SHARED WORLDS, PRIVATE MINDS: STRUCTURED MEMORY FOR LONG-FORM WRITING AS WORLD CREATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents that write long-form fiction need an explicit memory of the evolving storyworld to keep new events consistent with established facts. Such memory must keep heterogeneous narrative information distinct, integrate story developments across granularities, and recover dependencies that a writing request leaves implicit. We present NarraWorld, a structured memory system for long-form writing that treats memory construction as world creation. From a shared evidence-grounded graph, NarraWorld derives four connected views: world facts, per-character beliefs, open developments, and hypothetical branches (possible-world continuations). Hierarchical aggregation with atomic closure consolidates events into scenes, plotlines, and plots, keeping each higher-level node traceable to its constituent source spans. For retrieval, planned reconstruction infers a query's dependencies from the current narrative situation and a preview of memory, then assembles the relevant records within a token budget. Across three writing benchmarks, NarraWorld achieves the strongest aggregate results. Its memory also transfers to situated role-playing and largely preserves recall on a general-purpose long-term memory benchmark, paving the way for agents that sustain coherent storyworlds across diverse narrative tasks.

1 INTRODUCTION

LLM agents are increasingly used to assist long-form fiction writing: drafting stories (Yang et al., 2022), expanding outlines into full narratives (Yang et al., 2023), and co-authoring with human writers (Mirowski et al., 2023; Huot et al., 2025). At this scale, quality depends on narrative consistency, meaning that new events must agree with facts already established in the story. A story's history can quickly outgrow a single model context, and even within long contexts, access to relevant information degrades when it is distant or appears in the middle of the input (Liu et al., 2024). Long-form writing therefore requires an explicit memory for tracking the evolving story state.

Such memory must represent the latent storyworld rather than only the surface story text. This poses three challenges for existing agent memory systems. First, narrative information is heterogeneous: a passage can simultaneously mix objective world facts, character beliefs, plot developments, and foreshadowed possibilities. Although these elements belong to the same storyworld, they may differ or even conflict in ways that are difficult to detect from the surface text. For example, after one character places a key in a drawer, another moves it into a hollow book; the first character's belief that the key remains in the drawer then conflicts with its actual location. Memory that collapses these statuses (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2026) will conflate information that should remain distinct, leading to contradictions in the story (**Figure 1**).

Second, narrative development often accumulates gradually across many passages. For example, a betrayal may build up over ten chapters through small acts of concealment and misdirection. Memory must be able to summarize it as a single betrayal from multiple chapters without losing the specific facts. Integrating such evidence across multiple narrative granularities challenges existing methods, whose memories are typically organized as flat summaries, isolated facts, or event-level records (Piper et al., 2021; Lyu et al., 2025; Wang et al., 2025).

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.

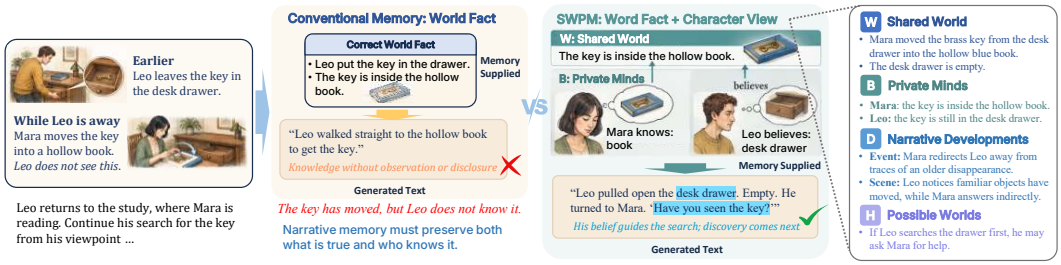


Figure 1: Illustrative comparison of conventional memory and NarraWorld. Leo leaves the key in the desk drawer, but Mara moves it into a hollow book while he is away. Conventional memory supplies the updated location, causing Leo to walk directly to the book. NarraWorld preserves both the shared fact that the key has moved and Leo's private belief that it remains in the drawer, so the continuation is: he checks the drawer, discovers it empty, and asks Mara about the key.

Third, a writing query often provides too little information to retrieve all the memory needed for the task. For example, a simple instruction such as "rewrite the betrayal scene" requires not only memories of the betrayal itself but also the character relationships that the revision may affect. Existing systems retrieve by similarity (Park et al., 2023; Zhong et al., 2024), so they are bounded by the request text itself and cannot extend to these implicit dependencies.

To this end, we present NarraWorld, a memory system that treats long-form writing as world creation. It maintains four kinds of records: world facts, per-character beliefs, open developments, and hypothetical branches. These four layers are connected views derived from a shared evidence-grounded graph, with the belief view indexed by character. A hierarchical aggregation mechanism with atomic closure consolidates events into scenes, plotlines, and plots. Every higher-level node can be traced through the hierarchy to all of its constituent events and source spans, so a multi-chapter development can be retrieved as a single unit and expanded back into its exact underlying facts. During memory retrieval, a planned reconstruction determines a query's dependency before retrieving. A planner reads the current narrative situation then checks this against a preview of memory, derives the dependencies, retrieves and assembles the related records within a token budget.

Experiments show that across three writing benchmarks, NarraWorld achieves the highest aggregate writing quality. The same memory transfers to situated role-playing, which likewise requires maintaining an underlying storyworld. On a general-purpose long-term memory task, the same memory largely preserves recall performance. Our contributions are fourfold:

(1) We propose a shared evidence-grounded memory substrate that separates world facts, per-character beliefs, open developments, and hypothetical branches.

(2) We propose a hierarchical aggregation mechanism with atomic closure that organizes events into scenes, plotlines, and plots while preserving links to constituent events and source spans.

(3) We propose a planned reconstruction procedure that derives implicit continuation dependencies from the narrative situation before retrieving and assembling memory.

(4) Experiments show that NarraWorld improves long-form consistency and writing quality while transferring to situated role-playing.

2 NarraWorld: Structured Memory as World Creation

As shown in Figure 2, NarraWorld has three components. First, an evidence-grounded substrate separates world facts, private character knowledge, narrative developments, and possible worlds. Second, a hierarchy organizes atomic events into scenes, plotlines, and plots. Atomic closure links every abstraction to its events and source evidence. Third, planned reconstruction uses the current narrative situation and a memory preview to infer implicit dependencies before retrieving and assembling an admissible context under a budget. The operational construction sequence is detailed in Appendix A.

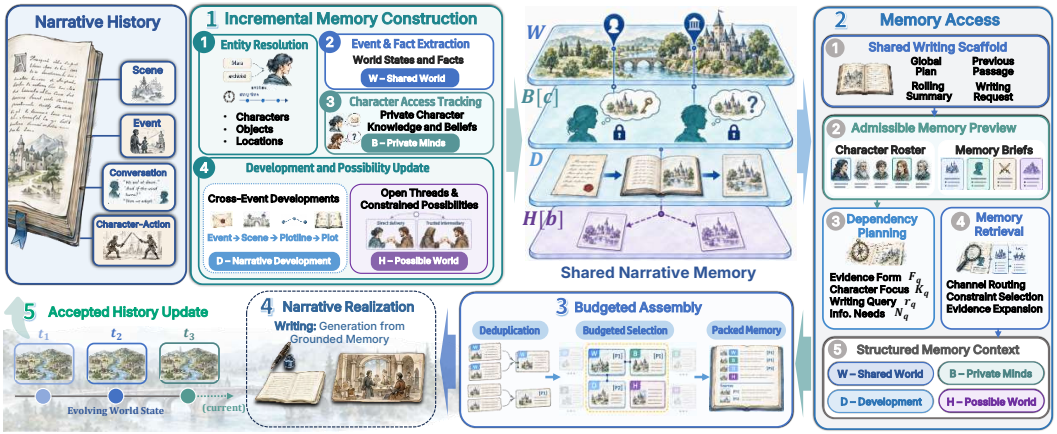


Figure 2: Overview of the NarraWorld continual world-creation cycle. Narrative history is formed into shared-world and character-belief memory, then reconstructed for the next writing act or situated agent action. The resulting continuation becomes the next source unit in the evolving narrative.

2.1 Task Formulation

We adopt the plan–write–revise setting of Re3 (Yang et al., 2022). At writing step τ, the agent receives a scaffold

$$x_\tau = (O, \rho_\tau, P_{\tau-1}, r_\tau, \kappa_\tau)\,, \tag{1}$$

where O is the global outline, ρ_τ a rolling summary, $P_{\tau-1}$ the preceding passage or chapter, r_τ the current request, and κ_τ the output constraints. We represent the complete narrative request as $q = (x_\tau, \eta_q)$, where η_q specifies the teller, recipient, focalizer, story time, discourse position, and applicable branch. NarraWorld reconstructs memory for the next writing action from this request.

2.2 Multi-View Memory over a Shared Narrative Graph

Long-form narrative memory must preserve two properties. World facts, character beliefs, narrative developments, and hypothetical continuations require different evidential status, yet they often refer to the same entities, events, times, and source passages. Narratology similarly treats story, narrative discourse, and narrating as distinct but interdependent levels (Genette, 1983). Inspired by this principle, NarraWorld therefore maintains one evolving narrative structure and forms four connected memory views over it.

Four views projected from the shared structure. At time τ, the shared structure G_τ connects recurring entities, events, times, and source evidence across views. The memory M_τ contains four views: W_τ represents established world state, including events that have occurred, currently valid facts, entity states, and relationships. $B_\tau[c]$ represents the current perspective, such as knowledge and beliefs, of character c. The two views refer to the same events and propositions but assign them different perspective-dependent status. Since long-form writing also depends on narrative developments (e.g., conflicts, relationships) that develop across passages and chapters, the development view D_τ connects the relevant events and state changes and records the current status of each development, such as open, advanced, blocked, or resolved. During story writing, an open development may suggest multiple future branches. H_τ stores bounded hypothetical continuations (possible-world branches) associated with open developments. All four views are obtained by projecting G_τ.

$$\begin{aligned} W_\tau &= \Pi_\tau^W(G_\tau), \quad & B_\tau[c] &= \Pi_\tau^{B[c]}(G_\tau), \\ D_\tau &= \Pi_\tau^D(G_\tau), & H_\tau &= \Pi_\tau^H(G_\tau). \end{aligned} \tag{2}$$

Each projection starts from a different set of nodes and expands as a different subgraph in G_τ. The world projection Π_τ^W contains source-supported events, facts, states, and relationships that are valid at time τ. The character projection $\Pi_\tau^{B[c]}$ starts from character c and follows action and

belief links (e.g., observation and communication) to propositions held by the character, which may be either world facts or false propositions that contradict them. The development projection Π_τ^D starts from nodes associated with narrative development, such as goals and conflicts, and organizes events through progression, causation, and state change (see Section 2.3). The possibility projection Π_τ^H starts from unresolved developments and collects their constrained candidate continuations and marks them with hypothetical status.

Incremental memory construction. During the writing process, given a new unit U_τ of the source text, NarraWorld updates the shared narrative structure:

$$G_\tau = \mathrm{Update}(G_{\tau-1}, U_\tau). \tag{3}$$

The update contains four stages. First, new character, location, and object entities in U_τ are linked to existing entities; unmatched entities create new records. Second, using these resolved entities, NarraWorld adds new events together with the facts and state changes supported by those events. When an event changes an existing state, NarraWorld preserves the validity period of the previous state and establishes a new current state instead of overwriting the earlier record. Third, NarraWorld determines each character's access to the established events and facts. When a character observes an event, receives information, or draws an inference, NarraWorld connects that character to the corresponding event or proposition and records the acquisition mode. Fourth, NarraWorld updates narrative developments from the new events and state changes. For developments that remain unresolved, it generates conditional candidates constrained by the current world state and character knowledge and adds them to H_τ.

An example for multi-view update. For the source text "Mara moved the key from the drawer into the hollow book while Leo was away," NarraWorld records the movement event and updates the key's location. Because the event and resulting location are source-supported and current, Π_τ^W includes them in W_τ. Mara participates in the event, so her access link allows $\Pi_\tau^{B[\text{Mara}]}$ to include the new location. Leo has no observation or communication link to the move, so $\Pi_\tau^{B[\text{Leo}]}$ retains his earlier belief that the key is in the drawer. The movement event advances the development concerning Leo's search; Π_τ^D therefore includes the event under that development. If the development remains unresolved, Π_τ^H retrieves candidates such as Mara informing Leo or Leo discovering the key himself.

2.3 Organizing Memory across Narrative Scales

NarraWorld organizes narrative memory from events to scenes, plotlines, and plots. This hierarchy connects local evidence with narrative developments that accumulate across the story:

$$\mathcal{E}_\tau \longrightarrow \mathcal{S}_\tau \longrightarrow \mathcal{L}_\tau \longrightarrow \mathcal{P}_\tau, \tag{4}$$

corresponding to events, scenes, plotlines, and plots.

