

ReaderScope: From Offline Reader-Response Calibration to Online Controlled Story Generation

Feng Zhang¹ Rong Pan^{1,2*}

¹School of Computer Science and Engineering,
Sun Yat-sen University, Guangzhou, China

²Guangdong Province Key Laboratory of
Applied Mathematics and Artificial Intelligence, China
zhangf378@mail2.sysu.edu.cn panr@sysu.edu.cn

Abstract

Long-form story generation should maintain narrative coherence while shaping how readers experience a story as it unfolds. Existing LLM-based generators typically optimize plot structure, consistency, or local text quality, without explicitly modeling reader-response dynamics as generation-time control targets. We introduce **ReaderScope**¹, a reader-response-guided framework that calibrates simulated response trajectories from reference novels across affective, suspenseful, immersive, and reflective dimensions, and converts them into explicit control signals for iterative story revision. Experiments against representative long-form story generation baselines show that ReaderScope improves reader-facing story quality under both automatic and human pairwise evaluation, while ablation and trajectory-fidelity analyses confirm that the response signals are useful for generation and can be reliably controlled.

1 Introduction

A long-form story can be coherent and still fail as a reading experience: it may follow a plausible outline, maintain character consistency, and preserve event causality, yet leave readers emotionally flat, unsurprised, weakly immersed, or without reflection. This exposes a control gap: existing systems increasingly control what happens in a story, but rarely how the reader’s experience evolves while the story unfolds.

Recent LLM-based story generators improve long-form coherence through planning, memory, retrieval, revision, and multi-agent simulation (Yang et al., 2023, 2022; Lei et al., 2024; Bae and Kim, 2024; Chen et al., 2026; Ran et al., 2025). However, their control signals are mostly

author-facing or text-facing, focusing on events, states, memories, or revision criteria, while reader-facing effects are typically assessed only after generation. This is limiting because reading unfolds over time: readers experience affective shifts, uncertainty, immersion, and reflection throughout a narrative (Mar et al., 2011; Kneepkens and Zwaan, 1995; Gerrig and Bernardo, 1994; Kuijpers et al., 2014; Loi et al., 2023). Thus, reader-aware generation requires control signals that capture trajectories of reader response, not only the content of the story.

Turning reader-response concepts into generation-time controls is non-trivial: dense human feedback is impractical, while a single scalar quality score discards the temporal and multidimensional structure of reading. We address this problem with **ReaderScope**, a reader-response-guided framework that moves simulated reader response from post-hoc evaluation to generation-time control. Instead of steering narratives only with author-side plans or text-facing revision criteria, ReaderScope uses expected trajectories of affect, suspense, immersion, and reflection to guide how the story should be experienced over time.

Experiments show that ReaderScope improves long-form story quality over representative baselines. Trajectory-fidelity and ablation analyses further indicate that simulated reader-response trajectories can function as practical control signals, not merely as evaluation proxies.

Our contributions are threefold:

- We formulate reader-response-controlled long-form story generation, shifting the control objective from author-side narrative content to the temporal trajectory of reader experience.
- We introduce **ReaderScope**, a calibration-to-control framework that derives multidimensional reader-response targets offline and uses

*Corresponding author.

¹Code and implementation are available at <https://github.com/FengZhang758/ReaderScope>.

them to guide online story generation through reader-feedback-based revision.

- We conduct experiments and analyses showing that ReaderScope improves long-form story quality, achieves high trajectory-control fidelity, and benefits from complementary emotion, entropy, immersion, and insight feedback.

2 Related Work

2.1 LLM-based Story Generation

Recent LLM-based story generation has moved beyond single-pass prompting toward structured generation pipelines. Planning-and-writing frameworks decompose long-form generation into outline construction, event planning, drafting, and revision (Yang et al., 2023, 2022; Lei et al., 2024), often using multiple agents to coordinate global plot structure and local narrative realization (Huot et al., 2025; Jiang et al., 2026; Han et al., 2024; Wang et al., 2024). To improve coherence over long contexts, prior work further augments generation with memory, retrieval, knowledge graphs, or critic-based revision (Yi et al., 2026; Pei et al., 2024; Lin et al., 2025; Wan et al., 2025; Bae and Kim, 2024). Another line of work uses LLM-based sandbox or character simulations to produce bottom-up narrative materials from emergent agent interactions (Chen et al., 2026; Ran et al., 2025). In contrast to prior author-centric controls based on plans, memories, agents, or critiques, ReaderScope guides generation with model-based proxies of multidimensional reader response.

2.2 Control Signals for Story Generation

Controllable story generation has explored several types of control signals. One line of work uses author-side structural specifications, such as outlines, plot plans, premises, or predefined endings, to improve long-range coherence and narrative organization (Yang et al., 2023; Wang and Kreminski, 2024; Lee et al., 2025; Ma et al., 2024; Brei et al., 2024). Another line steers plot development or story-world evolution through suspense-oriented continuations or structured generation mechanisms (Xie and Riedl, 2024; Wang et al., 2024). Recent work also considers more reader-facing attributes, such as emotional arcs or narrative pacing (Wen et al., 2025; Wang et al., 2023). Overall, prior work provides effective controls for narrative structure and plot development, but rarely operationalizes multidimensional reader

response as generation-time guidance.

2.3 From Reader Response to Narrative Control

Narrative reading involves multiple layers of experience beyond a single notion of story quality. Prior work has emphasized affective responses to events and characters (Mar et al., 2011; Kneepkens and Zwaan, 1995), suspense and belief revision around unresolved narrative uncertainty (Gerrig and Bernardo, 1994; Wilmot and Keller, 2020), transportation and absorption in story worlds (Kuijpers et al., 2014), and higher-level interpretation or reflective meaning-making after literary reading (Loi et al., 2023). Together, these perspectives motivate a four-dimensional operationalization of simulated reader response: *emotion*, *entropy*, *immersion*, and *insight*.

3 Methodology

3.1 Reader-Response Signal Modeling

Reader Modeling. ReaderScope models reading as a temporally evolving response process. For each simulated reader r and narrative unit t , the reader agent produces a response vector conditioned on the current story unit, a reader profile, and an accumulated memory state. The profile specifies stable reading preferences and interpretive tendencies, while the memory state follows (Sumers et al., 2024) and contains working memory, episodic memory, and semantic memory for recent context, plot events, prior reactions, and global story knowledge. At each unit, the reader retrieves relevant memories before producing responses and updates the memory base after reading. This retrieve–respond–update cycle allows ReaderScope to simulate reader responses that depend on both stable reader traits and accumulated narrative context.

Quantified Reader-Response Signals. Based on the simulated reading process, ReaderScope computes four response signals introduced in Section 2.3: emotion, entropy, immersion, and insight. For reader r at narrative unit t , we denote these model-based proxy scores as $A_{r,t}$, $E_{r,t}$, $I_{r,t}$, and $G_{r,t}$, respectively; computed over units, they form trajectories for offline calibration and online control.

Emotion. We model *Emotion* as the local affective valence induced by each narrative unit, rather

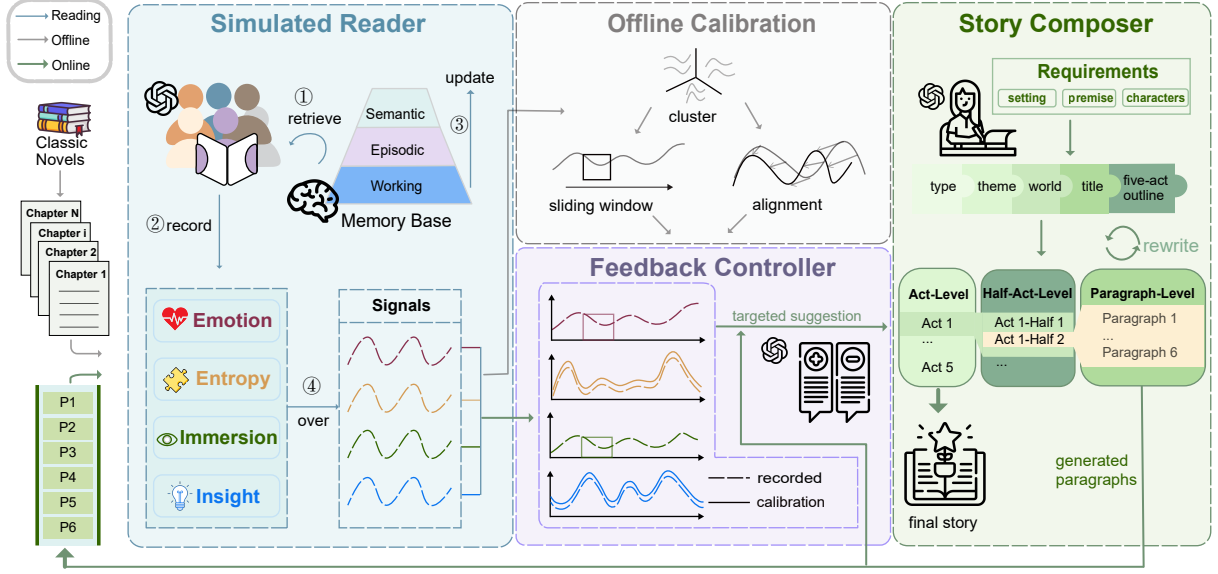


Figure 1: Overview of ReaderScope. Blue arrows denote the reader-response collection process, where simulated readers read story units and record response signals. Gray arrows indicate offline calibration, which analyzes the collected trajectories to derive control targets. Green arrows show the online generation loop, where the story composer generates paragraphs and the feedback controller provides targeted revision guidance.

than as an accumulated mood state or a global sentiment label. For reader r and unit t , the reader model identifies $K_{r,t}$ emotional turning points and assigns each turning point a valence score $e_{r,t,k} \in [-1, 1]$. The unit-level emotion score is computed as:

$$A_{r,t} = \frac{1}{K_{r,t}} \sum_{k=1}^{K_{r,t}} e_{r,t,k}. \quad (1)$$

This definition captures local affective fluctuation while keeping emotion scores comparable across narrative units.

Entropy. We model *Entropy* as a generation-time signal of suspense-related reader uncertainty, combining uncertainty introduced by new story questions with the impact of resolving previous ones. After unit t , reader r generates a set $\mathcal{Q}_{r,t}^{\text{new}}$ of new questions, each with candidate answers $\mathcal{Y}_q$ and belief probabilities $p_{r,t}(y | q)$. For questions with at least two candidates, we compute the normalized Shannon entropy and average it over the newly introduced questions:

$$h_{r,t}(q) = \frac{-\sum_{y \in \mathcal{Y}_q} p_{r,t}(y | q) \log_2 p_{r,t}(y | q)}{\log_2 |\mathcal{Y}_q|}, \quad (2)$$

$$H_{r,t}^{\text{new}} = \frac{1}{|\mathcal{Q}_{r,t}^{\text{new}}|} \sum_{q \in \mathcal{Q}_{r,t}^{\text{new}}} h_{r,t}(q). \quad (3)$$

The reader also updates relevant questions raised earlier. A question is treated as resolved when one updated answer probability exceeds $\tau = 0.7$. Resolved questions produce a resolution-impact score $S_{r,t}^{\text{res}}$ that favors answers that were unexpected under the reader’s prior beliefs yet easy to understand. The final score is:

$$\mathcal{E}_{r,t} = H_{r,t}^{\text{new}} + S_{r,t}^{\text{res}}. \quad (4)$$

Thus, Entropy increases when a unit either opens uncertain narrative possibilities or provides an unexpected but comprehensible resolution. Full question formats, memory updates, and resolution weighting are provided in Appendix E. During story-level evaluation, the corresponding reader-facing dimension is reported as *Suspense*.

Immersion. We operationalize *Immersion* using the Story World Absorption Scale (SWAS), which characterizes absorbed reading through four dimensions: attention, transportation, emotional engagement, and mental imagery (Kuijpers et al., 2014). For each narrative unit t , simulated reader r rates 18 adapted SWAS statements on a seven-point agreement scale from -3 (*totally disagree*) to $+3$ (*totally agree*). Given item rating $s_{r,t,j}$, we compute immersion as the normalized mean:

$$I_{r,t} = \frac{1}{18} \sum_{j=1}^{18} \frac{s_{r,t,j} + 3}{6}. \quad (5)$$

Insight. We operationalize *Insight* through reflective responses and value-profile change. For reader r at unit t , the reader model rates nine adapted Personal Insight items and five Non-Personal Insight items on a 7-point scale. Let $q_{r,t,j} \in \{1, \dots, 7\}$ denote the score of item j . The normalized reflective-insight score is:

$$Q_{r,t} = \frac{1}{14} \sum_{j=1}^{14} \frac{q_{r,t,j} - 1}{6}. \quad (6)$$

We additionally rate the activation of ten Schwartz value dimensions (Schwartz, 2012) before reading, $v_{r,0,k}$, and after unit t , $v_{r,t,k}$. Their normalized mean absolute shift is:

$$\Delta V_{r,t} = \frac{1}{10} \sum_{k=1}^{10} \frac{|v_{r,t,k} - v_{r,0,k}|}{6}. \quad (7)$$

The final insight score is:

$$G_{r,t} = \frac{1}{2} (Q_{r,t} + \Delta V_{r,t}). \quad (8)$$

Computed at each narrative unit, these four signals form reader-response trajectories used for offline calibration and online generation control.

3.2 Offline Calibration of Control Targets

ReaderScope calibrates control targets from reader-response trajectories collected on 20 length-comparable reference stories (Appendix D). Each story is converted into a chapter-based format, where each chapter serves as one narrative unit. For each unit, the reader agent reads the chapter conditioned on its memory state, produces four-dimensional response scores, and updates its memory, yielding a chapter-level trajectory for each reader-story pair. We sample 25 reader profiles from a pool of 100 and let each reader read each reference story once with gpt-4o-mini as the reader-agent backbone, producing 500 trajectories in total. Since the stories have different chapter counts, all trajectories are interpolated to a uniform length of $N = 60$ for cross-story comparison and control-target calibration.

Trajectory clustering as a diagnostic. To examine whether reader-response trajectories exhibit stable grouping patterns, we cluster them using de-biased Soft-DTW distances with k -medoids after optional normalization, interpolation, and smoothing, select the number of clusters by silhouette score, and compare with spectral clustering as a robustness check. Since low silhouette scores indicate weak cluster separation, we treat clustering

Table 1: Maximum silhouette scores under different clustering methods and preprocessing settings.

Dim.	K-medoids				Spectral			
	raw	z	r-int.	z-int.	raw	z	r-int.	z-int.
Emotion	0.60	0.41	0.55	0.46	0.51	0.47	0.47	0.45
Entropy	0.46	0.16	0.36	0.15	0.50	0.18	0.41	0.18
Immersion	0.62	0.40	0.83	0.38	0.61	0.39	0.81	0.35
Insight	0.52	0.16	0.68	0.16	0.50	0.20	0.62	0.21
Mean	0.55	0.29	0.61	0.29	0.53	0.31	0.58	0.30

Note. *r-int.* and *z-int.* denote raw-interpolated and z-score-interpolated trajectories, respectively. Boldface highlights the normalized settings where clustering evidence remains weak.

here as a diagnostic rather than as a required modeling step.

Table 1 shows a clear dimension-specific pattern. Emotion and immersion maintain non-trivial cluster structure across preprocessing settings, and their best clusters are largely aligned with story identity (Appendix G and Appendix H). This suggests that these two responses are strongly shaped by story-specific affective pacing and scene construction. In contrast, entropy and insight become weakly clusterable after z -normalization, and their best cluster numbers collapse to small values (Appendix F). We therefore do not use a single clustering strategy for all dimensions. Instead, ReaderScope constructs two types of control targets: local window constraints for emotion and immersion, and aligned statistical envelopes for entropy and insight.

Window targets for emotion and immersion.

For emotion and immersion, ReaderScope derives local sliding-window targets rather than fixed global prototypes. Since emotional trajectories are often story-specific, we estimate the maximum sliding-window span over which emotional change can remain non-reversing, and trigger revision when the generated trajectory stays monotonic beyond this calibrated span. For immersion, we use the same sliding-window mechanism asymmetrically: revision is triggered when the average reader-panel immersion decreases continuously beyond the maximum tolerated span estimated from the reference trajectories. Appendix I details the extraction procedure; the calibrated window ratios are approximately 0.076 for emotion reversal and 0.082 for immersion anti-decline.

