Narrative World Model: Narratology-Grounded Writer Memory for Long-Form Fiction

Anonymous authors
Paper under double-blind review

Abstract

Long-form fiction writers need memory that answers multi-hop questions about evolving story state: who knows a secret and when they learned it, whether an event preceded the narration that revealed it, whether a setup paid off, and how a relationship shifted. General-purpose retrieval and agent-memory systems represent entities and facts but not the narratological structure these questions turn on, so they surface the wrong evidence or none at all. We introduce the *Narrative World Model* (NWM), a writer-memory system that pairs a narratology-grounded typed temporal-state graph with query-conditioned hybrid retrieval. To measure memory rather than the answerer, we read every system through a single held-constant Opus 4.8 reader over only that system's chapter-safe evidence, on a reproducible public corpus and a validated multi-hop benchmark, and we compare against the strongest existing temporal-knowledge-graph agent-memory framework, Graphiti/Zep (Rasmussen et al., 2025). NWM substantially and significantly outperforms this baseline on multi-hop narratological QA across both corpora, and far exceeds GraphRAG and flat retrieval. The advantage is representational rather than an artifact of extraction: it survives rebuilding the baseline with NWM's own extractor, and traces to the typing of narratological structure, not graph size or extractor quality.

1 Introduction

Long-form fiction generation exposes a gap between local fluency and durable narrative state. A model can write an appealing scene while forgetting who knows a secret, where an object is, whether a promise paid off, or how a relationship changed. Chapter $N{+}1$ must respect finalized chapters $1 \ldots N$ while advancing the story.

A writer-facing narrative memory system must do more than surface nearby text. Rolling summaries discard details that later become plot-critical. RAG (Lewis et al., 2020; Guu et al., 2020) retrieves relevant prose, but a passage may say where an object *was* before it moved. Graph retrieval can expose entities and events, but generic graphs do not necessarily track who knows what, the difference between event order and reveal order, relationship deltas, unresolved promises, or dramatic function. These approaches can miss the current, typed narrative state a writer needs even when the source prose exists.

These failures concentrate on *multi-hop narratological* queries: who knew a fact at chapter n and when they learned it, whether an event preceded the narration that revealed it, whether a setup planted earlier paid off, and what dramatic function a beat serves. Answering such a query requires evidence from two or more distinct chapters and a representation that types narratological structure rather than generic entities.

We therefore ask: does a narratology-grounded memory system provide better chapter-safe evidence for multi-hop story-state questions than generic search, source-chunk RAG, GraphRAG, or the strongest existing temporal-knowledge-graph agent-memory framework, Graphiti/Zep (Rasmussen et al., 2025)? We propose *Narrative World Model* (NWM), which publishes finalized chapters into typed memory records, global registries, a temporal knowledge graph (KG), and a recursive language-model QA (RLM QA) verifier. In a scoring

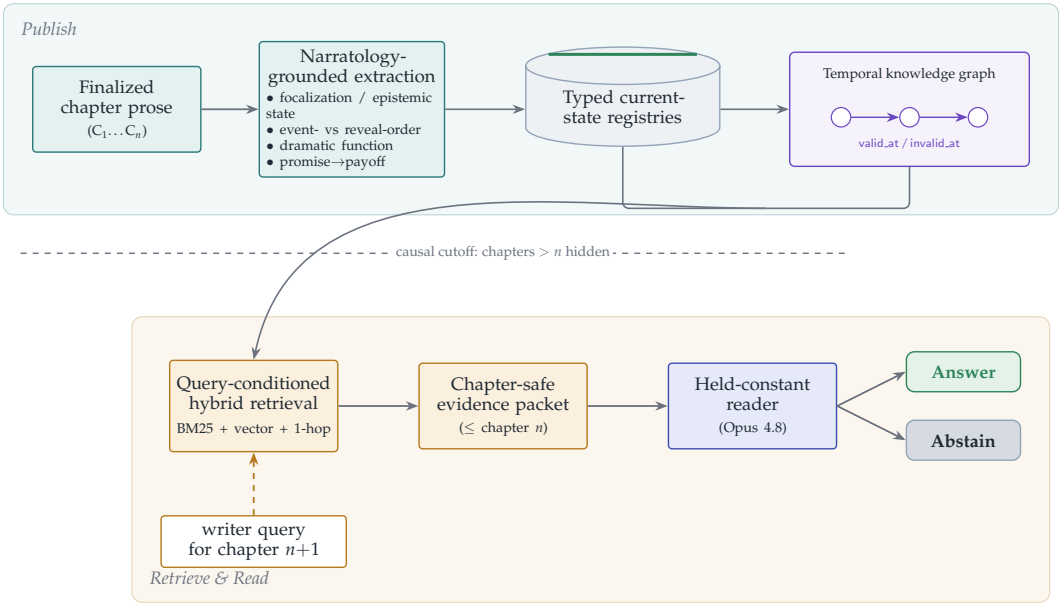

Figure 1: NWM pipeline. Finalized chapters are extracted into narratology-grounded typed records (focalization/epistemic state, event-vs-reveal order, dramatic function, promise/payoff) and a temporal knowledge graph with validity intervals; query-conditioned hybrid retrieval (BM25, vector, and one-hop graph expansion) assembles a chapter-safe evidence packet from chapters up to the checkpoint only, and a held-constant reader answers or abstains.

protocol over two corpora, each system answers nuanced writer questions from only its own chapter-filtered memory or retrieval evidence, isolating memory/retrieval behavior from generator behavior.

This paper makes the following contributions.

1. A writer-facing multi-hop narrative-memory QA protocol and a validated 176-item multi-hop slice built by curation, a single-passage hardness filter that discards questions a top-1 retrieved chunk can answer, and independent adjudication by a stronger model than the generator.

2. The NWM system: a narratology-grounded typed temporal-state graph with query-conditioned hybrid retrieval, with its publish/update flow, schema registries, temporal KG, and RLM QA.

3. NWM Graph Retrieval significantly beats Graphiti, the strongest existing temporal-KG framework, on two corpora under a held-constant Opus 4.8 reader: 0.898 versus 0.574 on the private multi-hop slice (paired McNemar 64–7, $p < 10^{-5}$) and 0.625 versus 0.516 on the public 576-item set ($p = 0.0001$), both far above GraphRAG and RAG.

4. An extractor-fairness control showing the gain is representational: Graphiti is extracted with NWM's own Sonnet 4.5 model, and re-ingesting it with a cheaper extractor does not change its accuracy ($p = 0.89$). A cross-system abstention analysis is the operative ablation.

5. A reproducible public protocol and benchmark, including a 110-item public multi-hop subset.

2 Related Work

Long-form story generation and revision. Neural story generation has long used intermediate representations to control global structure. Hierarchical Neural Story Generation decomposes generation into higher-level representations and lower-level realization (Fan et al., 2018); Plan-and-Write uses an explicit storyline as a control signal (Yao et al., 2019); Re3 and DOC use recursive prompting, revision, and detailed outlines to improve coherence (Yang et al., 2022; 2023). These systems control *intended* structure. NWM addresses a different object: *accepted* state after prose is finalized, and whether that state is actually available to the next continuation.

