Multilingual Needle in a Haystack: Evaluating Long-Context Retrieval in LLMs with English, Spanish, and Arabic

Omar Mohamed El-Gohary Omar Mohamed El-Abasery
Malak Mohamed Ibrahim Shahed Ahmed Kandil Rawan Ahmed Soliman
Dr. Mohamed Taher (Professor) Gehad Mustafa (TA)

May 26, 2025

Abstract

As large language models (LLMs) expand their context windows to 128K tokens and beyond, new questions emerge about their ability to retrieve specific information from deeply embedded positions—especially in multilingual documents. This paper evaluates the retrieval performance of GPT-4.1-mini on a 66,000-word trilingual input consisting of English, Arabic, and Spanish segments. We introduce a multilingual variant of the "Needle in a Haystack" benchmark by inserting six synthetic facts at controlled depths and test retrieval under three settings: monolingual prompting, cross-lingual querying, and segment-language misalignment. Results show consistent factual recall with high semantic accuracy across languages and scripts, even under disruptive structural conditions. While verbatim accuracy decreases with token depth, the model maintains strong semantic fidelity throughout. These findings highlight the capability of modern LLMs to perform language-agnostic retrieval in long-context settings, paving the way for advances in multilingual evaluation, cross-lingual QA, and scalable memory modeling.

1 Introduction

As large language models (LLMs) gain the ability to process increasingly long contexts, a critical challenge arises: can they accurately retrieve specific information embedded deep within such extended inputs? This retrieval task—often termed the "Needle in a Haystack" (NiH) problem—tests a model's capacity for precise memory over long documents. Although earlier studies have shown promising results in monolingual English contexts, it remains unclear whether these findings generalize to multilingual documents, especially those involving non-Latin scripts such as Arabic. This paper builds on the work of Kamradt (2023), who developed the NiH benchmark to evaluate factual recall in long English-only contexts using GPT-4. His findings highlighted that retrieval accuracy declines with increased context length and fact depth. We extend this line of research into the multilingual domain by constructing and evaluating a long-context benchmark in English, Spanish, and Arabic. Our evaluation spans different fact placements—beginning, middle, and end—within a 66,000-word document. We aim to examine whether retrieval accuracy varies across languages with differing levels of training resource availability and script complexity. In doing so, we contribute to the broader discussions around language equity, generalization in LLMs, and the design of multilingual evaluation benchmarks. Specifically, we investigate the following research questions:

- How accurately can large language models retrieve facts from long multilingual contexts?
- How do language resource levels and script types (Latin vs. non-Latin) affect retrieval performance?

2 Related Work

Research on long-context retrieval has expanded significantly with the advent of LLMs supporting 100K+ token windows. One of the most influential benchmarks in this space is the "Needle in a Haystack" (NiH) task by Kamradt (2023), which involves embedding a single factual sentence within a long, unrelated text

and testing whether a model can retrieve it. His findings revealed a sharp drop in accuracy as fact depth increased—especially when the fact was neither at the beginning nor end of the document. However, this benchmark remains monolingual, limited to English.

In parallel, multilingual evaluation frameworks like L-Eval and M4 have assessed LLM performance in summarization, QA, and translation, but have largely overlooked long-context retrieval, especially across non-Latin scripts. Existing studies often assume equal token efficiency across languages, yet Liu et al. (2024) show that Arabic and other morphologically rich languages incur greater token overhead, potentially skewing retrieval accuracy in multilingual settings.

Further, Xu et al. (2024) report that attention mechanisms in LLMs degrade unevenly across input lengths, with middle segments receiving less focus—an effect known as "middle blindness." This has significant implications for evaluating retrieval by token position.

While retrieval-augmented models (e.g., Schick et al., 2023) have improved factual accuracy in open-domain settings, they rely on external knowledge sources, whereas our work focuses solely on internal in-context memory.

To date, no benchmark simultaneously evaluates long-context retrieval across multiple languages within the same input, nor do existing datasets incorporate language-specific token inflation or script variation. Our study directly addresses this gap by constructing a multilingual benchmark and analyzing factual recall across English, Spanish, and Arabic.

