CROSS: ANALYZING THE TRADE-OFFS IN LONG-CONTEXT CROSS-LINGUAL RETRIEVAL

Sina Bagheri Nezhad & Ameeta Agrawal
Department of Computer Science
Portland State University, USA
{sina5, ameeta}@pdx.edu

ABSTRACT

Cross-lingual information retrieval in long-context settings faces challenges such as the "lost-in-the-middle" phenomenon and computational inefficiencies. We introduce CROSS (Cross-lingual Retrieval Optimized for Scalable Solutions), a two-phase retrieval framework that integrates multilingual embeddings with efficient candidate selection to enhance retrieval-augmented generation (RAG). Evaluating CROSS on the newly developed mLongRR-V2 benchmark—covering seven languages and 49 language pairs—we demonstrate substantial improvements in retrieval accuracy, scalability to 512,000-token contexts, and robustness across linguistic structures. Compared to baseline large language models (LLMs), CROSS significantly mitigates mid-context retrieval failures while reducing computational overhead. Our results establish CROSS as an efficient and scalable solution for multilingual long-context retrieval.

1 INTRODUCTION

The increasing need for multilingual information retrieval in today's globalized environment has brought to light several challenges, particularly when handling extremely long documents, mid-context target information, and diverse linguistic structures. In such settings, retrieval methods often suffer from a "lost-in-the-middle" phenomenon, where relevant details embedded deep within lengthy texts are overlooked, and from difficulties in generalizing across varied language pairs and domains. This work focuses on addressing the following core research questions:

1. **RQ1:** How can retrieval methods be designed to effectively capture and utilize long-context information across languages while balancing computational efficiency and scalability?
2. **RQ2:** How can the "lost-in-the-middle" problem be mitigated in long-context retrieval without sacrificing model expressiveness or adaptability to diverse linguistic domains?
3. **RQ3:** What scalable and robust frameworks can enhance context retrieval in scenarios with lengthy inputs, especially for low-resource languages and heterogeneous data settings?

Recent works have made promising strides toward these questions. For instance, Liu et al. (2023) and Xu et al. (2024a) highlight the degradation in performance when target information is located mid-context, while studies such as Hengle et al. (2024) reveal that even models with extended context windows (up to 8,000 tokens) suffer from reduced retrieval accuracy. Other approaches, including the parameter-efficient re-ranking method by Litschko et al. (2022), XAMPLER Lin et al. (2024), and OPTICAL for low-resource languages Huang et al. (2023), address aspects of computational efficiency or multilingual adaptation but leave open the challenge of building a unified, scalable solution that is robust to both long contexts and mid-context target positioning. Additionally, Yang et al. (2024) expose inherent trade-offs when using probabilistic structured queries for cross-lingual retrieval, underscoring the need for improved evaluation frameworks and scalable methodologies.

To tackle these challenges, we propose **CROSS** (Cross-lingual Retrieval Optimized for Scalable Solutions), a novel framework that integrates a two-phase retrieval mechanism with robust needle-positioning strategies. In the first phase, CROSS segments the extensive input context into manageable units and employs a multilingual embedding model to filter and rank these segments based

on semantic relevance. In the second phase, a language model processes only the top-ranked segments—irrespective of their position in the document—to extract or reason over the target information.

Recognizing the importance of cross-lingual capability, our evaluation emphasizes a wide spectrum of language combinations. We introduce mLongRR-V2, a benchmark dataset spanning seven languages (covering both Latin and non-Latin scripts) and varying context lengths. mLongRR-V2 is designed with 49 language combinations in mind, of which 42 are cross-lingual pairs—where the query language differs from the context language—and 7 are same-lingual pairs.

Our evaluation protocol covers both *single-target* (1-needle) scenarios—where a single target sentence is embedded within a monolingual context—and *multi-target* (3-needles) scenarios, in which multiple target sentences are present and the model must reason over them (e.g., by identifying the largest value among several candidates). These tests are conducted across different needle positions, context lengths, and sentence cap sizes to provide a comprehensive assessment of retrieval accuracy, scalability, and robustness. In addition, we perform a comprehensive failure analysis to identify the root causes of errors, distinguishing between failures in the embedding retrieval stage and those in the language model's reasoning process, particularly in complex multi-target scenarios.

2 Related works

The field of cross-lingual information retrieval (CLIR) has evolved with advancements in multilingual embeddings and transformer-based architectures.

Recent frameworks, such as LONGEMBED Zhu et al. (2024) and LongRAG Jiang et al. (2024), extend context windows, enhancing retrieval for lengthy documents. However, these primarily focus on monolingual or limited multilingual tasks. Approaches like DR-RAG Hei et al. (2024) and McCrolin Limkonchotiwat et al. (2024) add dynamic relevance scoring and multi-task learning to improve cross-lingual performance but are computationally intensive, limiting scalability.

Existing datasets have supported CLIR research but with notable gaps. mLongRR Agrawal et al. (2024) and BordIRlines Li et al. (2024) are limited in linguistic diversity and context length. While newer benchmarks like LONGEMBED Zhu et al. (2024) and DEBATEQA Xu et al. (2024b) address long-context evaluation, they fall short in comprehensive cross-lingual testing.

Despite progress, CLIR models still face challenges in long-context scenarios, often losing accuracy mid-document and struggling with diverse languages. Many solutions remain computationally expensive and lack the scalability needed for real-world multilingual applications.

3 Methodology

3.1 CROSS Framework

The CROSS framework (Cross-lingual Retrieval Optimized for Scalable Solutions) efficiently extracts "needles" of relevant information from extensive, multilingual "haystacks." Using a two-phase approach, CROSS improves retrieval accuracy, ensures cost efficiency, and overcomes the limitations of current models in handling long, cross-lingual contexts.

3.1.1 Two-Phase Retrieval Mechanism

CROSS employs a Retrieval-Augmented Generation (RAG) framework that leverages a two-phase retrieval process to enhance precision while minimizing computational overhead.

Phase 1: Tokenization and Embedding The context—potentially comprising hundreds of thousands of words in multiple languages—is segmented into sentences using the Punkt tokenizer Kiss & Strunk (2006). Each sentence is then embedded using the multilingual "bge-m3" model Chen et al. (2024), which effectively captures semantic nuances across languages (see Appendix A for more details about embedding model). Although operating at the sentence level might seem computationally demanding, particularly if one considers finer granularity, our cost analysis (see Part 4.6) demonstrates that the expense of embedding and retrieval is negligible relative to the cost of LLM processing.

