MedOdyssey: A Medical Domain Benchmark for Long Context Evaluation Up to 200K Tokens

Anonymous ACL submission

Abstract

Numerous advanced Large Language Models (LLMs) now support context lengths up to 128K, and some extend to 200K. Benchmarks in the general domain have also followed up on evaluating long-context capabilities. In medical domain, due to the unique contexts and need for domain expertise, more professional and further evaluations are necessitating. Long-context scenarios are common in medical domain tasks but lacks a long-context LLMs benchmark specifically for medical domain. In this paper, we propose MedOdyssey, the first medical long-context benchmark with seven length levels ranging from 4K to 200K tokens. MedOdyssey consists of two primary components: the medical "needles in a haystack" evaluation and a series of medical related long-context tasks, totally 10 datasets. The former includes challenges such as counter-intuitive reasoning and novel (unknown) facts injection to mitigate knowledge leakage and data contamination of LLMs. The latter confronts the challenge of requiring professional medical expertise. Especially, we design the "Maximum Identical Context" principle to improve fairness by guaranteeing that different LLMs observe as many identical contexts as possible. Our experiment evaluates advanced proprietary and open-source LLMs tailored for processing long-context and presents detailed performance analyses. This highlights that LLMs still face challenges to handle long-context in medical domain.

1 Introduction

Long-Context Large Language Models (LLMs) (OpenAI, 2023; Anthropic, 2023; 01.AI et al., 2024) have become a mainstream research topic. To deal with the long-context scenarios when encounter books, lengthy chat history or long documents, two major types of methods are applied. One type of methods using long context as external information for retrieval and summarization to reduce the actual input length for LLMs (Lewis et al., 2020). Meanwhile, another type of ways focuses on increasing the context length that LLMs can handle, thereby avoiding the errors that may arise from retrieval and summarization.

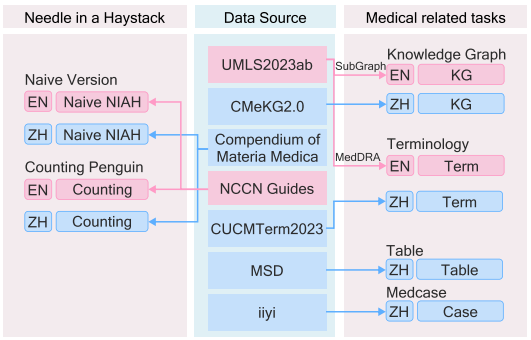


Figure 1: The overall architecture of the MedOdyssey.

Benefiting from various efficient Transformers architectures and positional embedding methodss (Huang et al., 2023; Peng et al., 2023; Jin et al., 2024; Ding et al., 2024), LLMs' context length (context window) is extended, and long-context prompts frequently encountered in practical scenarios can be supported to handle, such as books, lengthy chat history or documents retrieved from website. The LLMs currently available on the market generally support context lengths of 8k tokens. Advanced models have extended this capability to 128k tokens, with some even reaching 200k tokens or more. Researchers have swiftly responded by conducting evaluations of LLMs in long contexts, proposing numerous works in the generic domain to assess their performance. These include the classic needle-in-a-haystack experimental projects (Kamradt, 2024; Song et al., 2024) and several benchmarks (An et al., 2023; Yuan et al., 2024; Zhang et al., 2024) for evaluating and analyzing the long-context performance of LLMs.

In the medical domain, evaluating the medical capabilities of LLMs is often conducted independently due to the unique context and the need for professional knowledge (Tang et al., 2023; Jin et al., 2021; Zhu et al., 2023; Singhal et al., 2023). However, the long-context evaluations in this field (Saab et al., 2024) are relatively infrequent and lack of medical-context "needles in a haystack" experiment. Actually, there are some more difficult long-context scenarios that exist for medical practices, e.g., biomedical terminology normalization and electronic health record (EHR) analysis (Sarker et al., 2018; Shickel et al., 2017). There is a noticeable lack of benchmarks involving a package of basic and various long-context evaluation tasks.

