CEPO: CAUSAL ENTROPY POLICY OPTIMIZATION FOR LONG-FORM NARRATIVE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated remarkable capabilities in open-ended text generation; however, generating ultra-long novels with long-range causal dependencies, stable character personalities, and coherent narrative arcs remains a substantial challenge. The standard autoregressive generation process is prone to accumulating local inconsistencies, and conventional sequence-level reinforcement learning (e.g., GRPO) suffers from severe credit assignment problems when applied to long-form texts comprising tens of thousands of tokens. To address these challenges, we propose a hierarchical agent-based framework for novel generation that decomposes long-form writing into multiple verifiable stages, utilizing synopses, blueprints, a memory manager, and a Narrative World Graph. Furthermore, to tackle reward sparsity in the optimization of extended texts, we introduce a dense token-level reward mechanism guided by an LLM evaluator. This mechanism translates qualitative narrative feedback into a causal separation of coherent prefixes and contaminated suffixes. Building upon this mechanism, we propose Causal Entropy Policy Optimization (CEPO). By combining suffix-level penalties with an entropy-guided modulation mechanism, CEPO provides a heuristic approximation of token-level contributions around localized failure points, thereby effectively reducing the variance and bias of credit assignment in the generation of ultra-long narratives. Extensive automated and human evaluations on a large-scale dataset of online web novels demonstrate that, despite requiring significantly fewer parameters, the proposed method is competitive with strong baselines on six evaluated dimensions. Consequently, this work establishes an effective paradigm for the high-quality alignment and generation of complex, ultra-long narratives.

1 INTRODUCTION

Large language models (LLMs) enable creative-writing applications such as story drafting, script writing, and novel generation (Mirowski et al., 2023; Fan et al., 2018). Yet generating a complete novel remains harder than producing a short story or single response. Beyond fluent prose, it requires a persistent world model, stable characters, long-range causality, chapter-level rhythm, and meaningful progression, making novel generation a challenging testbed for long-context reasoning, planning, and alignment.

Recent advances in long-context LLMs (Liu et al., 2024b) have partially alleviated context-window limitations by allowing models to condition on tens or hundreds of thousands of tokens. Nevertheless, a longer input window does not solve the output-side difficulty of novel generation. Each response remains far shorter than a full novel, requiring generation through a sequence of partial outputs. This process readily accumulates four categories of macro-structural logical errors: *entity-consistency errors*, such as contradictions in character or object attributes; *outline-alignment errors*, such as omitted plot points, reversed causality, or unauthorized ending changes; *hard-constraint violations*, including shifts in point of view or tense and forbidden terms; and *spatiotemporal-logic errors*, such as impossible movements or timeline conflicts. These failures can propagate across chapters and destabilize later generation. Therefore, long-form novel generation requires not only long-context conditioning, but also explicit mechanisms for global planning, local continuity, and iterative verification.

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

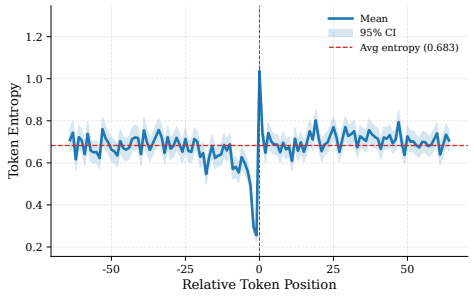


Figure 1: Token entropy dynamics around localized narrative errors. Failed generation trajectories produced by the Qwen3-8B model ($N = 2048$) are aligned by the first token of the LLM-identified problematic sentence, designated as relative position 0.

To address this challenge, we propose a hierarchical novel-generation agent that decomposes the creative process into multiple verifiable stages. Starting from a one-sentence inspiration, the Idea Agent expands the seed into a high-level synopsis that captures the central dramatic premise and global narrative objective. The Blueprint Agent then constructs a structured blueprint consisting of world settings, timelines, character profiles, stylistic constraints, and chapter-level outlines. The Writer Agent generates the complete novel chapter by chapter, while the Memory Manager maintains both the global blueprint and a localized history window of recent chapters. In parallel, the World Graph records evolving entities, relations, events, and state transitions (Park et al., 2023; Zhou et al., 2023), and the Critic Agent checks whether the generated content is causally plausible, faithful to the outline, consistent with the world graph, and coherent with previous chapters.

While such an agentic pipeline improves controllability, it does not by itself optimize the underlying model's generation policy. This is especially important for long-form fiction, where quality is often determined by sparse but pivotal textual decisions. A single character-breaking sentence, an unsupported causal transition, or a premature revelation may severely damage the reading experience, even though most tokens in the same chapter remain fluent and reasonable. Standard reinforcement learning methods such as PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024) typically assign a scalar reward to the entire generated response. When applied to long novel chapters containing thousands of tokens, this coarse sequence-level feedback leads to severe credit assignment errors (Poddar et al., 2024; Lightman et al., 2024): tokens unrelated to the failure may be over-penalized, while the few tokens that actually trigger narrative breakdown may not receive sufficiently precise learning signals.

We therefore introduce a reinforcement learning framework for narrative-aware token-level credit assignment. Instead of treating the generated chapter as uniformly good or bad, we use an LLM-based rubric evaluator to assess narrative quality and localize problematic spans. These rubric-guided judgments are converted into token-level reward signals, allowing the policy to reinforce expressions that improve coherence, engagement, and literary quality while penalizing localized violations of narrative logic. However, external error localization alone is insufficient; it overlooks the model's internal distributional shifts that accompany narrative collapse (Kuhn et al., 2023). To investigate this, we analyze token entropy dynamics in failed generation trajectories: we employ a powerful LLM judge to identify the onset of problematic spans within generated novel passages, then compute the token entropy in a neighborhood around these identified positions, with t=0 denoting the first token of a flagged error. As illustrated in Figure 1, we hypothesize a three-phase empirical signature:

- **Increasing Commitment ($t < 0$):** Declining entropy may reflect commitment to a trajectory that later proves inconsistent.
- **Predictive-Uncertainty Spike ($t = 0$):** The entropy spike correlates with increased uncertainty at the localized error onset.
- **Post-Error Uncertainty ($t > 0$):** Elevated entropy may arise from conditioning on a corrupted prefix.

Motivated by links between entropy and model confidence (Stolfo et al., 2024; Xiong et al., 2024), this is not a causal claim: narrative branching and context corruption are alternative explanations. Appendix E evaluates category-specific and null-baseline results.

Motivated by both dense external feedback and internal entropy dynamics, we further propose Causal Entropy Policy Optimization (CEPO), which uses entropy-guided modulation to estimate the relative contribution of tokens around structural failure points. The resulting optimization objective effectively reduces the variance and bias of credit assignment in ultra-long creative generation.

Our contributions are summarized as follows:

- We introduce a critic-guided reinforcement learning framework for long-form generation agents, in which a Critic Agent provides localized feedback that directly optimizes the Writer Agent's parameters.
- Through entropy analysis and a systematic comparison of sequence-level, span-level, and suffix-level strategies, we propose suffix-level credit assignment for long-form generation, preserving coherent prefixes while penalizing continuations conditioned on an identified error.
- We develop an entropy-modulated token-level credit-assignment mechanism that dynamically scales optimization weights using predictive uncertainty, enabling targeted correction around narrative failures while preserving generation diversity.

2 Related work

Ultra-Long Text and Narrative Generation Generating ultra-long narratives is challenging for LLMs because it requires long-term coherence and plot consistency. While RoPE (Su et al., 2024) scaling, RingAttention (Liu et al., 2024a), and memory architectures (Packer et al., 2023) expand context length, they cannot guarantee narrative quality. To address logical structuring, systems like RecurrentGPT (Zhou et al., 2023), Dramatron (Mirowski et al., 2023), Ex3 (Lei et al., 2024), StoryWriter (Xia et al., 2025), and StoryAgent (Hu et al., 2024) adopt hierarchical frameworks decomposing generation into planning and drafting stages. Although these inference-time strategies improve coherence, they rely primarily on zero-shot capabilities. Unlike our approach, they do not directly update the model parameters to improve narrative planning or long-range consistency.

Token-Level Reinforcement Learning for LLMs Standard RLHF frameworks like PPO (Schulman et al., 2017) and DPO (Rafailov et al., 2023) align LLMs predominantly through sequence-level rewards. For ultra-long text, this holistic evaluation creates severe reward sparsity and credit assignment problems, hindering the identification of specific tokens causing plot holes. To distribute signals densely, recent works explore token-level RLHF (Poddar et al., 2024), segment-level DPO (Kong et al., 2025), and dense rewards (Chan et al., 2024; Li et al., 2024; Yoon et al., 2024), alongside Process Reward Models (Lightman et al., 2024) for mathematical reasoning. Despite these advances, token-level RL remains unadapted for creative writing, where token validity depends on overarching contextual contributions rather than immediate accuracy. Our approach bridges this gap by proposing a token-level RL framework providing dense, narrative-aware reward signals to resolve the credit assignment bottleneck in long-text generation.

3 Novel Generation Agent

As illustrated in Figure 2, we implement long-form novel generation as a collaboration among specialized sub-agents. Each sub-agent handles one level of narrative abstraction and exposes a clear input-output interface to the next stage. This design avoids asking a single LLM call to simultaneously invent the global plot, maintain the world state, remember previous chapters, and produce polished prose. Instead, the system externalizes the writing process into easily inspectable and verifiable intermediate artifacts: a synopsis, a blueprint, chapter drafts, context memories, a world graph, and verifier feedback.

