SYNC: A Synthetic Long-Context Understanding Benchmark for Controlled Comparisons of Model Capabilities

Shuyang Cao, Kaijian Zou, and Lu Wang
University of Michigan
Ann Arbor, MI
{caoshuy, kjzou, wangluxy}@umich.edu

Abstract

Recently, researchers have turned to synthetic tasks for evaluating long-context capabilities of large language models (LLMs) , as they offer more flexibility than realistic benchmarks in scaling both input length and dataset size. However, existing synthetic tasks typically target narrow skill sets such as retrieving information from massive input, limiting their ability to comprehensively assess model capabilities. Furthermore, existing benchmarks often pair each task with a different input context, creating confounding factors that prevent fair cross-task comparison. To address these limitations, we introduce SYNC, a new evaluation suite of synthetic tasks spanning domains including graph understanding and translation. Each domain includes three tasks designed to test a wide range of capabilities—from retrieval, to multi-hop tracking, and to global context understanding that that requires chain-of-thought (CoT) reasoning. Crucially, all tasks share the same context, enabling controlled comparisons of model performance. We evaluate 14 LLMs on SYNC and observe substantial performance drops on more challenging tasks, underscoring the benchmark's difficulty. Additional experiments highlight the necessity of CoT reasoning and demonstrate that SYNC poses a robust challenge for future models.

1 Introduction

Large language models (LLMs) have extended their context lengths with recent advances, enabling them to accommodate more diverse and extensive user inputs (OpenAI et al., 2024b; AI, 2024). To understand LLMs' capabilities when consuming long contexts, it is crucial to develop benchmarks with sufficiently long inputs. Early benchmarks for long-context evaluation primarily focus on realistic tasks, where data is either sourced from human-annotated documents (Pang et al., 2022) or existing text corpora (Shaham et al., 2023; Dong et al., 2024). While these tasks reflect real-world use cases, they are limited in flexibility: Once constructed, the input contexts in such datasets are essentially fixed, extending which to longer contexts often requires sourcing and annotating new data samples (Bai et al., 2024b; Wang et al., 2025). Precisely controlling the difficulty of understanding the contexts is also challenging, as they depend on the available realistic contexts. As a result, realistic benchmarks are not suitable for controlled evaluation of models' long-context capabilities.

Recently, synthetic benchmarks have emerged as a more scalable way to probe model capabilities at controllable lengths. A notable example is the needle-in-a-haystack (NIAH) test, where key-value pairs to be retrieved are inserted into long passages of irrelevant text (Kamradt, 2023). While such tests are simple to construct and offer extremely lengthy contexts without the need for human annotations, they only evaluate whether the model is able to pinpoint the required content. Although extensions of NIAH have been proposed to test other capabilities (Hsieh et al., 2024), the range of capabilities evaluated remains narrow. Moreover, existing synthetic benchmarks curate tasks with different input contexts which become confounding factors when comparing model performance across tasks to understand model behaviors.

To address these limitations of existing *synthetic* benchmarks, we propose a new benchmark, SYNC, comprising SYNthetic Contexts for fine-grained comparisons of LLMs' long-context capabilities.[1] SYNC features synthetic contexts covering two domains: *graph understanding* and *unseen language translation*. Three tasks are composed for each domain and designed to evaluate *a broader range of capabilities*, from simple information retrieval to multi-hop state tracking, and to global context un-

[1]Our data and data generation code are available at `https://shuyangcao.github.io/projects/sync/`.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33616–33637
November 4-9, 2025 ©2025 Association for Computational Linguistics

derstanding that requires synthesizing and reasoning across multiple pieces of information scattered throughout the long context.

For example, in graph understanding tasks, the LLM is first tasked with finding nodes connected to a given node to examine its retrieval capability. To assess multi-hop state tracking, the model must then determine the shortest path between two given nodes. Finally, to demonstrate global context understanding, the LLM is queried to find the longest path within the graph, a task requiring a holistic comprehension of the entire graph structure. Importantly, by presenting tasks of varying complexity under the *same context* that describes the graph or translation rules, we ensure that differences in model performance across tasks are solely attributed to the task difficulty. This design choice allows us to accurately assess the models' capabilities associated with different tasks without introducing confounding factors that arise when each task has a different context.

On SYNC, we evaluate 12 open-source and 2 proprietary LLMs that support a context length of 128K or longer, which reveals a consistent degradation in performance as the complexity of tasks increases. Notably, on the most challenging tasks requiring global context understanding, no model surpasses 25% accuracy. Compared with existing synthetic benchmarks, SYNC proves to be more effective in differentiating model capabilities, offering clearer alignment between task difficulty and performance. We further conduct experiments without the usage of chain-of-thought (CoT), where models struggle to maintain reasonable performance, which indicates the necessity of CoT on SYNC and again demonstrates the difficulty of our tasks. Additionally, we examine the correlation between our tasks and realistic tasks, revealing that identifying the shortest path in a graph can be a good predictor of real-world performance.

Our contributions can be summarized as follows:

1. We propose a new long-context evaluation benchmark, SYNC, which comprises synthetic contexts including graphs and unseen languages. Tasks requiring different levels of capabilities are designed based on the same contexts, enabling accurate assessment of model capabilities.

2. We benchmark 14 LLMs (12 open-source and 2 proprietary) on SYNC, revealing that existing LLMs face a challenge when handling tasks beyond retrieval on long contexts.

3. We conduct thorough analyses of our benchmark, including comparisons with existing synthetic benchmarks, an investigation into the effect of chain-of-thought reasoning, and a study of correlation between our tasks and realistic tasks. These analyses validate the difficulty of SYNC, illustrate the benefit of shared contexts, and indicate that our tasks can predict real-world performance.

2 Related Work

Early long-context evaluation benchmarks (Shaham et al., 2023; Tay et al., 2021) are gradually falling behind the advancements of LLMs with long context windows (Ainslie et al., 2023; Liu et al., 2023a; Chen et al., 2023; Peng et al., 2023; Team et al., 2024), due to the insufficient coverage of model context lengths (128K and longer) by the included data (Kočiský et al., 2018; Zhong et al., 2021; Huang et al., 2021; Wang et al., 2022). Recent benchmarks aim to address this gap by developing tasks featuring significantly longer contexts.

