WritingBench: A Comprehensive Benchmark for Generative Writing

Anonymous ACL submission

Abstract

The rapid proliferation of large language models (LLMs) highlights an urgent need for evaluation frameworks that cover a wide range of writing tasks but also deliver reliable and nuanced evaluation results. However, current benchmarks are limited in scope, lacking both comprehensive coverage of specialized writing tasks and the granularity required for precise requirements. Moreover, existing static evaluation methods fall short in capturing stylistic and contextual fidelity, particularly when applied to diverse and complex writing tasks. To tackle these challenges, we present **WritingBench**, a comprehensive benchmark comprising 1,239 queries spanning 6 domains and 100 subdomains with diverse material contexts, designed to evaluate multi-dimensional requirements such as style, format, and length. We further propose a *query-dependent evaluation* framework enabling LLMs to dynamically generate task-specific assessment criteria. This framework is complemented by a fine-tuned critic model for criteria-aware scoring, ensuring fine-grained evaluations across a wide range of writing tasks. Leveraging the precise feedback from this evaluation process, we further filter synthesized data to train a writing-enhanced model, which demonstrates superior performance, achieving a 18% improvement in human evaluation over baseline models.

1 Introduction

In recent years, large language models (LLMs) have garnered significant attention due to their expanding capabilities, enabling applications across a diverse range of real-world writing tasks (DeepSeek-AI et al., 2025; Yang et al., 2024a; Anthropic, 2024; Reid et al., 2024; Dubey et al., 2024). These tasks include generating creative content (Mirowski et al., 2023; Marco et al., 2024; Karpinska et al., 2024; Yang et al., 2024b; Wang et al., 2024), enhancing professional workflows (Shao et al., 2024; Li et al., 2024), and etc. As

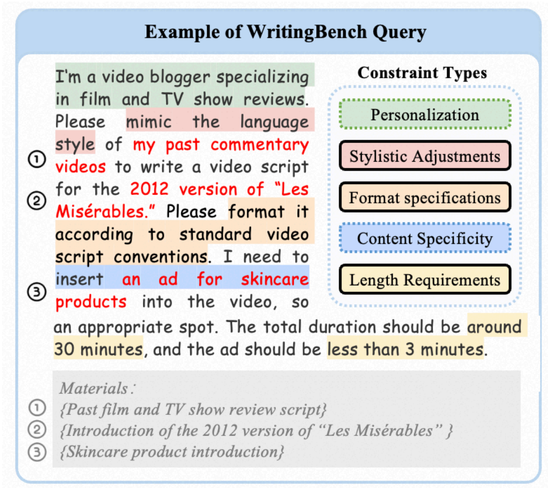

Figure 1: Example of Writing Query

LLMs play an increasingly prominent role in these domains, establishing comprehensive and reliable benchmarks is crucial for evaluating their current performance and guiding future improvements in writing proficiency.

Existing evaluation benchmarks exhibit two significant limitations. First, there is a notable scarcity of specialized benchmarks for various writing tasks. Most existing writing-oriented benchmarks are restricted to single domains, like fictions (Karpinska et al., 2024; Marco et al., 2024; Mirowski et al., 2023; Yang et al., 2024b). Their task formulations are typically simplistic, often limited to single-sentence queries (Bai et al., 2024; Karpinska et al., 2024) or constrained by a small set of instruction templates (Paech, 2023; Que et al., 2024). Furthermore, most test instances are based on homogeneous input materials (Que et al., 2024; Karpinska et al., 2024), which diminishes their ability to accommodate the complex and customized requirements inherent in real-world writing scenarios. Consequently, these benchmarks do not capture the diversity and intricacies of practical writing tasks. Second, current automatic evaluation metrics lack

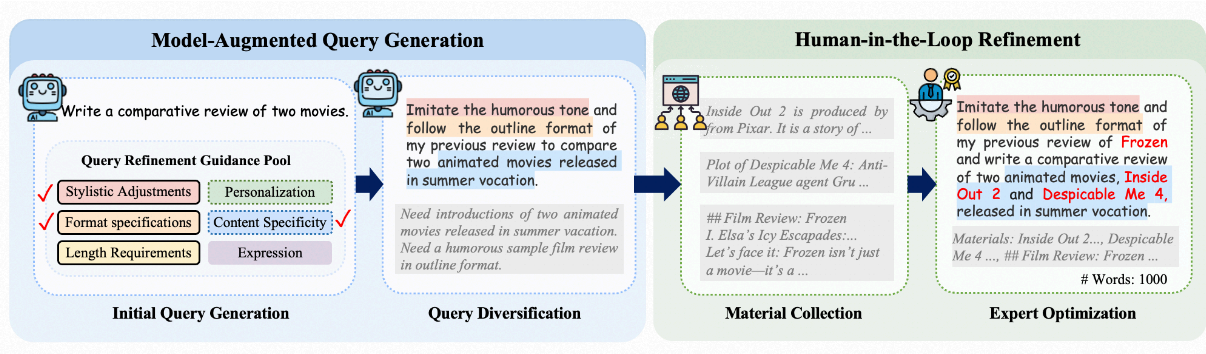

Figure 2: Construction pipeline of WritingBench.

the robustness needed for comprehensive assessment. Although LLM-based evaluation methods demonstrate strong capabilities in capturing semantic meanings (Shao et al., 2024; Que et al., 2024; Bai et al., 2024), they generally rely on a limited number of predefined dimensions (e.g., fluency and coherence). As LLMs continue to advance and exhibit increasingly sophisticated writing abilities, these static evaluation criteria are inadequate to measure diverse requirements and specifications of complex writing tasks.

To address these challenges, we propose **WritingBench**, a comprehensive benchmark and reliable framework for general-purpose writing evaluation. Our approach begins with the deliberate establishment of a secondary domain categorization, grounded in real-world writing requirements. We propose a four-stage query generation pipeline as illustrated in Figure 2. LLMs first generate various queries, which are followed by human material collection and refinement. This process results in a set of writing query that is characterized by broad domain coverage, varied requirements, and the integration of materials from diverse sources. To facilitate a more nuanced evaluation of generated responses across different domains, we design a *query-dependent evaluation* framework that dynamically generates five query-specific criteria using LLMs, which are then scored by a fine-tuned critic model. Finally, we integrate the aforementioned methods to synthesize and filter writing-specific data, which then is used to train a small-scale, writing-enhanced model.

Our primary contributions are as follows:

- We present **WritingBench**, an open-source writing benchmark comprising *1,239* queries across *6* primary domains and *100* subdomains, encompassing task requirements along the dimensions of *style*, *format*, and *length*. WritingBench facilitates extended-context generation, accommodating input lengths ranging from tens to thousands of words, thereby addressing the diverse input requirements in real-world scenarios.
- We propose a *query-dependent evaluation* framework that integrates automatic criteria generation with a criteria-aware scoring model. Our approach achieves a strong correlation with human judgments (87%).
- We fine-tune a *7B-parameter writing-enhanced model* using synthesized and filtered data, demonstrating performance comparable to the chatgpt-4o-latest model. The WritingBench is publicly released along with the query-dependent criteria, the scoring model, and the writing model, at: https://anonymous.4open.science/r/ACL-2025-2DD2 to enable and encourage further research in the field.