Phase 2: Candidate Selection and Model Input Within this RAG framework, CROSS calculates the semantic distance between each sentence embedding and the query, selecting the top k most relevant sentences based on a tunable hyperparameter. In our experiments, we evaluated k values of 3, 5, 10, 20, and 50. These selected sentences are then passed as concise, contextually rich inputs to the language model (e.g., GPT-4o-mini or Llama 3.2 90b) for final answer extraction. This design ensures that, despite the additional embedding and retrieval steps, the overall token processing by the LLM is drastically reduced, preserving both accuracy and efficiency.

3.1.2 Efficiency and Model Independence

CROSS is model-independent, enhancing retrieval accuracy with any language model used in Phase 2. Tested with GPT-4o-mini and Llama 3.2, it dynamically adjusts the number of retrieved sentences, ensuring consistent, cost-effective performance. By focusing on the most relevant context segments, CROSS avoids attention drop-offs in long texts and maximizes precision. Its fixed input length makes it scalable, effectively handling document lengths far beyond the native context limits of most language models.

3.2 Benchmark Dataset: mLongRR-V2

To evaluate CROSS, we introduce mLongRR-V2, an extended version of the mLongRR benchmark Agrawal et al. (2024). This dataset:

- Covers seven languages spanning Latin and non-Latin scripts.
- Includes 49 language pairs, with 42 cross-lingual settings.
- Extends context lengths up to 512,000 tokens.
- Evaluates retrieval using single-target (1-needle) and multi-target (3-needles) tasks.

A full dataset description is provided in Appendix B.

3.3 Evaluation Protocol

We assess CROSS using two retrieval scenarios:

- 1-Needle Retrieval: A single target sentence is embedded in a long document, and the model must retrieve the exact phrase.
- 3-Needles Retrieval: Three target sentences are embedded at different positions, requiring the model to reason over multiple candidates.

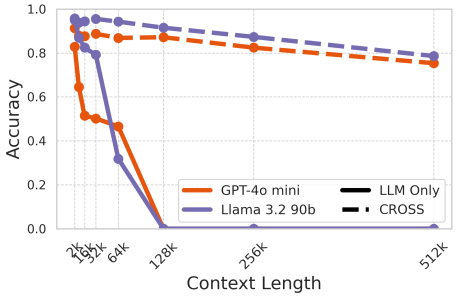
The primary evaluation metric is retrieval accuracy, defined as the percentage of cases where the correct target sentence is retrieved. To further assess performance, we analyze:

- Scalability: Accuracy across varying document lengths.
- Needle Position Sensitivity: Performance at different insertion points.
- Cross-lingual Robustness: Accuracy across diverse language pairs.
- Computational Efficiency: Reduction in token processing relative to full-document models.

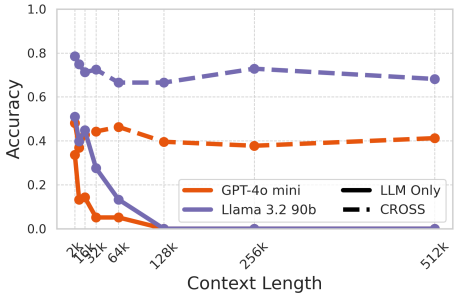
A more detailed breakdown of experimental settings, and evaluation conditions is available in Appendix C.

4 Results and Analysis

This section presents the experimental results of our approach, which combines CROSS's retrieval framework with Llama 3.2 and GPT-4o-mini as the underlying language model. We analyze its performance across retrieval accuracy, robustness to context length, needle position sensitivity, cross-lingual consistency, and cost efficiency, comparing it to the baseline performance of LLMs alone.



(a) Comparison of Retrieval Accuracy Across Context Lengths in 1-Needle Test



(b) Comparison of Retrieval Accuracy Across Context Lengths in 3-Needles Test

Figure 1: Retrieval Accuracy Across Different Context Lengths

4.1 Initial LLM Performance Evaluation without "Contexts"

Before integrating CROSS, we tested both GPT-4o-mini and Llama 3.2 90b independently to verify their ability to understand the prompts and correctly retrieve or reason answers without any provided context. This evaluation was conducted for both the 1-needle (retrieval) and 3-needles (reasoning) scenarios. Each model was tested 10 independent times using prompts and needles alone, without contextual interference. Both models demonstrated flawless performance, successfully identifying the correct answers in all tests. These results confirm that the prompts are clear and fully understandable to the LLMs, establishing a solid foundation for evaluating the impact of CROSS in more complex retrieval scenarios.

4.2 Retrieval Accuracy

CROSS achieved significant improvements in retrieval accuracy across all tested languages and language pairs when paired with both GPT-4o-mini and Llama 3.2 90b. Compared to using the language models alone, the CROSS-enhanced approach consistently retrieved the target sentence with higher exact match accuracy, especially in long contexts and complex cross-lingual pairs.

On average, across all 49 language combinations, CROSS with GPT-4o-mini achieved a retrieval accuracy of **87%**, significantly outperforming the baseline GPT-4o-mini, which achieved only **37%**. Similarly, CROSS with Llama 3.2 achieved a remarkable improvement, with accuracy increasing from **47%** for Llama 3.2 alone to **92%** when enhanced with CROSS.

Furthermore, for contexts under 64k words—the length supported by both models—CROSS-enhanced GPT-4o-mini maintained a retrieval accuracy of **88%**, compared to **59%** for GPT-4o-mini alone. Llama 3.2 also showed improvement under 64k words, with accuracy increasing from **75%** for the baseline model to **95%** with CROSS. These substantial improvements across both context lengths and models demonstrate CROSS's effectiveness in preserving high retrieval accuracy.

As illustrated in the radar graphs in Figure 11 in Appendix E, CROSS enhances retrieval performance across all prompt and context languages, indicating the robustness of this approach in varied multilingual scenarios. These results highlight the effectiveness of the CROSS framework in maintaining high accuracy across diverse linguistic contexts when paired with both GPT-4o-mini and Llama 3.2

4.3 Context Length Robustness

A key strength of CROSS is its robust performance across varied context lengths. Without CROSS, both models exhibit a notable decline in retrieval accuracy as context length increases, with sharp reductions observed beyond 64k words. In contrast, the CROSS-enhanced approach maintains consistent accuracy across context lengths up to 512k words for both models, showing only minimal reduction (Figure 1a). By narrowing the input to a fixed set of top-relevant sentences, CROSS effectively mitigates the typical accuracy drop-off associated with large contexts, enabling both GPT-4o-mini and Llama 3.2 to perform reliably on larger-scale retrieval tasks.