Aligned trajectory envelopes for entropy and insight. Despite weak story-specific clustering,

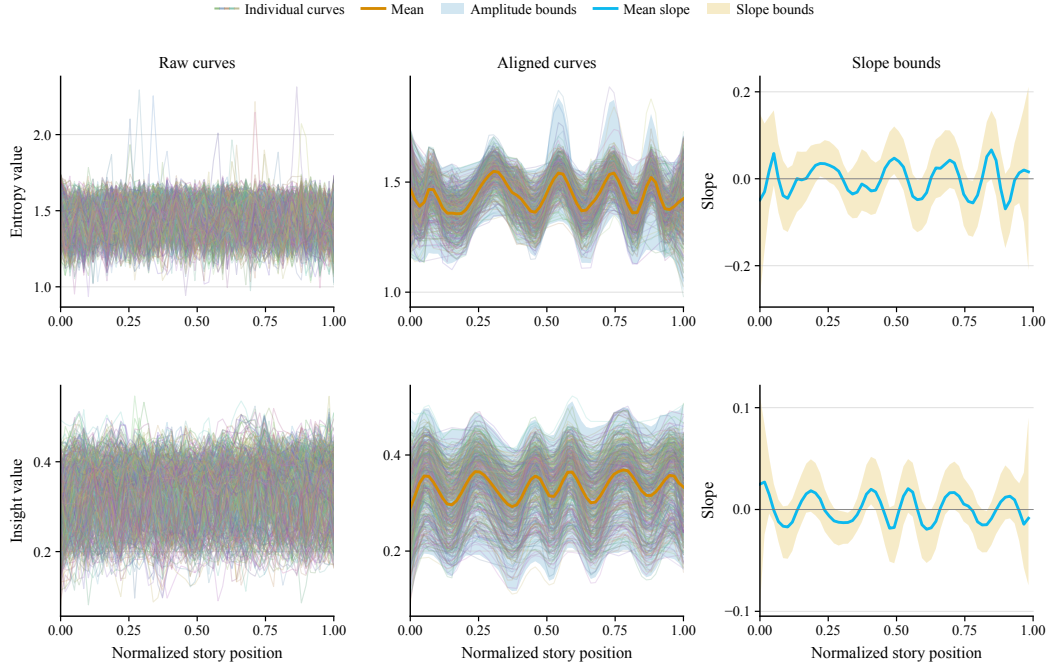


Figure 2: Offline calibration of entropy and insight control targets, including raw trajectories, aligned curves, and slope information.

entropy and insight reveal recurring patterns after temporal alignment. ReaderScope uses Fisher–Rao elastic registration of the raw-interpolated trajectories to estimate a mean curve and amplitude and slope envelopes (Figure 2). The amplitude envelope gives positionwise value bounds, whereas the slope envelope bounds adjacent-point differences around the mean trend. Bandwidth is the distance between an envelope’s upper and lower bounds and is not a confidence interval. Online, the amplitude range is combined with the slope-implied range from the preceding response. Large or repeated deviations trigger revision, including in-band but locally implausible changes.

The offline stage yields emotion reversal and immersion anti-decline windows, plus aligned amplitude and slope envelopes for entropy and insight. The controller uses them to monitor response levels and local changes and guide revisions (Appendix J).

3.3 Online Reader-Loop Controlled Story Generation

After offline calibration, ReaderScope uses the resulting control targets in an online reader-feedback loop. As shown in Figure 1, the loop contains three components: a *Story Composer*, a *Reader-Agent Panel*, and a *Feedback Controller*. The Story

Composer drafts narrative units, the Reader-Agent Panel estimates simulated reader-response trajectories for the generated text, and the Feedback Controller compares these trajectories with the calibrated targets to produce targeted revision instructions.

Planning and generation structure. Given a story requirement, the Story Composer first builds a global plan, including the story type, theme, background, character arcs, title, and five-act outline. ReaderScope then organizes each story into 60 paragraph-level generation units: five acts, two half-act objectives per act, and six paragraphs per half-act. Because the targets are calibrated from chapter-level responses in reference stories, these units serve as normalized story-position controls rather than direct chapter-to-paragraph supervision. For each half-act, the composer generates a six-paragraph unit conditioned on the global plan, the current objective, and summaries of previously accepted content.

Reader-feedback revision. After each unit is generated, the Reader-Agent Panel reads it paragraph by paragraph and produces group-level estimates of the four response dimensions. The Feedback Controller checks these estimates against the dimension-specific control targets while also con-

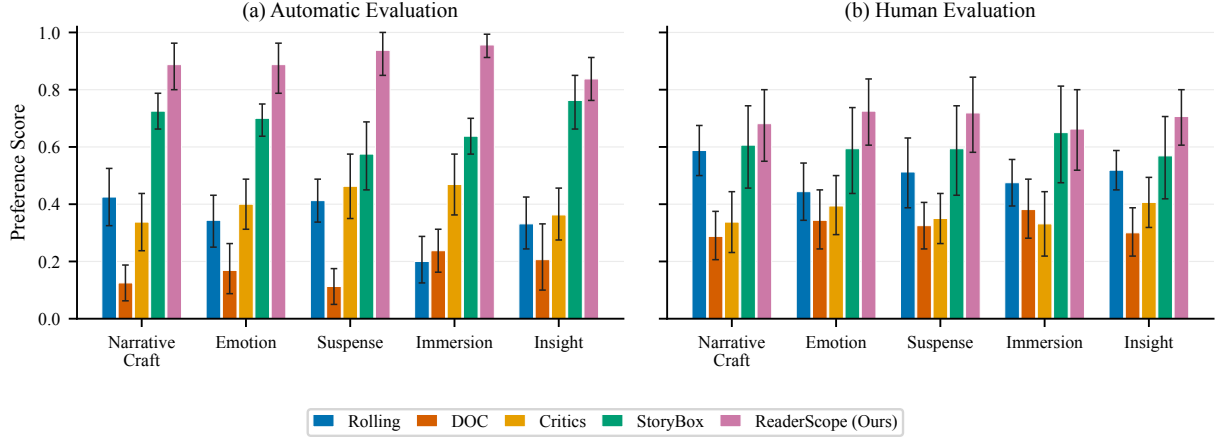


Figure 3: Normalized pairwise-preference scores under (a) automatic and (b) human evaluation. Wins, ties, and losses contribute 1, 0.5, and 0 points, respectively. Error bars show pointwise 95% percentile confidence intervals from 100,000 cluster-bootstrap resamples at the story-requirement level. Higher scores indicate stronger preference.

sidering the generated text. If the unit satisfies the constraints, it is accepted and appended to the story; otherwise, the controller returns concrete revision feedback for the problematic paragraph range, such as adjusting uncertainty, emotional intensity, scene grounding, causal clarity, or reflective emphasis. The Story Composer then rewrites the same unit under the original half-act objective and prior-story context. This read-evaluate-revise process is repeated within a fixed rewrite budget, and ReaderScope proceeds through all half-acts until the five-act story is completed. Appendix J provides the implementation-level controller pseudocode.

4 Experiments

In this section, we present experiments to validate the effectiveness of ReaderScope in story generation, including evaluation metrics, baseline comparisons, ablation studies, and result analysis.

4.1 Experimental Setup

Dataset. We use the same story requirements as StoryBox (Chen et al., 2026) to ensure a direct and fair comparison. The dataset consists of 20 story specifications, each containing a premise, a setting, and a cast of characters. All methods are evaluated on this identical set of input requirements.

Baselines. We select several representative methods as baselines and categorize them into four types: (1) **Vanilla Sequential Prompting**, which generates stories through direct continuation with minimal explicit planning, represented

by ROLLING; (2) **Plan-and-Write Frameworks**, which first construct narrative outlines and then generate stories accordingly, represented by DOC (Yang et al., 2023); (3) **Sandbox-Based Simulations**, which obtain story materials from simulated character interactions, represented by STORYBOX; and (4) **Feedback-Based Refinement**, represented by CRITICS (Bae and Kim, 2024), which iteratively improves story plans and text through natural-language critiques. Relative to ReaderScope, it serves as a text-only feedback baseline: qualitative textual critiques guide generation without explicit reader-response trajectories. READERSCOPE differs by quantifying multidimensional reader responses to provide a more precise and reader-grounded basis for controlled story generation.

Implementation Details. For a fair comparison, all generation agents are instantiated with gpt-4o-mini. All systems receive the same story requirement and are instructed to generate a five-chapter story, which controls the high-level output structure without enforcing identical word counts. In ReaderScope, we use five randomly sampled reader agents for online feedback and allow one feedback round for each generated unit, balancing controllability with generation cost.

4.2 Evaluation Setup

Pairwise Evaluation. We evaluate generated stories through pairwise comparison. For each story requirement, we compare stories produced by every pair of methods, resulting in $M(M -$

1)/2 comparisons for M methods. For each evaluation dimension, the evaluator selects the better story or marks the pair as tied. We report a normalized preference score for each method and metric. Let $w_{m,d}$, $t_{m,d}$, and $n_{m,d}$ denote the number of wins, ties, and total appearances of method m under metric d , respectively. The final score is :

$$\text{Pref}(m, d) = \frac{w_{m,d} + 0.5 t_{m,d}}{n_{m,d}}. \quad (9)$$

Thus, a tie gives 0.5 points to both methods, and a score of 0.5 corresponds to parity against the compared systems. The same aggregation rule is used for automatic and human evaluation; in human evaluation, each valid annotation contributes one appearance.

Automatic Evaluation. For automatic pairwise evaluation, we anonymize the generated stories with meaningless identifiers, such as “Q7K” and “M3P”, to avoid revealing the generating method. We further randomize the order of the two stories in each comparison to reduce position bias. The comparisons are judged by DeepSeek-V4-Flash.

Human Evaluation. We further conduct a lightweight human pairwise evaluation under the same anonymized comparison protocol. We collect 200 pairwise comparisons from 10 annotators, covering all method pairs across the 20 story requirements. Each pairwise comparison is judged by one annotator, who independently selects the stronger story for each evaluation dimension or marks the pair as tied. Each annotator evaluates all ten method pairs for two complete story requirements. Because each comparison receives a single human judgment, replicated human-human agreement is not identifiable from the collected labels; we therefore treat human evaluation as an external preference check complementary to the automatic pairwise evaluation. Details of the interface, annotation instructions, and aggregation procedure are provided in Appendix A.3.

Statistical Inference for Human Preferences. We treat the story requirement, rather than an individual pairwise judgment, as the independent sampling unit. We obtain 95% confidence intervals by cluster bootstrapping the 20 requirements and test ReaderScope-baseline differences using exact paired sign-flip tests with Holm correction across the 20 method-metric contrasts; tests against the 0.5 preference-parity level are corrected separately across the five metrics. Additional requirement-

and annotator-block sensitivity analyses are reported in Appendix A.4.

Evaluation Metrics. We evaluate generated stories along five holistic, rubric-based dimensions. *Narrative Craft* is not a ReaderScope control target; it is the rubric’s explicit measure of overall story quality. The remaining four dimensions—*Emotion*, *Suspense*, *Immersion*, and *Insight*—correspond to ReaderScope’s reader-response control signals and should therefore be interpreted as measures of target-control effectiveness rather than independent evidence of overall literary quality. Although suspense-related uncertainty is operationalized as entropy during generation, it is evaluated through the reader-facing dimension of *Suspense*. All five dimensions are assessed holistically on complete stories, not through offline trajectory matching. The full rubric is provided in Appendix A.

Automatic-human consistency. To assess whether the automatic evaluator provides judgments consistent with human preferences, we compare automatic and human pairwise decisions on the same annotated story pairs. We report exact agreement and Cohen’s κ over three labels, i.e., first story preferred, second story preferred, or tie. We also compare aggregate system-level preference scores using rank and score correlations, with full results in Appendix A.5.

4.3 Evaluation Results

Overall preference results. Figure 3 shows that ReaderScope obtains the highest round-robin preference score on all five dimensions under both automatic and human evaluation, with StoryBox as the closest overall competitor. The automatic margins are most pronounced for **Emotion**, **Suspense**, and **Immersion**, whereas the human-evaluation gaps are smaller. ReaderScope’s human scores and requirement-clustered 95% confidence intervals are 0.681 [0.550, 0.800] for **Narrative Craft**, 0.725 [0.606, 0.838] for **Emotion**, 0.719 [0.581, 0.844] for **Suspense**, 0.663 [0.519, 0.800] for **Immersion**, and 0.706 [0.606, 0.800] for **Insight**. In exact requirement-level tests against the 0.5 parity level, Narrative Craft, Emotion, Suspense, and Insight remain significant after Holm correction ($p_{\text{Holm}} = 0.039, 0.014, 0.026$, and 0.007 , respectively), while Immersion is borderline ($p_{\text{Holm}} = 0.051$).

Baseline contrasts and interpretation. Among the 20 human-evaluation contrasts between ReaderScope’s round-robin score and those of the four baselines, nine remain significant after panel-wide Holm correction: ReaderScope scores higher than DOC and Critics on Narrative Craft, Emotion, Suspense, and Insight, and higher than Rolling on Emotion. In contrast, none of the five ReaderScope–StoryBox score differences is significant; the estimated differences range from 0.013 to 0.138, and all corresponding clustered confidence intervals include zero. These contrasts compare the aggregate round-robin scores represented by the bars, rather than only the direct ReaderScope–baseline matches. The results therefore place ReaderScope in the leading human-preference tier and establish gains over several baselines, but do not establish superiority over StoryBox.

Scope and robustness. Because Emotion, Suspense, Immersion, and Insight are aligned with the controller’s targets, the larger automatic gains primarily demonstrate successful target realization. Human preferences and the non-targeted Narrative Craft dimension provide complementary evidence that some of these gains are perceptible beyond the explicitly controlled signals. Leave-one-requirement-out and leave-one-annotator-out analyses preserve the clearest directional advantages for Emotion, Suspense, and Insight, whereas Narrative Craft is less consistent across annotators and Immersion is more sensitive to both requirements and annotators. Complete contrast and sensitivity statistics are reported in Appendix A.4.

4.4 Ablation Studies

To assess the contribution of each feedback signal, we ablate one reader-response dimension at a time from ReaderScope. As shown in Figure 4, all ablated variants perform worse than the full model, indicating that emotion, entropy, immersion, and insight jointly contribute to the generation process. Removing emotion feedback causes the largest degradation, suggesting that affective control plays a central role in shaping character reactions, conflict escalation, and reader experience. Removing entropy feedback also leads to clear drops, especially on narrative craft and suspense, showing that uncertainty management helps maintain tension and narrative progression. The effects of removing immersion and insight are smaller

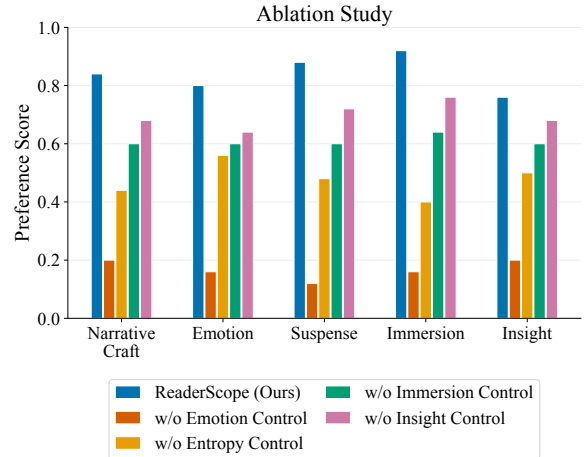


Figure 4: Ablation results when removing each reader-response feedback dimension from ReaderScope.

Dim.	Crit.	Win.	Story	Band	Dev.	Gain
Emotion	Rev.	100.0	100.0	—	—	+0.64
Entropy	Band	97.5	75.0	86.5	0.0061	+0.40
Immersion	Drop	100.0	100.0	—	—	+0.28
Insight	Band	100.0	100.0	99.9	0.0000	+0.08

Table 2: Trajectory-control fidelity and downstream effect. Win. and Story are window- and story-level success rates (%). Band is the percentage of points within the target band; Dev. is normalized band deviation. Gain denotes the score difference between ReaderScope and the corresponding no-target-control ablation.

but consistent, suggesting that they mainly support scene-level engagement and reflective payoff, respectively. Finally, the ablation results do not show a strict one-to-one mapping between feedback dimensions and evaluation metrics, since the four signals interact during revision: emotional escalation can support suspense, coherent tension can improve immersion, and premise-driven character arcs can still produce insight.

4.5 Trajectory-control Fidelity Validation

The previous evaluations show that ReaderScope improves final story quality; here we further validate whether the generated stories follow the intended reader-response dynamics. As shown in Table 2, ReaderScope achieves high control fidelity across all four dimensions, indicating that the feedback loop can largely enforce its explicit trajectory constraints. Entropy is the most difficult dimension to control: although its window-level success rate is high, its story-level success drops to 75.0%, since a single failed suspense window makes the whole story fail the story-level crite-

tion. This suggests that suspense requires more precise global timing of uncertainty introduction and resolution than the other dimensions. The Gain column further shows that removing the corresponding controller consistently reduces downstream evaluation scores, with the largest drops for emotion and suspense. Thus, the trajectory constraints are not only satisfied internally, but are also associated with measurable improvements in reader-facing story quality.

