Staying In Character: Perspective-Bounded Memory for Book-Based Role-Playing Agents

Anonymous ACL submission

Abstract

Recent LLM role-playing systems build character agents from novels by extracting characters, scenes, and relations. Yet long-narrative role playing suffers from two failures: *Factual Overreach*, where shared retrieval or parametric memory lets a character use facts outside its perspective, and *Stylistic Monotony*, where profile descriptions flatten a character into a fixed voice. To address these failures, we propose REVERIEMEM, a three-layer memory architecture for book-based character agents. The episodic layer stores first-person scene memories; the semantic layer stores visibility-tagged facts; and the personality layer stores situation-dependent speech and behaviour patterns. For evaluation, we construct KBF-QA, a 4,386-question benchmark over eight novels for testing knowledge boundaries. REVERIEMEM improves Knowledge Boundary Fidelity by 34.6 percentage points over the strongest prior method. On BOOKWORLD's five-dimension pairwise narrative protocol, REVERIEMEM achieves a $\sim 79\%$ win rate, suggesting that perspective-bounded memory improves both boundary fidelity and character-grounded narrative generation. The code of this paper can be found at: https://anonymous.4open.science/r/ReverieMem-7D5F/.

1 Introduction

"The divine gift does not come from a higher power, but from our own minds."
— Dr. Robert Ford, Westworld

LLM role playing agents have become a practical interface for character chat, narrative sandboxes, and interactive story generation (Shao et al., 2023; Wang et al., 2024a; Tu et al., 2024; Chen et al., 2024a). Recent work builds such agents from novels by extracting characters, scenes, and relations as narrative context (Zhao et al., 2024; Ran et al., 2025; Wang et al., 2025b); for example, BOOKWORLD constructs interactive agent societies from books and guides characters with profiles and story context (Ran et al., 2025). A credible character agent must reason within the character's knowledge boundary and dynamically adapt its voice and behaviour to the evolving narrative situation. In long novels, this constraint is often broken by two OOC (out-of-character) failures, illustrated in Figure 1: **Factual Overreach** and **Stylistic Monotony**.

The first failure, **Factual Overreach**, occurs when an agent states canonically true facts that are unavailable to the character it is playing. This can happen when long-narrative RAG systems retrieve from a shared book-level memory, or when the LLM's parametric memory supplies the fact directly (Gutiérrez et al., 2024; Wang et al., 2026). In Figure 1 (a), Lestrade should not confirm Holmes's deduction about the commissionaire's background at Baker Street: the deduction is correct, but Lestrade was not present.

The second failure, **Stylistic Monotony**, occurs when profiles collapse a character into one expressive mode. In Figure 1 (b), Holmes, for example, is not only a cold analytical reasoner: in Doyle's first chapter, he greets Watson as an excited experimenter (*"I've found it!"*), then moments later becomes the composed deducer who infers Watson's military past (*"Afghanistan, I perceive"*). A static profile may produce a recognisable Holmes while losing this situational shift.

Addressing these failures requires memory that is neither a shared book index nor a static profile. Factual access must be scoped to the character's narrative position, while expression must be grounded in situated behaviour across scenes. Three cognitive accounts guide the design: 1) *Complementary Learning Systems* (McClelland et al., 1995) models human memory as the cooperation between hippocampal encoding of individual experiences and neocortical consolidation of structured knowledge; 2) Conway's *Self-Memory System* (Conway and Pleydell-Pearce, 2000) casts au-

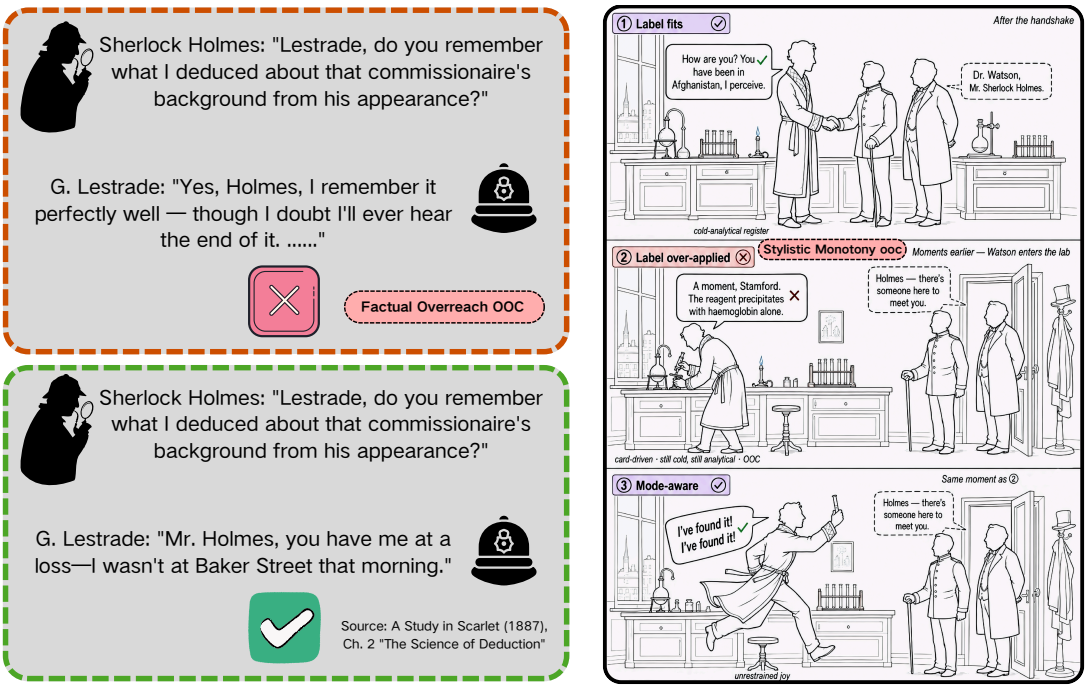

Figure 1: Two OOC failures in long-narrative role playing: **Factual Overreach** (a), where a character claims facts outside its perspective, and **Stylistic Monotony** (b), where a character's expression is flattened into one fixed profile.

tobiographical recall as a top-down reconstruction that locates event context first and reconstructs details within it; 3) *Narrative identity* (Mcadams and McLean, 2013) holds that identity is constituted by self-defining episodes rather than static traits.

We therefore present REVERIEMEM, a cognitive-psychology-inspired three-layer memory architecture for book-based role playing. The *Episodic Layer* stores first-person scene summaries from each character's perspective; the *Semantic Layer* stores structured facts with character-specific visibility; and the *Personality Layer* stores situation-dependent patterns distilled from canonical speech, behaviour, and emotion transitions. At inference time, REVERIEMEM anchors on the character's own scenes, retrieves only visibility-allowed facts, and conditions the response on an appropriate behavioural pattern.

