Plan-and-Write: Structure-Guided Length Control for LLMs without Model Retraining

Adewale Akinfaderin
akinfaa@amazon.com
Amazon Web Services
Seattle, WA, USA

Shreyas Subramanian
subshrey@amazon.com
Amazon Web Services
Seattle, WA, USA

Akarsha Sehwag
akshseh@amazon.com
Amazon Web Services
Seattle, WA, USA

Abstract

Length control in Large Language Models (LLMs) is a crucial but under-addressed challenge, with applications ranging from voice interfaces requiring concise responses to research summaries needing comprehensive outputs. Current approaches to length control, including Regularized DPO, Length-Instruction Fine-Tuning, and tool-augmented methods, typically require expensive model retraining or complex inference-time tooling. This paper presents a prompt engineering methodology that enables precise length control without model retraining. Our structure-guided approach implements deliberate planning and word counting mechanisms within the prompt, encouraging the model to carefully track and adhere to specified length constraints. Comprehensive evaluations across six state-of-the-art LLMs demonstrate that our method significantly improves length fidelity for several models compared to standard prompting when applied to document summarization tasks, particularly for shorter-to-medium length constraints. The proposed technique shows varying benefits across different model architectures, with some models demonstrating up to 37.6% improvement in length adherence. Quality evaluations further reveal that our approach maintains or enhances overall output quality compared to standard prompting techniques. Our approach provides an immediately deployable solution for applications requiring precise length control, particularly valuable for production environments where model retraining is impractical or cost-prohibitive.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Human-centered computing** → Natural language interfaces; • **Information systems** → Language models.

Keywords

Prompt Engineering, Large Language Models, Length Control & Fidelity

ACM Reference Format:
Adewale Akinfaderin, Shreyas Subramanian, and Akarsha Sehwag. 2025. Plan-and-Write: Structure-Guided Length Control for LLMs without Model Retraining. In *Proceedings of First International KDD Workshop on Prompt Optimization, 2025 (Prompt Optimization KDD '25).* ACM, Toronto, ON, Canada, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Length control is a fundamental yet frequently overlooked aspect of language model capabilities. As Large Language Models (LLMs) become increasingly integrated into real-world applications, the ability to precisely control response length emerges as a critical requirement. Different use cases demand different response lengths: voice interfaces require concise answers, research summaries need comprehensive detail, mobile applications have screen space constraints, and documentation systems must provide varying levels of information density [?]. Despite their remarkable progress in reasoning and knowledge tasks, even state-of-the-art LLMs struggle with this basic aspect of communication, often violating explicit length constraints in numerous cases [? ?].

The inability to control output length creates significant challenges for production deployments. Unpredictable response lengths lead to poor user experiences, increased computational costs, and reduced system reliability. Moreover, many LLMs exhibit a systematic bias toward verbosity, a phenomenon that compromises both their efficiency and utility [? ?]. While several methods have been proposed to address this issue, these approaches typically require expensive model retraining, complex fine-tuning procedures, or resource-intensive inference-time mechanisms [? ?]. These limitations make them impractical for many production environments, particularly those utilizing black-box LLMs where model parameters are inaccessible [?].

Our work introduces "Plan-and-Write," a prompt engineering methodology that enables precise length control without model retraining. Unlike existing approaches, our method operates entirely within the prompt space, making it immediately deployable across any LLM without additional training or tooling requirements. The core innovation of our approach lies in a structure-guided planning technique that encourages models to strategically allocate their word budget across content sections. By incorporating explicit planning and word-counting mechanisms within carefully designed prompts, we guide LLMs to maintain awareness of their output length while preserving response quality.

We evaluate our approach across seven state-of-the-art LLMs on document summarization tasks with varying target length requirements. Our results demonstrate that the Plan-and-Write methodology significantly improves length fidelity compared to standard prompting techniques, with minimal impact on response quality. We identify patterns in model behavior that suggest advanced LLMs possess untapped capabilities for length control that can be effectively activated through structured prompting. See Table 1 for examples of length-controlled generations using both standard prompting and our structure-guided approach. Our evaluation shows that our method achieves substantially higher precision in meeting exact word count targets while maintaining or enhancing the quality and relevance of model responses.

Prompt Optimization KDD '25, Toronto, ON, Canada
2025. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Example length-controlled document summarization using different prompting strategies. The vanilla approach often fails to meet exact target lengths, while our thinking approach demonstrates precise word count control.

Instruction	Target	Generation	Length
Summarize this document in exactly {target_words} words.	20	Amazon had a strong year in 2023, with revenue, operating income, and free cash flow growth. The company is investing in primitives and generative AI capabilities.	26
	50	Amazon had a strong year in 2023, with revenue growth, improved profitability, and continued customer experience enhancements. The company is focused on building foundational "primitive" services to rapidly innovate and empower both internal and external builders, with a particular emphasis on generative AI capabilities.	46
	100	Amazon's 2023 annual letter to shareholders highlights the company's strong financial performance, with 12% revenue growth and a dramatic improvement in operating income and free cash flow. The letter emphasizes Amazon's focus on customer experience, with enhancements in selection, pricing, and delivery speed. It also discusses the company's progress in Advertising, AWS, and newer business investments like Prime Video and Project Kuiper. The letter emphasizes Amazon's "primitives" approach, which involves building foundational services to enable rapid innovation. It also outlines the company's vision for Generative AI and its potential to transform various Amazon businesses. Overall, the letter conveys Amazon's optimism and conviction in its long-term growth and innovation potential.	118
YOUR ONLY TASK: Summarize this document in EXACTLY {target_words} WORDS ({target_words}). I will only accept summaries with EXACTLY {target_words} words - count carefully. STEP 1: Count each word as you write: <thinking> 1 First 2 word ... {target_words} lastword </thinking> STEP 2: Rewrite those SAME {target_words} WORDS in paragraph form: <final_answer> First word... [EXACTLY {target_words} WORDS TOTAL] </final_answer>	20	Amazon's strong financial results, customer experience, and focus on building primitive services to empower builders and innovation across businesses.	20
	50	Amazon saw strong growth in 2023, with revenue, operating income, and free cash flow improving significantly. The company is investing in customer experience, logistics, and advertising, while also making progress in AWS, Prime Video, and new initiatives like Generative AI and Project Kuiper. Amazon remains focused on long-term value creation.	50
	100	Amazon saw strong growth in 2023, with revenue increasing 12% to $575 billion. Operating income and free cash flow also improved significantly. The company attributed these results to its focus on customer experience, including expanded selection, competitive pricing, and faster delivery. Amazon continued investing in key initiatives like AWS, Prime Video, and Project Kuiper. The letter discusses the company's "primitives" approach to building flexible, reusable services that enable rapid innovation. It also highlights Amazon's focus on generative AI as a transformative technology. Overall, the letter conveys the company's enthusiasm and optimism for the future.	99

2 Related Work

Length control in LLMs has emerged as a critical challenge as these models become increasingly integrated into real-world applications with specific output constraints. Research on text generation control dates back to earlier work by Holtzman et al. [?], who explored cooperative discriminators to guide language generation. However, as noted in recent benchmarks, modern LLMs still struggle with this fundamental aspect of communication, often violating explicit length constraints in over 50% of cases [?]. Current research addressing this challenge has primarily explored three categories of approaches: model fine-tuning, inference-time modifications, and prompt engineering techniques.