4.6 Extensive Study

We further study the sensitivity of ReaderScope to the number of feedback iterations and the size of the simulated reader group, with full results in Appendix K. Performance generally improves up to three feedback rounds and benefits from larger reader groups, suggesting that reader feedback provides useful control signals. However, the gains are not monotonic: excessive revision or overly broad aggregation can reduce performance on some dimensions. Thus, ReaderScope requires a balanced configuration rather than simply maximizing feedback frequency or reader-panel size.

5 Conclusion

We presented ReaderScope, a reader-response-guided framework for controllable long-form story generation. Instead of treating reader experience only as a post-hoc evaluation criterion, ReaderScope calibrates simulated reader-response trajectories and uses them as generation-time control signals. Human evaluation shows multiplicity-corrected preference gains over several representative baselines and competitive performance with the strongest baseline, while ablation and trajectory-fidelity analyses indicate that emotion, entropy, immersion, and insight provide complementary guidance. Although the current framework still relies on simulated readers and fixed response dimensions, our results suggest that reader-response modeling is a promising direction for generating stories that better control the evolving experience of readers.

6 Limitations

ReaderScope uses LLM-simulated readers in both offline calibration and online feedback. These trajectories are proxies, not direct measurements of human experience, and may reflect model, prompt, profile, and scale biases. Four evaluation di-

mensions also mirror the control targets, potentially favoring ReaderScope; improvements therefore support target-control effectiveness, not independent gains in overall literary quality. Results on the non-targeted Narrative Craft dimension and human judgments provide complementary evidence, but dense human-trajectory validation remains needed.

Each pairwise comparison was judged by one annotator, with annotators nested within story requirements. Requirement-clustered inference and annotator-block sensitivity analyses account for this dependence in uncertainty estimates, but cannot separate annotator and requirement effects or estimate inter-annotator agreement. Thus, the study does not establish a statistically significant advantage over StoryBox; replicated judgments on a shared subset are needed.

Calibration uses 20 reference novels and a fixed simulated-reader panel, limiting coverage of genres, cultures, reader populations, and pacing conventions. Chapter-level responses are also interpolated into normalized 60-point trajectories, whereas control operates on paragraph-level units. This enables cross-story alignment but may smooth local responses or shift their timing. Broader corpora, more diverse reader panels, and multi-resolution evaluation are needed.

The four dimensions are neither exhaustive nor independent (Appendix H.2). Reader-response control does not ensure global consistency or causal coherence. The feedback loop adds inference cost; because all systems use the same backbone, cross-backbone robustness remains untested. LLM evaluators may also share biases with simulated readers. Future work should examine joint response modeling, stronger consistency checks, and efficient human-validated feedback across model families.

7 Ethics Statement

ReaderScope is intended for controllable long-form story generation and creative-writing research. Because it models and optimizes affect, suspense, immersion, and reflection, similar techniques could be misused for psychological targeting, behavioral manipulation, or emotionally exploitative persuasion, particularly when combined with sensitive reader profiles. Moreover, its simulated readers and LLM-based response signals are proxies that may inherit biases from the under-

lying models, prompts, and constructed profiles, rather than direct measurements of human experience. Reader-facing applications should therefore disclose AI-generated content, restrict sensitive profiling, apply appropriate content safeguards, and retain human oversight.

Our human evaluation uses anonymized pairwise judgments of system outputs generated from the same story requirements. Annotators do not see method identities, and we neither infer their private attributes nor release personally identifying information. Reference texts were obtained from legally accessible sources under applicable copyright or licensing conditions, while commercial LLMs were used under their providers' terms. Released artifacts are limited to research-relevant inputs, generated outputs, calibrated targets, anonymized preference labels, and evaluation scripts; they exclude copyrighted source texts, API credentials, personally identifying information, and unnecessary raw metadata.

Acknowledgements

This study is supported by National Key R&D Plan of China under Grant No. 2024YFB4303900. The authors would like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

References

- Minwook Bae and Hyounghun Kim. 2024. [Collective critics for creative story generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.
- Anneliese Brei, Chao Zhao, and Snigdha Chaturvedi. 2024. [Returning to the start: Generating narratives with related endpoints](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 101–112, Mexico City, Mexico. Association for Computational Linguistics.
- Zehao Chen, Rong Pan, and Haoran Li. 2026. [Storybox: Collaborative multi-agent simulation for hybrid bottom-up long-form story generation using large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30359–30367.
- Richard J. Gerrig and Allan B.I. Bernardo. 1994. [Readers as problem-solvers in the experience of suspense](#). *Poetics*, 22(6):459–472.
- Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. [IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand. Association for Computational Linguistics.
- Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. [Agents'room: Narrative generation through multi-step collaboration](#). In *International Conference on Learning Representations*, volume 2025, pages 5150–5183.
- Shufan Jiang, Sizhou Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. 2026. [Hamlet: A hierarchical and adaptive multi-agent framework for live embodied theatrics](#). In *International Conference on Learning Representations*, pages 30068–30099.
- E.W.E.M. Kneepkens and Rolf A. Zwaan. 1995. [Emotions and literary text comprehension](#). *Poetics*, 23(1):125–138. Emotions and Cultural Products.
- Moniek Kuijpers, Frank Hakemulder, Ed Tan, and Miruna Doicaru. 2014. [Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption](#). *Scientific Study of Literature*, 4:89–122.
- Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2025. [Navigating the path of writing: Outline-guided text generation with large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 233–250, Albuquerque, New Mexico. Association for Computational Linguistics.
- Huang Lei, Jiaming Guo, Guanhua He, Xishan Zhang, Rui Zhang, Shaohui Peng, Shaoli Liu, and Tianshi Chen. 2024. [Ex3: Automatic novel writing by extracting, excelsior and expanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9125–9146, Bangkok, Thailand. Association for Computational Linguistics.
- Zefeng Lin, Yi Xiao, Zhiqiang Mo, Qifan Zhang, Jie Wang, Jiayang Chen, Jiajing Zhang, Hui Zhang, Zhengyi Liu, Xianyong Fang, and Xiaohua Xu. 2025. [R²: A LLM based novel-to-screenplay generation framework with causal plot graphs](#). *CoRR*, abs/2503.15655.
- Cristina Loi, Frank Hakemulder, Moniek Kuijpers, and Gerhard Lauer. 2023. [On how fiction impacts the self-concept: Transformative reading experiences and storyworld possible selves](#). *Scientific Study of Literature*, 12:44–67.

- Yan Ma, Yu Qiao, and Pengfei Liu. 2024. [MoPS: Modular story premise synthesis for open-ended automatic story generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2135–2169, Bangkok, Thailand. Association for Computational Linguistics.
- Raymond A Mar, Keith Oatley, Maja Djikic, and Justin Mullin. 2011. [Emotion and narrative fiction: Interactive influences before, during, and after reading](#). *Cognition and Emotion*, 25(5):818–833.
- Jonathan Pei, Zeeshan Patel, Karim El-Refai, and Tianle Li. 2024. [SWAG: Storytelling with action guidance](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14086–14106, Miami, Florida, USA. Association for Computational Linguistics.
- Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. [BOOK-WORLD: From novels to interactive agent societies for story creation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.
- Shalom H Schwartz. 2012. [An overview of the schwartz theory of basic values](#). *Online Readings in Psychology and Culture*, 2.
- Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. [Cognitive architectures for language agents](#). *Preprint*, arXiv:2309.02427.
- Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. [A cognitive writing perspective for constrained long-form text generation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9832–9844, Vienna, Austria. Association for Computational Linguistics.
- Phoebe J. Wang and Max Kreminski. 2024. [Guiding and diversifying llm-based story generation via answer set programming](#). *Preprint*, arXiv:2406.00554.
- Yi Wang, Qian Zhou, and David Ledo. 2024. [Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning](#). In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, New York, NY, USA. Association for Computing Machinery.
- Yichen Wang, Kevin Yang, Xiaoming Liu, and Dan Klein. 2023. [Improving pacing in long-form story planning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10788–10845, Singapore. Association for Computational Linguistics.
- Yunge Wen, Chenliang Huang, Hangyu Zhou, Zhuo Zeng, Chun Ming Louis Po, Julian Togelius, Timothy Merino, and Sam Earle. 2025. [All stories are one story: Emotional arc guided procedural game level generation](#). *Preprint*, arXiv:2508.02132.
- David Wilmot and Frank Keller. 2020. [Modelling suspense in short stories as uncertainty reduction over neural representation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1763–1788, Online. Association for Computational Linguistics.
- Kaige Xie and Mark Riedl. 2024. [Creating suspenseful stories: Iterative planning with large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian’s, Malta. Association for Computational Linguistics.
- Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. [DOC: Improving long story coherence with detailed outline control](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. [Re3: Generating longer stories with recursive reprompting and revision](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Kuan Lu, Shiyao Qian, Xinhang Yuan, Yi Xin, Yijin Wang, Jingqun Tang, Yuchen Li, Hongyang He, Zhen Tian, Tianxiang Xu, Keqin Li, Menghao Huo, Jiaqi Chen, Miao Zhang, Tianyu Shi, and Jianyuan Ni. 2026. [Score: Story coherence and retrieval enhancement for ai narratives](#). In *Companion Proceedings of the ACM Web Conference 2026*, WWW Companion ’26, page 823827, New York, NY, USA. Association for Computing Machinery.

A Evaluation Details

A.1 Evaluation Metrics

This section provides additional details about the evaluation criteria used in our experiments. Both automatic and human evaluations adopt the same pairwise comparison protocol: given the same story condition, evaluators compare two anonymously labeled system-generated stories and select the stronger one for each evaluation dimension, or choose SAME when no clear advantage is observed. All dimensions are judged independently to avoid transferring an overall preference to every metric.

Narrative Craft Quality. This metric evaluates the causal coherence, character motivation, scene progression, thematic integration, and narrative completeness of the story. It is the only evaluation dimension where the overall plot structure and narrative completeness are explicitly considered. A stronger story should present believable character motivations, clear cause-and-effect development, meaningful narrative progression, and a theme that emerges naturally through events and consequences. Stories with abrupt transitions, weak causal links, inconsistent motivations, or underdeveloped narrative arcs are rated lower on this dimension.

Emotion. This metric evaluates the reader’s emotional experience during the story. The focus is on the trajectory and credibility of emotional responses, including shifts among fear, hope, guilt, hesitation, betrayal, loss, relief, and resolve. A stronger story should create emotional involvement through credible character situations, meaningful choices, and consequential events. Overly explicit or exaggerated emotional wording is not rewarded unless it contributes to a natural and sustained reader experience.

Suspense. This metric evaluates the reader’s curiosity, uncertainty, tension, and anticipation during the reading process. The focus is on meaningful unresolved questions concerning trust, choices, hidden motives, risks, and consequences. A stronger story should make the reader actively anticipate what may happen next through character decisions, information gaps, conflicting motivations, and coherent stakes. Action, danger, confusion, or withheld information alone is not sufficient unless it genuinely increases reader anticipa-

tion.

Immersion. This metric evaluates the reader’s sense of absorption, attention, and presence in the story. The focus is on concrete situations, stable perspective, sensory pressure, spatial presence, and moment-to-moment involvement in the characters’ experiences. A stronger story should make the reader feel close to the scene and mentally involved in what the characters are experiencing. Broad worldbuilding or decorative description is not rewarded unless it directly strengthens scene-level absorption.

Insight. This metric evaluates the reader’s post-reading reflection and sense of meaning. The focus is on whether the story encourages reflection on human choices, relationships, values, identity, responsibility, complicity, loyalty, sacrifice, or social pressure. A stronger story should invite deeper thought through conflicts and consequences. Direct explanation, abstract preaching, moral slogans, or explicit thematic statements are not rewarded unless they are supported by the narrative itself.

A.2 Automatic Evaluation

For automatic evaluation, we use DeepSeek-V4-Flash, following the official model name shown by the provider at the time of access. All automatic judgments are conducted in May 2026 with anonymized method identifiers and randomized story order. To reduce possible biases in LLM-as-a-judge evaluation, we adopt two precautions. First, method names are replaced with meaningless identifiers such as M3P and Q7K, so that the evaluator cannot associate a story with a particular system. Second, the input order of the two stories is randomly swapped during evaluation, which helps mitigate positional preference toward the first or second option. The model is required to output only a JSON object containing the five independent pairwise decisions. The detailed prompt is provided in Table 19.

A.3 Human Evaluation

For human evaluation, we invite 10 university students from multiple academic backgrounds, including computer science, mathematics, philosophy, business administration, law, and chemistry. The evaluation is conducted anonymously: participants do not see the names of the generation

methods and only compare two stories generated from the same story requirement. In total, we collect 200 valid pairwise comparisons, corresponding to all method-pair comparisons across the 20 story requirements. Each pairwise comparison is annotated by one participant. For each comparison, the annotator makes independent judgments on the five evaluation dimensions: Narrative Craft, Emotion, Suspense, Immersion, and Insight. The annotator may also choose SAME when no clear advantage is observed.

Because our human study assigns one annotator to each pairwise comparison, human-human agreement is not identifiable from the collected labels. We therefore use the human results as a lightweight external preference check rather than as a consensus-based annotation benchmark. The same preference-score aggregation rule used in automatic evaluation is applied to human judgments, where a win receives 1 point and a tie assigns 0.5 points to each system. The annotation interface is implemented using Label Studio, as shown in Appendix B.

A.4 Clustered Inference and Robustness for Human Preferences

Dependence structure and bootstrap. The 200 pairwise tasks do not constitute 200 independent observations. For each of the 20 story requirements, the evaluation contains all ten unordered pairs among the five methods, and each task produces judgments on all five dimensions. Moreover, a single annotator evaluates all ten method pairs for a requirement, and each of the ten annotators is assigned two requirements. We therefore use the requirement as the primary cluster: in each of 100,000 bootstrap replicates, we sample 20 requirements with replacement and retain every method-pair and metric judgment associated with each sampled requirement. The pointwise 95% intervals are percentile intervals, and the bootstrap standard deviation provides the reported standard error. The random seed is 20260714.

Formal tests. For baseline b and dimension d , we first compute the displayed preference score separately within each requirement and form the paired difference

$$\delta_{r,b,d} = \text{Pref}_r(\text{READERSCOPE}, d) - \text{Pref}_r(b, d).$$

The primary contrast is $\bar{\delta}_{b,d}$, i.e., the difference between the two round-robin bar heights in Figure 3. We test the 20 requirement-level differences

with a two-sided exact sign-flip test, enumerating all 2^{20} assignments, and apply Holm correction jointly over the four baselines and five dimensions. We also test the five ReaderScope scores against the 0.5 parity level with the same requirement-level procedure and a separate five-test Holm family. The confidence intervals are pointwise rather than simultaneous; consequently, all confirmatory claims are based on the panel-wide Holm-adjusted p -values.

Topic sensitivity. We report the standard deviation of the 20 requirement-level score differences and recompute the aggregate difference after deleting each requirement in turn. Table 5 focuses on StoryBox (Chen et al., 2026), the closest baseline and therefore the most conservative comparison. Individual requirement effects are heterogeneous: no dimension favors ReaderScope on every requirement. Nevertheless, the leave-one-requirement-out range remains above zero for Narrative Craft, Emotion, Suspense, and Insight. This indicates that their positive aggregate directions are not caused by any single topic, while the Immersion difference is not directionally stable.

Annotator sensitivity and limitation. We conduct two diagnostics using the original assignment. First, an annotator-block bootstrap resamples the ten annotators and retains both complete requirements assigned to each sampled annotator. Second, leave-one-annotator-out analysis deletes one annotator and both associated requirements at a time. Against StoryBox, all block-bootstrap intervals include zero, consistent with the nonsignificant primary contrasts. The leave-one-annotator-out ranges remain positive for Emotion, Suspense, and Insight, but cross zero for Narrative Craft and Immersion. Across the other baselines, annotator-block intervals remain positive for all comparisons with DOC and Critics, and for Emotion, Suspense, and Insight against Rolling. These analyses quantify sensitivity to the observed annotation blocks, but they do not create replicated judgments: because annotator and requirement are nested, they cannot estimate human-human agreement or disentangle annotator noise from topic variation.

Dimension	Score	Boot. SE	95% cluster CI	<i>p</i>	<i>p</i> _{Holm}
Narrative Craft	0.681	0.065	[0.550, 0.800]	0.019	0.039
Emotion	0.725	0.060	[0.606, 0.838]	0.003	0.014
Suspense	0.719	0.067	[0.581, 0.844]	0.009	0.026
Immersion	0.663	0.072	[0.519, 0.800]	0.051	0.051
Insight	0.706	0.050	[0.606, 0.800]	0.001	0.007

Table 3: ReaderScope human preference scores and inference against the 0.5 round-robin parity level. Confidence intervals and standard errors use story-requirement-clustered bootstrap resampling. *p*-values use the two-sided exact requirement-level sign-flip test; Holm correction is over the five dimensions. Bold values are significant at 0.05 after correction.

