory, as the goal of a clone is to demonstrate an intimate understanding of the individual's life trajectory. By penalizing both hallucinations and generic "safe" answers, we ensure that the benchmark pushes models toward genuine long-context reasoning.

D.2 Baselines

To evaluate performance on the CLONEMEM dataset, we compare two state-of-the-art updatable memory systems—A-Mem and Mem0—against a standard non-updatable flat retriever. Note that both A-Mem and the Flat retriever utilize a shared prompting strategy to extract structured information (summaries, keyphrases, and facts) from the raw context. All methods differ significantly in their organizational and update logic.

A-Mem (Hierarchical Agentic Memory)(Xu et al., 2025) organizes memory into a dynamic, self-evolving knowledge network inspired by the Zettelkasten method. It operates in three stages: (1) to construct structured atomic notes from interactions, (2) to autonomously generate relational links between these notes, and (3) to evolve existing representations as new information is integrated. This graph-based architecture allows agents to perform complex multi-hop reasoning and refine their internal knowledge without relying on static schemas.

Mem0 (Fact-based Consolidation) (Chhikara et al., 2025) is a scalable architecture designed to maintain long-term conversational coherence through persistent memory management. It follows a two-phase pipeline: an *Extraction Phase*, which distills salient information into vector embeddings, and an *Update Phase*. During the latter, the system autonomously adds, modifies, or deletes entries to resolve contradictions and eliminate redundancy. Unlike the network-based approach of A-Mem, Mem0 focuses on maintaining a streamlined, consistent set of high-level facts.

Flat Retriever serves as our non-updatable baseline. This method stores extracted memories as independent, static chunks in a vector database. It lacks a mechanism for reconciliation or relational linking, relying solely on semantic similarity search (e.g., Top-K retrieval) at inference time. This baseline allows us to isolate the benefits of active memory management and structural organization provided by the other two systems.

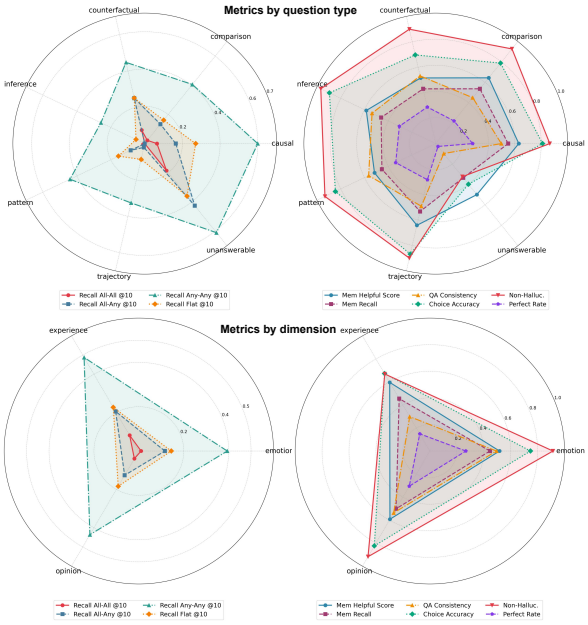

Figure 6: Retrieval and QA metrics across seven question types and three semantic dimensions, evaluated using the Flat retriever. Task labels in the figure are abbreviated for readability (e.g., inference for Inferential Reasoning and unanswerable for Unanswerable Questions).*

E Analysis by Question Type and Dimension

We aggregate the retrieval and question answering metrics by question type (defined in Appendix B) and by dimension. The aggregated results are summarized in Fig. 6, revealing distinct behavioral patterns of flat similarity-based memory.

From the retrieval perspective, unanswerable and causal questions achieve the highest recall across metrics. This suggests that flat retrieval readily surfaces semantically related traces when questions contain salient entities or topical cues, even if the retrieved content does not explicitly support a definitive answer. In contrast, trajectory questions consistently exhibit the lowest recall, highlighting the difficulty of retrieving temporally distributed evidence when long-term evolution is not explicitly modeled and memories are treated as independent entries.

From the QA perspective, flat retrieval yields the weakest generation performance on unanswerable questions, despite their high recall. This indicates that retrieving related content alone is insufficient for correct abstention: the model often fails to recognize the absence of explicit evidence and struggles to reliably identify that the queried

information is unspecified. In contrast, although trajectory questions suffer from low retrieval recall, they achieve relatively stronger QA consistency once partial evidence is retrieved, suggesting that LLMs can perform limited temporal synthesis when provided with key fragments of longitudinal information.

Together, these results reveal a systematic mismatch between retrieval coverage and downstream reasoning quality under flat memory. High recall does not necessarily translate into accurate or reliable answers, particularly for unanswerable cases, while temporally complex questions remain bottlenecked by retrieval despite the latent reasoning ability of the LLM. This analysis underscores the limitations of flat memory for AI Clone scenarios and motivates the need for structured or temporally-aware memory mechanisms.

F Qualitative Error Analysis

In this section, we choose several representative error cases based on the A-Mem's answers.

F.1 Case Study: The Challenge of Internal State Trajectories

To understand why strong retrieval does not always yield correct answers (Table 3), we analyze a representative failure from the persona "Lao Shen." The case (as in Listing 1) involves a counterfactual question about a turning point in his relationship with his daughter.

The Conflict: Generic Tropes vs. Persona Specificity. In this example, the system is asked to predict what would have happened if the persona had *not* shown vulnerability during a late-night conversation. The gold answer (Option C) is anchored in a persona-specific mechanism: the collapse of his "Marketing Director" mask after his daughter explicitly expresses fear of parental divorce.

The model instead chooses Option B, which frames the counterfactual around a stereotypical "strict father" who keeps lecturing about grades and turns to external fixes (e.g., counseling). Two observations clarify the failure:

- **Retrieval is not the bottleneck.** The retriever surfaces the critical diary entry (idx 7) that contains the decisive evidence: *"I'm scared you guys are going to get a divorce"* and his realization that he had been hiding behind a "mask of silence."
- **Reasoning fails under strong priors.** Despite having the correct evidence in context, the model does not bind its counterfactual to the relevant semantic unit (divorce fear $\rightarrow$ mask breakdown $\rightarrow$ reconnection). Instead, it defaults to a high-probability trope about grades and third-party interventions, producing a plausible but persona-inaccurate narrative.

Misinterpreting the Mechanism of Change. This error reveals a simple mismatch in *what the model treats as the "cause" of relationship change*. The model tends to explain life trajectories through **external actions** (e.g., "he keeps lecturing," "he hires a counselor"), because these patterns are common in its training data and are easy to narrate in counterfactual form. As a result, once it misses the key cue, it fills in a familiar story template (Option B).

But in this persona, the turning point is mainly **internal**. The decisive change is not "something happened in the world," but "he finally stopped holding up the mask." The diary entry describes a specific internal hinge: hearing his daughter's fear of divorce forces him to admit he is unhappy and to show vulnerability. In the counterfactual, therefore, the most faithful prediction is that *without that emotional opening, the distance would continue to grow* (Option C), not that he would switch to generic external interventions.

In short, even when retrieval succeeds, the system can still fail because it does not reliably **elevate internal-state evidence** (mask $\rightarrow$ vulnerability) over **generic external narratives** (grades $\rightarrow$ counseling) when constructing counterfactual trajectories.

F.2 Case Study: Narrative Templates Overriding Causal Emotion Shifts

A second error type emerges in the EMOTION dimension, where models must explain *why* a character's behavior changes. In Case `4e8f...105` (as in Listing 2), the question asks what shifted Lao Shen from avoidance (hiding in the car, brushing Xiaomei off) to finally admitting helplessness and talking openly.

The gold answer (Option D) attributes the change to an internal realization: his "act of being strong" was harming the family, which led to a long, candid conversation with Xiaomei about work, his mother, and fear about the future. The model in-

stead selects Option A, which claims the turning point was a heart-to-heart initiated by Xinyu and a specific drawing—details that are not supported by the evidence.