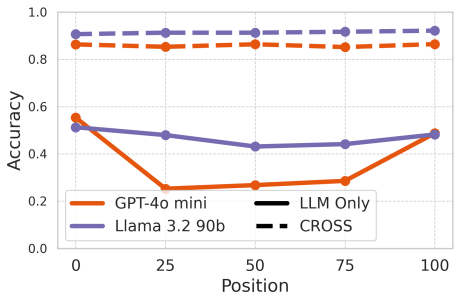


Figure 2: Retrieval Accuracy Comparison Across Needle Positions

Notably, this pattern persists in the more challenging 3-needles test. CROSS continues to stabilize retrieval accuracy across increasing context lengths for both models, as shown in Figure 1b, further emphasizing its robustness in scenarios with multiple target sentences.

4.4 Needle Position Sensitivity

To assess CROSS's effectiveness in addressing the "lost in the middle" issue, we measured retrieval accuracy across five needle positions (0%, 25%, 50%, 75%, and 100%). Both GPT-4o-mini and Llama 3.2 90b performed best when the needle was at the beginning (0%) or end (100%) of the context. However, GPT-4o-mini exhibited a significant drop in accuracy for mid-context positions (25%, 50%, 75%), with an average accuracy of only **27%**. Llama 3.2, being a newer model, handled mid-context positions better, achieving an average accuracy of **45%**, though it still showed a noticeable reduction compared to its performance at the boundaries.

When paired with CROSS, both models demonstrated a dramatic improvement in positional resilience. CROSS maintained stable performance across all needle positions, achieving an average mid-context accuracy of **86%** for GPT-4o-mini and **91%** for Llama 3.2. This indicates that CROSS effectively mitigates the loss in retrieval accuracy commonly associated with middle-positioned target information.

Notably, CROSS ensures consistent accuracy regardless of where the needle is located, addressing the challenges inherent in finding information deeply embedded within extensive contexts. This improvement underscores CROSS's ability to generalize effectively across models, resolving the "lost in the middle" problem even for a robust baseline like Llama 3.2 (Figure 2).

4.5 Cross-Lingual Consistency

Notably, CROSS demonstrates strong performance across linguistically dissimilar language pairs, such as Hindi-Russian, where the prompt and query are in Hindi and the context is in Russian. In these challenging cross-lingual scenarios, GPT-4o-mini alone exhibits a marked reduction in accuracy, while Llama 3.2 fares better. When paired with CROSS, both models maintain high retrieval accuracy, demonstrating robust consistency even across varied linguistic structures. Recent work has shown that cross-lingual effectiveness depends on intrinsic language features beyond data quantity (Bagheri Nezhad & Agrawal, 2024; Bagheri Nezhad et al., 2025), and CROSS's multilingual embeddings help capture these nuances.

For GPT-4o-mini, CROSS significantly boosts accuracy in cross-lingual pairs, bridging the gap between same-language and cross-language scenarios. Similarly, Llama 3.2 paired with CROSS achieves consistently strong performance, making CROSS a robust solution for multilingual applications. Figure 3 illustrates the performance of both models in the 1-needle test, comparing retrieval accuracy when the prompt and context languages are the same versus different. For the 3-needles test, a similar comparison is provided in Figure 4, highlighting CROSS's ability to maintain robust accuracy even in complex reasoning tasks.

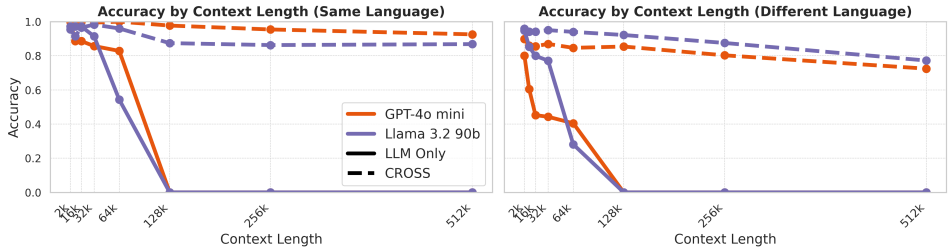


Figure 3: Comparison of Retrieval Accuracy in Same-Language and Cross-Lingual Settings in the 1-needle test.

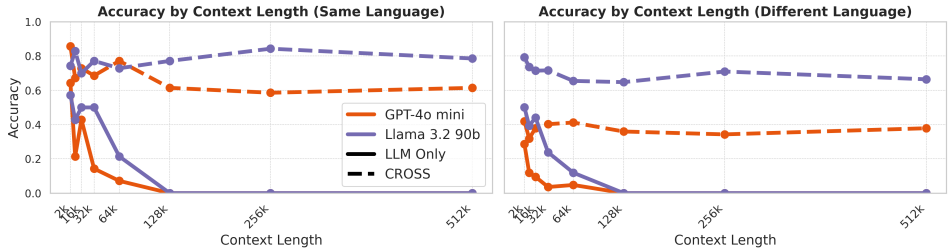


Figure 4: Comparison of Retrieval Accuracy in Same-Language and Cross-Lingual Settings in the 3-needles test.

4.6 Cost Efficiency Analysis

A key advantage of CROSS is its dramatic reduction in the number of tokens processed by the expensive language model. In a conventional setup, the entire context of T tokens is fed into the LLM, incurring a cost that scales linearly with T. In contrast, CROSS first processes the full context with a lightweight embedding model (BGE-M3) and then selects the top k sentences (each averaging roughly T/N tokens) to pass to the LLM. Thus, the LLM processes approximately

$$k \cdot \frac{T}{N} \quad \text{tokens},$$

instead of T tokens.

Assuming that the computational cost per token for the LLM is proportional to the model's parameter count, and noting that BGE-M3 (568M parameters) is roughly 160 times more efficient per token than Llama 3.2 (90B parameters), the cost incurred by the embedding stage is only a small fraction of that of the LLM. In our experiments, this two-phase approach resulted in an average reduction of token usage for the LLM by about 90% across various context lengths (see Figure 14 in Appendix E).

In addition to these costs, CROSS requires computing the semantic distances between the embedded query and each sentence's embedding. This involves a vector distance computation (typically a cosine or Euclidean similarity) for each of the N sentence embeddings. The computational cost of these distance calculations is generally:

$$\text{Cost}_{\text{distances}} \propto N \times d,$$

where d is the embedding dimension (e.g., 1024). In practice, since d is relatively small and these computations can be highly optimized (or even performed using approximate nearest-neighbor search techniques), the overall cost of the distance calculations is modest compared to the cost saved by significantly reducing the LLM's input size.