We evaluate REVERIEMEM on eight novels with two complementary tasks. For knowledge boundaries, we construct KBF-QA (Knowledge Boundary Fidelity-QA), a 4,386-question multiple-choice benchmark in which a character must answer facts it could know and refuse facts outside its perspective. For open-ended narrative quality, we follow the five-dimension pairwise comparison protocol of BOOKWORLD. REVERIEMEM improves KBF by 34.6 percentage points over the strongest prior method and achieves a $\sim 79\%$ win rate in pairwise narrative comparisons.

This work makes three contributions:

- To our knowledge, we are the first to formalise and address two OOC failure modes in book-based role-playing LLM agents: *Factual Overreach* and *Stylistic Monotony*.
- We propose REVERIEMEM, a three-layer memory architecture that separates scene experience, visibility-gated factual knowledge, and situation-dependent expression for book-based character agents.
- We construct KBF-QA, a large-scale benchmark for Knowledge Boundary Fidelity, and show that REVERIEMEM surpasses the strongest prior character-agent system on both KBF-QA and pairwise narrative comparisons.

2 Related Work

2.1 Role Playing Agents

Role-playing agents test whether LLMs can sustain consistent characters in dialogue and narrative tasks. One line of work improves role consistency through persona descriptions, dialogue histories, source-text grounding, synthetic personas, and role playing training (Shao et al., 2023; Wang

et al., 2024a, 2025a), with literary settings further grounding agents in book scenes and character internals (Wang et al., 2025b). Benchmarks accompany this line and evaluate multi-turn dialogue, personality fidelity, and emotional fidelity (Tu et al., 2024; Wang et al., 2024b; Feng et al., 2025), while a recent survey separates fictional role playing from personalization (Tseng et al., 2024). These works mainly test whether responses match the target persona, but provide less evidence on whether the character should know the expressed fact.

A second line builds story worlds populated by interacting agents through sandbox simulations, multi-agent frameworks, and novel-to-simulation systems (Park et al., 2023; Chen et al., 2024b; Ran et al., 2025), alongside multi-agent narrative-generation systems for long-story writing, screenwriting via role-played authoring, and autonomous plot progression (Xia et al., 2025; Chen et al., 2024a; Zhao et al., 2026). Characters in these systems are typically represented through structured profiles, relation states, or shared narrative context, and evaluation focuses on plot coherence and interaction quality rather than on whether the speaking character can access a given fact in the source narrative. Whether a character should be allowed to access a given fact remains unaddressed.

2.2 RAG & Memory for Narrative Reasoning

Memory determines what evidence is available during generation. Retrieval-augmented generation grounds outputs in external documents (Lewis et al., 2020), and adaptive retrieval, hierarchical summarisation, long-term memory, and narrative memory systems improve evidence selection and multi-hop or long-story reasoning (Asai et al., 2024; Sarthi et al., 2024; Wang et al., 2026). For role playing over novels, however, the question is not only whether a retrieved fact is relevant but also whether the character can access it.

A complementary line addresses the boundary of what an agent should commit to. General LLM knowledge boundaries are catalogued in a recent survey (Li et al., 2025); for role playing agents specifically, boundary-aware training (Tang et al., 2024) and representation-level refusal editing (Liu et al., 2025) aim to suppress out-of-role answers. These works define the boundary by what the model knows or what the role constraint forbids; our setting defines it by what the character could have witnessed in the source narrative, making per-character visibility a core design axis.

3 REVERIEMEM

In this section we present REVERIEMEM, a perspective-bounded three-layer memory architecture for character role playing agents that addresses the failures identified in §1. REVERIEMEM therefore serves two design objectives, one for each failure:

- **Epistemic Objective.** The agent retrieves only facts the character could plausibly know in the source text, and abstains from claims that no such fact supports.
- **Expressive Objective.** The agent's prose matches the style the character exhibits in comparable situations in the source text, rather than a flattened trait label.

REVERIEMEM instantiates these objectives through three memory layers (§3.2) and a perspective-bounded inference pipeline (§3.4). Figure 2 sketches the architecture.

3.1 Overview

REVERIEMEM builds on two cognitive principles: the episodic–semantic distinction (McClelland et al., 1995) structures its memory layers (§3.2), and reconstructive, and scene-anchored recall (Conway and Pleydell-Pearce, 2000) structures its inference pipeline (§3.4). In addition, we add a per-character visibility constraint enforced during memory construction: when speaking as character c, retrieval is bounded to the visibility-allowed subset $\mathcal{F}_c$, and the agent refuses rather than searches harder when $\mathcal{F}_c$ contains no answer. The full system runs in three phases: an offline construction phase, *Source-to-Memory* (§3.3), followed at inference by *CLS Memory Collaborative Reasoning* and *Fusion Injection* (both in §3.4).

3.2 Bounded Three-Layer Memory

Following the episodic–semantic distinction(Figure 2), REVERIEMEM splits factual memory into an *episodic layer* (scene memories) and a *semantic layer* (discrete facts with per-character visibility), and adds a *personality layer* of discrete behavioural patterns derived from each character's conduct in the source text, selected at inference by emotion transitions (Mcadams and McLean, 2013).

Episodic Layer. The episodic layer plays the hippocampus-like role of fast, scene-level encoding in CLS: it stores what the character has lived

