Leveraging Outline-Optimized Generative Interactions and Critique for Self-Refining Outlines with Reinforcement Learning

Hengwei Liu[1] **Haoyuan Ma**[1] **Qingqing Lyu**[1] **Daoxin Zhang**[2]
Yao Hu[2] **Yongliang Shen**[1] **Yin Zhang**[1] **Weiming Lu**[1†]
[1]Zhejiang University [2]Xiaohongshu
{hengweiliu, luwm}@zju.edu.cn

Abstract

Long-form outline generation requires satisfying multiple competing objectives simultaneously: outlines must be engaging, well-organized, topically relevant, and comprehensive while maintaining logical consistency across hierarchical structures. Current approaches either rely on expensive multi-turn interactions with large language models or employ procedural refinement pipelines that cannot systematically learn from critique. We present **LOGIC-RL**, a framework that transforms critique-guided outline refinement into a learnable policy through reinforcement learning. Our approach constructs refinement trajectories from teacher demonstrations, synthesizes explicit reasoning chains that decompose the critique-revision process, and optimizes a refinement policy using group relative policy optimization with structure-aware rewards. Experiments on FreshWiki and WikiOutline demonstrate that **LOGIC-RL** achieves substantial improvements over strong baselines, with the 0.6B model obtaining 79.17% relative gain and the 1.7B model achieving 8.67% improvement in average rubric scores compared to the best existing methods. Further analysis reveals that learned refinement policies generalize across domains and can be iteratively applied, with quality continuing to improve through three refinement rounds before diminishing returns.

1 Introduction

Large language models (LLMs) have shown strong potential for long-form generation when guided by *planning-then-writing* paradigms, where a high-level plan is produced first and the final text is generated by decomposing the task into structured subtasks (Lei et al., 2024; Yao et al., 2019; Rohman, 1965). This paradigm is especially important for expository writing such as Wikipedia-style articles, technical documentation, and report-like documents, including automatic Wikipedia-style generation settings (Fan and Gardent, 2022; Banerjee and Mitra, 2015). In these settings, the outline serves as the backbone of the document: it determines what to cover, how to group information, and how to allocate depth across sections. Consequently, improving outline quality is often a prerequisite for improving downstream article quality.

Recent systems have strengthened the pre-writing stage by integrating multi-perspective topic exploration, iterative reflection, and retrieval-based evidence gathering. STORM (Shao et al., 2024a) covers with perspectives and consolidates dialogue and retrieval into a rigorous hierarchy; RAPID (Gu et al., 2025) enhances efficiency and stability by using retrieved outlines to bootstrap structure and perform targeted retrieval for gaps; and Omni-Think (Xi et al., 2025) builds on STORM with iterative expansion-and-reflection to improve depth and compactness. Meanwhile, retrieval-enhanced outline generation has evolved from basic RAG (Lewis et al., 2020) to structure-first, evidence-aligned pipelines: WebWeaver (Li et al., 2025) couples outline evolution with an evidence memory, and ReSum (Wu et al., 2025c) enables longer-horizon planning via periodic summarization of interactions.

Despite these advances, producing *detailed, globally coherent, and consistently well-structured* long-form outlines remains challenging, especially under resource constraints. We highlight three persistent limitations. (1) **Brittle hierarchical control and global planning:** current LLMs still struggle to consistently enforce hierarchical constraints and make globally coherent structural decisions, which often leads to level skipping, missing key branches, and suboptimal depth allocation. (2) **Cost–quality tension in exploration and refinement:** multi-perspective exploration and reflective iterations can improve coverage and reduce redundancy (Shao et al., 2024a; Xi et al., 2025), but their effective-

†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33029–33046

ness depends on the quality of perspective sets and the reliability of reflection criteria, and they often incur substantial retrieval and reasoning costs. (3) **Retrieval–coherence and long-horizon consistency:** while structure-first and evidence-aligned retrieval pipelines improve factual grounding (Gu et al., 2025; Li et al., 2025), retrieval noise and partial evidence can still destabilize global coherence, and long-horizon planning remains difficult without robust mechanisms to maintain consistent intermediate states (Wu et al., 2025c). Crucially, existing refinement pipelines are largely *procedural*: they lack an effective *learnable* mechanism that systematically transfers critique-driven refinement behaviors to small models.

To address these challenges, we propose **LOGIC-RL**, a RL (reinforcement learning) framework that makes critique-guided outline refinement *learnable* for small models via an *early experience* paradigm (Zhang et al., 2025; Simmhan and Kulkarni, 2025; Abdollahi et al., 2025). Our core idea is to preserve high-quality refinement experiences produced by a strong teacher under the Logic interaction protocol (Liu et al., 2025), and convert them into explicit supervision and alignment signals. We collect refinement triplets (draft outline, structured feedback, improved outline), further elicit two-stage reasoning-and-reflection trajectories that explicitly model critique and revision, and apply quality filtering to retain low-noise training signals. We then adopt a two-stage optimization pipeline: we *cold-start* the student via SFT (supervised fine-tuning) on structured reasoning-and-reflection sequences (Wei et al., 2022), and then apply RL using GRPO (Group Relative Policy Optimization) (Shao et al., 2024b) to align the refinement policy with a format reward, a structure reward, and a quality reward.

We evaluate **LOGIC-RL** on FreshWiki (Shao et al., 2024a) and WikiOutline dataset, using heading-based metrics together with rubric grading over seven outline-quality dimensions. Experiments show that **LOGIC-RL** yields substantial and consistent gains for 0.6B and 1.7B models over direct generation and strong refinement baselines, demonstrating that reinforcement-learned critique and refinement can significantly strengthen small-model global planning and coherence.

Our main contributions are:

- We propose **LOGIC-RL**, an early-experience-driven framework that learns critique-guided refinement policies for long-form outline generation under resource constraints.
- We construct structured training signals based on Logic-style refinement triplets and two-stage reasoning-and-reflection trajectories, with quality filtering for reliable learning.
- We introduce a two-stage optimization pipeline and demonstrate strong gains on FreshWiki and WikiOutline across model scales.

2 Related Work

2.1 Automatic Expository Writing

Long-form text generation has long been recognized as a planning-intensive NLG problem: as generation length grows, models must maintain global coherence, avoid redundancy, and respect long-range constraints, which is difficult to achieve with purely local next-token modeling. Early work on narrative/story generation systematically exposed these bottlenecks and motivated the *planning-then-writing* paradigm, where a high-level plan is produced first and then expanded into fluent text (Riedl and Young, 2010; Mostafazadeh et al., 2016; Alhussain and Azmi, 2021; Fan et al., 2018; Yao et al., 2019). This line of research suggests that explicit structural planning is often a prerequisite for improving long-form generation quality.

Expository writing emphasizes factual explanation and knowledge organization rather than narrative coherence. With retrieval-augmented generation (RAG) (Lewis et al., 2020), systems increasingly incorporating web evidence to reduce hallucinations and improve traceability, often via research-agent pipelines that decompose questions, collect evidence, and synthesize answers (Nakano et al., 2021; Yao et al., 2022). Recent work further couples iterative information gathering with structured planning, for example STORM (Shao et al., 2024a) uses perspective-conditioned interactions and retrieval to build Wikipedia-style articles, and follow-up systems improve efficiency and long-horizon stability through structure-aware retrieval, evidence organization, and summarization-based memory (Qiao et al., 2025; Wu et al., 2025a,b).