At a high level, the generation process can be summarized as:

$$\text{Inspiration} \rightarrow \text{Synopsis} \rightarrow \text{Blueprint} \rightarrow \text{WorldGraph} \rightarrow \text{ChapterBlock} \rightarrow \text{Critique},$$

where each ChapterBlock contains one or more consecutive chapters depending on the underlying model's context length and output budget. The generated block then updates the memory and world graph before the next block is written.

Idea Agent. The Idea Agent is responsible for the transition from a minimal inspiration to a novel-level intention. Given a one-sentence seed, it produces a high-level synopsis that summarizes the central premise, main conflict, emotional trajectory, and expected payoff of the story. The synopsis is intentionally compact: it does not enumerate every event, but defines the global direction of the novel. In this way, the system first fixes what the story is about before deciding how the story should unfold.

Blueprint Agent. The Blueprint Agent turns the synopsis into a structured writing plan. It produces two complementary artifacts: a story setting and a chapter outline. The setting specifies the world background, timeline, character profiles, personality constraints, relationships, and style requirements.

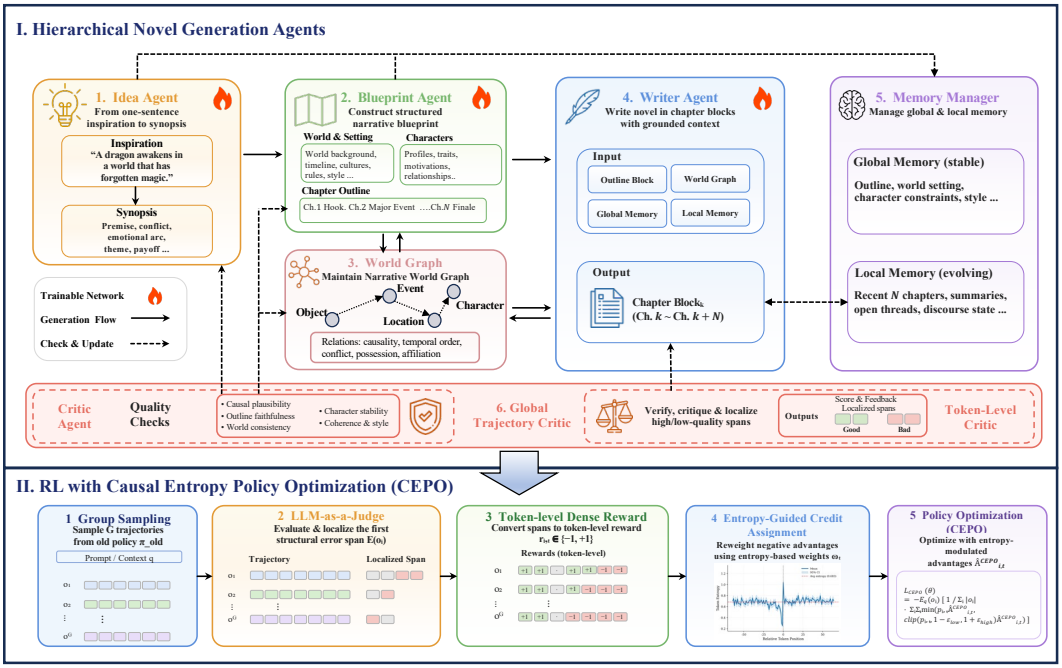


Figure 2: Overview of the proposed framework. A set of specialized sub-agents expands a one-sentence inspiration into a synopsis, a structured narrative blueprint, and chapter-level novel, while the Memory, World, and Critic Agents maintain continuity and provide localized feedback for reinforcement learning.

The outline specifies the main hook of the novel and the major event of each chapter in a short textual form. Since each chapter outline is concise, it acts as a stable control signal that keeps the long narrative aligned with the intended plot while leaving enough freedom for expressive writing.

World Graph. The World Graph maintains a structured representation of the evolving fictional world. After each generated segment or after a fixed number of chapters, it extracts newly introduced characters, locations, objects, events, relations, and state changes from the prose. These elements are merged into a Narrative World Graph, where nodes represent narrative entities and edges represent relations such as causality, temporal order, affiliation, conflict, possession, and emotional dependency. Compared with textual memory alone, this graph provides a more explicit reference for detecting contradictions and tracking how the fictional world changes over time.

Writer Agent. The Writer Agent expands the chapter outline into full paragraphs. It writes the novel sequentially in chapter blocks rather than necessarily one chapter at a time. Each block may contain K consecutive chapters, where K is determined by the available context window, the expected output length, and the amount of historical information that must be retained. For each block, the Writer Agent receives the corresponding chapter outlines, relevant global settings, the current world graph, and recent local memory. Its objective is not only to generate fluent language, but also to faithfully realize the planned events, preserve character behavior, and maintain the rhythm of the current narrative arc across multiple adjacent chapters.

Memory Manager. The Memory Manager manages the information provided to the Writer Agent. It maintains two types of memory. The global memory stores information that should remain stable throughout the novel, including the complete outline, world settings, character constraints, and style requirements. The local memory stores information that should evolve with the story, including the previous N chapters, recent summaries, unresolved conflicts, and short-term discourse states. By combining global and local memory, the system preserves long-range consistency while ensuring that adjacent chapters remain coherent.

Critic Agent. The Critic Agent verifies the outputs of other sub-agents. At the synopsis and blueprint stages, it checks whether the planned story is causally plausible, whether character motivations are coherent, and whether the chapter outline supports the intended narrative arc. At the chapter-writing stage, it checks whether the generated paragraph faithfully realizes the assigned

outline, respects the current world graph, and remains consistent with previous chapters. When a failure is detected, the Critic Agent can either trigger revision or record the problematic span as supervision for later policy optimization.

Agentic Feedback Loop. The novelty of this agentic design lies in treating novel generation as an iterative write–remember–verify process rather than a one-pass decoding problem. The Idea and Blueprint Agents define the global narrative target, the Writer Agent realizes it locally, the Memory Manager and World Graph preserve textual and symbolic continuity, and the Critic Agent closes the loop by identifying inconsistencies and localized quality failures. This feedback loop improves inference-time controllability and also provides the training signal used in the next section. In particular, the Critic Agent can localize both successful and problematic spans, which allows us to convert qualitative narrative judgments into dense token-level rewards for reinforcement learning.

4 Causal Entropy Policy Optimization (CEPO)

4.1 Reinforcement Learning with Verifiable Rewards

Formally, we consider an LLM parameterized by θ. Given a dataset of query prompts Q, standard RLVR learns a policy π_θ to minimize the objective: $\mathcal{L}_{\mathrm{RLVR}}(\theta) = -\mathbb{E}_{q\sim Q, o\sim\pi_\theta(\cdot|q)}[r(o,q)]$, where $r(o,q)$ is a verifiable reward function that evaluates the correctness of the response o given the query q. To enhance training stability and efficiency, GRPO (Shao et al., 2024) samples a group of responses $\{o_i\}_{i=1}^{G}$ for each query from the old policy $\pi_{\theta_{\mathrm{old}}}$ and optimizes the following surrogate objective loss (Shao et al., 2024; Yu et al., 2025):

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}_{q\sim Q,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}\cdot\hat{A}_{i,t}, \mathrm{clip}(\rho_{i,t}, 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}})\cdot\hat{A}_{i,t}\Big)\right], \quad (1)$$

where $\hat{A}_{i,t} = \frac{r_i - \mathrm{mean}(\{r_i\}_{i=1}^{G})}{\mathrm{std}(\{r_i\}_{i=1}^{G})}$ represents the group advantage, normalizing the outcome reward r_i. $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t}|q,o_{i,<t})}$ denotes the importance sampling ratio at step t of trajectory o_i, and $\epsilon_{\text{high/low}}$ indicates the clipping range of importance ratios.

4.2 Dense Credit Assignment via LLM-as-a-Judge

In long-form narrative generation, the objective is to maximize an implicit human preference function that evaluates coherence and overall quality. Since this function is unobservable, we employ an LLM-based evaluator as a scalable proxy.

In this task, logical inconsistency is a subtle yet critical issue that severely degrades generation quality. Although the LLM judge can also localize high-quality spans, we do not introduce additional span-specific weights for them in the final objective. A localized positive span provides a reliable credit signal mainly when it appears before the first structural failure, which occurs relatively infrequently. In addition, positive judgments in creative writing are more subjective than the localization of logical or narrative inconsistencies, so we focus primarily on the parameter updates for negative samples. Specifically, the LLM judge identifies the first critical position where logical consistency or narrative coherence breaks down. Due to the causal nature of autoregressive generation, any token produced after this initial failure is conditioned on a corrupted prefix. Consequently, we classify all subsequent tokens as a contaminated suffix.

Accordingly, we model the latent preference by decomposing it into global fluency and a structural penalty applied strictly to this contaminated suffix. This reflects the empirical observation that a single localized narrative breakdown usually triggers a cascading effect, compounding quality degradation across all downstream tokens.

