Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP

Omer Goldman*, **Alon Jacovi***, **Aviv Slobodkin***,
Aviya Maimon*, **Ido Dagan**, **Reut Tsarfaty**
Bar-Ilan University
omer.goldman@gmail.com

Abstract

Improvements in language models' capabilities have pushed their applications towards longer contexts, making long-context evaluation and development an active research area. However, many disparate use cases are grouped together under the umbrella term of "long-context", defined simply by the total length of the model's input, including – for example – Needle-in-a-Haystack tasks, book summarization, and information aggregation. Given their varied difficulty, in this position paper we argue that conflating different tasks by their context length is unproductive. As a community, we require a more precise vocabulary to understand what makes long-context tasks similar or different. We propose to unpack the taxonomy of long-context based on the *properties that make them more difficult* with longer contexts. We propose two orthogonal axes of difficulty: (I) *Dispersion:* How hard is it to find the necessary information in the context? (II) *Scope:* How much necessary information is there to find? We survey the literature on long context, provide justification for this taxonomy as an informative descriptor, and situate the literature with respect to it. We conclude that the most difficult and interesting settings, whose necessary information is very long and highly dispersed within the input, is severely under-explored. By using a descriptive vocabulary and discussing the relevant properties of difficulty in long context, we can implement more informed research in this area. We call for a careful design of tasks and benchmarks with *distinctly* long context, taking into account the characteristics that make it qualitatively different from shorter context.

1 Introduction

The ability to deal with ever-longer contexts has been one of the most notable trends among the emerging capabilities of large language models (LLMs). Starting with a few hundred tokens as the maximal input length of the first attention-based LLMs (Devlin et al., 2019; Raffel et al., 2020), contemporary models are – *technically* – able to process up to 128k and even 1M tokens (Gemini Team Google, 2024; OpenAI, 2024).

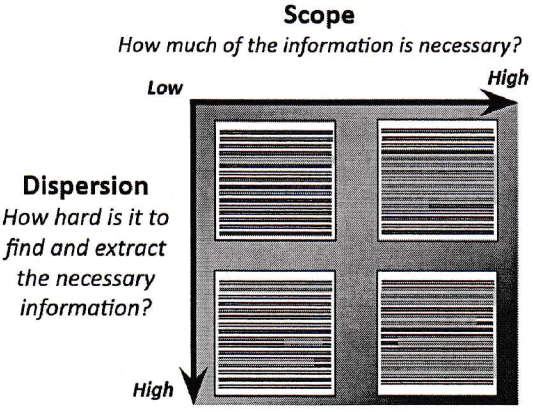

Figure 1: A taxonomy of long context tasks based on the distribution of the needed information in the text. Tasks with larger scope and higher dispersion are more difficult (indicated by shade) and more indicative of the long context capabilities of large language models.

The demand to evaluate LLMs in this setting has led to a line of research on designing long-context tasks and benchmarks, in order to systematically understand models' capabilities and drive their development. However, the field has generally a sole recurring descriptor to define such measurements by – simply, the length of the context. For example, long-context benchmarks group tasks mostly by length in words (e.g., Shaham et al., 2022; Bai et al., 2023; Zhang et al., 2024b). This leads to qualitatively different measurements being conflated together, with conclusions about long-context capabilities being extended from one class of tasks to others. The community is, of course, aware that, for example, tasks which require a small part of the input are different from tasks that require a large part of it. But we ask the more general ques-

*Equal contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16576–16586
November 12-16, 2024

tion: What are the properties that differentiate tasks when conditioned on their context length? What can we accomplish with such a distinction?

In this position paper, we claim that the current landscape of works on long-context evaluation will greatly benefit from a more fine-grained characterization of long-context task design. We argue that judging LLMs by their ability to process long sequences, while disregarding the task they process them for, overlooks the characteristics that make longer inputs more difficult, and interesting to research, to begin with (§2).