$\mathcal{E}_\tau$ consists directly of event records in canonical narrative memory. Each event forms an atomic node and retains its participants, temporal scope, source unit, and supporting evidence. $\mathcal{S}_\tau$ consists of derived scene bundles. Each bundle gathers the events, state changes, relationship evolution, and plan progress within one local narrative unit and summarizes them as a scene. $\mathcal{L}_\tau$ identifies persistent goals, conflicts, relationships, and plans across scenes. Candidate continuity comes from shared characters, continuous content, state transitions, developmental relations, and thread evolution. Given the selected scenes, an LLM produces a plotline summary and the connections among its scenes, with explicit references to the supporting scene records. $\mathcal{P}_\tau$ combines plotlines that share participants and exhibit common continuity or developmental signals. The resulting plot forms a connected narrative arc and aggregates its member plotline summaries in temporal order.

Each higher-level node retains an atomic closure of its member events, $W/B/D$ records, and supporting source spans. The hierarchy can therefore represent a multi-scene development as one plotline or plot while preserving every lower-level record from which that representation was derived. Retrieval can select the higher-level unit and then unfold it into the relevant actions, beliefs, and source evidence. These derived views evolve alongside atomic memory. Hypothetical H records remain outside the hierarchy until subsequent source text realizes them as $W/B/D$ evidence.

2.4 RECONSTRUCTING MEMORY FOR SITUATED WRITING

Writing requests are underspecified retrieval queries. A request may name an event but omit affected relationships, character knowledge boundaries, and unresolved consequences. NarraWorld reconstructs memory in four stages: admissible view projection, dependency planning, multi-scale retrieval, and budgeted assembly.

Situation-conditioned memory access. Relevant memory and usable memory differ by story time, disclosure state, focalization, and branch. For example, in a mystery, the detective may not yet know the truth, while the culprit's identity should not be disclosed prematurely. The narrative situation η_q defines these access conditions and situates memory in an occasion of telling (Herman, 2009; Piper et al., 2021). It determines the admissible view

$$V_q = \Pi_{\eta_q}(\mathcal{M}_\tau), \tag{5}$$

where Π identifies records available under the request's temporal, disclosure, and branch scope. It preserves record authority and belief holders. The author-facing view can include beliefs from several characters. Holder labels preserve their separation. Focalization determines which character's available information may be expressed in the narration.

Preview-guided dependency planning. Similarity search follows explicit query content. Implicit dependencies require an intermediate planning step. Appendix D compares this preview-conditioned design with deterministic and model-based query-centric retrieval on LoCoMo, and with scaffold-conditioned query planning on ConStory-Bench. The planner receives the request and $\sigma(V_q)$, a compact preview containing the full character roster, brief descriptions, developments, and $W/B/D/H$ records. The current passage and authorial brief provide immediate context. Outlines provide intended directions. Planning produces

$$(F_q, K_q, r_q, \mathcal{N}_q) = \Psi(x_\tau, \eta_q; \sigma(V_q)), \tag{6}$$

where F_q specifies the required evidence form and composition, K_q identifies focal characters, r_q states the content-focused retrieval intent, and $\mathcal{N}_q$ lists missing information needs and their dependencies. The system resolves K_q against the admissible character view and translates these semantic decisions into retrieval constraints. This character-centered planning can recover relevant entities even when they are absent from the surface request.

Scale-adaptive retrieval. Narrative dependencies map to different memory units: facts to atomic records, local interactions to events and scenes, and ongoing developments to scenes and plotlines. Retrieval uses these planning outputs to expand the admissible candidate set and search for records that satisfy the remaining needs:

$$E_q = \mathrm{Expand}(K_q, F_q; V_q) \cup \mathcal{R}(x_\tau, r_q, \mathcal{N}_q, K_q, F_q; V_q). \tag{7}$$

Expansion follows the focal characters and requested evidence composition to construct an admissible candidate set; search then satisfies the remaining needs. It combines lexical and semantic matching with temporal, epistemic, and developmental access. Higher-level matches unfold into their supporting records. The retrieval uses holder and time to distinguish world propositions, character stances, and acquisition events.

Budgeted assembly. Given a memory budget b, NarraWorld renders the retrieved evidence as a readable context

$$C_q^{\mathrm{mem}} = \mathrm{Reconstruct}(q, V_q; b), \qquad \ell(C_q^{\mathrm{mem}}) \le b, \tag{8}$$

where ℓ measures rendered length. Evidence is grouped into $W/B/D/H$ sections, with B organized by character. Hypothetical candidates retain their premises and conditions without presenting possibilities as accepted evidence. The writer agent receives $X_q = x_\tau \oplus C_q^{\mathrm{mem}}$ and uses it to produce the next passage. Before history exists, the scaffold supplies the authorial brief and constraints; subsequent memory updates use only accepted source text.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Tasks and evaluation data. Our primary evaluation covers three writing benchmarks: ConStory-Bench has 200 stories equally divided among generation, continuation, expansion, and completion (Li et al., 2026); TMAS has 168 source-grounded continuation checkpoints (Huot et al., 2025); and LSGC has five official starters, each extended to 14 chapters and at least 40,000 words (Migal et al., 2024). We retain the original writing tasks while adapting generation into incremental segments that better reflect writers' workflows and expose memory use across successive writing steps. For LSGC, every condition is evaluated in three complete runs; the reported scores average the resulting 15 story instances, each first aggregated over its 14 chapter judgments and final-piece judgment. We also use RMTBench's 2,880 multi-turn role-playing trajectories to assess whether NarraWorld transfers beyond writing (Xiang et al., 2025), and LoCoMo's 1,540 main recall questions to test whether its writing-oriented design compromises general-purpose memory performance (Maharana et al., 2024). Appendix B details the writing protocols.

Baselines and model configuration. We compare NarraWorld with Full Context, Scaffold Only, Mem0 (Chhikara et al., 2025), Graphiti (Rasmussen et al., 2025), Hindsight (Latimer et al., 2026), AriGraph (Anokhin et al., 2025), and StructMem (Xu et al., 2026), adding MemoryOS (Kang et al., 2025) for LoCoMo and Emotional RAG for RMTBench. Full Context gives the TMAS writer the visible source prefix directly; Scaffold Only omits retrieved memory from the shared writing scaffold. All benchmarks use GPT-5-mini for memory-system calls, GPT-5 for generation, and DeepSeek-V4-Pro for judging. Systems retain their native memory procedures, with embedding and reranking services fixed by benchmark. Rendered-memory budgets are 12k tokens for ConStory-Bench and LSGC, 4k for RMTBench, and 8k for LoCoMo (2k–16k in the sensitivity analysis). Throughout the paper, "Mem. tokens" denotes the average rendered memory length under the corresponding budget. Conditions share task inputs, writing models, and evaluation protocols, and exclude future text and evaluator-only references.

Metrics and evaluation protocol. ConStory-Bench reports Consistency Error Density (CED) across five categories and 19 subtypes, plus Words per Penalized Error (WPE), the story-level word count divided by one plus the subtype count (Li et al., 2026). TMAS reports four primary dimensions—Plot, Creativity, Development, and Language Use—following its source-grounded continuation setting (Huot et al., 2025). We also report their arithmetic mean, TMAS Quality, as a supplementary aggregate; auxiliary diagnostics appear in Appendix B.1. Because the original LSGC evaluation relies heavily on human annotation, we use the 16 Longform Writing Bench criteria for scalable quality evaluation (Paech, 2025), reporting direction-corrected 0–100 scores for Character, Plot, World/Coherence, Style/Creativity, and Prompt Adherence; Overall equally weights all criteria. RMTBench reports Emotional Expression (EE), Emotional Comprehension (EC), Plot Advancement (PA), Character Understanding (CU), Character Maintenance (CM), and User Preference Awareness (UPA) (Xiang et al., 2025). LoCoMo reports Overall Main and single-hop, multi-hop, temporal, and open-domain scores (Maharana et al., 2024). Appendix C provides their definitions.

3.2 OVERALL RESULTS

Quantitative results. Across all three writing benchmarks, NarraWorld provides the strongest overall results (Tables 1–3). It leads all six reported ConStory-Bench measures, including CED of 0.185 versus 0.200 for Full Context and 0.385 for Graphiti, as well as the highest WPE. On LSGC, averaged over three runs, it leads Character (91.12), Plot (90.50), Style/Creativity (88.97), and Overall (90.24), ahead of StructMem (86.92) and Full Context (85.03); the full rubric shows a remaining Full Context advantage on World/Coherence (Appendix 8). Run-level ranges and sample standard deviations are reported in Appendix 10. NarraWorld also leads 13 of the 16 LSGC criteria and the TMAS Creativity (84.90), Development (85.60), and aggregate Quality measures, while Full Context leads TMAS Plot (80.03) and Language (83.42). Graphiti uses more rendered memory than NarraWorld on both ConStory-Bench ($5,971$ vs. $3,231$) and LSGC ($10,542$ vs. $6,810$), yet remains weaker on consistency and long-form quality, indicating that more connected history is insufficient without perspective and development modeling. NarraWorld's gains over the strongest LSGC memory baseline, StructMem, further support narrative-specific views and hierarchical organization. Full

Table 1: ConStory-Bench consistency results. CED and task-specific errors are lower-is-better; WPE is an auxiliary higher-is-better word-per-error measure.

Method	CED ↓	WPE ↑	Generation	Continuation	Expansion	Completion	Mem. tokens
NarraWorld	**0.185**	**10,520.4**	**0.120**	**0.235**	**0.098**	**0.287**	3,231
Graphiti	0.385	9,480.5	0.290	0.415	0.310	0.525	5,971
StructMem	0.450	9,215.3	0.355	0.495	0.380	0.570	3,891
Mem0	0.955	7,845.2	0.812	0.985	0.920	1.103	2,179
AriGraph	1.125	7,250.6	0.945	1.180	1.055	1.320	3,050
Hindsight	1.215	6,985.4	1.050	1.265	1.140	1.405	3,280
Scaffold Only	1.350	–	1.220	1.410	1.280	1.490	–
Full Context	0.200	9,847.8	0.136	0.255	0.113	0.298	–

Table 2: LSGC long-form writing results (↑). Direction-corrected macro means over five starters and three runs; memory in tokens.

Method	Character	Plot	Style/ Creat.	Overall	Mem. tokens
NarraWorld	**91.12**	**90.50**	**88.97**	**90.24**	6,810
Mem0	84.34	81.71	73.69	80.18	2,876
Graphiti	82.44	87.11	67.34	79.10	10,542
Hindsight	75.32	79.46	67.73	74.91	3,412
AriGraph	72.56	76.87	67.56	73.03	3,850
StructMem	88.87	87.74	83.60	86.92	5,428
Scaffold Only	66.50	69.22	70.86	69.61	–
Full Context	87.17	87.27	78.79	85.03	–

Table 3: TMAS continuation results (↑). Primary dimensions are macro means; more diagnostic metrics are reported in Appendix Table 7.

Method	Plot	Creat.	Develop- ment	Lang.	Mem. tokens
NarraWorld	78.92	**84.90**	**85.60**	82.62	1,593
Mem0	77.60	84.27	82.19	83.08	941
Graphiti	79.13	83.65	82.40	82.92	1,873
Hindsight	75.58	79.65	76.83	79.42	1,481
AriGraph	76.00	77.56	75.79	77.75	981
StructMem	77.71	84.06	82.25	82.17	1,927
Scaffold Only	77.10	81.50	80.90	82.20	–
Full Context	**80.03**	83.33	83.73	**83.42**	–

Context remains a strong reference, but exposing complete history at every step is difficult for very long iterative writing and is not a practical deployment strategy.

Human evaluation. Human readers support the main writing conclusion: on ConStory-Bench, NarraWorld and Full Context have nearly identical human CED (0.095 and 0.096), both below Graphiti (0.289). For LSGC, readers evaluated complete stories across five quality dimensions. NarraWorld obtained the highest LSGC Human Overall (84.50 versus 80.50 for StructMem and 80.06 for Full Context). Judge–human agreement was moderate overall (Spearman $\rho = 0.55$), with reader variation in subjective literary judgments; evaluation protocols, alignment details, and human rubrics appear in Appendix F.

3.3 Ablation Studies

The component ablations show distinct, nonuniform contributions (Figure 3; Table 4). Removing character beliefs weakens perspective-sensitive continuity, while removing possible worlds lowers creativity and development; removing cross-event organization causes the broadest degradation across consistency and long-form quality. The largest decline occurs when cross-event organization is removed, suggesting that retaining individual facts is insufficient without connecting them into sustained narrative developments. The retrieval-planning ablation holds offline memory and budget fixed: preview-guided planning improves every ConStory-Bench task, with the largest gains on continuation and completion. Complete breakdowns appear in Appendix E, and retrieval comparisons in Appendix D.

3.4 Case Studies

The case studies show how NarraWorld supports writing-specific retrieval. In the selected ConStory-Bench continuation, it jointly retrieves an unfinished fox drawing, the narrator's knowledge of a condition attached to their mother's box, and the father's promise to give them space. The continuation completes the drawing while the father remains outside and refrains from asking to see it, combining

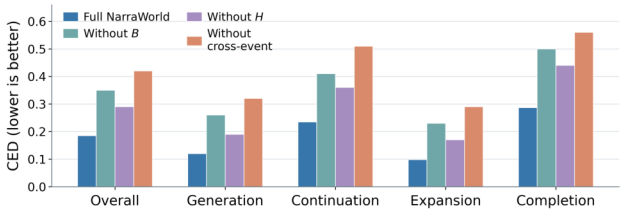


Figure 3: ConStory-Bench ablations across task types. CED is lower-is-better. The four settings are Full NarraWorld, without character beliefs (B), without possible worlds (H), and without cross-event organization.

