M[4]LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models

Wai-Chung Kwan[1,4*], **Xingshan Zeng**[2], **Yufei Wang**[2], **Yusen Sun**[2], **Liangyou Li**[2], **Yuxin Jiang**[3]
Lifeng Shang[2], **Qun Liu**[2], **Kam-Fai Wong**[1,4]
[1]The Chinese University of Hong Kong [2]Huawei Noah's Ark Lab
[3]The Hong Kong University of Science and Technology
[4]MoE Key Laboratory of High Confidence Software Technologies
{wckwan,kfwong}@se.cuhk.edu.hk
{zeng.xingshan,wangyufei44,sun.yusen1,liliangyou,Shang.Lifeng,qun.liu}@huawei.com
yjiangcm@connect.ust.hk

Abstract

Managing long sequences has become an important and necessary feature for large language models (LLMs). However, assessing their ability to handle long contexts remains a challenge. This paper introduces M^{4}LE, a **M**ulti-ability, **M**ulti-range, **M**ulti-task, **M**ulti-domain benchmark for **L**ong-context **E**valuation. It encompasses 36 NLP datasets, covering 11 types of tasks and 12 domains, providing a comprehensive test bed. To address the lack of tasks featuring naturally long sequences, we propose an automatic approach to convert short-sequence tasks into long-sequence scenarios. These scenarios evaluate LLMs' long-context understanding across five key abilities: understanding of single or multiple relevant spans in long contexts based on explicit or semantic hints, and global context understanding. This automatic approach allows us to create instances evenly distributed from 1k to 8k input length.[1] Our evaluation of 11 prominent LLMs reveals that 1) Current LLMs struggle to understand long context, particularly when tasks require multiple-span attention. 2) Semantic retrieval is more difficult for competent LLMs. 3) Models fine-tuned on longer text with position interpolation have comparable performance to those using Neural Tangent Kernel (NTK) aware scaling methods without fine-tuning. We make our benchmark publicly available to encourage future research in this challenging area [2].

1 Introduction

Large language models (LLMs) are gaining traction in addressing diverse NLP challenges. LLMs, mostly transformer-based models (Vaswani et al., 2017), are trained on a large amount of data with numerous parameters (Ouyang et al., 2022; Touvron et al., 2023b). These models have demonstrated impressive capabilities across a wide range of tasks (Brown et al., 2020; Schick et al., 2023; Shen et al., 2023; Bang et al., 2023). As LLMs continue to evolve, their ability to handle long-sequence tasks, such as extracting specific information from or summarizing lengthy documents, has become an important and competitive feature (Du et al., 2022; Chiang et al., 2023; Li et al., 2023). Therefore, a comprehensive, fair, and objective benchmark to evaluate the long-sequence capabilities of models is necessary for the progress of LLMs.

Despite numerous efforts to develop benchmarks for assessing the knowledge or reasoning ability of LLMs (Hendrycks et al., 2021; Suzgun et al., 2022; Huang et al., 2023), comprehensive evaluation of their long-context understanding ability has received limited attention. Recent concurrent works, such as L-Eval (An et al., 2023) and LongBench (Bai et al., 2023), primarily rely on existing long-sequence NLP datasets which usually limit the task diversity and flexibility in conducting length-control experiments. They lack an objective and comprehensive understanding of the model's capability across different dimensions of long sequences.

In this study, we aim to maximize the diversity of constructed tasks and analyze the long-context capabilities of LLMs from a user's practical perspective. We discovered that when processing instructions based on long sequences, the essential components for task completion can be classified as single-span, multiple-span, or global, based on relevance. Building on this and considering how to locate the relevant information, we categorize long-context understanding into five distinct abilities and introduce an automated method to transform short-sequence tasks into a comprehensive long-sequence scenario encompassing all these ca-

[*]Work done during an internship at Huawei Noah's Ark Lab.

[1]The released benchmark would contain samples up to 32k words. Even longer samples and other types of tasks can be constructed using our method.

[2]Code and data are available at https://github.com/KwanWaiChung/M4LE.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15568–15592
August 11-16, 2024

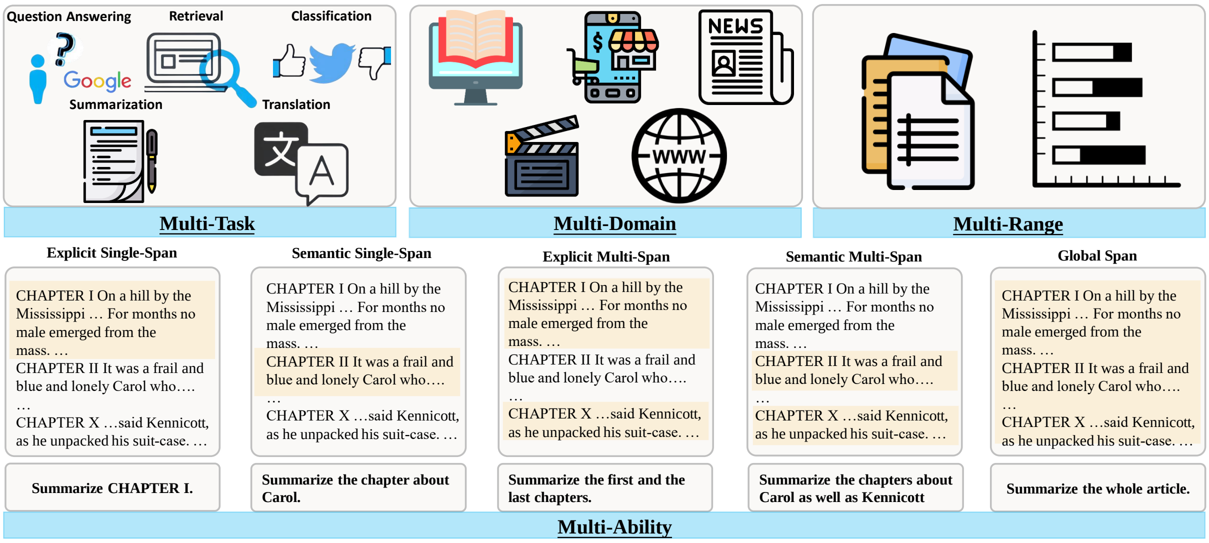


Figure 1: The illustration of M^{4}LE. M^{4}LE covers multiple task types, domains and length ranges, and introduces five long-context understanding abilities, each of which is exemplified with a summarization instance, to facilitate the long-context evaluation.

pabilities. As a result, M^{4}LE is proposed, a multi-ability, multi-range, multi-task, and multi-domain long-context evaluation benchmark for evaluating LLMs' ability to handle long inputs (Figure 1).

- Multi-ability: M^{4}LE includes tasks with five different types of understanding abilities, determined by whether single or multiple parts of the ongoing context are relevant to the current tasks and whether explicit or semantic hints are used in the question.
- Multi-range: Each task in M^{4}LE consists of samples with variable lengths, from 1K to 8K words, divided evenly into five buckets to measure the effect of length on model performance.
- Multi-task: M^{4}LE encompasses 36 datasets covering 11 task types, including original tasks such as classification and summarization, and their combination for more complex scenarios.
- Multi-domain: M^{4}LE spans a wide variety of domains, including Wikipedia, academic, news, E-Commerce, etc., prompting diversity and comprehensiveness.

Table 1 compares our benchmark with the existing similar benchmarks. M^{4}LE targets comprehensively evaluating LLMs' long-context understanding capabilities across different abilities and length ranges, rather than simply assessing naturally long input tasks. Therefore, the tasks in M^{4}LE are constructed from both existing long-context datasets and short-context datasets widely used in the NLP community, where short instances can be aggregated into long-context ones with designed procedures covering different abilities with varied instructions. Our approach is able to extend existing datasets to arbitrary context lengths. While the generated instances may not perfectly mimic natural long-form texts like lengthy reports, we believe that evaluating these instances effectively test model performance across the five defined abilities, thereby adequately reflects the model's long-context understanding capabilities. Moreover, this construction method can effectively prevent data leakage issues since the models are unlikely to have been trained on similarly constructed datasets.

We conducted a systematic evaluation over 11 well-known LLMs, especially those claimed to support long inputs, with M^{4}LE. This involves evaluating their long-context understanding ability across different length ranges and their performance in our proposed five different abilities. We also delve into the factors influencing long-context understanding capability, including LLMs performance under different languages and the positioning of relevant information (Liu et al., 2023). We find that current LLMs still struggle to understand long-context inputs, especially when multiple-span attention is required. While semantic retrieval is considered more complex than explicit, the consistent perfor-

Benchmarks	SCROLLS	ZeroSCROLLS	L-Eval	LongBench	M4LE
#Tasks	3	4	4	6	11
#Datasets	7	10	18	21	36
#Domains	7	9	10	10	12
Languages	en	en	en	en, zh	en, zh
Ranges	×	×	×	×	✓
Abilities	×	×	×	×	✓

Table 1: Comparison with other long context benchmarks.

mance drop in this scope can only be observed on competent models. A more effective fine-tuning approach deserves exploration, as current methods show no significant improvement over simple Neural Tangent Kernel (NTK) aware scaling methods. We also observe that language differences and the positioning of relevant information impact the long-context understanding capabilities.

2 Related Work

2.1 Long-Context Modelling for LLMs

To address length extrapolation challenges in LLMs beyond the training context window, several methodologies have emerged. Position embeddings such as Alibi (Press et al., 2022) and XPos (Sun et al., 2023) have been developed. Alibi employs an exponential decay on the attention matrix to mitigate out-of-distribution positions' influence, while XPos introduces a block-wise causal attention mask. While these techniques require integration during training, alternative approaches enhance existing RoPE-based LLMs (Su et al., 2021), notably LLaMA (Touvron et al., 2023a), LLaMA 2 (Touvron et al., 2023b), and PaLM (Chowdhery et al., 2022). Concurrently, kaiokendev (2023) and Chen et al. (2023) propose extending context length by modifying RoPE through Position Interpolation and subsequent limited data finetuning. Another line of research introduces fine-tuning free approaches (bloc97, 2023; emozilla, 2023; Peng et al., 2023), including NTK-aware and dynamic NTK interpolations.

2.2 Existing Evaluation Benchmarks for LLMs

As LLMs have demonstrated superior performance in a wide range of NLP tasks, comprehensively and effectively evaluating their ability becomes increasingly critical. Many of the research efforts focus on developing benchmarks for specific knowledge types (Hendrycks et al., 2021; Zhong et al., 2023) and specific task families (Chen et al., 2021; Cobbe et al., 2021). For more details, we refer readers to the recent LLMs evaluation survey (Chang et al., 2023; Wang et al., 2023). Several preliminary studies have begun to assess the model capability on long context input. Long Range Areana (Tay et al., 2020) verifies the capability of transformer-based models to handle various long sequence inputs, such as languages, vision tokens, and symbols. SCROLLS (Shaham et al., 2022) simply collects a set of naturally long NLP benchmarks covering multiple tasks and domains. Recently, ZeroSCROLLS (Shaham et al., 2023), L-Eval (An et al., 2023) and LongBench (Bai et al., 2023) are proposed to evaluate long text modelling capability of LLMs. However, these benchmarks are mainly compiled from a set of existing long NLP benchmarks, thereby suffering from data diversity (i.e., limited evaluation patterns) and data leakage (i.e., LLMs potentially already using these benchmarks for pre-training or alignment). In contrast, M4LE not only constructs evaluation instances from various tasks, domains, and length ranges but also covers three types of attention spans, offering a comprehensive evaluation of LLMs' long text capability.

3 M4LE

This section outlines the M4LE benchmark's rationale, design principles, data sources, and task construction methodologies. M4LE is designed to comprehensively evaluate large language models' (LLMs) abilities in understanding long contexts. It covers a wide range of tasks, domains, and context lengths, ensuring a thorough assessment of LLMs' competencies in this crucial area.