For example, Needle in a Haystack tasks (NIAH; Ivgi et al., 2023; Mohtashami and Jaggi, 2023) involve queries whose main challenge is finding the relevant information in a long context, without requiring much further processing. Synthetic NIAH datasets are, of course, easier than their natural equivalents (Ivgi et al., 2023), but the "natural vs. artificial" classification is not informative in our setting, since it applies equally for tasks regardless of context length. What, then, is an informative property? What makes long-context tasks different from each other? For example, multiple-needle variants of NIAH (Hsieh et al., 2024), or those that position the "needles" closer or farther apart (Levy et al., 2024). Evidently, "the number of tokens in the input" is not a sufficient descriptor.

To resolve this roadblock, we present a taxonomy of long-context tasks for the different factors that make them harder *when controlling for context length* (§3). This taxonomy is derived by surveying the long-context literature and surfacing the most salient points of distinction between various tasks. We focus on (I) how difficult it is to find and extract the required information from the input (its *dispersion* in the input), and (II) the absolute quantity of required information to solve the task (its *scope*). See Figure 1 for a summary.

To understand this categorization and its utility, we review the literature on long-context evaluation and position the works with respect to those factors. We find that the most challenging setting, in which a large quantity of required information is present in a dispersed manner that is difficult to extract, is significantly under-explored (§4).

Finally, acknowledging the inherent and legitimate reasons behind the focus on context length as the sole descriptor of difficulty, we discuss possible paths forward for designing more reliable measurements of long-context capabilities when utilizing a more nuanced vocabulary (§5).

2 Task Design in Long Context

Evaluating the performance of NLP models over very long contexts is a fast-changing area of research (Bishop et al., 2024; Wu et al., 2024). Measurements are regularly updated to account for new capabilities which improve with extrapolation architectures (Vaswani et al., 2017; Su et al., 2024) and training data (He et al., 2023; Chen et al., 2023). Evaluators were tasked with designing measurements of long-context capabilities cheaply, efficiently, and quickly, while matching realistic use cases as much as possible. The most common way of differentiating long-context tasks, besides the context's length, is whether they are naturally-constructed or synthetically-constructed (Tay et al., 2020; Bai et al., 2023; Hsieh et al., 2024).

Natural construction. A simple yet effective way of "moving the goalpost" for context length is by modeling long-context tasks based on short-context tasks. This was done, for example, with QA (Kočiský et al., 2018, cf. Dunn et al., 2017), summarization (Huang et al., 2021a, cf. Narayan et al., 2018), and NLI (Koreeda and Manning, 2021a, cf. Williams et al., 2018). Specialized domains like legal (Bruno and Roth, 2022; Nguyen et al., 2024) and literature (Wang et al., 2022; Kryscinski et al., 2022) often involve longer texts, turning typically short-context tasks such as QA and NLI into long-context scenarios. Another more native methodology is to create new tasks which inherently require a long context, such as multi-document summarization (Fabbri et al., 2019; Angelidis et al., 2021), survey generation (Gao et al., 2024), and structured data aggregation (Caciularu et al., 2024). Both methodologies share the constraint that, due to their natural construction (i.e., using natural text), once created, they are difficult to modify for longer contexts as models' long-context capabilities improve.

Synthetic construction. A more flexible approach, sacrificing natural construction for length control, is to use distractors to synthetically increase the context length. This method allows for cheap and efficient (in terms of task construction cost) evaluation of models' full context length capabilities, with difficulty adjusted by controlling the distractors. Tasks like Needle-in-a-Haystack (NIAH; Ivgi et al., 2023; Kamradt, 2023) and PassKey retrieval (Mohtashami and Jaggi, 2023) were created to evaluate a model's ability to pinpoint specific information amid lengthy distrac-

tors. Flexible and effective against existing models, they became standard benchmarks for evaluating new long-context models (GLM Team, 2024; Jiang et al., 2024). Followup studies have complicated these tasks by increasing the number of critical details to locate (Arora et al., 2023; Liu et al., 2024a) and changing their position within the input (Liu et al., 2024b; Levy et al., 2024).

Limitations of the status quo. NIAH-like tasks aim to assess information retrieval capabilities, yet many "naturally constructed" QA and reading-comprehension tasks with trivial questions about a long context accomplish the same goal. At the same time, "multiple needles" NIAH can increase difficulty not by increasing the quantity of needles or length of input, but by *adding distractors* between needles (Levy et al., 2024). What can systematically explain the different variables at play, in order to inform better task design in the future?

