From Long Inputs to Long Outputs: Why Generation Matters in Long-Context LLMs

Anonymous ACL submission

Abstract

Recent advancements in long-context Large Language Models (LLMs) have primarily concentrated on processing extended input contexts, resulting in significant strides in long-context comprehension. However, the equally critical aspect of generating long-form output has received comparatively less attention. This paper advocates for a paradigm shift in NLP research toward addressing the challenges of long-output generation. Tasks such as novel writing, long-term coding, and complex reasoning require models to understand extensive contexts and produce coherent, contextually rich, and logically consistent extended text. These demands highlight a critical gap in LLM capabilities. We underscore the importance of this under-explored domain and call for focused efforts to develop foundational LLMs tailored for generating high-quality, long-form outputs, which hold immense potential for real-world applications.

1 Introduction

Long-context Large Language Models (LLMs) refer to models capable of processing large amounts of input text within a fixed context window, commonly referred to as "*long-input LLMs*". Recent advances have significantly expanded this context window (Su et al., 2021; Ding et al., 2024), enabling models to handle inputs ranging from 8K to up to 1 million tokens (OpenAI, 2024c; Anthropic, 2024; Reid et al., 2024; GLM et al., 2024; Dubey et al., 2024). This ability to handle and process long input has become a core feature of modern general-purpose LLMs, leading to strong performance on long-context benchmarks (Kamradt, 2023; Bai et al., 2024b; Hsieh et al., 2024). These models enable practical applications such as retrieving information from lengthy documents (Hsieh et al., 2024; Yen et al., 2024) and summarizing content across multiple documents (Bai et al., 2024a; Chen et al., 2023).

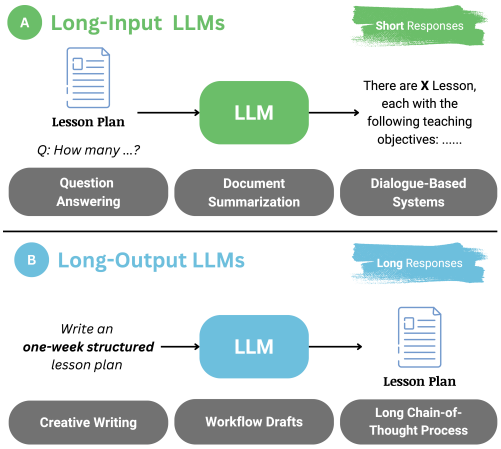

Figure 1: Difference between long-input and long-output LLMs.

While long-context LLMs have focused primarily on processing long input sequences, much less attention has been given to models that generate and evaluate long-form text, often referred to as "long output LLMs" (see Figure 1 for a comparison). However, the ability to produce long output is equally important, especially since more writing tasks now require models to generate long and well-organized content. These tasks often require not only long output generation but also *clear structure* and *coherent flow* across thousands of words. Recent studies have highlighted that many models still face significant challenges in generating high-quality long-form content beyond a few thousand words (Wu et al., 2024; Bai et al., 2024d; Ye et al., 2025). In addition, reasoning LLMs are recently introduced, where LLMs improve their reasoning ability through reinforcement learning. This reasoning process inadvertently increases output generation, highlighting the growing demand for long output generation. However, these reasoning models, especially DeepSeek-R1-Zero (Guo et al., 2025), suffer from poor readability and language mixing, showing the need to generate and evaluate long output.

This paper highlights the need to prioritize research on long-output capabilities in foundational LLMs, emphasizing the importance of advancing this relatively underexplored area. Real-world applications, such as automatic survey generation (Liang et al., 2025; Wang et al., 2024) and lengthy model reasoning processes (Qwen Team, 2025b; Comanici et al., 2025), require not only longer outputs but also a clear structure and coherent flow across thousands of words. Progress in long-output generation remains slow due to three main challenges. ① **Data limitations**: Instruction-following datasets mainly contain short input-output pairs, and there is a scarcity of high-quality long-sequence data (Bai et al., 2024a; Xiong et al., 2024; Chen et al., 2023), which restricts both research and real-world applications. ② **Task complexity**: Long-form tasks, such as novel or article writing, require coherence and logical consistency across long contexts, far exceeding the complexity of shorter tasks (Wu et al., 2024; Yang et al., 2024; Tan et al., 2024). ③ **Computational constraints**: Generating long text incurs linearly increasing computational costs in certain architectures (Gu and Dao, 2024; Dao et al., 2022), while proprietary models often impose restrictive token limits (e.g., 4K – 8K tokens) (OpenAI, n.d.; Anthropic, 2024; Reid et al., 2024). Addressing these challenges requires focused research and innovation to advance the capabilities of long-output LLMs.

Advancing long-output LLMs is essential to addressing real-world demands that require coherent and well-structured generation across long text spans. ① **Practical applications**: Fields such as healthcare, law, education, and the media heavily rely on long-form content generation, including research articles, legal documents, and detailed reports (Zhao et al., 2024b; Chiang et al., 2024). Automating these tasks with long-output LLMs streamlines workflows significantly. ② **Tackling Complex and Sustained Tasks**: While successful in long-form code generation, LLMs struggle to generalize this coherence to other complex, non-interactive tasks. This includes creating complex video game levels or composing technical patent applications (Anthropic, PBC, 2025; Anysphere, Inc., 2025; Wu et al., 2024). ③ **Efficiency—Single-pass vs. Multi-pass inference**: Figure 2 shows the efficiency of single-pass inference compared to multi-pass chained inference using the Llama-3.1-8B-instruction model. Evaluating 1,000 queries of 16K tokens each on 8×H800 GPUs, results demonstrate chained inference incurs significantly higher latency and costs, with the 32-stage configuration over seven times costlier than single-pass inference.

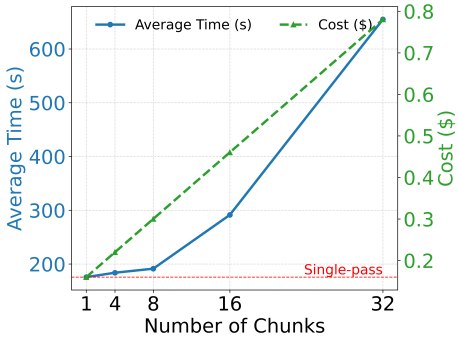

Figure 2: LLMs directly generate text in a single pass, while humans follow an iterative process of thinking, outlining, writing, and refining to ensure coherence and quality.

In a nutshell, improving the long-output capabilities of LLMs could be a rewarding and exciting opportunity for researchers.

