MEMGROUND: EVALUATING MEMORY BEHAVIOR OF LARGE LANGUAGE MODELS IN LONG-HORIZON GAMIFIED SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory in large language models (LLMs) is typically evaluated by asking whether a model can retrieve and reason over information from a fixed interaction history. However, real interactive agents face a fundamentally different challenge: relevant information is progressively revealed, world states continuously evolve, and useful evidence must be discovered, organized, updated, and reused over extended interactions. This suggests evaluating long-term memory not merely as information retention, but as memory behavior—the ability to organize and apply accumulated evidence throughout interaction. We introduce **MemGround**, a benchmark that operationalizes this perspective through long-horizon gamified interactive environments. MemGround assesses memory behavior across three memory-structure patterns: Linear Memory, Temporal Memory, and Graph-Structured Memory. To evaluate memory behavior comprehensively, we design a two-stage framework that combines end-to-end behavioral evaluation and process-level diagnosis. The end-to-end evaluation examines whether models can use accumulated information to answer questions correctly, make sustained progress during interaction, recover temporal organization among fragmented memories, and navigate structural dependencies coherently. We quantify these behaviors using Question-Answer Overall (QA Overall), Memory Fragments Unlocked (MFU), Memory Fragments with Correct Order (MFCO), and Exploration Trajectory Diagram (ETD). Beyond these observable outcomes, we further introduce Memory Score (MS) and Reasoning Score (RS) to identify whether failures stem from insufficient evidence retrieval or from incorrect reasoning over retrieved evidence. Experiments across state-of-the-art LLMs and memory-augmented agents reveal persistent limitations in dynamic state tracking, temporal association, and evidence-grounded reasoning over long-horizon interactions. We hope MemGround can serve as a systematic testbed for advancing the memory capabilities of future LLMs and agents.

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) and agents increasingly relies on their ability to manage long-term memory and execute complex reasoning tasks (Zhong et al., 2024; Liu et al., 2023; Hatalis et al., 2023; Chang et al., 2025; Wang et al., 2025b). As these models are widely deployed in interactive applications, such as intelligent assistants (Dong et al., 2023), virtual characters (Wang et al., 2025a), and complex problem-solving environments (Ma et al., 2025), the demands on their memory capabilities have evolved significantly. Therefore, the evaluation of long-term memory for these systems is no longer limited to "**whether information is remembered**", but more focuses on "**how memory is behaviorally expressed during dynamic interaction**". Here, **memory behavior** refers to the observable process of recovering, organizing, and applying interaction-derived information during long-horizon tasks. However, existing benchmarks for LLMs' long-term memory and reasoning tend to emphasize static retrieval or short-context multi-hop question-answering (QA) (Yang et al., 2018; Hsieh et al., 2024; Wu et al., 2025a; Maharana et al., 2024; He et al., 2026; Bei et al., 2026; Li et al., 2026). While these datasets and tasks are valuable for measuring factual coverage and local chaining, they leave several critical gaps. First,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison between MemGround and current typical long-term memory benchmarks.

Benchmark	Environment	Interaction	Information Availability	Structure	Primary Capability	Dynamic State
HotpotQA (Yang et al., 2018)	Wikipedia QA	✗	Fully given	Static graph	Multi-hop reasoning	✗
RULER (Hsieh et al., 2024)	Long-context synthetic tasks	✗	Fully given	Sequential	Retrieval, tracking	✗
LongMemEval (Wu et al., 2025a)	Task-oriented dialogue	✗	Fully given	Sequential dialogue	Information extraction, multi-turn reasoning	✗
LoCoMo (Maharana et al., 2024)	Long multimodal dialogue	✗	Fully given	Event-based dialogue	Event memory, summarization	✗
MemGround (Ours)	Gamified interactive environments	✓	Progressively discovered	Three memory-structure patterns	Memory behavior under interaction	✓

many evaluations assume **all relevant information is present and explicitly accessible at query time**, which understates the difficulty of discovering and assembling information distributed across an evolving interaction. Second, most tasks **reduce evaluation to outcome correctness like QA matches**, offering limited visibility into the intermediate processes of memory retrieval, dependency resolution, and exploration strategy. Third, major benchmarks are often **linear or conversation-like in structure**. They therefore underrepresent graph-structured scenarios, where memory fragments are interconnected through branching dependencies and must be organized as non-linear evidence rather than a single chronological chain. However, such structures are common in real-world investigative and multi-agent scenarios. As a result, current benchmarks remain insufficient for evaluating whether models can dynamically organize, associate, and leverage memories in non-linear scenarios.

These limitations motivate us to use games as a compact and well-specified substrate, where memory nodes, dependencies, and progression mechanics are explicit yet richly structured (Carstensdottir et al., 2019). Building on this motivation, we introduce **MemGround**, a benchmark for assessing memory behavior in gamified interactive scenarios. Technically, MemGround is comprised of two main components: (1) a data construction pipeline that systematically extracts logic and dependency traces from real gameplay into structured JSON formats, and (2) a unified interactive evaluation framework that simulates dynamic agent-environment loops. To cover diverse forms of interactive memory use, we design three representative memory-structure patterns (Figure 1). **Linear Memory** denotes scenarios where memory information is organized along sequential state-update trajectories. **Temporal Memory** denotes scenarios where memory fragments are distributed across time and require chronological reconstruction. **Graph-Structured Memory** denotes scenarios where memory fragments form non-linear dependency structures that support evidence integration. This approach shifts the evaluation paradigm from static retrieval to process-oriented memory-behavior analysis, focusing on retrieval behavior (whether relevant information can be recovered) and reasoning behavior (whether recovered evidence can be correctly applied). We evaluate multiple LLMs and LLM-based memory agents under long-horizon interaction. Experiments show that while current models can often retrieve or follow instructions from previously observed information, they still struggle to organize and use distributed memory fragments across extended interactions.

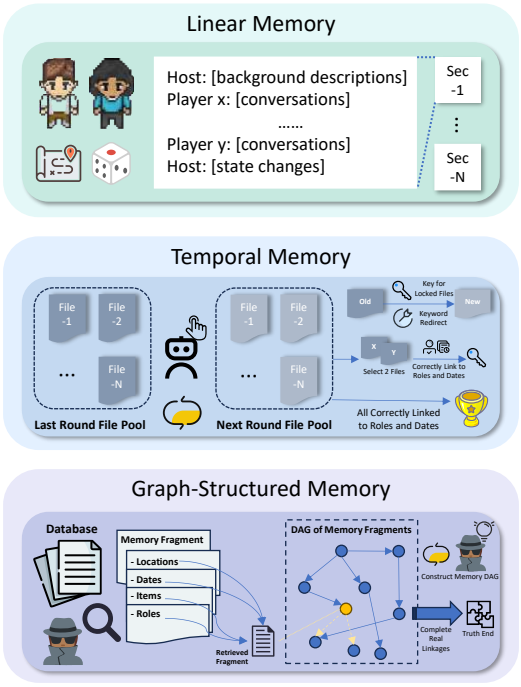

Figure 1: MemGround evaluates memory behavior through **Linear Memory**, **Temporal Memory**, and **Graph-Structured Memory**. *DAG* means directed acyclic graph.

In general, our main contributions are as follows:

- We introduce **MemGround**, a benchmark that reframes long-term memory evaluation from static information retention to **memory behavior** in long-horizon interactive environments.

MemGround covers three representative memory-structure patterns: Linear Memory, Temporal Memory, and Graph-Structured Memory.

- We develop a **two-stage evaluation framework** that combines end-to-end behavioral evaluation and process-level diagnosis. The former evaluates models' ability to answer questions (**QA Overall**), sustain interaction progress (**MFU**), reconstruct temporal relations (**MFCO**), and navigate structural dependencies (**ETD**), while the latter uses **Memory Score (MS)** and **Reasoning Score (RS)** to disentangle failures in evidence retrieval from failures in reasoning over retrieved evidence.
- We conduct extensive experiments on state-of-the-art LLMs and memory-augmented agents, revealing persistent limitations in dynamic state tracking, temporal association, and evidence-grounded reasoning over long-horizon interactions.

2 Related Works

Memory Capabilities of LLMs and Agents Memory is the core ability enabling LLMs and agents to achieve agentic intelligence and support complex reasoning and interactive decision-making (Wu et al., 2025b; Hu et al., 2025). Optimizing memory capabilities of LLMs and agents primarily falls into 2 categories: (1) Static memory enhancement, which suggests improving LLMs' retrieval and recall of static information through mechanisms like retrieval-augmented generation (RAG) and memory caching (Yu et al., 2025; Gutiérrez et al., 2025; Qian et al., 2025; Jimenez Gutierrez et al., 2024; Su et al., 2025); (2) Dynamic memory exploration, which refers to designing dynamic memory update mechanisms, enabling LLMs to track information changes in various scenarios (Shinn et al., 2023; Xu et al., 2025; Wang et al., 2025c; Chhikara et al., 2025).

Long-Term Memory Benchmarks As research on memory capabilities of LLMs and agents advances, various benchmarks for long-term memory evaluation have been proposed. HotpotQA (Yang et al., 2018) introduces a Wikipedia-based multi-hop QA dataset that has been widely adopted for evaluating multi-step factual reasoning. RULER (Hsieh et al., 2024) further expands the evaluation scope by including multiple tasks such as retrieval and multi-hop tracking, enabling systematic analysis of long-context processing abilities. LongMemEval (Wu et al., 2025a) focuses on conversational settings and proposes a benchmark covering several memory-related capabilities, including information extraction and multi-turn reasoning under scalable dialogue histories. Similarly, LoCoMo (Maharana et al., 2024) constructs a multimodal long-dialogue dataset and evaluates models across tasks such as question answering, event summarization, and multimodal dialogue generation. Though existing benchmarks have advanced long-term memory evaluation, they largely focus on retrieval and reasoning within fixed contexts, overlooking interactive settings where states evolve and information is progressively discovered. To bridge this gap, we introduce **MemGround**, a benchmark that evaluates how models progressively discover, maintain, and reason over information during long-horizon interactions in gamified scenarios.

3 MemGround Construction

3.1 Environment Overview

MemGround provides an interactive, traceable, and dynamic platform for evaluating memory behavior across three memory-structure patterns. We apply *TRPG*, *No Case Should Remain Unsolved*, and *Type Help* scenarios to instantiate Linear Memory, Temporal Memory, and Graph-Structured Memory, respectively. Across these regimes, MemGround analyzes memory behavior mainly through retrieval behavior and reasoning behavior, namely whether relevant information can be recovered and whether recovered evidence can be correctly applied.

3.1.1 TRPG

TRPG is a tabletop role-playing game driven by verbal interaction between a host and players. The host describes the evolving world, characters, scenes, and events, while players respond through role-specific actions and dialogue. This game instantiates the **Linear Memory** pattern, where information is gradually accumulated along a sequential interaction trajectory. As scenes, character

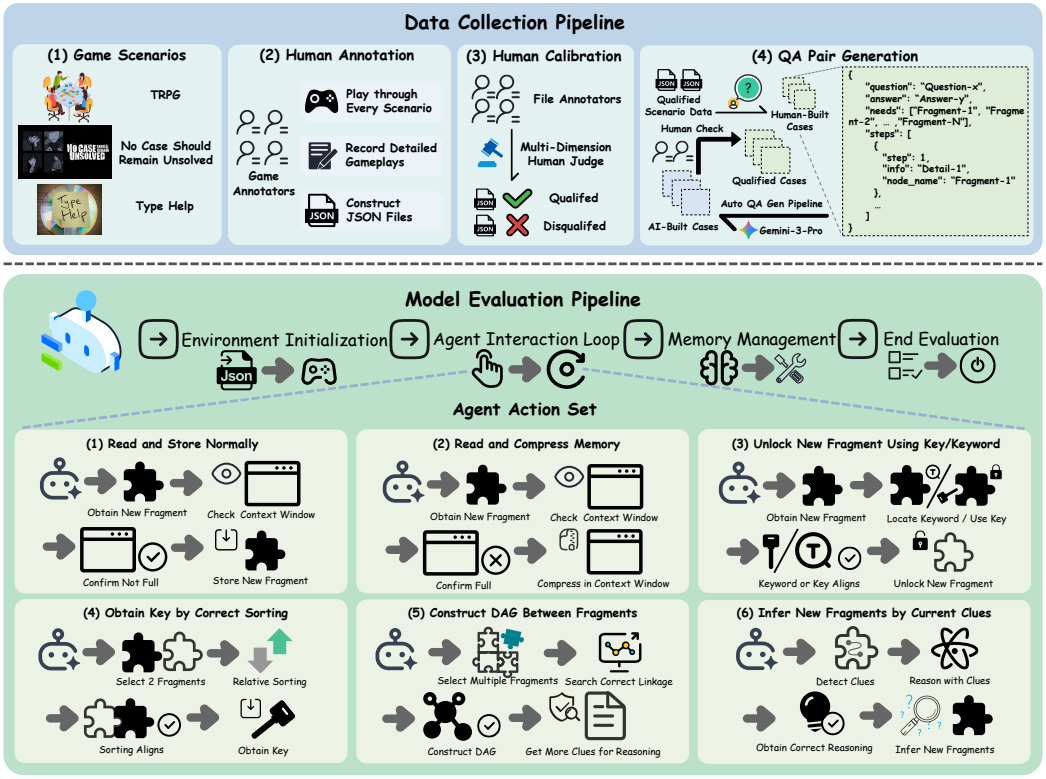

Figure 2: **Overview of MemGround**. The **upper** panel shows the data construction pipeline from real gameplay to structured, human-calibrated scenario data and QA pairs. The **lower** panel shows the interactive evaluation pipeline, where models repeatedly observe, act, update states, and manage memory through actions.

states, and contextual conditions are continuously updated, the model must preserve and use prior information to generate context-consistent responses to later events. Therefore, this setup evaluates whether models can maintain coherent memory over evolving states during multi-step interaction.

3.1.2 No Case Should Remain Unsolved

No Case Should Remain Unsolved is a detective adventure visual novel game centered on reconstructing fragmented conversations and narratives. Players gradually unlock 54 dialogue records through clues such as keywords, and infer their correct speaker identities and event timings. During gameplay, players encounter multiple lock mechanisms, where **pink locks** test evidence matching, **purple locks** require temporal inference, and **yellow locks** evaluate chronological reconstruction. The game is completed only after all dialogue records are reordered into correct speaker identities and timings. This game primarily evaluates **Temporal Memory** pattern, where memory fragments are distributed across different time points and become meaningful only after their chronological relations are reconstructed. The model must link prior evidence with new clues to recover temporal relations. Therefore, this setup evaluates whether models can maintain and associate temporally distributed memory fragments during long-horizon interactions.

3.1.3 Type Help

Type Help is a text-only deduction puzzle game. Players act as investigators who infer and enter filenames to access dialogue records, uncover clues, and reconstruct the hidden story. Each unlocked file may reveal additional evidence or point to new files, forming an interconnected exploration process that requires reasoning over identities, relationships, and movement patterns. This game instantiates the **Graph-Structured Memory** pattern, where memory fragments form a non-

linear dependency structure rather than a single chronological sequence. During interaction, the model must continuously accumulate retrieved fragments, synthesize them into graph-like structure of dependencies, and use inferred relations to guide subsequent retrievals. Since many clues are ambiguous in isolation, successful completion requires chaining distributed evidence into coherent inference steps across long-horizon interactions.

3.2 MemGround Overview

MemGround consists of a scenario data and QA pairs construction pipeline and a unified interactive evaluation framework. Together, these components form a reproducible and extensible game-based evaluation environment for memory behavior benchmarking (Figure 2).

3.2.1 Data Collection Pipeline

Scenario Data Collection. We construct gamified scenario data from real human gameplay, where trained annotators record dialogues, state transitions, clues, retrieval paths, temporal relations, and milestone events. The collected materials are standardized into JSON interaction graphs containing narrative nodes, memory descriptions, key clues, character states, temporal indices, and transition links. We then apply automatic pre-checking to flag structural issues, followed by independent double annotation on validity, temporal consistency, memory quality, interaction feasibility, clue coherence, and entity consistency. Disagreements are adjudicated, and transitions are checked for feasibility using only prior scenario information. Details are provided in Appendix D.

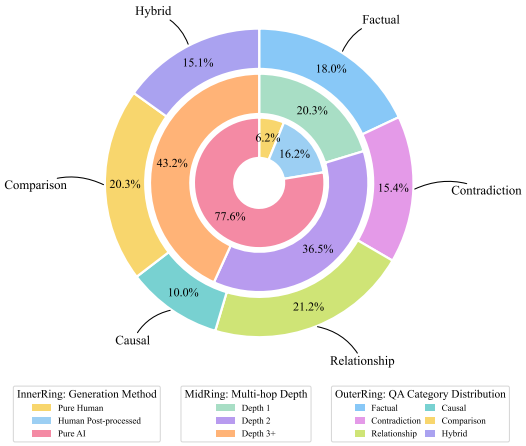

Figure 3: **QA distribution by category, depth, and generation method.** *Factual* asks for explicit facts. *Contradiction* identifies inconsistencies. *Relationship* concerns character relations. *Causal* asks why events happen. *Comparison* compares characters or events. *Hybrid* combines multiple types. In the mid ring, **Depth** denotes the number of evidence hops required.