3 Methodology

3.1 Dataset Construction

We constructed a 66,000-word (approx. 105,000-token) multilingual document composed of interleaved passages from *To Kill a Mockingbird*, translated into English, Arabic, and Spanish. The document was divided into six sections with deliberate language alternation, enabling controlled testing of retrieval across language boundaries and context depths. The token allocation for each section was as follows:

- English: 18,000 tokens (Section 1)
- Arabic: 20,000 tokens (Section 2)
- Spanish: 18,000 tokens (Section 3)
- Arabic: 18,000 tokens (Section 4)
- English: 16,000 tokens (Section 5)
- Spanish: 16,000 tokens (Section 6)

This six-switch structure allowed us to introduce depth variation while minimizing context fragmentation. Excessive switching across languages or scripts (Latin vs. Arabic) could disrupt attention flows or reduce usable context for each language. By maintaining two contiguous blocks per language, we ensured a balance between diversity and interpretability.

Within this document, we manually inserted six synthetic, language-specific factual statements—two per language—into different paragraphs. These statements (the "needles") were crafted to be semantically independent of the original source material and were not present in the pretraining data. This ensured that successful retrieval would require in-context memory rather than memorization or pattern recognition from pretraining.

Importantly, facts were unique to each language and were positioned at different token depths to enable position-based performance analysis. All insertions were done manually with careful attention to paragraph structure, ensuring natural embedding within the surrounding narrative.

3.2 Model Setup

We conducted all experiments using GPT-4.1-mini (version 2025-04-14) via the OpenAI API, which supports context windows up to 128K tokens. This model was selected for its balance between performance and availability, and it allowed us to evaluate retrieval behavior within a 66,000-word multilingual context without truncation.

Given that the OpenAI API is stateless, we submitted the entire document with each factual query to ensure that all relevant context remained accessible to the model. To comply with OpenAI's token-per-minute (TPM) limit of 200,000 tokens, we introduced a 15-second delay between queries to avoid rate-limiting errors.

All questions were asked using zero-shot prompting. In Experiment 1 and Experiment 3, each query was written in the same language as its corresponding inserted fact, isolating the model's monolingual retrieval capabilities. In Experiment 2, we introduced cross-lingual prompting, where questions were posed in a different language than the fact itself. This allowed us to evaluate the model's ability to perform semantic mapping across languages without explicit translation instructions. This combination of prompting strategies enabled us to test both language-specific and cross-language retrieval performance within a consistent input context.

3.3 Evaluation Design

To assess retrieval accuracy, we conducted manual evaluation of all model responses. For each query, the output was compared against the corresponding inserted fact ("needle") in the document. A binary evaluation metric was used: the response was considered either correct or incorrect, based on whether it matched the intended factual content.

We allowed for minor grammatical or surface-level variations—such as word order changes, synonyms, or diacritic usage—so long as the core factual meaning was preserved. Responses that were vague, partially correct, or hallucinated were marked as incorrect.

Given the multilingual nature of the task and the small number of queries, we opted for manual scoring rather than automated semantic similarity tools. This choice enabled greater interpretability and finer-grained error analysis, especially for non-Latin scripts and cross-lingual phrasing variations.

3.4 Experimental Conditions

Each experiment was conducted on the same 66,000-word multilingual document, with careful manual control over the structure and placement of inserted facts. While no tokenizer was used directly, we estimated token depth and section boundaries using LLM-based token feedback and structured text placement to ensure reasonably accurate positioning of each fact.

Inserted facts were manually placed within three distinct context zones:

- Early (approximately 0–30% of the document),
- Middle (approximately 40–60%),
- Late (approximately 70–100%).

Each experiment involved six queries, one for each inserted fact (two per language), and was designed to isolate retrieval behavior along key dimensions:

- Experiment 1: Monolingual prompting (query language matches fact language)
- Experiment 2: Cross-lingual prompting (query language differs from fact language)
- Experiment 3: Segment-language mismatch (fact placed in a different language section)

This setup enabled a controlled comparison of retrieval performance by language, token depth, and script type (Latin vs. non-Latin), while keeping the overall input constant. By modifying only the prompt language or placement strategy, we maintained a consistent experimental framework for evaluating multilingual retrieval robustness.

3.5 Scope and Limitations

This study focuses on evaluating the retrieval capabilities of large language models in long, multilingual input contexts. Specifically, we examine how well GPT-4.1-mini retrieves manually inserted factual statements embedded within a 66,000-word input composed of English, Spanish, and Arabic segments.