2 Long-Output LLMs

In this section, we first explore the prevalence of tasks involving these models in real-world applications, highlighting the significant neglect of related research, as revealed by statistical data. Next, we define long-output LLMs and outline the requirements for a model to qualify as such

2.1 High Demand, Low Research Focus

Despite the growing need for models capable of generating long-form content in real-world applications, significant gaps remain in targeted research. Tasks such as scientific writing, technical documentation, and AI-driven dialogues require models capable of producing coherent, high-quality outputs over extended spans. However, research has largely concentrated on input processing (An et al., 2024a; Hsieh et al., 2024; Vodrahalli et al., 2024; Reid et al., 2024), often neglecting the complexities of long-output generation (Wu et al., 2024; Bai et al., 2024d). To substantiate this claim, we provide statistical evidence that underscores the limitations of prevailing research trends.

High Demand. To quantify the increasing demand for long-output generation in natural language processing (NLP), we analyzed 100 K user requests from real-world scenarios(Zhao et al., 2024b), calculating the input-output length ratios using the Llama-3.3-instruction-70B model (Dubey et al., 2024) and employed Few-shot learning predictions (Brown et al., 2020) to estimate

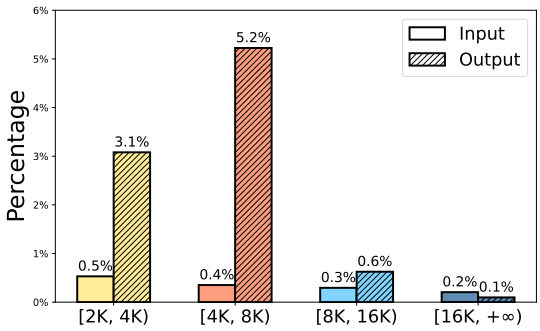

Figure 3: Proportion of real-user demand: The aforementioned 2K (words) range refers to the interval [2K, 4K), and similarly for the other ranges. Solid color fill for input demand, slash fill for output.

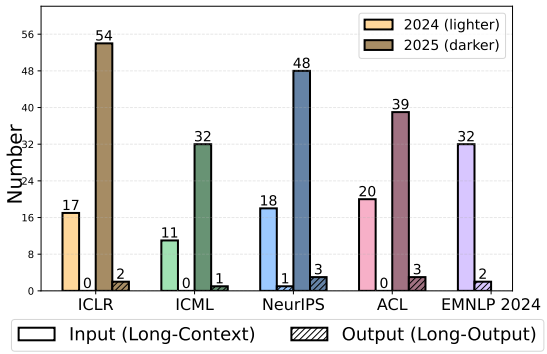

Figure 4: ML and NLP Conf Long-context Research Trends Statistics (sorted by conference date). Solid color fill for Input-paper, slash fill for Output-paper.

the output length required for real-user queries[1]. Specifically, we examined four output lengths ranges—$[2K, 4K)$, $[4K, 8K)$, $[8K, 16K)$, and $[16K, +\infty)$ words—and compared them with the distribution of input lengths. The results reveal that demand for long-output generation exceeds equivalent-length inputs by more than 2-3 times in all cases except those with outputs greater than 16K , with the ratio reaching nearly 15 times at $[4K, 8K)$ level(Figure 3). These findings underscore the necessity of long-output models for generating extensive, coherent content in practical applications[2].

Low Research Focus. While significant progress has been made in research on large input processing, particularly for long-context models, the area of long-output generation remains underexplored. This imbalance is evident in our analysis of papers from leading ML (ICML, ICLR, NeurIPS) and NLP (ACL, EMNLP, NAACL) conferences in 2024&2025[3]. Out of 271 papers addressing long-context tasks, only 12 papers specifically focused on long-output generation—a stark 23 : 1 ratio (Figure 4). This imbalance focus is particularly concerning given the real-world demand for long-output models, which often surpasses the demand for long-input models.

[1]In Appendix B, describes the specific implementation of our statistics.

[2]The user query statistics analysis of WildChat (Zhao et al., 2024b) is provided in Appendix C.

[3]Appendix E includes all paper titles. We focus solely on papers requiring LLM output of 4000 tokens or more; calculations are not performed for those falling below this threshold.

2.2 Defining Long-Output LLMs

We propose that long-output LLMs must satisfy two key requirements to effectively address the challenges of long-output generation.

Input - Context Handling Capabilities. A model's ability to handle instructions with complex context is critical for producing coherent and contextually relevant outputs over long spans. This challenge intensified when the model began generating longer text, as it had to reference and integrate the semantics of the generated segments to ensure instruction conformity and generation coherence. Transitioning from long-context models to long-output LLMs requires enhanced capabilities in managing long-range dependencies and understanding complex, long-range contextual relationships. Benchmarks like LongBench-V2 (Bai et al., 2024c) demonstrate that these capabilities go beyond mere context processing, requiring a deep understanding of long contexts to answer questions accurately.

Output - Length and Quality of Generated Text. Unlike conventional long-context models primarily designed to handle extensive input contexts—such as LLaMa 3.1 (Dubey et al., 2024) supporting up to 128K tokens or Gemini (Reid et al., 2024) accommodating 1M tokens—long-output LLMs explicitly focuses on generating coherent and high-quality outputs of considerable length. Specifically, long-output LLMs aims to produce outputs spanning thousands to millions of tokens, maintaining logical consistency, narrative coherence, creativity, and relevance throughout extended generations. This capability represents an advancement in natural language generation, addressing the challenges of preserving content quality and consistency over

substantially longer and more complex texts, thus distinguishing long-output LLMs as a fundamentally novel contribution to the field.

In this paper, we establish a performance baseline, starting with 4K tokens [4] as the effective length for long-content generation tasks.

Dataset	Input Length	Output Length
LongAlpaca	5,945	218
LongAlign	12,134	169
Suri	347	4,371
LongWriter	262	5,333

Table 1: Comparison of Average Input and Output Lengths (words) for Long-Context SFT Datasets.

3 Real-World Application

The advancement of long-output LLMs significantly broadens creative writing capabilities beyond traditional short-form tasks, effectively addressing complex and demanding scenarios. This highlights the versatile real-world applicability of long-output LLMs .

Policy Drafting and Legislative Writing. long-output LLMs excel in creating structured, comprehensive documents like policy drafts and legal texts. Unlike traditional models limited to brief communications or partial reports, long-output LLMs autonomously generates entire coherent documents. This automation streamlines repetitive tasks, enabling professionals to focus on higher-level analysis and decision-making, significantly enhancing productivity and output quality (Narendra et al., 2024; Guha et al., 2023).

Educational Content Development. Long-output LLMs greatly benefit creative writing, notably in genres such as children's literature and science fiction. Authors can use long-output LLMs to craft complete narratives, refine drafts, and ensure consistency throughout extended texts. This mitigates disjointedness common in incremental content creation, supporting seamless, cohesive creative workflows and empowering authors in ambitious literary projects (Bae and Kim, 2024; Tian et al., 2024).