In this paper, we propose MedOdyssey, the first medical-domain long-context evaluation benchmark for LLMs. MedOdyssey is comprised of two primary components: the medical-context needles in a haystack (NIAH) tasks and a series of medical-related tasks, containing 10 complex datasets and involving several medical domain professional corpora, e.g., medical books and guides, medical cases with electronic health records, medical knowledge graphs, medical terminology database and medical tables. Based on these corpora, we construct several evaluation tasks, as shown in Figure 2. Additionally, apart from the naive implementation, we introduced the latest Counting Stars (Song et al., 2024) to enhance the reliability of the "needle in a haystack" component. To ensure fairness, we propose a new "maximum identical context" principle to address the issue of varying contexts resulting from direct middle truncation (Zhang et al., 2024; Yuan et al., 2024). We also prevent data contamination and data leakage during evaluation by incorporating counter-intuitive reasoning problems and novel (unknown) facts questions.

We evaluate the performance of advanced LLMs remarkably supporting long-context prompts, including both proprietary and open-source models. The overall performance is shown in Figure 1 using a radar chart. Our experimental results demonstrate that the performance of LLMs in the medical long contexts is actually still lacking. Specifically, even the newest GPT-4o only performs well in the naive NIAH experiment, and is not a hexagonal warrior. Moreover, we perform a comprehensive analysis to provide insights and direction. We encourage further research by the NLP community to jointly address the more realistic settings presented in this benchmark.

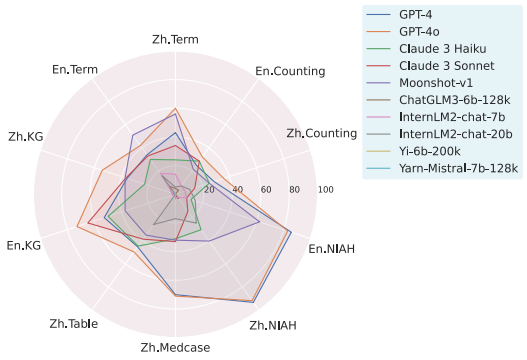


Figure 2: Radar chart of the overall performance of long-context LLMs on MedOdyssey.

The raw data, task data, evaluation results, and evaluation code for MedOdyssey benchmark are publicly available[1].

2 Related Work

Long-Context LLMs. The challenge of supporting long-context prompts for LLMs has been a focal research topic, leading to various innovative approaches. Numerous position embedding methods and efficient transformer architectures(Su et al., 2024; Press et al., 2022; Beltagy et al., 2020; Kitaev et al., 2020; Han et al., 2023) have been instrumental in extending the maximum context length of LLMs. Recently studies on LLMs have garnered interest in handling long-context. For instance, GPT-4 (OpenAI, 2023), Moonshot (MoonshotAI, 2023), Yarn-Mistral (Peng et al., 2023), and ChatGLM3 (THUDM, 2023) can handle up to 128K tokens. Furthermore, models such as Claude 3 (Anthropic, 2023) and Yi (01.AI et al., 2024) support context lengths up to 200K tokens.

Generic-domain Long-Context Evaluation for LLMs. Some research focuses on the capability of LLMs to process long contexts, proposing various datasets and benchmarks. For example, ZeroSCROLLS (Shaham et al., 2023) evaluates state-of-the-art LLMs through document summarization, question answering, and aggregation tasks. L-Eval (An et al., 2023) relabeled some public datasets and proposed additional evaluation metrics. However, most of these studies do not include evaluations in the medical domain.

Medical-domain Evaluation Benchmark for LLMs. LLMs are increasingly used in medical fields, where specialized context requires different

[1] https://anonymous.4open.science/r/MedOdyssey-F925

evaluation methods from general domains. Tang et al. (2023) assess LLMs with zero-shot medical evidence summarization, and Rydzewski et al. (2024) evaluate LLMs in specific medical areas. Primary data sources often include existing exams or benchmarks. Jin et al. (2021) created the MedQA dataset from medical board exams. Liu et al. (2023) uses questions from the Chinese National Medical Licensing Examination, while MultiMedQA (Singhal et al., 2023) combines six medical QA datasets from online searches. However, there is a lack of evaluation benchmarks with medical long-context.

3 The MedOdyssey Benchmark and Dataset

3.1 Benchmark Tasks in MedOdyssey

We define a total of ten tasks in two types, **needle in a haystack** for the general long-context scenario evaluation, and **medical-related tasks** for medical domain long-context scenario evaluation, as shown in Figure 2.