Realistic Tasks. Realistic tasks assess the practical performance of LLMs in applications closely aligned with real-world scenarios. Unlike synthetic tasks, these provide a more representative evaluation of long-context capabilities. However, realistic tasks are challenging to construct, and controlling the length and complexity of data points can be difficult.

Several benchmarks have emerged to comprehensively evaluate LLMs across diverse realistic applications. For instance, NovelQA (Wang et al., 2025), LongBench (Bai et al., 2024a), Nocha (Karpinska et al., 2024), ∞Bench (Zhang et al., 2024b), BABILong (Kuratov et al., 2024), BAMBOO (Dong et al., 2024), Loong (Wang et al., 2024), and LongCite (Zhang et al., 2024a) emphasize question-answering tasks involving lengthy narratives or multiple documents. For long-document summarization, benchmarks such as L-Eval (An et al., 2024) and LooGLE (Li et al., 2024a) provide a variety of relevant tasks. Furthermore, benchmarks like LongBench v2 (Bai et al., 2024b) and Long Code Arena (Bogomolov et al., 2024) assess repo-level code understanding, while LOFT (Lee et al., 2024) evaluates retrieval-augmented generation (RAG) tasks in extensive contexts. Finally, LongICLBench (Li et al., 2024b) and ManyICLBench (Zou et al., 2025) specifically target the evaluation of long-context models in many-shot in-context learning scenarios.

Synthetic Tasks. Synthetic tasks are specifically designed to rigorously test LLMs through artificially constructed contexts. Their primary advantage is the ease of generating data points, enabling precise control over context length and difficulty.

One of the most widely used synthetic tasks is Needle-in-a-Haystack (NIAH) (Kamradt, 2023), where models must retrieve a fact statement embedded within a large volume of random text. Synthetic tasks have also been incorporated into benchmarks that simutaineously contain realistic tasks, such as LongBench v2 (Bai et al., 2024a) and HELMET (Yen et al., 2024), yet they are mostly retrieval tasks similar to NIAH. Beyond extending NIAH, RULER (Hsieh et al., 2024) introduces tasks demanding more complex capabilities, such as state tracking.

Despite these efforts, current synthetic tasks remain limited in scope, primarily assessing retrieval capabilities. A broader limitation in multi-task benchmarks is the use of varying input contexts across tasks, which introduces confounding factors that hinder fair comparisons of model proficiency across different capabilities.

3 SYNC Task Creation

In this section, we introduce our benchmark, SYNC, a suite of synthetic tasks designed to evaluate varying capabilities of LLMs when processing long contexts. SYNC spans two domains: graph understanding (§3.1) and unseen language translation (§3.2). The domains are selected based on three criteria: (1) the domains can accommodate tasks that challenge different levels of model capabilities and can be adapted in difficulty for future models, (2) automatic sample construction is feasible, and (3) automatic evaluation is reliable. The contexts for these tasks are formed by descriptions of graphs and translation rules, respectively, and are supplemented with task-irrelevant information to increase the context length. For each domain, we consider three tasks of increasing complexity, evaluating the model's capabilities in information *retrieval*, state *tracking*, and *global* context understanding. We discuss key properties that distinguish our tasks from existing synthetic benchmarks in §3.3, including (1) targeting capabilities of varying levels; and (2) sharing contexts for controlled comparison.

Context
`You will answer a given question based on a directed acyclic graph. The edges of the graph are hidden within the following text. Make sure to memorize them. The nodes in the graph are: Node 1, Node 2, Node 3, Node 4, Node 5. [haystack] There is a directed edge from Node 1 to Node 2. [haystack] There is a directed edge from Node 2 to Node 3. [haystack] There is a directed edge from Node 2 to Node 4. ...`
Task Query
[Retrieval] Connected Node: `What are the nodes with directed edges from Node 1?`
[Tracking] Shortest Path: `What is the shortest path from Node 1 to Node 3?`
[Global] Longest Path: `What is the longest path in the graph?`

Table 1: Example of graph understanding tasks. The context interleaves essential graph details with haystack to extend the context length.

3.1 Tasks on Graph Understanding

We generate random directed acyclic graphs (DAGs) based on the number of nodes and the edge density (i.e., the probability that an edge exists between two nodes). To adjust the difficulty, we change the number of nodes in each graph. Each graph is presented to the model by listing its nodes and edges, as shown in Table 1. To extend the context length without overly increasing task complexity, we follow the needle-in-the-haystack approach and interleave the graph description with redundant but topically consistent information. Specifically, we repeatedly state that there is no self-cycle at each node (e.g., "There is no directed edge from Node 1 to Node 1"). Since the models are informed that the graphs are acyclic, the haystack does not provide extra clues for solving the tasks.

The **Connected Node** task requires the model to identify all nodes that have outgoing directed edges from a queried node. For each graph, a node is randomly selected as the query. This task tests the model's ability to retrieve relevant information with the query node as the key. For evaluation, we check if the set of nodes in the model response exactly match the reference set.

In the **Shortest Path** task, the model must find the shortest path from a given node to another. Intuitively, to effectively solve the problem, the model should maintain the state of its exploration in the graph, which keeps track of the explored and unex-

Context
Answer the question based on the given languages. These languages are created for special purposes and do not exist in the real world. Their vocabularies and bilingual dictionaries are as follows. Note that the vocabularies might be duplicated. The vocabulary of LL0: eszyci, dppu, ... [haystack] Dictionary from LL0 to LL1: dpamn -> aqdek; czrzib -> rqntu; ybzol -> ucc; ... [haystack] The vocabulary of LL1: ubmcrdu, erwkyr, ... [haystack] Dictionary from LL1 to LL2: ...
Task Query
[Retrieval] Single-hop Translation: What is the translation of "lrg eafi axry ikxxqq viw" from LL1 to LL2?
[Tracking] Multi-hop Translation: What is the translation of "ayg nrhu lsloiv mzg phx" from LL0 to LL2?
[Global] Letter Coverage: Find the three words in LL0 such that the union of the first letters of all their translations contains the maximum number of distinct letters.

Table 2: Example of translation-based tasks. Vocabularies are repeated to construct the haystack.

plored nodes during reasoning. If multiple shortest path exist, the model is allowed to return any valid shortest path. The model generated path is examined for the existence of each edge and the optimality of the path length. Note that the queried node pairs are chosen such that either no shortest path exists or the shortest path length is greater than 1, as a path length of 1 reduces the task to simple retrieval.