The scope of this work includes:

- Testing retrieval from in-context memory only, without using external tools or retrieval augmentation.
- Using a single shared multilingual input, with two synthetic facts inserted per language.
- Evaluating performance using manual, binary scoring based on exact or semantically faithful matches.
- Operating within a single model architecture (GPT-4.1-mini) using the OpenAI API, with no fine-tuning or system-level adjustments.

While the design ensures strong internal control, certain limitations remain:

- The study is limited to a single LLM; results may vary with larger or multilingual-optimized models such as GPT-4-turbo, Claude, or Mistral.
- Due to time and rate-limit constraints, only a single run per experimental condition was conducted, which limits statistical generalization.
- The manual nature of fact placement and scoring may introduce slight subjectivity, especially in edge cases of paraphrasing.
- Retrieval was tested on a fixed-length document (100K tokens); results may differ with shorter or much longer inputs.
- Only three languages were included, and while they vary by script and resource level, this does not capture the full diversity of global languages (e.g., agglutinative or logographic systems).
- The experiments emphasized monolingual and bilingual prompting, while other language tasks (e.g., many-to-one or zero-shot cross-lingual reasoning) remain open for future work.

These constraints help define the current study's boundaries and also inform several future directions for deeper multilingual long-context evaluation.

4 Experiments and Results

4.1 Experiment 1 – Monolingual Prompting

4.1.1 Experimental Setup

This experiment evaluated GPT-4.1-mini's ability to retrieve language-specific facts from a long multilingual context when the prompt and fact were in the same language. Six synthetic factual statements were manually inserted—two each in English, Arabic, and Spanish—into distinct paragraphs across the document. All queries were issued in zero-shot form and written in the same language as the corresponding fact. The facts were distributed across early and late positions within their respective language segments to test retrieval performance at varying depths.

4.1.2 Retrieval Results Summary

Language	Language Segment	Question	Inserted Fact (Needle)	Model Answer	Match Type
English	First Segment	Where was the group project done?	The group project was done in a cafe called EspressoLAB!	The group project was done in a cafe called EspressoLAB!	Exact
cibarA	First Segment	من أين كان عمر الجوهرى؟	عمر الجوهرى كان من البحيرة من قرية تسمى ابوحمص	عمر الجوهرى كان من البحيرة من قرية تسمى ابوحمص.	Exact
Spanish	First Segment	¿Qué come y bebe David el lunes?	David come manzanas y bebe leche en el lunes	David come manzanas y bebe leche en el lunes	Exact
cibarA	Second Segment	ما هي جامعة مصر للمعلوماتية؟	جامعة مصر للمعلوماتية هي جامعة مرموقة وموقعها العاصمة الإدارية الجديدة	جامعة مصر للمعلوماتية هي جامعة مرموقة تقع في العاصمة الإدارية الجديدة في مصر.	Semantic
English	Second Segment	What is Maryam's favorite fashion brand?	Dior is Maryam's favorite fashion brand	Maryam's favorite fashion brand is Dior.	Semantic
Spanish	Second Segment	¿Cuál es el mejor equipo de la liga de campeones?	El Real Madrid es el mejor equipo de la liga de campeones	El mejor equipo de la Liga de Campeones, según el texto proporcionado, es el Real Madrid.	Semantic

4.1.3 Analysis and Interpretation

The model successfully retrieved all six inserted facts, with 100% factual accuracy. Notably, facts placed in early segments were retrieved verbatim, while those located in later segments were returned with surface-level variation (e.g., reworded phrasing), though their semantic content remained intact. This outcome supports the presence of a token-depth sensitivity, where facts closer to the start of the document are more likely to be recalled exactly. However, the semantic preservation in all cases indicates that deep factual recall remains strong, even at greater context depths. Importantly, GPT-4.1-mini maintained high retrieval accuracy across three languages, including Arabic, which presents additional challenges due to script complexity and higher token overhead. This highlights the model's ability to handle language-specific memory retrieval without requiring translation or special prompts.

4.2 Experiment 2: Cross-Lingual Prompting

4.2.1 Experimental Setup

This experiment tested GPT-4.1-mini's ability to retrieve facts when the prompt language differed from the language in which the fact was written. Each of the six previously inserted facts (from Experiment 1) remained unchanged in the multilingual document, but queries were posed in a different language than the fact's original language.

For example:

- An English fact was queried in Arabic.
- An Arabic fact was queried in Spanish.
- A Spanish fact was queried in English.

