IS-CoT: Breaking the Long-form Generation Collapse via Interleaved Structural Thinking

Zechen Sun[1], Yuyang Sun[1], Zecheng Tang[1], Juntao Li[1]*, Wenpeng Hu[2]*, Wenliang Chen[1], Zhunchen Luo[2], Guotong Geng[2], Min Zhang[1]

[1]Institute of Computer Science and Technology, Soochow University
[2]Information Research Center of Military Science, PLA Academy of Military Science
zcsunc@stu.suda.edu.cn, ljt@suda.edu.cn, wenpeng.hu@pku.edu.cn

Abstract

Generating coherent and controllable long-form content remains a persistent challenge for Large Language Models (LLMs). While reasoning-enhanced models have demonstrated success in logic-intensive domains, our evaluation reveals that they suffer from a severe length collapse in open-ended writing, where performance degrades sharply as target lengths exceed 2,000 words. We attribute this failure to the limitation of static hierarchical planning, which struggles to provide dynamic guidance over extended contexts. To bridge this gap, we introduce the **Interleaved Structural Chain-of-Thought (IS-CoT)** framework. Unlike external agentic workflows, IS-CoT embeds a dynamic Plan-Write-Reflect cycle into the generation process, enabling continuous strategy adaptation and global alignment without additional assistance. Based on this framework, we construct a high-quality dataset of interleaved reasoning traces via a multi-teacher pipeline and train **IS-Writer-8B**. Experiments demonstrate that IS-Writer-8B achieves state-of-the-art performance on challenging long-form benchmarks (e.g., +3.08 vs. DeepSeek-V3.2 on LongBench-Write), exhibiting robust length compliance and coherence competitive with significantly larger proprietary models.

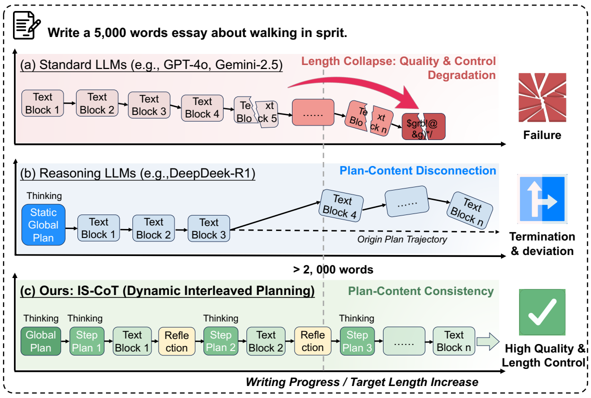

Figure 1: Comparison of three long-form generation paradigms. While existing methods degrade as target length increases due to static planning, our IS-CoT introduces a dynamic Plan-Write-Reflect cycle, maintaining coherence and length control over long horizons.

1 Introduction

With the remarkable success of large language models (LLMs) across various generation tasks (Comanici et al., 2025; DeepSeek-AI et al., 2025; Anthropic, 2024), there is a growing demand for generating long-form content, such as novels, technical reports, and screenplays (Wang et al., 2024). However, compared to short-text generation, producing coherent and strictly length-constrained long documents remains a significant challenge (Wu et al., 2025c; Yang et al., 2025c; Bai et al., 2025).

Recent research has focused on enhancing LLMs with deep reasoning capabilities (Shao et al., 2024; Wang et al., 2025c). While methods like DeepSeek-R1 (Guo et al., 2025) demonstrate impressive logic in mathematics and coding, their potential in open-ended long-form writing is less understood. In the writing domain, reasoning typically takes the form of static planning, where the model generates a comprehensive outline only at the initial stage (Gurung and Lapata, 2025; Wang et al., 2025b; Wu et al., 2025a). To evaluate the limits of this paradigm, we conducted a systematic preliminary study (Section 2). Our experiments reveal a critical phenomenon that we term *Length Collapse*: LLMs exhibit a sharp performance degradation as content length exceeds 2,000 words. We attribute this to the diminishing guidance of initial plans over long horizons, leading to a loss of narrative constraints. These findings suggest that one-shot reasoning is insufficient; effective long-form writing requires a mechanism to continuously update plans and reflect on progress.

To address this, we propose the **Interleaved**

*Corresponding Authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19874–19887
July 2-7, 2026 ©2026 Association for Computational Linguistics

Structural Chain-of-Thought (IS-CoT) framework. As illustrated in Figure 1, unlike static outlining methods, IS-CoT enforces a dynamic, recursive workflow comprising *global planning, local planning, content generation, and reflection*. This ensures that local execution remains aligned with global goals throughout the generation process. We implement this via a three-stage pipeline to construct a high-quality dataset of 5,000 samples with explicit interleaved reasoning traces, enabling the model to internalize this dynamic planning process.

By fine-tuning Qwen3-8B on this high-quality corpus, we develop IS-Writer-8B. Evaluations on LongBench-Write and WritingBench demonstrate that our model achieves state-of-the-art performance with average scores of 88.25 and 8.60, respectively. Our model not only surpasses the proprietary Gemini-2.5-Flash (+4.58) but also significantly outperforms larger open-source LLMs (e.g., DeepSeek-V3.2-671B) and other writing-enhanced LLMs (+10.28) on LongBench-Write. Notably, IS-Writer demonstrates consistent generalization across different length ranges without overfitting. This robustness is particularly strong in the challenging ultra-long [4k, 20k) range, where our method effectively mitigates generation collapse compared to baselines. Furthermore, our ablation studies verify the critical contributions of the *interleaved planning* and *reflection* modules, while case studies further highlight IS-Writer's precise controllability. In short, our contributions are:

- We empirically identify **Length Collapse** in long-form generation, demonstrating that static planning is insufficient for ultra-long horizons.
- We propose **IS-CoT** and a dataset with explicit `Plan-Write-Reflect` traces, enabling process supervision for dynamic structural planning.
- We train **IS-Writer-8B**, which achieves SOTA performance on long-form generation, demonstrating that a smaller model with dynamic reasoning can outperform much larger baselines.

2 Motivation and Preliminary Study

To investigate the limitations of LLMs in long-form generation, we conduct two preliminary studies: quantifying performance degradation across target lengths and evaluating different planning strategies.

2.1 *Length Collapse* in Long-Form Generation

While LLMs demonstrate strong performance in short text generation, their stability often decreases when generating very long content. In this study, we aim to measure this decline and explore the potential of reasoning capabilities.

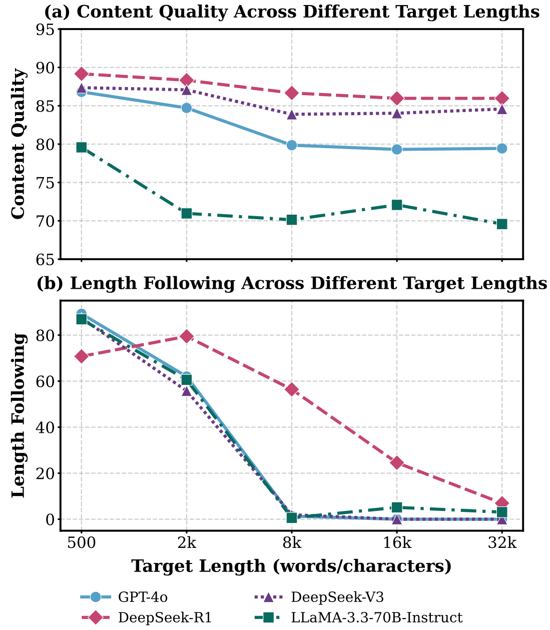

Figure 2: Comparison of quality and length-following scores for LLMs across increasing target generation lengths, with quality scores evaluated by GPT-4o-mini.