Thus, the overall computational cost of CROSS can be expressed as:

$$\text{Cost}_{\text{CROSS}} = T \cdot C_{\text{embed}} + \left(k \cdot \frac{T}{N}\right) \cdot C_{\text{LLM}} + N \cdot d \cdot C_{\text{dist}} \tag{1}$$

where:

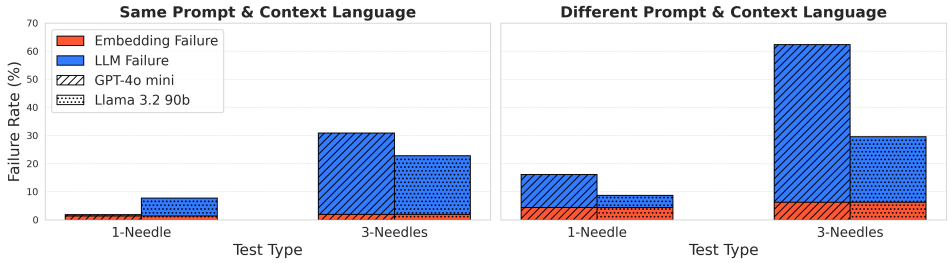


Figure 5: Failure rate in same vs different language

- C_{embed} is the per-token cost of the embedding model,
- C_{LLM} is the per-token cost of the LLM,
- C_{dist} is the per-dimension cost of computing distances.

Given that $C_{\text{embed}} \ll C_{\text{LLM}}$ and that the cost of the distance computations ($N \cdot d \cdot C_{\text{dist}}$) is relatively low, the overall efficiency gains are overwhelmingly driven by reducing the number of tokens fed into the LLM—an effect that is most pronounced when $k \ll N$.

5 Failure Analysis

While CROSS demonstrates significant improvements in retrieval accuracy, a closer examination of failure cases provides insights into its limitations, particularly in multi-target scenarios. This section analyzes the failure rates in both the 1-needle and 3-needles tests, distinguishing between failures arising in the embedding retrieval phase and those occurring within the language model's response generation.

We categorize retrieval failures into two types:

- **Embedding Failure**: Cases where the target label is absent from the retrieved sentence cap, indicating that the embedding model did not select the relevant sentences.
- **LLM Failure**: Cases where the language model fails to correctly extract or reason about the label, despite it being present in the retrieved sentence cap.

5.1 Failure Rates and Trends

Figure 6 illustrates the failure rates across different test scenarios. In the 1-needle test, both embedding and LLM failures remain relatively low. The embedding model correctly retrieves the relevant sentence in 96% of cases, with embedding failures accounting for only **4.0%**. Similarly, LLM failures remain low, at **10.1%** for GPT-4o-mini and **4.6%** for Llama 3.2.

However, in the 3-needles test, we observe a substantial increase in LLM failures. Although embedding failures remain marginal at **5.7%**, LLM failures escalate significantly. GPT-4o-mini exhibits a failure rate of **52.2%**, while Llama 3.2, though performing better, still registers a notable **22.9%** failure rate. This indicates that while CROSS reliably retrieves relevant sentences, the challenge in the 3-needles scenario primarily lies in the model's ability to reason over multiple retrieved labels and correctly extract the appropriate one.

5.2 Analysis of Increased LLM Failures in 3-Needles Test

The increased failure rate in the 3-needles test suggests that the presence of multiple target sentences creates ambiguity, making it more difficult for the language model to consistently select the correct answer. Possible contributing factors include:

- **Increased Distractors:** The presence of multiple similar sentences in the sentence cap introduces additional reasoning complexity, leading to incorrect selections.

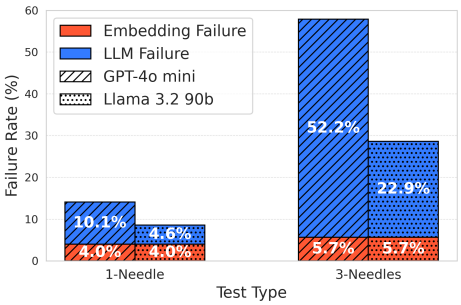


Figure 6: Failure rate analysis

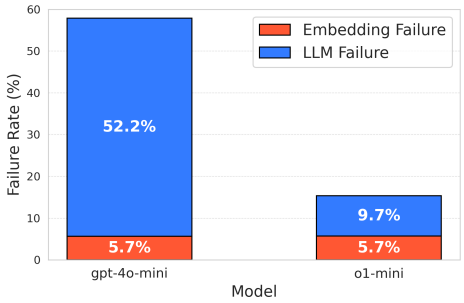


Figure 7: Comparison of failure rates between GPT-4o-mini and o1-mini in the 3-needles scenario.

- **Inconsistent Answer Prioritization:** The LLM may struggle to determine the "largest" or "most relevant" label when multiple valid answers exist.
- **Ambiguity in Sentence Ranking:** Despite successful embedding retrieval, the semantic similarity between different needle sentences can lead to incorrect prioritization when forming the final response.

To further investigate whether the LLM failure is due to a lack of reasoning capability, we tested the 3-needles scenario on the o1-mini model, which is specifically designed for reasoning tasks OpenAI (2024). The results, shown in Figure 7, indicate a significant reduction in LLM failure rates. GPT-4o-mini exhibited a LLM failure rate of **52.2%**, whereas o1-mini showed a much lower failure rate of **9.7%**.

5.3 Interplay Between Sentence Cap Size, Accuracy, and Failure Dynamics

Our experiments reveal a nuanced interplay between sentence cap size, retrieval accuracy, and failure dynamics, underscoring that more context is not always beneficial. We evaluated cap sizes of 3, 5, 10, 20, and 50 sentences under two scenarios: a single target needle (1-needle) and multiple target needles (3-needles).

In the 1-needle scenario, increasing the sentence cap size consistently improved accuracy (Figure 12 in Appendix E). This suggests that for simpler tasks, providing additional top-relevant sentences increases the likelihood of including the correct information, benefiting both GPT-4o-mini and Llama 3.2. However, when extending the task to the 3-needles scenario, a different trend emerged. Although a larger cap still increases the pool of potential relevant sentences, it also introduces more distractors, thereby reducing overall accuracy (Figure 13 in Appendix E). This inverse relationship highlights the challenge of balancing context expansion with the introduction of noise.

These trends are further illuminated by our failure rate analysis. As the sentence cap size increases, embedding failures—which occur when the correct sentence is omitted from the retrieved context—decline in both scenarios (Figure 8). This indicates that a larger cap reliably captures the target sentence. In contrast, LLM failures, defined by the language model's inability to accurately extract or reason about the correct label, tend to increase with a larger cap size, particularly in the more complex 3-needles scenario. This suggests that while more context aids the retrieval module, it simultaneously burdens the reasoning process with superfluous information.