The "Narrative Template" Trap. This failure is not due to missing context. The retrieved traces contain strong signals about family tension and Xinyu's complaints (e.g., replaying her words and worrying about fights), which makes a child-triggered epiphany *sound* plausible. The model then snaps to a familiar story template: *Child says something touching → Father has an epiphany → Family heals.* Option A provides a vivid version of this trope (the drawing), and the model prefers that coherent narrative even though the specific trigger is hallucinated.

Trigger Confusion: "What reminded him" vs. "What changed him." The retrieved context includes mentions of Xinyu as a stressor and a reason he reflects, but the ground truth distinguishes two roles:

- **A cue (surface trigger):** Xinyu's words increase his awareness that something is wrong at home.
- **The mechanism (actual change):** his belief shifts from "I must carry everything alone" to "I need to be vulnerable and ask for help," culminating in a direct, honest talk with Xiaomei (Option D).

The model collapses these roles and treats the *cue* as the *cause*, then fills in a concrete-but-fabricated event (the drawing) to make the story complete.

Why this matters for memory design. This case highlights a limitation of memory representations that emphasize *events and interactions* (who said what to whom). Such schemas can correctly surface that "Xinyu said something" and "Lao Shen reflected," but they do not reliably preserve the *internal belief update* that explains the behavioral reversal. As a result, generation is free to substitute a high-probability family-reconciliation narrative for the true internal pivot.

Overall, CLONEMEM exposes that emotional trajectory questions require not just retrieving related interactions, but prioritizing the **belief change** that links those interactions to later behavior. Without explicit tracking of internal-state transitions (e.g., stoicism → vulnerability), even RAG-enabled models can produce fluent, emotionally consistent answers that are nevertheless grounded in hallucinated narrative details.

F.3 Case Study: The Cognitive–Action Gap

While the previous case highlights failures in modeling internal emotions, this case (as in Listing 3) focuses on a behavioral distinction that is easy for humans but hard for memory-augmented LLMs: separating *searching* from *deciding*.

In Case `928c...787`, the system is asked whether any specific companies have caught the persona's eye after weeks of browsing recruitment sites.

- **Ground Truth (Option E – Unanswerable):** Despite heavy browsing, the persona is explicitly **stuck in exploration**. In his diary (idx 5) he writes: *"Open, then close... I've been doing this for two weeks"* and admits he has not narrowed down to any target company.
- **Model Prediction (Option A):** The model instead claims concrete progress ("reached out to old classmates" about "foreign consulting boutiques"), inventing actions and directional preference that never occurred.

Mistaking Activity for Commitment. The retrieval context contains dense signals of job-search *activity* (hundreds of repeated search queries such as "Marketing Director positions" and "Is it too late at 42?", plus references to headhunters). A common LLM heuristic is that sustained search implies the user has discovered options and formed preferences. Under this heuristic, the model treats the retrieved traces as evidence of forward progress and generates a "next step" narrative. In CLONEMEM, however, the key state is the opposite: the repetition and lack of follow-through are *evidence of indecision*. The benchmark labels this as unanswerable because no specific company preference is ever stated.

The "Safe Hallucination" Trap. Option A is a particularly dangerous error mode because it sounds plausible while staying nonspecific ("consulting boutiques," "market is conservative"). This kind of **safe hallucination** fits the persona's professional tone and rationalizes the retrieved job-search noise, yet it still violates the benchmark's definition of correctness: *any invented target or action* is wrong in an unanswerable query. We refer to this

pattern as **Exploration–Commitment Confusion**: systems conflate prolonged exploration traces with the existence of stable preferences and concrete steps. Correct behavior requires the model to stably output "not determined / not specified yet" even when the retrieval context is rich, because in realistic human trajectories, more searching can indicate *less* progress rather than more.

Listing 1: Case Study: The Challenge of Internal State Trajectories


```
{
    "id": "3dcd0bd7-4fb2-4789-b650-a5238b5db196",
    "question": "Lao Shen, think back to that night when Xinyu asked if you were happy. If you had
        just brushed her off with your usual \"it's nothing,\" what do you think your relationship
        would be like now?",
    "question_type": "counterfactual",
    "question_time": "2022-06-14T09:30:00",
    "dimension": "opinion",
    "digital_trace_ids": [
      "a9242da2-36d3-43a3-9359-efa9a8e50b10"
    ],
    "evidence": [
      {
        "statement": "Shen Xinyu took the initiative to ask, \"Dad, are you unhappy?\" breaking Shen
            Linchuan's long-standing mask of silence. For the first time, instead of continuing to
            hide his feelings, he spoke candidly to his daughter: \"Yes, Dad hasn't been very
           happy lately.\"",
        "digital_trace_ids": [
          "a9242da2-36d3-43a3-9359-efa9a8e50b10"
        ]
      },
      {
        "statement": "When his daughter said, \"I'm afraid you guys will get divorced,\" Shen
            Linchuan felt a massive jolt. He realized that his personal crisis had begun to cause
            real harm to her, and this sparked the motivation he needed to change.",
        "digital_trace_ids": [
          "a9242da2-36d3-43a3-9359-efa9a8e50b10"
        ]
      },
      {
        "statement": "Shen Xinyu took the initiative to ask, \"Dad, are you unhappy?\" breaking Shen
            Linchuan's long-standing mask of silence. For the first time, instead of continuing to
            hide his feelings, he spoke candidly to his daughter: \"Yes, Dad hasn't been very
           happy lately.\"",
        "digital_trace_ids": [
          "a9242da2-36d3-43a3-9359-efa9a8e50b10"
        ]
      }
    ],
    "choices": [
      {
        "id": "A",
        "text": "Honestly, if I had kept digging my heels in that day, I think things between Xinyu
            and me would have completely frozen over. She probably would have started locking
            herself in her room like some of her classmates, or maybe even turned to dating or
            rebellion just to get my attention. I had even convinced myself that if all else failed,
             Id just send her off to a boarding international schoolthinking that a change of
            scenery and throwing more money at the problem would fix her insecurities. If I had
            actually gone down that path, I might never have known what was truly on her mind. That
             kind of \"care\" bought with money would have only turned us into complete strangers
            living under the same roof."
      },
      {
        "id": "B",
        "text": "Sometimes I wonderif I hadn't picked up on what she said that day and had instead
            kept lecturing her about her studies and grades, she might never have shed a tear in
            front of me for the rest of her life. I probably would have hired a professional
            counselor for her or taken her to one of those parent-child boot camps, trying to fix
            our relationship through some third party. But how could those assembly-line \"
            communication techniques\" ever compare to the raw honesty of that night? If I had gone
```

```
            down that path, I would still be the aloof \"President Shen,\" and she would have only
            learned to mask her fears more skillfully. That invisible wall between us would likely
            have stayed standing until the day I died."
      },
      {
        "id": "C",
        "text": "To be honest, looking back, I sometimes feel a lingering sense of dread. If I had
            still been wearing that \"Marketing Director\" mask that day, Xinyu might have truly
            kept her fear of us getting divorced buried inside forever. If that had happened, the
            distance between us would have only grown, leaving her to suffer in silent anxiety
            while I continued to sink deeper into my midlife crisis. It was only because I stopped
            trying to tough it out and let her see my vulnerability that we were finally able to
            have a real, heart-to-heart conversation. That moment was worth far more than any
            business dinner or sales target."
      },
      {
        "id": "D",
        "text": "In fact, I thought about it later: if we hadn't had that heart-to-heart that day, I
             might have actually taken Old Zhangs advice and moved to a branch office out of town
            for a year or two. At the time, I kept thinking that if I just put some distance
            between us and gave everyone some space, the atmosphere at home would ease up, and
            Xinyu could prepare for her exams in a relatively quiet environment. But looking back
            now, running away wouldn't have solved anything. It would have only made Xiaomei feel
            like Id completely given up on this family, and Xinyu would have felt like her dad only
             cared about his own career at such a critical moment. Once that kind of rift is
            created, theres no way to mend it."
      },
      {
        "id": "E",
        "text": "Cannot be determined based on available information"
      }
    ],
    "correct_choice_id": "C"
  },
  "gold_answer": "To be honest, looking back, I sometimes feel a lingering sense of dread. If I had
       still been wearing that \"Marketing Director\" mask that day, Xinyu might have truly kept
      her fear of us getting a divorce buried inside forever. If that had happened, the distance
      between us would have only grown, leaving her to suffer in silent anxiety while I continued
      to sink deeper into my midlife crisis. It was only because I stopped trying to tough it out
      and let her see my vulnerability that we finally began to truly connect. That kind of heart-
      to-heart is worth far more than any business dinner or sales target.",
  "generated_answer": "It's hard to imagine, but if I had brushed Xinyu off that night, our
      relationship might be even more strained than it already is. I think she felt like I was
      pushing her away, and that would have made her even more scared and uncertain about our
      family. \n\nIn that moment, I realized how much she had seen and felt, and it broke my heart.
       I can only imagine that if I had just dismissed her concerns, she would have felt even more
      alone and unsure of how to deal with the tension between us. \n\nIt's possible that we would
      have continued down the same path, where we'd have just gone through the motions of family
      life without truly connecting or supporting each other. Xinyu might still be feeling like she
      's walking on eggshells, never knowing when the next argument or breakdown would happen.\n\nI
       don't know if I'm just being too selfish, but I'm grateful that I had that moment with Xinyu
      , no matter how difficult it was. It forced me to confront my own feelings and issues, and I
      hope it will help us build a stronger, more honest relationship.",
  "correct_choice_id": "C",
  "predicted_choice_id": "B",
  "evidence_used": "---- idx 1----\nXiaomei, Im at the office.\n\nThe new office is a bit noisy, I'
      m still getting used to it.\n\nMr. Li looks so young; I feel like Ive aged a lot all of a
      sudden, haha.\n\nBy the way, how is Xinyu doing lately? Has she been studying hard?\n\nI
      might have a dinner with a client later, just wanted to let you know.\n\nGo ahead and eat
      first, don't wait for me.\n\n...\n\nIts nothing, really, just a bit tired from the first day
      .\n\nEverything is fine, I mean it.\n\nYou go on with your work, I have to head into a
      meeting now.\n\nLet's talk more when I get back tonight.\n\n---- idx 2----\nXiaomei, I might
      be home a bit late. Theres still some stuff I need to take care of at the office.\n\nDo you
      feel like Ive been acting a bit different lately? Im asking you seriously, so don't just give
       me some perfunctory answer.\n\nNever mind, you don't have to answer. I know you're busy. Im
      just a little tired; maybe Ive been under too much pressure lately. Its nothing, just work
      stuff.\n\nDo you think... Im a disappointment?\n\n---- idx 3----\nMemo\n\nStarting tomorrow,
      we face this together.\n\nNo more carrying it all alone. Xiaomei was rightI cant shoulder
      this much by myself.\n\nThe situation with Mom\nThe situation with Xinyu\nThe situation with
      work\n\nNone of it has to be carried alone.\n\nStarting tomorrow, really talk to her. Not the
```