Table 4: LSGC ablation Overall scores (↑), averaged over five starters and three runs.

Setting	Overall ↑
Full NarraWorld	**90.24**
Without character beliefs (B)	86.04
Without possible worlds (H)	88.74
Without cross-event organization	81.05

Table 5: RMTBench multi-turn role-playing results (↑). EE/EC/PA/CU use 1–5 scores; CM/UPA are percentage success rates. Scores are averaged over applicable turns.

Method	EE ↑	EC ↑	PA ↑	CU ↑	CM ↑	UPA ↑	Mem. tokens
NarraWorld	**4.22**	**4.06**	**4.30**	**4.57**	87.15%	68.85%	704
Mem0	3.98	3.56	3.74	4.25	85.00%	67.31%	344
Graphiti	4.02	3.61	3.82	4.41	86.15%	53.85%	1,215
Hindsight	4.01	3.52	3.78	4.21	84.60%	65.40%	654
AriGraph	3.95	3.60	3.71	4.28	85.35%	68.10%	429
Emotional RAG	3.99	3.55	3.73	4.22	82.10%	67.30%	600
StructMem	3.96	3.49	3.67	4.38	85.71%	50.00%	2,357
Full Context	3.92	3.52	3.66	4.30	**87.86%**	**69.23%**	–

a factual callback with private access and relational commitment. The full comparison appears in Appendix H.1; further cases appear in Appendix H.3.

3.5 Extension to Situated Role-Playing

Because NarraWorld reconstructs the storyworld, it also supports role-playing tasks that depend on character beliefs. On RMTBench it leads all four 1–5 dimensions, including plot advancement and character understanding (Table 5); Full Context is slightly higher on the percentage-based CM and UPA measures. This transfer shows that separating shared events from character-specific access helps an agent preserve its role while extending the situation.

3.6 General-Purpose Memory

To test transfer beyond writing, we evaluate NarraWorld on LoCoMo. Under the shared main budget, it leads or ties four of five reported measures, with Graphiti ahead only on single-hop questions (Table 6); NarraWorld also ranks first throughout the budget sweep (Figure 4). Complete category-level results appear in Appendix C.5. For this QA task, the default memory view excludes hypothetical-world records (H): exposing that view lowers Overall Main from 0.614 to 0.570 at the same 8k budget (Appendix 18). This contrast suggests that factual recall and creative continuation require different views of the same memory: hypothetical alternatives can support writing, while QA benefits from established evidence.

4 Related Work

Long-form Narrative Generation and Storyworld Modeling Long-form writing systems use hierarchical outlines, recursive drafting and revision, multi-agent planning, or explicit story-state models. These families include retrieved-summary revision and detailed outline control (Yang et al., 2022; 2023), structured co-writing roles (Mirowski et al., 2023; Huot et al., 2025), and temporal, validity-aware, or character-grounded world models (Wang et al., 2025; Lyu et al., 2025; Aluru et al., 2026). They establish the need for state that persists beyond the model context. Most evaluations

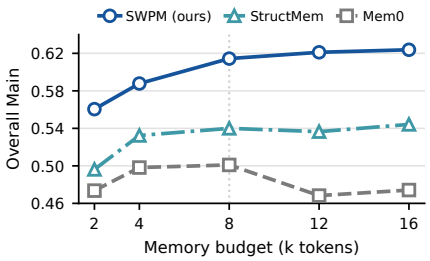


Figure 4: LoCoMo budget sensitivity. The vertical guide marks the 8k budget used in the main comparison.

Table 6: LoCoMo results at the shared 8k memory budget. Overall Main aggregates the four main question types (↑).

Method	Overall Main	Multi-hop	Single-hop	Temporal	Open-domain
NarraWorld	**0.614**	**0.458**	0.679	**0.649**	**0.392**
Graphiti	0.582	0.444	**0.698**	0.469	0.351
StructMem	0.540	0.435	0.575	0.595	0.362
Mem0	0.501	0.410	0.541	0.547	0.270
AriGraph	0.495	0.378	0.623	0.314	0.322
Hindsight	0.486	**0.458**	0.556	0.394	0.264
MemoryOS	0.453	0.371	0.537	0.354	0.292

still produce one complete story or continuation from a fixed prompt, prefix, or outline. NarraWorld instead treats writing as a sequence of accepted narrative acts: the agent generates one segment, updates the story state, and reconstructs memory for the next request. This makes memory responsible for consequences, perspective, and unresolved developments across steps.

Computational Narratology and Situated Character Minds Computational narratology distinguishes story events from their discourse-level selection and ordering, and narratorial voice from the consciousness through which events are perceived (Genette, 1983; Bal, 2009; Piper et al., 2021; Herman, 2009). Work on focalization, imperfect knowledge, false belief, and narrative time formalizes the associated perspective constraints (Bae et al., 2011; Teutenberg & Porteous, 2015; Sanghrajka et al., 2022; Kearns, 2020; Han et al., 2022). These distinctions motivate NarraWorld's four views: shared world state (W), private character beliefs (B), cross-event development (D), and retained possibilities (H), the last grounded in possible-world accounts of narrative (Ryan, 2013).

Persistent Memory for Situated Language-Model Agents Persistent-agent memories support storage, retrieval, reflection, and revision. Memory streams, working/external memory, and personalized retention represent three common designs (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024); graph and structured systems add associative traversal, event binding, and cross-event consolidation (Ji et al., 2026; Xu et al., 2026). Representative query-centric baselines span flat retained-memory stores (Mem0), temporal or entity-event graphs (Graphiti and AriGraph), and typed or multi-store organizations (Hindsight and MemoryOS) (Chhikara et al., 2025; Rasmussen et al., 2025; Latimer et al., 2026; Anokhin et al., 2025; Kang et al., 2025). These systems improve recall through query-matched records, but do not jointly model canonical facts, private knowledge, cross-event development, and possibilities. StructMem binds events but lacks epistemic and hypothetical views; graph systems do not resolve canonicity or private belief. NarraWorld addresses this gap with a shared substrate and task-conditioned reconstruction for writing and situated action.

5 Conclusion

We introduced NarraWorld, a narrative memory system framing long-form writing as world creation. It maintains shared world state, private character knowledge, cross-event development, and retained possible worlds within a source-grounded substrate. It then reconstructs request-specific views constrained by perspective, time, and branch. Across ConStory-Bench, Tell Me A Story, and LSGC, NarraWorld achieved the strongest aggregate writing results, with ablations showing complementary but nonuniform contributions from character beliefs, possible worlds, and cross-event organization. The same representation also supported situated character action and general-purpose memory use, while the LoCoMo ablation shows that the admissible memory view should be conditioned on the task. These results position narrative memory as the reconstruction of a dynamic, perspective-dependent world rather than relevance-based retrieval. The remaining challenges are the construction and planning cost of the richer representation, and deciding when speculative branches should be exposed; addressing these trade-offs in broader interactive settings is left for future work.

AI USE STATEMENT

Generative AI models perform memory construction, dependency planning, narrative generation, and evaluation under the reported protocols. AI assistants also supported conceptual and methodological refinement, implementation, experiment execution and analysis, literature review, figure preparation, and manuscript drafting and editing. The authors reviewed the resulting text, code, and analyses, checked reported results and citations, and take responsibility for the final manuscript and research artifacts.

REPRODUCIBILITY STATEMENT

Section 2 specifies the memory representation, update process, hierarchical organization, and online reconstruction procedure, while Section 3 reports the benchmark scope, model roles, memory budgets, and comparison controls. Appendix B documents benchmark-specific inputs and adaptations; Appendix C defines the metrics, aggregation rules, and evaluation prompts; Appendices D and E give the ablation protocols; and Appendices F and A.5 provide the human-evaluation procedure and selected system prompts. We retain model configurations, prompts, per-example outputs, and scoring procedures to support reproduction, including for models accessed through proprietary APIs. Code and supplementary materials will be released in anonymized form for review and made publicly available upon publication.

ETHICS STATEMENT

This work uses publicly available benchmark materials and generated fictional narratives. The human-evaluation analysis reports only anonymized aggregate judgments and does not expose personally identifiable or sensitive information. Evaluation materials and records are handled in accordance with the applicable requirements for the evaluation setting.

REFERENCES

Aayush Aluru, Chloe Ho, Muhammad Hammouri, Kerry Luo, Myra Malik, Ryan Lagasse, Arjun Bahuguna, and Vasu Sharma. From personas to plot: Character-grounded multi-agent story generation for long-form narratives, 2026. URL https://arxiv.org/abs/2607.00918.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for LLM agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 12–20, 2025. doi: 10.24963/ijcai.2025/2. URL https://www.ijcai.org/proceedings/2025/2.

Byung-Chull Bae, Yun-Gyung Cheong, and R. Michael Young. Toward a computational model of focalization in narrative. In *Proceedings of the 6th International Conference on the Foundations of Digital Games*, pp. 313–315, 2011. doi: 10.1145/2159365.2159423. URL https://doi.org/10.1145/2159365.2159423.

Mieke Bal. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, 2009.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Gérard Genette. *Narrative Discourse: An Essay in Method*. Cornell University Press, 1983.

Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1450–1470, 2022. doi: 10.18653/v1/2022.naacl-main.104. URL https://aclanthology.org/2022.naacl-main.104/.

David Herman. *Basic Elements of Narrative*. Wiley-Blackwell, 2009.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. Agents' room: Narrative generation through multi-step collaboration. In *International Conference on Learning Representations*, pp. 5150–5183, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0fbc8a83d93dd8021a4dd8d2d34138eb-Abstract-Conference.html.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for LLM agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/061241ca5a30c124478594f14b2e19955327b0c0.pdf.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, 2025. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Edward Kearns. Annotating and quantifying narrative time disruptions in modernist and hypertext fiction. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pp. 72–77, Online, 2020. Association for Computational Linguistics.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285, 2026. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Junjie Li, Xinrui Guo, Yuhao Wu, Roy Ka-Wei Lee, Hongzhi Li, and Yutao Xie. Lost in stories: Consistency bugs in long story generation by LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8400–8428, 2026. doi: 10.18653/v1/2026.findings-acl.410. URL https://aclanthology.org/2026.findings-acl.410/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638.

Zhiheng Lyu, Kevin Yang, Lingpeng Kong, and Dan Klein. FactTrack: Time-aware world state tracking in story outlines. 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. 2825–2848, 2025. doi: 10.18653/v1/2025.naacl-long.144. URL https://aclanthology.org/2025.naacl-long.144/.

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*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Aleksandr Migal, Daria Seredina, Ludmila Telnina, Nikita Nazarov, Anastasia Kolmogorova, and Nikolay Mikhaylovskiy. Overview of long story generation challenge (LSGC) at INLG 2024. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pp. 47–53, 2024. doi: 10.18653/v1/2024.inlg-genchal.4. URL https://aclanthology.org/2024.inlg-genchal.4/.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2023. doi: 10.1145/3544548.3581225. URL https://doi.org/10.1145/3544548.3581225.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

S. J. Paech. Longform creative writing benchmark, 2025. URL https://github.com/EQ-bench/longform-writing-bench. GitHub repository.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 298–311, 2021. doi: 10.18653/v1/2021.emnlp-main.26. URL https://aclanthology.org/2021.emnlp-main.26/.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Marie-Laure Ryan. Possible worlds. In Peter Hühn et al. (eds.), *The Living Handbook of Narratology*. Hamburg University, Hamburg, 2013. URL https://www-archiv.fdm.uni-hamburg.de/lhn/node/54.html.

Rushit Sanghrajka, R. Michael Young, and Brandon Thorne. HeadSpace: Incorporating action failure and character beliefs into narrative planning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pp. 171–178, 2022. doi: 10.1609/aiide.v18i1.21961. URL https://doi.org/10.1609/aiide.v18i1.21961.

Jonathan Teutenberg and Julie Porteous. Incorporating global and local knowledge in intentional narrative planning. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1539–1546, 2015. URL https://www.ifaamas.org/Proceedings/aamas2015/aamas/p1539.pdf.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pp. 1352–1391, 2025. doi: 10.18653/v1/2025.naacl-long.63. URL https://aclanthology.org/2025.naacl-long.63/.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, 2025. doi: 10.18653/v1/2025.findings-emnlp.730. URL https://aclanthology.org/2025.findings-emnlp.730/.

Buqiang Xu, Yijun Chen, Jizhan Fang, Ruobin Zhong, Yunzhi Yao, Yuqi Zhu, Lun Du, and Shumin Deng. StructMem: Structured memory for long-horizon behavior in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 122–146, 2026. doi: 10.18653/v1/2026.acl-short.12. URL https://aclanthology.org/2026.acl-short.12/.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4393–4479, 2022. doi: 10.18653/v1/2022.emnlp-main.296. URL https://aclanthology.org/2022.emnlp-main.296/.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 3378–3465, 2023. doi: 10.18653/v1/2023.acl-long.190. URL https://aclanthology.org/2023.acl-long.190/.