2.2 Automatic Outline Writing

Automatic outline generation aims to produce a hierarchical plan that specifies content and organization before writing. The simplest approach, direct generation, prompts an LLM to output an outline

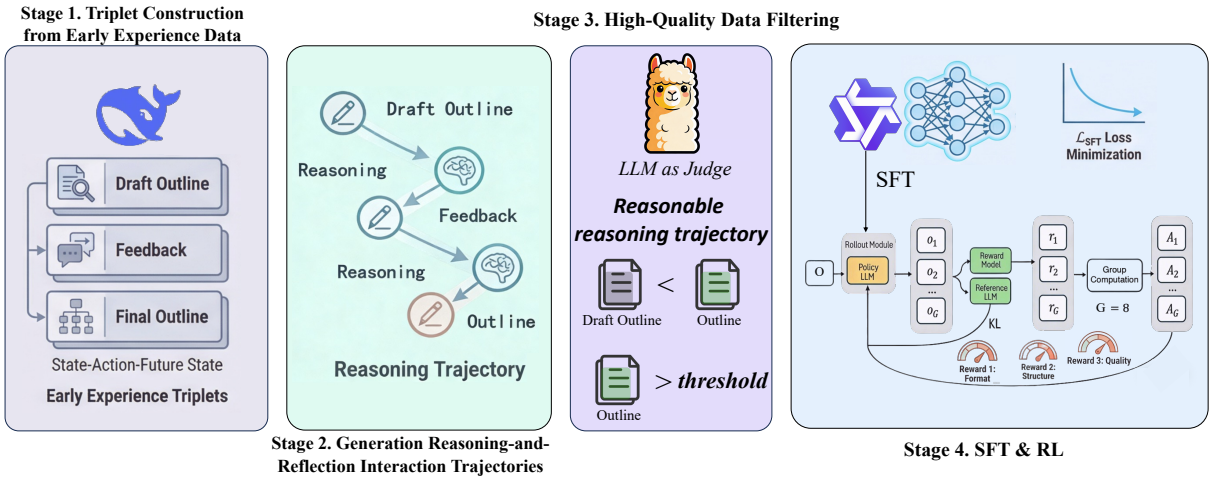

Figure 1: The overview of **LOGIC-RL**. In stage 1 (§3.1), we collect early experience data under the Logic protocol and normalize each interaction into refinement triplets (draft outline, feedback, refined outline). In stage 2 (§3.2), we elicit two-stage reasoning-and-reflection trajectories to expose critique and revision processes. In stage 3 (§3.3), we apply automatic quality filtering to retain high-consistency, low-noise samples for learning. In stage 4 (§3.4, §3.5), we optimize the student model with a two-stage training pipeline, including sft for cold-start initialization and rl for alignment.

in one pass; however, it often suffers from shallow coverage and unstable structure, particularly for long outlines (Yao et al., 2019). To address this, recent research explores *structure-building* through exploration, consolidation, and refinement.

A key direction in structure-building is multi-perspective topic exploration to improve coverage. STORM (Shao et al., 2024a) operationalizes coverage by gathering evidence through perspective-conditioned dialogue before consolidating the outline. OmniThink (Xi et al., 2025) builds on this by reorganizing knowledge for better depth. Another approach is retrieval-enhanced, structure-first outlining, where retrieved outlines or evidence guide the structure, with targeted retrieval filling gaps. RAPID (Gu et al., 2025) improves stability by bootstrapping structure from retrieved outlines and performing constrained retrieval. Evidence-aligned planning further strengthens the coupling between the outline and supporting sources, as seen in WebWeaver (Li et al., 2025), which co-evolves with an evidence memory, and ReSum (Wu et al., 2025c), which supports longer-horizon planning via periodic summarization. Beyond procedural refinement, critique-driven iterative outlining has been explored to improve coherence and hierarchy validity. Logic (Liu et al., 2025) uses structured interactions and imitation from similar-topic outlines to produce more coherent outlines. Knowledge distillation has also been widely used to transfer capabilities from larger teacher models to smaller students by matching output distributions (Hinton et al., 2015), enabling efficient deployment while preserving generation quality. Most existing pipelines lack a learnable mechanism for transferring critique and revision to smaller models. In contrast, our work makes critique-guided outline refinement learnable under resource constraints by distilling refinement experiences and aligning policies with RL.

3 LOGIC-RL

3.1 Triplet Construction from Early Experience Data

To train small language models for long-form outline refinement, we construct a structured dataset that turns a teacher model's refinement process into explicit, learnable supervision signals. Concretely, we use the framework Logic (Liu et al., 2025) and sample topic keywords from FreshWiki to improve domain diversity and generalization. We adopt DeepSeekR1 (Guo et al., 2025), a Mixture-of-Experts model (Dai et al., 2024), as the teacher to generate high-quality refinement trajectories.

Given a topic, the teacher-driven Logic interaction produces (i) an initial draft outline $\mathcal{O}_{draft}$, (ii) actionable feedback $\mathcal{F}$, and (iii) a revised outline $\mathcal{O}^*$. The feedback $\mathcal{F}$ explicitly targets seven outline-quality dimensions, including Interest, Organization, Relevance, Coverage, Logicality, Breadth, and Depth, and thus provides structured guidance for revision. For standardized training

and analysis, we normalize each interaction into a triplet: $(\mathcal{O}_{draft},\ \mathcal{F},\ \mathcal{O}^*)$. We additionally record metadata such as topic category, outline depth, heading count, and output length to ensure traceability and reproducibility.

3.2 Generation of Reasoning-and-Reflection Interaction Trajectories

Supervising only $(\mathcal{O}_{draft}, \mathcal{F}, \mathcal{O}^*)$ risks encouraging a shallow input-to-output mapping, without teaching the model *how* to diagnose defects, formulate feedback, and apply it to revise structure. To expose this process, we further elicit explicit reasoning-and-reflection trajectories in two stages. The first-stage trajectory $\mathcal{C}_1$ explains how the teacher identifies issues in $\mathcal{O}_{draft}$ and derives feedback $\mathcal{F}$, including the evidence for detected defects, the mapping to specific evaluation dimensions, and concrete improvement directions. The second-stage trajectory $\mathcal{C}_2$ explains how to transform $\mathcal{O}_{draft}$ into $\mathcal{O}^*$ conditioned on $\mathcal{F}$, covering structural edits, content completion, and logical reordering. The prompts used to prompt the teacher to generate $\mathcal{C}_1$ and $\mathcal{C}_2$ are provided in Appendix D.

This yields a process-supervision quintuple: $(\mathcal{O}_{draft},\ \mathcal{C}_1,\ \mathcal{F},\ \mathcal{C}_2,\ \mathcal{O}^*)$, as shown in Table 1. To reduce format variance and support automatic parsing and reward computation, we enforce a tag-based schema with `<reflective_critique>`, `<iterative_refinement>`, and `<answer>`, where the first two tags capture critique and refinement rationales and `<answer>` contains the refined outline used for structural checks and rubric grading. This structured formatting standardizes supervision signals, aligning with reasoning paradigms such as Search-R1 (Jin et al., 2025).

3.3 High-Quality Data Filtering

Reliable reward signals are crucial for stable RL (Christiano et al., 2017). To minimize noise and prevent reward misguidance, we use GPT-4 to assess $\mathcal{C}_1$ and $\mathcal{C}_2$, and Prometheus to score $\mathcal{O}^*$. The filtering consists of two parts. First, we assess the quality of $\mathcal{C}_1$ and $\mathcal{C}_2$, focusing on reasoning completeness and reflection specificity, and discard trajectories with missing rationale, weak defect-to-feedback grounding, or unfaithful refinement explanations. Second, we score the final outline $\mathcal{O}^*$ on the same seven rubric dimensions, using a 1–5 scale per dimension with a maximum total of 35. The detailed rubric definitions are provided in Appendix A. We additionally require $\text{score}(\mathcal{O}^*) > \text{score}(\mathcal{O}_{\text{draft}})$ to ensure the refinement yields a quality improvement. We remove samples that violate basic structural constraints (e.g., invalid hierarchy transitions or missing headings) or have insufficient overall quality (total score < 32). After filtering, we retain 749 high-quality samples, whose high consistency helps reduce reward noise and stabilize RL training.

3.4 Cold-Start for Supervised Fine-Tuning

The SFT stage aims to provide a strong initialization for small models by distilling structured decision-making behaviors from a stronger teacher and by teaching the model to explicitly perform "defect identification $\rightarrow$ feedback generation $\rightarrow$ outline revision."

Concretely, we represent each example as a process-supervision sequence that contains the draft outline, intermediate reasoning, feedback, refinement reasoning, and the final outline. Following CoT (Chain-of-Thought) distillation (Wei et al., 2022), we use the teacher (DeepSeekR1) to provide explicit reasoning traces so that each refinement step can be traced and learned. This formulation naturally matches an "initial state–action–future state" view of refinement, where the draft outline is the initial state, the feedback acts as the key action signal, and the revised outline is the future state. The model is trained to maximize the likelihood of generating the entire refinement chain conditioned on the draft outline. The SFT objective is defined in Eq. 1.

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathcal{O}_{\text{draft}},\mathcal{C}_1,\mathcal{F},\mathcal{C}_2,\mathcal{O}^*)\sim\mathcal{D}} \left[\log \pi_\theta(\mathcal{C}_1, \mathcal{F}, \mathcal{C}_2, \mathcal{O}^* \mid \mathcal{O}_{\text{draft}})\right]. \tag{1}$$

where π_θ denotes the student model, $\mathcal{C}_1$ and $\mathcal{C}_2$ are the two-stage reasoning-and-reflection traces, and $\mathcal{O}^*$ is the target refined outline. Unless otherwise specified, we perform full-parameter fine-tuning to better align the model's representation space and generation distribution with the structured refinement task, which we found beneficial for learning global planning and coherence.