Finally, we test the model's global understanding of the graph by asking the model to extract the **Longest Path** from the whole graph. Similar to Shortest Path, the model must generate a path with fully valid edges and its length must match the length of the reference longest path.

3.2 Tasks on Unseen Language Translation

We also explore a domain closer to natural language by simulating translation between synthetic languages. For each language, we generate a vocabulary by randomly determining word lengths and selecting letters from the English alphabet. The translation rules are provided as a sequence of bilingual dictionaries. For instance, given three constructed languages L0, L1, and L2, two bilingual dictionaries are created to map L0 to L1 and L1 to L2. Each dictionary entry is limited to word-to-word translation, because phrase-level translation proves to be too difficult for existing models in our pilot experiment. The difficulty of the task can be controlled by varying the number of languages constructed. Each bilingual dictionary, containing task-relevant information, appears only once in the context, while the vocabularies of created languages are taken as haystack.

In **Single-hop Translation**, the model is provided with a source text (more than one word) in one synthetic language and must translate it into another language. Importantly, there exists a direct bilingual dictionary between the selected source and target languages, so that the model can solve the task by retrieving corresponding translations using words in the source text as keys.

A more challenging task, **Multi-hop Translation**, further requires the model to perform a sequence of translations across multiple languages. Without a direct bilingual dictionary between the source and target languages, the model must correctly apply consecutive bilingual dictionaries while keeping track of the translated outputs across the intermediate languages. Both single-hop and multi-hop translation are evaluated with exact match of the reference target text.

Letter Coverage. In this task, the model needs to identify three words in the source language, such that the union of the first letters of all their translations across every language contains the maximum number of distinct letters. This task evaluates the model's ability to fully understand and leverage all provided translation rules. During evaluation, we obtain all translated words for the model-selected words and compare the size of the union of first letters with the reference value (exact match of the size).

3.3 Properties of SYNC

We highlight several important properties of SYNC that differentiates it from existing synthetic benchmarks for long-context evaluation.

Varying Levels of Capabilities. Each domain contains tasks that differs in the *number of hops* and the *hop range*, which estimate the levels of capabilities required for task solving. A hop refers to one piece of information within the context that is needed to solve the task. When multiple hops *must* be chained to correctly solve the task, the distance between consecutive hops is the hop range. A higher hop count requires the model to identify more pieces of information, while a large hop range tests the model's ability to combine information scattered across a wide range of the context.

Model	Connected Nodes 32K	64K	128K	Shortest Path 32K	64K	128K	Longest Path 32K	64K	128K
Llama-3.3-70B-Instruct	98.7	92.7	11.3	57.3	46.0	8.0	18.7	16.0	0.0
Mistral-Large-Instruct	90.0	52.7	12.7	57.3	42.0	8.7	22.0	4.0	0.0
DeepSeek-Distill-Llama-70B	90.7	74.0	2.7	52.7	52.7	0.0	13.3	11.3	0.0
DeepSeek-Distill-Qwen-32B	74.7	56.0	20.0	55.3	37.3	8.7	13.3	10.0	0.0
GPT-4o	94.7	93.3	**96.0**	77.3	**77.3**	70.7	9.3	8.0	2.7
Gemini-2.0-Flash	**100.0**	**98.7**	88.0	**80.0**	76.0	**71.3**	**37.3**	**26.0**	**26.7**

Model	Single-hop Translation 32K	64K	128K	Multi-hop Translation 32K	64K	128K	Letter Cover 32K	64K	128K
Llama-3.3-70B-Instruct	84.0	86.0	0.0	37.3	33.3	0.0	**1.3**	1.3	0.0
Mistral-Large-Instruct	68.7	40.7	3.3	18.7	12.7	0.0	**1.3**	0.0	0.0
DeepSeek-Distill-Llama-70B	82.7	72.0	0.0	38.7	26.7	0.0	**1.3**	0.7	0.0
DeepSeek-Distill-Qwen-32B	74.0	38.7	8.7	45.3	20.0	0.0	**1.3**	1.3	0.0
GPT-4o	80.7	78.0	69.3	70.0	71.3	**51.3**	**1.3**	0.7	0.0
Gemini-2.0-Flash	**98.7**	**90.7**	**80.0**	**93.3**	**80.7**	27.3	0.7	**2.0**	0.0

Table 3: Performance of models with more than 30B parameters on SYNC. Performance of other models is reported in Appendix B. The best model in each setup is **bolded**. Performance is visually represented by a color scale from white (0) to green (100). From left to right, the tasks demand retrieval, state tracking, and global context understanding capabilities. Existing LLMs struggle with tasks beyond simple retrieval, suffering significant degradation on state tracking and global context understanding tasks.

Although NIAH tasks can involve retrieving multiple values in the context (Kamradt, 2023), their *hop range* is effectively zero because each necessary key-value pair can be retrieved independently. In contrast, our suite spans simple retrieval tasks to more complex ones (Shortest Path and Multi-hop Translation) that demand both a higher number of hops and longer hop ranges. Furthermore, we include tasks that require understanding all relevant information in the context (Longest Path and Budget-Aware Translation), maximizing both hops and ranges. Based on the qualitative analysis of hop counts and ranges, we categorize our tasks into *retrieval*, *tracking*, and *global understanding* tasks.

Shared Context across Tasks. Within each domain, we reuse the *same context* for three tasks of varying difficulty. Sharing the same context decouples the difficulty stemming from the context itself from the complexity of the task, allowing for a *controlled comparison of model capabilities* at different levels. Existing synthetic benchmarks such as RULER use different contexts for tasks of varying complexity (Hsieh et al., 2024), which obscures the assessment of the gap between capabilities.

4 Experiments

Models. We benchmark 14 LLMs, including 12 open-source models and 2 proprietary models, all of which can consume 128K or more tokens. Details of the models are provided in Appendix A. All models are evaluated under the 0-shot setting, as our pilot experiments showed that human-aligned models perform worse with in-context learning demonstrations. Due to high computational cost, for each setup, we ran model inference with greedy decoding once.

Benchmark Configurations. For graph understanding tasks, we generate DAGs with varying sizes. Specifically, we consider graphs with 10, 15, and 20 nodes. For each node count configuration, 50 DAGs are generated with an edge density of 0.15, resulting in a total of 150 graphs. To guarantee that each graph is topologically unique, we compute the hash of its canonical form. Node labels are assigned sequentially, starting from 0.