2 Related Work

2.1 Writing Benchmarks

Existing evaluation benchmarks suffer from significant limitations in domain coverage and task granularity. For instance, HelloBench encompasses 5 domains using templated queries (Que et al., 2024), while LongWriter incorporates length constraints in 120 queries (Bai et al., 2024); however, they both lack hierarchical domain taxonomies and multi-dimensional requirement specifications (e.g., style and format). Furthermore, most current benchmarks rely on fixed instruction templates and short contexts (Paech, 2023), rendering them insufficient for addressing the complexity of real-world data needs. In contrast, our proposed benchmark fills these gaps by introducing 1,237 free-form queries distributed across 100 subdomains, with explicit controls over style, format, and length, paired with inputs ranging from tens to thousands of words.

Benchmark	Num	Domains		Requirement			Input Token		Free Query Template	Free Material Source
		Main	Sub	Style	Format	Length	Avg	Max		
EQ-Bench	241	1	/	✗	✗	✗	130	213	✗	/
LongBench-Write	120	7	/	✗	✗	✓	87	684	✓	/
HelloBench	647	5	38	✗	✗	✓	1,210	7,766	✗	✗
WritingBench (Ours)	1,239	6	100	✓	✓	✓	1,546	19,361	✓	✓

Table 1: Comparison of existing benchmarks.

2.2 Evaluation Methods

Using LLMs as judges has become a prevalent approach for evaluating the quality of generated responses. Typically, researchers pre-define a fixed set of evaluation dimensions applicable across all test instances. For example, SuperCLUE (Xu et al., 2023) employs three dimensions (e.g., creativity and coherence), while LongWriter (Bai et al., 2024) adopts six dimensions (e.g., relevance and accuracy). HelloBench (Que et al., 2024) introduces task-specific dimensions, but the dimensions remain consistent across all queries of a given task. Although the LLM-as-a-judge approach enhances scalability, static evaluation dimensions often fail to accommodate the diversity of writing styles and specifications. To address this limitation, recent work (Liang et al., 2024) proposes training a model to dynamically generate evaluation dimensions for individual queries. However, the total number of dimensions in such methods remains confined to a small predefined set. In contrast, our query-dependent evaluation framework leverages LLMs to generate diverse and query-specific criteria for different queries while fine-tuning a dedicated critic model to perform the evaluation.

2.3 Writing Models

Although existing LLMs demonstrate exceptional writing capabilities, researchers continue to strive for improvements in their overall writing proficiency. Recent models, such as Weaver (Wang et al., 2024) and LongWriter (Bai et al., 2024), have exhibited notable domain-specific strengths. For instance, Weaver benefits from over 200B parameter pretraining, supporting four distinct writing domains, while Suri specializes in generating technical content (Pham et al., 2024). However, these models experience substantial performance degradation when addressing cross-domain scenarios and multi-constraint tasks. In this work, we introduce a comprehensive writing-enhanced model that achieves competitive performance compared to chatgPT-4o-latest across various tasks.

3 WritingBench

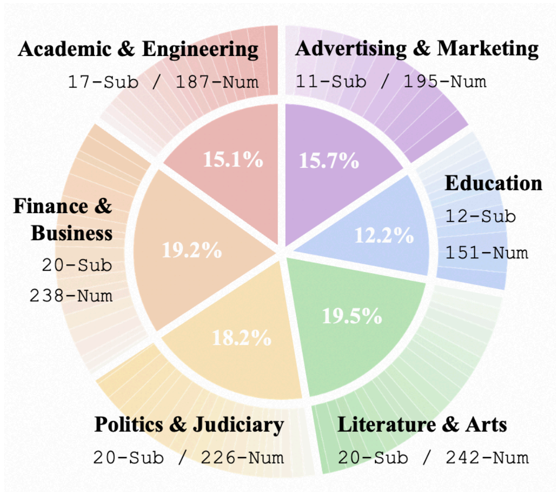

Figure 3: Domain categories in WritingBench.

In this section, we will mainly introduce the construction process of our WritingBench and the query-dependent evaluation framework. Furthermore, we train a critic model for criteria-aware evaluation and a writing-enhanced model to achieve superior writing performance.

3.1 Benchmark Construction

To construct WritingBench, we design a systematic pipeline combining model-generated data refinement and human annotation, ensuring both diversity and real-world alignment of the benchmark. The construction process consists of two phases: model-augmented query generation and human-in-the-loop refinement, as illustrated below.

3.1.1 Model-Augmented Query Generation

This phase focuses on leveraging the capabilities of LLMs to generate an initial set of writing queries and supported materials, which are enriched and diversified through systematic guidance.

Phase 1: Initial Query Generation

We begin by constructing a two-tiered domain pool grounded in real-world writing scenarios, consisting of 6 primary-level domains and 100 secondary-

Domain	CNT	Ave Token	Max Token
Academic&Engineering	187	1,915	15,534
Finance&Business	238	1,762	19,361
Politics&Judiciary	226	2,274	18,317
Literature&Arts	242	1,133	9,973
Education	151	1,173	10,737
Advertising&Marketing	195	886	6,504
Requirement			
Style	400	1,404	18,197
Format	342	1,591	18,197
Length	214	1,226	14,097
Length			
<1K	727	443	994
1K-3K	341	1,808	2,991
3K-5K	94	3,804	4,966
5K+	77	8,042	19,361

Table 2: Data statistics for WritingBench categorized by domain, requirement, and length.

level subdomains. The selected domains are designed to capture both traditional and emerging user needs for AI-assisted writing, encompassing categories such as academic & engineering, finance & business, politics & judiciary, literature & art, education, publicity & marketing. Leveraging the domain and subdomain tags, we prompt ChatGPT and Claude to generate initial writing queries that simulate realistic user requests.

Phase 2: Query Diversification

To improve the diversity and practical applicability of queries, we propose a set of query diversification strategies inspired by Xu et al. (2024), which include:

• Length constraints (e.g., “Generate a 500-word executive summary")

• Format specifications (e.g., “Follow the IEEE conference template")

• Stylistic adjustments (e.g., “Write in a formal tone for a corporate audience")

• Personalization (e.g., “Incorporate the user’s internship experience")

• Content specificity (e.g., “Detail the 2023 Q3 financial metrics")

• Conciseness requirements (e.g., “Summarize in one sentence")

Once the queries are refined, these diversified prompts are used to elicit material requirements from LLMs (e.g., requesting financial reports as input for market analysis queries). This approach results in enriched queries accompanied by corresponding recommended reference materials.

3.1.2 Human-in-the-Loop Refinement

This phase incorporates human expertise to verify model-generated queries and supplement model-generated requirements, thereby ensuring their alignment with real-world applications.

Phase 1: Material Collection

At this stage, we engage over 20 paid annotators with specialized expertise, who have undergone rigorous training tailored to the annotation tasks. Their primary responsibility is to collect open-source documents in response to queries that require supplementary external resources (e.g., public financial statements or legal templates), guided by material requirements generated by LLMs. To minimize errors arising from parsing documents in diverse formats, the annotators carefully extract and verify the most pertinent text segments.

Phase 2: Expert Screening & Optimization

Subsequently, we invite five experts to perform data screening. All experts have experience with the use of LLMs or are professionals in the related industry. The experts performed dual filtering: (1) query adaptation: rewrite ambiguous or unrealistic queries to better align with materials and practical scenarios (e.g., adjusting a legal opinion query to reference specific clauses from provided statutes). (2) material pruning: removed redundant or irrelevant content from collected materials, ensuring focused context for writing tasks.