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. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

APPENDIX CONTENTS

A NarraWorld Memory Construction **14**

A.1 Source units and canonical records . . . 14
A.2 Incremental update and view projection . . . 14
A.3 Hierarchical organization and evidence closure . . . 14
A.4 Worked multi-view update . . . 15
A.5 Selected Prompt Instructions . . . 15

B Writing Benchmark Adaptations **17**

B.1 TMAS: source-faithful rolling checkpoints . . . 18
B.2 ConStory-Bench: task-adaptive continual writing . . . 18
B.3 LSGC: chapter-continual long-story generation . . . 19

C Evaluation Protocols, Metrics, and Judge Prompts **21**

C.1 ConStory-Bench: consistency error density . . . 21
C.2 Tell Me A Story: adapted continuation quality . . . 21
C.3 LSGC: full-story quality under the Longform Writing Bench rubric . . . 22
C.4 RMTBench: turn-level role-play scores . . . 23
C.5 LoCoMo: category-aware deterministic scoring . . . 23

D Retrieval Planning Ablations **24**

D.1 LoCoMo: Question-Centric Retrieval . . . 25
D.2 ConStory-Bench: Writing-Conditioned Retrieval . . . 25

E Memory Ablations Across Tasks **26**

E.1 TMAS: continuation quality . . . 26
E.2 LSGC: long-form narrative quality . . . 26
E.3 LoCoMo: QA-time memory-view access . . . 26

F Human Evaluation and Judge Alignment **27**

F.1 Evaluation Protocol . . . 27
F.2 ConStory-Bench: Human Consistency Assessment . . . 28
F.3 LSGC: Human Quality and Score Alignment . . . 28
F.4 Human Evaluation Rubric for Long-Form Narrative . . . 29

G Limitations **30**

H Tracing Memory into Narrative: Qualitative Case Studies **30**

H.1 ConStory-Bench: Objects, Promises, and Reconciliation . . . 30
H.2 LSGC: Evidence, Perspective, and Reassessing Loyalty . . . 30
H.3 Further Examples of Constraints and Consequences . . . 32
H.4 Qualitative Limitations . . . 33

A NarraWorld Memory Construction

This section expands the construction procedure in Section 2. Each accepted source unit is parsed once, integrated with the existing narrative structure, and made available through four memory views. Section A.5 provides abridged modeling and planning instructions.

A.1 Source units and canonical records

The input to an update is a trusted source unit U_τ, consisting of the newly accepted narrative text together with its source order, optional time anchor, participants, discourse position, and branch or authority metadata. The system preserves exact source spans and segments the unit at verified boundaries. The language-model stages operate on this unit and its local evidence, while persistent identifiers and source references are assigned by the system rather than predicted by the model.

The first stage grounds entity mentions against the committed entity index. It links aliases to existing entities when the evidence supports the match and creates a new entity otherwise. The next stage extracts events and propositions: an event records its participants, preconditions, effects, and causes, while a proposition records its subject, predicate, object, polarity, and truth status. Each proposition remains independently retrievable. A third stage models perspective and access, distinguishing a proposition that holds in the shared world from information that a particular character has observed, received, or inferred. The fourth stage emits state, relation, and development deltas. State deltas cover properties such as location, possession, goal, emotion, and obligation; development deltas mark a thread as opened, advanced, blocked, or resolved.

After extraction, the system verifies local references, restores exact evidence spans, and reconciles new entities and propositions with historical records. Only records with resolved references and retained provenance are committed to the append-only store. A state change closes the previous validity interval and creates a new record; it does not overwrite the historical observation. Thus, the store retains both what was true earlier and what is currently supported.

A.2 Incremental update and view projection

Let $G_{\tau-1}$ be the committed narrative structure before a new unit. The update can be summarized as

$$G_{\tau-1} + U_\tau \longrightarrow \text{extract} \longrightarrow \text{reconcile} \longrightarrow \text{commit} \longrightarrow \{W_\tau, B_\tau, D_\tau, H_\tau\}. \tag{9}$$

The four views are projections of the same committed records, not four independent memories. W_τ exposes source-supported world facts, events, states, and relations. $B_\tau[c]$ follows the access links of character c and may therefore contain a false belief or an incomplete account of the world. D_τ links events and state changes to the developments they advance or block. For an unresolved development, H_τ stores bounded conditional continuations whose preconditions are checked against the current world and character views. A record can participate in several projections while retaining one provenance chain.

Before each writing or question-answering step, the system projects the records supported by the available narrative prefix, together with their derived nodes. It retrieves from this admissible view under the current temporal, perspective, and branch constraints.

A.3 Hierarchical organization and evidence closure

Hierarchical nodes are derived after canonical records are committed. Events are direct canonical records. A scene groups temporally and semantically related events with shared participants, state changes, or development signals; material bundles may receive a short model-generated summary, while the underlying events remain authoritative. Plotline candidates require more than a shared character: they must also exhibit content continuity, temporal adjacency, a state or transition signal, or an existing development link. Plots group connected plotlines and do not force isolated threads into a larger arc. If a consolidation attempt fails its checks, the lower-level records are retained and no unsupported summary is committed.

Every scene, plotline, and plot stores its member records, source units, supporting evidence, temporal range, participants, and uncertainty. The resulting atomic-closure operation expands any selected

high-level node back to its constituent events, propositions, and exact source spans. Retrieval can therefore use a compact summary as an entry point without losing the evidence needed to inspect or regenerate the underlying narrative state.

A.4 Worked multi-view update

Consider the source unit "Mara moved the key from the drawer into the hollow book while Leo was away." Entity grounding links Mara, Leo, the key, and the book to existing records. Event extraction adds a movement event with the location transition as its effect. The world projection records that the key is now in the book and closes the earlier drawer-location interval. Mara's belief view receives the new location through her participation, whereas Leo's belief view remains unchanged because the unit contains no observation, communication, or inference linking him to the move. The event advances the development in which Leo is searching for the key. If that development remains open, the possibility view may retain conditional candidates such as Mara informing Leo or Leo discovering the key; these candidates are marked as hypothetical and are never written back as established world facts. This single update illustrates why the four views share entities and events while preserving different evidential and epistemic status.

A.5 Selected Prompt Instructions

The following abridged instructions summarize memory formation and writing-time planning. Illustrative examples clarify the distinctions among events, facts, beliefs, developments, and possible worlds.

Narrative Memory Planning. The planner selects relevant evidence and focal characters, then identifies information still needed for the next writing step.

Narrative Memory Planning (abridged)

Inputs: the writing request, current passage, admissible memory preview, complete character roster, and reference examples.

Task
Select established narrative memory needed for the next writing step. W contains shared-world facts, events, states, and relationships. B contains character-specific knowledge, beliefs, doubts, and misunderstandings. D connects supported changes across events, including progress, reversal, cause, consequence, and relationship development. H contains conditional possibilities, their premises, and possible consequences.

Decision rules
Preserve exact names, quantities, object states, time expressions, knowledge boundaries, and unresolved commitments. Select related records together when they explain one developing situation. Keep the underlying details available alongside summaries.
Form a query describing what is needed from memory. Add information needs only for material not already covered by the selected evidence. Distinguish facts, temporal and causal relations, motivation, character knowledge, state history, unresolved threads, developments, and possible worlds.
Use only characters and evidence supplied in the admissible context. Briefs and outlines express intentions, not evidence that planned events occurred. Do not write the next passage.

Illustrative response
Selected evidence: Mara's promise and the related world, belief, and development records. Focal characters: Mara and Leon. Remaining information need: what Leon knows about the available route at this point. This example assumes that the selected evidence leaves that question unanswered.

Narrative Event and Fact Modeling. Event modeling separates consequential narrative beats from the precise facts that describe them, preserving detail without treating every movement as an event.

Narrative Event and Fact Modeling (abridged)

Inputs: the current passage, source context, identified entities, unresolved references, supporting passages, and reference examples.

Task
Extract narrative events and independently readable propositions. An event changes world state, character knowledge or commitment, a relationship, an active conflict, or an unresolved development. A proposition records factual content with its subject, predicate, object, polarity, and truth status.

Decision rules
Read the passage as a whole. Merge consecutive actions that constitute one consequential beat. Ask what remains different after the event. Preserve distinguishing names, objects, quantities, attributes, dates, and durations in propositions.
Support every event and proposition with a source passage. Earlier summaries provide continuity context, not evidence for new records. A plan is not its execution, chronology is not causation, and a spoken claim does not establish its truth. Retain false or disputed claims when needed to represent a character's belief. Leave unresolved identities explicit.

Illustrative example
Passage: "Mara handed the key to Leon."
Event: Mara transfers the key to Leon; participants are Mara, Leon, and the key.
Resulting fact: Leon possesses the key.
Both records cite the passage. The transfer is one event, while the resulting possession remains separately recoverable.

Character Knowledge Modeling. Perspective modeling distinguishes world propositions from each character's access to and stance toward them. Each belief remains linked to its supporting evidence and acquisition event.

Character Knowledge Modeling (abridged)

Inputs: the current passage, source context, identified entities, grounded events and propositions, and reference examples.

Task
Identify the supported discourse perspective and material changes or differences in character access to facts. For each belief, retain the holder, proposition, attitude, acquisition mode, and acquisition event when established.

Decision rules
Ground access in seeing, hearing, reading, discovery, communication, remembering, inference, doubt, or explicit lack of information. Being named in an event does not imply that a character knows every fact associated with it.
A speaker's assertion does not prove that the speaker believes it, that a listener accepts it, or that the claim is true. Preserve speakers, recipients, and distinct belief holders. A question indicates inquiry or uncertainty, not necessarily disbelief. Emotion and intention alone do not establish knowledge.
Record unawareness only when the passage explicitly establishes a character's lack of information. Missing evidence is not evidence of unawareness.

Illustrative example
Passage: "The lamp was broken. Iona falsely told Tomas it worked; Tomas believed her."
World proposition: "The lamp works" is false.
Character belief: Tomas believes that proposition after hearing it from Iona.
The belief is linked to Iona's statement as its acquisition event. The passage does not establish that Iona believes her own claim.

Development and Possible-World Modeling. State and thread formation preserves established commitments and their changing status. Possible worlds are conditional extensions of unresolved developments; consolidation organizes accepted developments across scenes.

State, Thread, and Possible-World Modeling (abridged)

Inputs: the current passage, entities, grounded events and facts, character knowledge, relevant prior state, and reference examples.

Task
Extract supported relationship and entity-state changes and unresolved narrative threads. Associate conditional possibilities with newly established or materially changed unresolved threads.

Decision rules
Keep goals, intentions, commitments, and observed actions distinct. Compare current evidence with prior state to identify change. Keep the same obligation identifiable across updates. Mark it resolved only when its promised outcome is achieved, abandoned when explicitly relinquished, and advanced when progress leaves that outcome unsettled. A scene change or missing mention does not close a thread.
Construct possibilities from the unresolved condition at the end of the passage. Do not repeat completed steps. Each possibility retains a supported premise, possible events, continuity constraints, and uncertainty. Respect character knowledge and distinguish different mechanisms or choices from paraphrases. Resolved or abandoned threads have no possibilities; an unresolved thread may remain without a warranted candidate continuation.

Illustrative example
Prior commitment: Mara promises to get Leon out tonight.
Current passage: "Mara finally unlocked the gate. Leon was still inside the locked infirmary, and she had not yet reached him."
Development: the promise has advanced but remains unfulfilled.
Possible continuation: if Mara gains entry to the infirmary, she could bring Leon outside.
Constraints: Leon has not been rescued, and entry to the infirmary remains uncertain. The conditional continuation is not an accepted event.

Scene-to-Plotline Consolidation.

Scene-to-Plotline Consolidation (abridged)

Inputs: source-grounded scene records and reference examples.

Task
Determine whether the supplied scenes form a continuing plotline: a persistent goal, conflict, relationship, plan, or state trajectory. Use accepted scenes, excluding hypothetical possibilities.

Decision rules
Shared characters alone do not establish continuity. Connect scenes through supported actions, changes, temporal order, or explicit relations. Distinct descriptions of the same events are not successive scenes. Distinguish plans from execution and chronology from causation. Preserve supported resolution, revelation, and setup/payoff links without inventing them from genre expectations.
Summarize the starting situation, the change, and its result. Retain exact names, quantities, unresolved conditions, and uncertainty when relevant. Support each connection with at least two distinct scenes.

Expected response
A concise account of the development, the relation connecting its scenes, the supporting scene references, and any unresolved uncertainty.

B Writing Benchmark Adaptations

This appendix specifies how the three writing benchmarks are presented to memory systems. They test related but distinct capabilities. TMAS is a single fixed-prefix continuation problem: the system must recover useful context from a visible source prefix before producing one continuation. ConStory-Bench is a task-adaptive continual protocol: the system must preserve a story state across several accepted writing units while the requirements changes across four task types. LSGC extends continual generation to 14 chapters and at least 40,000 words per story. All conditions receive the same task inputs.