Clearly, there are *underlying qualitative differences* that discriminate between these various tasks besides their natural and synthetic constructions, and besides their actual context length. Therefore, we require a more informative vocabulary to discuss the goals of each task design, what it accomplishes, and what it does not, in terms of measuring long-context capabilities.

3 What Makes Long Context More than Retrieval?

We require a taxonomy to capture task difficulty variations beyond mere "number of tokens". We focus on the information that is canonically *required* to solve the task as the conditioning variable. Our classification can be summarized via the following two questions, when asked about a given task:

(I) *How difficult is it to find and extract the required information?*

(II) *How much information is needed to be found?*

Assuming that some highlighting of the relevant information is needed to solve the task (see Figure 1), the latter question asks how much text is highlighted, while the former addresses the challenge of locating the relevant text for highlighting.

For instance, consider the task of summarizing a book, in comparison to a NIAH task of identifying a numerical detail in a long financial report (e.g., "how much did the company earn in 2015?"). Although both tasks involve long texts, the *information required* and its *accessibility* vary significantly. The NIAH task focuses on localized, identifiable information, while summarization requires extracting key details dispersed throughout the text, tangled together with irrelevant content. Therefore, we can say that the book summarization task is more difficult on both axes (I) and (II).

Below we give more formal descriptions of the two axes characterized by the questions above.

(I) Dispersion. Although the question above intuitively defines "difficulty of information finding", we offer a more concrete description. Between two similar tasks, we consider the information harder to find in one task compared to another if: (1) it is more obscured (e.g., linguistically, semantically, contextually, etc); (2) it is more sparse, such that it is interspersed with non-required information; (3) its indicators are less redundant, such that there are fewer places in the document where the same information is available.

(II) Scope. The property of scope is simpler, and refers to the minimal quantity of information needed to solve the task. Importantly, we are not concerned with precise metric for "quantity of information" at this stage – it can refer to quantity of tokens, sentences, relations, cells in a table, etc. Any metric that reliably captures difficulty for an established solver is sufficient for our purposes.

Illustrative example. To illustrate, consider the Wikipedia entry for *New York City* and a simple question: "What is the estimated population of the city?" Since the answer needs a small snippet of information, we say that the task has *small scope*. And since it is easily accessible, we say that it has *low dispersion*. Consider, instead, the question "how many syllables are in this document?" – since this question requires the entire document to answer, we say that it has *large scope*, but if we consider counting syllables as straightforward, then we say its *dispersion* is still *low*. Finally, with the question "Was the city's mayor elected before or after the city was affected by Hurricane Sandy?" – since it requires snippets from at least two different areas of the text, we can say that when compared to the question about the city's population, the *dispersion* is *higher*, but not as high as for the question "What makes the city a prominent place on the world stage?" which poses a challenge on both axes.

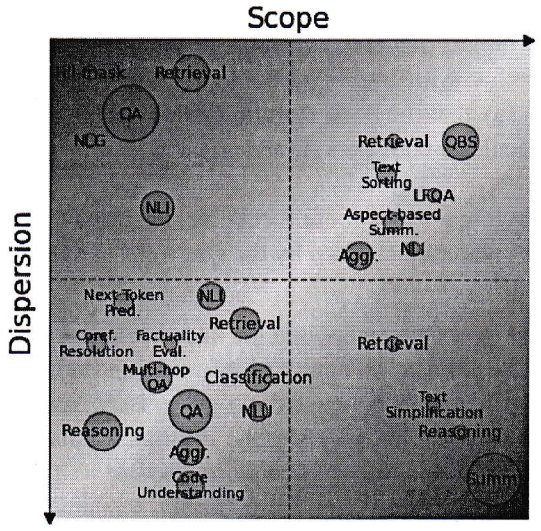

Figure 2: This figure illustrates our subjective judgment on the distribution of long-context benchmarks for each task, categorized by their scope and dispersion characteristics, with the four quadrants being marked by the dashed lines. Difficulty is expressed by shade, where red is more difficult and green in easier. Notably, some tasks, like Question-answering (QA), appear in multiple quadrants, as different benchmarks demand varying levels of scope and dispersion (e.g., a single fact versus multiple facts spread across a document). For a detailed breakdown of benchmarks and their task associations, refer to Appendix A.