We subsequently engage five domain experts to perform data screening, all of whom possess substantial experience with the use of LLMs or are professionals in relevant industries. The experts conducted a two-stage filtering process:(1) query adaptation: ambiguous or unrealistic queries are revised to better align with the provided materials and practical scenarios (e.g., adjusting a legal opinion query to reference specific clauses from the supplied statutes). (2) material pruning: redundant or irrelevant content is eliminated from the collected materials, ensuring that the context provided for writing tasks remained focused and relevant.

Finally, we construct WritingBench, a benchmark comprising 1,239 queries categorized using a two-tiered taxonomy, as depicted in Figure 3. In comparison to existing writing benchmarks summarized in Table 1, WritingBench exhibits notable advantages in terms of the number of instances, domain diversity, requirement coverage, and variability in input lengths. The detailed statistical distribution of WritingBench is shown in Table 2.

3.2 Evaluation Metric

Traditional LLM-as-a-judge evaluations typically rely on fixed evaluation criteria derived from general writing assessment conventions (Bai et al., 2024). However, such static criteria exhibit three critical limitations: (1) domain exhaustiveness: fixed criteria fail to adapt effectively to specialized domains, such as technical documentation or creative writing; (2) requirement specificity: fixed criteria lack the flexibility to capture specific requirements related to style, format, or length control; and (3) material dependency: fixed criteria are insufficient to verify whether responses appropriately utilize the provided reference materials.

To address these challenges, we propose a query-dependent criteria evaluation framework that enables dynamic adaptation to diverse writing scenarios. As illustrated in Figure 4, our approach comprises two phases:

Phase 1: Dynamic Criteria Generation
Given a query q in the WritingBench, the LLM is prompted to automatically generate a set of five evaluation criteria, $C_q = \{c_1, \ldots, c_5\}$, using a carefully designed instruction to ensure structural guidance during criteria specification (see Appendix C.4). Each criterion comprises three components: a concise name summarizing the criterion, an extended description elaborating on the evaluation focus, and detailed scoring rubrics, which provide fine-grained quality levels for the respective evaluation dimensions.

Phase 2: Rubric-based Scoring
For each criterion $c_i \in C_q$, the LLMs are instructed to independently assign a score on a 10-point scale to a given response r. During the scoring process, the model must provide both the numerical score and a detailed justification for its evaluation. The final overall score is computed by averaging the scores across all dimensions. Detailed prompts used are provided in Appendix C.4.

3.3 Critic Model

To alleviate the computational overhead with LLM-based evaluation, we develop a dedicated critic model, $\mathcal{M}$, designed to implement our rubric-driven scoring framework. Specifically, this model performs the mapping $\mathcal{M}_c : (q, r, C_i) \mapsto [1, 10] \times \mathcal{J}$, where the output consists of a numerical score and corresponding justification text, $\mathcal{J}$, both in accordance with the predefined evaluation rubric.

We fine-tune the critic model on a dataset comprising 50K instances, which are collected using LLMs in our experiments. The dataset encompasses diverse queries, evaluation criteria, and model responses to enhance the robustness of evaluation. The Training details are provided in Appendix B.3, and the experiments presented in Section 4.3 validate the consistency of the critic model.

3.4 Writing Model

To develop a writing-enhanced model, we integrate the two aforementioned methods for synthesizing and filtering training data. Specifically, we follow the initial three steps outlined in Section 3.1.1, leveraging LLMs to generate writing queries and produce extended supplemental materials, replacing the need for human annotators. This process yields a total of 24K training examples. Subsequently, we apply the query-dependent evaluation metric, utilizing our critic model described in Section 3.3, to filter and select a subset of 12K high-quality training samples. Fine-tuning experiments are conducted using the llama-3.1-8b-instruct and qwen-2.5-7b-instruct models. Both models demonstrate significant performance improvements over their previous versions and, in our experiments, even outperform larger models such as llama-3.3-70b-instruct and qwen-2.5-72b-instruct.

4 Experiment

4.1 Experiment Settings

In this section, we describe the comprehensive settings employed in our experiments to evaluate the effectiveness of the models using the WritingBench framework. Our approach is designed to ensure accuracy and consistency in performance assessment, leveraging both advanced AI-assisted scoring mechanisms and human evaluation for robust verification. We detail the dataset configuration, the evaluation protocol incorporating cutting-edge methodologies, and the training model configurations, which together comprise a rigorous experimental setup. These components are meticulously outlined to facilitate reproducibility and provide transparency, enabling other researchers to replicate and build upon our findings. Details of dataset configurationcan, evaluation protocol ,and training model configurations be found in Appendex.

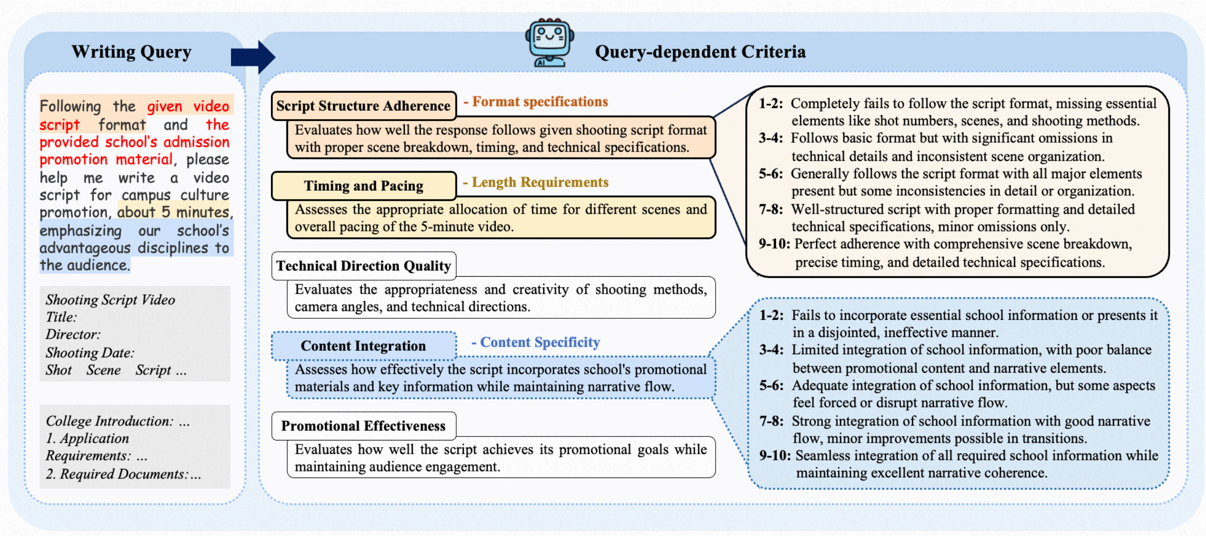

Figure 4: Example of dynamically generating criteria for a writing query.