For translation tasks, we construct vocabularies and bilingual dictionaries with different numbers of languages. We consider three configurations with 3, 5, and 7 languages, respectively. For each configuration, 50 distinct sets of vocabularies and dictionaries are generated. Each language comprises 250 words, with each word having a length between 3 and 7 letters. Phrases are limited to at most 5 words, and each bilingual dictionary contains 50 entries.

We insert haystack to create tasks with context lengths of 32K (32,768), 64K (65,536), and 128K (131,072) tokens, as counted by the tokenizer of Mistral-Large-Instruct.

These parameters are freely adjustable to alter data complexity and diversity. We do not expand the graph size or increase the number of languages further, because the complexity might surpass the model's capability for a meaningful evaluation.

5 Results

5.1 Main Results

Table 3 shows the performance of the six models with more than 30B parameters on SYNC. Results for other models are provided in Appendix B. Although current LLMs can consume long contexts, they still **struggle with tasks beyond simple retrieval**. Among the six LLMs, four achieve over 70% accuracy on listing connected nodes and performing single-hop translations with context lengths up to 64K. However, only the two proprietary models, Gemini-2.0-flash and GPT-4o, maintain over 70% accuracy on the tracking tasks (i.e., Shortest Path and Multi-hop Translation). On the most challenging global understanding tasks (i.e., Longest Path and Letter Coverage), all models achieve below 40% accuracy.

Open-source models experience **catastrophic degradation at 128K context length**. Even Llama-3.3-70B, which remains stable from 32K to 64K, suffers a sudden drop at 128K. Upon manual inspection, we find that at such extreme lengths, models often fail to follow prompt instructions, resulting in invalid responses. We hypothesize that instruction-following abilities diminish when models approach their maximum context length, likely because they have less exposure to very long inputs during training.

Comparisons with Other Synthetic Tasks. For comparisons, we include the extraction of corresponding keys with given values from JSON file (JSON KV) (Liu et al., 2023b; Yen et al., 2024). We also consider two tasks from RULER (Hsieh et al., 2024): NIAH tests augmented with multiple values, which are reported to be more challenging than other synthetic NIAH tests; and variable tracking, which targets the state-tracking capability.

Table 4 presents the performance of GPT-4o and Gemini-2.0-flash on SYNC and existing synthetic tasks. The tasks in SYNC are **more difficult overall**, especially for those requiring capabilities beyond retrieval. While existing benchmarks extend standard NIAH with additional key-value pairs and more distracting content, fundamentally they only test retrieval. In SYNC, state tracking and global context understanding tasks require gathering distant clues and synthesizing them, demanding more advanced reasoning. Furthermore, it is noteworthy that high performance on our retrieval tasks relies on the usage of CoT during inference (discussed in §5.2), whereas existing synthetic tasks generally do not. This demonstrates that SYNC offers *sufficient complexity* to challenge future, more advanced models.

Task	GPT-4o 64K	GPT-4o 128K	Gemini-2.0-flash 64K	Gemini-2.0-flash 128K
JSON KV	100.0	100.0	98.0	92.0
RULER				
NIAH Multi-Value	100.0	99.5	99.8	87.8
Variable Tracking	99.6	<u>99.8</u>	<u>100.0</u>	<u>100.0</u>
SYNC *Graph*				
Connected Nodes	94.0	96.0	92.0	88.0
Shortest Path	76.7	70.7	76.0	71.3
SYNC *Translation*				
Single-hop Trans.	78.0	69.3	90.7	80.0
Multi-hop Trans.	71.3	51.3	80.7	27.3

Table 4: Performance of GPT-4o and Gemini-2.0-flash on SYNC and other synthetic tasks. We highlight perfect performance with green. Within the same benchmark, state-tracking tasks that yield higher performance than retrieval tasks are <u>underlined</u>. Using the same context across tasks mitigates the counfounding effect of input context on task difficulty, ensuring tasks requiring more complex capabilities are more challenging.

Sharing the same input context across tasks enables **controlled comparisons of model capabilities**. RULER uses different input contexts for each task, making task performance attributable to both the difficulty of understanding the context and the complexity of the capabilities required to solve the task. Although state tracking is intuitively more complex than retrieval, the absolute performance on RULER's variable tracking task is often higher than that on the multi-value NIAH task. On SYNC, model performance declines as task complexity increases.

Controlled comparisons also allow for precise analysis of model capabilities. For example, Gemini-2.0-flash suffers a more significant decline in multi-hop translation than in single-hop translation when the same input context increases from 64K to 128K tokens. This reveals that while the model retains its ability to understand lengthened contexts and perform retrieval, it lacks the robust state-tracking capability needed to support multi-hop reasoning in longer contexts.

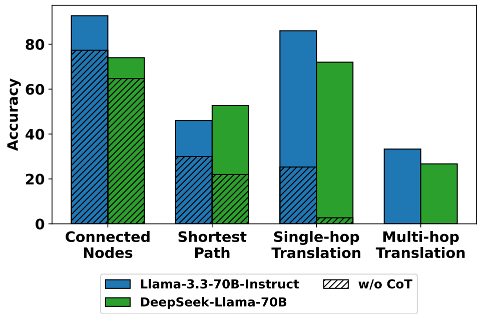


Figure 1: Model performance with and without CoT. Without CoT, performance decreases substantially on SYNC.

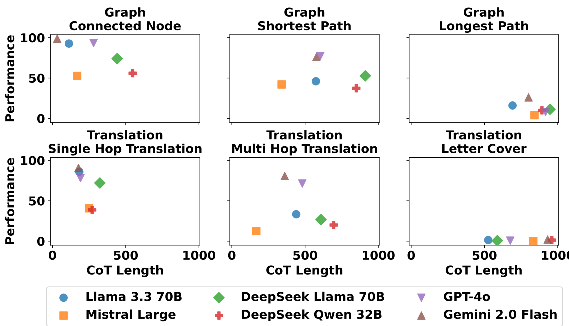


Figure 2: Accuracy and average CoT length for each task in SYNC at 64K context length. Complex tasks prompt longer CoTs, yet more CoT tokens do not necessarily lead to better performance.