3.5 GRPO-Based Reinforcement Learning Alignment Optimization

While SFT provides a reasonable initialization, small models may still fall into suboptimal refinement behaviors when facing complex structural

Draft Outline: {draft outline}, Feedback: {feedback}, Outline: {outline}, generate reflective reasoning about how to provide feedback for the draft outline based on the draft outline and feedback. You need to provide one sentence of feedback from the {feedback_dimension} dimension to improve the draft outline, and then improve the draft outline into the final outline based on the feedback. In <reflective>, generate reflective reasoning about how to provide feedback for the draft outline, and in <iterative_refinement>, generate the reflective reasoning of how to refine from draft outline to the final outline. Please strictly follow the following format. Example Format: <think> <reflective_critique> {reflective_critique} </reflective_critique>, Feedback, <iterative_refinement> {iterative_refinement} </iterative_refinement> </think>, <answer> {outline} </answer>.

Table 1: Data Generation Template Incorporating Reasoning and Reflection Processes.

decisions. Given a draft outline $\mathcal{O}_{draft}$, the current policy π_θ samples a group of G candidate refinement trajectories $\{\tau_i\}_{i=1}^{G}$ (we use G=8), each trajectory containing the full chain "draft $\rightarrow$ reasoning $\rightarrow$ feedback $\rightarrow$ refinement reasoning $\rightarrow$ refined outline." We compute a scalar reward for each trajectory and optimize the policy to assign higher probability to trajectories that perform better relative to others in the same group.

Reward design. We design a multiplicative reward that encourages (i) valid tagged outputs, (ii) structurally well-formed outlines, and (iii) high-quality refined outlines under a multi-dimensional rubric. For a trajectory τ, the format reward r_{format} in Eq. (3) encourages tag compliance, where $N_{\text{think}}(\tau)$, $N_{\text{reflective}}(\tau)$, and $N_{\text{iterative}}(\tau)$ count the occurrences of the corresponding tags in τ. The structure reward $r_{\text{structure}}$ in Eq. (4) penalizes empty or malformed outlines and discourages level skipping. The quality reward r_{quality} in Eq. (5) is designed to encourage high-quality refinements based on the rubric score. Specifically, "average_score" refers to the average score across the seven rubric dimensions, while "total_score" is the sum of the scores across these seven dimensions. When the total score reaches or exceeds 30, the reward increases exponentially. This exponential scaling ensures that high-quality refinements receive significantly higher rewards, incentivizing the model to focus on producing the highest-quality outlines.

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau\sim\mathcal{D}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min \Big(r_i(\theta)\hat{A}_i,\ \text{clip}\big(r_i(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\Big) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg]. \tag{6}$$

$$R(\tau) = \begin{cases} 0, & \text{No <answer> tag in } \tau, \\ r_{\text{format}}\, r_{\text{structure}}\, r_{\text{quality}}, & \text{Has <answer> tag in } \tau. \end{cases} \tag{2}$$

$$r_{\text{format}} = \frac{N_{\text{think}}(\tau) + N_{\text{reflective}}(\tau) + N_{\text{iterative}}(\tau)}{6}. \tag{3}$$

$$r_{\text{structure}} = \begin{cases} 0, & \text{empty or malformed outline,} \\ 0.5, & \text{valid format but level skipping,} \\ 1, & \text{valid format and hierarchy.} \end{cases} \tag{4}$$

$$r_{\text{quality}} = \begin{cases} 1, & average\_score < 2 \\ 2, & 2 \le average\_score < 3 \\ 4, & 3 \le average\_score < 4 \\ 8, & average\_score = 4 \\ 12, & total\_score = 29 \\ 2^{total\_score-26}, & total\_score \ge 30. \end{cases} \tag{5}$$

GRPO objective. For each group of sampled trajectories, we compute a normalized advantage using the group statistics, and update the policy with a clipped importance-weighted objective plus a KL regularization term to stabilize training, as defined in Eq. 6:

where $r_i(\theta) = \frac{\pi_\theta(\tau_i|\mathcal{O}_{draft})}{\pi_{\text{old}}(\tau_i|\mathcal{O}_{draft})}$ is the importance ratio, ϵ is the clipping coefficient, and π_{ref} is the reference policy (initialized as the SFT model). The KL penalty, scaled by β, prevents the policy from drifting too far from the SFT initialization while still allowing improvements driven by relative comparisons within each group. Overall, this group-based relative optimization enables stable policy refinement and encourages the model to consistently transform drafts into structurally valid and high-quality long-form outlines.

	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
			Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Direct Gen(GPT-4)	87.66	34.78	2.33	**2.34**	3.12	3.42	**3.20**	4.52	2.41	3.05
Direct Gen(DeepSeekR1)	**92.38**	**39.24**	**2.43**	2.31	**3.23**	**3.48**	3.18	4.63	**2.59**	**3.12**
Direct Gen(GPT-4o-mini)	87.24	33.67	2.24	2.27	3.02	3.32	3.19	**4.71**	2.36	3.02
Direct Gen(Qwen2-72B)	91.24	36.18	1.87	1.74	2.07	1.88	1.50	2.60	1.96	1.95
Direct Gen(Llama3.1-8B)	85.46	32.72	1.44	1.33	1.52	1.34	1.29	2.02	1.59	1.50
Direct Gen(Qwen3-1.7B)	84.73	30.28	1.68	1.53	1.35	1.23	1.19	1.76	1.43	1.45
Direct Gen(Qwen3-0.6B)	61.76	28.98	1.10	1.17	1.29	1.32	1.20	1.39	1.11	1.21
STORM(GPT-4o-mini)	90.56	35.51	2.93	3.59	3.45	3.77	3.83	4.79	2.90	3.64
OmniThink(GPT-4o-mini)	91.05	33.36	3.58	4.07	3.48	3.54	3.78	4.49	3.15	3.73
Logic(GPT-4o-mini)	**96.04**	**41.66**	**4.00**	**5.00**	**4.97**	**4.99**	**4.45**	4.86	**4.86**	**4.73**
STORM(Qwen2-72B)	76.31	39.47	2.79	3.12	3.14	3.14	2.78	3.96	2.72	3.09
OmniThink(Qwen2-72B)	62.56	32.71	2.92	3.53	2.69	3.07	2.89	3.93	3.45	3.21
Logic(Qwen2-72B)	92.00	40.56	3.90	4.90	4.48	4.67	4.41	**4.88**	4.63	4.55
STORM(Llama3.1-8B)	83.73	35.83	2.12	2.01	2.64	2.23	1.88	3.41	1.95	2.32
OmniThink(Llama3.1-8B)	93.64	32.98	3.00	3.26	2.44	1.98	2.72	3.59	3.31	2.90
Logic(Llama3.1-8B)	86.92	36.79	3.79	4.71	3.94	4.32	4.13	4.57	4.27	4.25
Logic(Qwen3-1.7B)	87.32	32.28	3.98	4.63	4.15	4.34	3.87	4.51	4.39	4.27
Logic(Qwen3-0.6B)	64.98	29.07	1.63	2.49	2.16	1.87	1.94	1.89	1.46	1.92
LOGIC-RL(Qwen3-1.7B)	**92.91**	**34.38**	**4.00**	**4.89**	**4.82**	**4.63**	**4.63**	**4.92**	**4.56**	**4.64**
LOGIC-RL(Qwen3-0.6B)	87.77	30.33	3.04	4.08	2.85	3.12	3.01	4.13	3.84	3.44

Table 2: Results of automatic outline quality evaluation of FreshWiki dataset.

3.6 Generating Expository Articles from Long-Form Outlines

Given a long-form outline $\mathcal{O} = \{o_1, \ldots, o_N\}$ and a topic keyword T, we first produce search queries $\mathcal{Q}_i = f_q(o_i, T) = \{q_{i,1}, \ldots, q_{i,K}\}$. A retrieval operator $\mathcal{R}(\cdot)$ then collects evidence $\mathcal{E}_i = \mathcal{R}(\mathcal{Q}_i) = \bigcup_{k=1}^{K} \mathcal{R}(q_{i,k})$, which includes definitions, representative results, data, and limitations relevant to the section.

Conditioning on $(o_i, T, \mathcal{E}_i)$, a language model generates a section draft $s_i^{(0)} = \mathcal{G}(o_i, T, \mathcal{E}_i)$, where $\mathcal{G}(\cdot)$ is prompted to follow o_i's structure and integrate evidence into key claims. To improve depth and readability, we further expand the draft without drifting from the section constraints:$s_i = \mathcal{H}(s_i^{(0)}, o_i, T)$. Finally, all sections are concatenated in outline order to form the full article: $\mathcal{A} = \mathrm{Concat}(s_1, s_2, \ldots, s_N)$. The prompt used to generate $s_i^{(0)}$ is provided in Appendix D.