Retrieval and long-term memory. RAG conditions generation on external evidence (Lewis et al., 2020), and REALM treats retrieval as latent memory (Guu et al., 2020). Long-running memory systems aim to carry information across interactions (Zhong et al., 2023; Wang et al., 2023; Packer et al., 2023; Park et al., 2023), and production agent-memory frameworks such as Mem0 (Chhikara et al., 2025) extract and consolidate salient facts across sessions, with a graph variant that captures relational structure. Even when a large context is available, models use the middle unevenly (Liu et al., 2024). These systems carry generic facts or conversational state; NWM differs in its *unit of state*: rather than prose chunks or generic facts, it stores typed, versioned, chapter-scoped narrative state and exposes the latest causally valid slice.

Knowledge graphs and narrative QA. Knowledge graphs represent relational facts for lookup and traversal (Hogan et al., 2021); GraphRAG-style systems build retrieval over source-derived entity/event graphs (Edge et al., 2024), and more recent graph-structured retrievers such as HippoRAG (Gutiérrez et al., 2024) and LightRAG (Guo et al., 2024) couple entity graphs with dense retrieval to support multi-document reasoning. For fiction the relevant graph is instead temporal and narrative-specific. Multi-hop QA benchmarks such as HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) establish that questions requiring evidence from two or more passages form a distinct, harder regime than single-passage retrieval; they motivate our multi-hop slice, though they target encyclopedic facts rather than evolving narrative state. NarrativeQA established QA over stories as a test of narrative understanding (Kočiský et al., 2018). Our RLM QA layer follows the recursive-decomposition spirit of Recursive Language Models (Zhang et al., 2025) but specializes it to evidence-backed narrative-state verification.

Narratology and temporal graph reasoning. Narratology distinguishes who sees from who speaks, story-event order from discourse order, and event function within an arc (Genette, 1980; Propp, 1968). Writer-facing craft taxonomies also track dramatic situations and motivation/reaction beats (Polti, 1921; Swain, 1965). Temporal-KG methods model time-scoped facts and query-relevant graph neighborhoods (Goel et al., 2020; Han et al., 2021), while temporal KGQA benchmarks test questions over time-scoped relations (Saxena et al., 2021). The closest agent-memory framework is Graphiti/Zep (Rasmussen et al., 2025), a bi-temporal knowledge graph that records when a fact was true and when it was ingested, and serves it as time-scoped evidence for agents. NWM is closest to this line in its use of temporal edges and graph neighborhoods, but Graphiti's edges are generic time-scoped facts, whereas NWM's schema types are narrative-specific: knowledge boundaries, reveal order, promise/payoff status, focalized observer, and dramatic function are first-class writer-memory fields rather than generic entity relations. We use Graphiti as our strongest baseline.

3 Narrative World Model

3.1 Causal Publish Flow

NWM treats memory as a record of finalized prose, not future intent. Let C_n be accepted chapter prose, M_n world state after chapter n, and O_{n+1} the next scaffold:

$$M_n = \text{Publish}(C_n, M_{n-1}), \tag{1}$$

$$X_{n+1} = \text{Retrieve}(M_n, O_{n+1}), \tag{2}$$

where X_{n+1} is the retrieved memory packet for chapter $n+1$ writer queries and continuation support. The causal constraint is strict: a query or continuation may use updates completed through chapter n, never future scaffolds. Publishing stores a prose digest, extracts typed memory records, merges registries, projects graph indexes, and exposes the completed state; republishing an edited chapter invalidates downstream edges from that chapter.

Publish is an extraction step, not a copy: an extractor (Claude Sonnet 4.5) reads the accepted prose and emits the typed records of §3.2 directly, each stamped with its source chapter and evidence span and merged into cumulative registries (see Appendix A).

3.2 Schema Updates and Global Registries

Instead of storing only summaries or prose chunks, NWM extracts a public writer-memory structure. Implementations may add bookkeeping, but the writer-facing retrieval state is this evidence-backed typed memory:

NWM writer-memory structure

Memory record	Writer-visible fields
Chapter digest	chapter id, source-grounded summary, evidence spans
Scene/event	location, participants, event order, reveal order, source chapter
Character state	goal, knowledge/unknowns, status, location, relationship deltas
Relationship state	character pair, relation type, polarity/status, validity
Object state	object, owner, location, condition, last evidence
Plot/promise	thread id, structural role, open/closed status, promised payoff
Narrative function	focalized observer, dramatic beat, turn/reversal, reader knowledge
World fact	fact, scope, valid chapter range, evidence

Each record is chapter-scoped and evidence-backed, and merges into cumulative registries (characters, relationships, locations, objects, active threads, promises, world facts) that expose the latest causally valid slice for chapter $n+1$. The writer query interface needs current slices for relevant entities, not all prior chapters; thus a registry retains an old-but-current object location that recent summaries omit.

Two properties make these records more than generic facts: every record is *evidence-backed* (it stores its source chapter and supporting span), and the narratological fields are *first-class typed slots* (focalized observer, reveal order vs. event order, epistemic knowledge/unknowns, open/closed promise-payoff status, dramatic function) rather than prose that happens to mention a craft concept. These are the fields a generic entity/edge graph lacks (see Appendix A).

3.3 Temporal Knowledge Graph

The KG projects schema state into temporal relations: characters know facts, objects are located at places, relationships change, threads resolve, and events cause later events. Each edge stores source chapter, evidence, validity interval, and confidence. A vector retriever returns a relevant mention; the graph returns the *latest valid* state as of chapter n while preserving history.

Edges are typed by the narrative relation they assert (knowledge, location, relationship-polarity change, thread resolution, causation), carrying the §3.2 semantics into the graph, and each edge's validity interval makes the graph temporal rather than a flat snapshot. A query for state "as of chapter n" selects the latest edge valid at n for each entity or relation while older edges remain as history, returning an old-but-still-current state a rolling summary cannot (see Appendix A).

3.4 Recursive Language-Model QA

RLM QA is a recursive verifier over NWM, not a new base model. It decomposes a narrative question, queries schema and graph state, gathers evidence, and reports a support trace. It assembles pre-continuation constraints and checks candidate chapters for unsupported state changes.

3.5 Query-Conditioned Hybrid Retrieval

Retrieval turns the store into a chapter-safe evidence packet for a writer query q at checkpoint n: it restricts to records whose source chapter is $\leq n$, ranks entity/seed nodes by a hybrid of BM25 (Robertson & Zaragoza, 2009) and dense vectors (Xiao et al., 2023) fused by reciprocal-rank fusion (Cormack et al., 2009), expands each anchor's one-hop typed neighborhood, and truncates to a bounded heterogeneous packet that is the only evidence the reader sees (full four-stage description in Appendix A.1). This is what distinguishes NWM Graph Retrieval from NWM State Memory, which serializes *all* state up to n with no query-specific step: conditioning, not store contents, is the operative difference between the two NWM conditions.