4 Challenging Long Context Is Under-Explored

Revisiting the works surveyed in §2, they clearly differ with respect to both scope and dispersion.

With respect to *dispersion*. The information needed for tasks ranges from easily accessible to highly dispersed and difficult to detect. On low dispersion we have NIAH (Kamradt, 2023; Mohtashami and Jaggi, 2023) and a myriad of factual single-hop QA datasets (Tseng et al., 2016; Kočiský et al., 2017; Kwiatkowski et al., 2019; Dasigi et al., 2021, inter alia) in which the answer is relatively accessible. Adding more snippets of information separated by distractors, either in the form of several needles (Arora et al., 2023; Hsieh et al., 2024) or of hops in a multi-hop question (Trivedi et al., 2022; Zhao et al., 2022), complicates the information detection due to the need to find at least two snippets (Levy et al., 2024), thereby increasing dispersion. Dispersion can also be increased by making the detection of the information less straightforward (e.g., Pang et al., 2022) or requiring aggregation (Shaham et al., 2023). Lastly, summarization tasks are of a very high dispersion (Huang et al., 2021a; Wang et al., 2022), as they require the non-trivial detection of salient facts that are interwoven with the irrelevant text.

With respect to *scope*. Tasks overwhelmingly target relatively small scope. In addition to the aforementioned NIAH tasks and their variants, many QA datasets apply as well (Li et al., 2023; Zhao et al., 2023; Reddy et al., 2024, inter alia). A somewhat higher scope is achieved by datasets for query-based summarization (Zhong et al., 2021; Wang et al., 2022), and QA datasets with more obfuscated answers that require reading the text surrounding the answer for its verification (An et al., 2023; He et al., 2023). Although much higher on the scope ladder, book summarization is still limited in its scope as datasets include texts that are only of up to $20k$ tokens (Huang et al., 2021a; Chen et al., 2022a; Shaham et al., 2023). Currently, tasks with the highest scope, requiring information across the entire input length, are artificial and of low dispersion, like common words extraction (Hsieh et al., 2024).

Conclusion. Figure 2 summarizes the above classification of tasks and datasets. Note that without a concrete definition of dispersion and scope, the plot is only an illustration that involves a good deal of subjective judgements. However, we conclude that (1) the majority of tasks designed to challenge LLMs in a long-context setting target either scope *or* dispersion, such that (2) tasks that push current models' capabilities on *both* axes are under-represented in the current landscape.

5 Discussion: Towards Genuinely Difficult Long-Context Task Design

Challenges. Designing meaningful long-context tasks amidst rapid model progress is profoundly challenging, making the deficiency in tasks representing difficulty on both the *dispersion* and *scope* axes unsurprising. One source of this challenge is the lack of diverse, coherent long texts, as models' context windows can now be comparable to the length of the New Testament[1] and the Odyssey.[2] The methodologies discussed in §2 for creating long context tasks – lengthening short context tasks and synthetically creating length-adjustable tasks

[1] www.readinglength.com/book/isbn-0190909005
[2] www.readinglength.com/book/isbn-0140268863

– are preferred for their straightforward definition and the incremental adjustments they require for existing data. They rely on the common understanding of machine comprehension as formulated with short context in mind (Dunietz et al., 2020), and therefore they are intuitive and easy to comprehend for NLP researchers without domain expertise (e.g., in law or biomedical fields that have long contexts).

Future work. The goals laid forward in this work are clear: For more durable and robust measurement of long-context capabilities, we must design tasks that explicitly target both the *dispersion* and *scope* capabilities of models. How can this be achieved? As mentioned, one possible avenue is to rely more on *tasks that require domain expertise*, such as legal documents (Bruno and Roth, 2022), financial reports (Reddy et al., 2024), biomedical publications (Stylianou et al., 2021), and so on. In specialized domains, it is common that *dispersion* will be naturally higher (Zhao et al., 2022). Tasks that involve *implicit aggregations over structured data*, such as table manipulation (Caciularu et al., 2024), are possible avenues for increasing both scope and dispersion synthetically by leveraging the data structure. In this work, we argue that an *explicit vocabulary* for such properties of difficulty is what can enable more informed techniques to design difficult tasks in the future.