5.2 Analysis of Chain-of-Thought (CoT)

In our main experiments, we allow models to use CoT in their responses. To study the effects of CoT, we force the model to output the answer directly by prepending the answer prefix of each task to the model response. As shown in Figure 1, both Llama-3.3-70B and its DeepSeek R1 distilled variant have significant performance drops without CoT. Notably, both models approach 0% accuracy on multi-hop translations when CoT is disabled, indicating the **necessity of CoT** on SYNC.

Figure 2 shows the CoT length and accuracy across different tasks at a 64K context length. The average CoT length evidences the proposed task complexity ordering, as models generally produce **longer CoTs for more complex tasks**. For retrieval tasks, most models generate fewer than 500 CoT tokens. In contrast, global context understanding tasks often elicit CoTs exceeding 500 tokens, with DeepSeek-distilled models reaching 1,000 tokens, though they still fail to solve these tasks.

We further investigate the scaling effect of CoTs by forcing models to generate CoTs of different lengths ranging from 0 to 8192 tokens (Figure 3). Llama-3.1-8B and Llama-3.3-70B are not explicitly trained for long CoTs and therefore do not consistently benefit from longer reasoning chains across all tasks. Their DeepSeek-distilled variants show performance gains with longer CoTs on the Shortest Path and Longest Path tasks, but not on translation tasks. Per human inspection, both models make translation errors at the intermediate steps (usually the first step) and do not self correct. The discrepancy in domain performance highlights the importance of incorporating multiple domains in our benchmark. On graph understanding tasks, the performance of DeepSeek-distilled models plateaus with long CoTs, suggesting that SYNC has *sufficient capacity* for evaluating reasoning models.

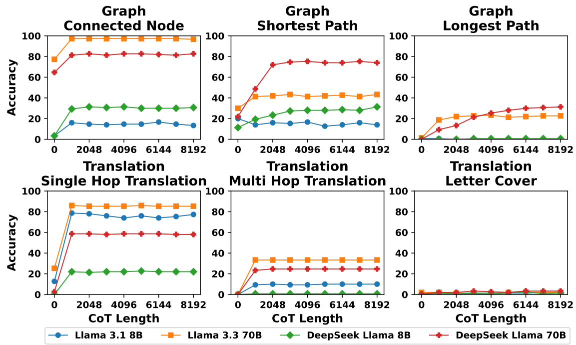


Figure 3: Accuracy under enforced CoT length for each task at 64K context length. DeepSeek-distilled models show improvements with longer CoT on graph understanding tasks, but not on translation tasks.

5.3 Correlation with Realistic Tasks

We study the correlation between SYNC and realistic tasks to understand how well our benchmark can predict real-world performance. We leverage the data released by Yen et al. (2024), which adapts existing datasets for long-context evaluation. Specifically, we include Natural Questions (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018) in the retrieval-augmented generation setup, single-document QA from InfiniteBench (Zhang et al., 2024b) and NarrativeQA (Kočiský et al., 2018), as well as single-document summarization using InfiniteBench. To explore in-context learning under long-context conditions, we use BANKING77 (Casanueva et al., 2020) and CLINC150 (Larson et al., 2019), where long contexts are formed by the demonstration examples. Additionally, we take the subsets from

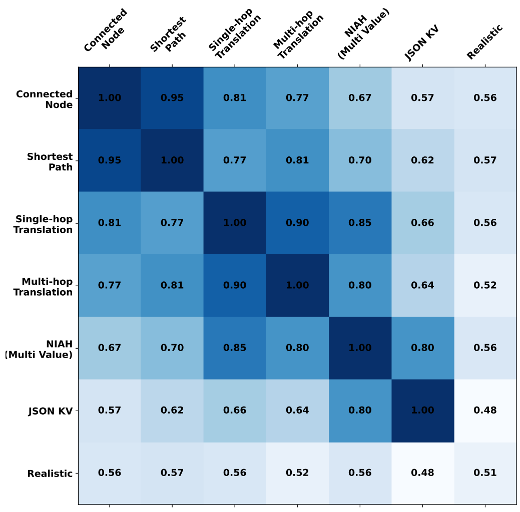


Figure 4: Spearman rank correlation between synthetic tasks and realistic tasks. Higher values indicate stronger alignment in the model rankings produced by two tasks. Correlations with realistic tasks are aggregated with macro average, with detailed breakdown in Appendix C.

LongBench v2 covering multi-document QA and code-based QA (Bai et al., 2024b). We follow the suggested evaluation metrics paired with the released data.

We compute the Spearman ranking correlation between each of the tasks in SYNC (excluding global context understanding tasks, which most models fail at) and the aforementioned realistic tasks. We measure the correlations based on all 14 models at 32K and 64K context lengths, then take the average across context lengths. We do not include 128K because most models perform near zero at that length, making ranking correlations uninformative. To compare with overall realistic tasks, we also aggregate correlations across realistic tasks. Note that the aggregated realistic task does not perfectly correlate with itself, as macro average is employed.

As shown in Figure 4, among all synthetic tasks, the Shortest Path task in SYNC achieves the highest overall correlation with realistic tasks. We think that the real-world tasks we study might be relying more on the state tracking capability. Although other tasks in SYNC do not surpass the NIAH task augmented with multi values, they still show higher correlation with realistic tasks than the other baseline synthetic task. Interestingly, SYNC tasks can even exceed the aggregated correlation that realistic tasks have with each other, suggesting that our tasks can serve as **proxies for real-world performance**.

	Shortest Path			Longest Path		
Model	**NA**	**IV**	**SO**	**NA**	**IV**	**SO**
Llama-3.3-70B-Instruct	8.0	44.7	1.3	5.3	45.3	33.3
Mistral-Large-Instruct	8.7	49.3	**0.0**	24.0	35.3	36.7
DeepSeek-Llama-70B	**4.7**	42.0	0.7	**0.0**	59.3	**29.3**
GPT-4o	11.3	**11.3**	**0.0**	18.0	22.0	52.0
Gemini-2.0-Flash	8.0	16.0	**0.0**	12.0	**15.3**	46.7

Table 5: Percentage of different error types on Shortest Path and Longest Path tasks at 64K context length. NA: No Answer; IV: Invalid Path; SO: Suboptimal Path. Invalid Path is the most common error on Shortest Path, while models start to produce more suboptimal paths on Longest Path.