Experimental Setup. We evaluate four representative models: GPT-4o (Hurst et al., 2024), LLaMA-3.3-70B-Instruct (Dubey et al., 2024), DeepSeek-V3 (Liu et al., 2024), and the reasoning-enhanced DeepSeek-R1 (Guo et al., 2025). We constructed a dataset of 30 bilingual prompts (15 English/15 Chinese) across five target lengths (500 to 32,000 words), yielding 150 test cases per model evaluated via two metrics:

- **Quality Score (0-100):** Following previous work (Bai et al., 2025), we adopt LLM-as-a-judge to evaluate content quality on six dimensions: Relevance, Accuracy, Coherence, Clarity, Breadth and Depth, and Reading Experience.
- **Length Following Score (0-100):** We calculate the length score S_l using a piecewise linear function based on target length l and output length l'. The score is 100 for exact matches and decays linearly to 0 at $l/3$ and $4l$, formulated as:

$$S_l = \begin{cases} 100 \cdot \max(0, 1 - (l'/l - 1)/3) & \text{if } l' > l \\ 100 \cdot \max(0, 1 - (l/l' - 1)/2) & \text{if } l' \leq l \end{cases} \quad (1)$$

Results. As shown in Figure 2, with the increasing of required generation length, we observe two

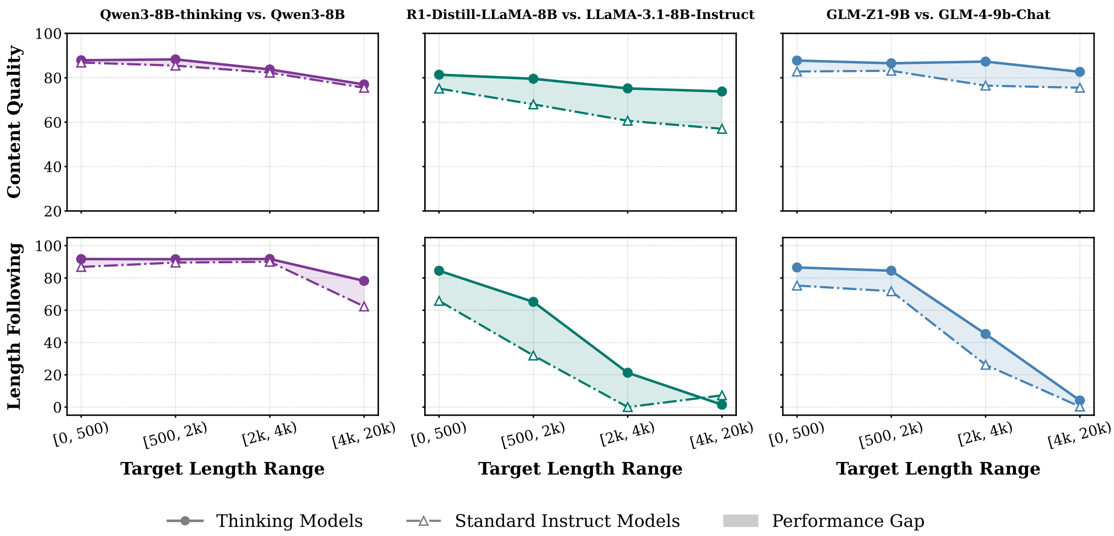

Figure 3: Performance comparison between reasoning-enhanced models ("Thinking") and standard instruct/chat models on LongBench-Write. Despite their superiority, "Thinking" models still degrade as length increases.

typical phenomena: (1) quality degrades; and (2) generation length collapse. Specifically, figure 2 (a) shows an inverse correlation between generation quality and target length, with the average Quality Score dropping by 5.83 across the [500, 32k] range. Notably, LLaMA-3.3 exhibits a more significant decline ($79.58 \rightarrow 69.58$) than GPT-4o ($86.81 \rightarrow 79.44$), highlighting the challenge of maintaining coherence in long-form generation. Figure 2 (b) indicates that length following degrades much more sharply than quality. The average score drops by 19.13 points from 500 to 2,000 words, followed by a critical collapse between 2,000 and 8,000 words, where the score falls to 15.07 ($\Delta \approx -68.5$). Models demonstrate complete failure at longer targets (16k, 32k) as average scores fall below 10 consistently.

Insight. DeepSeek-R1 outperforms all baselines on both metrics, exhibiting a significant advantage in length following for longer targets. This suggests the effectiveness of its internal "thinking" process in long-form generation. This observation motivates our core hypothesis: continuous planning throughout generation, rather than solely at the beginning, mitigates length and quality degradation. To explore this, we first examine the limitations of current planning methods in the following section.

2.2 Benefits and Limitations of "Thinking"

DeepSeek-R1's strong performance observed in Study 1 suggests that explicit reasoning capabilities (or "thinking" processes) are crucial for long-form generation. To confirm and verify this impact, we conducted a controlled comparison between models equipped with reasoning capabilities and their standard instruction-tuned versions.

Experimental Setup. We evaluated three pairs of models on the *LongBench-Write* benchmark. Each pair consists of a standard Instruct/Chat model and its reasoning-enhanced variant: (1) Qwen3-8B-Thinking vs. Qwen3-8B (Yang et al., 2025a); (2) R1-Distill-LLaMA-8B (Guo et al., 2025) vs. LLaMA-3.1-8B-Instruct (Dubey et al., 2024); (3) GLM-Z1-9B vs. GLM-4-9B-Chat (GLM et al., 2024). Consistent with Study 1, we report both the Quality Score and Length Following Score to assess performance across varying target lengths.

Results. Figure 3 demonstrate a significant advantage for the "think-before-you-write" paradigm, yet reveal the limitations of static planning at long horizons. As shown in Figure 3, the Thinking variants consistently outperformed their standard counterparts in both generation quality and length compliance across all three model families. This gap was particularly pronounced in the LLaMA series (average +14.57) and the GLM series (+9.15), indicating that initial planning leads to better performance than direct generation. However, despite outperforming baselines, the efficacy of this initial reasoning diminishes as the target length increases. For instance, the best-performing Qwen3-Thinking experienced a decrease in both Quality Score (-10.89) and Length Following Score (-13.53), with even more severe declines observed in the R1-

