ETHIC: Evaluating Large Language Models on Long-Context Tasks with High Information Coverage

Taewhoo Lee[1,3] **Chanwoong Yoon**[1] **Kyochul Jang**[2] **Donghyeon Lee**[1,3] **Minju Song**[1]
Hyunjae Kim[1,†] **Jaewoo Kang**[1,3,†]
[1]Korea University [2]Seoul National University [3]AIGEN Sciences
{taewhoo,cwyoon99,dong9733,thdalswn99}@korea.ac.kr
kyochul@snu.ac.kr {hyunjae-kim,kangj}@korea.ac.kr

Abstract

Recent advancements in large language models (LLM) capable of processing extremely long texts highlight the need for a dedicated evaluation benchmark to assess their long-context capabilities. However, existing methods, like the needle-in-a-haystack test, do not effectively assess whether these models fully utilize contextual information, raising concerns about the reliability of current evaluation techniques. To thoroughly examine the effectiveness of existing benchmarks, we introduce a new metric called information coverage (IC), which quantifies the proportion of the input context necessary for answering queries. Our findings indicate that current benchmarks exhibit low IC; although the input context may be extensive, the actual usable context is often limited. To address this, we present ETHIC, a novel benchmark designed to assess LLMs' ability to leverage the entire context. Our benchmark comprises 1,986 test instances spanning four long-context tasks with high IC scores in the domains of books, debates, medicine, and law. Our evaluations reveal significant performance drops in contemporary LLMs, highlighting a critical challenge in managing long contexts. Our benchmark is available at https://github.com/dmis-lab/ETHIC.

1 Introduction

The field of natural language processing (NLP) has made remarkable progress in developing models that can manage much longer texts. While earlier Transformer-based models (Vaswani, 2017) could only process 512 tokens at a time (Kenton and Toutanova, 2019; Raffel et al., 2020), modern large language models (LLM) have achieved a significant breakthrough, now capable of handling documents with up to two million tokens (Reid et al., 2024). In light of these advancements, recent efforts have focused on establishing benchmarks and tasks specifically designed to evaluate the performance of these long-context models (Shaham et al., 2023; Hsieh et al., 2024).

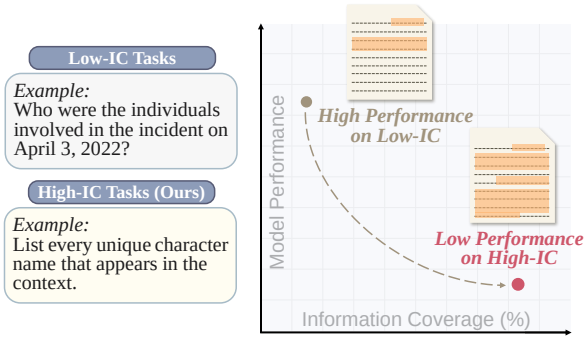

Figure 1: The variation in model performance with the level of information coverage (IC). Unlike low-IC tasks, which focus on specific parts of the input context, our benchmark features new high-IC tasks that demand the full utilization of all available information, posing a significant challenge for long-context models.

However, current long-context benchmarks often face challenges in assessing whether models are fully utilizing the information available in the provided context. One common research method, known as the needle-in-a-haystack test (Kamradt, 2023), aims to identify a specific piece of information within a lengthy context. However, excelling in these tasks does not guarantee that the model has effectively processed all the available information. Since the relevant information typically constitutes only a small portion of the entire text, much of the surrounding content is often irrelevant to the query. While several datasets have proposed tasks involving multiple key pieces of information scattered throughout the provided context (Dong et al., 2024; Li et al., 2024; Wang et al., 2024b), they still do not fully encompass the entire context. This raises concerns about whether models are adequately evaluated on their ability to fully incorporate the entire

†Corresponding authors.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5497–5512

context length (Goldman et al., 2024).

To this end, we propose ETHIC, a suite of long-context tasks specifically designed to assess whether LLMs can fully utilize the provided information. Our benchmark encompasses four distinct domains: books, debates, medicine, and law, each containing a set of tasks that require the use of all relevant information in the context to arrive at a solution. To measure this capability, we introduce the concept of *information coverage* (IC), which quantifies the proportion of the input context required to answer a query. Figure 1 shows that model performance significantly declines for high-IC tasks in our benchmark compared to low-IC tasks, even when the same input contexts are used.

We evaluated the latest LLMs that support at least 128k tokens, along with several training-free methods using our benchmark. Our findings revealed that recent models perform poorly across all tasks and domains, even when utilizing recent frameworks proposed for efficient long-context processing (Xiao et al., 2024b,a; Qian et al., 2024). This highlights a significant challenge for high-IC tasks and the need for further research in this area. We also conducted detailed analyses, comparing the performance gap between our task and traditional low-IC tasks, as well as identifying which parts of the input context the models typically fail to address. In summary, our contributions include:

- We introduce a new metric called information coverage (IC) to measure the proportion of input context required to answer a query.
- We propose ETHIC, the first benchmark of its kind, designed to assess whether LLMs can fully process the provided information. Our benchmark requires a higher IC than existing benchmarks, which presents a new challenge for the latest LLMs.
- We conduct a comprehensive analysis of how LLMs perform in high-IC tasks, establishing a foundation for future research on the development of advanced long-context models.

2 Preliminaries

In this section, we provide an overview of recent long-context LLMs and the benchmarks currently used to evaluate them (see Sections 2.1 and 2.2). We also present a formal description of *information coverage* and explain how our benchmark differs from existing ones, emphasizing the new aspects of LLMs that we aim to evaluate (see Section 2.3).