5.4 Types of LLM Failures

To better understand the nature of LLM failures, we further categorize them into:

- **Incorrect Answer Failures**: Cases where the model provides an incorrect label despite the correct label being present in the retrieved sentence cap.
- **Unanswerable Failures**: Cases where the model incorrectly responds with "UNANSWERABLE" even though the correct label is available.

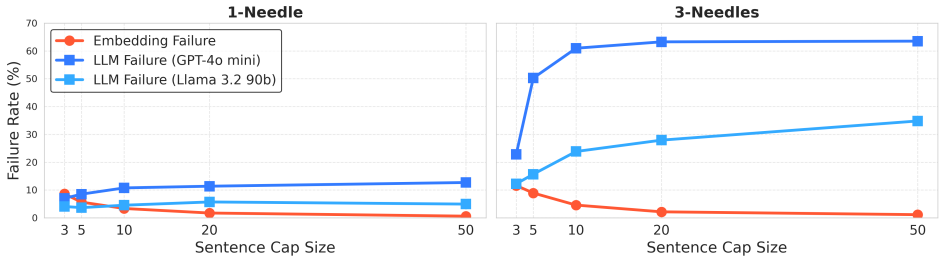


Figure 8: Effect of Sentence Cap Size on Embedding and LLM Failure Rates in both 1-Needle and 3-Needles Scenarios.

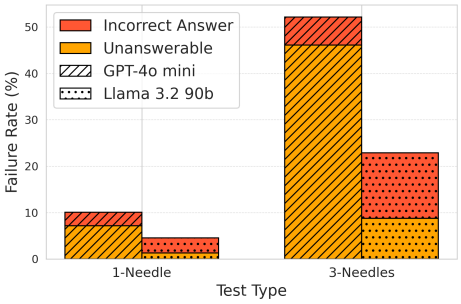


Figure 9: LLM Failure Breakdown

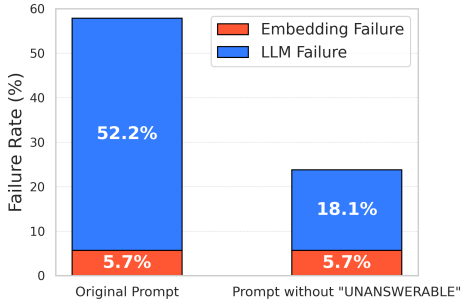


Figure 10: Effect of Prompt Modification on Failure Rates in the 3-Needles Scenario with GPT-4o-mini.

Figure 9 presents the breakdown of these failure types in both the 1-needle and 3-needles scenarios. In the 1-needle test, failure rates for both incorrect answers and unanswerable responses remain relatively low. However, in the 3-needles test, we observe a significant increase in unanswerable failures, particularly with GPT-4o-mini, where over 45% of total responses were classified as unanswerable despite the correct label being retrievable.

5.4.1 Effect of Prompt Modification on Failure Rates

To assess whether instructing models to return "UNANSWERABLE" when unsure contributes to failure rates, we conducted an experiment where we removed the sentence *"If the information is not available in the context, respond UNANSWERABLE."* from the prompt. The effect of this modification was tested on GPT-4o-mini in the 3-needle scenario. The results, shown in Figure 10, indicate that failure rates decreased when this instruction was omitted. In the original prompt setting, GPT-4o-mini exhibited an LLM failure rate of **52.2%**, which dropped to **18.1%** when the instruction was removed.

6 Conclusion

In this work, we introduced CROSS (Cross-lingual Retrieval Optimized for Scalable Solutions), a novel framework that addresses the challenges of long-context and cross-lingual retrieval. Through a two-phase retrieval mechanism leveraging multilingual embeddings and targeted segment ranking, CROSS significantly improves retrieval accuracy while maintaining computational efficiency. Our evaluations on the mLongRR-V2 dataset, spanning 49 language combinations and context lengths up to 512,000 words, demonstrate CROSS's robustness in mitigating the "lost-in-the-middle" problem and enhancing performance across diverse linguistic structures. Furthermore, CROSS reduces the token processing burden on language models, offering a scalable and cost-effective solution for multilingual information retrieval. Despite its strengths, failure analyses highlight areas for future work, particularly in improving multi-target reasoning and optimizing retrieval strategies for complex cross-lingual scenarios. These insights pave the way for further advancements in retrieval-augmented generation (RAG) frameworks, fostering more efficient and accurate multilingual retrieval systems.

REFERENCES

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. Evaluating multilingual long-context models for retrieval and reasoning, 2024. URL https://arxiv.org/abs/2409.18006.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Sina Bagheri Nezhad and Ameeta Agrawal. What drives performance in multilingual language models? In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov, and Jörg Tiedemann (eds.), *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pp. 16–27, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.vardial-1.2. URL https://aclanthology.org/2024.vardial-1.2/.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. Beyond data quantity: Key factors driving performance in multilingual language models. In Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage (eds.), *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pp. 225–239, Abu Dhabi, United Arab Emirates, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.loreslm-1.18/.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137.

Zijian Hei, Weiling Liu, Wenjie Ou, Juyi Qiao, Junming Jiao, Guowen Song, Ting Tian, and Yi Lin. Dr-rag: Applying dynamic document relevance to retrieval-augmented generation for question-answering, 2024. URL https://arxiv.org/abs/2406.07348.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models, 2024. URL https://arxiv.org/abs/2408.10151.

Zhiqi Huang, Puxuan Yu, and James Allan. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pp. 1048–1056. ACM, February 2023. doi: 10.1145/3539597.3570468. URL http://dx.doi.org/10.1145/3539597.3570468.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. Longrag: Enhancing retrieval-augmented generation with long-context llms, 2024. URL https://arxiv.org/abs/2406.15319.

Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 12 2006. ISSN 0891-2017. doi: 10.1162/coli.2006.32.4.485. URL https://doi.org/10.1162/coli.2006.32.4.485.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. Bordirlines: A dataset for evaluating cross-lingual retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2410.01171.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. McCrolin: Multi-consistency cross-lingual training for retrieval question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2780–2793, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.157.

Peiqin Lin, André F. T. Martins, and Hinrich Schütze. Xampler: Learning to retrieve cross-lingual in-context examples, 2024. URL https://arxiv.org/abs/2405.05116.

Robert Litschko, Ivan Vulić, and Goran Glavaš. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1071–1082, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.90.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172.

OpenAI. Openai o1 system card. https://cdn.openai.com/o1-system-card-20241205.pdf, 2024. Accessed: Month Day, Year.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=xw5nxFWMlo.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. Debateqa: Evaluating question answering on debatable knowledge, 2024b. URL https://arxiv.org/abs/2408.01419.