6 Conclusions

We present a taxonomy of factors that make long-context tasks more challenging compared to short ones. This is in contrast with the existing literature that refers only to the length of the input as the hallmark of long context, and as a result ends up conflating tasks of different character when assessing the ability of models to understand longer text. We reviewed works on evaluation in a long-context setting and found that the most challenging setting, in which the information needed is of large *scope* and is highly *dispersed* within the input, is under-explored. Finally, we suggested some leads for future work to tackle this imbalance towards a more informative long context evaluation.

7 Limitations

Formality. In the context of this work, we have deliberately adhered to a taxonomy based on an intuitive description, in the interest of utility to a wide diversity of research and flexibility for future work. Difficulty in searching for and extracting information, and quantity of information, are both vague terms that can only be grounded in the context of a specific family of tasks and use-cases. We intend for this work to serve as a call to action and a tool for a shared vocabulary in the interest of more informed long-context task design in the future, rather than to anchor the taxonomy to a specific and fragile point in time.

Retrieval is still interesting. Although we argue that small scope and low dispersion tasks are the least indicative of the model's ability to long-context capabilities, tasks that are well-served by implicit retrieval or by traditional retrieval-based pipelines are certainly relevant and useful in a variety of common use-cases (Stylianou et al., 2021; Bruno and Roth, 2022; Gao et al., 2023).

Other uses for a long-context window. This paper deals only with long inputs that serve as inputs to a task. The long context of course can have other purposes as well, like containing many in-context examples (Bertsch et al., 2024) or containing other modalities and structures (Jiang et al., 2023).

Acknowledgments

The authors would like to thank Gabriel Stanovsky for the fruitful discussions.

References

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. OpenAsp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991, Singapore. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *Preprint*, arXiv:2307.11088.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. 2023. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

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. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *Preprint*, arXiv:2405.00200.

Jennifer A Bishop, Qianqian Xie, and Sophia Ananiadou. 2024. Longdocfactscore: Evaluating the factuality of long document abstractive summarisation. *Preprint*, arXiv:2309.12455.

Odellia Boni, Guy Feigenblat, Guy Lev, Michal Shmueli-Scheuer, Benjamin Sznajder, and David Konopnicki. 2021. Howsumm: A multi-document summarization dataset derived from wikihow articles. *Preprint*, arXiv:2110.03179.

William Bruno and Dan Roth. 2022. Lawngnli: A long-premise benchmark for in-domain generalization from short to long contexts and for implication-based retrieval. *Preprint*, arXiv:2212.03222.

Avi Caciularu, Alon Jacovi, Eyal Ben-David, Sasha Goldshtein, Tal Schuster, Jonathan Herzig, Gal Elidan, and Amir Globerson. 2024. Tact: Advancing complex aggregative reasoning with information extraction tools. *Preprint*, arXiv:2406.03618.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022a. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022b. Summscreen: A dataset for abstractive screenplay summarization. *Preprint*, arXiv:2104.07091.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *ArXiv*, abs/2309.12307.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Preprint*, arXiv:1804.05685.

Anze Xie Ying Sheng Lianmin Zheng Joseph E. Gonzalez Ion Stoica Xuezhe Ma Dacheng Li*, Rulin Shao* and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. 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*, pages 4599–4610, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. *Preprint*, arXiv:1906.01749.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Dairui Liu, Tianwei She, Yuang Jiang, and Irene Li. 2024. Large language models on wikipedia-style survey generation: an evaluation in nlp concepts. *Preprint*, arXiv:2308.10410.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. *Preprint*, arXiv:2210.08726.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

GLM Team. 2024. GLM-4-9b-chat technical report.