Several works have focused on modifying model parameters through specialized fine-tuning procedures. Yuan et al. [?] proposed Length-Instruction Fine-Tuning (LIFT), which augments preference datasets with examples where responses violating length constraints automatically lose in preference pairs. Zhou et al. [?] introduced token-level reward regularization (T-REG) to improve preference optimization, which could benefit length control among other aspects. Park et al. [?] developed Regularized Direct Preference Optimization to address length exploitation in RLHF by adding a principled regularization term. Dubois et al. [?] identified systematic biases toward verbosity in LLMs and proposed a causal inference framework to debias evaluation metrics. These approaches show promising results but require expensive retraining procedures and access to model weights, making them impractical for many deployment scenarios involving black-box LLMs.

Inference-time methods attempt to control output length without modifying model parameters. Gu et al. [?] developed an iterative sampling framework based on the Metropolis-Hastings algorithm that treats length control as sampling from a target distribution. Their approach requires multiple inference passes. Nayab et al. [?] explored the relationship between output verbosity and reasoning quality, proposing Constrained Chain-of-Thought prompting to explicitly limit reasoning length. While these methods offer more flexibility than fine-tuning approaches, they often introduce additional computational overhead during inference.

More closely aligned with our approach, prompt engineering techniques attempt to achieve length control through careful instruction design without modifying models or using complex inference procedures. Zhang et al. [?] demonstrated that structured prompting can guide LLMs to better understand and follow formatting constraints in data representation tasks. Lyu et al. [?] explored zero-shot in-context learning with pseudo-demonstrations, showing how carefully constructed prompts can guide model behavior without parameter updates. Bai et al. [?] introduced constitutional AI techniques that improve instruction following in general, which indirectly benefits length constraint adherence. Wang et al. [?] explored a primal-dual approach for controlled question generation that incorporates specific constraints during generation. Despite these advances, there remains a significant gap between theoretical approaches and practical, deployment-ready solutions for precise length control. Most current methods either sacrifice response quality or work only for specific types of constraints. Our Plan-and-Write methodology addresses these limitations through a prompt engineering approach that enables precise word-count control without model retraining or additional inference overhead, while maintaining response quality, providing an immediately deployable solution for practitioners using black-box LLMs.

3 Methodology

3.1 Overview

We introduce Plan-and-Write, a prompt engineering methodology for precise length control in large language models. The core innovation of our approach is the decomposition of the generation process into two distinct phases:

(1) a planning phase where the model explicitly counts words as it drafts content, and
(2) a verification phase where the model reconstructs the content into a coherent output of exactly the specified length. This structure-guided approach leverages the LLM's inherent capabilities for metacognition and self-monitoring without requiring any model parameter modifications or additional inference passes.

Unlike traditional approaches that simply include a length constraint in the instruction (e.g., "Summarize in X words"), our method guides the model through a deliberate process of word counting and budget allocation. This creates what we term "length awareness"—an explicit tracking mechanism that helps the model maintain precise control over its output length while preserving content quality. The approach is model-agnostic and can be applied to any LLM capable of following multi-step instructions.

3.2 Framework

We formally define the length-controlled generation problem as follows: Given a document D, a target word count t, and an LLM M, find a summary S that maximizes the semantic relevance to D while strictly adhering to the target length:

$$S = \arg\max_{s} \text{Quality}(s, D) \quad \text{subject to} \quad |s| = t \tag{1}$$

where $|s|$ represents the word count of summary s, and $\text{Quality}(s, D)$ measures semantic similarity and relevance between the summary and the original document.

Traditional approaches attempt to solve this directly with a simple constraint in the prompt:

$$S_{\text{vanilla}} = M(D, P_{\text{vanilla}}(t)) \tag{2}$$

where $P_{\text{vanilla}}(t)$ represents a prompt instructing the model to generate a summary of length t.

Our Plan-and-Write approach decomposes this into two phases:

$$S_{\text{draft}} = f_{\text{planning}}(M, D, t) \tag{3}$$

$$S_{\text{final}} = f_{\text{verify}}(S_{\text{draft}}, t) \tag{4}$$

This decomposition enables the model to first plan a response with explicit word counting, followed by a verification step that ensures the exact target length is met. We measure the effectiveness of our approach using length fidelity error:

$$E = ||\ |S| - t\ || \tag{5}$$

with the goal of minimizing E to 0 while maintaining high quality output.

Phase 1: Planning with Explicit Word Counting. The model generates content while numbering each word sequentially, creating "length awareness" throughout the generation process.

Phase 2: Verification and Coherence. The model reformats the same content into a coherent paragraph while maintaining the exact word count, ensuring both precision in length and quality in expression.

The final response is extracted from between the <final_answer> tags, providing a clean output that meets the exact target length. This approach requires no model retraining or additional inference passes, making it immediately deployable with any LLM capable of following multi-step instructions.

3.3 Justification

Our approach is based on three simple but effective principles that explain why explicit word counting improves length control in LLMs:

Explicit Monitoring. By instructing the model to count words as it generates, we create a simple tracking mechanism that makes the length constraint concrete rather than abstract. This transforms the vague instruction "write exactly t words" into a procedural task with clear progress indicators.

Two-Phase Structure. Separating content generation from final formatting allows the model to first focus on meeting the word count exactly, then focus on making the text coherent and fluent. This division of the task reduces the cognitive burden on the model by addressing one constraint at a time.

Leveraging Existing Capabilities. Rather than requiring new model capabilities through fine-tuning, our method simply provides a process that helps the model apply its existing abilities more effectively. Modern LLMs can count and follow instructions—our prompt structure just guides them to apply these skills to solve the length control problem.

This straightforward approach explains why Plan-and-Write achieves better length fidelity than traditional prompting methods. By providing a clear process rather than just stating a constraint, we help the model systematically achieve exact word counts while maintaining response quality.