3.6 Extraction and Baseline Independence

The primary comparison deliberately separates baseline extraction from NWM extraction. Simple search and source-chunk RAG index only chapter chunks; GraphRAG builds its own entity/event graph directly from chapter text; Graphiti (Rasmussen et al., 2025) ingests the same chapter episodes into its bi-temporal knowledge graph and retrieves time-scoped facts; and the NWM rows use NWM's own current-state memory and graph-retrieval conditions. No external baseline reads NWM schemas, registries, graph state, or other NWM-derived memory records. Every row is answered by the same held-constant Opus 4.8 answerer over its own chapter-filtered evidence and evaluated with the same token-normalized support rule; baseline construction itself may use an independent LLM extractor.

To separate the contribution of the representation from the contribution of the extraction model, Graphiti is extracted with NWM's own model (Sonnet 4.5), so the comparison is not confounded by extraction quality. As a robustness check, we also re-ingest Graphiti with a cheaper extractor (GPT-4.1-mini); if a cheaper extractor changed Graphiti's accuracy, the comparison would be sensitive to extraction quality, whereas if it does not, the gap is representational. Both ingests are read by the same Opus 4.8 answerer as every other row, holding the answerer and the representation constant while varying only the extractor.

4 Evaluation: Narrative Memory QA

4.1 Protocol

We evaluate over two corpora. A public corpus of 12 public-domain books (six genres, ≥ 20 chapters each) makes the protocol reproducible. A private corpus of five production-style serialized books of 50 chapters each is a production-scale check; it is not redistributable.

External baselines use only chapter text and never consume NWM schemas, registries, graph state, or other NWM-derived memory records: **Simple Search** is lexical retrieval over chapter chunks; **RAG** uses 360-word source chunks with 80-word overlap, BGE-large dense embeddings (Xiao et al., 2023), BM25 (Robertson & Zaragoza, 2009), and reciprocal-rank fusion (Cormack et al., 2009); **GraphRAG** builds per-book source-derived entity/event

graphs with an independent LLM extractor and retrieves seed nodes plus local graph neighborhoods per question; and **Graphiti** (Rasmussen et al., 2025) ingests chapter episodes into a bi-temporal knowledge graph and retrieves time-scoped facts, extracted with Sonnet 4.5 to match NWM's extractor (a cheaper-extractor re-ingest is reported as a robustness check in §4).

The NWM rows evaluate two retrieval conditions over the same published memory: direct current-state memory and question-conditioned graph retrieval. Every system is filtered by book and chapter so it cannot read future chapters, and every row is answered by the same held-constant Opus 4.8 answerer over its own chapter-filtered evidence.

Private multi-hop benchmark. On the private corpus we curated a multi-hop benchmark: 176 validated multi-hop questions, each requiring evidence from at least two distinct chapters, plus 96 matched single-hop control questions. Items are balanced across four narratological families: focalization/epistemic (who knows what, and when they learned it), reveal-order-versus-event-order, dramatic-shape/setup-payoff, and combination. Questions were generated from chapter text with a strong language model, then filtered by a single-passage hardness check that keeps a question only if a top-1 retrieved 360-word chunk cannot support its answer, and finally adjudicated for genuine multi-hopness and answer correctness by a separate, stronger model than the generator. This curation isolates questions that demand cross-chapter reasoning over typed narrative state.

Public multi-hop subset. On the public corpus we used the frozen 576-question source-written set and defined a 110-question multi-hop subset as those questions whose gold evidence cites at least two distinct chapters.

NWM retrieval conditions. **NWM State Memory** is the direct current-state condition: all typed memory items published up to the checkpoint are merged and serialized into a bounded evidence context, without an additional query-specific graph search step, testing whether the published memory itself contains enough evidence. **NWM Graph Retrieval** is the question-conditioned graph condition: the query ranks relevant temporal-graph nodes and relations, expands local graph neighborhoods, and packages compact evidence (plus recursive verification traces) for the same answerer, combining structured current-state memory and query-conditioned graph retrieval rather than a single vector index.

Answerers and scoring. All rows are answered by the same held-constant Opus 4.8 answerer over *only* the system's chapter-filtered evidence, so the reported differences compare each system's memory representation, extraction, retrieval, and evidence packaging under one answer model; the answerer must abstain when evidence is insufficient and return evidence ids. Final correctness for all rows is then computed with a token-normalized support check against the reference answers, accepting direct answer-string support or overlap on answer-bearing content tokens. Thus score measures evidence support for writer questions, not generated-continuation behavior or privileged store access.

Query taxonomy. Queries span latest character state, knowledge boundary, relationship constraint, object location/state, event grounding, promise/payoff, temporal ordering, world-rule constraint, narratological function, and source-needle detail. The taxonomy is motivated by narratology and writing-craft categories such as focalization, event function, dramatic turns, and promise/payoff, not by the serialized NWM schema alone.

Public memory-QA comparison. The public 576-question set is curated from chapter text with GPT-5.5-assisted curation; each item stores evidence spans, required chapters, checkpoint, category, and curator provenance. The public corpus makes the protocol reproducible while the private five-book corpus tests the same systems on longer, production-style serialized stories; in both settings, accuracy gains require useful evidence from the evaluated memory or retrieval system rather than from the answerer alone.

Memory source	Overall	Multi-hop	Control
No Memory	0.000	0.000	0.000
Simple Search	0.107	0.085	0.146
RAG (BGE+BM25+RRF)	0.188	0.176	0.208
GraphRAG	0.199	0.188	0.219
NWM State Memory	0.294	0.358	0.177
Graphiti	0.496	0.574	0.354
NWM Graph Retrieval	**0.893**	**0.898**	**0.885**

Table 1: Private multi-hop comparison over 272 questions (176 validated multi-hop, 96 single-hop control), scored with a held-constant Opus 4.8 answerer. Graphiti is extractor-matched to NWM (Sonnet 4.5). NWM Graph Retrieval reaches 0.898 multi-hop accuracy (Wilson 95% interval $[0.844, 0.934]$), against 0.574 for Graphiti. On the 176 multi-hop items the paired exact McNemar test between NWM Graph Retrieval and Graphiti is 64 to 7, $p < 10^{-5}$.

Memory source	Full (576)	Multi-hop (110)
RAG	0.214	0.191
GraphRAG	0.351	0.355
NWM State Memory	0.464	0.573
Graphiti	0.516	0.582
NWM Graph Retrieval	**0.625**	**0.709**

Table 2: Public comparison over the frozen 576-question source-written set and its 110-question multi-hop subset, scored with the same held-constant Opus 4.8 answerer. NWM Graph Retrieval reaches 0.625 full-set and 0.709 multi-hop accuracy, above Graphiti at 0.516 and 0.582.

Metrics. We report item-weighted QA accuracy under a deterministic token-coverage rule: the gold answer's salient content tokens must be at least half-covered by the evidence-supported answer. We report Wilson 95% confidence intervals (Wilson, 1927) and paired exact McNemar tests between systems on identical items, which control for item difficulty and isolate where one representation answers what another misses.

5 Results

All numbers below are produced by the evaluation implementation. The private multi-hop comparison is the headline result; the public corpus replicates the ordering on redistributable data. These are narrative-memory retrieval results, not full-story continuation outcomes.

