RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

Peng Xu†, Wei Ping†, Xianchao Wu, Lawrence McAfee
Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina
Mohammad Shoeybi, Bryan Catanzaro
NVIDIA
†{pengx, wping}@nvidia.com

ABSTRACT

Extending the context window of large language models (LLMs) is getting popular recently, while the solution of augmenting LLMs with retrieval has existed for years. The natural questions are: *i) Retrieval-augmentation versus long context window, which one is better for downstream tasks? ii) Can both methods be combined to get the best of both worlds?* In this work, we answer these questions by studying both solutions using two state-of-the-art pretrained LLMs, i.e., a proprietary 43B GPT and LLaMA2-70B. Perhaps surprisingly, we find that LLM with 4K context window using simple retrieval-augmentation at generation can achieve comparable performance to finetuned LLM with 16K context window via *positional interpolation* on long context tasks, while taking much less computation. More importantly, we demonstrate that retrieval can significantly improve the performance of LLMs regardless of their extended context window sizes. Our best model, retrieval-augmented LLaMA2-70B with 32K context window, outperforms GPT-3.5-turbo-16k and Davinci003 in terms of average score on seven long context tasks including question answering and query-based summarization. It also outperforms its non-retrieval LLaMA2-70B-32k baseline by a margin, while being much faster at generation. Our study provides general insights on the choice of retrieval-augmentation versus long context extension of LLM for practitioners.

1 INTRODUCTION

The long context large language models (LLM) have recently received a lot of attention in production (e.g., Anthropic, 2023; OpenAI, 2023b), research community (e.g., Chen et al., 2023; Liu et al., 2023; Tworkowski et al., 2023), and open source community (e.g., Kaiokendev, 2023). Although the *approximate* attention methods have been studied for years (e.g., Tay et al., 2022) (due to the quadratic time and memory complexities of self-attention mechanism in sequence length), the recent advance for long context LLMs with *exact* attention is mainly driven by the development of faster GPU with more memory and memory-efficient exact attention (Dao et al., 2022; Dao, 2023).

An alternative and long-standing solution for handling long context is *retrieval*. Specifically, the LLMs only read relevant context retrieved from a standalone retriever (e.g., Karpukhin et al., 2020; Wang et al., 2022; Lin et al., 2023), which is much easier to scale [1] and runs orders of magnitudes faster than LLMs for selecting relevant context. Conceptually, the retrieval-augmented decoder-only LLM can be viewed as applying the sparse attention over its long context window, where the sparsity pattern is not predefined as Child et al. (2019) but determined by the standalone retriever. In other words, unretrieved context is treated as irrelevant and has zero-valued attention weights.

Given the surge of interest in long context LLM research and much more required computation at inference [2], it is still unclear for practitioners whether extending the context window of LLM

[1]The dense embedding retriever can easily retrieve context from billions of tokens using the fast similarity search library (Johnson et al., 2019).

[2]For example, the price of GPT-4 with 32k context length is twice the 8k context model.

provides higher accuracy than the retrieval augmentation for downstream tasks with informative queries. Moreover, it would be compelling if we could combine the strength of both methods and achieve even higher accuracies. In this work, we attempt to answer the above questions through a comprehensive study.

Specifically, we make the following contributions:

1. We perform comprehensive study using two state-of-the-art LLMs, a proprietary 43B pretrained GPT and LLaMA2-70B (Touvron et al., 2023b) on 7 downstream long context tasks, including single and multi document question answering (QA) as well as query-based summarization.
2. We demonstrate that retrieval-augmentation significantly improves the performance of 4K context LLMs. Perhaps surprisingly, we find this simple retrieval-augmented baseline can perform comparable to 16K long context LLMs, i.e., average score 29.32 vs. 29.45 by using GPT-43B, and 36.02 vs. 36.78 by using LLaMA2-70B, while using much less computation.
3. Furthermore, we demonstrate that the performance of long context LLM (i.e., 16K or 32K) can still be improved by retrieval, especially for the larger LLaMA2-70B. As a result, our best model, retrieval augmented LLaMA2-70B-32k-ret with 32K context window (avg. score 43.6), outperforms GPT-3.5-turbo-16k (avg. score 42.8) and Davinci-003 in terms of average score. It also largely outperforms its non-retrieval LLaMA2-70B-32k baseline (avg. score 40.9), while can be much faster at generation (e.g., 4× faster on NarrativeQA).