4 Experiments

4.1 Datasets and Automatic Metrics

We evaluate **LOGIC-RL** on two datasets to assess both in-domain performance and cross-domain generalization. We first use FreshWiki, a widely adopted benchmark for Wikipedia-style expository outline generation that has been used by prior systems such as STORM (Shao et al., 2024a) and OmniThink (Xi et al., 2025). FreshWiki provides high-quality reference outlines with diverse topics and stable evaluation protocols; we follow its original data split and settings to ensure direct comparability with existing results. In addition, we use WikiOutline, a multi-domain dataset introduced by Logic (Liu et al., 2025) to address limitations in domain diversity and timeliness.

Following STORM (Shao et al., 2024a), we report two heading-based recall metrics: Heading Soft Recall and Heading Entity Recall (Fränti and Mariescu-Istodor, 2023). Heading Soft Recall measures semantic matching between generated and reference headings without requiring exact string overlap, which better reflects structural alignment in practical outlining. Heading Entity Recall measures knowledge coverage by computing the proportion of reference core entities that appear in the generated outline, with duplicate entities removed to avoid rewarding entity repetition. We additionally use rubric grading with Prometheus (Kim et al., 2023) for 1–5 scoring.

4.2 Baseline and Implementation

We select four representative baselines for comparison, including Direct Gen, STORM (Shao et al., 2024a), OmniThink (Xi et al., 2025), and Logic (Liu et al., 2025). Direct Gen prompts an LLM to produce an outline in a single pass without retrieval or refinement, serving as a simple but often shallow baseline. STORM is a role-playing baseline for outline generation in multi-perspective

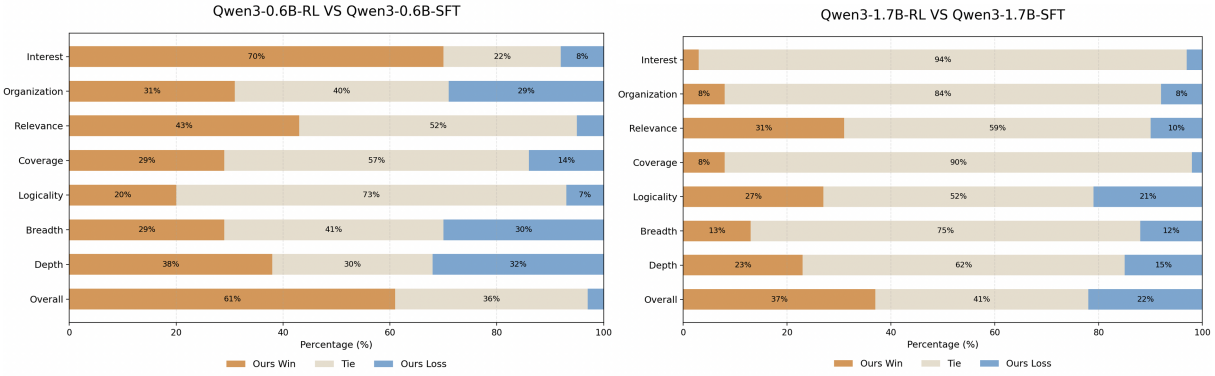

Figure 2: Ablation Experiment Statistics Plot of **LOGIC-RL** on the FreshWiki dataset.

	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
			Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3-0.6B	69.31	29.67	1.51	2.14	1.63	1.47	1.24	1.82	1.63	1.63
Qwen3-0.6B-SFT	85.38	31.19	1.79	**2.99**	1.32	1.70	1.34	3.16	2.62	2.13
Qwen3-0.6B-RL	**86.68**	**31.52**	**3.56**	2.90	**2.28**	**2.34**	**1.87**	**3.18**	**2.83**	**2.71**
Qwen3-1.7B	89.07	33.53	3.86	4.75	4.04	4.27	3.75	4.39	4.44	4.21
Qwen3-1.7B-SFT	**93.57**	31.58	3.98	4.78	4.12	4.43	4.17	4.65	4.37	4.36
Qwen3-1.7B-RL	92.53	**33.88**	**3.99**	**4.79**	**4.34**	**4.70**	**4.33**	**4.79**	**4.52**	**4.49**

Table 3: Ablation experimental results of Qwen3-0.6B and Qwen3-1.7B on the FreshWiki dataset.

dialogues. OmniThink expands knowledge and reorganize it for improved depth and compactness. Logic is a structured interactive outlining system that combines planning and critique-driven refinement to improve global coherence. We train our models in a two-stage pipeline: SFT using LLaMAFactory (Zheng et al., 2024), followed by RL alignment with VERL (Sheng et al., 2025). For inference, the model is set with temperature 1.0 and top_p 0.9.

5 Results and Analysis

5.1 Main Results

Tables 2 and 8 summarize the main results on FreshWiki and WikiOutline, including the performance of models DeepSeekR1 (Liu et al., 2024; Guo et al., 2025), GPT-4 (Achiam et al., 2023), GPT-4o-mini (Menick et al., 2024), Qwen (Team et al., 2024; Yang et al., 2025), and Llama (Dubey et al., 2024). We evaluate models using heading-based automatic metrics, namely Heading Soft Recall and Heading Entity Recall, together with rubric grading on seven dimensions, including Interest, Organization, Relevance, Coverage, Logicality, Breadth, and Depth, each scored on a 1–5 scale. Overall, **LOGIC-RL** achieves consistent gains in rubric quality across datasets and model scales, indicating that reinforcement-learned critique and refinement improves long-form outline planning beyond direct generation and prior refinement baselines.

As shown in Table 2 on FreshWiki, **LOGIC-RL** improves the average rubric score for both 0.6B and 1.7B models and outperforms the strongest direct-generation baseline, yielding relative gains of 79.17% and 8.67% over Logic at the same scale. The 1.7B variant is consistently strong across key dimensions such as Organization, Logicality, and Depth, and it also surpasses STORM and OmniThink in average rubric score. Heading Entity Recall does not always increase proportionally, suggesting that **LOGIC-RL** mainly improves global structure and discourse quality rather than surface-level heading overlap. As shown in Table 8, the 1.7B model raises the average rubric score from 3.24 to 4.37 relative to the strongest direct-generation baseline, and the 0.6B model improves from 1.23 to 3.10, indicating effective transfer of critique-driven refinement to smaller models. Overall, these results show that **LOGIC-RL** generates higher-quality long-form outlines with consistent gains across dimensions.

	Refinement Steps	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
				Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3 -0.6B-RL	1	86.68	**31.52**	**3.56**	2.90	2.28	2.34	1.87	3.18	2.83	2.71
	2	83.06	30.19	2.81	3.76	2.36	2.89	2.69	3.94	3.64	3.16
	3	**87.77**	30.33	3.04	4.08	2.85	3.12	**3.01**	**4.13**	**3.84**	**3.44**
	4	82.59	29.59	2.96	**4.09**	2.96	3.13	2.97	3.97	3.67	3.39
	5	78.60	29.26	2.92	3.90	**3.01**	**3.37**	2.98	4.02	3.48	3.38
Qwen3 -1.7B-RL	1	92.53	33.88	3.99	4.79	4.34	4.70	4.33	4.79	4.52	4.49
	2	92.74	33.88	**4.00**	4.87	4.53	**4.83**	4.24	4.81	4.52	4.54
	3	**92.91**	**34.38**	**4.00**	**4.89**	**4.82**	**4.63**	**4.63**	**4.92**	4.56	**4.64**
	4	91.97	32.88	3.98	4.44	4.54	4.53	4.56	4.72	4.62	4.47
	5	89.73	32.88	3.98	4.39	4.13	4.60	4.29	4.67	**4.70**	4.39

Table 4: Experimental results of Qwen3-0.6B-RL and Qwen3-1.7B-RL with different refinement steps on the FreshWiki dataset.

	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
			Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3-0.6B	69.31	29.67	1.51	2.14	1.63	1.47	1.24	1.82	1.63	1.63
Qwen3-0.6B-SFT	85.38	31.19	1.79	**2.99**	1.32	1.70	1.34	3.16	2.62	2.13
Qwen3-0.6B-RL (Exponential)	**86.68**	31.52	**3.56**	2.90	**2.28**	**2.34**	**1.87**	**3.18**	**2.83**	**2.71**
Qwen3-0.6B-RL (Linear)	86.53	**31.82**	3.14	2.71	1.98	1.87	1.69	3.11	2.59	2.44

Table 5: Ablation on Reward Formulation: Exponential vs. Linear Quality Reward in RL of Qwen3-0.6B on the FreshWiki dataset.

