HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

Anonymous authors
Paper under double-blind review

Abstract

There have been many benchmarks for evaluating long-context language models (LCLMs), but developers often rely on synthetic tasks like needle-in-a-haystack (NIAH) or arbitrary subsets of tasks. It remains unclear whether they translate to the diverse downstream applications of LCLMs, and the inconsistency further complicates model comparison. We investigate the underlying reasons behind current practices and find that existing benchmarks often provide noisy signals due to low coverage of applications, insufficient lengths, unreliable metrics, and incompatibility with base models. In this work, we present **HELMET** (**H**ow to **E**valuate **L**ong-context **M**odels **E**ffectively and **T**horoughly), a comprehensive benchmark encompassing seven diverse, application-centric categories. We also address many issues in previous benchmarks by adding controllable lengths up to 128k tokens, model-based evaluation for reliable metrics, and few-shot prompting for robustly evaluating base models. Consequently, we demonstrate that HELMET offers more reliable and consistent rankings of frontier LCLMs. Through a comprehensive study of 51 LCLMs, we find that (1) synthetic tasks like NIAH are not good predictors of downstream performance; (2) the diverse categories in HELMET exhibit distinct trends and low correlation with each other; and (3) while most LCLMs achieve perfect NIAH scores, open-source models significantly lag behind closed ones when the task requires full-context reasoning or following complex instructions—the gap widens with increased lengths. Finally, we recommend using our RAG tasks for fast model development, as they are easy to run and more predictive of other downstream performance; ultimately, we advocate for a holistic evaluation across diverse tasks.[1]

1 Introduction

Long-context language models (LCLMs) unlock a myriad of applications, from summarizing long documents to learning new tasks on the fly with thousands of examples. Many recent benchmarks sought to evaluate language models' long-context abilities (Zhang et al., 2024b; An et al., 2024; Shaham et al., 2023; Bai et al., 2024, *inter alia*). However, recent developments on long contexts (Chen et al., 2023; Xiong et al., 2023; Peng et al., 2024; Fu et al., 2024) still rely on either perplexity or synthetic needle-in-a-haystack tasks (NIAH; Kamradt, 2024; Hsieh et al., 2024). Frontier LCLMs (Dubey et al., 2024; Team et al., 2024c; OpenAI, 2023; Team et al., 2024a) also mostly report NIAH, sometimes with arbitrary subsets of other datasets, as shown in Table 1. Such inconsistency complicates comparisons between different models. It is also unclear whether synthetic NIAH tasks or the few chosen datasets can provide a holistic picture of models' long-context abilities and their performance in real-world applications.

Table 1: Most LCLMs evaluate on synthetic (Syn.) tasks. ♭: base models. ∞: ∞Bench.

Model	Syn.	PPL	∞		ZeroSCROLLS					RAG	ICL
			QA	All	NQA	QS	QL	SQ	All		
Gemini-1.5	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
GPT-4	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Claude-3.5	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Llama-3.1	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗
Phi-3	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗
Jamba-1.5	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Qwen2	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Command R	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Xiong et al.	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗
Chen et al.♭	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Peng et al.♭	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Fu et al.♭	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

[1] Our code and data will be made publicly available.

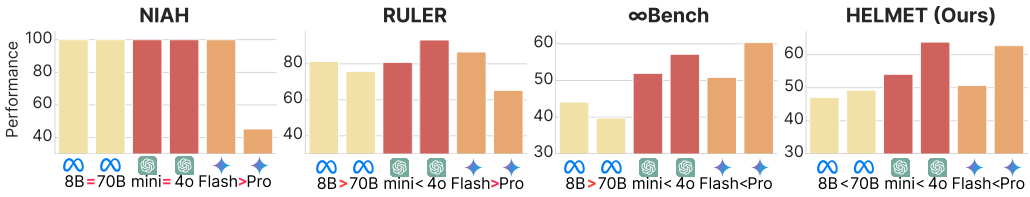

Figure 1: Long-context benchmark results of frontier LCLMs (Llama-3.1 8B/70B, GPT-4o-mini, GPT-4o-08-06, and Gemini-1.5 Flash/Pro) at 128k input length. NIAH is saturated for almost all models; RULER (Hsieh et al., 2024) and ∞BENCH (Zhang et al., 2024b) show unexpected trends for Llama-3.1 (Dubey et al., 2024). In contrast, HELMET demonstrates more consistent rankings of these frontier models.

Why don't model developers agree on these evaluations? We take a closer look and find that existing benchmarks have many critical design flaws, including:

- **Insufficient coverage of downstream tasks**: Existing benchmarks either focus on synthetic tasks (Hsieh et al., 2024) or add simple question answering (Zhang et al., 2024b). Other works study particular aspects of LCLMs, such as summarization (Chang et al., 2024), in-context learning (Li et al., 2024c), and retrieval-augmented generation (RAG; Lee et al., 2024), but they do not provide a holistic evaluation of LCLMs.
- **Inadequate lengths**: Most natural language datasets in existing benchmarks (Shaham et al., 2023; An et al., 2024, Table 4) are too short to test frontier long-context abilities (usually $\geq$128k).
- **Unreliable metrics**: For commonly used long-document QA and summarization tasks, most existing benchmarks still rely on metrics like ROUGE (Lin, 2004), which can be noisy and unreliable (Goyal et al., 2023; Deutsch et al., 2022; Chang et al., 2024).
- **Incompatibility with base models**: Many LCLM developments focus on base models without instruction tuning, but most existing benchmarks require models to be instruction-tuned—hence developers can only rely on synthetic tasks or perplexity.

Consequently, existing benchmarks either are not applicable to long-context works (inadequate lengths or incompatibility) or provide very noisy signals (insufficient coverage or unreliable metrics). We summarize the shortcomings of many of the benchmarks in Table 2. For the three benchmarks that support 128k[2] context length (NIAH, RULER, and ∞BENCH), we use them to evaluate 6 frontier models and report the numbers in Figure 1. We see that NIAH does not reflect differences across models; RULER and ∞BENCH show unexpected trends—on RULER, Gemini Flash is stronger than Gemini Pro; on ∞BENCH, the 70B Llama model is weaker than the 8B one. This raises concerns about the reliability of the benchmarks and how they can differentiate long-context models, which likely contributes to the lack of use of these benchmarks in model development.

To address these challenges, we present **HELMET** (**H**ow to **E**valuate **L**ong-context **M**odels **E**ffectively and **T**horoughly). We curate a *diverse* set of application-centric long-context tasks across seven categories. Beyond widely adopted categories like *synthetic recall*, *long-document question answering (QA)*, and *summarization*, we also add *many-shot in-context learning* (ICL), *retrieval-augmented generation* (RAG), *passage re-ranking*, and *generation with citations* (Gao et al., 2023). We further address the shortcomings of existing benchmarks: (1) we ensure all datasets support input lengths of 128k and are easily extendable to more tokens (§2.1); (2) we introduce reference-based model evaluation for QA and summarization and show that it drastically improves over n-gram overlap metrics (§2.2); (3) we refine the prompts and in-context demonstrations used for all tasks, reducing the evaluation noise by different output formats and allowing base models to be evaluated on most of the categories robustly (§2.3). Together, HELMET enables a holistic evaluation of long-context capabilities and provides more reliable signals for model development. Figure 1 shows that HELMET can clearly differentiate models at different capabilities and reflect comparisons consistent with human perception. More discussions on the improvement of HELMET over previous benchmarks and direct comparisons are in §A.

[2]Throughout the paper, we use binary prefixes k= 2^{10}.

Table 2: Comparison of long-context benchmarks: ZeroSCROLLS (Shaham et al., 2023), LongBench (Bai et al., 2024), L-Eval (An et al., 2024), RULER (Hsieh et al., 2024), ∞BENCH (Zhang et al., 2024b), and our HELMET. L: input tokens. †: All datasets have $L < 128$k except one dataset. ‡: L-Eval uses LLMs to compute reference-free, pairwise win-rates; we design reference-based model evaluation for specific tasks.

	Type of tasks							Benchmark features		
	Cite	RAG	Re-rank	Long-QA	Summ	ICL	Synthetic Recall	Robust Eval.	$L \geq 128$k	Controllable L
ZeroSCROLLS	✗	✗	✗	✓	✓	✗	✗	✗	✗†	✗
LongBench	✗	✓	✗	✓	✓	✓	✓	✗	✗†	✗
L-Eval	✗	✓	✗	✓	✓	✗	✗	✓‡	✗†	✗
RULER	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
∞BENCH	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓
HELMET (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

To understand the progress of LCLMs and how different long-context capabilities correlate with one another, we evaluate a thorough list of 51 LCLMs of various architectures, scales, and training strategies. Our analysis reveals that (1) synthetic tasks are poor indicators of downstream performance (§3.1), (2) different categories in HELMET show distinctive trends (§3.2), and (3) open-source models significantly lag behind closed ones on tasks that require reasoning over long contexts or following complex instructions—the gap further widens with increased lengths (§3.3). Finally, we find that the RAG category strikes a good balance between ease of use, better correlation with downstream tasks, and compatibility with base models. Ultimately, it is imperative to evaluate LCLMs over a diverse spectrum of categories. We hope that our insights offer a more effective way to evaluate LCLMs for future model development and benchmarking.

2 OUR BENCHMARK: HELMET

In this work, we seek to overcome the shortcomings of existing benchmarks by meeting the following desiderata: (1) diverse coverage across different tasks and capabilities of LCLMs, (2) controllable lengths that support more than 128k input tokens, and (3) reliable evaluation for both base and instruction-tuned models. In this section, we describe the datasets used in HELMET and how they improve upon existing evaluation benchmarks in terms of settings and metrics.

2.1 REALISTIC AND DIVERSE LONG-CONTEXT APPLICATIONS

Retrieval-augmented generation (RAG). We use open-domain question answering (ODQA)—which requires retrieving from a knowledge corpus and then generating correct answers (Chen et al., 2017)—as a representation of retrieval-augmented generation (RAG) applications. We use Natural Questions (NQ; Kwiatkowski et al., 2019), TriviaQA (TQA; Joshi et al., 2017), HotpotQA (Yang et al., 2018), and PopQA (Mallen et al., 2023). We use the gold passage information (the passage that can answer the question) from Petroni et al. (2021), or otherwise use any passage that contains the answer.

For a particular input length L, we first determine the number of passages k that can fit within L tokens, then retrieve k passages[3] from the corpus[4] that *do not contain the answer* as *distractors*. This differs from previous works that *randomly sample passages* from the corpus (Lee et al., 2024) and is more realistic and challenging. For NQ, TQA, and PopQA, we take the top $k-1$ distractors and insert the gold passage at six evenly distributed positions following Liu et al. (2023). For HotpotQA, which requires two gold passages, we combine them and the top $k-2$ distractors and randomly shuffle them into three permutations. We use substring exact match (SubEM; whether the

[3]We use `Alibaba-NLP/gte-large-en-v1.5` for retrieval (Zhang et al., 2024a).
[4]We use Wikipedia 2019-8-01 dump, split into 100-word passages (Petroni et al., 2021).

Table 3: Overview of evaluation datasets. We select datasets that cover various important long-context capabilities. SubEM: substring exact match.

Category	Dataset	Metrics	Description
Retrieval-augmented generation	Natural Questions	SubEM	Factoid question answering
	TriviaQA	SubEM	Trivia question answering
	PopQA	SubEM	Long-tail entity question answering
	HotpotQA	SubEM	Multi-hop question answering
Passage re-ranking	MS MARCO	NDCG@10	Rerank passage for a query
Generation with citations	ALCE ASQA	Recall, Cite	Answer ambiguous questions with citations
	ALCE Qampari	Recall, Cite	Answer factoid questions with citations
Long-document QA	NarrativeQA	Model-based	Book and movie script QA
	∞BENCH QA	ROUGE F1	Novel QA with entity replacement
	∞BENCH MC	Accuracy	Novel multiple-choice QA with entity replacement
Summarization	∞BENCH Sum	Model-based	Novel summarization with entity replacement
	Multi-LexSum	Model-based	Summarizing multiple legal documents
Many-shot in-context learning	TREC Coarse	Accuracy	Question type classification, 6 labels
	TREC Fine	Accuracy	Question type classification, 50 labels
	NLU	Accuracy	Task intent classification, 68 labels
	BANKING77	Accuracy	Banking intent classification, 77 labels
	CLINC150	Accuracy	Intent classification, 151 labels
Synthetic recall	JSON KV	SubEM	Retrieve a key in JSON dictionary
	RULER MK Needle	SubEM	Retrieve the needle (a number) within noisy needles
	RULER MK UUID	SubEM	Retrieve the needle (a UUID) within noisy needles
	RULER MV	SubEM	Retrieve multiple values for one needle (key)

answer is included in the output), following previous work (Mallen et al., 2023; Asai et al., 2024a). §B.1 offers more details.

Generation with citations (Cite). We leverage ALCE (Gao et al., 2023) to evaluate LCLMs on a realistic application of answering questions while providing correct attributions (Bohnet et al., 2022). Given multi-faceted questions and relevant passages, models are required to generate a long-text answer and cite supporting passage IDs at the end of each sentence. This tests models' ability to utilize the passages in the context and also to follow the instructions about citation formats. We use the ASQA (Stelmakh et al., 2022) and QAMPARI (Rubin et al., 2022) subsets from ALCE. For an input length L, we first determine the number of passages k, and use the top k retrieved passages from Wikipedia as contexts. The model's generations are evaluated on correctness and citation quality, and we report the average across all metrics. See §B.2 for more details.

Long-document question answering (LongQA). We use NarrativeQA (Kočiský et al., 2018) and the English book QA and multiple choice (MC) subsets from ∞BENCH (Zhang et al., 2024b) for evaluating long-document QA. We select those tasks for their abundant context lengths (Table 4). We truncate the document from the end based on L. We use ROUGE F1 for ∞BENCH QA (answers are mostly entity names) and accuracy for ∞BENCH MC. For NarrativeQA, as the answer can be long text and open-ended, we use a model-based evaluation (§2.2).

Table 4: Dataset lengths.

Datasets	Medium	Max
ZeroSCROLLS		
QASPER	6k	12k
GovReport	12k	33k
QuALITY	9k	11k
SQuALITY	8k	10k
HELMET		
NarrativeQA	73k	518k
∞BENCH QA	191k	835k
∞BENCH MC	167k	835k
∞BENCH Sum	154k	835k
Multi-LexSum	90k	5m

Summarization (Summ). Summarization tests LCLMs' ability to aggregate information across the contexts. We choose Multi-LexSum (legal document summarization) and the English summarization task from ∞BENCH (novel summarization) for their long lengths (Table 4). We truncate the document from the end based on the evaluation length L. We use our model-based evaluation (§2.2) for both datasets instead of the commonly used ROUGE, as it better reflects human judgment.

Passage re-ranking (Re-rank). Re-ranking retrieved passages based on their relevance to the query is an important application of LCLMs (Sun et al., 2023). The task requires the model to retrieve relevant information and compare and reason over different parts of the contexts. We use the MS

MARCO dataset (Bajaj et al., 2018), where each instance contains a query and passages retrieved by BM25 (Robertson & Zaragoza, 2009) from the Internet. Each passage has annotations of a relevance label—perfect, highly relevant, or not relevant. We determine the number of passages k from the input length L, and randomly sample k passages with balanced labels for each test query. The model is prompted with the query and the shuffled k passages and is instructed to output the top-10 document IDs ranked by relevance. We report NDCG@10. Details are in §B.3.

Many-shot in-context learning (ICL). In-context learning (ICL) is a key ability that enables LLMs to adapt to new tasks on the fly (Brown et al., 2020). Ratner et al. (2023); Xu et al. (2024); Li et al. (2024c); Bertsch et al. (2024) explore performing many-shot ICL (with thousands of examples) with LCLMs. Following Bertsch et al. (2024), we focus on datasets with large label spaces: TREC-coarse, TREC-fine (Li & Roth, 2002), BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and NLU (Liu et al., 2019). We adjust the number of shots to control the input length L, and the number of examples in each class is balanced. We report the accuracy on the test set for each dataset.

One *difference* from previous works is that we map original natural language labels (e.g., *location*) into numbered labels (i.e., *0*, *1*) to test how well a model can learn new tasks instead of relying on its pre-trained priors (Wei et al., 2023; Pan et al., 2023; Min et al., 2022). More details are in §B.4.

Synthetic recall. Synthetic recall tasks, such as needle-in-a-haystack (NIAH), stress test models' ability to recall relevant information (needle) from long contexts. They gain popularity for being easy to use (can test any arbitrary length) and easy to control (can put the "needle" at any position). For this category, we select multiple synthetic recall tasks from RULER (a more comprehensive version of NIAH; Hsieh et al., 2024) and also add a JSON KV retrieval task (Liu et al., 2023), which we find more challenging. Following previous works, we report the percentage of the ground truth answers that are substrings in the generation (SubEM). Refer to §B.5 for more details.

2.2 Reliable evaluation metrics

Existing long-context benchmarks (Zhang et al., 2024b; Shaham et al., 2023) largely rely on n-gram overlap metrics like ROUGE (Lin, 2004) that have been shown to correlate poorly with human judgment for tasks with long outputs, such as summarization (Goyal et al., 2023; Deutsch et al., 2022). L-Eval (An et al., 2024) uses LLMs to score reference-free "win-rates", which neglects the available answer annotations and always requires evaluating model pairs. Instead, we design *reference-based model evaluation* for long-document QA and summarization that are more reliable and easy to use.