5.4 Error Analysis

To better understand model behavior, we perform an error analysis on the Shortest Path and Longest Path tasks. Table 5 presents the distribution of error types at a 64K context length, categorized as follows: (1) *No Answer*—the model fails to produce a response (example: the model repeats the context *"The longest path in the graph is from Node 9 to Node 0. There is no directed edge from Node 9 to Node 9 ..."*); (2) *Invalid Path*—the predicted path includes at least one edge not present in the graph (example: the model generates *"Longest Path: Node 0, Node 9, Node 10, Node 12, Node 3, Node 8, Node 18, Node 11, Node 19"* while the edge from Node 10 to Node 12 does not exist in the graph); (3) *Suboptimal Path*—the predicted path is valid but its length differs from the reference shortest or longest path (example: the model generates *"Longest Path: Node 19, Node 15, Node 10, Node 8, Node 7, Node 5, Node 9"* while the path length is shorter than 9, the length of the longest path).

A notable portion of errors falls under *No Answer*, likely due to degraded instruction-following ability as context length increases. In the Shortest Path task, suboptimal paths are rare, whereas they are more common in the Longest Path task, where the absence of a fixed start or end node expands the solution space. This pattern suggests that models often resort to brute-force search rather than employing more efficient strategies (e.g., linear-time algorithms for DAGs).

Invalid Path is prevalent in both tasks. While models can often identify valid local connections, maintaining correctness over longer chains of reasoning remains difficult. Manual inspection of 20 invalid-path cases revealed that all errors occurred in the middle of the path, suggesting that while shallow reasoning is manageable, deeper multi-hop reasoning still poses a significant challenge.

6 Conclusions

We introduce SYNC, a long-context evaluation benchmark consisting of synthetic contexts based on graphs and translation rules. Our benchmark includes three tasks per constructed context, each of which targets a specific model capability among retrieval, state tracking, and global context understanding. By eliminating variation in the input context, SYNC achieves more controlled evaluation of model capabilities. Experiments with 14 LLMs show that SYNC is more challenging in two ways: (1) it includes tasks requiring more complex capabilities; (2) chain-of-thought (CoT) reasoning is needed to solve the tasks in SYNC. We also quantitatively illustrate the importance of sharing input contexts, via comparisons with a popular synthetic benchmark. Further analyses reveal the potentials of SYNC for predicting performance of realistic tasks.

7 Limitations

SYNC includes tasks targeting three capabilities per domain: retrieval, tracking, and global understanding. However, model capabilities can be more diverse and complex. For example, some capabilities might entangle each other, making it hard to separate them. We assume that retrieval, tracking, and global understanding are the most significant model capability, and we only consider them when building the benchmark. The tasks in SYNC evaluate specific skills of LLMs, such as math and algorithmic reasoning. While these skills are critical for their performance in many real-world situations, we recognize that the skills we cover are limited. We also note that our dataset does not aim to replace realistic datasets. Rather, our objective is to improve existing synthetic datasets and complement existing LLM evaluation.

Our benchmark is configured into a 0-shot setup, as we observe degraded performance with few-shot demonstrations. 0-shot prompting also allows the usage of longer input contexts. Nevertheless, we recognize that the 0-shot setup relies on models' instruction-following capabilities, and therefore might not be applicable for certain models (e.g., base LLMs).

8 Ethical Considerations

In §5.3, tasks in SYNC show higher correlations with realistic tasks than most baseline synthetic tasks and even some of the realistic tasks. However, we note that correlations varies a lot across tasks, as shown in the detailed breakdown (Figure 5 and 6). Therefore, performance on SYNC should never be the single criteria for determining the usage of models in real-world applications. In addition to experiments on SYNC, models to be deployed should also be experimented on proper real-world tasks to mitigate the potential risks.

Acknowledgments

This work is supported in part by the National Science Foundation through grant IIS-2046016. Shuyang Cao is supported by a Bloomberg Data Science Ph.D. Fellowship. We thank ARR reviewers for their feedback.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen

Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mistral AI. 2024. Mistral nemo.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Preprint*, arXiv:2305.13245.

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.

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. 2024a. 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)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, and Timofey Bryksin. 2024. Long code arena: a set of benchmarks for long-context code models. *Preprint*, arXiv:2406.11612.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099, Torino, Italia. ELRA and ICCL.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan

Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. *Preprint*, arXiv:https://arxiv.org/abs/2406.16264.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, pages 106519–106554. Curran Associates, Inc.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

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. 2024. Can long-context language models subsume retrieval, rag, sql, and more? *Preprint*, arXiv:2406.13121.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024b. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *Preprint*, arXiv:2310.01889.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang

Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo,

Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. *arXiv preprint 2205.11465*.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. 2025. NovelQA: Benchmarking question answering on documents exceeding 200k tokens. In *The Thirteenth International Conference on Learning Representations*.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of EMNLP*, pages 5627–5646.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *Preprint*, arXiv:2410.02694.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024a. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *Preprint*, arXiv:2409.02897.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *North American Association for Computational Linguistics (NAACL)*.

Kaijian Zou, Muhammad Khalifa, and Lu Wang. 2025. On many-shot in-context learning for long-context evaluation. *Preprint*, arXiv:2411.07130.

A Experiment Details

Models. We benchmark 7 open-source LLMs that are pre-trained from scratch, including Llama-3.2-3B, Llama-3.1-8B, Llama-3.3-70B (Grattafiori et al., 2024), Mistral-Nemo-2407, Mistral-Large-2411 (AI, 2024), Phi-3.5-mini (Abdin et al., 2024), and GLM-4-9B (GLM et al., 2024). We use

Model	# of Para.	Context Len
Open-source Models		
Llama-3.2-3B-Instruct	3B	131072
Llama-3.1-8B-Instruct	8B	131072
Llama-3.3-70B-Instruct	70B	131072
Mistral-Nemo-Instruct-2407	12B	131072
Mistral-Large-Instruct-2411	123B	131072
Phi-3.5-mini-Instruct	4B	131072
GLM-4-9B-Chat	9B	131072
DeepSeek Distilled Models		
DeepSeek-R1-Distill-Llama-8B	8B	131072
DeepSeek-R1-Distill-Llama-70B	70B	131072
DeepSeek-R1-Distill-Qwen-7B	7B	130172
DeepSeek-R1-Distill-Qwen-14B	14B	130172
DeepSeek-R1-Distill-Qwen-32B	32B	130172
Proprietary Models		
GPT-4o-2024-11-20	-	128000
Gemini-2.0-flash-001	-	1000000

Table 6: Information about the models used in our experiments. All models support 128K tokens.

their human-aligned variants which can better follow instructions. Besides models that are pre-trained from scratch, 5 models that are further fine-tuned with data distilled from DeepSeek-R1 are tested: DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Llama-70B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B, and DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025). We also include 2 proprietary models, GPT-4o (OpenAI et al., 2024a) and Gemini-2.0-flash (Team et al., 2024). Table 6 summarizes these models.