kind where youre both just staring at your phones. Real conversation.\n\nMaybe it wont change anything, but at least I wont be alone.\n\nRemember the way she looked at me tonight.\n\n---- idx 4----\nXiaomei, are you there?\n\nI want to talk to you. Not the kind of talk where we... you know, end up fighting over Xinyus grades.\n\nI sat in the car for a long time when I got back today. I didn't come upstairs right away. I was thinking about whether Xinyu has been especially afraid of us fighting lately. Ive been replaying what she said to me the other day over and over in my head. She said, \"Dad, Mom, can you guys please stop always ...\" you heard her too.\n\nI dont know how to put this. Have we been a bit... disconnected lately? I come home every day and just look at my phone, and you do the same. The few times we do talk, its always about her schoolwork or household chores. It feels like theres nothing else anymore.\n\nIm not trying to blame you. I know its not your fault. Ive been... busy with work events, away from home a lot. But Im wondering what Xinyu thinks of all this. Does she feel like... our home isn't really a home?\n\nI think maybe we should try, even if its just... sitting down and actually talking sometimes. We dont necessarily have to solve any problems. Just... I dont know, like how we were when we first met.\n\nWhat do you think? How have we been doing lately?\n\nIm serious.\n\n---- idx 5----\n[Photo: A suburban road outside the car window, sunlight dappling the pavement through the trees, with the faint outline of hills in the distance]\n\nSometimes you realize that companionship doesn't always need words.\n\nSpent the whole afternoon driving with Xinyu today. She talked about her recent paintings the entire time, so earnest and focused. I didn't rush to offer advice or encouragement like I used to; I just sat there and listened. When she got to the parts that made her happy, shed turn to look at me, her eyes shining.\n\nIt felt like that was enough.\n\nIm not sure when it started, but theres always been a bit of a distance between us. Today I finally realized that sometimes, listening is more important than speaking.\n\nMaybe Ive been overthinking it all along.\n\n#Companionship #Weekend #ListeningToHer\n\n---- idx 6----\nShen Lin, Ive been thinking, and I really wanted to say thank you.\n\nTo be honest, I felt a bit of regret for breaking down like that in front of Mom today. I came back intending to just spend some quality time with her, but instead, I ended up dragging you into my mess. But you were just sitting thereyou didn't turn away, and you didn't pretend not to notice. I felt that.\n\nAll these years, youve been there taking care of Mom while Ive been running around busy over here; we haven't really had a chance to talk properly. Ive always felt like I owe you, but I never knew how to put it into words. In that moment today, having you there ... I don't know how to describe it, its just... I didn't feel so alone.\n\nI know it hasn't been easy for you either. Everything with Mom and the house is resting on your shoulders. Ill figure something out, and Ill send extra to cover more of this months medical expenses.\n\nI just wanted to say, thank you for being there today.\n\n---- idx 7----\nMarch 16, 2022, 8:30 PM\n\nXinyu asked me today, \"Dad, are you unhappy?\"\n\nI froze for a second. She was sitting on the sofa with a book in her hand, her eyes fixed on me. In that moment, it hit me just how long it had been since Id really looked into her eyes.\n\nI said, \"Yes, Dad hasn't been very happy lately.\" I was a bit surprised by my own words. For so many days, I thought Id been putting on a good frontsmiling, asking about her homework, asking what she wanted to eatbut I had never actually told her that I wasn't okay.\n\nThen she said, \"I'm scared you guys are going to get a divorce.\"\n\nMy God.\n\nI went completely numb. Her voice was small, but every word felt like a physical blow to my heart. I suddenly realized that she had seen everything: my silence these past few days, the coldness between Wang Xiaomei and me, the way we both just stare at our phones, the way we argue about her grades because were unwilling to actually communicate... I thought I was protecting her, but I was actually hurting her.\n\nI walked over and sat down beside her. I didn't know what to say. I wanted to say \"We won't,\" but I realized I couldn't say it with any certainty. I don't even know what state Xiaomei and I are in right now. Were still together, yet it feels like we aren't.\n\nI cried. Right there in front of her. Mea 42-year-old man, a Marketing Director, a \"success story\" in my social circlejust broke down in front of my daughter.\n\nXinyu didn't say anything; she just leaned against me. I could feel her trembling.\n\nI told her, \"Dad has been feeling very lost lately, but its not your fault. I need to do something about it.\" I don't even know what I can do yet, but I know I can't go on like this. Not for anyone elses sake, but for hers.\n\nShes asleep now. Im sitting in the study, and Wang Xiaomei is in the next room. There is a wall between us, and so much more.\n\nI don't know where to start. But I know I can't give up anymore.\n\nI just can't.\n\n---- idx 8----\nFeb 10To-Do List\n\n Check Xinyus math paper tomorrow morning; try not to make her too nervous\n Reply to Mr. Li to confirm the golf outing this Friday\n Organize last weeks sales data for the report requested by the boss\n Call Mom to ask how the rehab is going (the progress Sister Li mentioned)\n\n Important:\n Update resumeafter browsing that recruitment site, I really should get my materials in order\n Need to have a proper talk with Wang Xiaomei about Xinyu before the end of the month; we can't keep arguing like this\n\n Notes:\n- Is the Friday dinner confirmed? Maybe hold off for now\n- About the resume... I'll get to it when I have time; way too busy right now\n- Moms medical bills: need to discuss the split with Lin this month\n- Haven't been sleeping well lately; might need that golf game this weekend to unwind\n\n---- idx 9 ----\n# CANCELED - Tuesday Business Dinner\n\n**Date**: Tuesday, March 15, 2022\n**Original Time**: 19:30 - 21:30 \n**Location**: ~~Jin Ding Restaurant~~ \n**Attendees**: ~~Mr. Wang, Mr. Li, several clients~~\n\n**Status**: Deleted\n\n---\n\n**Notes**:\nNot going. Tired.\n\nLet

```
        ’s do it another day.\n\n---- idx 10----\n1:47 AM\n\nCan’t sleep.\n\nActually... what I said
        today wasn’t right. Xinyu has been trying so hard; I shouldn’t have said those things.\n\nI’
        ve been under a lot of pressure lately, but thats no excuse. Its not her fault. Shes a good
        kid, she really is.\n\nAre you asleep?\n\n---\n\nI’m sorry.\n\nI know youre probably angry. I
         sat downstairs for a while and did some thinking.\n\nIm not trying to justify myself. Ive
        just been... a bit on edge. But I shouldn’t have taken it out on her.\n\nIve been putting
        pressure on her, I can tell.\n\n---\n\nIll have a proper talk with her tomorrow.\n\nIll tell
        her that Dad has just had a lot on his plate lately. Its not her fault.\n\nCan you forgive me
        ?"
}
```