Question answering. In NarrativeQA, we prompt GPT-4o[5] with the question, the ground truth, and the model output to check for fluency and correctness. The fluency score is either 0 (incoherent or repetitive) or 1 (fluent), and the correctness score takes on the value of 0 (incorrect), 1 (partly correct), 2 (correct but not fully relevant), and 3 (correct and relevant). We take the product of the two as the final score and normalize it to a range of $[0, 100]$.

Summarization. Following previous works (Kamoi et al., 2023; Zhang & Bansal, 2021), we first decompose the gold summary into atomic claims and use GPT-4o to check if each claim is supported by the generation (recall) and if each sentence in the generation is supported by the reference summary (precision). We then calculate the F1 score from the recall and precision scores. We also ask GPT-4o to check for fluency (0 or 1) and take the final score as the product of the fluency and the F1 score.

Empirically, our reference-based model evaluation reflects more consistent trends, as shown in Figure 2: (1) Llama-3.1-8B-Inst achieves similar ROUGE scores to GPT-4o, while our evaluation shows a significant gap. (2) Our metric better identifies incoherent generations and shows lower performance for models with smaller context windows, such as the Mistral models. (3) Our metric shows a substantially more positive trend for GPT-4o as input length increases, whereas ROUGE-L remains within a difference of 2 absolute points. We further validate the model-based evaluation through human studies, which suggests that our new metrics highly agree with human judgments:

[5] `GPT-4o-2024-05-13`

	Multi-LexSum										InfBench Sum									
	ROUGE ?					Ours					ROUGE ?					Ours				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o-08	23.1	23.7	24.6	24.9	24.9	44.0	48.2	55.7	55.3	56.1	15.8	16.6	17.1	17.1	16.8	13.9	23.2	26.1	28.4	31.1
Gemini-1.5-Pro	24.5	24.8	25.8	25.1	25.8	41.5	44.8	54.2	56.4	58.1	14.8	14.8	15.7	16.6	16.1	18.6	20.3	24.4	34.6	32.0
Claude-3.5-sonnet	23.5	22.1	22.5	21.9	22.0	41.6	43.3	50.7	46.0	50.9	13.9	14.8	15.9	14.5	14.2	13.2	22.8	29.6	28.3	30.5
Llama-3.2-1B-Inst	21.2	22.0	22.1	23.0	23.1	16.4	19.6	20.3	21.3	20.3	14.0	15.0	15.5	16.5	16.6	1.1	1.8	2.8	3.2	2.9
Llama-3.2-3B-Inst	21.3	21.2	22.8	23.3	23.7	32.9	37.5	42.4	40.2	43.6	15.6	16.4	16.3	16.5	16.5	5.0	8.5	9.9	11.9	12.4
Llama-3.1-8B-Inst	24.1	24.8	24.5	25.0	25.2	41.5	42.8	46.4	45.0	46.9	15.7	16.3	16.3	17.0	17.4	5.0	8.6	11.9	15.3	16.2
Llama-3.1-70B-Inst	24.2	24.3	24.9	25.7	25.2	44.0	46.4	50.3	51.4	51.6	15.8	16.5	17.0	16.9	17.5	11.5	17.1	20.6	20.1	19.8
Mistral-7B-Inst-v0.1	23.6	22.6	23.0	22.4	22.1	25.7	20.4	21.1	21.1	16.1	11.2	12.2	12.5	12.4	12.6	1.7	2.8	2.9	2.8	4.2
Mistral-7B-Inst-v0.3	22.7	23.3	23.8	21.6	19.0	30.9	37.2	37.4	32.3	24.4	14.7	15.3	15.2	13.4	12.3	9.1	11.9	10.6	1.8	0.1

Figure 2: Comparison between ROUGE-L F1 and our model-based evaluation metric on summarization tasks. Our metric shows more consistent trends: it reflects the performance gain on GPT-4o with increased input length, while ROUGE remains almost the same; our metric also clearly distinguishes models while ROUGE shows little difference.

for example, on ∞BENCH Sum, our metric reaches a human-model agreement of Cohen's $\kappa = 0.91$ and $\kappa = 0.76$ on summary precision and recall, respectively. More details on the human validation studies can be found in §B.

2.3 ROBUST PROMPTING AND CONTROLLED EVALUATION SETTINGS

Robust prompting reduces noise and enables evaluation on base models. Many long-context benchmarks require models to follow instructions and only support evaluating instruction-tuned models (Zhang et al., 2024b; Shaham et al., 2023). However, many model developments do not apply instruction tuning (Chen et al., 2023; Fu et al., 2024) and these models can only rely on perplexity or synthetic tasks. To support long-context research efforts, we design our benchmark so that at least a portion of the datasets supports evaluating base models.

Existing benchmarks mostly use zero-shot prompting (Shaham et al., 2023; Zhang et al., 2024b), which leads to noisy output formats, especially for base models. For example, the model may output a long answer in RAG when a short answer is required. We add two-shot demonstrations in the prompt for all tasks to address this problem.[6] For long-document QA and summarization, we replace the original document with a placeholder phrase to reduce the number of input tokens in the ICL example. As shown in Table 8, we observe that both base and instruct models substantially benefit from the demonstrations.

Furthermore, we employ the length-instruction-enhanced evaluation from L-Eval for long-generation tasks (i.e., summarization), which is shown to have substantially more consistent and reliable evaluations (An et al., 2024). As a result, we find that our reproduction of previous datasets, such as ∞BENCH QA tasks, better reflects the capabilities of LCLMs, as shown in Table 7. The use of demonstration and better instructions thus better depicts how models perform in real applications.

Controlled input length and difficulty. An important dimension to consider when evaluating LCLMs is the input length L, as longer inputs can provide more information while challenging the model's ability to process distracting contexts. As we discussed in §2.1, we can control the input length L for each task by either adjusting the number of retrieved passages, the number of demonstrations, or truncating the document text to fit within the specified lengths. Consequently, we are able to study model performance at or beyond the length of current frontier LCLMs ($\geq$ 128k).

As shown in Figure 1 and 2, HELMET provides rankings more consistent with human perception of model performance. The diverse tasks, controllable lengths, and robust metrics and prompting enable a thorough examination of LCLMs across multiple dimensions.

3 ANALYSIS

We evaluate 51 LCLMs with HELMET. To our best knowledge, this is the most thorough and controlled comparison of long-context models on diverse applications. These models cover closed-

[6]Except for ICL (the number of shots varies) and RULER (we follow the original formatting).

source models, such as GPT4, Claude, and Gemini, as well as open-source model families, such as Llama-3 (Dubey et al., 2024), Mistral (Jiang et al., 2023), and Phi (Abdin et al., 2024). We also consider models that use different architectures—full-attention transformers (Vaswani et al., 2017), sliding-window attention (Beltagy et al., 2020), and hybrid models with SSM modules (Dao & Gu, 2024). We also benchmark position extrapolation models such as YaRN (Peng et al., 2024) and LongRoPE (Ding et al., 2024). We list all the models evaluated in Table 15. We evaluate each model at input lengths: $L \in \{8\text{k}, 16\text{k}, 32\text{k}, 64\text{k}, 128\text{k}\}$, where L is the number of Llama-2 tokens (Touvron et al., 2023), and use greedy decoding for all models for consistency. We randomly sample 100 examples from each dataset, and more details are in §D.

3.1 Simple synthetic tasks are poor predictors of real-world performance

Many model developers rely on simple synthetic tasks, such as NIAH, for evaluating long-context language models, but it is unclear if these tasks are representative of real-world performance. To this end, we calculate Spearman's rank correlation ρ with the performance of 30 instruction-tuned models between synthetic and real-world tasks. First, Figure 3 shows that none of the synthetic tasks achieves an average correlation higher than 0.8. We also make the following observations.

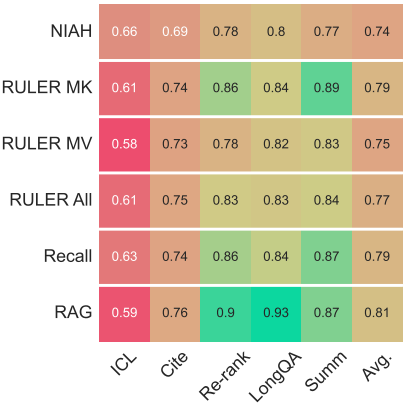

Figure 3: Spearman's rank correlation at 128k input length, calculated across 30 instruction-tuned models.

Not all synthetic tasks are created equal. The original NIAH, which places a needle in the middle of unrelated essays and asks the model to retrieve it, does not correlate well with real-world tasks: all correlations are ≤ 0.8. Using the popular RULER average score—which contains not only NIAH variants but also synthetic aggregation, multi-hop tracing, and QA—also does not give strong correlations (all < 0.85).

We take a closer look at different RULER tasks and find that harder recall-type tasks are more reflective of real-world categories—such as RULER MK, which places distracting needles around the target needle. Despite the overall low correlation, we believe they can still serve as a useful *sanity check* during model development. We put together several such RULER tasks, along with JSON KV, to form the HELMET synthetic recall set (more discussions on this in §E.1).

Tasks with noisier, more distracting contexts better differentiate models. To understand why synthetic tasks poorly correlate with real-world tasks, we plot the performance of different models on NIAH, RULER MK (one of our recall tasks), and HotpotQA (one of our RAG tasks) in Figure 4. We use ∞BENCH QA as a representative real-world task. We find that most models achieve either perfect or near-zero performance on the original NIAH, leaving few data points in the middle and resulting in poor separability between models. On the other hand, RULER MK, whose contexts are more distracting, better distributes the models between 0 and 100% performance.

RAG is a better proxy for real-world tasks. Finally, we find that RAG datasets, such as HotpotQA, consistently achieve higher correlation with other real-world tasks. We also see from Figure 4 that HotpotQA enjoys an almost linear relationship with the QA dataset. Similar to synthetic tasks, RAG tasks are easy to control and also test models' recall abilities; yet since all the passages are retrieved and relevant to the query, RAG contexts are more distracting and hence harder to saturate.

3.2 Diverse LCLM applications call for diverse evaluation

In long-context language modeling, realistic tasks are often only used in isolated settings (Karpinska et al., 2024; Li et al., 2024c; Dubey et al., 2024), which limits the understanding of LCLMs in a broader context. In this work, we cross-examine model performance over the wide range of real tasks, and find that different categories do not always correlate with each other, as shown in Figure 5.

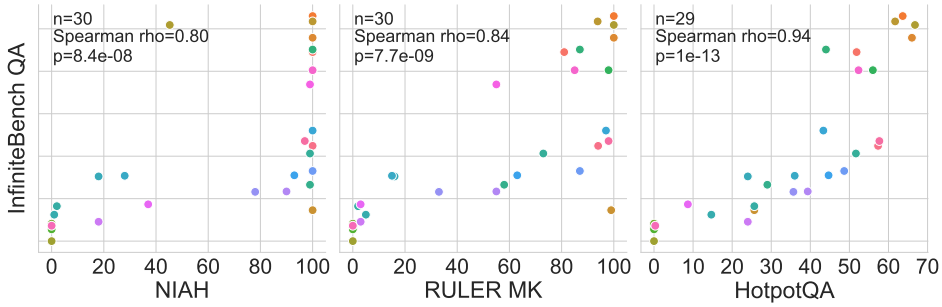

Figure 4: Distribution of instruction-tuned models' performance on ∞BENCH QA with respect to NIAH, RULER MK, and HotpotQA.

Naturally, RAG and passage re-ranking moderately correlate due to the shared retrieval component. However, the added complexity of generating citations in ALCE leads to lower correlation with other categories. As shown in Figure 9, generating correct answers and valid citations are not well correlated, which suggests that instruction following and recalling facts with long contexts are distinct capabilities.

Furthermore, some categories—summarization and in-context learning—do not correlate well with other categories at all. Intuitively, summarization tests for the model's ability to aggregate information across the whole input while ICL tests for the model's ability to learn new tasks from many examples. Such capabilities are orthogonal to recall facts in long contexts. Therefore, model developers should evaluate across these distinct axes to draw a more holistic picture of the model's capabilities. Additional analysis are in §E.2.

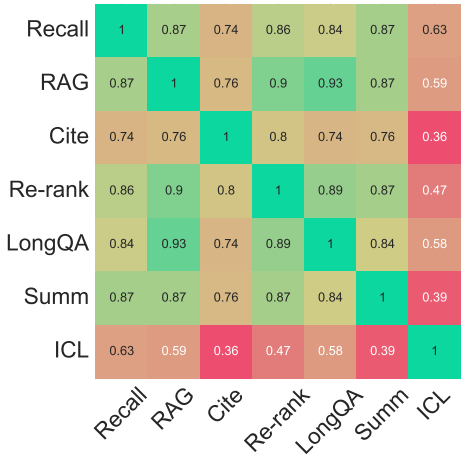

Figure 5: Spearman rank correlation between different categories at L =128k.

3.3 MODEL PERFORMANCE ACROSS TASKS AND LENGTHS

We show the performance of instruction-tuned models on HELMET at five different lengths in Figure 6, and the full results are illustrated in Figure 10. We analyze the model performance across two critical dimensions of long-context language modeling: task complexity and input length.

Open-source models lag behind closed-source models on complex tasks. First, we consider the performance of frontier LCLMs at the longest input length of 128k tokens. We find that the closed-source models—notably GPT-4o-08 and Gemini-1.5-Pro—stand out as the strongest LCLMs. Other than ICL, the closed-source models outperform the open-source models on all tasks. The gap is relatively small on synthetic recall and LongQA, where the task is to retrieve information from the context. There is a stark contrast in the generation with citations and re-ranking performance, where the closed-source models are 30 to 40 absolute points better than the best open-source models.

Performance degradation with longer inputs is category-dependent. Most frontier models largely retain performance on recall and RAG with longer inputs; however, even the best models significantly degrade with more contexts on tasks like re-ranking and generation with citations. As illustrated in Figure 7, with increased task complexity from left to right, performance degradation at longer lengths is more pronounced. On generation with citations, open-source models completely collapse at 128k while GPT-4o stays roughly the same. This underscores the importance of evaluating on more complex and realistic long-context applications.

No clear winner across all categories. As we observe from the previous sections, the different categories are not always correlated with each other. This is evident in the different winners across

	Recall					RAG					Cite					Re-rank				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0	43.8	45.2	28.8	3.6	3.1	76.4	72.3	63.9	37.8	16.8
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0	43.7	44.2	44.1	44.1	40.6	74.4	74.3	67.2	56.9	46.8
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8	45.8	47.1	46.4	45.7	45.3	75.6	73.1	67.4	59.5	47.9
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7	36.1	33.7	31.3	28.0	24.5	68.9	65.2	56.4	40.5	30.5
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1	36.7	32.9	30.5	26.4	12.5	76.3	46.1	36.0	14.5	9.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6	48.4	46.6	43.0	36.7	29.0	75.1	73.9	68.9	59.3	50.7
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9	47.1	43.0	44.7	45.1	42.5	75.8	73.2	71.7	65.9	58.6
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0	35.4	26.9	12.6	12.8	3.4	58.7	45.9	42.0	31.9	15.0
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8	44.5	42.1	39.5	30.9	7.6	73.3	69.7	58.4	40.0	19.4
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9	33.7	8.6	3.7	1.3	0.5	56.8	46.0	13.1	0.0	0.0
MegaBeam-Mistral	93.9	90.0	81.6	83.6	76.0	62.6	62.6	61.8	57.4	55.2	22.3	13.8	9.7	4.5	4.0	49.9	36.2	34.2	21.7	15.9
Phi-3-mini-128k	90.3	84.9	81.1	80.1	42.3	61.2	60.6	57.9	55.7	46.0	22.8	16.9	9.3	2.7	0.8	44.1	28.7	25.6	16.6	5.8
Phi-3-small-128k	91.0	89.3	73.5	66.7	59.0	66.5	65.8	62.5	61.3	58.1	18.9	15.9	8.9	4.6	2.9	38.3	32.1	28.1	17.2	6.5
Phi-3-med-128k	76.1	70.6	62.5	51.8	14.4	65.3	64.5	62.7	56.9	45.2	39.1	27.1	10.2	5.8	3.3	43.2	33.3	25.5	11.9	5.8
Phi-3.5-mini	95.8	90.7	83.1	77.2	40.7	59.8	57.9	55.6	51.0	41.4	22.1	17.2	7.1	2.0	2.5	42.4	29.6	23.2	18.0	9.1
Jamba-1.5-Mini	87.3	85.6	85.0	79.7	76.8	66.2	65.0	64.0	63.4	56.6	15.4	10.0	5.7	3.1	2.5	53.5	43.0	35.6	23.2	14.6
ProLong	96.6	95.8	93.0	92.9	85.5	68.8	69.3	66.6	66.5	64.8	33.7	24.0	11.0	2.3	1.2	53.8	43.9	39.3	33.3	25.0