Infrastructure. The inference is performed with vLLM (Kwon et al., 2023) on 8 A40 GPUs.

Usage of AI Assistant. We use ChatGPT (OpenAI et al., 2024a) for correcting grammar errors in our writing.

B Additional Results

We provide the full results of all models on the synthetic tasks in SYNC in Table 7.

C Correlation Breakdown

We report the detailed breakdown of ranking correlation between synthetic tasks and realistic tasks at 32K and 64K context length in Figures 5 and 6.

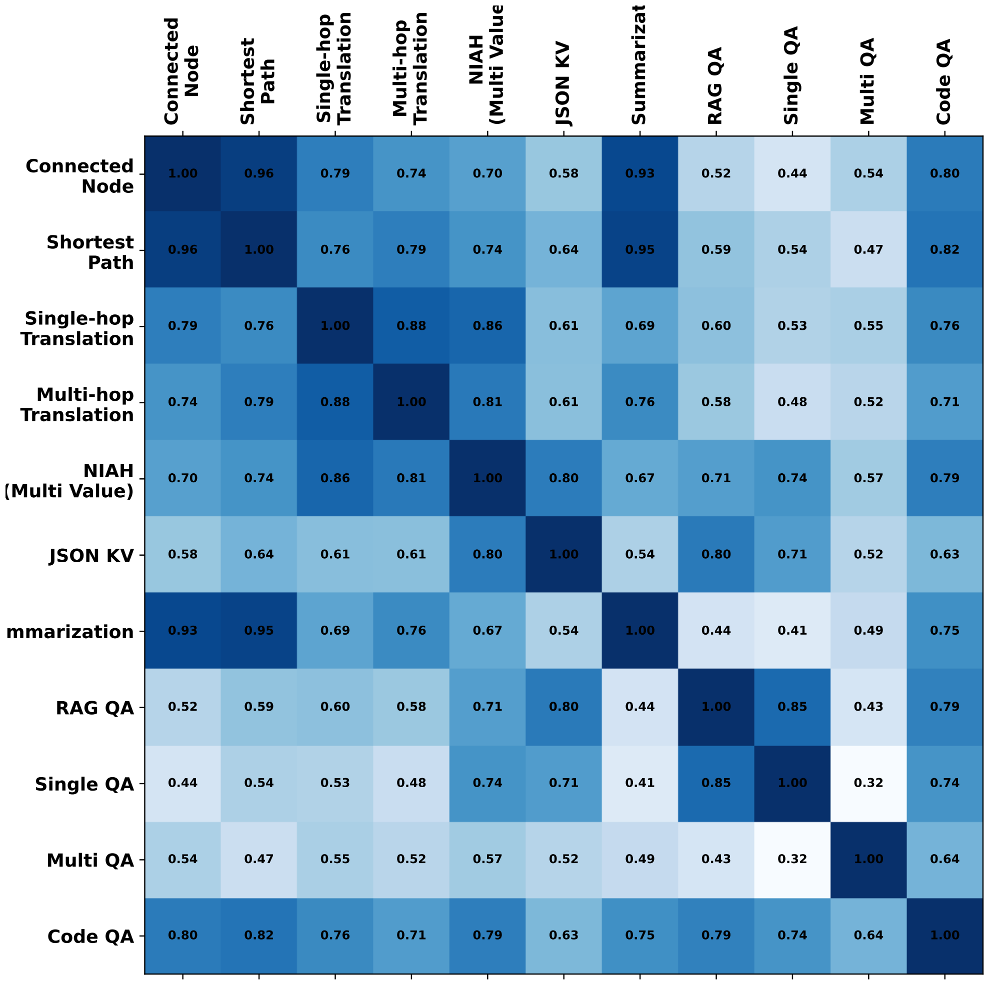
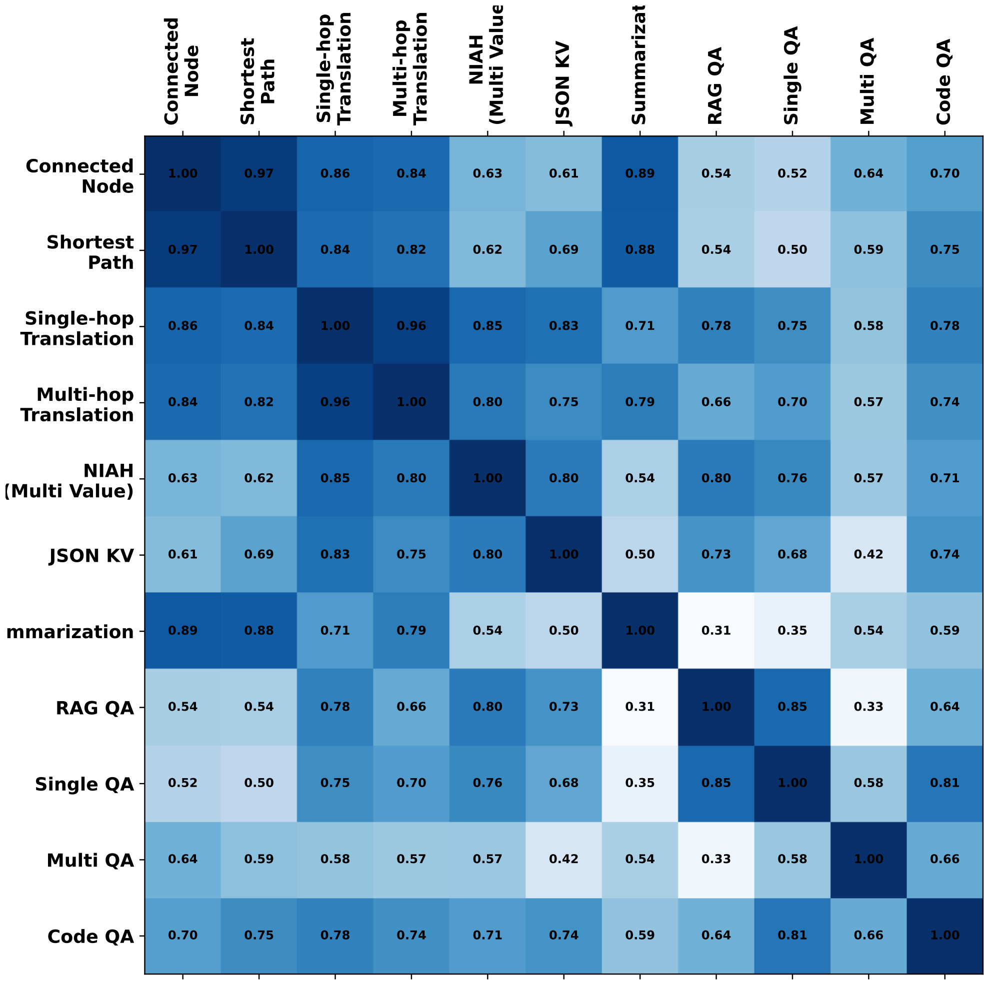