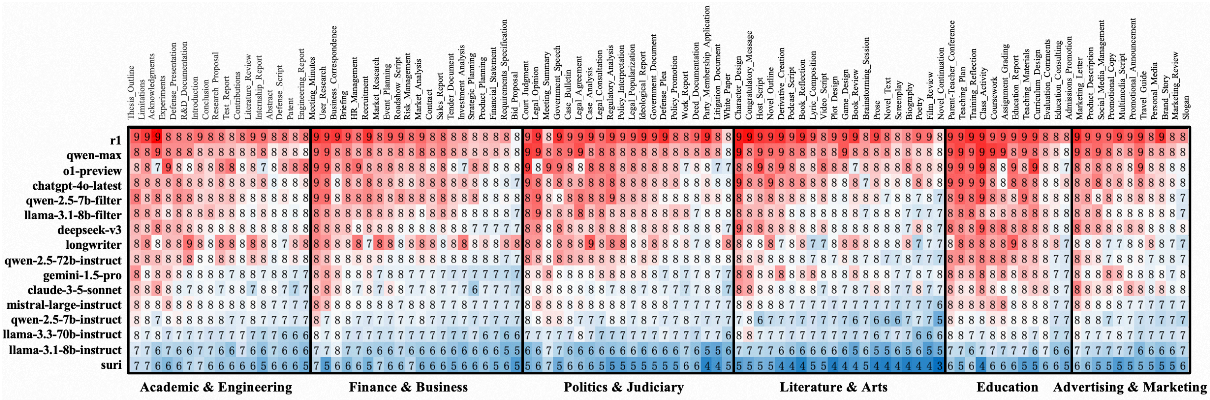

Figure 5: Scores of sub domains.

4.2 Comparison between LLMs

We evaluate 16 LLMs on WritingBench with 1,239 queries covering 6 domains and 3 core requirements. Each query is assessed by 5 criteria (10-point scale), with domain-specific subcategory heatmaps revealing task-level variations.

Key Insights from Domain Scores: Finance (D2) and Politics & Judiciary (D4) are areas where most models, such as Qwen-Max and Deepseek-R1, showed consistent high performance. Literature & Art (D5) had slightly more variance, with models like Deepseek-R1 outperforming others, indicating better handling of narrative and creative content. Difficulties were noted in niche and detailed content areas such as tender proposals and white papers, where models generally scored lower, highlighting potential areas for further enhancement to handle detailed and specialized documents better.

Key Insights from Requirement Scores: Across Format (R1), models like Qwen-Max excelled, indicating their robustness in structuring and presenting information accurately. The Style (R3) dimension revealed distinctions among models, where language nuances play a significant role in scoring, with Deepseek-R1 and Qwen-Max often leading due to their ability to adapt language style effectively. Models, such as Suri, scored lower across all dimensions, indicating potential enhancements needed in core capability for consistent performance across different requirements.

The overall analysis of the WritingBench main experiment highlights:

Deepseek-R1 consistently leads across both domain and requirement dimensions, showcasing its versatility and strong language model capabilities. The general weakness across models in specialized formats like tender proposals and white papers suggests an opportunity for development focus in creating specialized datasets to improve model training in these areas. There's a noticeable vari-

Model	Total	Language		Domain						Requirement					
		ZH	EN	D1	D2	D3	D4	D5	D6	R1	C	R2	C	R3	C
Proprietary LLM															
ChatGPT-4o-latest	8.2	8.3	8.1	8.1	8.1	8.2	8.1	8.4	8.1	8.2	8.9	8.2	8.3	8.3	8.7
o1-Preview	8.2	8.1	8.2	8.0	8.1	8.2	8.2	8.4	8.1	8.2	8.8	8.2	8.2	8.2	8.6
Claude-3-5-Sonnet	7.7	7.7	7.7	7.6	7.5	7.6	7.7	7.9	8.0	7.7	8.5	7.9	8.0	7.9	8.5
Gemini-1.5-Pro	7.8	7.8	7.7	7.7	7.5	7.8	7.9	8.0	7.9	7.9	8.8	7.9	8.0	7.9	8.6
Qwen-Max	8.4	8.4	8.3	8.3	8.3	8.4	8.4	8.5	8.4	8.4	9.0	8.4	8.5	8.5	8.7
Open LLM															
Deepseek-R1	8.6	8.7	8.5	8.5	8.5	8.6	8.6	8.7	8.6	8.6	9.0	8.6	8.7	8.7	8.9
Deepseek-V3	8.0	8.0	7.9	7.9	7.8	8.0	7.8	8.2	8.0	8.0	8.9	8.0	8.2	8.1	8.6
Mistral-Large-Instruct	7.6	7.6	7.7	7.7	7.6	7.8	7.3	7.9	7.6	7.7	8.7	7.7	7.9	7.7	8.2
Qwen-2.5-72B-Instruct	7.9	8.0	7.9	8.0	7.8	8.1	7.7	8.2	7.8	8.0	8.8	7.9	8.0	8.0	8.3
Qwen-2.5-7B-Instruct	7.4	7.3	7.5	7.7	7.4	7.6	6.9	7.8	7.3	7.6	8.6	7.4	7.5	7.5	7.9
Llama-3.3-70B-Instruct	7.0	6.7	7.3	7.0	6.9	7.0	6.8	7.3	7.3	7.1	8.2	7.0	7.2	7.1	7.8
Llama-3.1-8B-Instruct	6.4	5.7	6.9	6.6	6.4	6.1	6.0	6.7	6.6	6.4	7.6	6.3	6.4	6.4	7.0
Capability-Enhanced LLM															
Suri	5.0	4.4	5.5	5.6	5.3	5.0	4.1	5.0	5.1	5.0	5.4	4.5	4.0	4.8	5.2
Longwriter	7.9	7.9	7.9	8.0	8.1	8.1	7.7	8.1	7.6	8.1	8.8	7.7	7.7	7.9	8.2
Qwen-2.5-7B-SFT-Filter	8.0	8.2	7.9	8.0	7.9	8.1	7.8	8.3	7.9	8.1	8.9	7.9	8.1	8.0	8.5
Llama-3.1-8B-SFT-Filter	8.0	8.0	8.0	8.0	8.0	8.1	7.7	8.2	7.9	8.1	8.8	7.9	8.1	8.0	8.5

Table 3: WritingBenchmark Evaluation of LLM Performance Across 6 Domains and 3 Writing Requirements using Critic Model(Scale 0-10). Domains: (D1) Academic & Engineering, (D2) Finance & Business, (D3) Politics & Judiciary, (D4) Literature & Art, (D5) Education, (D6) Publicity & Marketing. Requirements: (R1) Format, (R2) Length, (R3) Style (C indicates category-specific scores)

ance among models in creative and style-intensive domains, where models like Claude-3-5-Sonnet sometimes falter compared to technical or factual domains, pointing towards a need for more nuanced language processing enhancements. This analysis not only benchmarks existing model capabilities and highlights the leading model performers but also underscores specific areas needing improvement for holistic future model development.

4.3 Human Consistency

Evaluation Metric	**Judge**	**Score**
Static Global	GPT-4o	69%
Static Domain-Specific	GPT-4o	40%
Dynamic Query-Dependent	GPT-4o	79%
Static Global	Claude	65%
Static Domain-Specific	Claude	59%
Dynamic Query-Dependent	Claude	**87%**
Dynamic Query-Dependent	Critic Model	83%

Table 4: Comparison of human agreement scores across different criteria generation methods.