We organize the rest of the paper as follows. We discuss related work in Section 2, and present the experimental setup in Section 3. We report results in Section 4 and conclude the paper in Section 5.

2 Related Work

In this section, we discuss the related work in long context LLM, efficient attention methods, and retrieval-augmented language models.

2.1 Long Context Large Language Models

Over the past few years, pretraining large language models (LLMs) with long context window becomes a viable solution thanks to faster GPU with more memory and memory-efficient exact attention (e.g., Dao et al., 2022). For example, the context window for pretrained LLM have been increased from 1024 of GPT-2 (Radford et al., 2019), 2048 of GPT-3 (Brown et al., 2020), 4096 of Llama 2 (Touvron et al., 2023b), to 8192 of GPT-4 (OpenAI, 2023a). However, further extending the context window in pretraining can be challenging, because, *i)* pretraining LLM from scratch with long context (e.g., >16K tokens) is very expensive due to the quadratic time and memory complexities of exact attention, and *ii)* most of documents in pretraining corpus (e.g., Common Crawl) are relatively short.

Most recently, researchers start to extend the context window of LLMs with continued training or fine-tuning (e.g., Kaiokendev, 2023; Nijkamp et al., 2023; Chen et al., 2023; Tworkowski et al., 2023; Mohtashami & Jaggi, 2023). Tworkowski et al. (2023) introduced LongLLaMA by fine-tuning the 3B and 7B OpenLLaMA checkpoints with contrastive training on 8K context length. Landmark attention (Mohtashami & Jaggi, 2023) extends the context length of LLaMA 7B from 4K to 32K by introducing "landmark tokens" to represent blocks of the context and fine-tuning the attention to use landmark tokens for selecting relevant blocks. Chen et al. (2023) and Kaiokendev (2023) introduced *positional interpolation* to extend the context window sizes of RoPE-based (Su et al., 2021) pretrained LLMs. In particular, Chen et al. (2023) demonstrates promising results on LLaMA 7B to 65B (Touvron et al., 2023a) with minimal fine-tuning effort (within 1000 steps). ALiBi (Press et al., 2021) extrapolates context window length by removing the positional embeddings while simply biasing the key-query attention scores with a linear penalty that is proportional to their distance, so one does not need finetuning for context window extrapolation. Ratner et al. (2023) chunks long context into multiple sub-windows and re-use the positional embeddings across these windows, thus can handle longer context without any further finetuning. In this work, we apply *positional interpolation* method to extend the 4K context window of a proprietary 43B pretrained LLM and LLaMA2-70B (Touvron et al., 2023b) to 16K and 32K, as they both use rotary position embedding at pretraining. In terms of evaluation, we focus on downstream task performance (e.g., Shaham et al., 2023; Bai et al., 2023) after instruction tuning (Wei et al., 2021).

There are other studies showing the interplay between retrieval-augmentation and long context LLM. Liu et al. (2023) performs the black-box evaluation for the long context capability of existing LLM products, including ChatGPT 3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023a), Claude (Anthropic, 2023), in retrieval-augmented setting, and identify the "lost in the middle" phenomenon in these models.

2.2 Efficient Attention Methods

In previous study, many approximate attention methods (Tay et al., 2022) have been introduced for dealing with the quadratic complexity of self-attention that becomes a computational bottleneck for long context. They can be grouped into the following categories: *i)* Sparse attention mechanisms with predefined sparsity patterns (e.g., Child et al., 2019; Parmar et al., 2018; Ho et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Zhu et al., 2021), *ii)* recurrence-based method (Dai et al., 2019; Bulatov et al., 2022), *iii)* low-rank projection attention (e.g., Wang et al., 2020; Xiong et al., 2021; Tay et al., 2021; Zhu et al., 2021), *iv)* memory-based mechanisms (e.g., Rae et al., 2020; Liu et al., 2018), *v)* similarity and clustering based methods (e.g., Kitaev et al., 2020; Tay et al., 2020; Roy et al., 2021). These approximate methods introduce inductive bias (e.g., predefined sparsity) that can fit well for specific domain, but may reduce model quality in general LLM training.

Most recently, FlashAttention (Dao et al., 2022; Dao, 2023) is introduced to speed up the exact attention computation by accounting for reads and writes between levels of GPU memory. FlashAttention is particularly useful for handling longer sequences.