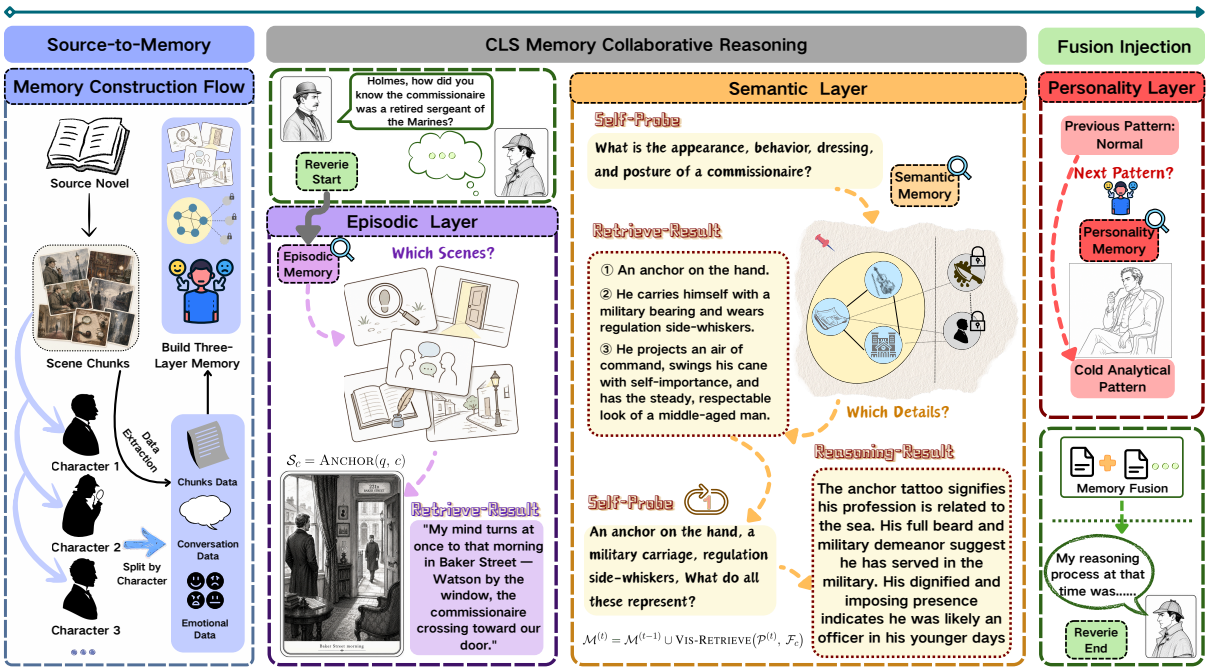

Figure 2: Overview of REVERIEMEM. Given a query on character c, the system runs three phases described in §3: 1) Source-to-Memory extracts per-scene text, dialogue, and emotion data via an LLM and constructs a perspective-bounded three-layer memory for each focus character; 2) CLS Memory Collaborative Reasoning anchors on the **Episodic Layer** scene memory $\mathcal{S}_c$, which both grounds and informs SELF-PROBE to iteratively expand the fact pool $\mathcal{M}^{(t)}$ from the visibility-bounded **Semantic Layer** subset $\mathcal{F}_c$, and finally synthesises a reasoning conclusion; 3) Fusion Injection selects a pattern m^* from the **Personality Layer** and injects it into the final memory fusion that integrates multiple memory components to produce the response from c's perspective.

through as a perspectival frame that anchors subsequent retrieval. For each focus character c, the layer maintains a corpus of first-person scene summaries, one per scene c was present for, recording in c's voice what c did, felt, perceived, and inferred about others. The corpus is indexed for similarity retrieval and serves as the substrate of the ANCHOR operation at inference time. Character-level scoping ensures every downstream retrieval starts from c's lived experience.

Semantic Layer. The semantic layer plays the neocortex-like role of slow, cross-episode integration: it consolidates the facts implied across scenes into a visibility-tagged knowledge graph. Each fact $f \in \mathcal{F}$ is a five-element *SPOCV* tuple $f = (s, p, o, \kappa, V)$: (s, p, o) is a standard subject–predicate–object triple; κ records the in-narrative cause when one is explicit, capturing that narrative reasoning hinges on *why* an event occurred, not only *that* it did; and $V \subseteq \mathcal{C}$ is the per-character visibility set defined below.

Visibility is granted along four routes — *direct experience*, *observation*, *organisational sharing*, and *common knowledge* — with assignment details deferred to §3.3. For each character c, the visibility-allowed subset $\mathcal{F}_c = \{f \in \mathcal{F} : c \in V(f)\}$ bounds retrieval when REVERIEMEM speaks as c: facts outside $\mathcal{F}_c$ cannot be surfaced by any amount of retrieval relevance.

Personality Layer. The personality layer addresses stylistic monotony: a character's behavioural style varies with situation, distinct in confrontation, in deference, and in retreat. We therefore represent each character not as a static profile of trait adjectives but as a finite set of discrete patterns abstracted from the character's conduct in the source text.

For each character c, the layer maintains two artefacts. The first is a set $\{m_1, \ldots, m_K\}$ of *personality patterns*, each pairing a short description with canonical excerpts from the source text (e.g. "Lestrade in cautious deference to Holmes"); the descriptions support selection and the excerpts serve as in-context style anchors. The second is a record of c's *emotion transitions*, derived from per-utterance dialogue annotations: each transition encodes an $(\text{emotion}, \text{intensity}, \text{trigger})$ triple together with the speaker's inferred intent. At infer-

ence time, the emotion record situates next-state prediction within c's observed trajectory, and the selected pattern provides the behavioural register that conditions memory fusion. Construction is described in §3.3.

3.3 Source-to-Memory

The three memory layers are constructed offline from the source novel by a scene-anchored procedure (Figure 2, left). The only human input is the focus-character list with canonical names and aliases; the remaining annotations are elicited from a language model.

Scene preparation. The novel is partitioned into scenes by an LLM-guided procedure that merges contiguous text spans falling within a single dramatic unit. Each scene is augmented with situational metadata (location, time, atmosphere) and a focus-character roster that records, for each focus character, whether they are present and active, present but silent, or only referenced. This scene-level decomposition is the shared substrate of the three layers.

Episodic summarisation. For each focus character c and each scene in which c is at least present, we elicit a first-person retrospective summary in c's voice. Summaries are constructed scene by scene, without access to subsequent narrative material, so that each summary reflects only the character's local experience at that point. The resulting per-character corpora are indexed by dense embedding for similarity-based retrieval at inference time.

Knowledge graph extraction. Following the two-stage extract-then-relate paradigm common in narrative-domain knowledge graphs (Gutiérrez et al., 2024; Wang et al., 2026), we first elicit per-scene entities spanning seven types, then induce (s, p, o, κ) relations over them, with κ recording the in-narrative cause when explicit. Entity aliases are unified across scenes, after which facts are deduplicated against the canonical entity table. The visibility component V is then assigned in two phases, completing the SPOCV tuple: a local phase grants direct experience to the fact's subject and object, observation to characters witnessing the scene of the fact, and common knowledge for world-level background; a propagation phase extends organisational sharing by tracing each fact through a character–organisation membership graph derived from the same scenes.

Personality clustering. We extract the dialogues attributed to each focus character on a per-scene basis and normalise speaker labels to canonical names. Each utterance is annotated with an (emotion, intensity, trigger) triple together with the speaker's inferred intent; concatenated chronologically, these annotations form the character's emotion-transition record. We then cluster the character's scenes by behavioural similarity into K personality patterns. For each pattern, we distil a one-line description (embedded for retrieval) and select canonical excerpts from the source dialogues that exemplify the pattern.

Per-layer artefacts are shown in Appendix B, with per-book layer sizes in Appendix A.

3.4 Inference Pipeline

REVERIEMEM's inference pipeline realises a perspective-bounded form of reconstructive recall. For a query q directed at character c, two phases run in coordination, with every retrieval restricted to the visibility-allowed subset $\mathcal{F}_c$: *CLS Memory Collaborative Reasoning* retrieves a scene-level frame from c's episodic memory and iteratively fills it with visibility-bounded facts; *Fusion Injection* closes inference with a memory fusion that incorporates a pattern m^* supplied by the parallel personality track (Figure 2, middle and right).

CLS Memory Collaborative Reasoning. The main reasoning chain interleaves two stages over $\mathcal{F}_c$.