Figure 5: Spearman rank correlation between synthetic tasks and realistic tasks at 32K.



	Connected Node	Shortest Path	Single-hop Translation	Multi-hop Translation	NIAH (Multi Value)	JSON KV	Summarization	RAG QA	Single QA	Multi QA	Code QA
Connected Node	1.00	0.97	0.86	0.84	0.63	0.61	0.89	0.54	0.52	0.64	0.70
Shortest Path	0.97	1.00	0.84	0.82	0.62	0.69	0.88	0.54	0.50	0.59	0.75
Single-hop Translation	0.86	0.84	1.00	0.96	0.85	0.83	0.71	0.78	0.75	0.58	0.78
Multi-hop Translation	0.84	0.82	0.96	1.00	0.80	0.75	0.79	0.66	0.70	0.57	0.74
NIAH (Multi Value)	0.63	0.62	0.85	0.80	1.00	0.80	0.54	0.80	0.76	0.57	0.71
JSON KV	0.61	0.69	0.83	0.75	0.80	1.00	0.50	0.73	0.68	0.42	0.74
mmarization	0.89	0.88	0.71	0.79	0.54	0.50	1.00	0.31	0.35	0.54	0.59
RAG QA	0.54	0.54	0.78	0.66	0.80	0.73	0.31	1.00	0.85	0.33	0.64
Single QA	0.52	0.50	0.75	0.70	0.76	0.68	0.35	0.85	1.00	0.58	0.81
Multi QA	0.64	0.59	0.58	0.57	0.57	0.42	0.54	0.33	0.58	1.00	0.66
Code QA	0.70	0.75	0.78	0.74	0.71	0.74	0.59	0.64	0.81	0.66	1.00

Figure 6: Spearman rank correlation between synthetic tasks and realistic tasks at 64K.

Model	Connected Nodes 32K	64K	128K	Shortest Path 32K	64K	128K	Longest Path 32K	64K	128K
Llama 3.2 3B	14.0	12.0	2.7	12.7	14.0	6.7	0.0	0.0	0.0
Llama 3.1 8B	14.7	12.0	2.7	15.3	20.0	3.3	0.7	0.0	0.0
Llama 3.3 70B	98.7	92.7	11.3	57.3	46.0	8.0	18.7	16.0	0.0
Mistral Nemo	10.7	1.3	8.7	5.3	4.0	1.3	0.0	0.0	0.0
Mistral Large	90.0	52.7	12.7	57.3	42.0	8.7	22.0	4.0	0.0
DeepSeek Llama 8B	45.3	22.0	13.3	16.7	28.7	0.0	1.3	0.7	0.0
DeepSeek Llama 70B	90.7	74.0	2.7	52.7	52.7	0.0	13.3	11.3	0.0
DeepSeek Qwen 7B	8.7	4.7	7.3	2.7	0.0	0.0	0.0	0.0	0.0
DeepSeek Qwen 14B	44.7	21.3	12.7	36.0	20.0	11.3	4.0	2.7	0.0
DeepSeek Qwen 32B	74.7	56.0	20.0	55.3	37.3	8.7	13.3	10.0	0.0
Phi 3.5 mini	12.7	7.3	8.7	12.7	15.3	13.3	0.7	0.0	0.0
GLM 4 9B	29.3	40.7	27.3	20.0	26.0	19.3	0.0	0.0	1.3
GPT-4o	94.7	93.3	**96.0**	77.3	**77.3**	70.7	9.3	8.0	2.7
Gemini 2.0 Flash	**100.0**	**98.7**	88.0	**80.0**	76.0	**71.3**	**37.3**	**26.0**	**26.7**

Model	Single-hop Translation 32K	64K	128K	Multi-hop Translation 32K	64K	128K	Letter Cover 32K	64K	128K
Llama 3.2 3B	35.3	8.0	1.3	0.0	0.0	0.0	0.0	0.7	0.0
Llama 3.1 8B	76.0	74.7	76.0	39.3	28.0	1.3	0.0	0.0	0.0
Llama 3.3 70B	84.0	86.0	0.0	37.3	33.3	0.0	**1.3**	1.3	0.0
Mistral Nemo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral Large	68.7	40.7	3.3	18.7	12.7	0.0	**1.3**	0.0	0.0
DeepSeek Llama 8B	38.7	24.0	0.0	5.3	0.7	0.0	0.7	0.7	0.0
DeepSeek Llama 70B	82.7	72.0	0.0	38.7	26.7	0.0	**1.3**	0.7	0.0
DeepSeek Qwen 7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepSeek Qwen 14B	31.3	20.7	2.7	11.3	6.7	0.0	0.7	**2.0**	0.0
DeepSeek Qwen 32B	74.0	38.7	8.7	45.3	20.0	0.0	**1.3**	1.3	0.0
Phi 3.5 mini	6.0	0.7	0.7	0.0	0.0	0.0	0.0	0.7	0.7
GLM 4 9B	86.0	74.0	65.3	48.0	43.3	2.7	0.0	0.0	0.0
GPT-4o	80.7	78.0	69.3	70.0	71.3	**51.3**	**1.3**	0.7	0.0
Gemini 2.0 Flash	**98.7**	**90.7**	**80.0**	**93.3**	**80.7**	27.3	0.7	**2.0**	0.0

Table 7: Performance of all models on the synthetic tasks in SYNC. The best model for each task setup is **bolded**.