Suffix-Level Reward as a Surrogate Risk Estimator. For a given model output $o = (o_1, \ldots, o_T)$, the evaluator identifies the first critical position t^* where the breakdown occurs. Based on this detected error onset, we formulate a dense token-level reward specifically for negative samples:

$$r_t = \begin{cases} -1, & t \geq t^*(o) \\ +1, & t < t^*(o) \end{cases} \quad (2)$$

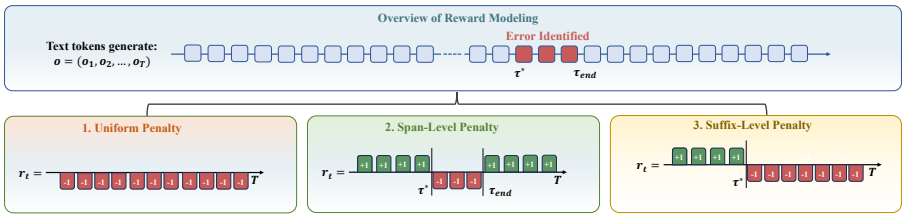


Figure 3: Illustration of uniform, span-level, and suffix-level penalty under the LLM-as-a-Judge framework.

This suffix-level assignment provides a tractable surrogate for the otherwise intractable structural risk. It yields a binary signal strictly aligned with the causal structure of the trajectory: tokens generated prior to the breakdown are deemed coherent, whereas all tokens from t^* onwards are penalized holistically as a causally contaminated continuation. This design is fundamentally grounded in the autoregressive generation process: once the conditioning context is corrupted, subsequent tokens lose their statistical independence from the initial error and thus cannot be reliably evaluated in isolation.

Comparison among Uniform, Span-Level, and Suffix-Level Penalties. Two natural alternatives fail to capture the causal structure of narrative failure. As shown in Figure 3, a uniform trajectory penalty, assigning negative credit to every token in a flawed output, systematically biases credit assignment by penalizing tokens generated under a clean prefix. In contrast, a span-level penalty that only penalizes a short window around t^* underestimates the true extent of contamination, as downstream tokens, although syntactically fluent, are conditioned on a corrupted context and should not be reinforced. The suffix-level scheme lies between these extremes and is more faithful to the autoregressive causal graph: the pre-error prefix is preserved as a positive learning signal, while the entire post-error continuation is uniformly suppressed. This transforms sparse sequence-level supervision into structured, causality-aware token-level feedback, yielding a lower-variance and less biased proxy for the latent preference function. To this end, we conduct comparative experiments, and the detailed experimental results are presented in Appendix D.

4.3 Entropy-Guided Credit Assignment

For negative samples, the suffix-level reward introduced in Section 4.2 provides a causally-aligned decomposition that separates the coherent prefix from the contaminated suffix, yet it implicitly assigns a uniform penalty magnitude to every token within the suffix and a uniform reward to every token within the prefix. In practice, however, the marginal contribution of individual tokens is highly non-uniform. As shown in Figure 1, within the prefix, tokens produced under misleading overconfidence may already commit the policy to a flawed reasoning path, while within the suffix, tokens generated under high-entropy confusion carry qualitatively different error signatures from those produced immediately after the breakdown. Treating these regimes uniformly therefore risks reinforcing deceptive confidence on one side and diluting the learning signal on the other. To recover this heterogeneity, we define the modulated advantage as

$$\hat{A}_t^{\text{CEPO}} = \omega_t \cdot A_t, \tag{3}$$

where A_t is the raw advantage induced by the suffix-level reward of Section 4.2 and ω_t is a token-level entropy modulation weight. In what follows, we construct ω_t in a progressive manner, motivated by empirical observations on failed trajectories and guided by two inductive biases: (i) the policy's confidence relative to a global reference should shape the magnitude of credit, and (ii) causal responsibility for an error decays smoothly with distance from the point of breakdown.

Entropy-Guided Credit Modulation. We first identify an empirical modulation anchor. Given the raw advantage sequence $\{A_t\}_{t=1}^T$, we define the *zero point* t^* as the first position where the advantage shifts from positive to negative, i.e., $A_{t^*-1} > 0$ and $A_{t^*} < 0$. As shown in Figure 1, the model exhibits a structured entropy evolution pattern around t^*: a gradual decline before t^* (overconfidence), a sharp spike at t^* (distribution shift), and a sustained plateau thereafter (confused generation). This consistent pattern indicates that tokens in different temporal regions exhibit significant heterogeneity in their contributions to the ultimate failure.

To characterize this phenomenon, we contrast the token-level entropy $H_t = H(\pi_\theta(\cdot \mid s_t))$ with the sequence-level average entropy $\bar{H}_{\text{seq}} = \frac{1}{T}\sum_{t=1}^T H_t$. This provides a stable global reference

to identify local anomalous deviations: $H_t \ll \bar{H}_{\text{seq}}$ indicates overconfidence, whereas $H_t \gg \bar{H}_{\text{seq}}$ signifies confusion. These deviations are particularly pronounced around the logical breakdown.

Building upon this, we further introduce a temporal decay to encode the locality of causal responsibility; that is, the impact of a token on the error should diminish as its distance from t^* increases. Integrating the entropy deviations with temporal locality yields the final weighting function:

$$\omega_t = \begin{cases} \left[1 + \alpha \cdot \max\left(0, \frac{\bar{H}_{\text{seq}} - H_t}{\bar{H}_{\text{seq}}}\right) \cdot \exp\left(-\frac{|t^* - t|}{\tau}\right)\right]^{-1}, & t < t^* \\ \beta, & t = t^* \\ 1 + \gamma \cdot \max\left(0, \frac{H_t - \bar{H}_{\text{seq}}}{\bar{H}_{\text{seq}}}\right) \cdot \exp\left(-\frac{|t^* - t|}{\tau}\right), & t > t^* \end{cases} \tag{4}$$

where $\alpha, \gamma > 0$ control the scaling of entropy deviations, $\beta > 1$ amplifies the penalty at the error onset, and τ dictates the range of temporal decay. Due to the smoothing mechanism introduced by τ, CEPO is expected to be less sensitive to minor localization errors of t^*.

This design assigns distinct roles to the three temporal regions. Before the error, low-entropy tokens are discounted to suppress reinforcement of an increasingly committed but flawed path; at the localized onset, the root-error token receives a strong penalty; afterward, penalties for high-entropy tokens are amplified so that the contaminated continuation is not treated as weak evidence. Temporal decay returns ω_t smoothly to 1 away from the error, restricting modulation to its local neighborhood. We then mini-batch-whiten $\hat{A}_t^{\text{CEPO}} = \omega_t A_t$. Since whitening is the affine transform $(\hat{A}_t^{\text{CEPO}} - \mu)/\sigma$, it preserves the relative ordering induced by ω_t rather than removing the entropy signal, while improving optimization stability.

Final Optimization Objective. Finally, we integrate the proposed CEPO into the policy optimization framework. Let $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$ denote the importance sampling ratio for token t in the i-th trajectory. For a prompt q, given a group of G trajectories $\{o_i\}_{i=1}^G$ sampled from the old policy $\pi_{\theta_{\text{old}}}$, the overall objective is to minimize the clipped surrogate loss:

$$\mathcal{L}_{\text{CEPO}}(\theta) = -\mathbb{E}_{q\sim Q, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min\Big(\rho_{i,t} \cdot \hat{A}_{i,t}^{\text{CEPO}}, \text{clip}(\rho_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}) \cdot \hat{A}_{i,t}^{\text{CEPO}} \Big) \right], \tag{5}$$

where $\hat{A}_{i,t}^{\text{CEPO}}$ is our entropy-modulated advantage for token t in trajectory i, and $\epsilon_{\text{low}}, \epsilon_{\text{high}}$ are the clipping thresholds that prevent destructively large policy updates. By replacing standard uniform advantages with $\hat{A}_{i,t}^{\text{CEPO}}$, the policy is specifically regularized to avoid confident failure paths and aggressively penalize the exact onsets of structural breakdown.

5 Experiment

5.1 Experiment Setup

We collect 81,496 licensed web novels across 12 genres and split them into 80,396 SFT, 1,000 RL, and 100 test examples. We SFT the Qwen3-32B (Yang et al., 2025) Idea and Blueprint Agents and apply RL only to the Writer Agent. RL uses ground-truth plans, but the suffix reward and ω_t depend only on Writer generations and rewards; moreover, these plans were produced by Gemini-2.5-Flash summarization, reducing the train–inference gap. Both critics use Gemini-2.5-Flash-Lite. Unless noted otherwise, training follows DAPO (Yu et al., 2025) in VeRL (Sheng et al., 2025), with $\alpha = 1$, $\beta = 2$, $\gamma = 1$, $\tau = 100$, and maximum response length 8,192. Appendix A provides further setup details; Appendix B analyzes critic stability.

5.2 End-to-End Main Results

Evaluation Criteria and Methods. We evaluate web-novel generation quality along six dimensions: Pacing for smooth and balanced plot progression, Appeal for sustaining attention through conflicts, suspense, and climaxes, Completeness for world, character, and spatiotemporal construction, Coherence for logical and character consistency, Style for natural and grammatically varied language, and Hook for early conflict and suspense that attract readers. We apply both automatic and human evaluation to validate the effectiveness of our framework. Additionally, we report the average token count of the generated novel bodies.