Scene recall. The ANCHOR operation queries the episodic-layer corpus for c by similarity to q and returns the matched first-person scene summaries as the scene memory:

$$\mathcal{S}_c = \text{ANCHOR}(q,\ c). \tag{1}$$

This is not the response to q but the frame within which the next stage reconstructs detail, following the Self-Memory System's hierarchical retrieval cascade (Conway and Pleydell-Pearce, 2000). Mainstream narrative RAG queries a novel-level index directly, surfacing evidence outside the speaker's perspective. Anchoring on c's own scenes first restricts consideration to c's lived experience.

Scene-guided detail reconstruction. The scene memory $\mathcal{S}_c$ guides which factual details to recall, rather than treating $\mathcal{F}_c$ as a flat search space. Given q and $\mathcal{S}_c$, SELF-PROBE generates a set of focused sub-queries $\mathcal{P}^{(1)}$ targeting the factual gaps in the

scene memory, and the candidate-fact pool starts empty: $\mathcal{M}^{(0)} = \emptyset$. The reconstruction then loops over rounds $t = 1, \ldots, N$: each round issues the current probes against $\mathcal{F}_c$ via dense retrieval, expanding the pool

$$\mathcal{M}^{(t)} = \mathcal{M}^{(t-1)} \cup \textsc{Vis-Retrieve}\big(\mathcal{P}^{(t)},\ \mathcal{F}_c\big). \tag{2}$$

SUFFICIENT then checks whether $(\mathcal{S}_c, \mathcal{M}^{(t)})$ already supports a response from c's perspective; if so (or if $t = N$), the loop exits at round τ. Otherwise SELF-PROBE generates follow-up probes targeting the residual gap, and the loop continues. The terminal pool $\mathcal{M}^{(\tau)}$ is handed to the memory-fusion stage.

This retrieve-evaluate-deepen schedule resembles existing narrative-domain RAG; what is specific to a character agent is that detail reconstruction is anchored on the speaker's own scene memory rather than a novel-level summary, and that every retrieval is bounded to $\mathcal{F}_c$: facts c could not have witnessed are never admissible into the pool.

Fusion Injection. The personality track runs in parallel with the reasoning chain and supplies a pattern m^* to the closing memory fusion through three operations: EMO-GATE, EMO-TRANSITION, and PATTERN.

Pattern selection (parallel). EMO-GATE judges whether the current dialogue turn carries a salient emotional event; routine exchanges leave the emotional state in place. When the gate fires, EMO-TRANSITION reads c's emotion-transition record and proposes a new state, and PATTERN selects m^* from $\{m_1, \ldots, m_K\}$ by description-embedding match to the new state and dialogue context.

Perspective-bounded memory fusion. The closing memory-fusion stage operates over the terminal workspace $(\mathcal{S}_c, \mathcal{M}^{(\tau)})$, conditioned on the injected pattern m^*. The selected pattern's canonical excerpts S_{m^*} and description d_{m^*} are concatenated into the memory-fusion prompt as in-context style anchors, and MEMORY-FUSION produces a response from c's perspective:

$$r = \textsc{Memory-Fusion}(q,\ c,\ \mathcal{S}_c,\ \mathcal{M}^{(\tau)};\ m^*). \tag{3}$$

Because the workspace is bounded by $\mathcal{F}_c$ throughout, the only claims the fusion can ground are claims c could plausibly know; when $\mathcal{M}^{(\tau)}$ contains no relevant material, c acknowledges the absence in character. The memory-fusion stage and the personality layer thus jointly determine *whether* and *how* the agent conducts itself. Algorithm 1 (Appendix C) gives the procedure end-to-end, with the three LLM prompts driving it reproduced in Appendix F.

4 Knowledge Boundary Benchmark

We construct KBF-QA, a multiple-choice benchmark for Knowledge Boundary Fidelity (KBF). It tests whether a character agent answers facts visible from the character's narrative position and refuses facts outside it. The dataset spans eight novels and contains 4,386 multiple-choice questions in total (2,442 KRF questions and 1,944 KR questions). Each question is posed to a character c and offers five options: four candidate answers $\{A, B, C, D\}$ and a refusal option E ("I cannot answer this from my own knowledge"). The visibility component V (§3.2) determines the split: KRF items have the queried fact in $\mathcal{F}_c$ with gold answer in $\{A, B, C, D\}$; KR items have it in $\mathcal{F} \setminus \mathcal{F}_c$ with gold answer E. The two splits target complementary epistemic skills: recalling what the character could witness, and refusing what it could not. The KR split is what distinguishes this from closed-book QA: an LLM that "knows" the fact still fails by committing to it while speaking as c. Construction process, example items, and per-book composition are in Appendix D.

The KBF metric. KBF summarises performance on the two KBF-QA splits.Let $\text{KRF}_{\text{acc}}, \text{KR}_{\text{acc}}$ be the recall and refusal accuracies, with sample counts $n_{\text{KRF}}, n_{\text{KR}}$. We define

$$\text{KBF} = \frac{n_{\text{KRF}} + n_{\text{KR}}}{\dfrac{n_{\text{KRF}}}{\text{KRF}_{\text{acc}}} + \dfrac{n_{\text{KR}}}{\text{KR}_{\text{acc}}}}, \tag{4}$$

the sample-weighted harmonic mean. The harmonic mean (as in F1) penalises systems that collapse one split to gain the other; sample-weighting accounts for per-book variation in KRF:KR ratio. Each method's free-form response is deterministically matched to one of $\{A, B, C, D, E\}$.

5 Experiments

We evaluate REVERIEMEM on two tasks over the same eight novels. KBF-QA tests whether agents respect character knowledge boundaries, addressing factual overreach. Pairwise narrative generation tests whether agents preserve character voice and behaviour, addressing stylistic monotony.

5.1 Evaluation Metrics

Boundary task. On KBF-QA,We report three scores on the benchmark in §4: KRF accuracy on visible items, KR accuracy on invisible items, and their sample-weighted harmonic mean KBF (Eq. 4).

Narrative-generation task. We follow the five-dimension pairwise evaluation protocol of BOOKWORLD. An LLM judge compares two narratives generated from the same scripts and the same character cast. Four dimensions are shared across both modes: **Anthropomorphism** (An), **Character Fidelity** (CF), **Immersion & Setting** (IS), and **Writing Quality** (WQ). The fifth dimension depends on the generation mode: **Storyline Quality** (SQ) when a script is supplied (*with script*) and **Creativity** (Cr) when generation is open-ended (*without script*).

5.2 Experimental Setup

Boundary task. All methods are evaluated on the benchmark in §4 with `gpt-5-mini` under greedy decoding. REVERIEMEM is compared against six reference methods: *direct*, character-profile prompting without retrieval; four narrative-domain RAG methods, *naive RAG* (Lewis et al., 2020), *RAPTOR* (Sarthi et al., 2024), *HippoRAG* (Gutiérrez et al., 2024), and *ComoRAG* (Wang et al., 2026); and BOOKWORLD (Ran et al., 2025), a retrieval-augmented character-agent system.