To validate the alignment between automated evaluation and human judgment, we conducted human evaluation on 300 queries, covering all 100 subdomains. Five professionally trained annotators with linguistic backgrounds perform pairwise comparisons of model responses. For each query, two responses are randomly selected from different models. were evaluated based on requirement of the query and material utilization. Annotators selected the preferred response or declared equivalence based on the query's requirements, yielding 1,500 total judgments (5 annotators × 300 queries). The experiment compared two baselines: static globally uniform criteria with LLM scoring, static domain-specific customized criteria with LLM scoring. Static criteria are designed by domain experts.)

As shown in Table 4, our dynamic query-dependent criteria achieve superior human alignment compared to static, both globally uniform criteria or domain-specific customized criteria. We observe that human disagreement often occurs on queries requiring multi-dimension balancing, precisely where dynamic criteria show strongest gains (21% over static). Notably, domain-specific criteria underperform despite customization, suggesting our queries' diversity exceeds tradional category boundaries. These findings confirm that context-sensitive query-denpedent evaluation better captures real-world writing complexity compared to conventional static approaches. Furthermore, the critic model attains 83% agreement, confirming its practical viability.

4.4 Ablation of Writing Model

In this subsection, we present an in-depth ablation analysis of our WritingBench model to assess the efficacy of different data selection across model architectures and benchmarks.

In the evaluation results of WritingBench, models trained on the curated 12K subset outperforms full 24K data both on qwen-7b-instruct and Llama-8b-instruct. This suggests that quality-driven curation outweighs quantity(, particularly crucial for specialized writing tasks). Our critic-guided filtering demonstrates remarkable effectiveness. This approach not only significantly outperforms the baseline models but also exceeds the capabilities of larger models such as llama-3.3-70b-instruct and qwen2.5-72b-instruct (see Main Table 3). Furthermore, we evaluate on another writing benchmark, LongBench-Write. Our filtered models maintain performance advantages, demonstrating generalizability beyond the training domain. Detailed cross-dataset analysis can be found in Appendix.

These outcomes underscore the effectiveness of our data construction and selection pipelines across model architectures and benchmarks. This critic-guided filtering validate the robustness of our query-dependent evaluation strategy and the utility of our critic model.

4.5 Ablation of Writing Model

Further validation of our approach was conducted on an alternative writing benchmark, LongBench-Write. Our writing model achieved results consistent with previous observations, surpassing baseline performances (detailed in the Appendix). These outcomes underscore the effectiveness of our data construction and selection pipelines, validating the robustness of our query-dependent evaluation strategy and the utility of our critic model.

This comprehensive analysis confirms that a systematic approach to data filtering can substantially enhance model performance, enabling smaller models to rival and even outperform larger counterparts across diverse writing tasks.

4.6 Ablation of Length

In this ablation study, we compared the performance of models across different input and output lengths. Our findings indicate that most state-of-the-art models exhibit insensitivity to varying input lengths and maintain strong performance, thanks to the enhanced ability of contemporary large models to understand long texts effectively.

However, when it comes to output length, the performance of current models tends to decline as the output becomes longer. This decline manifests in aspects such as coherence, accuracy, and stylistic consistency of the generated text. Notably, models like r1 and o1, which incorporate Chain of Thought (CoT) techniques, show less performance degradation with longer outputs. The integration of CoT helps these models maintain logical coherence and improve the quality of lengthy text generation by facilitating step-by-step reasoning.

This analysis underscores the need for further optimization of models' ability to handle extended outputs. Incorporating advanced reasoning and structured approaches during generation can enhance overall performance. This finding provides valuable insights for researchers and practitioners in implementing strategies for model development.

5 Conclusion

In this paper, we propose WritingBench, emerging as a crucial innovation in evaluating large language models' writing capabilities across a diverse array of real-world tasks. By establishing a comprehensive benchmark with 1,239 queries spanning 6 primary domains and 100 subdomains, WritingBench bridges the gap in current writing evaluations by accommodating a wide range of requirements, including style, format, and length. Our proposed query-dependent evaluation framework not only aligns closely with human judgments but also enhances the assessment with dynamic criteria generation and scoring. Moreover, the development of a fine-tuned, 7-billion-parameter writing-enhanced model marks a significant step forward, offering writing performance on par with leading models like ChatGPT-4o-latest. By making WritingBench and its associated resources publicly available, we aim to foster further research and advancements in the field of writing evaluation. In the future, we will explore the adaptability of our query-dependent evaluation method to a wide range of subjective tasks (e.g., question answering and role-play agent) and see its effectiveness in downstream evaluation and SFT data filtering.

Limitations

In this study, several limitations of our approach were identified, which open avenues for future work. Firstly, both our writing model and critic model were primarily trained and evaluated on straightforward SFT data, without extensive optimization of training strategies. This limited experimentation may have restricted the potential performance gains that could be achieved with more advanced techniques.

One of the enduring challenges lies in controlling the generation length of the models. Despite utilizing our criteria for evaluation, the effectiveness of managing output length remains limited, as discussed in the Appendix. This indicates a need for more sophisticated scoring strategies, ideally incorporating rule-based evaluations to better guide the models' output.

Furthermore, conducting pair-wise preference annotations for writing tasks remains a significant challenge. When two responses are otherwise well-constructed, human annotators often exhibit subjective biases based on personal preferences. These biases can complicate pair-wise comparison tasks, introducing variability and potential inconsistencies in the annotations.

Addressing these limitations requires intensive research efforts to refine training methodologies, develop more nuanced evaluation frameworks, and establish clearer guidelines for human annotations to enhance the reliability and consistency of evaluations.

References

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2024-10-22.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *CoRR*, abs/2408.07055.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan

Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17048–17085. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard H. Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10158–10177. Association for Computational Linguistics.

Xiaobo Liang, Haoke Zhang, Helan hu, Juntao Li, Jun Xu, and Min Zhang. 2024. Fennec: Fine-grained language model evaluation and correction extended through branching and bridging. *CoRR*, abs/2405.12163.

Guillermo Marco, Julio Gonzalo, María Teresa Mateo Girona, and Ramón Santos. 2024. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19654–19670. Association for Computational Linguistics.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 355:1–355:34. ACM.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *CoRR*, abs/2312.06281.

Chau 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, Miami, Florida, USA, November 12-16, 2024*, pages 1722–1753. Association for Computational Linguistics.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *CoRR*, abs/2409.16191.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6252–6278. Association for Computational Linguistics.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Weaver: Foundation models for creative writing. *CoRR*, abs/2401.17268.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *CoRR*, abs/2307.15020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhenyuan Yang, Zhengliang Liu, Jian Zhang, Lu Cen, Tai Jiaxin, Zhong Tianyang, Li Yiwei, Zhao Siyan, Yao Teng, Liu Qing, Yang Jinlin, Liu Qixin, Li Zhaowei, Wang Kexin, Ma Longjun, Dajiang Zhu, Ren Yudan, Ge Bao, Zhang Wei, Qiang Ning, Zhang Tuo, and Tianming Liu. 2024b. Analyzing nobel prize literature with large language models. *CoRR*, abs/2410.18142.

A Benchmark Statistics

A.1 Overview of Six Main Domains

1. **Academic & Engineering:** Covers academic writing workflows, including paper outlines, abstracts, literature reviews, experiment reports, and technical documents (e.g., patents, test reports).

2. **Finance & Business:** Encompasses corporate documentation such as contracts, market analyses, investment reports, strategic plans, and operational materials (e.g., product specifications, sales reports).