Model	Method	Agent	Length	Pacing	Appeal	Completeness	Coherence	Style	Hook	Average
Gemini-2.5-Flash	-	×	8k	3.52	3.45	3.66	3.73	3.61	4.86	3.81
Qwen3-32B	-	×	2k	2.22	2.06	2.12	2.30	3.61	4.89	2.87
GLM-5.1	-	✓	39k	3.38	3.42	3.24	2.65	3.07	4.71	3.41
DeepSeek-R1	-	✓	33k	3.60	3.87	3.39	2.82	**3.69**	4.78	3.70
Gemini-2.5-Flash	-	✓	31k	<u>4.29</u>	**4.13**	**4.11**	**3.48**	3.31	4.80	<u>4.02</u>
Qwen3-8B	-	✓	33k	1.75	1.90	2.38	1.68	2.29	4.25	2.38
Qwen3-8B	GRPO	✓	31k	2.71	2.78	2.82	2.07	2.85	4.83	3.01
Qwen3-8B	DAPO	✓	36k	2.65	2.83	2.71	2.22	2.84	4.72	2.99
Qwen3-8B	A3PO	✓	36k	2.72	2.76	2.95	2.11	2.93	4.82	3.05
Qwen3-8B	CEPO	✓	33k	3.04	2.91	3.12	2.22	3.00	4.89	3.20
Qwen3-32B	-	✓	33k	2.51	2.62	2.91	2.05	2.96	4.63	2.94
Qwen3-32B	DAPO	✓	36k	3.74	3.68	3.52	3.11	3.31	<u>4.93</u>	3.71
Qwen3-32B	CEPO	✓	34k	**4.30**	<u>4.01</u>	<u>3.98</u>	<u>3.25</u>	<u>3.63</u>	**4.96**	**4.03**

Table 1: Automated evaluation (length in body-text tokens; best/second-best agent results in bold/underline). We report means without standard deviations because scores are ordinal $(1/3/5)$; Appendix H gives Mann–Whitney U tests.

Comparison	Pacing	Appeal	Completeness	Coherence	Style	Hook
Qwen3-32B vs Ours	22.0/**91.0**	14.0/**94.0**	29.0/**93.0**	34.0/**84.0**	27.0/**84.0**	40.0/**87.0**
GLM-5.1 vs Ours	34.0/**86.0**	21.0/**88.0**	41.0/**85.0**	40.0/**75.0**	34.0/**78.0**	60.0/**88.0**
DeepSeek-R1 vs Ours	90.0/**98.0**	53.0/**82.0**	31.0/**83.0**	62.0/**85.0**	73.0/**83.0**	66.0/**96.0**
Gemini-2.5-Flash vs Ours	64.0/**84.0**	**80.0**/77.0	77.0/**98.0**	**71.0**/70.0	66.0/**86.0**	82.0/**91.0**

Table 2: Human evaluation results. Each cell reports the baseline score and our Qwen3-32B CEPO score in the format "baseline/ours". Values indicate the percentage of annotators who judged the generated novels to satisfy each criterion.

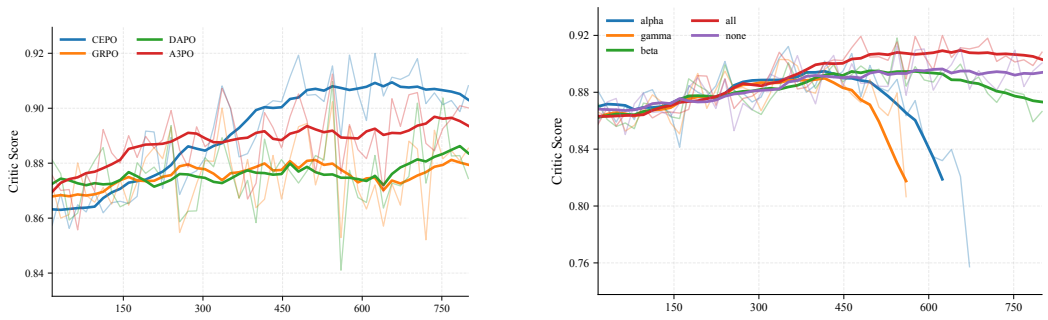
Automated Evaluation. To comprehensively quantify the performance gap, we design an automated evaluation protocol for the aforementioned evaluation criteria. For each novel, we employ LLMs to assign scores according to these criteria. For each criterion, the LLM outputs a score of 1, 3, or 5. To avoid potential bias where an LLM may favor its own responses, for each criterion of a novel, we simultaneously use Gemini-2.5-Flash, DeepSeek-R1, and GLM-5.1 to produce scores, and report the mean as the final result for that criterion, as shown in Table 1. Detailed scoring rubrics for each numerical level are provided in Appendix K.

Human Evaluation. To complement the automated evaluation and further validate the alignment with human perception, we also conduct a human evaluation. Annotators are presented with a premise and two corresponding stories, for each evaluation criterion, they are asked to judge which story is better, or whether both stories are equally good (All Good, AG) or equally bad (All Bad, AB). We then compute the proportion of annotators who consider that the story from a given method satisfies the respective criterion. Formally, the score for Method A is calculated as $(\mathrm{A}+\mathrm{AG})/(\mathrm{A}+\mathrm{B}+\mathrm{AG}+\mathrm{AB})$, and the score for Method B is calculated as $(\mathrm{B}+\mathrm{AG})/(\mathrm{A}+\mathrm{B}+\mathrm{AG}+\mathrm{AB})$, under identical experimental conditions, as shown in Table 2. Detailed descriptions of the human evaluation are provided in Appendix L.

Evaluation Result. As shown in Tables 1 and 2, Qwen3-32B CEPO achieves the highest automated-evaluation average among the evaluated systems. The hierarchical agent framework extends output from roughly 2k to 33k tokens, while CEPO improves over other RL methods on the same Qwen3-32B backbone. Relative to non-agent Gemini-2.5-Flash, CEPO obtains significant gains in pacing, appeal, completeness, style, and hook, although Gemini remains stronger in coherence (Appendix H). Thus, the evidence supports competitive overall performance rather than uniform dominance across all dimensions. Human evaluation likewise favors CEPO over the base model and advanced baselines on most criteria. Error-distribution, length-stratified, and training-dynamics analyses appear in Appendices F, J, and C.

5.3 Comparative Experiment

To isolate the contributions of CEPO, we compare its training dynamics on the Qwen3-8B model against GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025), which employ response-level or uniform credit, and A3PO (Tang et al., 2025), which entropy-reweights token advantages but does not separate positive prefixes from negative suffixes within one trajectory. We evaluate performance using a Token-Level Critic score, defined as 1 minus the fraction of low-quality tokens, where higher



(a) Across different RL algorithms (b) Under various modulation weight configurations

Figure 4: Training dynamics of the Token-Level Critic score.

scores indicate superior narrative quality. Figure 4a illustrates their different training dynamics in ultra-long narrative credit assignment.

CEPO differs from A3PO in three respects. First, its weights are anchored by localized Critic Agent feedback rather than entropy alone. Second, its suffix formulation can retain positive credit for a coherent prefix while assigning negative credit to the contaminated continuation. Third, entropy modulation refines these sign-separated advantages near the localized failure. Consequently, the suffix-without-entropy configuration—not A3PO—is the direct isolation of suffix-level credit assignment. Appendix G reports the controlled 2×2 analysis. Baseline algorithms exhibit severe credit dilution and gradient noise. GRPO and DAPO yield near-flat learning curves bounded below 0.89, with extreme batch-level volatility dropping to 0.84. This instability stems from confounded gradient signals under sequence-level or heuristic rewards, which erroneously penalize sound prefixes alongside contaminated suffixes. Consequently, models cannot identify the causal origin of narrative collapse. Although A3PO marginally reduces variance, it converges to a sub-optimal plateau trapped by deceptive high-confidence errors. Conversely, CEPO achieves robust, monotonic convergence. Despite starting from a lower baseline near 0.86, CEPO bypasses optimization bottlenecks and stabilizes above 0.90. This stability supports our hypothesis of CEPO: using token entropy to modulate the advantage $\hat{A}_t$ effectively decouples coherent prefixes from cascading suffix confusion. By penalizing the exact singularity of structural breakdown and suppressing high-entropy noise, CEPO drastically reduces policy gradient variance, guiding the actor network toward causally aligned optimization and higher narrative fidelity.

5.4 Ablation Study

To evaluate the three phases of our piecewise entropy modulation, we ablate parameters α, β, and γ. Figure 4b compares Critic score trajectories of isolated components against the full formulation. Applying either α or γ alone causes severe late-stage performance drops. Using only β mitigates this instability but still slightly degrades final performance. Conversely, combining all weights yields the highest scores and robust stability.

Dynamically, isolating α (suppressing overconfidence where $t < t^*$) or γ (amplifying penalties for high-entropy confusion where $t > t^*$) unbalances updates along the error trajectory, creating excessively sharp or weak gradients that destabilize training. The β weight provides a safe but localized signal, penalizing only the exact collapse point without broader context guidance, thereby slightly underperforming late in training. Operating in tandem, these mechanisms globally balance gradient updates for negative samples, achieving optimal performance.

6 Conclusion

We propose a fine-grained credit assignment framework for long-form narrative generation. Using an LLM-as-a-Judge, we transform sparse supervision into localized token-level rewards via suffix-based decomposition. Our entropy-guided approach (CEPO) adaptively reweights token advantages near critic-identified failures. By addressing token-impact heterogeneity and temporal error locality, CEPO stabilizes policy optimization. Empirically, CEPO improves the overall automated score and most evaluated dimensions, while error breakdowns associate its gains with fewer macro-structural narrative failures. This work establishes a foundation for integrating structured feedback and uncertainty-aware learning in LLM alignment.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, we disclose the use of generative AI tools in the research and preparation of this manuscript. Specifically, AI assistance was utilized in the following capacities:

Writing and Polishing: We employed large language models to aid in refining the academic prose, improving grammatical structures, and enhancing the overall clarity of the manuscript. The application of these tools was strictly confined to stylistic editing; no novel scientific claims, theoretical derivations, or core arguments were generated by AI.