2.3 Retrieval-augmented Language Models

Retrieval has been integrated into language models for years to improve perplexity (Borgeaud et al., 2022; Wang et al., 2023), factual accuracy (Nakano et al., 2021), downstream task accuracy (Guu et al., 2020; Izacard & Grave, 2021; Izacard et al., 2022; Lewis et al., 2020), and in-context learning capability (Huang et al., 2023). Combined with a standalone retriever (Karpukhin et al., 2020; Wang et al., 2022; Lin et al., 2023), retrieval-augmented LLM is well established for handling question answering with long document and in open-domain. In previous study, language models have been augmented with retrieval at inference (Khandelwal et al., 2019; Yogatama et al., 2021), fine-tuning (Izacard et al., 2022; Lewis et al., 2020; Guu et al., 2020), and pretraining (Borgeaud et al., 2022; Izacard et al., 2022; Wang et al., 2023). There are also methods that try to integrate LLM and retriever in a single model and build the end-to-end solution (e.g., Jiang et al., 2022; Shi et al., 2023). However, most of previous works mainly study retrieval-augmentation for LLMs that have around 10 billion parameters, except a few recent ones (e.g., Shi et al., 2023).

In this work, we focus on decoder-only LLMs with 43B and 70B parameters trained on trillions of tokens, because the LLMs at such scale exhibit strong zero-shot capability to incorporate context after instruction tuning (Wei et al., 2021; 2022).

2.4 Concurrent work

When we are preparing this manuscript, we notice that a concurrent work (Bai et al., 2023) (arXived on 28 Aug 2023) also studies the impact of retrieval on long context LLM, including black-box model GPT-3.5-Turbo-16k (OpenAI, 2022), white-box model Llama2-7B-chat-4k (Touvron et al., 2023b), and ChatGLM2-6B-32k (Zeng et al., 2022). Different from our findings, they find that retrieval is only helpful for Llama2-7B-chat-4k with 4K context window, but not helpful for long context model, i.e., GPT-3.5-Turbo-16k and ChatGLM2-6B-32k. We hypothesize the major reasons are: *i)* it is challenging to do controlled experiments using black-box APIs, *ii)* the white-box LLMs used in their study are relatively small, thus they have limited zero-shot capability of incorporating context through retrieval. Our conclusions are drawn from much larger LLMs. In particular, our best long context model LLaMA2-70B-32k performs as well as ChatGPT-3.5, while it can still be further enhanced by retrieval (see Table 3).

3 Experimental Setup

In this section, we present the details of our experimental setup.

3.1 Large Language Models

We focus on comparing the zero-shot capability of integrating long context information for generative QA or summarization tasks via retrieval or LLM's own self-attention mechanism. In contrast to most existing works that focus on relatively small models (e.g., 3B or 7B) (Kaiokendev, 2023; Nijkamp et al., 2023; Tworkowski et al., 2023; Mohtashami & Jaggi, 2023), we gather the insights by exploring model sizes that are larger than 40B after instruction tuning, as previous study suggests that instruction tuning becomes effective when the decoder-only LLM has around 50B parameters (Wei et al., 2021; 2022).

Specifically, we experimented with two pretrained GPT models, a proprietary Nemo GPT-43B and LLaMA2-70B. GPT-43B is a 43 billion parameter model that is trained with 1.1T tokens with 70% English corpus and the other 30% for multilingual and code data. For the English pretraining corpus, GPT-43B used Common Crawl web archive (WARC), Wikipedia, Reddit, Books, Gutenberg, ArXiv, StackExchange, PubMed, etc. It contains 48 layers with the hidden dimension of 8,192. It is trained with a sequence length of 4,096 and RoPE embeddings (Su et al., 2021). The other LLaMA2-70B is a public available 70B GPT model trained on 2T tokens using around 90% English data. It contains 80 layers with the hidden dimension of 8,192. It also has the context window size of 4,096 and trained with RoPE embeddings.

3.2 Datasets and Metrics

In this study, we include seven datasets ranging from single document QA, multi document QA, to query-based summarization for our zero shot evaluations. Specifically, we include four datasets from the validation set of the Scroll benchmark (Shaham et al., 2022).