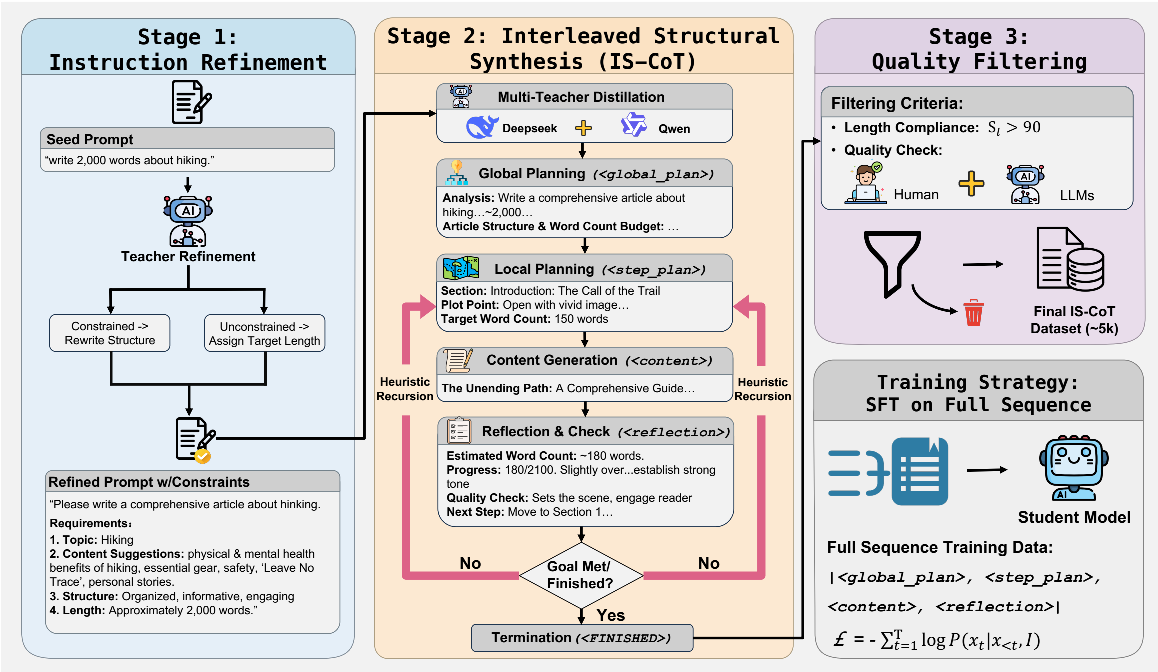

Figure 4: The overall framework of IS-CoT. We construct the IS-CoT Dataset through a three-stage pipeline: (1) Instruction Refinement, (2) Interleaved Structural Synthesis via a Multi-Teacher Distillation Framework, and (3) Quality Filtering. Finally, we perform SFT to internalize these reasoning capabilities.

Distill and GLM-Z1 models. This suggests that while initial reasoning provides a strong starting point, it is insufficient to maintain coherence and constraint adherence throughout the generation of 2,000+ words, highlighting a critical issue.

Takeaway: The necessity of interleaved guidance. These findings highlight a critical limitation in current reasoning paradigms: the reliance on *one-shot planning*. In standard reasoning models, "Thinking" occurs entirely at the beginning. However, the model tends to "forget" this initial plan during extensive generation. We argue that a static initial plan is insufficient for effective long-form generation. Instead, the model requires a mechanism to continuously guide and adjust its writing trajectory. This motivates our proposed Interleaved Structural CoT framework, which introduces a `Plan-Write-Reflect` cycle to support high-quality generation, as detailed in Section 3.

3 Methodology

This section details our methodology for constructing the **Interleaved Structural CoT (IS-CoT)** dataset and the subsequent training strategy. Embedding a `Plan-Write-Reflect` cycle into the training data enables the model to achieve fine-grained control over structure and length in long-form generation. As illustrated in Figure 4, our data construction pipeline proceeds in three sequential stages: Instruction Refinement, Interleaved Structural Synthesis, and Quality Filtering.

3.1 Stage I: Instruction Refinement

Developing models capable of controllable long-horizon planning demands a standardized set of challenging writing tasks. We start by sampling 6,000 seed prompts from the *DeepWriting* (Wang et al., 2025a) dataset, balanced between English and Chinese. This subset consists of 3,000 prompts with explicit length constraints and 3,000 without. To ensure high-quality supervision, we utilize DeepSeek-V3.2 to enhance these prompts. The process specifically addresses two scenarios:

- **Constrained Prompts:** It enhances clarity and structure by rephrasing the instructions.
- **Unconstrained Prompts:** It assigns a challenging target length, with a preference for long-form tasks to facilitate long-context training.

The outcome is a set of 6,000 clear writing instructions, each with an explicit length target.

3.2 Stage II: Interleaved Structural Synthesis

Standard CoT typically generates reasoning only at the beginning, which often leads to guidance decay in long-form generation. To address this, we design a recursive generation workflow that interleaves planning, writing, and reflection.

Multi-Teacher Distillation Framework. We employ DeepSeek-V3.2 and Qwen3-235B-A22B-Instruct as teacher models for data synthesis. This dual-source strategy exposes the student model to diverse writing styles and distinct reasoning patterns, effectively preventing it from overfitting to the specific biases of a single teacher.

The Generation Workflow. We formulate long-form writing as an iterative decision-making process. To enable explicit planning and reflection, the generation is structured into four components defined by specific tokens:

1. **Global Planning (`<global_plan>`)**: Constructs a high-level outline based on the prompt. It decomposes the task into sections and assigns a length budget to each, ensuring the global structure satisfies the total length requirement.
2. **Local Planning (`<step_plan>`)**: Designs the trajectory for the current segment. This includes outlining points, specifying constraints, and defining the length for the current segment.
3. **Content Generation (`<content>`)**: Generates content based on the local plan, strictly adhering to the required format (e.g., report structure).
4. **Reflection (`<reflection>`)**: Performs a reflection after each segment. The model calculates length progress, verifies logical coherence, and adjusts the subsequent plan if necessary.

Heuristic-Guided Recursive Generation. To ensure strict adherence to the interleaved format and length constraints, we implement a heuristic control loop during synthesis. We continuously inspect the output for the `<reflection>` token to immediately evaluate the current completion status:

- **Continuation:** If the accumulated length is below the target, the system forces the generation of a `<step_plan>` token, prompting the model to initiate planning for the subsequent section.
- **Termination:** Conversely, if the content is complete and satisfies all constraints, the loop concludes with a `<FINISHED>` token.

This active intervention ensures structural integrity and effectively mitigates the premature termination commonly observed in standard LLMs generation.

3.3 Stage III: Quality Filtering

Following data synthesis, we apply a rigorous filtering pipeline to guarantee both generation quality and length compliance. Quality assessment combines LLM-as-a-Judge with manual verification. For length compliance, we utilize the metric S_l from Section 2.1, retaining only samples where $S_l > 90$. This yields the final `IS-CoT` dataset of approximately 5,000 high-quality samples.

3.4 Training Strategy

We train our student models via Supervised Fine-Tuning (SFT) on the `IS-CoT` dataset. Unlike standard methods that remove intermediate rationale, we train on the complete sequence including `<global_plan>`, `<step_plan>`, and `<reflection>` tokens. The objective function is the standard autoregressive negative log-likelihood:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t | x_{<t}, \mathcal{I}), \tag{2}$$

where x denotes the unified sequence of reasoning and content tokens, and $\mathcal{I}$ is the instruction. This explicitly forces the model to learn planning and reflection as core components of a coherent writing process, rather than as additional tasks.

4 Experiments

This section presents the experimental setup, comparative results against advanced baselines, ablation study and case study for our **IS-Writer-8B**.

4.1 Experimental Setup

Implementation Detail. IS-Writer-8B is fine-tuned from Qwen3-8B on our `IS-CoT` dataset. Training utilizes DeepSpeed ZeRO-3 across 64 NVIDIA H800 GPUs with a global batch size of 64 and a learning rate of 2×10^{-5}. To capture long-range dependencies, we extend the context window to 32,768 tokens and train for 5 epochs.

Baseline. We compare IS-Writer-8B against three categories of strong competitors:

- **Proprietary LLMs:** Representative closed-source models, including GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet (Anthropic, 2024), and Gemini-2.5-Flash (Comanici et al., 2025).