Across all three evaluations, reference continuations, future text, and evaluation annotations are withheld from memory construction and generation. Each story and condition maintains independent memory. The writer receives the task, available recent history, and retrieved memory as read-

Table 7: TMAS aggregate and auxiliary diagnostics. TMAS Quality is the arithmetic mean of the four reported dimensions; Prefix Continuity, word counts, and Unique are descriptive measures.

Method	TMAS Quality	Prefix Continuity ↑	Words Generated	Unique
NarraWorld	**83.01**	87.29	905.6	0.3372
Mem0	81.79	86.98	931.6	0.3582
Graphiti	82.04	85.83	914.7	0.3599
Hindsight	77.87	80.48	920.9	0.3573
AriGraph	76.78	77.88	901.9	0.3595
StructMem	81.80	86.83	936.8	0.3414
Scaffold Only	81.05	68.50	922.3	0.3532
Full Context	82.87	88.23	913.4	0.3621

able text. Full Context receives the complete available history under the same task and writing model, while memory systems retain their native construction and retrieval procedures and process source text or accepted writing units in order.

B.1 TMAS: SOURCE-FAITHFUL ROLLING CHECKPOINTS

The released source contains 168 checkpoints: 120 from the development split and 48 from the held-out split. We evaluate all 168 checkpoints using a paired rolling protocol. Each checkpoint contains an English writing instruction, the source story up to a designated boundary, and an evaluator-only human continuation. The task instruction is retained verbatim, including its requested setting, characters, constraints, style, form, and any example-specific whole-story length instruction. The visible source history is represented as ordered lines or paragraphs with source positions. The boundary is non-empty and is selected from the source story rather than from the human continuation.

Each checkpoint is evaluated as follows:

1. initialize independent memory for the checkpoint;
2. ingest the visible prefix in source order, without reading the evaluator-only continuation;
3. issue the memory request at the continuation boundary; and
4. give the writer the original instruction, the available recent prefix, and the memory system's rendered context.

The writer returns only the new prose continuation and must not repeat the visible prefix. The human continuation is reserved for evaluation. We measure generated word count after completion, with length requirements inherited from each original instruction rather than imposed as a universal target. TMAS therefore evaluates one continuation from a fixed source history.

Each memory system constructs and retrieves context from the visible prefix; the Full Context control receives the same task and full visible source prefix directly, without memory construction or retrieval. The reported primary TMAS dimensions are Plot, Creativity, Development, and Language Use. Table 7 also reports the arithmetic-mean aggregate, TMAS Quality, together with Prefix Continuity, generated word counts, and Unique.

B.2 CONSTORY-BENCH: TASK-ADAPTIVE CONTINUAL WRITING

The evaluation dataset contains 200 English stories, with 50 stories for each of four task types: generation, continuation, expansion, and completion. Task construction uses only the original prompts, excluding reference stories and evaluation annotations. DeepSeek-V4-Pro restructures each prompt into a task-specific global context. Writing schedules, local objectives, and word targets are fixed by task type and shared across conditions.

The four task types use distinct writing schedules:

- **Generation.** The writer starts from a premise and a set of fixed narrative requirements, together with style and form requirements. The story is developed over 8 units, targeting approximately 1,100 words per unit and 8,800 words in total.
- **Continuation.** The writer receives a supplied opening fragment, an explicit continuation boundary, and requirements inherited from the original prompt. New prose begins after the opening and must not repeat it. This task uses 5 units of approximately 1,760 words, again targeting 8,800 words in total.
- **Expansion.** The writer receives an ordered compact outline whose beats define the material to be expanded, together with expansion, style, and form requirements. The outline is a planning input, not a reference story. The task uses 6 units of approximately 1,467 words.
- **Completion.** The writer receives a beginning anchor and an ending anchor and must construct the missing bridge while satisfying the bridge requirements. The two anchors constrain the endpoint of the story, so this is not equivalent to ordinary forward continuation. It uses 6 units of approximately 1,467 words.

The writer receives task-specific instructions: a premise, an opening and continuation boundary, an ordered outline, or beginning and ending anchors. Across all four types, the dataset contains 1,250 writing units. The first unit of each story establishes the initial writing state, so 1,050 later units can retrieve material from previously accepted writing.

For each story, the continual protocol is:

1. initialize independent story memory and expose the task's global context and first unit's requirements;
2. generate and accept the next narrative unit;
3. ingest the accepted unit into the memory system before the next unit; and
4. retrieve a new memory view for the next unit using the global task requirements, the current unit goal, and the available generated history.

The current unit's instructions state its local goal and any required material; the global context preserves the original requirements and the task-specific anchors or outline. The output is narrative prose with the unit-level word target. The final story is the ordered concatenation of the accepted units. Future units, evaluator targets, and other conditions are not visible when an earlier unit is generated.

This protocol makes ConStory a test of memory under changing writing states. The memory system must retain characters, events, world facts, private knowledge, and unresolved developments that remain useful later, while the writer must also respect the current unit's requirements. The four task types differ in the initial state and in the constraints that must remain available; they are therefore reported separately as well as in aggregate.

The official ConStory evaluator is applied after all units are assembled. We report its Consistency Error Density (CED, lower is better) and the official category-level consistency breakdown.

B.3 LSGC: CHAPTER-CONTINUAL LONG-STORY GENERATION

The LSGC memory track uses all five released official starters. Each system develops every starter into a 14-chapter story, with two sequential generation segments per chapter, and repeats the complete protocol for three runs. The target is 42,000 words per story, and a completed story must contain at least 40,000 words. All starters and runs use the same chapter-by-chapter writing protocol.

Before generation, the writer receives the starter and a reviewed coarse-grained outline shared by all memory conditions. For chapter t, the legal input consists of the starter, the same outline, the current chapter objective, the complete text of chapter $t-1$, and the memory system's rendered view of chapters earlier than t. The memory ceiling is 12,000 rendered tokens and does not include the shared starter, outline, chapter objective, or complete preceding chapter. Memory is retrieved once at the start of a chapter and reused for its two segments. After both segments are accepted, the new chapter is incorporated into memory before the next chapter.

Table 8: LSGC final-rubric scores by criterion. Values are direction-corrected macro means over five starters and three runs on a 0–100 scale (↑); negative criteria are inverted.

Criterion	Mem0	AriGraph	Hindsight	Graphiti	StructMem	Narra World	Scaffold Only	Full Context
Believable Character Actions	89.07	78.13	80.84	88.69	90.02	**91.04**	70.96	86.80
Nuanced Characters	87.62	72.27	75.78	88.51	88.22	91.22	69.11	**91.53**
Pacing	70.93	71.02	72.98	81.29	85.20	**87.71**	69.09	81.76
World Building	86.00	74.80	79.06	88.29	87.29	87.49	66.56	**92.62**
Compelling Plot	80.22	75.62	79.91	**88.76**	85.60	87.25	67.18	88.42
Emotionally Engaging	87.58	80.53	82.38	89.00	89.84	**94.60**	71.02	91.09
Coherent	87.56	78.22	81.49	88.93	91.11	**92.02**	66.18	91.13
Weak Dialogue	74.73	72.00	72.64	73.18	85.60	**91.00**	72.20	77.56
Tell-Don't-Show	65.16	69.24	66.78	67.58	80.09	**86.87**	67.80	74.51
Unsurprising or Uncreative	77.33	71.60	73.07	60.82	78.18	**84.60**	65.76	77.53
Amateurish	77.60	76.02	76.29	71.89	87.98	**93.16**	74.80	87.16
Purple Prose	74.00	55.98	56.33	63.44	85.16	**89.67**	71.73	78.44
Forced Poetry or Metaphor	73.29	60.53	61.27	67.13	84.62	**88.56**	72.87	77.56
Unearned Transformations	76.33	67.27	69.33	70.11	88.38	**91.09**	59.42	83.18
Well-earned Lightness or Darkness	88.09	80.31	82.58	89.40	90.31	**92.42**	69.60	87.82
Faithful to Writing Prompt	87.40	85.00	87.87	88.51	93.07	**95.09**	79.53	93.42

Table 9: LSGC Overall scores by starter, averaged over three runs. Each value equally weights the 16 direction-corrected final criteria.

Starter	Mem0	AriGraph	Hindsight	Graphiti	StructMem	NarraWorld	Scaffold Only	Full Context
01	79.33	73.32	74.17	85.78	84.83	**91.95**	74.55	87.26
02	65.67	73.07	75.60	85.36	**89.35**	86.78	61.74	86.35
03	85.47	71.56	73.36	54.53	86.61	**91.05**	65.81	88.03
04	82.93	74.26	77.20	84.69	87.52	**90.38**	84.13	87.48
05	87.51	72.97	74.24	85.11	86.27	**91.01**	61.85	76.04
Macro mean	80.18	73.03	74.91	79.10	86.92	**90.24**	69.61	85.03

We retain the Longform Writing Bench aggregation structure while adapting the final-piece context to the much longer LSGC stories. A condition-blind Judge scores the 16 criteria for each chapter. After the fourteenth chapter is accepted, the system produces a final rolling summary over the complete story, and the Judge scores the final piece from the writing prompt, starter, and this summary rather than the full 40,000-word story. We then average the 14 chapter judgments and the final-piece judgment within each story instance. Raw criteria retain their original 0–20 scales and directions: positive criteria are multiplied by five, while negative criteria are mapped to $(20 - s) \times 5$.

For compact reporting, Character averages Believable Character Actions, Nuanced Characters, and Unearned Transformations. Plot averages Pacing, Compelling Plot, Emotionally Engaging, and Well-earned Lightness or Darkness. World/Coherence averages World Building and Coherent. Style/Creativity averages Weak Dialogue, Tell-Don't-Show, Unsurprising or Uncreative, Amateurish, Purple Prose, and Forced Poetry or Metaphor. Prompt Adherence contains Faithful to Writing Prompt. Overall is the equal-weight mean of all 16 criteria, not an equal-weight mean of the five groups. The completed story, after this within-story aggregation, is the primary statistical unit; chapter-level judgments are not reported separately or treated as independent samples.

Let $y_{m,r}$ denote the direction-corrected Overall score for method m in run r, after averaging that run over the five starters and the 16 criteria. With $R = 3$ runs, the reported mean, range, and sample standard deviation are respectively

$$\bar{y}_m = \frac{1}{R} \sum_{r=1}^{R} y_{m,r}, \qquad \text{Range}_m = \max_r y_{m,r} - \min_r y_{m,r}, \tag{10}$$

Table 10: Run-to-run variability of LSGC Overall scores. Each run is first aggregated over five starters and the 16 criteria; range and SD are computed over the three independent runs. SD is the sample standard deviation.

Method	Overall mean ↑	Run range	Run SD
NarraWorld	**90.24**	2.59	1.35
Mem0	80.18	5.38	2.69
Graphiti	79.10	3.03	1.51
Hindsight	74.91	4.06	2.10
AriGraph	73.03	3.57	1.90
StructMem	86.92	3.34	1.69
Scaffold Only	69.61	4.55	2.30
Full Context	85.03	2.13	1.10

$$s_m = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(y_{m,r} - \bar{y}_m)^2}. \tag{11}$$

Range and SD describe variation among the three complete runs after within-run aggregation; they are not confidence intervals.

C Evaluation Protocols, Metrics, and Judge Prompts

This appendix specifies the metric sources, scoring procedures, and task-specific adaptations. Judges evaluate anonymized outputs against the task evidence; reference answers and evaluation annotations are reserved for scoring.

C.1 ConStory-Bench: consistency error density

We use the official ConStory-Bench taxonomy and checker (Li et al., 2026). The five categories cover 19 error subtypes, and CED counts whether each subtype has at least one supported error in the assembled story. For story i, let L_i be its whitespace-delimited word count and z_{ic} indicate the presence of subtype c:

$$E_i = \sum_{c=1}^{19} z_{ic}, \qquad \mathrm{CED}_i = 10{,}000\frac{E_i}{L_i}, \qquad \mathrm{CED} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CED}_i. \tag{12}$$

Multiple instances of one subtype therefore contribute one unit. Task-specific columns use the same story-level mean within each task type.

Words per Penalized Error (WPE) is an auxiliary measure introduced for this paper, adapted from the benchmark's length-normalized intermediate score:

$$\mathrm{WPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{L_i}{1+E_i}. \tag{13}$$

Both measures depend on story length and are reported as consistency diagnostics rather than literary-quality scores.

C.2 Tell Me A Story: adapted continuation quality

The original Tell Me A Story evaluation uses pairwise comparisons and a Bradley–Terry aggregation. We use its Plot, Creativity, Development, and Language Use dimensions for the fixed-prefix continuation task. Each Judge request contains the original writing prompt, the shared prefix, the candidate continuation, and the evaluator-only human continuation. The reference calibrates execution quality; it is not the only acceptable plot or ending.