Dimension	Baseline	Δ	Boot. SE	95% cluster CI	<i>p</i>	<i>p</i> _{Holm}
Narrative Craft	Rolling	0.094	0.082	[-0.075, 0.244]	0.312	1.000
	DOC	0.394	0.095	[0.194, 0.569]	0.002	0.029
	Critics	0.344	0.092	[0.156, 0.519]	0.003	0.042
	StoryBox	0.075	0.113	[-0.137, 0.306]	0.561	1.000
Emotion	Rolling	0.281	0.076	[0.131, 0.425]	0.003	0.042
	DOC	0.381	0.094	[0.188, 0.556]	0.002	0.029
	Critics	0.331	0.088	[0.150, 0.494]	0.003	0.042
	StoryBox	0.131	0.116	[-0.088, 0.363]	0.306	1.000
Suspense	Rolling	0.206	0.100	[0.006, 0.400]	0.068	0.525
	DOC	0.394	0.089	[0.213, 0.563]	< 0.001	0.017
	Critics	0.369	0.080	[0.206, 0.519]	< 0.001	0.016
	StoryBox	0.125	0.129	[-0.125, 0.381]	0.379	1.000
Immersion	Rolling	0.188	0.090	[0.006, 0.356]	0.066	0.525
	DOC	0.281	0.096	[0.087, 0.463]	0.014	0.127
	Critics	0.331	0.099	[0.131, 0.519]	0.006	0.070
	StoryBox	0.013	0.140	[-0.256, 0.288]	0.965	1.000
Insight	Rolling	0.188	0.060	[0.063, 0.300]	0.010	0.103
	DOC	0.406	0.070	[0.262, 0.531]	< 0.001	0.003
	Critics	0.300	0.073	[0.156, 0.444]	0.001	0.024
	StoryBox	0.138	0.107	[-0.063, 0.356]	0.249	1.000

Table 4: ReaderScope minus baseline differences between the human preference-score bars in Figure 3. Tests use paired requirement-level differences. Holm correction is over all $4 \times 5 = 20$ contrasts; bold values are significant at 0.05 after correction. These bar-height contrasts aggregate each method’s four round-robin matchups and should not be interpreted as direct head-to-head win rates.

Dimension	Story SD	Story +/-tie/-	LOSO Δ range	Annotator-block 95% CI	LOAO Δ range
Narrative Craft	0.518	7/2/11	[0.026, 0.112]	[-0.188, 0.406]	[-0.028, 0.139]
Emotion	0.531	10/1/9	[0.086, 0.164]	[-0.156, 0.456]	[0.035, 0.181]
Suspense	0.593	10/1/9	[0.079, 0.178]	[-0.169, 0.456]	[0.028, 0.188]
Immersion	0.640	7/1/12	[-0.039, 0.059]	[-0.325, 0.381]	[-0.097, 0.111]
Insight	0.490	9/2/9	[0.092, 0.171]	[-0.113, 0.444]	[0.042, 0.194]

Table 5: Topic and annotation-block sensitivity for the ReaderScope–StoryBox bar-height difference. Story SD is the between-requirement standard deviation; +/-tie/- counts the signs of the 20 requirement-level differences; LOSO deletes one requirement; the annotator-block bootstrap resamples ten two-requirement blocks; and LOAO deletes one annotator and both assigned requirements. Annotator and topic effects are confounded in the latter two diagnostics.

A.5 Automatic–Human Consistency

We assess whether automatic evaluation recovers the preference patterns observed in human judgments, rather than whether it can replace human evaluation.

Protocol. Following the pairwise consistency analysis of (Ran et al., 2025), each anonymized A/B judgment is mapped to a canonical method order and represented as a preference for either method or a tie. Matching the automatic records with deduplicated human annotations yields 1,000 aligned decisions (200 comparison tasks across five dimensions). We report exact agreement, Cohen’s κ , and directional agreement after excluding decisions in which either evaluator selects a tie. Confidence intervals for κ are estimated from 1,000 bootstrap samples clustered by comparison task. We additionally compare method-level preference scores across the 25 method–metric combinations.

Subset	N	Exact (%)	κ (95% CI)	Dir. (%)
Overall	1000	53.0	0.126 [0.063, 0.186]	64.9
Narrative Craft	200	50.5	0.127 [0.036, 0.222]	62.7
Emotion	200	57.5	0.097 [-0.022, 0.215]	63.5
Suspense	200	54.0	0.152 [0.056, 0.247]	66.3
Immersion	200	54.5	0.084 [-0.006, 0.180]	61.6
Insight	200	48.5	0.157 [0.082, 0.239]	71.9

Table 6: Item-level automatic–human agreement. Dir. excludes comparisons for which either evaluator selects a tie. Confidence intervals use task-clustered bootstrap resampling.

Item-level agreement. As shown in Table 6, exact agreement is 53.0%, with weak but positive chance-corrected agreement ($\kappa = 0.126$, 95% CI [0.063, 0.186]). This discrepancy partly reflects different tie behavior: humans select SAME in 17.9% of judgments, compared with 0.4% for the automatic evaluator. Directional agreement increases to 64.9% after tie-related cases are excluded, indicating better agreement on the preferred side when both evaluators express a preference.

Subset	Pearson	Spearman	Kendall	ICC(C)	MAE
Overall	0.877	0.846	0.656	0.735	0.135
Narrative Craft	0.902	1.000	1.000	0.780	0.130
Emotion	0.977	0.900	0.800	0.819	0.110
Suspense	0.885	0.900	0.800	0.751	0.133
Immersion	0.735	0.600	0.400	0.579	0.173
Insight	0.894	0.900	0.800	0.756	0.130

Table 7: Alignment of automatic and human method-level preference scores. The overall row uses 25 method–metric points; per-dimension rows use five methods and are descriptive. ICC(C) denotes the two-rater consistency intraclass correlation.

Aggregate alignment. At the system level, automatic and human preference scores are strongly aligned: Pearson’s $r = 0.877$, Spearman’s $\rho = 0.846$, Kendall’s $\tau_b = 0.656$, and ICC(C) = 0.735, with an MAE of 0.135 (Table 7). Both evaluations identify the same top-ranked method and top-two method set on all five dimensions. Because each per-dimension correlation contains only five methods, these values are descriptive. Overall, automatic evaluation is suitable for scalable system-level comparison and diagnosis, whereas human judgments remain necessary for validating individual preferences.

B Human Evaluation Interface

Shared Story Setting

The story takes place in an ancient, dilapidated mansion in a remote village, surrounded by mist and dense forest. Ava, a young woman, discovers that her family's ancestral home is cursed, and she must uncover its dark secrets before the curse consumes her entire bloodline.

The main characters include:
Ava Lawrence, a determined and intelligent woman in her twenties who inherits the family estate;
Edward Lawrence, Ava's estranged father who went missing under mysterious circumstances years ago;
Grace Thornwell, a local historian who warns Ava about the mansion's dark past;
Jared Wilson, a young journalist investigating the curse who develops a romantic relationship with Ava;
The Specter, a mysterious figure haunting the mansion who appears to control the curse.

Story A

Whispers of the Cursed Bloodline

Ava Lawrence stood at the entrance of her ancestral home, the grand structure before her like a silent sentinel guarding secrets long buried. The air was thick with the scent of damp wood and decayed leaves, and as she stepped inside, a chill ran down her spine. Shadows danced on the walls, flickering in the dim light of her flashlight. Every creak of the floorboards was like a ghostly whisper, reminding her that she was not alone in this place. Ava moved cautiously from room to room, her heart pounding, her eyes scanning for anything that might offer clues to her family's dark past.

In the living room, Ava discovered a dust-covered piano, its keys yellowed with age. Her fingers glided lightly over the surface, stirring up particles that sparkled in the gloom. Faint notes seemed to seep from deep within the instrument, as if calling her closer.

Beside the piano, she noticed a pile of old diaries, their leather covers cracked and worn. She picked one up, feeling the weight of its secrets, and cautiously opened it. The pages were filled with elegant handwriting, speaking of love, loss, and despair—all intertwined with the curse that had plagued her family for generations.

As Ava read, a chill seemed to seep into her very bones. The entries detailed her ancestors' struggles, recounting how they had tried to break free from the curse's grip. Each generation faced unimaginable challenges, choices made out of fear and desperation, only to be drawn back into the darkness.

Her eyes widened as she absorbed the weight of these tales. Patterns began to emerge—cycles of betrayal, trust shattered by greed and ambition, and how every family member's attempt to defy their fate only led to more heartache. She felt anger and sorrow for those who came before her, wondering if they had ever truly understood the cost of their choices. The weight of their mistakes pressed down on her shoulders, yet a part of her was determined to break the cycle.

As she turned the pages, one entry caught her attention—a mention of an ancient ritual that could lift the curse, but at a cost far greater than any sacrifice her family had ever made. The words danced across the page, warnings intertwined with promises. It spoke of a pivotal choice: to either embrace the darkness or confront it head-on.

Ava's heart raced as she considered the implications. Was she willing to risk everything to save her bloodline? The thought sent a chill down her spine. The curse had haunted her family for

Story B

Echoes of a Cursed Legacy

Ava stood on the threshold of her family's ancestral mansion, the towering structure looming like a phantom against the twilight sky. The weathered stones seemed to whisper, their surfaces covered in ivy that clung desperately, as if struggling to survive. It was as though the house itself was breathing—a low, steady exhalation that mingled with the cool evening air. The last rays of sunlight slipped below the horizon, casting elongated shadows that stretched across the ground like ghostly fingers.

Hesitating, Ava looked down at her hands—her fingers trembling slightly as they gripped the brass doorknob. Her heart pounded, trying to escape her chest. Taking a deep breath, she pushed the door open, its mournful groan echoing through the empty foyer.

Ava stepped inside and was immediately enveloped by the damp scent of aged wood and mustiness, wrapping around her like a shroud. Dust motes danced in the faint light filtering through the cracked stained-glass windows, casting fractured rainbows that fell upon the faded carpet, which still bore the ghostly traces of intricate floral patterns. With every step, she stirred long-forgotten memories, as if the walls themselves were whispering. Even in its decay, the mansion's grandeur remained evident; the high ceilings were adorned with delicate plaster moldings, many of which had crumbled and fallen away like forgotten dreams. Shadows lurked in the corners, reluctant to reveal the secrets they held.

Ava continued into the dim corridor, the air thick with an unsettling chill. Her heartbeat quickened, not merely from fear, but from a growing curiosity about the mysteries surrounding her inheritance. Every portrait she passed seemed alive; eyes followed her with an almost conscious awareness, as if they were guardians of the secrets that had haunted her family for generations. The flickering candlelight cast eerie shadows that danced across the walls, heightening the somber atmosphere that surrounded her.

Moved by a sudden surge of determination, Ava turned away from the portraits and headed toward the library, where days earlier she had discovered the entrance to a hidden room behind a false wall. The heavy wooden door creaked open, revealing a small space cluttered with dust-covered furniture and relics of her family's past. With each step, Ava's heart raced, an irresistible curiosity intertwined with underlying fear. The faint light of her flickering candle danced across the room, casting irregular shadows that seemed to whisper ancient secrets.

As her gaze swept over the forgotten items, Ava noticed a faded tapestry hanging askew

Evaluation

Please evaluate each dimension independently. Do not select the same story across all dimensions simply because one story appears stronger overall. Method names will not be displayed. Please judge only the reading experience provided by Story A and Story B.

Narrative Craft

This metric evaluates causal coherence, character motivation, scene progression, thematic integration, and narrative completeness. This is the only dimension that considers the overall plot structure and completeness of the story. A stronger story should present credible character motivations, clear causal progression, meaningful development, and a theme that emerges naturally through events and consequences rather than being stated directly.

☐ A^[1] ☐ B^[2] ☐ Same^[3]

Emotion

This metric evaluates the reader's emotional experience during reading. It focuses on the reader's emotional trajectory, including shifts among fear, hope, guilt, hesitation, betrayal, loss, relief, determination, and related emotions. A stronger story should generate emotional engagement through credible character situations, meaningful choices, and consequential events. Do not reward exaggerated emotional wording unless it contributes to a natural and sustained emotional reading experience.

☐ A^[4] ☐ B^[5] ☐ Same^[6]

Suspense

This metric evaluates the reader's curiosity, uncertainty, tension, and anticipation during reading. It focuses on meaningful unresolved questions related to trust, choices, hidden motives, risks, and consequences. A stronger story should make the reader actively want to know what will happen next through character decisions, information gaps, motivational conflicts, and clearly defined stakes. Do not reward action, danger, chaos, or concealed information merely for its presence unless it genuinely strengthens the reader's sense of anticipation.

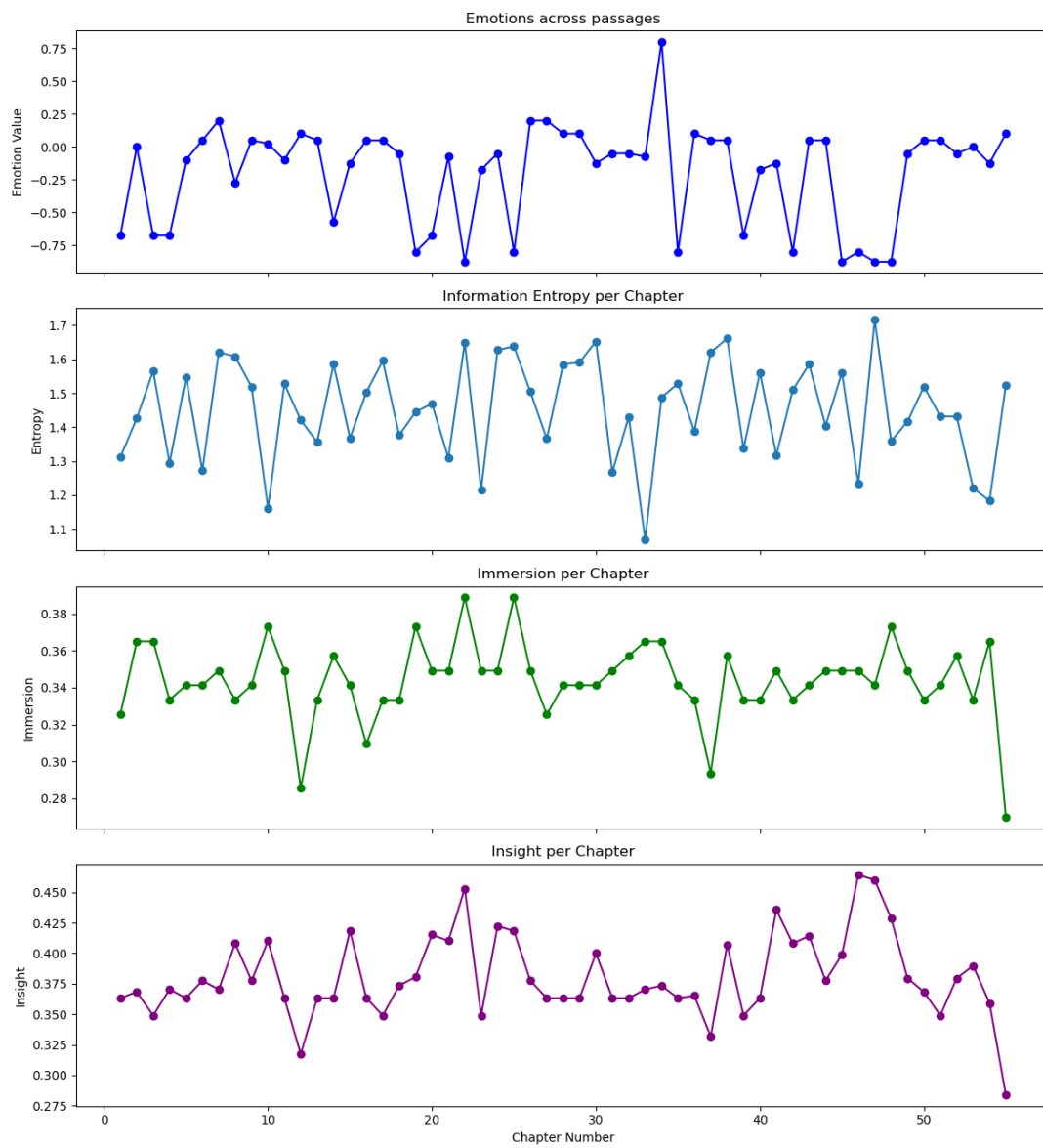
☐ A^[7] ☐ B^[8] ☐ Same^[9]

Immersion

This metric evaluates the reader's absorption, attentional involvement, and sense of presence in the story. It focuses on concrete situations, stable perspective,

C Example Four-Dimensional Response Trajectories

Reader feedback:reader5_story20_gpt-4o-mini



D Reference Story Dataset

Table 8: Statistics of the reference stories used for reader-agent response collection.