5.1 Private Multi-Hop Comparison

Table 1 is the headline comparison. NWM Graph Retrieval reaches 0.898 on the 176 multi-hop items, against 0.574 for the extractor-matched Graphiti. The paired exact McNemar test between NWM Graph Retrieval and Graphiti is 64 to 7, $p < 10^{-5}$. Direct NWM State Memory, which serializes current state without a query-conditioned graph search, reaches only 0.358 multi-hop and does not beat Graphiti; against NWM Graph Retrieval the paired test is 101 to 6, $p < 10^{-5}$. The advantage appears only when the query ranks graph nodes and expands local neighborhoods. NWM Graph Retrieval also leads on the single-hop control (0.885), showing the gain is not specific to multi-hop items (per-system bars in Figure 2, Appendix A).

5.2 Public Comparison

Table 2 replicates the ordering on redistributable data. NWM Graph Retrieval reaches 0.625 on the full 576-item set and 0.709 on the 110-item multi-hop subset, above Graphiti at 0.516 and 0.582 and far above GraphRAG and RAG. The paired exact McNemar test between NWM Graph Retrieval and Graphiti is 156 to 93 on the full set ($p = 0.0001$) and 30 to 16 on the multi-hop subset ($p = 0.054$); the subset test is underpowered at $n = 110$.

Extractor fairness. The advantage is representational, not an artifact of NWM's extractor. Graphiti is extracted with NWM's own model (Sonnet 4.5), so the comparison is not confounded by extraction quality. As a robustness check, re-ingesting Graphiti with a cheaper extractor (GPT-4.1-mini) yields a sparser fact graph (Appendix 3)— entities rise from 443 to 491 while fact edges fall from 3,600 to 2,226 across the ingested chapter episodes—yet statistically indistinguishable accuracy (0.585 versus 0.574; paired exact McNemar $p = 0.89$), confirming the gap is representational, not an extraction artifact. Matching or cheapening the extractor does not close the gap to NWM Graph Retrieval.

Representation ablation. The cross-system comparison is itself the ablation. On the 64 private multi-hop items that NWM Graph Retrieval answered and Graphiti missed, Graphiti abstained on all 64, and in every case the gold fact was absent from Graphiti's retrieved evidence. These were narratological-structure questions, dramatic irony, reveal order, setup and payoff, and knowledge-boundary shifts, for which Graphiti's generic entity/edge schema has no representation. The gap is not extraction quality but the absence of narratological typing in the representation.

Representation-size footprints (Table 3, Appendix A) further show NWM's typed graph is smaller than the source GraphRAG graph yet answers far more multi-hop questions: typing of narrative structure, not graph density or extractor, is the driver (per-family breakdown in Figure 3).

6 Discussion

The results admit a single mechanistic reading: the advantage is where narratological structure is *typed*, not where the graph is larger or the extractor is stronger.

Why typed representation wins on multi-hop. A multi-hop narratological query is answered not by one fact but by composing *typed temporal relations across chapters*: an epistemic boundary that shifts, a reveal whose discourse position differs from its story position, a promise opened and discharged. NWM's edges carry exactly these types with validity intervals, so query-conditioned retrieval assembles the cross-chapter structure the query presupposes. A generic entity/edge graph (Edge et al., 2024; Rasmussen et al., 2025) has no slot for that structure, which is why the gap concentrates on multi-hop items and persists on the single-hop control. The failure mode confirms this: Graphiti almost always *abstained* rather than erring, with the gold fact simply absent from its evidence; the held-constant reader was told to abstain without support, so this is the signature of representational absence, not reasoning error. A better (denser) extractor cannot manufacture a narratological type the schema does not contain, which is why the extractor-fairness control left the gap unchanged.

Retrieval conditioning is part of the representation. Typing is necessary but not sufficient: the same typed memory dumped as a serialized current-state context (NWM State Memory) badly underperforms query-conditioned NWM Graph Retrieval and, on the private corpus, does not even beat the Graphiti baselines. The hop-count interaction reinforces this (Figure 4): on both corpora the multi-hop subset is exactly where graph-structured retrieval opens its largest margin, whereas flat retrieval stays low regardless of hop count. Structure in the store pays off in proportion to how much cross-chapter composition the query demands; removing either the typing (cross-system comparison) or the conditioning (State Memory) collapses the advantage.

These results establish that NWM supplies the chapter-safe evidence a writer question needs, a necessary-but-not-sufficient step before NWM-conditioned *generation*; the regime where this should matter most, very long serialized stories, and the writer-loop evaluation that would test it are discussed in Appendix A and Section 7.

7 Toward Generation and Writer-Workflow Evaluation

The next step is whether NWM-conditioned *generation* is better. A generation evaluation would place NWM inside the writing loop: continue story worlds for 10–15 chapters under matched memory conditions, then score contradictions, entity/character/relationship consistency, harder composite QA, writer correction burden, and blind pairwise human preference, testing whether KG/RLM verification changes temporal disambiguation, contradiction detection, and how retrieved memory shapes prose.

8 Limitations

Within-system ablation. We isolate the contribution of narratological typing through the cross-system comparison rather than a within-NWM field ablation: masking fields in the evidence packet is confounded by retrieval richness, since the held-constant reader can recover masked content from surviving graph rows. A retrieval-controlled within-system field ablation is left to future work.

Statistical power. The public multi-hop subset is underpowered at $n = 110$, where the NWM-versus-Graphiti advantage (30 to 16) is suggestive but not significant ($p = 0.054$). The private multi-hop slice ($n = 176$, $p < 10^{-5}$) and the full public set ($n = 576$, $p = 0.0001$) carry the significant results.

Benchmark design. The private multi-hop benchmark is narratology-tilted by design, balanced across focalization/epistemic, reveal-order, dramatic-shape, and combination families to stress exactly the structure NWM types. This is the intended contribution, disclosed so the result is read as a comparison on narratological multi-hop questions rather than arbitrary story QA.

Automatic scoring. The scorer is deterministic and reproducible but not a human-equivalence claim: a token-coverage support rule can penalize correct paraphrases and accept shallow lexical matches, so we treat the scores as evidence-support accuracy under a fixed rule. A stronger benchmark should report stratified human or LLM-judge agreement over accepted, rejected, and abstained answers.

Data and workflow scope. The public corpus is reproducible and the five-book private corpus is production-style but not redistributable; the key control is that all question writing and external baseline construction are separate from NWM memory records. The no-memory row is an evidence-free abstention control, not a closed-book parametric-memory probe; a closed-book run should be added if the claim concerns pretrained recall. The current results measure memory/retrieval support; conditioned *generation* remains future work (Section 7).

Baseline configuration. The GraphRAG row is one concrete chapter-level configuration (source-only LLM extraction, per-book stores, seed-node retrieval, local neighborhood expansion); it does not use Microsoft GraphRAG's global community-summary mode. Graphiti is run in its default retrieval configuration, extractor-matched to NWM (Sonnet 4.5), with a cheaper-extractor re-ingest reported only as a robustness check. The main claim is a comparison against these reproducible configurations, not every possible graph-RAG or temporal-KG system.