- **QMSum (QM)** (Zhong et al., 2021) is a query-based summarization dataset, consisting of 232 meetings' transcripts and their corresponding summaries from multiple domains such as academic, industrial product. Annotators were tasked with writing queries basing on the contexts and ensuring that the relevant text for answering each query spans contains at least 200 words or 10 turns.
- **Qasper (QASP)** (Dasigi et al., 2021) is a question answering dataset over NLP papers filtered from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020). Qasper contains abstractive, extractive, and yes/no questions, as well as unanswerable ones.
- **NarrativeQA (NQA)** (Kočiský et al., 2018) is an established question answering dataset over entire books from Project Gutenberg[3] and movie scripts from a list of websites. Summaries of the books and scripts obtained from Wikipedia were given to the annotators to produce question-answer pairs, resulting in approximately 30 questions and answers for each of the 1,567 books and scripts. Each question was answered by providing two reference answers.
- **QuALITY (QLTY)** (Pang et al., 2022) is a multiple-choice question answering dataset over stories and articles sourced from several resources, such as Project Gutenberg and the Open American National Corpus[4]. 50% of the questions in QuALITY are labeled as *hard* to ensure the whole given document must be read slowly to conclude a correct answer, i.e., a skim of the document always yields wrong answers.

We take another three datasets from LongBench (Bai et al., 2023).

- **MuSiQue (MSQ)** (Trivedi et al., 2022) stands for Multihop Questions via Single-hop Question Composition aiming at multihop reasoning question answering. A bottom–up process of constructing multihop from single-hop questions allows systematic exploration of a large space of multihop candidates and greater control over which questions that are composed manually. In order to correctly generate the answers, LLMs require connected reasoning by reducing potential

[3] https://www.gutenberg.org/

[4] https://anc.org/

	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
# of samples	200	1,726	2,000	2,000	200	200	150
avg doc length	14,140	4,912	84,770	6,592	16,198	13,319	7,185
avg top-5 chunks	2,066	2,071	2,549	2,172	2,352	2,322	2,385
avg top-10 chunks	4,137	3,716	5,125	4,018	4,644	4,554	4,305
avg top-20 chunks	8,160	4,658	10,251	5,890	9,133	8,635	6,570

Table 1: Statistics of seven datasets used for zero-shot evaluation. All lengths are counted by the number of tokens using LLaMA2-70B tokenizer, and "avg top N chunks" denotes the average number of tokens from the top N retrieved chunks. Figure 1 gives more details.

reasoning shortcuts, minimizing train-test leakage, and including harder distractor contexts. Thus, MuSiQue is significantly less cheatable via disconnected reasoning than previous datasets.

- **HotpotQA (HQA)** (Yang et al., 2018) is a Wikipedia-based question-answer dataset with several key features. First, multiple supporting documents are required to be read for answering and reasoning. Second, the questions are diverse and not constrained to any pre-existing knowledge bases. Third, sentence-level supporting are provided with strong supervision to support LLM's requirement for reasoning. Finally, new types of factoid comparison questions are provided to test LLMs' ability to extract and compare various entity properties in text.
- **MultiFieldQA-en (MFQA)** (Bai et al., 2023) was manually curated to better test the model's long context understanding ability across diverse fields. Documents and articles from multiple sources, including legal documents, government reports, encyclopedias, and academic papers are collected. Ph.D. students were asked to annotate the questions and answers for each article. The evidences are fairly randomly located in the documents to avoid biases that might occur at the beginning or ending of the documents.

The full details of the dataset can be found in Table 1. We can see that our evaluation datasets have a wide range of average document length from 4.9k (QASP) to 84k (NQA). Therefore, for the baseline model without retrieval, we truncate the document accordingly to fit into the input sequence length.

Following the official metrics, we report the geometric mean of ROUGE scores (i.e., ROUGE-1/2/L) (Lin, 2004) for QM, the exact matching (EM) score for QLTY, and F1 scores for the remaining five datasets QASP, NQA, MSQ, HQA and MFQA.

3.3 Context Window Extension

We extend the context window length with position interpolation method (Chen et al., 2023), as it is simple and effective for RoPE embeddings. We extend the 4K context window to 16K for GPT-43B. For LLaMA2-70B, we extend its 4K context window to 16K and 32K. We follow Chen et al. (2023) and finetune both LLMs on the Pile dataset (Gao et al., 2021) with batch size as 128, constant learning rate of 5e-6 to adapt the position embeddings.