2.1 Long-Context LLMs

Pre-training LLMs on long contexts requires significant computational resources. Early open-source models, such as LLaMA (Touvron et al., 2023a), could manage input lengths of about 2K tokens, while LLaMA 2 (Touvron et al., 2023b) increased this limit to 4K tokens. Even commercial models like GPT-3.5 initially supported input lengths of 16K tokens. More recent models have vastly improved these capabilities, now supporting input lengths that range from 128K (Dubey et al., 2024) to two million tokens (Reid et al., 2024). However, the techniques employed to achieve this efficiency, along with the actual amount of input that models can effectively utilize, remain largely disclosed.

Several fine-tuning techniques have been explored to effectively extend the context window of pre-trained LLMs (Zhu et al., 2023; Chen et al., 2024; Peng et al., 2024). For instance, Chen et al. (2023) noted that simply training a model on longer contexts is both computationally intensive and often ineffective. They proposed a position interpolation method as a more efficient solution. Furthermore, approaches to extend the context window during inference—without requiring additional training—have also been investigated (Jin et al., 2024; Xiao et al., 2024a,b; Han et al., 2024).

2.2 Long-Context Benchmarks

Researchers have been evaluating how well long-context LLMs handle extensive text. A common approach is the needle-in-a-haystack (NIAH) task, where the goal is to locate key information (the "needle") within a large volume of text (the "haystack") (Kamradt, 2023; Mohtashami and Jaggi, 2023). Some studies manipulate the number of needles and the haystack's length to increase the complexity (Hsieh et al., 2024; Song et al., 2024). Additionally, several studies have adapted traditional NLP tasks–such as retrieval, single-document QA, and summarization–to serve as long-context evaluation scenarios (Shaham et al., 2023; An et al., 2024; Bai et al., 2024; Zhang et al., 2024). Some benchmarks target long-dependency or multi-hop reasoning and distribute information throughout the context (Dong et al., 2024; Li et al., 2024; Wang et al., 2024b). However, these benchmarks often utilize contexts in which a significant portion of the text is irrelevant to the query, and

Benchmark	Newly Curated	Input Text Structure	Document Relevance	Information Coverage (%)
NIAH (Kamradt, 2023)	Yes	Multi	Low	N/A
RULER (Hsieh et al., 2024)	Partial	Multi	Low	N/A
Counting-Stars (Song et al., 2024)	Yes	Multi	Low	N/A
ZeroSCROLLS (Shaham et al., 2023)	No	Single & Multi	High	56.1
L-Eval (An et al., 2024)	Partial	Single & Multi	High	35.4
InfiniteBench (Zhang et al., 2024)	Yes	Multi	Mixed	16.5
BAMBOO (Dong et al., 2024)	Yes	Single	High	41.4
Loong (Wang et al., 2024b)	Yes	Multi	High	14.4
LooGLE (Li et al., 2024)	Yes	Single	High	9.6
ETHIC (**Ours**)	Yes	Single & Multi	High	91.0

Table 1: Comparison of existing long-context benchmarks and our dataset. "Newly Curated" indicates whether the input text and queries/instructions are reused from existing datasets or newly created. "Input Text Structure" specifies whether the input context consists of a single document or multiple documents. "Document Relevance" assesses whether the different documents in multi-document tasks are unrelated and noisy, or if they are connected and coherent. "Information Coverage" quantifies the amount of information within the input context that is necessary to answer the query. Note that information coverage is marked as "N/A" if the proportion of required information varies depending on custom settings. Please refer to Sections 2 and 3 for information on the datasets.

only a small segment contains useful information. Consequently, they do not fully assess how well LLMs understand and integrate different parts of the given context.

In contrast, our benchmark requires models to make extensive use of the provided context. To measure the necessary amount of information, we introduce a metric called information coverage, which is explained in detail in Section 2.3. Table 1 illustrates the differences between existing benchmarks and our proposed benchmark.

2.3 Information Coverage

Let $\mathcal{D} = \{(\mathbf{C}_i, \mathbf{q}_i\, \mathbf{a}_i)\}_{i=1}^{N}$ be a dataset, where $\mathbf{C}_i$, $\mathbf{q}_i$, and $\mathbf{a}_i$ represent the i-th input long context, query, and output (answer), respectively. The input context can be a single document, such as a book, or a collection of related documents, like a set of papers on a similar research topic. We omit the subscript i for simplicity. We divide the input context into chunks, each with a length of up to 512, denoted as $\{\mathbf{c}_1, \ldots, \mathbf{c}_T\}$. For each data example, we calculate the IC score by using an evaluator model to determine whether each text chunk is (potentially) necessary for answering the query, as follows:

$$s(\mathbf{C}, \mathbf{q}) = \frac{1}{T} \sum_{j=1}^{T} \mathcal{M}(\mathbf{c}_j, \mathbf{q}), \qquad (1)$$

where s is the IC scoring function, $\mathbf{q}$ is the query, and $\mathcal{M}$ represents the evaluator model that returns 1 if the text chunk should be taken into account and 0 otherwise. The IC score for the entire dataset is calculated as the average of all individual IC scores from the examples. We used GPT-4o as the evaluator because of its higher consistency compared to other models. Please refer to Appendix A for the detailed prompt.

A few concurrent studies have also sought to establish the criteria for evaluating long-context LLMs. Goldman et al. (2024) defined the aspect of "scope" as "how much necessary information is there to find?", which is similar to our definition. However, they did not provide a specific metric and systematically evaluate existing benchmarks. In contrast, we present a novel approach for measuring the quantity of information, marking the first time this has been done in this area.

3 ETHIC

In this section, we outline the document collection process (see Section 3.1) and the task construction process (see Section 3.2) in detail. Additionally, we describe the label annotation process for each task in Section 3.3.