3. **Politics & Judiciary:** Includes government documents (policy interpretations, white papers), legal writings (legal opinions, litigation files), and political communications (speeches, work reports).

4. **Literature & Art:** Spans creative writing (novels, poetry, scripts), artistic design (character/game concepts), and critical reviews (book/movie analyses).

5. **Education:** Focuses on pedagogical materials (lesson plans, course designs), student-teacher interactions (feedback, assignments), and institutional communications (admission promotions, parent-teacher meeting scripts).

6. **Publicity & Marketing:** Addresses modern digital content needs, including social media scripts, advertising copy, brand narratives, and multimedia campaign materials.

A.2 Overview of 100 Subdomains

See Table 7.

1. **Academic & Engineering:** Covers academic writing workflows, including paper outlines, abstracts, literature reviews, experiment reports, and technical documents (e.g., patents, test reports).

2. **Finance & Business:** Encompasses corporate documentation such as contracts, market analyses, investment reports, strategic plans, and operational materials (e.g., product specifications, sales reports).

3. **Politics & Judiciary:** Includes government documents (policy interpretations, white papers), legal writings (legal opinions, litigation files), and political communications (speeches, work reports).

4. **Literature & Art:** Spans creative writing (novels, poetry, scripts), artistic design (character/game concepts), and critical reviews (book/movie analyses).

5. **Education:** Focuses on pedagogical materials (lesson plans, course designs), student-teacher interactions (feedback, assignments), and institutional communications (admission promotions, parent-teacher meeting scripts).

6. **Publicity & Marketing:** Addresses modern digital content needs, including social media scripts, advertising copy, brand narratives, and multimedia campaign materials.

B Experiment Settings

B.1 Dataset Configuration

The dataset used for experimentation comprises 1,239 queries from the WritingBench framework. To specifically assess human consistency in evaluation, a subset of 300 queries was isolated, ensuring thorough representation across domains. Each subdomain contains three selected queries, totaling 30 queries per subdomain. For this subset, we randomly selected two models to generate responses for each query. These responses were then evaluated to not only score them but also provide detailed reasoning for the scores assigned. This

Domain	Description
Academic and Engineering	
Thesis Outline	Structured framework for organizing dissertation chapters and content
Abstract	Concise summary of research objectives, methods, and findings
Introduction	Contextual background and problem statement presentation
Contribution	Clear articulation of original research value and innovations
Literature Review	Critical synthesis of existing scholarly works
Experiment	Detailed documentation of scientific procedures and results
Conclusion	Comprehensive summary of research outcomes and implications
Limitations	Objective analysis of study constraints and validity boundaries
Acknowledgments	Formal recognition of contributors and funding sources
Defense PPT	Visual presentation structure for academic viva voce
Defense Speech Draft	Oral argumentation framework for research validation
Dissertation Proposal	Detailed plan outlining research objectives and methodology
Internship Report	Documentation of professional training experiences
R&D Documentation	Records of research processes and technological innovations
Engineering Report	Technical analysis of engineering projects and systems
Patent	Technical documentation for intellectual property protection
Test Report	Systematic evaluation of product/process performance
Finance & Business	
Contract	Legally binding agreement outlining business terms
User Survey	Design and analysis of market feedback instruments
Minutes of Meeting	Official record of corporate discussions and decisions
Briefing	Condensed executive summary of business situations
Financial Statement	Formal records of economic activities and positions
Invitation to Bid	Solicitation document for procurement opportunities
Bid Document	Competitive proposal for project acquisition
Requirements Specification	Detailed technical needs documentation
Product Planning	Strategic roadmap for product development lifecycle
Investment Analysis	Financial evaluation of capital allocation options
Risk Management	Documentation of risk assessment and mitigation strategies
Market Analysis	Comprehensive evaluation of industry trends and competitors
Market Research	Systematic investigation of consumer behavior patterns
Human Resource Management	Personnel policy and procedure documentation
Recruitment	Talent acquisition strategy and process documentation
Pitch Deck Script	Narrative structure for investment presentations
Event Planning	Organizational framework for corporate activities
Business Letter	Formal corporate communication and correspondence
Sales Report	Analytical documentation of revenue performance
Strategic Planning	Long-term organizational development blueprints

Domain	Description
Politics and Law	
Application for Party Membership	Formal petition for political organization affiliation
Ideological Report	Documentation of political belief system alignment
Policy Interpretation	Analysis and explanation of government regulations
Government Document	Official administrative correspondence and records
Policy Promotion	Public communication strategies for legislative changes
Government Speech Draft	Rhetorical framework for official addresses
Work Report	Performance documentation of governmental operations
Achievement Material	Compilation of administrative accomplishments
White Paper	Authoritative report on complex policy issues
Legal Consultation	Professional advice documentation on juridical matters
Regulation Analysis	Critical examination of legislative frameworks
Legal Opinion Letter	Professional interpretation of legal implications
Legal Agreement	Binding contractual documentation between parties
Litigation Document	Formal paperwork for legal proceedings
Judgment Document	Court-issued resolution of legal disputes
Defense Brief	Structured argumentation for legal protection
Case Analysis	Detailed examination of legal precedents and scenarios
Case Report	Comprehensive documentation of legal proceedings
Legal Propaganda	Public education materials about legal systems
Literature and Art	
Idea Brainstorming	Creative concept development documentation
Essay	Structured exploration of literary themes and ideas
Biography	Narrative documentation of individual life stories
Novel Outline	Framework for fictional narrative construction
Novel Main Text	Primary narrative composition in prose form
Novel Continuation	Extended narrative development strategies
Plot Design	Architectural planning of story progression
Creative Derivative	Adaptation documentation for existing works
Book Review	Critical analysis of literary works and themes
TV and Film Review	Analytical critique of visual media productions
Script	Narrative structure for theatrical or cinematic productions
Video Script	Sequential planning for audiovisual content
Poetry	Creative composition with rhythmic and metaphorical language
Lyric Writing	Poetic composition for musical interpretation
Character Design	Development of fictional personas and backstories
Game Design	Interactive narrative and rule system documentation
Reading Reflection	Personal interpretation of literary experiences
Hosting Script	Structured framework for event presentation
Blessing Words	Ritualistic or ceremonial language composition
Podcast Script	Audio program structure and dialogue planning

Domain	**Description**
Education	
Lesson Plan	Structured outline for instructional sessions
Course Design	Curriculum development and learning objective mapping
Education Consultation	Professional advice documentation for pedagogy
Course Assignment	Learning task specification and guidelines
Assignment Grading	Evaluation criteria and feedback documentation
Teaching Materials	Educational resources and pedagogical tools
Training Reflection	Post-instructional analysis and improvement plans
Recruitment Pamphlet	Institutional promotional materials for enrollment
Class Activity	Structured learning exercise documentation
Comment	Constructive feedback on academic performance
Education Report	Analytical documentation of pedagogical outcomes
Parent-Teacher Meeting	Documentation of academic progress discussions
Publicity and Marketing	
Slogan	Memorable phrase encapsulating brand identity
Promotional Pitch	Persuasive messaging for product/service adoption
Travel Guide	Destination marketing and itinerary planning
Promotional Copy	Persuasive text for advertising campaigns
Multimedia Production Script	Cross-platform content development framework
Social Media Content	Engaging copywriting for digital platforms
Marketing Comment	Strategic response to market trends and feedback
Brand Story	Narrative development for corporate identity
Marketing Letter	Targeted communication for customer engagement
Product Description	Technical specifications and feature highlights
Self Media	Personal branding and content creation strategies

approach allows us to analyze consistency and variance in human judgment across different domains and tasks.