5.2 Ablation Study

To quantify the contributions of SFT and RL, we conduct ablations on Qwen3-0.6B and Qwen3-1.7B on FreshWiki and WikiOutline. We compare three variants, the base model, SFT, and RL, and report Heading Soft Recall, Heading Entity Recall, and rubric grading over seven dimensions. As shown in Tables 3 and 11, SFT consistently improves both structural matching and rubric quality, with larger gains for Qwen3-0.6B, for example on FreshWiki the average rubric score increases from 1.63 to 2.13 after SFT, along with a clear rise in Heading Soft Recall from 69.31 to 85.38.

RL further improves rubric quality beyond SFT, mainly on discourse-level dimensions rather than uniformly increasing heading recalls. On FreshWiki, Qwen3-0.6B improves from 2.13 to 2.71 after RL, while Qwen3-1.7B shows smaller but consistent gains, such as 4.36 to 4.49 on FreshWiki and 4.09 to 4.29 on WikiOutline. The win-rate analysis in Figure 2 corroborates this pattern, where RL outperforms SFT more often for Qwen3-0.6B with an overall win rate of about 61%, whereas for Qwen3-1.7B the outcomes are more tie-dominated, suggesting RL mainly provides fine-grained improvements once the model has been well-initialized by SFT.

To further analyze the impact of reward design, we conduct an ablation study comparing the original exponential quality reward with a linear alternative while keeping the rest of the RL pipeline unchanged. As shown in Table 5, RL with a linear quality reward still significantly outperforms the SFT baseline across most evaluation dimensions, indicating that the performance gains mainly stem from the RL optimization framework itself. Nevertheless, the original multiplicative reward with exponential quality scaling consistently achieves stronger improvements and a higher overall score, suggesting that the proposed reward shaping further enhances the effectiveness of RL rather than being the sole source of improvement.

5.3 Result of Expository Articles

We further evaluate downstream expository writing by feeding outlines from Qwen3-0.6B and Qwen3-1.7B under Base, SFT, and RL with one refinement step into a fixed writer model, GPT-4o-mini (Menick et al., 2024), and scoring the resulting articles with Prometheus on seven rubric dimensions. As shown in Table 13, article quality improves consistently with outline quality: Qwen3-0.6B shows a clear increase from 1.66 to 1.88 and 2.20, while Qwen3-1.7B increases from 3.59 to 4.37 and 4.47, with notable improvements across the overall rubric. The detailed rubric definitions are provided in Appendix A.

5.4 Detailed Analysis

We analyze the effect of iterative refinement steps using the RL-aligned models. As shown in Tables 4 and 12, increasing refinement steps improves most metrics at first, but the gains saturate and can reverse when refinement becomes too frequent. Domain-wise breakdowns on WikiOutline are reported as shown in Tables 9 and 10.

On FreshWiki, Qwen3-0.6B-RL performs best with three refinement steps, reaching a peak Heading Soft Recall of 87.77 and achieving strong rubric scores on Logicality, Breadth, and Depth, indicating improved global coherence beyond surface coverage, as shown in Table 4. Refinement beyond three steps leads to a decline in the average rubric score, indicating potential redundancy or topic drift. A similar trend holds on WikiOutline, where performance improves from one to three steps, with three steps providing the best balance in Organization and Depth. Further steps cause larger fluctuations and weaker results, as shown in Table 12. These results consistently suggest that a moderate number of refinement steps achieves the best trade-off between improved structure and over-refinement.

6 Conclusion

We presented **LOGIC-RL**, a RL framework for critique-guided refinement of long-form outlines in expository writing. **LOGIC-RL** converts teacher-driven refinement into learnable supervision by constructing refinement triplets and explicit reasoning-and-reflection trajectories, applies high-quality filtering to reduce training noise, cold-starts the policy with SFT, and then aligns refinement behavior using group-relative policy optimization with a format reward, a structure reward, and a quality reward. This design is especially beneficial for small models, since it explicitly distills structured critique and revision behaviors that are hard to acquire from direct generation alone, leading to better hierarchical control, more coherent global organization, and higher-quality outlines compared with procedural refinement pipelines. More broadly, the framework provides a practical approach for improving structural planning and critique-driven refinement in long-form generation under limited model capacity.

Limitations

While **LOGIC-RL** significantly improves long-form outline generation, there are several limitations to consider. The dataset used, although diverse, remains limited in scope and could benefit from expansion to cover more domains and topics, enhancing the model's generalization capabilities. Additionally, the effectiveness of **LOGIC-RL** heavily relies on the quality of the teacher model, which may not always be feasible in resource-constrained settings. Moreover, while iterative refinement shows improvements, it is subject to diminishing returns after a certain number of iterations, raising concerns about its scalability for longer outlines. The transferability of the framework to other long-form generation tasks, such as creative or highly technical writing, also requires further investigation, as the current methods may not be directly applicable. Finally, the reliance on rubric-based evaluation, while useful, might not capture task-specific nuances, suggesting the need for more specialized evaluation metrics. Future work should address these challenges by enhancing dataset diversity, refining iterative techniques, and adapting the framework for broader applications and more precise evaluations.

Ethic Statements

This work studies critique-guided refinement for long-form outline generation and aligns small models using supervised fine-tuning and reinforcement learning with LLM-based automatic evaluation. The datasets used in our experiments are sourced from publicly available resources, but we cannot guarantee they are free from harmful or toxic language. The data for training our models are synthesized using the teacher model DeepSeekR1, which generates critique-guided refinement trajectories. Since automated rubric scoring and teacher-generated trajectories may introduce evaluator or teacher-specific preferences and biases, we adopt a consistent rubric and evaluation protocol across all methods and interpret results as comparative rather than absolute.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62376245), the Key Research and Development Program of Zhejiang Province, China (No. 2024C03255), China Knowledge Centre for Engineering Sciences and Technology (CKCEST-2022-1-7), and MOE Engineering Research Center of Digital Library.

References

Sina Abdollahi, Mohammad Maheri, Sandra Siby, Marios Kogias, and Hamed Haddadi. 2025. An early experience with confidential computing architecture for on-device model protection. *arXiv preprint arXiv:2504.08508*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Siddhartha Banerjee and Prasenjit Mitra. 2015. Wikikreator: Improving wikipedia stubs automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 867–877.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Angela Fan and Claire Gardent. 2022. Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

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

Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognit. Lett.*, 167:115–121.

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. 2025. RAPID: efficient retrieval-augmented long text generation with writing planning and information discovery. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16742–16763. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, and 1 others. 2025. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hengwei Liu, Yongliang Shen, Zhe Zheng, Haoyuan Ma, Xingyu Wu, Yin Zhang, and Weiming Lu. 2025. Logic: Long-form outline generation via imitative

and critical self-refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18119–18144, Suzhou, China. Association for Computational Linguistics.

Jacob Menick, Kevin Lu, Shengjia Zhao, E Wallace, H Ren, H Hu, N Stathas, and F Petroski Such. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, and 1 others. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

D Gordon Rohman. 1965. Pre-writing: The stage of discovery in the writing process. *College Composition & Communication*, 16(2):106–112.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024a. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pages 1279–1297. ACM.

Yogesh Simmhan and Varad Kulkarni. 2025. Towards ai agents for course instruction in higher education: Early experiences from the field. *arXiv preprint arXiv:2510.20255*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025b. WebWalker: Benchmarking LLMs in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10290–10305, Vienna, Austria. Association for Computational Linguistics.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, and 1 others. 2025c. Resum: Unlocking long-horizon search intelligence via context summarization. *arXiv preprint arXiv:2509.13313*.

Zekun Xi, Wenbiao Yin, Jizhan Fang, Jialong Wu, Runnan Fang, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2025. Omnithink: Expanding knowledge boundaries in machine writing through thinking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 956–976.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, and 1 others. 2025.

Agent learning via early experience. *arXiv preprint arXiv:2510.08558*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Details of LLM-Based Evaluation

We use Prometheus as an automatic judge to evaluate both long form outlines and downstream expository articles, which provides a consistent and reproducible assessment protocol across methods. For outline evaluation, Prometheus assigns 1 to 5 scores on seven rubric dimensions, Interest, Organization, Relevance, Coverage, Logicality, Breadth, and Depth, and the detailed scoring criteria are listed in Table 6. For expository article evaluation, we apply an aligned seven dimension rubric, shown in Table 7, to measure how well a given outline supports coherent and informative writing. For all systems, we use the same evaluation prompt and the same scoring procedure, then we report averaged scores over the test set, which enables a direct comparison between outline level improvements and downstream article quality.