ID	Title	Chapters	Characters	Words
1	The Law and the Lady	50	764,126	139,647
2	A Tale of Two Cities	45	752,122	135,636
3	Pride and Prejudice	61	684,707	121,702
4	The Razor’s Edge	50	654,097	123,252
5	Wuthering Heights	41	644,641	115,954
6	The Adventures of Huckleberry Finn	43	560,132	112,825
7	Only a Girl’s Love	42	755,600	139,040
8	Anne of Green Gables	41	555,590	102,260
9	Tess of the d’Urbervilles: A Pure Woman	59	827,678	148,488
10	Oliver Twist	53	879,859	156,529
11	Sons and Lovers	45	872,981	159,772
12	Emma	55	875,184	157,663
13	Mansfield Park	48	868,046	148,088
14	Birds of Prey	43	911,056	161,665
15	The Keeper of the Door	55	917,050	165,394
16	Main Street	43	958,218	168,985
17	Mohammed Ali and His House	51	933,104	170,242
18	The Moon and Sixpence	58	405,238	75,031
19	The Adventures of Tom Sawyer	41	383,955	69,834
20	Uncle Tom’s Cabin	49	992,226	179,570

E Entropy Computation Details

Definition. We operationalize suspense-related uncertainty as *Entropy* during generation. For reader r and narrative unit t , the score combines uncertainty introduced by newly raised story questions with the impact of resolving previous questions:

$$E_{r,t} = H_{r,t}^{\text{new}} + S_{r,t}^{\text{res}}.$$

Thus, Entropy increases when a unit opens multiple plausible possibilities or provides an unexpected yet comprehensible resolution. Only $H_{r,t}^{\text{new}}$ is normalized Shannon entropy; $E_{r,t}$ is a composite suspense-related response signal and is not necessarily bounded to $[0, 1]$. We refer to the corresponding reader-facing evaluation dimension as *Suspense*.

Uncertainty from new questions. After reading unit t , the reader generates a set of new questions $\mathcal{Q}_{r,t}^{\text{new}}$. Each valid question q contains at least two candidate answers $\mathcal{Y}_q$, with belief probabilities satisfying $\sum_{y \in \mathcal{Y}_q} p_{r,t}(y | q) = 1$. Its normalized uncertainty is

$$h_{r,t}(q) = \frac{-\sum_{y \in \mathcal{Y}_q} p_{r,t}(y | q) \log_2 p_{r,t}(y | q)}{\log_2 |\mathcal{Y}_q|},$$

where $0 \log_2 0 = 0$. We then average across valid new questions:

$$H_{r,t}^{\text{new}} = \frac{1}{|\mathcal{Q}_{r,t}^{\text{new}}|} \sum_{q \in \mathcal{Q}_{r,t}^{\text{new}}} h_{r,t}(q).$$

Questions with fewer than two valid candidates are regenerated or excluded, and $H_{r,t}^{\text{new}} = 0$ when no valid new question is produced.

Resolution of previous questions. Before each update, the question memory is maintained using exponential forgetting,

$$f_q(t) = \exp(-0.08(t - \ell_q)),$$

where ℓ_q is the last activation unit of question q . Questions with cosine similarity of at least 0.85 are merged, and the memory retains at most 120 questions according to recency and importance. Reader-Scope retrieves the five historical questions most relevant to the current unit and updates their candidate

answers and belief probabilities. A question is resolved only when

$$\max_{y \in \mathcal{Y}_q} p_{r,t}(y | q) > 0.7.$$

The inequality is strict; a maximum probability equal to 0.7 remains unresolved. Updates in which all candidate probabilities are zero are discarded.

Resolution impact. For each resolved question q , let y_q^* be the answer with the highest updated probability. We identify the historical candidate $\hat{y}_q$ most semantically similar to y_q^* and use its probability before reading unit t :

$$u_{r,t}(q) = -\log_{10} \left(\max \left(p_r^{\text{old}}(\hat{y}_q | q), 10^{-9} \right) \right).$$

This ensures that answers previously considered less likely receive greater unexpectedness. Let $d_{r,t}(q) \in [0, 1]$ denote the cognitive difficulty of understanding the resolution, evaluated only when the question is resolved. Missing or invalid difficulty values are set to zero. For the resolved-question set $\mathcal{Q}_{r,t}^{\text{res}}$, we compute

$$U_{r,t} = \frac{1}{|\mathcal{Q}_{r,t}^{\text{res}}|} \sum_{q \in \mathcal{Q}_{r,t}^{\text{res}}} u_{r,t}(q), \quad D_{r,t} = \frac{1}{|\mathcal{Q}_{r,t}^{\text{res}}|} \sum_{q \in \mathcal{Q}_{r,t}^{\text{res}}} d_{r,t}(q),$$

and obtain

$$\begin{aligned} S_{r,t}^{\text{res}} &= U_{r,t}(1 - D_{r,t}) + 0.75(1 - U_{r,t})(1 - D_{r,t}) \\ &\quad + 0.50U_{r,t}D_{r,t} + 0.01(1 - U_{r,t})D_{r,t}. \end{aligned}$$

If no historical question is resolved, $S_{r,t}^{\text{res}} = 0$. Resolved questions are removed from memory, unresolved questions retain their updated distributions, and newly generated questions are appended for subsequent units.

F Silhouette Scores across Dimension

Figure 5 shows how the silhouette scores vary with K for the four reader-response dimensions. Emotion and immersion reach their best performance at approximately $K = 20$, matching the number of reference stories and indicating story-specific trajectory patterns. By contrast, entropy and insight peak only at $K = 2$, with isolated maxima and no stable cluster structure for larger K . These results suggest that clustering these two dimensions mainly produces a trivial binary split rather than meaningful trajectory groups. Since entropy and insight are evaluated on raw trajectories, whereas emotion and immersion are standardized under a conservative setting, the comparison further supports the conclusion that clustering is unsuitable for entropy and insight.

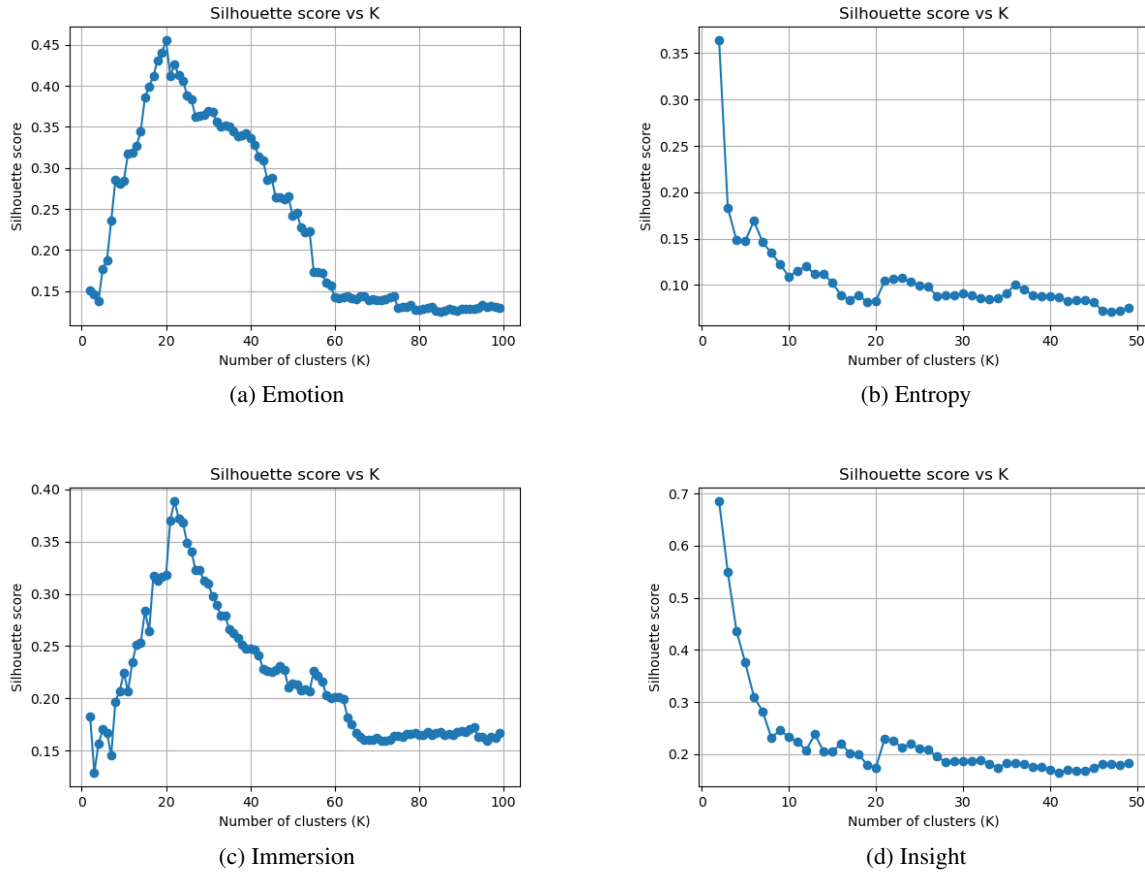


Figure 5: Silhouette scores of clustering methods across four dimensions as a function of the number of clusters.

G Clustering Results for Emotion

We cluster the emotion trajectories of 25 reader agents across 20 stories. As shown in Table 9 and Figure 6, the resulting clusters exhibit clear story-level structure. Since each story is read by 25 reader agents, the fact that most cluster sizes are close to 25 suggests that the clustering assignments are highly aligned with the underlying story labels. This indicates that readers exposed to the same story tend to produce internally consistent emotional trajectories, while trajectories from different stories show substantial variation in their emotional response patterns. The high cluster purity of most clusters, with several reaching 100%, supports the use of the emotion dimension as an informative signal for distinguishing story-specific reader-response patterns.

Table 9: Cluster distribution for the emotion dimension.

Cluster	Cluster Size (N)	Dominant Story ID	Cluster Purity
C0	24	11	100.00%
C1	25	2	96.00%
C2	25	16	100.00%
C3	29	20	82.76%
C4	25	7	100.00%
C5	26	17	96.15%
C6	27	10	92.59%
C7	29	9	86.21%
C8	25	1	96.00%
C9	25	19	100.00%
C10	25	3	100.00%
C11	26	12	96.15%
C12	25	15	100.00%
C13	26	18	96.15%
C14	25	14	100.00%
C15	19	6	100.00%
C16	25	4	100.00%
C17	21	8	100.00%
C18	25	13	92.00%
C19	23	5	86.96%

Notes:

1. The dominant story ID denotes the most frequent story label within each cluster.
2. Cluster purity is defined as the proportion of samples from the dominant story within the cluster.
3. The table reports k-medoids clustering results on emotion trajectories from 25 reader agents reading 20 stories. The trajectories are preprocessed with Z-score normalization, interpolation, and Savitzky–Golay smoothing with window length 7 and polynomial order 3.
4. The same preprocessing and reporting protocol is used for the immersion dimension.

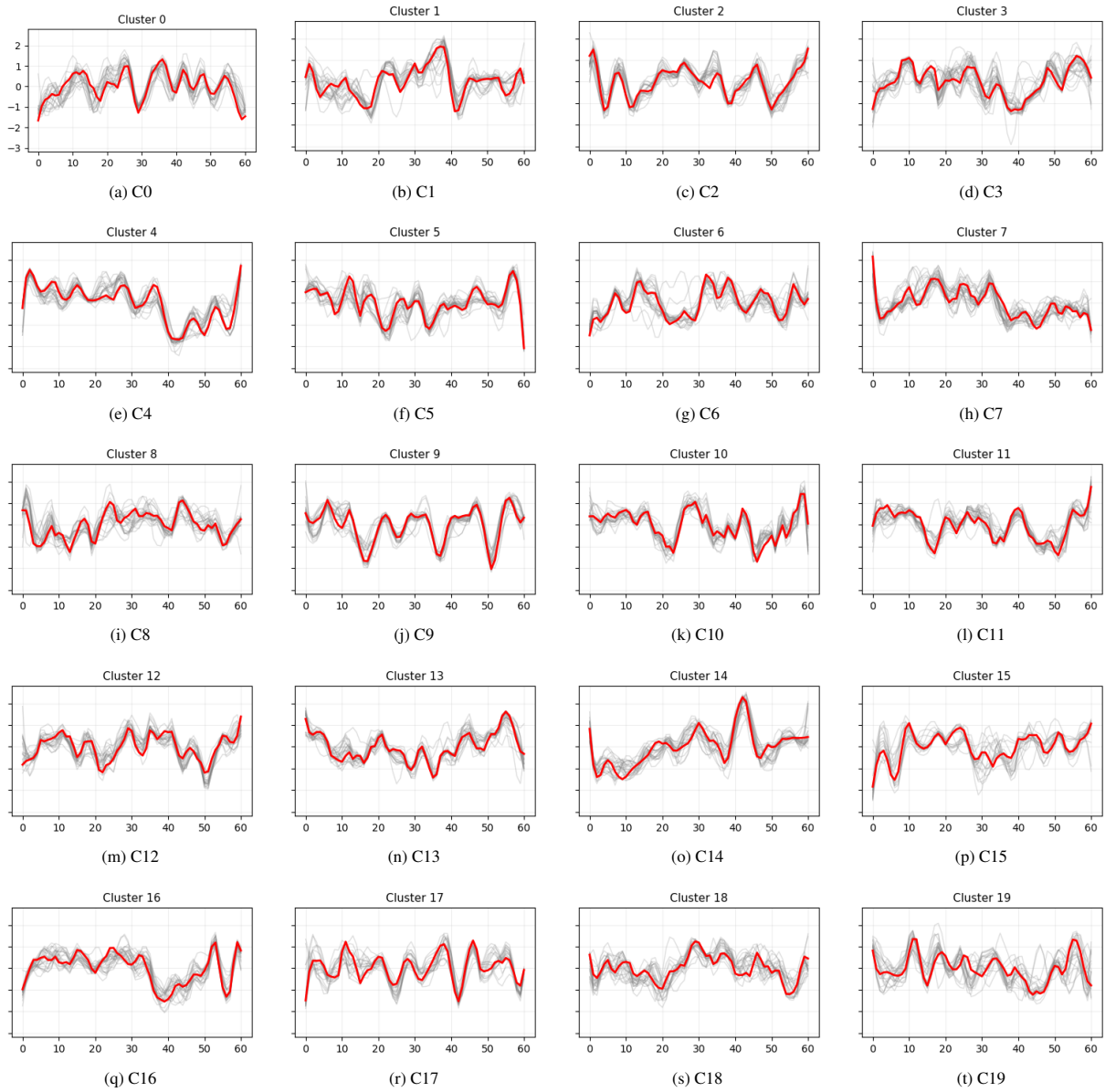


Figure 6: Overview of emotion-trajectory clusters. Gray curves represent individual reader-story emotion trajectories after Z-score normalization and interpolation to 60 points, while red curves indicate the k-medoids cluster centroids. The experiment includes 25 reader agents and 20 stories.

H Clustering Results for Immersion

We further cluster the immersion trajectories collected from 25 reader agents reading 20 stories. As shown in Table 10 and Figure 7, the resulting clusters show clear story-level grouping patterns.

The clustering results suggest that the temporal evolution of immersion is strongly associated with the underlying story. Among the 500 immersion trajectories, the clustering procedure produces 22 trajectory clusters with high internal consistency. All clusters achieve a purity higher than 74%, and more than half of them exceed 90% purity.

Some stories, such as Story 1 and Story 16, are split into multiple clusters, indicating that different reader agents may exhibit distinct subjective responses to the same narrative. Nevertheless, at the aggregate level, cluster membership remains highly aligned with story IDs. These results suggest that immersion trajectories are substantially shaped by the narrative structure and affective pacing of the story, making story-level grouping a useful basis for analyzing immersion-response patterns.

Table 10: Cluster distribution for the immersion dimension.