9 Conclusion

Narratology-grounded writer memory significantly outperforms the strongest existing temporal-knowledge-graph framework on multi-hop story-state QA. Under a held-constant Opus 4.8 reader, NWM Graph Retrieval scored 0.898 versus Graphiti 0.574 on a validated 176-item private multi-hop slice (paired McNemar 64–7, $p < 10^{-5}$) and 0.625 versus 0.516

on the public 576-item set ($p = 0.0001$), far above GraphRAG and RAG. The gain is representational: Graphiti uses NWM's own Sonnet 4.5 extractor, and re-ingesting it with a cheaper extractor leaves its accuracy unchanged ($p = 0.89$), and on the multi-hop items NWM answers and Graphiti misses, Graphiti abstains on all 64 because its generic schema cannot represent narratological structure. NWM-conditioned generation is future work.

References

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

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

Gérard Genette. *Narrative Discourse: An Essay in Method*. Cornell University Press, 1980.

Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. Diachronic embedding for temporal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *International Conference on Learning Representations*, 2021.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, Jose Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 2021.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

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*, 2024.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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.

Georges Polti. *The Thirty-Six Dramatic Situations*. J. K. Reeve, 1921.

Vladimir Propp. *Morphology of the Folktale*. University of Texas Press, 1968.

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

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Dwight V. Swain. *Techniques of the Selling Writer*. University of Oklahoma Press, 1965.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, 2023.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

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*, 2022.

Kevin Yang et al. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

A Additional Method Details

This appendix expands the method summaries in §3 and collects the supporting figures and tables relocated from the body. The body retains the terse §3.1–3.4 prose and the schema table; the fuller descriptions below add no new claims.

Extraction at publish (expands §3.1). Publish is an extraction step, not a copy. When a chapter is finalized, an extractor (Claude Sonnet 4.5) reads the accepted prose and emits the typed memory records of §3.2 directly, rather than a free-text summary: each emitted record is tied to the chapter that produced it and carries the evidence span that licenses it. Records do not overwrite the store; they merge into the cumulative registries, keyed by the entity, relationship, object, or thread they describe, so that a single entity accretes a chapter-ordered history of typed states. The causal cutoff is a property of the store, not of a single query: it is enforced at write time, because every record is stamped with its source chapter, and again at read time, because retrieval restricts to records stamped at or before the checkpoint. Editing and republishing a chapter re-runs extraction for that chapter and invalidates the downstream edges that depended on the superseded state, so a correction propagates forward rather than leaving stale assertions in the graph.

Evidence-backed narratological fields (expands §3.2). Two properties make these records more than generic facts, and both are visible in the writer-memory structure (§3.2). First, every record is *evidence-backed*: it stores the source chapter it was extracted from and the evidence span that supports it, so a downstream answer can cite where a state was established rather than asserting it unsupported. Second, the narratological fields are *first-class typed slots*, not prose that happens to mention a craft concept. A focalized observer names through whose perception a scene is rendered; reveal order is stored separately from event order, so the chapter in which the reader learns a fact is distinguished from the chapter in which the underlying event occurred; character knowledge and unknowns record an epistemic boundary, the set of facts a character does and does not yet hold at a checkpoint; a plot or promise record carries an open-versus-closed payoff status, marking whether a setup planted earlier has yet been discharged; and a narrative function record names the dramatic beat or turn a scene performs. These are the fields a generic entity/edge graph lacks: it can record that two entities are related, but not who is licensed to know it, when it was revealed as opposed to when it happened, or what arc function its disclosure serves.

Narrative-typed temporal edges (expands §3.3). Edges in the KG are typed by the narrative relation they assert rather than by a generic association. An edge records that a character knows a fact, that an object is located at or owned by a place or agent, that a relationship has changed polarity, that a thread has resolved, or that one event causes a later event; the same edge type therefore carries the narratological semantics of §3.2 into the graph. Each edge stores its source chapter, its evidence, a validity interval over chapters, and a confidence. The validity interval is what makes the graph temporal rather than a flat snapshot: a fact that held from the chapter it was established until the chapter it was superseded is a closed interval, while a fact that still holds is left open.

State as of a checkpoint (expands §3.3). Because edges carry chapter-scoped validity, a query for state "as of chapter n" is answered by selecting, among the edges valid at n, the latest one for each entity or relation, while older edges remain in the graph as history rather than being discarded. This is what a rolling summary cannot do: it lets the store return an old-but-still-current state (for example, an object's last known location) even when more recent chapters do not mention it, and it lets a later query reconstruct the same entity's earlier state without re-reading the prose.

A.1 Query-Conditioned Hybrid Retrieval (expands §3.5)

The publish flow builds the store; retrieval is what turns it into a chapter-safe evidence packet for a specific writer query. Given a writer query q and a checkpoint chapter n, Retrieve(M_n, q) proceeds in four stages.

- **Causal restriction.** Retrieval first restricts the searchable store to records and edges whose source chapter is at most n, so no candidate evidence can originate in a future chapter. This is the same cutoff enforced at write time, re-applied as a hard filter at read time.
- **Hybrid node ranking.** Within that restricted store, entity nodes and seed nodes are ranked against q by a hybrid score that combines a lexical signal (BM25 (Robertson & Zaragoza, 2009)) with a dense-vector signal over embedded node text (Xiao et al., 2023); the two rankings are merged by reciprocal-rank fusion (Cormack et al., 2009) so neither signal dominates. The result is a small set of top-ranked entity and seed nodes that anchor the query in the graph.
- **One-hop graph expansion.** Each anchor node's one-hop neighborhood is expanded to pull in the typed edges incident to it (and their endpoints), bounded by a neighbor limit. This is the step that converts a ranked set of nodes into connected typed structure: it surfaces the knowledge, location, relationship-delta, reveal, and thread-resolution edges that a node-only ranking would leave implicit.
- **Bounded packet assembly.** The ranked vector hits, the expanded graph nodes, and the expanded graph edges are merged, re-ranked, and truncated to a size-bounded evidence packet. The packet is therefore heterogeneous by construction, ranked entity/seed vector hits together with the graph nodes and typed edges they induce, and it is the only evidence the held-constant reader sees for that query. RLM QA (§3.4) may optionally decompose the query and verify the assembled evidence before the reader answers.

Why conditioning matters. The contrast with NWM State Memory makes the role of conditioning explicit. NWM State Memory serializes *all* typed memory items published up to n into a single bounded context, with no query-specific search step; it tests whether the published state, taken whole, contains the answer. Query-conditioned hybrid retrieval instead uses q to decide *which* parts of the store to surface and then follows the graph one hop from those parts. The two share the same underlying store and the same causal cutoff, and differ only in conditioning. A multi-hop narratological query needs a few specific typed edges from different chapters, an epistemic boundary established early, a reveal recorded later, a relationship delta in between, and these are easy to bury when the entire state is serialized at once but easy to surface when the query ranks the relevant nodes and the graph supplies their typed neighborhoods. Conditioning, not store contents, is therefore the operative difference between the two NWM conditions.