3.1 Design Principle

Each sample in M4LE is a tuple of ⟨Task description, Context, Instruction, Response⟩. To follow the instructions, LLMs must identify relevant information within a lengthy context. This information can be a single text segment (*single-span*), multiple text segments (*multiple-span*), or the entire context (*global*). The models locate these segments either through direct hints (*explicit*) or inferred meaning (*semantic*) in the instructions. We categorize the understanding ability into five distinct types: *explicit single-span*, *semantic single-span*, *explicit multiple-span*, *semantic multiple-span*, and *global* context understanding (Figure 1). This classification helps in assessing the models' comprehension capabilities.

To ensure a comprehensive evaluation, we prioritize task diversity in two aspects:

- Data Source: We select widely-used Chinese and English datasets in NLP which cover a variety of representative task types (e.g., QA, Summarization) and domains (e.g., News, Wiki, Web). In addition, we introduce tasks that integrate multiple task types, like Classification and Retrieval. These newly integrated tasks help measure LLMs' ability to solve more complex tasks.

- Length Range: It is important to reveal how LLMs perform on various lengths of contexts. In our benchmark, we evenly divide samples into buckets according to their context lengths. In addition, in order to alleviate the effects of the location of relevant parts in context (Liu et al., 2023), we intentionally construct instances with the relevant paragraphs uniformly distributed in the input context.

By focusing on these five core abilities and maximizing task diversity, M^{4}LE offers a comprehensive assessment of LLMs' long-context understanding capabilities.

3.2 Data Collection

We collect established datasets, both in English and Chinese, to cover a broad range of tasks and domains. We not only select datasets featuring long inputs, but also include datasets with shorter inputs for our customized construction, and at the same time, enriching the domain variety. The short-context datasets can be adapted to longer contexts using our designed process, which will be introduced in the next subsection. Below we describe the datasets selected in the benchmark briefly.

Question-Answering (QA): We include TriviaQA (Joshi et al., 2017), a single-document QA dataset based on web snippets and Wikipedia, with documents extended to 12k words. Additionally, NQ-Open (Lee et al., 2019), HotpotQA (Yang et al., 2018), and DRCD (Shao et al., 2019) are included, all of which are based on Wikipedia articles. We further collect NewsQA (Trischler et al., 2017) and DuoRC (Saha et al., 2018), both in English and constructed from news articles and movie plots. We also add C3 (Sun et al., 2021), a Chinese dataset comprising textbook questions.

Classification: We incorporate BIGPATENT (Sharma et al., 2019) which includes long patent documents, and MNDS News (Petukhova and Fachada, 2023) in English and THUCNews (Hu et al., 2019) in Chinese which would be further processed for different abilities. We also utilize a sentiment classification dataset collected from e-commerce platforms (SophonPlus, 2013).

Summarization: For English, we include Arxiv, Pubmed (Cohan et al., 2018), BIGPATENT (Sharma et al., 2019), and Booksum (Kryscinski et al., 2022), where the corresponding domains span across academic, medical, patent documents and books. We also introduce shorter summarization datasets enabling extension, such as CNNNews (See et al., 2017) and MNDS News, featuring news articles, and Wikihow (Koupaee and Wang, 2018). For Chinese, we incorporate CNewsum (Wang et al., 2021), CLTS+ (Liu et al., 2022), and News2016 (Xu, 2019), all constructed from long news articles. The LCSTS (Hu et al., 2015) dataset contains shorter news articles, while CEPSUM (Li et al., 2020) comprises product descriptions from e-commerce platforms. We also use NCLS (Zhu et al., 2019) to establish a bilingual task that generates a Chinese summary for a specific English news article.

Natural Language Inference (NLI): We construct two tasks using English and Chinese Wikipedia articles from WikiText-103 (Merity et al., 2016) and Wiki2019zh (Xu, 2019), respectively.

Translation: Three translation datasets are included, depending on sentence-level translation alignments to form long contexts, including Tedtalks (Qi et al., 2018), OpenSubtitles (Lison and Tiedemann, 2016), and News commentary (Tiedemann, 2012).

Retrieval: Lastly, we construct two retrieval tasks from the same datasets used for the NLI task for both languages. Since M^{4}LE comprises numerous tasks combined with retrieval capability, we do not construct additional standalone retrieval datasets.

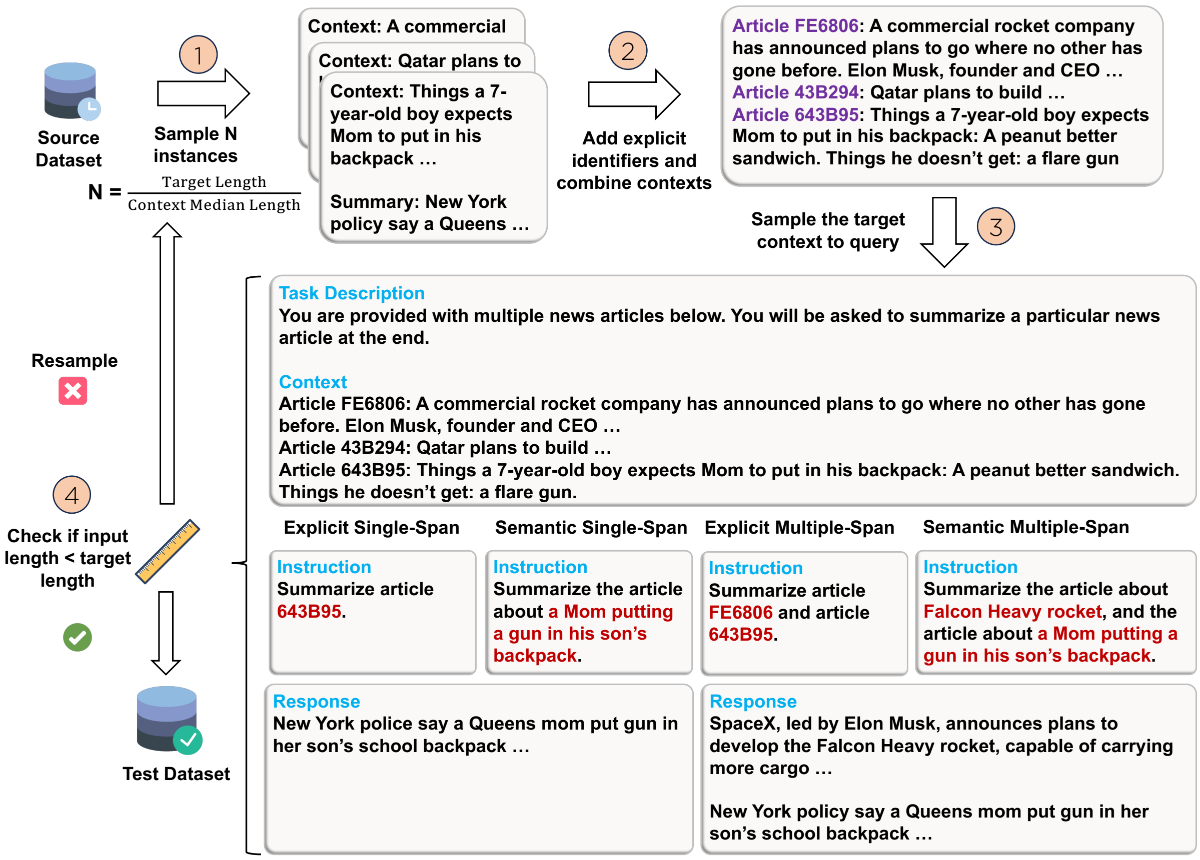


Figure 2: The illustration of the process of constructing a test instance with a target length from a source dataset. Each instance comprises a tuple containing the task description, context, instruction, and response. (1): The process begins by estimating the number of samples needed to achieve the desired target length. This is accomplished by dividing the median length of the context in the dataset by the target length. Subsequently, N instances are sampled from the source dataset. (2): The context of each sample is then marked with an explicit identifier and combined. (3): For single-span tasks, we uniformly sample one context to construct the query. For multi-span tasks, multiple contexts are sampled. We incorporate the explicit identifiers for explicit tasks and semantic hints for semantic tasks in the instruction. (4): If the total length exceeds the target length, the process returns to step one. Otherwise, the constructed sample is added to the test dataset.

3.3 Task Construction

This subsection details the dataset construction process of the evaluation benchmark. We construct test instances with diverse length ranges by transforming instances from collected datasets.

Figure 2 illustrates the construction process. To construct a test instance for a specific task with a target length range K, we first sample N instances from a single source dataset. These original instances contain context, such as an article, a talk transcript, or several text segments. We then concatenate their context paragraphs into a single sequence as "Context", marking each paragraph with an explicit identifier at the beginning for indexing. The value of N is determined by dividing K by the dataset's median context length. For each task, we manually craft a description and make sure LLaMA2-7B-Chat can understand it through preliminary testing with a few examples. We further provide instructions to guide the model to locate relevant information within the context using paragraph identifiers for explicit tasks and semantic hints for semantic tasks. This approach extends existing datasets with short contexts to accommodate arbitrary context lengths. Table 2 provides an overview of the constructed datasets in M^{4}LE. Appendix A provides the detailed statistics of the datasets used. In the following sections, we elaborate on the instruction construction process for each of the five abilities.

Explicit Single-Span Understanding. Instructions for tasks within this scope should direct models to complete the task based on a specific paragraph, with explicit hints to be located. For instance, in a question-answering task, the model might be asked to answer a question based on paragraph II. This approach has been used to construct ten unique datasets covering a wide range of task types and domains for the ability. Consequently, the task types are a fusion of retrieval and their original task, such as classification, which is labeled as "CLS + RET".

Semantic Single-Span Understanding. Analogous to explicit single-span understanding, the instructions for the tasks long to this ability instruct models to complete tasks based on a designated paragraph. Rather than using explicit identifiers, we provide hints about the paragraph, and models are tasked with retrieving it based on semantic information. For example, in a translation task, the model might be prompted to translate a paragraph associated with sports. Tasks within this ability are designed to introduce increased complexity and challenges since semantic-level retrieval necessitates the model to understand all paragraphs to pinpoint the right one. We have constructed nine distinct datasets aligned with this ability.

Explicit Multiple-Span Understanding. We add further complexities to the tasks within this ability. Specifically, models are tasked with handling assignments related to multiple, disjoint paragraphs within the lengthy input context. This could necessitate addressing several original instances, for example, summarizing the first and the third paragraphs. Despite these complexities, the instructions for this ability continue to utilize explicit hints. We have constructed four distinct datasets to align with this ability.

Semantic Multiple-Span Understanding. We replace the explicit hints in explicit multiple-span understanding with semantic ones, resulting in the instructions for tasks in this scope. We've developed three distinct datasets of high complexity in line with this. Within this ability, we've incorporated counting tasks (labeled as "CNT"), which demand the counting of relevant paragraphs. Such tasks pose a challenge since counting is not an innate function of language models.

Global Context Understanding. Finally, we present tasks in global context understanding, which is a special case within our construction process. When the original instances have sufficiently extensive context, such that the target length range K can be attained with $N = 1$, we directly employ them for the associated tasks, indicating that the entire context is relevant to the task completion, and global understanding is required. Within this category, we have included ten different datasets.

3.4 Models

We introduce the five families of LLMs evaluated in this study, comprising a total of 11 models.

LLaMA 2: It is a family of LLMs that support a maximum 4k input length (Touvron et al., 2023b). These models use rotary positional embeddings (RoPE) (Su et al., 2021). LLaMA 2 has 7B, 13B and 70B variant. We focus on its 7B and 13B models in this section. We also include their aligned versions: LLaMA2-7B-Chat and LLaMA2-13B-Chat.

Vicuna: We employ Vicuna-7B-v1.5-16K and Vicuna-13B-v1.5-16K (Chiang et al., 2023), fine-tuned based on the LLaMA2 models with 125k conversational data, collected from ShareGPT with context length up to 16K tokens using linear positional interpolation (Chen et al., 2023).

LongChat: We leverage LongChat-7B-v1.5-32K and LongChat-13B-16K (Li et al., 2023), fine-tuned on 80K and 18K conversations respectively, with context lengths up to 32K and 16K tokens, respectively. They utilize linear positional interpolation.