For checkpoint i, dimension d has K_d integer subscores $a_{idk} \in \{0, 1, 2, 3, 4\}$. Plot, Creativity, and Development each have four subcriteria; Language Use has five. We normalize each dimension and

define our arithmetic-mean aggregate as

$$S_{id} = \frac{100}{4K_d} \sum_{k=1}^{K_d} a_{idk}, \qquad Q_i = \frac{1}{4} \sum_{d \in \{\text{Plot,Creativity,Development,Language}\}} S_{id}. \tag{14}$$

Reported dimension scores and TMaS Quality are arithmetic means over completed checkpoints. The four dimensions receive equal weight; Language's fifth subcriterion does not give Language extra weight in Q_i.

Prefix Continuity is an auxiliary measure introduced for our continuation task, not a source-benchmark metric. Its four subscores measure event, character/motivation, perspective/voice, and factual continuity with the supplied prefix. It does not enter Q_i. Generated Words counts only new continuation text, and Unique is the distinct-word ratio; both are descriptive diagnostics.

Tell Me A Story Judge: scoring instructions

Evaluate one continuation in the context of the original writing prompt and shared prefix. Use the human continuation only as a fixed quality and task-interpretation anchor. Do not select a winner or compare system identities. Score every named subcriterion with one integer from 0 to 4, where 0 indicates failure, 2 competent execution with a clear limitation, and 4 exceptional execution without a meaningful weakness. Support scores with brief grounded evidence and also evaluate event, character, perspective, and factual continuity with the prefix.
Inputs: the original writing prompt, shared story prefix, candidate continuation, and evaluator-only human continuation.

Table 11: TMAS evaluation subcriteria, each scored from 0 to 4.

Dimension	Subcriteria
Plot	Connected structure; plot progression; logical coherence; intentional turns.
Creativity	Engaging specificity; non-generic ideas; purposeful use of tropes; original contribution.
Development	Character context; setting context; believable detail; motivated development.
Language Use	Sentence variety; lexical richness; literary effect; repetition control; appropriateness of voice.
Prefix Continuity	Continuity of events, characters, perspective and voice, and facts.

C.3 LSGC: FULL-STORY QUALITY UNDER THE LONGFORM WRITING BENCH RUBRIC

LSGC specifies a long-form generation protocol but does not prescribe a native quality metric (Migal et al., 2024). We therefore evaluate its completed stories with the separate 16-criterion rubric from Longform Writing Bench. We retain the criterion names and raw 0–20 scores, and report direction-corrected 0–100 scores in five groups alongside an equal-weight Overall.

For story instance i, let $b_{ijc} \in [0, 20]$ be the raw score for criterion c at chapter or final-piece position j, where $j \in \{1, \ldots, 14, \text{final}\}$. For the seven negative criteria—Weak Dialogue, Tell-Don't-Show, Unsurprising or Uncreative, Amateurish, Purple Prose, Forced Poetry or Metaphor, and Unearned Transformations—lower raw values indicate fewer problems. We first apply the direction correction to every chapter and final-piece score and then average the resulting 15 values within each story instance:

$$T_{ijc} = \begin{cases} 5(20 - b_{ijc}), & c \text{ is negative}, \\ 5b_{ijc}, & c \text{ is positive}, \end{cases} \qquad \bar{T}_{ic} = \frac{1}{15} \sum_{j=1}^{15} T_{ijc}, \qquad O_i = \frac{1}{16} \sum_{c=1}^{16} \bar{T}_{ic}. \tag{15}$$

Each reported column is then a macro mean over five starters and three runs. A grouped column averages its assigned criteria within each story instance before taking the same macro mean. The reported unit is therefore the completed story instance, not an individual chapter. Character, Plot, World/Coherence, Style/Creativity, and Prompt Adherence use the group membership specified in Appendix B.3; Overall averages all 16 criteria rather than the five group scores.

For run-level variability, let u_{mrc} be the mean score for method m, run r, and criterion c after averaging the five starters, and let $o_{mr} = 16^{-1} \sum_{c=1}^{16} u_{mrc}$ be that run's Overall score. Across the three runs, the reported mean, range, and sample standard deviation are

$$\mu_m = \frac{1}{3} \sum_{r=1}^{3} o_{mr}, \qquad R_m = \max_r o_{mr} - \min_r o_{mr}, \qquad s_m = \sqrt{\frac{1}{2} \sum_{r=1}^{3} (o_{mr} - \mu_m)^2}. \tag{16}$$

The criterion-level range and standard deviation use the same formulas with o_{mr} replaced by u_{mrc}. Thus the reported Overall variability is computed after averaging criteria, whereas criterion-level variability retains variation specific to each rubric item.

LSGC Judge (core instructions)

Evaluate the current chapter or the final story-level view against the supplied writing prompt and all 16 criteria. For the final-piece judgment, use the official starter and the final rolling summary produced after Chapter 14. Score every criterion from 0 to 20 and briefly justify the scores with concrete strengths and weaknesses. For the seven negative criteria listed above, lower scores indicate fewer problems.
Inputs: the chapter text for chapter judgments, or the official starter and final rolling summary for the final-piece judgment, together with the 16 Longform Writing Bench criteria.

C.4 RMTBench: turn-level role-play scores

We follow RMTBench's native mixed-scale evaluation (Xiang et al., 2025). EE, EC, PA, and CU use integer scores from 1 to 5; CM and UPA use binary success values. The four Likert dimensions apply to all scenarios, CM applies to the character-maintenance scenario, and UPA applies to the final preference-recognition turns.

For metric d, let $\mathcal{A}_d$ be its applicable turns and r_{td} its value. We use the turn-level mean

$$\bar{R}_d = \frac{1}{|\mathcal{A}_d|} \sum_{t \in \mathcal{A}_d} r_{td}. \tag{17}$$

Likert dimensions remain on their 1–5 scale, while binary dimensions are reported as $100\bar{R}_d\%$. Inapplicable dimensions are excluded from their own denominator, and no cross-metric average is reported.

RMTBench Role-Play Judge (core instructions)

Evaluate only the dimensions supplied as applicable. Use the original 1–5 rubrics for EE, EC, PA, and CU. Score CM and UPA as binary success or failure according to their applicable RMTBench criteria, and report null for every inapplicable dimension.

C.5 LoCoMo: category-aware deterministic scoring

We use the released category-aware LoCoMo answer scorer (Maharana et al., 2024). The main comparison evaluates the C1–C4 subset: 282 multi-hop, 321 temporal, 96 open-domain, and 841 single-hop questions, for 1,540 questions in total. The 446 adversarial C5 questions are excluded. Each category score follows the official scorer, while Overall Main is the question-weighted mean over these 1,540 main questions.

Comparison controls and length measures. Comparisons use the same task set and evaluation protocol across conditions. Scores include only completed generations with valid assessments. Each condition develops its own history throughout generation. ConStory and LSGC word counts use whitespace-delimited English words. TMaS counts only newly generated continuation words. Memory length is measured in tokens and excludes the writing instruction, recent legal text supplied outside memory, and generated output. Retrieved memory may occupy less than the specified ceiling.

Table 12: LoCoMo budget sensitivity for NarraWorld, StructMem, and Mem0. All scores are higher-is-better.

Method	Budget	Overall Main	Multi-hop	Single-hop	Temporal	Open-domain
Mem0	2k	0.4736	0.3718	0.5241	0.4919	0.2691
	4k	0.4982	0.4047	0.5392	0.5372	**0.2834**
	8k	**0.5010**	**0.4100**	**0.5410**	**0.5470**	0.2700
	12k	0.4683	0.3984	0.5125	0.4758	0.2620
	16k	0.4742	0.4047	0.5125	0.4963	0.2682
StructMem	2k	0.4962	0.3565	0.5275	0.5827	0.3435
	4k	0.5324	0.4109	0.5659	**0.6020**	0.3631
	8k	0.5400	0.4350	0.5750	0.5950	0.3620
	12k	0.5365	0.4311	0.5725	0.5897	0.3526
	16k	**0.5442**	**0.4352**	**0.5836**	0.5892	**0.3680**
NarraWorld	2k	0.5605	0.4225	0.5985	0.6405	0.3655
	4k	0.5878	0.4386	0.6310	0.6567	0.4173
	8k	0.6144	0.4580	0.6790	0.6490	0.3920
	12k	0.6211	0.4499	0.6853	0.6515	**0.4594**
	16k	**0.6238**	**0.4712**	**0.6897**	**0.6577**	0.3810

D Retrieval Planning Ablations

We isolate whether retrieval planning benefits from access to a memory preview on both LoCoMo and ConStory-Bench. Figure 5 places no-preview controls within the query-centric paradigm and contrasts them with NarraWorld's preview-conditioned reconstruction. The generic diagram depicts the full NarraWorld memory interface; the benchmark-specific experiments below hold their respective offline memories, context budgets, and downstream models fixed.

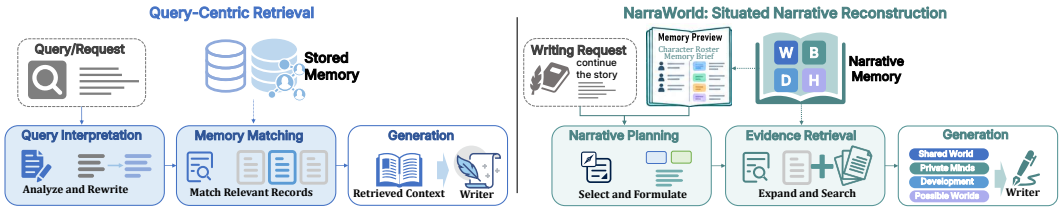


Figure 5: Conceptual comparison of retrieval paradigms. Query-centric methods interpret the request and match the resulting query against stored memory; both deterministic parsing and question-only LLM planning instantiate this left-hand path. NarraWorld instead conditions narrative planning on an admissible memory preview before expanding and searching for supporting evidence. The diagram summarizes the two information flows.

Deterministic Retrieval. Rules parse surface entities, temporal expressions, relation terms, and question forms. Temporal questions activate time and event channels, for example, while causal questions additionally activate causal and development access. This setting requires no model call for planning and is fully deterministic, but it cannot reinterpret implicit intent beyond the encoded rules. Because these rules map directly to LoCoMo question forms, we use this setting only as a LoCoMo control rather than as a general writing baseline.

Retrieval w/o Dependency Planning. The planner receives the question but no memory content. It extracts target entities, assigns $W/B/D$ channels, rewrites the query, and specifies temporal, perspective, or multi-hop constraints. This setting tests semantic question interpretation without dependency planning grounded in records that actually exist in memory.

Preview-Guided Reconstruction (NarraWorld). Before formal retrieval, the planner receives a lightweight admissible preview of the $W/B/D$ directory. It may select records that already address

the question or use visible events and entities as anchors for retrieving missing links. Preview-Guided Reconstruction therefore grounds dependency planning in the available memory structure before downstream answering.

D.1 LoCoMo: Question-Centric Retrieval

All three LoCoMo settings use the same offline $W/B/D$ memory, 8k final memory budget, and answer-generation model. They differ only in how the question is converted into retrieval operations and whether the planner can inspect an admissible summary of stored memory before formal retrieval.

Table 13: LoCoMo retrieval ablation with shared offline $W/B/D$ memory, an 8k final memory budget, and the same answer model. All metrics are higher-is-better.

Retrieval setting	Overall Main	Multi-hop	Single-hop	Temporal	Open-domain
Deterministic Retrieval	0.5963	0.4479	0.6467	**0.6513**	0.4069
Retrieval w/o Dependency Planning	0.5982	0.4487	0.6488	0.6449	**0.4383**
Preview-Guided Reconstruction (NarraWorld)	**0.6144**	**0.4580**	**0.6790**	0.6490	0.3920

Table 13 shows that Preview-Guided Reconstruction achieved the highest Overall Main score (0.6144), improving over Deterministic Retrieval by 0.0181 and Retrieval w/o Dependency Planning by 0.0162. It also led multi-hop and single-hop accuracy. Relative to the LLM query-centric control, the gains were 0.0093 on multi-hop and 0.0302 on single-hop, consistent with the benefit of anchoring retrieval decisions in the available memory structure rather than deriving them from the question alone.

Replacing Deterministic Retrieval with Retrieval w/o Dependency Planning changed Overall Main by only 0.0019 under the matched protocol. The two query-centric variants nevertheless showed different category strengths: deterministic rules obtained the highest temporal score (0.6513), whereas LLM planning obtained the highest open-domain score (0.4383). Preview-Guided Reconstruction combined the strongest aggregate result with the best multi-hop and single-hop scores.

D.2 ConStory-Bench: Writing-Conditioned Retrieval

Writing does not provide the lexical question forms required by deterministic retrieval: a continuation request rarely states which prior facts or developments should be recovered. We therefore compare retrieval without dependency planning against the full NarraWorld retrieval pipeline. Both receive the same global plan, rolling summary, recent text, and current writing objective. The first setting retrieves directly from this scaffold, whereas Preview-Guided Reconstruction uses the admissible $W/B/D/H$ preview to plan dependencies before retrieval. The offline memory, retrieved-memory budget, writing model, and generation protocol are otherwise held fixed.

Table 14: ConStory-Bench retrieval ablation under matched writing scaffolds, offline memory, memory budgets, and generation settings. All columns report CED (↓).