	LongQA					Summ					ICL					Avg.				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	47.8	48.1	49.5	47.6	45.7	28.0	29.6	33.6	36.2	35.8	78.6	81.6	70.2	55.0	40.1	64.2	63.4	58.6	48.1	39.9
GPT-4o-05	45.2	46.9	51.8	56.0	61.4	29.5	34.9	40.7	42.1	45.1	77.8	79.2	76.2	65.4	47.7	62.8	63.7	63.3	60.5	56.3
GPT-4o-08	40.4	45.7	50.1	53.3	57.4	29.0	35.7	40.9	41.8	43.6	83.2	85.6	88.8	86.6	84.6	63.9	65.8	66.3	65.0	63.8
GPT-4o-mini	37.8	39.6	46.0	51.0	52.2	28.3	33.8	36.7	39.2	40.8	73.2	77.8	80.0	79.6	80.0	59.5	60.1	59.9	57.0	54.1
Claude-3.5-sonnet	29.8	25.8	32.5	18.2	19.9	27.4	33.0	40.1	37.2	40.7	86.0	87.8	88.4	89.6	59.8	59.5	53.7	53.5	45.8	39.5
Gemini-1.5-Flash	33.1	40.2	44.9	51.9	57.9	25.2	30.0	33.2	36.2	39.9	70.4	67.0	53.6	39.5	21.9	59.6	60.2	58.1	55.0	50.7
Gemini-1.5-Pro	33.9	40.2	48.9	53.4	61.7	30.1	32.6	39.3	45.5	45.0	77.3	79.6	81.0	79.5	76.3	59.8	60.7	63.5	64.1	62.7
Llama-3.1-8B	24.6	32.3	39.4	43.3	45.6	23.2	25.7	29.2	30.2	31.5	69.8	74.6	77.0	78.6	83.4	54.3	53.3	51.7	51.4	47.0
Llama-3.1-70B	31.5	38.9	45.8	55.4	58.4	27.7	31.7	35.5	35.8	35.7	71.6	74.2	77.0	79.4	83.6	60.2	61.2	60.8	57.0	49.3
Mistral-Nemo	32.2	32.0	26.8	24.6	24.2	26.0	23.2	25.3	21.5	20.1	66.8	75.4	80.0	81.6	84.4	53.9	47.4	36.9	28.3	25.9
MegaBeam-Mistral	23.1	30.6	34.0	36.3	34.7	21.6	24.3	27.9	30.4	29.4	72.0	77.2	78.4	82.8	85.0	49.3	47.8	46.8	45.2	42.9
Phi-3-mini-128k	24.4	31.7	31.6	34.1	27.7	20.8	24.2	26.9	28.1	28.8	61.6	71.4	73.2	77.0	80.0	46.4	45.5	43.7	42.0	33.1
Phi-3-small-128k	22.8	29.0	33.4	40.3	36.4	18.1	20.7	25.4	25.3	26.6	67.6	73.2	79.2	82.6	84.2	46.2	46.6	44.4	42.6	39.1
Phi-3-med-128k	22.6	21.5	20.4	20.4	27.1	23.1	24.5	25.4	31.4	28.6	58.8	61.2	70.6	72.4	72.0	46.9	43.2	39.6	35.8	28.1
Phi-3.5-mini	25.2	28.3	29.9	31.2	27.8	21.2	23.9	24.6	29.4	27.9	61.2	69.0	74.2	77.8	78.4	46.8	45.2	42.5	41.0	32.5
Jamba-1.5-Mini	34.8	41.1	46.2	52.2	52.7	18.0	19.0	19.1	19.6	19.3	77.6	82.0	85.6	88.4	91.2	50.4	49.4	48.7	47.1	44.8
ProLong	27.2	34.1	36.2	42.3	43.5	21.1	25.9	26.1	27.2	29.0	66.4	72.2	78.0	81.4	84.0	52.5	52.2	50.0	49.4	47.6

Figure 6: Results of HELMET. All models are instruction-tuned and have a claimed context window of 128k tokens or more.

	NIAH					RAG					Re-rank					Cite				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o-08	100.0	100.0	100.0	100.0	100.0	73.4	73.8	72.4	71.1	70.8	75.6	73.1	67.4	59.5	47.9	45.8	47.1	46.4	45.7	45.3
Gemini-1.5-Pro	97.9	97.0	98.0	54.6	45.3	73.0	72.9	71.6	71.9	70.9	75.8	73.2	71.7	65.9	58.6	47.1	43.0	44.7	45.1	42.5
Llama-3.1-8B-Inst	100.0	100.0	100.0	100.0	100.0	69.1	67.9	64.8	64.6	59.0	58.7	45.9	42.0	31.9	15.0	35.4	26.9	12.6	12.8	3.4
Llama-3.1-70B-Inst	100.0	100.0	100.0	100.0	100.0	73.0	72.2	71.5	70.3	55.8	73.3	69.7	58.4	40.0	19.4	44.5	42.1	39.5	30.9	7.6
Jamba-1.5-Mini	100.0	100.0	100.0	100.0	100.0	66.2	65.0	64.0	63.4	56.6	53.5	43.0	35.6	23.2	14.6	15.4	10.0	5.7	3.1	2.5
Qwen2-7B-Inst	100.0	100.0	100.0	97.0	18.0	62.6	57.8	56.6	45.2	28.3	44.5	30.3	13.3	0.6	0.0	18.2	5.0	2.3	2.6	2.1
Qwen2-57B-Inst	100.0	100.0	100.0	90.0	37.0	65.3	64.6	63.5	55.3	11.6	44.3	37.7	17.1	5.2	0.0	35.5	11.0	4.0	3.4	0.9

Figure 7: Results of select instruct models on various lengths and increasing complexity of tasks. Notably, Qwen2 relies on position extrapolation, while other open-models are trained at or greater than 128k context window.

different categories: for instance, GPT-4o is stronger on recall and generation with citations, while Gemini is stronger on passage re-ranking and long-document QA. Furthermore, many open-source models outperform the closed-source models on ICL, potentially because heavy instruction tuning hurts ICL, and we provide qualitative examples in Table 19. Thus, it is critical to evaluate models across multiple axes. We also present additional analysis, such as the performance of positional extrapolation methods (§E.3), the lost-in-the-middle phenomenon (§E.4), and the performance of Claude (§E.6).

4 Related Works

Long-context language models. Frontier models such as GPT-4 (OpenAI, 2023), Gemini, and Claude claim to have expanded their context window beyond 100k tokens. In the open-source

community, there are also efforts to train models with longer input lengths (Dubey et al., 2024; Fu et al., 2024; AI et al., 2024), explore position extrapolation techniques (Peng et al., 2024; Chen et al., 2023; Ding et al., 2024), and experiment with efficient architectures (Gu & Dao, 2023; Dao & Gu, 2024; Bertsch et al., 2023; Beltagy et al., 2020; Yen et al., 2024; Lieber et al., 2024, *inter alia*).

Synthetic tasks. Synthetic tasks are often used to evaluate LCLMs since they can be procedurally generated, thereby enabling arbitrarily long input lengths and controlled "needle" placement (Tay et al., 2021; Liu et al., 2023). In particular, Needle-in-a-Haystack (NIAH; Kamradt, 2024) inserts a "needle" at specific depths of a long essay (i.e., the haystack) and asks the model to recall the fact. Recent works have expanded upon it and designed new procedures to test different aspects of LCLMs (Hsieh et al., 2024; Li et al., 2024b; Levy et al., 2024; Arora et al., 2023; Laban et al., 2024; Goldman et al., 2024). However, they do not study how results on synthetic tasks transfer to real applications. In contrast, we evaluate both synthetic and downstream tasks and investigate how they correlate with each other.

Long-context benchmarks. As discussed in the main paper, existing benchmarks either are limit to relatively short context lengths (Shaham et al., 2022; 2023; Li et al., 2024a; An et al., 2024; Bai et al., 2024; Dong et al., 2024) or lack rigorous evaluation methods (Zhang et al., 2024b; Yuan et al., 2024). Many works instead focus on specific domains, such as question answering (Wang et al., 2024b; Karpinska et al., 2024; Wang et al., 2024a), in-context learning (Li et al., 2024c; Bertsch et al., 2024; Xu et al., 2024; Anil et al., 2024; Agarwal et al., 2024), summarization (Chang et al., 2024; Kim et al., 2024; Shen et al., 2022), or RAG (Lee et al., 2024). In this work, we construct a comprehensive benchmark that tests across diverse downstream tasks at long input lengths and also present a unified comparison and analysis across 51 LCLMs.

5 Conclusion

In this work, we first identify the shortcomings of long-context evaluation settings and existing benchmarks—over-reliance of synthetic tasks, low coverage of realistic applications, and unreliable metrics among others. We seek to address these issues by constructing HELMET, an application-centric benchmark with diverse domains and reliable evaluation settings. We then present a comprehensive evaluation of 51 frontier LCLMs across multiple dimensions, including different tasks, input lengths, and model types. Our analysis shows that synthetic tasks are poor predictors of downstream performance and different categories exhibit distinct capabilities and trends. Thus, it is critical to evaluate models on a diverse set of real-world tasks. Furthermore, open-source models still lag behind closed-source models on complex tasks at long lengths. Finally, we hope that our benchmark and subsequent comprehensive evaluation can serve as a valuable resource for future research in long-context language modeling.

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. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning, 2024. URL https://arxiv.org/abs/2404.11018.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.776.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Anthropic, April*, 2024.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=hSyW5go0v8.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. Reliable, adaptable, and attributable language models with retrieval, 2024b. URL https://arxiv.org/abs/2403.03187.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.172.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document Transformer. *arXiv:2004.05150*, 2020. URL https://arxiv.org/abs/2004.05150.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-range transformers with unlimited length input. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=lJWUJWLCJo.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. In *First Workshop on Long-Context Foundation Models @ ICML 2024*, 2024. URL https://openreview.net/forum?id=4KAmc7vUbq.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022. URL https://arxiv.org/pdf/2212.08037.pdf.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah (eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Ttk3RzDeu.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171.

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

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2019 deep learning track, 2020. URL https://arxiv.org/abs/2003.07820.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

Daniel Deutsch, Rotem Dror, and Dan Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6038–6052, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.442. URL https://aclanthology.org/2022.naacl-main.442.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2086–2099, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.188.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony

Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun

Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). 2024.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp, 2024. URL https://arxiv.org/abs/2407.00402.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News Summarization and Evaluation in the Era of GPT-3, 2023. URL https://arxiv.org/abs/2209.12356.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv preprint arXiv:2303.01432*, 2023. URL https://arxiv.org/abs/2303.01432.

Garrett Kamradt. Needle in a haystack - pressure testing llms, 2024. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

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

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. Summary of a haystack: A challenge to long-context llms and rag systems, 2024. URL https://arxiv.org/abs/2407.01370.

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. An evaluation dataset for intent classification and out-of-scope prediction. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131.

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

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Association for Computational Linguistics (ACL)*, 2024.

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

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window?, 2024b. URL https://arxiv.org/abs/2407.11963.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060, 2024c. URL https://doi.org/10.48550/arXiv.2404.02060.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model, 2024. URL https://arxiv.org/abs/2403.19887.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023. URL https://api.semanticscholar.org/CorpusID:259360665.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. *ArXiv*, abs/1903.05566, 2019. URL https://api.semanticscholar.org/CorpusID:76660838.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Association for Computational Linguistics (ACL)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9844–9855, 2022. URL https://aclanthology.org/2022.emnlp-main.669.

OpenAI. GPT-4 Technical Report, 2023.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. URL https://aclanthology.org/D16-1264/.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6383–6402, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.352. URL https://aclanthology.org/2023.acl-long.352.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Samuel Joseph Amouyal Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. QAMPARI: An Open-domain Question Answering Benchmark for Questions with Many Answers from Multiple Paragraphs. *arXiv preprint arXiv:2205.12665*, 2022. URL https://arxiv.org/abs/2205.12665.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 12007–12021, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.823.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.536. URL https://aclanthology.org/2023.findings-emnlp.536.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=z1d8fUiS8Cr.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, December

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

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 Kepa, 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. Gemini: A family of highly capable multimodal models, 2024a. URL https://arxiv.org/abs/2312.11805.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng,

Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, Hyun-Jeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis,

Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C.

Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang

Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024b. URL https://arxiv.org/abs/2403.05530.

Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Opher Lieber, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shaked Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Yehoshua Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba-1.5: Hybrid transformer-mamba models at scale, 2024c. URL https://arxiv.org/abs/2408.12570.

Together. Llama-2-7b-32k, July 2023. URL https://huggingface.co/togethercomputer/LLaMA-2-7B-32K.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3712–3724, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.205. URL https://aclanthology.org/2024.naacl-long.205.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. Novelqa: Benchmarking question answering on documents exceeding 200k tokens, 2024b. URL https://arxiv.org/abs/2403.12766.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023. URL https://arxiv.org/abs/2303.03846.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models, 2023.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. Stress-testing long-context language models with lifelong ICL and task haystack. In *First Workshop on Long-Context Foundation Models @ ICML 2024*, 2024. URL https://openreview.net/forum?id=5ltgEKXnZt.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2588–2610, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.142. URL https://aclanthology.org/2024.acl-long.142.

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

Shiyue Zhang and Mohit Bansal. Finding a balanced degree of automation for summary evaluation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6617–6632, 2021. doi: 10.18653/v1/2021.emnlp-main.531. URL https://aclanthology.org/2021.emnlp-main.531.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, 2024a. URL https://arxiv.org/abs/2407.19669.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞bench: Extending long context evaluation beyond 100k tokens, 2024b. URL https://arxiv.org/abs/2402.13718.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on*

Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

A COMPARISON WITH OTHER BENCHMARKS

Due to the lack of large-scale human studies on long-context language models, it may be challenging to discern the true rankings of LCLMs. Thus, we rely on previous works and intuition to guide our expectations. Specifically, the Gemini 1.5 report observes a trend of large models performing better at long contexts—the larger models beating the smaller models both qualitatively and quantatively across the Gemini, GPT, and Claude model families (Section 5.2, Tables 4 and 5 Team et al., 2024b). Although the evaluation suite is not publicly available, the results align with our intuition, which is commonly observed in other uses of language models. Since we do not have access to a ground truth ranking of models and acquiring one via human evaluation is nearly impossible due to the associated costs, relying on such a commonly-accepted intuition is reasonable. In this section, we further examine the improvement of HELMET over existing benchmarks through direct comparison of model rankings and performance as well as careful ablation studies.

A.1 RESULTS

We build upon previous datasets and update evaluation settings to reflect the capabilities of the model better. Specifically, we evaluate models on ∞BENCH, another benchmark that evaluates long-context at 128k tokens. We show the results reproduced from the original authors' code in Table 6. The numbers for Figure 1 are shown in Table 5.

We take a closer look at a set of open-sourced models' ranking and performance on HELMET and ∞BENCH in Table 9. In addition to the previously noted performance discrepancy between Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct, we observe that all Llama-3.2 models degenerate in performance on ∞BENCH at input length 128k. However, on HELMET, we find that the Llama-3.2 models, especially Llama-3.2-3B-Instruct, rank well against other open-source models. Upon qualitative examination, we find that the Llama 3.2 models are able to produce coherent and useful generation at long contexts with better prompting strategies from HELMET, such as adding in-context learning examples.

We further ablate the impact of few-shot demonstrations on model performance in Table 8, and find that the 2-shot demonstrations significantly improve the performance of base models, which is a more realistic evaluation of how the models are used in practice. The improvement in instructions and prompting strategies are illustrated in 7, where we show that the smaller Llama-3.2 models are able to compete against some of the larger models when the instructions are improved. Thus, HELMET provides a better and more realistic reflection of how these models would be used in practice over previous benchmarks.

Table 5: Results with 128k input length across different benchmarks.

Model	Claimed Length	NIAH	RULER	∞BENCH	HELMET
GPT-4o-mini	128000	**100.0**	80.8	51.9	54.1
GPT-4o-08	128000	**100.0**	**93.3**	57.1	**63.8**
Gemini-1.5-Flash	1048576	**100.0**	86.6	50.8	50.7
Gemini-1.5-Pro	2097152	45.3	65.3	**60.3**	62.7
Llama-3.1-8B-Inst	131072	**100.0**	81.3	44.1	47.0
Llama-3.1-70B-Inst	131072	**100.0**	75.8	39.7	49.3

A.2 DISCUSSIONS

Many existing long-context language modeling benchmarks are studied in isolated settings, such as synthetic tasks (Hsieh et al., 2024; Li et al., 2024b; Levy et al., 2024; Laban et al., 2024), in-context learning (Anil et al., 2024; Bertsch et al., 2024; Li et al., 2024c; Agarwal et al., 2024), summarization (Chang et al., 2024; Kim et al., 2024), and retrieval-augmented settings (Lee et al., 2024). However, these works lack a unified evaluation across diverse downstream tasks.

There are also benchmarks that sought to unify different datasets, such as ∞BENCH (Zhang et al., 2024b), LongBench (Bai et al., 2024), L-Eval (An et al., 2024), LV-Eval (Yuan et al., 2024), and

Table 6: Results on ∞BENCH(Zhang et al., 2024b). We report numbers from running the original authors repo: https://github.com/OpenBMB/InfiniteBench. The original code did not support evaluation of the Gemini and Llama 3 models at the time of access, so we evaluate them by following the script and template for GPT-4 and open-source models, respectively. We exclude coding tasks from the evaluation suite since it is out of the scope for general long-context language modeling tasks.