ChatGLM2: ChatGLM2-6B and ChatGLM2-6B-32K are based on the GLM (Du et al., 2022) models. Similar to LLaMA2, ChatGLM2 leverage RoPE. Both models are further refined on 8K and 32K input data, respectively, using linear positional interpolation.

GPT-3.5-Turbo: It is a closed-source language model developed based on InstructGPT (Ouyang et al., 2022). Analogous to LLaMA 2, it is fine-tuned with instruction data and refined by RLHF. We use the GPT-3.5-Turbo-16K variant [3], which supports a 16K context length.

3.5 Inference Details

Apart from the tuples introduced in Section 3.1, we also employ a concise and short in-context example, from the same dataset, to demonstrate the desired output format. Several full examples used in this work can be found in Appendix E. The main goal of M^4LE is to evaluate the performance variations across different context length buckets and abilities. We did not perform extensive prompt engineering

[3] We use the GPT-3.5-Turbo-16K-0613 api from https://cuhk-api-dev1-apim1.developer.azure-api.net.

for each task to obtain the optimal performance. Instead, we focus on analyzing performance changes of particular LLMs with longer input contexts.

Since LLaMA 2 models were trained on data within 4k tokens, we used dynamic NTK-aware RoPE scaling (emozilla, 2023; Peng et al., 2023) for context longer than 4k. We used 16 floating points precision during inference. To facilitate fair comparisons across various tasks with different metrics, we normalized the raw performance score $r(M, l)$ (i.e., the performance of LLM M at context length l) as follows:

$$\hat{r}(M, l) = \frac{r(M, l)}{r(\text{GPT-3.5-Turbo-16K}, 1000) + r(M, l)}$$

$\hat{r}(M, l)$ provides a measure of how other models perform relative to GPT-3.5-Turbo-16K in the length range bucket of 0-1000 tokens, and how their performance deteriorates with longer input.

3.6 Results

Figure 3 illustrates the changes in normalized average scores for various evaluated models as context lengths extend, and Figure 4 depicts their ability in the context length range of 0-1000, 1000-4000, and 4000-8000 (the full results for each task can be found in Appendix C). Based on the figures, several key observations emerge:

The performance of all models significantly deteriorates with increasing context lengths. This trend is expected, given that a longer context might necessitate more sophisticated modelling capabilities. It suggests that these LLMs struggle with understanding extensive context. The performance gap between ChatGPT and most open-source models widens as context length increases. This is largely because open-source models tend to exhibit a steeper decline, particularly when the context length exceeds 4k. For example, Vicuna-13B-v1.5-16K achieves competitive performance, compared to GPT-3.5-Turbo-16K, in the 0-4K length range, but its performance drops significantly after that. A notable exception is ChatGLM2-6B-32k which achieves similar performance when testing on 6K and 8K instances and is only surpassed by GPT-3.5-Turbo-16K on 8K instances.

Fine-tuning with additional long context data does not offer a significant advantage over simply NTK scaling for understanding long contexts. Both Vicuna and LongChat models are claimed to support long context as they are directly fine-tuned with longer context data. However, their performance still drops quickly when context length exceeds 4k, with no additional advantage compared to LLaMA2 models, which are trained only on 4k data and merely equipped with NTK scaling method when context length exceeds 4k. This suggests that existing long-context fine-tuning methods contribute minimally to improving long context understanding and a more efficient and effective way to enhance long context understanding ability is needed.

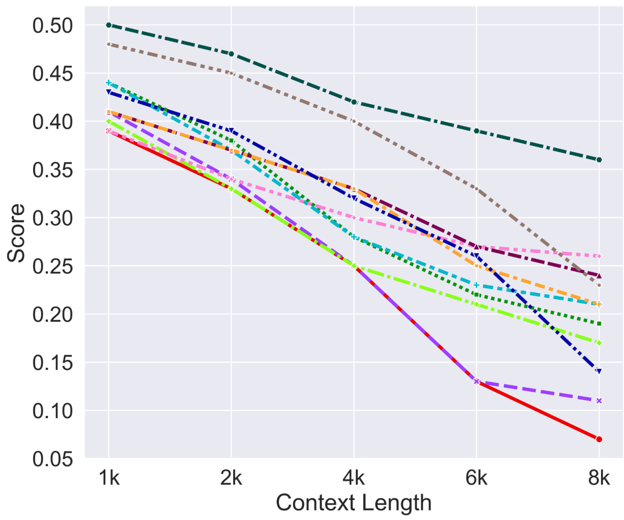
Multiple-span understanding is more difficult, and semantic retrieval is even harder for competent models. There is a significant drop in performance on tasks requiring multiple-span attention as context lengthens. This is expected since attending to multiple positions is naturally harder than a single position, and it might require additional ability to distinguish and determine compared to global understanding. Surprisingly, semantic retrieval is only more challenging for GPT-3.5-Turbo-16K, the most competent model in the experiment. We hypothesize that this is because explicit retrieval, looking for relative information by an identifier, is an unnatural task for less competent and generalized LLM. On the contrary, semantic retrieval is more similar to tasks like QA that these models experienced during instruction fine-tuning.

3.7 Ablation Study

We perform further analysis to understand how models behave in different languages and locations of the supporting document.

Impact of language differences on long-context understanding. Tasks in different languages may have distinct ability requirements due to the nature of languages and the effects of tokenization. While most models presented in this study are primarily trained on English data, we aim to assess the influence of language differences on the results. In Figure 5, we compare the performance of the top-performing models (namely ChatGPT, ChatGLM2, Vicuna, and LongChat) in both Chinese and English tasks to determine if their long-context understanding abilities differ across languages.

We observe a comparable decline in performance for both GPT-3.5-Turbo-16K and ChatGLM2-6B-32K across the two languages. However, the Vicuna and LongChat models exhibit a more pronounced performance drop in Chinese. This sug-



Model	Slope
LLaMA2-7B	-0.084
LLaMA2-7B-Chat	-0.081
LLaMA2-13B	-0.066
LLaMA2-13B-Chat	-0.060
ChatGLM2-6B	-0.058
ChatGLM2-6B-32K	-0.033
LongChat-7B-v1.5-32K	-0.044
LongChat-13B-16K	-0.052
Vicuna-7B-v1.5-16K	-0.071
Vicuna-13B-v1.5-16K	-0.062
GPT-3.5-Turbo-16K	-0.036

Figure 3: The normalized scores of various models in different context lengths (left), accompanied by the slopes of the corresponding best-fit lines (right). The performance of all models deteriorates with increasing context length.

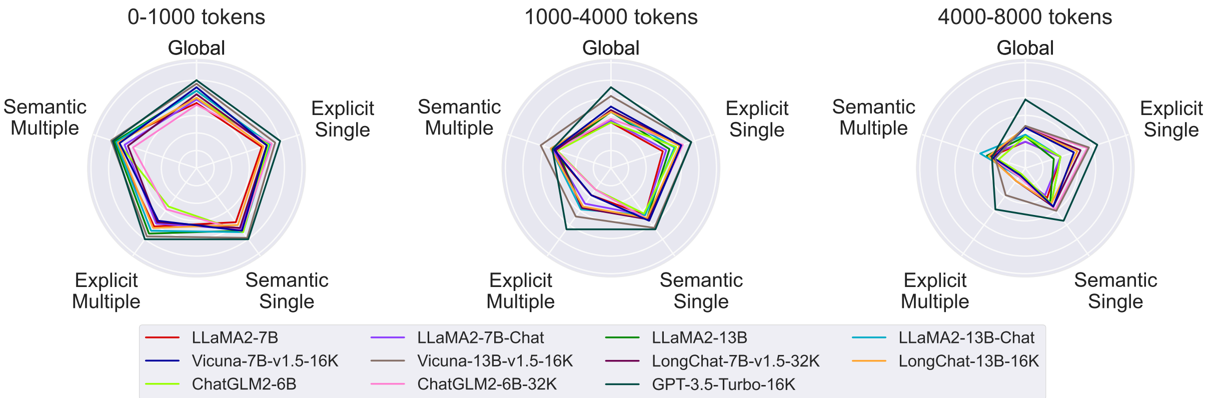


Figure 4: The comparison of abilities of various models in three context length ranges, respectively. It shows that multi-span understanding is more difficult in general. While semantic retrieval appears to be intuitively more challenging, our findings indicate that it is only more demanding for competent models such as GPT-3.5-Turbo-16K at longer lengths.

gests that the degradation of understanding ability when the context length increases is not unique to English. Furthermore, the diversity of data employed during fine-tuning, as highlighted by ChatGLM2's emphasis on its bilingual (Chinese and English) proficiency during its tuning process, appears to be a successful strategy in handling bilingual long context input.

Lost-in-the-middle exists in other NLP long sequence tasks. Recently, Liu et al. (2023) find that LLMs tend to ignore the information in the middle of long input context for the task of question-answering and retrieval. In this section, following the setup in Liu et al. (2023), we conduct a comprehensive experiment to study the impact of positions of the supporting paragraphs within the context based on our proposed M^{4}LE benchmark. Specifically, we generate additional instances from the tasks in M^{4}LE, each containing an identical input but with the supporting paragraph placed at different locations. We employ four datasets for question-answering and summarization, and two datasets for retrieval tasks. The setup details are in Appendix B.

The average score for each relative position of the supporting document across the three tasks is presented in Figure 6, demonstrating that models typically perform better when the supporting document is positioned either at the beginning or the end of the context, a finding consistent with Liu et al. (2023). Consequently, this suggests that the tendency for LLM to ignore information in the middle of the context is ubiquitous across various languages, models, and tasks. This also shows the potential of M^{4}LE in discovering interesting and unique LLMs behavior in the long context scenario.

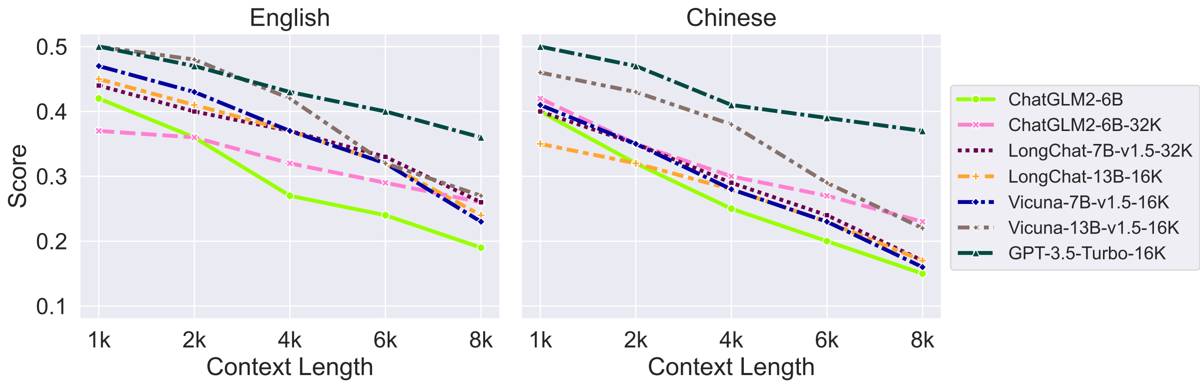


Figure 5: The normalized performance of the models fine-tuned in longer data for English and Chinese tasks, respectively. While GPT-3.5-Turbo-16K and ChatGLM2-6B-32K exhibit a similar trend in the decline of performance in both languages, other models demonstrate a more pronounced performance drop in Chinese tasks with increasing context lengths.

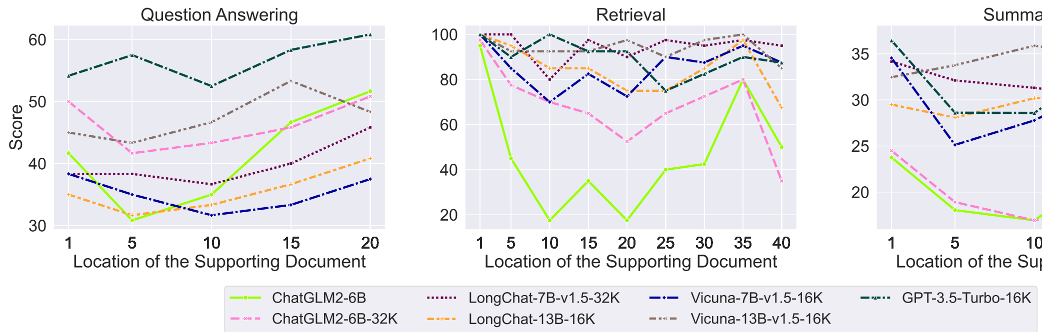


Figure 6: The performance of various models across three tasks, with the supporting document located at different relative positions. It shows higher performance is often obtained when the supporting document is positioned either at the beginning or the end, consistent with Liu et al. (2023).

