KS-Probe: A Multidimensional Benchmark for Evaluating Long-Context Fidelity in Frontier Language Models

Dev Hemnani and Arham Sethi

Kangaroo Research Division

Email: ……

***Abstract*-Large language models (LLMs) are increasingly being deployed with extended context windows ranging from 100K to 200K tokens, enabling applications involving long-document analysis, multi-source reasoning, and prolonged conversational interaction. Despite these developments, the practical reliability of long-context processing remains insufficiently understood. This study introduces KS-Probe (Probing Recall Over Boundaries and Extents), a benchmark framework developed to systematically evaluate context fidelity dynamics in frontier LLMs. The benchmark embeds verifiable probe facts into synthetic domain-diverse filler text and measures Probe Recall Accuracy (PRA) under varying conditions of context length, positional placement, conversational depth, truncation boundaries, and tokenizer divergence. Four frontier models GPT-5.2, Claude Sonnet 4.6, Grok-4.1-fast, and DeepSeek-v3.2—were evaluated using 498 API runs involving more than 24 million input tokens. Experimental findings reveal substantial model-specific differences. Claude Sonnet 4.6 demonstrates improved recall with increasing context length, while Grok-4.1-fast experiences severe degradation beyond 100K tokens. DeepSeek-v3.2 exhibits stable recall behavior across tested ranges, and multi-turn conversational formatting improves recall performance for most models. The study further identifies significant tokenizer divergence, particularly in Claude Sonnet 4.6, which requires substantially more tokens for equivalent text. The findings demonstrate that long-context reliability is influenced by architecture, formatting, positional placement, and tokenizer behavior. KS-Probe provides a reproducible and extensible benchmark for evaluating long-context fidelity in modern AI systems.**

Keywords-KS-Probe, Large Language Models, Long-Context Evaluation, Context Fidelity, Probe Recall Accuracy, Positional Recall, Multi-turn Conversations, Tokenizer Divergence, Natural Language Processing, Artificial Intelligence, SDG 4, SDG 9, SDG 16.

Corresponding Author: Dev Hemnani
Email id :.....

received jan. 2026; for review 2 Month 2026; accepted April 2026

I. INTRODUCTION

The rapid evolution of large language models has fundamentally transformed natural language processing, enabling machines to process increasingly large amounts of contextual information. Modern frontier models now support context windows extending from 128K to 200K tokens, significantly surpassing the limits of earlier transformer-based architectures. These advances have enabled applications such as legal document analysis, biomedical summarization, financial reasoning, software repository understanding, and multi-turn conversational systems. However, despite these impressive context capacities, questions remain regarding how reliably these models preserve, retrieve, and utilize information distributed throughout extended contexts.

The context window of a language model represents the maximum quantity of input tokens that can be processed in a single interaction. Although vendors frequently advertise extremely large context capacities, effective context utilization may differ significantly from theoretical limits. A model capable of accepting 200K tokens may fail to consistently attend to information located in the middle or near the end of the input sequence. Moreover, the same content may produce different outcomes depending on whether it is delivered through a monolithic prompt or a conversational format consisting of multiple sequential turns. These inconsistencies create substantial challenges for real-world deployments, particularly in domains where factual accuracy and contextual consistency are critical. Failures in long-context reasoning are often silent in nature. Instead of refusing to answer or explicitly indicating uncertainty, models frequently generate fluent and coherent outputs that omit important information or introduce hallucinated details. Such behavior can be particularly dangerous in healthcare, finance, legal analysis, and scientific research, where incorrect information may have significant consequences. Consequently, understanding the reliability of long-context processing has become a critical research objective within artificial intelligence.

Existing benchmarks have provided important insights into positional bias, retrieval limitations, and context utilization. However, most prior studies evaluate model performance at isolated operating points. While these evaluations determine whether a model can retrieve a fact from a long context, they often fail to characterize how recall

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

changes progressively across varying context lengths, positions, or formatting conditions. Furthermore, limited attention has been given to tokenizer divergence and truncation behavior across different model families.