Model	MC	QA	Sum	Diag	Calc	Find	Number	PassKey	KV	Avg.
GPT4	75.9	23.7	10.7	18.5	0.0	47.1	100.0	100.0	58.8	44.9
GPT-4o-05	88.2	37.9	23.7	28.5	0.0	58.6	100.0	100.0	94.2	55.4
GPT-4o-08	86.5	26.0	21.5	51.0	0.0	58.9	100.0	100.0	99.8	57.1
GPT-4o-mini	78.2	19.1	24.8	21.5	0.0	69.7	100.0	100.0	80.4	51.9
Gemini-1.5-Flash	76.0	42.1	30.0	55.8	0.0	47.4	100.0	100.0	31.4	50.8
Gemini-1.5-Pro	77.5	27.7	29.0	97.5	0.0	58.0	100.0	100.0	70.4	60.3
Llama-3.1-8B	56.8	8.8	14.3	0.5	0.0	22.0	99.7	100.0	18.8	33.0
Llama-3.1-8B-Inst	67.2	15.5	26.7	23.0	0.0	33.1	99.5	100.0	55.0	44.1
Llama-3.1-70B	66.4	9.2	17.5	8.5	0.0	32.3	100.0	100.0	13.4	35.1
Llama-3.1-70B-Inst	75.5	23.3	31.3	18.0	0.0	43.1	99.7	100.0	2.6	39.7
Llama-3.2-1B	2.2	1.4	8.6	4.5	0.0	0.0	1.5	0.0	0.0	2.0
Llama-3.2-1B-Instruct	3.5	1.4	12.5	5.5	0.0	0.0	0.0	0.0	0.0	2.4
Llama-3.2-3B	1.3	1.2	7.6	4.5	0.0	0.0	1.7	0.0	0.0	1.9
Llama-3.2-3B-Instruct	2.2	1.6	13.3	4.5	0.0	0.0	1.7	1.7	0.0	2.8

Table 7: We compare the results on ∞BENCH multiple-choice (MC) and question answering (QA) tasks between the original authors and HELMET. In our implementation, we leverage refined prompts and carefully crafted parsing scripts to ensure robust and fair evaluation. For example, we find that Llama-3.2 models are actually much better long-context language models than ∞BENCH would suggest.

		Original		**HELMET**	
Model	**Claimed Length**	**MC**	**QA**	**MC**	**QA**
GPT-4o-mini	128000	88.2	37.9	71.0	44.5
GPT-4o-05	128000	86.5	26.0	76.0	53.0
GPT-4o-08	128000	78.2	19.1	73.0	47.9
Gemini-1.5-Flash	1048576	76.0	42.1	79.1	51.7
Gemini-1.5-Pro	2097152	77.5	27.7	83.3	50.9
Llama-3.1-8B	131072	56.8	8.8	54.0	38.8
Llama-3.1-8B-Inst	131072	67.2	15.5	49.0	40.3
Llama-3.2-1B	131072	2.2	1.4	21.0	17.6
Llama-3.2-1B-Inst	131072	3.5	1.4	16.0	13.3
Llama-3.2-3B	131072	1.3	1.2	36.0	31.0
Llama-3.2-3B-Inst	131072	2.2	1.6	42.0	20.6

ZeroSCROLLS (Shaham et al., 2023). Many of these benchmarks are still limited by the context length, evaluation metrics, or both. Most similar to our work, ∞BENCHZhang et al. (2024b) also evaluates models at context lengths at 128k tokens. However, their evaluation settings are limited to a few domains—synthetic, QA, and summarization. Although they also evaluate coding tasks, the domain is limited to code-specialized models outside of the realm of general-purposed language models. Furthermore, their summarization evaluations lack robust evaluation and still rely on ROUGE scores. Similarly, LV-Eval (Yuan et al., 2024) evaluates long-context models across different lengths but is limited to the QA tasks. Furthermore, we are the first to evaluate existing LCLMs comprehensively—we evaluate over 50 models of different sizes and architectures, which enables previously unavailable insights into the correlation across tasks and the landscape of current models.

To summarize, HELMET improves upon existing benchmarks by providing a comprehensive evaluation across diverse tasks and carefully designed evalution metrics and prompting strategies. Consequently, HELMET provides a more accurate reflection of the models' capabilities in practice.

Table 8: We evaluate the performance of models on a subset of HELMET tasks with 0-shot and 2-shot demonstrations at 128k input length to understand the impact of ICL on model performance, averaged across three random seeds. The standard deviation across three runs are shown in the subscript. We observe that the performance is generally higher for 2-shot demonstrations compared to 0-shot demonstrations. Crucially, the 2-shot examples enable base models to achieve higher results that reflect the model's long-context capabilities in realistic settings, such as for MSMARCO and JSON KV.

	JSON KV		NQ		MSMARCO		∞BENCH MC		∞BENCH QA	
Model	**0 shot**	**2 shot**	**0 shot**	**2 shot**	**0 shot**	**2 shot**	**0 shot**	**2 shot**	**0 shot**	**2 shot**
GPT-4o-mini	$92.3_{5.5}$	$93.7_{2.1}$	$57.8_{1.1}$	$60.4_{2.6}$	$24.8_{0.7}$	$31.4_{1.1}$	$72.0_{2.6}$	$69.3_{4.7}$	$28.8_{1.9}$	$47.1_{4.4}$
GPT-4o-05	$99.3_{0.6}$	$36.7_{7.1}$	$59.4_{4.0}$	$63.0_{2.1}$	$42.7_{0.9}$	$46.8_{1.3}$	$73.7_{0.6}$	$76.3_{0.6}$	$49.1_{1.9}$	$51.3_{3.4}$
GPT-4o-08	$100.0_{0.0}$	$100.0_{0.0}$	$58.5_{2.0}$	$61.9_{4.5}$	$46.6_{0.3}$	$48.7_{1.0}$	$75.3_{2.5}$	$72.7_{0.6}$	$29.8_{0.9}$	$48.6_{1.5}$
Gemini-1.5-Flash	$98.3_{0.6}$	$99.0_{1.0}$	$53.7_{5.9}$	$51.5_{4.5}$	$44.9_{3.0}$	$49.9_{0.9}$	$73.9_{2.3}$	$77.1_{3.2}$	$48.2_{4.7}$	$49.9_{2.6}$
Gemini-1.5-Pro	$97.7_{0.6}$	$93.7_{2.3}$	$62.2_{2.6}$	$59.2_{2.7}$	$57.1_{0.9}$	$58.4_{0.3}$	$83.6_{2.8}$	$83.3_{0.7}$	$46.4_{1.8}$	$47.2_{3.2}$
Llama-3.1-8B$^{\flat}$	$77.3_{2.1}$	$98.0_{1.0}$	$45.7_{2.1}$	$44.9_{3.9}$	$0.1_{0.0}$	$7.5_{0.5}$	$55.3_{5.5}$	$53.3_{8.0}$	$36.1_{2.0}$	$36.6_{2.1}$
Llama-3.1-8B-Inst	$96.0_{1.0}$	$95.7_{0.6}$	$48.4_{1.3}$	$48.9_{2.6}$	$5.4_{0.0}$	$13.8_{1.1}$	$49.7_{2.1}$	$52.0_{2.6}$	$29.1_{2.0}$	$37.9_{2.6}$
Llama-3.2-1B$^{\flat}$	$34.0_{3.5}$	$34.3_{2.1}$	$25.1_{0.7}$	$23.2_{2.8}$	$0.3_{0.0}$	$4.8_{0.5}$	$23.0_{2.6}$	$22.7_{1.5}$	$18.9_{2.5}$	$18.9_{3.3}$
Llama-3.2-1B-Inst	$6.3_{1.5}$	$9.3_{3.2}$	$23.3_{4.4}$	$22.9_{1.4}$	$0.0_{0.0}$	$0.5_{0.3}$	$16.0_{2.6}$	$16.3_{2.5}$	$12.6_{0.4}$	$12.7_{0.5}$
Llama-3.2-3B$^{\flat}$	$30.3_{2.1}$	$54.0_{8.9}$	$40.2_{2.0}$	$42.3_{4.1}$	$0.1_{0.0}$	$6.0_{1.3}$	$43.0_{1.0}$	$40.7_{7.2}$	$31.3_{1.1}$	$31.6_{2.2}$
Llama-3.2-3B-Inst	$31.3_{2.3}$	$36.7_{2.5}$	$44.2_{2.5}$	$42.8_{3.8}$	$0.2_{0.0}$	$0.9_{0.2}$	$36.0_{2.0}$	$40.0_{3.5}$	$17.8_{1.2}$	$19.7_{0.8}$

Table 9: Comparison of model rankings on HELMET and ∞BENCH.

Model	**HELMET**	**Model**	**InfBench**
Llama-3.1-70B-Inst	49.3	Llama-3.1-8B-Inst	46.7
Llama-3.1-8B-Inst	47.0	Llama-3.1-70B-Inst	43.7
Llama-3.1-70B	41.3	Yi-34B-200k	43.1
Yi-34B-200k	38.3	Llama-3.1-70B	38.6
Llama-3.2-3B-Inst	36.9	Yi-9B-200k	37.6
Llama-3.1-8B	35.6	Llama-3.1-8B	35.7
Yi-9B-200k	33.0	Yi-6B-200k	32.0
Llama-3.2-3B	31.9	Llama-3.2-3B-Inst	2.8
Yi-6B-200k	26.3	Llama-3.2-1B-Inst	2.6
Llama-3.2-1B-Inst	24.6	Llama-3.2-1B	2.0
Llama-3.2-1B	21.2	Llama-3.2-3B	1.8

B DATASETS

B.1 RETRIEVAL-AUGMENTED GENERATION

Natural Questions (NQ; Kwiatkowski et al., 2019) is a large-scale dataset for open-domain questions answering with real user queries. TriviaQA (TQA; Joshi et al., 2017) is a dataset of trivia questions with answers. HotpotQA (Yang et al., 2018) is a dataset of questions that require reasoning over multiple passages. We obtain these datasets from KILT (Petroni et al., 2021), where each query is annotated with a set of gold passages that the answer is located in.

PopQA (Mallen et al., 2023) is a dataset with questions about long-tail entities. We filter the original dataset to only include the subject entities with a popularity of less than 1000 to ensure that models are unlikely to have memorized the fact during training. We do not have access to the gold passages for PopQA, so we use the ground truth answer to classify each retrieved passage into positive or negative by simply checking if the gold answer is in the passage.

It is important to note that we fill the context with *hard negative passages* that are retrieved from the same corpus as the positive passages with a real retriever, making it much harder for the model to identify the answer than randomly sampled passages filled the context. We believe that this is a more realistic setting for retrieval-augmented generation tasks, as models are often required to distinguish between relevant and irrelevant information in real-world applications.

B.2 GENERATION WITH CITATIONS

Generation with citations is an important and realistic task for making language models more trustworthy and verifiable (Bohnet et al., 2022; Gao et al., 2023; Asai et al., 2024b). Thus, we use ALCE to evaluate the models' ability to generate text with correct citations. Following the original work, we use Wikipedia as the retrieval corpus and GTR (Ni et al., 2022) as the retriever. We do not evaluate the fluency of the generated text using MAUVE, as we find that models are often able to generate fluent text, and when it does not, the other scores are also near zero. Thus, we only evaluate correctness and citation quality and report their average.

B.3 PASSAGE RE-RANKING

We test the models' ability to process and re-rank passages in the context (Sun et al., 2023) with the MS MARCO passage ranking dataset (Bajaj et al., 2018). We use the datasets with annotations from the TREC Passage Re-ranking challenge (Craswell et al., 2020). Each instance in the dataset is consisted of a query and a list of passages with relevance scores. The queries are sourced from Bing user queries and the passages are retrieved by BM25 from Internet web pages. The relevance labels are either perfect, highly relevant, or not relevant. We first calculate the number of passages k from the input L. We balance the number of labels between the passages and shuffle them to prevent positional biases. We then shuffle the k passages into three different permutations to construct the inputs. The performance is evaluated with NDCG@10. Although we may evaluate NDCG@k, we find it expensive to generate hundreds of passage IDs during inference, and thus we only evaluate NDCG@10.

B.4 IN-CONTEXT LEARNING

For in-context learning, we use a label mapping to force the model to leverage the in-context learning examples as opposed to relying on their learned priors for classification tasks. To this end, each label is first randomly mapped into an integer $l \in \{0, 1, \ldots, n-1\}$ where n is the number of unique labels in the dataset, following previous works (Pan et al., 2023). Then, we simply replace all label texts with their corresponding integers in the dataset. Furthermore, we put examples into round of demonstrations similar to (Li et al., 2024c), where each round contains exactly one example from each label and the order is shuffled. We then construct the input by concatenating rounds of demonstrations together until we meet a certain input length L.

B.5 SYNTHETIC TASKS

For the RULER tasks, we generate the data using the original authors' scripts (Hsieh et al., 2024). To standardize the amount of texts input to all models, we use the Llama-2 tokenizer (Touvron et al., 2023) when generating the data. The RULER suite of tasks are consisted of these tasks:

- Three variants of NIAH Single: NIAH (variant 2, which is the most similar to the original NIAH), NIAH Single Repeat (variant 1, where the context is repeated phrase instead of Paul Graham essays), and NIAH Single UUID (variant 3, same as variant 2 but with UUIDs as the value to retrieve)
- Three variants of NIAH MultiKey (MK): NIAH MK Essay (variant 1, where there are 3 irrelevant keys and 1 gold key to retrieve and the context is Paul Graham Essays), NIAH MK Needle (variant 2, where the context are all needles), and NIAH MK UUID (variant 3, same as variant 2 but with UUIDs as the value to retrieve for every needle)
- NIAH MultiValue (MV): There is one key but is associated with four different numbers that the model needs to retrieve. The context is irrelevant essays.
- NIAH MultiQuery (MQ): There are four keys, each of which is associated with a different value, and the model needs to retrieve the correct value for each key. The context is irrelevant essays.
- Variable Tracking (VT): the model needs to track the value of a variable through a series of operations.

- Common word extraction (CWE) and frequent word extraction (FWE): The model needs to extract the words that occurs the most often.
- Question answering: SQuAD (variant 1; Rajpurkar et al., 2016) and HotpotQA (variant 2; Yang et al., 2018). The key difference between our RAG tasks and these QA datasets is that our context is filled with *retrieved* passages from the corpus instead of a random sample of the corpus. Thus, this makes the task more realistic and challenging.

For more details, please refer to the original paper (Hsieh et al., 2024).

For the JSON KV tasks, we randomly generate a JSON dictionary where each key and value is a randomly generated UUID. Then, we query the model with the key at six evenly spaced depths in the context, asking it to retrieve the corresponding value.

B.6 Model-based evaluation

Automatic evaluation metrics, such as ROUGE, are known to be unreliable and uncorrelated with human judgments (Goyal et al., 2023; Chang et al., 2024). However, existing long-context benchmarks still largely rely on these metrics (Bai et al., 2024; Zhang et al., 2024b). In this work, we seek to more reliably evaluate language models at long-contexts by leveraging LLMs as judges, inspired by previous works (Zheng et al., 2023). For all model-based evaluations, we use `GPT-4o-2024-05-13` as the judge.

Long document question answering. For NarrativeQA, we found that the models can often output answers that are correct but have little surface form or lexical overlaps with the ground truth answers in preliminary experiments, resulting in lower-than-expected performance. This is often due to the long lengths of the ground truth answers, which gives more possibilities of how to write it than a simple named entity that is often the case of other QA datasets.

Therefore, we employ an LLM to judge if the model output is fluent and semantically similar to the ground truth. Given the question, the ground truth answer, and the model-generated output, we ask the LLM to judge if (1) the model output is fluent and (2) the generated output is correct and relevant to the question with the ground truth as a reference. The detailed prompts, precise definitions, and instructions are found in Table 10.

In Table 14, we also find that model evaluation can be useful in catching extremely cases. For example, Claude scores low in terms of F1 due to its tendency to output extra, assistant-like text, which is penalized by the n-gram overlap metric, and appears to be worse than the much smaller Llama-3.2-3B-Inst model. However, the model-based evaluation is able to catch this issue, and Claude scores higher than Llama-3.2-3B-Inst, which users might have expected given the model sizes.

Summarization. At a high level, we ask the model to check for three properties: fluency, precision, and recall. Fluency can take on a value of either 0 (incoherent, incomplete, and/or repetitive) or 1 (fluent and coherent). Precision is the percentage of model-generated sentences that are supported by the gold summary (we use the long summary from Multi-LexSum here). Recall is the percentage of the key points that are supported by the model-generated output. We calculate the F1 score from precision and recall and multiply it with the fluency score for the final score.

We first ask the model to generate a list of key points or atomic claims from the gold summary, following previous works that show that LLMs can accurately decompose long texts (Kamoi et al., 2023; Gao et al., 2023). We manually checked 105 claims originating from 25 Multi-LexSum summaries and found that the claims were all factually correct. Although we found one out of 25 instances where the model missed a possible key point in the given summary, we found that GPT-4o is almost always reliable for the decomposition task. For Multi-LexSum, we use the short summary to obtain the key point as the annotation contains both a long and a short summary. These key points are then saved for Multi-LexSum and ∞Bench Sum.

We show the detailed prompts for the summarization tasks in Table 11, 12, and 13, which are modeled after previous works (Kamoi et al., 2023; Chang et al., 2024; Kim et al., 2024). These previous works have shown that LLMs can effectively judge model outputs (Zheng et al., 2023), but we conduct human analysis to further verify the evaluation metric.