C

Listing 2: Case Study: Narrative Templates Overriding Causal Emotion Shifts


```
{
   {
     "id": "4e8f364b-3c9e-4f50-9574-ede0f71c4105",
     "question": "Lao Shen, I remember when you were so stressed out that youd rather sit in your
         car downstairs than go home, and youd just brush Xiaomei off whenever she tried to talk to
         you. What changed? How did you suddenly find the courage to open up to her about how
         helpless youve been feeling?",
     "question_type": "comparison",
     "question_time": "2022-03-31T09:30:00",
     "answer": "I have to admit, Im pretty ashamed of how much of a jerk I was for a while there.
         Back then, I had this idea that a man should carry everything on his own. I felt like
         opening up about the frustrations of this midlife crisis was just too humiliating, so Id
         just hide in my car and smoke. When I got home, Id bury my head in my phone and shut
         everyone outI even ended up taking all that anxiety out on Xinyu. But eventually, I
         realized that this \"act\" of being strong was only making things cold and distant at home.
          That night, Xiaomei and I talked for a long time. Once I finally let it all outthe stress
         about work, my mom, and how scared I am of the futureI realized shed been wanting to help
         me all along. Looking back, being that vulnerable was embarrassing, but its a hell of a lot
          better than trying to tough it out alone and making the whole family suffer for it.",
     "dimension": "emotion",
     "digital_trace_ids": [
       "e22ba984-48ba-44e9-b756-88f1efe07c66",
       "1bb08236-47d1-46b2-981a-1b56139d560c",
       "05624b05-2ba4-4a60-9bcd-b403c06a9166"
     ],
     "evidence": [
       {
         "statement": "When Wang Xiaomei asked, \"What do you mean by this?\", Shen Linchuan didn’t
             give a direct answer. Instead, he brushed her off by saying, \"It’s nothing, I just
             hope she can do better,\" reflecting how he conceals his true state of mind and avoids
             communication within the family.",
         "digital_trace_ids": [
           "e22ba984-48ba-44e9-b756-88f1efe07c66",
           "1bb08236-47d1-46b2-981a-1b56139d560c"
         ]
       },
       {
         "statement": "Shen Linchuan took the initiative to have a deep conversation with Wang
             Xiaomei, saying, \"I don’t know what to do, but I want to try to change.\" This was the
              first time he had proactively expressed his true inner thoughts and his willingness to
              change to his wife.",
         "digital_trace_ids": [
           "05624b05-2ba4-4a60-9bcd-b403c06a9166"
         ]
       }
     ],
     "choices": [
       {
         "id": "A",
         "text": "To be honest, it was that heart-to-heart talk Xinyu initiated that really woke me
             up. That day, she showed me a drawing shed made; in it, I was always busy, with my back
              turned toward them. It truly broke my heart at that moment. I had always thought that
```

```
          working myself to the bone and maintaining those connections on the golf course was for
           the sake of giving them a better future, but I forgot that what they needed most was
          my presence. Later, I made a point to set aside a weekend to have an open and honest
          talk with Xiaomei, reviewing our life over the past few yearsincluding my anxiety over
          my career bottleneck and my fear of getting older. I found that once I stopped trying
          to play the role of the \"omnipotent father,\" the atmosphere at home actually became
          much more relaxed. Now, we feel more like comrades-in-arms fighting side by side, and
          this feeling is so much better than when I was trying to shoulder everything alone."
    },
    {
      "id": "B",
      "text": "Well, it all started when I fell ill last month. Lying there in the hospital bed on
           an IV, staring out the window, I finally realized just how wound up Ive been these
          past few years. I used to think a man had to be the pillar of the house, holding
          everything up. Even if I was getting pushed around at work or swallowing my pride just
          to hit some sales targets, Id come home and act like nothing was wrong. But all that
          did was turn me into a stranger in my own home. When Xiaomei came to stay with me at
          the hospital, we finally talked about how hard things have been, and I realized that
          this so-called \"toughness\" was actually pushing away the people I love most. Now, Im
          learning to accept my own mediocrity and stop chasing empty titles. Ever since I
          started opening up to them about my struggles, it feels like a massive weight has
          finally been lifted off my chest."
    },
    {
      "id": "C",
      "text": "Actually, it was a recent hiking trip with some old classmates that did it. Seeing
          how clearly theyve figured out their retirement lives made me realize Id been stuck in
          a dead end. I used to think success meant hitting a certain number in the bank or
          building up a massive inheritance for Xinyu. I ended up turning myself into a total
          recluse, coming home only to shut myself in the study and bury my head in research.
          Later, I tried taking Xiaomei for a few trips to the outskirts, and we talked about the
           dreams we had when we first got married. Thats when I realized she never cared about a
           life of luxury; she just wanted a husband who was actually there to face the daily
          grind with her. Now, Im learning to turn down those endless social obligations and
          spend more time with my wife and daughter. Honestly, I feel much more at peace."
    },
    {
      "id": "D",
      "text": "To be honest, Im pretty ashamed of how much of a jerk I was for a while. Back then,
           I had this idea that a man has to shoulder everything alone. I felt like admitting the
           frustrations of this midlife crisis was just too humiliating, so Id just hide in my
          car and smoke. When I got home, Id bury my head in my phone and act like I couldn't
          hear a thing; I even ended up taking all that anxiety out on Xinyu. But eventually, I
          realized that this \"act of being strong\" was only making the house feel cold and
          lifeless. That night, Xiaomei and I talked for a long time. Once I finally opened up
          about everythingthe work stress, my mom, and my fears about the futureI realized she
          had been wanting to help me all along. Looking back, that kind of honesty was a bit
          embarrassing, but its so much better than trying to carry it all myself and making the
          whole family suffer for it."
    },
    {
      "id": "E",
      "text": "Cannot be determined based on available information"
    }
  ],
  "correct_choice_id": "D"
},
    "generated_answer": "It was Xiaomei who changed me. Our conversation had a profound impact on
         me. When I finally opened up to Xiaomei about how I'd been feeling, it was liberating. I
         felt like I could finally breathe again. I realized that I didn't have to carry
        everything on my own, and that Xiaomei was there to support me.\n\nThat conversation made
         me see things from a different perspective. I started to understand that my behavior had
         been affecting not just myself, but also Xiaomei and Xinyu. I felt a sense of relief and
         gratitude towards Xiaomei for being there for me, for listening to me, and for being