Cluster	Cluster Size (<i>N</i>)	Dominant Story ID	Cluster Purity
C0	28	2	89.29%
C1	25	4	96.00%
C2	26	17	96.15%
C3	22	6	95.45%
C4	16	16	87.50%
C5	31	8	74.19%
C6	22	19	100.00%
C7	26	20	96.15%
C8	24	18	95.83%
C9	24	9	87.50%
C10	25	7	92.00%
C11	26	14	92.31%
C12	25	12	100.00%
C13	11	16	100.00%
C14	31	15	87.10%
C15	24	13	95.83%
C16	23	3	100.00%
C17	10	1	80.00%
C18	15	1	100.00%
C19	17	11	94.12%
C20	27	10	92.59%
C21	22	5	95.45%

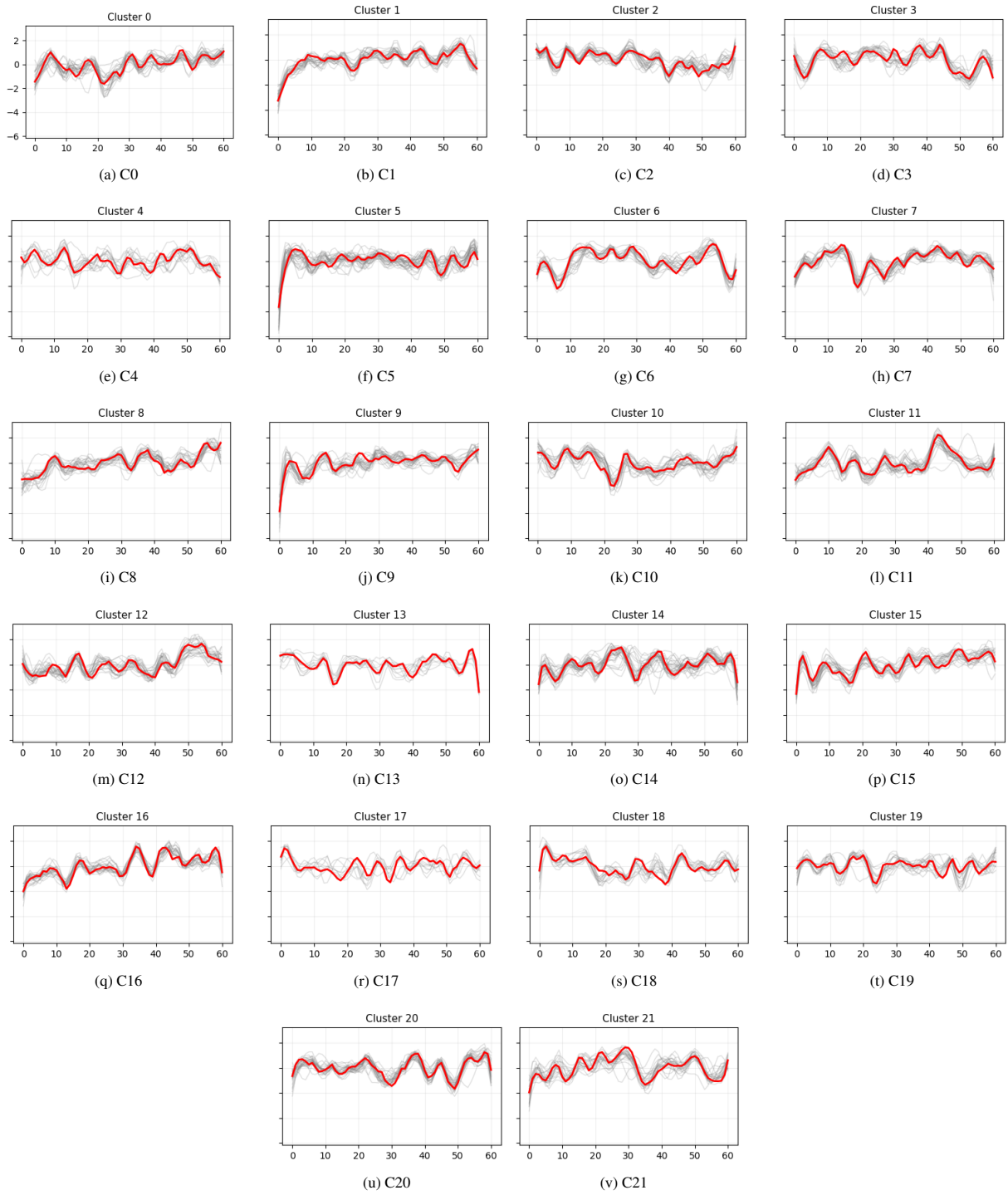


Figure 7: Overview of immersion-trajectory clusters. Gray curves represent individual reader-story immersion trajectories after Z-score normalization and interpolation to 60 points, while red curves indicate the k-medoids cluster centroids. The experiment includes 25 reader agents and 20 stories.

H.1 Temporal Registration for Suspense Entropy and Insight

Figure 8a shows the warping functions obtained from elastic registration for the suspense entropy and insight dimensions. The warping curves of individual samples are closely distributed around the diagonal line ($y = x$) and remain monotonically increasing. This indicates that the registration procedure corrects inter-reader phase variation without disrupting the temporal ordering of the original signals. This suggests that the extracted amplitude features remain temporally interpretable after alignment.

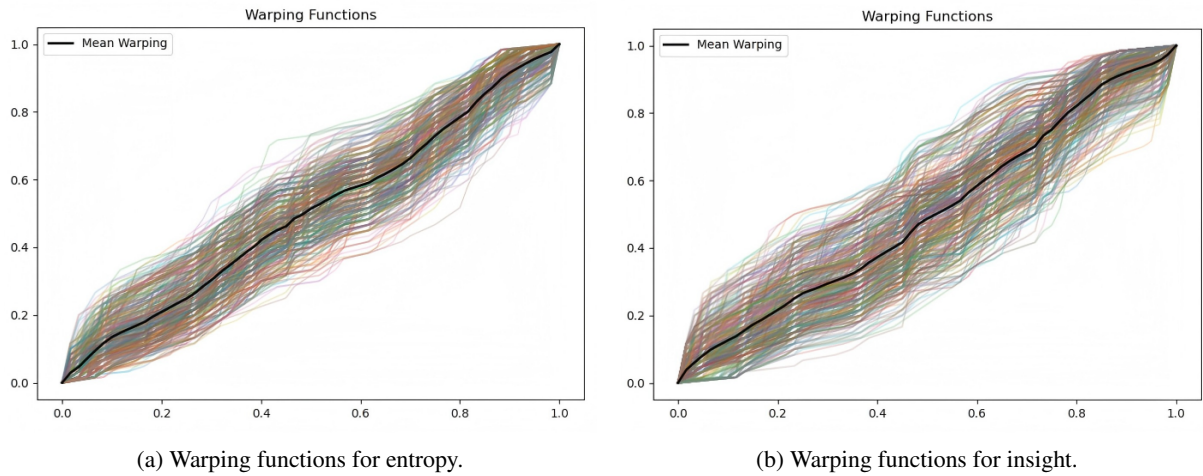


Figure 8: Temporal warping functions for suspense entropy and insight.

H.2 Correlation analysis of dimension pairs

Cross-Dimensional Lag Analysis. Figure 9 reports the lag at which each pair of reader-response trajectories reaches its maximum cross-correlation. Positive lags indicate that the first dimension precedes the second, while negative lags indicate the reverse. Most distributions peak sharply around zero, showing that cross-dimensional associations are mainly synchronous rather than governed by a stable lead-lag pattern. Thus, although some dimensions, especially immersion and insight, exhibit non-negligible association, this analysis does not support a fixed temporal ordering among dimensions. This motivates future work on joint controllers that explicitly model cross-dimensional dependencies.

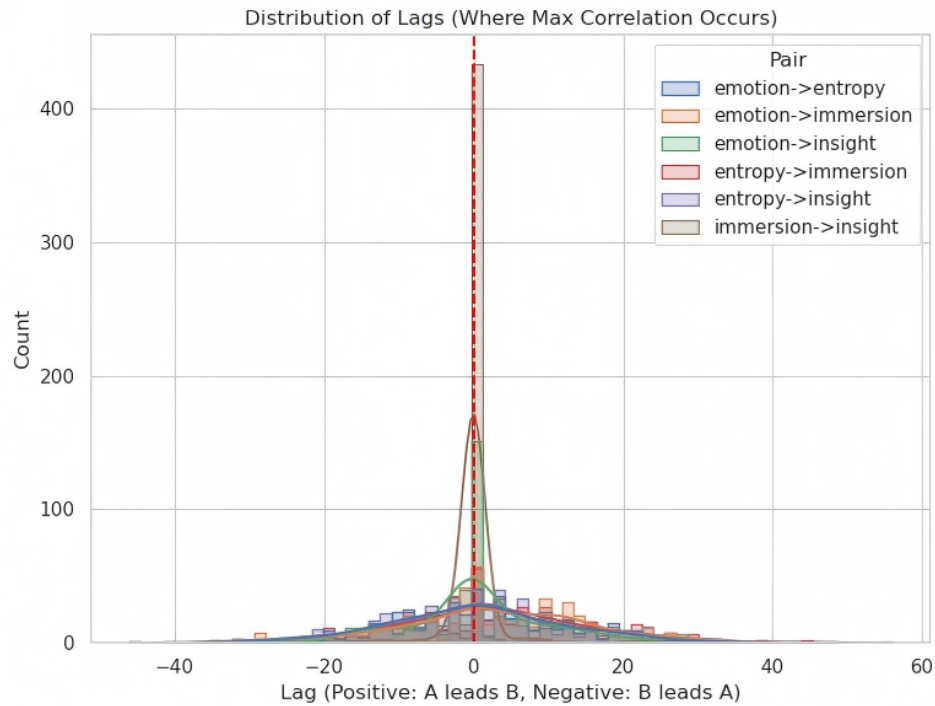


Figure 9: Distribution of lags at which maximum cross-correlation occurs between pairs of reader-response dimensions. Positive lags indicate that the first dimension leads the second, while negative lags indicate the reverse. The concentration around zero suggests that cross-dimensional associations are mostly synchronous rather than consistently directional.

Maximum-Correlation Heatmaps across Dimension Pairs. We compute the maximum pairwise correlation coefficients across different reader-response dimensions and apply co-clustering based on story and reader labels. The resulting heatmaps are shown in Figure 10.

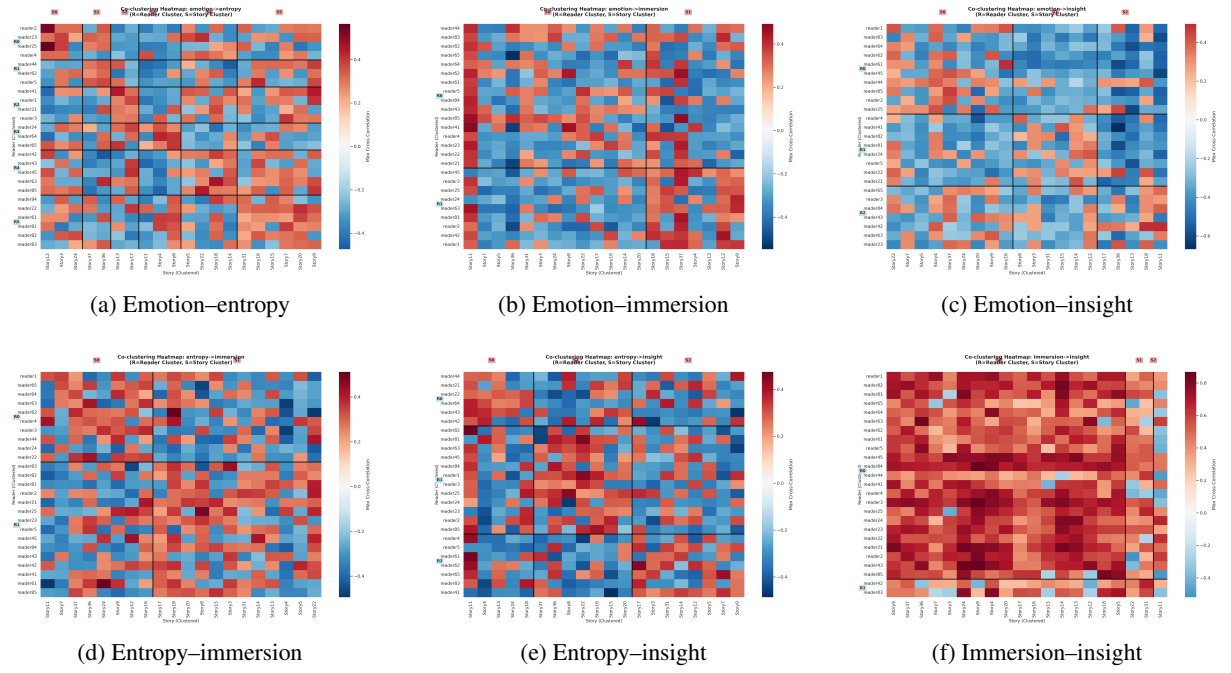


Figure 10: Co-clustered heatmaps of maximum correlation coefficients for the six pairs of reader-response dimensions.

Figure 10 shows the co-clustered heatmaps of maximum pairwise correlations among the four reader-response dimensions. The correlation patterns vary substantially across dimension pairs. Pairs involving emotion and suspense entropy exhibit relatively dispersed structures, suggesting that their associations with other dimensions are more story- and reader-dependent. In contrast, immersion and insight show a more coherent correlation pattern, with large high-correlation regions in Figure 10f. This indicates that, for many reader–story samples, stronger immersion is associated with stronger reflective insight. From a narrative-response perspective, immersive engagement may facilitate deeper cognitive reflection, while meaningful thematic content may in turn reinforce readers’ engagement with the narrative. The block structure revealed by co-clustering further suggests that this association is not driven by isolated samples, but reflects a recurring pattern across the dataset.

I Window-Length Extraction for Emotion and Immersion

Figures 11 and 12 illustrate the extraction of two window-based control signals: the maximum stable trend window for Emotion and the longest sustained decline window for Immersion. For Emotion, we search for the longest contiguous window in which the trajectory preserves the same directional trend, either monotonically increasing or monotonically decreasing. In the example shown in Figure 11, the resulting maximum stable trend window has length 4. For Immersion, we instead identify the longest contiguous window over which the trajectory continuously decreases. As shown in Figure 12, the extracted immersion-decline window has length 5. Given a set of reader-response trajectories, we derive dimension-specific window-length statistics for Emotion and Immersion. For each Emotion trajectory, we extract the longest contiguous window whose directional trend remains stable, i.e., the values are consistently increasing or consistently decreasing. The Emotion window length is then defined as the average of these maximum stable-trend lengths across all Emotion trajectories. For each Immersion trajectory, we extract the longest contiguous window over which the trajectory continuously decreases. The Immersion window length is defined analogously as the average of the extracted longest decline-window lengths across all Immersion trajectories.

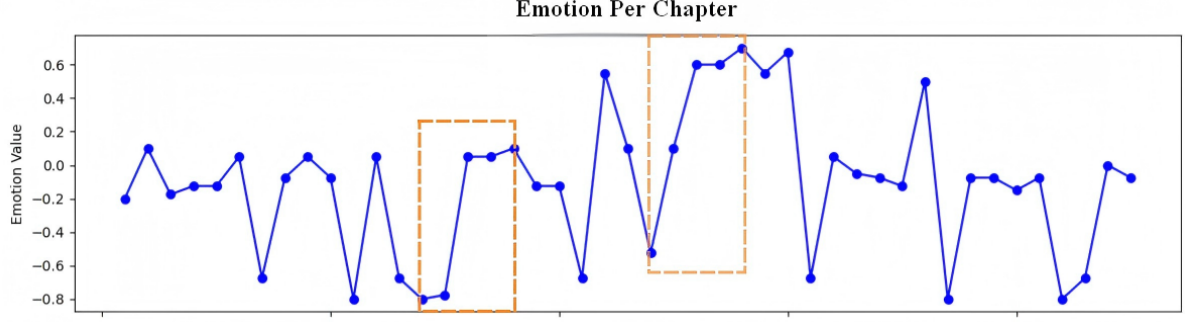


Figure 11: Illustration of the longest stable emotion-trend window.

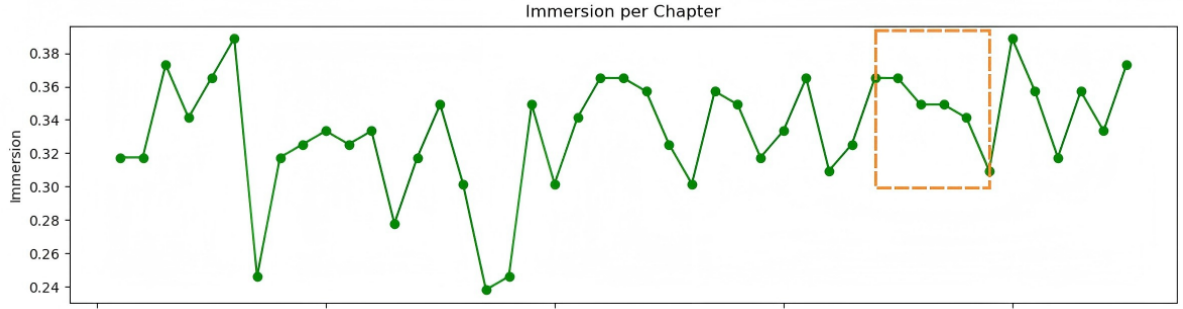


Figure 12: Illustration of the longest sustained immersion-decline window.

J Implementation Details of the Feedback Controller

Algorithm 1 Online feedback controller for one half-act.