3.4 Retrieval

For the retriever, we experimented with three retrievers: 1) *Dragon* (Lin et al., 2023) as it achieves state-of-the-art results on both supervised and zero-shot information retrieval benchmarks (Thakur et al., 2021). Dragon is a dual encoder model that consists of a query encoder and a context encoder. 2) a widely used *Contriever* model (Izacard et al., 2021). Following the MoCo technique (He et al., 2020), Contriever used a simple contrastive learning framework to pre-train models for information retrieval. It was trained without supervision and achieved competitive results with BM25 for R@100 on the BEIR benchmark (Thakur et al., 2021), and 3) *OpenAI embedding*[5]. For the OpenAI embedding model, we use the latest "text-embedding-ada-002" as recommended by OpenAI. It accepts 8,191 maximum input tokens for one sequence with an output vector of 1,536 dimensions. The cosine similarities are then computed between the questions and the list of contexts for retrieval ranking.

[5]https://platform.openai.com/docs/guides/embeddings

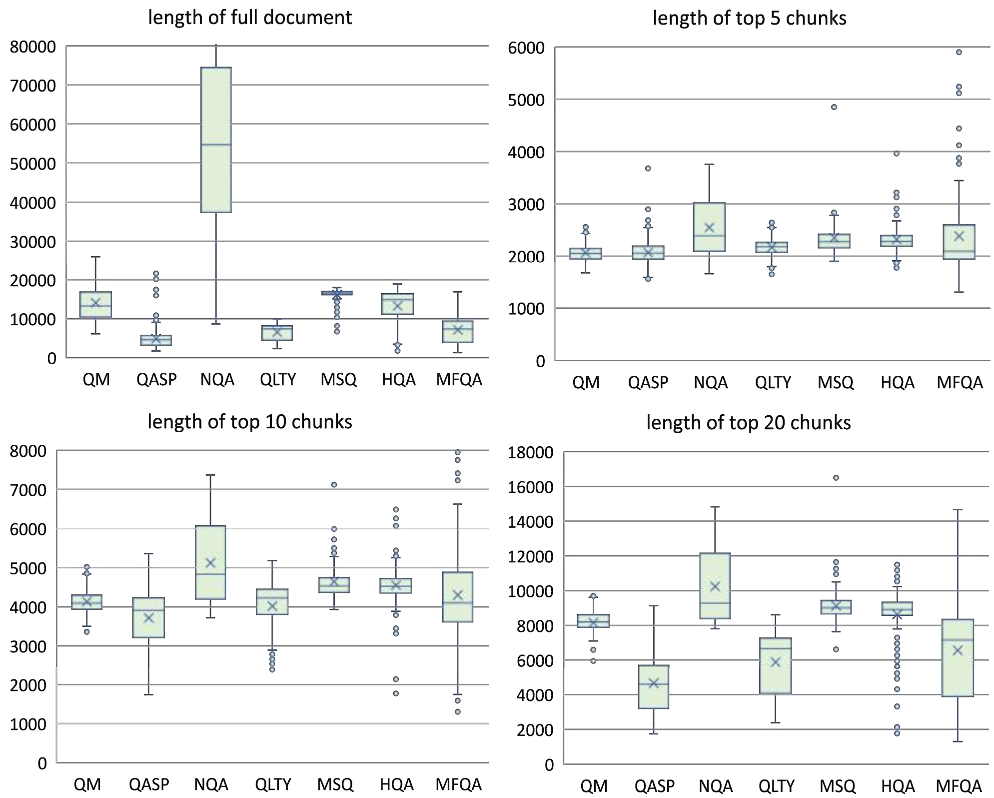

Figure 1: Token length distribution of the full document and the top-5, 10, 20 chunks of the seven datasets.

To use these retrievers, we first chunk each context document with 300 words, and then we encode both the questions and all chunks independently with corresponding encoders. The most relevant N chunks, ranked by the dot product of the question embedding and chunk embedding, are then concatenated together (following the left to right order from the most relevant to least relevant) as the context of the prompt for generation. Table 1 shows the statistics of the top N retrieved chunks while Figure 1 gives more details of the token length distribution of all seven datasets. Note that, some dataset like Qasper (QASP) is relatively short and don't have up to 20 chunks, so the average length of top-10 chunks and top-20 chunks are close. We can see that top-5 chunks can all fit into 4k sequence length (except few outliers) while top-10 and top-20 chunks can fit into 16k sequence length.