Models	LongBench-Write				WritingBench							
	Avg.	S_l	S_q	[4k, 20k)	L_R	L_C	D1	D2	D3	D4	D5	D6
Proprietary LLMs												
GPT-4o	71.61	57.29	85.94	41.09	7.35	7.16	7.39	6.86	7.03	7.45	7.48	7.70
GPT-4o-mini	72.22	59.45	84.98	43.60	7.47	7.41	7.41	7.01	7.18	7.56	7.73	7.78
Claude-3.5-sonnet	72.90	60.17	85.63	45.74	7.34	7.08	7.23	7.19	6.64	7.41	7.71	7.80
Gemini-2.5-flash	83.67	79.35	87.99	82.11	8.47	8.49	8.44	8.41	8.48	8.55	8.49	8.43
Open-source LLMs												
DeepSeek-R1-671B	81.87	75.31	88.44	70.81	8.48	8.23	**8.58**	8.52	8.57	8.50	8.46	8.27
DeepSeek-V3-671B	75.10	62.76	87.43	49.92	8.05	7.57	7.93	7.77	8.11	8.10	8.11	8.10
DeepSeek-V3.2-671B	85.17	80.24	**90.10**	71.03	**8.60**	**8.59**	8.53	8.55	8.59	**8.73**	**8.63**	**8.50**
Qwen3-235B-A22B-Instruct	87.15	85.07	89.24	71.91	8.45	8.45	**8.58**	8.43	8.40	8.60	8.44	8.26
LLaMA3.3-70B-Instruct	66.45	55.58	77.33	35.45	6.58	6.40	6.52	6.10	6.26	6.43	7.13	7.01
Capability-enhanced LLMs												
Suri-I-ORPO-7B	44.25	42.42	46.08	29.67	2.24	2.05	2.45	2.70	1.83	2.00	2.41	2.43
LongWriter-8B	77.97	77.25	78.68	75.37	4.44	4.14	4.50	3.66	4.12	4.13	5.24	5.00
Writing-Model-Qwen-7B	70.24	58.77	81.70	39.77	8.35	8.39	8.57	8.26	8.38	8.21	8.49	8.30
Deepwriter-8B	57.94	41.98	73.90	37.04	5.81	5.44	6.29	4.69	5.23	5.79	6.23	6.46
IS-Writer-8B (ours)	**88.25**	**88.31**	88.19	**85.46**	**8.60**	8.44	8.56	**8.61**	**8.60**	8.68	8.60	8.47

Table 1: Main performance comparison on LongBench-Write (scale: 0-100; see Appendix B for details) and WritingBench (scale: 1-10). The six domains evaluated in WritingBench include: (D1) Academic & Engineering, (D2) Finance & Business, (D3) Politics & Law, (D4) Literature & Art, (D5) Education, and (D6) Advertising & Marketing. The **best** and second-best results are highlighted in bold and underlined, respectively.

- **Open-Source LLMs:** Large-scale open-weight models, including DeepSeek-R1, DeepSeek-V3, DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Qwen3-235B-Instruct (Yang et al., 2025a), and LLaMA3.3-70B-Instruct (Dubey et al., 2024).
- **Capability-enhanced LLMs:** Models specialized in long-context generation, including Suri-I-ORPO (Pham et al., 2024), LongWriter-8B (Bai et al., 2025), Writing-Model-Qwen-7B (Wu et al., 2025c), and DeepWriter-8B (Wang et al., 2025a).

Benchmark and Evaluation. We employ *LongBench-Write* (Bai et al., 2025) to assess comprehensive capabilities in ultra-long generation, measuring both length compliance and quality metrics. Additionally, we leverage the length-constrained subset of *WritingBench* (Wu et al., 2025c) and assess performance across six writing domains. Following official recommendations, we adopt the LLM-as-a-Judge with GPT-4o-mini to ensure fair and scalable evaluation.

4.2 Main Results

Table 1 presents the comprehensive performance comparison. Our proposed **IS-Writer-8B** achieves state-of-the-art performance among writing-enhanced LLMs, and remarkably surpasses several proprietary LLMs and much larger open-source models across both benchmarks.

Superiority in Comprehensive Long-Form Generation. IS-Writer-8B achieves an average score of 88.25 on LongBench-Write, outperforming the best proprietary model Gemini-2.5-Flash (+4.58) and Qwen3-235B (+1.10). Notably, in the challenging [4k, 20k) length range detailed in Table 1, IS-Writer demonstrates a substantial advantage, surpassing the best open-source LLM (+13.55) and writing-enhanced LLM (+10.09). While some baselines show higher quality scores (S_q), they suffer in length compliance (S_l), often benefiting from the judge's bias toward shorter outputs. In contrast, IS-Writer-8B achieves the highest length score (S_l: 88.31) without compromising coherence. Similarly, on WritingBench, our model matches the much larger DeepSeek-V3.2 (8.60) and achieves SOTA results in Finance (D2) and Politics (D3). This confirms that our Interleaved Structural CoT paradigm enhances general writing capabilities.

Surpassing Teacher Models via Dynamic Planning. A significant finding is that IS-Writer outperforms the teacher models (Qwen3-235B and DeepSeek-V3.2) used to synthesize its training data. While teacher models possess vast knowledge, they struggle to organize it over ultra-long contexts in a single inference pass. By distilling the writing process into explicit steps (Plan-Write-Reflect) and utilizing a multi-teacher framework for data robustness, IS-Writer-8B effectively learns

Models	Overall			[0, 500)		[500, 2k)		[2k, 4k)		[4k, 20k)	
	$\bar{S}$	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q
Qwen3-8B	83.52	83.23	83.82	86.82	86.07	**89.53**	**89.05**	**90.00**	86.67	62.36	69.67
IS-Writer-8B	**88.25** (+4.73)	**88.31**	**88.19**	**89.60**	**91.67**	88.56	87.98	87.68	**88.13**	**86.75**	**84.17**
w/o Reflection	85.97 (+2.45)	85.90	86.04	88.95	86.85	85.47	86.34	87.09	86.46	81.80	84.17
w/o Interleaved	84.07 (+0.55)	83.03	85.11	81.04	85.55	85.05	85.95	80.49	86.67	84.15	81.88

Table 2: Ablation results on LongBench-Write. **IS-Writer-8B** represents the full model, while *w/o Reflection* and *w/o Interleaved* denote the variants without the reflection module and interleaved planning, respectively.

a dynamic planning logic that proves more robust than the unstructured generation of its teachers.

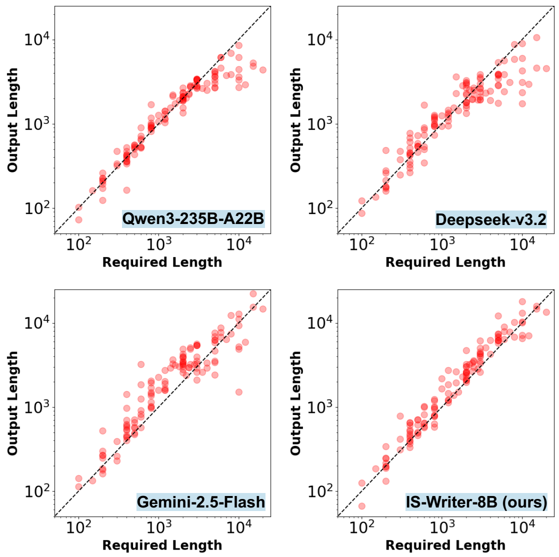

Figure 5: Model response length w.r.t. instruction required length on LongBench-Write.

Robustness in Ultra-Long Length Compliance. Figure 5 visualizes the length compliance in the range of [0, 20k). IS-Writer demonstrates strong robustness across various length constraints, especially in the challenging [4k, 20k) range. In this specific range, open-source LLMs exhibit a "lazy" behavior, tending to under-generate as requirements increase. While proprietary LLMs such as Gemini-2.5 perform better, they still show high variance. In contrast, IS-Writer consistently meets or slightly exceeds target lengths. This precision is attributed to the interleaved reasoning, where the model dynamically verifies its progress and adjusts next plans, effectively mitigating the "early stoping" common in long-form generation.

4.3 Ablation Study

To verify the contributions of core components within the IS-CoT framework, we conduct ablation studies on LongBench-Write, comparing the full model against variants with components removed.