B Details Results on WikiOutline

We evaluate Logic-RL on the WikiOutline dataset, and the results in Table 8 demonstrate that Logic-RL significantly improves long-form outline generation compared to both direct generation baselines and existing refinement methods like STORM and OmniThink. Across all metrics, including Interest, Organization, Relevance, and Logicality, Logic-RL provides substantial improvements, particularly for the Qwen3-0.6B and Qwen3-1.7B models. These gains indicate that Logic-RL excels at improving the overall quality and coherence of generated outlines. Additionally, the improvements in Heading Soft Recall and Heading Entity Recall suggest that the system not only enhances the structure but also better covers the content of the outlines. Table 11 presents an ablation study to further explore the effectiveness of different components of the Logic-RL framework. It shows that SFT and RL contribute significantly to improving outline quality, particularly in Heading Recall and rubric scores across the dimensions.

Further analysis of the results in Tables 12, 9, and 10 explores the impact of different refinement steps on Logic-RL's performance. We find that iterative refinement consistently improves the model's performance, with specific gains in Logicality, Depth, and Breadth. However, the improvements plateau after a certain number of iterations, and excessive refinement leads to diminishing returns. This suggests that the number of iterative refinement steps plays a crucial role in optimizing long-form outline generation, with the most significant improvements occurring within the first few iterations.

C Detail Results of Expository Articles

Table 13 presents the detailed results of expository articles generated from different long-form outlines. Both Qwen3-0.6B and Qwen3-1.7B show significant improvements in all rubric dimensions with RL compared to the base model and SFT. For Qwen3-0.6B, RL achieves the highest average score of 2.20, with notable improvements in organization, relevance, and depth. For Qwen3-1.7B, RL outperforms SFT, reaching an average score of 4.47, with substantial gains in coverage, depth, and logicality, demonstrating the effectiveness of critique-guided refinement in enhancing article quality.

D LOGIC-RL inference prompt

This section summarizes the prompts used at inference time to support the **LOGIC-RL** pipeline. As shown in Table 14, we employ three professional prompt templates covering (i) reflective reasoning for feedback provision, which explains how critique targets draft deficiencies under Wikipedia-style outline norms, (ii) reflective reasoning for outline refinement, which details the feedback-conditioned revision rationale and concrete editing actions from draft to final outlines, and (iii) Wikipedia-style section drafting, which generates section-level article text by strictly following the given outline point and integrating retrieved evidence while maintaining neutrality, topical focus, and structural consistency.

Criteria Description	**Interest Level:** How engaging and thought-provoking is the outline?
Score 1 Description	Not engaging at all; no attempt to capture the reader's attention.
Score 2 Description	Fairly engaging with a basic narrative but lacking depth.
Score 3 Description	Moderately engaging with several interesting points.
Score 4 Description	Quite engaging with a well-structured narrative and noteworthy points that frequently capture and retain attention.
Score 5 Description	Exceptionally engaging throughout, with a compelling narrative that consistently stimulates interest.
Criteria Description	**Coherence and Organization:** Is the outline well-organized and logically structured?
Score 1 Description	Disorganized; lacks logical structure and coherence.
Score 2 Description	Fairly organized; a basic structure is present but not consistently followed.
Score 3 Description	Organized; a clear structure is mostly followed with some lapses in coherence.
Score 4 Description	Good organization; a clear structure with minor lapses in coherence.
Score 5 Description	Excellently organized; the outline is logically structured with seamless transitions and a clear argument.
Criteria Description	**Relevance and Focus:** Does the outline stay on topic and maintain a clear focus?
Score 1 Description	Off-topic; the content does not align with the headline or core subject.
Score 2 Description	Somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
Score 3 Description	Generally on topic, despite a few unrelated details.
Score 4 Description	Mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
Score 5 Description	Exceptionally focused and entirely on topic; the outline is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.
Criteria Description	**Broad Coverage:** Does the outline provide an in-depth exploration of the topic and have good coverage?
Score 1 Description	Severely lacking; offers little to no coverage of the topic's primary aspects, resulting in a very narrow perspective.
Score 2 Description	Partial coverage; includes some of the topic's main aspects but misses others, resulting in an incomplete portrayal.
Score 3 Description	Acceptable breadth; covers most main aspects, though it may stray into minor unnecessary details or overlook some relevant points.
Score 4 Description	Good coverage; achieves broad coverage of the topic, hitting on all major points with minimal extraneous information.
Score 5 Description	Exemplary in breadth; delivers outstanding coverage, thoroughly detailing all crucial aspects of the topic without including irrelevant information.
Criteria Description	**Logical Structure of the Outline:** Is the logical sequence of ideas presented in the outline clear?
Score 1 Description	Completely disorganized. Ideas are listed randomly without any logic and cannot be connected.
Score 2 Description	There is a basic structure, but the logic often breaks down, transitions are unclear, and the organization is chaotic.
Score 3 Description	The structure is relatively clear. Most ideas are logical, but some connections are not well-defined.
Score 4 Description	The logical structure is excellent. Ideas are presented clearly and orderly, with smooth transitions.
Score 5 Description	The logical structure is perfect. Ideas progress step by step, and the argument is rigorous.
Criteria Description	**Breadth of Coverage in the Outline:** To what extent does the outline cover different aspects and dimensions of the topic?
Score 1 Description	The scope is extremely narrow, focusing on only one or two aspects and ignoring the most relevant dimensions.
Score 2 Description	The breadth is limited, covering a few aspects, missing important content, and presenting an incomplete view.
Score 3 Description	The breadth is moderate, covering most major aspects, but some content is briefly mentioned or not fully covered.
Score 4 Description	The breadth is good, covering many aspects, providing a comprehensive overview with only minor omissions.
Score 5 Description	The coverage is comprehensive, exploring every aspect and dimension of the topic without omission.
Criteria Description	**Depth of Analysis in the Outline:** How thoroughly does the outline analyze the details and implications of each aspect of the topic?
Score 1 Description	The analysis is only superficial, merely listing facts without exploring underlying causes and relationships.
Score 2 Description	The analysis is shallow, providing basic details but not delving into meanings, consequences, and potential connections.
Score 3 Description	The analysis has some depth, but some content can be further explored for subtleties.
Score 4 Description	In-depth analysis, comprehensively exploring details, implications, and connections of each aspect, with only a few areas for improvement.
Score 5 Description	The analysis is extremely in-depth, digging into every detail, implication, and subtle relationship without blind spots.

Table 6: Scoring rubrics on a 1-5 scale for long-form outline by the evaluator LLM.

Criteria Description	**Interest Level:** How engaging and thought-provoking is the expository article?
Score 1 Description	Not engaging at all; no attempt to capture the reader's attention.
Score 2 Description	Fairly engaging with a basic framework but lacking appeal.
Score 3 Description	Moderately engaging with several interesting explanatory points.
Score 4 Description	Quite engaging with vivid examples and clear explanatory logic.
Score 5 Description	Exceptionally engaging throughout, with a compelling narrative rhythm.
Criteria Description	**Coherence and Organization:** Is the expository article well-organized and logically structured?
Score 1 Description	Disorganized; lacks logical structure and coherence.
Score 2 Description	Fairly organized; a basic structure is present but not consistently followed.
Score 3 Description	Organized; a clear structure is mostly followed with minor coherence lapses.
Score 4 Description	Well-organized; a clear structure with smooth logic and natural transitions.
Score 5 Description	Excellently organized; logically structured with seamless transitions.
Criteria Description	**Relevance and Focus:** Does the expository article stay on topic and maintain a clear focus?
Score 1 Description	Off-topic; the content does not align with the core subject.
Score 2 Description	Somewhat on topic but with several digressions; focus is unclear.
Score 3 Description	Generally on topic, despite a few unrelated details.
Score 4 Description	Mostly on topic and focused; consistent relevance to the core subject.
Score 5 Description	Exceptionally focused and entirely on topic; tightly centered on the subject.
Criteria Description	**Broad Coverage:** Does the expository article provide an in-depth exploration of the topic and have good coverage?
Score 1 Description	Severely lacking; offers little coverage of the topic's primary aspects.
Score 2 Description	Partial coverage; includes some main aspects but misses others.
Score 3 Description	Acceptable breadth; covers most main aspects of the topic.
Score 4 Description	Good coverage; achieves broad coverage of all major points.
Score 5 Description	Exemplary in breadth; thoroughly details all crucial aspects of the topic.
Criteria Description	**Logical Structure of the Expository Article:** Is the logical sequence of ideas presented in the expository article clear?
Score 1 Description	Completely disorganized; ideas are listed randomly without any logic.
Score 2 Description	Basic structure exists, but logic often breaks down; transitions are unclear.
Score 3 Description	Structure is relatively clear; most ideas are logical with minor gaps.
Score 4 Description	Logical structure is excellent; ideas are presented clearly and orderly.
Score 5 Description	Logical structure is perfect; ideas progress step by step with rigorous argument.
Criteria Description	**Breadth of Coverage in the Expository Article:** To what extent does the expository article cover different aspects and dimensions of the topic?
Score 1 Description	Extremely narrow scope; focuses on only one or two aspects.
Score 2 Description	Limited breadth; covers a few aspects, missing important content.
Score 3 Description	Moderate breadth; covers most major aspects, some briefly mentioned.
Score 4 Description	Good breadth; covers many aspects with a comprehensive overview.
Score 5 Description	Comprehensive coverage; explores every aspect and dimension.
Criteria Description	**Depth of Analysis in the Expository Article:** How thoroughly does the expository article analyze the details and implications of each aspect of the topic?
Score 1 Description	Analysis is only superficial; merely lists facts without exploration.
Score 2 Description	Analysis is shallow; provides basic details but no in-depth exploration.
Score 3 Description	Analysis has some depth; some content can be further explored.
Score 4 Description	In-depth analysis; comprehensively explores details and implications.
Score 5 Description	Analysis is extremely in-depth; digs into every detail and subtle relationship.