Retrieval setting	CED ↓	Generation	Continuation	Expansion	Completion
Retrieval w/o Dependency Planning	0.315	0.210	0.380	0.185	0.485
Preview-Guided Reconstruction (NarraWorld)	**0.185**	**0.120**	**0.235**	**0.098**	**0.287**

Preview-Guided Reconstruction reduced overall CED by 0.130 and improved every task type. The largest absolute reduction occurred on Completion (0.198), followed by Continuation (0.145), where selecting earlier constraints and developments is especially important. These results isolate dependency planning over the memory preview rather than the contribution of the writing scaffold itself, which is shared by both conditions.

E MEMORY ABLATIONS ACROSS TASKS

The ablation study uses the same four configurations on ConStory-Bench and TMAS. Full NarraWorld is compared with removal of character beliefs, possible worlds, or cross-event organization. The belief ablation excludes character-specific epistemic assertions from both planner previews and final memory, including assertions carried by derived summaries. The possible-world ablation disables H formation and access while retaining unresolved accepted developments in D. The cross-event ablation disables derived scene, plotline, and plot access while preserving atomic events, world facts, beliefs, and development records. Models, writing instructions, decoding settings, memory caps, and evaluation protocols are held fixed across conditions. Contexts are not padded to fill the cap. TMAS uses the same fixed source histories with the corresponding memory views excluded. ConStory conditions generate independent complete trajectories and incorporate their own accepted units; later memory states from Full NarraWorld are not reused for another condition. Comparisons are paired by story or checkpoint.

ConStory-Bench consistency and task coverage. Figure 3 reports Overall CED and the four task-specific scores for each configuration.

E.1 TMAS: CONTINUATION QUALITY

The matched continuation evaluation complements consistency measurement: possible-world memory may affect creativity and development even when it does not change consistency error density. Table 15 reports the four quality dimensions, their equal-weight TMAS Quality aggregate, and the auxiliary Prefix Continuity diagnostic.

Table 15: TMAS ablations. All quality dimensions are higher-is-better.

Setting	Plot	Creativity	Development	Language	TMAS Quality	Prefix Continuity
Full NarraWorld	**78.92**	**84.90**	**85.60**	**82.62**	**83.01**	**87.29**
Without character beliefs (B)	76.54	82.15	81.33	81.58	80.40	83.21
Without possible worlds (H)	77.20	78.85	80.15	82.10	79.57	85.65
Without cross-event organization	74.35	81.12	78.90	80.45	78.70	79.42

E.2 LSGC: LONG-FORM NARRATIVE QUALITY

The LSGC ablation uses the same five-starter protocol and model configuration as the main evaluation. Each condition removes one NarraWorld component while keeping the writing scaffold, memory cap, decoding settings, and accepted-history update process fixed. The table reports the complete 16-criterion breakdown; Overall is the equal-weight mean across criteria.

The ablation pattern is consistent across the aggregate and the detailed criteria. Removing cross-event organization produces the largest Overall drop, from 90.24 to 81.05 (a decrease of 9.19 points), with especially large losses in pacing, world building, and coherence. Removing character beliefs reduces Overall to 86.04 (a decrease of 4.20 points) and is most visible in the character-related criteria, while removing possible worlds has a smaller aggregate effect (88.74) but changes several creativity and world-modeling criteria. The full system therefore benefits from the complementary views: cross-event structure supports long-range organization, character beliefs preserve perspective, and possible worlds retain unresolved alternatives. The run-level statistics also reveal a stability trade-off: removing H produces the smallest range and SD (1.53 and 0.80), whereas the full system is more variable (2.59 and 1.35); removing cross-event organization is both the weakest and the most variable ablation (4.02 and 2.09).

E.3 LOCOMO: QA-TIME MEMORY-VIEW ACCESS

LoCoMo uses a task-specific memory view because hypothetical-world records are not necessarily admissible evidence for a factual answer. The default QA view therefore excludes H, while the comparison view exposes it. The remaining rows remove the same components used in the writing

Table 16: LSGC ablations by criterion. Scores are direction-corrected macro means over 15 samples (5 starters $\times$ 3 runs) on a 0–100 scale (↑).

Criterion	Full NarraWorld	Without B	Without H	Without cross-event
Believable Character Actions	91.04	83.33	**92.02**	82.51
Nuanced Characters	**91.22**	81.93	90.09	81.16
Pacing	**87.71**	84.22	85.36	72.44
World Building	87.49	85.96	88.44	74.84
Compelling Plot	**87.25**	82.49	84.33	76.82
Emotionally Engaging	**94.60**	87.36	88.27	82.49
Coherent	92.02	88.36	**92.11**	75.05
Weak Dialogue	**91.00**	80.07	90.20	83.22
Tell-Don't-Show	**86.87**	85.49	86.07	79.56
Unsurprising or Uncreative	**84.60**	83.33	79.38	77.69
Amateurish	**93.16**	89.64	91.47	85.44
Purple Prose	89.67	87.80	**90.69**	84.64
Forced Poetry or Metaphor	88.56	88.18	**90.29**	87.89
Unearned Transformations	**91.09**	83.49	89.56	78.84
Well-earned Lightness or Darkness	**92.42**	88.71	86.07	82.22
Faithful to Writing Prompt	95.09	**96.33**	95.49	92.02
Overall	**90.24**	86.04	88.74	81.05

Table 17: Run-to-run variability for LSGC ablations. Overall means are direction-corrected scores; range and sample SD are computed over the three independent runs after aggregating each run over five starters and 16 criteria.

Setting	Overall mean ↑	Run range	Run SD
Full NarraWorld	**90.24**	2.59	1.35
Without B	86.04	3.26	1.68
Without H	88.74	**1.53**	**0.80**
Without cross-event	81.05	4.02	2.09

ablations. All conditions use the shared 8k memory budget and the same category-aware scorer; the table reports the complete subset breakdown.

The QA-view comparison isolates access to H. Excluding H improves Overall Main by 4.4 points and improves every reported subset. This task dependence complements the writing ablations, where H remains available to preserve unresolved alternatives during continuation.

F Human Evaluation and Judge Alignment

F.1 Evaluation Protocol

We paired human ratings with automatic scores on the same anonymized story sample. For ConStory-Bench, we randomly sampled 12 prompts across the four task types with equal allocation, comparing NarraWorld, Graphiti, and Full Context on 36 stories. For LSGC, we evaluated NarraWorld, StructMem, and Full Context on all five official starters, yielding 15 complete 14-chapter stories. This human-evaluation sample is separate from the three-run aggregate used for the main LSGC results.

Four native-English-speaking assessors with backgrounds in narrative analysis evaluated anonymized texts without access to system identities or automatic scores. Two assessors independently read each complete story. For ConStory, they labeled the presence of each of the 19 error subtypes and supplied textual evidence. A coordinator, also blind to system identities and automatic results, adjudicated disagreements. For LSGC, readers evaluated complete stories on five 0–100 dimensions: Character, Plot, World/Coherence, Style/Creativity, and Prompt Adherence. The human rubric groups the 16 automatic criteria into these five dimensions (Section F.4). Human Overall

Table 18: LoCoMo memory-view and component ablations at the shared 8k budget. The default QA view excludes hypothetical-world records (H); the exposed-H condition includes them. Overall Main aggregates the four reported question subsets (↑).

Setting	Overall Main	Multi-hop	Single-hop	Temporal	Open-domain
NarraWorld (QA view: H excluded)	**0.614**	**0.458**	**0.679**	**0.649**	**0.392**
NarraWorld (QA view: H included)	0.570	0.398	0.655	0.559	0.366
Without character beliefs (B)	0.586	0.415	0.650	0.635	0.370
Without cross-event organization	0.537	0.388	0.618	0.512	0.341

preserves the relative weighting of the original 16-criterion rubric: the five dimension scores receive weights of 3, 4, 2, 6, and 1, respectively, and the weighted sum is divided by 16.

Across the 684 ConStory story–subtype pairs, both assessors marked seven positive and 666 negative; five were positive only for assessor A and six only for assessor B. Pooled agreement was 98.4%, with Cohen's $\kappa = 0.55$. Positive agreement was 56.0% ($2 \times 7/(2 \times 7 + 5 + 6)$), distinguishing agreement on errors from the many shared negative labels. Adjudication yielded five positive labels.

F.2 ConStory-Bench: Human Consistency Assessment

Table 19 reports human and automatic CED on the same 12 prompts. Human assessment identified one positive error-subtype label for each of NarraWorld and Full Context, compared with three for Graphiti. Human CED was 0.095, 0.289, and 0.096 for NarraWorld, Graphiti, and Full Context, respectively. NarraWorld and Full Context therefore showed similar observed consistency, with fewer confirmed error subtypes than Graphiti. The automatic judge flagged more errors than human readers.

Table 19: ConStory-Bench human evaluation and judge alignment on 12 prompts (12 stories per method). Δ is Auto CED minus Human CED. Detection metrics use adjudicated error-subtype presence as the reference.

Method	Human CED ↓	Auto CED ↓	Δ	TP	FP	FN	Precision	Recall	F1
NarraWorld	0.095	0.185	+0.090	1	1	0	50.0%	100.0%	66.7%
Graphiti	0.289	0.385	+0.096	2	2	1	50.0%	66.7%	57.1%
Full Context	0.096	0.200	+0.104	1	1	0	50.0%	100.0%	66.7%

Across the 36 stories, the automatic judge produced four false positives and one false negative relative to the adjudicated human labels.

F.3 LSGC: Human Quality and Score Alignment

Human readers rated NarraWorld highest in long-form quality (Table 20), with an Overall score of 84.50, followed by StructMem at 80.50 and Full Context at 80.06. NarraWorld led Character, Plot, and Style/Creativity, whereas Full Context led World/Coherence and Prompt Adherence across the five official starters.

Table 20: LSGC human scores across all five starters on a 0–100 scale (↑). Auto Overall is shown alongside the human ratings for comparison.

Method	Character	Plot	World/Coh.	Style/Creat.	Prompt	Human Overall	Auto Overall
NarraWorld	**86.00**	**84.00**	86.00	**82.00**	94.00	**84.50**	**90.24**
StructMem	84.00	81.00	82.00	76.00	92.00	80.50	86.92
Full Context	83.00	82.00	**88.00**	72.00	**96.00**	80.06	85.03

Table 21 summarizes the alignment between the automatic judge's aggregated 5-dimension scores and the humans' direct 5-dimension assessments for the 15 complete stories. Overall correlation

was moderate ($\rho = 0.55$), with an average automatic score 5.71 points higher than the human score. Correlation was lowest for Style/Creativity ($\rho = 0.38$).

Table 21: LSGC human–judge alignment across 15 stories. Bias is the mean automatic minus human score. Both MAE columns use the 0–100 scale.

Dimension	Spearman ρ	MAE	Bias	Human–Human MAE
Character	0.52	5.40	+4.89	6.20
Plot	0.48	6.80	+6.42	7.50
World/Coherence	0.61	5.10	+4.67	5.50
Style/Creativity	0.38	7.60	+7.17	9.00
Prompt	0.68	3.20	+2.33	4.20
Overall	0.55	6.20	+5.71	6.48

Human readers also differed on Plot and Style/Creativity, with human–human MAE of 7.50 and 9.00. These differences reflect the subjective component of literary assessment and provide context for the moderate judge alignment.

F.4 Human Evaluation Rubric for Long-Form Narrative

The human rubric groups the 16 Longform Writing Bench criteria (Paech, 2025) into five dimensions. Assessors read each complete story and assigned a score from 0 to 100 for each dimension using the following guidelines.

Human Evaluator Guidelines: 5-Dimension Rubric

1. Character (0-100): Evaluates the depth, consistency, and psychological realism of the cast.

- *High score criteria:* Characters have nuanced, distinct voices and multi-dimensional personalities. Their actions, decisions, and emotional responses are believable and strictly grounded in their established traits and past experiences.
- *Penalties:* Cardboard or interchangeable characters; actions that break character logic purely to advance the plot; sudden, unearned psychological transformations or skill acquisitions that lack proper narrative buildup.

2. Plot (0-100): Evaluates the narrative arc, emotional engagement, and structural pacing.

- *High score criteria:* The story presents a compelling, emotionally engaging trajectory with well-earned stakes. The pacing feels natural—slowing down for character moments and accelerating for tension. Tonal shifts (lightness or darkness) feel earned and impactful.
- *Penalties:* Meandering, rushed, or sluggish pacing; disconnected episodes lacking an overarching drive; emotional beats that feel manipulative, unearned, or hollow.

3. World/Coherence (0-100): Evaluates the solidity, logic, and internal consistency of the storyworld.

- *High score criteria:* The setting is vividly established and operates under clear, consistent rules. The narrative flows logically from scene to scene, tracking physical objects, locations, and time accurately across the entire text.
- *Penalties:* Contradictory facts (e.g., dead characters reappearing, locations shifting inexplicably); thin, generic world-building that lacks sensory detail or internal logic; confusing spatial or temporal jumps.