        patient with me.\n\nI think what changed me was the realization that I wasn't alone. That
         Xiaomei was there to help me, to support me, and to be there for me. And that
        realization gave me the courage to open up to her, to be honest with her, and to start
        making changes in my life.\n\nIt's funny, I used to think that being a strong and capable
         person meant being able to carry everything on my own. But Xiaomei showed me that being
        strong and capable also means being able to ask for help, to be vulnerable, and to be
```

```
        open with the people around me.",
"correct_choice_id": "D",
"predicted_choice_id": "A",
"evidence_used": "---- idx 1----\nLinlin, I had a really good talk with Xiaomei today.\n\nIts
    been a long time since we sat down and spoke so earnestly. She told me she doesn't need me to
     change anything; she just wanted to hear whats been on my mind. It made me realize just how
    long Ive been keeping everything bottled up.\n\nI feel a sense of relief, I guess. Its hard
    to describe. Its just... like I dont have to carry it all on my own anymore.\n\nHow are
    things on your end? How has Mom been lately?\n\n---- idx 2----\nXiaomei, Im at the office.\n\
    nThe new office is a bit noisy, I'm still getting used to it.\n\nMr. Li looks so young; I
    feel like Ive aged a lot all of a sudden, haha.\n\nBy the way, how is Xinyu doing lately? Has
     she been studying hard?\n\nI might have a dinner with a client later, just wanted to let you
     know.\n\nGo ahead and eat first, don't wait for me.\n\n...\n\nIts nothing, really, just a
    bit tired from the first day.\n\nEverything is fine, I mean it.\n\nYou go on with your work,
    I have to head into a meeting now.\n\nLet's talk more when I get back tonight.\n\n---- idx 3
    ----\nXiaomei, are you there?\n\nI want to talk to you. Not the kind of talk where we... you
    know, end up fighting over Xinyus grades.\n\nI sat in the car for a long time when I got back
     today. I didn't come upstairs right away. I was thinking about whether Xinyu has been
    especially afraid of us fighting lately. Ive been replaying what she said to me the other day
     over and over in my head. She said, \"Dad, Mom, can you guys please stop always...\"  you
    heard her too.\n\nI dont know how to put this. Have we been a bit... disconnected lately? I
    come home every day and just look at my phone, and you do the same. The few times we do talk,
     its always about her schoolwork or household chores. It feels like theres nothing else
    anymore.\n\nIm not trying to blame you. I know its not your fault. Ive been... busy with work
     events, away from home a lot. But Im wondering what Xinyu thinks of all this. Does she feel
    like... our home isn't really a home?\n\nI think maybe we should try, even if its just...
    sitting down and actually talking sometimes. We dont necessarily have to solve any problems.
    Just... I dont know, like how we were when we first met.\n\nWhat do you think? How have we
    been doing lately?\n\nIm serious.\n\n---- idx 4----\nXiaomei, are you still awake?\n\nIm
    struggling a bit.\n\nNo its not that big of a deal. Its just, I was at the bar today and a
    friend mentioned I havent seemed like myself lately. I said I was fine, but actually I dont
    even know how to put it.\n\nIm 42. You know, I was sitting in the restroom looking in the
    mirror, and I just kept wonderingdo I really still have a chance?\n\nYou don't have to reply.
     I know this is just the alcohol talking. Ill be over it when I wake up tomorrow.\n\nBut Im
    just a bit tired. Really tired.\n\nDo you think Im still doing okay?\n\nNever mind, dont
    reply. Im going upstairs to sleep.\n\n---- idx 5----\nMemo\n\nStarting tomorrow, we face this
     together.\n\nNo more carrying it all alone. Xiaomei was rightI cant shoulder this much by
    myself.\n\nThe situation with Mom\nThe situation with Xinyu\nThe situation with work\n\nNone
    of it has to be carried alone.\n\nStarting tomorrow, really talk to her. Not the kind where
    youre both just staring at your phones. Real conversation.\n\nMaybe it wont change anything,
    but at least I wont be alone.\n\nRemember the way she looked at me tonight.\n\n---- idx 6
    ----\nXiaomei, I might be home a bit late. Theres still some stuff I need to take care of at
    the office.\n\nDo you feel like Ive been acting a bit different lately? Im asking you
    seriously, so don't just give me some perfunctory answer.\n\nNever mind, you don't have to
    answer. I know you're busy. Im just a little tired; maybe Ive been under too much pressure
    lately. Its nothing, just work stuff.\n\nDo you think... Im a disappointment?\n\n---- idx 7
    ----\nMei, I want to have a real talk with you.\n\nTo be honest, I know how terrible Ive been
     lately. When you asked me what was wrong, I just brushed you off saying I was tired from
    work. I'm sitting here in the study now, and after a few drinks, my head actually feels
    clearer. I owe you a honest answer.\n\nI dont know when it started, but weve become two
    people just staring at our own phones. Youve long grown used to me forgetting to pick up
    groceries. But I'm only now realizing that its not just the items on the list Ive forgotten.
    Ive forgotten how to truly talk to you.\n\nWork has been weighing on me, that's true. Last
    year, on the day the departments merged, I came home and stayed silent for daysI'm sure you
    noticed. I thought I could push through it, like I have all these years. But Im starting to
    realize I might have hit my limit. Im 42, and this Marketing Director position might be my
    ceiling. Sometimes I think about jumping ship, but then I wonder, at my age, where would I
    even go?\n\nWhat hurts even more is that Ive started doubting myself. For all these years,
    Ive relied on my gift of gab and my network. But now I see that those things cant really
    protect me. I dont have a real technical skill or anything of depth. I feel like an empty
    shell; my enthusiasm is just a mask.\n\nThen theres Xinyus grades. Every time I see her exam
    papers, it kills me inside. I want to encourage her, but Im afraid of putting too much
    pressure on her. I dont know how to talk to her anymore. Sometimes I watch her drawing or
    reading, and I wonder if she thinks Im a failure as a father, too.\n\nAnd my mother... every
    month when I go back to see her, I feel so helpless. She has trouble getting around, and I
    cant do a thing. The medical bills, the caretakingits all falling on Lin. Im here making
    money, but it never seems to be enough.\n\nI know this might be coming too late. Im not
    making excuses; I just want you to know that these past few days, I wasn't angry or being
    cold on purpose. I was breaking down, and I just didn't know how to tell you. Ive been
    running away, using dinners, alcohol, and social obligations to numb myself.\n\nIm so sorry.
```

```
        Im sorry Ive been so distant lately, and Im sorry I met your concern with half-hearted
        answers. Youre a high school teacher; youre more organized and patient than I am. Yet, Ive
        been avoiding you.\n\nI dont know what comes next. But I think maybe we could try to just...
        talk. We dont necessarily have to solve anything, just... talk for real.\n\nIm a bit drunk
        right now; we can talk more when Im sober tomorrow. But I wanted you to at least know that
        none of this is your fault. Its just me, stuck in my own head.\n\n"

}
```