4 Conclusion

In this paper, we propose M⁴LE for LLMs assessing their capability of long-context understanding. To establish a benchmark with diverse NLP tasks, rather than just those that are inherently lengthy, we propose a systematic method to convert short NLP task instances into long context inputs, encompassing five distinct abilities. We collect and construct in total of 36 tasks from different sources and domains covering multiple length ranges to maximize the diversity of the tasks in benchmark, with our customized construction methods which enable flexibility to extend arbitrary context lengths. We evaluate 11 well-known LLMs with our benchmark and find that current models struggle to understand long-context inputs and the corresponding performance related to ability types, data used when fine-tuning, and positions of the relevant information.

Limitations

Due to computational constraints, our experiments are restricted to smaller open-source models and lengths of up to 8K. Nevertheless, our method can create instances of arbitrary length (the released benchmark will include instances up to 32,000 words) and the analyses in this paper reveal meaningful observations concerning long-context understanding capabilties. Additionally, our study focuses on English and Chinese, the two most commonly used languages. We suggest future research to apply our methodology to construct long instances in additional languages.

Acknowledgements

This research work is partially supported by CUHK direct grant No. 4055209 and CUHK Knowledge Transfer Project Fund No. KPF23GWP20.

References

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

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.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

bloc97. 2023. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

emozilla. 2023. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: A Chinese Machine Reading Comprehension Dataset from Real-world Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Weixiong Hu, Zhaoquan Gu, Yushun Xie, Le Wang, and Keke Tang. 2019. Chinese text classification based on neural networks and word2vec. In *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pages 284–291.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

kaiokendev. 2023. Extending context is hard…but not impossible.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. WikiHow: A Large Scale Text Summarization Dataset.

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.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

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

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Xiaojun Liu, Shunan Zang, Chuang Zhang, Xiaojun Chen, and Yangyang Ding. 2022. CLTS+: A New Chinese Long Text Summarization Dataset with Abstractive Summaries. In *Artificial Neural Networks and Machine Learning – ICANN 2022*, Lecture Notes in Computer Science, pages 73–84, Cham. Springer International Publishing.

Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. 2020. CLTS: A New Chinese Long Text Summarization Dataset. In *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, pages 531–542, Cham. Springer International Publishing.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models.

Alina Petukhova and Nuno Fachada. 2023. MN-DS: A Multilabeled News Dataset for News Articles Hierarchical Classification. *Data*, 8(5):74.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*.

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.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. DRCD: A Chinese Machine Reading Comprehension Dataset.

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.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

SophonPlus. 2013. Chinesenlpcorpus. `https://github.com/SophonPlus/ChineseNlpCorpus`.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv.org*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do Long-Range Language Models Actually Use Long-Range Context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2023. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14590–14604, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, and Lei Li. 2021. CNewSum: A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural Cross-Lingual Summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

A Datasets

This section describes the datasets used in M^{4}LE. Table 2 provides an overview of the constructed datasets.

A.1 MNDS News

MNDS News (Petukhova and Fachada, 2023) is an English hierarchical news category classification dataset comprising 10,917 news articles from 260 sources. We only use the 17 first-level categories as the labels for this study. For multiple retrieval tasks, we randomly sample a class label that appears in the instance.

A.2 THUCNews

THUCNews (Hu et al., 2019) is a Chinese classification dataset containing 74 million news articles from Sina, with each article belonging to one of the ten categories. We filter out the articles with the number of words less than 20. The multiple retrieval task is built similarly to MNDS News.

A.3 MARC

MARC (Keung et al., 2020) is a dataset for the bilingual (English and Chinese) setting. It contains multilingual Amazon reviews with star ratings from 1 to 5, where 5 is the best. We use 1-star and 5-star reviews for negative and positive reviews respectively, and ask models to return all positive reviews.

A.4 Online Shopping

Online Shopping (SophonPlus, 2013) is a Chinese sentiment dataset containing 60K product reviews from Chinese e-commerce platforms. Each review is marked as positive or negative.

A.5 BIGPATENT

BIGPATENT (Sharma et al., 2019) consists of 1.3 million records of U.S. BIGPATENT documents across nine technological areas. The abstract of the document is used as the golden document summary.

A.6 CEPSUM

CEPSUM (Li et al., 2020) is a dataset containing product descriptions and summary pairs collected from a popular Chinese e-commerce platform. We removed instances with less than 60 words in the product description.

A.7 CNNNews

CNNNews (See et al., 2017) contains English online news articles from CNN, where each of it is paired with a multi-sentence summary. We preprocess the data using the script from See et al. (2017) and select the instances with at least 30 words in the article.

A.8 LCSTS

LCSTS (Hu et al., 2015) is a Chinese summarization dataset consisting of over 2 million posts and short summary pairs collected from the Chinese microblogging website Sina Weibo. We use part two of the data, which consists of 10,666 (text, summary) pairs with a human-labeled score to indicate the relevance between the post and the summary. The score ranges from 1 to 5, where 5 indicates the most relevant. We select only the samples with a score of 5 in the relevance score.

A.9 NCLS

NCLS (Zhu et al., 2019) is a cross-lingual summarization dataset consisting of pairs of articles in one language and summaries in another language (Chinese or English), constructed from the CNNNews and LCSTS datasets.

A.10 WikiHow

WikiHow (Koupaee and Wang, 2018) comprises 230,000 English articles that describe a procedural task along with corresponding summaries. Each article has a title that starts with "How to". The procedures described in the article are separated into multiple steps, where each step corresponds to a paragraph. Each paragraph has a short summary. These summaries are concatenated to form the summary of the article. We remove instances with articles that have less than 60 words.

A.11 News2016

News2016 (Xu, 2019), encompassing over 2 million Chinese news articles. Each article contains a title and keywords. The title is used as the golden summary of the news article. We remove instances with the number of words less than 200 and more than 800.

A.12 Arxiv

Arxiv (Cohan et al., 2018) consists of 215K academic papers from arXiv.org. The abstracts of the papers are used as the golden summary.

Ability	Dataset	Task Type	Language	Domain	Metric	Ave. Len.
Explicit Single	MNDS News	CLS + RET	En	News	Acc	3805
	THUCNews	CLS + RET	Zh	News	Acc	3650
	NewsQA	QA + RET	En	News	Acc	3679
	C3	QA + RET	Zh	Textbook	Acc	3797
	WoW	RET	En	Wiki	Acc	3434
	DRCD	RET	Zh	Wiki	Acc	3617
	CNNNews	SUM + RET	En	News	Rouge-L	3754
	CEPSUM	SUM + RET	Zh	E-Commerce	Rouge-L	4003
	LCSTS	SUM + RET	Zh	News	Rouge-L	4102
	NCLS	SUM + RET	En,Zh	News	Rouge-L	3470
Explicit Multiple	MNDS News	CLS + RET	En	News	F1	3772
	THUCNews	CLS + RET	Zh	News	F1	3721
	MARC	CLS + RET	En,Zh	E-Commerce	F1	3543
	Online Shopping	CLS + RET	Zh	E-Commerce	F1	3714
Semantic Single	WikiText-103	NLI + RET	En	Wiki	Acc	3278
	Wiki2019zh	NLI + RET	Zh	Wiki	Acc	3723
	DuoRC	QA	En	Movie	Acc	3572
	NQ-Open	QA	En	Wiki	Acc	3128
	DuReader	QA	Zh	Web	Rouge-L	3261
	DRCD	QA	Zh	Wiki	Acc	3300
	WikiHow	SUM + RET	En	WikiHow	Rouge-L	3514
	News2016	SUM + RET	Zh	News	Rouge-L	3785
	TedTalks	TRAN + RET	En,Zh	TedTalks	BLEU	2956
Semantic Multiple	MNDS News	CLS + CNT	En	News	Acc	3791
	THUCNews	CLS + CNT	Zh	News	Acc	3699
	HotpotQA	QA	En	Wiki	Acc	1060
Global	BIGPATENT	CLS	En	Patent	Acc	3407
	TriviaQA	QA	En	Web	Acc	3329
	Arixv	SUM	En	Academic	Rouge-L	3748
	BIGPATENT	SUM	En	Patent	Rouge-L	3293
	Pubmed	SUM	En	Medical	Rouge-L	3678
	Booksum	SUM	En	Book	Rouge-L	2643
	CNewsum	SUM	Zh	News	Rouge-L	1883
	CLTS+	SUM	Zh	News	Rouge-L	3158
	OpenSubtitles	TRAN	En,Zh	Movie	BLEU	2048
	News Commentary	TRAN	En,Zh	News	BLEU	3585

Table 2: The overview of the evaluated tasks in M^4LE, categorized by abilities. CLS, QA, RET, SUM, TRAN, and CNT denote classification, question-answering, retrieval, summarization, translation, and counting respectively. Acc in metric stands for accuracy.

A.13 Booksum

Booksum (Kryscinski et al., 2022), which includes 405 English books including plays, short stories, and novels with human-written summaries for each chapter. We combine the consecutive chapters and the corresponding summaries to construct instances for any context length range.

A.14 CNewsum

CNewsum (Wang et al., 2021) contains 304,307 Chinese news articles from different press publishers with human-written summaries.

A.15 CLTS+

CLTS+ (Liu et al., 2022) is an improved Chinese new articles summarization dataset based on CLTS (Liu et al., 2020). CLTS contains more than 180,000 Chinese long articles with human-written summaries. CLTS+ utilizes back translation to enhance the abstractiveness of the summaries.

A.16 NewsQA

NewsQA (Trischler et al., 2017) is an English QA dataset based on 12,744 news articles from CNN. Crowdsourced workers are recruited to generate 119,633 questions and answers.

A.17 C3

C3 (Sun et al., 2021) is a Chinese textbook-based machine comprehension dataset. The questions are multiple-choice questions collected from exams for second-language Chinese learners.

A.18 DuoRC

DuoRC (Saha et al., 2018) is an English question-answer dataset based on 7680 movie plots collected from IMDb and Wikipedia. Crowdsourced workers are hired to create 186,089 unique question-answer pairs.

A.19 NQ-Open

NaturalQuestions-Open (NQ-Open) (Lee et al., 2019) is an open-domain question-answering dataset based on Wikipedia documents. The questions are collected from Google Search queries. We directly use the processed version from Liu et al. (2023).

A.20 DuReader

DuReader (He et al., 2018) is an open-domain Chinese machine reading comprehension dataset, consisting of 200K questions collected from Baidu Search.

B Experiment Details for Lost-In-The-Middle

For the experiment in Figure 6, which explores the effects of the positions of the relevant paragraphs, we additionally construct the following instances:

In the QA task, 100 instances, each comprising 20 paragraphs, are generated from NQ-Open and DuoRC for English, and from DRCD and C3 for Chinese. Similarly, for the summarization task, we generate 100 instances each from WikiHow and CNNNews for English and News2016, and LCSTS for Chinese. For the retrieval task, we formulate 200 instances each using WoW for English and DRCD for Chinese. The supporting paragraph will be evenly placed at different locations.

C Main Results

We report the results used for plotting Figure 3.

D Task Results

We show the results of each task in Table 7 to 45

E Prompts

In this section, we describe the prompts used in M^{4}LE. The prompt begins with the task definition, followed by the in-context example and the testing instance. Below we show the prompt examples used for each of the five abilities. Other tasks' prompts are constructed similarly.

	1k	2k	4k	6k	8k
LLaMA2-7B	0.39	0.33	0.25	0.13	0.07
LLaMA2-7B-Chat	0.41	0.34	0.25	0.13	0.11
LLaMA2-13B	0.44	0.38	0.28	0.22	0.19
LLaMA2-13B-Chat	0.44	0.37	0.28	0.23	0.21
ChatGLM2-6B	0.40	0.33	0.25	0.21	0.17
ChatGLM2-6B-32K	0.39	0.34	0.30	0.27	0.26
LongChat-7B-v1.5-32K	0.41	0.37	0.33	0.27	0.24
LongChat-13B-16K	0.41	0.37	0.33	0.25	0.21
Vicuna-7B-v1.5-16K	0.43	0.39	0.32	0.26	0.14
Vicuna-13B-v1.5-16K	0.48	0.45	0.40	0.33	0.23
GPT-3.5-Turbo-16K	0.50	0.47	0.42	0.39	0.36

Table 3: The average normalized performance of different models in various lengths.

	Explicit Single	Semantic Single	Explicit Multiple	Semantic Multiple	Global
LLaMA2-7B	0.39	0.38	0.41	0.49	0.37
LLaMA2-7B-Chat	0.43	0.43	0.39	0.43	0.39
LLaMA2-13B	0.44	0.44	0.46	0.49	0.44
LLaMA2-13B-Chat	0.44	0.45	0.44	0.48	0.42
ChatGLM2-6B	0.43	0.45	0.27	0.47	0.40
ChatGLM2-6B-32K	0.45	0.44	0.29	0.38	0.36
LongChat-7B-v1.5-32K	0.42	0.42	0.38	0.41	0.42
LongChat-13B-16K	0.40	0.40	0.42	0.47	0.40
Vicuna-7B-v1.5-16K	0.42	0.44	0.37	0.46	0.46
Vicuna-13B-v1.5-16K	0.47	0.49	0.48	0.51	0.48
GPT-3.5-Turbo-16K	0.50	0.50	0.50	0.50	0.50

Table 4: Performance comparison of various models in different abilities over the 0-1000 tokens.

	Explicit Single	Semantic Single	Explicit Multiple	Semantic Multiple	Global
LLaMA2-7B	0.31	0.33	0.19	0.35	0.28
LLaMA2-7B-Chat	0.33	0.33	0.25	0.35	0.27
LLaMA2-13B	0.34	0.35	0.28	0.32	0.32
LLaMA2-13B-Chat	0.38	0.33	0.28	0.36	0.29
ChatGLM2-6B	0.39	0.32	0.15	0.32	0.26
ChatGLM2-6B-32K	0.43	0.36	0.15	0.33	0.28
LongChat-7B-v1.5-32K	0.42	0.36	0.28	0.33	0.33
LongChat-13B-16K	0.41	0.35	0.27	0.36	0.32
Vicuna-7B-v1.5-16K	0.42	0.37	0.19	0.34	0.35
Vicuna-13B-v1.5-16K	0.48	0.42	0.34	0.42	0.41
GPT-3.5-Turbo-16K	0.48	0.43	0.43	0.35	0.46

Table 5: Performance comparison of various models in different abilities over the 2000-4000 tokens
.

	Explicit Single	Semantic Single	Explicit Multiple	Semantic Multiple	Global
LLaMA2-7B	0.13	0.20	0.06	0.21	0.16
LLaMA2-7B-Chat	0.13	0.15	0.04	0.22	0.11
LLaMA2-13B	0.16	0.25	0.07	0.18	0.15
LLaMA2-13B-Chat	0.16	0.26	0.05	0.20	0.17
ChatGLM2-6B	0.24	0.24	0.04	0.16	0.18
ChatGLM2-6B-32K	0.40	0.31	0.06	0.22	0.23
LongChat-7B-v1.5-32K	0.30	0.24	0.09	0.17	0.22
LongChat-13B-16K	0.23	0.24	0.06	0.21	0.23
Vicuna-7B-v1.5-16K	0.24	0.23	0.05	0.13	0.22
Vicuna-13B-v1.5-16K	0.33	0.22	0.10	0.18	0.23
GPT-3.5-Turbo-16K	0.43	0.37	0.29	0.20	0.39

Table 6: Performance comparison of various models in different abilities over the 4000-8000 tokens
.

	1k	2k	4k	6k	8k
LLaMA2-7B	54.00	50.75	34.48	32.37	23.08
LLaMA2-7B-Chat	64.50	62.19	40.89	18.84	16.83
LLaMA2-13B	58.00	55.22	42.36	31.40	24.37
LLaMA2-13B-Chat	64.00	62.19	44.83	36.23	25.32
ChatGLM2-6B	49.00	37.81	31.53	23.67	16.83
ChatGLM2-6B-32K	46.50	46.27	36.95	28.99	35.10
LongChat-7B-v1.5-32K	59.50	57.21	49.75	47.34	37.50
LongChat-13B-16K	59.00	52.74	49.75	48.31	24.39
Vicuna-7B-v1.5-16K	61.00	59.70	50.74	44.93	31.73
Vicuna-13B-v1.5-16K	65.00	59.20	54.19	51.21	24.39
GPT-3.5-Turbo-16K	62.00	59.70	55.17	51.69	46.63

Table 7: NQ-Open (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	78.00	71.00	45.00	47.26	33.50
LLaMA2-7B-Chat	83.00	76.00	43.00	43.28	34.52
LLaMA2-13B	82.00	81.00	74.00	50.40	42.70
LLaMA2-13B-Chat	88.00	83.00	77.50	51.84	45.32
ChatGLM2-6B	79.00	74.00	67.50	56.22	41.00
ChatGLM2-6B-32K	81.50	74.50	69.50	72.14	67.00
LongChat-7B-v1.5-32K	81.00	77.50	70.50	77.61	72.00
LongChat-13B-16K	66.00	60.00	51.50	54.73	47.45
Vicuna-7B-v1.5-16K	85.00	84.50	80.50	83.58	73.50
Vicuna-13B-v1.5-16K	88.50	91.50	84.50	82.59	74.32
GPT-3.5-Turbo-16K	89.00	90.50	85.50	86.57	79.50

Table 8: DRCD (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	98.50	95.52	64.00	35.91	23.12
LLaMA2-7B-Chat	97.50	99.50	82.50	46.38	32.02
LLaMA2-13B	98.50	99.50	54.00	26.57	18.00
LLaMA2-13B-Chat	97.50	99.00	84.00	58.45	48.92
ChatGLM2-6B	97.00	93.03	65.00	32.37	15.87
ChatGLM2-6B-32K	93.50	91.54	86.50	74.88	54.33
LongChat-7B-v1.5-32K	98.50	99.50	97.50	97.10	76.92
LongChat-13B-16K	98.00	99.00	94.00	90.82	74.93
Vicuna-7B-v1.5-16K	98.50	99.50	94.50	91.79	64.42
Vicuna-13B-v1.5-16K	98.50	99.00	98.50	92.27	26.92
GPT-3.5-Turbo-16K	98.50	98.51	97.50	90.82	87.98

Table 9: WoW (RET)

	1k	2k	4k	6k	8k
LLaMA2-7B	99.00	99.50	62.50	46.43	31.96
LLaMA2-7B-Chat	100.00	97.51	65.00	42.37	32.39
LLaMA2-13B	99.50	99.50	52.00	48.70	35.88
LLaMA2-13B-Chat	98.50	99.50	75.50	52.56	41.02
ChatGLM2-6B	94.00	94.03	81.00	50.24	31.10
ChatGLM2-6B-32K	94.50	89.55	81.50	70.53	61.72
LongChat-7B-v1.5-32K	100.00	99.00	98.00	93.72	92.82
LongChat-13B-16K	98.00	94.03	91.00	85.51	81.49
Vicuna-7B-v1.5-16K	99.00	99.50	97.00	90.82	83.35
Vicuna-13B-v1.5-16K	100.00	99.50	98.00	96.14	85.79
GPT-3.5-Turbo-16K	100.00	98.51	99.00	89.37	87.08

Table 10: DRCD (RET)

	1k	2k	4k	6k	8k
LLaMA2-7B	11.62	12.96	11.72	8.46	3.57
LLaMA2-7B-Chat	14.19	14.68	16.79	8.40	4.59
LLaMA2-13B	13.51	13.24	12.34	9.38	5.86
LLaMA2-13B-Chat	13.47	13.56	13.96	11.46	5.93
ChatGLM2-6B	12.88	13.22	12.63	10.32	6.81
ChatGLM2-6B-32K	13.71	14.28	14.24	12.39	8.00
LongChat-7B-v1.5-32K	14.14	14.80	14.39	10.81	8.11
LongChat-13B-16K	11.94	13.42	13.48	8.75	7.15
Vicuna-7B-v1.5-16K	15.14	15.35	15.29	11.63	6.47
Vicuna-13B-v1.5-16K	14.28	14.81	14.07	8.37	6.92
GPT-3.5-Turbo-16K	18.00	16.98	15.65	12.18	10.86

Table 11: Booksum (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	87.50	88.50	84.00	73.00	65.00
LLaMA2-7B-Chat	86.00	86.50	76.00	64.00	63.50
LLaMA2-13B	90.50	92.00	82.00	75.50	61.00
LLaMA2-13B-Chat	90.50	89.00	80.50	73.00	66.00
ChatGLM2-6B	78.50	66.00	52.00	54.00	32.50
ChatGLM2-6B-32K	77.50	76.00	61.50	58.50	45.50
LongChat-7B-v1.5-32K	87.50	84.50	80.00	75.50	68.50
LongChat-13B-16K	85.00	86.50	75.00	75.50	50.00
Vicuna-7B-v1.5-16K	91.00	87.50	84.50	78.50	56.50
Vicuna-13B-v1.5-16K	88.50	85.00	80.00	77.00	50.00
GPT-3.5-Turbo-16K	89.50	83.00	82.00	77.00	73.50

Table 12: TriviaQA (QA)

	1k	2k
LLaMA2-7B	47.50	36.50
LLaMA2-7B-Chat	44.50	42.00
LLaMA2-13B	52.50	39.50
LLaMA2-13B-Chat	51.50	41.00
ChatGLM2-6B	43.50	31.50
ChatGLM2-6B-32K	41.50	35.00
LongChat-7B-v1.5-32K	49.50	40.50
LongChat-13B-16K	55.00	43.50
Vicuna-7B-v1.5-16K	50.00	44.50
Vicuna-13B-v1.5-16K	56.00	52.00
GPT-3.5-Turbo-16K	55.00	41.50

Table 13: HotpotQA (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	10.03	8.71	8.08	4.69	7.55
LLaMA2-7B-Chat	21.91	16.95	13.00	0.25	0.52
LLaMA2-13B	19.99	15.91	16.73	3.07	0.29
LLaMA2-13B-Chat	19.19	13.48	11.73	2.38	0.49
ChatGLM2-6B	16.82	14.48	11.78	10.35	7.01
ChatGLM2-6B-32K	20.76	20.18	18.22	14.43	14.97
LongChat-7B-v1.5-32K	22.18	23.60	23.81	14.81	18.46
LongChat-13B-16K	24.11	25.46	22.97	16.20	13.20
Vicuna-7B-v1.5-16K	23.59	23.39	21.28	19.06	8.22
Vicuna-13B-v1.5-16K	24.22	23.99	18.65	12.49	10.83
GPT-3.5-Turbo-16K	21.64	21.20	20.33	17.66	14.84

Table 14: Arxiv (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	24.17	23.81	25.28	19.44	14.66
LLaMA2-7B-Chat	29.89	26.48	24.41	14.14	13.02
LLaMA2-13B	30.95	32.29	21.61	16.36	13.32
LLaMA2-13B-Chat	25.05	21.74	20.69	12.94	11.92
ChatGLM2-6B	28.45	25.07	20.27	19.86	19.71
ChatGLM2-6B-32K	19.25	18.86	20.35	15.16	13.04
LongChat-7B-v1.5-32K	27.57	28.78	26.30	18.98	23.14
LongChat-13B-16K	24.77	26.33	24.47	23.34	28.07
Vicuna-7B-v1.5-16K	32.52	31.99	26.03	21.18	20.79
Vicuna-13B-v1.5-16K	33.41	31.40	26.63	14.40	12.54
GPT-3.5-Turbo-16K	28.65	23.13	19.25	16.97	17.36

Table 15: BIGPATENT (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	20.47	18.38	17.41	5.82	4.20
LLaMA2-7B-Chat	24.83	21.68	22.95	9.53	8.96
LLaMA2-13B	22.50	19.58	14.88	13.18	9.00
LLaMA2-13B-Chat	23.99	20.99	20.95	14.80	10.58
ChatGLM2-6B	23.07	20.42	16.81	16.39	15.74
ChatGLM2-6B-32K	22.13	19.25	18.57	17.72	17.53
LongChat-7B-v1.5-32K	25.92	23.51	20.52	14.96	17.83
LongChat-13B-16K	23.57	21.52	19.94	11.62	16.14
Vicuna-7B-v1.5-16K	27.63	23.65	23.53	19.24	16.77
Vicuna-13B-v1.5-16K	25.10	24.43	24.15	17.77	10.95
GPT-3.5-Turbo-16K	27.06	25.13	24.97	23.25	22.79

Table 16: Wikihow (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	16.70	29.24	19.15	4.42	2.08
LLaMA2-7B-Chat	13.50	17.87	4.11	2.18	1.93
LLaMA2-13B	36.68	31.98	25.90	4.44	1.21
LLaMA2-13B-Chat	22.73	22.09	11.42	7.06	3.12
ChatGLM2-6B	16.90	15.23	13.05	13.65	12.20
ChatGLM2-6B-32K	20.92	21.94	18.73	16.93	15.77
LongChat-7B-v1.5-32K	19.33	25.59	18.80	11.03	7.14
LongChat-13B-16K	22.55	32.76	23.39	9.13	4.25
Vicuna-7B-v1.5-16K	15.87	21.25	8.34	10.64	5.55
Vicuna-13B-v1.5-16K	23.44	27.54	18.40	9.45	9.60
GPT-3.5-Turbo-16K	16.91	20.81	15.95	13.68	12.40

Table 17: Pubmed (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	18.87	16.27	10.21	8.20	4.92
LLaMA2-7B-Chat	22.50	21.35	21.86	4.63	4.43
LLaMA2-13B	23.48	20.28	18.81	9.18	5.56
LLaMA2-13B-Chat	26.83	27.89	23.37	8.03	6.12
ChatGLM2-6B	24.96	20.87	9.54	2.28	0.53
ChatGLM2-6B-32K	23.39	22.91	24.64	22.35	25.76
LongChat-7B-v1.5-32K	24.47	24.58	24.07	19.53	13.33
LongChat-13B-16K	21.19	21.30	20.91	15.22	26.33
Vicuna-7B-v1.5-16K	24.71	25.92	24.31	17.50	18.67
Vicuna-13B-v1.5-16K	29.12	27.90	26.79	24.69	41.10
GPT-3.5-Turbo-16K	30.23	28.84	27.19	23.07	22.60

Table 18: NCLS (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	10.00	17.00	16.50	10.00	9.50
LLaMA2-7B-Chat	8.00	11.00	17.00	14.00	12.00
LLaMA2-13B	7.00	15.50	16.50	12.63	11.00
LLaMA2-13B-Chat	17.50	24.00	18.50	13.42	11.00
ChatGLM2-6B	14.00	21.50	14.50	9.00	5.00
ChatGLM2-6B-32K	6.00	7.00	6.50	5.50	4.00
LongChat-7B-v1.5-32K	16.50	15.00	13.00	11.00	6.50
LongChat-13B-16K	23.50	22.50	21.50	23.50	12.00
Vicuna-7B-v1.5-16K	22.00	14.50	17.00	10.00	6.00
Vicuna-13B-v1.5-16K	13.00	16.00	16.50	11.00	13.04
GPT-3.5-Turbo-16K	19.50	19.50	20.00	18.50	14.50

Table 19: BIGPATENT (CLS)

	1k	2k	4k	6k	8k
LLaMA2-7B	7.78	0.01	0.00	0.00	0.03
LLaMA2-7B-Chat	4.03	0.28	0.01	0.00	0.00
LLaMA2-13B	13.19	0.90	2.89	0.19	0.00
LLaMA2-13B-Chat	7.48	1.19	0.01	0.00	nan
ChatGLM2-6B	5.54	0.64	0.00	0.00	0.00
ChatGLM2-6B-32K	1.06	0.68	0.56	0.06	0.08
LongChat-7B-v1.5-32K	7.88	3.45	2.25	0.05	0.00
LongChat-13B-16K	5.60	1.82	0.59	0.00	0.00
Vicuna-7B-v1.5-16K	12.71	3.39	0.00	0.00	0.00
Vicuna-13B-v1.5-16K	15.56	11.69	6.55	0.02	0.00
GPT-3.5-Turbo-16K	21.60	20.01	19.40	16.32	11.17

Table 20: OpenSubtitles zh2en (TRAN)

	1k	2k	4k	6k	8k
LLaMA2-7B	6.14	0.95	0.00	0.00	0.00
LLaMA2-7B-Chat	8.30	3.04	0.73	0.20	0.00
LLaMA2-13B	9.17	3.68	1.40	0.21	0.01
LLaMA2-13B-Chat	12.77	8.00	0.97	0.00	0.00
ChatGLM2-6B	9.67	1.62	0.00	0.00	0.00
ChatGLM2-6B-32K	5.64	2.49	1.96	0.23	0.23
LongChat-7B-v1.5-32K	7.15	4.28	0.75	0.03	0.00
LongChat-13B-16K	4.69	2.61	2.06	0.58	0.00
Vicuna-7B-v1.5-16K	12.84	9.99	2.88	0.00	0.07
Vicuna-13B-v1.5-16K	15.60	13.52	10.05	2.23	0.00
GPT-3.5-Turbo-16K	20.61	21.18	23.13	21.28	19.57

Table 21: OpenSubtitles en2zh (TRAN)

	1k	2k	4k	6k	8k
LLaMA2-7B	9.50	4.98	3.50	2.46	0.48
LLaMA2-7B-Chat	14.50	4.98	0.50	0.00	0.00
LLaMA2-13B	11.50	8.96	1.00	0.99	0.00
LLaMA2-13B-Chat	15.50	3.48	0.00	0.99	0.00
ChatGLM2-6B	30.00	2.49	0.00	0.00	0.00
ChatGLM2-6B-32K	17.00	5.47	3.00	0.00	0.00
LongChat-7B-v1.5-32K	6.50	3.98	4.00	2.46	0.97
LongChat-13B-16K	13.50	4.98	6.00	5.91	0.00
Vicuna-7B-v1.5-16K	14.00	10.95	6.00	2.46	0.00
Vicuna-13B-v1.5-16K	40.00	23.88	7.00	0.00	0.00
GPT-3.5-Turbo-16K	38.00	22.89	11.50	5.91	5.31

Table 22: WikiText-103 (NLI)

	1k	2k	4k	6k	8k
LLaMA2-7B	24.00	22.00	1.49	0.50	0.49
LLaMA2-7B-Chat	39.00	30.00	0.50	0.50	0.00
LLaMA2-13B	47.50	22.50	0.50	4.00	0.00
LLaMA2-13B-Chat	66.00	5.00	0.00	0.00	0.00
ChatGLM2-6B	47.00	15.50	2.49	8.00	0.00
ChatGLM2-6B-32K	51.50	25.00	6.97	5.00	1.96
LongChat-7B-v1.5-32K	47.00	15.00	1.49	2.50	0.98
LongChat-13B-16K	21.50	23.50	1.00	5.00	0.00
Vicuna-7B-v1.5-16K	37.50	4.50	0.00	0.50	0.00
Vicuna-13B-v1.5-16K	75.00	26.00	3.48	0.00	0.00
GPT-3.5-Turbo-16K	77.50	58.00	4.98	12.50	4.41

Table 23: Wiki2019zh (NLI)

	1k	2k	4k	6k	8k
LLaMA2-7B	57.44	33.21	13.73	6.94	6.45
LLaMA2-7B-Chat	31.62	18.03	17.74	9.19	5.43
LLaMA2-13B	54.87	35.51	19.43	2.58	1.12
LLaMA2-13B-Chat	59.10	45.35	22.91	7.89	3.13
ChatGLM2-6B	45.35	34.06	9.15	8.68	5.87
ChatGLM2-6B-32K	20.92	10.02	17.49	15.33	12.09
LongChat-7B-v1.5-32K	48.47	43.76	32.78	25.66	21.02
LongChat-13B-16K	55.77	50.73	37.16	26.45	23.00
Vicuna-7B-v1.5-16K	52.23	43.40	30.19	18.55	10.60
Vicuna-13B-v1.5-16K	61.13	54.82	43.19	33.21	21.38
GPT-3.5-Turbo-16K	73.07	63.61	48.60	39.22	22.59

Table 24: MNDS News (CLS, Explicit Multiple)

	1k	2k	4k	6k	8k
LLaMA2-7B	60.00	36.32	17.16	16.18	10.29
LLaMA2-7B-Chat	30.00	27.86	22.55	19.12	12.00
LLaMA2-13B	50.50	20.90	16.18	16.18	11.72
LLaMA2-13B-Chat	43.50	43.78	26.96	28.89	19.57
ChatGLM2-6B	47.50	34.33	17.65	15.20	15.69
ChatGLM2-6B-32K	14.00	32.84	16.18	15.20	19.61
LongChat-7B-v1.5-32K	32.50	18.41	23.04	24.02	12.25
LongChat-13B-16K	50.50	41.79	21.08	22.55	37.50
Vicuna-7B-v1.5-16K	39.50	31.84	25.98	20.10	10.78
Vicuna-13B-v1.5-16K	55.00	47.76	26.96	13.24	10.30
GPT-3.5-Turbo-16K	54.50	39.80	17.65	19.61	12.25

Table 25: MNDS News (CLS, Semantic Multiple)

	1k	2k	4k	6k	8k
LLaMA2-7B	29.00	19.92	12.00	6.35	2.19
LLaMA2-7B-Chat	38.05	29.21	19.89	8.34	0.02
LLaMA2-13B	47.18	43.22	16.05	2.65	0.00
LLaMA2-13B-Chat	48.73	42.74	25.36	4.91	0.00
ChatGLM2-6B	20.88	7.60	4.67	2.46	2.55
ChatGLM2-6B-32K	20.54	8.85	6.01	0.22	0.00
LongChat-7B-v1.5-32K	34.88	30.98	26.39	6.88	0.00
LongChat-13B-16K	51.43	44.99	30.75	7.94	0.00
Vicuna-7B-v1.5-16K	33.63	29.48	6.49	0.23	0.00
Vicuna-13B-v1.5-16K	66.40	45.69	32.44	21.28	11.65
GPT-3.5-Turbo-16K	65.58	49.92	33.37	23.50	14.25

Table 26: MARC (CLS)

	1k	2k	4k	6k	8k
LLaMA2-7B	31.60	21.02	22.52	17.92	15.95
LLaMA2-7B-Chat	32.01	27.26	18.19	15.48	11.87
LLaMA2-13B	40.79	33.70	27.80	16.87	12.38
LLaMA2-13B-Chat	31.89	25.69	22.84	18.72	11.76
ChatGLM2-6B	31.44	22.57	20.92	17.84	15.68
ChatGLM2-6B-32K	37.68	30.31	29.33	22.77	24.71
LongChat-7B-v1.5-32K	30.79	28.92	23.22	15.25	9.19
LongChat-13B-16K	26.88	24.92	23.17	14.93	12.08
Vicuna-7B-v1.5-16K	32.74	29.45	25.10	16.76	11.08
Vicuna-13B-v1.5-16K	35.06	32.61	31.64	23.05	19.37
GPT-3.5-Turbo-16K	32.28	29.77	25.12	23.19	23.04

Table 27: DuReader (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	23.80	6.10	0.72	0.09	0.05
LLaMA2-7B-Chat	26.39	17.88	11.14	4.67	0.00
LLaMA2-13B	43.50	22.64	10.20	2.85	0.00
LLaMA2-13B-Chat	32.73	23.59	14.12	3.59	0.00
ChatGLM2-6B	1.69	0.37	0.57	0.00	0.00
ChatGLM2-6B-32K	10.22	3.87	0.89	0.00	0.00
LongChat-7B-v1.5-32K	28.13	19.17	10.14	4.72	0.00
LongChat-13B-16K	27.78	16.21	3.11	1.28	0.00
Vicuna-7B-v1.5-16K	19.58	6.93	0.20	0.10	0.43
Vicuna-13B-v1.5-16K	40.92	27.95	7.15	4.18	3.76
GPT-3.5-Turbo-16K	34.84	31.15	19.03	14.29	10.23

Table 28: Online Shopping (CLS)

	1k	2k	4k	6k	8k
LLaMA2-7B	67.17	33.62	20.27	7.54	4.00
LLaMA2-7B-Chat	64.12	31.26	14.43	1.29	0.00
LLaMA2-13B	58.83	34.57	16.17	4.71	0.00
LLaMA2-13B-Chat	49.83	19.02	3.03	0.37	0.00
ChatGLM2-6B	51.08	36.49	25.11	10.41	2.07
ChatGLM2-6B-32K	67.03	40.79	16.10	10.50	5.99
LongChat-7B-v1.5-32K	39.75	22.85	9.40	2.97	0.00
LongChat-13B-16K	44.00	15.12	6.97	1.10	2.96
Vicuna-7B-v1.5-16K	45.75	21.52	5.87	1.33	0.00
Vicuna-13B-v1.5-16K	55.33	36.70	27.50	23.34	13.70
GPT-3.5-Turbo-16K	75.75	77.28	59.08	47.32	44.98

Table 29: THUCNews (CLS, Explicit Multiple)

	1k	2k	4k	6k	8k
LLaMA2-7B	54.00	50.00	21.50	21.08	16.67
LLaMA2-7B-Chat	59.50	35.00	30.50	19.61	20.59
LLaMA2-13B	63.50	38.50	24.50	20.76	19.52
LLaMA2-13B-Chat	60.50	24.00	26.50	18.82	17.00
ChatGLM2-6B	60.00	46.50	14.00	8.33	4.90
ChatGLM2-6B-32K	61.00	38.00	23.00	13.24	15.69
LongChat-7B-v1.5-32K	38.50	29.50	30.00	13.73	0.00
LongChat-13B-16K	46.50	38.50	22.00	10.78	16.67
Vicuna-7B-v1.5-16K	58.50	30.00	17.00	6.37	0.00
Vicuna-13B-v1.5-16K	64.50	56.50	27.50	7.84	0.00
GPT-3.5-Turbo-16K	61.00	44.50	18.50	14.71	11.27

Table 30: THUCNews (CLS, Semantic Multiple)

	1k	2k	4k	6k	8k
LLaMA2-7B	31.00	21.50	23.50	14.00	9.45
LLaMA2-7B-Chat	45.00	31.50	21.00	19.00	4.49
LLaMA2-13B	62.50	45.00	32.00	10.00	5.08
LLaMA2-13B-Chat	63.00	49.00	34.50	14.50	3.07
ChatGLM2-6B	38.00	26.50	16.00	7.50	4.50
ChatGLM2-6B-32K	55.50	52.00	42.50	33.50	29.85
LongChat-7B-v1.5-32K	24.00	27.50	21.00	19.00	10.31
LongChat-13B-16K	26.50	34.50	30.00	20.00	12.42
Vicuna-7B-v1.5-16K	37.00	36.00	32.50	19.00	11.39
Vicuna-13B-v1.5-16K	61.00	61.00	61.50	36.50	20.00
GPT-3.5-Turbo-16K	67.50	68.50	69.50	51.50	38.31

Table 31: THUCNews (CLS, Explicit Single)

	1k	2k	4k	6k	8k
LLaMA2-7B	18.50	14.00	5.00	4.48	3.50
LLaMA2-7B-Chat	30.50	22.50	10.50	11.44	3.50
LLaMA2-13B	33.50	35.50	8.50	7.46	2.00
LLaMA2-13B-Chat	35.50	36.50	15.00	10.95	5.50
ChatGLM2-6B	17.00	17.50	6.00	3.48	3.00
ChatGLM2-6B-32K	26.00	29.50	22.00	19.40	22.00
LongChat-7B-v1.5-32K	29.00	31.00	20.50	23.88	17.00
LongChat-13B-16K	32.00	34.00	31.00	15.47	11.00
Vicuna-7B-v1.5-16K	30.00	27.50	21.50	17.41	15.00
Vicuna-13B-v1.5-16K	40.50	38.50	34.50	20.40	16.50
GPT-3.5-Turbo-16K	41.50	41.50	33.00	26.37	17.50

Table 32: MNDS News (CLS, Explicit Single)

	1k	2k	4k	6k
LLaMA2-7B	17.67	12.48	9.66	3.04
LLaMA2-7B-Chat	22.57	12.09	11.03	4.18
LLaMA2-13B	18.69	13.45	10.59	5.72
LLaMA2-13B-Chat	23.09	15.51	11.46	9.70
ChatGLM2-6B	28.61	14.23	10.56	9.45
ChatGLM2-6B-32K	28.13	18.41	11.73	7.54
LongChat-7B-v1.5-32K	21.11	14.99	11.63	7.21
LongChat-13B-16K	19.61	12.55	10.20	10.57
Vicuna-7B-v1.5-16K	17.09	14.54	12.07	20.21
Vicuna-13B-v1.5-16K	20.76	15.95	13.31	11.92
GPT-3.5-Turbo-16K	28.32	18.11	14.85	13.74

Table 33: CNewsum (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	37.12	26.96	24.15	10.31	8.68
LLaMA2-7B-Chat	36.83	31.13	12.40	11.31	7.94
LLaMA2-13B	33.86	28.09	20.15	12.96	9.20
LLaMA2-13B-Chat	34.12	26.76	23.76	17.05	10.34
ChatGLM2-6B	37.26	23.70	10.97	8.89	10.06
ChatGLM2-6B-32K	38.11	34.49	32.31	29.36	26.12
LongChat-7B-v1.5-32K	39.25	32.58	26.72	23.24	19.26
LongChat-13B-16K	37.34	32.63	26.10	23.62	19.00
Vicuna-7B-v1.5-16K	34.73	30.68	27.81	17.40	20.11
Vicuna-13B-v1.5-16K	34.16	30.03	27.68	10.56	9.88
GPT-3.5-Turbo-16K	37.81	32.25	30.26	26.23	25.09

Table 34: CLTS+ (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	20.99	20.96	16.51	9.00	8.88
LLaMA2-7B-Chat	20.58	19.72	16.87	10.08	7.75
LLaMA2-13B	21.30	20.92	14.27	7.71	4.00
LLaMA2-13B-Chat	21.22	19.83	17.50	8.50	3.83
ChatGLM2-6B	25.08	24.62	20.53	17.22	14.85
ChatGLM2-6B-32K	22.77	23.36	22.19	21.99	21.69
LongChat-7B-v1.5-32K	21.28	21.16	21.08	15.63	4.56
LongChat-13B-16K	20.48	21.11	20.57	12.52	8.00
Vicuna-7B-v1.5-16K	22.21	21.05	19.97	15.67	4.99
Vicuna-13B-v1.5-16K	21.70	21.72	21.98	21.65	11.29
GPT-3.5-Turbo-16K	25.08	24.56	24.52	22.51	22.19

Table 35: CEPSUM (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	18.75	15.32	13.38	11.23	9.84
LLaMA2-7B-Chat	16.69	9.00	3.98	2.12	3.23
LLaMA2-13B	17.71	15.68	7.67	5.06	5.31
LLaMA2-13B-Chat	9.90	9.37	5.14	4.48	3.12
ChatGLM2-6B	10.84	18.96	14.35	14.14	10.39
ChatGLM2-6B-32K	18.86	18.26	19.39	18.49	17.89
LongChat-7B-v1.5-32K	12.74	15.36	17.57	29.64	3.59
LongChat-13B-16K	10.41	11.74	16.29	12.32	4.85
Vicuna-7B-v1.5-16K	14.15	19.49	21.00	12.65	5.52
Vicuna-13B-v1.5-16K	18.46	21.13	19.08	17.37	15.32
GPT-3.5-Turbo-16K	13.39	12.35	11.70	14.23	11.27

Table 36: CNNNews (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	5.00	10.15	11.64	7.03	4.20
LLaMA2-7B-Chat	10.04	7.44	3.49	2.13	2.88
LLaMA2-13B	11.75	9.14	10.58	8.88	7.06
LLaMA2-13B-Chat	9.84	6.27	8.39	8.34	5.12
ChatGLM2-6B	13.91	13.99	15.63	12.42	12.93
ChatGLM2-6B-32K	13.21	15.08	12.26	12.10	11.86
LongChat-7B-v1.5-32K	10.14	9.95	9.84	6.96	3.08
LongChat-13B-16K	8.78	9.17	13.77	7.53	1.21
Vicuna-7B-v1.5-16K	10.89	11.51	12.07	8.88	2.16
Vicuna-13B-v1.5-16K	9.75	13.49	20.83	12.42	10.70
GPT-3.5-Turbo-16K	16.72	15.51	15.88	15.35	16.45

Table 37: News2016 (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	22.34	18.79	9.45	8.31	4.36
LLaMA2-7B-Chat	19.78	20.01	11.21	9.41	5.39
LLaMA2-13B	20.62	16.49	5.05	3.26	4.31
LLaMA2-13B-Chat	22.19	19.63	11.53	8.41	7.12
ChatGLM2-6B	23.78	26.44	17.99	11.52	8.16
ChatGLM2-6B-32K	21.24	14.47	16.09	14.31	11.50
LongChat-7B-v1.5-32K	21.34	19.03	19.89	20.91	7.03
LongChat-13B-16K	19.41	17.68	15.97	14.25	12.02
Vicuna-7B-v1.5-16K	21.70	20.32	22.22	14.91	9.46
Vicuna-13B-v1.5-16K	21.93	21.84	20.60	28.16	23.15
GPT-3.5-Turbo-16K	27.46	27.34	21.02	12.98	11.97

Table 38: LCSTS (SUM)

	1k	2k	4k	6k	8k
LLaMA2-7B	31.50	28.00	23.88	16.00	5.45
LLaMA2-7B-Chat	35.50	30.50	19.90	14.50	8.98
LLaMA2-13B	37.50	34.00	29.85	7.50	5.00
LLaMA2-13B-Chat	44.50	42.50	34.33	16.83	13.21
ChatGLM2-6B	71.00	66.50	61.19	58.00	53.43
ChatGLM2-6B-32K	72.50	70.50	63.18	65.00	68.14
LongChat-7B-v1.5-32K	30.00	30.00	25.87	26.50	10.26
LongChat-13B-16K	23.00	29.00	24.38	30.00	18.96
Vicuna-7B-v1.5-16K	34.50	27.50	26.87	21.00	12.82
Vicuna-13B-v1.5-16K	56.00	49.50	52.74	50.00	30.98
GPT-3.5-Turbo-16K	85.00	84.00	81.09	76.00	74.02

Table 39: C3 (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	2.50	1.00	0.00	1.99	4.50
LLaMA2-7B-Chat	7.50	2.00	0.50	2.49	7.50
LLaMA2-13B	6.00	3.50	2.00	1.00	0.00
LLaMA2-13B-Chat	7.50	7.50	4.50	3.30	0.00
ChatGLM2-6B	8.50	7.00	8.00	5.47	4.00
ChatGLM2-6B-32K	9.50	8.00	9.00	6.97	8.00
LongChat-7B-v1.5-32K	13.50	16.00	15.50	14.93	4.84
LongChat-13B-16K	8.50	7.50	16.00	11.44	7.50
Vicuna-7B-v1.5-16K	7.50	11.00	8.00	6.97	1.61
Vicuna-13B-v1.5-16K	11.00	19.00	24.50	14.93	4.00
GPT-3.5-Turbo-16K	18.00	16.00	13.00	14.43	18.50