Impact of Components. As shown in Table 2, IS-Writer-8B demonstrates a substantial overall improvement, outperforming the base Qwen3-8B by 4.73 points. Removing the reflection module (*w/o Reflection*) reduces the gain to +2.45, while further removing the local planning module (*w/o Interleaved*) drops the improvement to just +0.55. This trend confirms that while global planning offers a basic direction, the combination of step-wise local planning and dynamic reflection is essential for maintaining high-quality long-form generation.

Robustness in Ultra-Long Contexts. The advantages of our method are most significant in the ultra-long [4k, 20k) range. Without explicit planning, Qwen3-8B performs poorly with an average score of 66.02. In contrast, IS-Writer-8B achieves an average score of 85.46 (+19.44). Notably, both ablation variants also score above 80 in this range, validating that explicit thinking processes, whether global or interleaved are critical for mitigating generation collapse in extended contexts.

Analysis of Short-Range Performance. We observe a small performance drop for IS-Writer in the [500, 2k) interval compared to the baseline. This is likely due to the data distribution in our training set because it contains fewer samples in this range. It may also relate to the bias of LLM judges toward standard responses for shorter texts. Nevertheless, considering the significant gains achieved in complex, long-form generation tasks, this slight decline is an acceptable trade-off for the model's specialized capability in handling extended outputs.

4.4 Case Study

We analyze the narrative task "Debut" (Figure 6) to demonstrate the efficacy of IS-CoT. While baselines like Gemini-2.5-Flash and DeepSeek-V3.2 produce coherent text, they often fail to meet specific length or narrative constraints due to the

Prompt

On stage, opera actors have their moments of appearance. Life also has many moments of appearance: speeches under the national flag, reports on research findings, launches of new products… Every appearance attracts attention; every appearance also tells a story. Please write a narrative essay entitled "Debut." Requirements: Healthy themes, rich and reasonable …….; fluent language, clear handwriting; **no less than 800 words.**

Gemini-2.5-flash

Response : On stage, opera performers adorned in elaborate costumes freeze in a single, meaningful pose—a flicked sleeve, a dramatic turn—amidst the clanging of gongs and drums, drawing thunderous applause. It is not merely a momentary stillness, but a concentrated explosion of character emotion and narrative depth……

Judge: effectively captures the essence of the user's request by narrating a personal experience that embodies the theme of debut, **however,** while the response is strong in many areas, there are moments where the narrative **could be tightened for better impact.**
Quality/Length Score: 83.25/62.75

DeepSeek-V3.2

Response : Unveiling\nOn the third day of dismantling the stage, the weathered dressing mirror in the backstage area finally lay exposed beneath the scorching July sun. The gilded characters "Clouds evoke thoughts of robes" on its frame had largely peeled away……

Judge: well-crafted narrative that effectively explores the theme of appearance, **However**, while it is artistically strong, it may **not fully align with the user's request** for a straightforward narrative……
Quality/Length Score: 87.50/91.63

IS-Writer (ours)

Response: ***<global_plan>*** Establish a four-part structure for the narrative piece titled "Debut" with an 800-word word count…… ***<step_plan>*** Plan the opening scene. ***<content>*** "Make your entrance!" That was the bell for first period this morning—and the first true public debut of my life……***<reflection>*** Check progress (120/800) and confirm thematic resonance…… ***<reflection>*** Complete the final quality check. ***<FINISHED>***

Judge: "The response **effectively captures the essence of 'debut'** by narrating a personal story of a public speaking event. **aligns well with the user's request**......
Quality/Length Score: 95.75/99.63

Figure 6: Case study on a long-form narrative task from LongBench-Write, where IS-Writer uses interleaved reasoning, with additional case studies presented in Appendix C.

lack of explicit continuous planning. In contrast, IS-Writer-8B initiates with a <global_plan> and interleaves generation with <local_plan> and <reflection>. As shown in Figure 6, the model explicitly checks its progress (e.g., "word count: 120/800") in the reflection phase and adjusts the next writing plan accordingly. This dynamic mechanism ensures strict constraint adherence, yielding a length compliance score of 99.6. This demonstrates that controllable long-form generation requires a unified process of continuous planning and reflection rather than a large context window.

5 Related Work

Long-Form Generation and Planning. While Large Language Models (LLMs) excel in short-context tasks, they struggle to maintain coherence in open-ended long-form generation (Ping et al., 2025; Lei et al., 2025; Que and Rong, 2025). Extended sequences often suffer from progressive quality decay (Yang et al., 2025b,c; Liu et al., 2025; He et al., 2025) and the "lost-in-the-middle" phenomenon (An et al., 2024; Baker et al., 2024), leading to a negative length-quality correlation. To address these issues, hierarchical "Plan-then-Write" paradigms (Bai et al., 2025; Wu et al., 2025b; Zhao et al., 2025) have been proposed to decompose complex tasks into global outlines followed by sequential decoding. However, these frameworks typically rely on static planning that cannot adapt to content development during generation. In contrast, our approach replaces inflexible planning with an interleaved dynamic framework, enabling continuous alignment between plan and content.

Chain-of-Thought and Generation with Reasoning. Chain-of-Thought (CoT) prompting has achieved significant success in logic-intensive domains like mathematics and coding (Wei et al., 2022; Shao et al., 2024; Yang et al., 2025d; Xu et al., 2025). Recent studies advance this by distilling reasoning capabilities into smaller models (Hsieh et al., 2023; Wang et al., 2023; Feng et al., 2024; Dai et al., 2024; Wang et al., 2025a; Zhuang et al., 2025; Yin et al., 2025). Despite these advancements, explicit reasoning in creative long-form generation remains underexplored. Existing methods primarily confine reasoning to a pre-generation planning phase (Liang et al., 2024), failing to leverage reasoning capabilities for intermediate adjustments. We bridge this gap by introducing an interleaved structural CoT framework that combines global planning, local reasoning, and reflection, allowing the model to dynamically refine its strategy based on the generated context.

6 Conclusion

In this work, we systematically evaluate current LLMs on long-form generation, revealing that even reasoning-enhanced models suffer from "Length Collapse" due to the diminishing effectiveness of static planning over extended horizons. To address this limitation, we present the `IS-CoT` framework, which introduces a dynamic "Plan-Write-Reflect" cycle to ensure continuous alignment with global constraints. By training on the `IS-CoT` dataset with explicit intermediate reasoning steps, our model learns to execute dynamic planning strategies through process supervision. Our experiments demonstrate that **IS-Writer-8B** achieves superior performance with remarkable data efficiency, surpassing significantly larger LLMs. This work underscores the necessity of dynamic guidance for long-context tasks and provides a robust framework for enhancing the controllability of LLMs.

Limitations

Despite the promising performance of IS-Writer in controllable long-form generation, we acknowledge several limitations in our current work:

Dependency on Teacher Quality. Our training data relies on a multi-teacher synthesis pipeline. Although IS-Writer demonstrates performance that matches or even exceeds the teacher models in long-form generation tasks, the diversity and knowledge scope of the model are still influenced by the source supervision. We believe that integrating significantly stronger or a more diverse set of teachers in the future could further unlock the model's potential and generalization capabilities.

Inference Overhead. The core mechanism of `IS-CoT` involves generating interleaved "Plan" and "Reflect" tokens alongside the content. While this significantly enhances coherence and length compliance, it inevitably increases token consumption during inference. Compared to direct generation models, IS-Writer requires more computation time to produce the same amount of content. Future work could explore token-efficient mechanisms or reasoning approaches to mitigate this overhead.

Acknowledgments

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by the National Science Foundation of China (NSFC No. 62576232), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX25_3467), Key Laboratory of Data Intelligence and Advanced Computing, Soochow University.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188.