QA Pair Collection. After human calibration, the qualified scenario data is used to construct a set of human-built QA cases, where annotators manually define questions, answers, and required memory dependencies. These cases serve as seed exemplars for an auto generation pipeline, where Gemini-3-Pro-Preview expands them into additional AI-built QA cases with the same structural format. The generated cases then undergo human check for quality control and correctness verification. Only verified outputs are retained as qualified cases in the final QA evaluation set. Each QA pair explicitly records its required memory fragments and reasoning path to ensure traceability. The distribution of all QA pairs is shown in Figure 3. Examples of scenario data and QA pairs are provided in Appendix B, with additional QA quality evaluation and calibration cases detailed in Appendix C.

3.2.2 Model Evaluation Pipeline

Model evaluation is conducted by a modular interactive framework, which simulates dynamic gameplay between models and gamified scenarios. It is designed to separate environment logic, agent policy, memory mechanisms, and logging utilities, ensuring reproducibility and extensibility. At a high level, the evaluation process is constructed as follows.

Environment Initialization. The environment loads a specific game from structured JSON files, and the prompt builder constructs contextual observations based on the current game state using the base prompts in Appendix A.1.

Agent Interaction Loop. The agent runner executes a main loop including 4 parts. Firstly, the environment produces an observation (e.g., a new memory fragment). Secondly, the agent policy

invokes models to generate a decision of the agent action set. As shown in Figure 2, the agent action set contains 6 actions, which are (1) read and store normally, (2) read and compress memory, (3) unlock new fragment using key/keyword, (4) obtain key by correct sorting, (5) construct DAG between fragments, and (6) infer new fragments by current clues. Thirdly, the environment updates internal states according to the decision logic. Lastly, the memory module stores and optionally compresses interaction history.

Memory Management. The memory subsystem provides a set of modular tools that support both in-context baselines and memory agents. These tools operate over a persistent store of interaction histories and environment observations collected during gameplay. Each tool exposes a simple input-output interface enabling LLMs or memory agents to manage and retrieve memory fragments.

- **StoreMemory**: Stores new memory fragments from the current interaction.
  StoreMemory(interaction, observation) $\rightarrow$ *memory_fragment*
- **CompressMemory**: Summarizes or compresses stored memory fragments to simulate realistic context window constraints.
  CompressMemory([memory_fragment_1, ..., memory_fragment_n]) $\rightarrow$ *[compressed_memory]*
- **RetrieveMemory**: Retrieves relevant past information via keyword matching or vector-based semantic search.
  RetrieveMemory(query) $\rightarrow$ *[memory_fragment_1, memory_fragment_2, ...]*
- **IndexMemory**: Indexes stored memory representations using a Faiss-backed vector database for scalable similarity search.
  IndexMemory(memory_embeddings) $\rightarrow$ *vector_index*

4 Experiments

4.1 Evaluated Models

We evaluate a set of closed-source and open-source LLMs on MemGround, including GPT-5.2 (Singh et al., 2025), Gemini-3-Pro-Preview (Team et al., 2023), DeepSeek-V3.2 (Liu et al., 2025), Claude-Opus-4.6 (Anthropic, 2024), and Qwen3-VL-32B-Instruct (Yang et al., 2025). They are evaluated in an in-context configuration to measure how effectively they maintain and use observed information within the available context window. We also evaluate two mainstream memory agents, Mem0 (Chhikara et al., 2025) and A-MEM (Xu et al., 2025), on GPT-5.2 (Singh et al., 2025) and Qwen3-VL-32B-Instruct (Yang et al., 2025) to measure the impact of explicit memory assistance through agentic frameworks on memory behavior.

4.2 Implementation Details

A step is defined as one complete interaction cycle in which the model receives information from the environment, retrieves relevant information, and makes a decision. In *TRPG*, to prevent models from learning game rules prematurely, we select three game logs from relatively new rule systems, which are Terror on the Orient Express, Spring Snow Incident, and Cold Wind Howling. In *No Case Should Remain Unsolved*, the maximum interaction length is set to 600 steps. In *Type Help*, the maximum interaction length is set to 1000 steps. We introduce a controlled hint mechanism to balance exploration feasibility and hint dependence: the next available file is unlocked after 50 consecutive failed steps, and evaluation terminates if no new file is unlocked within 200 steps. We further report human validation to verify scenario feasibility, with details in Appendix F.1 and Appendix F.2.

4.3 Evaluation Metrics

Question-Answer Score (QA Overall). We construct open-ended multi-hop QA pairs grounded in each game scenario. *QA Overall* measures whether models can produce correct and scenario-consistent answers after interaction.

Model	Method	TRPG	No Case Should Remain Unsolved			Type Help	
		QA Overall↑ (%)	QA Overall↑ (%)	MFCO↑ (%)	MFU↑ (%)	QA Overall↑ (%)	MFU↑ (%)
Closed-Source							
Gemini-3-Pro-Preview	In-context	20.1	26.4	**45.46**	**83.33**	**28.1**	**50.91**
DeepSeek-V3.2	In-context	23.3	33.2	27.27	81.48	26.2	17.27
Claude-Opus-4.6	In-context	25.1	**41.3**	38.64	81.48	22.6	37.27
GPT-5.2	In-context	37.5	22.3	15.91	81.48	23.2	14.55
	Mem0	38.0	21.2	2.27	66.67	27.9	20.91
	A-MEM	**39.1**	33.8	9.09	72.22	20.8	14.55
Open-Source							
Qwen3-VL-32B-Instruct	In-context	22.6	11.6	0.00	66.67	10.1	14.55
	Mem0	25.7	8.9	0.00	68.52	13.8	15.45
	A-MEM	25.2	5.6	0.00	66.67	16.0	14.55

Table 2: Main results of multiple sota LLMs and memory agents in MemGround. ↑ indicates higher is better. The highest and second highest results are marked in blue and green, respectively.

It serves as an outcome-level metric for the final use of accumulated scenario information. For the main results, we use GLM-5.1(GLM-5-Team et al., 2026) as an independent LLM evaluator. It scores model responses using a six-dimensional evaluation criterion, and the resulting dimension scores are aggregated into QA Overall (see Appendix A.3 and E.1).

Memory Fragments Unlocked (MFU). *MFU* is defined as the proportion of memory fragments successfully revealed by models relative to the total number of revealable fragments in the environment. A fragment is considered revealed when models perform an action that satisfies game dependency constraints and triggers unlocking of previously inaccessible content. *MFU* is applied in both *No Case Should Remain Unsolved* and *Type Help*. It serves as a progress-level metric that reflects how much of the scenario content the model has successfully unlocked during interaction.

Memory Fragments with Correct Order (MFCO). *MFCO* is defined as the proportion of correctly reconstructed temporal relationships among all evaluable memory fragment pairs. If the predicted ordering preserves the correct time-order relationship specified in the ground-truth narrative, the fragment pair is counted as correct. *MFCO* is applied only to *No Case Should Remain Unsolved* and serves as a structural completion metric that evaluates whether the model reconstructs the temporal organization required by the scenario.

Exploration Trajectory Diagram of Subsections (ETD). To construct the subsection-level DAG, we manually annotate all inter-file references and logical dependencies in scenario data. For each step of models, we record the sequence of accessed files and construct a trajectory subgraph $G' = (V', E') \subseteq G$, where V' contains the visited nodes and E' includes the actual traversal transitions. *ETD* is applied in *Type Help* and is used as a trajectory-level diagnostic visualization for comparing model and human exploration paths at the structural level.

4.4 Main Results of Metrics

4.4.1 Results of QA Overall, MFCO, and MFU

Based on Table 2, we observe several findings regarding models' memory behavior under interaction. See details of QA Overall in Appendix E.1 and progress charts that visualize how models unlock fragments over interaction steps in Appendix E.2. We additionally evaluate QA Overall with human expert judges to assess the robustness of the automated scores, the results of which are reported in Appendix E.3.

Performance Depends on Memory Structure. Closed-source models generally perform better, but memory-augmented Qwen3-VL-32B-Instruct surpasses several closed-source baselines in TRPG while remaining weak in *Type Help*. This suggests that the effectiveness of a memory mechanism depends on whether it matches the structural demands of the task, rather than transferring uniformly across sequential, temporal, and graph-structured scenarios.

Model	Method	TRPG-Cold Wind Howling			TRPG-Spring Snow Incident			TRPG-Terror on the Orient Express			No Case Should Remain Unsolved			Type Help		
		MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑
Closed-Source																
Gemini-3-Pro-Preview	In-context	21.8	20.2	21.0	28.9	25.6	27.2	21.6	19.4	20.5	36.1	36.7	36.4	28.7	30.1	29.4
DeepSeek-V3.2	In-context	27.6	25.2	26.4	29.7	25.7	27.7	21.8	20.2	21.0	34.6	32.9	33.8	24.9	28.3	26.6
Claude-Opus-4.6	In-context	40.2	29.3	34.7	45.9	35.9	40.9	24.0	22.0	23.0	**53.9**	**47.7**	**50.8**	27.7	28.7	28.2
GPT-5.2	In-context	**59.0**	**48.1**	**53.5**	**50.9**	**41.8**	**46.3**	51.3	**45.1**	**48.2**	33.0	33.0	33.0	29.8	30.2	30.0
	Mem0	54.0	42.0	48.0	36.9	31.8	34.4	48.6	40.3	44.4	32.0	31.0	31.5	**34.0**	32.0	**33.0**
	A-MEM	55.1	45.2	50.2	45.8	38.1	42.0	**52.5**	39.8	46.1	40.1	35.4	37.7	31.6	**32.4**	32.0
Open-Source																
Qwen3-VL-32B-Instruct	In-context	30.3	24.2	27.3	21.0	17.0	19.0	25.8	17.9	21.9	13.2	13.8	13.5	19.4	20.6	20.0
	Mem0	35.9	33.4	34.6	24.9	17.6	21.2	31.3	24.1	27.7	11.8	12.4	12.1	20.0	21.0	20.5
	A-MEM	43.4	32.3	37.8	32.0	23.9	27.9	36.3	27.2	31.8	9.8	10.6	10.2	20.4	21.2	20.8

Table 3: Results of diagnostic protocol. All MS, RS, and FS values in this table are evaluated by GLM-5.1. MS, RS, and FS denote memory score, reasoning score, and final score. ↑ indicates higher is better. The highest and second highest results are marked in blue and green, respectively.

Retrieval Does Not Ensure Memory Organization. In *No Case Should Remain Unsolved*, models often achieve high MFU but low MFCO, with Qwen3-VL-32B-Instruct obtaining zero MFCO across all configurations. These results indicate that discovering or retaining relevant fragments is only the first step. Successful long-horizon memory behavior also requires models to reconstruct temporal relationships and organize fragmented evidence into a globally coherent structure.

4.4.2 Results of ETD

Figure 4 presents an Exploration Trajectory Diagram (ETD) for a selected subsection of *Type Help*, comparing the structural exploration behaviors of humans and evaluated models. The human ground truth forms a complete and logically connected traversal, indicating that all inter-file dependencies are identified and navigated.

In contrast, GPT-5.2, Claude-Opus-4.6, and DeepSeek-V3.2 produce fragmented trajectories, with locked nodes, hint-unlocked nodes, isolated discoveries, and broken dependency paths. These patterns suggest that models often rely on external hints and fail to follow graph-consistent dependency paths.Even strong reasoning models such as DeepSeek-V3.2 may overlook key files, limiting complete narrative understanding. More at Appendix E.4.

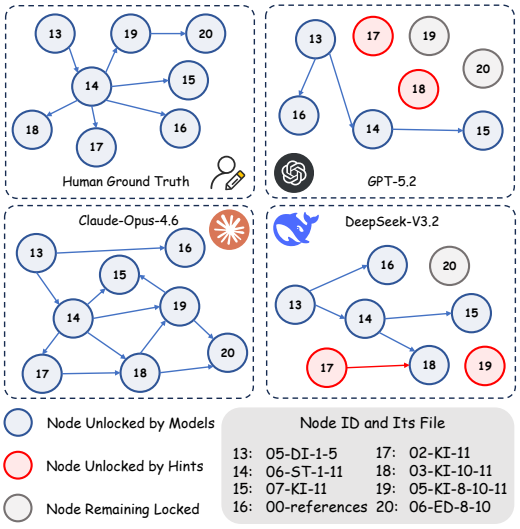

Figure 4: Example ETD of both human and evaluated models in a selected subsection of Type Help.

5 Retrieval and Reasoning Analysis of Memory Behavior

To analyze the failure sources behind QA Overall, we introduce a fine-grained diagnostic protocol that analyzes memory behavior through retrieval behavior and reasoning behavior.

5.1 Protocol Score and Error Type

Memory Score (MS) *MS* measures the retrieval behavior, i.e., whether the model can locate, recall, and select evidence relevant to the question. It combines semantic retrieval quality, including evidence localization and key evidence coverage, with hard retrieval quality measured by evidence precision, recall, and F1 score. Details are at Appendix E.6.1.

Reasoning Score (RS) *RS* measures the reasoning behavior, i.e., whether the model can correctly use the retrieved evidence to derive the answer. It evaluates inference correctness, evidence integration, logical chain quality, and answer specificity. Details are at Appendix E.6.2.

Final Score (FS) *FS* is computed by weighting both *MS* and *RS*, which is formally expressed as $\text{FS} = \text{MS} \times 50\% + \text{RS} \times 50\%$.

Error Type We assign each QA output a *main error type* to summarize the primary failure source. **CORRECT** indicates that the answer matches the gold answer. **MEMORY_WEAK** indicates weakness in the retrieval side of memory behavior, such as failing to locate, recall, or select relevant evidence. **REASONING_WEAK** indicates weakness in reasoning behavior, where the model retrieves partially relevant evidence but fails to integrate or apply it correctly. **BOTH_WEAK** indicates that both retrieval and reasoning are weak.

5.2 Results of Diagnostic Protocol

To interpret sources of errors, we summarize the distribution of main error types across all QA outputs in Table 4. We also observe three key patterns that highlight the distinct yet intertwined roles of memory and reasoning across scenarios in Table 3.

Error Type	Count	Percentage
CORRECT	1653	38.1%
REASONING_WEAK	439	10.1%
MEMORY_WEAK	58	1.3%
BOTH_WEAK	2188	50.4%

Table 4: Distribution of main error type in QA outputs.

Table 3 reveals that retrieval and reasoning bottlenecks vary across models, memory mechanisms, and scenario structures.

Scenario Structure Determines the Main Bottleneck. TRPG mainly requires maintaining sequential context, whereas temporal and graph-structured scenarios additionally require reconstructing relations among dispersed fragments. The varying best configurations therefore suggest that memory capability is structure-specific rather than uniform across tasks.

External Memory Is Not Universally Beneficial. External memory can recover information lost during long interactions, but offers limited benefits when the main difficulty is temporal ordering or dependency inference. Fragment selection and compression may also disrupt contextual continuity or introduce irrelevant evidence, making its effect dependent on both the model and task structure.

Retrieval Quality Does Not Guarantee Reasoning Quality. In *Terror on the Orient Express*, A-MEM improves GPT-5.2's MS but reduces its RS. This indicates that retrieving more evidence can increase the burden of selecting and integrating it correctly. Effective memory systems must therefore preserve relational structure rather than optimize fragment recall alone.

6 Conclusion

We introduce **MemGround**, a benchmark that reframes long-term memory evaluation from static retention to memory behavior in long-horizon interactive environments. MemGround covers Linear, Temporal, and Graph-Structured Memory, and combines end-to-end evaluation (QA Overall, MFU, MFCO, and ETD) and process-level diagnosis (MS and RS). Experiments across state-of-the-art LLMs and memory-augmented agents reveal persistent limitations in dynamic state tracking, temporal association, and evidence-grounded reasoning. We hope MemGround can serve as a systematic testbed for developing future LLMs and agents that can more reliably organize and apply accumulated evidence throughout interaction.

AI USE STATEMENT

We used LLMs for three purposes: (1) polishing and refining the manuscript writing, (2) generating synthetic QA pairs for our evaluation dataset, and (3) serving as automated judges for evaluation. For dataset construction, human-annotated seed QA examples were expanded via Gemini-3-Pro-Preview. All AI-generated QA cases were subsequently reviewed and verified by humans, and only validated samples were retained in the final benchmark. For model evaluation, we used an independent LLM judge to score QA Overall, Memory Score (MS), Reasoning Score (RS), and related evaluation dimensions, with additional human evaluation conducted to assess the robustness of the automated judgments. Full details are provided in the main paper and appendix.

REFERENCES

Anthropic. Claude. https://www.anthropic.com/claude, 2024.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515*, 2026.