From qualitative analysis, we found that the model is consistently able to distinguish between fluent and non-fluent outputs, where we agree with the model judgments 100% of the time for randomly sampled Multi-LexSum and ∞BENCH Sum outputs. We then check if the GPT-4o judgments for precision and recall agree with human judgments. To this end, we sample 10 generated summaries for Multi-LexSum and ∞BENCH each (from Gemini-1.5-Pro and Llama-3.1-70B-Inst) and check five key point evaluations for each summary. We follow a similar procedure where we check if the model output supports the key point. We observed Cohen's $\kappa = 0.76$ for ∞BENCH Sum and $\kappa = 0.72$ for Multi-LexSum, suggesting substantial agreement. For precision, we conduct a similar human evaluation, and found a $\kappa = 0.91$ for ∞BENCH Sum and $\kappa = 0.83$ for Multi-LexSum, suggesting near-perfect agreement.

Inspecting the disagreements, we find that that most of them arise from the partially supported cases. For instance, the key point may include specific details, such as the names of government departments or Court Justices' names, that are not explicitly mentioned in the generated summary, and the model judge is typically more lenient about the exclusion of these small details while humans are more strict. However, this is also subjective to the preference of the human.

Qualitatively, We find that model-based evaluation can catch cases where the model is overly repetitive and scores high ROUGE-L score as a result, such as Mistral-7B-Inst-v0.3 on ∞BENCH Sum. For instance, the Mistral model may generate an output consisted of the sentence "The author's object is to show that the study of grammar is necessary part of education" repeated for hundreds of times. This summary would receive an ROUGE-L score of 12.3 while the GPT-4o judge would penalize the model for being overly repetitive and incoherent and assign the output a final score of 0.0. Thus, our GPT-4o judge penalizes the model for being overly repetitive, and the final metric reflects this issue.

C MODELS

Please see Table 15 for the detailed information of the models used in this work.

D EXPERIMENTAL SETUP

As previously described, we are able to evaluate models across different input lengths. Thus, we evaluate all models at $L \in \{8192, 16384, 32768, 65536, 131072\}$. For the proprietary models, GPT, Gemini, and Claude, we rely on the provider's API. For all open-source models, we evaluate on a H100 GPUs with 80GB of memory. We use the HuggingFace framework (Wolf et al., 2020) to load and generate model outputs. We apply instruction-tuned models' chat template whenever applicable. We use FlashAttention2 (Dao, 2023) and BF16 for faster inference. Our compute is limited to 8 H100 GPUs; thus, we are not able to run some of the larger models, such as Command-R or Jamba-1.5-Large, at 128k tokens. Although we sample 100 test queries for each dataset, this actually results in 600 examples for JSON KV, NQ, PopQA, and TQA, and 300 examples for the MSMARCO and HotpotQA. This is due to the fact that we evaluate at either six different depths or three different permutations.

E ADDITIONAL RESULTS

E.1 CORRELATION BETWEEN SYNTHETIC AND DOWNSTREAM TASKS

We show the correlation between all synthetic and RAG datasets with other downstream tasks in Figure 8. We see that the synthetic tasks are generally less correlated with the downstream tasks, whereas the RAG datasets are more correlated with the downstream tasks. Furthermore, there appears to be a pattern between the complexity of the synthetic task and its correlation with other tasks—the more complex the task, the more correlated it is with the downstream tasks. Furthermore, noisier variants tend to be more reflective of real-world applications, as seen in the difference between RULER MK Essay, where the context is irrelevant essays, and RULER MK Needle, where the context is other distracting needles. Upon closer inspection, we note that Needle MK Needle/UUID, JSON KV, and NIAH MV generally have relatively high correlation with other realistic datasets, and are able to test the recall abilities of the model. Thus, we select these four datasets as

Table 10: Model-based evaluation prompt for long-document question answering.

```
Please act as an impartial judge and evaluate the quality of the
provided answer which attempts to answer the provided question based on
a provided context.  Although you are not given the context, you will be
given a set of correct answers that achieves full scores on all metrics,
and you need to assess the provided answers using the correct answers.

Below is your grading rubric:


Fluency:  - Score 0 (incoherent, repetitive, or incomplete):  Incoherent
sentences, repetitive sentences (even if not by exact words), incomplete
answers, or gibberish.  Note that even if the answer is coherent, if it
is repetitive or incomplete, it should be given a score of 0.  - Score 1
(coherent, non-repetitive answer):  Coherent, non-repetitive, fluent,
grammatically correct answers.


Correctness:  - Score 0 (Incorrect):  The answer does not agree with the
provided correct answers at all.  - Score 1 (partly correct):  Partly
agree with one of the provided correct answers (for example, the
question asks for a date and a person; the answer gets the date right
but the person wrong).  - Score 2 (correct but not fully relevant):
Fully agrees with one of the provided correct answers but mentions other
completely irrelevant information.  Note that extra details provided in
the answer, even if not mentioned in the correct answers, should NOT be
seen as irrelevant as long as they are relevant to the question to a
reasonable extend.  - Score 3 (correct and relevant):  Fully agrees with
one of the provided correct answers and only provides information
relevant to the question.  Note that if the answer is longer than the
correct answer, as long as everything in the answer is relevant to the
question, it should still be given score 3.  For example, if the correct
answer is ''the North Pole'' and the answer is ''They are headed for the
North Pole'', it should still be given a score of 3.


Now, read the following question, answer, and correct answers.  First
think step-by-step and provide your reasoning and assessment on the
answer.  Then output your score in the following json format:
{{''fluency'':  0, ''correctness'':  1}}.

Question:  {question}
Correct answers:  {correct_answers}
Answer:  {parsed_output}
```

part of the HELMET synthetic recall subset. We validate this selection by checking the correlation between this set and other downstream tasks, and found that it generally has higher correlation than using synthetic datasets individually. There may be more optimal methods for selecting synthetic datasets, and we leave this as future work.

E.2 Correlation between datasets

We plot the correlation between all HELMET datasets in Figure 9. In general, the datasets in each category are strongly correlated with each other. One exception to this observation is the lack of correlation between the ALCE datasets. This suggests that writing the correct citations is a different skill from answering questions with facts. We also observe some low corrlation between datasets for long-document QA and ICl, but this is likely due to the diversity within each categories.

Table 11: Model-based evaluation prompt for summarization fluency score.

```
Please act as an impartial judge and evaluate the fluency of the
provided text.  The text should be coherent, non-repetitive, fluent, and
grammatically correct.


Below is your grading rubric:  - Score 0 (incoherent, repetitive, or
incomplete):  Incoherent sentences, repetitive sentences (even if not by
exact words), incomplete answers, or gibberish.  Note that even if the
answer is coherent, if it is repetitive or incomplete, it should be
given a score of 0.  - Examples:  - Incomplete:  ''Summary:''  -
Incoherent:  ''Summary:  The plaintiff the the the the able the the the
the the the the the the the able the the the the the the '' - Repetitive:
''Summary:  The U.S. government brought a criminal case against four
defendants.  Summary:  The U.S. government brought a criminal case
against four defendants.  Summary:  The U.S. government brought a
criminal case against four defendants.  Summary:  The U.S. government
brought a criminal case against four defendants.''


- Score 1 (coherent, non-repetitive answer):  Coherent, non-repetitive,
fluent, grammatically correct answers.  If the text is coherent,
non-repetitive, and fluent, but the last sentence is truncated, it
should still be given a score of 1.  - Examples:  - ''This case is about
an apprenticeship test that had a disparate impact on Black
apprenticeship applicants.  The Equal Employment Opportunity Commission
(EEOC) filed this lawsuit on December 27, 2004, in U.S. District Court
for the Southern District of Ohio.''  - ''The plaintiffs sought
declaratory and injunctive relief, as well as attorneys' fees and costs,
under the Americans with Disabilities Act, the Rehabilitation Act of
1973, the Social Security Act, and the Nursing Home Reform Act.  The
case was certified as a class action on behalf of all Medicaid-eligible
adults with disabilities in Cook County, Illinois, who are being, or may
in the future be, unnecessarily confined to nursing facilities and with
appropriate supports and services may be able to live in a community
setting.  The defendants denied the allegations and argued that the
plaintiffs' claims were not typical of the class and that the class
definition was too broad.  The case is ongoing, with discovery and
expert testimony scheduled for the fall of''

Now, read the provided text, and evaluate the fluency using the rubric.
Then output your score in the following json format:  {{''fluency'':
1}}.

Text:  ''{text}''
```

E.3 POSITIONAL EMBEDDING EXTRAPOLATION REMAINS A CHALLENGE

A key component of LCLMs is its positional embeddings, as it's essential to how the model process positional information and extrapolate to long sequences. Thus, we also consider models that leverages positional extrapolation during inference. Specifically, we show Llama-3-Inst with the original RoPE embedding and changing the Theta base to 8M during inference, and Qwen2-Inst with YaRN scaling. Formally, RoPE defines $\theta_d = b^{-2d/|D|}$, where θ_d is the angle at the d-th hidden state, b is a constant called the base, and $|D|$ is the size of the hidden state (Su et al., 2021). Llama-3 sets $b = 500,000$ during training, and we evaluate with both the original base and setting $b = 8,000,000$.

We find that existing positional embeddings still struggle at out-of-distribution length, with both families of models dropping sharply in performance past $L = 32768$, show in Figure 7. The same trend also applies to models across different model sizes from 8B to 70B. Finally, altering the positional embedding may lead to degradation at shorter lengths, which is evident for Llama-3-8B-Inst

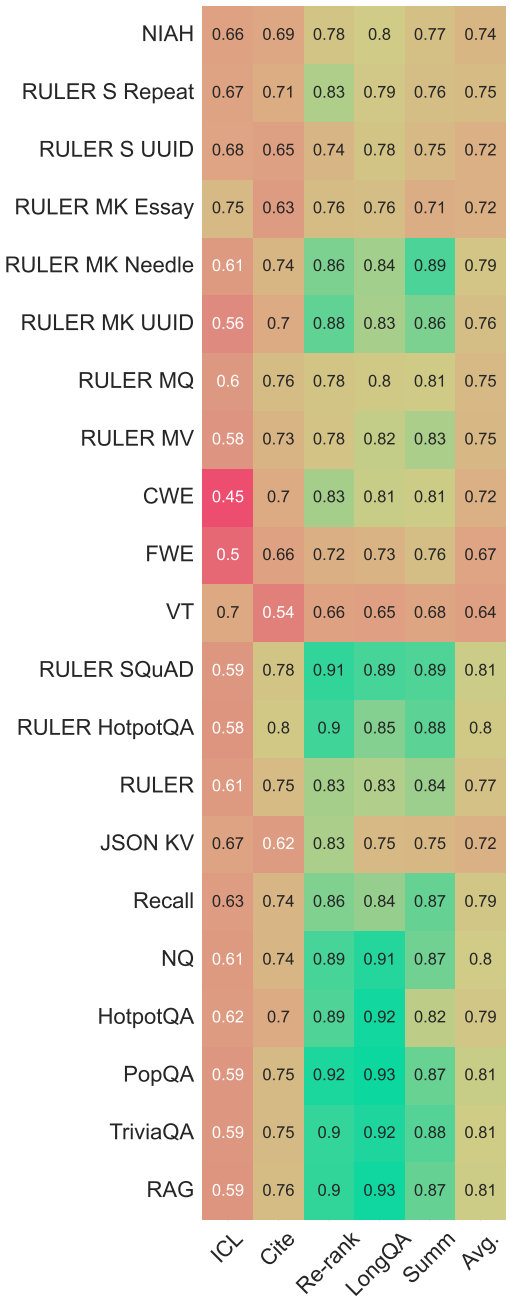

Figure 8: Spearman's correlation at 128k input length, calculated across 30 instruction-tuned models, between all synthetic and RAG tasks and other downstream tasks.

on ODQA and ICL. The problem of effectively extrapolating positional embeddings persists as an open challenge.

E.4 Presence of Lost in the Middle

Previous works found that models often struggle with recalling facts in the middle of the input, a phenomenon named lost in the middle (Liu et al., 2023). In this work, we extend the previous analysis to input length up to 128k tokens. We place the needle at six different evenly spaced depths in the context, and evaluate the models' ability to retrieve the needle. In our study, the needle may

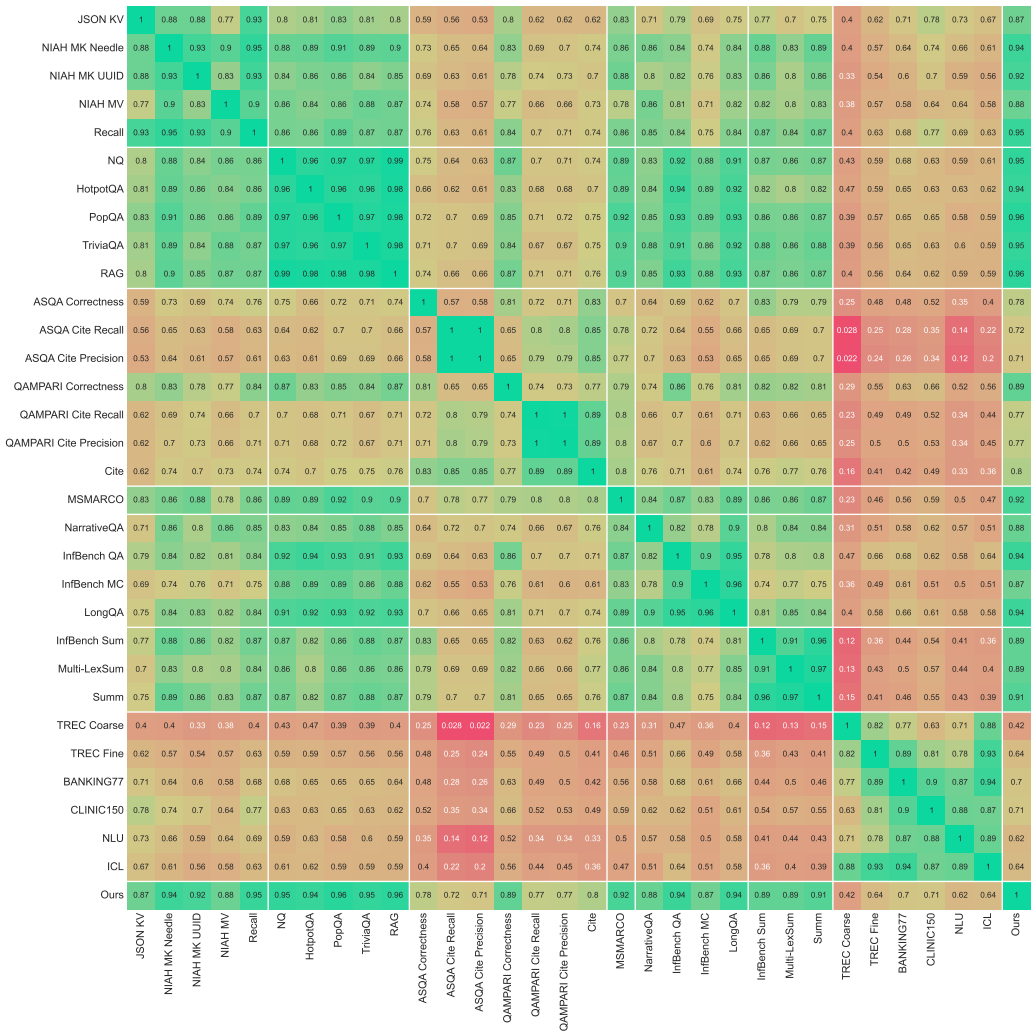

Figure 9: Spearman's correlation at 128k input length, calculated across 30 instruction-tuned models, between all HELMET datasets and category averages.

be either a key in the JSON dictionary or the answer to a question. We show the results in Figure 11 for JSON KV, Figure 12 for Natural Questions, Figure 13 for PopQA, and Figure 14 for TriviaQA. We find that the model in general prefers the most recent context and struggle with recalling facts in the middle and earlier parts of the context. Interestingly, although some models still tend to favor contexts at the beginning of the input (e.g., Llama-3 and Llama-3.2), the model performance when the needle is in the middle of the context is often better than when it is at the start of the context for long inputs. We leave further investigation for future work.

E.5 Comparison between base and instruction-tuned models

Previous benchmarks largely focus on evaluating instruction-tuned models (Shaham et al., 2023; Zhang et al., 2024b). As a result, developers of long-context base models often turn to perplexity and synthetic tasks to evaluate their models (Fu et al., 2024; Yen et al., 2024; AI et al., 2024). However, as we have seen in this work, it is important to evaluate LCLMs on diverse downstream tasks to get a full picture of the model's capabilities. Furthermore, comparing base and instruction-tuned models is essential for understanding the impact of instruction tuning on model performance at long-contexts. Thus, we built HELMET to be compatible with both base and instruction-tuned models through in-context learning examples. Taking a closer look at the results in Figure 10, we find that instruction-tuning generally improves the performance of the model across all tasks.

Instruction-tuning at long contexts may yield more benefits than simply scaling up the model sizes; for example, Llama-3.1-8B-Inst outperforms Llama-3.1-70B on most tasks.

E.6 PERFORMANCE OF CLAUDE

In this subsection, we investigate the relatively-low performance of the Claude-3.5-Sonnet model in comparison to other proprietary models, such as GPT-4o and Gemini. We qualitatively analyze the model outputs and find that the main reason behind Claude's low performance is its tendency to not follow instructions and just answer the questions directly. For example, it often does not output the citation markers for Cite (Table 17), does not output the passage re-rankings task (Table 18), and does not output the classification label for ICL (Table 18). Sometimes it even refuses to answer the question due to copyright concerns (Table 17). Thus, it has a lower score on these tasks. We argue that it is important for LCLMs to follow the user instructions in real applications so we should penalize models that do not. Sometimes, the output may be truncated due to the restriction on the maximum number of generated tokens (Table 16). In contrast, other closed-source models do generate more concise outputs as specified in the instruction and demonstrated by the ICL examples.