3.1.1 Needle in a Heystack.

To evaluate the performance in handling long-context in a whole length level and align with existing benchmarks, we build a needle in a heystack task dataset.

Naive NIAH. The naive needle in a haystack, inserting a fragment of unrelated knowledge (the needle) within a lengthy context (the haystack) and then prompting the LLM to answer questions about the unrelated knowledge.

Counting. A more challenging variation of the NIAH task. Within the context of a virtual story, dispersed counting fragments are embedded throughout a lengthy context. The LLM is then prompted to identify and output the sequence of these counting fragments.

3.1.2 Medical Related Tasks.

In medical domain, many tasks such as clinical decision support (Papadopoulos et al., 2022) and diagnosis (Wang et al., 2020), involves querying long-context with high accuracy, such as terminology, medical records, and tables.

Term Norm. The medical terminology normalization task, requires LLMs to identify the corresponding standard term for a medical phrase from a large standard terminology database.

KG QA. The LLM is prompted to answer questions derived from a medical knowledge graph presented in triplet form, concentrating on the relationships of entities and relationships.

Table QA. This task involves the LLM responding to questions based on medical tables that are formatted in Markdown.

Case QA. Here, the LLM addresses questions related to provided medical cases, which include details of patient EHR information and the treatment processes.

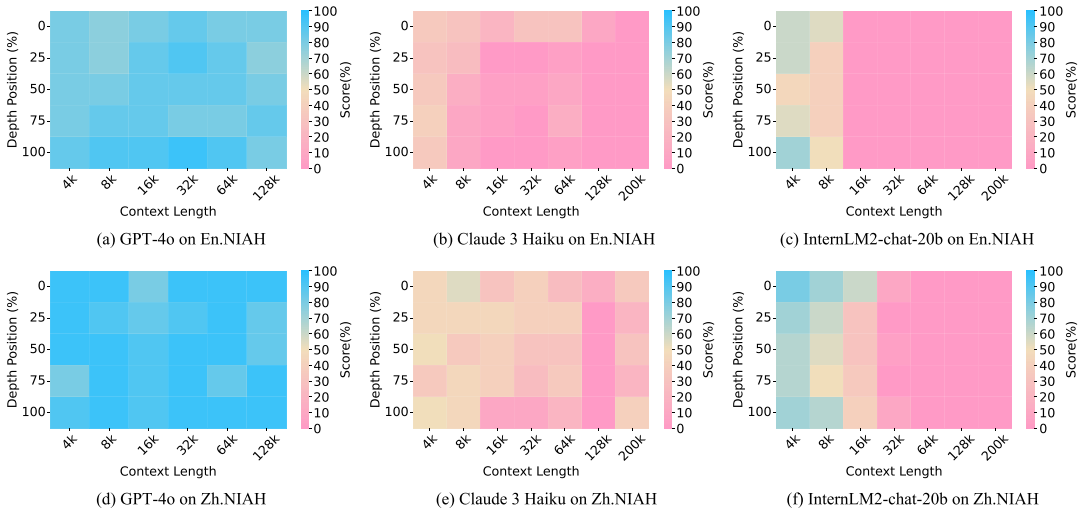
We use some Chinese books and English guides as the haystack in NIAH and Counting tasks. Additionally, all QA tasks are based on closed-ended, text-based questions. Figure B1 to Figure B4 in the appendix show the examples of input and output.

3.2 Dataset Collection

To meet the professional needs in medical domain, we prefer to collect real scenario data rather than through simulation, self-building, or distillation techniques. However, due to the copyright and privacy protection concerns, collecting diverse and valuable corpora is challenging. Consequently, we dedicated significant effort to finding academic open-source, formal application pathways, and copyright-free medical data and knowledge.