4. Style/Creativity (0-100): Evaluates prose quality, narrative voice, and originality.

- *High score criteria:* The writing is engaging, professional, and employs "show, don't tell" effectively. Dialogue is sharp and subtextual. The narrative offers fresh, unsurprising twists or creative interpretations of tropes.
- *Penalties:* "Amateurish" prose (stilted phrasing, repetitive sentence structures); "purple prose" or overly forced, melodramatic metaphors; heavy exposition dumps instead of organic action; generic, highly predictable, or cliché-ridden writing.

5. Prompt Adherence (0-100): Evaluates fidelity to the original writing premise and constraints.

- *High score criteria:* The story fully embraces and explores the core concept, characters, and specific constraints dictated by the original prompt or starter.

- *Penalties:* Ignoring key constraints; drifting entirely away from the requested genre or premise; failing to utilize the provided opening effectively.

G LIMITATIONS

Modeling and construction cost. NarraWorld's richer structured modeling is more expensive than a compact passage store, but its construction overhead remains comparable to other structured or graph-based memory systems: baseline workloads span roughly 20 to 800 successful modeling calls per story, while NarraWorld averages about 200. This reflects a quality–cost trade-off: the current system prioritizes explicit narrative structure, while reducing extraction and planning overhead is an important direction for future work.

The contribution of possible worlds is conditional. The H view is designed for long-form writing, where retaining unresolved alternatives can support continuation quality, but hypothetical content is not equally useful in general-purpose tasks such as factual QA. NarraWorld therefore uses task-conditioned access to H; learning a more adaptive gate for this writing-specific view is a promising direction for future work.

H TRACING MEMORY INTO NARRATIVE: QUALITATIVE CASE STUDIES

We examine how remembered details, relationships, and perspectives reappear in ConStory-Bench and LSGC continuations from Mem0, Graphiti, StructMem, and NarraWorld. The figures visualize selected textual memory records as condensed structures; these diagrams were not writer inputs. Each method continues its own generated history under a shared task and also receives recent text. Selected for clear memory–narrative correspondences and contrasting outcomes, these cases illustrate how memory is used within individual trajectories. They do not estimate how often these patterns occur or isolate retrieval's causal contribution. ConStory examples pair stored memory with continuations; LSGC examples also examine the retrieved text presented to the writer.

H.1 CONSTORY-BENCH: OBJECTS, PROMISES, AND RECONCILIATION

Figure 6 shows prompt 128, continuation unit 5. A young artist returns to their childhood home after abandoning drawing following an attack on their father. Earlier drawings and family promises acquire new significance as the story approaches reconciliation.

Mem0 retains the father's attachment to an old drawing and the narrator's difficulty expressing affection; the continuation returns to these details through mutual recognition. Graphiti links the father, a recovered sketch, and an earlier confrontation. The sketch becomes an object the narrator can resume, turning retrieval of a physical artifact into renewed artistic practice. StructMem connects collaborative drawing to growing trust with Kyle; the continuation expresses that relationship through the wish to keep drawing together. These examples show distinct forms of continuity across statements, entity relations, and developments.

NarraWorld connects the unfinished fox drawing, the narrator's knowledge of their mother's condition for receiving her box, and the father's promise to give them space. The continuation completes the fox while the father remains outside the threshold and does not ask to see the drawing. A remembered commitment thus appears as restraint in the scene, alongside a concrete artistic callback. A remaining inconsistency is the sketchbook cover changing from green to blue within NarraWorld's story. The case illustrates relational and thematic continuity while leaving object-state tracking as a separate challenge.

H.2 LSGC: EVIDENCE, PERSPECTIVE, AND REASSESSING LOYALTY

We examine starter-01, Chapter 10, selected for its contrasting chapter-level outcomes. The shared objective reveals that someone close to Ron helped conceal Fred's fate, challenging his understanding of loyalty. Figure 7 pairs each method's memory from the preceding nine chapters with its continuation.

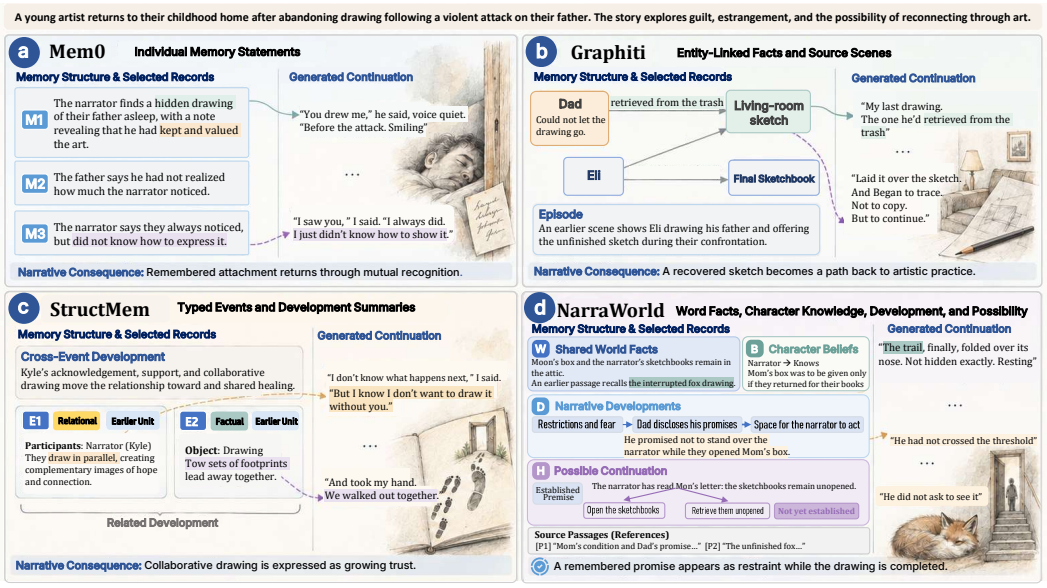


Figure 6: Memory and continuation in ConStory-Bench prompt 128, unit 5. Solid colored arrows connect recurring details; dashed arrows connect related developments.

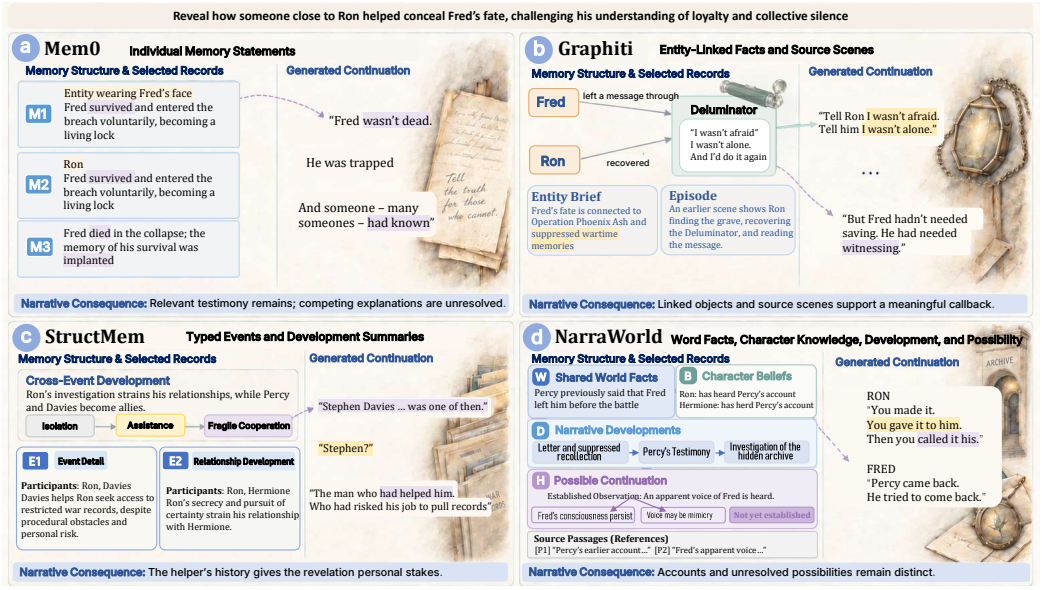


Figure 7: Memory and revelation in LSGC starter-01, Chapter 10. Solid arrows connect recurring content; dashed arrows connect related developments and reinterpretations.

Mem0 retrieves attributed claims that Fred survived, that he died, and that his apparent voice exploits Ron's grief, without explicit revision relations. The continuation opens with Ron's conviction that Fred is trapped, but leaves the competing accounts insufficiently reconciled. Graphiti exposes entity-linked evidence and earlier scenes. Its continuation follows records and testimony, returning to Fred's earlier message with a new emotional meaning: Ron recognizes a need to bear witness. The object and its message connect the investigation to a change in Ron's understanding.

StructMem preserves relationship changes across events. Davies's earlier willingness to risk helping Ron gives the revelation of his complicity personal stakes; Hermione's hurt over Ron's secrecy informs their renewed cooperation. NarraWorld separately presents Percy's earlier account, identifies who heard it, and tracks the developing investigation and unresolved interpretations of Fred's apparent voice. Its continuation reframes Percy's concealment as a coerced choice. Ron's response—"You made it. You gave it to him. Then you called it his."—shifts responsibility toward those who imposed that choice. The revelation develops through a reassessment of prior testimony and relationships.

Evidence updates remain a challenge: Graphiti's successive accounts of Fred's fate are incompletely reconciled, while NarraWorld's continuation only partly distinguishes evidence from Ron's interpretation (Section H.4). Mem0 also exhibits an unexplained location transition, and StructMem reverses a decision to stay behind without explanation. These observations distinguish the availability of historical context from its use in a new scene.

H.3 Further Examples of Constraints and Consequences

The following cases extend the analysis to physical rules, responsibility, evidence revision, and changing access to another character's inner state. Each comparison follows the constraints established by its own story.

World rules constrain rescue actions. In ConStory prompt 1939, unit 6, NarraWorld's memory preserves two unusual safety requirements: a containment indicator must remain red, and the capsule must stay sealed. The continuation administers treatment through a sealed channel, keeps the indicator red, and maintains separation while Mara reassures Nia through the glass. Physical safety and emotional commitment remain compatible. Graphiti's version connects a transmission network to an intervention through the host, while StructMem develops cooperation with the quarantined girl. Each trajectory follows its own established rescue mechanism.

Commitments become observable responsibility. In ConStory prompt 199, unit 8, memory records Cam changing a household reminder from "Kitchen reset. Do not make Nora ask." to "Kitchen reset." It also retains his promise to take responsibility for chores. The continuation makes that responsibility explicit: Cam treats reminders as his own obligations, while Eli acknowledges that listening should not depend on Nora remaining calm. Later actions, including laundry and hanging a wet towel, carry the conversation into daily behavior. Graphiti and StructMem also depict concrete follow-through; NarraWorld's distinctive emphasis here is the allocation of different responsibilities among three people.

New evidence changes the meaning of old clues. In ConStory prompt 461, unit 5, saved records distinguish Froshi's hopeful interpretation of a bridle rein from Mara's caution. Witness accounts and a recognized pendant subsequently change the interpretation of the rein, smoke, and red wool: they guide people toward one another without establishing that Froshi's sister remains alive. His grief and request for companionship persist after that revision. Attributed interpretations in world records and their connected developments support this revision.

Local success leaves consequences unresolved. In LSGC starter-02, Chapter 11, retrieved memory preserves the three tokens' states and earlier conditions involving voluntary contributions and a witness. The continuation uses the objects to establish a temporal anchor, but leaves the warden undefeated and the house beginning to dissolve. This preserves the requested crisis at the chapter boundary. Mem0 and Graphiti also depict the initial confrontation, then advance toward reconciliation or departure. The continuations differ in how far they advance beyond the chapter's crisis.

A relationship can continue after automatic access ends. In LSGC starter-04, Chapter 13, NarraWorld's memory connects coercive family arrangements, warnings about free choice, and the protagonists' loss of privacy. After the ritual, Pansy no longer automatically feels Theo's emotions and expresses her choice through speech: "This time, no magic carried the words." The continuation changes access to another mind while preserving the relationship. Other methods develop stronger or transformed bonds, offering different resolutions to the shared objective of attempting to break the bond.

H.4 Qualitative Limitations

Representing uncertainty does not ensure scrutiny. The LSGC starter-01 case exposes a gap between retaining uncertainty and acting on it during writing. Retrieved memory leaves open whether Fred's voice reflects surviving consciousness or mimicry. A source passage also warns that a reflector may show either truth or what its observer fears. Later, images associated with Carrow and partly corroborating testimony rapidly reshape Ron's judgment of Bill and Kingsley. The scene advances the emotional conflict, but only partly distinguishes what each source supports and what remains Ron's provisional interpretation. The scene illustrates the gap between retaining uncertainty in memory and examining new evidence during generation.

Remembered constraints still require execution. The sketchbook-color change in the ConStory case illustrates that thematic continuity can coexist with inconsistent object state. Similarly, returning relevant people and unresolved questions does not specify how a scene should test competing explanations. Rapid revelation through dialogue can achieve a chapter's emotional objective while leaving little space for investigation. These observations motivate evaluating whether selected evidence becomes appropriate action, belief revision, and state change, rather than measuring retrieval alone. Targeted checks of consequential constraints are a direction for future work.

Shared writing summaries should likewise preserve claim attribution and uncertainty, since stronger summary wording can override distinctions retained in memory.