Anthropic. 2024. Introducing claude 3.5 sonnet. `https://www.anthropic.com/research/claude-3-5-sonnet`.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context llms. In *The Thirteenth International Conference on Learning Representations*.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa. *arXiv preprint arXiv:2412.10079*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Beyond imitation: Learning key reasoning steps from dual chain-of-thoughts in reasoning distillation. *arXiv preprint arXiv:2405.19737*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. 2024. Keypoint-based progressive chain-of-thought distillation for llms. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13241–13255.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Alexander Gurung and Mirella Lapata. 2025. Learning to reason for long-form story generation. In *Second Conference on Language Modeling*.

Jacqueline He, Howard Yen, Margaret Li, Shuyue Stella Li, Zhiyuan Zeng, Weijia Shi, Yulia Tsvetkov, Danqi Chen, Pang Wei Koh, and Luke Zettlemoyer. 2025.

Precise information control in long-form text generation. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xuanyu Lei, Chenliang Li, Yuning Wu, Kaiming Liu, Weizhou Shen, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2025. Writing-rl: Advancing long-form writing via adaptive curriculum reinforcement learning. *arXiv preprint arXiv:2506.05760*.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and 1 others. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

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*.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. Verifact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17919–17936.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1722–1753, Miami, Florida, USA. Association for Computational Linguistics.

Bowen Ping, Jiali Zeng, Fandong Meng, Shuo Wang, Jie Zhou, and Shanghang Zhang. 2025. LongDPO: Unlock better long-form generation abilities for LLMs via critique-augmented stepwise information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7613–7632, Vienna, Austria. Association for Computational Linguistics.

Haoran Que and Wenge Rong. 2025. Pic: Unlocking long-form text generation capabilities of large language models via position id compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6982–6995.

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

Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, and 1 others. 2025a. Reverse-engineered reasoning for open-ended generation. *arXiv preprint arXiv:2509.06160*.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025b. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391.

Ruida Wang, Yuxin Li, Yi R. Fung, and Tong Zhang. 2025c. Let's reason formally: Natural-formal hybrid reasoning enhances LLM's math capability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16820, Suzhou, China. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

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.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2025a. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning. *arXiv preprint arXiv:2506.18841*.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Juanzi Li, and Roy Ka-Wei Lee. 2025b. Superwriter: Reflection-driven long-form generation with large language models. *arXiv preprint arXiv:2506.04180*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025c. Writingbench: A comprehensive benchmark for generative writing. *Preprint*, arXiv:2503.05244.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.

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

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025b. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. *arXiv preprint arXiv:2502.12583*.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025c. Uncle: Uncertainty expressions in long-form generation. *arXiv preprint arXiv:2505.16922*.

Wen Yang, Minpeng Liao, and Kai Fan. 2025d. Markov chain of thought for efficient mathematical reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7132–7157.

Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang, Weihua Luo, and 1 others. 2025. Marco-o1 v2: Towards widening the distillation bottleneck for reasoning models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23506–23516.

Xueguan Zhao, Wenpeng Lu, Chaoqun Zheng, Weiyu Zhang, Jiasheng Si, and Deyu Zhou. 2025. Plan dynamically, express rhetorically: A debate-driven rhetorical framework for argumentative writing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9584.

Xianwei Zhuang, Zhihong Zhu, Zhichang Wang, Xuxin Cheng, and Yuexian Zou. 2025. Unicott: A unified framework for structural chain-of-thought distillation. In *The Thirteenth International Conference on Learning Representations*.

A `IS-CoT` Dataset Statistics

We provide detailed statistics of the 4,988 high-quality training samples in the `IS-CoT` dataset to demonstrate its diversity and coverage.

A.1 Domain Distribution

As shown in Figure 7, the dataset covers a broad range of topics. "Artistic" writing (e.g., novels, scripts) makes up the largest part (48.4%), helping the model learn narrative structures. Moreover, professional fields like "Finance & Business" and "Science & Engineering" are also well-represented. This mix ensures IS-Writer performs effectively for both creative stories and technical reports.

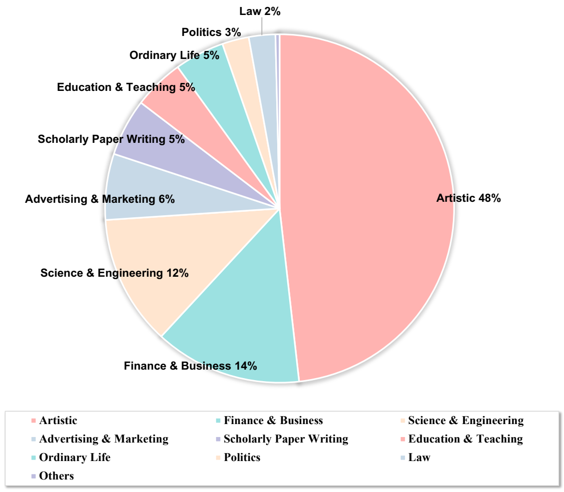

Figure 7: Domain distribution of the `IS-CoT` dataset. The diversity of topics ensures robust generalization.

A.2 Length Distribution

To address the "Length Collapse" problem mentioned earlier, our dataset focuses on long-form generation. Figure 8 shows the length distribution. Over 60% of the samples are longer than 2,000 words, with the [4,000, 8,000) range being the most common (26.36%). We also include a significant number of very long samples (13.05%) in the [8,000, 16,000) range. This helps the model learn the dynamic *Plan-Write-Reflect* cycle for consistent and coherent long-document generation.

B Detailed Results on LongBench-Write

We present detailed performance results of the models across different target length ranges on the LongBench-Write benchmark in Table 3.

Comparison with Baselines. IS-Writer-8B significantly outperforms all other capability-enhanced LLMs (e.g., LongWriter-8B, DeepWriter-8B) across every length. Furthermore, it generally surpasses leading proprietary models in overall scores, including GPT-4o and Gemini-2.5-Flash.

Performance by Length Range. In the medium-length range (e.g., [500, 4k)), our model performs slightly below the best open-source models, such as DeepSeek-V3.2 and Qwen3-235B. We attribute

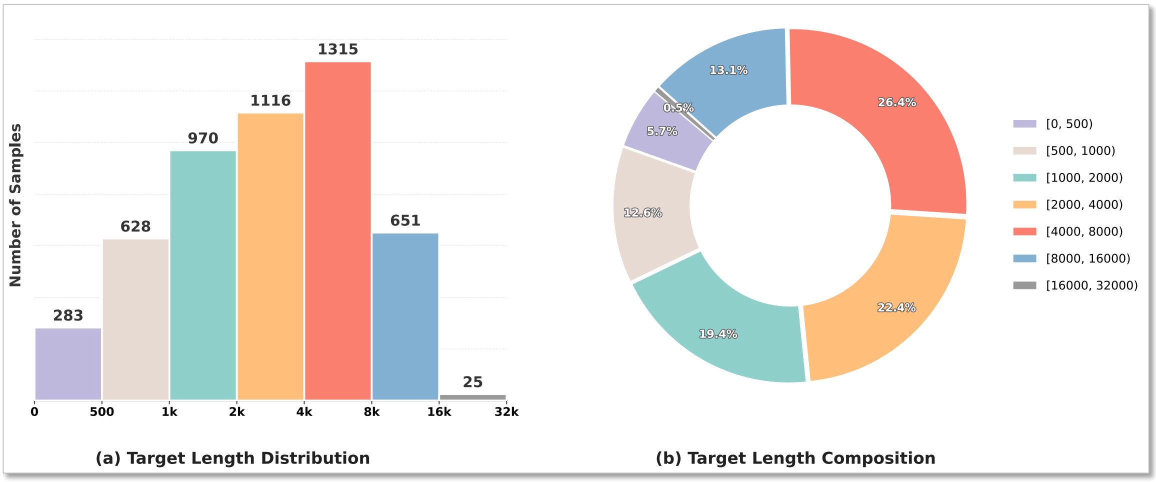

Figure 8: Target length distribution of the IS-CoT dataset. The dataset is concentrated on lengths over 2,000 words to support long-form generation.