Eugene Yang, Suraj Nair, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Kevin Duh. Efficiency-effectiveness tradeoff of probabilistic structured queries for cross-language information retrieval, 2024. URL https://arxiv.org/abs/2404.18797.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024. URL https://arxiv.org/abs/2308.07107.

A Embedding Model: bge-m3 for Cross-Lingual Compatibility

The bge-m3 embedding model, with a 1024-dimensional embedding size, is key to CROSS's multilingual capabilities. It embeds sentences from different languages into a shared vector space, enabling CROSS to assess sentence relevance across languages and significantly boosting cross-lingual accuracy. By capturing both syntactic and semantic features, bge-m3 ensures robustness across diverse linguistic families, supporting accurate retrieval in languages like Persian, Hindi, Russian, and Arabic.

B Dataset: mLongRR-V2

The mLongRR-V2 dataset builds on the original mLongRR by Agrawal et al., which evaluated multilingual long-context models on retrieval tasks using five Latin script languages. However, the original mLongRR was limited to a maximum context length of 64,000 tokens and lacked diversity in script types Agrawal et al. (2024). mLongRR-V2 addresses these limitations by extending the context length to **512,000 words** and expanding the language set to include seven languages: **English, Vietnamese, Swahili, Persian, Russian, Hindi, and Arabic**. This expansion not only enhances linguistic diversity by incorporating non-Latin scripts such as Cyrillic, Devanagari, and Arabic, but also introduces a crucial **cross-lingual** dimension, allowing for more robust evaluations of retrieval models in multilingual and cross-script scenarios.

The cross-lingual aspect of mLongRR-V2 is designed to rigorously test retrieval models across different language combinations. In this dataset, **the haystack is always monolingual**, meaning

that all context within a given test case is written in a single language. However, the **needle (target sentence) is embedded within the haystack in the same language as the haystack itself**. The cross-lingual challenge arises from the fact that the **query is presented in a different language** from the haystack, requiring the model to bridge linguistic differences to retrieve the correct information.

Needle Structure In this task, the model's objective is to locate and extract information from a single target sentence hidden within the context. We adopt the same needle pattern as used in previous studies Agrawal et al. (2024); Team (2024); Anthropic (2024), which takes the form: **"The special magic {city} number is: {number}"**. Here, {city} is randomly chosen from a list of 23 unique cities worldwide, and {number} is a randomly generated 7-digit number. The list of cities was translated into all the dataset languages to ensure accuracy and linguistic consistency.

Cross-Lingual Language Pairs and Needle Positioning To provide a rigorous assessment, mLongRR-V2 includes **49 cross-lingual language pairs**, pairing each language in the prompt and query with every other language in the context. This setup simulates real-world scenarios where queries and contexts are often in different languages, adding complexity to the retrieval task.

Building on the original mLongRR, mLongRR-V2 positions the target information (the "needle") at **five distinct locations within the context**: the beginning (0%), near the start (25%), in the middle (50%), near the end (75%), and at the end (100%). This systematic positioning tests model robustness across varying depths, addressing challenges like the "lost in the middle" problem, where retrieval accuracy typically drops for mid-context information.

To test the reasoning capability of CROSS, we introduced a **3-needles setup**, where three needles are placed randomly within the context. The task requires the model to identify and reason about the needles to answer a query related to the largest one, further evaluating CROSS's ability to process complex multilingual scenarios.

Context Length mLongRR-V2 significantly extends the context length to a maximum of **512,000 words**, enabling the evaluation of models on much longer texts compared to the original mLongRR dataset. The dataset is carefully designed to test models across varying context lengths, consisting of **2k, 8k, 16k, 32k, 64k, 128k, 256k, and 512k words**. This range allows for a comprehensive assessment of a model's scalability and performance under diverse conditions.

C Evaluation Protocol

To comprehensively evaluate the effectiveness of CROSS, we designed two distinct evaluation tasks: the 1-needle test and the 3-needles test. These tests assess retrieval and reasoning capabilities across diverse cross-lingual scenarios.

1-Needle Test The 1-needle test evaluates the model's ability to retrieve specific information embedded within extensive multilingual contexts. In this task, a single "needle" (a target piece of information) is placed in the context at one of five predefined positions: the beginning (0%), near the start (25%), in the middle (50%), near the end (75%), or at the end (100%).

The prompt asks the model: "What is the special magic number?", written in different languages to assess cross-lingual retrieval. The model must locate the relevant information and provide the correct answer, ensuring retrieval accuracy across both language pairs and varying context positions.

3-Needles Test The 3-needles test evaluates the model's reasoning capability in addition to retrieval. In this setup, three needles are randomly distributed throughout the context. The model is prompted to identify and reason about the needles to answer the query: "What is the largest special magic number?"

This task challenges the model to not only locate multiple relevant pieces of information but also reason over them to produce the correct answer. The random placement of the needles tests robustness against varying context complexity. Each case was tested three times to account for the variability introduced by the random distribution of needles, ensuring a more reliable evaluation of the model's reasoning capabilities.

Metric: Retrieval Accuracy We use retrieval accuracy as the primary metric to evaluate model performance in both tasks. Accuracy is defined as the percentage of test cases where the model correctly identifies and retrieves the required information. In the 1-needle test, this means correctly locating the "special magic number." In the 3-needles test, accuracy measures the model's ability to reason and correctly identify the largest "special magic number."

Prompts and Queries The prompts and queries used in both the 1-needle and 3-needles tests are carefully crafted to ensure clarity and fairness across languages. Detailed template of the prompts are provided in the Appendix D for reference.

D Prompt and Needle Templates

This appendix presents the English versions of the templates and data used in our experiments. For the corresponding versions in other languages (Swahili, Vietnamese, Persian, Hindi, Arabic, and Russian), please refer to the anonymous repository [1].

D.1 Needle Template

`"The special magic {city} number is {number}."`

Here, `number` represents a randomly generated 7-digit number, and `city` is selected at random from the list below (with the city names translated into the context language):

- Chicago
- Yangon
- Antananarivo
- Colombo
- Almaty
- Sydney
- Mexico City
- Seattle
- Lagos
- Amsterdam
- Belgrade
- Cairo
- Baghdad
- Damascus
- Kigali
- Dakar
- Sofia
- Victoria
- Tashkent
- Mumbai
- Barcelona
- Amman
- Toronto

D.2 Prompt Templates

The following boxes display the prompt templates used in our experiments. These prompts instruct the LLM to generate concise and direct answers based on a provided context. Each prompt delineates the context (enclosed between `#CONTEXT` and `#ENDCONTEXT`) followed by a specific question.