B.2 Evaluation Protocol

The evaluation was conducted using a dynamic protocol where the criteria were generated and scored using Claude-3.5-Sonnet. In addition to general scoring, the requirement evaluation (column C in Table 3) included specific assessments for three specialized subsets: Style, Format, and Length. For each subset, we calculated the average score for the criteria related to its specific capabilities, ensuring a focused evaluation of each area's strengths across different models.

B.3 Training Model Configurations

For our experimental setup, we utilized a configuration featuring 8 NVIDIA A100 GPUs. The training process was conducted with a learning rate set to 7e-6, and we enabled ZeRO-3 optimization to efficiently manage memory and computational resources. Leveraging the Llama-factory framework, the writing model was trained for five epochs, while the critic model underwent three epochs of training. This setup ensured a robust training process to refine both models' performance on the tasks.

C Prompts

C.1 Initial Query Generation Prompt

Generate 10 different writing requests (in English) under {domain2} within the context of {domain1}. Ensure the requests are as detailed and specific as possible, and reflect realistic user tone and needs.

Please return in the following JSON format, and do not include anything outside of JSON:

```
[
  "Writing Request 1",
  "Writing Request 2",
  ...
]
```

C.2 Guidance Pool

- Add a requirement for generating specific lengths
- Include format adherence requirements, such as writing according to a prescribed outline or outputting in a specific format
- Add style requirements, like drafting a speech suitable for a particular occasion or adopting the style suitable for a specific audience or mimicking a particular tone
- Incorporate user personalization needs, such as considering the user's identity or integrating personal experiences
- Include more specific content requirements, like details about a particular event or focusing on specific content
- Express concisely in one sentence

C.3 Query Refine Prompt

Please refine the original writing request for {domain2} under {domain1} based on the provided modification guidance to enhance details.
Original Writing Request
{query}
Modification Guidance
{guidance}
Output Requirement
Return in the following JSON format, and do not include anything outside of JSON:

```
{
  "query": "Refined writing request (in English)"
}
```

C.4 Evaluation

1. Evaluate system: You are an expert evaluator with extensive experience in evaluating response of given query.

2. Criteria generation prompt: Please generate five strict evaluation criteria for assessing the response given the following query. Each criterion should include the following fields: name, criteria_description, score1_description, score2_description, score3_description, score4_description, score5_description.

   The criteria should be designed to emphasize detailed assessment and distinguish subtle differences in quality. Ensure that the criteria can discern issues such as relevance, coherence, depth, specificity, and adherence to the query context.

   Do not include any additional text. Only output the criteria in the specified JSON format.

   Query
   {query}

Output format
{ "name": "first criteria name",
"criteria description": "Description for the first criteria, emphasizing detailed and critical assessment.",
"1-2": "Low score description: Clearly deficient in this aspect, with significant issues.",
"3-4": "Below average score description: Lacking in several important areas, with noticeable problems.",
"5-6": "Average score description: Adequate but not exemplary, meets basic expectations with some minor issues.",
"7-8": "Above average score description: Generally strong but with minor shortcomings.",
"9-10": "High score description: Outstanding in this aspect, with no noticeable issues." },

3. Score prompt: Evaluate the Response based on the Query and criteria provided.

Criteria
`{criteria}`

Query
`{query}`

Response
`{response}`

Provide your evaluation based on the criteria:

- Provide reasons for each score, indicating where and why any strengths or deficiencies occur within the Response
- Reference specific passages or elements from the text to support your justification
- Ensure each reason is concrete with explicit references to the text
- Scoring Range: Assign an integer score between 1 to 10

Output format
Return the results in the following JSON format: { "score": an integer score between 1 to 10, "reason": "Specific and detailed justification for the score using text elements." }

Model	Total	Language		Domain						Requirement					
		ZH	EN	D1	D2	D3	D4	D5	D6	R1	C	R2	C	R3	C
Proprietary LLM															
ChatGPT-4o-latest	8.1	8.2	8.0	8.1	8.1	8.1	8.1	8.3	8.1	8.2	8.9	8.1	8.2	8.2	8.6
o1-Preview	8.1	8.1	8.2	8.1	8.1	8.1	8.2	8.3	8.1	8.2	8.8	8.1	8.2	8.2	8.6
Claude-3-5-Sonnet	7.7	7.7	7.7	7.6	7.5	7.6	7.6	7.9	8.0	7.7	8.5	7.9	7.9	7.9	8.5
Gemini-1.5-Pro	7.7	7.8	7.7	7.7	7.4	7.7	7.8	8.0	7.8	7.8	8.7	7.8	7.8	7.9	8.5
Qwen-Max	8.3	8.4	8.3	8.2	8.3	8.3	8.3	8.5	8.3	8.4	9.0	8.3	8.4	8.4	8.7
Open LLM															
Deepsekk_R1	8.5	8.7	8.4	8.5	8.4	8.6	8.6	8.6	8.6	8.6	9.0	8.6	8.4	8.6	8.9
Deepseek-V3	8.0	8.0	7.9	8.0	7.8	8.0	7.8	8.2	8.0	8.0	8.8	8.0	8.2	8.0	8.5
Mistral-Large-Instruct	7.6	7.6	7.6	7.7	7.6	7.7	7.2	7.9	7.5	7.7	8.7	7.6	7.8	7.7	8.2
Qwen-2.5-72B-Instruct	7.8	7.9	7.8	8.0	7.8	8.0	7.5	8.1	7.7	8.0	8.8	7.8	7.8	7.9	8.3
Qwen-2.5-7B-Instruct	7.3	7.2	7.4	7.6	7.3	7.5	6.6	7.7	7.2	7.5	8.5	7.2	7.2	7.3	7.8
Llama-3.3-70B-Instruct	7.1	6.8	7.4	7.1	6.9	7.1	6.9	7.4	7.3	7.2	8.3	7.1	7.3	7.2	7.9
Llama-3.1-8B-Instruct	6.3	5.6	6.8	6.5	6.3	6.0	5.9	6.7	6.5	6.3	7.5	6.2	6.3	6.3	6.9
Capability-Enhanced LLM															
Suri	5.3	4.8	5.8	6.1	5.7	5.3	4.3	5.4	5.3	5.4	5.8	4.8	4.4	5.1	5.4
Longwriter	7.9	7.9	7.9	8.1	8.1	8.1	7.7	8.2	7.6	8.1	8.7	7.7	7.7	7.9	8.3
Qwen-2.5-7B-SFT-Filter	8.1	8.2	8.0	8.1	8.1	8.1	7.9	8.3	8.0	8.2	8.8	8.0	8.1	8.1	8.5
Llama-3.1-8B-SFT-Filter	8.0	8.0	8.0	8.0	8.0	8.1	7.7	8.2	7.9	8.1	8.8	7.9	7.9	8.0	8.4

Table 8: WritingBenchmark Evaluation of LLM Performance Across 6 Domains and 3 Writing Requirements using Claude-3-5-Sonnet (Scale 0-10).