Narrative-generation task. For each of the eight novels we adopt the scripts released by BOOKWORLD as experimental presets, yielding 226 scripts in total. REVERIEMEM is evaluated with five generator LLMs:`gpt-5-mini`, `gemini-3.1`, `qwen-3.5-plus`, `deepseek-3.7`, and `deepseek-v4pro` (Singh et al., 2026; Team, 2026; DeepSeek-AI et al., 2025), two closed modules and three open modules. Reference methods are *direct* (character-profile prompting), *HoLLMwood* (Chen et al., 2024a), and BOOKWORLD. Pairwise judging uses `gpt-4o`.This protocol follows Ran et al. (2025), who reported substantial agreement between model-based pairwise judgments and human evaluations using Cohen's Kappa coefficient(κ).

5.3 Main Results

Knowledge-boundary fidelity. Table 1 reports the boundary results. REVERIEMEM achieves the highest score on every KBF metric, with the gain coming jointly from recall and refusal. Upon closer examination, the RAG baselines split into two failure modes: *naive RAG*, *RAPTOR*, and *HippoRAG* retrieve any matching fact regardless of perspective and reach moderate recall but near-zero refusal; *ComoRAG* inverts this profile, refusing too aggressively at the cost of recall. *Direct* and BOOKWORLD fall in the middle on both axes but score low overall.

Method	KRF acc	KR acc	KBF
direct (no retrieval)	36.4	38.2	37.2
naive RAG	63.8	10.0	18.9
RAPTOR	66.4	8.7	16.8
HippoRAG	65.4	11.3	20.9
ComoRAG	33.0	45.9	37.7
BOOKWORLD	41.6	35.5	38.7
REVERIEMEM	**68.1**	**81.2**	**73.3**

Table 1: Results on the boundary benchmark: REVERIEMEM against six reference methods.

In-character narrative generation. Table 3 reports pairwise win rates of REVERIEMEM against each reference method across five generator LLMs. REVERIEMEM consistently wins the majority of comparisons under both modes, with the largest margins against *Direct* and *HoLLMwood* and a clear lead even against BOOKWORLD, the strongest baseline. The largest gains appear on WQ, SQ/Cr, and IS.

Comparison	KRF	KR	KBF
REVERIEMEM	68.1	81.2	**73.3**
w/o Episodic Layer	79.6	47.0	60.9
w/o Semantic Layer	10.8	93.1	17.8

Table 2: Ablation results of REVERIEMEM on the boundary benchmark.

5.4 Ablation Study

Boundary ablation. Table 2 reports boundary results when one factual layer is removed at a time. Removing the episodic layer pushes recall up but drops refusal: without scene-level gating, the agent over-commits to retrieved facts that the character could not have witnessed. Removing the semantic layer inverts the trade-off: with no fact pool to query, the agent collapses into refusal. Neither variant matches the full system's balanced KBF, confirming that the two factual layers are jointly necessary.

Model	Method	*with script*					*without script*				
		An.	**CF.**	**IS.**	**WQ.**	**SQ.**	**An.**	**CF.**	**IS.**	**WQ.**	**Cr.**
closed-source models											
gpt-5-mini	RM vs Direct	86.8	98.6	97.9	99.0	99.4	95.2	85.5	99.6	99.4	98.8
	vs HW	83.1	97.5	90.9	94.7	97.6	71.0	76.5	97.4	98.4	97.3
	vs BW	68.3	93.1	67.7	90.5	96.3	65.3	67.6	78.2	94.4	92.6
gemini-3.1	RM vs Direct	97.7	95.8	93.5	99.2	99.1	97.9	97.4	100.0	99.9	99.6
	vs HW	96.0	89.9	94.6	99.6	96.8	97.1	91.9	98.8	100.0	99.9
	vs BW	74.8	82.7	68.1	85.4	95.1	67.7	62.8	70.4	92.9	96.4
open-source models											
qwen-3.5-plus	RM vs Direct	91.9	93.1	96.3	97.8	96.2	97.5	88.3	94.7	99.5	99.7
	vs HW	89.0	86.0	95.3	97.0	88.3	89.5	78.5	100.0	98.9	99.9
	vs BW	63.0	82.6	71.2	87.2	90.0	64.1	61.4	72.7	89.5	97.3
deepseek-3.7	RM vs Direct	91.5	98.7	98.9	97.0	98.1	97.6	95.8	100.0	99.8	99.5
	vs HW	89.5	97.6	96.9	98.1	95.6	94.8	93.6	98.3	99.8	99.0
	vs BW	72.0	90.7	72.0	80.8	94.9	79.2	68.5	72.2	94.9	90.2
deepseek-v4pro	RM vs Direct	38.3	57.4	73.9	98.3	53.2	33.9	30.4	62.6	80.9	92.2
	vs HW	42.0	79.1	84.3	98.3	96.5	52.2	60.0	87.0	96.5	78.3
	vs BW	56.5	73.0	70.4	91.3	84.3	60.0	50.4	80.0	96.5	75.7

Table 3: Pairwise win rates (%) of REVERIEMEM against Direct, HoLLMwood (HW), and BOOKWORLD (BW) under two narrative-generation modes. Protocol: Appendix E.

Narrative ablation. Table 4 reports the full system's pairwise win rate against each ablated variant under both modes. The full system wins or ties on every dimension, with the largest margins on WQ, SQ, and Cr. Removing the personality layer leaves the agent factually competent but stylistically flat, with the biggest drop on WQ. Removing either factual layer most weakens IS, the dimension that depends on scene-anchored knowledge.

Comparison	An.	CF.	IS.	WQ.	SQ./Cr.
with script					
w vs w/o Episodic	61.7	66.7	83.3	86.7	75.0
w vs w/o Semantic	66.7	65.0	81.7	86.7	81.7
w vs w/o Personality	61.7	58.3	66.7	93.3	86.7
without script					
w vs w/o Episodic	53.3	51.7	68.3	91.7	76.7
w vs w/o Semantic	51.7	53.3	80.0	96.7	75.0
w vs w/o Personality	55.0	50.0	75.0	93.3	85.0

Table 4: Pairwise win rates (%) of REVERIEMEM against ablated variants that remove one memory layer at a time.

5.5 Discussion

The two tasks share a single architectural explanation. The boundary task isolates factual overreach in knowledge-access terms: unbounded RAG baselines retrieve any matching fact and answer aggressively, while direct prompting and BOOKWORLD under-retrieve; only when retrieval is gated by per-character visibility do recall and refusal improve together. The narrative-generation task shows that the same gating mechanism also produces stronger character-anchored stories, with the largest wins on the dimensions most affected by cross-character knowledge leaks. The ablations sharpen this picture: each layer is independently necessary, and together they address the two failure modes of factual overreach and stylistic monotony. Avoiding these failures in book-based role playing is therefore governed less by how much an agent retrieves than by what it is allowed to retrieve, which is what perspective-bounded memory enforces.