This design evaluates the model's internal semantic alignment across languages, without any explicit translation instructions or cross-lingual fine-tuning.

4.2.2 Retrieval Results Summary

Lang.	Segment	Question Language	Match Type	Inserted Fact (Needle)	Model Answer
EN	First Segment	Arabic	Semantic	The group project was done in a cafe called EspressoLAB!	تم عمل المشروع الجماعي في مقهى يسمى !BALosserpsE
AR	First Segment	Spanish	Semantic	الطائرة المصرية هبطت بنجاح في مطار القاهرة الدولي	El avión egipcio aterrizó con éxito en el Aeropuerto Internacional de El Cairo.
ES	First Segment	English	Semantic	David come manzanas y bebe leche en el lunes	David eats apples and drinks milk on Monday.
AR	Second Segment	English	Semantic	جامعة مصر للمعلوماتية هي جامعة مرموقة وموقعها العاصمة الإدارية الجديدة	Egypt University of Informatics is described as a prestigious university located in the New Administrative Capital of Egypt.
EN	Second Segment	Spanish	Semantic	Dior is Maryam's favorite fashion brand	La marca de moda favorita de Maryam es Dior.
ES	Second Segment	Arabic	Semantic	El Real Madrid es el mejor equipo de la liga de campeones	وفقًا للنص الذي قدمته، فإن نادي ريال مدريد هو أفضل فريق في دوري أبطال أوروبا.

4.2.3 Analysis and Interpretation

The model achieved 100% factual recall, retrieving all six facts semantically correctly despite the language mismatch between prompt and fact. While none of the responses matched the original insertions verbatim, the core factual meaning was preserved in each case.

This performance indicates that GPT-4.1-mini has developed a robust internal multilingual representation space, enabling it to resolve cross-lingual queries and trace the intended fact without explicit translation steps. The experiment demonstrates that the model can:

- Understand a question in one language.
- Navigate through a multilingual context.
- Locate and return the correct fact in a different language.

These findings suggest strong potential for applications in cross-lingual question answering, multilingual document understanding, and global-scale retrieval systems.

4.3 Experiment 3: Segment-Language Mismatch

4.3.1 Experimental Setup

This experiment assessed GPT-4.1-mini's ability to retrieve facts when the language of the inserted fact did not match the language of the surrounding context segment. Each of the six previously inserted facts was manually placed into a paragraph written in a different language than the fact itself. For example:

- An English fact was inserted into an Arabic section.
- An Arabic fact was inserted into a Spanish section.
- A Spanish fact was inserted into an English section.

All queries were issued in the original language of the fact, not the surrounding segment's language, ensuring that only the segment placement was altered.

4.3.2 Retrieval Results Summary

Lang.	Actual Section	Question Language	Match Type	Inserted Fact (Needle)	Model Answer
EN	Arabic	English	Exact	The group project was done in a cafe called EspressoLAB!	تم عمل المشروع الجماعي في مقهى **.BALosserpsE**
AR	Spanish	Arabic	Exact	عمر الجوهرى كان من البحيرة من قرية تسمى ابوحمص	عمر الجوهرى كان من البحيرة من قرية تسمى ابوحمص.
ES	English	Spanish	Exact	David come manzanas y bebe leche en el lunes	David come manzanas y bebe leche el lunes.
AR	English	Arabic	Semantic	جامعة مصر للمعلوماتية هي جامعة مرموقة وموقعها العاصمة الإدارية الجديدة	جامعة مصر للمعلوماتية هي جامعة مرموقة تقع في العاصمة الإدارية الجديدة في مصر.
EN	Spanish	English	Semantic	Dior is Maryam's favorite fashion brand	Maryam's favorite fashion brand is Dior.
ES	Arabic	Spanish	Semantic	El Real Madrid es el mejor equipo de la liga de campeones	Según el texto proporcionado, "El Real Madrid es el mejor equipo de la Liga de Campeones."

4.3.3 Analysis and Interpretation

GPT-4.1-mini successfully retrieved all six inserted facts, with three returned verbatim and three in paraphrased but semantically faithful forms. Notably, the model was able to identify and extract facts even when they were surrounded by foreign-language content and written in a different script. This demonstrates that the model does not depend heavily on the linguistic consistency of the surrounding segment. Instead, it appears to rely on semantic representations of the inserted content and is capable of navigating across language and script boundaries. The experiment reinforces the model's capability for language-agnostic retrieval and its robustness to structural and contextual noise, such as mismatched scripts or grammar. These findings support GPT-4.1-mini's effectiveness in complex multilingual documents where facts may not appear within clearly segmented language blocks—a common scenario in bilingual reports, mixed-language transcriptions, or interleaved data sources.