Complex Planning and Decision Making. Long-output LLMs facilitate complex planning and strategic decision-making, including project and detailed itinerary development. By generating thorough plans that integrate multiple variables, these models deliver comprehensive and contextually informed solutions, especially valuable when outputs exceed typical token limits (e.g., beyond 4K tokens) (Tang et al., 2024; Hao et al., 2025).

Scientific Research Surveys and Academic Paper Writing. Extensive literature reviews, meta-analyses, and state-of-the-art reports often require synthesizing numerous publications, a task challenging for traditional summarization models. Advanced agent-based methods, such as Deep Research[5], leverage the capabilities of long-output LLMs. These models effectively manage lengthy source materials, maintaining thematic consistency and structuring comprehensive reviews, thus accelerating knowledge dissemination and fostering interdisciplinary collaboration.

In summary, long-output LLMs substantially enhance existing language models by generating extensive, coherent, high-quality content, thereby becoming essential tools for automating and improving complex writing and strategic planning tasks across various professional domains[6].

4 Current State of Long-Output LLMs

This section provides an overview of the current landscape of long-output LLMs , organized into three key areas: Data, Benchmarks, and Models. These dimensions collectively represent the core elements driving progress in the field—Data provides the foundation, Benchmarks set the evaluation standards, and Models showcase the cutting-edge advancements in long-output generation.

4.1 Data

During the long-context continual-pretraining phase, the datasets used for Long-Input LLMs and Long-Output LLMs overlap. However, significant

[4]The selection of 4K as the starting point for long-output LLMs is based on two key reasons. First, it aligns with the starting point established in long-context benchmarks such as Ruler (Hsieh et al., 2024). Second, it aligns with the length of real-world requirements, as shown in Figure 3. While our work provides a preliminary and simplified definition, similar to the evolving notion of long-context LLMs, the precise boundaries of long-output models will require continued efforts from the research community, including the development of diverse benchmarks to evaluate their capabilities comprehensively.

[5]OpenAI-DeepResearch: https://openai.com/index/introducing-deep-research/

[6]Long-output LLMs also have a wide range of other applications, such as long-horizon code generation and extended chain-of-thought reasoning. These use cases are discussed in more detail in the Appendix D.

divergence occurs during supervised fine-tuning. Early research primarily focused on datasets featuring extended input sequences while limiting output sequences to shorter lengths (Xiong et al., 2024; Xu et al., 2024). For instance, datasets like LongAlpaca-12k (Chen et al., 2023) and LongAlign-10k (Bai et al., 2024a) were designed for tasks such as summarization and question answering, where outputs length remained relatively constrained.

More recently, datasets have been developed to support the generation of longer, more detailed outputs. Suri (Pham et al., 2024), for example, employs backtranslation to transform long-content data into comprehensive instructions as input. Similarly, LongWriter-6k (Bai et al., 2024d) uses an agent-based methodology to generate a plan for user queries and then produces responses in segments, ensuring coherence in long-form outputs. Another work, Self-Lengthen (Quan et al., 2024), uses iterative expansion to progressively extend responses through repeated elaboration, resulting in more detailed and lengthy outputs.

This research represents a significant evolution in the field, as more datasets are specifically constructed to facilitate long-output generation—an essential capability for tasks that demand extensive reasoning or the production of extended text. Table 1 summarizes several key datasets' average input and output lengths, underscoring the growing focus on generating longer outputs to fine-tune long-context LLMs.

4.2 Benchmarks and Evaluation

Long-output benchmarks for long-context LLMs are designed to assess both the length and quality of outputs exceeding 4K tokens (approximately 2.6K words). These benchmarks differ significantly from traditional benchmarks that primarily focus on processing long input contexts ($\geq$16K tokens) (Kamradt, 2023; Hsieh et al., 2024; Bai et al., 2024c) or generating moderate-length outputs of around 1K tokens (Xu et al., 2020; Stelmakh et al., 2022; Xu et al., 2022; Tan et al., 2024). However, because of the considerable length of the generated text, manual evaluation of its quality becomes time-consuming and impractical. To address this, three evaluation approaches have been developed.

The first approach is rule-based evaluation, which focuses on verifying output length by counting tokens or words (Bai et al., 2024d; Quan et al., 2024; Liu et al., 2024). While this method ensures compliance with predefined length requirements, it provides little to no insight into the quality, coherence, or depth of the generated content. Consequently, it is limited in its ability to offer a holistic evaluation of long-output models.

The second approach is LLM-based evaluation, which leverages the capabilities of LLMs to evaluate outputs in two distinct ways. One method involves using an LLM to assess the entire output holistically (Bai et al., 2024d; Quan et al., 2024; Ye et al., 2025), while the other relies on a predefined checklist to determine whether the output meets specific criteria (Pham et al., 2024; Que et al., 2024). Although LLM-based evaluations provide more comprehensive insights compared to rule-based methods, they are computationally expensive and heavily reliant on the model's ability to understand and evaluate long, complex texts.

The third approach is segment-based evaluation, exemplified by frameworks like LongGenBench (Wu et al., 2024). This method divides the output into smaller, more manageable segments, allowing for detailed and interpretable assessments of each portion. However, this approach is best suited to tasks that involve structured outputs and is less applicable to unstructured or narrative-based long-output tasks.

4.3 Models

While many recent models claim strong long-context capabilities, they often focus on handling long inputs rather than generating extended outputs. Benchmarks like LongGenBench (Wu et al., 2024) and LongWrite-Ruler (Bai et al., 2024d) reveal that current models struggle to maintain quality and coherence in outputs exceeding 4,000 tokens[7]. This limitation persists despite advancements in model architectures and training methods.

Three models demonstrate potential in generating extended outputs: Bai et al. (2024d), Pham et al. (2024), and Quan et al. (2024). These models share common methodologies, including the use of specialized datasets (as outlined in Section 4.1) and fine-tuning techniques to optimize long-output performance. Additionally, approaches like Direct Preference Optimization (DPO) (Rafailov et al., 2024) are used to refine output length control. However, despite these innovations, current models still face significant challenges in generating coherent, high-quality outputs at longer lengths, as

[7]The details of the results from existing models are presented in Appendix G.

evidenced by suboptimal performance on benchmarks like LongGenBench (Wu et al., 2024) and LongBench-Write (Bai et al., 2024d).

5 Challenges and Opportunities

Shifting the research focus from long-context understanding to long-form generation reveals a distinct set of challenges across the entire LLM development lifecycle. This section explores these critical hurdles and corresponding opportunities in three key areas: the data used for training, the benchmarks for evaluation, and the model architectures for efficient training and inference.