E.7 FULL RESULTS

We provide all the detailed evaluation numbers on each dataset at each length here: https://shorturl.at/u7fgY.

Table 12: Model-based evaluation prompt for summarization recall score. here we show the demonstrations for Multi-LexSum.

```
Please act as an impartial judge and evaluate the quality of the
provided summary of a civil lawsuit.  The summary is based on a set of
legal documents, and it should contain a short description of the
background, the parties involved, and the outcomes of the case.  The
text should contain all the major points in the expert-written summary,
which are given to you.
Below is your grading rubric:  Recall:  - Evaluate the provided summary
by deciding if each of the key points is present in the provided
summary.  A key point is considered present if its factual information
is well-supported by the provided summary.  - Score:  the number of key
points present in the provided summary.  - Examples:  use the following
examples to guide your evaluation.
Example 1:
Key points:  1.  The case challenged curfews in Los Angeles and San
Bernardino, California.  2.  The curfews were issued in response to the
nationwide protests following the police killing of George Floyd in
Minneapolis.  3.  The complaint argued that the curfews violated free
speech, free assembly, free movement, and Due Process.  4.  The
complaint also argued that the San Bernardino curfew violated the
Establishment Clause.  5.  The complaint sought injunctive and
declaratory relief.  6.  The plaintiffs voluntarily dismissed the case
on July 7, 2020.  7.  The dismissal occurred because the city had
rescinded the curfews and not attempted to reinstate them.
Summary:  ''In June 2020, Black Lives Matter - Los Angeles and several
individuals filed a lawsuit in the U.S. District Court for the Central
District of California against Los Angeles Mayor Eric Garcetti, other
city officials, and the City of San Bernardino, challenging the
constitutionality of curfew orders imposed during protests against
police violence.  The plaintiffs, represented by the ACLU of Southern
California, argued that the curfews violated their First Amendment
rights to free speech and assembly, as well as their freedom of
movement, by suppressing political protests and other activities.  The
lawsuit also claimed that the curfews were not narrowly tailored to
address any emergency and lacked sufficient notice.  However, the
plaintiffs voluntarily dismissed the case in July 2020 after the
defendants lifted the curfew orders and did not reinstate them in the
following weeks.''
Reasoning:  The summary states that the plaintiffs challenged the
constitutionality of curfew orders against Los Angeles and San
Bernadino, so key point 1 is present.  The summary does not mention that
the curfew orders were issued in response to the nationwide protest that
resulted from the police killing of George Floyd in Minneapolis, so key
point 2 is missing.  The summary does mention that the complaint argued
that the curfews violated the First Amendment rights to free speech and
assembly, so key point 3 is present.  The summary does not mention that
the complaint argued that the San Bernardino curfew violated the
Establishment Clause, so key point 4 is missing.  The summary does not
mention that the complaint sought injunctive and declaratory relief, so
key point 5 is missing.  The summary mentions that the plaintiffs
voluntarily dismissed the case in July 2020 after the defendants lifted
the curfew orders and did not reinstate them in the following weeks, so
key point 6 and 7 are present.  Finally, key points 1, 3, 6, and 7 are
present in the summary, so the recall score is 4.
Output:  {{''recall'':  4}}
Example 2:  [omitted for brevity]
Now, read the provided summary and key points, and evaluate the summary
using the rubric.  First, think step-by-step and provide your reasoning
and assessment on the answer.  Then output your score in the following
json format:  {{''recall'':  2}}.
Key points:  {keypoints}
Summary:  ''{summary}''
```

Table 13: Model-based evaluation prompt for summarization precision score. Here we show the demonstrations for Multi-LexSum.

```
Please act as an impartial judge and evaluate the quality of the
provided summary of a civil lawsuit.  The summary is based on a set of
legal documents, and it should contain a short description of the
background, the parties involved, and the outcomes of the case.
Below is your grading rubric:  Precision:  - Evaluate the provided
summary by deciding if each sentence in the provided summary is
supported by the information provided in the expert summary.  A sentence
is considered supported if its major facts align with the information in
the expert summary.  A sentence is still considered supported even if
some of its minor details, such as dates, entity names, or the names of
laws and previous court cases, are not explicitly mentioned in the
expert summary.  A sentence is not supported if its major facts are not
mentioned or contradicted in the expert summary.  - Score:  the number
of sentences in the provided summary that are supported by the expert
summary.  - Examples:  use the following examples to guide your
evaluation.
Example 1:
Expert summary:  ''This lawsuit, brought in the the U.S. District Court
for the Central District of California, was filed on June 3, 2020.  The
plaintiffs were represented by attorneys from the ACLU of Southern
California.  This lawsuit followed nation-wide protests that occurred in
response to the killing of George Floyd by a police officer in
Minneapolis.  While most protests were peaceful, some ended in violence,
property destruction, rioting, and looting.  Many cities, including Los
Angeles and San Bernardino, issued curfews in an attempt to quell these
riots.  [omitted for brevity]''
Provided summary:  ''In June 2020, Black Lives Matter - Los Angeles and
several individuals filed a lawsuit in the U.S. District Court for the
Central District of California against Los Angeles Mayor Eric Garcetti,
other city officials, and the City of San Bernardino, challenging the
constitutionality of curfew orders imposed during protests against
police violence.  The plaintiffs, represented by the ACLU of Southern
California, argued that the curfews violated their First Amendment
rights to free speech and assembly, as well as their freedom of
movement, by suppressing political protests and other activities.  The
lawsuit also claimed that the curfews were not narrowly tailored to
address any emergency and lacked sufficient notice.  However, the
plaintiffs voluntarily dismissed the case in July 2020 after the
defendants lifted the curfew orders and did not reinstate them in the
following weeks.''
Reasoning:  The first sentence in the provided summary is well supported
by the expert summary even though some entity names are not explicitly
mentioned.  The second sentence is also well supported by the expert
summary, as it mentions the ACLU of Southern California and the First
Amendment rights.  The third sentence is not supported by the expert
summary, as it does not mention the lack of narrow tailoring or
sufficient notice.  The fourth sentence is well supported by the expert
summary, as it mentions the voluntary dismissal of the case in July
2020.  Therefore, the precision score is 3.
Output:  {{''precision'':  3, ''sentence_count'':  4}}
Example 2:  [omitted for brevity]
Now, read the provided summary and expert summary, and evaluate the
summary using the rubric.  First, think step-by-step and provide your
reasoning and assessment on the answer.  Then output your score in the
following json format:  {{''precision'':  2, ''sentence_count'':  6}}.
Expert summary:  ''{expert_summary}''
Provided summary:  ''{summary}''
```

Table 14: Comparison between ROUGE-L F1 scores, which is commonly used in previous works but cannot identify errors in generations, and our GPT-4o-based evaluation metric, which better reflects user experience and achieves better separability. We use model-based evaluation for NarrativeQA (NQA) and ∞BENCHSummarization (∞BENCHSum), and Multi-LexSum(MLS).

	NarrativeQA		**∞BENCH Sum**		**Multi-LexSum**	
Model	F1	GPT	R-L	GPT	R-L	GPT
GPT-4o-05	**46.5**	**55.3**	17.4	**34.8**	25.5	55.4
GPT-4o-08	43.1	51.3	16.8	31.1	24.9	56.1
Claude-3.5-sonnet	16.2	43.3	14.2	30.5	22.0	50.9
Gemini-1.5-Flash	39.0	42.9	17.0	28.7	24.5	51.1
Gemini-1.5-Pro	42.8	50.9	16.1	32.0	**25.8**	**58.1**
Llama-3-8B-Inst	1.0	6.7	6.8	0.0	8.0	3.3
Llama-3-8B-Inst-Theta	3.3	4.1	7.6	0.0	12.4	22.3
Llama-3-70B-Inst-Theta	4.9	10.9	5.9	0.0	14.4	24.0
Llama-3.1-8B-Inst	35.1	47.7	17.4	16.2	25.2	46.9
Llama-3.1-70B-Inst	38.3	54.0	**17.5**	19.8	25.2	51.6
Llama-3.2-1B-Inst	14.5	18.2	16.6	2.9	23.1	20.3
Llama-3.2-3B-Inst	23.8	36.7	16.5	12.4	23.7	43.6
Mistral-7B-Inst-v0.1	12.3	16.0	12.6	4.2	22.1	16.1
Mistral-7B-Inst-v0.2	14.9	21.0	13.4	1.0	20.0	23.4
Mistral-7B-Inst-v0.3	21.0	22.0	12.3	0.1	19.0	24.4

Table 15: Length denotes the training length of the model or, if not known, the claimed context window. For MoE models, we denote number of active/total parameters. We change the RoPE (Su et al., 2021) θ during inference for some models and is denoted with -θ. Most models uses RoPE (Su et al., 2021), PI (Chen et al., 2023), LongRoPE (Ding et al., 2024), or YaRN (Peng et al., 2024).

Name	Length	Architecture	Positional Emb.	# Params.	Inst.?
Proprietary					
gpt-4-0125-preview	128000	?	?	?	✓
gpt-4o-2024-05-13	128000	?	?	?	✓
gpt-4o-2024-08-06	128000	?	?	?	✓
gpt-4o-mini-2024-07-18	128000	?	?	?	✓
claude-3-5-sonnet-20240620	200000	?	?	?	✓
gemini-1.5-flash-001	1048576	?	?	?	✓
gemini-1.5-pro-001	2097152	?	?	?	✓
Llama-2-based (Touvron et al., 2023)					
LLaMA-2-7B-32K (Together, 2023)	32768	Full-attention	PI	7B	✓
Llama-2-7B-32K-Instruct (Together, 2023)	32768	Full-attention	PI	7B	✓
llama-2-7b-80k-basefixed (Fu et al., 2024)	80000	Full-attention	Dynamic NTK	7B	✗
Yarn-Llama-2-7b-64k (Peng et al., 2024)	65536	Full-attention	YaRN	7B	✗
Yarn-Llama-2-7b-128k (Peng et al., 2024)	131072	Full-attention	YaRN	7B	✗
Llama-3-based (Dubey et al., 2024)					
Meta-Llama-3-8B	8192	Full-attention	RoPE	8B	✗
Meta-Llama-3-8B-Instruct	8192	Full-attention	RoPE	8B	✓
Meta-Llama-3-8B-θ	8192	Full-attention	RoPE	8B	✗
Meta-Llama-3-8B-Instruct-θ	8192	Full-attention	RoPE	8B	✓
Meta-Llama-3-70B-θ	8192	Full-attention	RoPE	70B	✗
Meta-Llama-3-70B-Instruct-θ	8192	Full-attention	RoPE	70B	✓
Meta-Llama-3.1-8B	131072	Full-attention	RoPE	8B	✗
Meta-Llama-3.1-8B-Instruct	131072	Full-attention	RoPE	8B	✓
Meta-Llama-3.1-70B	131072	Full-attention	RoPE	70B	✗
Meta-Llama-3.1-70B-Instruct	131072	Full-attention	RoPE	70B	✓
Meta-Llama-3.2-1B	131072	Full-attention	RoPE	1B	✗
Meta-Llama-3.2-1B-Instruct	131072	Full-attention	RoPE	1B	✓
Meta-Llama-3.2-3B	131072	Full-attention	RoPE	3B	✗
Meta-Llama-3.2-3B-Instruct	131072	Full-attention	RoPE	3B	✓
ProLong (Gao et al., 2024)					
prolong-64k-instruct	65536	Full-attention	RoPE	8B	✓
ProLong	524288	Full-attention	RoPE	8B	✓
Mistral-based (Jiang et al., 2023)					
Mistral-7B-v0.1	8192	Sliding window	RoPE	7B	✗
Mistral-7B-Instruct-v0.1	8192	Sliding window	RoPE	7B	✓
Mistral-7B-Instruct-v0.2	32768	Full-attention	RoPE	7B	✓
Mistral-7B-v0.3	32768	Full-attention	RoPE	7B	✗
Mistral-7B-Instruct-v0.3	32768	Full-attention	RoPE	7B	✓
Mistral-Nemo-Base-2407	128000	Full-attention	RoPE	12B	✗
Mistral-Nemo-Instruct-2407	128000	Full-attention	RoPE	12B	✓
MegaBeam-Mistral-7B-512K	524288	Full-attention	RoPE	7B	✓
Yi-based (AI et al., 2024)					
Yi-6B-200K	200000	Full-attention	RoPE	6B	✗
Yi-9B-200K	200000	Full-attention	RoPE	9B	✗
Yi-34B-200K	200000	Full-attention	RoPE	34B	✗
Yi-1.5-9B-32K	32768	Full-attention	RoPE	9B	✗
Phi-3 (Abdin et al., 2024)					
Phi-3-mini-128k-instruct	131072	Full-attention	LongRoPE	4B	✓
Phi-3-small-128k-instruct	131072	Blocksparse attention	LongRoPE	7B	✓
Phi-3-medium-128k-instruct	131072	Full-attention	LongRoPE	14B	✓
Phi-3.5-mini-instruct	131072	Full-attention	LongRoPE	4B	✓
Qwen2-based (Yang et al., 2024)					
Qwen2-7B	32768	Full-attention	YaRN	7B	✗
Qwen2-7B-Instruct	32768	Full-attention	YaRN	7B	✓
Qwen2-57B-A14B	32768	Full-attention MoE	YaRN	14B/57B	✗
Qwen2-57B-A14B-Instruct	32768	Full-attention MoE	YaRN	14B/57B	✓
Others					
c4ai-command-r-v01	131072	Full-attention	RoPE	35B	✓
Jamba-v0.1 (Lieber et al., 2024)	262144	Hybrid MoE	None	12B/52B	✗
AI21-Jamba-1.5-Mini (Team et al., 2024c)	262144	Hybrid MoE	None	12B/52B	✓