3.5 Instruction Tuning

To train the pretrained LLMs to follow instructions for question answering or text summarization, we also performed instruction tuning. We first construct a blend of instruction tuning datasets consisting of 102K training samples from the Soda dataset (Kim et al., 2022), ELI5 dataset (Fan et al., 2019), FLAN dataset (Wei et al., 2021) , Open Assistatant dataset (Köpf et al., 2023), Dolly (Conover et al., 2023) and a proprietary sourced conversational dataset, to adapt both GPT-43B and LLaMA2-70B to follow instructions. In terms of the template, we use "System: {System}\n\nUser: {Question}\n\nAssistant: {Answer}" as the format to support multi-turn dialogue training. As all of the tasks contain the context information for reasoning over at inference time, we add the context before the dialogue, i.e. "System: {System}\n\n{Context}\n\nUser: {Question}\n\nAssistant: {Answer}".

We finetune the LLM by taking the loss only on the {Answer} part with batch size 128 and learning rate of 5e-6 for 1000 steps. For the rest of the paper, results are all reported using the instruction tuned chat model on top of the foundational GPT-43B and LLaMA2-70B.

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
LLaMA2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
LLaMA2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
LLaMA2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**

Table 2: Comparison of model variants (GPT-43B, LLaMA2-70B) with sequence length ranging from 4k to 32k under seven datasets. "ret" denotes using the best retriever (Dragon or Contriever) and here we used top-5 for the retriever.

4 RESULTS

In this section, we report the results and provide detailed analysis.

4.1 MAIN RESULTS

In Table 2, we compare different model variants with context lengths ranging from 4K to as long as 32K using GPT-43B and LLaMA2-70B. First, we find that baseline models without retrieval of 4k sequence length achieve the worst results for both GPT-43B and LLaMA2-70B. This is because the minimum average sequence length of all seven tasks exceeds 4096, the context window of the foundation models and therefore valuable texts get truncated randomly. As a result, retrieval is especially helpful for 4K LLMs e.g., LLaMA2-70B-4K is improved from 31.61 to 35.73 while GPT-43B-4K is improved from 26.44 to 29.32. Second, we observe that HotpotQA (HQA) especially favors long sequence models as the score improves from 34.64 to 43.97 for LLaMA2-70B and from 28.91 to 37.48 for GPT-43B when the sequence length increases from 4k to 16k. This is because Hotpot QA is a multi-hop dataset where the questions are not hard to answer but all intermediate hops are necessary to get correct answer. Therefore, long context are beneficial to increase the recall of incorporating all intermediate hops.

It is quite interesting that the retrieval-augmented long context LLM (e.g., 16K and 32K) can obtain better results than retrieval-augmented 4K context LLM, even they are feed with the same top 5 chunks of evidence. We hypothesize this interesting observation is related to the "lost in the middle" phenomenon (Liu et al., 2023), where the LLMs has such "U-shaped" performance curve. Specifically, LLMs are better at utilizing relevant information that occurs at the beginning or end of its input context window. Due to this reason, the 4K context LLM tends to ignore the information in the middle of 4K input, while 32K context LLM tend to ignore the information in the middle of 32K input. From Figure 1, the length of top 5 chunks is about 2K tokens, which can be in the middle and ignored by 4K context LLM, but is only at the beginning part of 16K and 32K context and may not be ignored by the 16K or 32K context LLM.

Note that, we have very different observation from the conclusion drawn from LongBench work (Bai et al., 2023): *"Retrieval brings improvement for model with weak ability on long contexts, but the performance still lags behind models that have strong long context understanding capability"*. Here, we demonstrate retrieval can significantly improve the performance of both GPT-43B and LLaMA2-70B regardless their context window size. For example, our best retrieval-augmented LLaMA2-70B-32k-ret outperforms its baseline w/o retrieval by a margin, i.e., 39.60 vs. 37.36. We think the major reason for such different conclusion is that Bai et al. (2023) uses much smaller LLM with 6B and 7B parameters, which usually has relatively worse zero-shot capability to incorporate the retrieved chunked context. In contrast, the larger instruction tuned LLMs like LLaMA2-70B has much stronger zero-shot capability to incorporate retrieved evidence. This observation is becoming more clear when one compares the gain of retrieval-augmentation between GPT-43B and LLaMA2-70B, where LLaMA2-70B enjoys larger benefit of incorporating context through retrieval.