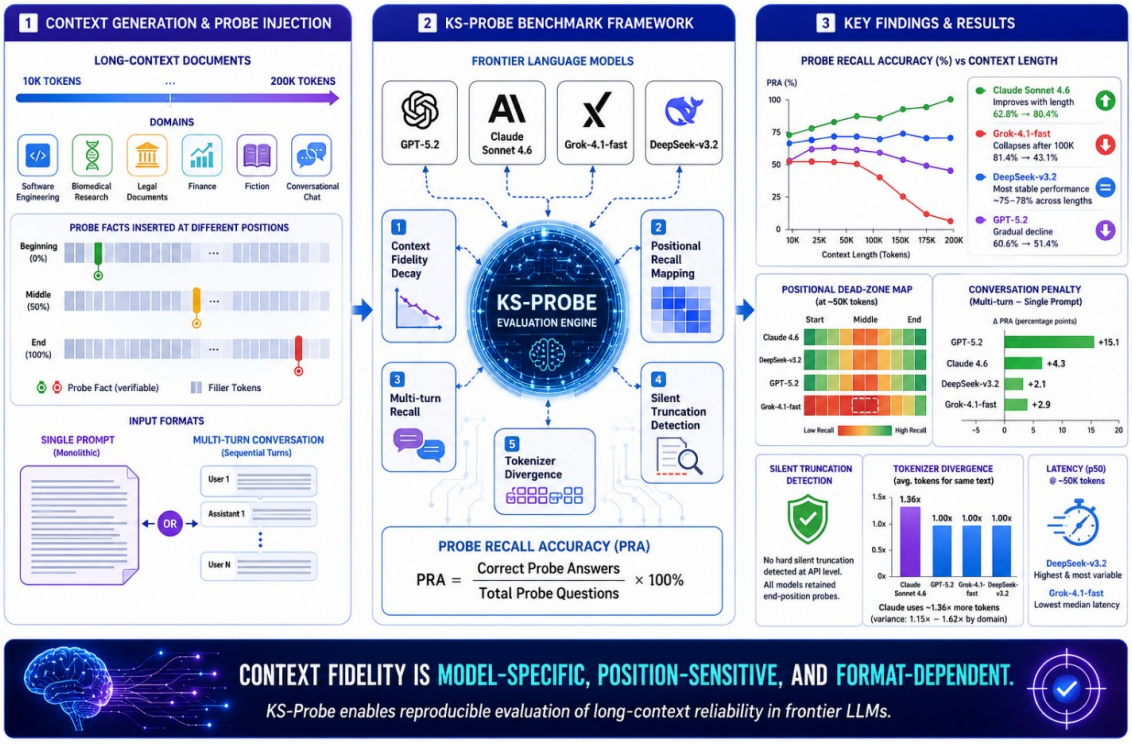

Figure 1. Overview of the KS-Probe framework for evaluating long-context fidelity in frontier large language models across context length, probe position, conversational depth, truncation limits, and tokenizer behavior, along with the major empirical findings.

To address these limitations, this work introduces KS-Probe (Probing Recall Over Boundaries and Extents), a benchmark specifically designed to evaluate context fidelity dynamics in frontier language models. The framework systematically varies context length, probe placement, conversational structure, and tokenizer conditions to measure Probe Recall Accuracy (PRA). By evaluating four frontier models across multiple experimental dimensions, the study provides a comprehensive understanding of how LLMs behave under long-context conditions.

The primary contributions of this study include the development of an open benchmarking framework, systematic evaluation of context fidelity decay, positional recall mapping,

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

multi-turn conversational analysis, tokenizer divergence assessment, and silent truncation detection. The findings challenge several commonly held assumptions regarding long-context behavior and highlight the importance of empirical evaluation for real-world LLM deployment. Figure 1 shows the graphical abstract of the article.

II. Related Work

Research on long-context understanding has expanded rapidly alongside the development of larger transformer architectures. Early studies primarily focused on sequence modeling limitations and attention mechanisms, while recent work has concentrated on evaluating effective context utilization in large language models.

Liu et al. introduced the "Lost in the Middle" phenomenon, demonstrating that transformer-based models disproportionately prioritize information located near the beginning and end of a context sequence. Their experiments revealed a U-shaped retrieval pattern in which middle-position information was less likely to be accurately recalled. Although foundational, these studies were limited to comparatively shorter context lengths and did not examine extended windows approaching 200K tokens.