5.1 Data

High-quality data is the bedrock of capable models, yet current datasets for long-output generation suffer from two primary limitations: a misalignment with user needs and an over-reliance on synthetic data.

User Demand Alignment. Real-world user demands are often not reflected in the inputs provided by existing datasets. A cross examination of user requests for long-output demands in the WildChat dataset (Zhao et al., 2024b), as described in Appendix B, against those of LongWriter and Suri demonstrates the discrepancies. We used the all-mpnet-base-v2 model (Reimers and Gurevych, 2020) to embed the request input and applied T-SNE (Van der Maaten and Hinton, 2008) for visualization. As shown in Figure 5, LongWriter only partially aligns with user demands, while Suri shows minimal overlap due to its synthetic instruction generation. This misalignment suggests that models trained on these datasets may struggle to generalize effectively to real-world scenarios, highlighting a critical gap in meeting user needs.

Synthetic Data. Given the scarcity of natural long-text data, synthetic datasets such as Longskywork (Zhao et al., 2024a) and FILM (An et al., 2024b) have been widely used. However, as demonstrated by Wu et al. (2024), synthetic data often introduce artificial dependencies that fail to capture the nuanced contextual relationships of real-world text. Hence, while these datasets may support performance on input-centric benchmarks, they fall short in enabling coherent and meaningful long-form text generation.

Opportunities. A fundamental limitation of current datasets is that they capture only the final text, not the dynamic, non-linear process of its creation. Human writing, especially for complex tasks, is not a linear stream of thought but an iterative process of "Concomitant Thinking"—a cycle of brainstorming, drafting, and revision. To train models that can replicate this procedural reasoning, we propose two data-centric opportunities. First, data collection must shift from static final products to dynamic "Process Logs." By creating datasets that include edit histories, discarded drafts, and brainstorming notes, we can train models on the workflow of writing, not just its output. Second, to scale this effort, we can use agent-based simulations to generate synthetic process data that mimics the "draft-reflect-revise" loop. A hybrid approach, combining authentic "Process Logs" with agent-generated data, offers a path to building large-scale SFT datasets that balance quality with quantity, better equipping models for real-world long-form generation.

5.2 Evalution/Benchmarks

Long-output benchmarks aim to evaluate both the length and quality of outputs exceeding 4K tokens. However, they face several challenges that hinder their effectiveness, particularly in model evaluation and applicability.

Limited Scope of Benchmarks. There is a significant mismatch between existing benchmarks and real-world user demands, often exceeding the gap observed between training data and real-world user demands to an even greater extent. As illustrated in Figure 6 through UMAP results, while benchmarks like LongWriter (Bai et al., 2024d) exhibit reasonable alignment with actual user needs, others, such as LongGenBench (Wu et al., 2024) and HelloBench (Que et al., 2024), demonstrate considerable divergence. This misalignment stems from the narrow scope of these benchmarks, which tend to focus on a limited subset of long-output tasks, leaving a vast range of real-world applications unaddressed. Consequently, the results derived from such benchmarks lack generalizability, impeding the development of models capable of tackling the diverse and complex requirements of long-output scenarios.

Evaluating Output Quality. Assessing the quality of long-form text remains an open challenge. Current methods fall short in several ways. Rule-based evaluations can precisely measure narrow aspects like mathematical reasoning (Liu et al., 2024) or instruction following (Wu et al., 2024), but they

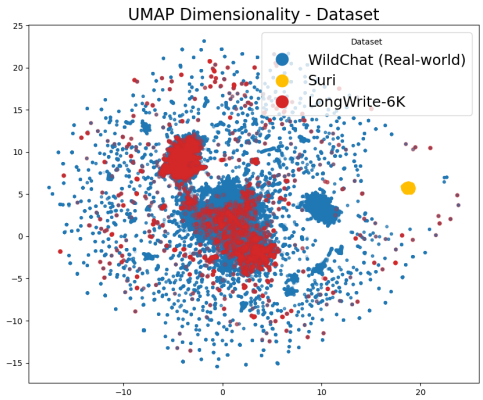

Figure 5: UMAP visualization results for different SFT datasets. WildChat is derived from the long output demands of real users, filtered and referenced in Section 2.1.

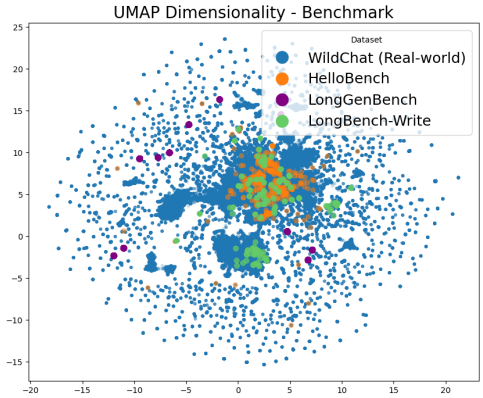

Figure 6: UMAP visualization results for different benchmark. We use the instructions from the benchmark to evaluate whether the benchmark assesses a wide range of long-output demand.

cannot capture holistic qualities like coherence, narrative flow, or logical consistency. Conversely, LLM-based evaluators, while more flexible, often function as "black boxes." They lack interpretability, rarely explaining *why* an output is judged poorly, which hinders targeted model improvement. Furthermore, their reliability diminishes on longer texts, as their own long-context understanding is a known weakness (Bai et al., 2024c). Our experiment in Appendix F, where a logical inconsistency [8] was introduced into a 3,000-word story, showed that most LLM evaluators failed to detect the flaw, despite easily identifying it in a 300-word version. Combined with high API costs, these limitations highlight the urgent need for more robust, interpretable, and scalable evaluation frameworks.

Opportunities. Evaluating high-level qualities in long-text generation, such as coherence, logical consistency, and creativity, presents a key opportunity for the evolution of long-output LLMs 's capabilities. We advocate for elevating existing rule-based benchmarks to a new level by constructing a structured evaluation paradigm. One implementation path is to establish explicit heuristic rules for these high-level qualities, for instance, by quantifying coherence through the construction of narrative graphs[9] or by leveraging existing research(Zhao et al., 2024c) to measure creativity using metrics of novelty and originality. Building on this foundation, we propose a more forward-looking objective: to distill and unify the "greatest common divisor" of human preferences in evaluating long texts. This aims to build an inclusive framework that captures the core, universal principles humans rely on when judging complex narratives. The framework aims to move beyond the traditional Reinforcement Learning from BT-Reward model**??**, mitigating its tendency toward stylistic convergence due to reward model biases, thereby preventing reward hacking and fostering genuinely diverse, high-quality content.

5.3 Train & Inference

While the inference processes for long-output LLMs share similarities with those of long-context LLMs, our experimentation reveals unresolved challenges.