Daya Guo, Canwen Xu, Nan Duan, Jian Yin, and Julian McAuley. 2023. Longcoder: A long-range pre-trained language model for code completion. *Preprint*, arXiv:2306.14893.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Yuxin Liang, Hao Wang, Qianguo Sun, Songxin Zhang, Zejian Xie, and Jiaxing Zhang. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. *Preprint*, arXiv:2311.09198.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *Preprint*, arXiv:2103.06268.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

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

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021b. Efficient attentions for long document summarization. *Preprint*, arXiv:2104.02112.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing LLMs. GitHub.

Yuta Koreeda and Christopher Manning. 2021a. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuta Koreeda and Christopher D. Manning. 2021b. Contractnli: A dataset for document-level natural language inference for contracts. *Preprint*, arXiv:2110.01799.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Preprint*, arXiv:1712.07040.

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

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. *Preprint*, arXiv:2105.08209.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *Preprint*, arXiv:2010.12694.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. In search of needles in a 11m haystack: Recurrent memory finds what llms miss. *Preprint*, arXiv:2402.10790.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *Preprint*, arXiv:2311.04939.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*.

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

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023a. Long text and multi-table summarization: Dataset and method. *Preprint*, arXiv:2302.03815.

Tianyang Liu, Canwen Xu, and Julian McAuley. 2023b. Repobench: Benchmarking repository-level code auto-completion systems. *Preprint*, arXiv:2306.03091.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Chau Nguyen, Phuong Nguyen, Thanh Tran, Dat Nguyen, An Trieu, Tin Pham, Anh Dang, and Le-Minh Nguyen. 2024. Captain at coliee 2023: Efficient methods for legal information retrieval and entailment tasks. *Preprint*, arXiv:2401.03551.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

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

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

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *Preprint*, arXiv:2304.04358.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *Preprint*, arXiv:1911.05507.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *Preprint*, arXiv:2401.06915.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *Preprint*, arXiv:2206.05802.

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

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

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Nikolaos Stylianou, Panagiotis Kosmoliaptsis, and Ioannis Vlahavas. 2021. Improved biomedical entity recognition via longer context modeling. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17*, pages 45–56. Springer.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *Preprint*, arXiv:2011.04006.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by llms. *Preprint*, arXiv:2309.07382.

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

Jiebin Zhang, Eugene J. Yu, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Xiaoguang Li, Qun Liu, and Sujian Li. 2024a. Retrieval-based full-length wikipedia generation for emergent events. *Preprint*, arXiv:2402.18264.

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. 2024b. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Linyong Nan, Lyuhao Chen, Ryo Kamoi, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2023. Docmath-eval: Evaluating numerical reasoning capabilities of llms in understanding long documents with tabular data. *ArXiv*, abs/2311.09805.

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

Yijie Zhou, Kejian Shi, Wencai Zhang, Yixin Liu, Yilun Zhao, and Arman Cohan. 2023. Odsum: New benchmarks for open domain multi-document summarization. *Preprint*, arXiv:2309.08960.

A Benchmark Scope-Dispersion Classification

In Table 1 we delineate the different long-context benchmarks, as well as classify them according to how challenging they are scope-wise and dispersion-wise.