Models	Overall			[0, 500)		[500, 2k)		[2k, 4k)		[4k, 20k)	
	$\bar{S}$	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q
Proprietary LLMs											
GPT-4o	71.61	57.29	85.94	**94.39**	89.58	79.28	86.43	22.24	83.75	0.00	82.17
GPT-4o mini	75.56	65.84	85.28	93.65	87.76	86.70	85.76	52.08	84.58	5.37	81.83
Claude 3.5 Sonnet	72.90	60.17	85.62	93.06	88.54	75.40	85.27	39.44	85.00	8.47	83.00
Gemini-2.5-Flash	83.67	79.35	87.99	85.51	90.89	72.12	88.28	84.80	86.88	79.54	84.67
Open-source LLMs											
DeepSeek-R1	81.87	75.31	88.44	81.34	90.76	80.04	87.98	81.32	87.50	54.63	87.00
DeepSeek-V3	75.10	62.76	87.43	90.65	90.63	81.18	88.57	35.55	85.83	17.17	82.67
DeepSeek-V3.2	85.17	80.24	**90.10**	89.03	90.23	87.02	**90.50**	85.38	**90.63**	53.22	**88.83**
Qwen3-235B-A22B-Instruct	87.15	85.07	89.24	92.09	89.06	**93.12**	89.44	**93.73**	90.00	55.31	88.50
LLaMA3.3-70B-Instruct	66.45	55.58	77.33	87.13	83.85	78.06	77.23	24.88	76.46	1.08	69.83
Capability-enhanced LLMs											
Suri-I-ORPO (7B)	44.25	42.42	46.08	56.70	55.34	50.19	50.29	23.19	38.33	26.17	33.17
LongWriter-8B	77.78	77.25	78.30	86.63	77.86	68.41	77.62	84.82	82.92	74.40	76.33
Writing-Model-Qwen-7B	70.24	58.77	81.70	75.69	81.38	81.07	85.47	52.38	81.67	3.86	75.67
DeepWriter-8B	61.26	47.93	74.58	76.13	80.99	62.77	75.29	21.79	72.50	7.24	66.83
IS-Writer-8B	**88.25**	**88.31**	88.19	89.60	**91.67**	88.56	87.98	87.68	88.13	**86.75**	84.17

Table 3: Detailed evaluation results on LongBench-Write across different length ranges. The best results are highlighted in **bold**, demonstrating that IS-Writer-8B achieves the highest overall average score.

this to our training data distribution, which is heavily weighted towards the [4k, 8k) range to focus on long-form generation. Nevertheless, our model remains competitive in these shorter length ranges.

Overall, IS-Writer-8B achieves the highest overall average score ($\bar{S} = 88.25$) among all evaluated models. This advantage is mainly due to its strong performance in the challenging ultra-long [4k, 20k) range. For instance, the length score of DeepSeek-V3.2 experiences a sharp decline to 53.22 in this specific range. While other models suffer from severe degradation (e.g., LLaMA3.3 dropping to 1.08 in length score), IS-Writer maintains a high length score (S_l) of 86.75. This demonstrates that our approach effectively prevents the loss of control during long-form generation, thereby validating the effectiveness of the IS-CoT framework in maintaining precise controllability and coherence.

C Case Studies

To demonstrate the performance of IS-Writer, we compare it against two strong baselines: Gemini-2.5-Flash and DeepSeek-V3.2. We select two distinct representative cases, covering Chinese creative writing and English technical writing, to evaluate generation quality and length control.

C.1 Chinese Creative Writing

The first case involves writing a popular financial article titled *"Do not miss these top-tier products!"* The prompt requires a story-like structure to attract readers on a recommendation feed, with a minimum length of 1,500 words, creating a significant challenge for maintaining coherence. The comparison is shown in Figure 9. **Analysis:** Gemini-2.5-Flash fails to meet the length requirement, resulting in a low length score of 51.71. While the content is structured, it lacks the necessary depth. DeepSeek-V3.2 performs better (Quality: 87.50, Length: 82.33) but misses opportunities to connect with a broader audience. In contrast, IS-Writer achieves the highest scores in both aspects (Quality: 95.75, Length: 90.07). The evaluation shows that IS-Writer produces a "compelling narrative" with "relatable stories" while keeping the content clear. This proves that IS-Writer can handle long writing tasks effectively, maintaining engagement consistently throughout the entire document.

C.2 English Technical Writing

The second case requires a technical introduction to *"Convolutional Long- and Short-Term Neural Networks (ConvLSTM)"* for travel time prediction, with a specific target length of about 600 words. The comparison is shown in Figure 10. **Analysis:** DeepSeek-V3.2 (Quality: 79.17) gives a detailed explanation but is considered "overly technical," leading to a lower quality score. Gemini-2.5-Flash provides a strong response (Quality: 87.50) but lacks sufficient practical examples. IS-Writer outperforms both baselines with a Quality score of 91.67 and a near-perfect Length score of 99.07. The evaluation highlights the response's "solid understanding" and "well-structured" format. The high length score confirms the robust effectiveness of the `IS-CoT` framework in precise control, ensuring the model follows the instructions precisely without generating unrelated content. This demonstrates that our method effectively optimizes content depth while respecting strict constraints.

Prompt

任务:\n以“这些顶流产品，千万别错过了！”为标题写一篇爆文级别的财经类文章，内容结构要遵循叙述模型，能够引起读者情感共鸣，同时适合在信息流平台上推荐。要求:\n1.文章开头应该让引人入胜，能够吸引读者的兴题\n2.文章内容应该引发读者的情感共鸣。\n3.语言表达要清晰流畅，符合信息流推荐的写作风格，适合快速阅读和吸引读者。\n4. 文章内容请严格按照所发的主题来，不要省略\n5、文章字数不低于1500字

Gemini-2.5-flash

Response :这些顶流产品，千万别错过了！**\n\n### **引子：你是否还在为“错过”而懊悔？**\n\n亲爱的朋友们，你是否曾有过这样的经历？\n\n在某个不经意的瞬间，你听到朋友们在谈论某项投资，当时你或许不以为意，或许觉得风险太大，或许只是单纯地“没时间研究”。然而，几年后，当那项投资的价值翻了几番，当那些当初“敢于吃螃蟹”的朋友们已经实现了财富的跃迁，你是否会感到一丝丝的懊悔？\n\n“如果当时我买了那只股票……”\n“如果我早点关注那个赛道……”\n“如果我没有把钱放在银行里睡大觉……”这些“如果”像一根根细针，扎在心头，提醒着我们，在这个瞬息万变的时代，错过一次机会，可能就错过了改变命运的契机。财富的列车呼啸而过，你，是选择做那个在站台旁，眼睁睁看着机会溜走的人，还是那个果断跳上车，与时代同行，共享繁华的智者……

Judge: The response effectively addresses the user's request by providing a well-structured, engaging, and emotionally resonant article on trending financial products. **However**, while it presents a compelling narrative and covers various investment themes, it **lacks some specificity and depth in certain areas**, which could enhance its overall quality. Additionally, certain segments could benefit from clearer transitions to improve flow. Overall, it is a strong response **but has room for improvement in detail and coherence**.
Quality/Length Score: 79.17/51.71

