Evaluating Multilingual Long-Context Models for Retrieval and Reasoning

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, Russell Scheinberg
Department of Computer Science
Portland State University, USA
{ameeta, andang, sina5, pokharel, rschein2}@pdx.edu

Abstract

Recent large language models (LLMs) demonstrate impressive capabilities in handling long contexts, some exhibiting near-perfect recall on synthetic retrieval tasks. However, these evaluations have mainly focused on English text and involved a single target sentence within lengthy contexts. Our work investigates how LLM performance generalizes to multilingual settings with multiple hidden target sentences. We create a new dataset – mLongRR – to comprehensively evaluate several multilingual long-context LLMs on retrieval and reasoning tasks across five languages: English, Vietnamese, Indonesian, Swahili, and Somali. These languages share the Latin script but belong to distinct language families and resource levels. Our analysis reveals a significant performance gap between languages. The best-performing models such as Gemini-1.5 and GPT-4o, achieve around 96% accuracy in English to around 36% in Somali with a single target sentence. However, this accuracy drops to 40% in English and 0% in Somali when dealing with three target sentences. Our findings highlight the challenges long-context LLMs face when processing longer contexts, an increase in the number of target sentences, or languages of lower resource levels.

1 Introduction

The ability to model long context sequences spanning tens of thousands of tokens is crucial for tasks such as summarization and question answering based on long documents such as books and reports, and code generation at the repository level. Recent advancements in large language models (LLMs) have focused on improving their capabilities in processing long context information (Dai et al., 2019; Chen et al., 2023; Ding et al., 2024).

Long-context language models, particularly multilingual ones, have the potential to enable remarkable progress in various applications by understanding lengthy textual data across different languages. An example of this potential was recently demonstrated by the newly introduced Gemini-1.5 Pro model (Reid et al., 2024) which leveraged its long-context window for in-context learning. By including a grammar manual in its context window, the model was able to learn to translate from English to Kalamang, an extremely low-resource language with fewer than 200 speakers (Visser, 2020). Such examples highlight the potential of long-context models in tackling challenging tasks in low-resource languages, where data scarcity has traditionally been a barrier.

Current methods for evaluating long-context LLMs primarily focus on English text. This has led to a severe lack of insights into their performance across diverse languages. Evaluating multilingual performance is crucial, not only for informing the development of effective models that serve diverse communities (Lai et al., 2023a; Ahuja et al., 2023), but also for developing safer models as research suggests that LLMs tend to generate more unsafe and irrelevant responses to malicious prompts in lower-resource languages (Shen et al., 2024). However, there is a notable lack of multilingual benchmarks hindering our understanding of how long-context LLMs perform across different linguistic contexts.

To address this gap, we present the first comprehensive study of long-context LLMs in multilingual settings leveraging evaluation frameworks relying on synthetic tasks (Mohtashami and Jaggi, 2023; Chen et al., 2023; Liu et al., 2024; Kamradt, 2023; Reid et al., 2024; Anthropic, 2024). Although the task is partially synthetic, we create a new dataset – mLongRR[1] – consisting of naturally occurring text and human translated data, making the setup as close to real-world setting while creating a con-

[1]The code is available at https://github.com/PortNLP/mLongRR.

Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), pages 216–231
November 16, 2024 ©2024 Association for Computational Linguistics

Language	ISO 639-3 Code	Resource Level	Language Family	Script
English	eng	Level 5	Indo-European	Latin
Vietnamese	vie	Level 4	Austro-Asiatic	Latin
Indonesian	ind	Level 3	Austronesian	Latin
Swahili	swa	Level 2	Niger-Congo	Latin
Somali	som	Level 1	Afro-Asiatic	Latin

Table 1: Languages studied and their details.

trolled environment for comparing model performance across languages. In addition to retrieval tasks, we introduce a new reasoning task where the models not only need to retrieve relevant items but also compare them with each other. For this, the models must keep track of these items in a long-context scenario, allowing us to analyze the models' reasoning capabilities.

We conduct a systematic evaluation of six different LLMs across five languages of varying resource levels. Our research aims to answer the following two questions:
(1) How do the long context capabilities of LLMs compare in retrieval and reasoning tasks in multilingual contexts?
(2) Are there significant performance differences between LLMs in multilingual contexts?

Some of our key findings are summarized as follows:

- The performance rapidly declines as we increase the context lengths for all languages.
- The performance also rapidly decreases as we move from higher-resource to lower-resource languages.
- Reasoning tasks are more challenging than retrieval tasks for all languages.
- There is a significant gap between the performance of different LLMs.
- Even seemingly simple "needle in the haystack" evaluation is able to expose limitations in current models when dealing with multilingual contexts.

We hope that the findings of our study will contribute to a deeper understanding of current long-context evaluation in multilingual contexts and encourage the development of more effective long-context multilingual models.

2 Related Work

Recent advancements in language models have focused on improving their ability to recall and reason over fine-grained information from tens of thousands of tokens of context (Achiam et al., 2023; Jin et al., 2024). Due to shortage of really long-context benchmarks, evaluation is typically focused on synthetic tasks such as passkey retrieval (Mohtashami and Jaggi, 2023; Chen et al., 2023; Liu et al., 2024; Ding et al., 2024; Jin et al., 2024) or needle in a haystack (Kamradt, 2023; Reid et al., 2024; Anthropic, 2024) which measure a model's ability to accurately recall information from a vast corpus of data.

Recently, Gemini 1.5 (Reid et al., 2024) and Claude-3 (Anthropic, 2024) models reported near-perfect recall on the needle in a haystack task. Prior work has also studied perplexity but a low perplexity score has shown to not necessarily indicate proficiency in handling long contexts or reflect the model's performance on sequence-level tasks in real applications (Sun et al., 2021; Pal et al., 2023; Jin et al., 2024). Furthermore, most of these studies have been limited to English only texts.