	LOW SCOPE	HIGH SCOPE
LOW DISPERSION	QA Qasper (Dasigi et al., 2021) NarrativeQA (Kočiský et al., 2018) Short-dependency QA (Li et al., 2023) MultiFieldQA (Bai et al., 2023) LitM (QA) (Liu et al., 2024b) L-eval (MC QA) (An et al., 2023) NQ (Kwiatkowski et al., 2019) RULER (single-hop QA) (Hsieh et al., 2024) MeetingQA (Prasad et al., 2023) BABILong (tasks 1,4-6,9-10) (Kuratov et al., 2024) Giraffe (2 tasks) (Pal et al., 2023) Retrieval LitM (Key-value Retrieval) (Liu et al., 2024b) MultiDoc2Dial (GSP) (Feng et al., 2021) TopicRet (Dacheng Li* and Zhang, 2023) Wiki-GenBen (Zhang et al., 2024a) RULER (S-NIAH & MK-NIAH) (Hsieh et al., 2024) LongChat-Lines (Pal et al., 2023) NLI LawngNLI (Bruno and Roth, 2022) ContractNLI (Koreeda and Manning, 2021b) Hallucination Detection (Dong et al., 2024) FLenQA (3 tasks) (Levy et al., 2024) Fill-mask Cloze (Li et al., 2023) NLG MultiDoc2Dial (ARG) (Feng et al., 2021)	QBS QMSum (Zhong et al., 2021) SQuALITY (Wang et al., 2022) Related Work Summarization (An et al., 2023) SPACE (Angelidis et al., 2021) WebBrain-G (Qian et al., 2023) AquaMuse (Kulkarni et al., 2020) FINDSum-Liquidity (Liu et al., 2023a) Aggregation ZeroSCROLLS (SpaceDigest & BookSumSort) (Shaham et al., 2023) PassageCount (Bai et al., 2023) FINDSum-ROO (Liu et al., 2023a) Aspect-based Summarization ACLSum (Takeshita et al., 2024) OpenAsp (Amar et al., 2023) Text Sorting Bamboo (ShowsSort & ReportSumSort) (Dong et al., 2024) Retrieval PassageRetrieval (Bai et al., 2023) LFQA LongFQA (An et al., 2023) NLI Legal Case Entailment (Nguyen et al., 2024)
HIGH DISPERSION	QA QuALITY (Pang et al., 2022) Long-dependency QA (Li et al., 2023) DuReader (Bai et al., 2023) SFcition QA (An et al., 2023) ExpertQA (Malaviya et al., 2024) DocFinQA (Reddy et al., 2024) BABILong (tasks 2-3,12) (Kuratov et al., 2024) Bamboo (QA) (Dong et al., 2024) Multi-hop QA MuSiQue (Trivedi et al., 2022) HotpotQA (Yang et al., 2018) Multi-hop Tracing (Hsieh et al., 2024) RULER (multi-hop QA) (Hsieh et al., 2024) 2WikiMultihopQA (Ho et al., 2020) NLI FLenQA (3 rand. placement tasks) (Levy et al., 2024) Legal Textual Entailment (Nguyen et al., 2024) Code Understanding LCC (Guo et al., 2023) RepoBench-P (Liu et al., 2023b) CodeU (An et al., 2023) PrivateEval (Dong et al., 2024) Classification LRA (tasks 2, 4-6) (Tay et al., 2020) Retrieval COLIEE (tasks 1,3,4) (Nguyen et al., 2024) RULER (MV-NIAH & MQ-NIAH) (Hsieh et al., 2024) Next Token Prediction PG-19 (Rae et al., 2019) Bamboo (LM) (Dong et al., 2024) Reasoning DocMath-Eval (Zhao et al., 2023) BABILong (tasks 14-20) (Kuratov et al., 2024) Aggregation RULER (2 Aggr. tasks) (Hsieh et al., 2024) BABILong (tasks 7-8) (Kuratov et al., 2024) NLU Academic Feedback Generation (An et al., 2023) CUAD (Hendrycks et al., 2021) Factuality Evaluation LongSciVerify (Bishop et al., 2024) Coreference Resolution BABILong (tasks 11,13) (Kuratov et al., 2024)	Summarization GovReport (Huang et al., 2021b) SummScreenFD (Chen et al., 2022b) Loogle (Summarization) (Li et al., 2023) VCSUM (Bai et al., 2023) Self-critiquing (Saunders et al., 2022) Abstract Generation (An et al., 2023) Multi-News (Fabbri et al., 2019) BigPatent (Sharma et al., 2019) Scientific Summarization (Cohan et al., 2018) BillSum (Kornilova and Eidelman, 2019) HowSumm (Boni et al., 2021) ODSum (Zhou et al., 2023) Klexikon (Summarization) (Aumiller and Gertz, 2022) Booksum (Kryściński et al., 2022) MeetingBank (Hu et al., 2023) Text Simplification Klexikon (Simplification) (Aumiller and Gertz, 2022) Reasoning Long ListOps (Tay et al., 2020) Retrieval LRA (task 3) (Tay et al., 2020)

Table 1: Classification of long-context benchmarks in terms of Scope and Dispersion.