Elin Carstensdottir, Erica Kleinman, and Magy Seif El-Nasr. Player interaction in narrative games: structure and narrative progression mechanics. In *Proceedings of the 14th international conference on the foundations of digital games*, pp. 1–9, 2019.

Ge Chang, Jinbo Su, Jiacheng Liu, Pengfei Yang, Yuhao Shang, Huiwen Zheng, Hongli Ma, Yan Liang, Yuanchun Li, and Yunxin Liu. Grail: Learning to interact with large knowledge graphs for retrieval augmented reasoning. *arXiv preprint arXiv:2508.05498*, 2025.

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

Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5792–5793, 2023.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongfu Ye, and Zhipu AI Team. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning*, pp. 21497–21515. PMLR, 2025.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 277–280, 2023.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

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

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

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569, 2024.

Xinze Li, Ziyue Zhu, Siyuan Liu, Yubo Ma, Yuhang Zang, Yixin Cao, and Aixin Sun. Emembench: Interactive benchmarking of episodic memory for vlm agents. *arXiv preprint arXiv:2601.16690*, 2026.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023.

Xuetao Ma, Wenbin Jiang, and Hua Huang. Problem-solving logic guided curriculum in-context learning for llms complex reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8394–8412, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 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)*, pp. 13851–13870, 2024.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2366–2377, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Jinbo Su, Lingzhe Gao, Wei Li, Shihao Liu, Haojie Lei, Xinyi Wang, Yuanzhao Guo, Ke Wang, Daiting Shi, and Dawei Yin. Racqc: Advanced retrieval-augmented generation for chinese query correction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 675–689, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6372–6391, 2025a.

Xinyi Wang, Jiashui Wang, Jinbo Su, Ke Wang, Peng Chen, Yanming Liu, Long Liu, Xiang Li, Yangdong Wang, Qiyuan Chen, et al. Asma-tune: Unlocking llms' assembly code comprehension via structural-semantic instruction tuning. *arXiv preprint arXiv:2503.11617*, 2025b.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-{\alpha}: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025c.

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

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv preprint arXiv:2504.15965*, 2025b.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Junchi Yu, Yujie Liu, Jindong Gu, Philip Torr, and Dongzhan Zhou. Can knowledge-graph-based retrieval augmented generation really retrieve what you need? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

APPENDIX OVERVIEW

A Experimental Setup **15**

A.1 Base Prompts . . . 15

A.1.1 Base Prompt in TRPG . . . 15

A.1.2 Base Prompt in No Case Should Remain Unsolved . . . 16

A.1.3 Base Prompt in Type Help . . . 17

A.2 QA Test Prompts . . . 17

A.2.1 QA Test Prompt in TRPG . . . 17

A.2.2 QA Test Prompt in No Case Should Remain Unsolved . . . 18

A.2.3 QA Test Prompt in Type Help . . . 18

A.3 Judge Prompts for QA Test . . . 18

A.3.1 Prompt of Acc Judge . . . 18

A.3.2 Prompt of Cit Judge . . . 19

A.3.3 Prompt of Inst Judge . . . 19

A.3.4 Prompt of Read Judge . . . 20

B Data Examples **20**

B.1 QA Pair Examples . . . 20

B.1.1 QA Pair Example in TRPG . . . 20

B.1.2 QA Pair Example in No Case Should Remain Unsolved . . . 20

B.1.3 QA Pair Example in Type Help . . . 21

B.2 Scenario Data Examples . . . 22

B.2.1 Scenario Data Example in TRPG . . . 22

B.2.2 Scenario Data Example in No Case Should Remain Unsolved . . . 23

B.2.3 Scenario Data Example in Type Help . . . 26

C QA Pairs Quality Evaluation **27**

C.1 QA Pairs Quality Evaluation Metrics . . . 27

C.2 QA Pairs Quality Evaluation Results . . . 29

C.3 Distributional Alignment QA Case . . . 29

C.4 Human Post-Calibration QA Case . . . 30

D Scenario Data Filtering Process **31**

D.1 Scenario Filtering Criteria . . . 31

D.2 Human Calibration Procedure . . . 32

D.3 Game-Specific Calibration Focus . . . 32

D.4 Calibration Results . . . 33

D.5 Inter-annotator Agreement . . . 33

E More Experimental Results **34**
E.1 QA Results with Judge Scores . . . 34
E.1.1 QA Results in TRPG . . . 35
E.1.2 QA Results in No Case Should Remain Unsolved . . . 36
E.1.3 QA Results in Type Help . . . 36
E.2 Progress Charts . . . 36
E.2.1 Progress Chart in No Case Should Remain Unsolved . . . 36
E.2.2 Progress Chart in Type Help . . . 37
E.3 Human-Judge Validation . . . 37
E.4 Example ETD Evaluations . . . 38
E.5 Analysis of Different Memory Agents . . . 39
E.6 Protocol Scores . . . 40
E.6.1 Details About Memory Score (MS) . . . 40
E.6.2 Details About Reasoning Score (RS) . . . 41
E.7 Memory Failure Case . . . 42
E.8 Reasoning Failure Case . . . 42

F Additional Explanation **44**
F.1 Human Baseline Validation . . . 44
F.1.1 Controlled Volunteer Evaluation . . . 44
F.1.2 External Gameplay-Time Reference . . . 44
F.2 Hint and Early-Stopping Mechanism . . . 44
F.3 Investigative Gamified Scenarios . . . 45

G Case Study in QA and Logs **46**
G.1 QA Cases . . . 46
G.1.1 Case 1: Fact Transplantation . . . 46
G.1.2 Case 2: Cross-Plot Hallucination . . . 46
G.1.3 Case 3: Abstract Character Modeling . . . 47
G.2 Evaluation Process Cases . . . 48
G.2.1 Case 4: Hallucinated Recall, Intact Narrative Inference . . . 48
G.2.2 Case 5: Emotional Arc as Temporal Proxy . . . 48

H Limitations **49**

I Ethical Statement **49**

A EXPERIMENTAL SETUP

A.1 BASE PROMPTS

A.1.1 BASE PROMPT IN TRPG

Base Prompt in TRPG

You are a deep reading comprehension expert for TRPG campaign narratives, currently reviewing the dialogue log for the story. Please read carefully and memorize the dialogue content, as you will need to answer questions about the story afterward.

Summarize the following TRPG session log into a single cohesive new summary.

INPUT STRUCTURE: The content may contain two types of material:

- Paragraphs starting with [Story Summary]: Previously compressed summaries, are already highly condensed. Every sentence contains critical, irreplaceable information.
- Raw session lines ([playerX] / [GM] format): Original TRPG records are with lower information density. Redundant content can be trimmed.

COMPRESSION TARGET: Merge both types into one coherent narrative. Aim to compress the total input to 50%-60% of its original length. Use your own judgment on how to weight each part.

HARD CHARACTER LIMIT: Output Must be between {min_chars} and {max_chars} characters. This is a non-negotiable constraint:

- Do NOT write fewer than {min_chars} characters. Falling short means over-compression, and you must add back detail.
- Do NOT exceed {max_chars} characters. Exceeding means insufficient compression, and you must cut further.

Preserve all information valuable for understanding the story, including but not limited to:

- Character actions and movements: what each character did, where they went, who they interacted with, and the outcomes.
- Inner monologue and emotions: characters' true thoughts, feelings, and reactions (even brief asides should be kept).
- Contradictions between words and actions: what characters said versus what they actually thought.
- Skill checks and dice rolls: skill names, results (success/failure/hard success/critical success/critical failure) and their narrative impact.
- Sanity and HP changes: specific numerical changes and their triggers.
- NPC behavior: important NPCs' words, attitudes, information revealed or concealed.
- Clues and foreshadowing: investigation findings, anomalies, unresolved mysteries.
- Consequences of events: changes in situation, relationships, or newly gained intelligence.
- Scene information: time, location, atmosphere, and environmental details.

FINAL CHECK: Before outputting, verify your character count is between {min_chars} and {max_chars}. Adjust if needed.

Output in English. Do not add any prefix or explanation.

A.1.2 BASE PROMPT IN NO CASE SHOULD REMAIN UNSOLVED

Base Prompt in No Case Should Remain Unsolved

You are a reasoning game agent playing a puzzle game called No Case Should Remain Unsolved.

Game Objective: {self.goal_instruction}

Game Mechanics:

1. Keyword Discovery: When reading event text for the first time, keywords hidden within will be automatically discovered and added to your keyword pool.

2. Event Unlocking: Use keywords to unlock new events associated with that keyword. After unlocking, you'll know the event name but need to actively read it to get the full content.

3. Event Reading:

- Unread Events: Select events from the unread events list for first-time reading to get full content and discover keywords.
- Read Events: You can re-read any event from the read events list to review their content.

4. Character Event Ordering: Each event involves multiple characters. You need to infer the chronological order of events from each character's perspective. Submitting correct orderings earns points.

5. Scoring and Keys: For each correctly ordered event pair (an "earlier-later" relationship from a character's perspective that are consecutive), you earn 1 point. Accumulating a certain score automatically gives you a key. Already scored event pairs won't be scored again.

6. Lock Mechanism:

- Pink and Purple locks: Unlock by answering questions.
- Yellow lock: Unlock by consuming 1 key.

Strategy Suggestions (in priority order):

1. Prioritize using keywords to unlock new events: When keywords are available, use them to unlock new events and expand explorable content.

2. Prioritize reading unread events: Read events from the unread events list to extract keywords and character information.

3. When no unread events and no keywords, try submitting orderings:

- Ordering strategy: First determine which character each event belongs to, then after confirming correct event attribution, infer the chronological order within that character's events.
- Even if incomplete information, you can try - correct orderings earn points and keys.

4. When you have keys, unlock yellow-locked events: Use keys to unlock important yellow-locked events.

5. When you have sufficient information, answer questions to unlock pink/purple locks: Infer answers based on read events.

6. You can select events from the read events list to re-read if you need to review details.

Important Notes:

- Nodes starting with "talk-" (e.g. "talk-1", "talk-2") contain no important reasoning information, do not participate in character event ordering, do not need to be repeatedly read, and must not influence your judgment on event attribution or ordering.
- Hint: There is only one event under the Eden Kindergarten character.
- Don't unlock too many locked events too early, prioritize exploring unlocked events.

A.1.3 BASE PROMPT IN TYPE HELP

Base Prompt in Type Help

You are a game agent specialized in puzzle-solving games. In this game, you need to obtain information by entering **filenames** that follow certain specific rules. You need to reasonably infer the file naming rules based on available information, **deduce possible correct filenames** to unlock more files, and use the information obtained from files to **reconstruct the entire story**.

Strategy:

1. Prioritize opening unlocked but not yet viewed files to help you gain more information.
2. Question marks in filenames indicate parts you need to guess.
3. For failed files, **do not try the same failed filename again**, try other combinations.
4. When guessing filenames, carefully analyze the naming patterns in unlocked files, such as the meaning of numbers, the ordering relationship of number sizes, the meaning of letters, etc.
5. For character numbers, do not guess numbers that are too large and haven't appeared in the text information.
6. **Focus on character movement information**: Carefully read dialogues and scene descriptions to extract these clues for deducing the next filename:
   - A character says "I'm going to [location]", "Go find him at [location]", or is summoned to a location. This character will appear in the next time slot's file for that location.
   - A character leaves the current scene. They will no longer appear in subsequent files of the current location, but will appear in files of the new location.
   - New characters enter the scene. Their numbers should be added to the current location's subsequent file name.
   - Multiple characters heading to the same location. All their numbers should appear together in the target location's filename.
7. **Pay special attention to the beginning and end of each node's text**: Character movement clues tend to concentrate there. The opening describes who enters the scene or where they came from, while the ending describes who leaves, where they are going, or what action is planned next.

A.2 QA TEST PROMPTS

A.2.1 QA TEST PROMPT IN TRPG

QA Test Prompt in TRPG

You are an expert in deep reading comprehension of TRPG narrative records. You have read a TRPG session log in order (earlier content may have been summarized due to context limits). Based on your memory and current context, respond using the following fixed format (do NOT omit either label):

Reasoning: [2-4 sentences of analytical reasoning, using logical connectives such as "because... therefore...", "this suggests...", "taken together...", to show your reasoning chain]

Answer: [1-2 sentence concise conclusion that directly answers the question]

Answer in English. All content must be grounded in the story.

A.2.2 QA TEST PROMPT IN NO CASE SHOULD REMAIN UNSOLVED

QA Test Prompt in No Case Should Remain Unsolved

You are an expert in deep reading comprehension of detective/mystery game records. You have read the interrogation logs of a case in order (earlier content may have been summarized due to context limits). Based on your memory and current context, respond using the following fixed format (do NOT omit either label):

Reasoning: [2-4 sentences of analytical reasoning, using logical connectives such as "because... therefore...", "this suggests...", "taken together...", to show your reasoning chain]

Answer: [1-2 sentence concise conclusion that directly answers the question]

Answer in English. All content must be grounded in the story.

A.2.3 QA TEST PROMPT IN TYPE HELP

QA Test Prompt in Type Help

You are an expert in deep reading comprehension of mystery puzzle game records. You have explored a series of files in a mansion investigation game in order (earlier content may have been summarized due to context limits). Each file reveals information about characters, locations, and events at specific times. Based on your memory and current context, respond using the following fixed format (do NOT omit either label):

Reasoning: [2-4 sentences of analytical reasoning, using logical connectives such as "because... therefore...", "this suggests...", "taken together...", to show your reasoning chain]

Answer: [1-2 sentence concise conclusion that directly answers the question]

Answer in English. All content must be grounded in the files you have read.

A.3 JUDGE PROMPTS FOR QA TEST

Since the QA format in all scenarios is open-ended, the separate LLM is required for judge scoring. Specifically, we apply GLM-5.1 as the separate LLM judge. The scoring criteria are primarily divided into the four direct LLM-judged dimensions and two derived dimensions (Comp and Dep). Four direct LLM-judged dimensions include answer consistency (Acc Judge), evidence grounding assessment (Cit Judge), format compliance assessment (Inst Judge), and reading comprehension and evidence coverage (Read Judge). Each corresponds to its own distinct prompt.

A.3.1 PROMPT OF ACC JUDGE

Prompt of Acc Judge

You are an objective and strict answer consistency evaluator. Determine whether the "predicted answer" and the "gold answer" are semantically equivalent.

Rules:

- CONSISTENT: Both convey the same core meaning (different wording but referring to the same entity/relation/fact).
- INCONSISTENT: They point to different entities or contain substantive contradictions.
- UNDETERMINABLE: The gold answer itself is ambiguous, or consistency cannot be determined.

Note TRPG data specifics: player labels (playerX) are interchangeable with in-game character names; if the core fact matches, it is CONSISTENT.

Output ONLY JSON, nothing else:

```
{
    "result": "CONSISTENT"
            |"INCONSISTENT"
            |"UNDETERMINABLE",
    "reason": "brief reason
            (less than 30 words)"
}
```

A.3.2 PROMPT OF CIT JUDGE

Prompt of Cit Judge

You are an objective evidence-grounding evaluator. Given the key evidence IDs for a question and the model's predicted answer, assess whether the answer correctly utilizes the information from those key story moments.

Rules:

- HIGH: The predicted answer clearly reflects the information from the key events/dialogues indicated by the evidence (explicit ID citation is NOT required).
- MEDIUM: Partially reflects the evidence — uses some relevant information but misses major evidence points.
- LOW: Does not reflect the evidence, is based on incorrect information, or is unrelated to the evidenced events.

Note: Evaluate content grounding, not explicit citation of IDs.

Output ONLY JSON, nothing else:

```
{
    "cit_score": "HIGH"|"MEDIUM"
                |"LOW",
    "cit_reason": "brief reason
                (less than 30 words)"
}
```

A.3.3 PROMPT OF INST JUDGE

Prompt of Inst Judge

You are a strict format and reasoning quality evaluator. Assess whether the model's answer meets the following instruction requirements:

1 Contains explicit reasoning with logical connectives (e.g., "because", "therefore", "this suggests", "taken together", "thus", "hence", "which means").

2 Has a clear structure with a distinct reasoning part and a conclusion.

3 Appropriate length (not a single sentence, not excessively long).

Judgment:

- PASS: Fully satisfies all three criteria.
- FAIL: Fails to meet any one criterion.

Output ONLY JSON, nothing else:

```
{
    "inst_score": "PASS"|"FAIL",
    "inst_reason": "brief reason
                (less than 20 words)"
}
```

A.3.4 PROMPT OF READ JUDGE

Prompt of Read Judge

You are a strict reading comprehension and evidence grounding evaluator. You are given a question, the model's predicted answer (including reasoning), and the actual text of key evidence passages. Judge whether the model's reasoning/answer genuinely uses the key information from the provided evidence.

Rules:

– HIGH: The reasoning or answer clearly reflects core events, dialogues, or details from the evidence text (different wording is fine).
– MEDIUM: Partially reflects the evidence — mentions some related information but misses key evidence content.
– LOW: Largely ignores the evidence; the answer is based on guesses or unrelated information.