4 Experiments

We designed a comprehensive evaluation framework to assess the effectiveness of our Plan-and-Write methodology across multiple dimensions, including model capabilities, task types, and target lengths. Our experiments focused on comparing the length fidelity of our structure-guided approach against standard prompting techniques while maintaining output quality.

4.1 Experimental Setup

4.1.1 Models. We evaluated our approach on six state-of-the-art large language models representing diverse training methodologies and architectural designs: Claude 3 Haiku, Claude 3.5 Haiku, Claude 3.5 Sonnet, Claude 3.7 Sonnet, Mistral Large, and Meta's Llama 3.1 70B. These models span different parameter sizes and capabilities, allowing us to assess the generalizability of our approach across the current landscape of LLMs.

4.1.2 Tasks and Target Lengths. We conducted experiments on document summarization, where models were provided with a PDF document (Amazon's 2023 Shareholder Letter) and instructed to create summaries of varying specified lengths.. To evaluate length control across different magnitudes, we tested eight target word counts: 20, 50, 100, 200, 500, 1000, 2000, and 5000 words. This range enabled us to observe model performance on both extremely concise outputs and more extensive generations.

4.1.3 Evaluation Metrics. Our primary metric was Mean Absolute Percentage Deviation (MAPD), which measures the percentage error between generated and target word counts:

$$\text{MAPD} = \frac{|\text{generated\_words} - \text{target\_words}|}{\text{target\_words}} \tag{6}$$

MAPD provides a normalized measure of length fidelity independent of target size, enabling fair comparison across different length targets.

4.2 Implementation Details

4.2.1 Document Processing. For the summarization task, we provided the document as a PDF file through the models' document understanding capabilities. This allowed us to test length control in a realistic setting where models must process and condense complex multi-page documents.

4.2.2 Prompt Variants. We designed and evaluated four distinct prompting strategies to systematically test different implementations of length control. The first two variants represent conventional approaches that simply state the length constraint (with slight variations in phrasing), while the latter two implement our Plan-and-Write methodology in different forms. The thinking variants differ in their framing and structure: Thinking V1 emphasizes explicit word counting with a procedural approach, while Thinking V2 employs a scientific framing with hierarchical information organization. This design allows us to evaluate both the effectiveness of explicit planning compared to conventional prompting and the impact of different planning frameworks on length control:

(1) **Vanilla V1**: A straightforward instruction to summarize in the target word count.

Vanilla V1 Prompt

Summarize this document into exactly {target_words} words.

(2) **Vanilla V2**: A variation of the standard prompt using different phrasing.

Vanilla V2 Prompt

Transform this document into exactly {target_words} words.

(3) **Thinking V1**: Our Plan-and-Write approach with explicit word counting.

Thinking V1 Prompt

```
YOUR ONLY TASK: Summarize this document in EXACTLY {target_words} WORDS ({target_words}). I will only accept summaries with EXACTLY {target_words} words - count carefully.
STEP 1: Count each word as you write:
<thinking>
1 First
2 word
...
{target_words} lastword
</thinking>
STEP 2: Rewrite those SAME {target_words} WORDS in paragraph form:
<final_answer>
First word... [EXACTLY {target_words} WORDS TOTAL]
</final_answer>
```

(4) **Thinking V2**: A scientific framing of our approach with structured planning.

Thinking V2 Prompt

TASK: Transform this document to EXACTLY {target_words} words while maximizing information preservation.
SCIENTIFIC METHODOLOGY: 1. First, identify the core information hierarchy and key points 2. Then perform controlled expansion to EXACTLY {target_words} words by: a) Preserving primary information structures b) Including supporting details proportionally c) Maintaining relative emphasis on topics from original document d) Adding clarifying context where needed to reach target length
EXECUTION:
<thinking>
• First outline core information structure
• Draft initial version (likely shorter than target)
• Systematically expand by adding:
- Supporting examples
- Contextual details
- Clarifying explanations
• Count words meticulously: 1, 2, 3... until reaching {target_words}
</thinking>
Final {target_words}-word document:

4.2.3 Generation Protocol. For each model and target length combination, we conducted five independent attempts to account for generation variance. This resulted in 960 individual generations (6 models × 8 target lengths × 5 attempts × 4 prompt variants). Between attempts, we implemented a delay to manage API rate limits and ensure model availability.

4.2.4 Word Counting Methodology. We employed NLTK's word_tokenize function to standardize word counting across all outputs, excluding punctuation marks. For the "thinking" prompt variants, we extracted only the final answer text from between the designated tags to evaluate word count accuracy.

4.2.5 Analysis Framework. After collecting results, we computed statistical measures including mean and standard deviation of word count accuracy across models and prompt variants. We also conducted significance testing to determine whether improvements from the Plan-and-Write methodology were statistically significant compared to baseline approaches.

This experimental design allows us to systematically evaluate both the effectiveness of our approach across different models and tasks and to identify patterns in how modern LLMs respond to different length control strategies.

5 Results

Our evaluation reveals significant differences in length control capabilities across models and prompting strategies. We present results from document summarization task, examining how our Plan-and-Write approach compares to standard prompting methods.

Figures 1 and 2 illustrate the ratio of generated to target length across different models. With vanilla prompting (Figure 1), most models exhibit considerable variation in output length, with a systematic bias toward over-generation, particularly for shorter targets. In contrast, our Plan-and-Write approaches (Figure 2) demonstrate tighter clustering around the ideal ratio of 1.0 across all length targets, providing visual evidence that explicit word counting significantly improves length fidelity.

Table 2 presents the Mean Absolute Percentage Deviation (MAPD) for each model across four prompting strategies. Lower values indicate better performance. Several key findings emerge from our analysis:

- **Plan-and-Write Effectiveness:** Our structure-guided approaches (Thinking V1 and V2) outperformed standard prompting for four out of six models, demonstrating their broad applicability across model architectures. For Claude 3.7 Sonnet, Thinking V1 reduced MAPD to 0.088, representing a 37.6% improvement over the best vanilla approach. For Claude 3 Haiku, Thinking V2 achieved the best results with an MAPD of 0.120.
- **Model-Specific Performance:** Llama 3.1 70B demonstrated exceptional length control across all prompting strategies, achieving the lowest overall MAPD of 0.027 with the Vanilla V2 prompt. This suggests that some models may have already developed strong length control capabilities during their training.
- **Length Scaling:** As shown in Figures 1 and 2, most models demonstrated better length adherence for longer target lengths (500+ words) compared to shorter targets. This pattern was consistent across prompting strategies but was less pronounced with our Plan-and-Write approaches.
- **Variance Reduction:** The standard deviation of MAPD was typically lower for our thinking approaches compared to vanilla prompts, especially for more advanced models. This indicates that explicit word counting leads to more consistent and predictable length control.
- **Scientific Framing Benefits:** Our results suggest that for some models like Claude 3 Haiku, the scientific framing of the task in Thinking V2 provides additional benefits, indicating that both counting mechanisms and task framing play important roles in length control.