Prompt Template: 1 Needle test

You are a helpful AI bot that answers questions for a user. Keep your response short and direct.

#CONTEXT
{text}
#ENDCONTEXT

#QUESTION
What is the special magic number? Don't give information outside the document or repeat your findings.
If the information is not available in the context, respond UNANSWERABLE.

[1] `https://anonymous.4open.science/r/prompts-template-D729`

Prompt Template: 3 Needles test

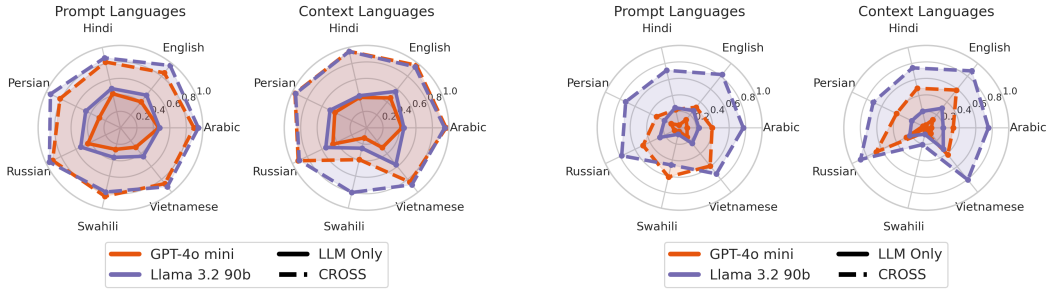
You are a helpful AI bot that answers questions for a user. Keep your response short and direct.

#CONTEXT
{text}
#ENDCONTEXT

#QUESTION
What is the largest special magic number? Don't give information outside the document or repeat your findings.
If the information is not available in the context respond UNANSWERABLE.

E LANGUAGE SPECIFIC FIGURES

This section presents additional figures that provide further insight into the performance and failures of CROSS in different retrieval settings and languages.



(a) Comparison of Average Accuracy for Each Prompt and Context Language in 1-Needle Test

(b) Comparison of Average Accuracy for Each Prompt and Context Language in 3-Needles Test

Figure 11: Radar Plots Comparing Average Accuracy in Different Tests

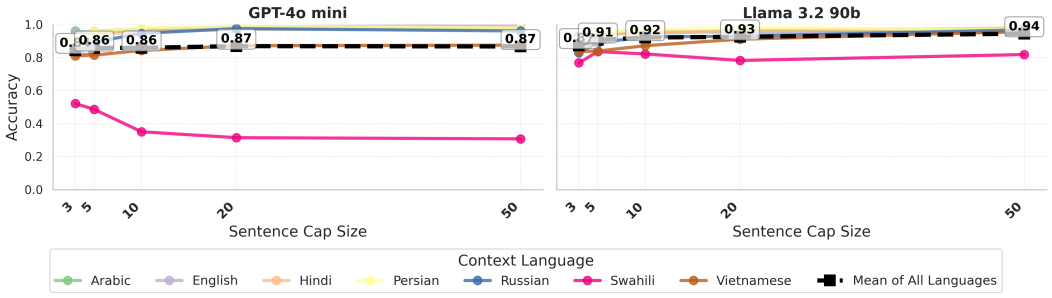


Figure 12: Accuracy of CROSS with varying sentence cap sizes in the 1-needle scenario.

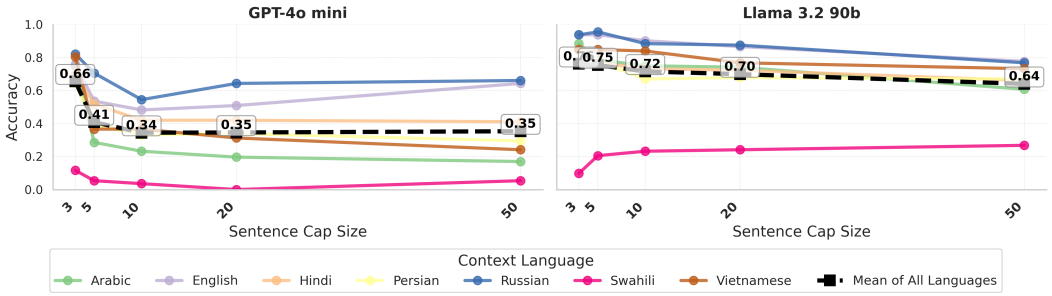


Figure 13: Accuracy of CROSS with varying sentence cap sizes in the 3-needles scenario.

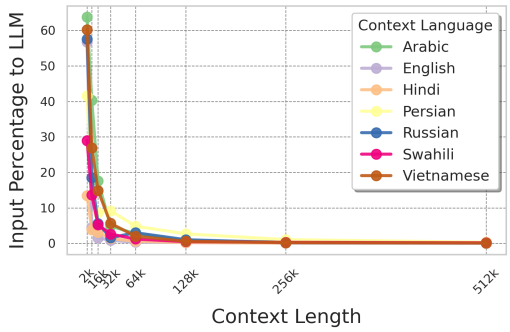


Figure 14: Mean Input Reduction Using CROSS by Context Length for Each Context Language