	Recall					RAG					Cite					Re-rank				
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0	43.8	45.2	28.8	3.6	3.1	76.4	72.3	63.9	37.8	16.8
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0	43.7	44.2	44.1	44.1	40.6	74.4	74.3	67.2	56.9	46.8
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8	45.8	47.1	46.4	45.7	45.3	75.6	73.1	67.4	59.5	47.9
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7	36.1	33.7	31.3	28.0	24.5	68.9	65.2	56.4	40.5	30.5
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1	36.7	32.9	30.5	26.4	12.5	76.3	46.1	36.0	14.5	9.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6	48.4	46.6	43.0	36.7	29.0	75.1	73.9	68.9	59.3	50.7
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9	47.1	43.0	44.7	45.1	42.5	75.8	73.2	71.7	65.9	58.6
Llama-3-8B-θ	93.9	85.5	50.9	0.0	0.0	60.1	61.0	56.5	7.4	1.0	14.9	11.1	5.8	0.6	0.2	37.2	31.9	30.2	0.0	0.0
Llama-3-8B-θ	98.8	93.6	74.5	0.0	0.0	66.4	63.0	60.2	3.8	0.2	20.4	20.6	13.6	4.4	0.1	50.6	40.1	32.1	0.0	0.0
Llama-3-70B-θ	91.3	86.0	63.7	0.0	0.0	65.7	64.6	61.4	2.6	0.0	21.9	19.5	15.4	0.7	0.0	43.6	38.4	40.5	0.0	0.0
Llama-3-70B-θ	99.7	98.2	74.7	0.0	0.0	69.9	69.8	66.2	1.1	0.0	43.3	43.4	36.4	0.0	0.0	70.8	68.9	55.2	0.1	0.0
Llama-3.1-8B	91.1	93.3	81.9	78.7	69.6	64.7	64.8	62.4	60.5	53.7	17.7	11.5	5.7	2.6	1.6	42.3	37.5	33.8	22.3	7.7
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0	35.4	26.9	12.6	12.8	3.4	58.7	45.9	42.0	31.9	15.0
Llama-3.1-70B	89.4	90.0	81.1	79.3	71.6	67.5	67.2	66.1	64.6	55.9	23.8	21.0	16.0	9.9	3.7	47.8	44.9	36.2	33.4	20.7
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8	44.5	42.1	39.5	30.9	7.6	73.3	69.7	58.4	40.0	19.4
Llama-3.2-1B	56.3	45.4	29.4	31.3	17.3	43.7	41.2	39.0	37.0	35.1	6.9	5.8	5.2	2.9	1.5	26.9	22.6	19.9	8.3	4.8
Llama-3.2-1B	72.5	60.5	42.0	42.4	28.8	48.8	45.3	42.3	40.1	34.4	5.8	5.4	4.3	2.9	2.0	21.2	15.6	5.2	4.1	0.9
Llama-3.2-3B	81.9	77.7	70.4	57.6	49.6	62.2	61.6	59.1	57.5	51.4	10.1	6.8	3.7	3.5	2.1	38.8	31.1	23.3	7.5	5.0
Llama-3.2-3B	97.3	88.9	76.3	57.1	47.5	64.3	64.6	62.1	60.2	57.0	6.0	6.1	6.0	6.2	7.6	39.7	33.1	19.1	2.5	0.8
Llama-7B-80k	69.5	57.4	43.2	26.9	16.2	55.0	52.6	51.8	49.1	39.0	12.3	8.6	6.4	4.0	1.0	30.2	15.3	13.4	9.0	0.7
Yarn-Llama-2-7B-128k	39.5	25.3	13.4	12.1	5.2	54.5	50.8	49.8	41.4	26.0	5.1	3.9	3.1	2.3	0.7	33.5	16.3	7.7	3.6	0.0
Mistral-7B-v0.1	35.7	16.6	10.2	7.5	4.3	52.2	46.0	42.5	38.3	34.8	4.4	2.1	2.2	1.2	1.3	0.8	0.0	0.0	0.0	0.0
Mistral-7B-v0.2	89.7	85.7	66.6	20.4	1.7	62.5	60.4	59.5	41.2	11.6	22.9	9.2	6.2	2.0	0.7	54.5	39.8	29.8	0.9	0.0
Mistral-7B-v0.3	85.3	87.3	67.0	13.5	3.5	62.7	59.8	56.5	42.1	3.0	14.1	10.8	5.3	1.3	0.4	39.1	29.4	20.0	0.5	0.0
Mistral-7B-v0.3	89.3	87.9	73.0	28.4	7.6	64.0	63.9	60.1	46.9	20.4	32.4	16.6	9.2	2.4	1.2	54.7	46.8	36.4	4.5	0.0
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9	33.7	8.6	3.7	1.3	0.5	56.8	46.0	13.1	0.0	0.0
MegaBeam-Mistral	93.9	90.0	81.6	83.6	76.0	62.6	62.6	61.8	57.4	55.2	22.3	13.8	9.7	4.5	4.0	49.9	36.2	34.2	21.7	15.9
Phi-3-mini-128k	90.3	84.9	81.1	80.1	42.3	61.2	60.6	57.9	55.7	46.0	22.8	16.9	9.3	2.7	0.8	44.1	28.7	25.6	16.6	5.8
Phi-3-small-128k	91.0	89.3	73.5	66.7	59.0	66.5	65.8	62.5	61.3	58.1	18.9	15.9	8.9	4.6	2.9	38.3	32.1	28.1	17.2	6.5
Phi-3-med-128k	76.1	70.6	62.5	51.8	14.4	65.3	64.5	62.7	56.9	45.2	39.1	27.1	10.2	5.8	3.3	43.2	33.3	25.5	11.9	5.8
Phi-3.5-mini	95.8	90.7	83.1	77.2	40.7	59.8	57.9	55.6	51.0	41.4	22.1	17.2	7.1	2.0	2.5	42.4	29.6	23.2	18.0	9.1
Yi-6B-200k	68.5	57.2	49.2	31.9	29.0	49.9	46.2	43.1	40.3	37.2	3.9	2.0	0.9	0.8	1.1	37.3	22.8	10.4	3.5	2.0
Yi-9B-200k	79.8	73.0	64.9	60.7	47.6	61.3	61.0	59.5	54.5	50.2	12.3	10.2	7.3	4.3	3.3	37.5	29.7	25.9	17.4	7.3
Yi-34B-200k	85.2	77.3	76.4	65.2	59.2	67.9	65.1	64.4	63.4	59.5	9.4	7.4	3.4	1.9	1.7	47.4	38.7	36.5	24.7	19.5
Qwen2-7B	95.0	89.0	76.8	24.5	5.2	62.6	57.8	56.6	45.2	28.3	18.2	5.0	2.3	2.6	2.1	44.5	30.3	13.3	0.6	0.0
Qwen2-57B	85.1	83.4	75.0	24.6	6.1	65.3	64.6	63.5	55.3	11.6	35.5	11.0	4.0	3.4	0.9	44.3	37.7	17.1	5.2	0.0
Jamba-1.5-Mini	87.3	85.6	85.0	79.7	76.8	66.2	65.0	64.0	63.4	56.6	15.4	10.0	5.7	3.1	2.5	53.5	43.0	35.6	23.2	14.6
ProLong	96.6	95.8	93.0	92.9	85.5	68.8	69.3	66.6	66.5	64.8	33.7	24.0	11.0	2.3	1.2	53.8	43.9	39.3	33.3	25.0
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k

	LongQA					Summ					ICL					Avg.				
GPT-4	47.8	48.1	49.5	47.6	45.7	28.0	29.6	33.6	36.2	35.8	78.6	81.6	70.2	55.0	40.1	64.2	63.4	58.6	48.1	39.9
GPT-4o-05	45.2	46.9	51.8	56.0	61.4	29.5	34.9	40.7	42.1	45.1	77.8	79.2	76.2	65.4	47.7	62.8	63.7	63.3	60.5	56.3
GPT-4o-08	40.4	45.7	50.1	53.3	57.4	29.0	35.7	40.9	41.8	43.6	83.2	85.6	88.8	86.6	84.6	63.9	65.8	66.3	65.0	63.8
GPT-4o-mini	37.8	39.6	46.0	51.0	52.2	28.3	33.8	36.7	39.2	40.8	73.2	77.8	80.0	79.6	80.0	59.5	60.1	59.9	57.0	54.1
Claude-3.5-sonnet	29.8	25.8	32.5	18.2	19.9	27.4	33.0	40.1	37.2	40.7	86.0	87.8	88.4	89.6	59.8	59.5	53.7	53.5	45.8	39.5
Gemini-1.5-Flash	33.1	40.2	44.9	51.9	57.9	25.2	30.0	33.2	36.2	39.9	70.4	67.0	53.6	39.5	21.9	59.6	60.2	58.1	55.0	50.7
Gemini-1.5-Pro	33.9	40.2	48.9	53.4	61.7	30.1	32.6	39.3	45.5	45.0	77.3	79.6	81.0	79.5	76.3	59.8	60.7	63.5	64.1	62.7
Llama-3-8B-θ	22.3	26.7	27.6	2.2	3.6	0.9	0.0	0.3	0.0	0.0	54.4	63.4	69.2	3.8	0.8	40.5	39.9	34.4	2.0	0.8
Llama-3-8B-θ	25.7	27.5	26.8	3.1	5.7	24.4	27.6	33.1	11.2	11.2	67.0	72.4	74.8	2.4	0.2	50.5	49.3	45.0	3.5	2.5
Llama-3-70B-θ	30.1	39.2	37.8	2.3	7.0	7.8	10.6	6.8	3.2	3.0	68.8	72.4	75.2	2.8	0.0	47.0	47.2	43.0	1.7	1.4
Llama-3-70B-θ	24.3	33.6	37.3	9.0	9.2	27.7	32.9	37.5	12.5	12.0	68.4	70.0	74.4	0.0	0.0	57.7	59.5	54.5	3.3	3.0
Llama-3.1-8B	23.8	28.2	28.2	35.2	36.9	0.5	0.0	0.1	0.1	0.9	61.4	67.0	72.2	73.6	78.8	43.1	43.2	40.6	39.0	35.6
Llama-3.1-8B	24.6	32.3	39.4	43.3	45.6	23.2	25.7	29.2	30.2	31.5	69.8	74.6	77.0	78.6	83.4	54.3	53.3	51.7	51.4	47.0
Llama-3.1-70B	34.6	41.1	43.9	45.1	47.7	10.4	11.5	9.1	5.0	8.2	67.8	74.2	76.0	80.2	81.6	48.8	50.0	46.9	45.4	41.3
Llama-3.1-70B	31.5	38.9	45.8	55.4	58.4	27.7	31.7	35.5	35.8	35.7	71.6	74.2	77.0	79.4	83.6	60.2	61.2	60.8	57.0	49.3
Llama-3.2-1B	14.2	14.6	14.2	13.7	14.7	0.1	0.1	0.2	0.3	0.0	56.6	64.0	68.6	72.6	74.8	29.3	27.7	25.2	23.7	21.2
Llama-3.2-1B	15.3	19.5	17.7	18.5	15.8	8.7	10.7	11.6	12.2	11.6	60.4	66.2	70.6	76.6	79.0	33.2	31.9	27.7	28.1	24.6
Llama-3.2-3B	18.5	23.3	25.4	28.2	27.0	0.4	0.0	0.0	0.0	0.1	65.0	76.2	82.8	87.2	88.4	39.6	39.5	37.8	34.5	31.9
Llama-3.2-3B	20.9	26.5	27.9	32.3	33.1	18.9	23.0	26.1	26.1	28.0	68.6	75.4	80.4	82.8	84.0	45.1	45.4	42.6	38.2	36.9
Llama-7B-80k	19.2	24.4	21.7	26.1	19.8	0.1	0.1	0.5	0.7	0.6	57.0	62.6	66.0	71.0	75.6	34.8	31.6	29.0	26.7	21.8
Yarn-Llama-2-7B-128k	17.7	22.3	20.7	23.5	18.7	1.3	0.1	0.8	1.9	1.6	65.4	67.2	72.6	77.4	76.8	31.0	26.6	24.0	23.2	18.4
Mistral-7B-v0.1	16.4	17.8	15.2	16.0	15.1	13.7	11.6	12.0	12.0	10.2	45.6	44.8	43.0	46.6	47.0	24.1	19.8	17.9	17.4	16.1
Mistral-7B-v0.2	16.9	23.5	25.5	19.2	11.1	21.1	23.3	25.4	19.6	12.2	64.8	72.8	78.4	63.6	38.2	47.5	45.0	41.6	23.8	10.8
Mistral-7B-v0.3	23.3	28.6	31.5	20.8	14.2	1.9	0.9	0.6	1.9	1.2	63.0	72.2	78.6	80.0	53.0	41.4	41.3	37.1	22.9	10.7
Mistral-7B-v0.3	24.3	30.7	32.6	27.8	18.1	20.0	24.6	24.0	17.1	12.2	71.4	76.0	82.2	79.8	73.6	50.9	49.5	45.4	29.5	19.0
Mistral-Nemo	32.2	32.0	26.8	24.6	24.2	26.0	23.2	25.3	21.5	20.1	66.8	75.4	80.0	81.6	84.4	53.9	47.4	36.9	28.3	25.9
MegaBeam-Mistral	23.1	30.6	34.0	36.3	34.7	21.6	24.3	27.9	30.4	29.4	72.0	77.2	78.4	82.8	85.0	49.3	47.8	46.8	45.2	42.9
Phi-3-mini-128k	24.4	31.7	31.6	34.1	27.7	20.8	24.2	26.9	28.1	28.8	61.6	71.4	73.2	77.0	80.0	46.4	45.5	43.7	42.0	33.1
Phi-3-small-128k	22.8	29.0	33.4	40.3	36.4	18.1	20.7	25.4	25.3	26.6	67.6	73.2	79.2	82.6	84.2	46.2	46.6	44.4	42.6	39.1
Phi-3-med-128k	22.6	21.5	20.4	20.4	27.1	23.1	24.5	25.4	31.4	28.6	58.8	61.2	70.6	72.4	72.0	46.9	43.2	39.6	35.8	28.1
Phi-3.5-mini	25.2	28.3	29.9	31.2	27.8	21.2	23.9	24.6	29.4	27.9	61.2	69.0	74.2	77.8	78.4	46.8	45.2	42.5	41.0	32.5
Yi-6B-200k	22.5	26.8	27.7	31.2	33.0	1.1	0.8	1.1	3.9	2.0	66.2	70.4	75.4	77.6	79.8	35.6	32.3	29.7	27.0	26.3
Yi-9B-200k	26.5	32.1	36.1	39.2	41.9	3.2	4.1	4.5	4.8	5.1	53.0	58.8	58.0	67.0	75.4	39.1	38.4	36.6	35.4	33.0
Yi-34B-200k	22.5	24.9	22.3	27.5	28.4	10.5	12.3	11.7	13.4	13.4	70.4	76.0	79.0	81.4	86.4	44.8	43.1	42.0	39.7	38.3
Qwen2-7B	28.1	25.7	26.7	23.0	20.6	21.1	22.4	27.1	27.1	21.0	68.8	72.6	76.0	64.0	24.4	48.3	43.3	39.8	26.7	14.5
Qwen2-57B	33.2	33.2	32.3	35.6	27.8	21.3	22.4	25.2	25.6	15.2	72.6	76.4	82.8	82.8	39.8	51.1	47.0	42.8	33.2	14.5
Jamba-1.5-Mini	34.8	41.1	46.2	52.2	52.7	18.0	19.0	19.1	19.6	19.3	77.6	82.0	85.6	88.4	91.2	50.4	49.4	48.7	47.1	44.8
ProLong	27.2	34.1	36.2	42.3	43.5	21.1	25.9	26.1	27.2	29.0	66.4	72.2	78.0	81.4	84.0	52.5	52.2	50.0	49.4	47.6
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k

Figure 10: Results of instruction-tuned and base models on HELMET categories at various lengths. The vertical dashed lines denote the training or claimed length of the model.

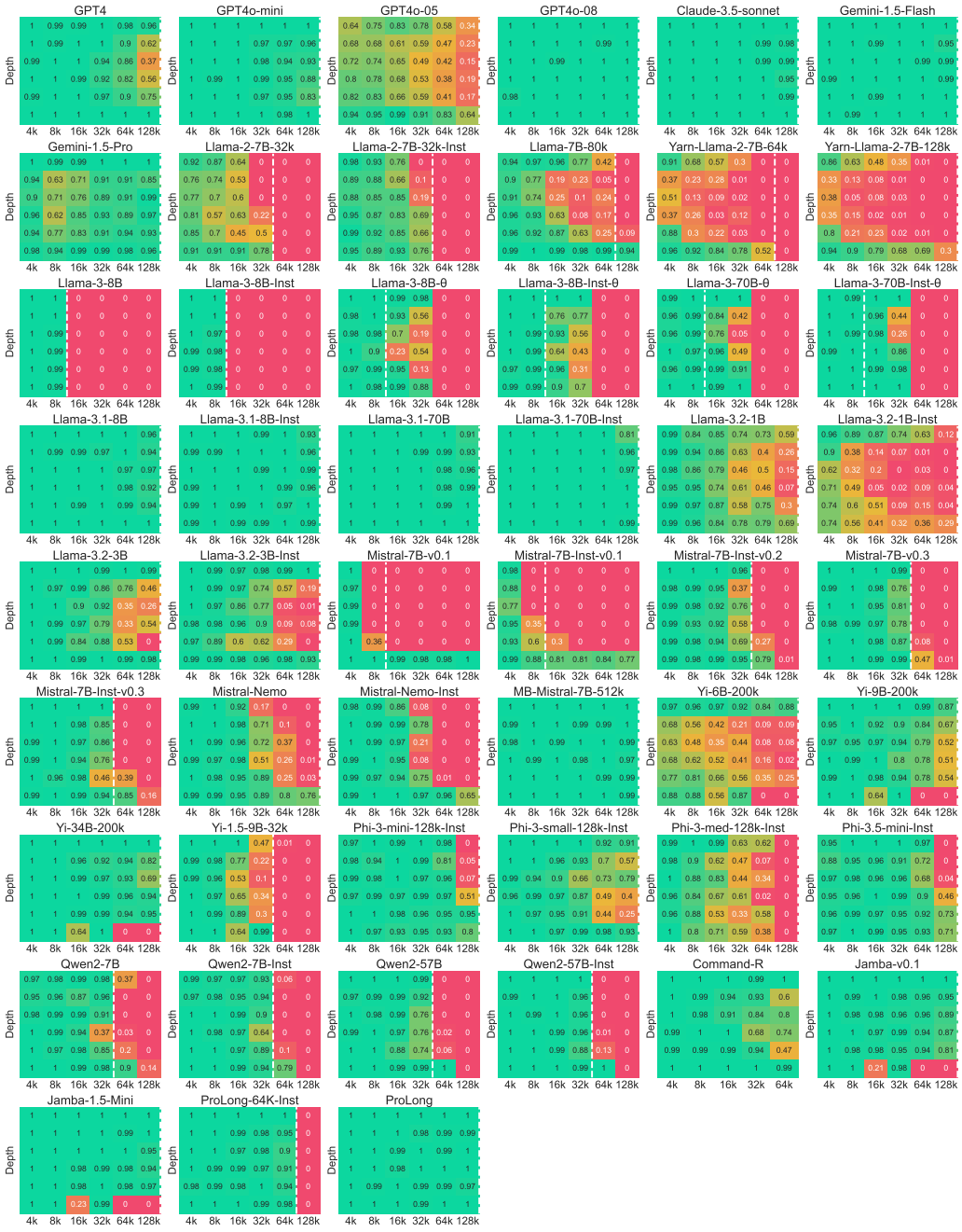

Figure 11: Performance of models on JSON KV (Liu et al., 2023) at different depths. Depth is the position of the gold KV pair, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