Our results shows a correlation between model capabilities and length control performance. More advanced models like Claude 3.7 Sonnet and Llama 3.1 70B demonstrated superior length fidelity across all prompting strategies. However, the relative improvement from structure-guided prompting was most pronounced in mid-tier models like Claude 3.5 Sonnet and Claude 3.5 Haiku, suggesting that explicit planning mechanisms are particularly beneficial for these models.

5.1 Quality Evaluation with LLM-as-a-Judge

While length fidelity is our primary focus, it is essential to ensure that improved length control does not come at the expense of output quality. To systematically evaluate content quality across prompting strategies, we employed LLM-as-a-Judge (LLMaaJ), an increasingly common evaluation method that leverages large language models to assess semantic properties beyond surface-level metrics. This approach allows us to evaluate whether our structure-guided

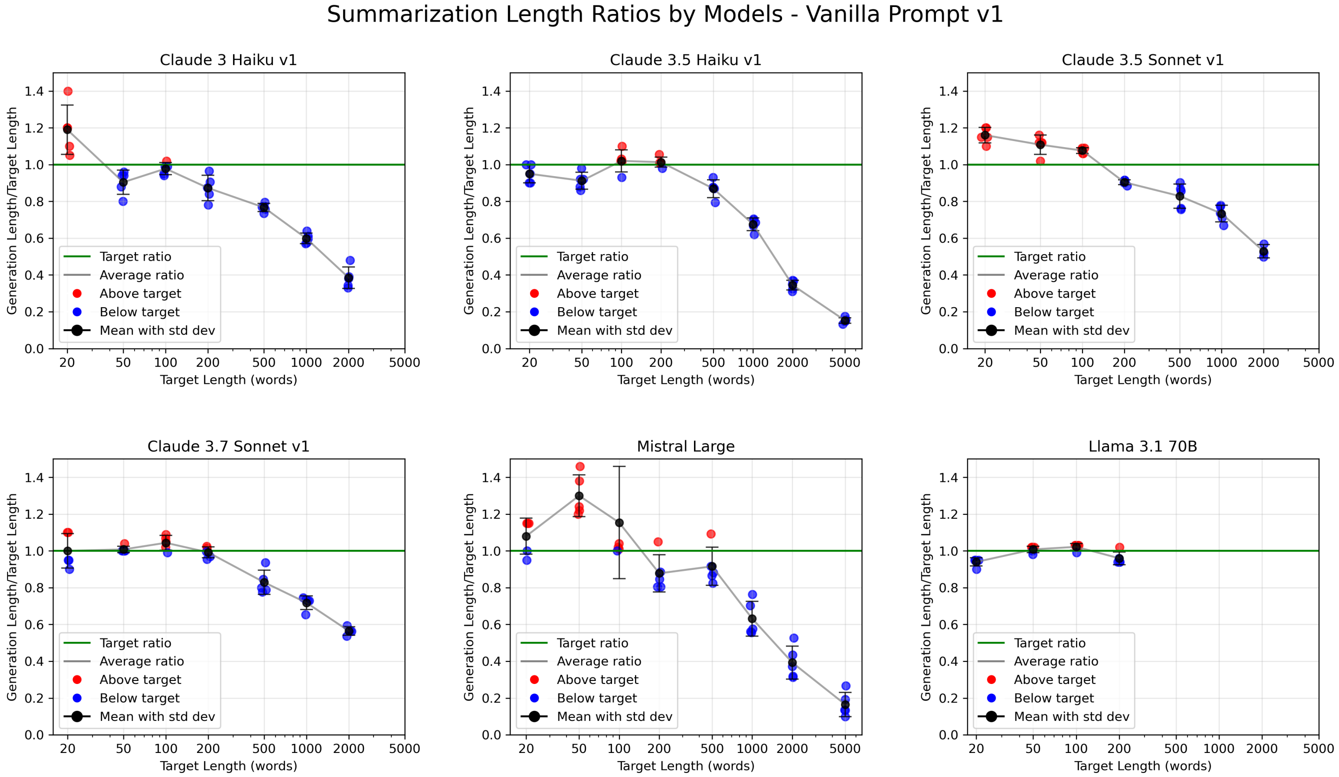

Figure 1: Length fidelity with vanilla prompting. The vertical axis shows the ratio of generated length to target length, with 1.0 representing perfect adherence. Each point represents an individual generation attempt, with color indicating over-generation (red) or under-generation (blue). Black points with error bars show the mean and standard deviation for each target length.

Table 2: Mean Absolute Percentage Deviation (MAPD ± standard deviation) across models and prompting strategies for document summarization. Lower values indicate better length control. Best method for each model is highlighted in bold.

Model	Vanilla V1 MAPD	Vanilla V2 MAPD	Thinking V1 MAPD	Thinking V2 MAPD	Best
Claude 3 Haiku v1	0.242 ± 0.200	0.130 ± 0.083	0.297 ± 0.273	**0.120 ± 0.097**	Thinking V2
Claude 3.5 Haiku v1	0.271 ± 0.301	0.330 ± 0.363	**0.225 ± 0.270**	0.316 ± 0.352	Thinking V1
Claude 3.5 Sonnet v1	0.176 ± 0.119	0.308 ± 0.291	**0.159 ± 0.172**	0.192 ± 0.115	Thinking V1
Claude 3.7 Sonnet v1	0.141 ± 0.146	0.268 ± 0.307	**0.088 ± 0.079**	0.177 ± 0.149	Thinking V1
Llama 3.1 70B	0.037 ± 0.023	**0.027 ± 0.019**	0.032 ± 0.020	0.036 ± 0.042	Vanilla V2
Mistral Large	**0.328 ± 0.279**	0.361 ± 0.274	0.349 ± 0.283	0.402 ± 0.262	Vanilla V1

prompting methods maintain or enhance the semantic quality of outputs while improving length adherence.

We assessed four key dimensions of quality across all model outputs:

- **Correctness**: Measures the factual accuracy of information presented in the summary relative to the source document, evaluating whether statements accurately reflect information from the original text.
- **Completeness**: Assesses whether the summary captures all essential information from the original document proportionate to its length target, including key points, arguments, and conclusions.
- **Faithfulness**: Evaluates whether the summary contains information that is consistent with the source document without introducing facts or claims not present in the original.
- **Relevance**: Measures how well the summary focuses on information that matters to the core message of the document, avoiding tangential details while highlighting central themes.