As shown in Figure 2, for the "needles in a haystack" part, we have collected 30 volumes of Chinese medical books "Compendium of Materia Medica" from an open-source repository[2], and three English clinical guides[3] in PDF format were converted to meet long text requirements. And there are four knowledge bases involved in medical-related tasks. We converted and organized the "Chinese Common Clinical Medical Terminology 2023 Edition" (CUCMTerm2023) from PDF format to obtain four types of standard terms: disease diagnosis, clinical examination, procedure operation, and symptom. We used MedDRA terms from the UMLS2023ab version (Bodenreider, 2004)[4] as the foundational terminology bases. Additionally, we used CMeKG2.0[5] and extracted MedDRA subgraphs from the UMLS2023ab version as the basic knowledge graphs. We also obtained 500 medical cases with EHR information from an open-source medical forum iiyi[6], and crawled 100 medical ta-

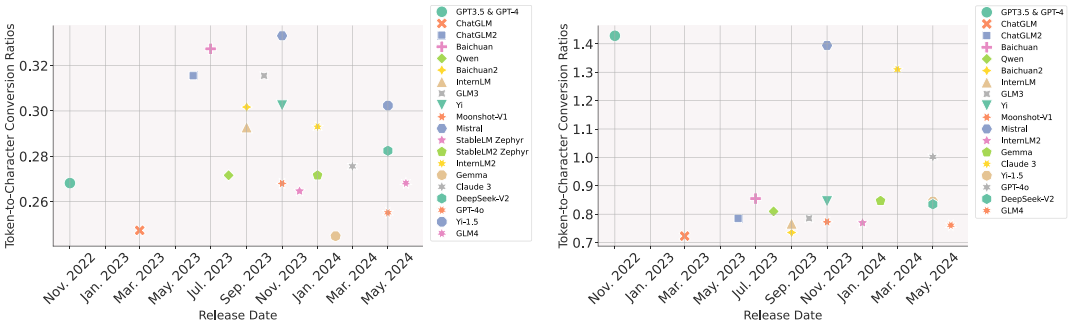
[2] https://github.com/lab99x/tcmoc/tree/master
[3] https://www.nccn.org/guidelines/
[4] https://www.nlm.nih.gov/research/umls/
[5] http://cmekg.pcl.ac.cn/
[6] https://bingli.iiyi.com/



(a) GPT-4o on En.NIAH (b) Claude 3 Haiku on En.NIAH (c) InternLM2-chat-20b on En.NIAH

(d) GPT-4o on Zh.NIAH (e) Claude 3 Haiku on Zh.NIAH (f) InternLM2-chat-20b on Zh.NIAH

Figure 3: Heatmaps of GPT-4o, Claude 3 Haiku and InternLM2-chat-20b on NIAH task.



(a) Token-to-Character Conversion Ratios in English (b) Token-to-Character Conversion Ratios in Chinese.

Figure 4: Trends in token-to-character conversion rates for advanced LLMs over time.

bles from an open-source medical website MSD[7].

3.3 Dataset Construction

Our benchmark is primarily designed to evaluate the long-context capabilities of LLMs within medical texts. By examining the context windows supported by advanced LLMs, we have selected seven token lengths: 4k, 8k, 16k, 32k, 64k, 128k, and 200k.

To improve the fairness between LLMs with different tokenizers, we propose **maximum identical context**. To avoid evaluation data contamination, we apply **novel facts injection**. And to guarantee the answer from the LLMs is indeed from the long-context provided but not due to the the implicit knowledge that the LLMs have trained, we follow a **counter-intuitive reasoning** manner. The following part will introduce these principles in detail.

[7] https://www.msdmanuals.cn/professional/pages-with-widgets/tables?mode=list

Maximum Identical Context (MIC). It is worth noting that the current work aims to reach the maximum number of tokens for a given model, employing intermediate truncation when performing long-context evaluation. In practice, this strategy results in different models receiving different contextual texts, which ultimately lacks fairness.

In our work, we present the "Maximum Identical Context" principle and convert a fixed number of tokens to a fixed range. By analyzing the token-to-character conversion ratios of advanced LLMs in Table A1, we select a fixed conversion ratio for both Chinese and English to ensure that all LLMs can see the same context while accepting the maximum number of tokens. Formally, our goal is to optimize the formula 1 for each sample to obtain the maximum text length L' corresponding to a certain number of tokens N, where C is the predefined token length list and γ is the specific maximum token-to-character conversion ratio analyzed from Figure 4 and Table A1. In practice, all our dataset builds adopt this principle to get the

Table 1: Dataset statistics. The columns indicate the annotation method, the number of examples, average text length (input/output), use of the construction strategy from Section 3.3, and the evaluation metrics. **MIC** stands for Maximum Identical Context, **NFI** stands for Novel Facts Injection, and **CIR** stands for Counter-intuitive Reasoning.