Retrieval and Discovery: AI-assisted search and summarization tools were utilized to facilitate the discovery of relevant literature and streamline the review of related work. While these tools aided in identifying pertinent prior research, the critical analysis, selection, and synthesis of the literature were conducted entirely by the human authors.

Synthetic Dataset Generation and Evaluation: As detailed in Section 5 and the Appendices, we employed LLMs (specifically Qwen3 and Gemini-2.5-Flash-Lite) to construct the hierarchical training pipeline, generate synthetic artifacts (e.g., synopses and blueprints), and produce token-level critic evaluations for our reinforcement learning framework. Additionally, LLMs were utilized within the automated evaluation protocol to score narrative quality.

Verification and Responsibility: All AI-assisted outputs were subject to rigorous human oversight. The synthetic data and token-level critic annotations underwent manual validation by human annotators to ensure precision and reliability, as documented in our stability analysis (Appendix B). Furthermore, the authors meticulously reviewed, edited, and verified all AI-polished text and retrieved citations for accuracy and scientific validity. The authors affirm full and final responsibility for all content, data, analyses, and conclusions presented in this work.

ETHICS STATEMENT

This research involves the collection and utilization of a large-scale web-novel dataset and human evaluation studies, both of which require rigorous ethical oversight. All 81,496 web novels utilized in the training and evaluation sets were procured with full copyright ownership and compliance, ensuring that intellectual property rights are respected and the data is strictly constrained to academic research purposes. For the human evaluation phase, we recruited a total of 10 human annotators to conduct pairwise comparisons. We strictly adhered to ethical labor practices; every annotator was fairly and explicitly compensated for each individual data point they labeled.

Regarding broader societal impacts, we recognize that advancing long-form narrative generation presents dual-use risks. While it democratizes storytelling, it poses potential socio-economic challenges for creative professionals and introduces the risk of generating highly coherent synthetic disinformation. To proactively mitigate these risks, our training corpus underwent rigorous heuristic filtering to eliminate personally identifiable information (PII), toxic language, and inherently biased content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide a detailed description of the CEPO objective, including its token-level causal credit assignment, entropy modulation, temporal decay, advantage whitening, and clipped policy optimization formulation. The model architecture, agent roles, training setup, key hyperparameters, computational resources, and evaluation protocols are documented in the main paper and the appendix. The appendix additionally reports dataset statistics and splits, implementation and training details, the complete Token-Level Critic prompt, automated evaluation rubrics, human-evaluation procedures, statistical significance tests, critic validation, ablation studies, and analyses of entropy dynamics and potential evaluation biases.

REFERENCES

Alex J Chan, Hao Sun, Samuel Holt, and Mihaela Van Der Schaar. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12409–12423, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Huang Lei, Jiaming Guo, Guanhua He, Xishan Zhang, Rui Zhang, Shaohui Peng, Shaoli Liu, and Tianshi Chen. Ex3: Automatic novel writing by extracting, excelsior and expanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9125–9146, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.494.

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. Reinforcement learning with token-level feedback for controllable text generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1704–1719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.111. URL https://aclanthology.org/2024.findings-naacl.111/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581225. URL https://doi.org/10.1145/3544548.3581225.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 52516–52544. Curran Associates, Inc., 2024. doi: 10.52202/079017-1664.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125019–125049. Curran Associates, Inc., 2024. doi: 10.52202/079017-3970.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Rethinking sample polarity in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2512.21625*, 2025.

Haotian Xia, Hao Peng, Yunjia Qi, Bin Xu, Juanzi Li, Hou Lei, and Xiaozhi Wang. Storywriter: A multi-agent framework for long story generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6559–6563, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

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

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Nam, Daejin Jo, Kyoung-Woon On, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang Yoo. Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14969–14981, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi

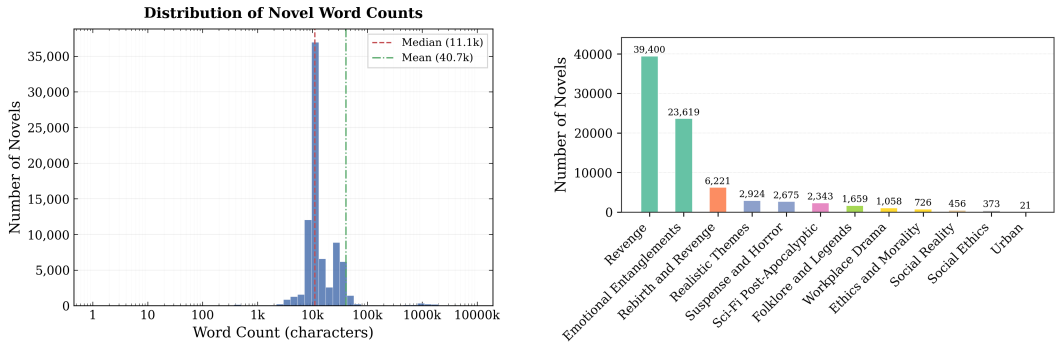
Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*, 2023.

A Additional Experimental Details

A.1 Statistics of Dataset



(a) Distribution of novel word counts.

(b) Distribution of novel genres.

Figure 5: Statistics of the collected web novel dataset, including the distributions of novel word counts and genres.

We collected 81,496 high quality web novels with full copyright ownership and analyzed their word count and genre distributions, as shown in Figure 5. The dataset exhibits a mean word count of 40,713 and a median word count of 11,117 per novel, reflecting the diverse length distribution of the crawled content. In terms of genre coverage, the corpus spans 12 distinct categories, including Emotional Entanglements, Ethics and Morality, Folklore and Legends, Realistic Themes, Rebirth and Revenge, Revenge, Sci-Fi Post-Apocalyptic, Social Ethics, Social Reality, Suspense and Horror, Urban, and Workplace Drama.

A.2 Training Configuration and Implementation Detail

For RL training, each batch contains 64 prompts, with 8 responses sampled per prompt, yielding 512 training samples per iteration. Following the settings of DAPO (Yu et al., 2025), we set the learning rate to 1e-6, $\epsilon_{\text{low}} = 0.2$, and $\epsilon_{\text{high}} = 0.28$. We keep the remaining optimization settings consistent with DAPO.

All experiments are implemented with the VeRL framework (Sheng et al., 2025). Training is conducted on a server with 2 NVIDIA H200 nodes, each containing 8 GPUs, with 4TB total RAM and 368 CPU cores. A full training cycle takes approximately 30 hours for Qwen3-8B and 60 hours for Qwen3-32B.

B Stability Analysis of the Token-Level Critic

Model	Precision	Recall	F1
Gemini-2.5-Flash-Lite	79.53%	64.10%	71.00%
GLM-5.1	94.59%	57.70%	71.68%
DeepSeek-R1	83.83%	28.00%	41.97%
Gemini-2.5-Flash	96.95%	19.10%	31.91%

Table 3: Stability analysis of different models serving as the token-level critic based on human evaluation.

To validate the reliability of the token-level critic, we conduct a manual stability analysis on a sampled dataset of 200 trajectories. Because the evaluation of narrative quality and logical consistency is inherently subjective, we prioritize precision while also reporting recall and F1. Human annotators examined the critic's error spans and the underlying trajectories to determine true and missed errors. As presented in Table 3, Gemini-2.5-Flash-Lite provides the highest recall among the evaluated critics and a competitive F1 score, despite lower precision than GLM-5.1 and Gemini-2.5-Flash. Stronger

critics appear more tolerant of narrative ambiguity: they frequently rationalize minor inconsistencies as creative liberties and therefore flag fewer spans, producing an inverse precision–recall trade-off. We select Gemini-2.5-Flash-Lite because its greater coverage supplies a denser reward signal while retaining 79.53% precision and offering the inference efficiency needed to scale training.

C RL Training Dynamics for Qwen3-32B

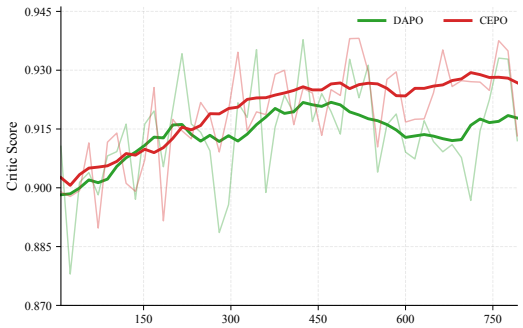


Figure 6: Token-Level Critic score trajectories of CEPO and the DAPO baseline during RL training on the Qwen3-32B backbone.

As illustrated in Figure 6, we compare the training dynamics of our proposed CEPO method against the DAPO baseline. Both methods initialize at a similar performance level (around 0.900) and exhibit initial learning progress. However, as training advances into the mid-to-late stages, their trajectories diverge significantly. The DAPO baseline (green curve) struggles with optimization instability; its performance plateaus early and even shows slight degradation and high variance after 900 steps, struggling to break through the 0.920 threshold.