Write-Side Performance Bottleneck. Long-output generation introduces significantly higher inference costs than input processing. As shown in Fig. 7, increasing the proportion of output tokens — even with the same total sequence length — causes latency to rise sharply, often several times slower than input-side processing (Reid et al., 2024; OpenAI, n.d.; GLM et al., 2024). This slowdown stems from the sequential dependency of decoding: each token relies on previous outputs, which severely limits parallelism and places heavy pressure on memory bandwidth. While optimizations such as FlashAttention-2 (Dao, 2023) and paged KV man-

[8] At a critical point in the story, a sentence is introduced that states the magic mirror has shattered and can no longer speak. As a result, any subsequent dialogue from the mirror becomes logically inconsistent.

[9] For example, if a character is stated to have died, subsequent parts of the story should not depict them as living with their spouse.

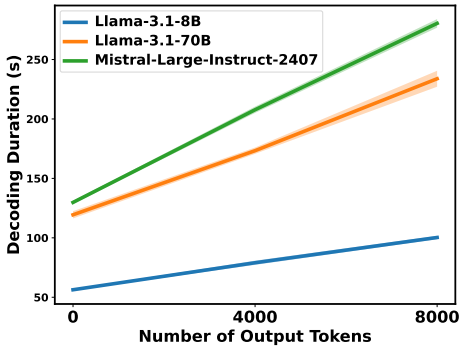

Figure 7: We set the total context length to 12,000 and gradually increased the proportion of output tokens.

agement (Kwon et al., 2023) improve utilization and mitigate some runtime costs, they cannot eliminate the token-by-token dependency that dominates write-side latency. Moreover, when speculative decoding is enabled, it can reduce the number of decoding steps but introduces additional verification overhead (Leviathan et al., 2023). Most KV compression or quantization schemes (Hooper et al., 2024) primarily reduce memory footprint but have limited impact on bandwidth demand, since the active KV state still grows with output length

Read–Write Imbalance. Even with the same total number of tokens, output generation remains inherently slower and more resource-intensive than input processing. Prefill benefits from highly parallel matrix–matrix operations, whereas decoding devolves into sequential matrix–vector computation, constrained by memory bandwidth (NVIDIA Developer Blog, 2023). Speculative decoding can reduce the number of decoding steps but at the cost of additional verification overhead (Cai et al., 2024). Meanwhile, production runtimes such as vLLM may still encounter KV fragmentation and scheduling hotspots under long decoding and high concurrency (Kwon et al., 2023). Despite continuous kernel-level advances (Hong et al., 2023), output-side latency remains the dominant barrier to scaling long-form applications.

Opportunities. To address the write-side bottleneck and the read–write imbalance in long-output LLMs, we argue for a memory-first, write-optimized architecture that embeds memory management into the model. A practical direction combines the compute/bandwidth relief of linear attention (as in Qwen3-Next(Qwen Team, 2025a,b)), which reduces complexity from $O(n^2)$ to near-linear, with the locality benefits of sparse attention (as in DeepSeek-V3.2(DeepSeek-AI, 2025)), which concentrates computation on information-dense regions; together, these mechanisms decouple cost from sequence length and curb per-step KV traffic during decoding. At the systems level, a persistent memory plane should store document outlines, entity states, and long-term constraints across sessions, with lightweight segment recurrence at decode time to query these commitments without replaying long histories. Paging, compression, and placement need to be co-designed with the runtime (e.g., paged attention) so long-lived state is compacted, quantized, and scheduled to avoid fragmentation and verification hot spots under high concurrency. Overall, this memory-centric view unifies efficient attention with structured access to stable knowledge, improving latency, memory pressure, and sustained coherence for extended long-form generation.

6 Conclusion

In this paper, we have defined and explored the potential of long-output LLMs . Despite their significant real-world applications, such models have yet to receive the attention they deserve in both academic research and practical implementations. As artificial intelligence and natural language processing continue to evolve, generating long-form content is becoming increasingly crucial, particularly for automated content creation, intelligent assistants, and complex information processing. We have identified the defining features, challenges, and emerging trends surrounding long-output LLMs , calling for greater focus on advancing this domain. Future research should prioritize improving the quality, efficiency, and controllability of these models, while also developing new evaluation metrics and exploring diverse application scenarios. The development of long-output LLMs has the potential to drive substantial change across industries, ushering in a new era of intelligent transformation.

Limitations

The main limitations of this work are as follows:

- Our experiments are confined to the tasks and genres we define, so generalization to other domains, languages, and multimodal settings remains to be established.
- The data used are primarily public datasets

from 2024, which may diverge from 2025 real-world user distributions; this gap can limit the external validity of our findings.

- Evaluation relies on automatic metrics and LLM-as-a-judge with limited human auditing; even with released prompts and fixed seeds, small residual bias and noise may persist.

- As a position paper, several components are presented conceptually rather than as full implementations, and large-scale empirical validation is deferred to future work.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024a. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024b. Make your llm fully utilize the context. *Preprint*, arXiv:2404.16811.

Anthropic. 2024. Anthropic: Introducing claude 3.5 sonnet.

Anthropic, PBC. 2025. Claude. https://www.claude.ai. Accessed: 2025-10-01.

Anysphere, Inc. 2025. Cursor - the ai-first code editor. https://www.cursor.so. Accessed: 2025-10-01.

Minwook Bae and Hyounghun Kim. 2024. Collective critics for creative story generation. *Preprint*, arXiv:2410.02428.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1376–1395, Miami, Florida, USA. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024c. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

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

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention. Hugging Face. Model card.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

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

Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. 2025. Large language models can solve real-world planning rigorously with formal verification tools. *Preprint*, arXiv:2404.11891.

Ke Hong, Guohao Dai, Jiaming Xu, Qiuli Mao, Xiuhong Li, Jun Liu, Kangdi Chen, Yuhan Dong, and Yu Wang. 2023. FlashDecoding++: Faster large language model inference on gpus. *arXiv preprint arXiv:2311.01282*.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*. NeurIPS 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *ICML*. PMLR.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, and 1 others. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

MoonshotAI-KiMi. 2025. Kimi k1.5: Scaling reinforcement learning with llms. Accessed: 2025-01-21.

Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. 2024. Enhancing contract negotiations with LLM-based legal document comparison. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153, Miami, FL, USA. Association for Computational Linguistics.

NVIDIA Developer Blog. 2023. Mastering llm techniques: Inference optimization. https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Learning to reason with llms.

OpenAI. 2024c. Openai: Hello gpt-4o.

OpenAI. n.d. Overview. Accessed: 2024-12-21.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation. *arXiv preprint arXiv:2406.19371*.