6 Conclusion

We propose REVERIEMEM, a three-layer memory architecture for character role playing over long narrative texts that addresses OOC failure along its two faces: Factual Overreach and Stylistic Monotony. Inspired by cognitive psychology, REVERIEMEM pairs a visibility-tagged knowledge graph with a perspective-bounded reasoning pipeline and a personality layer abstracted from in-source conduct. Across eight novels, REVERIEMEM surpasses the strongest prior character-agent system on both KBF-QA and pairwise narrative comparisons. These results indicate that avoiding both failures requires bounding what a character knows and grounding how it conducts itself in patterns extracted from the character's actual speech and behaviour.

Limitations

REVERIEMEM and KBF-QA target the cognitive side of perspective: what a character could plausibly know given their narrative position. We have not yet constructed targeted stress tests built around literary works in which perspective itself is the central narrative device, such as the multi-witness contradictions of *Rashomon*-style narratives, deliberately information-asymmetric detective plots, or novels carried by an unreliable narrator. Designing specialised evaluations that probe the architecture under these extreme perspective dynamics is left to future work.

Multi-character interaction. REVERIEMEM does not provide a dedicated mechanism for orchestrating multi-character interactions. In multi-agent scenes, the system relies on simple judgment over the current scene context to decide how each character responds, rather than a separate framework for inter-agent coordination, turn arrangement, or mutual reasoning over what other agents have inferred. Combining perspective-bounded memory with existing multi-agent character-orchestration frameworks is a natural future direction, where each agent reasons within its own visibility while a coordination layer manages turn-taking and inter-agent inference.

Ethics Statement

In this paper, we introduce REVERIEMEM, a perspective-bounded memory architecture for character role-playing agents, together with the KBF-QA benchmark. The development and use of both are guided by ethical principles to ensure responsible and beneficial outcomes.

Data sources. Our experiments cover eight novels: six English works in the public domain (*A Study in Scarlet*, *The Adventures of Tom Sawyer*, *Treasure Island*, *Around the World in Eighty Days*, *Dracula*, *Paradise Lost*) and two Chinese works (*Solaris*, in its Chinese translation, and *Ball Lightning*). We use these texts strictly for academic and non-commercial research purposes; no full text of any work is redistributed with the paper, and all references in the benchmark are presented as short, scene-anchored excerpts.

Use of Human Annotations. KBF-QA is constructed jointly by an LLM and human annotators who have read each of the eight novels in their entirety. The annotators are the authors and collaborators of this work; no external workers were recruited and no compensation was distributed, so concerns around minimum-wage standards, recruitment fairness, and demographic data collection do not apply. The annotation task is restricted to two operations: (i) verifying that each LLM-drafted multiple-choice item is factually grounded in the source text, and (ii) discarding items whose visibility assignment is borderline or ambiguous. No personal data, demographic information, or third-party content beyond the eight novels is collected, stored, or released at any stage. The annotators were informed of the purpose of the verification task and of how their annotations would be used in the released benchmark.

Risk. REVERIEMEM inherits the general risks of LLM-based systems: hallucination, bias propagated from pretraining data, and the possibility of generating misleading or harmful content. Our perspective-bounded design directly mitigates one specific failure mode — factual overreach across character knowledge boundaries — but does not address these broader base-model risks. Like other character-agent systems, REVERIEMEM could in principle be misused to produce content that misrepresents a real person; we intend it to be used only with fictional characters drawn from published literary works, and we discourage any application to the simulation of real, identifiable individuals. We encourage the responsible use of REVERIEMEM for academic, educational, and creative purposes, and discourage any application aimed at deception, harassment, or impersonation.

Use of AI Assistants. AI writing assistants (large language models) were used for grammar polishing, sentence-level rewording, and code-level debugging support during the preparation of this paper. All research design, experimental decisions, analysis, and final phrasing were made by the human authors. AI assistants were not used to generate scientific content, citations, or experimental results.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi,

Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024a. HoLLMwood: Unleashing the creativity of large language models in screenwriting via role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8075–8121, Miami, Florida, USA. Association for Computational Linguistics.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pages 20094–20136.

Martin A Conway and Christopher W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological review*, 107(2):261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. 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)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 59532–59569. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157, Vienna, Austria. Association for Computational Linguistics.

Wenhao Liu, Siyu An, Junru Lu, Muling Wu, Tianlong Li, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, Di Yin, Xing Sun, and Xuanjing Huang. 2025. Tell me what you don't know: Enhancing refusal capabilities of role-playing agents via representation space analysis and editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5983–6005, Vienna, Austria. Association for Computational Linguistics.

Dan Mcadams and Kate McLean. 2013. Narrative identity. *Current Directions in Psychological Science*, 22:233–238.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

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

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *Preprint*, arXiv:2504.14538.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *Preprint*, arXiv:2409.14710.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

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

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Juyuan Wang, Rongchen Zhao, Wei Wei, Yufeng Wang, Mo Yu, Jie Zhou, Jin Xu, and Liyan Xu. 2026. Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33557–33565.

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

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *Preprint*, arXiv:2501.15427.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025b. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Haotian Xia, Hao Peng, Yunjia Qi, Bin Xu, Juanzi Li, Hou Lei, and Xiaozhi Wang. 2025. Storywriter: A multi-agent framework for long story generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6559–6563, New York, NY, USA. Association for Computing Machinery.

Qi Zhao, Jian Li, Yizhang Jin, Dongqi Liu, Hang Ding, Qi Song, and Xiangyang Li. 2026. Rolearena: A multi-agent role-playing environment for long multi-turn dialogues with autonomous plot progression.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. *Narrative-Play: Interactive Narrative Understanding*, pages 82–93. EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations. Association for Computational Linguistics (ACL). Publisher Copyright: © 2024 Association for Computational Linguistics.; 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 ; Conference date: 17-03-2024 Through 22-03-2024.

A Three-Layer Memory Statistics

We build the three memory layers for eight novels (six English, two Chinese), covering 50 focus characters across 360 scenes. The Semantic Layer (§3.2) holds 12,468 SPOCV tuples and the Personality Layer (§3.2) holds 1,536 behavioural patterns; per-book breakdowns are in Table 6. Per-book KBF-QA composition and scores live in §D.

B Source-to-Memory Examples

Tables 7–9 show one real artefact from each of the three memory layers, all anchored to the same scene — Scene 1 of *A Study in Scarlet*, where Watson and Holmes first meet in the Bart's chemical laboratory. Each artefact is taken verbatim from the output of our Source-to-Memory pipeline (§3.3).