In contrast, CEPO (red curve) maintains a highly stable and continuous upward trajectory throughout the entire training process. It successfully overcomes the plateau, ultimately converging to a significantly higher Critic score (approaching 0.930). This marked difference demonstrates that CEPO's entropy-guided and suffix-level credit assignment mechanisms provide more precise and stable gradient signals. By effectively penalizing the exact onsets of narrative collapse, CEPO prevents the policy from reinforcing deceptive overconfidence, leading to a more robust optimization for ultra-long narrative generation.

D Comparative Experiments: Uniform, Span-Level, and Suffix-Level Penalties

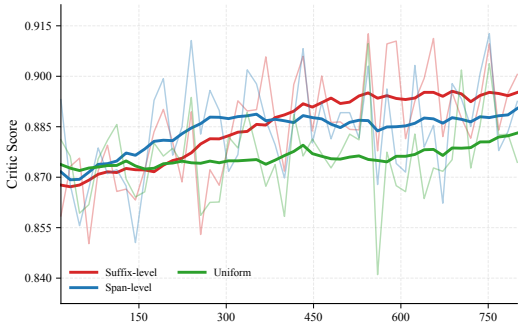


Figure 7: Token-Level Critic score trajectories under uniform, span-level, and suffix-level penalty schemes.

To validate the efficacy of our proposed credit assignment mechanism, we conduct comparative experiments on Qwen3-8B to evaluate three penalty strategies: uniform, span-level, and suffix-level penalties. In the uniform setting, a negative reward of -1 is assigned to every token in the generated trajectory if any error is detected. In the span-level setting, a reward of -1 is assigned exclusively to the tokens within the localized problematic segment. In contrast, our suffix-level strategy applies a uniform -1 reward starting from the first erroneous token and continuing through to the end of the sequence.

As illustrated by the Critic score trajectories in Figure 7, the suffix-level penalty demonstrates significantly faster convergence and achieves the highest performance plateau, followed by the span-level penalty, while the uniform penalty performs the worst.

The uniform penalty struggles because it systematically biases credit assignment by indiscriminately penalizing high-quality tokens generated under a clean prefix prior to the error, thereby sending a confusing and high variance learning signal to the policy. The span-level penalty performs slightly better but is ultimately limited by its flawed assumption of state recoverability. It implicitly assumes that once a problematic segment concludes, the model can seamlessly return to generating a high quality, coherent narrative. However, in the context of long form fiction, an initial logical error, plot hole, or character deviation inherently corrupts the entire subsequent conditioning context. Applying a span-level penalty often induces reward hacking: the policy learns to quickly terminate the penalized flawed sequence merely to resume gathering positive rewards from syntactically fluent but contextually compromised downstream tokens.

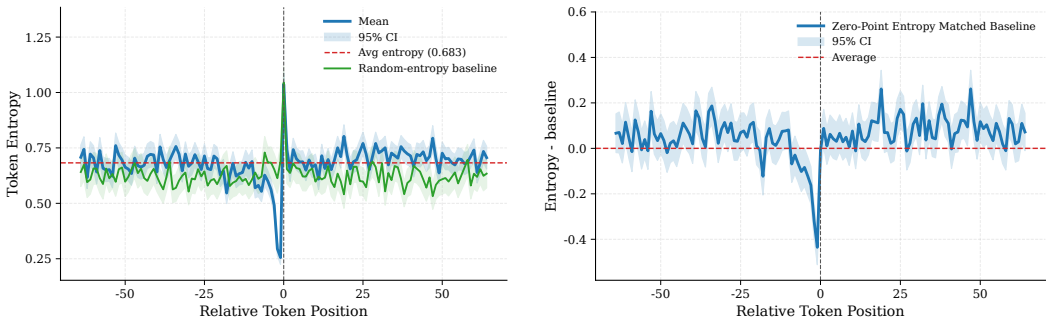
Conversely, the suffix-level penalty effectively models the corrupted context as an absorbing state in the autoregressive generation process. By uniformly penalizing all tokens generated after the breakdown, it preserves the clean prefix as a positive learning signal while propagating a dense, unambiguous negative advantage directly back to the exact decision-making step that triggered the logical collapse. This cascading penalty forces the policy to strictly avoid the initial fatal error, more accurately aligning the token-level optimization with the macro-structural integrity required for ultra-long narratives.

E Token Entropy Dynamics Across Error Categories and Null Baselines

We test whether the entropy signature in Figure 1 is an artifact of aligning trajectories at critic-selected positions. We sample 2,048 Qwen3-8B trajectories, each conditioned on a distinct prompt, and aggregate every localized error (one trajectory can contain multiple errors). The first error occurs at mean token position 911 (median 752), substantially later than short-continuation settings. Inference uses temperature 0.9, top-$p = 0.7$, top-$k = 20$, and repetition penalty 1.03. We compare real errors with three controls: random punctuation boundaries, random offsets in $[-100, 100]$, and non-error tokens matched to the zero-point entropy. For the matched baseline, we select the ten nearest-entropy candidates per error and break ties with seed 42.

The error taxonomy contains four macro-structural classes and two surface-level classes. *Entity Consistency* covers contradictions in character or item attributes; *Outline Alignment* covers missing plot points, reversed causality, or unauthorized ending changes; *Hard Constraints* covers violations of point of view, tense, or forbidden terms; and *Spatiotemporal Logic* covers impossible movements or timeline conflicts. *Repetition* and *Garbage Leakage* capture repeated blocks and leaked prompts, markup, or model meta-discourse, respectively.

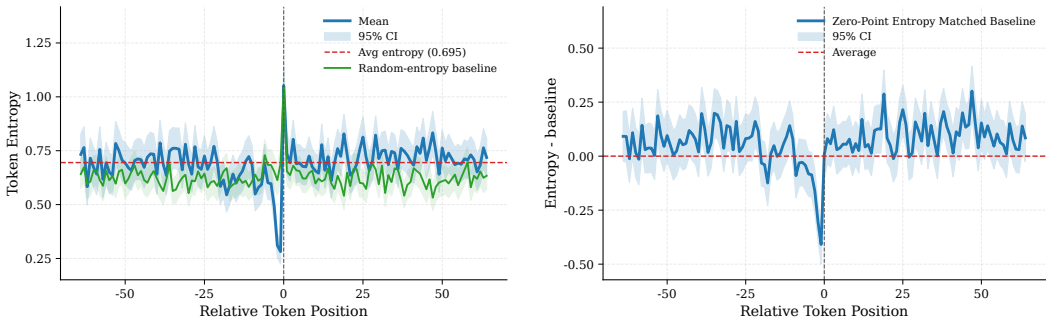
Figure 8 first shows the aggregate pattern across all logical errors: entropy declines locally before the identified error, spikes at the zero point, and remains elevated immediately afterward relative to entropy-matched non-error positions. Figures 9–12 then report the same comparison separately for Entity Consistency, Outline Alignment, Hard Constraints, and Spatiotemporal Logic. In each figure, the left panel compares the error-aligned trajectory with its zero-point-entropy-matched baseline, while the right panel reports their difference. The differential panels remove the zero-point entropy by construction and make the asymmetric local pattern directly visible. Because the magnitude and variance differ by category, we interpret the pattern as a local empirical association rather than a universal mechanistic law.



(a) All logical errors: entropy and matched baseline.

(b) All logical errors: difference.

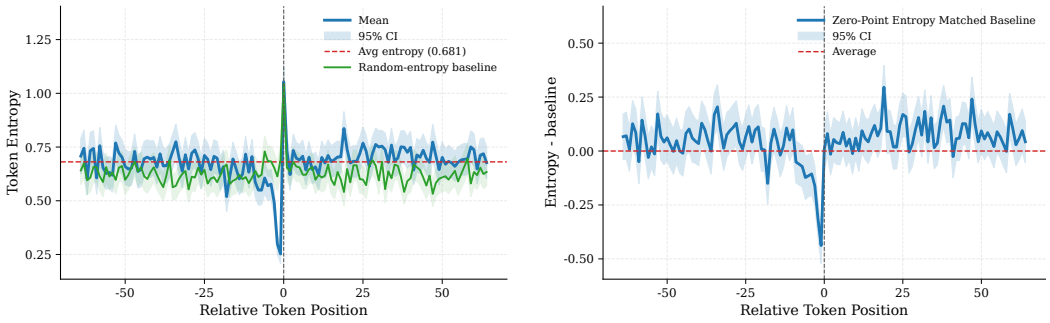
Figure 8: Overall token-entropy dynamics across all four logical-error categories. The left panel compares the error-aligned entropy trajectory with the zero-point-entropy-matched non-error baseline; the right panel shows their difference (error minus matched baseline).



(a) Entropy trajectory and matched baseline.

(b) Difference: error minus matched baseline.

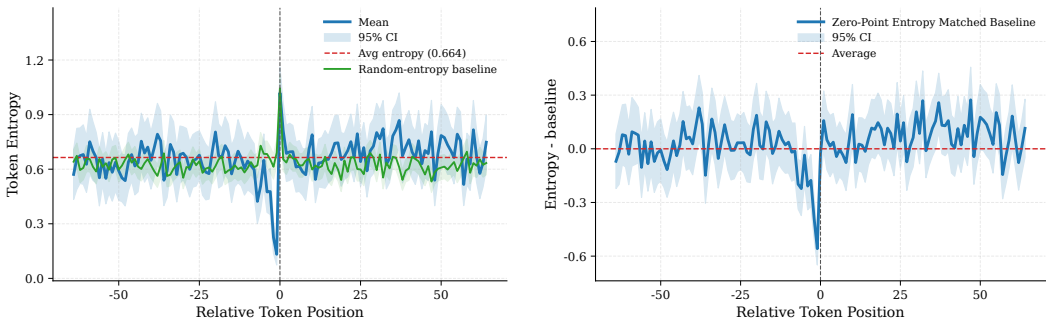
Figure 9: Token-entropy dynamics for Entity Consistency errors.