Shanghaoran Quan, Tianyi Tang, Bowen Yu, An Yang, Dayiheng Liu, Bofei Gao, Jianhong Tu, Yichang Zhang, Jingren Zhou, and Junyang Lin. 2024. Language models can self-lengthen to generate long texts. *arXiv preprint arXiv:2410.23933*.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Noah 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.

Qwen Team. 2025a. Qwen3-next-80b-a3b-instruct. Hugging Face. Model card.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. ProxyQA: An alternative framework for evaluating long-form text generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6806–6827, Bangkok, Thailand. Association for Computational Linguistics.

Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, and Wei Ma. 2024. ItiNera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1413–1432, Miami, Florida, US. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *Preprint*, arXiv:2407.13248.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, and 1 others. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*.

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

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024. Longgenbench: Benchmarking long-form generation in long context llms. *arXiv preprint arXiv:2409.02076*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, and 1 others. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5180–5197. Association for Computational Linguistics.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2831–2845. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. 2024. Logu: Long-form generation with uncertainty expressions. *arXiv preprint arXiv:2410.14309*.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025. Longproc: Benchmarking long-context language models on long procedural generation. *Preprint*, arXiv:2501.05414.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, Yimeng Gan, Rui Hu, Shuicheng Yan, Han Fang, and Yahui Zhou. 2024a. Longskywork: A training recipe for efficiently extending context length in large language models. *Preprint*, arXiv:2406.00605.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024b. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, and 1 others. 2024c. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A Alternative Views

While this paper advocates for prioritizing long-output generation in LLMs, there are alternative perspectives that challenge this position and propose different research priorities.

Long-Output Generation is Not Always Necessary. An alternative view suggests that long-output generation, where a model produces extended text in one pass, may not be essential. Instead, long-context LLMs can use chained inference, generating text iteratively by using each output as input for the next step. This approach builds long texts incrementally, leveraging the model's ability to maintain coherence over extended contexts without requiring specialized training for long outputs. Chained inference reduces computational challenges like memory bottlenecks and maintains local coherence, making it suitable for tasks like novel writing or report generation. It also allows for human feedback or external knowledge integration at each step, enhancing control in applications like technical documentation. Researchers could focus on optimizing chaining—improving context retention and minimizing cumulative errors—to ensure high-quality outputs. However, challenges like context drift and repeated context processing costs remain. This line of work offers a modular, practical alternative to direct long-output generation, balancing quality and efficiency.

Long-Context Input Optimization Over Long-Output Generation. A counterargument prioritizes optimizing long-context input processing over long-output generation, arguing that robust input comprehension is a prerequisite for coherent extended outputs. Without effective handling of long, diverse inputs, outputs risk being incoherent or inaccurate, especially in tasks like summarizing reports or generating narratives from large datasets. Current challenges, including attention inefficiencies and memory constraints, highlight the need for continued focus on input processing.

Advancements in techniques like sparse attention or memory-augmented architectures could improve input scalability, benefiting tasks beyond long outputs, such as question answering or knowledge synthesis. Critics note that this focus may delay progress on long-output tasks, and input optimization alone doesn't ensure high-quality generation. Still, prioritizing input processing builds a strong foundation for downstream tasks, ensuring long outputs are grounded in accurate context understanding.

B Proportion of real-user demand

B.1 Input Length Statistics and Classification

This subsection introduces a method for classifying the length of input text based on language and content size. The process involves two main steps: language detection and length calculation.

1. **Language Detection**: Use regular expressions to identify the language of the text based on character counts. The text is classified as Chinese, Japanese, Korean, or English, depending on the dominant script.

2. **Length Calculation**: The length of the text is calculated differently for different languages:

   - For Chinese, Japanese, and Korean, character count is used.
   - For English, word count is used.

3. **Length Classification**: Text is categorized into length buckets based on word count:

   - 2K-4K: 2,000 to 4,000 words
   - 4K-8K: 4,000 to 8,000 words
   - 8K-16K: 8,000 to 16,000 words
   - 16K+: More than 16,000 words

B.2 Predicting User's Length Requirement with LLaMA 3.3-70B

We describe a method for predicting the length of a user's input requirement using the LLaMA 3.3-70B model for few-shot learning. The process involves two main steps: predicting whether the input exceeds 2,000 words, and predicting the exact length requirement based on the first prediction.

1. **Step 1: Predicting Length Exceedance (Prompt 1)**: The first prediction is made by checking whether the input exceeds 2,000 words. A carefully crafted prompt (Prompt 1) is provided to the model to predict if the content's expected word count will surpass the 2K threshold. The model utilizes few-shot learning with example inputs to classify the task into either "above 2K" or "below 2K" based on the nature of the input.

2. **Step 2: Predicting Exact Length Requirement (Prompt 2)**: Once the model predicts whether the task exceeds 2,000 words, a second prediction is made to determine the exact length category. Based on the result from Step 1, Prompt 2 is designed to predict whether the content is in the 2K-4K, 4K-8K, 8K-16K, or 16K+ category. The model provides the final prediction by analyzing the contextual hints and the input length characteristics.

Prompt-1

Guidelines:
To determine whether the expected output will exceed 2000 words, consider the following factors:

1. **Depth and Complexity:** Does the task require detailed explanations, in-depth analysis, or comprehensive coverage of complex topics?
2. **Scope and Breadth:** Does the task cover multiple subtopics, perspectives, or extensive subject matter?
3. **Structure and Sections:** Does the output need to include multiple sections such as introductions, literature reviews, methodologies, results, discussions, and conclusions?
4. **Research and References:** Does the task require extensive research, citations, and referencing of multiple sources?

Response Format:

- Answer with either "#*# Yes" or "#*# No".
- Provide a concise justification based on the guidelines above.

Example 1:
Query: Is Sanskrit the oldest language?
Answer: This question requires a concise factual answer, not an extensive output. #*# No
***** END**
Example 2:
Query: Create a detailed business plan for a new cat litter product.
Answer: Creating a detailed business plan involves multiple sections such as market research, product development, financial projections, marketing strategy, and competitive analysis, all of which require in-depth exploration and explanation. #*# Yes ***** END**
.....
.....
Assess the following statement and decide whether the expected response is likely to require more than 2000 words. Answer with either "#*# Yes" or "#*# No," and include a brief justification, like above example.
Query: **User Query**
Answer:

Prompt-2

Guidelines:
To estimate the expected length of the output, consider the following factors:

1. **Depth and Complexity:** Does the task require detailed explanations, in-depth analysis, or complex reasoning?
2. **Scope and Breadth:** Does the task cover multiple subtopics, perspectives, or an extensive subject matter?
3. **Structure and Sections:** Does the output require multiple sections (e.g., introduction, literature review, methodologies, results, discussions, conclusions)?
4. **Research and References:** Does the task require significant research, citations, or references to multiple sources?
5. **Detail Level:** Is the task expected to be highly detailed, or can it be summarized concisely?