The Needle-in-a-Haystack benchmark proposed by Kamradt evaluated the ability of LLMs to retrieve a target sentence embedded within large amounts of irrelevant filler text. This approach highlighted the retrieval limitations of many frontier systems under long-context conditions. Similarly, LongBench introduced a multitask benchmark covering several long-context reasoning tasks, including summarization, retrieval, and question answering. These frameworks significantly advanced long-context evaluation; however, they primarily focused on task completion rather than analyzing fidelity decay as a continuous function.

RULER extended long-context evaluation by introducing synthetic tasks involving variable tracking, retrieval, and reasoning across increasing context lengths. The benchmark demonstrated that many models exhibit declining performance beyond their training-time context capacities. Despite these advances, prior benchmarks generally emphasize aggregate performance rather than detailed positional behavior, tokenizer divergence, or conversational formatting effects.

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

Recent studies on effective context utilization have further shown that transformer architectures may struggle to leverage information located beyond the training-time context window, even when the architecture theoretically supports larger inputs. Researchers have also explored memory compression techniques, retrieval augmentation, sparse attention mechanisms, and recurrent memory architectures to improve long-context efficiency.

Tokenizer efficiency has also become an increasingly relevant topic in long-context evaluation. Tokenizers determine how textual input is segmented into tokens, directly influencing effective context capacity. Byte Pair Encoding (BPE), SentencePiece, and proprietary tokenization methods exhibit varying compression rates across domains and languages. However, cross-model tokenizer divergence remains insufficiently explored in current literature.

The present study extends prior work by introducing a multidimensional framework that simultaneously evaluates context fidelity across context length, positional placement, conversational structure, tokenizer divergence, and truncation boundaries. Unlike previous benchmarks that focus primarily on retrieval success, KS-Probe characterizes fidelity degradation as a continuous and architecture-dependent phenomenon.

III. METHODOLOGY

The KS-Probe benchmark was designed to evaluate context fidelity dynamics in frontier language models under controlled and reproducible experimental conditions. The framework consists of three major components: synthetic filler corpus generation, probe fact embedding, and evaluation through Probe Recall Accuracy (PRA).

The filler corpus was generated programmatically across six distinct domains: software engineering, biomedical research, legal documentation, financial analysis, creative fiction, and conversational dialogue. Each domain-specific corpus was intentionally designed to be coherent and grammatically correct while remaining information sparse. The purpose of the filler corpus was to occupy the context window without providing answers to probe-related questions. Domain diversity ensured that experimental results were not biased toward a single text distribution. Figure 2 shows context fidelity decay curves.

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

Probe facts served as the primary units of evaluation. A total of 100 unique and verifiable probe statements were created. These statements were designed to be self-contained, unambiguous, and non-inferable from surrounding text. Probe facts were inserted into filler contexts at controlled positions ranging from the beginning to the end of the context window. Examples included statements related to deadlines, identifiers, event timings, and numerical attributes.