These metrics were evaluated on a 0-1 scale using Claude 3.5 Sonnet v2 as the judge model, with detailed rubrics provided for each quality dimension.

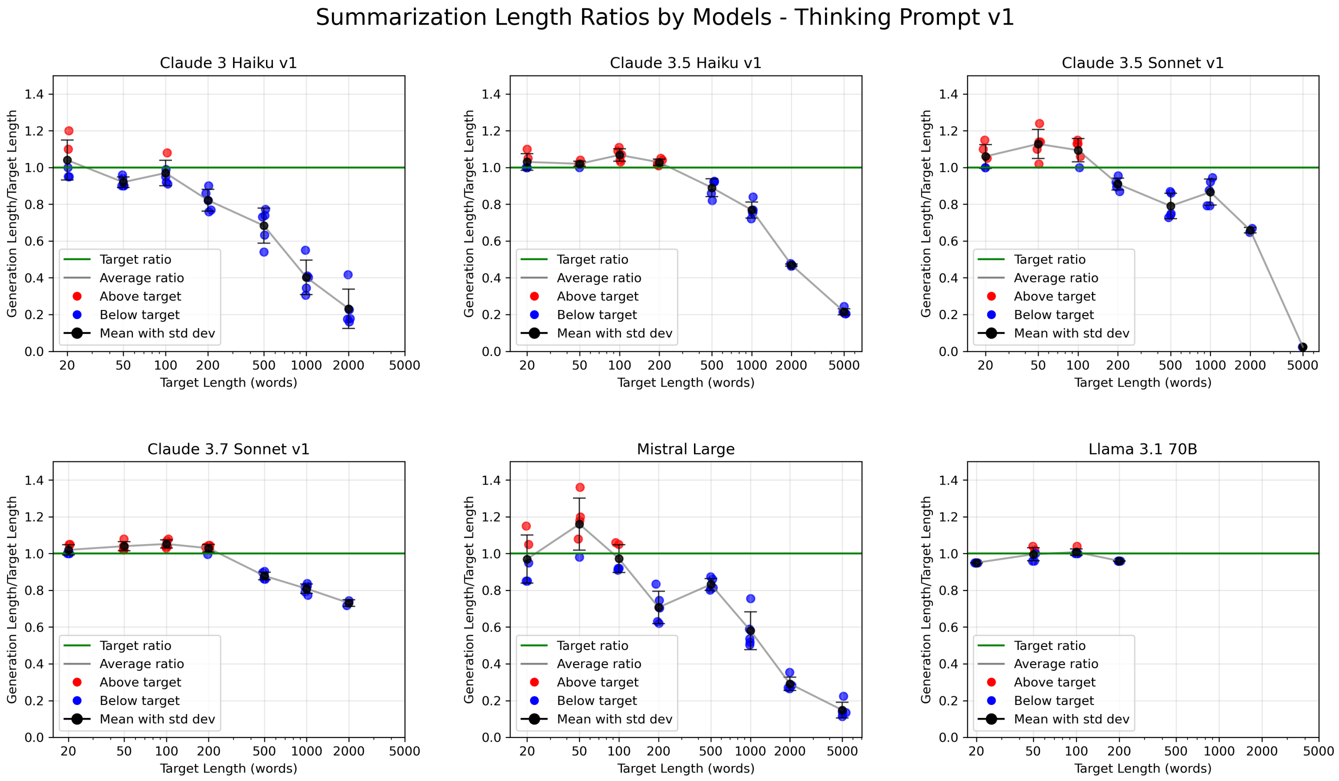

Figure 2: Length fidelity with Plan-and-Write prompting. Note the tighter clustering around the target ratio of 1.0 compared to vanilla prompting, indicating improved length control across most models.

Table 3: Quality evaluation results across different prompting strategies using LLM-as-a-Judge. Higher scores indicate better performance (0-1 scale). Best method for each quality dimension is highlighted in bold.

Prompting Strategy	Correctness	Faithfulness	Completeness	Relevance
Vanilla V1	**0.91**	0.95	0.81	0.71
Vanilla V2	**0.91**	0.93	0.73	0.69
Thinking V1	0.90	**0.96**	**0.85**	**0.87**
Thinking V2	0.87	0.94	0.84	0.73

Table 3 presents the quality evaluation results across all prompting strategies. Notably, our Plan-and-Write approach (Thinking V1) not only improved length fidelity but also enhanced summary quality in several dimensions. While Vanilla approaches scored marginally higher on correctness (0.91 vs. 0.90), Thinking V1 achieved the highest scores in faithfulness (0.96), completeness (0.85), and relevance (0.87). These results challenge the assumption that there must be a trade-off between length control and content quality, suggesting instead that structure-guided prompting can simultaneously improve length adherence and enhance output quality, particularly in dimensions related to information organization and relevance.

5.2 Open-Weight Model Evaluation

To evaluate generalizability across model families, we conducted additional experiments with the open-weight Qwen 2.5 7B model [?] deployed on AWS SageMaker, following identical experimental protocols as our primary evaluation. Our findings revealed that the Thinking V2 approach achieved the lowest overall MAPD (0.280 ± 0.636), marginally outperforming Vanilla V2 (0.281 ± 0.169), as illustrated in Figure 3. However, the high standard deviation in the Thinking V2 results indicates considerable variability in performance—suggesting that while structured prompting can improve length control in open-weight models, its effects may be less consistent than in more advanced proprietary models.

5.3 Trade-off Analysis

While structure-guided prompting improves length fidelity for several models, production deployments must consider computational trade-offs. We conducted a detailed cost-benefit analysis using Qwen 2.5 7B deployed on an AWS ml.g5.12xlarge instance