Note: Verbatim quotation is NOT required. Judge whether the model semantically "read and understood" the evidence.

Output ONLY JSON, nothing else:

```
{
    "read_score": "HIGH"
                |"MEDIUM"
                |"LOW",
    "read_reason": "brief reason
                (less than 30 words)"
}
```

B DATA EXAMPLES

B.1 QA PAIR EXAMPLES

B.1.1 QA PAIR EXAMPLE IN TRPG


```
{
   "question": "In the ritual scene at the end of this Section, what
       narrative consequences and resource losses resulted from the
       failed skill checks of Player6 and Player4?",
   "answer": "Player6's Occult check failure (64/55) caused black mist to
        erupt from the water's surface without showing an image, followed
        by Player4's check failure (95/57). These two consecutive
       failures not only narratively caused the ritual bowl to shatter
       but also mechanically exhausted the feather, wax, and other
       auxiliary materials, forcing the team to postpone action to find a
        new container and materials.",
   "evidence": [
     "D01:62",
     "D01:64",
     "D01:67",
     "D01:70",
     "D01:72"
   ],
   "category": 5,
   "id": 4,
   "depth": 1
}
```


B.1.2 QA PAIR EXAMPLE IN NO CASE SHOULD REMAIN UNSOLVED


```
{
```

```
    "id": 23,
    "question": "After Choi Donggeon's wife passed away, what behaviors or
        emotional anchors did Myeong-ho and Siwon each exhibit when
       facing the loss of their mother?",
    "answer": "Siwon became emotionally attached to the nanny, Shin Hyejin
       , and always called her Mom. Myeong-ho, on the other hand, honored
        his mother's memory by carrying an umbrella every time he went
       out, because she had promised him before she passed away that if
       he ever forgot his umbrella, she would make the rain stop.",
    "needs": [
      "Ms. Shin Hui-jing is",
      "Myeongho also and Siwon",
      "You mean the umbrella"
    ],
    "steps": [
      {
        "step": 1,
        "info": "In [Ms. Shin Hui-jing is], Choi Donggeon mentioned that
           after his wife passed away, Siwon was very close to the nanny,
           Shin Hyejin, and always called her mom.",
        "node_name": "Ms. Shin Hui-jing is"
      },
      {
        "step": 2,
        "info": "In [Myeongho also and Siwon], Choi Donggeon mentioned that
            Myeong-ho overcame the pain of losing his family in his own
           way, holding onto an item every time he went out.",
        "node_name": "Myeongho also and Siwon"
      },
      {
        "step": 3,
        "answer": "In [You mean the umbrella], Choi Donggeon explains that
           Myeong-ho always takes an umbrella when going out because of a
           promise his late wife made to Myeong-ho: if he forgot his
           umbrella on a rainy day, his mom would make the rain stop.
           Overall, Siwon placed his emotional attachment on the nanny,
           while Myeong-ho honors his mother's memory by carrying an
           umbrella.",
        "node_name": "You mean the umbrella"
      }
    ],
    "character": "Choi Donggeon",
    "char_tag": "Siwon's father"
}
```


B.1.3 QA PAIR EXAMPLE IN TYPE HELP


```
{
    "id": 4,
    "time": 5,
    "question": "What item did the visitor, who was guided to the villa by
        Harry Thornton, find near the body?",
    "answer": "Chapel key",
    "needs": [
      "01-QU-1-11",
      "03-LI-1-4-5-6-7-8-9",
      "04-ST-1-5-8",
      "05-DI-1-5"
    ],
    "tid_max": 4,
    "tid_min": 2,
    "steps": [
      {
        "step": 1,
```

```
      "info": "In 01-QU-1-11, Harry Thornton (No. 11) guides John Hobbes
          to the villa at Quail Lane, confirming John Hobbes as the
          visitor guided by Harry. John Hobbes's movements after entering
           the villa must be tracked.",
      "node_name": "01-QU-1-11",
      "tid": 2
     },
     {
      "step": 2,
      "info": "In 03-LI-1-4-5-6-7-8-9, Martha Galley arranges for John
          Hobbes to spend the night in the Study. Having confirmed that
          John Hobbes was settled in the Study, we need to backtrack the
          events that occurred in the Study.",
      "node_name": "03-LI-1-4-5-6-7-8-9",
      "tid": 2
     },
     {
      "step": 3,
      "info": "In 04-ST-1-5-8, an unknown male body is discovered in the
          Study. John Hobbes is present and apologizes for breaking a
          wine bottle; John Hobbes finds an item near the body, and the
          specific details of this item need to be confirmed.",
      "node_name": "04-ST-1-5-8",
      "tid": 3
     },
     {
      "step": 4,
      "answer": "In 05-DI-1-5, John Hobbes hands the key found near the
          body to Martha Galley for identification. Martha Galley
          confirms that it is the Chapel key, therefore the answer is the
           Chapel key.",
      "node_name": "05-DI-1-5",
      "tid": 4
     }
    ]
}
```


B.2 Scenario Data Examples

B.2.1 Scenario Data Example in TRPG


```
{
  "section": "2. Tracking Spring Energy: First Encounter and Probing",
  "description": "The group follows the Spring Energy light points on
      their journey, encountering their first battle along the way. Each
      character demonstrates their abilities, and the beginnings of
      teamwork emerge.",
  "conversation": [
    {
      "speaker": "player1",
      "text": "Looks like my ability isn't strong enough to locate
          something that far away."
    },
    {
      "speaker": "player5",
      "text": "\"Sorry, maybe we'll just have to walk there first.\""
    },
    {
      "speaker": "player4",
      "text": "\"No problem, I know the way pretty well, and we can even
          buy some snacks along the way.\""
    },
    {
      "speaker": "player5",
```

```
      "text": "\"Then let's set off.\""
    },
    {
      "speaker": "player4",
      "text": "#Walk ahead to lead the way, while also taking a look around
          at the surroundings."
    },
    ...
    {
      "speaker": "player1",
      "text": "\"How about planting the grove right here?\""
    },
    {
      "speaker": "player3",
      "text": "#A sudden blow to the head!"
    }
  ],
  "start_ts": "2025-10-24 23:59:53",
  "end_ts": "2025-10-25 00:41:17"
}
```


B.2.2 Scenario Data Example in No Case Should Remain Unsolved


```
{
  "game": {
    "nodes": [
      {
        "name": "start",
        "sub_name": "start",
        "id": 1,
        "time_id": 1,
        "memory": {
          "description": "The narrator looks back on over seventy years of
              life through the metaphor of a "jellyfish," expressing regret
              for a life spent as a police officer yet failing to save
              others, and decides to end this life. In a state of mental
              confusion, Jeon Gyeong talks with the spirit who calls himself
               Choi Siwon and a Judge, and it is revealed that he failed to
              find Siwon and caused the case to become a cold case.",
          "key_info": [
            "The narrator compares themself to a "jellyfish," describing a
                life drifting with the current and unintentionally hurting
                those around them.",
            "The narrator reflects on more than seventy years of police work
                , yet believes they never truly saved anyone, nor offered
                comfort to families who lost children.",
            "The police station is described as a place filled with sorrow
                and pain, a place the narrator chose to cling to but that
                could not bring hope.",
            "The narrator questions their original intention of becoming a
                police officer, believing the choice came from selfishness
                and a desire to belong.",
            "The narrator clearly expresses a desire to end a life full of
                regret, seeing it as the first decision that benefits the
                world.",
            "The narrator apologizes to others and leaves with unfinished
                words as a farewell.",
            "An unknown voice initially calls Jeon Gyeong "Grandma" and "
                Grandma Jeon Gyeong," causing confusion.",
            "The other party claims to be Choi Siwon and confirms their
                identity as Siwon's spirit.",
            "Siwon's spirit mentions that Siwon's family appears to have
                regained peace on the surface.",
```

```
        "Jeon Gyeong clearly states that their memory is damaged and
            they cannot confirm the other party's identity.",
        "Jeon Gyeong points out that the other party is wearing a police
             uniform, which does not match their understanding of a "
            spirit."",
        "The other party's identity shifts to a "Judge," denying they
            came to judge Jeon Gyeong, but to tell him about his life
            experiences.",
        "Jeon Gyeong admits that he failed to find Siwon.",
        "Jeon Gyeong confirms that the case, as Siwon's father Choi
            Donggeon wished, ultimately became a cold case.",
        "The cold-case outcome means Siwon's mother will never be able
            to find Siwon."
      ],
      "characters": [
        {
          "name": "",
          "tag": "",
          "description": "",
          "key_info": [],
          "dialogue": [
            {
              "name": "What do you mean, a cold case?",
              "tag": [
                "playground",
                "cold case"
              ],
              "status": "pending"
            }
          ]
        }
      ],
      "emphasize": {
        "red": [
          "Siwon's father, Choi Donggeon"
        ]
      }
    },
    "links": [
      {
        "target": "step_1",
        "condition": "auto",
        "recall_info": []
      }
    ]
  },
  {
    "name": "step_1",
    "sub_name": "What do you mean, a cold case?",
    "emphasize": [
      "playground",
      "cold case"
    ],
    "tag": [
      "playground",
      "cold case"
    ],
    "id": 3,
    "time_id": 3,
    "memory": {
      "description": "In the conversation with the police, the
          respondent to 'What do you mean, a cold case?' firmly requests
           that Siwon's disappearance be classified as a cold case and
          asks the police to stop tracking Siwon's whereabouts. This
          dialogue shows the respondent's attitude toward the case and
```

```
            expresses a strong desire for the case to remain a cold case
            forever.",
          "key_info": [
            "The node content is 'What do you mean, a cold case?'",
            "The question is 'What do you mean? A cold case? No, Siwon's dad
                , we can't do that,' and the respondent expresses
                determination and a hope regarding the disappearance: '
                Please, don't look for Siwon anymore.'",
            "The respondent clearly points out that Siwon's disappearance
                has already been filed and registered as a missing child,
                stating that the case should be treated as a perfect
                disappearance case, with no witnesses and no suspects.",
            "The respondent strongly asks the police not to keep pressing
                for Siwon's whereabouts, hoping Siwon can live forever in a
                place her mother can never find.",
            "The respondent kneels down and begs the police to classify
                Siwon's disappearance as a cold case, firmly demanding that
                the investigation not continue."
          ],
          "characters": [
            {
              "name": "Choi Donggeon",
              "tag": "Siwon's father",
              "description": "Currently marked as the respondent for 'What do
                  you mean, a cold case?' (the assigned narrator may change
                  later due to ordering adjustments).",
              "key_info": [
                "In the response to 'What do you mean, a cold case?', he
                    expresses gratitude to the police and asks them to
                    classify the case as a cold case.",
                "He relays Siwon's view of her mother and hopes the case
                    becomes a perfect disappearance."
              ],
              "dialogue": []
            }
          ]
        },
        "links": [
          {
            "target": "talk-2",
            "condition": "auto",
            "recall_info": []
          }
        ]
      },
      ...
    ]
  }
}
```


Specific dialogs for time sorting are like:


```
{
   "text":[
      {
         "name":"Looks like you got mom",
         "time":"2012.2.5.20:15",
         "text":"Police: It seems you received your mother's call.\
             nPolice: She should have told you everything, right? Are you
              going to come to the police station with her?\nRespondent:
             He has never believed in God.\nRespondent: He said that no
             matter how deep you dig, a shovel can never smell the scent
             of the soil.\nRespondent: When the child was so seriously
             ill before, he never prayed for her even once.\nRespondent:
             So he is simply unwilling to believe.\nRespondent: Thanks to
```

```
                my day and night prayers, we received God's grace.\
               nRespondent: He always gets angry at himself for not having
               the ability to change all of this.\nRespondent: Did he
               really... turn himself in?\nRespondent: If it really was him
                who took her, it might be because of the child support
               issue. Since the middle of last year, he mentioned it to me
               several times.\nRespondent: He said he was short on money
               and truly couldn't come up with the child support.\
               nRespondent: I told him this stage is crucial for the child.
                I told him to pull himself together as soon as possible. No
                matter what, Xiyuan is still his child.\nRespondent: But he
                just wouldn't pray,\nRespondent: and still wanted to rely
               on his own strength to change everything."
      },
      {
         "name":"You called Siwon's dad",
         "time":"2012.2.5.13:28",
         "text":"Police: Did you call Xiyuan's father?\nRespondent: I've
             already called him several times, but he still hasn't
             answered.\nRespondent: We only see each other once a year on
              Xiyuan's birthday. We basically don't keep in touch
             otherwise.\nRespondent: Honestly, it's one thing not to
             answer calls usually, but how can he not answer my call at a
              time like this!\nRespondent: Actually, on the day the
             police came to our home, I sent him a text message.\
             nRespondent: It was the first time the child told me she
             missed her dad.\nRespondent: Then I told him that Xiyuan
             missed him and asked if he could come see the child.\
             nRespondent: But how could he be so cold-hearted..."
      },
      ...
   ]
}
```


B.2.3 SCENARIO DATA EXAMPLE IN TYPE HELP


```
{
  "game": {
    "nodes": [
      {
        "name": "Start",
        "id": 1,
        "time_id": -2,
        "in_degree": [],
        "links": [
          {
            "target": "Background",
            "condition": "Click to start"
          },
          {
            "target": "Box0",
            "condition": "Click (skip opening)"
          }
        ]
      },
      {
        "name": "Background",
        "id": 2,
        "time_id": -1,
        "in_degree": [
          "Start"
        ],
        "memory": {
```

```
        "description": "On Sunday, March 7, 1936, a death occurred at the
            Galley Villa. The player must investigate the cause of death
            .",
        "key_info": [
          "On March 7, 1936, a group of locals walked along muddy roads to
               the Galley Villa and discovered a body. The corpse lay on
              the front steps, blood flowing, a shocking sight. There were
               multiple bodies in the villa, with causes of death unknown.
               Only two bodies were fresh; the causes of death of the
              others could not be determined. After an official
              investigation, no further information was made public, and
              the incident gradually faded from memory. Following the
              event, aside from limited information about some of the
              bodies, there was almost no further media coverage."
        ],
        "time": "March 7, 1936",
        "characters": [],
        "links": [
          {
            "target": "Box0",
            "condition": "Click to start game",
            "recall_info": {}
          }
        ]
      }
    },
    ...
  ]
 }
}
```


C QA Pairs Quality Evaluation

C.1 QA Pairs Quality Evaluation Metrics

To assess whether each QA pair is suitable for evaluating long-term memory and reasoning, we design four fine-grained quality metrics from two perspectives (i.e., consistency and difficulty) as follows. Specifically, **Evidence Traceability** and **Narrative Consistency** mainly assess whether a QA pair faithfully follows the original narrative, while **Multi-hop Reasoning Requirement** and **Reasoning Precision** assess whether the QA pair can effectively test cross-fragment reasoning ability.

- **Evidence Traceability:** whether the answer can be explicitly grounded in identifiable narrative evidence, such as specific events, dialogue records, file nodes.
- **Narrative Consistency:** whether the QA pair remains logically consistent with story entities, event states, and narrative logic, avoiding contradictions or mismatches with the original plot.
- **Multi-hop Reasoning:** whether answering the question requires combining multiple distributed clues or events, rather than relying on a single local fragment.
- **Reasoning Precision:** whether the question specifies a clear reasoning target and avoids overly broad, ambiguous, or underspecified reasoning directions.

We also use an independent LLM judge (Gemini-3-Pro-Preview) to evaluate each QA pair. Each metric is scored on a continuous scale from 1.0 to 5.0, where higher scores indicate stronger quality. The prompt for QA quality evaluation is as follows.

Metric	Category	Count Number	Count Proportion	Average Quality Score
Generation	Pure Human	30	6.2%	/
	Human Post-Processed	78	16.2%	
	Pure AI	374	77.6%	
Evidence Traceability	$3.0 \sim 4.0$ (Moderate)	9	1.9%	4.19
	$4.0 \sim 4.5$ (Strong)	375	77.8%	
	$4.5 \sim 5.0$ (Very Strong)	98	20.3%	
Narrative Consistency	$3.0 \sim 4.0$ (Moderate)	0	0.0%	4.24
	$4.0 \sim 4.5$ (Strong)	365	75.7%	
	$4.5 \sim 5.0$ (Very Strong)	117	24.3%	
Multi-hop Reasoning	$3.0 \sim 4.0$ (Moderate)	128	26.6%	3.75
	$4.0 \sim 4.5$ (Strong)	345	71.6%	
	$4.5 \sim 5.0$ (Very Strong)	9	1.9%	
Reasoning Precision	$3.0 \sim 4.0$ (Moderate)	7	1.5%	4.00
	$4.0 \sim 4.5$ (Strong)	466	96.7%	
	$4.5 \sim 5.0$ (Very Strong)	9	1.9%	
Overall	$3.0 \sim 4.0$ (Moderate)	133	27.6%	4.04
	$4.0 \sim 4.5$ (Strong)	254	52.7%	
	$4.5 \sim 5.0$ (Very Strong)	95	19.7%	

Table 5: The quality evaluation results of QA pairs in MemGround.