Response Format:
- Choose the most likely word count category: "Less than 2000 words", "2000 words", "4000 words", "8000 words", or "16000 words". Using (### Category: "Chosen category") as the response format.
- Provide a brief justification based on the guidelines above.
Example 1: Less than 2000 words
Query: Is Sanskrit the oldest language?
Answer: This is a factual question requiring a brief answer with no complex analysis or subtopics. Likely to be less than 2000 words. ### Category: Less than 2000 words
Explanation: Similar to a short blog post or brief news article, this task needs minimal detail and is concise. ***** END**
Example 2: 2000 words (2000 to 4000 words)
Query: Describe the key differences between classical and quantum computing.
Answer: This question requires moderate detail, comparing classical and quantum computing without exhaustive technical exploration. Likely to be around 2000 words. ### Category: 2000 words
Explanation: Similar to a moderate-length essay or a detailed blog post, this task covers key points with enough depth but remains manageable. ***** END**
.....
.....
Example 5: More than 16000 words

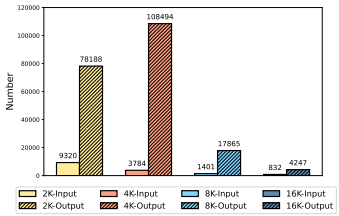

Figure 8: Proportion of real-user demand: The aforementioned 2K range refers to the interval [2K, 4K), and similarly for the other ranges. Solid color fill for input demand, slash fill for output demand in the Wildchat dataset.

C WildChat Long-Output Query Statistics and analysis

We analyzed input-output length ratios using the WildChat: 1M ChatGPT Interaction Logs dataset (Zhao et al., 2024b) and the Llama-3.3-70B model (Dubey et al., 2024). Using few-shot learning (Brown et al., 2020), we categorized output lengths into four ranges—$[2K, 4K)$, $[4K, 8K)$, $[8K, 16K)$, and $[16K, +\infty)$ words—and compared them to input lengths. The results show that demand for long-output generation is more than five times higher than for equivalent-length inputs, peaking at nearly 20 times at 4K-8K levels (Figure 8). However, due to the lack of file upload support in WildChat, the statistics for long-input queries are likely underestimated.

D Long Chain-of-Thought Task

One of the most impactful applications of long-output generation is its ability to support long chain-of-thought (CoT) tasks, which require extended sequences of reasoning to solve complex problems. These tasks serve as a key benchmark for evaluating and advancing the capabilities of LLMs.

The long CoT approach, as exemplified by the OpenAI o1 model (OpenAI et al., 2024), has demonstrated remarkable success across a range of domains. In mathematics, for instance, this method enables LLMs to tackle challenging problems, such as those encountered in Math Olympiads, while also excelling in tasks like complex code generation. These achievements underscore the transformative potential of long CoT techniques in domains that demand rigorous, systematic reasoning.

A critical enabler of long CoT success is the advancement of long-context scaling (MoonshotAI-KiMi, 2025). Complex reasoning tasks often result in extended outputs, requiring models to effectively manage both lengthy input and output sequences while maintaining coherence, relevance, and accuracy throughout. This necessitates innovations in scaling techniques to accommodate extended sequences without compromising performance. By refining these techniques, researchers ensure that LLMs can continue to perform effectively as the complexity and length of their outputs increase.

The progress in long CoT training highlights the importance of prioritizing long-output generation in LLM research. By tailoring models to meet the demands of long CoT applications, researchers unlock new possibilities in areas such as advanced problem-solving, strategic planning, and decision-making. As the field of long-output generation continues to evolve, its integration with long CoT tasks will play a pivotal role in shaping the future capabilities of LLMs, enabling them to address increasingly complex challenges across disciplines.

E Long-input & output paper list

E.1 ICML 2024

1. Linguistic Calibration of Long-Form Generations
2. Short-Long Convolutions Help Hardware-Efficient Linear Attention to Focus on Long Sequences
3. A Human-Inspired Reading Agent with Gist Memory of Very Long Contexts
4. Memory Consolidation Enables Long-Context Video Understanding
5. QUEST: Query-Aware Sparsity for Efficient Long-Context LLM Inference
6. LLM Maybe LongLM: SelfExtend LLM Context Window Without Tuning
7. LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens
8. Training-Free Long-Context Scaling of Large Language Models
9. LoCoCo: Dropping In Convolutions for Long Context Compression
10. Benchmarking and Building Long-Context Retrieval Models with LoCo and M2-BERT
11. Data Engineering for Scaling Language Models to 128k Context

E.2 ICLR 2024

1. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models
2. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis
3. BooookScore: A Systematic Exploration of Book-Length Summarization in the Era of LLMs
4. In-Context Pretraining: Language Modeling Beyond Document Boundaries
5. Functional Interpolation for Relative Positions Improves Long Context Transformers
6. RingAttention with Blockwise Transformers for Near-Infinite Context
7. Hierarchical Context Merging: Better Long Context Understanding for Pre-trained LLMs
8. FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores
9. CLEX: Continuous Length Extrapolation for Large Language Models
10. Retrieval Meets Long Context Large Language Models
11. IceFormer: Accelerated Inference with Long-Sequence Transformers on CPUs
12. PoSE: Efficient Context Window Extension of LLMs via Positional Skip-wise Training
13. Efficient Streaming Language Models with Attention Sinks
14. YaRN: Efficient Context Window Extension of Large Language Models
15. Parallelizing Non-linear Sequential Models over the Sequence Length
16. In-context Autoencoder for Context Compression in a Large Language Model
17. HyperAttention: Long-context Attention in Near-Linear Time

E.3 NIPS 2024

1. Streaming Long Video Understanding with Large Language Models
2. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack
3. Perceiving Longer Sequences with Bi-Directional Cross-Attention Transformers
4. Video Token Merging for Long Video Understanding
5. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks
6. LoTLIP: Improving Language-Image Pre-training for Long Text Understanding
7. Stress-Testing Long-Context Language Models with Lifelong ICL and Task Haystack
8. MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention
9. MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations
10. An Efficient Recipe for Long Context Extension via Middle-Focused Positional Encoding
11. Found in the Middle: How Language Models Use Long Contexts Better via Plug-and-Play Positional Encoding
12. Mini-Sequence Transformers: Optimizing Intermediate Memory for Long Sequences Training
13. Mixture of In-Context Experts Enhance LLMs' Long Context Awareness
14. MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding
15. StreamingDialogue: Prolonged Dialogue Learning via Long Context Compression with Minimal Losses
16. InfLLM: Training-Free Long-Context Extrapolation for LLMs with an Efficient Context Memory
17. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding

18. Rethinking Transformer for Long Contextual Histopathology Whole Slide Image Analysis

E.4 ACL 2024

1. L-Eval: Instituting Standardized Evaluation for Long Context Language Models
2. Analyzing Temporal Complex Events with Large Language Models? A Benchmark Towards Temporal, Long Context Understanding
3. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression
4. Making Long-Context Language Models Better Multi-Hop Reasoners
5. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding
6. Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models
7. CoCA: Fusing Position Embedding with Collinear Constrained Attention in Transformers for Long Context Window Extending
8. NextLevelBERT: Masked Language Modeling with Higher-Level Representations for Long Documents
9. RelayAttention for Efficient Large Language Model Serving with Long System Prompts
10. Marathon: A Race Through the Realm of Long Context with Large Language Models
11. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens
12. FinTextQA: A Dataset for Long-form Financial Question Answering
13. Long Context is Not Long at All: A Prospector of Long-Dependency Data for Large Language Models
14. M4LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models
15. Chunk, Align, Select: A Simple Long-sequence Processing Method for Transformers
16. LooGLE: Can Long-Context Language Models Understand Long Contexts?
17. Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training
18. DocFinQA: A Long-Context Financial Reasoning Dataset
19. SumSurvey: An Abstractive Dataset of Scientific Survey Papers for Long Document Summarization
20. Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization

E.5 EMNLP 2024

1. LongEmbed: Extending Embedding Models for Long Context Retrieval
2. Forgetting Curve: A Reliable Method for Evaluating Memorization Capability for Long-Context Models
3. LUQ: Long-text Uncertainty Quantification for LLMs
4. Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA
5. CItruS: Chunked Instruction-aware State Eviction for Long Sequence Modeling
6. Attribute or Abstain: Large Language Models as Long Document Assistants
7. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems
8. AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies
9. Where am I? Large Language Models Wandering between Semantics and Structures in Long Contexts
10. FinDVer: Explainable Claim Verification over Long and Hybrid-content Financial Documents
11. LONGAGENT: Achieving Question Answering for 128k-Token-Long Documents through Multi-Agent Collaboration

12. Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP
13. SEGMENT+: Long Text Processing with Short-Context Language Models
14. One Thousand and One Pairs: A "Novel" Challenge for Long-Context Language Models
15. LLoCO: Learning Long Contexts Offline
16. Enhancing Post-Hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition
17. Memorize Step by Step: Efficient Long-Context Prefilling with Incremental Memory and Decremental Chunk
18. LongRAG: A Dual-perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering
19. Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts
20. LongWanjuan: Towards Systematic Measurement for Long Text Quality
21. LongHeads: Multi-Head Attention is Secretly a Long Context Processor
22. Insights into LLM Long-Context Failures: When Transformers Know but Don't Tell
23. LongGenBench: Long-context Generation Benchmark
24. Can't Remember Details in Long Documents? You Need Some R&R
25. GraphReader: Building Graph-based Agent to Enhance Long-Context Abilities of Large Language Models
26. More Bang for Your Context: Virtual Documents for Question Answering over Long Documents
27. LongAlign: A Recipe for Long Context Alignment of Large Language Models
28. Long Sequence Modeling with Attention Tensorization: From Memory-Efficient Design to Long-context OpenQA
29. LSM1K: Large Scale Memory-based Dataset for Long Text Modeling

E.6 NAACL 2024

1. RST-LoRA: A Discourse-Aware Low-Rank Adaptation for Long Document Abstractive Summarization
2. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks
3. Effective Long-Context Scaling of Foundation Models
4. WESOME: GPU Memory-constrained Long Document Summarization using Memory Mechanism and Global Salient Content

E.7 Long-output

1. Suri: Multi-constraint Instruction Following in Long-form Text Generation
2. LongGenBench: Long-context Generation Benchmark

F Broken Mirror Test

We conducted a simple test, and almost all models (OpenAI, 2024c; Dubey et al., 2024; OpenAI, 2024a; Reid et al., 2024; GLM et al., 2024) failed to correctly answer the Prompt-long (only O1 (OpenAI, 2024b) was able to answer correctly, but not stable). We identified logical errors in handling long text cases, while the models were almost always able to correctly respond to the Prompt-short.

Prompt-long (3089 words)

This is a fairy tale. Please carefully consider whether there are any logical issues in the following story. If there are, point out just one significant logical flaw?

Once upon a time in midwinter, when the snowflakes were falling like feathers from heaven, a queen sat sewing at her window, which had a frame of black ebony wood. As she sewed she looked up at the snow and pricked her finger with her needle.

.....

.....

Mirror, mirror, on the wall, Who in this land is fairest of all? To this the mirror answered: You, my queen, are fairest of all.

.....

.....

With that, she slammed the mirror with all her strength, shattering it into pieces, her fury burning in her heart. The mirror was instantly destroyed, unable to speak again.

.....

.....

Back at home she asked her mirror: Mirror, mirror, on the wall, Who in this land is fairest of all? **It finally answered: You, my queen, are fairest of all.** ✗

.....

.....

Prompt-short (390 words)

This is a fairy tale. Please carefully consider whether there are any logical issues in the following story. If there are, point out just one significant logical flaw?

Once upon a time in midwinter, when the snowflakes were falling like feathers from heaven, a queen sat sewing at her window, which had a frame of black ebony wood. As she sewed she looked up at the snow and pricked her finger with her needle.

.....

Mirror, mirror, on the wall, Who in this land is fairest of all? To this the mirror answered: You, my queen, are fairest of all.

When the queen heard the mirror say this, she shook and trembled with anger, shouting, "Snow-White must die."

With that, she slammed the mirror with all her strength, shattering it into pieces, her fury burning in her heart. The mirror was instantly destroyed, unable to speak again.

"Tell me, mirror," she said, her voice softer now, "will I always be the fairest in the land?"

"True beauty comes not from the surface, but from the soul," the mirror answered. "What you seek is fleeting. Find peace, and you will see a beauty that lasts beyond your own." ✗

The Queen's eyes hardened again. "I do not seek peace. I seek power."

And so, with a cold smile, she turned away from the mirror, her heart set on a path that would lead her further from the light.

G Current model in Long-output Benchmark

The Fig 9 is from Longwrite (Bai et al., 2024d), and it reveals that most existing LLMs are unable to meet the output requirements for long instructions.

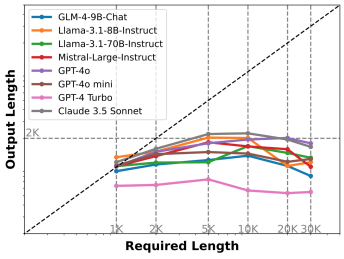

Figure 9: LongWriter-Ruler test demonstrates a maximum output length limitation of approximately 2k words for all models tested.