Although some long-context real-world benchmarks have been recently introduced, they are also limited to English (An et al., 2024), and while some bilingual English/Chinese (Bai et al., 2023; Qiu et al., 2024; Yuan et al., 2024) datasets offer a slight improvement, due to the effort-intensive nature of dataset creation, they are limited to a very small number of languages.

3 Multilingual Needles in a Haystack for Retrieval and Reasoning Evaluation

3.1 Languages and Models

Languages We selected five languages to study: English, Vietnamese, Indonesian, Swahili, and Somali. These languages span different resource lev-

els from high to extremely low[2] allowing us to gain insight into how language resource levels affect models' ability to work over long context windows.

Furthermore, we deliberately control for script-related variables as they have been shown to have a considerable impact on the performance of a model (Bagheri Nezhad and Agrawal, 2024b). We study languages that use the Latin script for three reasons: models perform significantly better on Latin-script languages than non-Latin languages (Chau and Smith, 2021; Bang et al., 2023), the fragmentation rate of Latin script is lower than other scripts allowing Latin-script languages to be represented with substantially fewer tokens as compared to languages in other scripts (Ács, 2019; Ahia et al., 2023) – a disparity that becomes even more pronounced over long contexts, and, lastly, the fragmentation rate of Latin-script languages remains comparable which is helpful when considering considerably long input texts. Our selection of languages, shown in Table 1, has the added benefit of including less-studied languages and language families[3], providing a more comprehensive view of the latest generation of multilingual capabilities of long-context LLMs.

Models We consider four proprietary and two open-source long-context LLMs.

- **GPT-4** is a proprietary multilingual LLM from OpenAI (Achiam et al., 2023) that has been shown to perform a wide range of tasks. We used the *gpt-4-0125-preview* version, which is the latest one at the time of our experiments. It has a context window of 128K tokens and was trained with the data until Dec 2023. We also study the recently introduced **GPT-4o** model.

- **Gemini-1.5** is another proprietary LLM withh a context window of 10M tokens (Reid et al., 2024). We used the *gemini 1.5 pro* version which is built on top of mixture-of-experts transformer-based architecture.

- **Claude-3** is yet another proprietary model released by Anthropic with a context window of length 200K (Anthropic, 2024) but is claimed to accept up to 1M tokens. We used *claude-3-sonnet-20240229* variant of the Claude family.

- **YaRN-Llama-2-7b** (Peng et al., 2024) is an open-source model that extends Llama 2 model (Touvron et al., 2023) to accept a larger context window. It is available in different model sizes with varying context windows. We selected the 7B model with the maximum context window of 128K tokens, accessed via Huggingface[4].

- **Llama-3-8B** is a robust open-source model (Dubey et al., 2024). We selected the instruction-tuned version of the model with a context window of 8k[5].

3.2 Retrieval and Reasoning Tasks

Language models with the ability to handle long context rely heavily on their capacity to retrieve relevant information from the given text and reason based on that information to interpret and follow human instructions effectively. Although synthetic tasks alone may not provide a comprehensive assessment of a language model's long-context capabilities, they offer the advantage of being easily adaptable to specific scenarios and languages. This is particularly important given that the most recent long-context real-world benchmarks are limited to English (An et al., 2024) or bilingual English/Chinese (Bai et al., 2023; Qiu et al., 2024; Yuan et al., 2024). Moreover, there is some evidence to suggest that performance on synthetic retrieval tasks can, to a certain extent, generalize to real-world datasets (Qiu et al., 2024). Therefore, carefully designed synthetic tasks can serve as a valuable tool for evaluating a language model's long-context capabilities across a diverse range of languages.

The "needle in a haystack" task (Kamradt, 2023), similarly to the passkey retrieval task (Mohtashami and Jaggi, 2023; Chen et al., 2023; Liu et al., 2024), evaluates a model's ability to extract relevant information from lengthy documents. Typically, a target sentence (the "needle") is inserted into a corpus of documents (the context or "haystack"), followed by a question designed to retrieve the fact in the needle. As the input text grows longer, this task typically becomes increasingly challenging.

[2]The linguistic diversity taxonomy (Joshi et al., 2020) is used to identify the resource levels.

[3]Language families were obtained from Ethnologue: Languages of the World, available at https://www.ethnologue.com/.

[4]https://huggingface.co/NousResearch/Yarn-Llama-2-7b-128k

[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

We can formalize the problem of needles in a haystack as follows: Given the needle n, a context (or haystack) c, and a question q, the model is expected to generate an answer a. Usually, n, q, and a are short, while c represents a long sequence of text that can span thousands of tokens. The task can involve either a single needle $n = 1$ or multiple needles $n > 1$. With one or more needles, we can create **retrieval** tasks, whereas with multiple needles we can construct **reasoning** tasks that require the model to draw connections between different pieces of information.

3.2.1 Retrieving a Needle ($n = 1$)

In this task, the model's objective is to locate and extract information from a single target sentence hidden somewhere in the haystack. We adopt the same needle pattern as used in previous studies (Dhinakaran, 2024; Reid et al., 2024; Anthropic, 2024), which takes the form: "*The special magic {city} number is: {number}*". Here, {city} is randomly chosen from a list of 69 unique cities from around the world, and {number} is a randomly generated 7-digit number. The list of cities were automatically translated and then post-edited into all the languages.

In English, this yields needle sentences such as "*The special magic Paris number is: 2243738*" or in Indonesian, "*Nomor ajaib khusus untuk kota Sydney adalah 9347172*".

The needle is then placed at different depths within the context. We experiment with five depth positions: 0% (near the beginning), around the 25% mark, 50% (in the middle), about 75% of the way through, and 100% (towards the end of the context). The needle is placed after the first complete sentence at each specified depth to ensure a linguistically meaningful position. Finally, the model is asked to retrieve some information (e.g., the magic number or the city) found in the needle. It is worth noting that all languages in this study use the same Hindu-Arabic numeral system.