C Inference Algorithm

Algorithm 1 gives the full procedure of REVERIEMEM's inference pipeline (§3.4), end-to-end. The pipeline holds two invariants throughout: every retrieval is bounded to the visibility-allowed Semantic-Layer subset $\mathcal{F}_c$ (so facts character c could not have witnessed are never admissible into the pool), and reasoning is anchored on c's own scene memory $\mathcal{S}_c$ before any detail reconstruction is attempted. The scene-guided loop runs for at most N rounds and exits early once the workspace $(\mathcal{S}_c, \mathcal{M}^{(t)})$ already supports a response.

D KBF-QA Construction

We build KBF-QA jointly with human annotators who have read each novel: an LLM drafts candidate items grounded in the source text, and the annotators verify each item against the source. Figure 3 illustrates one example from each split (posed to Dr. John Seward), and Table 10 gives per-book composition together with per-book KBF scores for all seven Task 1 methods (§5).

E Pairwise Narrative Evaluation Protocol

We adopt the eight novels and their plot scripts directly from the released setup of Ran et al. (2025) without introducing additional books or scripts. Table 5 gives the per-book counts, totalling 226 scripts.

For each generator LLM, we collect head-to-head pairwise comparisons of REVERIEMEM *vs.* BOOKWORLD on all 226 plot scripts, with AB+BA counterbalancing per item. To obtain rates against *direct* and *HoLLMwood* (Table 3), we combine these measurements with the BW-vs-{Direct, HW} pairwise rates reported in Ran et al. (2025) via a Bradley–Terry transitivity step, fitting one latent strength per method.

Book	Language	# Scripts
A Study in Scarlet	English	12
Tom Sawyer	English	36
Treasure Island	English	34
Around the World in 80 Days	English	37
Dracula	English	42
Paradise Lost	English	25
Solaris	Chinese	13
Ball Lightning	Chinese	27
Total	—	**226**

Table 5: Per-book number of plot scripts in Task 2, adopted from Ran et al. (2025).

F Inference Prompts

Algorithm 1 invokes three LLM prompts, reproduced in Tables 11–13: SELF-PROBE (lines 2 and 7) generates the next retrieval anchor from the scene memory; PATTERN (line 10) selects a behavioural pattern from the Personality Layer; and MEMORY-FUSION (line 15) writes the final response under the visibility-bounded fact pool. The exit condition SUFFICIENT (line 5) is decided within the same SELF-PROBE call from the second round on, rather than as a separate prompt. ANCHOR and VIS-RETRIEVE are dense-retrieval operations that do not call an LLM.

Book	Language	# Characters	# Scenes	# SPOCV facts	# Patterns
A Study in Scarlet	English	5	25	904	197
Tom Sawyer	English	3	92	2,298	170
Treasure Island	English	3	46	1,517	126
Around the World in 80 Days	English	10	41	1,836	175
Dracula	English	9	33	973	166
Paradise Lost	English	7	38	1,926	167
Solaris	Chinese	4	50	1,657	134
Ball Lightning	Chinese	9	35	1,357	401
Total	—	**50**	**360**	**12,468**	**1,536**

Table 6: Per-book size of the three memory layers.

Episodic Layer · *first-person scene summary for Dr. Watson, Scene 1*

Scene	1 — Chemical laboratory at St. Bartholomew's Hospital, London (1878–79)
Atmosphere	Intellectually intense, slightly eccentric; contrasts Watson's weariness with Holmes's energetic obsession
Scene understanding	*"I had recently returned to England after being wounded during the Second Anglo-Afghan War, and was seeking affordable lodgings. By chance I met Stamford at the Criterion Bar; over lunch he mentioned a man named Sherlock Holmes who was also looking for someone to share rooms… At Bart's, Holmes was ecstatic over a reagent that detects haemoglobin, and demonstrated it by pricking his own finger. Stamford introduced us, and Holmes startled me by correctly deducing I had been in Afghanistan."*
My actions	*"I asked Stamford what he'd been doing; pressed him on Holmes's character when he seemed evasive; agreed to visit the lab; expressed polite interest in the demonstration; was taken aback by Holmes's abrupt observation about Afghanistan."*
My emotions	*"Initial surprise and warmth at a familiar face; cautious optimism about the lodgings; mild suspicion at Stamford's hesitation; strong surprise at Holmes's Afghanistan deduction; restrained admiration mixed with unease."*

Table 7: Episodic Layer artefact: Watson's first-person summary of Scene 1 of *A Study in Scarlet*.

Semantic Layer · *six SPOCV tuples from Scene 1*

s	p	o	κ **(cause)**	V **(visibility)**
Holmes	demonstrated	haemoglobin test with his own blood	—	{W, H, S}, org={hospital}
Murray	threw	Watson across a pack-horse	to bring him safely to British lines	{W}
Holmes	references	Von Bischoff, Mason of Bradford, Muller… cases	—	{W, H, S}, org={Scotland Yard}
Watson	gravitated to	London	had neither kith nor kin in England	{W, H}
Watson	took degree of	Doctor of Medicine	—	{W, H}
Second Anglo-Afghan War	took place in	1878–1880	—	COMMON

Table 8: Semantic Layer artefacts: six SPOCV tuples from Scene 1, spanning the four visibility routes. W = Watson; H = Holmes; S = Stamford; `org` = organisational visibility; COMMON = common knowledge.

Personality Layer · *two of fifty personality patterns for Sherlock Holmes*			
Pattern	**Scene**	**Description**	**Canonical excerpt**
m_1	1	*Escalating, unrestrained joy — shouting, running, chuckling; voice bright and urgent, gestures expansive and physically energetic; tone exuberant and self-absorbed.*	"I've found it! I've found it!"
m_{49}	25	*Calm detachment; flat intonation and minimal physical emphasis; grants permission and redirects attention without warmth or irritation.*	"You may do what you like, Doctor."

Table 9: Personality Layer artefacts: two of fifty patterns for Sherlock Holmes, drawn from opposite ends of *A Study in Scarlet*.

Algorithm 1 REVERIEMEM inference for query q on character c.