(a) Entropy trajectory and matched baseline.

(b) Difference: error minus matched baseline.

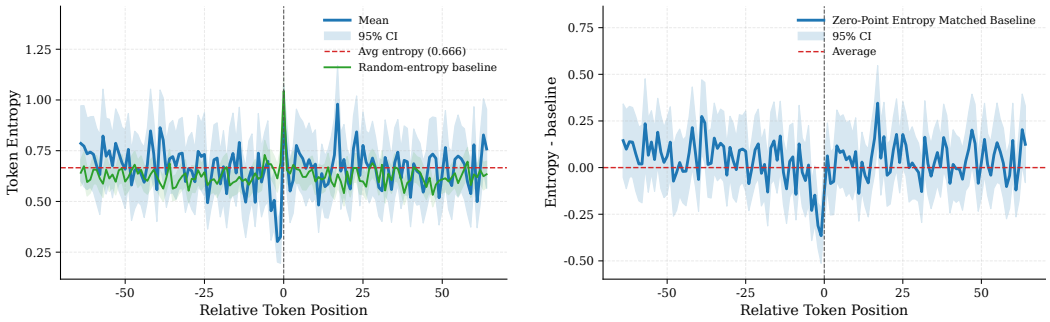
Figure 10: Token-entropy dynamics for Outline Alignment errors.



(a) Entropy trajectory and matched baseline.

(b) Difference: error minus matched baseline.

Figure 11: Token-entropy dynamics for Hard Constraint errors.



(a) Entropy trajectory and matched baseline.

(b) Difference: error minus matched baseline.

Figure 12: Token-entropy dynamics for Spatiotemporal Logic errors.

F Error Distribution and Post-Error Recovery

Method	Total	Entity Consistency	Outline Alignment	Hard Constraints	Spatiotemporal Logic	Repetition	Garbage Leakage
Original Model	2532	757	1412	31	265	53	14
GRPO	1955	568	1159	23	167	30	8
DAPO	1957	583	1143	17	175	28	11
A3PO	1857	431	1103	61	235	21	6
CEPO	1689	394	1085	51	135	20	4

Table 4: Error counts under identical evaluation settings.

To determine what CEPO improves, we compare error frequencies under fixed inference hyperparameters. Repetition and garbage leakage constitute only 4.03% of baseline errors and remain uncommon after every RL method. CEPO instead achieves the largest reductions in Entity Consistency and Spatiotemporal Logic errors and the lowest total error count, suggesting that its gains are associated primarily with macro-structural narrative consistency rather than only surface-level cleanup.

Category	Entity Consistency	Outline Alignment	Hard Constraints	Spatiotemporal Logic	Repetition	Garbage Leakage	ALL
Total Trajectories	1141	78	418	16	53	14	1720
Recovered Trajectories	343	23	127	3	0	0	496
Recovery Rate	30.0%	29.5%	30.4%	18.8%	0.0%	0.0%	28.8%

Table 5: Trajectory-level recovery after the first critic-identified error.

The suffix formulation assumes that a first error tends to contaminate later generation. We test this by asking the Critic Agent whether the same error type recurs after its first occurrence; a trajectory

is counted as recovered only when no later error of that type is detected. Logical-error recovery is low (18.8%–30.4%), while repetition and garbage leakage show no recovery, supporting the contaminated-suffix approximation. This operational definition does not imply that every later token is individually incorrect.

G Factorial Ablation of Entropy and Temporal Decay

We conduct a 2×2 ablation on Qwen3-8B while holding the zero-point weight, advantage normalization, and overall scale fixed. Performance is the Token-Level Critic score, defined as one minus the fraction of low-quality tokens. Entropy alone initially converges quickly but becomes unstable after 400 steps; temporal decay alone is more stable but degrades late. Combining both avoids the late-stage degradation and reaches 0.913, showing that entropy contributes beyond temporal locality while decay regularizes global entropy weighting.

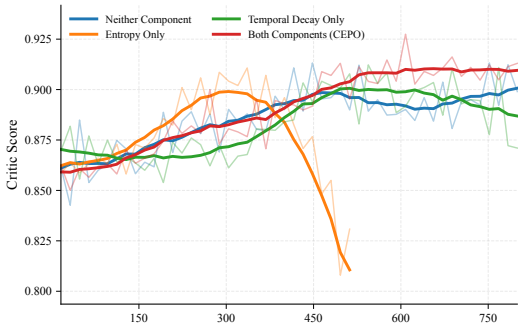


Figure 13: Training dynamics for the 2×2 factorial ablation.

H Statistical Significance of Automated Evaluation

Each test sample is independently evaluated five times using an ensemble of Gemini-2.5-Flash, DeepSeek-R1, and GLM-5.1 evaluators. Because scores lie on the discrete ordinal scale $\{1, 3, 5\}$, we use a two-sided Mann–Whitney U test rather than a parametric test. Table 6 compares Qwen3-32B CEPO with the non-agent Gemini-2.5-Flash baseline. All six distributions differ significantly, but the direction is not uniformly favorable to CEPO: Gemini scores higher in coherence, while CEPO scores higher in the other five dimensions.

Metric	**U-Statistic**	**P-Value**
Pacing	1,449,090.0	< 0.001
Appeal	1,297,241.0	< 0.001
Completeness	1,188,195.0	< 0.01
Coherence	958,086.0	< 0.001
Style	1,228,424.5	< 0.001
Hook	1,152,210.0	< 0.001

Table 6: Two-sided Mann–Whitney U tests for Qwen3-32B CEPO versus Gemini-2.5-Flash. Point estimates are reported in Table 1.

I Biases of LLM-Based Evaluation

LLM evaluators can exhibit position, style, self-enhancement, and length biases (Zheng et al., 2023). Position and style biases arise most strongly in open-ended pairwise ranking; our Token-Level Critic is instead restricted to localized error detection and minimal-span extraction within a fixed schema. Self-enhancement is reduced by using different model families for generation (Qwen3) and criticism

(Gemini-2.5-Flash-Lite). Length bias remains a limitation for ultra-long outputs, although chapter-wise generation and evaluation constrain the context of each individual decision. The automated main evaluation further aggregates three heterogeneous judge families rather than relying on the training critic alone.

J LENGTH-STRATIFIED EVALUATION

To assess whether output length drives the main results, we partition Qwen3-32B CEPO outputs into equal-count length quintiles. Longer narratives provide more space for plot development and correlate with greater Appeal, whereas Pacing and Completeness decline with length. Coherence, Style, and Hook remain comparatively stable. These trends show that length is a meaningful covariate but do not explain CEPO's results through uniformly increasing scores.

Length Quantile	Average Length	Pacing	Appeal	Completeness	Coherence	Style	Hook
Overall	34k	4.30	4.01	3.98	3.25	3.63	4.96
0%–20%	24k	4.49	3.85	4.24	3.27	3.54	4.94
20%–40%	29k	4.33	3.92	4.03	3.24	3.64	4.97
40%–60%	34k	4.34	4.07	3.99	3.19	3.62	4.97
60%–80%	39k	4.23	4.12	3.88	3.34	3.70	4.95
80%–100%	46k	4.11	4.08	3.78	3.22	3.63	4.97

Table 7: Automated evaluation of Qwen3-32B CEPO stratified by output-length quintile.

K AUTOMATED EVALUATION DETAIL

To ensure the objectivity and consistency of the Large Language Model (LLM) evaluators, we developed a standardized scoring rubric for each evaluation criterion. For every novel, the LLMs assign a score of 1, 3, or 5 based on the following granular descriptions.

K.1 PACING

Assessment: Whether the plot progresses smoothly with a balanced tempo, avoiding sluggishness or excessive haste.

- **5 (Excellent):** The pacing is compact and well-layered, with a masterfully balanced rhythm of tension and relaxation. The transitions between intense and slow-paced moments are highly natural, perfectly aligning with reader expectations.
- **3 (Fair):** The pacing is relatively stable but lacks variety. There are minor instances of redundant plots or slightly abrupt transitions.
- **1 (Poor):** The pacing is completely chaotic, being either excessively sluggish or overly rushed. The narrative focus is blurred, failing to establish an effective reading rhythm.

K.2 APPEAL

Assessment: Whether the story can sustain the reader's attention, featuring conflicts, suspense, and climaxes.

- **5 (Excellent):** The plot is highly creative and engaging with rigorous logic. Narrative twists are surprising yet reasonable. Characters are three-dimensional with complex, authentic emotions. Suspense and climaxes are exceptionally well-crafted, ensuring an immersive reading experience.
- **3 (Fair):** The plot is coherent with a clear theme but lacks novelty. The logic is generally sound, though some details feel forced. Character designs are basic, and the suspense/climaxes provide only moderate engagement.

- **1 (Poor):** The plot lacks highlights, and the narrative is dull or logically contradictory. Characters are flat and stereotypical. There is no discernible suspense or climax, failing to sustain any reader interest.

K.3 COMPLETENESS

Assessment: Whether the world setting, character designs, and spatiotemporal background are complete, providing a clear and comprehensive framework for the story.