Task	Annotation	# Examples	Avg. Len	MIC	NFI	CIR	Eval Metrics
En.NIAH	Auto & Human	20×7×5	179.2k/32	✔	✔	✗	Acc.
Zh.NIAH	Auto & Human	20×7×5	45.6k/10.2	✔	✔	✗	Acc.
En.Counting	Auto	4×7	179.0k/13.6	✔	✗	✔	Acc.
Zh.Counting	Auto	4×7	45.6k/12.3	✔	✗	✔	Acc.
En.KG	Auto & Human	100	186.4k/68.8	✔	✗	✔	P., R., F1.
Zh.KG	Auto & Human	100	42.5k/2.0	✔	✗	✔	P., R., F1.
En.Term	Auto	100	183.1k/11.7	✔	✗	✗	Acc.
Zh.Term	Auto	100	32.6k/7.0	✔	✗	✗	Acc.
Zh.Case	Auto & Human	100	47.7k/1.3	✔	✗	✗	Acc.
Zh.Table	Auto & Human	100	53.6k/1.4	✔	✗	✗	P., R., F1.

maximum identical context shared across LLMs.

We acknowledge that when evaluating a new LLM with our benchmark could impact the token-to-character conversion ratio and the dataset. Nonetheless, we remain committed to this approach and have identified effective measures through risk analysis to address these challenges. As shown in Figure 4, a clear trend is that the token-to-character conversion ratio of advanced LLMs is decreasing, which will keep our benchmark robust. Meanwhile, we tend to integrate **MedOdyssey** into periodic evaluation platforms, adjusting it by periodically adapting to new token-to-character conversion ratios, replacing old questions with new ones, and using code automation to complete the build. This approach will help further ensure fairness and prevent data leakage.

$$\min_{N \in C} \left(\frac{N}{\gamma} - L' \right), L' \leq \frac{N}{\gamma}, \quad \text{where } C = \{4k, 8k, \ldots, 200k\} \tag{1}$$

Novel Facts Injection (NFI). To prevent data leakage and contamination, i.e., to ensure that LLMs have not been trained on question-related data, we employ a novel fact injection method in the naive needle-in-a-haystack task. Specifically, we manually and meticulously crafted needles and their corresponding questions for the needle-in-a-haystack task, including ten non-medical questions and ten medical questions. These twenty questions are based on the latest information, with the general portion drawn from the newest plot and setting of the "Honkai: Star Rail" game, and the medical portion sourced from the latest literature in The Lancet and some real doctor-patient dialogues. Meanwhile, in this task, we measure the effect of five different depths at which the needle is located and seven different lengths of the haystack, achieved through automated code execution. Eventually, we get the datasets **En.NIAH** and **Zh.NIAH**.

Counter-intuitive Reasoning (CIR). Acquiring systematic medical knowledge, such as knowledge graphs, is challenging due to the slow accumulation of medical information. To address the difficulty in ensuring that the model hasn't been trained on this type of knowledge, we introduced counter-intuitive designs to test the LLM's reasoning with long contexts. For example, in the KG task, we ask the model to find all the triples that can answer a question instead of directly providing an answer. We randomized some questions involving three cases from the graph: head-entity to tail-entity, head-entity to relationship, and relationship to tail-entity, and generated questions using pre-constructed templates. For a given sample, we identify all relevant triples as the correct answer based on all input triples, resulting in the dataset **En.KG** and **Zh.KG**.

Similarly in the counting task, we designed a counter-intuitive story setting, i.e., we have a little star count penguins, where the LLM must retain the memory of the task goal regardless of the context length. Additionally, For the "Counting Penguin" task, four different difficulty types were designed, including counting a penguin repeatedly, counting penguins incrementally, counting penguins disorderly, and counting penguins with corrections. As in the original project, we use the correct counting order as the answer, and we get the dataset **En.Counting** and **Zh.Counting**.

We adopt SMM4H-17[8] (Sarker et al., 2018) to construct our English terminology normalization task dataset **En.Term**. We constructed for Chinese terminology normalization task dataset, **Zh.Term**, based on the synonyms and previously utilized phrases in CUCMTerm2023 corpus, which includes the same four term categories present in our established standard terminology database.