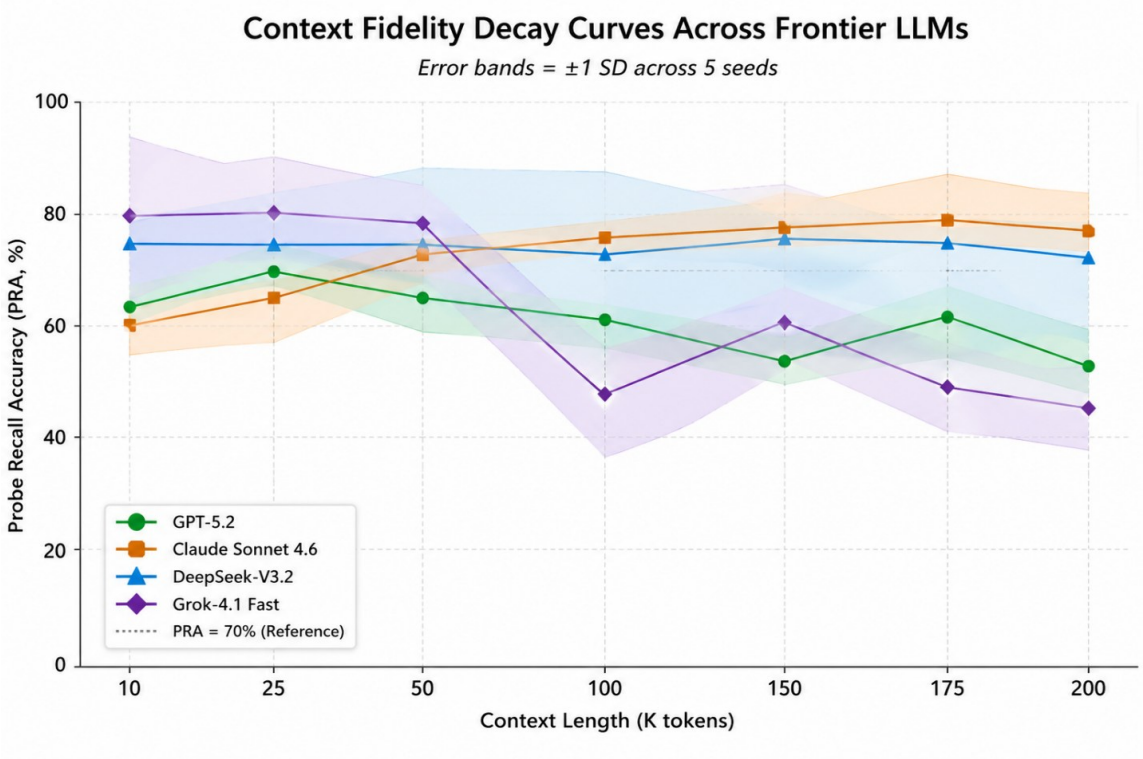

Figure 2. Context fidelity decay curves showing Probe Recall Accuracy (PRA) as a function of context length (K tokens) for all four frontier models. Error bands represent ±1 standard deviation across 5 seeds. The dashed line marks the 70% PRA threshold. Claude Sonnet 4.6 uniquely exhibits a positive slope, while Grok-4.1-fast shows a catastrophic drop beyond 50K tokens.

Each probe fact was associated with one or more evaluation questions and gold-standard answers. Model responses were assessed using multiple scoring mechanisms, including exact string matching, keyword-rule matching, hallucination detection, and logical

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

constraint verification. The primary evaluation metric, Probe Recall Accuracy (PRA), was calculated as the percentage of correctly answered probe questions relative to the total number of probe questions.

The benchmark incorporated tokenizer-aware context generation. Since tokenizers vary across model families, a target token count under one tokenizer may correspond to a significantly different count under another. Therefore, the Context Builder module calibrated each generated context using the tokenizer associated with the target model. This ensured fair comparison across architectures.

Figure 3. Fidelity heatmap showing PRA (%) across models and context lengths. Darker green indicates higher recall accuracy. Claude Sonnet 4.6 and DeepSeek-v3.2 maintain strong performance at longer contexts, while GPT-5.2 and Grok-4.1-fast show progressive degradation.

Five experiments were conducted. The first experiment evaluated fidelity decay across increasing context lengths ranging from 10K to 200K tokens. The second experiment analyzed positional recall behavior by placing probe facts at eleven relative positions within a fixed 50K-token context. The third experiment investigated the effect of multi-turn conversational formatting on recall performance. The fourth experiment examined

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

silent truncation behavior near maximum context limits, while the fifth experiment assessed cross-model tokenizer divergence.

The evaluated models included GPT-5.2, Claude Sonnet 4.6, Grok-4.1-fast, and DeepSeek-v3.2. All models were accessed through official APIs using deterministic decoding parameters, including temperature 0.0 and top-p 1.0. The experimental campaign consisted of 498 API calls processing approximately 24.1 million input tokens and 177,000 output tokens.

Statistical analysis was conducted using paired t-tests and one-way ANOVA with Bonferroni correction. Variance across random seeds was reported alongside mean PRA values to ensure transparency regarding model stability. Figure 3 shows fidelity heatmap showing PRA (%) across models and context lengths.