DeepSeek-V3.2

Response: 这些顶流产品，千万别错过了！你是否曾在深夜刷手机时，突然被一条财经新闻击中内心——某款产品悄然崛起，成为市场黑马，而你却因犹豫错过了最佳入场时机？那种“只差一步”的遗憾，是否让你在无数个夜晚反复琢磨：“如果当时……”？在瞬息万变的财经世界里，机会常如流星划过，唯有敏锐的眼光与果断的行动，才能抓住那些即将改变格局的“顶流产品”。今天，就让我们一同揭开这些潜力之星的神秘面纱……一、时代浪潮下的产品革命：为何“顶流”总能改写规则？\n\n回顾过去十年，从智能手机到电动汽车，从移动支付到人工智能芯片，每一个“顶流产品”的诞生，都不仅仅是一次技术升级，更是一场深刻的社会经济变革。它们共同的特质是什么？正是切中了时代脉搏，解决了未被满足的需求，并以颠覆性的体验重塑了用户习惯。比如，当特斯拉初代电动车面世时，多数人视其为小众玩物；但那些早早看到清洁能源与智能驾驶趋势的投资者，已随其股价飙升收获了数十倍回报……

Judge:The response is well-structured and adheres closely to the user's request for a compelling financial article. It effectively engages the reader's emotions, provides relevant examples, and maintains a clear narrative flow. **However**, there are areas where it could improve, particularly in terms of clarity and connecting with a broader audience. Overall, it is a strong response **with some minor weaknesses**.

Quality/Length Score: 87.50/82.33

IS-Writer (ours)

Response： ***<global_plan>***分析请求类型：这是一篇财经类爆文，主题为“这些顶流产品，千万别错过了！”，目标是吸引读者并适合信息流平台推荐。文章需要具备情感共鸣、清晰流畅的语言和结构化的叙述模型。\n\n概述整体结构：1. **引言部分（约200字）**：以一个生动、具象的场景切入，比如一位普通家庭主妇或年轻上班族在使用顶流产品后的巨大改变，迅速抓住读者注意力，提出核心问题：为什么有些产品能真正改变生活？它们有什么共同点？***<step_plan>***当前部分：引言部分。\n概述：以一个具体、有画面感的家庭场景开篇……提出问题“这些顶流产品，到底值不值得买？”。\n目标字数：约200字。\n***<content>***\n凌晨两点，窗外还下着雨。李芳蜷缩在床上，怀里抱着发烧的女儿，额头全是冷汗。她翻遍了药箱，试了各种方法，可孩子的体温一直没降下来。就在她几乎要崩溃时，手机突然震动——是她刚入手的智能健康手环发出的紧急提醒：“检测到异常心率，建议立即就医。”\n\n她抱着孩子冲进急诊，经过一系列检查后，医生确诊是幼儿急疹，但因为发现及时，避免了高烧惊厥的风险。那一刻，李芳看着手环上跳动的绿色数据，眼眶湿润了……***<reflection>***\n本部分预估字数：约250字……\n进度检查：250 / 1500。已完成约17%。\n质量检查：逻辑通顺……***<FINISHED>***

Judge: The response effectively addresses the user's request by presenting a compelling narrative that engages readers emotionally while discussing top trending financial products. It follows a coherent structure, includes relatable stories, and maintains clarity throughout. **Overall, it is a strong response that aligns well with the user's requirements.**
Quality/Length Score: 95.75/90.07

Figure 9: Comparison on a Chinese creative writing task. Baselines struggle with length or depth, while IS-Writer generates a coherent article that strictly follows the $>$1500 word requirement.

Prompt

Introduction to convolutional long- and short-term neural networks, introduction to convolutional, pooling, and fully connected layers, and extraction of ……for travel time prediction, **600 words.**

Gemini-2.5-flash

Response: Accurate travel time prediction is a cornerstone of modern intelligent transportation systems, impacting everything from navigation apps to logistics and urban planning. The inherent complexity of traffic flow, driven by intricate spatial and temporal dependencies, poses a significant challenge for traditional prediction models. Deep learning, particularly the fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) into Convolutional LSTMs (ConvLSTMs), offers a powerful solution by effectively capturing these dynamic patterns.\n\nAt the heart of this approach are the fundamental building blocks of CNNs: convolutional, pooling, and fully connected layers.\n\n**Convolutional Layers** are the primary feature extractors. They apply a set of learnable filters (or kernels) that slide across the input data, performing dot products and generating feature maps. This process leverages three key principles……

Judge: The response provides a detailed and relevant overview of convolutional long- and short-term neural networks, particularly focusing on their application in travel time prediction. It covers the essential components of CNNs and LSTMs and explains how their integration into ConvLSTMs allows for effective spatio-temporal feature extraction……**However**, while it is informative, it could be enhanced with more examples or practical applications to deepen the reader's understanding. Overall, it is a strong response with minor areas for improvement.
Quality/Length Score: 87.50/90.89

DeepSeek-V3.2

Response: Introduction to Convolutional Long- and Short-Term Neural NetworksConvolutional Long- and Short-Term Neural Networks (ConvLSTMs) represent a powerful fusion of two cornerstone deep learning architectures: Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). This hybrid model is specifically engineered to handle **spatiotemporal data**—information that varies across both space and time. Traditional LSTMs excel at modeling temporal sequences (like a time series of travel speeds) but treat input features as independent points, losing crucial spatial relationships. CNNs, conversely, are masters at extracting hierarchical spatial patterns from grid-like data (such as images or road network maps) but lack inherent temporal memory. ConvLSTMs solve this by replacing the fully connected layers inside an LSTM cell with **convolutional operations**……

Judge: The response provides a thorough explanation of Convolutional Long- and Short-Term Neural Networks (ConvLSTMs) and their application in travel time prediction. It effectively covers the core components of the architecture and illustrates how the model processes spatial and temporal information. **However,** while the content is generally relevant and accurate, some sections could benefit from clearer connections to the user's request, particularly in relation to travel time prediction. Overall, it is a well-structured and informative response, though it may be overly technical for some readers.

Quality/Length Score: 79.17/92.67

IS-Writer (ours)

Response: ***<global_plan>***Request Type:** Technical Introduction/Explanatory Article.\n **Overall Structure**: A single, cohesive article introducing the concepts in a logical sequence.\n **Word Count Allocation**:\n 1. Introduction: Setting the stage for CNNs and their application in travel time prediction. (~100 words)\n 2. Core Concepts: Explanation of Convolutional, Pooling......***<step_plan>***Current Section: Introduction.\n **Overview**: Open the article by stating the importance of ……Target Word Count:100 words……***<content>***\n In the realm of smart transportation and urban mobility, accurate travel time prediction stands as a cornerstone for optimizing traffic flow, enhancing public transit scheduling, and reducing congestion. Traditional statistical models often fall short due to the complex, spatially and temporally correlated nature of traffic data……This article introduces the fundamental architecture of CNNs…….***<reflection>***\n **Estimated Word Count**: ~120 words.Progress:120 / 600 words.\n **Quality Check**: Introduction is concise and sets the necessary context…….**Next Step Adjustment**: Proceed to the core concepts section, ensuring a balanced explanation of each layer's function……***<FINISHED>***

Judge: The response provides a thorough introduction to convolutional neural networks (CNNs) and their application in travel time prediction, effectively covering the requested topics. It demonstrates a solid understanding of the subject matter……Overall, the response is well-structured and informative.
Quality/Length Score: 91.67/99.07

Figure 10: Comparison on an English technical writing task. IS-Writer achieves better content quality and near-perfect length precision (Score: 99.07), outperforming baselines in both accuracy and control.