3.2.2 Reasoning over Multiple Needles ($n > 1$)

In real-world applications, tasks often require not just accurate text retrieval but also the ability to reason with the recalled information. To increase the challenge, we introduce a setup where multiple needles are placed within the context, requiring the model to track and reason about these different pieces of information.

The needle format remains similar to the one used in retrieval task. We discretize the positions of the target needles into four intervals: near the top (0-25%), in the middle (25-50% and 50-75%), and closer to the end (75-100%) of the context. For instance, in the 25-50% bucket, the first needle is placed around the 25% depth and the remaining needles are randomly placed between somewhere within the 25-50% depth. We explore two variations of this task, with $n = 2$ and $n = 3$. Finally, the model is asked to generate a response based on the information (e.g., the larger or the largest magic number, or the city with the larger/largest number) found in the needles.

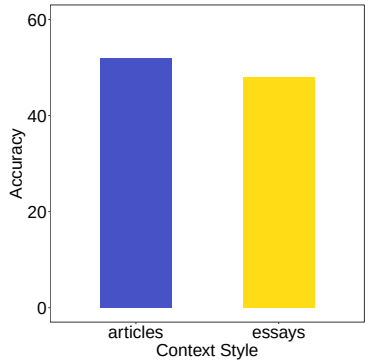

Figure 1: Ablation results of comparing Paul Graham's essays and news articles serving as haystacks for English experiments tested using GPT-4 model.

3.3 Creating `mLongRR` Dataset

Prior work has extensively relied on Paul Graham's essays in English to make up the haystacks (Kamradt, 2023; Dhinakaran, 2024; Anthropic, 2024). Translating these essays into multiple languages, however, could potentially introduce translation errors. Instead, we create a new dataset – `mLongRR` – by collecting BBC news articles published in the five languages[6], inspired by recent work (Bagheri Nezhad and Agrawal, 2024a). This approach allows us to work with sufficiently long, naturally occurring, with the added benefit that this recently published data is less likely to have been encountered by the models during training. It is worth noting that this does not result in a parallel dataset as the news articles are often specific to their respective regions.

We assess the impacts of different data for haystacks by conducting an ablation study using the GPT-4 model. As shown in Figure 1, we observed

[6]Our language selection was primarily constrained by the availability of authentic texts, especially those within the same script but varying in resource levels.

no noticeable differences in the model performance when the haystacks consisted of Paul Graham's essays or news articles in English. To statistically confirm this observation, we applied McNemar's test (McNemar, 1947), which yielded a p-value of 1.0, indicating no significant difference between the two datasets.

The haystacks in mLongRR were created by drawing on sufficient numbers of articles to fill up the target context window length. There are enough articles in each language to provide non-repeating text for all window lengths. Thus, for example, the first half of the 8K haystack is the same as the 4K haystack, but its second half is composed of different articles. An example of an English haystack of 8K tokens with the needle "*The special magic Doha number is 9121372.*" located at 50% depth is shown below:

```
Star dunes - or pyramid dunes - are
named after their distinctive...
[continues to about 4K tokens]
The special magic Doha number is
9121372.
In our dark laboratory, we see light
from these sand grains...
[context continues until 8K tokens]
```

3.4 Prompts

In our initial run of experiments, we explored two existing prompt templates used in previous work:

- **prompt 1** (Kamradt, 2023; Dhinakaran, 2024; Anthropic, 2024), and
- **prompt 2** (Reid et al., 2024).

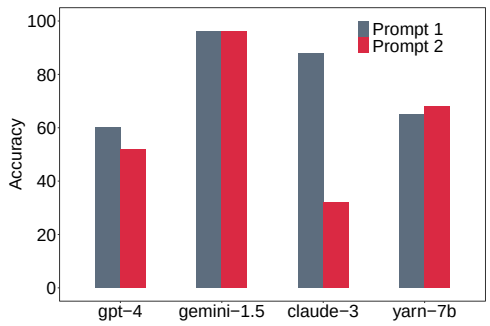

Figure 2: Ablation results of comparing two different prompts.

As shown in Figure 2, prompt 1 performed better in most of the models, so we decided to continue with it for the rest of the experiments. The prompt for the $n = 1$ retrieval task is as follows:

```
You are a helpful AI bot that answers
questions for a user. Keep your response
short and direct. The following is a
set of context and a question that will
relate to the context.
#CONTEXT
{context}
#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.
```

For the reasoning with multiple needles task, the prompt format remains largely similar except the question. Recall that each needle has two random variables: the number and the city. Therefore, we explore two possible reasoning questions, one about the larger number, and the other about the city with the larger number.

- For reasoning about numbers, the question is either "*What is the larger magic number?*" when $n = 2$, or "*What is the largest magic number?*" when $n > 2$.
- For reasoning over cities, the question is either "*Which city has the larger magic number?*" when $n = 2$, or "*Which city has the largest magic number?*" when $n > 2$.

Although the articles in mLongRR are already in the native languages, we hired four professional human translators from the freelancing website fiverr.com to translate the needles, city names, and prompts from English into Vietnamese, Indonesian, Swahili, and Somali. In our preliminary experiments with the GPT-4 model, we experimented with English and language-specific prompts, and found that English prompts worked better than language-specific prompts, in line with previous studies (Etxaniz et al., 2024; Bareiß et al., 2024; Lai et al., 2023b), for three of four non-English languages (with the exception of Swahili). As a result, the remaining experiments were conducted using English prompts.