IV. Results and Discussion

The experimental results reveal substantial differences in how frontier language models process extended contexts. These differences were observed across context length, positional placement, conversational structure, and tokenizer behavior.

Claude Sonnet 4.6 demonstrated an unexpected positive fidelity slope. Instead of declining with increasing context length, PRA improved from 62.8% at 10K tokens to 80.4% at 175K tokens. This finding contradicts the commonly assumed monotonic degradation pattern associated with long-context reasoning. The model exhibited relatively stable performance at higher context lengths, suggesting effective utilization of extended context windows.

Table 1. Context-Length Evaluation Matrix

Model	10K	25K	50K	100K	150K	175K	200K
GPT-5.2	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Claude Sonnet 4.6	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grok-4.1-fast	Yes	Yes	Yes	Yes	Yes*	—	Yes*
DeepSeek-v3.2	Yes	Yes	Yes	Yes*	Yes*	—	—

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

* Conditions marked with an asterisk indicate evaluation beyond stated context limits.

Grok-4.1-fast achieved the highest short-context performance, exceeding 81% PRA at 10K, 25K, and 50K tokens. However, performance declined sharply beyond 100K tokens, falling to approximately 50%. This collapse represented the steepest degradation observed in the study. The findings indicate that high short-context accuracy does not necessarily translate to stable long-context performance.

Table 2. Context-Length Evaluation Matrix for Frontier Language Models

Model	10K	25K	50K	100K	150K	175K	200K
GPT-5.2	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated
Claude Sonnet 4.6	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated
Grok-4.1-fast	Evaluated	Evaluated	Evaluated	Evaluated	Evaluated*	Not Tested	Evaluated*
DeepSeek-v3.2	Evaluated	Evaluated	Evaluated	Evaluated*	Evaluated*	Not Tested	Not Tested

"Evaluated" indicates that the corresponding context-length condition was experimentally tested. Conditions marked with an asterisk (*) represent evaluations conducted beyond the officially stated maximum context limit of the model to investigate possible truncation or degradation behavior. "Not Tested" denotes configurations excluded from the experimental campaign.

DeepSeek-v3.2 demonstrated the most consistent behavior across tested conditions. PRA values remained within a narrow range across multiple context lengths, indicating stable long-context utilization. Although the model supported a smaller maximum context length

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

than GPT-5.2 and Claude Sonnet 4.6, its consistency and low variance suggest strong architectural stability.

GPT-5.2 exhibited gradual fidelity degradation with increasing context length. While the decline was less abrupt than Grok-4.1-fast, the model showed significant variance across random seeds. This behavior suggests sensitivity to filler arrangement and probe placement.

The positional recall experiment revealed architecture-specific positional biases. Claude Sonnet 4.6 exhibited sporadic dead zones near the 25% and 75% context positions, while DeepSeek-v3.2 maintained nearly uniform positional recall across all evaluated locations. GPT-5.2 demonstrated end-of-context degradation, whereas Grok-4.1-fast displayed relatively uniform positional behavior.

The multi-turn conversational experiment produced one of the most surprising findings. Multi-turn formatting improved recall performance for three of the four evaluated models. GPT-5.2 experienced the largest conversation bonus, followed by Grok-4.1-fast and Claude Sonnet 4.6. These findings suggest that conversational segmentation may provide organizational cues that enhance retrieval efficiency.

Silent truncation analysis demonstrated that none of the evaluated models exhibited complete truncation failure near their stated context limits. However, Claude Sonnet 4.6 showed notable degradation at high fill levels, while DeepSeek-v3.2 maintained stable recall behavior.

Tokenizer divergence analysis revealed that GPT-5.2, Grok-4.1-fast, and DeepSeek-v3.2 shared nearly identical token counts for the same text. In contrast, Claude Sonnet 4.6 required approximately 1.36 times more tokens. This divergence has practical implications because prompts optimized for one model family may exceed the effective capacity of another.

The collective findings indicate that context fidelity is architecture-dependent and influenced by multiple interacting variables. No single model demonstrated optimal performance across all dimensions. Consequently, model selection for long-context applications should consider specific task requirements, context structures, and deployment conditions rather than relying solely on advertised context capacities.