A.2 Extended Discussion: Implications for Long-Form Generation

These results establish that NWM supplies the chapter-safe evidence a writer question needs; they do not by themselves establish that NWM-conditioned *generation* is better, a separate, necessary-but-not-sufficient step. Answering correctly is a precondition for using the answer to constrain the next chapter, but the writer-loop study (Section 7) is where one would measure whether better evidence yields more consistent continuations. We expect the regime where structured memory matters most to be the production regime of very long serialized stories: as story length grows, full-context prompting degrades both in cost and in the model's uneven use of long contexts (Liu et al., 2024), while a query-conditioned typed store returns a bounded, chapter-safe slice whose size is set by the query rather than the story. The margins here are on memory QA over corpora up to fifty chapters per book; the gap to full-context prompting should, if anything, widen as stories lengthen beyond what a single context can faithfully hold.

Table 3 quantifies representation size. NWM's typed graph is smaller than the independently built source GraphRAG graph while answering far more multi-hop questions (Tables 1 and 2), and Graphiti's denser matched-extractor graph does not close the gap. Read together, these two facts separate *what* is represented from *how much*: if accuracy tracked graph size, the larger GraphRAG and denser Graphiti graphs would lead, yet the smaller narratology-typed graph wins. The result therefore points to the typing of narrative structure, not graph density or extractor, as the driver.

Representation	Nodes / Entities	Edges
NWM temporal graph (public)	7,136	16,951
Source GraphRAG graph (public)	18,345	72,768
Hybrid RAG source chunks (public)	2,347	—
Graphiti (cheaper extractor) (private)	491	2,226
Graphiti (matched extractor) (private)	443	3,600

Table 3: Representation footprints from saved evaluation outputs; counts measure representation size, not QA accuracy. On the 12-book public corpus the source-only GraphRAG graph has $\sim 2.6\times$ as many nodes and $\sim 4.3\times$ as many edges as the deployed NWM graph. On the private corpus, Graphiti's matched (Sonnet 4.5) extractor yields a denser fact graph (more edges, fewer entities) than the cheaper-extractor robustness ingest, yet the cheaper ingest does not change its multi-hop accuracy.

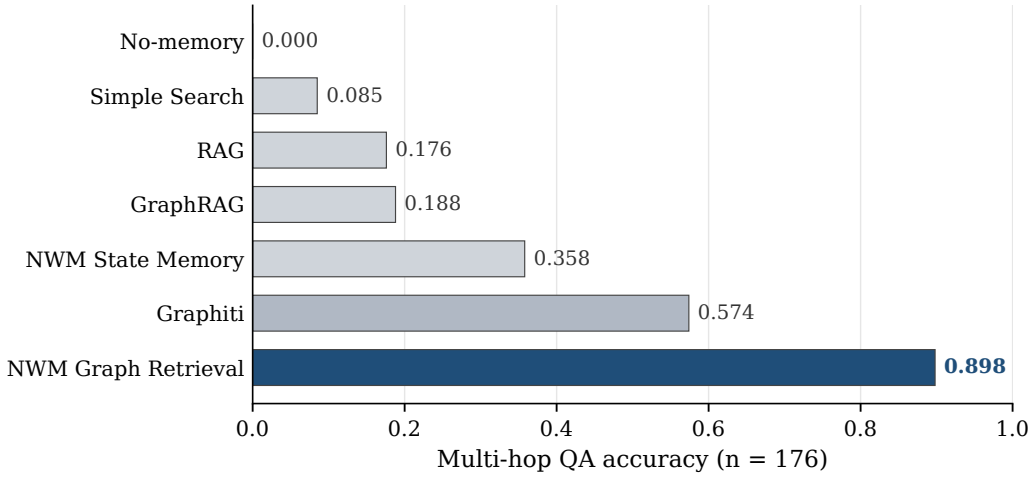

Figure 2: Multi-hop accuracy for all systems on the private benchmark, plotting the multi-hop column of Table 1.

B Memory-QA Examples

Table 4 shows twenty representative questions from the public 576-item memory-QA set used in Table 2. Gold answers are abbreviated for space; the released evaluation set contains the full question, answer, evidence spans, curator id, book id, and checkpoint for every item.

Table 5 shows representative questions from the private five-book corpus used for the multi-hop benchmark in Table 1. To preserve the confidentiality of the production corpus, the five serialized books are de-identified as Books A–E with their genre, and character names are replaced by fixed neutral pseudonyms; the questions and gold answers are otherwise written from chapter text rather than NWM memory.

C Private Multi-Hop QA Examples

Table 6 shows eight representative items from the 176-question private multi-hop benchmark used in Table 1, two from each of the four narratological families. Each question requires evidence from at least two distinct chapters (the *Chapters* column lists the gold required chapters). Book titles and character names are anonymized; gold answers are abbreviated to roughly one line.

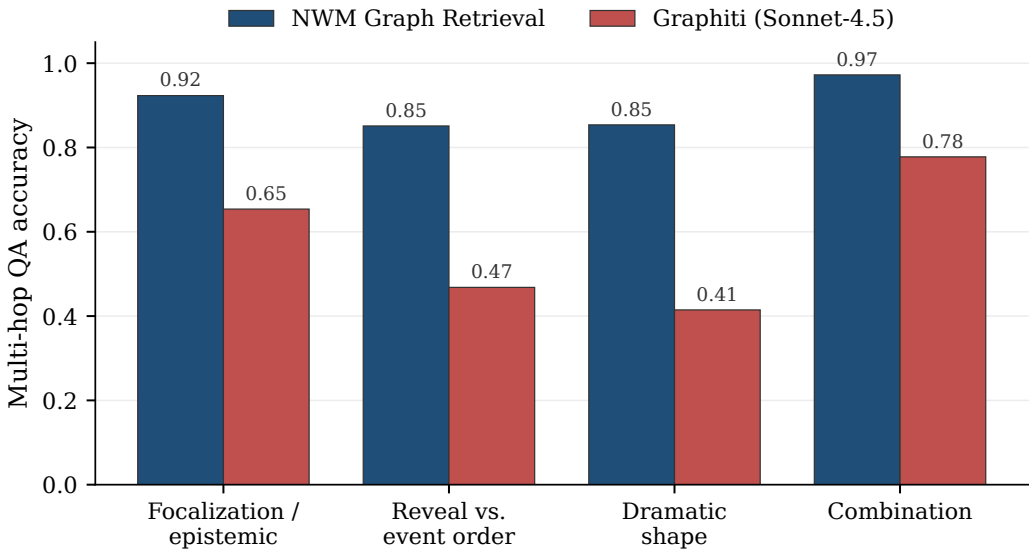

Figure 3: Per-narratology-family multi-hop accuracy on the private benchmark, NWM Graph Retrieval versus Graphiti (extractor-matched, Sonnet 4.5).