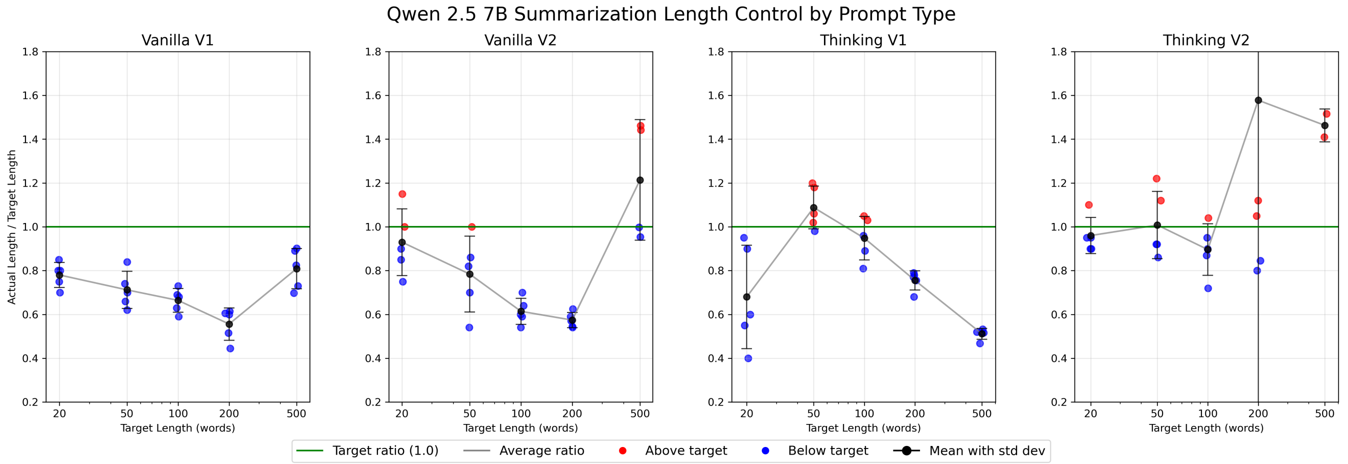

Figure 3: Length fidelity comparison across four prompting strategies for Qwen 2.5 7B.

($7.09/hour), measuring both token consumption and inference latency, as shown in Table 4.

Table 4: Cost-benefit analysis of different prompting strategies with Qwen 2.5 7B on AWS.

Metric	Vanilla Prompts	Thinking Prompts
Average token usage	7,914 tokens	8,046 tokens (1.02×)
Average inference time	1,000.1 ms	1,573.8 ms (1.57×)

Our analysis reveals that thinking prompts require only marginally more tokens (1.02×) but substantially longer inference time (1.57×) compared to vanilla prompts. For this specific model, the increased computational cost did not translate to improved length fidelity, suggesting that simpler prompting strategies may be more cost-effective in this case. However, this trade-off analysis may differ for other models where thinking prompts demonstrated significant improvements in length control. These findings highlight the importance of model-specific evaluation when deciding between prompting strategies in resource-constrained production environments.

6 Limitations

While our Plan-and-Write methodology shows promising results, several limitations warrant acknowledgment. First, the approach does not benefit all models equally—some models like Llama 3.1 70B already exhibit strong length control capabilities with standard prompting, limiting the relative improvement from our method. Second, we observed that for longer target lengths (beyond 500 words), length fidelity tends to decrease across all prompting strategies, suggesting fundamental limitations in LLMs' ability to maintain precise counting over extended outputs. This indicates our approach may be more effective for shorter-to-medium length constraints rather than very long generations. Third, the two-phase generation process introduces additional computational overhead by requiring models to generate more tokens during the planning phase, potentially increasing inference costs. Finally, while we observed no significant quality degradation in model outputs, our quality evaluation relies on LLM-as-a-Judge, which may introduce its own biases [?], and future work should systematically evaluate potential tradeoffs between length fidelity and response quality, especially at extreme target lengths.

7 Conclusion

This paper introduces Plan-and-Write, a prompt engineering methodology for precise length control in large language models without requiring model retraining. Through comprehensive evaluation across seven state-of-the-art LLMs, we demonstrated that explicit word counting and structured planning can significantly improve length fidelity compared to standard prompting techniques. Our results show that five of seven tested models benefit from our approach, with improvements in Mean Absolute Percentage Deviation of up to 37.6%. Our quality evaluation using LLM-as-a-Judge demonstrates that the methodology not only improves length control but also maintains or enhances overall output quality. Our findings reveal that many modern LLMs possess untapped capabilities for length control that can be activated through carefully designed prompting strategies.

Our work offers immediate practical value for developers and researchers working with black-box LLMs in production environments where model retraining is impractical or cost-prohibitive. The Plan-and-Write methodology provides a straightforward, deployable solution for applications requiring precise length control, from voice interfaces to mobile applications with display constraints. The approach's effectiveness across complex prompting patterns demonstrates its generalizability beyond simple instructions. Future research directions include exploring hybrid approaches that combine prompt engineering with lightweight inference-time modifications, investigating the generalizability of structure-guided prompting to other types of constraints beyond length, and developing adaptive prompting strategies that adjust based on model capabilities and specific length targets.

References

[] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint* arXiv:2212.08073 (2022).

[] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *COLM* (2024).

[] Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, Kun Zhu, Lei Huang, Tat-Seng Chua, and Bing Qin. 2024. Length Controlled Generation for Black-box LLMs. *arXiv preprint* arXiv:2412.14656 (2024).

[] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to Write with Cooperative Discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1638–1649.

[] Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865* (2022).

[] Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost. *CoRR* (2024).

[] Jongwoo Park, Chen Chen, Serena Singh, Ziang Jiang, William Wang, Ruoxi Chen, Zhangyang Wu, and Xiaodong Yi. 2024. Regularized Direct Preference Optimization: Addressing Length Bias and Outperformance in RLHF. *arXiv preprint* arXiv:2402.07319 (2024).

[] Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[] Qifan Wang, Li Yang, Xiaojun Quan, Fuli Feng, Dongfang Liu, Zenglin Xu, Sinong Wang, and Hao Ma. 2022. Learning to generate question by asking question: A primal-dual approach with uncommon word generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 46–61.

[] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. [n. d.]. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *Neurips Safe Generative AI Workshop 2024*.

[] Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2024. Following length constraints in instructions. *arXiv preprint* arXiv:2406.17744 (2024).

[] Chi Zhang, Daochen Zha, Zhuosheng Zeng, Shanghang Wang, and Chenguang Wang. 2023. REST: Prompting LLMs to Understand and Generate Tables and Charts. *arXiv preprint* arXiv:2310.16445 (2023).

[] Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. 2024. T-REG: Preference Optimization with Token-Level Reward Regularization. *arXiv preprint* arXiv:2412.02685 (2024).