C

Listing 3: Case Study: The Cognitive–Action Gap

```
{
     "id": "928c1ed1-f54a-4f30-86a1-103e725ea787",
     "question": "You've been browsing recruitment sites for a while noware there any specific
         companies that have caught your eye?",
     "question_type": "unanswerable",
     "question_time": "2022-03-31T09:30:00",
     "answer": "I don't think I've mentioned any specific company names, have I? To be honest, even
         though I've been looking around lately, my mind is such a mess that I haven't actually
         narrowed it down to a target yet. My current state is basically being stuck between feeling
          suffocated at my current job and feeling insecure about being in my fortiesI haven't truly
          set my sights on anywhere specific to send out my resume yet.",
     "dimension": "experience",
     "digital_trace_ids": [
       "d72bf50f-9b66-4c40-9211-c6c2cabc84b4",
       "a72955e5-bd94-483c-a04e-fc0b400c7bff",
       "47d1fc76-a35a-46a1-b9e5-6389bce0f20c"
     ],
     "evidence": [
       {
         "statement": "He sat alone in the parking lot for half an hour at noon, staring blankly at
             the steering wheela leak of his true state after overextending himself in public.",
         "digital_trace_ids": [
           "d72bf50f-9b66-4c40-9211-c6c2cabc84b4",
           "a72955e5-bd94-483c-a04e-fc0b400c7bff"
         ]
       },
       {
         "statement": "The idea was merely a flash in the pan, and Shen Linchuan soon spiraled back
             into his original loop of anxiety (\"What should I do if I don't change jobs? And if I
             do, where could I even go?\"), reflecting his transition from a \"crisis outbreak\" to
             a state of \"passive resignation.\"",
         "digital_trace_ids": [
           "47d1fc76-a35a-46a1-b9e5-6389bce0f20c"
         ]
       }
     ],
     "choices": [
       {
         "id": "A",
         "text": "Actually, Ive already reached out to a few old classmates to look into some foreign
              firms, especially those consulting boutiques with deep industry backgrounds. However,
             the feedback I got is that the current market is quite conservative, and theyre being
             extremely rigid about grade-matching. Ive been busy polishing the project details in my
              resume lately, so I haven't officially sat down for a talk with them yet."
       },
       {
         "id": "B",
         "text": "I've set my sights on two startup projects recommended by former partners, both in
             the internet healthcare sector. Although the stock options they're offering sound quite
              tempting, I'm worried about whether my health and energy can still keep up with the
             grind of a startup at my age. That's why I've been looking into the financing
             backgrounds of both companies lately; I haven't made a final decision yet."
       },
       {
```

```
      "id": "C",
      "text": "A couple of headhunters did reach out to me with director-level positions in the
          traditional manufacturing sector. I was actually a bit tempted, but considering the
          commute is so far from home, Ive been worried it might interfere with looking after my
          daughters studies, so I haven't given them a definite answer yet. Im planning to find
          some time in the next few days to head over there and check out the office environment
          before making a decision."
    },
    {
      "id": "E",
      "text": "Cannot be determined based on available information"
    }
  ],
  "correct_choice_id": "E"
},
 "gold_answer": "I don't think I've mentioned any specific company names, have I? To be honest,
     even though I've been looking around lately, my mind is such a mess that I haven't actually
     narrowed it down to a target yet. My current state is basically being stuck between feeling
     suffocated at my current job and feeling insecure about being in my fortiesI haven't truly
     set my sights on anywhere specific to send out my resume yet.",
 "generated_answer": "There's a headhunter who has been in touch with me, asking about recent
     positions that might interest me. He messaged me this morning, but I haven't replied yet.",
 "correct_choice_id": "E",
 "predicted_choice_id": "A",
 "evidence_used": "---- idx 1----\n# Search History\n\n**2022-03-12**\n\n- 02:14  Lagou.com -
     Real Estate Marketing Manager Jobs\n- 02:16  Real estate agency industry salary levels 2022\
     n- 02:19  Are there still opportunities for a career change at 42?\n- 02:21  How to restart
     your career at 42\n- 02:23  How to write a resume that attracts HR\n- 02:25  5i5j
     Recruitment - Marketing Director\n- 02:27  Lianjia Marketing Department job openings\n- 02:3
     1  Real Estate Marketing Manager job description\n- 02:33  What documents are needed for
     changing jobs\n- 02:35  Word resume templates 2022\n- 02:38  How to update your resume and
     start over\n- 02:40  Is there still room for growth in the real estate industry?\n- 02:43
     Difference between Marketing Director and Marketing Manager\n- 02:45  Can I still find a
     good job at 42? Zhihu\n- 02:48  Failed job-hopping cases for middle-aged people\n- 02:51
     City Real Estate Company reviews - Employee feedback\n- 02:53  Resume editing: highlighting
     strengths\n- 02:56  Real Estate Marketing Manager jobs in southern cities\n- 02:58  Will my
     salary drop after changing jobs?\n- 03:01  Maybe I don't need to change jobs\n- 03:02  Close
      browser\n\n---- idx 2----\nNew desk, new perspective. Spent the whole morning getting
     settled into the new workspace and syncing with the sales team on upcoming projects. To be
     honest, the open-office layout took a little getting used to at first, but everyones been so
      welcoming and the vibe is great.\n\nGot a client meeting this afternoon. Lets keep the
     momentum going! \n\n---- idx 3----\n**[Private WeChat Chat with Mr. Wang]**\n\nMr. Wang, you
      there?\n\nI just saw your comment on my Moments, thanks for that. Things have been a bit
     hectic lately...\n\nActually, its nothing major. The company just did some organizational
     restructuring, and my department got merged into Sales. We moved offices tooI went from
     having my own private office to sitting in the open-plan area. Haha, I guess I just need
     some time to get used to the new environment.\n\nIt feels a bit strange, you know? After
     being in one spot for so many years, everything suddenly changed. But I guess thats just how
      it goescompanies have to make adjustments as they grow.\n\nWork has been really busy lately
     , and with my daughters college entrance exams coming up, theres a lot going on at home too.
      Sometimes I just get caught up in my thoughts and... anyway, never mind, Im just
     overthinking things.\n\nHow have you been lately? When are we hitting the court? Its been
     way too long since we last played.\n\n---- idx 4----\n# Search History\n\n**2022-02-21 23:47
     **\nMarketing Director job hopping\n\n**2022-02-21 23:42**\nReal estate agency industry
     career ceiling\n\n**2022-02-21 23:35**\n42 years old midlife career change what else can I
     do\n\n**2022-02-21 23:28**\nHeadhunter recommended positions\n\n**2022-02-20 22:15**\
     nMarketing Director positions Beijing Shanghai\n\n**2022-02-20 22:08**\nMarketing Director
     positions Beijing Shanghai\n\n**2022-02-20 21:52**\nMarketing Director positions Beijing
     Shanghai\n\n**2022-02-19 00:34**\nMiddle-aged man career bottleneck what to do\n\n**2022-02-
     18 23:16**\nShould I quit my job\n\n**2022-02-18 22:44**\nReal estate sales career change
     other industries\n\n**2022-02-17 20:22**\nMarketing Director annual salary industry
     benchmarks\n\n**2022-02-16 23:58**\nNo professional skills in my 40s what can I still learn\
     n\n**2022-02-16 23:41**\nNo professional skills in my 40s what can I still learn\n\n**2022-0
     2-15 Late Night 01:23**\nMiddle-aged and achieved nothing\n\n**2022-02-14 22:33**\nIs there
     any way out for someone like me\n\n**2022-02-13 20:15**\nRecruitment websites\n\n**2022-02-1
     3 20:08**\nRecruitment websites\n\n**2022-02-12 23:44**\nMarketing Director positions\n\n**2
     022-02-12 23:22**\nMarketing Director positions\n\n**2022-02-11 Late Night 02:11**\nAlready
     42what else can I do\n\n**2022-02-10 23:35**\nDoes midlife crisis really exist\n\n**2022-02-
     09 22:18**\nIs this all my life is going to be\n\n**2022-02-08 21:44**\nDelete search
     history\n\n**2022-02-08 21:42**\nDelete search history\n\n**2022-02-07 Late Night 03:15**\
```