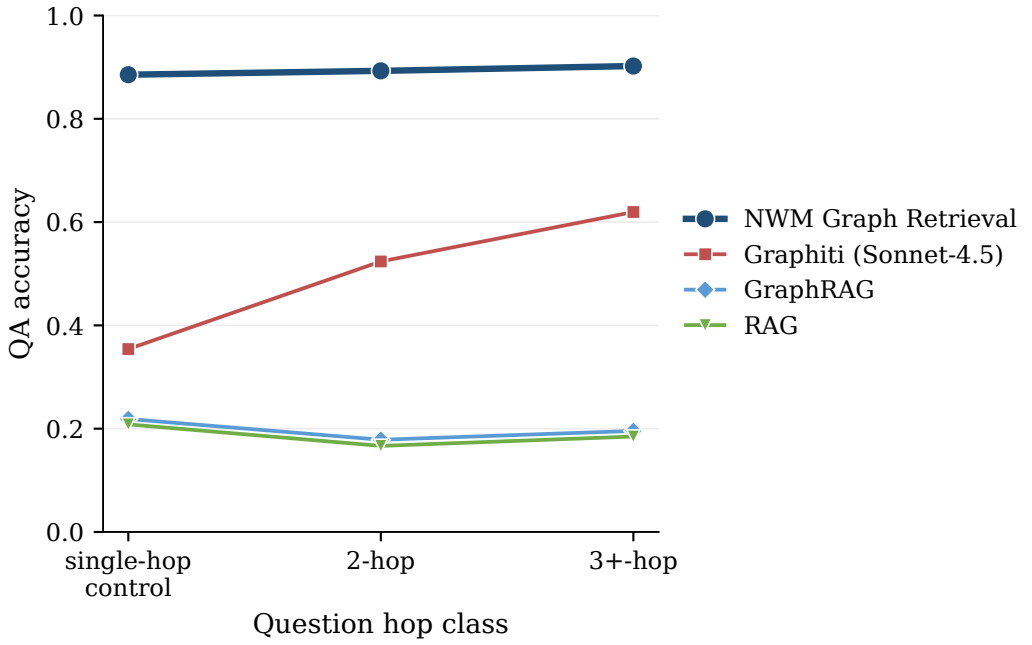

Figure 4: Accuracy by hop class (single-hop control, 2-hop, 3+-hop). Graph-structured systems rise or stay high as the hop count grows, while flat retrievers stay low across hop classes.

D Qualitative Case Study

We illustrate the representation gap with a single discordant item from the *dramatic-shape* family in Book D (romance/fantasy), spanning chapters 3 and 9, on which NWM Graph Retrieval answered correctly and Graphiti (extractor-matched, Sonnet 4.5) abstained. This

Slice	Book	Example question	Gold answer
Continuity	pg9746	Why does Juliet Byrne not know who her birth parents are by the opening chapters?	She was adopted by Lady Byrne and Sir Arthur, and Lady Byrne kept her parentage secret.
Continuity	pg2883	Who narrates the investigation in the opening chapters?	Herbert Burroughs, a young but already experienced detective.
Continuity	pg26514	Who is the narrator of the adventure?	Mark Strong, owner of the yacht *Celsis* and friend of Roderick and Mary.
Continuity	pg25186	By checkpoint 5, who is secretly watching the Indian fleet and village?	Henry Ware and his four forest-runner comrades are watching from cover.
Continuity	pg1214	Why is Harmony left alone in the old lodge at the start?	Scatch and the Big Soprano leave Vienna, while Harmony stays for music study.
Narratology	pg9746	How is Gimblet characterized when he first appears?	A celebrated, observant private detective who loves strange and inexplicable cases.
Narratology	pg2883	How does the opening use Fleming Stone to define the detective ideal?	Stone's brilliant deductions set a standard that Burroughs admires but cannot match.
Narratology	pg26514	How is Martin Hall characterized in the opening chapters?	Comic and foolish on the surface, but his private request and manuscript reveal courage.
Narratology	pg25186	What mood does the opening river description establish?	A vast, ancient, ominously quiet wilderness mood.
Narratology	pg1214	What mood does the old stucco house create in chapter 1?	Faded grandeur mixed with poverty, cold, and loneliness.
Arc	pg9746	What is the main inciting turn for Juliet by checkpoint 5?	The solicitors' letter pulls Juliet from adopted life toward a family revelation.
Arc	pg2883	What event launches Burroughs into the case?	He is sent to West Sedgwick to investigate Joseph Crawford's murder.
Arc	pg26514	What launches the main pursuit arc?	Hall's request, disappearance, and manuscript push Mark toward Captain Black.
Arc	pg25186	What arc function does the passing fleet serve in chapter 1?	It launches the surveillance plot by leading the five to the Wyandot village.
Arc	pg1214	What arc function does Scatchy's under-pillow gift serve?	It forces Harmony's stay-or-return decision and makes independence precarious.
Needle	pg9746	What color was the envelope Lord Ashiel had Higgs bring to Gimblet?	Blue.
Needle	pg2883	What kind of bag is found on Crawford's desk?	A gold-mesh woman's bag.
Needle	pg26514	At what Paris hotel do Mark, Roderick, and Mary stay?	The Hotel Scribe.
Needle	pg25186	What crops are named in the cleared land around the Wyandot village?	Corn and pumpkins.
Needle	pg1214	What instrument does Harmony carry and study?	The violin.

Table 4: Representative examples from the 576-question memory-QA set.

item is one of the 64 private multi-hop questions in the NWM-wins cell of the paired McNemar test against Graphiti; Graphiti abstained on every one of those 64.

Question. *What reversal occurs in the governess's narrative function between Chapters 3 and 9?*

Gold answer. In Chapter 3 the governess functions as the protagonist's abusive tormentor, threatening to beat her and wishing she had drowned her. By Chapter 9 she becomes a desperate mother begging the protagonist—the girl she despises—for help saving her son, and the protagonist responds with sympathy rather than gratification.

NWM Graph Retrieval (correct). Query-conditioned retrieval over the NWM temporal graph surfaced two chapter-anchored scene-beat nodes that together encode the role reversal: an early beat in which the governess is the powerful abuser, and a chapter-9 beat in which she is an injured, helpless supplicant begging for help with her abducted son. Reading only these chapter-safe rows, the held-constant answerer recovered the dramatic-function shift: "The governess shifts from being the protagonist's mocking, powerful abuser to an injured, helpless supplicant who begs her for help saving her abducted son, reversing their power dynamic." The narratological link—a single character's *dramatic function* inverting across chapters— was carried by the graph's typed scene/situation structure, not by any one surface fact.

Graphiti (abstained). Graphiti's retrieval returned a single entity/fact edge for the same query, and the answerer responded "Not enough evidence is available to answer from the