For both the medical table QA dataset **Zh.Table** and medical case QA dataset **Zh.Case**, we use a manual querying strategy by randomly selecting a medical table or case and formulating questions based on the relevant information it contains. For example, when working with a medical table, we ask questions related to the specific medical knowledge presented in the table. In the context of medical cases, our questions cover aspects such as the patient's chief complaint, symptoms, result of imaging studies, findings of complete checkup.

When design the QA pairs manually in **NIAH**, **KG**, **Case** and **Table** tasks. The design procedure of the QA pairs including initial designing, checking, and revising. All the participants in the manually design procedure are the authors of this work. In each length level, we firstly design several QA pairs according the principles above. Then other participants that not designed the QA pairs implemented a validation process to confirmed the matching between the questions and answers, and they also confirm whether the principles are followed or not. After the checking, we will have a discussion on the conflict between the designers and checkers to determine a final version of the QA pairs.

3.4 Dataset Statistics.

We present the dataset statistics and the general overview in Table 1. We totally build a dataset with 2,056 long-context samples. The average length of the context in the sub-set various from 32.6k to 186.4k, cover a integrated length range.

4 Experiments

4.1 Baseline Models

We researched current state-of-the-art long-context LLMs and presented the performance of two kinds of baseline LLMs in MedOdyssey. For closed-source commercial LLMs, we call the official APIs to get the responses for each task. We also deployed open-source models for inference on our own. The LLMs and versions we selected are as follows:

GPT-4 (OpenAI, 2023): Released in March 2023 by OpenAI. The context length of GPT-4 has been extended to 128k in the November 2023 update. (gpt-4-turbo-2024-04-09)

GPT-4o (OpenAI, 2024): The latest LLM of OpenAI, GPT-4o was introduced in May 2024, with a 128k context window, and has a knowledge cut-off date of October 2023. (gpt-4o-2024-05-13)

Claude 3 (Anthropic, 2023): Launched by Anthropic in March 2024, we use two versions of Claude, Haiku and Sonnet. Claude offers a 200k context window upon launch. (claude-3-haiku-20240307 and claude-3-sonnet-20240229)

Moonshot-v1 (MoonshotAI, 2023): Released in 2023 by Moonshot AI, it emphasizes scalability and supports a context window of 128k tokens for generating very long texts. (moonshot-v1-128k)

ChatGLM3-6b-128k (THUDM, 2023): Developed by ZHIPU·AI in 2024, it builds based on ChatGLM3-6B and better handles long contexts up to 128K tokens.

InternLM2 (Cai et al., 2024): An open-source LLM is introduced in 2024 by Shanghai AI Lab, including 7b and 20b sizes. It initially trained on 4k tokens before advancing to 32k tokens in pre-training and fine-tuning stages, and has supported up to 200k when inference.

Yi-6b-200k (01.AI et al., 2024): Yi series models are trained from scratch by 01.AI and the 6B version is open-sourced and available to the public in November 2023 and supports a context window length of 200k.

Yarn-Mistral-7b-128k (Peng et al., 2023): Developed by NousResearch and released in November 2023. It is further pretrained on long context data for 1500 steps using the YaRN extension method based on Mistral-7B-v0.1 and supports a 128k token context window.

4.2 Implementation Details

We inferred open-source LLMs using the official deployment method on a single NVIDIA A100 80GB GPU. Yarn-Mistral-7b-128k and Yi-6B-200K, as base models (non-chat), completed tasks via text completion but showed some limitations in following instructions and formats. We set the inference temperature to 0 to eliminate randomness.

In MedOdyssey, seven context lengths were considered in MedOdyssey: 4k, 8k, 16k, 32k, 64k, 128k, and 200k. The naive needle-in-a-haystack

[8] https://data.mendeley.com/datasets/rxwfb3tysd/1

Table 2: The main experiment results of medical-related tasks based on exact string matching.