$\mathcal{S}_c \leftarrow$ ANCHOR$(q,\ c)$ ▷ scene recall
$\mathcal{P}^{(1)} \leftarrow$ SELF-PROBE$(q,\ \mathcal{S}_c)$; $\mathcal{M}^{(0)} \leftarrow \emptyset$
for $t = 1$ **to** N **do** ▷ scene-guided detail reconstruction
  $\mathcal{M}^{(t)} \leftarrow \mathcal{M}^{(t-1)} \cup$ VIS-RETRIEVE$(\mathcal{P}^{(t)},\ \mathcal{F}_c)$
  $\sigma^{(t)} \leftarrow$ SUFFICIENT$(q,\ \mathcal{S}_c,\ \mathcal{M}^{(t)})$
  if $\sigma^{(t)} =$ true **or** $t = N$ **then break**
  end if
  $\mathcal{P}^{(t+1)} \leftarrow$ SELF-PROBE$(q,\ \mathcal{S}_c,\ \mathcal{P}^{(1:t)},\ \mathcal{M}^{(t)})$
end for
$\tau \leftarrow t$ ▷ exit round
if EMO-GATE$(q,\ c) =$ FIRE **then** ▷ parallel personality track
  $e^* \leftarrow$ EMO-TRANSITION(c); $m^* \leftarrow$ PATTERN$(c,\ e^*)$
else
  $m^* \leftarrow$ previous pattern of c ▷ emotional inertia
end if
return $r \leftarrow$ MEMORY-FUSION$(q,\ c,\ \mathcal{S}_c,\ \mathcal{M}^{(\tau)};\ m^*)$ ▷ perspective-bounded memory fusion

Book	**Language**	$n_{\text{KRF}}:n_{\text{KR}}$	**direct**	**BOOKWORLD**	**naive RAG**	**RAPTOR**	**HippoRAG**	**ComoRAG**	**REVERIEMEM**
A Study in Scarlet	English	344:256	35.2	30.3	19.0	20.7	30.8	37.1	**68.8**
Tom Sawyer	English	142:118	20.5	22.3	11.5	8.5	13.2	32.0	**63.5**
Treasure Island	English	161:90	28.3	32.5	10.8	10.7	37.3	28.7	**65.1**
Around the World in 80 Days	English	485:342	38.1	36.5	11.0	12.2	11.6	36.3	**73.7**
Dracula	English	502:419	29.1	39.9	18.8	18.5	22.4	35.1	**75.5**
Paradise Lost	English	62:66	29.1	13.0	5.6	0.0	2.9	27.9	**73.0**
Solaris	Chinese	202:97	39.1	32.0	28.5	24.9	34.5	42.2	**71.0**
Ball Lightning	Chinese	544:556	32.7	35.0	24.5	18.7	16.9	28.4	**78.0**

Table 10: Per-book composition and KBF scores for the Task 1 boundary benchmark (KBF-QA). Bold indicates the highest value in each row.

Source: Dracula *Asked of:* Dr. John Seward

KRf · *fact visible to Seward* ($f \in \mathcal{F}_{Seward}$)

Inquirer: "Dr. Seward, you saw your patient slip away one night — where did he go?"

Options:
(A) the ruined churchyard
(B) the quay at Whitby
(C) a deserted house
(D) an empty lodging in Munich
(E) "I cannot answer this from my own knowledge."

GOLD (**C**) BOOKWORLD (A) ✗ REVERIEMEM (C) ✓

Why: Seward, as the asylum physician, personally witnesses Renfield escape one night and follows him to the deserted house next door (Carfax). BOOKWORLD defaults to a vampire-trope answer ("ruined churchyard") without anchoring on what Seward actually saw.

KR · *fact not visible to Seward* ($f \in \mathcal{F} \setminus \mathcal{F}_{Seward}$)

Inquirer: "Doctor Seward, whom did your dear friend mimic in private before her illness worsened?"

Options:
(A) a Saxon child
(B) Mr. Swales
(C) a Magyar woman
(D) Mina Harker
(E) "I cannot answer this from my own knowledge."

GOLD (**E**) BOOKWORLD (B) ✗ REVERIEMEM (E) ✓

Why: Lucy's private imitation of Mina took place between the two friends alone, with no medical witness present. Seward attends Lucy as her physician but is not part of these intimate exchanges; the fact is visible only to Lucy and Mina.

Figure 3: Two example KBF-QA items from *Dracula*, posed to Dr. John Seward.

Self-Probe Prompt

You are {character}.

Current situation
{current_context}

Recalled scenes from your own first-person perspective (Episodic Layer)
{scene_context}

Probes already issued in this query (when present)
{prior_probes}

Facts already retrieved (when present)
{retrieved_facts}

Task
Step 1 — Read the scenes above and use them to sharpen what the current situation is really asking. In one short sentence, in your own voice, state your refined understanding of what is being asked, anchored to specific names, events, or relations that the scenes literally surfaced.

Step 2 — Name the single concrete **anchor** your inner attention now rests on as the next thing to look up. The anchor must be a specific name, object, event, relation, time, or place — *not* a vague theme, and *not* a full question. Write it as a short noun phrase. It must **not** semantically overlap with any probe already issued, and it must be a new angle that the scenes (or already-retrieved facts) make worth pulling on.

Output format
Return a single JSON object, double quotes only, no markdown:
{
"refined_query": "*one short sentence sharpening what is being asked*",
"self_probe": "*one concrete noun-phrase anchor for the next retrieval round*"
}

Table 11: SELF-PROBE prompt (Algorithm 1, lines 2 and 7).

Pattern Selection Prompt

You are {character}.

Current situation
{current_context}

Your current emotional state
{current_emotion}

Candidate behavioural patterns (Personality Layer)
Below are {num_modes} behavioural patterns distilled from your past behaviour — recurring patterns in how you speak and act under different emotional arcs. Each pattern has an identifier and a short description:

{modes_list}

Task
Which pattern best fits how you would naturally respond to the current situation, given your current emotional state? Pick **one**. If no pattern fits perfectly, pick the closest.

Output format
Return a single JSON object, double quotes only, no markdown:
{
"chosen_mode_id": "*the identifier of the selected pattern*",
"brief_reason": "*one sentence explaining why this pattern fits*"
}

Table 12: PATTERN prompt (Algorithm 1, line 10).

Memory-Fusion Prompt

You are {character}.

Current situation
{current_context}

Scenes you have lived through (Episodic Layer, first-person)
{scene_context}

Specific facts your reasoning surfaced (Semantic Layer, visibility-bounded)
{retrieved_facts}

Behavioural pattern conditioning your voice this turn (Personality Layer)
- Description: {pattern_description}
- Canonical excerpts: {pattern_samples}

Task
Sit with everything above end to end and respond to the current situation as {character} would. In 1–3 first-person sentences, meet the moment with what you actually recall — name who, what, where, or when, with the certainty your recollection affords. If a small step is needed to go from what the facts literally show to what the situation asks, take that step visibly in your own voice.

Constraints
1. Treat the scenes and the specific facts as **authoritative**; do not add detail they do not carry.
2. Restrict yourself to what {character} could plausibly know from these scenes and facts. If the facts do not contain the answer to the situation, acknowledge the gap honestly in your own voice rather than fabricating a response.
3. Let the behavioural pattern shape your cadence, tone, and gesture: imitate the canonical excerpts' style, not their literal content.
4. Stay within {character}'s perspective; do not draw on knowledge of stories, books, or events outside this character's lived experience.

Output format
Output the response directly — first person, 1–3 sentences. No JSON, no preamble, no header.

Table 13: MEMORY-FUSION prompt (Algorithm 1, line 15).