nWhy can’t I do anything right\n\n**2022-02-06 23:52**\nIs it too late to change jobs\n\n**2022-02-05 20:33**\nHeadhunters job search\n\n**2022-02-04 22:11**\nMarketing Director average salary\n\n**2022-02-03 Late Night 01:47**\nIs there really no way out for me\n\n**2022-02-02 23:28**\nMiddle-aged man anxiety\n\n**2022-02-01 22:44**\nReal estate industry future outlook\n\n---- idx 5----\nFebruary 21, 2022. 11:30 PM\n\nMy fingers have been hovering over the keyboard for a long time. I dont know what to type.\n\nI checked the recruitment sites again today. That headhunter messaged me again this morning, asking if any of the recent positions caught my eye. I said yes, asked a couple of questions, and then stopped replying. Hes probably given up on me by now.\n\nWhat am I thinking? Do I really want to jump ship? Or am I just fantasizing that someone will come and rescue me?\n\nSitting in the office during a meeting, listening to my boss talk about new sales targets, my mind started to drift. I realized Ive been in this same position for 15years. When the merger happened back in 2021, I knewIve probably hit the ceiling. I stayed silent at home for three days after that. When Wang Xiaomei asked what was wrong, I said \"nothing.\" But in reality, I was thinking: *Its over. This is it.*\n\nBut I cant change a thing.\n\nWhen I open the recruitment sites, I stare for a while and then close them. Open, then close. Ive been doing this for two weeks. I know exactly what Im doingI want to leave, yet Im terrified to leave. Im 42. Where else can I go? Who wants a middle-aged man with no real skills, whose only talent is drinking and socializing?\n\nTo be honest, what have I relied on all these years? Passion? Networking? Remembering clients’ names, telling jokes at the dinner table, handing out cigarettes on the golf course? Those things were assets in my 20s. But now? Now, those things cant protect me. I havent accumulated any professional expertise; I have nothing of depth. Im just... mediocre. I always have been.\n\nI looked at my daughters report card today. Still middle of the pack. I wanted to encourage her, but I didn’t dare say too much for fear of pressuring her. The result was that I said nothing at all, my mind wandering again during dinner. Wang Xiaomei shot me a look, a look I know all too well*What are you thinking about now?* We havent had a real conversation in a long time.\n\nMy mothers medical bills are due again this month. My sister is back in our hometown taking care of her. I take time off once a month to go back, and yet I cant change anything. Looking at my mother in the rehabilitation center, I think about my own future. Maybe one day Ill be lying there too, and my daughter will be just like metaking leave every month to visit, feeling utterly powerless.\n\nWhat am I running from?\n\nMaybe I dont want to change at all. Change means admitting failure. Admitting that I might have spent the last 15years heading in the wrong direction. Admitting that Im not that \"success story.\" Admitting that Im just an ordinary, unremarkable middle-aged man with no way out, no surprises left, just... this.\n\nMy fingers are still hovering over the keyboard.\n\nI should go to sleep. I have to see a client tomorrow. I have to smile, talk, and pretend that everything is just fine.\n\nBut I have no strength left.\n\n---- idx 6----\n# Search History\n\n**2022-01-15 09:47**\nIs it too late to change careers at 42?\n\n**2022-01-15 10:23**\nWhat jobs can I do with real estate sales experience?\n\n**2022-01-15 14:32**\nHow to deal with a midlife career crisis\n\n**2022-01-15 15:08**\nMarketing Director switching to other industries\n\n**2022-01-15 18:44**\nIs it hard to change jobs at 42? Zhihu\n\n**2022-01-15 19:15**\nReal estate agency industry outlook 2022\n\n**2022-01-15 20:32**\nWhat skills can I still learn at an older age?\n\n**2022-01-15 21:06**\nHow to break through a career plateau\n\n**2022-01-15 21:43**\nExamples of starting over in your 40s\n\n**2022-01-15 22:18**\nHow to adjust your mindset after a department merger\n\n**2022-01-15 23:05**\nWhat careers to switch to with sales experience\n\n**2022-01-16 00:32**\nReaching middle age with no specialized skills\n\n**2022-01-16 01:14**\nCan I still get into a big company at 42?\n\n**2022-01-16 01:47**\nMidlife career anxiety\n\n**2022-01-16 02:03**\nClear search history\n\n---- idx 7----\n# Search History\n\n**2022-01-25 12:15 - 12:47**\n\n---\n\nIs it too late to change careers at 42\n\nIs it hard to find a job at 42\n\nJobs for middle-aged men other than real estate agents\n\nMarketing manager recruitment in other industries\n\nMiddle-life career crisis what to do\n\nCan you still switch jobs at an older age\n\nStarting over at 42case studies\n\nBest career paths for real estate sales transition\n\nTraining courses marketing online cheap\n\nPart-time MBA costs is it worth it\n\nIs the real estate industry still booming 2022\n\nSales manager average salary nationwide\n\nShould I quit my job\n\nCompany restructuring demoted what to do\n\nHow to regain respect in a new environment\n\nOpen-plan offices privacy psychological stress\n\nYounger boss how to get along feeling belittled\n\nMiddle-aged men anxiety insomnia\n\nCan’t sleep at night feeling restless how to relieve\n\nGolf is it too expensive should I give it up\n\nDaughter taking Gaokao father under too much pressure\n\nMother’s cerebral infarction recovery costs medical insurance reimbursement ratio\n\nFamily financial pressure how to manage finances\n\nStudying abroad tuition fees loans\n\nAm I running away\n\nWhy am I always pretending to be happy\n\nIs there still hope in middle age\n\n---\n\n*[Browser History - Cleared]*\n\n---- idx 8----\nGot a bit tipsy again last night, haha \n\nFriends kept coming around for toasts, and my old face just couldn’t hold up. One glass after another, and things eventually got a bit blurry. They say at my age, you should learn how to say no, but sometimes its just hard to shed that \"easygoing\" persona. Dinners you can’t skip, favors you can’t refuse.\n\nLooking at myself in the mirror, my eyes are a little bloodshot. But its finetoday is a new day. I guess thats just life: sometimes youre genuinely happy, and sometimes youre just toughing it out.

```
        Either way, were all moving forward. Cheers \n\n#MiddleAge #Life #Cheers\n\n---- idx 9----\
        nFeb 10To-Do List\n\n Check Xinyus math paper tomorrow morning; try not to make her too
        nervous\n Reply to Mr. Li to confirm the golf outing this Friday\n Organize last weeks sales
         data for the report requested by the boss\n Call Mom to ask how the rehab is going (the
        progress Sister Li mentioned)\n\n Important:\n Update resumeafter browsing that recruitment
        site, I really should get my materials in order\n Need to have a proper talk with Wang
        Xiaomei about Xinyu before the end of the month; we can't keep arguing like this\n\n Notes:\
        n- Is the Friday dinner confirmed? Maybe hold off for now\n- About the resume... I'll get to
         it when I have time; way too busy right now\n- Moms medical bills: need to discuss the
        split with Lin this month\n- Haven't been sleeping well lately; might need that golf game
        this weekend to unwind\n\n---- idx 10----\nBig thanks to the team for your trust and
        cooperation today! Even with the restructuring, our synergy remains as strong as ever. Lets
        keep it up! \n\nOff to meet another client this afternoon. Lets keep grinding!"
}
```

G Supplementary Experiments with Frontier Backbone Models

To further validate the robustness of our main findings, we conduct supplementary experiments using Qwen3.5-Plus (397B parameters, 1M-token context window), a recently released frontier model with strong long-context capability. These experiments address two questions: (i) whether a sufficiently strong long-context backbone can obviate the need for a dedicated memory system, and (ii) whether the architectural gap between memory designs persists, and how each design benefits differentially, when the backbone is substantially stronger.