3.1 Corpus Selection

We selected four popular domains and gathered publicly available documents online without license restrictions for research purposes: books, debates, medicine, and law. The first two domains are used for single-document settings, while the latter two are for multi-document settings.

Books Book-sourced corpora have been widely used across various benchmarks (Kočiský et al.,

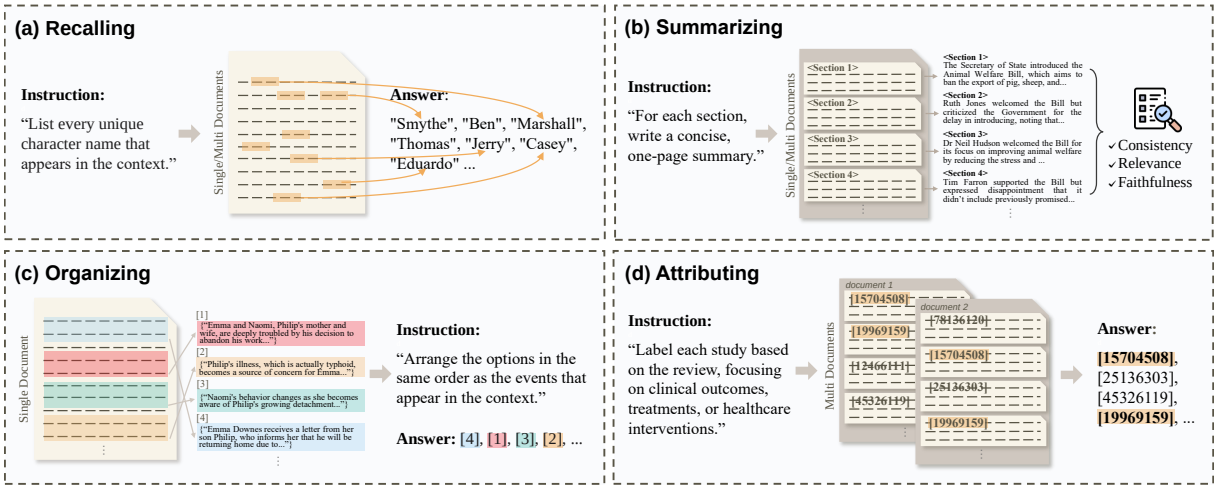

Figure 2: Overall description of ETHIC. Our benchmark includes four tasks: (a) the recalling task involves identifying specific types of entities in the text, (b) the summarizing task involves writing a summary for each section of the input, (c) the organizing task involves arranging mixed contents in the correct order, and (d) the attributing task focuses on identifying the underlying point of view within medical studies or legal documents.

2018; Kryscinski et al., 2022; Chang et al., 2024). We collected 100 English books from Project Gutenberg,[1] all of which are no longer under copyright as of 2024.

Debates We collected 229 debate transcripts from Hansard online,[2] which provides up-to-date records of debates held in the UK Parliament. Parliamentary debates cover a wide range of subjects, including political issues and legislative proposals. We manually selected debates by filtering out those that are either short (fewer than 10k tokens) or not considered debates, such as maiden speeches.

Medicine We collected 230 samples from the test set of MSˆ2 (DeYoung et al., 2021), a multi-document summarization dataset in which each sample comprises relevant medical abstracts used for systematic reviews. We excluded any samples with fewer than 10K tokens.

Law We gathered legal cases using the API from CourtListener[3], which provides up-to-date legal documents for research purposes. We grouped each target case with up to 15 related cases that cite it, resulting in a total of 103 samples.

3.2 Task Construction

When designing tasks, we prioritized two key aspects: (1) maximizing the use of the provided context and (2) ensuring they are grounded in well-defined categorization standards. By emphasizing these criteria, we enhanced the clarity and effectiveness of our benchmark, distinguishing it from existing long-context benchmarks. To achieve this, we drew on insights from Anderson et al. (2000), an authoritative source in the education field that provides a systematic approach to classifying educational objectives. From this framework, we adopted the following three distinct cognitive categories that align with our goals. (i) Remember: this cognitive process involves retrieving relevant information from the provided context in its original form. It serves as a crucial step for addressing more complex tasks that require integrating the knowledge gained during this process. (ii) Understand: this category involves constructing meaning by interpreting and making sense of the knowledge obtained from the provided context. (iii) Analyze: this process entails breaking down the provided context into its constituent parts and examining their relationships. Based on these categories, we developed four tasks–recalling, summarizing, organizing, and attributing–each corresponding to one of the three categories. Figure 2 illustrates the four tasks included in ETHIC.

Recalling In this task, models should retrieve all specific types of entity mentions from the input context, similar to the named entity recognition task (Sang and De Meulder, 2003). This includes identifying character names from books, names of individuals from debates, numbers of patients or populations from medical studies, and legal refer-

[1] https://www.gutenberg.org
[2] https://hansard.parliament.uk/
[3] https://www.courtlistener.com

Task	Cognitive Process	Domain	# Instances	# Avg. Tokens	Information Coverage (%)
Recalling	Remember	Books	100	76,812	93.8
		Debates	229	28,524	82.1
		Medicine (†)	230	29,037	86.6
		Law (†)	103	59,938	85.4
Summarizing	Understand	Books	100	76,893	92.3
		Debates	229	28,598	97.7
		Medicine (†)	230	29,065	91.7
		Law (†)	103	59,970	98.1
Organizing	Analyze	Books	100	94,403	74.4
		Debates	229	35,751	87.7
Attributing	Analyze	Medicine (†)	230	29,166	97.5
		Law (†)	103	56,905	82.5

Table 2: A summary of our dataset construction. ETHIC covers four domains—books, debates, medicine, and law—with a total of 1,986 test instances. In the medicine and law domains (marked with †), the inputs consist of multiple documents, while the inputs in the books and debates domains consist of a single long document. We randomly sampled 50 instances per domain and task to report information coverage.

ences from legal cases, all of which consistently appear throughout the document. To eliminate ambiguity in the answer format, models are guided to return a list of single words or numbers only.

Summarizing This task requires the model to summarize and rephrase the key ideas from the given context. While current LLMs are known to excel at traditional summarization tasks (Pu et al., 2023), evaluating this ability under long-context settings remains a challenge (Wu et al., 2024). We extended the existing summarization task to a high-IC setting. Instead of summarizing the long context all at once, we divided it into smaller text chunks and required models to summarize each chunk individually (e.g., sections in books). This encourages models to capture essential information from each section without missing important details.

Organizing In this task, models are provided with the entire context along with summaries of each chunk in a random order. Models should then rearrange these summaries into their original sequence. The output consists of the document IDs in the correct order. Conventional tasks generally require filling in missing parts in the correct order (Wang et al., 2024a) or reordering a limited number of information chunks (Dong et al., 2024), all of which can be solved by attending to specific areas instead of the entire context. Our task prevents these potential bypasses by instructing models to organize summary chunks that altogether represent the entire context. This task applies only to the single-document setting.

Attributing Unlike the organizing task, this task applies only to the multi-document setting. Models are tasked with inferring the underlying point of view within the given context. In the medicine domain, models receive a set of abstracts used in the same systematic review, along with a "background" paragraph. The background section is taken from a different target review, and some of these abstracts are also referenced in that review. The models must then identify the IDs of the medical abstracts that were included in the target review. For the law domain, we first grouped multiple pages from each legal case into segments. Models are given a target case, along with a set of segments from other cases that cite the target case. Note that the specific citations are masked using a citation mark. Models must go over each segment and check whether the context surrounding each citation mark aligns with the target case. Overall, this task involves actively integrating information from each document and understanding the reasoning behind citing a particular study or case.

3.3 Annotation Process

This section provides details of our annotation process for each task, except for the summarizing task, where we adopted a reference-free method to evaluate the generated summaries (Liu et al., 2023a). Table 2 provides a summary of our benchmark.

Recalling Since annotating long contexts all at once can reduce accuracy, we processed each context by dividing it into smaller chunks (up to 1,024 tokens). We used GPT-4o to annotate each small

chunk, and merged the labels without duplicates to obtain the final label set for the full input document. We initially instructed GPT-4o to review its own answers, but empirically found that this process often led the model to misjudge correct answers as incorrect. We manually reviewed the accuracy of the model's annotations by examining 100 chunk-label pairs from each domain and found that they were highly accurate (see Appendix E for details).

Organizing Generating labels for this task involves generating multiple summary chunks from the original context and shuffling them into random orders. Using the same small chunks used when annotating the recalling task, we prompted GPT-4o to briefly summarize each chunk with up to five sentences, and then we randomly shuffled them.

Attributing For the medical domain, each sample (i.e., a set of medical abstracts used in the same systematic review) was inspected to identify the PMIDs of abstracts that were also included in another sample, which became the label set. If multiple samples contained overlapping abstracts, we chose the one with the highest overlap. If no abstract was used in any other sample, we randomly selected background paragraphs, and the label set was labeled as "none." For the law domain, each sample included one target case and multiple citing cases. We first generated a summary of the target case using GPT-4o. Then, for each page of a citing case, the model was instructed to identify all spans referring to the target case based on its title and summary. Any spans referring to the target case were replaced with target citation markers ("[TARGET CITATION]"), while references to other cases were replaced with generic citation markers ("[CITATION]"). Finally, we grouped pages into segments, identified the segment IDs containing target citation markers, and replaced every target citation marker with a generic citation marker.

4 Experiments

In this section, we outline different metrics used to evaluate each task (see Section 4.1). We introduce baseline models and methods in Section 4.2, and provide the results in Section 4.3.

4.1 Metrics

For the recalling task, we used the F1 score to evaluate how well the predicted entities matched the ground truth. For the summarizing task, we followed the approach of Wu et al. (2024), prompting GPT-4o to rate the generated summaries on consistency, relevance, and faithfulness, using a scale from 1 to 5. The final score was calculated by multiplying the probability by the assigned score, similar to the method used by Liu et al. (2023b). For the organizing task, we found that models can hardly obtain any scores when evaluated using Exact Match. Therefore, we used Longest Common Subsequence (LCS), which measures the proportion of the longest matching subsequence between the prediction and the ground truth, relative to the total sequence length. The subsequences do not need to be consecutive. Lastly, for the attributing task, we used the F1 score to compare the predicted document IDs with the ground truth.

4.2 Baselines

Long-Context Models We used the current best LLMs that support a context window of over 128k tokens on ETHIC. This included three powerful proprietary models–Gemini Pro 1.5 (Reid et al., 2024), GPT-4o, and GPT-4o mini (OpenAI, 2024)–as well as open-source models such as Phi-3.5-mini-instruct (Abdin et al., 2024), Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct (Yang et al., 2024), GLM4-9B-Chat (GLM et al., 2024), Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct (Dubey et al., 2024). Gemini Pro 1.5 supports a length of 2M, while the other models support up to 128K.

Training-Free Methods To investigate promising methods, we tested three training-free frameworks specifically designed to efficiently manage long input contexts. These frameworks can be applied to any LLM without modification; for our experiments, we used Llama-3.1-8B-Instruct as the backbone LLM. (1) StreamingLLM (Xiao et al., 2024b) leverages the key-value caches of initial tokens within the input context and a finite attention window, capitalizing on the phenomenon where the model's attention heavily "sinks" into these initial tokens. (2) InfLLM (Xiao et al., 2024a) selects relevant token sequences from a part of the input that is distant from the current tokens, and it combines these with the initial tokens and local context. This method has proven effective in existing long-context benchmarks, including InfiniteBench (Zhang et al., 2024) and LongBench (Bai et al., 2024). (3) MemoRAG (Qian et al., 2024) is a retrieval-augmented generation (RAG) framework (Lewis et al., 2020) that consists of a lightweight memory model and a more

Model	Recalling	Summarizing	Organizing	Attributing
	F1 (%)	Score (1-5)	LCS (%)	F1 (%)
Proprietary				
Gemini Pro 1.5 (Reid et al., 2024)	**69.1**	2.9	**54.5**	39.4
GPT-4o (OpenAI, 2024)	49.5	**3.1**	39.0	**41.3**
GPT-4o mini (OpenAI, 2024)	32.3	2.7	21.7	30.5
Open-Source				
Qwen2.5-72B-Instruct (Yang et al., 2024)	**45.4**	**3.0**	**27.9**	**45.1**
Llama-3.1-70B-Instruct (Dubey et al., 2024)	37.7	2.4	25.7	41.1
Qwen2.5-7B-Instruct (Yang et al., 2024)	14.3	2.7	19.0	24.2
Llama-3.1-8B-Instruct (Dubey et al., 2024)	18.0	2.3	20.2	28.8
GLM4-9B-Chat (GLM et al., 2024)	18.3	2.3	22.1	28.2
Phi-3.5-mini-instruct (3.8B) (Abdin et al., 2024)	11.7	2.2	15.9	25.9
Training-Free Methods (built upon Llama3.1-8B-Instruct)				
StreamingLLM (Xiao et al., 2024b)	15.8	1.6	1.7	10.6
InfLLM (Xiao et al., 2024a)	17.0	1.8	13.8	12.7
MemoRAG (Qian et al., 2024)	16.8	1.7	22.1	16.7

Table 3: The performance of models and training-free methods on ETHIC. The best scores are highlighted in bold. We used instruction-tuned versions of open-source models. Please refer to Section 4.1 for the details of the metrics.

resource-intensive answer generator. The memory model retrieves answers from a long-context database, which the answer generator then uses as clues to produce the final response.

4.3 Results

Table 3 shows the performance of models evaluated on ETHIC. Both open-source and commercial models demonstrated weak overall performance. Noticeably, Gemini Pro 1.5 consistently outperformed GPT-4o in our benchmark. This may be attributed to GPT-4o's 128K context limit, which is significantly lower than the 2M context capacity of Gemini Pro 1.5. This result suggests that models with superior long-context handling perform better in our benchmark.

Among the four tasks, the performance gap between the models was most pronounced in the recalling task. Gemini Pro 1.5 achieved 69.1%, while GPT-4o, the second-best model, scored 49.5%. The lowest-performing model managed just 11.7%. Despite the recalling task involving straightforward retrieval queries, the models struggled as the volume of information increased. In the summarizing task, the models achieved only moderate performance, which contrasts with the strong performance of recent models in traditional summarization tasks. Additionally, all models showed room for improvement in both organizing and attributing tasks.

When we applied long-context encoding methods to Llama-3.1-8B-Instruct, there was no performance improvement; in fact, overall performance generally decreased. Although these methods have been presented as effective for long-context tasks, they demonstrated limitations in our high-IC tasks, underscoring the need for further research.

5 Analysis

We conducted further analysis to better understand the models' limitations within our benchmark. We used GPT-4o mini as the main model throughout Sections 5.1 to 5.3, and all models for Section 5.4.

5.1 Comparison of Model Performance Between Low-IC and High-IC Tasks

As illustrated in Figure 1, we observed that model performance can vary significantly based on the amount of information required to answer the query, even when using the same input contexts. To explore this further, we constructed a set of low-IC tasks, including traditional NLP tasks such as single-document QA, multi-document QA, and query-focused summarization (QFS). Specifically, we extracted one or more chunks from the contexts in our benchmark and asked GPT-4o to generate queries and their corresponding answers. This method followed a framework similar to that used for developing our benchmark, ensuring high annotation accuracy. Figure 3 shows the performance was consistently better on low-IC tasks compared to high-IC tasks. This highlights that, even when tasks involve the same documents and similar cog-

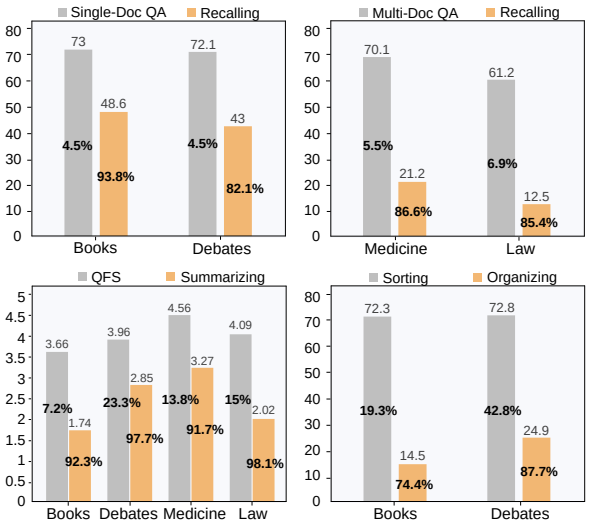

Figure 3: The model's performance on low-IC and high-IC tasks. Low-IC tasks were created by generating new queries and answers using the same input context from our benchmark, which are represented by the gray bars on the left side of the graph (please refer to Section 5.1 for details). The yellow bars on the right represent high-IC tasks from our benchmark. The numbers (%) displayed in the bar graphs represent the IC values of the tasks. The y-axis indicates the model performance.

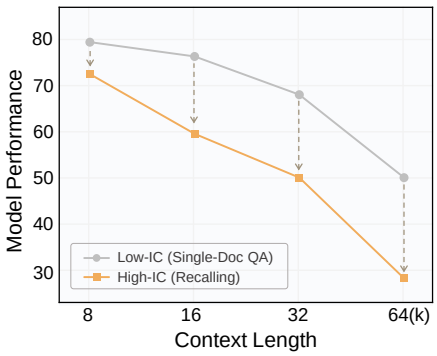

Figure 4: The performance with varied context lengths on low- and high-IC tasks. We used the single-document QA and recalling tasks from the books and debates domains for low- and high-IC tasks, respectively.

nitive demands, information coverage is a key factor that significantly impacts model performance.

5.2 Effect of Context Length

We examined how model performance in our benchmark is affected by increasing context lengths. For this analysis, we utilized recalling queries from the book and debate domain corpora. Figure 4 illustrates a consistent decline in model performance as the context length increased. This trend was also evident in low-IC tasks, as shown in the figure; however, the drop in performance was more

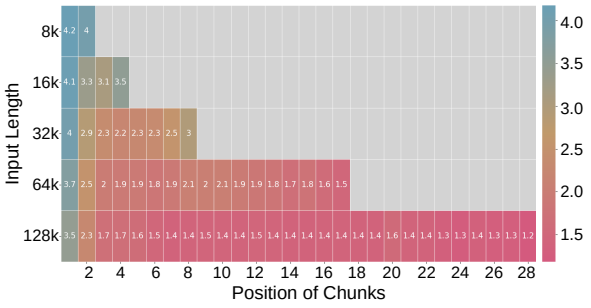

Figure 5: The effect of the position of information in the summarizing task. The x-axis represents the position of the chunks within the input context, while the y-axis represents the total length of the input context. Blue (red) chunks indicate summaries with high (low) scores.

pronounced in high-IC tasks compared to low-IC tasks, highlighting their distinct characteristics.

5.3 Effect of Position of Information

A recent study demonstrated that when relevant information is located in the middle of the input sequence, both QA and retrieval performance significantly declined (Liu et al., 2024). To investigate whether a similar position-dependent phenomenon occurs in our tasks, we visualized the model scores for each chunk in the summarizing task. Figure 5 shows that the model effectively processed information at the beginning of the provided context, but performance decreased toward the end. For context lengths ranging from 8K to 32K, the performance in the middle sections was the lowest, which is consistent with findings from Liu et al. (2024). When the context length exceeds 32K, we consistently observed a decline in performance as the context length increased. This might occur because, during decoding, the summaries of the later chunks are influenced by the outputs from the earlier sections, which increases the likelihood of errors in those later parts. However, this does not imply that the errors in the later sections are solely due to decoding issues. The performance of the earlier sections also decreased as the context length increased, suggesting that the model had inherent limitations in its ability to handle long-context encoding.

5.4 Error Analysis: Degeneration

Through a close examination of model responses, we noticed that models mainly suffer from degeneration (Welleck et al., 2020; Holtzman et al., 2020; Fu et al., 2021; Li et al., 2023), i.e. generation of unreasonably repetitive texts. Following previous works, we used Rep-n and Rep-r to quantify

Model	Recalling	Summarizing
	Rep-n (%)	Rep-r (%)
Proprietary		
Gemini Pro 1.5	**27.9**	**43.0**
GPT-4o	90.9	65.2
GPT-4o mini	69.8	53.8
Open-Source		
Qwen2.5-72B-Instruct	87.3	61.0
Llama-3.1-70B-Instruct	70.7	67.7
Qwen2.5-7B-Instruct	82.9	65.5
Llama-3.1-8B-Instruct	76.8	66.0
GLM4-9B-Chat	89.9	63.6
Phi-3.5-mini-instruct	**55.4**	**58.3**

Table 4: Rep-*n* and Rep-*r* scores across different models on the Recalling and Summarizing tasks. Please refer to Appendix F for the details of the metrics.

this phenomenon across different models (see Appendix F for details of the metrics). Specifically, we applied Rep-*n* on the Recalling task to account for the portion of repeated *n*-grams, and Rep-*r* on the Summarizing task to account for the portion of repetitive summary snippets. As shown in Table 4, all models exhibited severe degeneration issues across both tasks. This indicates that tasks requiring high information coverage adversely impacted model behavior during generation, ultimately hindering performance. Furthermore, notably high repetition scores observed from models such as GPT-4o indicates that degeneration remains a critical issue even for bigger models. Exploring different strategies to mitigate degeneration would be crucial for improving model performance on tasks that demand high information coverage.

6 Conclusion

In this study, we introduced ETHIC, a novel benchmark designed to evaluate LLMs' ability to fully process information in long-context settings. We also introduced a new metric, information coverage (IC), to quantify how much of the provided context is required to answer a query. Compared to existing benchmarks, ETHIC has significantly higher IC values, which underscores its unique aspect. Our experiments revealed that current commercial/open-source LLMs and long-context encoding methods perform poorly on our benchmark, emphasizing the need for future research to address long-context processing with high IC.

Limitations

We calculated the IC value for long contexts by breaking them into smaller chunks and assessing whether each chunk is necessary to answer the given query. However, in cases requiring multi-hop reasoning, some chunks may mislead the evaluator into believing they do not directly contribute to answering the query, even if they are crucial for intermediate reasoning when connected to other chunks. Consequently, the evaluator might overlook the potential usefulness of these chunks when evaluating them individually. While our study did not thoroughly analyze this aspect, we encourage future research to conduct additional analyses and improvements in this area.

In Figure 4, we show that as the context length increases, the performance of the models on our benchmark significantly declines. This drop is likely caused by a combination of difficulties in encoding long contexts and decoding lengthy output sequences. We have not analyzed these two factors separately. In future work, we aim to develop a more advanced evaluation framework that either fixes the output length while increasing the input context or maintains a constant input context while increasing the output length.

Data contamination and leakage (Magar and Schwartz, 2022; Xu et al., 2024) can lead to overestimating model performance and undermine the reliability of benchmarks. Recent LLMs are trained on extensive datasets during their pre-training phase, but they do not specify which data was used. Although we cannot guarantee that we have completely eliminated the possibility of data leakage in our benchmark, we would like to emphasize the steps we have taken to minimize it: (1) When selecting the book corpus, we chose data from sources where licensing issues were resolved recently (in 2024). This reduces the risk of exposure compared to books that were licensed long ago. (2) We only collected debate transcripts and legal cases that took place in late 2023 or early 2024. (3) In the medical domain, we used only the test split of the MSˆ2 dataset exclusively and did not incorporate any data from the training split.

Acknowledgements

This research was supported by (1) the National Research Foundation of Korea (NRF2023R1A2C3004176, RS-2023-00262002), (2) the Ministry of Health & Welfare, Republic of

Korea (HR20C0021), (3) ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-20200-01819), and (4) Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency(KOCCA) grant funded by the Ministry of Culture, Sports and Tourism(MCST) in 2023(Project Name: Development of storytelling AI technology for cultural heritage tailored to the various interests of users, Project Number: RS-2023-00220195, Contribution Rate: 100%).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2000. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives.

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. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

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

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12848–12856.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *arXiv preprint arXiv:2407.00402*.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. Lm-infinite: Zero-shot extreme length generalization for large language models. 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 3991–4008.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. LLM maybe LongLM: SelfExtend LLM context window without tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22099–22114. PMLR.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. Accessed: 2024-10-04.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

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.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective.

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

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 54567–54585. Curran Associates, Inc.

OpenAI. 2024. Hello gpt-4o.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

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.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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.

Mingyang Song, Mao Zheng, and Xuan Luo. 2024. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models.

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

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024a. Ada-LEval: Evaluating long-context LLMs with length-adaptable benchmarks. 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 3712–3724, Mexico City, Mexico. Association for Computational Linguistics.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. 2024a. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. *arXiv*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024b. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*.

A Details on Information Coverage

Below is the instruction used to measure IC.

You are given a passage and a query. The given query originally requires a reader to answer based on a longer context. This task splits the long context into multiple small passages, and aims to identify whether a certain passage needs to be taken into account in order to answer the query.

Passage:
(Note: This passage is a small portion of the original context, and may not be provided in the format specified in the query.)
<passages>

Query:
(Note: This query is for evaluation only. DO NOT answer it yourself.)
<query>

Score Rubrics:
0 : The passage is an unnecessary part of the original context, which does not need to be taken into account to answer the query.
1 : The passage is a necessary part of the original context, which should be taken into account to answer the query.

Follow the following format:

Query Understanding : {{demonstrate a clear understanding of the query and its requirements}}
Passage Understanding : {{demonstrate a clear understanding of the passage, focusing on how it relates to the query}}
Assessment : {{demonstrate your assessment based on the Score Rubrics provided above}}
Final Score : {{a single score only}}

B Details on Annotation Process

Below are the instructions used to evaluate models across different tasks. Note that for the Recalling task, the same instructions were used in the annotation process, but with smaller chunks of text.

B.1 Recalling

B.1.1 Books

Context:
<passages>

Now, respond to the instruction.
Instruction:
List every unique character name that appears in the context. Each name should be a single word. Remove any titles (e.g. Mr., Mrs., Captain, etc.) attached. If a character is mentioned with a full name, list each part of the name separately. Use commas to separate each name. If no such names are present, return "None".

Answer:

B.1.2 Debates

Context:
<passages>

Now, respond to the instruction.
Instruction:
List every real name that refers to a person. Remove any titles (e.g. Mr., Mrs., Speaker, Lady, etc.) attached to a name. If a person is mentioned with a full name, list each part of the name separately. Use commas to separate each name. If no such names are present, return "None".

Answer:

B.1.3 Medicine

Context:
<passages>

Now, respond to the instruction.
Instruction:
List every integer that refers to the count of people (or a group of people) mentioned in the passage. Do not include percentages, proportions, or integers related to non-human entities. If an integer is written in words, convert it to digits. Only return integers, without explanations. Use semicolons to separate each integer. Return "None" if no such integer is mentioned.

Answer:

B.1.4 Law

Context:
<passages>

Now, respond to the instruction.
Instruction:
List every number that appears directly in the names of legal citations, including case law numbers, volumes, series, statutes, sections, or codes. Do not include numbers that appear separately from the full citation name, such as standalone paragraph numbers. Remove any special characters that appear with the numbers, and return each number separately. Only return the numbers, separated by semicolons, without explanations or details. If no such numbers are present, return "None".

Answer:

B.2 Summarizing

Context:
<passages>

Now, respond to the instruction.
Instruction:
The context is split into <num_sections> section(s). For each section, write a concise, one-page summary. Prepend the appropriate section header (e.g. <Section 1>) to the summary, and use newlines if there are multiple sections to summarize. Only return the summary, without additional explanations, context, or commentary.

B.3 Organizing

Context:
<passages>

Now, respond to the instruction.
Instruction:
The following options summarize different parts of the given context. Arrange the options in the same order as the events that appear in the context. Only enumerate each option number surrounded by square brackets, without explanations. Use commas to separate your answer.