Table 40: NewsQA (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	38.00	31.00	26.50	19.00	10.50
LLaMA2-7B-Chat	41.00	37.00	34.00	24.00	10.00
LLaMA2-13B	41.00	36.00	29.00	24.00	12.50
LLaMA2-13B-Chat	42.50	42.50	34.50	30.50	18.00
ChatGLM2-6B	35.50	27.00	12.00	12.00	14.00
ChatGLM2-6B-32K	31.50	32.50	29.50	27.00	27.50
LongChat-7B-v1.5-32K	43.50	42.50	37.50	33.00	16.50
LongChat-13B-16K	43.00	37.00	35.50	32.00	17.50
Vicuna-7B-v1.5-16K	42.00	40.50	35.00	31.00	20.00
Vicuna-13B-v1.5-16K	39.50	38.00	36.00	9.50	11.50
GPT-3.5-Turbo-16K	39.50	36.50	32.50	31.00	32.50

Table 41: Duorc (QA)

	1k	2k	4k	6k	8k
LLaMA2-7B	10.27	6.66	2.20	2.01	0.69
LLaMA2-7B-Chat	8.83	5.13	1.37	1.13	0.40
LLaMA2-13B	20.99	12.85	2.92	1.78	0.72
LLaMA2-13B-Chat	15.93	9.24	3.64	2.58	1.32
ChatGLM2-6B	12.85	7.61	0.28	0.69	0.38
ChatGLM2-6B-32K	13.44	5.05	3.60	3.37	3.22
LongChat-7B-v1.5-32K	14.10	10.97	8.00	6.39	4.78
LongChat-13B-16K	10.40	8.85	5.13	4.54	3.24
Vicuna-7B-v1.5-16K	19.88	20.31	8.61	7.74	3.17
Vicuna-13B-v1.5-16K	27.31	22.04	13.88	9.82	5.13
GPT-3.5-Turbo-16K	33.30	28.38	24.33	23.94	18.48

Table 42: News Commentary en2zh (TRAN)

	1k	2k	4k	6k	8k
LLaMA2-7B	13.28	7.42	0.89	0.22	0.01
LLaMA2-7B-Chat	8.16	4.01	0.50	0.32	0.09
LLaMA2-13B	20.28	13.89	2.43	1.38	0.34
LLaMA2-13B-Chat	8.83	7.19	2.53	1.56	0.58
ChatGLM2-6B	6.80	7.51	0.16	0.04	0.02
ChatGLM2-6B-32K	5.55	7.32	1.14	2.26	2.21
LongChat-7B-v1.5-32K	15.01	9.61	7.31	2.91	3.08
LongChat-13B-16K	12.82	9.55	4.18	2.30	1.13
Vicuna-7B-v1.5-16K	17.64	15.14	10.58	6.76	2.35
Vicuna-13B-v1.5-16K	20.17	17.43	12.88	11.32	7.35
GPT-3.5-Turbo-16K	26.23	22.22	17.99	15.94	13.12

Table 43: News Commentary zh2en (TRAN)

	1k	2k	4k	6k	8k
LLaMA2-7B	9.30	6.21	1.01	0.91	1.05
LLaMA2-7B-Chat	15.20	9.40	3.05	2.17	0.88
LLaMA2-13B	14.58	10.47	2.71	3.00	2.14
LLaMA2-13B-Chat	13.94	10.78	2.16	3.09	2.32
ChatGLM2-6B	14.86	0.98	0.07	0.02	0.00
ChatGLM2-6B-32K	13.67	5.19	1.84	1.17	1.18
LongChat-7B-v1.5-32K	20.43	9.78	4.23	2.93	3.03
LongChat-13B-16K	6.43	5.50	2.91	2.06	2.83
Vicuna-7B-v1.5-16K	23.75	11.36	5.93	2.01	3.23
Vicuna-13B-v1.5-16K	22.52	20.22	9.77	4.03	3.12
GPT-3.5-Turbo-16K	25.84	22.48	13.99	9.84	9.39

Table 44: Tedtalks en2zh (TRAN)

	1k	2k	4k	6k	8k
LLaMA2-7B	13.82	5.32	0.25	0.00	0.00
LLaMA2-7B-Chat	17.49	5.26	1.99	0.93	0.00
LLaMA2-13B	19.94	5.55	1.75	0.00	0.00
LLaMA2-13B-Chat	17.37	5.74	2.64	0.00	0.00
ChatGLM2-6B	13.22	4.26	1.03	0.19	0.05
ChatGLM2-6B-32K	9.72	2.91	1.53	1.77	1.31
LongChat-7B-v1.5-32K	12.06	2.01	0.43	0.09	0.00
LongChat-13B-16K	14.78	2.05	0.99	1.11	0.82
Vicuna-7B-v1.5-16K	20.46	5.97	1.97	2.83	1.32
Vicuna-13B-v1.5-16K	24.07	11.94	7.27	5.74	3.13
GPT-3.5-Turbo-16K	16.14	10.86	9.32	7.85	4.46

Table 45: Tedtalks zh2en (TRAN)

You are given multiple news articles below. Each of them belongs to one of the following categories:
1. crime, law and justice
2. arts, culture, entertainment and media
3. economy, business and finance
4. disaster, accident and emergency incident
5. environment
6 education
7. health
8. human interest
9. lifestyle and leisure
10. politics
11. labour
12. religion and belief
13. science and technology
14. society
15. sport
16. conflict, war and peace
17. weather

You will be asked to return the category of a news article I specified at the end.
Article AD3258: Rarely do the worlds of art and science intersect, but they did with famed Dutch artist Escher. Even if you if do not recognize his name, it is likely you have seen his work without knowing it. One of the largest collections of his work is now on display in the US. Article D55E47: On Sunday, NBC's Meet The Press will air an interview with President Donald Trump, conducted by the network's political director, Chuck Todd... Article 5675E9: The full extent of the ferry disaster in the Iraqi city of Mosul is becoming clearer... Question: What is the category of article AD3258?
Answer: arts, culture, entertainment and media

Article 11BD15: Read the full article by Catherine Frompovitch at NaturalBlaze
Abstract
The human cytochrome P450 (CYP) superfamily comprises 57 genes. These genes code for enzymes that can have a role in: metabolism of drugs, foreign chemicals, arachidonic acid and eicosanoids; cholesterol metabolism and bile-acid biosynthesis; steroid synthesis and metabolism; vitamin D(3) synthesis and metabolism; retinoic acid hydroxylation;...
Article 92FF60: Undoubtedly, this latest flooding crisis in Iran reveals the highly vicious nature of the current U.S. Administration with regards to the application of collective punishment of a target nation...
...
Question: What is the category of article 11BD15?
Answer:

Figure 7: An example prompt for the explicit single retrieval task based on MNDS.

Below are some articles from wikihow. I will ask you to summarize a particular article at the end.
Article 1: It's style that will leave you looking classy and feminine. This is a twist on the classic messier and more mermaid-like. This more obscure braid requires skill, but results in an interesting look. It's cute, classic, and easy to do. It looks bit medieval and very eye-catching. It's ideal for weddings or other elegant occasions.
Article 2: It's a green app that contains a white phone icon inside a white text bubble. It's at the topcenter of the screen. Select the chat with the attachment you wish to download. Select the attachment you wish to download. It's in the upper-right corner of the screen. The attachment has been saved to your Android device.
Question: Summarize the article related to "How to Style Very Long Hair" using a few instructive sentences.
Summary: Do a French braid. Make an intricate fishtail braid. Try a Dutch braid. Do a triple braid. Make a crazy braid. Do a cascading waterfall braid.

Article 1: Make sure that the shoe is the appropriate length and width for your child's foot. If a shoe squeezes a child's foot too much, it can cause the child to have foot conditions such as blisters and calluses. Remember that your child's foot will grow at a rapid pace and that he may need to be fitted every few months for a new size. So, if your child takes off his shoe and you notice that there are red marks on your child's foot, it may be time to take your child in for a new fitting and buy him a new shoe...
Article 2: Learning and studying shouldn't be stressful. Being stressed out can actually make it harder to learn and remember things. Think about the reasons why you're stressed out and try to resolve those reasons (remove them from your life). For example, if you get stressed out about assignments because you leave them to the last minute to finish, create yourself a study schedule. Build enough time into the study schedule so that you finish your assignments well enough in advance of the due dates to eliminate any of the stress you were feeling. If the grades you're receiving aren't that great it can be easy to let negativity take over.
Question: Summarize the article related to "How to Fit Your Kid for Shoes" using a few instructive sentences.
Summary:

Figure 8: An example prompt for the semantic single retrieval task based on Wikihow.

You are given multiple news articles below where each of them belongs to one of the 17 categories. Each article is prefixed with a article id. You will be asked to return the article ids of all articles belong to a particular category.
The article id is 0A1A04. Rarely do the worlds of art and science intersect, but they did with famed Dutch artist Escher. Even if you do not recognize his name, it is likely you have seen his work without knowing it. One of the largest collections of his work is now on display in the US.
The article id is 95A4BF. On Sunday, NBC's Meet The Press will air an interview with President Donald Trump, conducted by the network's political director, Chuck Todd. While Todd's interviews with 2020 Democratic contenders have consisted largely of challenges from the left interspersed with the odd softball, Trump is unlikely to receive the same friendly treatment.
The article id is A7D6BE. The full extent of the ferry disaster in the Iraqi city of Mosul is becoming clearer. Civil Defence says the number of dead is now at least 120, while 100 people are still missing. Iraq's Prime Minister Adel Abdul Mahdi is formally requesting a local governor be sacked over the incident.
Question: Provide me the article id of all the news articles related to 'arts, culture, entertainment and media.'
Answer: 0A1A04, 95A4BF.

The article id AE8707. CNN contributor Ana Navarro accused President Donald Trump of being the "enemy" of conservative principles and, indeed, the "American presidency" altogether Friday on CNN...
The article id 5BC439. A widely-anticipated exchange of prisoners between Russia and Ukraine is under way, according to reports. Buses from a Moscow prison believed to be carrying Ukrainian prisoners arrived at the capital's Vnukovo airport on Saturday. . .
The article id 638F02. Dec. 18 (UPI) Thousands of nurses in Northern Ireland walked out in a 12hour labor strike Wednesday, rallying for better pay and greater patient safety. A total of about 15,000 nurses participated in the walkout...
...
Question: Provide me the article id of all the news articles related to 'society'.
Answer:

Figure 9: An example prompt for the explicit multiple retrieval task based on MNDS.

Answer the question based on the given paragraphs. Note that some paragraphs might be irrelevant.
Paragraph 1: Pratia is a genus of flowering plants in the family Campanulaceae, native to Asia, Australia and New Zealand.
Paragraph 2: Sutherlandia is a genus of flowering plants in the family Fabaceae.
Question: Are Sutherlandia and Pratia in the same family?
Answer: no.

Paragraph 1: The Stresa Festival Orchestra is a formation composed by young and talented musicians, coming from renewed european orchestras, calling by Gianandrea Noseda to perform every year some original production for the Stresa Festival. The debut of the Orchestra, on 26 August 2003 with Mozart' "Don Giovanni", began the project of the concert performances of different operas: "Così fan tutte" (2004), "Le nozze di Figaro" (2005), "The magic flute" (2006), "La clemenza di Tito" (2007), ...
Paragraph 2: The Metropolitan City of Messina (Italian: "Città metropolitana di Messina") is a metropolitan city in Sicily, Italy. Its capital is the city of Messina. It replaced the Province of Messina and comprises the city of Messina and other 107 municipalities ("comuni"). According to Eurostat the FUA of the metropolitan area of Messina has in 2014 277,584 inhabitants.
Paragraph 3: Pompei is a city and "comune" in the Metropolitan City of Naples in Italy, home of the ancient Roman ruins part of the UNESCO World Heritage Sites.
Paragraph 4: Banca di Credito Popolare S.C.p.A (BCP) is an Italian cooperative bank based in Torre del Greco, in Metropolitan City of Naples, Campania. Most of the revenue of the bank came from the Metropolitan City of Naples, which the bank had 44 branches in the metropolitan city.
...
Question: What Metropolitan City was Massimo Giordano born in?
Answer:

Figure 10: An example prompt for the semantic multiple retrieval task based on HotpotQA.