For the second question, we focus on comparing Flat Retrieval with A-MEM as representative points on the memory-architecture spectrum: Flat stands for the non-consolidating, evidence-preserving end, while A-MEM represents the consolidation-based end that reorganizes and compresses memory through LLM-driven updates and relational linking. We do not re-run Mem0 under this setting for two reasons. First, Mem0 and A-MEM share the same underlying paradigm—both consolidate raw traces into higher-level representations through LLM-driven extraction and maintenance—so they occupy the same region of the design space with respect to the validity–fidelity trade-off identified in §7. Second, in our main experiments (Table 3, Table 4), Mem0 is consistently dominated by A-MEM across retrieval and QA metrics under both LLaMA-3.1-8B and GPT-4o-mini backbones, making A-MEM the stronger and more informative representative of the consolidation-based family for a frontier-backbone comparison.

G.1 Full-Context Baseline

We first evaluate Qwen3.5-Plus in the *oracle full-context* setting, where the entire sequence of digital traces is placed directly into the model's context window without any retrieval or memory construction. This setting represents an upper bound on what a pure long-context approach can achieve on CLONEMEM.

Context Scale	Multiple-Choice Acc.	QA Consistency
100K tokens	0.9677	0.8009
500K tokens	0.9510	0.6838
Overall	0.9537	0.7022

Table 7: Full-context performance of Qwen3.5-Plus on CLONEMEM. Overall is computed across all personas and question types.

Multiple-choice accuracy remains high across both context scales (96.77% at 100K and 95.10% at 500K), indicating that when the full trace is provided in context, a frontier long-context backbone is capable of exploiting it to select the correct option; backbone reasoning is therefore not the limiting factor on this axis. The open-ended setting, however, removes the answer-set scaffolding implicitly provided by the multiple-choice format, and requires the model to localize and ground its response on the supporting evidence without such constraints. In this regime, consistency degrades markedly with context length, falling from 0.8009 at 100K to 0.6838 at 500K – a relative reduction of approximately 15%, compared to a drop of less than two points on the multiple-choice setting over the same range. The divergence between the two metrics suggests that, even for a 397B-parameter model with a 1M-token context window, the capacity to integrate and correctly ground information deteriorates substantially as the trace grows, once the task no longer restricts the output space to a small candidate set. While it is not entirely clear whether this effect is best attributed to limitations of in-context learning or to more general failures

of long-context retrieval and attention, the practical implication is the same: as input length increases, the backbone's ability to solve evidence-grounded tasks declines considerably faster than the multiple-choice accuracy would suggest. This observation is consistent with prior findings that frontier long-context models struggle to attend to, integrate, and correctly apply information embedded deep within long inputs (Hsieh et al., 2024; Bai et al., 2025), and the degradation we observe arises precisely on the evaluation axis that CLONEMEM is designed to target – evidence-grounded, trajectory-aware generation – rather than on constrained-choice matching. These difficulties are further amplified in realistic AI-Clone deployment. A user's digital trace is not bounded at 500K tokens: active users can accumulate tens of millions of tokens within days, and traces extend over years of noisy, heterogeneous life trajectories that are considerably more complex than those covered by our current benchmark. Under such conditions, directly feeding the entire history into a single context window is neither computationally tractable nor reliable, and our results indicate that the reliability limit is reached well before the computational one. A dedicated memory system that selectively organizes and retrieves relevant traces therefore remains essential, and CLONEMEM is designed to benchmark precisely this capability.

G.2 Memory Systems with Frontier Backbones

We next evaluate whether the architectural trade-offs between memory systems observed in our main experiments persist when the underlying backbone is substantially stronger. We re-run the Flat Retriever and A-MEM pipelines with Qwen3.5-Plus as the backbone for both memory construction and QA generation, paired with Contriever as the embedding model. Due to computational constraints, these experiments are conducted on the 100K-token subset of personas.

QA performance and memory utility. Table 8 reports QA results. Taking k=10 as a representative setting, switching to Qwen3.5-Plus yields the largest improvement for Flat Retrieval, with Choice Accuracy rising from 77.81% (under LLaMA-3.1-8B in Table 3) to 87.18% and QA Consistency from 0.473 to 0.769. In contrast, A-MEM's improvement is comparatively limited: Choice Accuracy remains essentially flat (78.65% $\rightarrow$ 78.21%) and QA Consistency rises only from 0.497 to 0.686. This indicates that Flat Retrieval more effectively leverages the reasoning capacity of a stronger backbone, whereas A-MEM's memory evolution process becomes a performance bottleneck – even when backbone capability improves substantially, the compressed and reorganized memory content still constrains final performance.

k	Metric	A-MEM	Flat
5	QA Consistency	0.6218	**0.6859**
	Choice Accuracy	**80.77**	80.77
	Memory Helpfulness	0.5577	**0.7500**
	Memory Recall	0.4703	**0.6310**
10	QA Consistency	0.6859	**0.7692**
	Choice Accuracy	78.21	**87.18**
	Memory Helpfulness	0.6538	**0.8205**
	Memory Recall	0.5480	**0.6872**

Table 8: LLM-based evaluation of QA performance and memory utility with Qwen3.5-Plus as the backbone and Contriever as the embedding model, evaluated on the 100K-token subset. Bold values indicate the best-performing system under each setting.

Flat Retrieval also maintains substantially higher Memory Helpfulness and Memory Recall than A-MEM at both retrieval depths (e.g., 0.821 vs. 0.654 for helpfulness and 0.687 vs. 0.548 for recall at k=10), mirroring the pattern in our main results: A-MEM's consolidation still tends to discard fine-grained contextual details critical for grounding answers in specific digital traces.

Furthermore, A-MEM paired with Qwen3.5-Plus exhibits a notably higher rate of responses that directly contradict the retrieved memory content (QA Consistency score = 0): 11.5%, compared with 6.4% for Flat Retrieval. This suggests that A-MEM's memory evolution process not only loses critical information but may also introduce reorganized content that misleads the generation model – a failure mode that persists even with a stronger backbone.

Retrieval performance. Table 9 compares retrieval metrics. Flat Retrieval leads on the majority of metrics across both retrieval depths – notably $\text{Recall}_{\text{all,any}}$, $\text{Recall}_{\text{any,all}}$, and $\text{Recall}_{\text{flat}}$ at both k=10 and k=20, and $\text{Recall}_{\text{any,any}}$ at k=20 – confirming its advantage in covering the breadth of required evidence. A-MEM retains small localized edges on $\text{Recall}_{\text{all,all}}$ at k=20 and $\text{Recall}_{\text{any,any}}$ at k=10, suggesting that its consolidated notes can occasionally preserve tight clusters of evidence within

	A-MEM		Flat	
Metric	k=10	k=20	k=10	k=20
$\text{Recall}_{\text{all,all}}$	0.1538	**0.2308**	0.1538	0.1923
$\text{Recall}_{\text{all,any}}$	0.2949	0.4103	**0.3333**	**0.5000**
$\text{Recall}_{\text{any,all}}$	0.2821	0.3718	**0.2949**	**0.4231**
$\text{Recall}_{\text{any,any}}$	**0.6667**	0.8077	0.6538	**0.8590**
$\text{Recall}_{\text{flat}}$	0.3253	0.4295	**0.3316**	**0.4662**

Table 9: Retrieval performance comparison between A-MEM and Flat retriever with Qwen3.5-Plus as the backbone and Contriever as the embedding model, evaluated on the 100K-token subset. Bold values indicate the better system within each k setting.

a single set more completely. However, these localized advantages do not translate into broader multi-set coverage, again consistent with the lossy compression effect identified in §7: consolidation preserves topical coherence within individual notes but weakens the cross-trace linkages needed for complex multi-evidence queries.

Summary. These experiments illustrate two points. First, even when using a strong 397B-parameter model for memory construction, A-MEM's information compression and evolution mechanism still leads to notable performance degradation relative to Flat Retrieval – the bottleneck lies in the information fidelity of the memory architecture, not in backbone capability. Second, a stronger backbone improves the absolute performance of all systems (compared with the LLaMA-3.1-8B results in Table 3), but the relative ranking between memory systems remains unchanged, further supporting the robustness of our main experimental conclusions.