3.5 Experiments

We conducted experiments on six models: GPT-4, Gemini-1.5, Claude-3, Yarn-7b, Llama 3, and GPT-4o. The context lengths varied from $2k$, $8k$, $16k$,

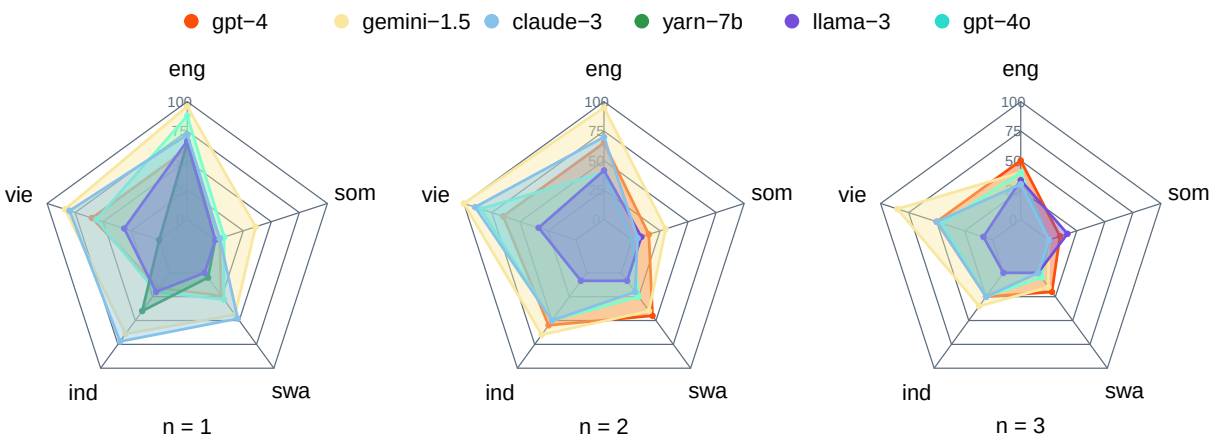

Figure 3: Radar plots showing the performance of six language models (GPT-4, Gemini-1.5, Claude-3, Yarn-7b, Llama-3, GPT-4o) across five languages (English, Vietnamese, Indonesian, Swahili, Somali) in retrieval and reasoning tasks involving one, two, and three target sentences ("needles"). The three plots represent different task complexities: single needle retrieval ($n = 1$, left plot), two needle reasoning ($n = 2$, center plot), and three needle reasoning ($n = 3$, right plot).

$32k$, to $64k$ tokens, and the needles were placed at different depths/positions: 0%, 25%, 50%, 75%, and 100%. For the retrieval task, we experiment with one needle ($n = 1$), whereas for the reasoning tasks, we investigate setups of needles $n = 2$, and $n = 3$. To enhance the robustness of our evaluation, we used a diverse corpus of recently published news articles and a combination of random cities and random numbers resulting in a vast number of possible needle variations. Furthermore, we conducted multiple runs for a subset of our experiments and consistently observed a variance close to 0 across these runs. Each model was evaluated using its default configuration, and the maximum output token size was set to 50.

3.6 Evaluation

The responses generated by the models were used to calculate the accuracy. For both the retrieval and reasoning tasks, the models generated a short, straightforward text containing the 7-digit number (for number-based reasoning) (Dhinakaran, 2024) or the city name (for city-based reasoning). For example, a typical output looked like this: `"3210496"` or `"The larger magic number is 8134445"`. We extracted the number/city and compared it to the ground truth to check whether the model's response was correct or not. For languages other than English, the models occasionally responded with the city name in English or the target language and both were acceptable.

4 Results and Discussion

This section presents our results of four main experiments: (1) performance of different models, (2) performance with respect to varying needle depths and haystack lengths, (3) performance across five different languages, and (4) reasoning over magic numbers and world cities. For the first three experiments, we analyze the models' responses when asked to retrieve and reason about the magic number. In the last experiment, we compare the models' performance when asked about the magic number or the city.

4.1 Performance of different models across languages and tasks

Figure 3 presents the radar plots summarizing the the average accuracy of each model for all tasks and languages.

Across the languages evaluated, English generally demonstrates strong performance across all models and tasks, particularly in the simpler retrieval task ($n = 1$), likely due to the extensive amount of English data available for model training. Vietnamese also performs relatively well, especially in the more complex reasoning tasks ($n = 2$ and $n = 3$), which may be attributed to effective tokenization (more discussion in section 4.3). In contrast, performance drops significantly for Indonesian, Swahili, and Somali, particularly as task complexity increases. While this decline is not surprising and highlights the ongoing challenge in multilingual NLP models trained predominantly on

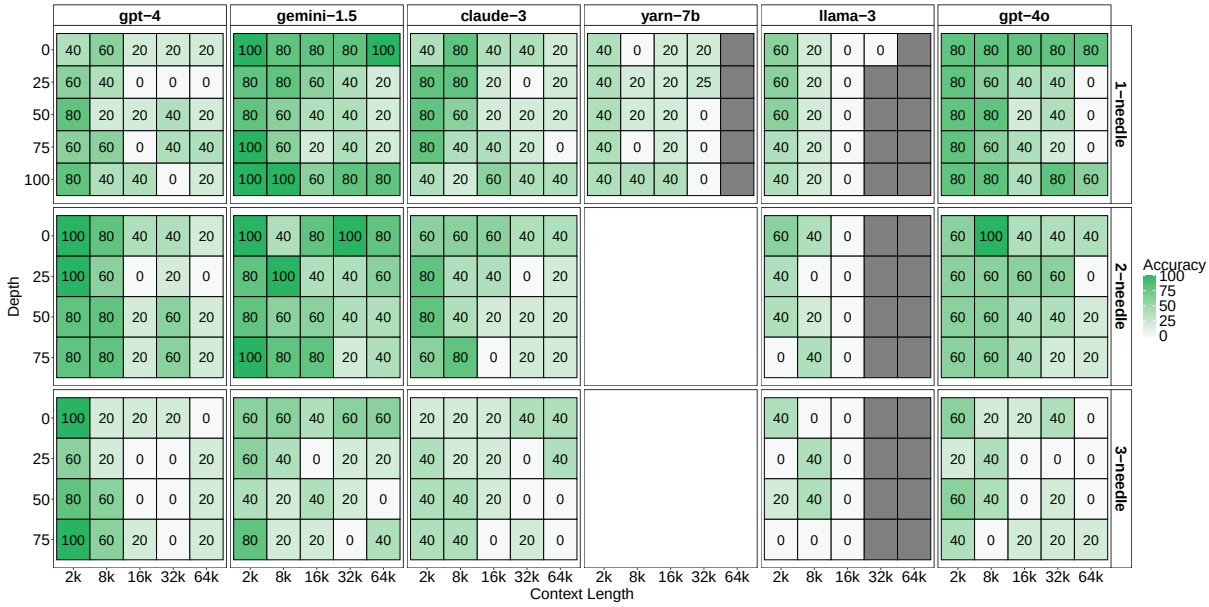

Figure 4: Heatmap visualizations with varying depths on the y-axis and context lengths on the x-axis, showing average model performance over all the languages for both retrieval (top panel) and reasoning tasks (middle and bottom panels). The color gradient from white to dark green represents accuracy levels, with darker green indicating higher accuracy.

high-resource languages tending to perform well in those languages but faltering in low-resource languages, the extent of the decline remains noteworthy.

Gemini-1.5 and GPT-4o exhibit strong performance across all tasks and languages, maintaining the most balanced results overall, particularly in English and Vietnamese. However, their performance declines in more complex tasks for low-resource languages like Swahili and Somali. In contrast, other models display more variability, with certain strengths in specific languages but generally lower performance in reasoning tasks, particularly when multiple needles are involved.

As task complexity increases (from $n = 1$ to $n = 3$), extending from retrieval to reasoning, all models experience a performance drop, particularly in low-resource languages. This indicates that while models can handle simple retrieval tasks reasonably well, they struggle significantly with reasoning tasks that require understanding and processing long contexts in less-resourced languages.

4.2 Performance across varying needle depths and haystack lengths

Figure 4 presents a detailed heatmap analysis of each model's performance with varying context lengths and needle depths. For all models, performance is better in shorter contexts, or when the needle is either near the top or the bottom of the context, suggesting that the "lost in the middle" phenomenon which was previously observed in English settings (Liu et al., 2024) extends to multilingual contexts as well.

The heatmaps clearly show that longer context lengths and greater depths negatively impact accuracy. This suggests that current LLM architectures struggle to use relevant information effectively when processing large amounts of data or when reasoning requires multiple steps. As the task complexity increases (from retrieval to 3-needle reasoning), model performance declines across the board. This decline is particularly pronounced in models like Yarn-7b and Llama-3, which fail to handle the increased cognitive load of deeper reasoning tasks with longer contexts. Gemini-1.5 is the most resilient model across all tasks, maintaining relatively high accuracy even in complex scenarios. However, its performance also suffers as depth and context length increase, highlighting the challenges of scaling reasoning abilities in LLMs.

4.3 Performance across different languages

The results presented in Figure 5 provide a fine-grained analysis of language-specific performance. English consistently performs well across tasks,

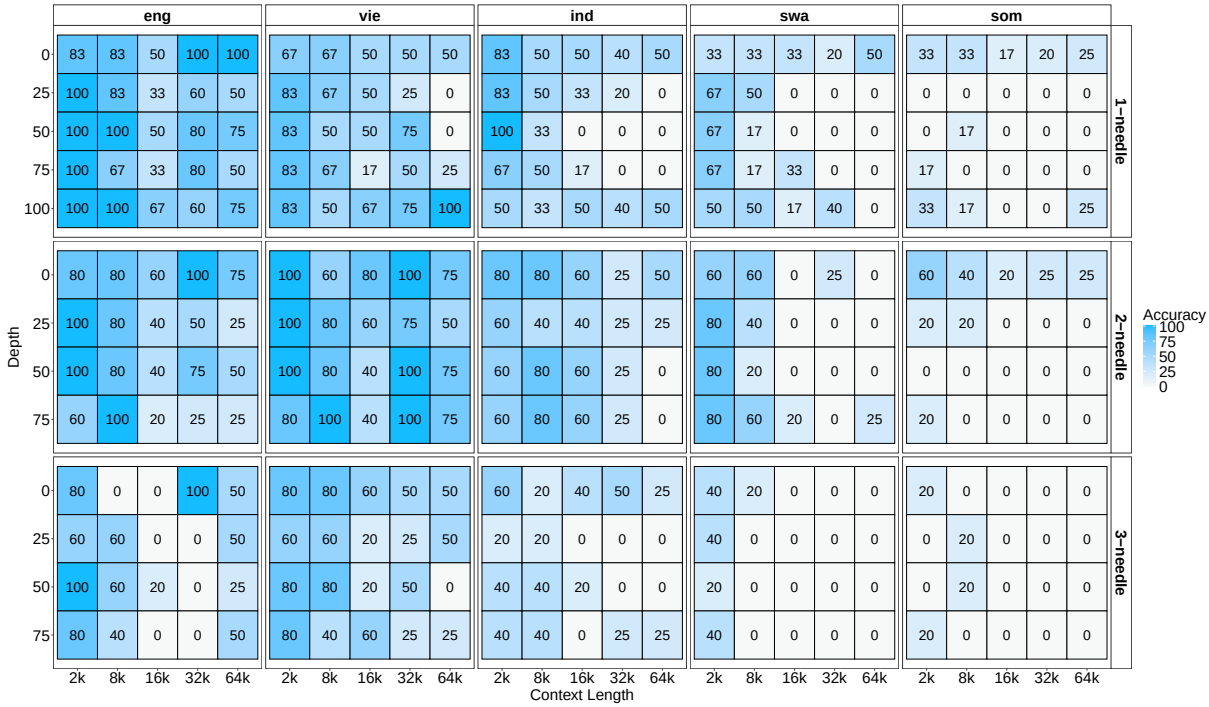

Figure 5: Language-specific heatmap visualizations with varying depths on the y-axis and context lengths on the x-axis, averaged over all the models, when $n = 1$, $n = 2$, and $n = 3$.

	GPT-4	Gemini-1.5	Claude-3	YaRN-7b	Llama-3	GPT-4o
English	1.13	1.15	1.15	1.32	1.13	1.11
Vietnamese	2.08	1.20	2.89	2.75	1.27	1.29
Indonesian	1.92	1.40	2.33	2.48	1.91	1.55
Swahili	2.23	1.85	2.36	2.48	2.21	1.68
Somali	2.37	2.09	2.47	2.70	2.36	1.79
Average	1.94	1.53	2.24	2.34	1.77	1.48

Table 2: Tokenization rate for each language using different model tokenizers.

with near 100% accuracy in simpler tasks but declining with increased complexity, particularly at greater depths and longer contexts. Vietnamese also maintains high accuracy, though it declines similarly with complexity. Indonesian starts reasonably well but drops significantly in more complex scenarios. Swahili shows weaker overall performance, struggling with all tasks, especially complex ones. Somali performs the poorest, often reaching zero accuracy as task complexity increases, highlighting challenges in handling this low-resource language. In short, performance degrades progressively as we move from high-resource languages to low-resource languages. The detailed results of each model and language are included in Appendix A.

The strong performance of English and, to a lesser extent, Vietnamese, reflects the availability of ample training data in these languages. However, access to the exact language distributions of training data are not readily available for most models, including open-source model like Llama-3.

4.4 Impact of tokenization

We further analyze the tokenization rate, also known as fertility rate, which is the average number of tokens generated per word for the different languages. The results are presented in Table 2. Unsurprisingly, English consistently shows the lowest tokenization rates across all models. Vietnamese has varying rates, with Gemini-1.5 being the most efficient, while Claude-3 and YaRN-7b tokenize more heavily. Indonesian exhibits moderate rates with some variability across models. Swahili and

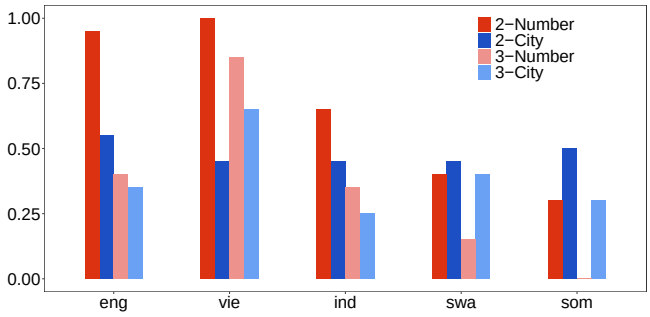

Figure 6: Comparing reasoning over magic numbers and random cities, when $n = 2$ and $n = 3$ (results obtained using Gemini-1.5).

Somali, the two lowest resource level languages in our study, generally have higher tokenization rates, suggesting these are more challenging for models to process effectively.

We can make two interesting observations: (i) the performance of LLMs is influenced by the way the models tokenized text across languages, with lower fragmentation leading to improved performance, and (ii) the models with overall lower fragmentation scores, such as Gemini-1.5 followed by GPT-4o, achieved better results across all languages and tasks.

4.5 Reasoning about magic numbers and world cities

Lastly, we compare the performance of the models in reasoning tasks with 2 and 3 needles for two types of question prompts: identifying the larger/largest magic number (e.g., 4281932) or the city with the larger/largest magic number (e.g., Doha). From the results presented in Figure 6, we observe that the models yield generally better performance in the "number" tasks compared to the "city" tasks implying that they may be more adept at handling numerical reasoning than reasoning over geographic entities, however, this trend is reversed for Swahili and Somali.

5 Conclusion

We introduce a new dataset designed to study long-context retrieval and reasoning tasks across multiple languages. By evaluating six LLMs on their ability to process text in five languages with varying resource levels, using naturally occurring text and a needle-in-a-haystack paradigm with different numbers of needles, we discovered key insights. Notably, we observed a significant decline in performance, particularly when dealing with longer contexts, an increased number of needles, or lower resource levels. Even seemingly simple synthetic tasks like needle-in-a-haystack revealed substantial performance disparities. Our findings highlight the need to develop not only more effective long-context models but also improved tokenization schemes for the effective processing of low-resource languages.

Limitations

While our current focus has been on languages that use Latin script, we are eager to expand our horizons and explore the diversity of languages from other scripts in the future. Furthermore, our investigation was restricted to three needles. It would be interesting to explore whether addition of more needles continues to increase the task complexity.

Ethics Statement

We did not implement any filtering of the haystack data, it is possible that there are inherent biases towards certain groups within the dataset. The impact of such biases on our findings remains unclear and fall outside the scope of this study. For annotation in Vietnamese, Indonesian, Swahili, and Somali, we hired translators and paid USD 15 to each translator as the short translation tasks took less than one hour each.

Acknowledgments

We thank the anonymous reviewers as well as the members of PortNLP lab for their insightful comments that helped improve this paper. This research was partially supported by the National Science Foundation grant HNDS-R 2242205.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. [Online; accessed 20-March-2024].

Sina Bagheri Nezhad and Ameeta Agrawal. 2024a. Exploring the maze of multilingual modeling. *Preprint*, arXiv:2310.05404.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024b. What drives performance in multilingual language models? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Aparna Dhinakaran. 2024. Llmtest_needleinahaystack. [Online; accessed 20-March-2024].

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

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

Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook

Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Greg Kamradt. 2023. Llmtest needleinahaystack: Doing simple retrieval from llm models. `https://github.com/gkamradt/LLMTest-NeedleInAHaystack/tree/main`. [Online; accessed 20-March-2024].