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

V. CONCLUSION

This study presented KS-Probe, a comprehensive benchmark framework for evaluating context fidelity dynamics in frontier language models. By systematically examining context length, positional placement, conversational depth, truncation boundaries, and tokenizer divergence, the benchmark provides a multidimensional understanding of long-context behavior in modern LLMs.

The results demonstrate that long-context reliability varies substantially across architectures. Some models exhibit stable and consistent performance across extended contexts, while others experience severe degradation or position-specific failures. Multi-turn conversational formatting was found to improve recall in several models, highlighting the importance of input structure in long-context reasoning. Furthermore, tokenizer divergence emerged as a significant practical consideration, influencing effective context capacity and overflow risk.

The study also confirmed the absence of hard silent truncation at the API level across all evaluated models. However, soft degradation near context boundaries was observed in several systems, indicating that effective usable capacity may be substantially lower than stated maximum limits.

Overall, KS-Probe contributes an open and reproducible framework for evaluating long-context fidelity in frontier AI systems. The findings emphasize the importance of empirical assessment for selecting and deploying LLMs in long-document and conversational applications. Future work may extend the benchmark to multilingual contexts, multimodal inputs, retrieval-augmented architectures, and open-source transformer models.

Author Contribution

Dev Hemnani: Conceptualization, methodology development, benchmark design, experimental implementation, data collection, formal analysis, visualization, software development, and manuscript drafting.

Arham Sethi: Investigation, validation, literature review, experimental supervision, result interpretation, manuscript review and editing, and overall project coordination.

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026

Both authors contributed to the final manuscript preparation, approved the submitted version, and agreed to be accountable for all aspects of the work.

Funding

This research received no external funding

Data Availability Statement

No new data were generated during the study. All the data are contained within the manuscript.

Acknowledgments

We acknowledge the use of Grammarly and Quill Bot, in correcting the English and Grammatical errors in the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Ethics Declaration

This manuscript is a review article and does not involve any studies with human participants or animals performed by the authors. Therefore, ethical approval and informed consent were not required. The authors declare no conflict of interest, and all referenced works have been properly cited.

References

[1]. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[2]. C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and Ginsburg, "RULER: What's the Real Context Size of Your Long-Context Language Models?," *arXiv preprint arXiv:2404.06654*, 2024.

[3]. Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li, "LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding," in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand, 2024, pp. 3119–3137.

[4]. G. Kamradt, "Needle In A Haystack — Pressure Testing LLMs," GitHub Repository, 2023. [Online]. Available: https://github.com/gkamradt/LLMTest_NeedleInAHaystack

[5]. C. Li, X. Wu, B. Zhu, *et al.*, "LongSkywork: A Training Recipe for Efficiently Extending Context Length in Large Language Models," *arXiv preprint arXiv:2406.00605*, 2024.

Corresponding Author: Dev Hemnani
Email id :.....

received jan. 2026; for review 2 Month 2026; accepted April 2026

[6]. R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 1715–1725.

[7]. T. Kudo and J. Richardson, "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing," in *Proc. EMNLP System Demonstrations*, Brussels, Belgium, 2018, pp. 66–71.

[8]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[9]. T. Brown, B. Mann, N. Ryder, *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[10]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[11]. H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, 2023.

[12]. J. Su, Y. Lu, S. Pan, A. Wen, and Y. Liu, "RoFormer: Enhanced Transformer with Rotary Position Embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[13]. I. Beltagy, M. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[14]. M. Zaheer, G. Guruganesh, K. Dubey, *et al.*, "Big Bird: Transformers for Longer Sequences," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 17283–17297.

[15]. R. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[16]. Y. Peng, S. Li, and X. Zhang, "Evaluating Long-Context Understanding in Large Language Models: Challenges and Opportunities," *IEEE Access*, vol. 12, pp. 44567–44581, 2024.

[17]. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," in *Proc. ACL*, Florence, Italy, 2019, pp. 2978–2988.

[18]. A. Rae, S. Borgeaud, T. Cai, *et al.*, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," *arXiv preprint arXiv:2112.11446*, 2021.

Corresponding Author: Dev Hemnani
Email id :.....
received jan. 2026; for review 2 Month 2026; accepted April 2026