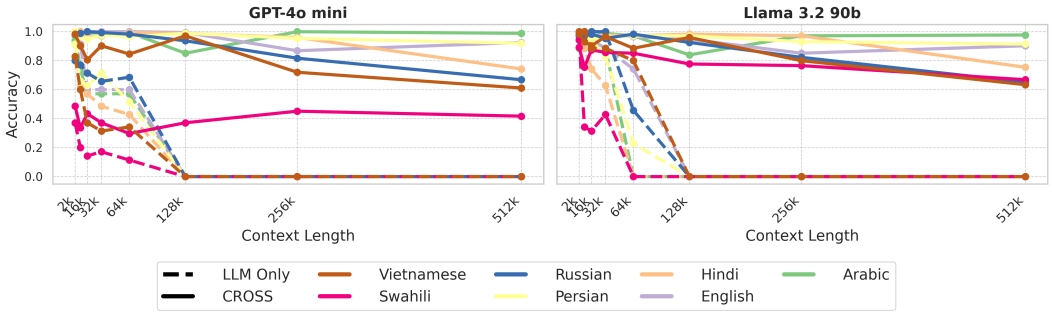


Figure 15: Accuracy by context length in 1-needle task

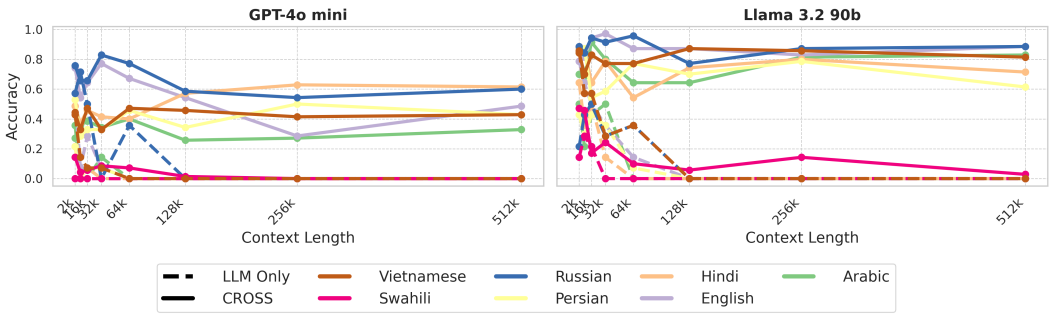


Figure 16: Accuracy by context length in 3-needles task

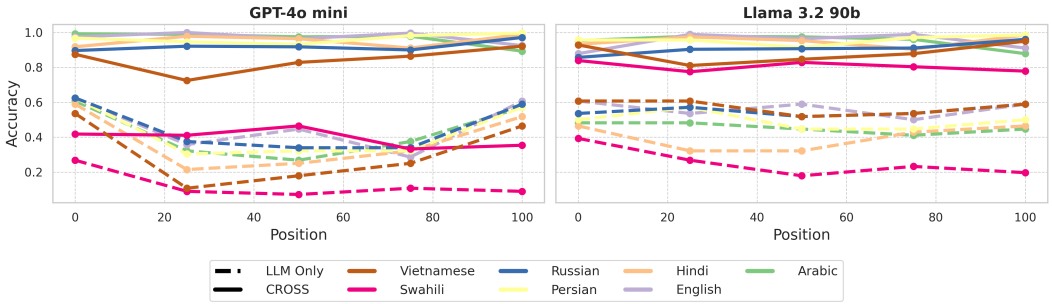


Figure 17: Accuracy by needle position in 1-needle task

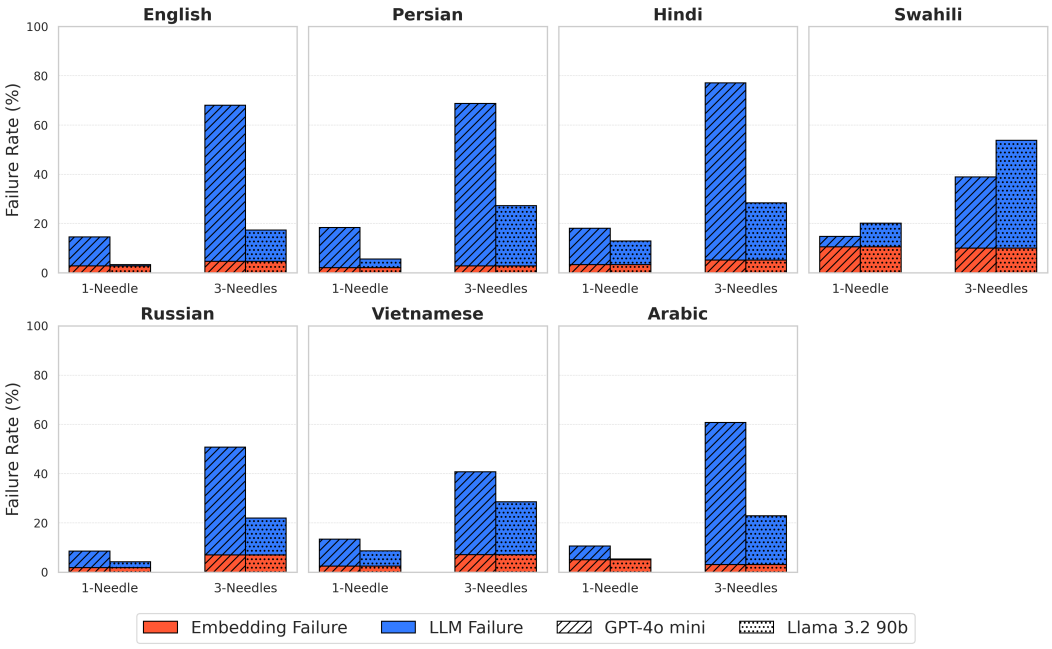


Figure 18: Failure rate by prompt language

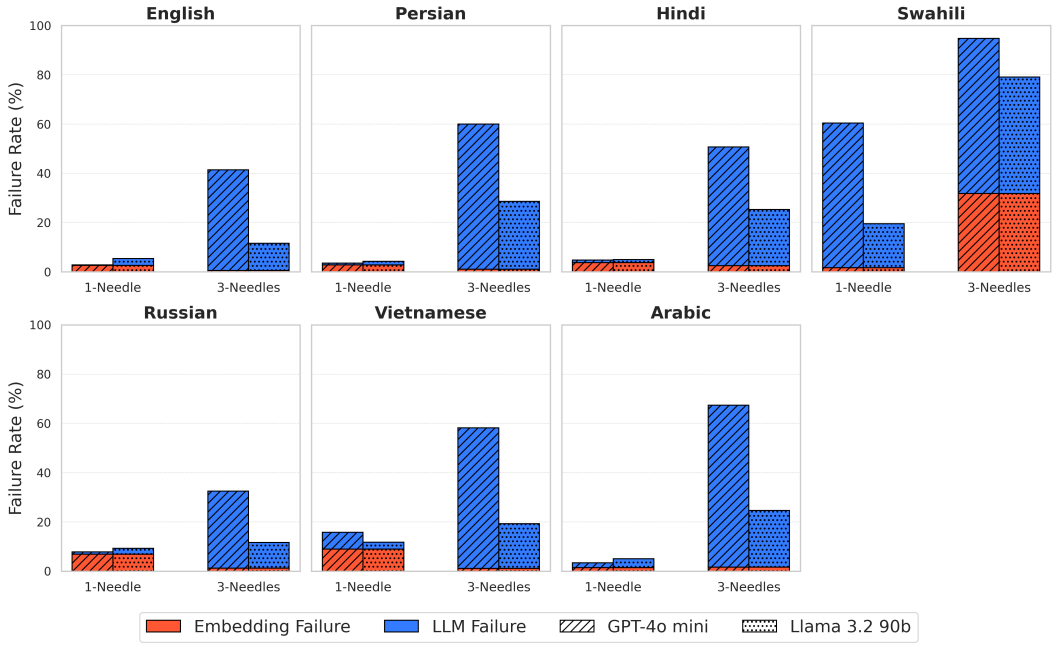


Figure 19: Failure rate by context language