5 Comparative Insights Across Experiments

Across all three experimental conditions, GPT-4.1-mini retrieved 100% of the inserted facts with either exact or semantically correct responses:

Experiment	**Prompt-Fact Language Match**	**Segment-Fact Language Match**	**Retrieval Accuracy**	**Verbatim Matches**	**Semantic Matches**
Experiment 1: Monolingual	Same	Same	6/6 (100%)	3	3
Experiment 2: Cross-Lingual	Different	Same	6/6 (100%)	0	6
Experiment 3: Segment Mismatch	Same	Different	6/6 (100%)	3	3

Table 1: Summary of retrieval accuracy and match types across the three experiments

5.1 Key Observations

The results across the three experiments demonstrate several consistent patterns in the model's retrieval behavior. In the first experiment, where prompts and inserted facts were in the same language and placed within matching segments, GPT-4.1-mini displayed a clear sensitivity to token depth. Facts located near the beginning of the input were often retrieved verbatim, while those embedded deeper were more likely to be paraphrased. This aligns with prior studies suggesting that attention fidelity degrades over long contexts, particularly in middle or late segments. In the second experiment, which introduced cross-lingual prompting, the model maintained full semantic fidelity even when asked to retrieve a fact in a different language from the one in which it was written. Although none of the responses were verbatim—owing to the necessary shift in language—all six facts were accurately retrieved in translated or paraphrased form. This finding reveals that GPT-4.1-mini possesses strong internal multilingual alignment, allowing it to resolve meaning across language boundaries without explicit translation mechanisms. The third experiment further tested the model's resilience by placing each fact into a segment written in a different language and script. Remarkably, the model was still able to retrieve all facts correctly, with several returned verbatim. This indicates that the model does not rely solely on the linguistic characteristics of the surrounding context and can instead retrieve based on internal semantic structure, even when the host segment is written in an unrelated language or script. Together, these observations point to a model that is not only capable of deep multilingual retrieval but also resilient to various types of structural noise—whether that be prompt-language mismatch, positional variation, or segment-script divergence. While verbatim accuracy does appear to decline with token depth, the preservation of factual meaning across all settings suggests that GPT-4.1-mini can maintain stable and language-agnostic memory within long and heterogeneous contexts.

6 Discussion

The results of the three experiments highlight the increasing maturity of long-context LLMs like GPT-4.1-mini in handling complex multilingual retrieval tasks. The model's consistent ability to retrieve all six inserted facts—regardless of prompt language, fact position, or surrounding segment—demonstrates that it possesses a strong form of semantic memory that is not bound by surface-level cues such as language or script alignment. This behavior departs significantly from earlier generations of LLMs, which often struggled with retrieval even under monolingual and well-aligned conditions. One particularly notable finding is the model's resilience to cross-lingual and structural variation. Even when facts were placed deep within a segment written in a different script and language, the model was able to locate and retrieve them accurately. This suggests that the model builds language-agnostic latent representations of factual content, allowing it to operate over heterogeneous documents in a way that does not require explicit translation or fine-tuned prompting. These capabilities have important implications for tasks such as multilingual document understanding, cross-lingual question answering, and retrieval in mixed-language corpora—all of which are increasingly common in globalized information systems. The results also reinforce the idea that context position plays a role not in whether retrieval is successful, but in how the retrieved information is phrased: earlier facts are more likely to be returned verbatim, while later ones tend to be paraphrased. This aligns with prior findings (e.g., Xu et al., 2024; Balachandran et al., 2023) on attention degradation over long contexts. At the same time, the controlled nature of this study also reveals some boundaries. With only two inserted facts per language and a single prompt per fact, it is not yet possible to generalize findings to high-density retrieval scenarios or test the model's robustness to factual interference (e.g., decoy statements or contradictions). Moreover, the use of a relatively high-resource trio of languages (English, Spanish, Arabic) leaves open the question of whether these results would generalize to lower-resource or morphologically complex languages, such as Swahili or Uyghur. In short, the experiments demonstrate that GPT-4.1-mini exhibits a surprising degree of semantic alignment and retrieval reliability across languages and contexts—especially notable given the inclusion of a non-Latin script language like Arabic, which typically poses greater tokenization and structural challenges for LLMs. However, future research will need to stress-test these capabilities under more adversarial or ambiguous conditions to better understand the limits of multilingual long-context memory.