- **5 (Excellent):** The settings are exceptionally detailed, covering all key elements including world-view, character backgrounds, and core rules. The framework is systematic and mutually supportive, allowing readers to easily immerse themselves in the background.
- **3 (Fair):** The core settings are generally complete, but the supporting details are thin. Some non-essential elements are missing, making the world-building feel slightly underdeveloped.
- **1 (Poor):** The settings are fragmented and lack a systematic structure. Key elements are missing, and the world-view or character backgrounds are too vague to support the narrative development.

K.4 COHERENCE

Assessment: Whether character designs and the spatiotemporal background conflict with the established world, and whether the internal logic is consistent and self-justifying.

- **5 (Excellent):** The logic is rigorous with no loopholes. Core rules are unified, and the environment, character motivations, and spatiotemporal settings are deeply compatible with the world-view. All dimensions support each other to provide a perfect logical foundation.
- **3 (Fair):** The overall logic is acceptable, but there are minor ambiguities or forced interactions (e.g., abrupt motivations or slight contradictions between rules and environment). The compatibility between different dimensions is average.
- **1 (Poor):** The settings lack internal logic. Core elements openly conflict with each other, and key dimensions such as characters and space-time are disjointed, failing to provide a valid logical basis for the story.

K.5 STYLE

Assessment: Whether the language is sufficiently grounded and natural, with grammatical accuracy and flexible sentence structures.

- **5 (Excellent):** The grammar is flawless with precise punctuation. The language is fluent, natural, and utilizes a diverse vocabulary and flexible sentence structures. The expression is concise, unambiguous, and demonstrates strong literary appeal.
- **3 (Fair):** There are several grammatical or punctuation errors that occasionally disrupt reading fluency. The expression is somewhat rigid, the vocabulary is limited, and minor semantic ambiguities may require re-reading.
- **1 (Poor):** The grammatical system is broken with non-standard punctuation. The sentence logic is collapsed, and the vocabulary usage is confused, rendering the text virtually unreadable.

K.6 HOOK

Assessment: Whether the first three chapters quickly establish conflicts, suspense, and other elements to capture and retain readers.

- **5 (Excellent):** The first three chapters rapidly establish strong conflicts and high suspense. They consistently provide clear points of interest, creating a strong desire in the reader to continue.

- **3 (Fair):** Identifiable conflicts and hooks are present, but the overall appeal is moderate and may struggle to maintain long-term engagement.
- **1 (Poor):** There are no effective core conflicts or hooks. The content is bland and fails to stimulate any initial reading interest.

L HUMAN EVALUATION DETAIL

To rigorously evaluate the generated long-form narratives, we conduct a human evaluation over 100 complete novels. We recruited a total of 10 annotators with rich experience in reading web novels; each annotator reviewed approximately 10 complete novels per evaluated method or variant. Each story pair is evaluated via pairwise comparison, and each annotator is compensated at a rate of 11.75 CNY per instance. We adopt a single-blind setup where the model names are hidden and the presentation order of the generated novels is randomized. Evaluators are required to read the complete novels before making their assessments. For each specific dimension, evaluators assign one of the following labels:

- **A (A Wins)**: Text A satisfies the criteria for a specific dimension, whereas Text B does not.
- **B (B Wins)**: Text B satisfies the criteria for a specific dimension, whereas Text A does not.
- **AG (All Good):** Both texts satisfy the criteria.
- **AB (All Bad):** Neither text satisfies the criteria.

The evaluation is conducted across the following six dimensions:

- **Pacing.** Whether the plot progresses smoothly with a balanced tempo, avoiding sluggishness or excessive haste.
- **Appeal.** Whether the story can sustain the reader's attention, featuring conflicts, suspense, and climaxes.
- **Completeness.** Whether the world setting, character designs, and spatiotemporal background are complete, providing a clear and comprehensive framework for the story.
- **Coherence.** Whether character designs and the spatiotemporal background conflict with the established world, whether the internal logic is consistent, and whether different setting dimensions support each other, justifying the story's development.
- **Style.** Whether the language is sufficiently grounded and natural, with grammatical accuracy and flexible sentence structures.
- **Hook.** Whether the first three chapters quickly establish conflicts, suspense, and other elements to capture and retain readers.

L.1 INTER-ANNOTATOR AGREEMENT

Given the subjectivity of long-form narrative evaluation, we compute Fleiss' κ over the four categorical assignments (A Wins, B Wins, All Good, All Bad). Table 8 shows moderate agreement for the more subjective Appeal and Style dimensions and substantial agreement for Pacing, Completeness, Coherence, and Hook.

	Pacing	**Appeal**	**Completeness**	**Coherence**	**Style**	**Hook**
Fleiss' κ	0.61	0.54	0.72	0.68	0.51	0.64
Agreement	Substantial	Moderate	Substantial	Substantial	Moderate	Substantial

Table 8: Inter-annotator agreement across human-evaluation dimensions.

M QUALITATIVE ERROR CASES

We provide representative logical errors identified by the Critic Agent. The examples illustrate the error taxonomy rather than claiming that every error has an identical entropy profile.

M.1 Entity Consistency

Entity setting. Mo Jingshen is a powerful head of a wealthy family who silently protects the protagonist through two lifetimes and supports her revenge.

Generated text. I covered my mouth, tears falling uncontrollably. I finally understood that, turns out, all my suffering was part of his plan. He was the one silently guarding me all along.

Error. Claiming that he planned all of her suffering contradicts his established role as her protector.

M.2 Outline Alignment

Outline. Years later, Lu Yanchen matures in a remote region. Mo Jingshen gives him an opportunity to start over and build a new career.

Generated text. He told Mo Jingshen that Lu Yanchen had established himself in the remote town. The old man sighed and said the Lu family had not abandoned him and had given him a chance to start over.

Error. The outline assigns the opportunity to Mo Jingshen, whereas the narrative assigns it to the Lu family.

M.3 Hard Constraints

Entity setting. Su Chen is the first-person protagonist, a gifted appraiser seeking revenge for his father.

Generated text. Tianqiong Pavilion's reputation began to decline. Qin Wan tried to suppress the news, but Su Chen used internet anonymity to spread it. I stood before the window, looking at the city lights.

Error. The text abruptly switches from the required first-person perspective to third person and then back again.

M.4 Spatiotemporal Logic

Generated text. The screen behind me lit up and played a recording: "Do not worry; I have balanced the accounts. Once the company goes public, the money is ours." It was Shen Yue's voice. "Did you also arrange the fire three years ago?" I pressed.

Error. A present-time question is inserted into a pre-existing recording, making the temporal interaction impossible.

N Critic Agent Prompt Template

The following is the principal Critic Agent template used to obtain localized, machine-readable error spans. The Idea Agent, Blueprint Agent, and automated-evaluation prompts use analogous role, input, rubric, and structured-output components; their evaluation rubrics are detailed in Appendix K.

```
# Role
You are a senior novel editor. Identify severe issues in novel text.

# Inputs
1. <UserRequirements>: settings, outline, and constraints.
2. <TextToEvaluate>: the novel text to check.

# Goal
Compare the requirements with the text. Return only a strict JSON
report. Retain only the minimum exact source snippet required for
localization; never return an entire paragraph or chapter.
```

```
# Negative checks
1. Entity consistency: contradictions in character or item attributes.
2. Outline alignment: missing points, reversed causality, or changes to
   the specified ending.
3. Hard constraints: violations of point of view, tense, or forbidden
   terms.
4. Spatiotemporal logic: impossible movement or timeline confusion.
5. Repetition: repeated blocks, sentences, plots, or chapters.
6. Garbage leakage: prompts, system instructions, evaluation workflow,
   model meta-discourse, markup, or placeholders.

# Workflow
1. Read the text together with the requirements.
2. Locate each severe issue and copy only its minimum exact span.
3. Return the following JSON object and no other text:
{
  "revision_summary": "one-sentence overall evaluation",
  "modifications": [
    {
      "subitem": [1],
      "issue_type": "short issue tag",
      "reason": "brief reason",
      "error_chapter_index": 1,
      "original_text": "minimum exact source snippet"
    }
  ]
}
```

O LIMITATIONS

The token-level policy optimization introduces additional memory constraints during training due to the need to compute and store fine-grained advantage signals across thousands of tokens. In addition, while the automated and human evaluation protocols attempt to quantify literary quality, narrative assessment remains inherently subjective and strongly influenced by cultural contexts. Our human panel consists of experienced web-novel readers but not a dedicated cohort of professional authors; future evaluation should include writers with professional experience to assess craft and structural design. LLM-based evaluation may also retain residual length and style biases despite heterogeneous judges, structured rubrics, and chapter-wise criticism (Appendix I). Finally, the entropy signature remains a correlational observation in the evaluated narrative setting; its mechanistic origin and generalization to other tasks require further controlled study.

P BROADER IMPACTS

The ability to generate coherent long-form narratives presents significant opportunities for creative industries by providing writers with intelligent drafting assistants and democratizing storytelling for individuals without formal literary training. However, scaling automated novel generation raises critical ethical and socio-economic concerns. The widespread deployment of highly capable narrative agents poses a risk of economic displacement for professional authors and screenwriters. Beyond the creative domain, the underlying techniques for maintaining long-range causal consistency could be misappropriated to generate sophisticated long-form disinformation or synthetic propaganda that is difficult for human readers to distinguish from authentic content. Mitigating these risks requires the development of robust watermarking techniques and transparent disclosure protocols for machine-generated media.