Slice	Book	Ch.	Representative internal question	Gold answer
Continuity	Book A (drama)	10	At the chapter 10 checkpoint, what leverage does Vera use to draw Dana to the Hotel Finest bar at 8pm?	Vera claims she can tell Dana where her abandoned child is.
Continuity	Book B (urban fantasy)	20	After biting Lena, what secrecy problem does Reyes identify before she wakes up?	He worries Lena may remember, and that exposure of his rare ability would draw private companies and the military before he can protect himself.
Continuity	Book C (dark fantasy)	20	By chapter 20, why is Sable's True Name dangerous to reveal?	Because the Shadow Bond would bind him to anyone who knows his True Name and recites it aloud.
Continuity	Book D (paranormal romance)	20	After Lorne questions her about the forest, what does Selene conclude about her hidden secret?	Lorne has not discovered it; her secret and her life are still safe for now.
Continuity	Book E (business drama)	20	What investment deal does Ellison close with Hale for Nova TV in chapter 20?	Ellison invests fifty million dollars for thirty percent of Nova TV's shares.
Narratology	Book A (drama)	35	How does Mara's arrival reframe Dana's family standing in front of Royce's household?	Mara's poise immediately outclasses the Royce household, and Royce drops his arrogance to ingratiate himself.
Narratology	Book D (paranormal romance)	10	What does the narration reveal during the dragon fight that Selene herself does not notice?	Her father arrives with Eclipse warriors, and another group arrives with a young man who shifts into a massive golden dragon.
Narratology	Book E (business drama)	35	What reversal does Carver experience after Rourke explains Ellison's status in chapter 35?	Carver realizes Ellison is someone even Rourke cannot afford to offend, so he becomes worried about insulting him.
Arc	Book A (drama)	10	What separate 8pm plans are being set up at Hotel Finest in chapter 10?	Aldo is staging a bar surprise while Cleo uses a walk with Mrs. Vane as cover to meet Cassian at the first-floor cafe.
Arc	Book B (urban fantasy)	35	How does Reyes's blood bank change the Drall fight from a skill mismatch into a winnable fight?	It heals Reyes mid-fight, letting him take damage recklessly and keep pressing Drall until the blood swipes break through.
Arc	Book C (dark fantasy)	50	Why do Nessa, Sable, and Cael decide they must attack the bone-scythe monster before morning in chapter 50?	They are trapped on the cliffs until morning; once sunrise lets the monster see them, they lose surprise, so they choose to strike first.
Arc	Book D (paranormal romance)	50	What immediate political outcome has Selene engineered by chapter 50?	The Alphas Summit has been cancelled, keeping Luna Coriel and Linus out of danger.
Needle	Book A (drama)	10	Where exactly does Cassian tell Cleo to meet him?	At Table 28 in the cafe on the first floor.
Needle	Book C (dark fantasy)	50	What object saves Sable from being swept away by the rising sea during the storm?	Nessa's golden rope.
Needle	Book E (business drama)	35	What does Alina realize Ellison's plain black card actually is in chapter 35?	It is the legendary Apex Club black card, the club's highest-level membership card, with only three in existence.

Table 5: Representative questions and gold answers from the 60-question internal writer-memory QA set, with the five production books de-identified as Books A–E and character names replaced by fixed pseudonyms.

retrieved context" and abstained. Graphiti's generic entity/relationship schema records that the governess and the protagonist co-occur and interact, but has no representation of a character's evolving *narrative role*; the chapter-3-versus-chapter-9 functional reversal was simply absent from its retrieved evidence. This pattern is typical of the discordant cases: across the 64 multi-hop items NWM answered and Graphiti missed, Graphiti abstained on all 64, and in every case the gold fact was missing from its retrieved evidence altogether.

E Benchmark Construction Details

The private multi-hop benchmark was built over the five production-style serialized books (50 chapters each) in three stages.

Curation. Candidate questions were generated from chapter text with a strong language model (GPT-5.5, high-reasoning setting), instructed to target the four narratological families—focalization/epistemic state, reveal order versus event order, dramatic shape (setup and payoff), and combinations of these—and to require evidence that spans more than one chapter.

Single-passage hardness filter. Each candidate was passed through a single-passage hardness check: the question was retained only if a top-1 retrieved 360-word chunk *could not* support the answer. This removes questions that a flat passage retriever could solve from one local window and biases the surviving set toward genuinely multi-chapter reasoning.

Family	Book	Chapters	Question	Gold answer (abbreviated)
Focalization / epistemic	Book A (drama)	26, 38	How does a side character's knowledge of the twin secret evolve from confusion to partial understanding?	At first he is baffled because the child playing with him and the child doing homework both seem to be the same boy; after seeing both children he understands that the protagonist is the boy's biological mother and why the twins keep the truth secret for now.
Focalization / epistemic	Book B (progression fantasy)	11, 12	How does the protagonist keep a peer from confirming a suspicion the peer already holds about the protagonist's hidden ability?	The peer already suspects a hidden ability after seeing the protagonist throw an opponent with surprising ease; when the protagonist feels his wounds healing he insists the peer leave, since watching the wound close would confirm it.
Reveal vs. event order	Book A (drama)	1, 11	How does an antagonist's chapter-11 confession reframe the earlier explanation of the protagonist's childhood obesity?	Chapter 1 first explains the obesity as an accidental incorrect hormone prescription, but the antagonist later reveals it was deliberate sabotage meant to weaken the protagonist's marriage prospects.
Reveal vs. event order	Book B (progression fantasy)	1, 18	How is the nature of the protagonist's inherited book withheld at first and then reinterpreted by chapter 18?	In chapter 1 the system's wording is cut off when the protagonist passes out; only by chapter 18 does he reinterpret the accumulated signs and ask whether he has been transformed into a supernatural being.
Dramatic shape	Book E (drama)	1, 5	How does the story reverse a love interest's chapter-1 dismissal of the protagonist by chapter 5?	In chapter 1 she rejects him as too poor; in chapter 5 it is publicly revealed that he secretly bought the high-end restaurant, leaving her shocked and unable to grasp what she misjudged.
Dramatic shape	Book D (romance/ fantasy)	3, 9	What reversal occurs in the governess's narrative function between Chapters 3 and 9?	In Chapter 3 the governess is the protagonist's abusive tormentor; by Chapter 9 she becomes a desperate mother begging the protagonist—the girl she despises—for help saving her son, and the protagonist answers with sympathy rather than gratification.
Combination	Book B (progression fantasy)	7, 8	How do chapters 7 and 8 turn a roommate's "fate" claim into dramatic irony?	Chapter 7 marks the roommate with a power-level-5 wrist number; in chapter 8 he calls the shared room assignment "fate," but a separate scene reveals someone forced a student to swap rooms and that the attacker bore a 5, so the reader can suspect engineered manipulation.
Combination	Book A (drama)	2, 11, 12	How does a suitor's airport mistake set up his later humiliation at the bar?	At the airport he fails to recognize the now-glamorous protagonist and denies any tie to the "ugly lady"; later he courts the same woman under an alias, only for her to reveal it is merely her middle name and that she is the very person he scorned.

Table 6: Representative private multi-hop QA items, two per narratological family. Each requires evidence from the listed chapters. Titles and names anonymized; gold answers abbreviated.

Independent adjudication. The surviving questions were adjudicated by a separate, stronger model than the generator (Opus 4.8) for two properties: genuine multi-hopness (the gold answer demonstrably depends on evidence from at least two distinct chapters) and answer correctness against the cited evidence spans. Items failing either check were dropped or flagged.

Final set. The procedure yielded 176 validated multi-hop questions plus 96 matched single-hop control questions, balanced across the four narratological families (focalization/epistemic 52, reveal-vs.-event-order 47, dramatic-shape 41, combination 36 among the multi-hop items). The control questions are single-chapter by construction and are used to confirm that systems are not simply rewarded for verbosity: a system must still abstain when the chapter-filtered evidence is insufficient. Every released item carries its full question, gold answer, evidence spans with chapter anchors, narratological family, hop count, and the curator and adjudicator identifiers.