# Is It Really Long Context or Do You Need Retrieval? Towards Genuinely Difficult Long Context NLP

Omer Goldman*[1], Alon Jacovi†[1], Aviv Slobodkin‡[1], Aviya Maimon§[1], Ido Dagan[1], and Reut Tsarfaty[1]

[1]Bar-Ilan University
[1]omer.goldman@gmail.com

February 8, 2026

**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 (**??**), contemporary models are—technically—able to process up to 128k and even 1M tokens (**??**).

*Equal contribution
†Equal contribution
‡Equal contribution
§Equal contribution

IMAGE NOT PROVIDED
Taxonomy of long context tasks based on the distribution of needed information
(Scope vs. Dispersion)

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 (**???**). 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 question: 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 (§**??**).

For example, Needle in a Haystack tasks (NIAH; **?**; **?**) 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 (**?**), 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 (**?**), or those that position the "needles" closer or farther apart (**?**). 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 (§**??**). 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 **??** 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 (§**??**).

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 (§**??**).

## 2 Task Design in Long Context

Evaluating the performance of NLP models over very long contexts is a fast-changing area of research (**??**). Measurements are regularly updated to account for new capabilities which improve with extrapolation architectures (**??**) and training data (**??**). 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 (**???**).

### 2.1 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 (**??**), summarization (**??**), and NLI (**??**). Specialized domains like legal (**??**) and literature (**??**) 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 (**??**), survey generation (**?**), and structured data aggregation (**?**). 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.

### 2.2 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; **?**; **?**) and PassKey retrieval (**?**) were created to evaluate a model's ability to pinpoint specific information amid lengthy distractors. Flexible and effective against existing models, they became standard benchmarks for evaluating new long-context models (**??**). Followup studies have complicated these tasks by increasing the number of critical details to locate (**??**) and changing their position within the input (**??**).

### 2.3 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 (**?**). 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:

[label=()]

1. How difficult is it to find and extract the required information?
2. How much information is needed to be found?

Assuming that some highlighting of the relevant information is needed to solve the task (see Figure **??**), 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.

## 3.1 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.

## 3.2 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.

## 3.3 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

IMAGE NOT PROVIDED
Distribution of long-context benchmarks by scope and dispersion characteristics

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 is 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.

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.

# 4 Challenging Long Context Is Under-Explored

Revisiting the works surveyed in §**??**, 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 (**??**) and a myriad of factual single-hop QA datasets (**????**) in which the answer is relatively accessible. Adding more snippets of information separated by distractors, either in the form of several needles (**??**) or of hops in a multi-hop question (**??**), complicates the information detection due to the need to find at least two snippets (**?**), thereby increasing dispersion. Dispersion can also be increased by making the detection of the information less straightforward (**?**) or requiring aggregation (**?**). Lastly, summarization tasks are of a very high dispersion (**??**), 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 (**???**). A somewhat higher scope is achieved by datasets for query-based summarization (**??**), and QA datasets with more obfuscated answers that require reading the text surrounding the answer for its verification (**??**). 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 (**???**). Currently, tasks with the highest scope, requiring information across the entire input length, are artificial and of low dispersion, like common words extraction (**?**).

Conclusion. Figure **??** 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

## 5.1 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 §**??** for creating long context tasks—lengthening short context tasks and synthetically creating length-adjustable tasks—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 (**?**), 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).

## 5.2 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 (**?**), financial reports (**?**), biomedical publications (**?**), and so on. In specialized domains, it is common that dispersion will be naturally higher (**?**). Tasks that involve implicit aggregations over structured data, such as table manipulation (**?**), 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.

## 5.3 Formality

Dispersion and scope, as defined here—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.

## 5.4 Retrieval is Still Interesting

Although we argue that small scope and low dispersion tasks are the least indicative of the model's ability to handle 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 (**???**).

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

### 5.5 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 (**?**) or containing other modalities and structures (**?**).

## 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

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.

## Acknowledgments

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

## A Benchmark Scope-Dispersion Classification

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

Table 1: Classification of long-context benchmarks in terms of scope and dispersion.

| LOW SCOPE | HIGH SCOPE |
|---|---|
| **Retrieval** | |
| NIAH (Ivgi et al., 2023) | |
| NaturalQA (Kočiskỳ 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) | |
| LitM (Key-value Retrieval) (Liu et al., 2024b) | |
| MultiDoc2Dial (CSP) (Feng et al., 2021) | |
| TopicRet (Dacheng Li* and Zhang, 2023) | |
| Wiki-GenBen (Zhang et al., 2024a) | |
| RULER (S-NIAH & MK-MAH) (Hsieh et al., 2024) | |
| LongBench (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 (Feng et al., 2021) | |
| QuALITY (Pang et al., 2022) | Long-dependency QA (Li et al., 2023) |
| | DuReader (Bai et al., 2023) |
| | Fiction 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) | |
| 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) | |