A Length Fidelity Plots for Summarization

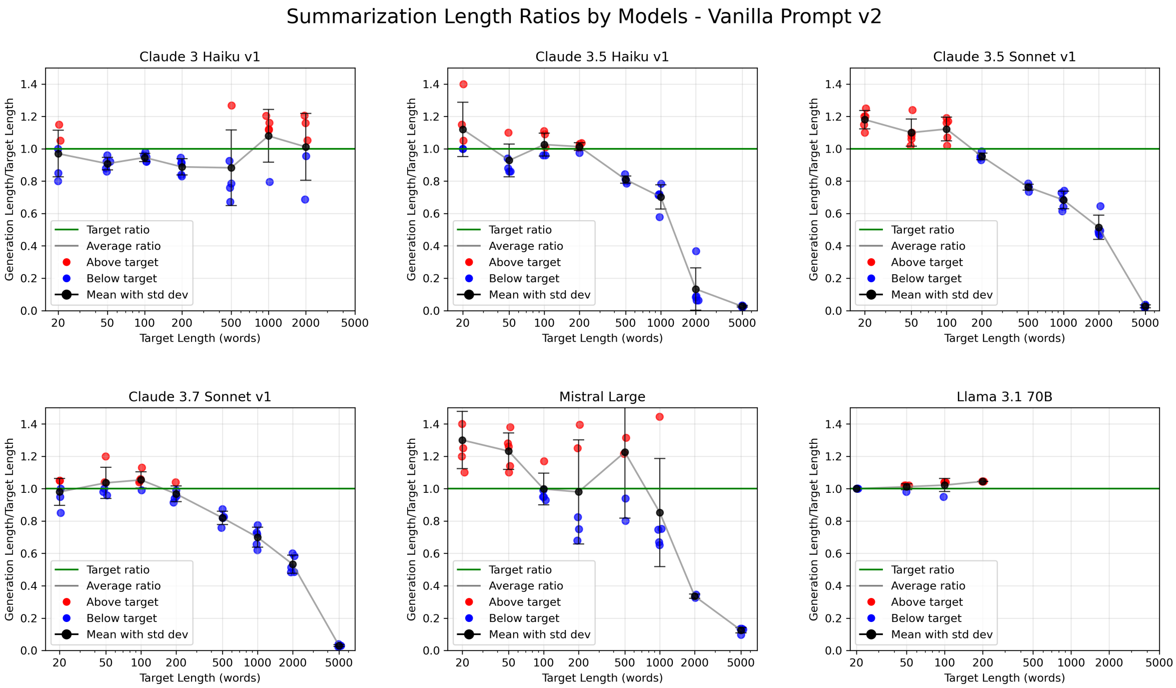

Figure 4: Length fidelity with vanilla prompting v2.

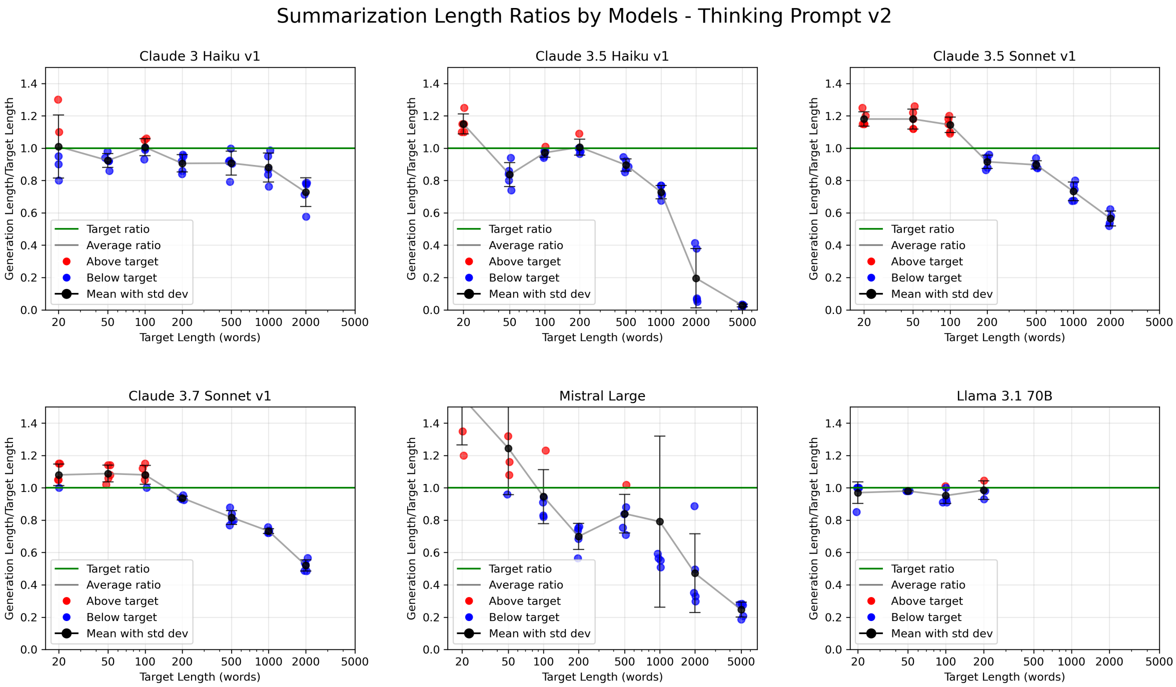

Figure 5: Length fidelity with Plan-and-Write prompting v2.

B Story Generation Length Fidelity test

In addition to document summarization, we conducted experiments on creative story generation tasks to evaluate whether our structure-guided Plan-and-Write approach would yield similar benefits for open-ended content creation. We prompted models to write stories about a young boy discovering a magical book, providing specific target word counts. Unlike document summarization, where models must extract and condense existing information, story generation requires models to create original content, potentially making length control more challenging as the model must simultaneously manage creativity and constraints. We used two distinct prompting strategies for the story generation task:

Vanilla Prompt for Story Generation

Write a story about a young boy who discovers a magical book in his attic and learns how to harness the power of magic within himself in exactly {target_words} words.

Thinking Prompt for Story Generation

TASK: Write a story about a young boy who discovers a magical book in his attic and learns how to harness the power of magic within himself using EXACTLY {target_words} words.
SCIENTIFIC METHODOLOGY: 1. First, outline the key story elements and narrative arc (10% of effort) 2. Then perform controlled story development to EXACTLY {target_words} words by: a) Establishing setting, characters, and conflict b) Developing plot points proportionally c) Maintaining narrative coherence and flow d) Including appropriate details to reach target length
CONSTRAINTS: - Output MUST contain EXACTLY {target_words} words - Story should be engaging and complete - Narrative complexity should scale with target length
EXECUTION:
<thinking>
• First sketch the narrative arc with key plot points
• Outline main character development arcs
• Draft initial version (likely shorter or longer than target)
• Systematically adjust by adding/removing:
- Descriptive details
- Character moments
- Plot developments
• Count words meticulously: 1, 2, 3... until reaching {target_words}
</thinking>
Final {target_words}-word story:

B.1 Results

Our results for story generation tasks reveal a notable contrast with document summarization findings. Table 5 and Figures 6 and 7 demonstrate that Plan-and-Write did not provide consistent benefits for creative generation tasks—in fact, for five of six models, vanilla prompting achieved better length fidelity. This suggests that structure-guided prompting may be more effective for information condensation than for creative content generation.