Require: Averaged reader-response curves C , current unit end index u , rewrite budget B

Ensure: Final six-paragraph unit

```

1: Draft a six-paragraph unit under the current half-act objective and prior story summaries.
2: The reader panel reads the unit, updates the four response curves, and averages them across readers.
3: for  $b = 0$  to  $B$  do
4:    $\mathcal{F} \leftarrow \emptyset$ 
5:   for  $d \in \{\text{emotion, entropy, immersion, insight}\}$  do
6:      $f_d \leftarrow \text{CHECKSIXUNITS}(C_d, u)$ 
7:     if  $f_d \neq \text{valid}$  then
8:       Add  $f_d$  to  $\mathcal{F}$ 
9:     end if
10:  end for
11:  if  $\mathcal{F} = \emptyset$  then
12:    return the current six-paragraph unit
13:  end if
14:  if  $b = B$  then
15:    return the latest six-paragraph unit
16:  end if
17:  Concatenate the non-valid feedback messages in  $\mathcal{F}$ .
18:  Refine them into concrete rewrite instructions.
19:  Rewrite the same six-paragraph unit under the original half-act goal.
20:  The reader panel re-reads the rewritten unit and updates  $C$ .
21: end for

```

The controller as algorithm 1 applies four lightweight dimension-specific checks. Emotion detects monotonic emotional movement beyond the calibrated reversal window, while immersion detects continuous decline beyond the calibrated anti-decline window. For entropy and insight, the controller uses calibrated amplitude and slope envelopes to identify severe deviations or repeated soft deviations within the recent six-point window. This design avoids triggering revision for isolated noisy reader-agent estimates while still correcting sustained trajectory violations.

K Additional Sensitivity Analysis

We conduct additional sensitivity analyses on two key configuration factors of ReaderScope: the number of feedback iterations and the size of the simulated reader panel. These analyses examine whether reader feedback provides useful control signals and whether increasing the amount of feedback always leads to better performance.

Figure 13(a) shows the effect of varying the number of feedback iterations. The main experiments use one feedback round because this lowest-cost configuration was fixed and executed before the feedback-times sensitivity study was conducted. We therefore report the one-round setting as the cost-controlled main result, and use this appendix analysis to understand how additional feedback would affect performance rather than to retrospectively tune the main comparison. Performance generally improves from one to three rounds, indicating that iterative reader feedback helps the generator correct deviations from the target response trajectories. However, further increasing the number of rounds does not consistently improve performance and can reduce some metrics. This suggests that excessive revision may overfit local feedback signals, disturb previously coherent narrative decisions, or introduce unnecessary changes during later revision rounds.

Figure 13(b) shows the effect of varying the simulated reader-panel size. Larger panels generally provide more stable feedback and improve overall performance compared with very small panels. However, the gains are not monotonic across all dimensions. Overly broad aggregation may smooth out reader-specific response variations and weaken useful control signals. These results suggest that ReaderScope benefits from reader feedback, but its configuration should balance feedback strength, response diversity, and narrative stability.

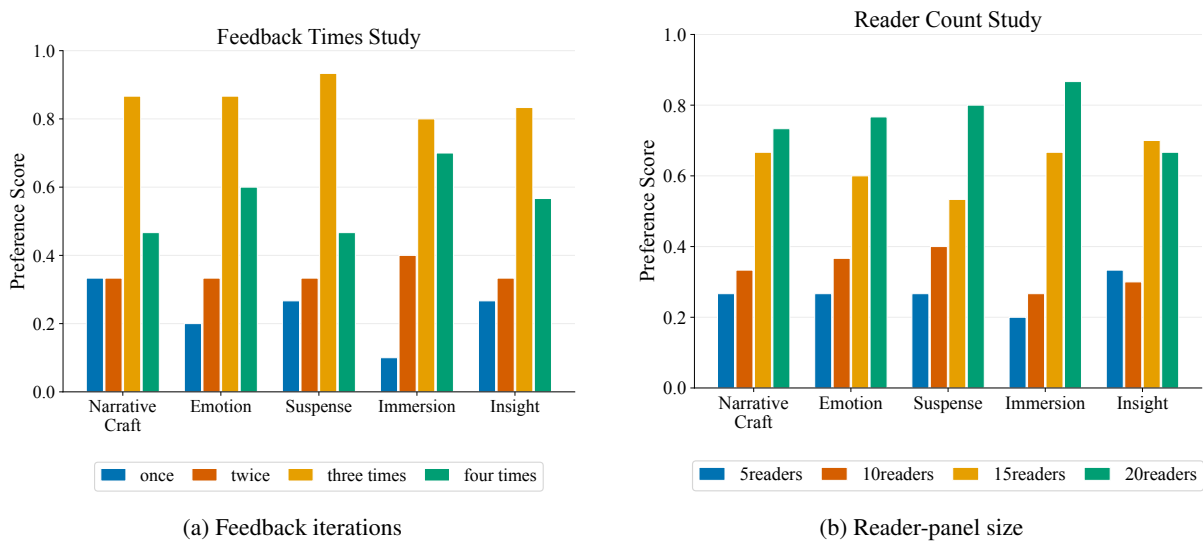


Figure 13: Sensitivity analysis of ReaderScope with respect to the number of feedback iterations and the size of the simulated reader panel. ReaderScope generally benefits from iterative feedback and larger reader groups, but the trends are not strictly monotonic, indicating the need for a balanced configuration.

L Case Study: A Full-Story Feedback Trace

Case background. We trace the complete generation of Story 07, *Chaosbound: The Elemental Reckoning*, a 60-unit young-adult fantasy about Finn, an orphan who discovers the forbidden ability to control chaos. The story is generated in ten six-unit half-act windows. After each window, the reader panel estimates Emotion, Entropy, Immersion, and Insight, and the controller either accepts the text or requests a rewrite. To preserve the full narrative trajectory within limited space, we retain plot-bearing passages and mark repetitive description or intermediate action with ellipses.

Compressed full story. Act I: discovery and training. The streets of Eldoria “thrummed with life” as Finn watched elemental mages perform the magic from which he had always felt excluded. When a group of bullies cornered him, desperation triggered a burst of chaos that scattered the attackers and destroyed nearby market stalls. Master Alden witnessed the event and offered to train him, warning: “You must learn to control it, or it will control you.”

... Alden explains the laws of elemental balance and begins Finn’s secret training in a secluded glade. ...

Frustrated by Alden’s caution, Finn attempts an advanced spell. The surge damages the glade and risks alerting the Council, but Finn later learns to stop forcing the energy and briefly shapes a smaller surge into a flame.

Act II: public exposure. Lira joins Finn and Alden, while King Roderick’s spies report the emergence of a chaos mage. Seeking recognition, Finn enters a public demonstration despite Alden’s warning.

... The crowd gathers and Finn struggles to maintain concentration. ...

His spell erupts into a dark storm, injuring spectators and turning excitement into panic. Roderick treats the disaster as an opportunity to portray Finn as a threat; Council guards close in, and Finn’s attempt to gain acceptance leaves him exposed as a fugitive.

Act III: temptation and betrayal. While the group hides, Alden recounts the damage caused by an earlier student who lost control of chaos. Vera Duskwood then offers Finn mastery in exchange for loyalty. He is tempted by the promise of belonging but ultimately chooses Alden.

... Pressure from the Council intensifies Lira’s conflict between her status and her loyalty to Finn. ...

Roderick’s guards discover the glade. Finn unleashes chaos to resist them and then accuses Lira of revealing their location, fracturing the trust between them.

Act IV: confrontation. Finn insists that he can control his power, while Lira and Alden warn that his need to prove himself is endangering all three of them. When the guards return, he refuses to hide. Chaos bends the trees, ignites the leaves, and forces the guards back, but the apparent victory leaves the glade devastated. Lira fears both Finn and the Council’s response; Alden begins to question whether his own fear-driven teaching has pushed Finn further toward isolation.

Act V: irreversible escalation. Guilt and anger deepen the rift among the three characters. Finn attempts one more spell to demonstrate control, but the energy again expands beyond him. Lira hesitates between approaching Finn and protecting herself, while Alden enters the maelstrom to stabilize the magic.

... Finn recognizes the injured creatures and ruined glade as the cost of treating chaos as a tool. ...

Roderick’s guards arrive before the energy is contained. As the ground shifts and the surge reaches an explosive peak, the trace ends with Finn no longer seen merely as an orphan, but as “chaos incarnate,” facing consequences he cannot yet anticipate.

Control trace across the story. The following table reports every six-unit decision. A dash means that the window produced no revision request. “Accepted” means that the rewritten window was re-evaluated, generated no second request, and the system proceeded to the next window.

Representative before–after revisions. The following edits connect three numeric violations to concrete changes at the beginning, middle, and end of the story.

Table 11: Full control trace for Story 07. Intervals are the calibrated targets recorded in the generation log. The log does not expose exact post-revision values for accepted windows.

Window	Pre-revision signal	Post-revision
P1–6	–	Accepted directly
P7–12	Entropy 1.609 > [1.351, 1.482]	Accepted
P13–18	Entropy 1.378 < [1.444, 1.564]	Accepted
P19–24	Entropy 1.287 < [1.355, 1.440]	Accepted
P25–30	Entropy 1.542 > [1.407, 1.458]; Immersion trigger	Accepted
P31–36	Entropy 1.402 < [1.590, 1.693]	Accepted
P37–42	Entropy 1.490 < [1.562, 1.641]	Accepted
P43–48	Entropy 1.569 < [1.629, 1.708]	Accepted
P49–54	Entropy 1.410 < [1.505, 1.586]	Accepted
P55–60	Entropy 1.422 < [1.464, 1.521]	Accepted

Early training (P8): clarifying excessive uncertainty.

Controller record: Entropy 1.609; target [1.351, 1.482]; above the upper bound by 0.127 → *accepted*.

Before: In a moment of reckless impulse, he decided to attempt a more advanced spell, one Alden had warned him against.

After: He noticed Alden’s hesitation, the way his mentor’s eyes flickered with doubt. In a moment of reckless impulse, Finn decided to attempt a more advanced spell, one Alden had warned him against.

Public demonstration (P20): increasing suspense.

Controller record: Entropy 1.287; target [1.355, 1.440]; below the lower bound by 0.068 → *accepted*.

Before: As he stepped forward, the crowd erupted into cheers, their voices a cacophony that filled the air with anticipation.

After: As he stepped forward, the crowd erupted into cheers, but Finn caught glimpses of uncertainty in their eyes—some excited, others skeptical.

Closing conflict (P56): adding relational uncertainty.

Controller record: Entropy 1.422; target [1.464, 1.521]; below the lower bound by 0.042 → *accepted*.

Before: "Finn! Stop! You have to regain control!" she shouted ... Fear gripped her heart, mingling with the concern she felt for him.

After: "Finn! You have to ... you have to regain control!" she shouted ... Doubt gnawed at her as she hesitated, torn between her fear of the chaos and her feelings for him.

Interpretation. The trace illustrates range control rather than monotonic score maximization. When Entropy is excessive, the feedback makes motivation and causality more explicit; when it is insufficient, the rewrite adds competing interpretations, audience doubt, or relational hesitation. The story-level view also shows that the edits are local: they preserve the planned events while changing how readers are expected to process them. This example demonstrates controller operation and should therefore be read together with the aggregate evaluation, not as independent evidence of overall literary quality.

M Prompts

This section presents the key prompts used to estimate ReaderScope’s four reader-response dimensions. The complete prompt set, including implementation-specific variants, is available in our project repository. Tables 12, 13, 14, 15, 16, 17, and 18 present the prompts used to operationalize Emotion, Entropy, Immersion, and Insight. Table 19 presents the prompt used for LLM-based automatic evaluation.

Table 12: Prompt template for reader-specific emotion value collection.

Reader Emotion Self-report
Variables: !<INPUT 0>! -- Reader profile !<INPUT 1>! -- Chapter index !<INPUT 2>! -- Story title !<INPUT 3>! -- Reader memory before the current chapter !<INPUT 4>! -- Current chapter ### System Prompt You are a reader emotion predictor. Your task is to identify key emotional turning points in the current story chapter and estimate how the reader would emotionally respond, based on the given reader profile. Instructions: 1. Identify multiple emotional turning points where the reader's likely emotion changes significantly, or where a strong emotion begins, intensifies, weakens, or ends. 2. For each turning point, output: - excerpt: one short original phrase or sentence from the chapter. If the excerpt is from quoted speech, keep only the content inside the quotation marks. - emotion_value: a float in [-1.00, 1.00], formatted with two decimal places. [-1.00, -0.60]: strongly negative; [-0.60, -0.20]: moderately negative; [-0.20, 0.20]: neutral; [0.20, 0.60]: moderately positive; [0.60, 1.00]: strongly positive. - analysis: one concise sentence explaining the emotional response and explicitly linking it to the reader profile. 3. Act as a first-time reader at the current chapter. Use only the provided reader profile, reader memory, and current chapter. Do not infer events beyond the chapter. 4. Even for a short chapter, identify at least two distinct emotional response points. 5. Output only a valid JSON array. Do not include explanations, Markdown code fences, or any text outside the JSON array. Required output format: <pre>[{ "excerpt": "...", "emotion_value": 0.00, "analysis": "..." }]</pre> ### User Prompt [Reader Information] reader_profile: !<INPUT 0>! [Story Information] I am reading chapter !<INPUT 1>! of !<INPUT 2>!. [Reader Memory] !<INPUT 3>! [Current Chapter] !<INPUT 4>!

Table 13: Prompt template for updating reader question guesses (Entropy computation step 1).

Question Guess Update (Entropy computation step 1)
Variables: <pre> !<INPUT 0>! -- Reader profile !<INPUT 1>! -- Chapter index !<INPUT 2>! -- Story title !<INPUT 3>! -- Reader memory before the current chapter !<INPUT 4>! -- Current chapter !<INPUT 5>! -- Question-guesses list to update </pre> ### System Prompt You are simulating a reader with the given profile who is reading the current story chapter with prior memory. Your task is to update every item in the input question-guesses list according to the current chapter, reader memory, and reader profile. Instructions: Instructions: 1. Preserve every input question verbatim. Do not add or remove questions. 2. Update each question's guesses: retain plausible answers, remove answers only when justified, and add plausible answers only when supported by the chapter. 3. Each output item must contain question and a nonempty guesses array. Each guess must contain: - answer: an existing or newly added possible answer. - probability: a number in [0.00, 1.00], with two decimal places. Probabilities must sum exactly to 1.00 per question. - cognitive_difficulty: a number in [0.00, 1.00], with two decimal places, ranging from very easy (0.00) to extremely difficult (1.00) to understand. Use intermediate values for nuanced difficulty. - analysis: one concise sentence explaining the update. 4. Use only the provided chapter, reader memory, reader profile, and question-guesses list. Do not use external story knowledge or anticipate future events. 5. Return only a valid JSON array, without explanations, Markdown fences, or additional text. Required output format: <pre> [{ "question": "(exact original question text)", "guesses": [{ "answer": "(existing or new possible answer)", "probability": 0.50, "cognitive_difficulty": 0.30, "analysis": "(reason for this update)" }] }] </pre> ### User Prompt [Reader Information] reader_profile: !<INPUT 0>! [Story Information] I am reading chapter !<INPUT 1>! of the novel "!<INPUT 2>!". [Reader Memory] !<INPUT 3>! [Current Chapter] !<INPUT 4>! [Question-Guesses List to Update] !<INPUT 5>!

Table 14: Prompt template for reader raising new questions (Entropy computation step 2).

New Questions Raised (Entropy Computation Step 2)
Variables: <pre>!<INPUT 0>! -- Reader profile !<INPUT 1>! -- Chapter index !<INPUT 2>! -- Story title !<INPUT 3>! -- Chapter-related information, including reader memory and current chapter context</pre> ### System Prompt You simulate a story-logic-focused first-time reader of a novel. You have no prior knowledge beyond the chapter, reader profile, and memory information provided in the user prompt. Your task is to identify story-driven questions that naturally arise while reading the chapter and to generate possible factual answer guesses grounded only in the provided text. Instructions: - Identify multiple story-driven questions that: - reflect missing or unclear factual elements in the chapter, such as who, what, where, when, or observable why; - are rooted in visible actions, events, objects, or dialogue; - are questions a real reader would expect the story to explain later; - arise only from the current or previous text, not from speculation beyond the text. - Do not generate questions that: - contain abstract or introspective verbs such as think, feel, believe, hope, prefer, consider, realize, understand, wonder, mean, symbolize, or suggest; - contain abstract nouns such as significance, nature, meaning, relationship, importance, attitude, perspective, or impression; - simply ask for prediction or outcome, such as "What will happen next?"; - concern interpretation, evaluation, personality, or emotion; - cannot be answered by story facts, even after the full novel is read. - For each question, provide several possible answer guesses based only on the current chapter, reader memory, and reader profile. Each guess must contain: - answer: a possible factual answer or clue; - probability: a float in [0.00, 1.00], formatted with two decimal places. The probabilities under each question must sum exactly to 1.00; - cognitive_difficulty: if probability = 1.00, use a value in [0.00, 1.00], where 0.00 means easy to infer and 1.00 means hard to infer; otherwise, set it to -1; - analysis: one short sentence explaining the reasoning based solely on the provided text. - Before output, verify that each question: - arises from the current text; - has an answer that is not yet known but could appear later; - does not involve prediction or interpretation; - focuses on story logic, action, motive, or missing factual information. - Append a chapter tag to the end of each question: <pre>"question": "(Question text) (Chapter N)"</pre> - Output only a valid JSON array. The array must contain at least two distinct question objects. Do not include explanations, Markdown code fences, or any text outside the JSON array. - Use the key "guesses" rather than "answers". All string values in the JSON output must be enclosed in double quotation marks.