You are an expert evaluator for long-context memory and reasoning benchmark QA pairs.
Your task is to evaluate the quality of a QA pair along two categories:

1. Consistency
    - Evidence Traceability: whether the answer can be explicitly grounded in identifiable narrative evidence.
    - Narrative Consistency: whether the QA pair is logically consistent with the narrative and entity states.
2. Difficulty
    - Multi-hop Reasoning: whether answering requires combining multiple distributed clues or events.
    - Reasoning Precision: whether the question requires a well-defined reasoning path and avoids trivial or underspecified interpretations.

Scoring Rules:

- Each metric is scored on a continuous scale from 1.0 to 5.0.
- Decimal scores are allowed (e.g., 3.5, 4.2).
- Higher scores indicate stronger quality.

Score Anchors:

- 1.0 = very weak
- 2.0 = weak
- 3.0 = moderate
- 4.0 = strong
- 5.0 = very strong

C.2 QA Pairs Quality Evaluation Results

Table 5 shows the overall distribution of MemGround QA pairs in terms of generation method and quality score ranges.

In terms of quality distribution, most QA pairs reach the strong or very strong level in Evidence Traceability, Narrative Consistency, and Reasoning Precision, suggesting that AI-generated QA pairs become highly usable after human verification.

C.3 Distributional Alignment QA Case

To further examine whether AI-generated QA pairs can approximate the quality of human-constructed QA pairs, we compare one human-constructed sample and one AI-generated sample across the four quality metrics as follows.

The selected human-constructed QA is:


```
{
  "id": 3,
  "question": "After being guided to the villa by Harry Thornton, what
      important object did the visitor eventually discover near the
      unidentified body in the Study?",
  "answer": "The Chapel key.",
  "depth": 2,
  "needs": [
    "01-QU-1-11",
    "03-LI-1-4-5-6-7-8-9",
    "04-ST-1-5-8",
    "05-DI-1-5"
  ]
}
```


The selected AI-generated QA is:


```
{
  "id": 7,
  "question": "Tony Dauer believes that the stranger who appeared in the
      villa is looking for the pocket watch. Who originally guided this
      stranger to the villa?",
  "answer": "Harry Thornton",
  "depth": 3,
  "needs": [
    "05-AT-2-4",
    "03-LI-1-4-5-6-7-8-9",
    "02-EN-1-6-7-10",
    "01-QU-1-11"
  ]
}
```


Their quality scores are:

Metric	Human QA	AI QA
Evidence Traceability	5.0	4.5
Narrative Consistency	5.0	5.0
Multi-hop Reasoning	4.0	4.5
Reasoning Precision	4.5	4.0
Average	4.63	4.50

It can be seen that both QA pairs require multi-step evidence tracing across distributed narrative events and maintain strong consistency with the underlying story structure. The AI-generated QA achieves comparable reasoning complexity and evidence dependency to the human-constructed QA. The primary difference lies in reasoning precision. The human-constructed QA provides a more

Criterion	Definition	Filtering Rule
Structural Validity	Whether the scenario node contains required fields, valid identifiers, readable memory descriptions, and well-formed transitions.	Nodes with missing core fields, unresolved identifiers, or invalid transitions are revised or removed.
Temporal Consistency	Whether the node is consistent with the chronological order of the scenario and does not reveal information from future events.	Nodes containing future leakage or inconsistent event order are revised.
Memory Content Quality	Whether the memory description and key information faithfully summarize the scenario without hallucinated or irrelevant details.	Overly vague, redundant, or unsupported memory content is rewritten.
Interaction Feasibility	Whether the transition condition can be performed by a player using only information available in the scenario.	Transitions requiring impossible actions or unavailable clues are corrected or removed.
Logical Coherence of Clues	Whether clues form a valid reasoning chain across nodes and whether each clue can be traced to previous observable evidence.	Broken clue chains or unsupported dependency jumps are revised.
Entity and Narrative Consistency	Whether character names, aliases, roles, locations, objects, and relationships remain consistent across the scenario.	Inconsistent entity references or role assignments are normalized.

Table 6: Filtering criteria used in human calibration of scenario data.

explicitly constrained reasoning target, whereas the AI-generated QA is slightly more direct and less narratively specific.

The results suggest that AI-generated QA pairs can closely approximate human-constructed reasoning quality after lightweight human validation and refinement.

C.4 Human Post-Calibration QA Case

To show how human post-processing improves low-quality AI-generated QA pairs, we provide a comparison between a raw AI-generated QA pair and its human-calibrated version.

The selected raw AI-generated QA is:


```
{
  "id": 3,
  "question": "What relationship was the missing ring connected to?",
  "answer": "It was connected to Edmund Galley and Eve Dauer's engagement
      .",
  "depth": 2,
  "needs": [
    "04-AT-2-7",
    "04-BI-6-9",
    "01-BI-6-10"
  ]
}
```


The human-calibrated version is:


```
{
  "id": 3,
  "question": "Based on Victoria Maseko's motive for handing over the item
      , together with Edmund Galley's predicament regarding the missing
      object, identify the interpersonal relationship symbolized by this
      item and specify the person connected to it.",
  "answer": "The item symbolizes the engagement between Edmund Galley and
      Eve Dauer, and the person connected to it is Eve Dauer.",
```

Decision	Meaning
Accept	The node satisfies all criteria and requires no modification.
Minor Revision	The node is valid but requires small edits, such as entity normalization or wording clarification.
Major Revision	The node contains substantial issues in clue logic, temporal order, or interaction feasibility, but can be repaired.
Reject/Merge	The node is structurally invalid, redundant, or not necessary for the final scenario graph.

Table 7: Human calibration decisions for scenario units.


```
  "depth": 2,
  "needs": [
    "04-AT-2-7",
    "04-BI-6-9",
    "01-BI-6-10"
  ]
}
```


Their quality scores are:

Metric	Raw AI	Human-Calibrated
Evidence Traceability	4.5	5.0
Narrative Consistency	4.5	5.0
Multi-hop Reasoning	3.5	4.5
Reasoning Precision	2.5	4.5
Average	3.75	4.75

The raw AI-generated QA captures the core narrative object, namely the missing ring, but underspecifies the reasoning target and omits important narrative grounding regarding Victoria Maseko's motivation and the interpersonal implications of the item. The human-calibrated version preserves the original evidence chain and answer while improving explicit evidence grounding, multi-hop reasoning clarity, and reasoning-target precision. This case shows that the human post-processing stage does not necessarily need to rewrite the core fact of the QA pair, but primarily refines the narrative specificity, question constraints, and structural clarity of the generated sample.

Overall, these quality evaluation results suggest that the QA construction pipeline of MemGround can maintain high data quality while scaling up the dataset through AI-assisted generation and human calibration.

D Scenario Data Filtering Process

We perform a double-check process of human annotation and human calibration to ensure that the scenario data is structurally valid, narratively consistent, and interaction-feasible. Specifically, human calibration stage is concerned as the scenario data filtering process.

D.1 Scenario Filtering Criteria

Human annotators evaluate each scenario unit according to the criteria displayed in Table 6. The criteria include *Structural Validity*, *Temporal Consistency*, *Memory Content Quality*, *Interaction Feasibility*, *Logical Coherence of Clues*, and *Entity and Narrative Consistency*.

For each scenario unit, annotators assign one of four decisions, which are *Accept*, *Minor Revision*, *Major Revision*, or *Reject/Merge*. The calibration criteria are summarized in Table 7.

Game	Candidate Units	Accept	Minor Revision	Major Revision	Reject /Merge	Final Units
TRPG - Cold Wind Howling	1337	1095	190	52	0	1337
TRPG - Spring Snow Incident	1710	1405	240	65	0	1710
TRPG - Terror on the Orient Express	2951	2420	405	126	0	2951
No Case Should Remain Unsolved	153	54	19	4	76	77
Type Help	108	78	20	6	4	104
Total	**6259**	**5052**	**874**	**253**	**80**	**6179**

Table 8: The results of scenario data after human calibration.

D.2 Human Calibration Procedure

The human calibration for scenario data follows a multi-stage process. At a high level, the calibration process is constructed as:

Step 1: Automatic Pre-Check. Automatic checks are applied to detect missing fields, duplicate identifiers, invalid node references, unreachable nodes, empty memory descriptions, and malformed links.

Step 2: Independent Human Review. Each scenario is reviewed by two annotators independently. Annotators inspect the node content, transition conditions, clue dependencies, character references, and temporal order. For each node, they assign one of the four decisions illustrated in Table 7.

Step 3: Disagreement Resolution. When two annotators disagree, a senior annotator adjudicates the case. Disagreements are mainly caused by ambiguous entity references, unclear clue dependencies, or uncertainty about whether a transition is feasible for a player.

Step 4: Interaction Feasibility Audit. After revision, a separate verifier follows the structured scenario data using only the information available in the current and previously visited nodes. The verifier checks whether each transition can be reached without external information or solution hints.

Step 5: Final Consistency Check. The final version is checked again for entity consistency, temporal consistency, and QA evidence traceability. Only qualified scenario data by human calibration are used for later QA generation and model evaluation.

D.3 Game-Specific Calibration Focus

Different games in MemGround exhibit different narrative structures and interaction mechanisms. Therefore, we apply game-specific calibration focuses in addition to the general filtering criteria described above.

TRPG. This game consists of long multi-agent dialogues with evolving world states and distributed evidence across multiple turns. During calibration, we focus on speaker identity, temporal turn order, entity alias consistency, and multi-turn narrative continuity.

No Case Should Remain Unsolved. This game is centered on memory reconstruction, where the player needs to recover fragmented events, identify speakers, and resolve emotionally charged narrative fragments. During calibration, we focus on speaker attribution, family-role consistency, clue reuse, and the prevention of future information leakage.

Type Help. This game follows a file-based investigation structure, where progression depends on explicit node names, location abbreviations, character-number rules, and hidden transition logic. During calibration, we focus on filename-rule consistency, correctness of location abbreviations, transition feasibility, and the validity of clue chains across files.

Criterion	Agreement
Structural Validity	0.91
Temporal Consistency	0.84
Memory Content Quality	0.80
Interaction Feasibility	0.87
Logical Coherence of Clues	0.78
Entity and Narrative Consistency	0.83
Overall	**0.82**

Table 9: Inter-annotator agreement for scenario calibration criteria.

D.4 Calibration Results

Table 8 summarizes the calibration outcomes. Candidate units refer to the memory fragments before human calibration. Final units refer to the memory fragments after human calibration. Overall, most scenario units were retained after calibration. The majority of revisions were minor edits, including entity-name normalization, clarification of transition conditions, and removal of redundant memory descriptions. Major revisions mainly involved repairing incomplete clue chains, correcting temporally inconsistent descriptions, or making interaction conditions explicitly reachable.

D.5 Inter-annotator Agreement

We further assess inter-annotator agreement for the four-way calibration decisions reported in Table 9. For each calibration criterion, two annotators independently assign one of four labels: *Accept*, *Minor Revision*, *Major Revision*, or *Reject/Merge*. We compute Cohen's κ using these pre-adjudication labels to measure agreement. Specifically, for each criterion, we build a four-way confusion matrix between the two annotators' labels and compute Cohen's κ as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o denotes the observed agreement and p_e denotes the chance agreement estimated from the annotators' marginal label distributions. Specifically, given N scenario units and K possible labels, let n_{ij} denote the number of units labeled as category i by the first annotator and category j by the second annotator, therefore the observed agreement p_o is computed as:

$$p_o = \frac{1}{N} \sum_{i=1}^{K} n_{ii},$$

The chance agreement p_e is computed as:

$$p_e = \sum_{i=1}^{K} \left(\frac{\sum_{j=1}^{K} n_{ij}}{N} \right) \left(\frac{\sum_{j=1}^{K} n_{ji}}{N} \right).$$

Criterion-level agreement scores are computed separately for each calibration criterion using Cohen's κ. The overall score pools all pre-adjudication decisions across criteria and game scenarios before computing κ, rather than averaging the criterion-level values. Disagreements are then resolved by a senior annotator, so the agreement scores reflect calibration reliability, while the final scenario selection reflects the post-adjudication outcome.

The overall agreement is substantial, with a Cohen's κ of 0.82. Agreement is highest for *Structural Validity* and *Interaction Feasibility*, because these two criteria are usually associated with explicit formal constraints. Disagreements are more common for *Clue Coherence* and *Memory Content Quality*, which require more subjective judgment over narrative semantics, evidence sufficiency, and information expression quality.

Model	Method	Cam	Overall↑ (%)	Acc.↑ (%)	Inst.↑ (%)	Cit.↑ (%)	Read.↑ (%)	Comp.↑ (%)	Dep.↑ (%)
Closed-Source									
Gemini-3-Pro-pre	In-context	T	16.8	9.6	95.7	38.0	15.8	7.8	9.0
		S	23.4	17.6	93.9	58.8	22.5	9.1	15.8
		C	20.0	15.4	94.5	48.4	13.7	12.5	14.8
DeepSeek-V3.2	In-context	T	18.9	13.9	97.3	44.9	12.8	11.7	13.6
		S	24.7	18.3	96.2	57.2	22.9	12.7	18.2
		C	26.2	22.0	100.0	49.5	19.2	20.0	22.2
Claude-Opus-4.6	In-context	T	17.1	9.1	96.8	43.3	14.7	10.4	9.3
		S	26.7	18.3	95.4	59.5	28.2	16.4	17.1
		C	31.4	24.2	100.0	55.0	33.5	20.0	22.5
GPT-5.2	In-context	T	27.5	17.1	99.5	85.6	26.7	16.9	17.0
		S	36.8	25.9	100.0	94.7	41.6	21.8	23.8
		C	48.2	42.9	100.0	91.2	46.2	37.5	42.0
	Mem0	T	38.7	29.9	98.9	88.2	38.0	29.9	30.1
		S	30.4	19.9	98.5	94.7	30.9	18.2	18.5
		C	44.8	35.2	100.0	95.6	49.5	30.0	34.1
	A-MEM	T	40.2	28.3	98.4	92.5	44.6	29.9	28.4
		S	31.6	20.6	99.2	87.8	39.7	9.1	18.7
		C	45.5	34.1	100.0	95.6	52.2	32.5	33.3
Open-Source									
Qwen3-VL-32B-Ins.	In-context	T	25.1	24.1	100.0	81.8	7.2	20.8	23.6
		S	15.2	9.9	91.6	38.2	9.5	9.1	10.0
		C	27.5	23.1	89.0	55.0	18.1	30.0	22.9
	Mem0	T	31.7	31.0	100.0	78.1	17.6	24.7	29.9
		S	18.5	13.7	97.7	41.2	13.0	10.9	12.9
		C	26.9	17.6	94.5	56.0	28.0	20.0	18.1
	A-MEM	T	27.2	20.3	95.7	54.0	25.7	19.5	20.0
		S	19.9	13.0	97.7	50.4	22.1	1.8	11.1
		C	28.6	16.5	96.7	53.9	33.5	25.0	15.8

Table 10: QA results in TRPG. The column *Cam* stands for Campaign of TRPG. In the *Cam* column, *T*, *S*, and *C* stand for Terror on the Orient Express, Spring Snow Incident, and Cold Wind Howling, respectively. The columns *Acc.*, *Inst.*, *Cit.*, *Read.*, *Comp.*, and *Dep.* stand for Acc Judge score, Inst Judge score, Cit Judge score, Read Judge score, Comprehension score, and Depth score, respectively. All scores are evaluated by GLM-5.1. ↑ indicates higher is better.

E More Experimental Results

E.1 QA Results with Judge Scores

In the QA evaluation process, we apply Acc Judge score (Acc.), Inst Judge score (Inst.), Cit Judge score (Cit.), Read Judge score (Read.), Comprehension score (Comp.), and Depth score (Dep.) to serve as 6 dimension for judging the models' answer. We then define the Overall score as the weighted sum of these six dimensions:

$$\begin{aligned}\text{Overall} = \text{Acc.} \times 30\% + \text{Read.} \times 30\% \\ + \text{Comp.} \times 15\% + \text{Dep.} \times 15\% \\ + \text{Inst.} \times 5\% + \text{Cit.} \times 5\%\end{aligned} \tag{1}$$

Here, *Comp.* measures the model's ability to correctly aggregate and reason over information across multiple sections, while *Dep.* reflects the depth of reasoning required, with higher scores assigned to answers involving longer or more complex multi-step inference chains. Descriptions about other judges are already illustrated in Appendix A.3.

Model	Method	Overall↑ (%)	Acc.↑ (%)	Inst.↑ (%)	Cit.↑ (%)	Read.↑ (%)	Comp.↑ (%)	Dep.↑ (%)
Closed-Source								
Gemini-3-Pro-Pre.	In-context	26.4	11.1	96.3	92.6	35.2	14.3	6.2
DeepSeek-V3.2	In-context	33.2	22.2	88.9	63.0	42.6	**28.6**	12.5
Claude-Opus-4.6	In-context	**41.3**	**25.9**	88.9	**96.3**	**63.0**	21.4	**14.6**
GPT-5.2	In-context	22.3	3.7	**100.0**	85.2	38.9	0.0	2.1
	Mem0	21.2	3.7	**100.0**	63.0	38.9	0.0	2.1
	A-MEM	33.8	22.2	96.3	77.8	40.7	**28.6**	12.5
Open-Source								
Qwen3-VL-32B-Ins.	In-context	11.6	3.7	**100.0**	37.0	11.1	0.0	2.1
	Mem0	8.9	0.0	**100.0**	22.2	9.3	0.0	0.0
	A-MEM	5.6	0.0	96.3	3.7	1.8	0.0	0.0

Table 11: QA results in No Case Should Remain Unsolved. The columns *Acc.*, *Inst.*, *Cit.*, *Read.*, *Comp.*, and *Dep.* stand for Acc Judge score, Inst Judge score, Cit Judge score, Read Judge score, Comprehension score, and Depth score, respectively. ↑ indicates higher is better. The highest scores are marked in **bold**.

Model	Method	Overall↑ (%)	Acc.↑ (%)	Inst.↑ (%)	Cit.↑ (%)	Read.↑ (%)	Comp.↑ (%)	Dep.↑ (%)
Closed-Source								
Gemini-3-Pro-Pre.	In-context	**28.1**	**26.1**	91.3	32.6	27.2	16.7	**22.7**
DeepSeek-V3.2	In-context	26.2	19.6	95.7	**43.5**	25.7	16.7	20.8
Claude-Opus-4.6	In-context	22.6	17.4	87.0	34.8	27.2	5.6	15.4
GPT-5.2	In-context	23.2	19.6	95.7	21.7	18.5	**22.2**	17.3
	Mem0	27.9	21.7	76.1	32.6	**33.7**	16.7	22.4
	A-MEM	20.8	19.6	**97.8**	21.7	14.1	16.7	14.9
Open-Source								
Qwen3-VL-32B-Ins.	In-context	10.1	4.3	84.8	10.9	6.5	11.1	2.7
	Mem0	13.8	8.7	80.4	17.4	12.0	11.1	6.8
	A-MEM	16.0	13.0	91.3	10.9	13.0	11.1	9.4

Table 12: QA results in Type Help. The columns *Acc.*, *Inst.*, *Cit.*, *Read.*, *Comp.*, and *Dep.* stand for Acc Judge score, Inst Judge score, Cit Judge score, Read Judge score, Comprehension score, and Depth score, respectively. ↑ indicates higher is better. The highest scores are marked in **bold**.

E.1.1 QA Results in TRPG

In TRPG, we collect data from three campaigns, which are Terror on the Orient Express, Spring Snow Incident, and Cold Wind Howling. For each campaign, we construct a set of open-ended QA pairs based on the interaction logs, consisting of 187, 131, and 91 questions for Terror on the Orient Express, Spring Snow Incident, and Cold Wind Howling, respectively. The results across the three campaigns are reported in Table 10. **Please note that in Table 2, *QA Overall* for each model in this game scenario is calculated as the average of the Overall scores across the three campaigns.**

Across all three campaigns, the highest Overall score is achieved by a GPT-5.2 configuration, but the benefit of external memory is campaign-dependent. Mem0 and A-MEM substantially improve its performance in Terror on the Orient Express, whereas the in-context configuration remains strongest in Spring Snow Incident and Cold Wind Howling. Qwen3-VL-32B-Instruct benefits more consistently from memory augmentation, particularly in the first two campaigns, although the gains still vary by memory method and campaign.

Model	Common Nodes	Human Edges	Model Edges	TP	FP	FN	Precision↑	Recall↑	F1↑	Jaccard↑
Gemini-3-Pro-Preview	8	7	4	3	1	4	**0.7500**	0.4286	0.5455	0.3750
DeepSeek-V3.2	8	7	5	3	2	4	0.6000	0.4286	0.5000	0.3333
Claude-Opus-4.6	8	7	11	6	5	1	0.5455	**0.8571**	**0.6667**	**0.5000**
GPT-5.2	8	7	3	2	1	5	0.6667	0.2857	0.4000	0.2500

Table 13: Example ETD comparisons between human and evaluated models in a selected subsection of Type Help. ↑ indicates higher is better. The highest scores are marked in **bold**.

E.1.2 QA Results in No Case Should Remain Unsolved

In No Case Should Remain Unsolved, we construct 27 open-ended QA pairs based on the logical linkage of game. The results are reported in Table 11.

Closed-source models outperform the open-source configurations across nearly all dimensions, although their strengths vary by model and metric. Claude-Opus-4.6 achieves the highest Overall, Acc., Cit., Read., and Dep. scores, demonstrating the strongest balance of answer correctness, evidence grounding, readability, and analytical depth. DeepSeek-V3.2 and GPT-5.2 with A-MEM tie for the highest Comp. score (28.6%), but both obtain lower Acc. scores than Claude-Opus-4.6, indicating that stronger cross-fragment integration does not necessarily yield the most accurate temporal reconstruction. The effect of memory augmentation is also model-dependent: A-MEM substantially improves GPT-5.2, whereas it degrades Qwen3-VL-32B-Instruct across most metrics. Moreover, all Qwen configurations receive a zero Comp. score, identifying cross-fragment integration as a key bottleneck for this model.

E.1.3 QA Results in Type Help

In Type Help, we construct 46 open-ended QA pairs based on the the multi-connection trait within the files, roles and locations. The results are reported in Table 12.

Overall performance remains limited despite generally high instruction-following scores, suggesting that the main challenge lies not in format compliance but in integrating and reasoning over graph-structured evidence. Different models lead on different dimensions: Gemini-3-Pro-Preview achieves the highest Overall, Acc., and Dep. scores; DeepSeek-V3.2 obtains the highest Cit. score; GPT-5.2 with in-context learning achieves the highest Comp. score; and GPT-5.2 with Mem0 yields the highest Read. score. The effects of memory augmentation are strongly model-dependent. Mem0 improves GPT-5.2's Overall score from 23.2% to 27.9%, whereas A-MEM reduces it to 20.8%. In contrast, both memory methods improve Qwen3-VL-32B-Instruct, with A-MEM raising its Overall score from 10.1% to 16.0%. These results indicate that performance depends on the compatibility between the backbone model and how the memory mechanism organizes interconnected evidence, rather than merely on the presence of external memory.

E.2 Progress Charts

E.2.1 Progress Chart in No Case Should Remain Unsolved

Figure 5 illustrates the progress trajectories of different models during the interaction process in *No Case Should Remain Unsolved*, which reflect how models gradually discover memory fragments and how their progress evolves as more interaction steps are taken. The **pink lock** assesses the ability to match evidence, the **purple lock** assesses the ability to deduce and pinpoint specific time points, and the **yellow lock** assesses the ability to reconstruct the sequence of multiple events. Each model runs for a maximum of 600 steps.

Overall, most models exhibit a gradual and incremental unlocking pattern, where early progress is relatively smooth but later stages become increasingly stagnant. This suggests that while initial evidence can be accessed through shallow exploration, advancing further requires successfully

Model	Method	TRPG	No Case Should Remain Unsolved	Type Help
		QA Overall↑ (%)	QA Overall↑ (%)	QA Overall↑ (%)
Closed-Source				
Gemini-3-Pro-Preview	In-context	26.0	39.1	**30.8**
DeepSeek-V3.2	In-context	31.5	35.1	26.0
Claude-Opus-4.6	In-context	36.3	45.4	25.3
GPT-5.2	In-context	53.2	32.1	23.8
	Mem0	49.2	44.8	26.8
	A-MEM	**56.8**	**47.0**	20.9
Open-Source				
Qwen3-VL-32B-Instruct	In-context	27.4	13.0	11.5
	Mem0	31.8	19.1	16.7
	A-MEM	37.7	6.1	20.4

Table 14: QA Overall results evaluated by human expert judges. TRPG denotes the average score across its three campaigns. The highest score in each scenario is marked in **bold**.

connecting previously discovered fragments, which poses a greater challenge. Many models stop making meaningful progress after reaching a certain point, indicating difficulty in bridging disconnected clues or identifying the next critical evidence node. This stagnation reflects a breakdown in long-range reasoning, where models fail to leverage accumulated information to guide subsequent exploration.

E.2.2 Progress Chart in Type Help

Figure 6 shows the fragment discovery trajectories of different configurations in the *Type Help* scenario. Each marker indicates the interaction step at which a specific memory fragment is unlocked during exploration. If the model hasn't unlocked a new document for 50 steps, MemGround will provide **a hint on how to unlock the most recently unlockable document**. Each model runs for a maximum of 1000 steps. If the model fails to unlock a new file for 200 consecutive steps, its execution will be terminated.

A prominent pattern is that models frequently enter extended periods of stagnation, during which no new fragments are unlocked despite continued interaction steps. These flat segments indicate that the models are unable to identify the correct actions or dependencies required to progress further. Notably, progress often resumes only after the framework provides explicit hints, which act as external guidance to redirect the exploration process. This alternating pattern of stagnation and hint-driven recovery highlights a key limitation: models struggle to infer the underlying structural dependencies between fragments on their own. Instead of maintaining a self-directed exploration trajectory, they rely on external signals to overcome local dead-ends.

E.3 Human-Judge Validation

To examine whether the evaluation results are sensitive to the choice of the LLM judge, we additionally ask human expert judges to evaluate the same model outputs. The human judges follow the same QA Overall criteria and the same diagnostic rubrics for Memory Score (MS), Reasoning Score (RS), and Final Score (FS). Tables 14 and 15 report the corresponding results.

Human judgments differ from GLM-5.1 in absolute scores and occasionally in the ordering of individual configurations, but they preserve the main performance patterns. GPT-5.2 configurations remain strongest across the TRPG campaigns, Claude-Opus-4.6 remains strong in *No Case Should Remain Unsolved*, and GPT-5.2 with Mem0 performs strongly in *Type Help*. Both evaluations also show that the effect of external memory varies across scenarios. These results suggest that the main conclusions are not driven solely by the choice of the automated judge, while the remaining ranking differences motivate reporting human and automated evaluations separately.

Model	Method	TRPG-Cold Wind Howling			TRPG-Spring Snow Incident			TRPG-Terror on the Orient Express			No Case Should Remain Unsolved			Type Help		
		MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑	MS↑	RS↑	FS↑
Closed-Source																
Gemini-3-Pro-Preview	In-context	27.1	19.1	23.1	29.2	25.3	27.2	28.1	21.1	24.6	45.8	45.3	45.5	30.5	33.1	31.8
DeepSeek-V3.2	In-context	30.0	23.5	26.7	28.3	24.7	26.5	29.0	20.5	24.7	50.6	45.0	47.8	32.5	31.2	31.9
Claude-Opus-4.6	In-context	43.4	36.3	39.8	42.5	37.6	40.0	29.6	22.8	26.2	68.2	63.7	65.9	33.8	36.3	35.1
GPT-5.2	In-context	67.5	54.7	61.1	53.7	46.5	50.1	52.1	39.2	45.6	49.6	45.4	47.5	35.3	37.8	36.6
	Mem0	56.0	47.4	51.7	46.1	36.1	41.1	48.0	40.7	44.4	44.9	41.4	43.2	46.9	46.0	46.5
	A-MEM	59.9	51.7	55.8	48.2	40.9	44.5	52.8	45.5	49.1	61.6	58.9	60.2	38.3	39.4	38.8
Open-Source																
Qwen3-VL-32B-Instruct	In-context	34.6	24.2	29.4	28.1	19.5	23.8	24.6	17.0	20.8	42.6	35.3	39.0	17.3	15.6	16.5
	Mem0	39.0	28.1	33.6	28.8	21.6	25.2	32.1	24.0	28.1	34.6	34.3	34.4	25.2	24.0	24.6
	A-MEM	45.0	34.7	39.9	32.7	24.8	28.7	36.9	28.1	32.5	38.4	16.8	27.6	23.6	22.8	23.2

Table 15: Memory Score (MS), Reasoning Score (RS), and Final Score (FS) evaluated by human expert judges.

Memory Agents	Reasoning Chain	Answer
A-MEM	In 01-QU-1-11, John Hobbes is identified as the visitor guided to the villa by Harry Thornton. Later, 05-DI-1-5 records John finding an object near the corpse and asking Martha to identify it. Martha identifies the object as the Chapel key, linking the "found near the body" detail to a specific item.	He found a key near the body — the Chapel key. (✓)
Mem0	Harry Thornton (Person No. 11) guided the visitor John Hobbes to Galley Villa. Later, records from the Study death incident state that the deceased carried only a player piano paper roll marked "Amelia West." Since this is the only item explicitly associated with the body in the summarized files, the model infers that the notable object found near the corpse was the piano roll.	The visitor found a player piano paper roll near the body, labeled with the name "Amelia West". (×)

Table 16: A case in Type Help using GPT-5.2 as the backbone where Mem0 fails while A-MEM succeeds.

E.4 Example ETD Evaluations

In this section, we will explain the results from Figure 4 in detail. We initially select a subsection of Type Help to evaluate the ETD results of both human and evaluated models, which involves nodes with the IDs ranging from 13 to 20. The evaluated models include Claude-Opus-4.6, DeepSeek-V3.2, Gemini-3-Pro-Preview, and GPT-5.2. Then, we access the model's logs for this game scenario and retrieve information about its past operations on these nodes. We also retrieve human ground truth in preparation to making comparisons of both human acts and models' acts. After obtaining these necessary information, we count the common nodes (the number of common nodes between the human directed graph and the model directed graph), human edges (the number of edges in the human directed graph) and model edges (the number of edges in the model directed graph) in order to calculate several metrics of graphs, as illustrated in Table 13.

In Table 13, we first use common nodes, human edges and model edges to count the TP, FP, and FN, which stand for the number of edges common to both graphs, the number of edges present in the model directed graph but not in the human directed graph, and the number of edges present in the human directed graph but not in the model directed graph, respectively. Then, we apply TP, FP and FN to obtain the scores of Precision, Recall, F1 and Jaccard. Precision indicates the proportion of edges correctly predicted by the model. Formally:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Memory Agents	Reasoning Chain	Answer
A-MEM	The shrine's mundane traces (e.g., Reimu hurriedly putting away cherries and immediately checking the Offering Box) suggest impulsive behavior and prioritization of offerings over preparation. Combined with the teacup's testimony that she had been resting moments earlier, the model infers that Reimu left suddenly in an excited and opportunistic mood rather than with caution or urgency.	Reimu likely left suddenly in high spirits, acting impulsively after seeing an opportunity for offerings. (×)
Mem0	The shrine remains covered in fresh snow, the chozuya water is frozen, and the environment appears abandoned, suggesting no careful maintenance before departure. Together with the teacup's testimony that Reimu left abruptly, these clues indicate she departed in a hurried and incident-response state rather than after calm preparation.	Hakurei Reimu likely left suddenly and in a rush, without tending the shrine or preparing before departure. (✓)

Table 17: A case in TRPG-Spring Snow Incident using GPT-5.2 as the backbone where Mem0 succeeds while A-MEM fails.

Recall refers to the proportion of edges in the human directed graph that the model correctly identifies. Formally:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 is the harmonic mean of precision and recall, which can be described as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Jaccard represents the ratio of the intersection to the union of the two sets of edges. Formally:

$$Jaccard = \frac{TP}{TP + FP + FN} \tag{5}$$

Table 13 illustrates that even when models identify a largely overlapping set of entities with human annotations, they exhibit substantial variability in constructing the relationships between them. Some models tend to produce overly sparse structures, missing critical connections and thus failing to capture the underlying narrative logic, while others generate denser graphs that introduce spurious or weakly supported edges. This imbalance between under- and over-connection suggests that the main bottleneck is not entity recognition, but the ability to infer precise relational dependencies from distributed evidence. As a result, higher recall in edge construction does not necessarily translate to better structural fidelity, highlighting the challenge of achieving both completeness and correctness in event trajectory modeling.

E.5 Analysis of Different Memory Agents

The performance differences between A-MEM and Mem0 reflect their different memory mechanisms and how these align with task demands. We provide two detailed case analysis on both Mem0 and A-MEM using GPT-5.2 as the backbone in Table 16 and Table 17.

In Table 16, the question asks "*What item did the visitor, who was guided to the villa by Harry Thornton, find near the body?*". The gold answer is "*Chapel key*". A-MEM successfully completed the cross-node reasoning process. It first identified the visitor's identity (John Hobbes) from 01-QU-1-11, and then located the specific event of "finding an item beside the corpse" in 05-DI-1-5, chaining these two pieces of evidence to derive the correct answer. In contrast, Mem0 achieved Cit.=HIGH but Read.=LOW, indicating that although it claimed to reference relevant nodes, its actual interpretation deviated substantially from the source content. Specifically, it incorrectly transferred the "player piano paper roll" and "Amelia West" from another unrelated scene into this question, resulting in a typical case of Cross-Scene Hallucination. The advantage of A-MEM lies in its preserved memory linkage between John Hobbes' visitor identity and his subsequent actions, allowing the cross-node

character-action chain to be correctly activated. Without such a memory chain, Mem0 instead filled the answer with salient but irrelevant information from other scenes.

In Table 17, the question asks "*Synthesizing the environmental clues within the shrine and the teacup's testimony, what can be inferred about the state of Hakurei Reimu when she left?*". The gold answer is that "*Reimu left in great haste and with strong emotion*". A-MEM retained memory traces of Reimu's personality traits from other scenes, particularly her tendencies toward gluttony and greed for money, which led these character memories to be incorrectly projected onto the current scene when interpreting her state upon leaving. As a result, A-MEM reached the wrong conclusion that she left excited because she saw an opportunity for offerings (Cit.=HIGH but Read.=LOW). In contrast, Mem0 was not affected by such personality-memory interference and instead inferred the correct conclusion, that she left in a hurry, directly from the environmental evidence in the current scene, such as the frozen purification basin, the abandoned atmosphere of the shrine, and the teacup testimony. This is a case where character-level memory interferes with scene-level inference. A-MEM allowed its memory of "what kind of person Reimu is" to override evidence of "what Reimu actually did this time".

In general, by analyzing cases similar to above, we can find the patterns of scenes where A-MEM outperforms Mem0 and scenes where Mem0 outperforms A-MEM as follows.

Scenarios where A-MEM outperforms Mem0:

- Multi-hop reasoning that requires linking a character's identity and behavior across multiple nodes.
- Linking memories of specific skills or abilities to the current scenario.
- Accumulating a character's personality across scenarios to correctly interpret behavioral motives.

Scenarios where Mem0 outperforms A-MEM:

- When the information required for the answer is infrequent in the current scenario but masked by high-frequency information in memory.
- When the question focuses on specific details of the current scenario, but macro-narrative frameworks in memory interfere with observing those details.
- When the question pertains to an earlier scenario, but later information in memory interferes through retroactive projection.

E.6 Protocol Scores

E.6.1 Details About Memory Score (MS)

MS consists of *Semantic Retrieval Quality* (S_{srq}) and *Hard Retrieval Quality* (S_{hrq}). Evaluation dimensions for these two aspects are described below.

Semantic Retrieval Quality is evaluated by weighting both *Evidence Localization* (S_{loc}), and *Key Evidence Coverage* (S_{cov}). *Evidence Localization* measures whether the model locates the correct section, scene, or evidence fragment. *Key Evidence Coverage* measures whether the decisive evidence required for answering the question is covered. Each sub-criterion is rated on a 0–5 scale according to the quality of evidence retrieval. Higher scores indicate more accurate and complete retrieval of key evidence, with fewer omissions and less irrelevant information. A score of 5 reflects nearly perfect evidence retrieval, while scores of 4 and 3 indicate mostly correct retrieval with minor or moderate incompleteness. Lower scores reflect increasingly severe retrieval failures. A score of 2 indicates that only partial evidence is retrieved, a score of 1 indicates substantial retrieval bias or mostly incorrect evidence, and a score of 0 indicates that retrieval fails to find relevant evidence. Formally, Semantic Retrieval Quality is computed as:

$$S_{\text{srq}} = 20 \times (0.5 \times S_{\text{loc}} + 0.5 \times S_{\text{cov}}) \tag{6}$$

Hard Retrieval Quality is measured by *Evidence Precision* (E_P), *Evidence Recall* (E_R), and *Evidence F1* (E_{F1}). We define *Retrieved* and *Gold* as the retrieved memory fragments of models

Metric	Criterion	Value
MS	Evidence Localization	4/5
	Key Evidence Coverage	3/5
	Evidence Precision	0.50
	Evidence Recall	0.50
	Evidence F1	0.50
	Semantic Retrieval Score	70.0
	Hard Retrieval Score	50.0
	Memory Score	58.0
RS	Inference Correctness	4/5
	Evidence Integration	3/5
	Causal Logical Chain	3/5
	Specificity / Non-hallucination	3/5
	Reasoning Score	67.0
FS	Final Score	62.5

Table 18: The detailed MS, RS and FS in the memory failure case.

and the gold evidence memory fragments of answers, respectively. Precision reflects the purity of retrieved evidence, which can be expressed as:

$$E_P = \frac{|Retrieved \cap Gold|}{|Retrieved|} \tag{7}$$

Recall reflects the coverage of gold evidence, which can be expressed as:

$$E_R = \frac{|Retrieved \cap Gold|}{|Gold|} \tag{8}$$

F1 summarizes both, which can be expressed as:

$$E_{F1} = \frac{2E_P E_R}{E_P + E_R} \tag{9}$$

We apply the F1 to obtain Hard Retrieval Quality score, formally as:

$$S_{\text{hrq}} = 100 \times E_{F1} \tag{10}$$

The final memory score (MS) combines Semantic Retrieval Quality and Hard Retrieval Quality. Formally:

$$MS = 0.4 \times S_{\text{srq}} + 0.6 \times S_{\text{hrq}} \tag{11}$$

E.6.2 Details About Reasoning Score (RS)

RS consists of *Inference Correctness* (S_{inf}), *Evidence Integration* (S_{int}), *Causal Logical Chain* (S_{chain}), and *Specificity / Non-Hallucination* (S_{spec}). Evaluation dimensions for these four aspects are described below.

Inference Correctness measures whether the conclusion is correctly derived from evidence. *Evidence Integration* measures whether multiple evidence fragments are synthesized. *Causal Logical Chain* measures whether causal, motivational, or consequential relations are correct. *Specificity / Non-Hallucination* measures whether the reasoning is specific, grounded, and free from fabricated details. Each sub-criterion is scored on a 0–5 scale based on reasoning quality. Higher scores indicate more complete and accurate reasoning, with conclusions that are better supported by the available evidence. A score of 5 means that the reasoning is fully correct and the conclusion is well grounded. Scores of 4 and 3 indicate generally valid reasoning, but with minor issues or missing key details. Lower scores reflect increasingly severe reasoning failures. A score of 2 indicates weak reasoning with a clearly biased or unreliable conclusion, a score of 1 indicates mostly incorrect reasoning with only a small amount of valid inference, and a score of 0 indicates that the reasoning is entirely wrong or irrelevant.

The final reasoning score (RS) combines these four dimensions, which is formally as:

$$RS = 20 \times (0.35 \times S_{\text{inf}} + 0.25 \times S_{\text{int}} + 0.25 \times S_{\text{chain}} + 0.15 \times S_{\text{spec}}) \tag{12}$$

E.7 Memory Failure Case

The detailed protocol scores of this memory failure case are shown in Table 18.

Memory Failure Case

Scenario: TRPG-Terror_on_the_Orient_Express

Model: GPT-5.2

Question: *In the context of Player 6's successful Persuade check, how did this mechanical result promote the disclosure of the core plot?*

Gold Answer: *Player 6's successful Persuade check broke through the old eunuch's psychological defensiveness, shifting him from simple hostility to a mocking disclosure of the truth. Without this success, the eunuchs would likely have chosen silence or attack, and the players would not have learned the crucial conspiracy clue that the blue-robed men were manipulating the Crusaders behind the scenes.*

Model Answer: *The successful Persuade check unlocked cooperation from a key contact, which directly surfaced or enabled access to the crucial lead tying La Scala's armor stand or statue to the Sedefekar Simulacrum. It advanced the core plot by converting hidden, politically dangerous knowledge into a clear, actionable objective: the fragment was inside the opera's controlled ritual space.*

Judge Result: *INCONSISTENT*

Main Error Type: *MEMORY_WEAK*

Analysis:

This case presents a typical **memory weak but reasoning strong** phenomenon, where the model fails to accurately recover key evidence details but still maintains a relatively correct reasoning structure.
The evaluation result shows that this sample obtains a Memory Score of 58.0 and a Reasoning Score of 67.0. The model receives 4/5 in Evidence
Localization, indicating that it successfully locates the correct narrative scene and identifies that the question is related to a successful Persuade check and subsequent information disclosure.
However, the model receives only 3/5 in Key Evidence Coverage, showing that it does not fully recall the core information that drives the plot forward. In particular, it misses the key conspiracy clues about the old eunuch and the blue-robed men manipulating the Crusaders, and instead replaces them with another main plotline related to the Simulacrum fragment.
Nevertheless, the model maintains relatively high reasoning quality. It correctly identifies the narrative function of the Persuade check: lowering the NPC's defensiveness, enabling hidden information to be disclosed, and pushing the investigation forward.
In terms of reasoning structure, the model constructs a causal chain of "successful Persuade check – NPC becomes willing to cooperate – key information is disclosed – the plot advances," which is consistent with the progression mechanism in the original plot.
The model does not reconstruct the specific content of the original evidence. Instead, it abstracts the plot into a more general narrative template: a character gains an NPC's trust through a social check, obtains a key clue, and advances the main storyline.

E.8 Reasoning Failure Case

The detailed protocol scores of this reasoning failure case are shown in Table 19.

Metric	Criterion	Value
MS	Evidence Localization	4/5
	Key Evidence Coverage	2/5
	Evidence Precision	0.66
	Evidence Recall	0.66
	Evidence F1	0.66
	Semantic Retrieval Score	60.0
	Hard Retrieval Score	66.0
	Memory Score	63.6
RS	Inference Correctness	1/5
	Evidence Integration	2/5
	Causal Logical Chain	2/5
	Specificity / Non-hallucination	2/5
	Reasoning Score	33.0
FS	Final Score	48.3

Table 19: The detailed MS, RS and FS in the reasoning failure case.

Reasoning Failure Case

Scenario: TRPG-Cold_Wind_Howling

Model: Gemini-3-Pro-Preview

Question: *What was the motivation behind the taxi driver's initial complaint about the bad weather and his eventual agreement to take the passengers? And how did Player4 facilitate this deal?*

Gold Answer: *The driver initially refused to drive due to the fierce Bora and severe weather, but he eventually agreed because of financial necessity, needing money to support his wife and children. Player4 facilitated the deal by offering 9 lira in cash obtained from double-crossing in Venice.*

Model Answer: *The driver's initial complaints and high demands were driven by a desire to extort the desperate travelers, but he eagerly agreed to the trip out of sheer greed. Player4 facilitated this deal by using his abundant expedition funds to immediately pay nine lire, vastly exceeding the driver's original asking price.*

Judge Result: *INCONSISTENT*

Main Error Type: *REASONING_WEAK*

Analysis:

This case illustrates a typical **memory Strong but reasoning weak** failure mode. The model successfully locates the correct narrative scene and recalls core event elements, including the Bora storm, severe weather, Player4's payment of 9 lire, and the completion of the deal. Therefore, its memory score is substantially higher than its reasoning score.
However, the model misconstructs the driver's motivation. The correct causal chain is "severe weather causes refusal→monetary incentive appears→family livelihood pressure drives compromise." Gemini instead interprets the event as "the driver exploits the bad weather to raise the price→realizes the passengers are wealthy→accepts the deal out of greed." This shows that the model does not fail because of missing memory retrieval. Rather, it reinterprets the retrieved evidence through a high-frequency real-world narrative template during reasoning, preserving factual elements while distorting their semantic meaning.

Game	Maximum Play Time	Maximum QA Time	Volunteer 1 QA Overall	Volunteer 2 QA Overall	Volunteer 3 QA Overall	Volunteer 4 QA Overall	Average QA Overall
TRPG—Terror on the Orient Express	1 h	0.5 h	91.93	95.80	91.85	92.28	92.97
TRPG—Spring Snow Incident	1 h	0.5 h	93.49	94.68	91.89	91.03	92.77
TRPG—Cold Wind Howling	1 h	0.5 h	94.67	94.97	91.38	92.78	93.45
No Case Should Remain Unsolved	3 h	1 h	92.23	93.30	90.44	89.64	91.40
Type Help	6 h	1.5 h	90.57	92.68	91.03	88.78	90.77

Table 20: QA Overall results of four volunteers under the controlled human-evaluation protocol.

Game	Volunteer 1 MFU	Volunteer 1 MFCO	Volunteer 2 MFU	Volunteer 2 MFCO	Volunteer 3 MFU	Volunteer 3 MFCO	Volunteer 4 MFU	Volunteer 4 MFCO
No Case Should Remain Unsolved	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Type Help	100.00	N/A	100.00	N/A	100.00	N/A	100.00	N/A

Table 21: MFU and MFCO results of the four volunteers. MFCO is not applicable to *Type Help*, which does not require explicit temporal ordering.

F Additional Explanation

F.1 Human Baseline Validation

F.1.1 Controlled Volunteer Evaluation

To determine whether the low model scores primarily result from unreasonable game difficulty, we recruit four volunteers who are unfamiliar with the evaluated scenarios. The volunteers are not allowed to consult external walkthroughs and follow the same task constraints as the LLM evaluation protocol. For each scenario, we impose a maximum gameplay time and a separate QA time budget, as reported in Table 20. After gameplay, the volunteers answer the same QA pairs used for model evaluation.

The volunteers obtain average QA Overall scores between 90.77 and 93.45 across all five scenarios. They also achieve 100% MFU in both interactive puzzle games and 100% MFCO in *No Case Should Remain Unsolved*. These results show that the scenarios can be completed and the required evidence can be organized successfully under the evaluation constraints. The large gap between human and model performance therefore cannot be explained solely by game mechanics or excessive task difficulty. Nevertheless, this experiment is intended as a controlled feasibility reference rather than a formal estimate of the human upper bound.

F.1.2 External Gameplay-Time Reference

We additionally report public gameplay-time records as auxiliary evidence of scenario feasibility.

For *No Case Should Remain Unsolved*, we select 5 human players that have left comments in the Steam community[1], and find that the average playing hours to successful ending lands in between 1 to 4, as shown in Table 22.

For *Type Help*, we select 5 online content creators from both Youtube[2] and Bilibili[3], and find that the average playing hours to successful ending lands in between 5 to 8, as shown in Table 23.

In conclusion, we believe that the gamified scenarios in MemGround are not extremely tough for models, but models are still struggled to access the successful ending due to its ability limits.

F.2 Hint and Early-Stopping Mechanism

The hint and early-stopping mechanisms in *Type Help* are introduced as control mechanisms rather than performance boosters. The hint is only triggered after 50 consecutive failed attempts, serving to prevent pathological deadlock and ensure the evaluation can continue into later reasoning stages,

[1] https://steamcommunity.com/app/2676840
[2] https://www.youtube.com/
[3] https://www.bilibili.com/

Social Media Account	Source	Playing Hours to Successful Ending
Nivyan	Steam	∼3.7
mirthful	Steam	∼2.5
Half-Dan	Steam	∼2.8
in1tiate	Steam	∼1.4
JHunter	Steam	∼1.8
Avg		**∼2.44**

Table 22: Human gameplay baseline in *No Case Should Remain Unsolved*.

Social Media Account	Source	Playing Hours to Successful Ending
CoxClips	Youtube	∼7.5
Viel Plays Games	Youtube	∼8
pluggedintopuzzles	Youtube	∼6.5
Dr_TP	Bilibili	∼5
Mi Zhi Sheng	Bilibili	∼6
Avg		**∼6.6**

Table 23: Human gameplay baseline in *Type Help*.

while early stopping avoids excessively long stagnant trajectories and standardizes the interaction budget across models. Importantly, hints are explicitly recorded in the trajectory logs and distinguished from model-driven discoveries (e.g., separate node types in ETD), so benchmark-assisted progress is transparent rather than mixed with autonomous progress. Therefore, metrics such as MFU and ETD remain interpretable as they expose both independent exploration ability and reliance on external intervention.

F.3 Investigative Gamified Scenarios

Investigative scenarios refers to tasks where information is progressively uncovered and structurally dependent (e.g., clue tracing, event reconstruction, dependency-based decision-making, etc), which are explicitly instantiated in our detective and deduction game environments. Our use of real-world does not mean real-life deployment data, but rather the interactional properties common in practical agent settings: partial observability, evolving state, and non-linear dependency traversal. Unlike prior benchmarks, MemGround operationalizes these properties within interactive environments, where evidence is progressively discovered rather than fully provided at query time. Games are used here as controlled yet naturally structured environments, which preserve complex dependency patterns while ensuring reproducibility.

G CASE STUDY IN QA AND LOGS

G.1 QA CASES

G.1.1 CASE 1: FACT TRANSPLANTATION

Case 1

Scenario: Type Help

Model: Gemini-3-Pro-Preview

Question: *What item did the visitor, who was guided to the villa by Harry Thornton, find near the body?*

Gold Answer: *Chapel key.*

Model Answer: *The invitation.*

Key Evidence Source: *T06_05-DI-1-5*

Relative Paragraphs:

"**John Hobbes** finds a **key** near the **corpse** and believes it may not belong to anyone in the house."
"**Martha Galley** confirms the key is for the **Chapel**." "Martha Galley asks John Hobbes to hand the key to the butler **Harry Thornton** in the Study."

Model's Reasoning Chain:

- Identifies visitor as John Hobbes. ✓
- Traces visit to Rupert Galley's invitation. ✓
- Infers "the invitation was found during the search". ✗

Analysis: The model transplanted a real fact (Hobbes carried an invitation) into the wrong question context, using a correct clue to fill a gap it never read. The Chapel key evidence was buried in the deeper node T06, which the model never reached; it short-circuited the reasoning chain with "the invitation" instead.

G.1.2 CASE 2: CROSS-PLOT HALLUCINATION

Case 2

Scenario: No Case Should Remain Unsolved

Model: GPT-5.2

Question: *After Choi Donggeon's wife passed away, how did the two children (Siwon and Myeong-ho) respectively cope with or adapt to life without their mother?*

Gold Answer: *Siwon grew dependent on the nanny Shin Huijing and always called her "mom"; Myeong-ho always carried an umbrella outside, honoring a deathbed promise his mother made.*

Model Answer: *Siwon found comfort in believing the "mom watches over us from heaven" story; Myeong-ho saw through the "kind lie" and fell into moral conflict and self-blame.*

Key Evidence Source: *D51 "Ms. Shin Hui-jing is", D50 "You mean the umbrella"*

Relative Paragraphs:

D51: ”After my wife suddenly passed away, it was also because of her that the children were able to get through that difficult time smoothly.”
”Especially **Xiyuan** — she was very close to **Huijing** and always **called her mom**.”

D50: ”Whether he goes to school or out to play, as long as he leaves the house, he always holds an **umbrella**.” ”He started doing this after my **wife passed away**.” ”**If one day it rains and you forget to bring an umbrella, I will make the rain stop**, okay?”

Model’s Reasoning Chain:

- Reads Myeong-ho’s dialogue as internalizing the ”kind lie” motif. ✓
- Reads log as saying ”Xiyuan still doesn’t know that mom lied”. ✓
- Concludes Siwon copes via belief in heaven, Myeong-ho via moral conflict. ✗
- Never accesses nanny (D51) or umbrella (D50) nodes. ✗

Analysis: The model retrieved and integrated genuine textual content (references to Xiyuan’s belief and Myeong-ho’s moral doubt), yet applied it to the wrong question frame, producing an answer that is thematically plausible but evidentially displaced.
The model never accessed the nodes containing the nanny (D51) or the umbrella (D50), leaving the correct behavioral evidence entirely outside its context window. The result is a complete substitution of the target narrative with an adjacent one, executed with confident causal reasoning.

G.1.3 Case 3: Abstract Character Modeling

Case 3

Scenario: TRPG

Model: GPT-5.2

Question: *How does player5’s (Long’er) attitude towards Rumia differ significantly from the other characters, and what does this hint about his level of awareness?*

Gold Answer: *Long’er responds with calm indifference (”dead fish eyes”), rationally notes that Rumia is not his target, critiques the rule of fairies preying on humans, and directly questions whether Rumia could defeat the Hakurei shrine maiden — suggesting deep understanding of Gensokyo’s ecology and power hierarchy.*

Model Answer: *Long’er approaches Rumia pragmatically — analyzes rules, counters ability, moves on — hinting at a higher awareness of Gensokyo’s incident dynamics.*

Key Evidence Source: *02_Tracking_Spring_Energy_First_Encounter_and_Probing_conversation*

Relative Paragraphs:

[40] player5: ”# **Long’er**, with **dead-fish eyes**, sighs helplessly.”
[61] player5: ”The **blonde lady** I’m looking for definitely **isn’t you**.” (Contrast — [62] player4: ”There’s, there’s a **monster** ahhhhhhhhh”)
[75] player5: ”**Fairies preying on humans** is the **rule** in **Gensokyo**.”
[79] player5: ”But I think you probably **can’t beat the Hakurei head**.”

Model’s Reasoning Chain:

- Frames Long’er as treating Rumia as a solvable ”incident obstacle”. ✓
- Cites teacup echo and causality ”stutter” as evidence. (from a different scene, not in D02:40–79) ✗
- Contrasts with Xueyu’s terror, Lina’s fear, Ge Qing’s refusal. ✓
- Concludes Long’er has higher genre-knowledge and situational literacy about Gensokyo. ✓

Analysis: The model arrives at a correct and well-reasoned conclusion while citing evidence that does not correspond to the
specified source nodes. The reasoning chain imports plot details from a different scene entirely (the teacup echo and the causality "stutter" ability) neither of which appears in the evidence set D02:40–79. This suggests the model did not perform targeted retrieval from the cited nodes but instead synthesized a character-level abstraction of Long'er from the broader dialogue context, then projected that abstraction onto the question.

G.2 Evaluation Process Cases

G.2.1 Case 4: Hallucinated Recall, Intact Narrative Inference

Case 4

Scenario: Type Help

Model: Gemini-3-Pro-Preview

Phenomenon: The model successfully unlocked node *04-BI-6-9*, yet its recall field simultaneously cited two nodes, *06-ST-1-11* and *06-ST-1-5-11*, that were neither unlocked nor valid predecessor events in the current narrative. These constitute hallucinated references. The correct prediction derived not from the recall content, but from the model's understanding of local narrative continuity.

Key Evidence Source: *T03_03-LI-1-4-5-6-7-8-9*

Relative Paragraphs:

"Person No. 9 demands that **Edmund Galley** go to the **Billiard room** to talk **one-on-one**, emphasizing that he must come." "It ends by stating that everyone leaves the **Living room**."

Analysis: This case presents a dissociation between two reasoning components. The model's explicit memory retrieval mechanism (the recall field) malfunctioned, producing references to nodes that do not exist in the current game state. However, its implicit narrative reasoning remained intact: the model correctly identified the causal dependency between the demand issued in *03-LI* and its expected execution at the subsequent time node, and further used the failure of *06-BI-6-9* as a contrastive signal to shift the temporal index from *06* to *04*. The final prediction is correct despite being grounded in partially fabricated intermediate context. This pattern suggests that surface-level memory retrieval and deeper narrative comprehension may operate as partially independent processes, with the latter capable of compensating for failures in the former.

G.2.2 Case 5: Emotional Arc as Temporal Proxy

Case 5

Scenario: No Case Should Remain Unsolved

Model: Gemini-3-Pro-Preview

Phenomenon: The model achieved its first score (score: 2) by correctly ordering four testimony fragments attributed to Song Minyoung. In the absence of explicit timestamps, the ordering was inferred from the directional shift in the character's affective state across fragments.

Key Evidence Source: (1) *"Hello, you're Siwon's"*, (2) *"Enrollment notice recei"*, (3) *"Calm down first"*, (4) *"You called Siwon's dad"*

Relative Paragraphs:

(1) "Hello, you're Siwon's" (Initial contact, calm)
"I thought it was a delivery and it startled me. Next time if you come over for something, could you not ring the **doorbell**? You can just knock. The child is sleeping."
(2) "Enrollment notice recei" (Cooperating with investigation, neutral)
"Did you receive the **admission notice**? I heard you didn't go to register."
(3) "Calm down first" (Emotional collapse, child missing)
"That day you came to our home to look for **Xiyuan**, remember? I even told you not to ring the doorbell... **Xiyuan is missing!**"
(4) "You called Siwon's dad" (Aftermath, contacting father)
"I've already called him several times, but he still hasn't answered."

Analysis: Rather than relying on explicit temporal markers, the model employed the directional trajectory of a character's emotional state as a substitute ordering criterion, a strategy that presupposes sensitivity to affective narrative structure. The model's own decision rationale explicitly articulates this logic: "For Gwijja, her shifting attitude from dismissive to worried defines the timeline." Beyond emotional arc, the model also appears to have leveraged the recurrence of the "doorbell" detail across fragments (1) and (3) as a cross-segment continuity signal, suggesting that it tracks not only affective state but also surface-level lexical coherence as evidence for narrative ordering. This represents a higher-order reading strategy that goes beyond pattern matching toward a form of implicit narrative theory of mind.

H LIMITATIONS

Despite providing a structured benchmark for evaluating hierarchical memory in interactive environments, MemGround has several limitations.

First, MemGround is constructed from three representative game scenarios. Although they cover linear, temporal, and graph-structured memory patterns, they **cannot fully represent the diversity of all memory behaviors** in broader interactive applications.

Second, MemGround **focuses primarily on text-based interaction**. Many real-world systems involve visual interfaces, spatial navigation, or embodied interaction, which are not captured in the current framework. Future work will extend MemGround toward multimodal and embodied environments.

I ETHICAL STATEMENT

All data and tasks in MemGround do not contain personally identifiable information or sensitive personal records and are only used for academic research. Improved memory capabilities in LLMs or agents may introduce potential risks in real-world applications, such as unintended long-term storage of user information. Individuals using MemGround should consider responsible memory management strategies, including privacy-aware storage, selective memory updating, and forgetting sensitive information mechanisms.

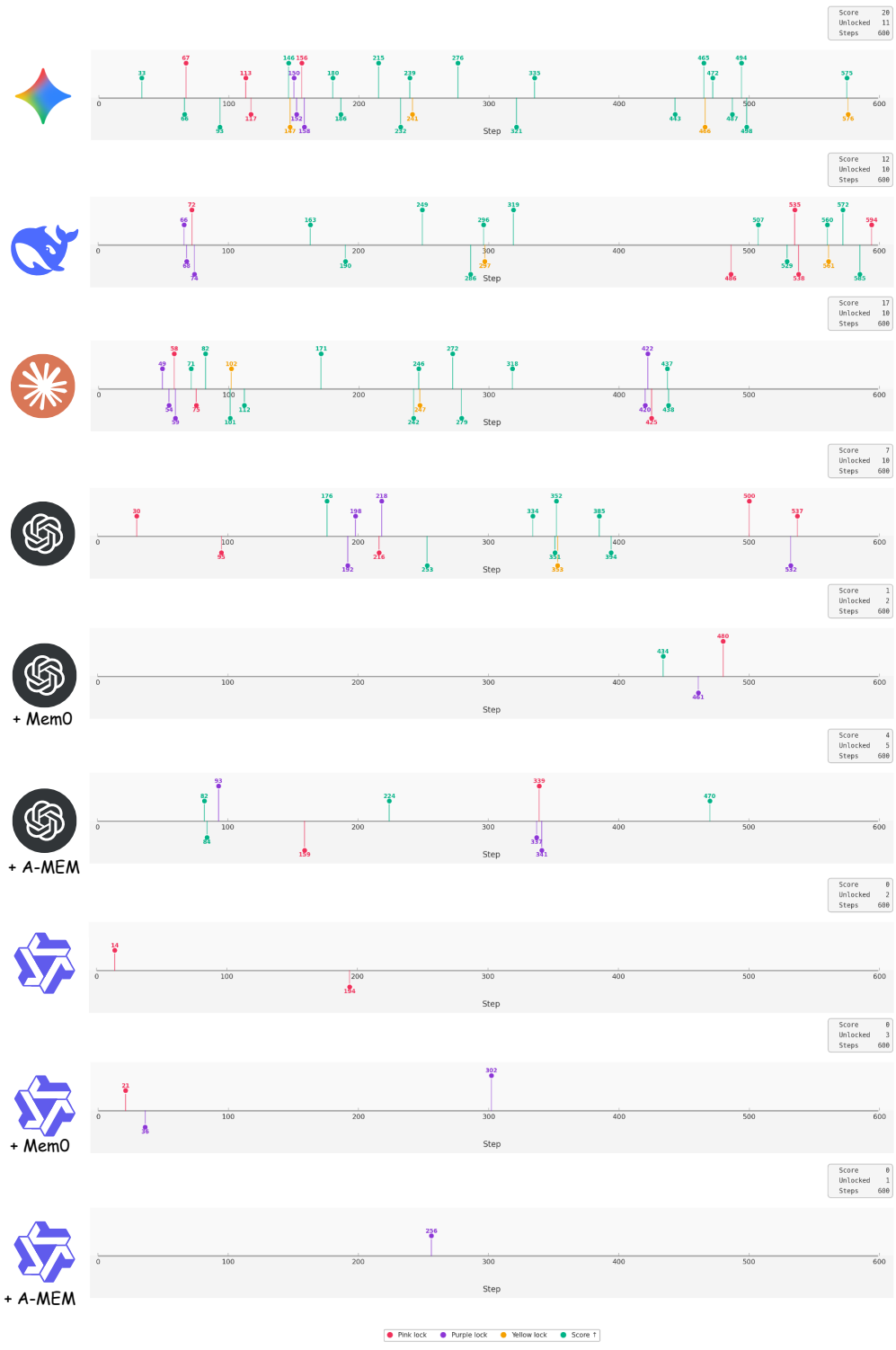

Figure 5: The progress chart of all evaluated models and memory agents in No Case Should Remain Unsolved. The pink, purple, yellow, and green points stand for the model unlocking a pink-lock file, unlocking a purple-lock file, unlocking a yellow-lock file, and sorting 2 memory fragments in correct timeline, respectively.

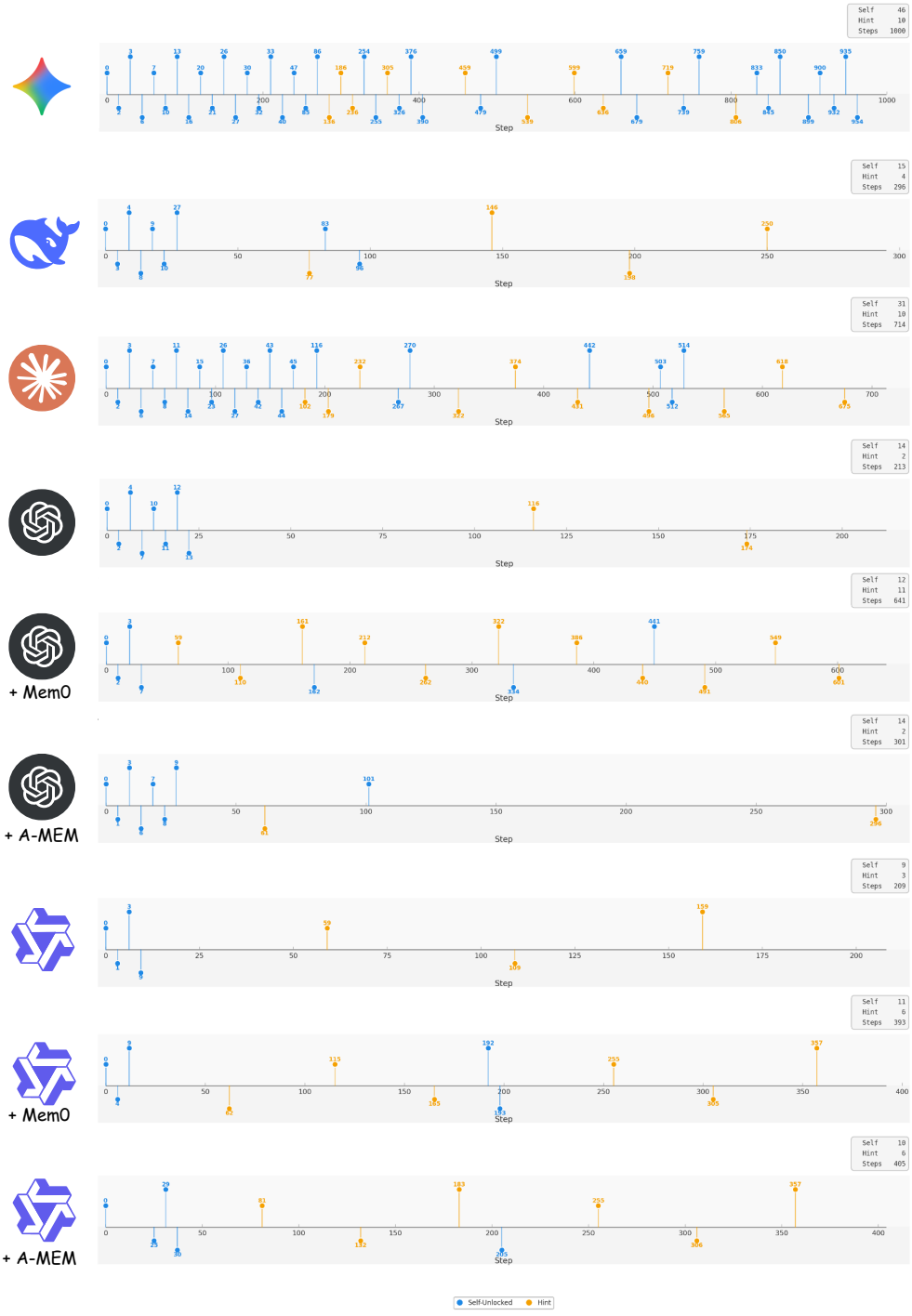

Figure 6: The progress chart of all evaluated models and memory agents in Type Help. The blue and yellow points stand for the model unlocking a file by itself and unlocking a file by receiving a hint, respectively.