Table 7: Scoring rubrics on a 1-5 scale for explanatory articles by the evaluator LLM.

	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
			Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Direct Gen(GPT-4)	96.21	68.78	2.21	**2.87**	**3.17**	3.42	3.67	4.69	2.37	3.20
Direct Gen(DeepSeekR1)	**98.29**	**73.21**	**2.32**	2.58	3.10	**3.56**	**3.73**	4.68	**2.73**	**3.24**
Direct Gen(GPT-4o-mini)	95.75	68.20	2.18	2.57	3.15	3.35	3.51	**4.71**	2.46	3.13
Direct Gen(Qwen2-72B)	93.27	67.90	1.92	1.80	2.17	2.09	1.61	2.73	2.01	2.05
Direct Gen(Llama3.1-8B)	87.73	59.31	1.52	1.41	1.65	1.42	1.39	2.14	1.66	1.60
Direct Gen(Qwen3-1.7B)	85.67	62.87	1.54	1.47	1.43	1.29	1.31	1.98	1.57	1.51
Direct Gen(Qwen3-0.6B)	73.43	44.98	1.13	1.19	1.27	1.29	1.25	1.33	1.15	1.23
STORM(GPT-4o-mini)	95.91	68.89	3.05	4.11	3.68	3.72	3.92	4.62	3.32	3.77
OmniThink(GPT-4o-mini)	96.45	67.83	3.43	3.81	3.58	3.38	3.84	4.33	3.07	3.63
Logic(GPT-4o-mini)	**98.52**	**72.24**	**3.99**	**5.00**	**5.00**	**4.97**	**4.40**	**4.86**	**4.76**	**4.71**
STORM(Qwen2-72B)	87.63	68.33	2.92	3.76	3.28	3.34	2.79	4.01	3.11	3.46
OmniThink(Qwen2-72B)	93.24	68.87	3.58	3.77	3.59	4.63	2.09	4.37	3.71	3.68
Logic(Qwen2-72B)	95.42	69.82	3.83	4.89	4.62	4.73	4.28	4.73	4.68	4.54
STORM(Llama3.1-8B)	89.12	60.87	2.17	3.29	2.60	2.44	1.96	3.73	2.09	2.61
OmniThink(Llama3.1-8B)	89.09	61.19	2.57	3.47	3.31	2.94	2.01	2.68	2.07	2.72
Logic(Llama3.1-8B)	91.43	62.73	3.67	4.69	4.19	4.57	3.88	4.37	4.21	4.23
Logic(Qwen3-1.7B)	97.33	67.93	3.38	4.33	3.89	4.28	3.68	3.99	3.98	3.93
Logic(Qwen3-0.6B)	85.67	56.55	1.67	2.58	2.22	1.88	2.17	2.00	1.55	2.01
LOGIC-RL(Qwen3-1.7B)	**97.66**	**67.81**	**3.97**	**4.37**	**4.29**	**4.58**	**4.39**	**4.60**	**4.36**	**4.37**
LOGIC-RL(Qwen3-0.6B)	94.27	64.28	2.75	3.95	2.28	2.75	2.63	3.74	3.59	3.10

Table 8: Results of automatic outline quality evaluation of WikiOutline dataset.

	Refinement Steps	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
				Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Characters	1	92.32	**43.33**	**3.40**	2.20	1.50	1.00	1.70	2.60	2.30	2.10
	2	93.83	40.00	2.30	3.00	2.00	2.10	2.00	3.50	2.80	2.53
	3	**95.10**	**43.33**	2.80	3.70	**2.60**	2.60	**2.70**	3.50	**3.30**	**3.03**
	4	83.53	**43.33**	2.50	**3.80**	2.50	**2.70**	2.20	**3.60**	3.20	2.93
	5	76.85	**43.33**	1.90	3.10	2.30	1.80	1.50	3.50	2.60	2.39
Events	1	92.64	61.67	3.05	1.79	1.84	1.42	1.00	2.11	1.89	1.87
	2	90.89	61.67	2.25	3.85	2.20	2.40	2.05	3.35	2.95	2.72
	3	**94.47**	61.67	2.70	4.20	2.25	2.50	**2.60**	3.65	3.65	3.08
	4	87.12	61.67	2.89	**4.25**	2.50	3.15	3.10	3.95	3.68	3.36
	5	86.15	**63.33**	**3.20**	4.05	**2.60**	**3.30**	2.55	**4.25**	**3.80**	**3.39**
Films	1	92.18	**95.00**	**3.55**	3.40	**2.35**	1.45	1.60	2.75	2.80	2.56
	2	88.86	**95.00**	2.90	3.95	1.80	**3.68**	**2.30**	3.65	3.30	3.08
	3	**94.93**	**95.00**	2.70	3.90	1.95	3.10	2.10	3.65	3.15	2.94
	4	89.45	**95.00**	2.60	**4.75**	**2.35**	2.65	2.20	**4.05**	**3.35**	**3.14**
	5	82.72	**95.00**	2.15	4.15	2.30	2.95	1.90	3.55	**3.35**	2.91
Disasters	1	**96.09**	**75.00**	**3.16**	2.05	1.74	1.79	1.84	2.63	2.32	2.22
	2	92.18	70.00	2.60	3.20	1.40	2.10	2.30	3.90	3.45	2.71
	3	93.16	70.00	2.95	3.53	2.32	2.74	**2.84**	3.89	**3.95**	3.17
	4	88.16	70.00	2.70	3.80	**2.75**	2.60	3.25	**4.15**	3.60	**3.19**
	5	87.10	70.00	3.00	**3.75**	2.55	**3.05**	2.75	3.75	3.15	3.14
Places	1	93.19	**51.43**	**3.70**	2.60	2.20	2.20	1.10	2.80	2.90	2.50
	2	**95.54**	**51.43**	2.50	**4.60**	1.70	**2.90**	2.50	**4.50**	3.50	3.17
	3	93.69	**51.43**	2.60	4.40	2.30	2.80	**2.90**	4.00	**3.90**	**3.27**
	4	91.14	**51.43**	2.50	3.10	**2.50**	2.70	2.70	3.40	3.20	2.87
	5	89.64	**51.43**	2.00	3.20	2.40	2.40	2.60	3.80	2.70	2.73

Table 9: Detailed results of various domains for Qwen3-0.6B-RL with different refinement steps on the WikiOutline dataset.