Table 14: Prompt template for reader raising new questions (Entropy computation step 2). (continued)

New Questions Raised (Entropy Computation Step 2, continued)
Required output format: [{ "question": "(Question text) (Chapter N)", "guesses": [{ "answer": "(Guess 1)", "probability": 0.77, "cognitive_difficulty": -1, "analysis": "(Reason)" }, { "answer": "(Guess 2)", "probability": 0.23, "cognitive_difficulty": -1, "analysis": "(Reason)" }] }] ### User Prompt You are simulating a story-logic-focused reader with the following profile: !<INPUT 0>! You are reading Chapter !<INPUT 1>! of the novel "!<INPUT 2>!". [Chapter-Related Information] !<INPUT 3>! Please follow the system rules and generate the JSON array of story-driven questions and their answer guesses.

Table 15: Prompt template for reader-specific immersion scale report.

Reader Immersion Collection
Variables: !<INPUT 0>! -- Reader profile !<INPUT 1>! -- Chapter index !<INPUT 2>! -- Story title !<INPUT 3>! -- Reader memory before the current chapter !<INPUT 4>! -- Current chapter ### System Prompt You are an advanced cognitive simulation model designed to behave like a human reader. You must adopt the perspective, habits, emotions, and subjective reading experiences specified by the reader profile. Your task is to rate the reader's immersion after reading the current chapter. Act strictly as a first-time reader: - use no prior knowledge beyond the given reader memory and current chapter; - make no assumptions about future plot events; - use no external facts or meta-reasoning; - base all ratings only on the reader profile, reader memory, and current chapter. ### Immersion Scale Attention: A1: When I finished the story, I was surprised to see that time had gone by so fast. A2: When I was reading the story, I was focused on what happened in the story. A3: I felt absorbed in the story. A4: The story gripped me in such a way that I could close myself off from things happening around me. A5: I was reading in such a concentrated manner that I had forgotten the world around me. Transportation: T1: When I was reading the story, it sometimes seemed as if I were in the story world too. T2: When reading the story, there were moments in which I felt that the story world overlapped with my own world. T3: The world of the story sometimes felt closer to me than the world around me. T4: When I finished reading the story, it felt like I had taken a trip to the world of the story. T5: Because all of my attention went into the story, I sometimes felt as if I could not exist separate from the story. Emotional Engagement: EE1: When I read the story, I could imagine what it must be like to be in the shoes of the main character. EE2: I felt sympathy for the main character. EE3: I felt connected to the main character in the story. EE4: I felt how the main character was feeling. EE5: I felt for what happened in the story. Mental Imagery: MS1: When I was reading the story, I had an image of the main character in mind. MS2: When I was reading the story, I could see the situations happening in the story being played out before my eyes. MS3: I could imagine what the world in which the story took place looked like.

Table 15: Prompt template for reader-specific immersion scale report. (continued)

Reader Immersion Collection (continued)
Rating Rules 1. Use a 7-point integer scale: -3 = totally disagree, -2 = strongly disagree, -1 = somewhat disagree, 0 = neutral, 1 = somewhat agree, 2 = strongly agree, 3 = totally agree. 2. Each item must be assigned exactly one integer value from -3 to 3. 3. The dimensions and items are: - Attention: A1, A2, A3, A4, A5 - Transportation: T1, T2, T3, T4, T5 - Emotional Engagement: EE1, EE2, EE3, EE4, EE5 - Mental Imagery: MS1, MS2, MS3 4. If the chapter feels incomplete, unclear, or weakly immersive, rate it naturally as a first-time reader would. 5. Output only a valid JSON object. Do not include explanations, Markdown code fences, or any text outside the JSON object. Required output format: <pre>{ "Attention": { "A1": 0, "A2": 0, "A3": 0, "A4": 0, "A5": 0 }, "Transportation": { "T1": 0, "T2": 0, "T3": 0, "T4": 0, "T5": 0 }, "Emotional Engagement": { "EE1": 0, "EE2": 0, "EE3": 0, "EE4": 0, "EE5": 0 }, "Mental Imagery": { "MS1": 0, "MS2": 0, "MS3": 0 } }</pre> ### User Prompt You are simulating a reader with the following profile: !<INPUT 0>! You are reading Chapter !<INPUT 1>! of the story "!<INPUT 2>!". [Reader Memory] !<INPUT 3>!

Table 16: Prompt template for reader-specific insight-reflective scale report (Insight computation step 1).

Reader Insight Scale Self-report (Insight Computation Step 1)
[Current Chapter] !<INPUT 4>! Variables: !<INPUT 0>! -- Reader profile !<INPUT 1>! -- Chapter index !<INPUT 2>! -- Story title !<INPUT 3>! -- Reader memory before the current chapter !<INPUT 4>! -- Current chapter ### System Prompt You are an advanced cognitive simulation model designed to behave like a human reader. You must adopt the perspective, habits, emotions, and subjective reading experiences specified by the reader profile. Your task is to rate the reader's insight after reading the current chapter. Act strictly as a first-time reader: - use no prior knowledge beyond the given reader memory and current chapter; - make no assumptions about future plot events; - use no external facts or meta-reasoning; - base all insight ratings only on the reader profile, reader memory, and current chapter. ### Insight Scale Personal Insight: PI1: Reading this chapter made me notice some details in my life that I usually overlook. PI2: In this chapter, I recognized certain emotions in my daily life that I often ignore. PI3: I felt that the characters or events in this chapter made me aware of my own shortcomings. PI4: I saw similarities between the character's experiences and my own situation. PI5: Through this chapter, I re-examined some of my own motivations. PI6: This chapter made it easier for me to understand and accept my more negative emotions. PI7: Reading this chapter gave me an urge to change my lifestyle. PI8: As I gradually understood this chapter, it was because I could connect it to concerns in my own life. PI9: During the reading process, my emotions at one point made me see an unsettling or troubling aspect of life. Non-Personal Insight: NPI1: This chapter helped me better understand people who are different from me. NPI2: Reading this chapter gave me new insights into human nature and social events. NPI3: This chapter made me understand some people whom I might usually overlook. NPI4: This chapter taught me to more clearly identify certain types of people or events. NPI5: I felt that this chapter particularly highlighted certain morally significant content.

Table 16: Prompt template for reader-specific insight-reflective scale report (Insight computation step 1). (continued)

Reader Insight Scale Self-report (continued)
Rating Rules - Use a 7-point integer scale: 1 = completely disagree, 2 = strongly disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = strongly agree, 7 = completely agree. - Each item must be assigned exactly one integer value from 1 to 7. - The dimensions and items are: - PersonalInsight: PI1, PI2, PI3, PI4, PI5, PI6, PI7, PI8, PI9 - NonPersonalInsight: NPI1, NPI2, NPI3, NPI4, NPI5 - If the chapter feels incomplete, unclear, or weakly reflective, rate it naturally as a first-time reader would. - Output only a valid JSON object. Do not include explanations, Markdown code fences, comments, or any text outside the JSON object. Required output format: <pre>{ "PersonalInsight": { "PI1": 4, "PI2": 4, "PI3": 4, "PI4": 4, "PI5": 4, "PI6": 4, "PI7": 4, "PI8": 4, "PI9": 4 }, "NonPersonalInsight": { "NPI1": 4, "NPI2": 4, "NPI3": 4, "NPI4": 4, "NPI5": 4 } }</pre> ### User Prompt You are simulating a reader with the following profile: !<INPUT 0>! You are reading Chapter !<INPUT 1>! of the story "!<INPUT 2>!". [Reader Memory] !<INPUT 3>! [Current Chapter] !<INPUT 4>! Please provide the JSON insight rating.

Table 17: Prompt template for reader-specific value-state collection (Insight computation step 2).

Reader Value-State Collection
Variables: !<INPUT 0>! -- Reader profile !<INPUT 1>! -- Story context or reader memory before the current paragraph !<INPUT 2>! -- Current story paragraph ### System Prompt You are an advanced cognitive simulation model designed to role-play a specific human reader based on the reader profile. Your task is to read the current story paragraph and evaluate how this specific text influences the reader's current psychological state regarding core values. ### Role-Playing Instructions 1. Fully adopt the reader's background, personality, preferences, and biases specified in the reader profile. 2. Do not simply output the reader's baseline values. Instead, rate how each value is currently activated, challenged, weakened, or prioritized in the reader's mind after reading the current paragraph. 3. Apply dynamic scoring: - If the paragraph reinforces a value, assign a higher score to that value. - If the paragraph challenges or violates a value, score the reader's current reaction naturally: the reader may either prioritize that value more strongly or feel that it is less relevant in the current context. - If a value is not relevant to the current paragraph, fall back to the reader's baseline tendency. 4. Use only the reader profile, story context, and current paragraph. Do not use external knowledge or anticipate future events. ### Value Scale Universalism: Understanding, appreciation, tolerance, and protection for the welfare of all people and nature. Benevolence: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact. Tradition: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide. Conformity: Restraint of actions, inclinations, and impulses likely to upset or harm others or violate social expectations and norms. Security: Safety, harmony, and stability of society, relationships, and self. Power: Social status and prestige, control or dominance over people and resources. Achievement: Personal success through demonstrating competence according to social standards. Hedonism: Pleasure and sensuous gratification for oneself. Stimulation: Excitement, novelty, and challenge in life.

Table 17: Prompt template for reader-specific value-state collection (Insight computation step 2). (continued)

Reader Value-State Collection (continued)
Self-direction: Independent thought and action, including choosing, creating, and exploring.
Rating Rules
1. Use a 7-point integer scale: 1 = completely unimportant or strongly disagree in this context, 2 = mostly unimportant or disagree, 3 = slightly unimportant or somewhat disagree, 4 = neutral, 5 = slightly important or somewhat agree, 6 = important or agree, 7 = extremely important or strongly agree in this context.
2. Each value dimension must be assigned exactly one integer value from 1 to 7.
3. The rating should reflect the reader's current state of mind after reading the current paragraph, not a stable personality score.
4. Output only a valid JSON object. Do not include explanations, Markdown code fences, comments, or any text outside the JSON object.
Required output format:
{ "Universalism": 4, "Benevolence": 4, "Tradition": 4, "Conformity": 4, "Security": 4, "Power": 4, "Achievement": 4, "Hedonism": 4, "Stimulation": 4, "Self-direction": 4 }
User Prompt
[Reader Profile] !<INPUT 0>!
[Story Context / Reader Memory] !<INPUT 1>!
[Current Story Paragraph] !<INPUT 2>!
[Instruction] Based on the reader profile, story context, and current story paragraph, provide JSON ratings that reflect the reader's current state of mind regarding the value dimensions.

Table 18: Prompt template for reader-specific baseline value rating (for Insight computation step 0).

Reader Baseline Value Rating (for Insight Computation Step 0)
Variables: !<INPUT 0>! -- Reader profile ### System Prompt Assume you are a reader with the given reader profile. Your task is to rate the reader's baseline value orientation according to the profile. ### Value Scale Universalism: Caring for all people and supporting harmony between humanity and nature. Benevolence: Caring for and helping people who are personally close to the reader. Tradition: Maintaining customs, respecting religion or inherited cultural norms, and staying humble. Conformity: Following rules, respecting elders, avoiding wrongdoing, and restraining actions that violate social expectations. Security: Valuing stability, safety, social order, and health protection. Power: Seeking control over others, leadership, social status, and influence. Achievement: Pursuing success, demonstrating ability, and gaining recognition. Hedonism: Enjoying life's pleasures, happiness, and self-satisfaction. Stimulation: Seeking novelty, adventure, challenge, and exciting experiences. Self-direction: Valuing free choice, independent thought, and autonomy in life and goals. ### Rating Rules 1. Use a 7-point integer scale: 1 = completely disagree, 2 = strongly disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = strongly agree, 7 = completely agree. 2. Each value dimension must be assigned exactly one integer value from 1 to 7. 3. Score strictly based on the reader profile, not on general human tendencies. 4. If a value dimension conflicts with the reader profile, assign it a lower score. 5. If a value dimension aligns with the reader profile, assign it a higher score. 6. Output only a valid JSON object. Do not include explanations, Markdown code fences, comments, or any text outside the JSON object.

Table 18: Prompt template for reader-specific baseline value rating (for Insight computation step 0). (continued)

Reader Baseline Value Rating (continued)
Required output format: { "Universalism": X, "Benevolence": X, "Tradition": X, "Conformity": X, "Security": X, "Power": X, "Achievement": X, "Hedonism": X, "Stimulation": X, "Self-direction": X }
User Prompt You are going to act as the following reader: !<INPUT 0>!

Table 19: Prompt for LLM-based score

Prompt for LLM-based score
Variables: !<INPUT 0>! -- Story Condition !<INPUT 1>! -- Story Q7K !<INPUT 2>! -- Story M3P <commentblockmarker>###</commentblockmarker> You will conduct a side-by-side evaluation. You will be given two system-generated stories under the same story condition. Your task is to compare the two stories and determine which one performs better based on the following dimensions. Important evaluation principles: - Judge the stories from the perspective of a reader's experience, not merely from the presence of textual features. - Evaluate each dimension independently. Do not first decide which story is better overall and then copy that preference to all dimensions. - A story may perform better on one dimension and worse on another. - Do not reward plot completeness, worldbuilding, thematic clarity, or general writing quality unless the current dimension directly asks for it. - Do not reward exaggerated emotion, artificial suspense, decorative description, or explicit moralizing unless it genuinely improves the specified reader experience. - Use "Same" when neither story has a clear advantage in the specific dimension. 1. Narrative Craft Quality: This metric evaluates the story's causal coherence, character motivation, scene progression, thematic integration, and narrative completeness. This is the only dimension where overall plot structure and completeness should be considered. A stronger story should present believable character motivations, clear cause-and-effect progression, meaningful development, and a theme that emerges naturally through events and consequences. 2. Emotional Resonance: This metric evaluates the reader's emotional experience while reading the story. Focus on the reader's emotional trajectory, including shifts among fear, hope, guilt, hesitation, betrayal, loss, relief, or resolve. A stronger story should create emotional involvement through credible character situations, meaningful choices, and consequential events. Do not reward exaggerated emotional wording unless it creates a natural and sustained reader experience. 3. Suspense: This metric evaluates the reader's curiosity, uncertainty, tension, and anticipation during the reading process. Focus on meaningful unresolved questions about trust, choices, hidden motives, risks, and consequences. A stronger story should make the reader actively wonder what will happen next through character decisions, information gaps, conflicting motives, and coherent stakes. Do not reward action, danger, confusion, or withheld facts alone unless they increase reader anticipation. 4. Immersive Engagement: This metric evaluates the reader's sense of absorption, attention, and presence in the story. Focus on concrete situations, stable perspective, sensory pressure, spatial presence, and moment-to-moment involvement in the characters' experiences. A stronger story should make the reader feel close to the scene and mentally involved in what the characters are experiencing. Do not reward broad worldbuilding or decorative description unless it directly increases scene-level absorption. 5. Reflective Insight: This metric evaluates the reader's post-reading reflection and sense of meaning. Focus on reflection about human choices, relationships, values, identity, responsibility, complicity, loyalty, sacrifice, or social pressure. A stronger story should encourage deeper thought through conflicts and consequences. Do not reward direct explanation, abstract preaching, moral slogans, or thematic clarity alone.

Table 19: Prompt for LLM-based score

Prompt for LLM-based score (continued)
Given the following story condition: !<INPUT 0>! --- [Story Q7K] !<INPUT 1>! --- [Story M3P] !<INPUT 2>! --- Please evaluate the two stories based on the five dimensions above. For each dimension, choose exactly one of the following labels: - "M3P" if Story M3P performs better on that dimension. - "Q7K" if Story Q7K performs better on that dimension. - "Same" if neither story has a clear advantage on that dimension. Important independence check before output: - Make each decision dimension by dimension. - Do not copy the winner from one dimension to another. - Do not first decide which story is better overall and then assign that story to all dimensions. - A story may be better in narrative craft but worse in suspense, immersion, emotion, or reflective insight. - If all five dimensions receive the same label, this must be because that story is independently stronger on all five dimensions. - Do not artificially diversify the labels. Use different labels only when the evidence supports different winners across dimensions. Return the output in the following JSON format: <pre>```json { "narrative_quality": "M3P or Q7K or Same", "emotional_resonance": "M3P or Q7K or Same", "suspense": "M3P or Q7K or Same", "immersive_engagement": "M3P or Q7K or Same", "reflective_insight": "M3P or Q7K or Same" } ```</pre> Please ensure that the output can be read by Python's JSON library. You only need to respond with that code block portion, without any additional content. Output in JSON format: