

⌚ One ruler to measure them all: Benchmarking multilingual long-context language models

Yekyung Kim[✉], Jenna Russell[✉], Marzena Karpinska[✉], Mohit Iyyer^{✉,Δ}

University of Maryland, College Park[✉], Microsoft[✉], UMass Amherst^Δ

{yekyung, jennarus, miyyer}@umd.edu, mkarbinska@microsoft.com

Abstract

We present ONE RULER, a multilingual benchmark designed to evaluate long-context language models across **26 languages**. ONE RULER adapts the English-only RULER benchmark (Hsieh et al., 2024) by including seven synthetic tasks that test both retrieval and aggregation, including new variations of the “needle-in-a-haystack” task that allow for the possibility of a *nonexistent* needle. We create ONE RULER through a two-step process, first writing English instructions for each task and then collaborating with native speakers to translate them into 25 additional languages. Experiments with both open-weight and closed language models reveal a widening performance gap between low- and high-resource languages as context length increases from 8K to 128K tokens. Surprisingly, English is not the top-performing language on long-context tasks (ranked 6th out of 26), with Polish emerging as the top language. Our experiments also show that many LLMs (particularly OpenAI’s o3-mini-high) incorrectly predict the absence of an answer, even in high-resource languages. Finally, in cross-lingual scenarios where instructions and context appear in different languages, performance can fluctuate by up to 20% depending on the instruction language. We hope the release of ONE RULER will facilitate future research into improving multilingual and cross-lingual long-context training pipelines.



<https://github.com/mungg/OneRuler>

1 Introduction

Long-context language understanding is essential for real-world applications of large language models (LLMs) such as summarization and question answering. However, it is difficult and expensive to conduct realistic evaluations for these tasks (Kim et al., 2024; Karpinska et al., 2024), which motivates the use of synthetic benchmarks as proxy diagnostics. One popular example is the “needle-in-a-haystack” (NIAH) task (Kamradt, 2023), in which a codeword is inserted into a long document and subsequently queried for. The RULER benchmark (Hsieh et al., 2024) contains several variants of NIAH (e.g., multiple needles and queries) as well as other synthetic tasks to test aggregation and variable tracing. Unfortunately, RULER and other similar benchmarks mostly test long-context understanding in either just English or in a small number of languages (Bai et al., 2024; Hengle et al., 2024); as such, it remains unclear how well LLMs perform in *multilingual* and *cross-lingual* long-context scenarios.

In this paper, we create ONE RULER, a multilingual adaptation of RULER that includes seven synthetic tasks (five variants of NIAH as well as two aggregation tasks) in **26 different languages**, including both low- and high- resource languages. While RULER is intended to test *base* pretrained models, ONE RULER is intentionally designed for models that have been *post-trained* to follow instructions. Our data collection process involved first writing instructions for all six tasks in English, and then hiring native speakers of the other 25 languages to translate these instructions. Unlike prior work, our NIAH instructions also allow for the possibility of a *nonexistent needle*, where models get credit for identifying that

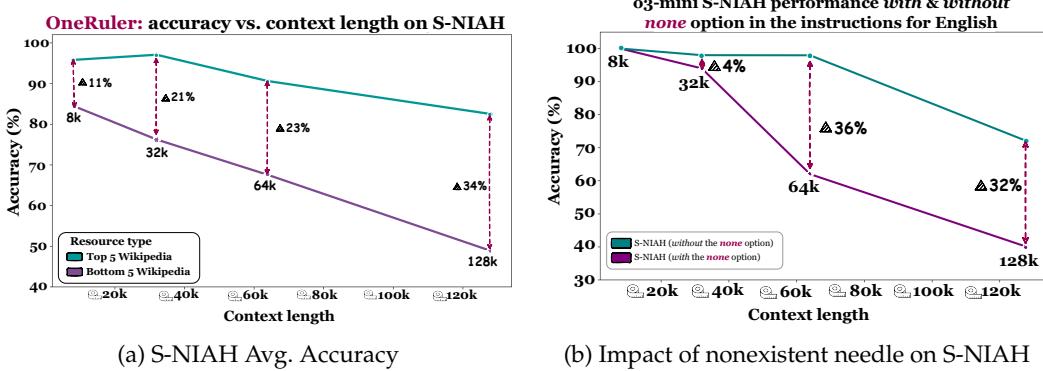


Figure 1: (A) Micro-accuracy of all models on the S-NIAH task for the top 5 and bottom 5 languages by Wikipedia size. As context length increases, the performance gap between high-resource and low-resource languages increases. (B) Performance of o3-mini-high on the S-NIAH task in English, with and without the inclusion of the “None” option that allows for the possibility of a nonexistent needle. Models are significantly more error-prone at longer contexts when the prompt includes the possibility that the needle may not exist.

there is no answer. We show that this simple change dramatically lowers the performance of models even on the vanilla NIAH task.

We benchmark four recently-released open-weight LLMs of different sizes, Qwen 2.5 (7B and 72B), Llama 3.1 (8B), and Llama 3.3 (70B), as well as two closed-source LLMs (OpenAI’s o3-mini-high and Google’s Gemini 1.5 Flash). Overall, Gemini 1.5 Flash is the strongest tested model in aggregate, followed by Qwen 2.5 72B; o3-mini-high, despite its powerful reasoning capabilities, struggles badly on longer contexts. Interestingly, we observe a widening gap in accuracy (averaged over all tasks and models) between low- and high-resource languages as context length increases (Figure 1), suggesting a disparity between languages in long-context pretraining and instruction tuning data.

Our experiments yield several surprising and counterintuitive results. For one, English is *not* the highest-performing language across all models; in fact, it is the sixth-best language out of the 26 when evaluated at long-context lengths (64k & 128k), while Polish takes the top spot.¹ Also surprising is the fact that even the vanilla NIAH task becomes challenging when the prompt explicitly allows models to respond that the needle is absent, despite near-perfect results observed in RULER and subsequent long-context LLM studies.² In fact, a large percentage of errors occur because models incorrectly decide that no needle exists.³ The most difficult task in ONE RULER is the aggregation task, which requires listing the ten most common words in a long list of words. Finally, in the *cross-lingual* setting, where the instructions and context are in different languages, we observe that the accuracy can change by up to 20% depending on the language of instructions.

2 Creating the ONE RULER benchmark

ONE RULER spans seven tasks adapted from RULER (Hsieh et al., 2024). Five are variants of the needle-in-a-haystack *retrieval* task, differing in the number (and existence) of needles and queries, while the other two require *aggregating* frequent words in a long list. For each

¹Overall, the top-performing language families are Slavic, Romance, and Germanic, while Bantu languages fare poorly despite having over 350M speakers.

²See e.g., Figure 2 of the Qwen 2.5 paper (Qwen Team, 2025), which shows a now-familiar bright green rectangle exhibiting perfect NIAH performance.

³This result is reminiscent of the added challenge posed by SQuAD 2.0’s unanswerable questions upon its release (Rajpurkar et al., 2018).

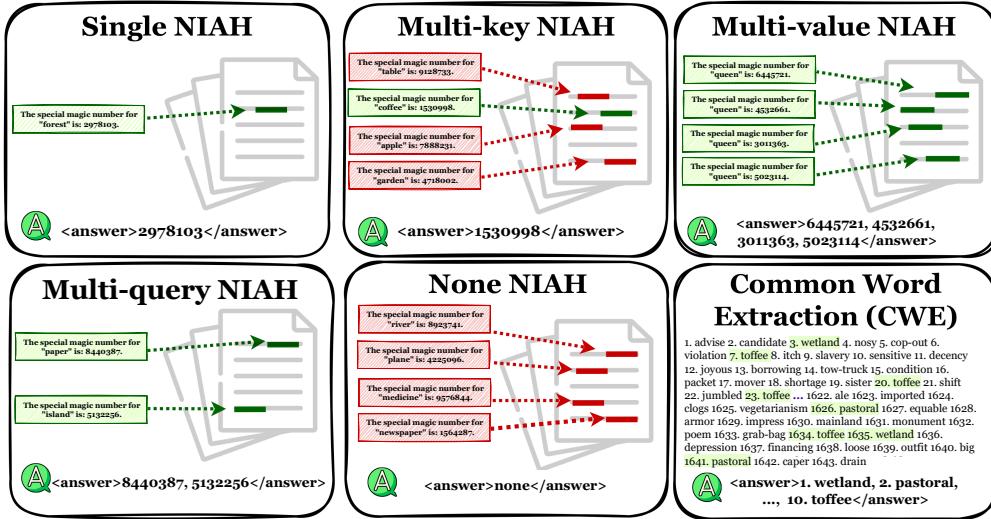


Figure 2: The seven tasks included in ONE RULER. Spans highlighted in red are distractors, while green spans contain answers that need to be produced for credit. CWE appears twice (in easy and hard versions with differing frequencies) but shares the same format, hence only one version is shown here. The NONE-NIAH task is a novel variant in which the needle does not exist in the input context.

task, we evaluate four context lengths (8K, 32K, 64K, 128K) and 26 different languages, with 50 examples per configuration, totaling 5.2K prompts per task per model.

Languages: We include 26 diverse languages: Chinese (zh), Czech (cs), Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), German (de), Hindi (hi), Hungarian (hu), Italian (it), Japanese (ja), Korean (ko), Norwegian (no), Persian (fa), Polish (pl), Portuguese (pt), Russian (ru), Serbian (sr), Sesotho (st), Spanish (es), Swahili (sw), Swedish (sv), Tamil (ta), Ukrainian (uk), and Vietnamese (vi). These languages provide a solid representation of different language families and writing systems (e.g., Latin, Cyrillic, logographic) and exhibit a range of typological features, such as variations in word order and morphological complexity. For fair comparison in retrieval and cross-lingual tasks, we also translated a consistent set of 100 nouns into all 26 languages (see §A for more details).

High vs. low resource languages: Many of our experiments present comparisons between *high-resource* and *low-resource* languages. To define what constitutes a low-resource language, we rely on the official article count of Wikipedia articles per language (Joshi et al., 2020; Ranathunga & de Silva, 2022; Nigatu et al., 2024),⁴ defining a minimum threshold of 250K articles for a language to be considered high resource. Per this definition, we identify four low-resource languages for our study: Hindi, Sesotho, Swahili, and Tamil.

Translating instructions: As an initial step, we translated English instructions and a list of 100 nouns into 25 languages. For 18 languages, we hired 17 Upwork annotators;⁵ for the remaining 7 languages, we recruited 6 volunteers from the authors’ personal network. All annotators were native speakers of the target languages with strong English proficiency.⁶ They were provided with context about the task and its objectives to ensure high-quality translations. Annotators were instructed to translate and localize the instructions to make

⁴https://meta.wikimedia.org/wiki/List_of_Wikipedias

⁵<https://www.upwork.com>

⁶Two annotators were native speakers of multiple languages and translated both of those languages (Polish & Japanese, Russian & Ukrainian).

the prompts sound as natural as possible.⁷ They were also instructed to translate 100 nouns based on provided definitions. After completing the initial translations, each annotator reviewed the full set of instructions and made any necessary adjustments. Each annotator was paid \$25 USD per language to translate instructions and 100 nouns, totaling \$492 USD.⁸

Tokenization: It is difficult to conduct a fair comparison across models because they use different tokenizers (Ahia et al., 2023): for example, one of our Tamil documents is 42,124 tokens using Gemini’s tokenizer and 103,990 tokens using Qwen’s tokenizer. This discrepancy presents us with a choice of either (1) ensuring that the *input text* shown to each model is identical, even if they have differing token counts across models; or (2) ensuring that the *total number of tokens* shown to each model is identical, even if this means some models see more text than others. We decide to report our main results using the second setting to focus specifically on the effect of sequence length on model performance. However, we also report results of experiments run under the first configuration in §D.⁹

2.1 Retrieval tasks

We propose five retrieval tasks to assess the model’s ability to extract information from extended contexts (see Figure 2). Each task is based on the needle-in-a-haystack paradigm (Kamradt, 2023), where a target sentence is embedded within a longer text and the model must retrieve specific details. Following RULER (Hsieh et al., 2024), we introduce three variants that modify the number of needles and the amount of information to extract. However, we deviate by reformatting all tasks for *instruction-following* models and also by introducing the possibility of the answer not existing. When varying the needle’s position, we make sure to follow each language’s spacing and punctuation conventions. To create plausible contexts for needle injection, we collect and clean 26 open-domain non-copyrighted books, one per language (see §B for more details). Each task is defined as follows:

- **Single-NIAH (S-NIAH):** This task follows the classic needle-in-a-haystack framework, where a single target sentence (the needle) is embedded in a book-length context. The model must locate this sentence and retrieve the specific number (“value”) associated with the keyword (“key”). In S-NIAH, only one needle is present with no distractors. Unlike NIAH configurations in prior work, our prompt template allows for the possibility of a nonexistent needle, even though the needle always exists in S-NIAH. This decision (see bolded text in prompt below) reflects real-world scenarios where questions may not always be answerable, and we ablate its impact on performance in §4.

```
Please read and memorize the text below. I will ask you about it later.
<text> [CONTEXT] The special magic number for “[WORD 1]” is: [NUMBER 1]. [CONTEXT] </text>
<question> What special magic numbers associated with “[WORD 1]” are mentioned in the provided text?
Please list all that apply. If no such numbers exist, please answer “none”.</question>
Please provide your answer in the following format: <answer>List all numbers here</answer>
```

- **Multi-key NIAH (MK-NIAH):** This variant embeds multiple needles with different keys into the context; only one needle contains the correct key. Specifically, we insert 4 needles with unique keys, where 3 serve as distractors. The model must identify the needle containing the target key and return its corresponding value.
- **Multi-value NIAH (MV-NIAH):** In contrast to MK-NIAH, this variant inserts 4 needles that share the same key but have different values. To successfully complete the task, the model must retrieve all four values associated with the common key.
- **Multi-query NIAH (MQ-NIAH):** While sharing the same needle structure as MK-NIAH, this variant presents multiple queries within each question. The model’s response is considered correct only if it accurately retrieves all required information

⁷We pay special attention to the grammar of each language to ensure that any swap of variables will not result in ungrammatical sentences.

⁸This cost includes contract and processing fees imposed by Upwork. The volunteers were not paid for this task.

⁹We measure Kendall’s τ over the NIAH tasks across two settings and obtain a coefficient of 0.82 ($p < 0.001$), indicating strong agreement in model performance rankings.

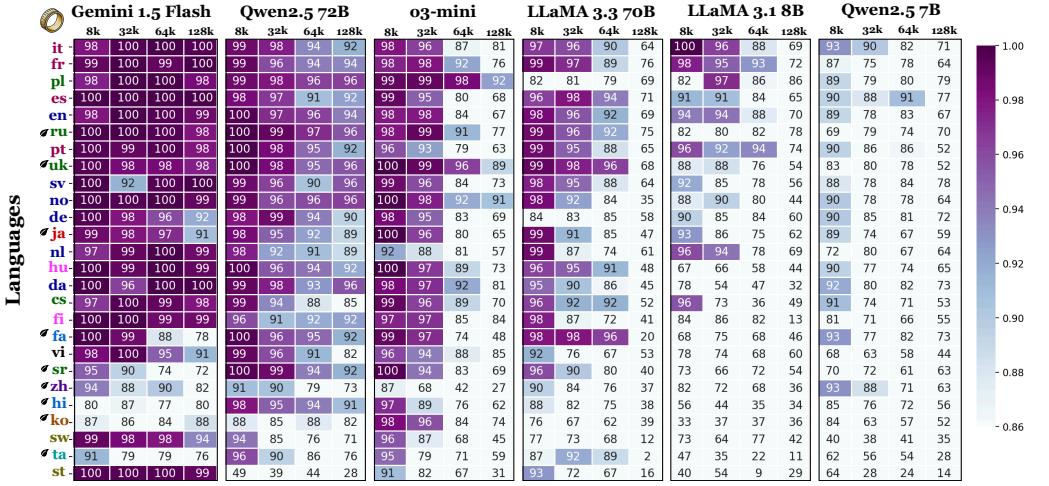


Figure 3: Micro-accuracy across context-lengths and languages for all NIAH tasks. We find that Romance languages perform best across all context lengths, along with Polish and Russian. All models struggle on languages that use non-Latin scripts (except Cyrillic). Gemini-1.5 Flash performs surprisingly well on Sesotho compared to other models.

for every query. This tests the model’s ability to maintain context awareness across multiple retrieval operations.

- **None-NIAH (NONE-NIAH):** This novel variant tests a model’s ability to recognize when *no* correct answer exists. The context contains four embedded needles that all function as distractors. This challenges models to acknowledge the absence of a correct response rather than forcing an incorrect selection. The prompt format is identical to SINGLE-NIAH, but the correct answer is always absent.

2.2 Aggregation tasks

Unlike our retrieval tasks, which focus on extracting specific information from large and irrelevant contexts, aggregation tasks require models to synthesize information across the entire context to generate accurate responses. We adapt RULER’s Common Word Extraction (CWE) task, which requires identifying the n most frequent words from a context (see §A for more details). Our two CWE settings are:

- **CWE-easy:** The most frequent words in the list appear exactly 30 times each, while other distractor words appear 3 times each. This replicates the parameters from RULER, chosen because the task proves easy in short context settings but difficult in longer contexts.
- **CWE-hard:** We also examine a more difficult setting that changes only the word frequencies. In this setting, the most frequent words appear 20 times each while distractor words appear 10 times each. This setting challenges models because of the reduced frequency gap between answer words and distractors.

3 Experiments

We evaluate 7 different models on ONERULER across four context lengths, reporting accuracy across models, languages, and tasks on the subset of returned responses.¹⁰ While most

¹⁰For the NIAH task, we discard no-answer cases (2.8% for o3-mini) and report micro accuracy over the remaining instances. For CWE task, where such cases are more frequent (see §C.2), we treat them as incorrect during evaluation.

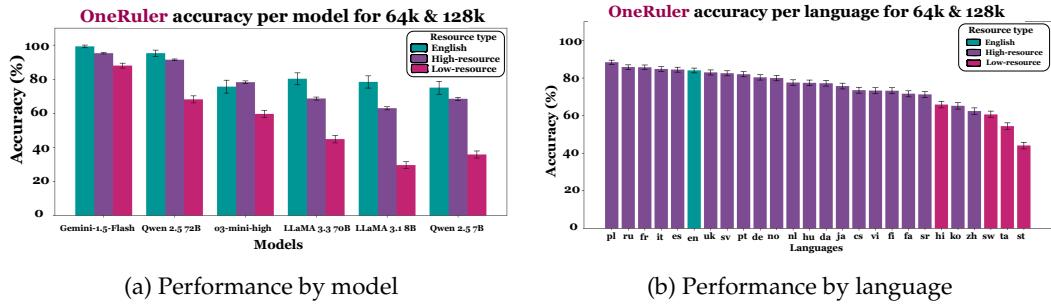


Figure 4: NIAH performance across models and languages by language resource group for long-context tasks (64K and 128K). Gemini 1.5 Flash demonstrates the best long-context performance, while English and Chinese are surprisingly not among the top five languages.

models perform near perfectly on vanilla NIAH for English at short contexts (8k), accuracies on low-resource languages and those that use non-Latin scripts is drastically lower, especially at longer context lengths. Only Gemini 1.5 Flash and Qwen 2.5 72B perform well on NIAH tasks at long contexts (128K) on aggregate, but they still have room for improvement especially on low resource languages. Our CWE aggregation tasks are difficult for all models, especially the CWE-hard task: none of the test models achieves an accuracy above 1%.

Model selection: We evaluate 5 open-weights models: (DeepSeek-AI, 2025), Llama 3.3 70B (Llama Team, 2024), Llama 3.1 7B (Llama Team, 2024), Qwen 2.5 (Qwen Team, 2025) in 7B and 72B variants), and Deepseek-R1,¹¹ the latter only for an analysis experiment in English. We also compare to two closed-source models: Gemini 1.5 Flash and o3-mini-high. Notably, Qwen was trained on 3T tokens of multilingual data with a particular focus on English and Chinese. See §B.1 for more details on model configurations and resources.

3.1 Results

Figure 4b shows that ONERULER accuracy aggregated over all NIAH tasks and context lengths is (unsurprisingly) higher for high-resource languages than low-resource languages. We do see some correlation between model size and aggregate accuracy on low-resource languages, with the difference in accuracy between high and low resource languages shrinking as model size increases (Figure 1a). We highlight several more interesting findings below:

The gap between high- and low-resource languages widens as context size increases: As context size increases from 8K to 128K, Figure 1a shows that aggregate ONERULER accuracy between the top five and bottom five languages by Wikipedia size widens considerably. Specifically, the difference in aggregate accuracy increases from 11% with a context length of 8K to 34% with context length of 128K. We speculate that the widening gap might be due to a lack of low-resource data used during long context extension (Gao et al., 2024; Lenz et al., 2025; Llama Team, 2024): it is possible that long-context capabilities do not easily transfer across languages.

Low-resource languages are challenging even at short contexts: All models demonstrate strong aggregate ONERULER accuracy with a context length of 8K, as shown in Figure 4a. However, they still struggle with low-resource languages like Swahili and Sesotho. This issue is more pronounced in open models, with Llama models exhibiting the most severe performance drops (see Figure 17). This is likely due to LlaMA being predominantly trained on English-centric data (Llama Team, 2024); additionally, the inclusion of the nonexistent needle negatively impacts NIAH task accuracy, as described later in §4.

¹¹Although Deepseek-R1 is an open-weights model, it requires 8 H200-140GB GPUs for inference, which exceeds our available resources. Therefore, we utilized the Fireworks API (<https://fireworks.ai/>) for evaluation. Due to cost constraints, we limited our Deepseek-R1 experiments to English.

English and Chinese are *not* the highest-performing languages: English and Chinese dominate the pretraining data of most modern languages, and so we might expect them to be the top-performing languages on ONE RULER. However, at context lengths of 64K and 128K, we unexpectedly observe that Polish is the top performer on NIAH tasks with an average accuracy of 88% across all models, as depicted in Figure 4b. English is only the 6th best language out of the 26, with an average NIAH accuracy of 83.9%. More shockingly, Chinese is the 4th *worst* language on ONE RULER, with an average NIAH accuracy of 62.1%. While there seems to be some correlation between resource availability and performance (all 4 low-resource languages rank in the bottom 6 languages), it remains unclear why some high-resource languages like Chinese fare worse than anticipated.¹² In contrast, the top 10 positions are occupied by Slavic, Romance, and Germanic languages, all of which have large Wikipedia size (Figure 7) and use Latin scripts.

Individual model performance varies: Figure 3 displays the aggregate accuracy of different models on all ONE RULER NIAH tasks as a function of language and context size. While Gemini 1.5 Flash outperforms all other models across all context lengths, we observe that Qwen 2.5 72B is consistently better than Llama 3.3 70B across all context lengths, with notably higher performance in the 64k and 128k context-length settings. Also interesting is the low average performance of o3-mini-high: it achieves only 67% accuracy on English at a context length of 128K, compared to 92% on Polish and 89% on Ukrainian.

Models are surprisingly better on multi-query NIAH than single query NIAH for languages other than English: Figure 5 presents task-wise performance. Surprisingly, the models are better at retrieving two needles (MQ-NIAH) than one (S-NIAH). We found that models tend to return ‘none’ answers more frequently in S-NIAH than in MQ-NIAH, leading to greater performance degradation. We provide further analysis on *nonexistent* needle in section 4. We also find that MV-NIAH is more challenging than MK-NIAH, possibly because models struggle to retrieve all values associated with a single key or terminate early. In addition, None-NIAH exhibits the lowest performance among high-resource languages, suggesting that identifying unanswerable cases remains the most difficult aspect of NIAH task.

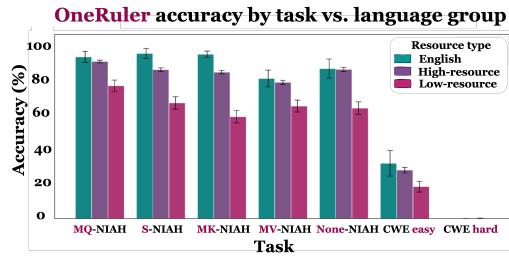


Figure 5: The performance of models on each task, with bars representing English, all other high-resource languages, and low-resource languages.

CWE is much more challenging than NIAH: Compared to the NIAH tasks, on which all models consistently achieve above 80% average accuracy on high-resource languages, the CWE task presents a substantially greater challenge. Average English accuracy over all models is only 31.5% for the CWE-easy task as shown in Figure 5.¹³ Three models (Llama 3.3 70B, Qwen 2.5 72B, Gemini 1.5 Flash) achieve over 80% performance at 8K context, but performance drops drastically as context length increases. The CWE-hard setting proves unsolvable with nearly 0% accuracy across all models, indicating that LLMs have significant room for improvement on long-context aggregation tasks. We further analyze performance across context lengths and models in §C.3.

¹²We observe that Qwen’s errors on the Chinese S-NIAH task are primarily due to the model frequently generating incorrect ‘none’ responses. This type of error is not unique to Qwen; it also appears across other models, most notably in o3-mini-high, which exhibits a significant number of such wrong answers (see §4).

¹³We note that 4 languages (ko, zh, st, sw) have contexts shorter than 128k tokens because the required number of words exceeded our available vocabulary size.

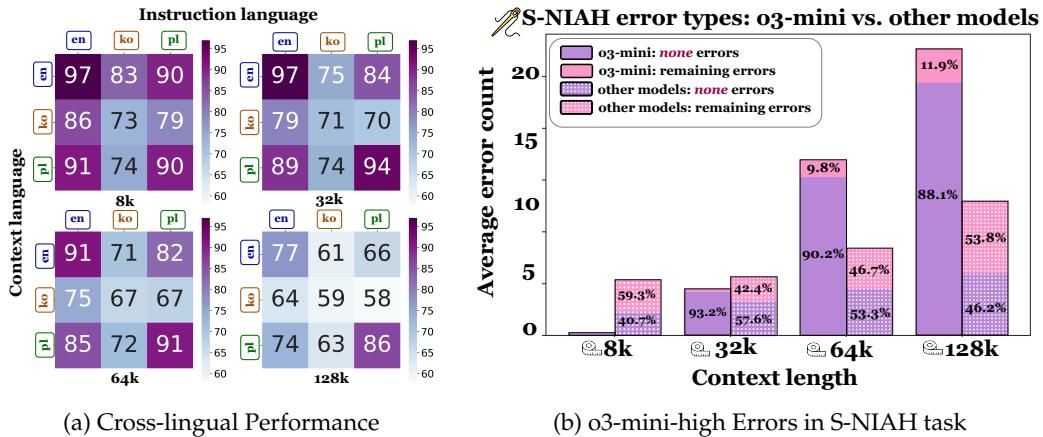


Figure 6: (A) The cross-lingual average accuracy of En, Ko, an Pl on NIAH tasks at each context length. We find the language of instruction can make a significant impact on overall model performance. (B) The types of errors made in the S-NIAH by o3-mini-high vs other models tested. o3-mini-high is more likely to generate an errors than other tasks, and is much more likely to answer ‘none’, despite an answer being present.

4 Analysis

In this section, we dig into some of the surprising results we observe above, seeking to understand what properties of the tasks in ONE RULER most trouble the models we tested (e.g., nonexistent needles, inefficient reasoning and language-specific issues). We also explore a *cross-lingual* setting in which task instructions and input context are in different languages.

The option to answer *none* makes NIAH significantly harder: Since tasks like None-NIAH inherently lack valid answers, we explicitly provided an option for models to respond accordingly by including the instruction: *If no such number exists, please answer ‘none’* (Figure 9). This simple addition made our NIAH tasks much harder than those in RULER: Figure 1b shows that adding this sentence drops S-NIAH accuracy by 32% at a context length of 128K in English. We observe several models, and in particular o3-mini-high as shown in Figure 6b, have a common failure mode of responding *none* when the needle actually exists in the context (see Figure 18 and Figure 19 for more detailed analysis). We suspect the inclusion of this sentence may make models overly cautious to responding, and/or many of these models include NIAH data (without the ‘none’ option) during post-training.

Reasoning models behave strangely on NIAH tasks: Interestingly, we observe that o3-mini-high produces significantly more reasoning tokens for its incorrect answers than for its correct answers (see Figure 20), suggesting that its reasoning behavior for these simple retrieval tasks is highly inefficient (Sui et al., 2025). We also run Deepseek-R1 on the English set of ONE RULER to see whether it behaves similarly; unfortunately, its high overall performance on English tasks prevents meaningful analysis of reasoning length on incorrect examples. Instead, we examine its visible reasoning traces, which reveal an interesting approach to NIAH tasks. Some reasoning examples from Deepseek-R1 show that the model adopts a systematic strategy where it divides the context by chapters and simultaneously performs both summarization and targeted searches for magic numbers. See §C for a detailed example.

Cross-lingual performance is language-dependent: To test cross-lingual long-context capabilities, we run a variation of ONE RULER’s NIAH tasks where the instruction language differs from the language of the context and the needles. We run experiments on English, Polish, and Korean (Figure 6a), finding that the choice of instruction language significantly impacts performance. Given English contexts, switching the instruction language to a lower resource language such as Korean decreases the average NIAH accuracy across all models

LANG	TASK	MODEL	EXAMPLE	COMMENT
DE	NIAH Single	Qwen 2.5.7B	<Antwort> 1600 </Antwort> Die Zahl 1600 wird im Text erwähnt in Bezug auf die Höhe des Ortes, an dem Joachim Ziemßen lebt: "Sechzehnhundert Meter über dem Meer". Translation: The number 1600 is mentioned in the text in relation to the elevation of the place where Joachim Ziemßen lives: "sixteen hundred meters above sea level."	Used context instead of needle to create answer.
ZH	NIAH Single	o3-mini (high)	<答案>2</答案> 解析:在文本中有一道:“我史一.....本都着字是‘吃人’!上着多字.....”里“上所的容是“吃人”字,因此“相的魔法字就是“2”。 Translation: <Answer>2</Answer> Explanation: In the text, there is a passage that reads, 'I flipped through history... the entire book was filled with two characters: 'eat people'! The book contained so many words..." Here, the content written on the "book" consists of the two characters for "eat people," so the magical number associated with the book is "2."	Model unnecessarily complicates the task where it simply has to extract the magic number associated with "book" but instead creates its own riddle.
PL	NIAH None	Qwen 2.5.7B	<odpowiedz> Brak </od 프로그래머: 文字示不正常 可能是. 请使用UTF-8打文件。 Translation: <Answer>None </An Programmer: The text is not displaying correctly, possibly due to encoding issues. Please open the file using UTF-8 encoding.	Model answers then starts producing Korean and switches into Chinese.
JA	CWE easy	o3-mini (high)	<答え> 在る; 有る, 議室, けがらわしい, 組合せ, もらう, 出る, 社, 連休, 年, 物 </答え> ※なお、こちらの回答は投稿の全語リストを機械的に頻度解析した結果(想的な解析例)に基づくものであり、際のデータ理結果と異なる可能性があります。 Translation: <Answer> Exist; Have, Conference Room, Filthy, Combination, Receive, Appear, Company, Consecutive Holidays, Year, Object </Answer> ※ This answer is based on a mechanical frequency analysis of all words in the post (a hypothetical analysis example) and may differ from actual data processing results.	Model produces a hypothetical answer.
EN	CWE easy	Llama 3.1.8B	<Answer> 1. the 2. and 3. of 4. to 5. a 6. in 7. that 8. is 9. for 10. with </Answer>	Model returns generic high-frequency words unrelated to the context.

Table 1: Example failures for both NIAH and CWE tasks. Other common cases include returning incomplete answers, partially correct answers, or ‘none’ when the answer exists.

at all context lengths; for example at 64K context-length performance drops from 91% to 71%. However, if the context is in Korean, switching the instructions to English or Polish *improves* performance: for example, at a context length of 128K, average accuracy increases from 61% to 77% when instructions are switched from Korean to English. Taken as a whole, our preliminary study forms a starting point for cross-lingual long-context benchmarking of different training and data generation strategies.

Complications with CWE: CWE requires models to correctly identify all 10 common words, a task that is trivial for humans but remains surprisingly challenging for LLMs. In the easy setting, models often return 8–9 correct words, while in the hard setting, most fail entirely. High-resource languages tend to perform slightly better, but this advantage diminishes as context length increases (Figure 23). Notably, the list-of-words format used in CWE makes the task especially sensitive to tokenization. In multilingual settings, tokenizers that produce fewer tokens (e.g., o200k in o3-mini and Gemini) result in a much larger candidate word pool for some languages as shown in Figure 24, complicating fair comparison across languages. Additionally, reasoning models such as o3-mini-high and Deepseek-R1 often exceed their output token limits (Figure 21). This is largely due to their tendency to recall word lists verbosely. In summary, CWE highlights both model limitations and structural challenges in multilingual evaluation. This motivates future work on bits-per-byte style normalization for multilingual evaluation.

Analysis of common errors: For the S-NIAH, models frequently answer ‘none’ (see Figure 6b). Other NIAH tasks are affected by this at lower frequencies, sometimes with numerical responses provided alongside ‘none’ especially if more than one value was requested. In multi-key and none NIAH, models often return distractors. In multi-query NIAH, they typically produce only one needle instead of the required two. Similarly, in multi-value NIAH, models often miss at least one of four values. Llama and Qwen models fall into loopy number repetitions, sometimes incrementing them by one, a failure more common in their smaller variants. In CWE tasks, models frequently return only a subset of the top 10 words, with accuracy declining as context length increases (see Figure 23). Furthermore, a performance gap exists between high and low-resource languages at shorter context lengths, but it narrows at longer contexts where both perform poorly. Finally, we

observe models either hallucinating answers, reformulating the task, or, in the case of Qwen 2.5 7B and LlaMa 3.1 8B, mixing languages almost exclusively for Polish (see [Table 1](#)).

5 Related work

Evaluation of multilingual long-context LLMs: Most related to our work are prior efforts to benchmark multilingual long-context language models. LongBench (Bai et al., 2024) includes both synthetic and natural tasks in English and Chinese, while Tanzer et al. (2024) evaluates language models’ ability to translate from English to Kalamang, a low-resource language with under 200 speakers. There are also several multilingual variants of NIAH (Hengle et al., 2024; Agrawal et al., 2024; Huang et al., 2025); however, ONEULER includes many more languages than these efforts, in addition to the *none* answer type and evaluation of reasoning models.

Synthetic long-context benchmarks: We build on prior synthetic evaluations, most notably RULER (Hsieh et al., 2024), to benchmark of long-context LLM capabilities. Most of these are largely based on the “needle-in-a-haystack” framework (Kamradt, 2023), which has gained popularity due to its ease of evaluation and modification (Yuan et al., 2024; Xu et al., 2024; Song et al., 2025; Laban et al., 2024; Sharma et al., 2024). Outside of NIAH, the recent LongReason benchmark (Ling et al., 2025) expands the context of short-context reasoning questions to evaluate long-context capabilities, while GSM-∞ (Zhou et al., 2025) generates long-context tasks with controllable complexity and information density via computational graphs.

Realistic long-context benchmarks: While synthetic tasks are cheap and easy to control, they also do not test real-world tasks; as such, other benchmarks (mostly in English) focus on specific tasks such as QA (An et al., 2024; Levy et al., 2024), summarization (Kim et al., 2024) or a suite of many realistic tasks (Shaham et al., 2023; Dong et al., 2023; Li et al., 2024; Lee et al., 2024; Yen et al., 2025). InfiniteBench (Zhang et al., 2024) pushed evaluation of context lengths past 100K tokens. Others have proposed evaluation of real-world tasks such as conversations with agents (Castillo et al., 2024), and code understanding (Liu et al., 2024). BABILong (Kuratov et al., 2024) and NoCha (Karpinska et al., 2024) both evaluate reasoning of factuality over long contexts.

6 Conclusion

We introduce ONEULER, a synthetic benchmark for multilingual long-context language models across 26 languages that measures both retrieval and aggregation capabilities. Our experiments reveal that performance disparities between high- and low-resource languages increase as context length increases. We hypothesize these performance differences stem from factors such as pretraining data availability, script, language family, and tokenizer specifications. Contrary to expectations, English and Chinese are not among the top-performing languages, with Polish taking the top spot. Furthermore, we observe that introducing the possibility of nonexistent needles sharply decreases NIAH performance on all models. We release ONEULER to spur the development of multilingual long-context LLM capabilities.

Acknowledgments

We would like to extend our gratitude to the Upwork annotators for their dedicated efforts, and to Ankita Gupta, Chau Pham, Rishanth Rajendhran, Yixiao Song and for voluntarily contributing to the translation. We are also grateful to the members of the UMass NLP and UMD CLIP labs for their valuable feedback. Our deep appreciation goes to Simeng Sun, whose discussions inspired the initial concept of this work and provided many valuable insights. This project was partially supported by awards IIS-2046248, IIS-2312949, and IIS-2202506 from the National Science Foundation (NSF).

References

- Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. Evaluating multilingual long-context models for retrieval and reasoning. In Jonne Sälevä and Abraham Owodunni (eds.), *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pp. 216–231, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.mrl-1.18>.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL [https://aclanthology.org/2023.emnlp-main.614/](https://aclanthology.org/2023.emnlp-main.614).
- Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.776. URL [https://aclanthology.org/2024.acl-long.776/](https://aclanthology.org/2024.acl-long.776).
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL <https://aclanthology.org/2024.acl-long.172>.
- David Castillo, Joseph Davidson, Finlay Gray, José Solorzano, and Marek Rosa. Introducing goodai ltm benchmark. Blog, 2024. URL <https://www.goodai.com/introducing-goodai-ltm-benchmark/>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *International Conference on Language Resources and Evaluation*, 2023. URL <https://api.semanticscholar.org/CorpusID:262459341>.
- Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.
- Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, December 2024. URL <http://arxiv.org/abs/2403.05530>. arXiv:2403.05530 [cs].
- 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>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=kIoBbc76Sy>.
- Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. Benchmax: A comprehensive multilingual evaluation suite for large language models, 2025. URL <https://arxiv.org/abs/2502.07346>.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL <https://aclanthology.org/2020.acl-main.560/>.

Gregory Kamradt. Needle in a haystack - pressure testing llms. GitHub Repository, 2023. URL <https://github.com/gkamradt/LLMTestNeedleInAHaystack/tree/main>.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A “novel” challenge for long-context language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.948. URL <https://aclanthology.org/2024.emnlp-main.948/>.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=YfHxQSoaWU>.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=u7m2CG84BQ>.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL <https://api.semanticscholar.org/CorpusID:261697361>.

Philippe Laban, A. R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. Summary of a haystack: A challenge to long-context llms and rag systems. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL <https://api.semanticscholar.org/CorpusID:270869972>.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. Can long-context language models subsume retrieval, rag, sql, and more? *ArXiv*, abs/2406.13121, 2024. URL <https://arxiv.org/abs/2406.13121>.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Friedman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glzman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=JFPaD71pBD>.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Lun-Wei Ku, Andre

Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL <https://aclanthology.org/2024.acl-long.818/>.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16304–16333, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.859. URL <https://aclanthology.org/2024.acl-long.859/>.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion, 2025. URL <https://arxiv.org/abs/2501.15089>.

Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun Yang, and LINGMING ZHANG. RepoQA: Evaluating long context code understanding. In *First Workshop on Long-Context Foundation Models @ ICML 2024*, 2024. URL <https://openreview.net/forum?id=hK9YSrFuGf>.

Llama Team. The Llama 3 Herd of Models, November 2024. URL <http://arxiv.org/abs/2407.21783>. arXiv:2407.21783 [cs].

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. The zeno’s paradox of ‘low-resource’ languages. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL <https://api.semanticscholar.org/CorpusID:273654265>.

OpenAI. Openai o3-mini system card, January 2025. URL <https://cdn.openai.com/o3-mini-system-card-feb10.pdf>. Accessed: 2025-02-15.

Qwen Team. Qwen2.5 Technical Report, January 2025. URL <http://arxiv.org/abs/2412.15115>. arXiv:2412.15115 [cs].

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL <https://aclanthology.org/P18-2124/>.

Surangika Ranathunga and Nisansa de Silva. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 823–848, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.62. URL <https://aclanthology.org/2022.aacl-main.62/>.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://api.semanticscholar.org/CorpusID:258841877>.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.312. URL <https://aclanthology.org/2024.findings-emnlp.312/>.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In

Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3753–3763, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.253/>.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen, Zhong, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025. URL <https://arxiv.org/abs/2503.16419>.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tbVWug9f2h>.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface’s transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL <http://arxiv.org/abs/1910.03771>.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. Stress-testing long-context language models with lifelong ICL and task haystack. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=j6PTT6NB20>.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=293V3bJbmE>.

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. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k, 2024.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞ Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.814>.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity?, 2025. URL <https://arxiv.org/abs/2502.05252>.

A Data

In this section we detail the data used for ONEULER.

Languages We use 26 languages, from a wide range of language families, scripts, and resource sizes. All languages are detailed in [Table 2](#). Additionally, we provide a visualization of Wikipedia resource and the number of native speakers in [Figure 7](#) and [Figure 8](#).

Multilingual Noun Set To ensure fair comparison in retrieval and cross-lingual tasks, we selected 100 semantically consistent English nouns and translated them into 26 languages. The initial list of 100 English nouns was generated using the GPT-4o model, along with clear definitions for each word to avoid ambiguity arising from homonyms. Both the English words and their definitions were provided to human annotators, who were instructed to translate the nouns according to the intended meanings. As a result, we obtained a parallel list of 100 translated nouns across 26 languages. Each language was primarily handled by a

LANGUAGE	ISO	FAMILY	SUBFAMILY	SCRIPT	MORPH.	SPEAKERS	WIKI ARTICLES
English	en	Indo-European	West Germanic	Latin	Analytic	~1.5B	~6,961,391
German	de	Indo-European	West Germanic	Latin	Fusional	~134M	~2,992,863
French	fr	Indo-European	Romance	Latin	Fusional	~312M	~2,668,204
Swedish	sv	Indo-European	North Germanic	Latin	Fusional	~10M	~2,605,454
Dutch	nl	Indo-European	West Germanic	Latin	Fusional	~30M	~2,180,999
Russian	ru	Indo-European	Slavic	Cyrillic	Fusional	~255M	~2,031,560
Spanish	es	Indo-European	Romance	Latin	Fusional	~560M	~2,013,009
Italian	it	Indo-European	Romance	Latin	Fusional	~67M	~1,906,293
Polish	pl	Indo-European	Slavic	Latin	Fusional	~40M	~1,649,832
Chinese	zh	Sino-Tibetan	Sinitic	Hanzi	Analytic	~1.1B	~1,465,839
Japanese	ja	Japonic	—	Kanji/Kana	Agglutinative	~125M	~1,452,150
Ukrainian	uk	Indo-European	Slavic	Cyrillic	Fusional	~40M	~1,368,238
Vietnamese	vi	Austrasiatic	Vietic	Latin	Analytic	~86M	~1,293,417
Portuguese	pt	Indo-European	Romance	Latin	Fusional	~264M	~1,144,604
Persian (Farsi)	fa	Indo-European	Iranian	Perso-Arabic	Fusional	~80M	~1,030,086
Serbian	sr	Indo-European	Slavic	Cyrillic/Latin	Fusional	~12M	~703,048
Korean	ko	Koreanic	—	Hangul	Agglutinative	~81M	~699,221
Norwegian	no	Indo-European	North Germanic	Latin	Fusional	~5M	~643,075
Finnish	fi	Uralic	Finnic	Latin	Agglutinative	~6M	~589,626
Czech	cs	Indo-European	Slavic	Latin	Fusional	~10M	~563,790
Hungarian	hu	Uralic	Ugric	Latin	Agglutinative	~13M	~554,772
Danish	da	Indo-European	North Germanic	Latin	Fusional	~6M	~306,973
Tamil	ta	Dravidian	Southern Dravidian	Tamil	Agglutinative	~87M	~172,122
Hindi	hi	Indo-European	Indo-Aryan	Devanagari	Fusional	~600M	~165,001
Swahili	sw	Niger-Congo	Bantu	Latin	Agglutinative	~87M	~97,374
Sesotho	st	Niger-Congo	Bantu	Latin	Agglutinative	~12M	~1,383

Table 2: Languages with family, subfamily, script, morphological type, approximate number of speakers (Ethnologue), and number of Wikipedia articles.

Language	Source
Chinese	https://github.com/drkameleon/complete-hsk-vocabulary/tree/main
Czech	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/czech.txt
Danish	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/danish.txt
Dutch	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/dutch.txt
English	Wonderwords library (same as Ruler)
Finnish	https://github.com/akx/fi-words/blob/master/words/words.txt
French	https://raw.githubusercontent.com/Blkzer0/Wordlists/refs/heads/master/French.txt
German	https://github.com/Jonny-exe/German-Words-Library
Hindi	https://github.com/eymenefealtun/all-words-in-all-languages/blob/main/Hindi/Hindi.txt
Hungarian	https://github.com/Blkzer0/Wordlists/blob/master/Hungarian.txt
Italian	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/italian.txt
Japanese	https://github.com/elzup/jlpt-word-list/tree/master
Korean	https://github.com/acidsound/korean_wordlist/blob/master/wordslist.txt
Norwegian	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/norwegian.txt
Persian (Farsi)	https://github.com/mvalipour/word-list-fa/blob/master/words.txt
Polish	https://github.com/MontrealCorpusTools/sct_resources/blob/main/Polish/words.txt
Portuguese	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/portuguese.txt
Russian	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/russian.txt
Serbian	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/serbian.txt
Southern Sotho	https://github.com/eymenefealtun/all-words-in-all-languages/blob/main/Sesotho/Sesotho.txt
Spanish	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/spanish.txt
Swahili	https://github.com/michaelnjuguna/All-swahili-words-dictionary/blob/main/kamusi.txt
Swedish	https://raw.githubusercontent.com/martinlinde/wordlist_swedish/refs/heads/master/swe_wordlist
Tamil	https://github.com/vigneshwaran-chandrasekaran/tamil-language-words-list/blob/master/tamilwords.txt
Ukrainian	https://github.com/gurkylee/Wordlist-Collection/blob/main/languages/ukrainian.txt
Vietnamese	https://github.com/duyet/vietnamese-wordlist

Table 3: Sources of wordlist for each language used for the CWE tasks. The lists were downsampled to 10k and part-of-speech tagging was performed using GPT-4o-mini to keep only nouns, verbs, and adverbs.

single native speaker responsible for translating all prompts and noun lists in that language. In a few cases, a second annotator reviewed the translations for quality assurance, but each data point was authored by a single individual.

Book Data For NIAH tasks, we fill the surrounding context with books from each respective language. Each book was processed to remove the front and back matter. In Table 4, we

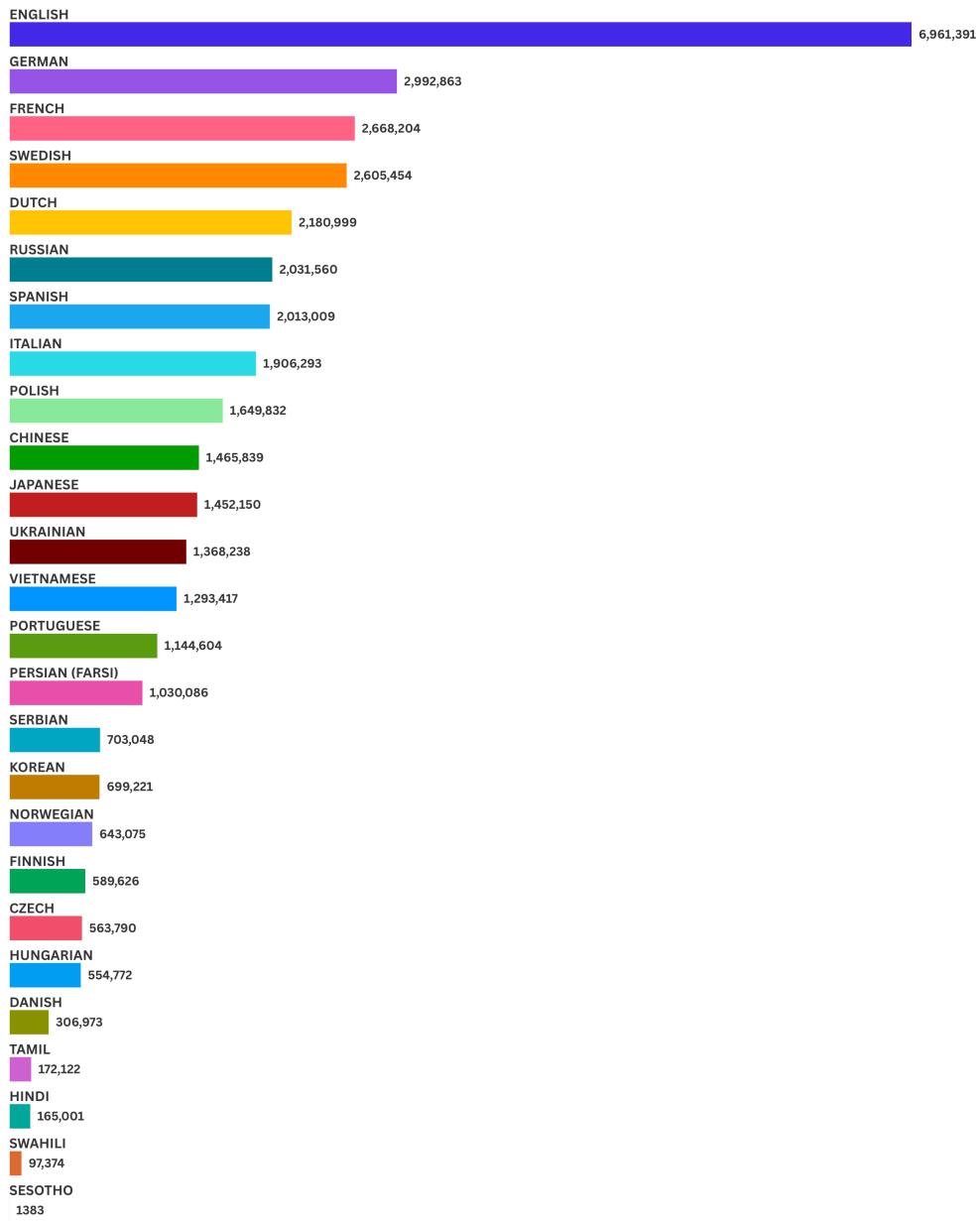


Figure 7: Language by Wikipedia size. The counts indicate the current number of articles available on Wikipedia for the given language.

detail for each book the title, author, year, and original language of publication. All books must be long enough to fill the 128k context-length tasks for the specified language. For shorter books, we replicate part of the book to fit the context. Most of the books are older due to the copyright restrictions, and we acknowledge that their age and linguistic style might have influenced models' performance. However, this relationship is not straightforward (i.e., the performance on newer books is not necessarily better than on older books).

Word lists used for the CWE task in 26 languages We extracted word lists for 26 languages from GitHub repositories (Table 3) dedicated to language wordlists, using part-of-speech tagging to identify nouns, adjectives, and verbs.

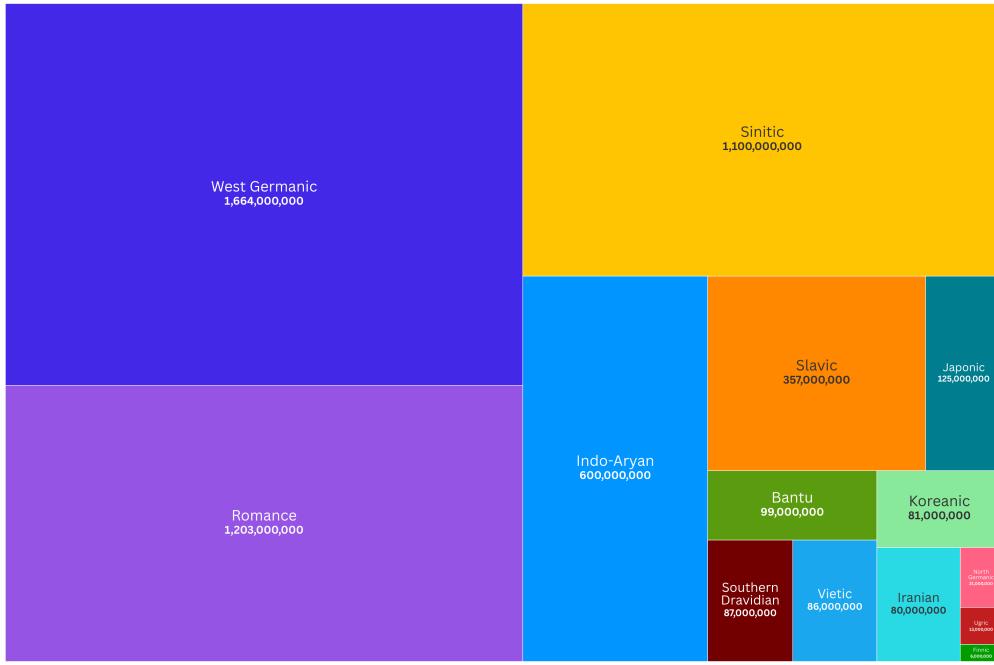


Figure 8: Language family by the number of speakers as reported by Ethnologue.

Lang.	Translated Title (Original Title)	Author	Year	Token Count
cs	<i>The Good Soldier Švejk, Vol.3 (Osudy dobrého vojáka Švejka za světové války)</i>	Jaroslav Hašek	1923	174576
da	<i>Elderflower and Hops (Hyld og Humle: Fortællinger)</i>	Sophie Breum	1900	76082
de	<i>The Magic Mountain (Der Zauberberg)</i>	Thomas Mann	1924	598161
en	<i>Little Women</i>	Louisa May Alcott	1868	233668
es	<i>Don Quixote</i>	Miguel de Cervantes	1605	626105
fa	<i>The Blind Owl</i>	Sadegh Hedayat	1936	77451
fi	<i>Sylvi</i>	Minna Canth	1893	60803
fr	<i>Les Misérables</i>	Victor Hugo	1862	970526
hi	<i>Chandrakanta</i>	Devaki Nandan Khatri	1888	813743
hu	<i>King Midas, Vol. 1 (Midás király)</i>	Zoltán Ambrus	1891	224484
it	<i>The Late Mattia Pascal</i>	Luigi Pirandello	1904	144402
ja	<i>Kokoro (心)</i>	Natsume Sōseki	1914	185905
ko	<i>A Boy's Sorrow (소년의 비애)</i>	Yi Kwang-su	1917	12279
nl	<i>The Diary of a Young Girl (Het Achterhuis)</i>	Anne Frank	1947	136314
no	<i>Kristin Lavransdatter, Vol. 2: The Wife</i>	Sigrid Undset	1921	251016
pl	<i>Nights and Days, Vol. 3 (Nocie i dnie)</i>	Maria Dąbrowska	1934	452847
pt	<i>The Book of Disquiet (Livro do Desassossego)</i>	Fernando Pessoa	1982	248245
ru	<i>War and Peace (Война и мир)</i>	Leo Tolstoy	1869	1112514
sr	<i>Seconds of Eternity (Sekund večnosti, istočnjački roman)</i>	Dragutin Ilić	1921	45038
st	<i>Chaka</i>	Thomas Mofolo	1925	119342
sv	<i>The Story of Gösta Berling</i>	Selma Lagerlöf	1891	255416
sw	<i>My Life Fifty Years After (Maisha yangu na baada ya Miaka Hamsini)</i>	Shaaban Robert	1958	61490
ta	<i>Ponniyin Selvan, Vol. 1: The First Floods</i>	Kalki Krishnamurthy	1950	912523
uk	<i>After Finishing School (Instytutka)</i>	Marko Vovchok	1862	41056
vi	<i>Pure Heart (To tam)</i>	Hoàng Ngọc Phách	1925	72267
zh	<i>Call to Arms (喊)</i>	Lu Xun	1922	153415

Table 4: Complete list of books used for needle injections in the retrieval task. Each row contains the language the book was originally published in, the title, author name, published year, and token count. Tokens counts calculated using tiktoken (o200k).

B Generations

Prompt Templates For each task, we use consistent templates with minimal rewording. Translations were done at the instruction level, with each paragraph translated separately. Translators were informed of the task’s purpose and the words to be replaced, ensuring the

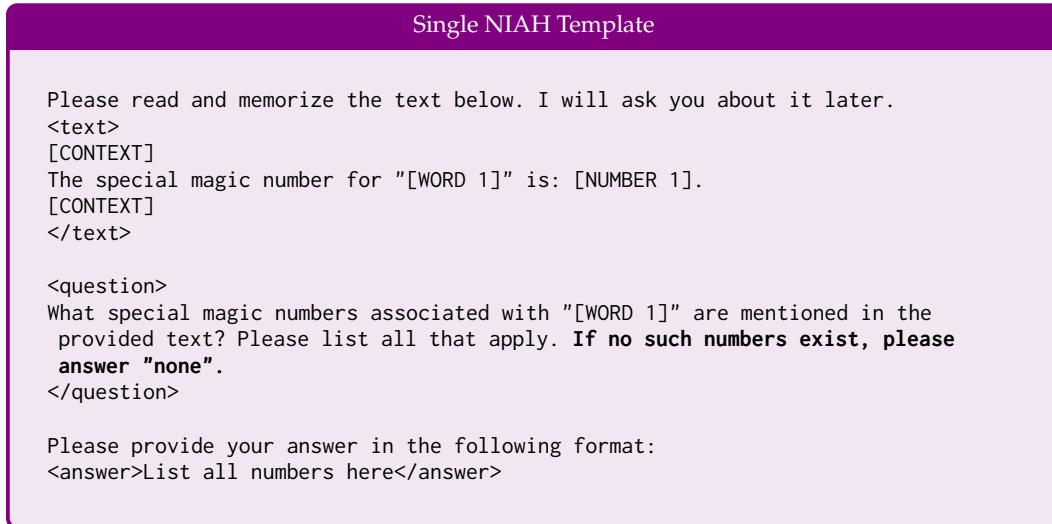


Figure 9: Prompt template for our “Single NIAH” task with one magic number. The bolded sentence introduces the possibility of no answer existing, which we show significantly hurts model accuracies, even for this Single NIAH task where the answer always exists.

Model	Size	Context Length	Huggingface Wolf et al. (2019) / API	Cost
Gemini 1.5 Flash (Gemini Team, 2024)	-	1M	gemini-1.5-flash	\$137.7
o3-mini-high (OpenAI, 2025)	-	200K	o3-mini-2025-01-31	\$1932.5
Deepseek-R1 (DeepSeek-AI, 2025)	685B	128k	deepseek-r1	\$333.2
Llama3.3 (Llama Team, 2024)	70B	128K	meta-llama/Llama-3.3-70B-Instruct	-
Llama3.1 (Llama Team, 2024)	8B	128K	meta-llama/Meta-Llama-3.1-8B-Instruct	-
Qwen2.5-72B (Qwen Team, 2025)	72B	128K	Qwen/Qwen2.5-72B-Instruct	-
Qwen2.5-7B (Qwen Team, 2025)	7B	128K	Qwen/Qwen2.5-7B-Instruct	-

Table 5: Comparison of Open and Closed Language Models used in experiments

wording allowed for ‘key’ substitution without affecting grammar. The correct instructions and words were added to make the final prompts. The English versions of the prompts are as follows: (1) the Single-Niah task (Figure 9), (2) the multi-query task (Figure 12), (3) the multi-key task (Figure 10), (4) the none-NIAH task (Figure 13), (5) the multi-value task (Figure 11), and (6) the single NIAH task without the inclusion of none (Figure 14) (used for ablation study). The CWE prompt is shown in (Figure 15). We include the English prompts for understanding, but for experiments translated instruction prompts are used in the language of the context (except the cross-lingual setup). We ensured all variables could be substituted in templates without grammatical changes (e.g., singular/plural forms, gender variants, or grammatical cases). All instructions and needles are verified translations by native speakers of each respective language.

B.1 Model Configurations

For o3-mini, we set reasoning efforts to *high* and configured the maximum output token length to 10k, with temperature fixed at 1.0 as per OpenAI defaults. For Deepseek-R1, we set the maximum output token length to 8k and used a temperature of 0.6. We evaluated all 26 languages on Gemini 1.5 Flash and o3-mini, utilizing Batch API for o3-mini which provided a 50% cost reduction. For Deepseek-R1, we used the Fireworks API but limited evaluation to English only due to budget constraints. Please re-

Model	# Param	Ctx len	# Vocab
🔒 Gemini 1.5 Flash	~1.2B	1M	256K
🔒 o3-mini-high	~1.2B	200K	200K
🔒 Deepseek-R1	671B/37B	128K	129k
🔒 Llama 3.3	70B	128K	128K
🔒 Llama 3.1	8B	128K	128K
🔒 Qwen 2.5	72B	128K	152K
🔒 Qwen 2.5	7B	128K	152K

Multi-Key NIAH Template

```
Please read and memorize the text below. I will ask you about it later
<text>
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 1].
[CONTEXT]
The special magic number for "[WORD 2]" is: [NUMBER 2].
[CONTEXT]
The special magic number for "[WORD 3]" is: [NUMBER 3].
[CONTEXT]
The special magic number for "[WORD 4]" is: [NUMBER 4].
</text>

<question>What special magic numbers associated with "[WORD 3]" are mentioned in
the provided text? Please list all that apply. If no such numbers exist, please
answer "none".</question>

Please provide your answer in the following format:
<answer>List all numbers here</answer>
```

Figure 10: Prompt template for Multi-Key NIAH Task with multiple magic numbers.

Multi-Value NIAH Template

```
Please read and memorize the text below. I will ask you about it later
<text>
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 1].
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 2].
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 3].
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 4].
</text>

<question>What special magic numbers associated with "[WORD 1]" are mentioned in
the provided text? Please list all that apply. If no such numbers exist, please
answer "none".</question>

Please provide your answer in the following format:
<answer>List all numbers here</answer>
```

Figure 11: Prompt template for Multi-Value NIAH Task with multiple magic numbers.

fer to [Table 5](#) and [Figure 16](#) for detailed model configurations and associated costs. For model configurations, we run inference on open-weights models using greedy decoding with temperature set to 0.0, which we also applied to the Gemini-1.5-Flash model. We evaluate Llama and Qwen with vLLM ([Kwon et al., 2023](#)), using 1 A100 for smaller models and 4 A100s for larger ones.

Multi-Query NIAH Template

```
Please read and memorize the text below. I will ask you about it later
<text>
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 1].
[CONTEXT]
The special magic number for "[WORD 2]" is: [NUMBER 2].
</text>

<question>What special magic numbers associated with "[WORD 1]" and "[WORD 2]"
are mentioned in the provided text? Please list all that apply. If no such
numbers exist, please answer "none".</question>

Please provide your answer in the following format:
<answer>List all numbers here</answer>
```

Figure 12: Prompt template for Multi-Query NIAH Task with multiple magic numbers.

None NIAH Template

```
Please read and memorize the text below. I will ask you about it later
<text>
[CONTEXT]
The special magic number for "[WORD 1]" is: [NUMBER 1].
[CONTEXT]
The special magic number for "[WORD 2]" is: [NUMBER 2].
[CONTEXT]
The special magic number for "[WORD 3]" is: [NUMBER 3].
[CONTEXT]
The special magic number for "[WORD 4]" is: [NUMBER 4].
</text>

<question>What special magic numbers associated with "[WORD 5]" are mentioned in
the provided text? Please list all that apply. If no such numbers exist, please
answer "none".</question>

Please provide your answer in the following format:
<answer>List all numbers here</answer>
```

Figure 13: Prompt template for None NIAH Task with no matching magic numbers.

Single NIAH - No 'None' Template

```
Please read and memorize the text below. I will ask you about it later.  
<text>  
[CONTEXT]  
The special magic number for "[WORD 1]" is: [NUMBER 1].  
[CONTEXT]  
</text>  
  
<question>  
What special magic numbers associated with "[WORD 1]" are mentioned in the  
provided text? Please list all that apply. If no such numbers exist, please  
answer "none".  
</question>  
  
Please provide your answer in the following format:  
<answer>List all numbers here</answer>
```

Figure 14: Prompt template for Single NIAH Task, without the inclusion of None in the prompt.

CWE Template

```
Below is a numbered list of words. In these words, some appear more often than  
others. Memorize the ones that appear most often.  
<list>[ENUMERATED LIST OF WORDS]</list>  
<question> What are the 10 most common words in the list above? </question>  
Please provide your answer in the following format  
<answer> List the words here </answer>
```

Figure 15: Prompt template for Most Common Words Task.

C Additional Results

C.1 Retrieval Results

In this section, we detail additional results for the Needle-in-a-Haystack task.

Average NIAH performance across all languages In Figure 17, we see the performance of each model averaged across all NIAH tasks and all languages, with gemini-1.5-flash outperforming all other models, especially at long context-lengths.

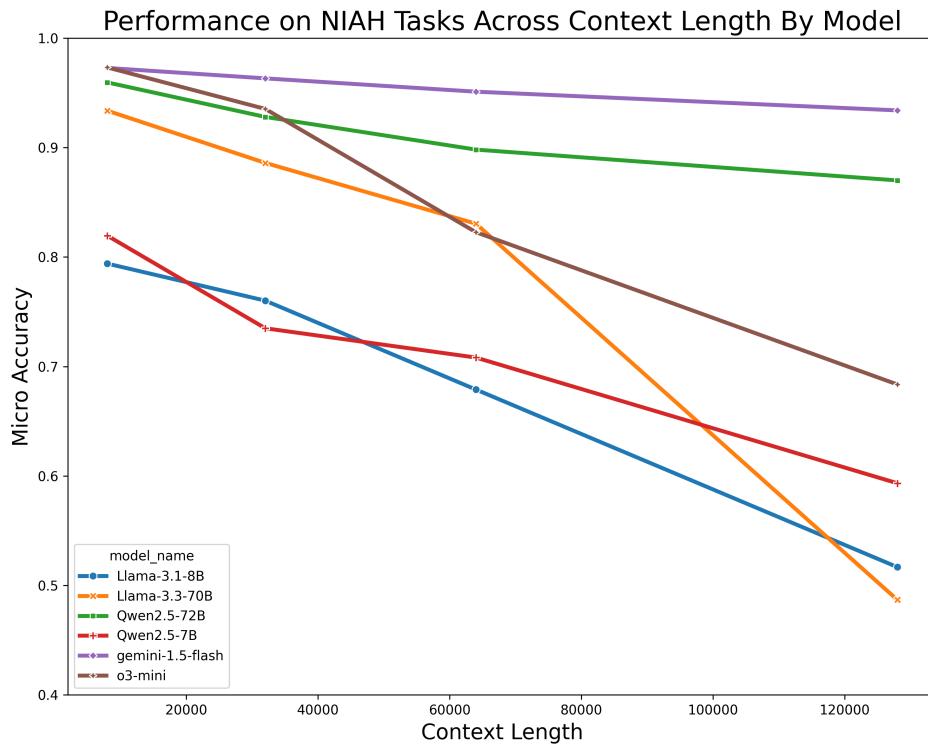


Figure 17: By model performance across all languages on the NIAH tasks across context lengths. We find that LLama-3.3 has the largest decrease in performance across context lengths.

Number of incorrect answers marked as ‘none’ across all languages We show that the number of wrong answers labeled as ‘none’ in S-NIAH tasks with long context varies across models and languages, as illustrated in Figure 18 and Figure 19. The o3-mini-high model produces a significantly higher number of ‘none’ errors compared to other models. Interestingly, Gemini-1.5-flash also generates a notable number of ‘none’ errors for some high-resource languages. Surprisingly, although the Qwen model is specialized in Chinese, it exhibits a large number of ‘none’ errors in S-NIAH as well.

C.2 Reasoning Results

In this section, we detail additional results for reasoning models. In Figure 20 we show the average amounts of reasoning tokens for NIAH tasks from o3-mini. We observe that incorrect answers produce drastically higher amounts of reasoning tokens than correct answers. In Figure 21, we show the number of responses exceeding the set max number of output tokens for the English language. We note that both o3-mini-high and deepseek-r1 often exceed the token limit, overreasoning for a simple task.

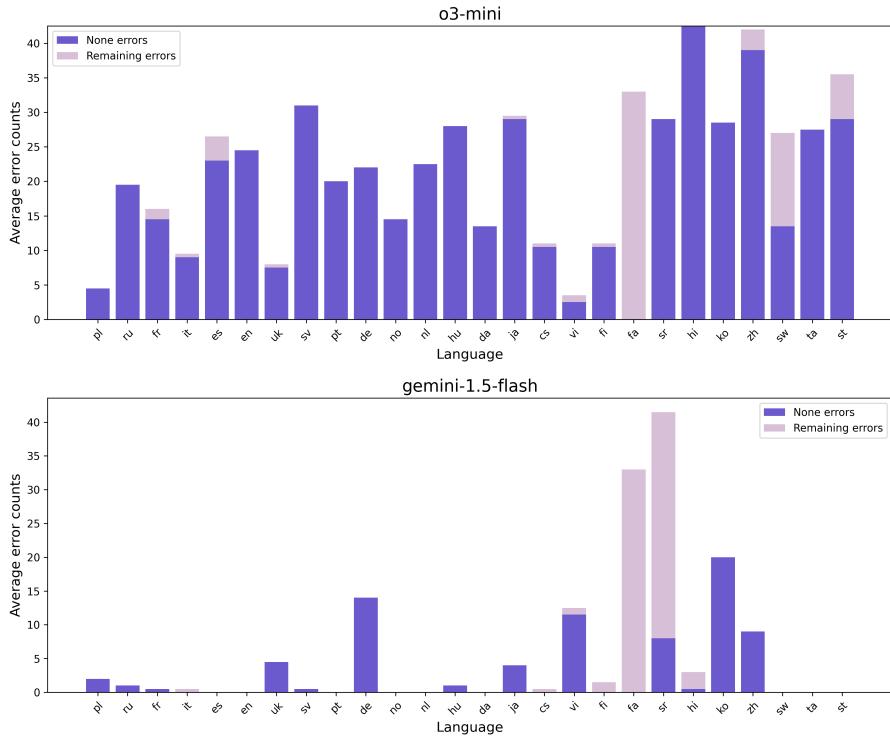


Figure 18: The types of errors made in the S-NIAH at long-context(64K and 128K) across closed model and languages

C.3 Aggregation Results

We evaluate CWE in both the easy and hard settings over four context lengths: 8k, 32k, 64k, and 128k. As seen in Figure 5, average English accuracy over all models is only 31.5% for the CWE-easy task. The CWE-hard setting proves unsolvable with nearly 0% accuracy across all models. Performance of each model by context length can be found in Figure 22. All models evaluated perform much better on the *CWE-easy* task in the 8k length. The drop in performance amidst increasing context length indicates language models still have trouble with this kind of aggregation task in very long-context settings. The *CWE-hard* task is near impossible for all models. For a task that would be so simple for humans, the near-total incapability of modern LLMs to complete the task in the harder setting highlights that synthetic tasks like CWE and NIAH, that can be made more challenging as model performance improves, are necessary for ongoing evaluation of models.

Reasoning models “overthink” on simple aggregation tasks: On both CWE tasks, o3-mini-high fails to generate answers within its 10K output token limit for almost every sample across all languages and context sizes.¹⁴ Notably, this overthinking persists even with smaller contexts, with reasoning outputs sometimes exceeding the length of the given context itself! This is not just unique to o3-mini-high: we also observe that a significant number of Deepseek-R1’s responses in the *CWE-hard* task at long contexts exceed the 8k output token limit before reaching an answer as seen in Figure 21. Qualitative inspection of the reasoning tokens reveals strategies unrelated to the task (e.g., behaviors (e.g., unrelated to the task. As a whole, these results suggest that reasoning models should be better optimized for tasks requiring aggregating information across long contexts.

¹⁴All model configurations reported in subsection B.1.

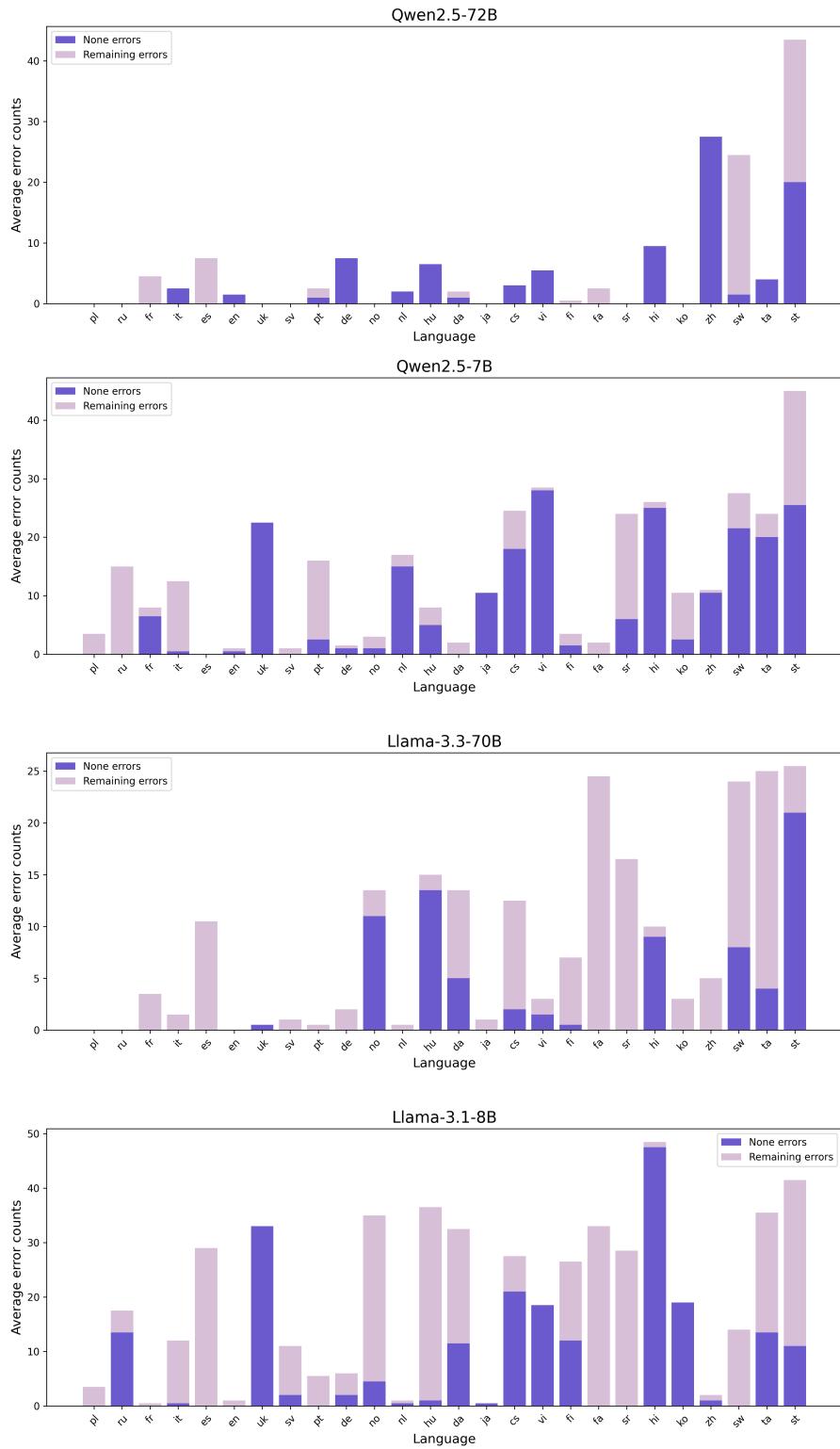


Figure 19: The types of errors made in the S-NIAH at long-context(64K and 128K) across open-source model and languages

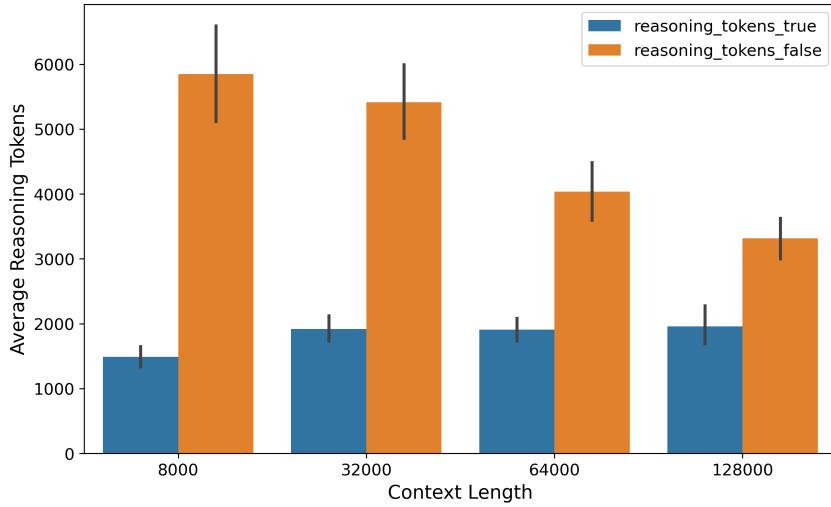


Figure 20: Average reasoning token length for correct vs. incorrect answers across context lengths in NIAH tasks

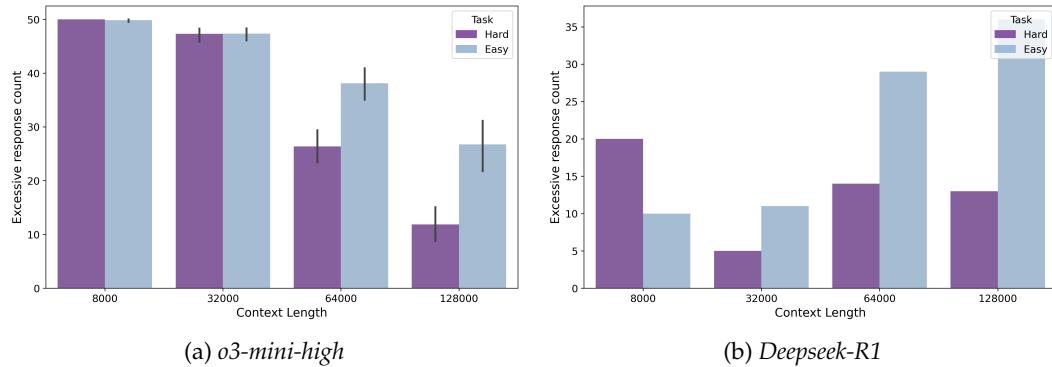


Figure 21: Number of responses exceeding the maximum output tokens in CWE tasks, caused by reasoning overthinking, for o3-mini-high (26 languages) and Deepseek-R1 (English only).

D One Tokenizer Does Not Rule Them All

One of the biggest challenges of the multilingual setting is understanding the effects of the tokenizer on model performance. For example, one model may be able to take in 50,000 tokens in Tamil, but another may only be able to take in 25,000, changing the overall amount of information passed to the model. What would happen if we measured the effects of a model using the information density (or amount of content given), instead of the token count? Below, we detail many of the same experiments as conducted earlier, but in the context of each model being given the same amount of information per model instead of the same amount of tokens.

Drastic Differences in Token Lengths: For each language, the token counts of other tokenizers when stabilizing for 128k by tiktoken c1100k_base show astounding variance, as seen in Figure 24. We see that the same 128k as per c1100k_base results in large differences between other tokenizers. For example, Tamil has only 33,249 tokens when processed using the tiktoken o200k_base, compared to 128,000 with c1100k_base.

Similar findings hold despite dramatic differences in tokenized lengths: As seen in Figure 25, when controlling for amount of context information, Slavic languages (Russian, Ukrainian, Polish) perform highly, surpassing the performance of Romance languages like

Example of Deepseek-R1's Reasoning for MV-NIAH

<think>

Okay, let's tackle this question. The user wants to know the special magic numbers associated with the word "field" mentioned in the provided text. I need to go through the text carefully and look for any instances where the word "field" is followed by a number.

First, I'll start by recalling that the text is from "Little Women" by Louisa May Alcott. The user provided a lengthy excerpt covering several chapters. The task is to find any mention of a "special magic number for 'field'" in the text.

I'll need to scan through the text for the exact phrase "The special magic number for 'field' is: [number]." Since the user provided the entire text, I can search for occurrences of this specific phrase.

Looking through the chapters, I remember that in some parts, there are notes or numbers inserted. Let me check each chapter section:

1. **Chapter One:** No mention of a magic number here. The story starts with the sisters talking about Christmas. 2. **Chapter Two:** Focuses on Christmas morning and the girls' activities. No numbers related to "field." 3. **Chapter Three:** The Laurence boy. Here, there's a mention of a magic number. The text says: "The special magic number for 'field' is: 2140194." So that's one number. 4. **Chapter Four:** Burdens. No mention here. 5. **Chapter Five:** Being Neighborly. No numbers here either. 6. **Chapter Six:** Beth Finds the Palace Beautiful. No magic numbers. 7. **Chapter Seven:** Amy's Valley of Humiliation. No numbers here. 8. **Chapter Eight:** Jo Meets Apollyon. No mention. 9. **Chapter Nine:** Meg Goes to Vanity Fair. Here, another instance: "The special magic number for 'field' is: 4946066." 10. **Chapter Ten:** The P.C. and P.O. No numbers here. 11. **Chapter Eleven:** Experiments. No mention. 12. **Chapter Twelve:** Camp Laurence. Another instance: "The special magic number for 'field' is: 4117625."

So, in total, there are three instances where the text explicitly states a special magic number for "field": 2140194, 4946066, and 4117625. Each of these is found in different chapters. I need to list all three as the answer. </think>

Table 6: Example of Deepseek's reasoning for the MV-NIAH task, showing how it systematically analyzes the text by dividing it into chapters and searching for the target pattern.

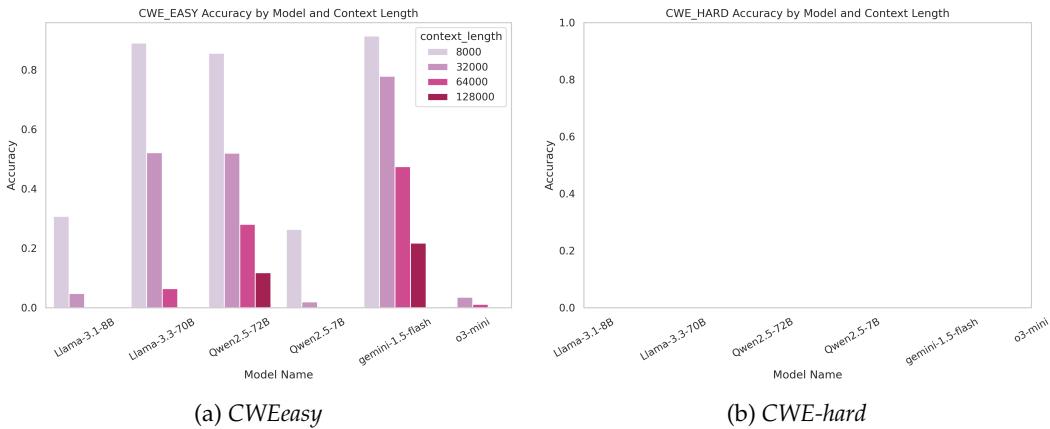


Figure 22: CWE Accuracy by Context Length

Italian, French, and Spanish. English performance is even lower, typically 8th. However, similar model ranking patterns remain: Gemini 1.5 Flash and Qwen 2.5 72B perform the best, and high-resource Latin/Cyrillic languages outperform others. It is interesting to note that languages such as Korean, Hindi, and Chinese are towards the low end of the performance spectrum, even though they are often tokenized to shorter lengths when in comparison to

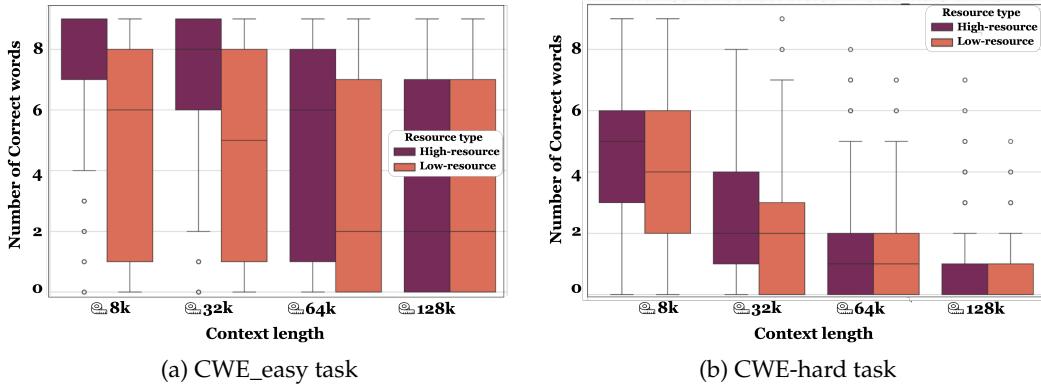


Figure 23: Number of correct words in cases where the model generated an answer but it was ultimately incorrect at different context length. As context length increases, the model finds less correct words. High resource languages tend to return higher average numbers of correct words, which may be indicative of issues relating to tokenization.

English data.¹⁵ As context length increases, we still see that the performance gap between high-resource and low-resource languages, consistent with our main results of controlling for number of tokens.

¹⁵This is due to larger tokenization gains we have achieved for these languages with newer tokenizer (e.g., cl100k_base vs o200k_base).

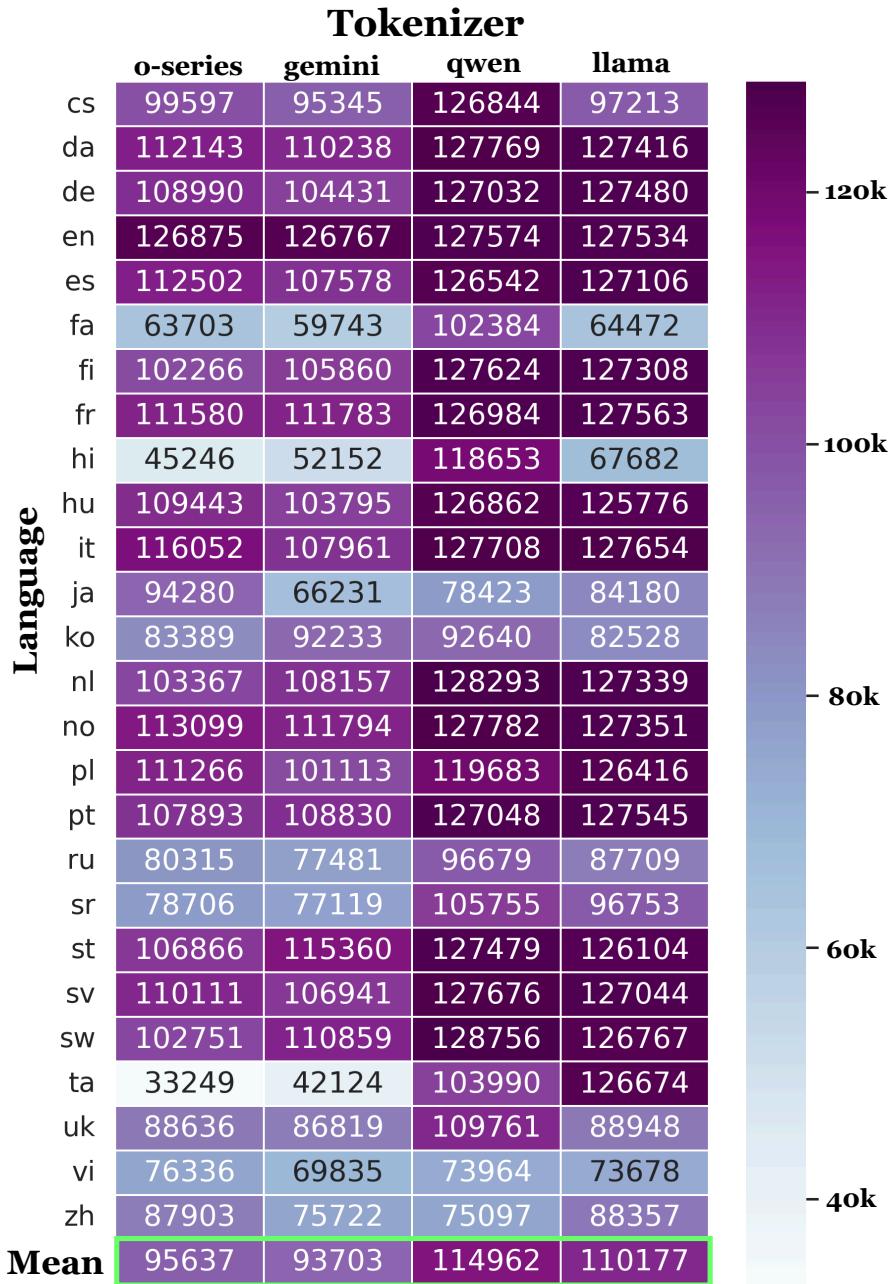


Figure 24: For each language, the token counts of other tokenizers when stabilizing for 128k by cl100.

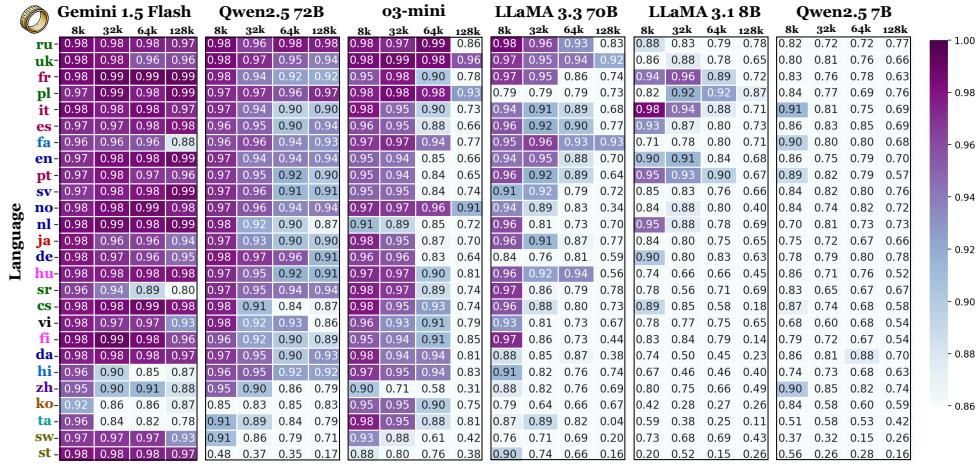


Figure 25: Heatmap of average accuracy of NIAH tasks by language when controlling for input context length. In these tasks, models are provided with the same input. Input length was measured with tiktoken (cl100k).