7 Future Work

While the present study confirms GPT-4.1-mini's ability to retrieve factual information from long, multilingual contexts, it also opens several avenues for further investigation. First, a key limitation of this work is the low factual density: only two inserted facts per language were tested in each run. Future experiments should explore higher-density retrieval scenarios, where multiple facts are embedded per language block, and where the model must discriminate between similar or partially overlapping content. This would allow researchers to assess the model's precision under factual interference and evaluate its resistance to semantic confusion. Second, we plan to expand the analysis of token depth effects by inserting facts at more granular positions across the document—beyond just early and late zones. Prior work (e.g., Balachandran et al., 2023) has pointed to a phenomenon known as "middle blindness", in which model attention degrades disproportionately in mid-context regions. Testing retrieval fidelity across continuous token intervals would help validate or challenge that claim in a multilingual setting. Another important direction is to deepen the exploration of cross-lingual prompting. While our second experiment demonstrated that facts can be retrieved when prompted in a different language, future work could test more complex scenarios—such as many-to-one retrieval, zero-shot cross-lingual chains of thought, or code-switched prompts—to evaluate how robust the model is to mixed-language reasoning. Additionally, introducing distractor facts—plausible but incorrect alternatives placed near the true fact—would test the model's selectivity and resistance to hallucination, especially under ambiguous or adversarial conditions. Finally, comparing retrieval behavior across different LLM architectures (e.g., GPT-4o, Claude, Mistral, LLaMA 3) and language-resource configurations (e.g., high-resource vs. extremely low-resource languages) would help determine whether the multilingual retrieval strength observed in GPT-4.1-mini is a general property of transformer-based LLMs, or the result of specific training scale, tokenizer design, or data diversity. Together, these directions would deepen our understanding of multilingual long-context memory and move us closer to developing models that are not only capable of deep recall but also robust, equitable, and precise in complex multilingual environments.

8 Conclusion

This study investigated the ability of large language models to retrieve factual information from long, multilingual contexts without the aid of external retrieval systems. By constructing a 66,000-word document containing interleaved English, Arabic, and Spanish segments, we evaluated GPT-4.1-mini's performance across three distinct experimental conditions: monolingual prompting, cross-lingual prompting, and segment-language mismatch. Across all experiments, the model successfully retrieved all six inserted facts, achieving full factual recall with either exact or semantically faithful responses. While verbatim accuracy tended to decline with token depth, core factual meaning was consistently preserved. These results highlight GPT-4.1-mini's ability to perform language-agnostic memory retrieval within long, structurally complex inputs. The inclusion of Arabic—a non-Latin script language with token inflation characteristics—further underscores the model's capacity to generalize across scripts and language-resource disparities. The model's success in both aligned and misaligned conditions suggests that modern LLMs are not only capable of multilingual retrieval but are also increasingly robust to structural and linguistic variability. These findings contribute to a growing body of research that explores how LLMs manage and retrieve information at scale, particularly in globally relevant, multilingual settings. As models continue to expand in context length and deployment reach, understanding their long-context memory behavior across languages will remain a central challenge for NLP research and real-world AI applications.

References

[1] V. Balachandran et al., "Lost in the Middle: How Language Models Use Long Contexts," 2023. Available: `https://arxiv.org/abs/2307.03172`.

[2] G. Kamradt, "LLMTest_NeedleInAHaystack," GitHub repository, 2023. [Online]. Available: `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

[3] OpenAI, "GPT-4 Technical Report," 2023. [Online]. Available: `https://arxiv.org/abs/2303.08774`.

[4] T. Schick et al., "Evaluating Multilingual Long-Context Models for Retrieval and Reasoning," 2024. [Online]. Available: `https://arxiv.org/abs/2409.18006`.

[5] Y. Xu et al., "Token-length Bias in Language Models: An Analysis and Mitigation Strategy," 2024. [Online]. Available: `https://arxiv.org/abs/2408.10151`.

[6] X. Zhou et al., "Multilingual Needle in a Haystack: Investigating Long-Context Behavior of Multilingual Large Language Models," 2024. [Online]. Available: `https://arxiv.org/abs/2405.15102`.