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023b. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 54567–54585. Curran Associates, Inc.

Arka Pal, Deep Karkhanis, Manley Roberts, Samuel Dooley, Arvind Sundararajan, and Siddartha Naidu. 2023. Giraffe: Adventures in expanding context lengths in llms. *arXiv preprint arXiv:2308.10882*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Zexuan Qiu, Jingjing Li, Shijue Huang, Wanjun Zhong, and Irwin King. 2024. Clongeval: A chinese benchmark for evaluating long-context large language models. *arXiv preprint arXiv:2403.03514*.

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

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2668–2680, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eline Visser. 2020. Kalamang dictionary.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. 2024. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *Preprint*, arXiv:2402.05136.

Judit Ács. 2019. Exploring bert's vocabulary.

A Detailed model and language-specific results

Figures 7, 8, 9, 10 and 11 show detailed results of the five models: GPT-4, Gemini-1.5, and Claude-3, Llama 3, and GPT-4o.

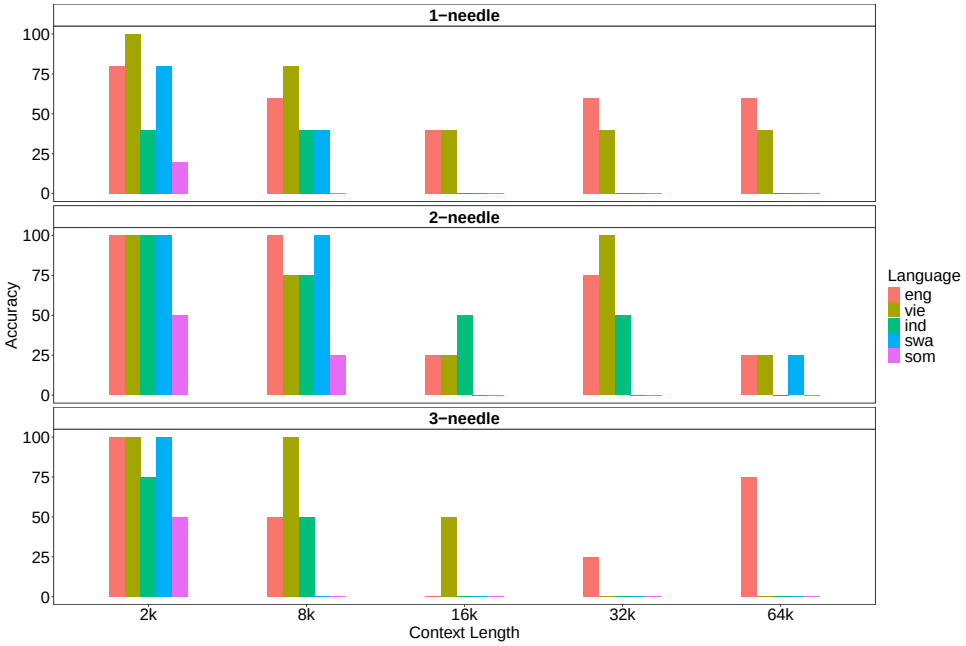

Figure 7: GPT-4

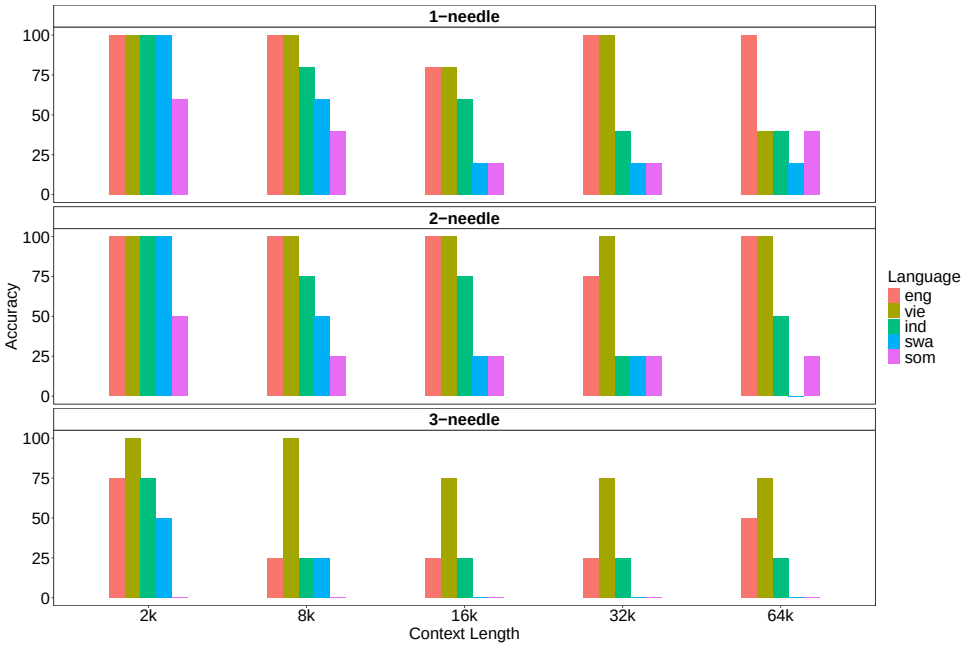

Figure 8: Gemini-1.5

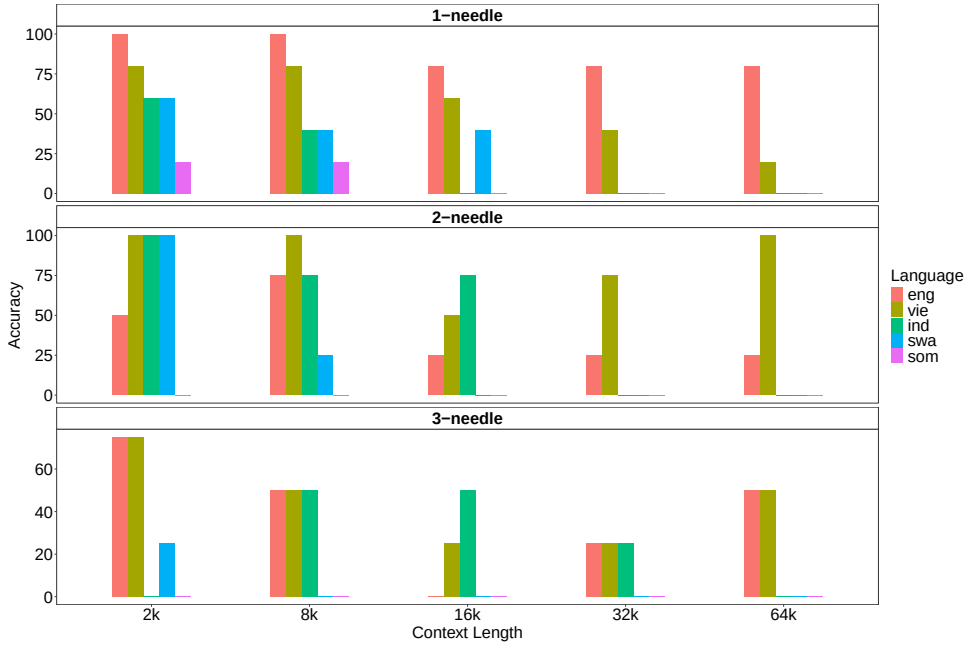

Figure 9: Claude-3

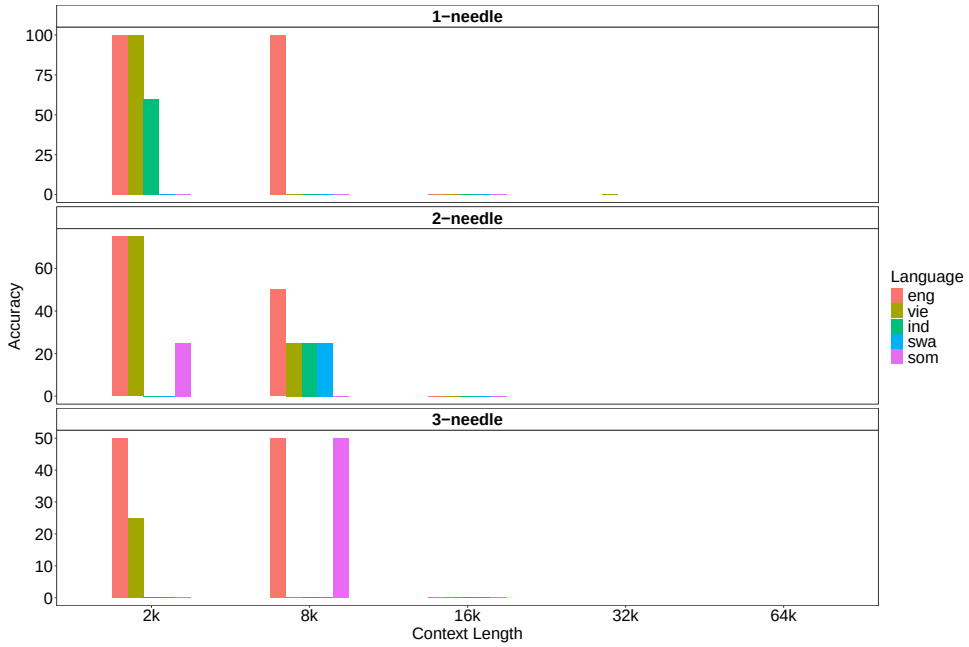

Figure 10: Llama 3

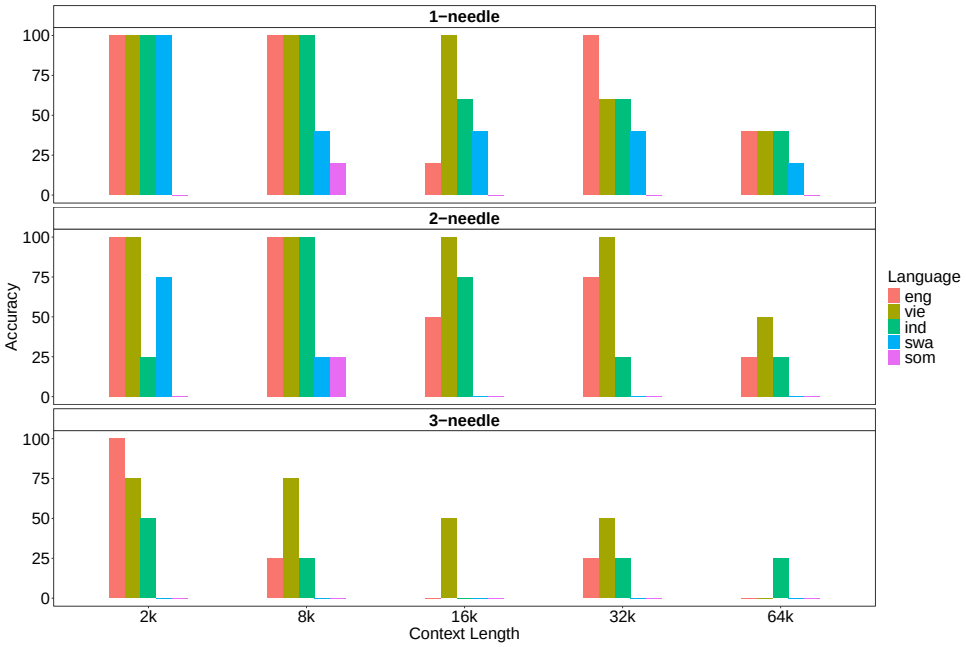

Figure 11: GPT-4o