	Refinement Steps	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
				Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Characters	1	**97.39**	**51.67**	3.90	**5.00**	4.20	3.00	3.30	**4.70**	**4.60**	4.10
	2	96.78	**51.67**	4.00	4.20	4.00	**4.60**	**4.40**	4.50	4.22	4.27
	3	97.21	**51.67**	**4.00**	**4.40**	**4.50**	4.50	4.00	4.60	4.40	**4.34**
	4	96.83	**51.67**	3.60	4.10	3.90	3.80	4.30	4.60	4.30	4.08
	5	97.21	**51.67**	3.60	4.40	3.50	2.80	3.30	4.40	4.00	3.71
Events	1	96.44	61.67	**4.00**	**4.85**	4.38	**5.00**	4.54	**4.92**	**4.92**	**4.66**
	2	96.82	61.67	3.97	4.63	3.79	4.53	4.70	4.60	4.65	4.41
	3	**97.67**	**62.00**	3.97	4.60	4.40	4.95	4.75	4.90	4.65	4.60
	4	97.19	**62.00**	3.95	4.25	**4.46**	4.30	**4.84**	**4.63**	4.37	4.39
	5	97.01	61.67	3.50	3.85	4.35	4.05	4.50	4.25	4.10	4.09
Films	1	94.54	**95.00**	**4.24**	**5.00**	4.00	**4.67**	4.00	**5.00**	**4.64**	**4.46**
	2	95.23	**95.00**	3.90	4.90	3.95	4.35	4.50	4.80	4.20	4.37
	3	**96.73**	**95.00**	3.94	4.86	**4.79**	4.42	**4.72**	4.92	4.42	4.58
	4	96.12	**95.00**	3.85	4.05	4.75	4.20	4.50	4.40	4.40	4.31
	5	95.87	**95.00**	3.80	4.40	3.90	4.40	3.70	4.55	4.60	4.19
Disasters	1	**99.56**	**75.00**	3.95	4.00	3.89	4.37	4.16	4.37	4.21	4.14
	2	99.12	**75.00**	3.85	**4.55**	**4.00**	4.25	4.00	4.20	4.15	4.14
	3	99.38	**75.00**	3.95	4.20	3.84	**4.65**	4.39	4.20	4.05	4.18
	4	98.89	**75.00**	**4.00**	4.35	3.90	4.50	**4.65**	**4.50**	**4.35**	**4.32**
	5	99.01	**75.00**	3.95	4.05	3.80	4.74	3.90	4.45	4.40	4.18
Places	1	98.15	**55.36**	**4.00**	4.10	4.00	4.50	**4.10**	4.00	4.10	4.11
	2	96.79	**55.36**	3.96	3.98	3.97	4.20	3.89	4.22	4.22	4.06
	3	97.32	**55.36**	**4.00**	3.80	3.90	4.40	**4.10**	4.40	4.30	4.13
	4	98.27	**55.36**	3.80	4.20	3.90	3.90	3.50	4.10	4.10	3.93
	5	**98.34**	**55.36**	3.80	**4.60**	**4.10**	**4.80**	4.00	**4.80**	**4.50**	**4.37**

Table 10: Detailed results of various domains for Qwen3-1.7B-RL with different refinement on the WikiOutline dataset.

	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
			Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3-0.6B	79.88	45.44	2.11	1.43	1.83	1.15	1.10	1.38	1.58	1.51
Qwen3-0.6B-SFT	91.77	64.29	3.17	1.80	1.79	**1.57**	1.15	2.38	2.02	1.98
Qwen3-0.6B-RL	**93.28**	**65.29**	**3.37**	**2.41**	**1.93**	**1.57**	**1.45**	**2.58**	**2.44**	**2.25**
Qwen3-1.7B	95.39	68.45	3.66	4.12	3.59	3.88	3.40	4.14	3.97	3.82
Qwen3-1.7B-SFT	96.75	69.74	3.88	4.41	3.89	3.91	3.86	4.31	4.37	4.09
Qwen3-1.7B-RL	**97.22**	**67.74**	**4.02**	**4.59**	**4.09**	**4.31**	**4.02**	**4.60**	**4.49**	**4.29**

Table 11: Ablation experimental results of Qwen3-0.6B and Qwen3-1.7B on the WikiOutline dataset.

	Refinement Steps	Heading Soft Recall	Heading Entity Recall	Rubric Grading							
				Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3 -0.6B-RL	1	93.28	**65.29**	**3.37**	2.41	1.93	1.57	1.45	2.58	2.44	2.25
	2	92.26	63.62	2.51	3.72	1.82	2.64	2.23	3.78	3.20	2.84
	3	**94.27**	64.28	2.75	**3.95**	2.28	2.75	2.63	3.74	**3.59**	**3.10**
	4	87.88	64.29	2.64	3.94	**2.52**	2.76	**2.69**	**3.83**	3.31	**3.10**
	5	84.49	64.62	2.45	3.65	2.43	**2.70**	2.26	3.77	3.12	2.91
Qwen3 -1.7B-RL	1	97.22	67.74	**4.02**	**4.59**	4.09	4.31	4.02	4.60	4.49	4.29
	2	96.95	67.74	3.96	4.45	3.94	4.39	4.30	4.46	4.29	4.25
	3	**97.66**	**67.81**	3.97	4.37	**4.29**	**4.58**	**4.39**	**4.60**	**4.36**	**4.37**
	4	97.46	**67.81**	3.84	4.19	4.18	4.14	4.36	4.45	4.30	4.21
	5	97.49	67.74	3.73	4.26	3.93	4.16	3.88	4.49	4.32	4.11

Table 12: Experimental results of Qwen3-0.6B-RL and Qwen3-1.7B-RL with different refinement steps on the WikiOutline dataset.

	Rubric Grading							
	Interest	Organization	Relevance	Coverage	Logicality	Breadth	Depth	Average
Qwen3-0.6B	1.72	2.23	1.45	1.37	1.29	2.09	1.48	1.66
Qwen3-0.6B-SFT	1.83	2.41	1.52	1.58	1.66	2.23	1.92	1.88
Qwen3-0.6B-RL	**1.99**	**2.73**	**2.21**	**1.80**	**2.09**	**2.43**	**2.17**	**2.20**
Qwen3-1.7B	3.17	3.69	3.24	3.29	3.87	3.91	3.97	3.59
Qwen3-1.7B-SFT	3.87	4.02	4.26	4.87	4.01	**4.76**	4.81	4.37
Qwen3-1.7B-RL	**4.00**	**4.29**	**4.53**	**4.88**	**4.09**	4.67	**4.82**	**4.47**

Table 13: Comparative experimental results of expository articles generated from different long-form outlines.

Task Type	Prompt Template
Reflective Reasoning on Feedback Provision	You are a specialist in Wikipedia-style entry outline evaluation and feedback design. Given the following materials: Draft Outline ({draft_outline}), Feedback ({feedback}), and Final Outline ({outline}), generate a professional reflective reasoning document. This document should focus on the methodological logic of feedback provision, specifically analyzing how the feedback is constructed to target the core deficiencies of the draft outline, align with the normative requirements of Wikipedia-style entry frameworks (e.g., logical hierarchy, comprehensiveness of key elements, relevance to the topic, structural rigor), and guide the refinement of the draft outline toward the final outline. The reasoning should clarify: (1) the criteria for identifying problematic points in the draft outline that the feedback focuses on; (2) how the feedback articulates revision directions while adhering to Wikipedia's editorial norms for outlines; (3) the logical connection between each feedback point and the corresponding improvements in the final outline.
Reflective Reasoning on Outline Refinement	You are a specialist in Wikipedia-style entry outline construction and refinement. Given the following materials: Draft Outline ({draft_outline}), Feedback ({feedback}), and Final Outline ({outline}), generate a professional reflective reasoning document. This document should systematically elaborate on the refinement logic and implementation path from the draft outline to the final outline, centered on the feedback. The reasoning should include: (1) the interpretation and decomposition of the feedback (clarifying the core demands and normative requirements implied in each feedback point); (2) the specific revision strategies adopted to address each feedback point (e.g., adjusting logical hierarchy, supplementing missing key sections, merging redundant content, optimizing terminology to conform to Wikipedia conventions); (3) how the revisions ensure the final outline meets the standards of a comprehensive, logically coherent Wikipedia-style entry (e.g., coverage of essential elements, hierarchical clarity, topic relevance); (4) the rationale for any trade-offs or prioritizations made during the refinement process (if applicable).
Wikipedia-Style Explanatory Section Drafting	You are tasked with writing a draft section of a Wikipedia article about the topic {T}. This section corresponds to the outline point: {o_i} (follow the structure of this outline point strictly, e.g., subheadings, logical flow, core focus as specified by {o_i}). To ensure accuracy and comprehensiveness, integrate the following evidence into your writing: {Evidence_Ei}. Your draft must adhere to Wikipedia's editorial guidelines: (1) maintain a neutral, objective tone; avoid subjective opinions, promotional language, or overly casual phrasing; (2) structure content logically to fully align with the outline point {o_i}; (3) embed the provided evidence into key claims: use definitions to clarify core concepts, use data/results to support factual statements, and explicitly note limitations where relevant; (4) use formal academic language consistent with Wikipedia's style; explain necessary jargon briefly and ensure accessibility to general readers; (5) do not introduce content irrelevant to {o_i} or the topic {T}; strictly stay within the scope of this section. Output only the draft section text.

Table 14: Inference-time prompt templates used for critique reflection, outline refinement reflection, and Wikipedia-style section drafting.