Model	Avg-7	Avg-4*	QM*	QASP*	NQA*	QLTY*	MSQ	HQA	MFQA
Davinci003 (175B)	-	40.8*	16.9*	52.7*	24.6*	69.0*	-	-	-
GPT-3.5-turbo (4k)	-	39.2*	15.6*	49.3*	25.1*	66.6*			
GPT-3.5-turbo-16k	42.8	42.4	17.6	50.5	28.8	72.6	26.9	51.6	52.3
LLaMA2-70B-32k	40.9	42.4	15.6	45.9	28.4	79.6	19.1	49.5	48.4
LLaMA2-70B-32k-ret	**43.6**	**43.0**	18.5	46.3	31.5	75.6	26.7	53.9	52.9

Table 3: Comparison of our best retrieval-augmented LLaMA2-70B-32k-ret with GPT-3.5-turbo-16k and Davinci-003 (175B parameters). For QMSum (QM), Qasper (QASP), NarrativeQA (NQA), QuALITY (QLTY), we used the test set from the ZeroSCROLLS leaderboard as the organizers have prepared the scores of GPT-3.5-turbo (4k) and Davinci-003 (highlighted with *). Avg-7 refers to the average score of all 7 datasets, and Avg-4* refers to the average of 4 datasets from ZeroSCROLLS.

Seq len	Setting	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
4k	baseline (w/o ret)	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
	Dragon	35.73	18.14	29.20	23.39	70.30	20.09	41.54	47.45
	Contriever	**36.02**	17.41	28.74	23.41	70.15	21.39	42.06	48.96
	OpenAI-embedding	35.79	17.76	28.85	23.57	70.70	19.92	41.76	47.99
32k	baseline (w/o ret)	37.36	15.37	31.88	23.59	73.80	19.07	49.49	48.35
	Dragon	**39.60**	18.34	31.27	24.53	69.55	26.72	53.89	52.91
	Contriever	38.85	17.60	31.56	23.88	69.00	26.61	49.65	53.66
	OpenAI-embedding	39.34	18.24	32.07	24.36	69.45	24.90	51.64	54.75

Table 4: Comparisons of adding top 5 retrieved chunks from different retrievers to the context under LLaMA2-70B. Public available retriever can be better than OpenAI-embedding.

Seq len	Setting	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
4k	base	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
	top-5	**35.73**	18.14	29.20	23.39	70.30	20.09	41.54	47.45
	top-10	34.62	16.54	28.67	24.38	68.70	19.00	42.18	42.84
	top-20	34.61	16.52	28.67	24.38	68.70	19.00	42.18	42.84
16k	base	36.78	16.72	30.92	22.32	76.10	18.78	43.97	48.63
	top-5	37.23	18.70	29.54	23.12	70.90	23.28	44.81	50.24
	top-10	**38.31**	18.41	30.20	25.53	73.60	22.78	47.72	49.91
	top-20	36.61	17.26	29.60	25.81	72.30	22.69	41.36	47.23
32k	base	37.36	15.37	31.88	23.59	73.80	19.07	49.49	48.35
	top-5	**39.60**	18.34	31.27	24.53	69.55	26.72	53.89	52.91
	top-10	38.98	17.71	30.34	25.94	70.45	22.80	55.73	49.88
	top-20	38.38	16.36	30.42	24.42	69.60	24.51	54.67	48.65

Table 5: Comparisons of adding top-5/10/20 retrieved chunks to the context under 4k, 16k, and 32k input sequence lengths using LLaMA2-70B. More context does not always give better results.

4.2 Comparing to OpenAI models

To further understand how good is our best model, i.e., augmenting LLaMA2-70B-32k with retrieval, we also compare it to GPT-3.5-turbo(4k), GPT-3.5-turbo-16k and Davinci-003 on those seven datasets.[6] We found that LLaMA2-70B-32k-ret achieves better results than GPT-3.5-turbo-16k in terms of the average accuracy over seven datasets, while better than Davinci-003 (w/ 175B parameters) on the average over 4 tasks. This indicates LLaMA2-70B-32k with retrieval is a strong model for these long context tasks, and our conclusion is built on the state-of-the-art results.