Options:
<options>

Answer:

B.4 Attributing

B.4.1 Medicine

Context:
<passages>

Now, respond to the instruction.
Instruction:
For each study, assess how the information presented aligns with the selection criteria for systematic reviews focused on clinical outcomes, treatments, or healthcare interventions. Consider whether the study's methodology, findings, or broader implications contribute to the robustness of a systematic review. Label its ID accordingly under one of the following categories:
- "Core IDs": Studies that meet the necessary criteria for inclusion in systematic reviews due to their direct contribution to clinical evidence or intervention outcomes.
- "Supplementary IDs": Studies that provide additional insights, but may not meet the primary inclusion criteria for systematic reviews.

Background:
<background>

Now, label each study under the correct category. Ensure that EVERY study is labeled under at least one category. Use square brackets to surround each ID, without explanations, and separate them by commas. Follow the following format.

- Core IDs:
- Supplementary IDs:

B.4.2 Law

Context:
<passages>

Now, respond to the instruction.
Instruction:
For the given context and target case, assume that the [citation]s within the context are replaced by the target case, and categorize each SEGMENT based on how the citations could assist in understanding the segment:
- "Related Segments": Segments where the "Target Case" provides a clear and direct connection to the citation marks within them, based on legal reasoning, evidence, or laws, making it a valuable reference.
- "Supporting Segments": Segments where the "Target Case" may provide some indirect relevance to understanding the context, but it does not have a direct connection to the citation marks in terms of legal reasoning, evidence, or laws.

Target Case:
<target_case>

Now, label each SEGMENT under the correct category. Ensure that EVERY study is labeled under at least one category. Use square brackets to surround each SEGMENT, without explanations, and separate them by commas. Follow the following format.

- Related Segments:
- Supporting Segments:

C Performance per Domain

In Table A, we report the performance of models and training-free methods for each domain.

D Implementation Details

All experiments were done using Nvidia A100 with 80GB memory. For inference, we used vLLM (Kwon et al., 2023) for efficiency. We adopted a greedy decoding strategy with temperature set to 0 and top_p set to 1.0.

Model	Books			Debates			Medicine			Law		
	Rec (%)	Sum.	Org (%)	Rec.	Sum.	Org	Rec.	Sum.	Att.	Rec.	Sum.	Att.
Proprietary												
Gemini Pro 1.5 (Reid et al., 2024)	**61.8**	**2.3**	**38.6**	**75.4**	2.9	**62.3**	**79.2**	3.3	**30.8**	**39.5**	2.5	38.1
GPT-4o (OpenAI, 2024)	54.4	2.1	30.2	69.6	**3.2**	42.8	44.4	**3.6**	30.4	11.7	**2.6**	**55.7**
GPT-4o mini (OpenAI, 2024)	48.6	1.7	14.5	43.0	2.9	24.9	23.4	3.3	22.4	12.5	2.0	42.0
Open-Source												
Qwen2.5-72B-Instruct (Yang et al., 2024)	**53.2**	**1.8**	**19.9**	**52.2**	**3.1**	**31.4**	**49.3**	**3.5**	**40.4**	14.1	**2.6**	48.0
Llama-3.1-70B-Instruct (Dubey et al., 2024)	50.9	1.7	19.4	38.0	2.5	28.5	42.0	2.7	30.5	**14.9**	2.3	**55.2**
Llama-3.1-8B-Instruct (Dubey et al., 2024)	29.2	1.5	13.6	11.2	2.3	23.1	23.4	2.6	24.9	10.3	2.2	36.6
Phi-3.5-mini-instruct (Abdin et al., 2024)	13.9	1.4	7.8	1.0	2.4	19.5	21.2	2.3	24.7	13.2	1.9	17.5
GLM4-9B-Chat (GLM et al., 2024)	33.5	1.6	16.5	13.9	2.4	24.6	18.7	2.6	35.6	12.6	2.2	9.7
Qwen2.5-7B-Instruct (Yang et al., 2024)	21.4	1.4	14.8	1.5	2.8	20.8	28.2	3.4	29.6	5.2	2.3	4.7
Training-Free Methods (built upon Llama3.1-8B-Instruct)												
StreamingLLM (Xiao et al., 2024b)	30.3	1.2	2.9	5.9	1.6	1.1	22.9	1.8	11.0	7.9	1.6	9.6
InfLLM (Xiao et al., 2024a)	29.1	1.2	5.9	9.5	1.8	17.2	23.6	2.1	12.4	6.9	1.7	13.5
MemoRAG (Qian et al., 2024)	29.8	1.2	15.7	10.4	1.8	25.0	25.9	1.8	17.9	7.9	1.9	14.0

Table A: Model performance per domain.

E Validation of Data Quality

To ensure the quality of labels generated by GPT-4o, we randomly selected 100 samples from each domain. The resulting annotations were evaluated against GPT-4o-generated labels using the F1 score, as shown in Table B.

Dataset	F1(%)
Books	96.5
Debates	99.0
Medicine	96.8
Law	95.2

Table B: Comparison of F1 scores (%) between GPT-generated labels and human-labeled data across different datasets.

F Details on Degeneration Analysis

We used two metrics to measure degeneration: Rep-*n* and Rep-*r*. Following the notations from Li et al. (2023), each metric is calculated as follows:

$$\text{Rep-n} = 1.0 - \frac{|\text{UniqueNgrams}(x, n)|}{L - n + 1}$$

$$\text{Rep-r} = \frac{1}{L}\left|\left\{i \mid (x_i = x_j \wedge x_{i+1} = x_{j+1}, \exists j \neq i) \vee (x_i = x_k \wedge x_{i-1} = x_{k-1}, \exists k \neq i)\right\}\right|$$

where x, L, and n refers to the sentence, its length, and the length of n-gram within the sentence, respectively. Rep-n measures repetition based on the portion of repeated n-grams, whereas Rep-r quantifies repetition based on the portion of repeated snippets measured by sentence length.