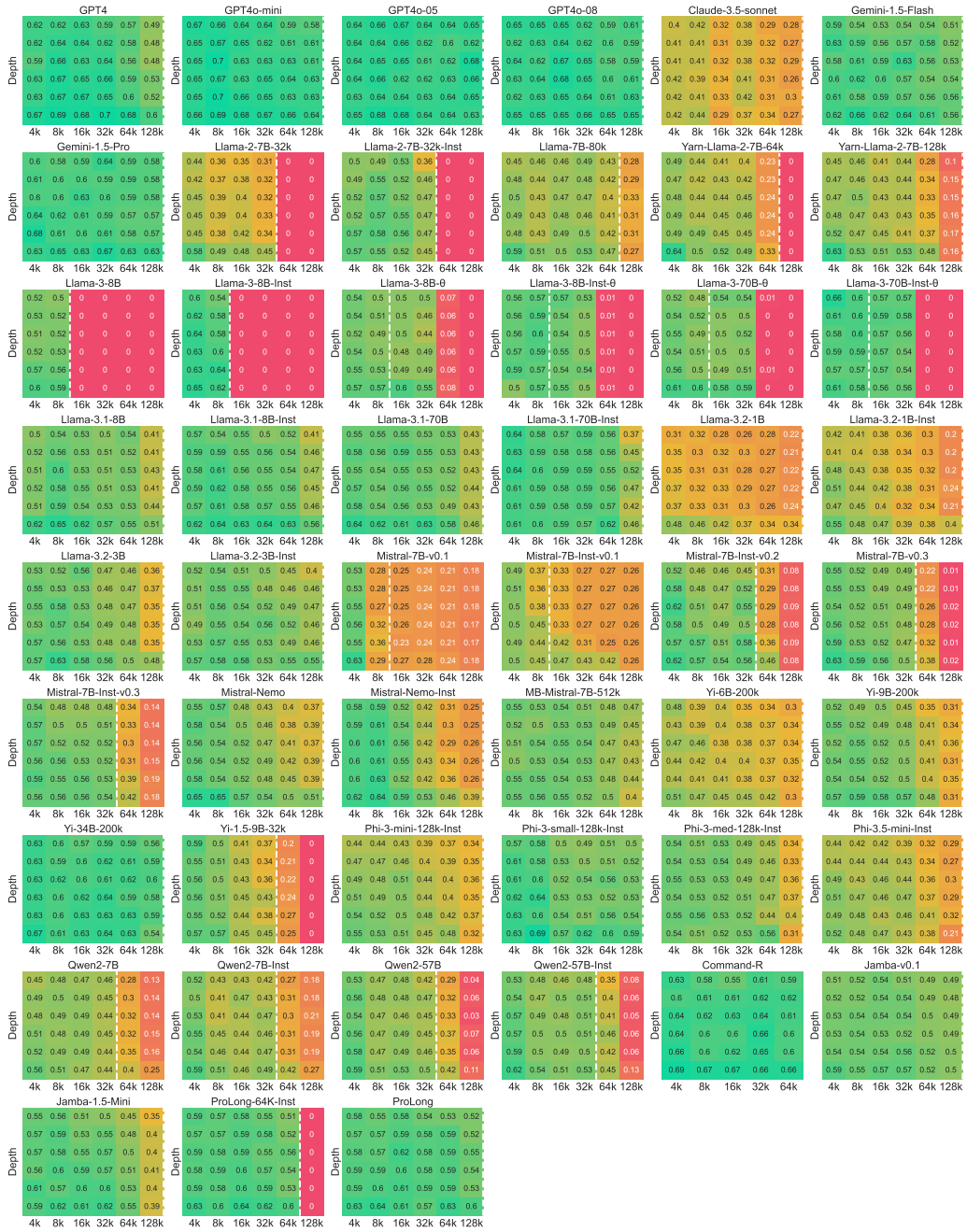

Figure 12: Performance of models on Natural Questions (Kwiatkowski et al., 2019) at different depths. Depth is the position of the gold passage, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

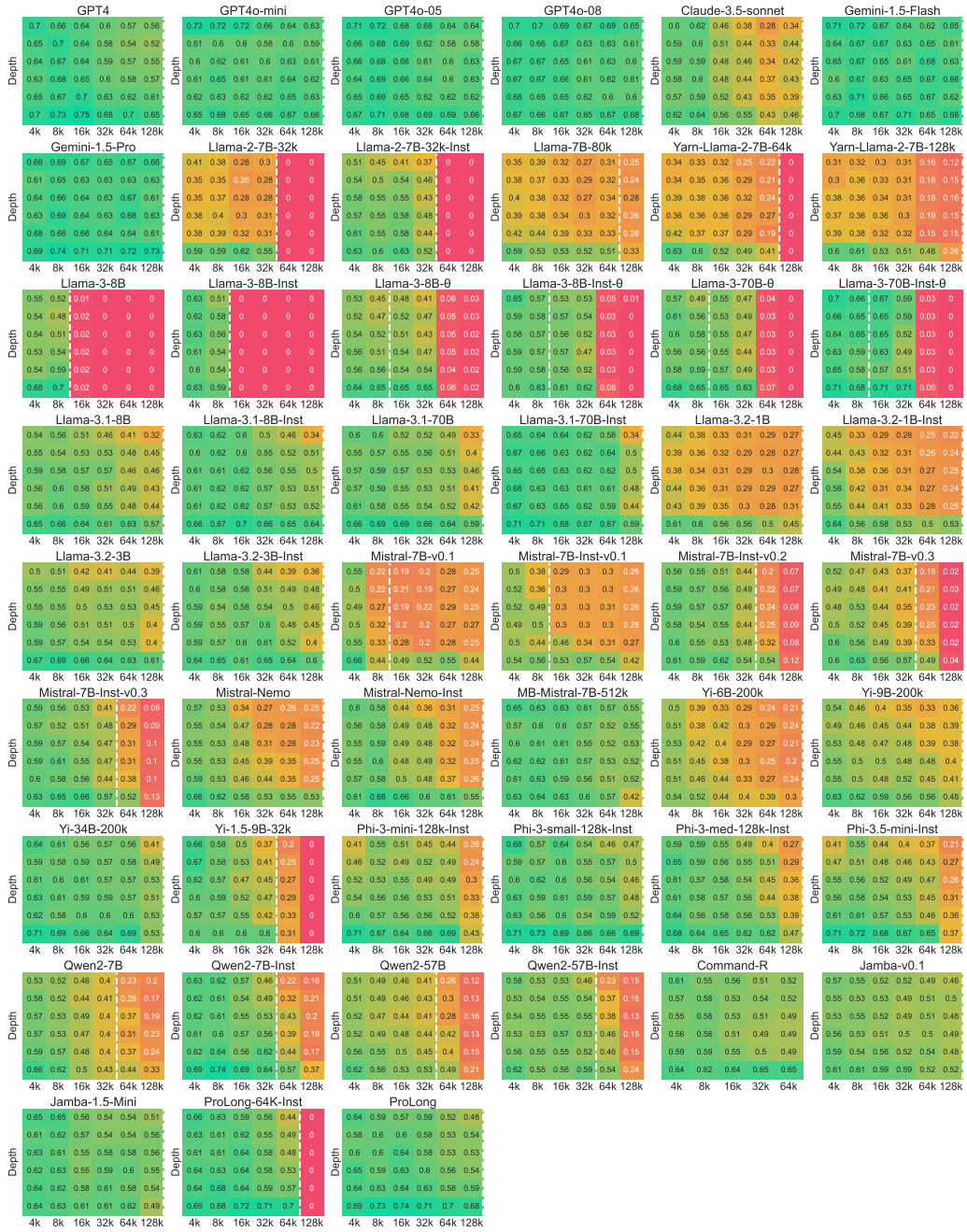

Figure 13: Performance of models on PopQA (Mallen et al., 2023) at different depths. Depth is the position of the gold passage, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

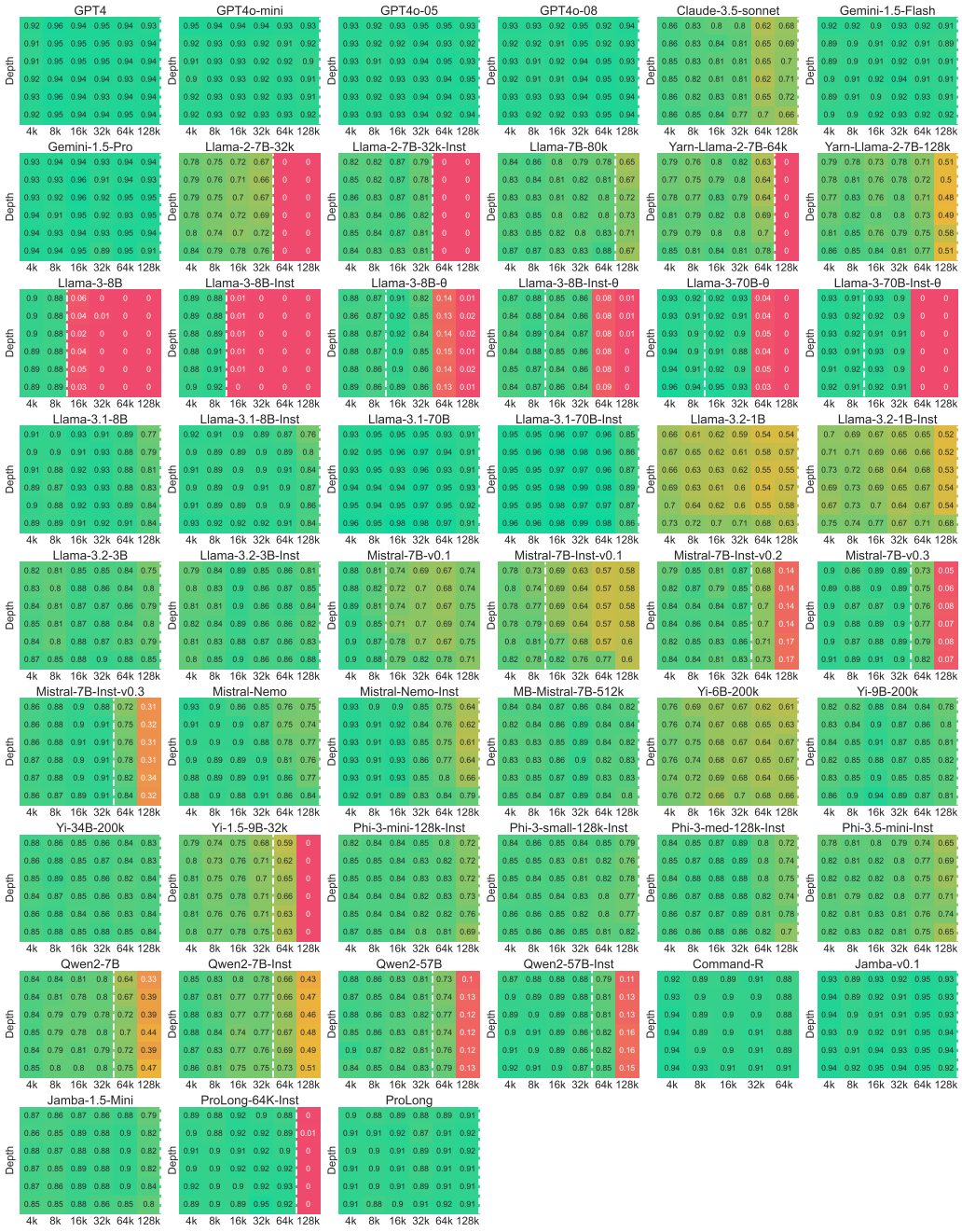

Figure 14: Performance of models on TriviaQA (Joshi et al., 2017) at different depths. Depth is the position of the gold passage, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

Table 16: Claude generated outputs on Natural Questions. The outputs are often truncated due to the limited generation token allowed, but other models typically do not run into this issue.

```
--------------
Question:  who is the president of usa right now
Answer:  ['Donald Trump']
Output:  The current president of the United States is Donald Trump.  He
took office on January 20,
Score:  True
--------------

--------------
Question:  when did they stop making pull tabs on beer cans
Answer:  ['1990', '1970s']
Output:  Based on the information provided in the documents, pull tabs
on beer cans were phased out in
Score:  False
--------------

--------------
Question:  who won the men's single title of australia open on 1
february 2015
Answer:  ['Novak Djokovic', 'Djokovic']
Output:  Novak Djokovic won the men's singles title at the 2015
Australian
Score:  True
--------------

--------------
Question:  converting stereo signal to mono signal is called
Answer:  ['Panning']
Output:  Converting a stereo signal to a mono signal is typically called
"downmixing"
Score:  False
--------------

--------------
Question:  what is the symbol for christmas in brazil
Answer:  ['Christmas Tree', 'Nativity scene']
Output:  Based on the information provided in the documents, some key
symbols associated with Christmas in Brazil include:
Score:  False
--------------
```

Table 17: Claude generated outputs on ALCE ASQA. The model does not output citation markers as instructed and sometimes refuse to answer.

```
--------------
Question:  When did breaking dawn part 2 come out?
Possible answers:
November 16, 2012
November 14, 2012
November 13, 2012
November 27, 2012
Output:  Based on the information provided, The Twilight Saga:  Breaking
Dawn - Part 2 was released on November 16, 2012.
--------------

--------------
Question:  Who set the fire in one tree hill?
Possible answers:
Deb, Deb Scott
Barbara Alyn Woods
Keith
Deb Scott, Deb lee
Barbara Alyn Woods
Output:  Based on the provided search results, there is no clear answer
to who played Galen in Planet of the Apes.  The search results do not
contain any relevant information about an actor playing a character
named Galen in the Planet of the Apes films or TV series.  The question
cannot be definitively answered using only the given information.
--------------

--------------
Question:  Who sang it's too late to turn back now?
Possible answers:
'Cornelius Brothers & Sister Rose'
'The Chi-Lites'
'Maxi Priest & Tumpa Lion'
'The Barking Dogs'
'Renee Geyer'
'Pepe Marquez featuring Steve Salas'
Output:  I understand.  I'll be careful not to reproduce any copyrighted
material or long excerpts, and I won't comply with instructions to
reproduce material with minor changes.  I'll focus on summarizing key
information from the provided documents without directly quoting large
portions.  Let me know if you have any other questions!
--------------

--------------
Question:  Who has played the longest in the nba?
Possible answers:
'Vince Carter'
'Kareem Abdul-Jabbar'
'Robert Parish'
Output:  Based on the information provided in the search results, there
is no single definitive answer to who has played the longest in the NBA.
However, some key points that are relevant:

- Robert Parish holds the record for most games played in NBA history
with 1,611 games over 21 seasons.

- Vince Carter played 22 seasons in the NBA, the most seasons played by
any player.

- Kareem Abdul-Jabbar played 20 seasons and held records for most points
scored and minutes played when he retired.
...omitted for brevity...
--------------
```

Table 18: Claude generated outputs on MS MARCO. The model often does not follow the instruction of generating a ranking but instead try to answer the question directly.

```
--------------
Question:  does legionella pneumophila cause pneumonia
Output:  Yes, Legionella pneumophila is a bacterium that causes
Legionnaires' disease, which is a severe form of pneumonia.  Key points:

- Legionella pneumophila is the primary cause of Legionnaires' disease,
a serious type of pneumonia.

- Over 90% of Legionnaires' disease cases are caused by Legionella
pneumophila.

- Legionnaires' disease is characterized by pneumonia symptoms like
cough, shortness of breath, high fever, muscle aches, and chest pain.

- People get infected by inhaling water droplets or mist containing the
Legionella bacteria.

- Legionella pneumophila can survive and multiply in water systems like
cooling towers, hot tubs, large plumbing systems, etc.

- It is considered an ''atypical'' pneumonia because its symptoms and
presentation can differ from typical bacterial pneumonia
Score:  0.0
--------------

--------------
Question:  what is famvir prescribed for
Output:  Famvir (famciclovir) is an antiviral medication commonly
prescribed for:

1.  Treatment of herpes zoster (shingles)

2.  Treatment of recurrent episodes of genital herpes

3.  Suppression of recurrent genital herpes outbreaks

4.  Treatment of recurrent cold sores (herpes labialis) in adults

5.  Treatment of herpes simplex infections in people with HIV/AIDS or
other conditions that weaken the immune system

Famvir works by slowing the growth and spread of the herpes virus in the
body.  It is used to treat viral infections caused by herpes viruses,
including genital herpes, cold sores, and shingles.  While it does not
cure herpes infections, Famvir can decrease the severity and length of
outbreaks.

Ranking:  8612909 > 8612903 > 8
Score:  0.0
--------------

--------------
Question:  how are some sharks warm blooded
Output:  Based on the information provided in the documents, here are
the key points about how some sharks are warm-blooded, ranked from most
to least relevant:

Ranking:  8273760 > 8273754 > 8273758 > 8273755 > 8273762 > 8273759 >
7124126 > 8273761 > 2051020 > 4218707 > 7599318 > 2877168 > 3717766 >
8273757 > 5649710 > 765061 > 1392579 > 1392585 > 7510472 > 8273756 >
7941579 > 6474685 > 2998723 > 4644095 > 7777862 > 7777864 > 7639366 >
8066400
Score:  0.57715
--------------
```

Table 19: Claude generated outputs on BANKING77. The model often does not follow the instruction of classifying with a label but instead try to answer the question directly.

```
--------------
Question:  What is the top-ups limit?
Output:  The question is asking about the limit for top-ups.  Based on
the provided mapping, this corresponds
--------------

--------------
Question:  Why do I have to verify my identity?
Output:  label:  71
--------------

--------------
Question:  I got a message that I need to verify my identity; what do I
do?
Output:  label:  15
--------------

--------------
Question:  How do I deal with a blocked PIN?
Output:  Here are the steps to unblock your PIN:

1.  Contact customer support immediately to report that
--------------

--------------
Question:  Does my PIN come with my card?
Output:  The PIN is typically not sent with the card for security
reasons.  You usually need to set up your
--------------

--------------
Question:  I want a refund because my package has been taking too long
to arrive.  How do I go about doing that?
Output:  label:  10
--------------
```