[6]For QMSum (QM), Qasper (QASP), NarrativeQA (NQA), QuALITY (QLTY), we used the test set from the ZeroSCROLLS leaderboard as the organizers have prepared the scores of GPT-3.5-turbo (4k) and Davinci-003 there.

4.3 Ablation on Different Retrievers

To investigate the impacts of different retrievers on top of LLaMA2-70B, we compare Dragon, Contriever, and OpenAI embeddings on top of LLaMA2-70B-4k and LLaMA2-70B-32k. The results in Table 4 confirms that our finding, i.e., *retrieval can boost the performance of both short context and long context LLMs*, is consistent across different retrievers. Also, we observe that public available retrievers can do better than the commercially closed OpenAI embeddings.

4.4 Increasing the number of retrieved chunks

To study the impact of adding more retrieved chunks to the context, we increase the number of retrieved chunks from 5 to 20 using Dragon retriever and the results can be found in Table 5. We observe that for different sequence lengths, the best averaged results are obtained either from top 5 or top 10. Even if 20 chunks can still fit into the 16K and 32K context window (as shown in Figure 1), adding more chunks up to 20 is not helpful and will sometime hurt the performance. We believe this is related to the "lost in the middle" phenomenon (Liu et al., 2023) or the model is getting distracted by irrelevant information and therefore needs further research.

5 Conclusion

In this work, we systematically study the retrieval-augmentation versus long context extension using the state-of-the-art LLMs after instruction tuning for various long context QA and query-based summarization tasks. After study, we have the following interesting findings: *i)* Retrieval largely boosts the performance of both 4K short context LLM and 16K/32K long context LLMs. *ii)* The 4K context LLMs with simple retrieval augmentation can perform comparable to 16K long context LLMs, while being more efficient at inference. *iii)* After context window extension and retrieval-augmentation, the best model LLaMA2-70B-32k-ret can outperform GPT-3.5-turbo-16k and Davinci003 in terms of average score on a suit of downstream tasks with informative queries. Our study shed light on the promising direction of combining retrieval and long context techniques together to build better LLM.

References

Anthropic. Introducing 100k context windows. https://www.anthropic.com/index/100k-context-windows, 2023.

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

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *ICML*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *NeurIPS*, 2022.

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

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 2022.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. REALM: Retrieval augmented language model pre-training. In *ICML*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023.

Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, 2021.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021. URL https://arxiv.org/abs/2112.09118.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Zhengbao Jiang, Luyu Gao, Jun Araki, Haibo Ding, Zhiruo Wang, Jamie Callan, and Graham Neubig. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. In *EMNLP*, 2022.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Kaiokendev. Things I'm learning while training SuperHOT. https://kaiokendev.github.io/til#extending-context-to-8k, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*, 2022.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

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, 05 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00023. URL https://doi.org/10.1162/tacl_a_00023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 2020.

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.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*, 2023.

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. *arXiv preprint arXiv:2307.03172*, 2023.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *ICLR*, 2018.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447.

Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Erik Nijkamp, Hiroaki Hayashi, Tian Xie, Congying Xia, Bo Pang, Congying Xia, and et al. Long sequence modeling with XGen: A 7b LLM trained on 8k input sequence length. https://blog.salesforceairesearch.com/xgen/, 2023.

OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

OpenAI. Gpt-4. https://openai.com/research/gpt-4, 2023a.

OpenAI. Function calling and other API updates (longer context). https://openai.com/blog/function-calling-and-other-api-updates, 2023b.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.391. URL https://aclanthology.org/2022.naacl-main.391.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, pp. 4055–4064, 2018.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive Transformers for long-range sequence modelling. In *ICLR*, 2020.

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 *ACL*, 2023.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 2021.

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. *arXiv preprint arXiv:2305.14196*, 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. RePlug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. In *ICML*, 2020.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *ICML*, 2021.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 2022.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *arXiv preprint arXiv:2307.03170*, 2023.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, et al. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. *arXiv preprint arXiv:2304.06762*, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *AAAI*, 2021.

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 *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.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373, 2021.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for longer sequences. In *NeurIPS*, 2020.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472.

Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision. *NeurIPS*, 2021.