Table 5: Mean Absolute Percentage Deviation (MAPD ± standard deviation) across models and prompting strategies for story generation. Lower values indicate better length control. Best method for each model is highlighted in bold.

Model	Vanilla MAPD	Thinking MAPD	Best
Claude 3 Haiku v1	**0.080 ± 0.093**	0.113 ± 0.082	Vanilla
Claude 3.5 Haiku v1	**0.204 ± 0.220**	0.844 ± 2.015	Vanilla
Claude 3.5 Sonnet v1	**0.045 ± 0.031**	0.084 ± 0.054	Vanilla
Claude 3.7 Sonnet v1	**0.047 ± 0.030**	0.061 ± 0.042	Vanilla
Llama 3.1 70B	0.133 ± 0.184	**0.117 ± 0.191**	Thinking
Mistral Large	**0.266 ± 0.229**	0.322 ± 0.286	Vanilla

These contrasting results between summarization and story generation suggest that the benefits of structure-guided prompting may be task-dependent. Specifically, explicit planning and word counting appear more beneficial for tasks that require condensing existing

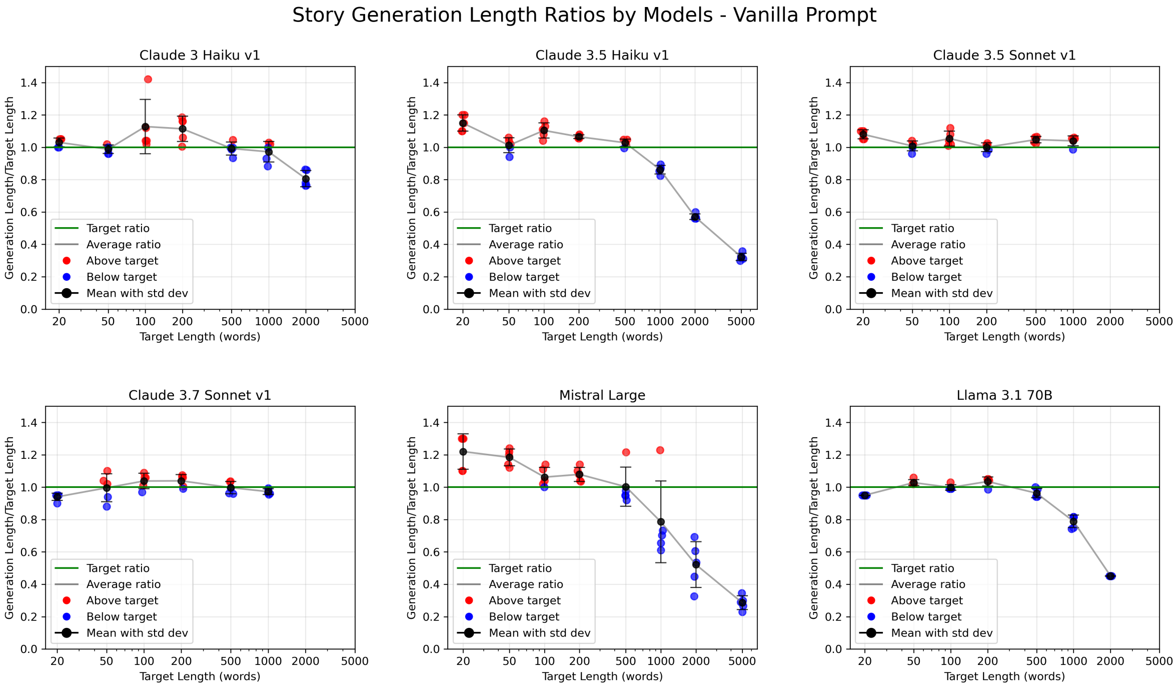

Figure 6: Length fidelity with vanilla prompting for story generation.

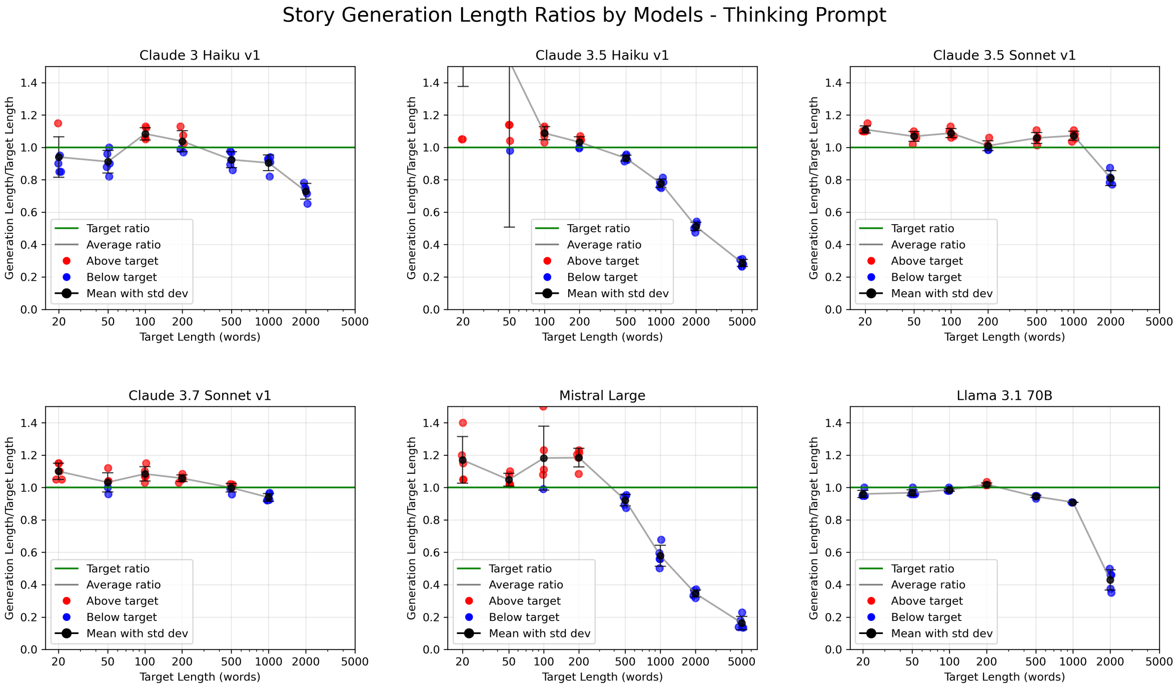

Figure 7: Length fidelity with Plan-and-Write prompting for story generation.

information (like summarization) than for tasks requiring creative content generation. This finding highlights the importance of task-specific prompt engineering and suggests that different cognitive processes may be involved in length control for different types of generation tasks.