Model	Total	Language		Domain						Requirement					
		Zh	EN	D1	D2	D3	D4	D5	D6	R1	C	R2	C	R3	C
Qwen-2.5-7B-Instruct	7.4	7.3	7.5	7.7	7.4	7.6	6.9	7.8	7.3	7.6	8.6	7.4	7.5	7.5	7.9
Llama-3.1-8B-Instruct	6.4	5.7	6.9	6.6	6.4	6.1	6.0	6.7	6.6	6.4	7.6	6.3	6.4	6.4	7.0
Suri	5.0	4.4	5.5	5.6	5.3	5.0	4.1	5.0	5.1	5.0	5.4	4.5	4.0	4.8	5.2
Longwriter	7.9	7.9	7.9	8.0	8.1	8.1	7.7	8.1	7.5	8.1	8.7	7.7	7.7	7.9	8.2
Qwen-2.5-7B-SFT	7.9	8.0	7.9	7.9	7.9	8.1	7.8	8.3	7.9	8.0	8.9	7.9	8.1	8.0	8.5
Llama-3.1-8B-SFT	7.9	8.0	7.9	7.9	7.9	8.0	7.7	8.2	7.9	8.0	8.7	7.9	8.0	8.0	8.5
Qwen-2.5-7B-SFT-Filter	8.0	8.1	7.9	8.0	8.0	8.1	7.8	8.3	7.8	8.1	8.9	7.9	8.0	8.0	8.5
Llama-3.1-8B-SFT-Filter	8.0	8.0	8.0	8.0	8.0	8.1	7.8	8.2	7.9	8.1	8.8	7.9	8.1	8.0	8.5

Table 9: WritingBenchmark Evaluation of LLM Performance Across 6 Domains and 3 Writing Requirements using Critic Model (Scale 0-10).

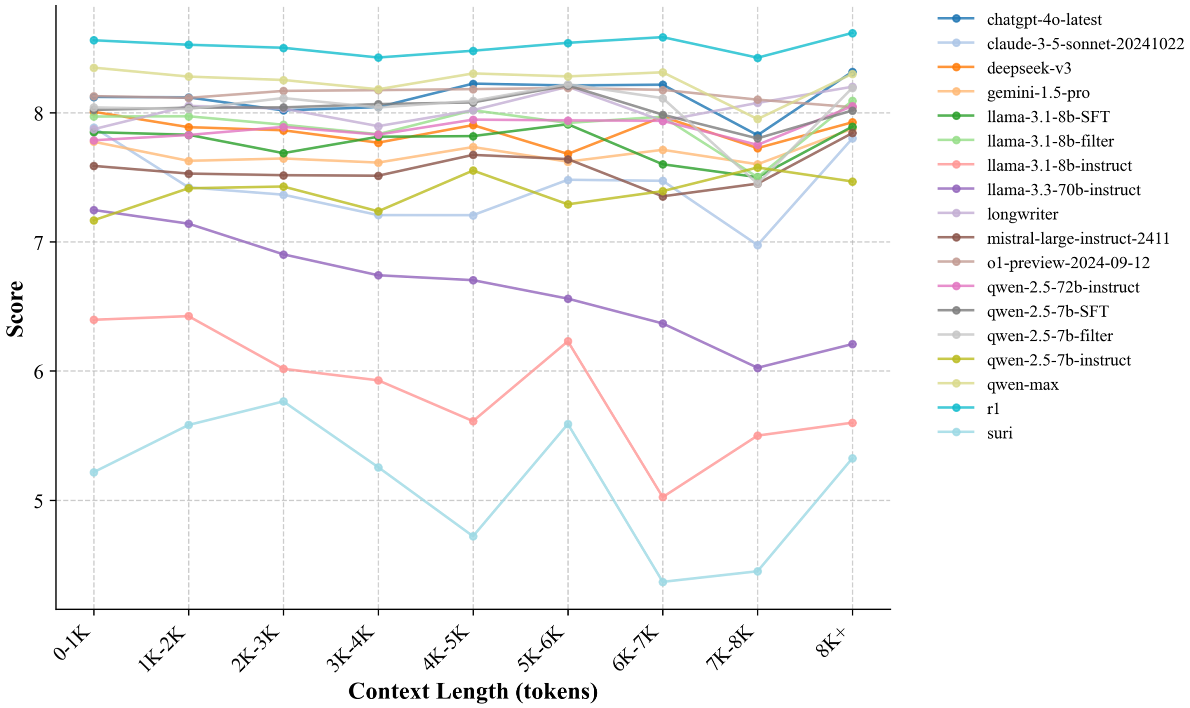

Figure 6: Scores of different model input lengths on the WritingBench.

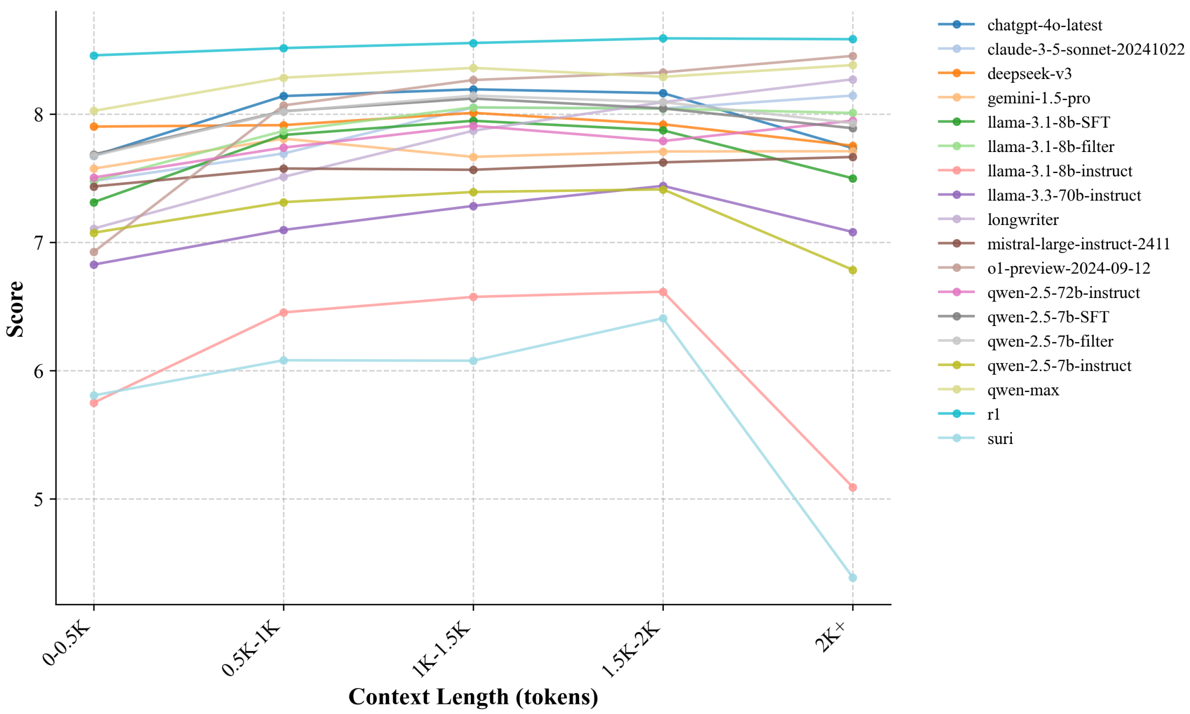

Figure 7: Scores of different model output lengths on the WritingBench.