Models	En.KG			Zh.KG			En.Term	Zh.Term	Zh.Case	Zh.Table		
	P.	R.	F1.	P.	R.	F1.	Acc.	Acc.	Acc.	P.	R.	F1.
GPT-4	59.34	47.37	52.68	42.28	31.03	35.80	34.00	43.00	70.00	46.27	44.29	45.26
GPT-4o	**76.70**	**69.30**	**72.81**	**76.58**	**41.87**	**54.14**	42.00	**60.00**	**71.00**	**48.00**	**51.43**	**49.66**
Claude 3 Haiku	53.54	46.49	49.77	21.19	24.63	22.78	30.00	24.00	31.00	45.86	43.57	44.69
Claude 3 Sonnet	72.04	58.77	64.73	48.39	29.56	36.70	33.00	34.00	33.00	39.55	37.86	38.69
Moonshot-v1	33.33	42.11	37.21	62.07	26.60	37.24	**51.00**	56.00	32.00	36.15	34.31	35.21
ChatGLM3-6b-128k	0.00	0.00	0.00	7.89	1.48	2.49	7.00	4.00	1.00	0.00	0.00	0.00
InternLM2-chat-7b	2.90	1.75	2.19	5.45	1.48	2.33	18.00	14.00	3.00	0.00	0.00	0.00
InternLM2-chat-20b	0.00	0.00	0.00	0.00	0.00	0.00	16.00	5.00	17.00	31.63	22.14	26.05
Yi-6b-200k	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Yarn-Mistral-7b-128k	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

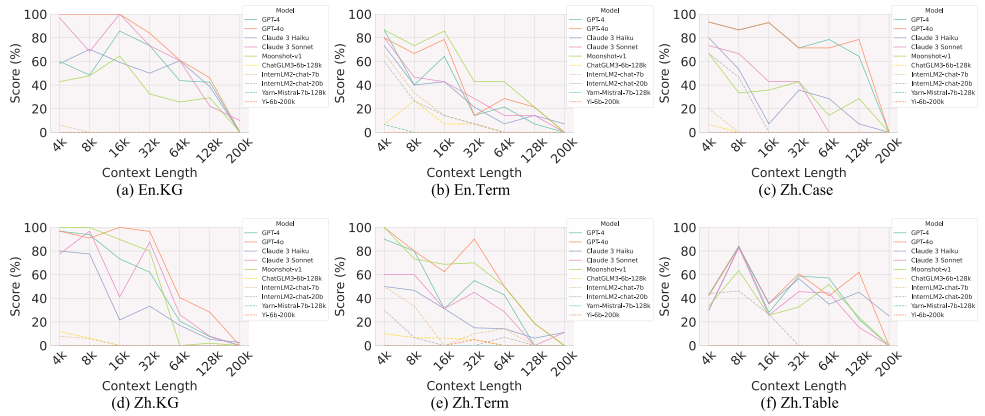


Figure 5: Trends in the performance variations of LLMs on medical-related tasks across different context lengths.

experiment evaluated five needle depths: 0%, 25%, 50%, 75%, and 100%. Also, ground truths are mainly context-based and close-ended. We used standard prompts, clearly defining tasks and requiring outputs in JSON format. Specific prompts are in Appendix Figure B5 to B11. Table 1 outlines evaluation metrics, computed using exact string matching (ESM).

4.3 Results and Analysis

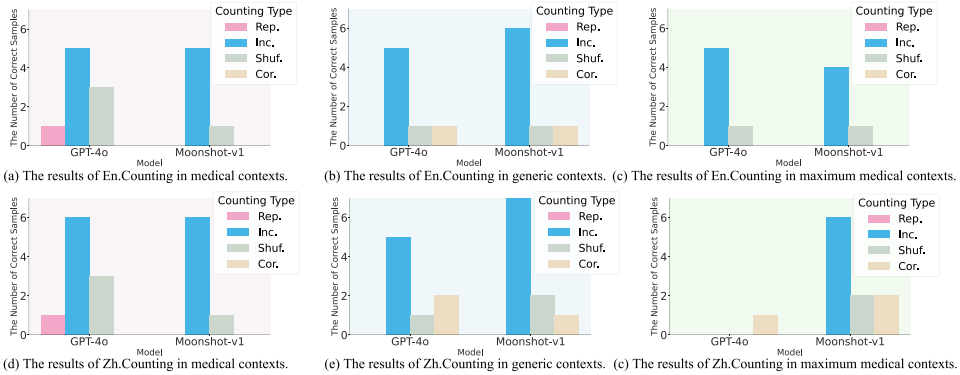
NIAH Results and Analysis. Figure 3 shows the results of the naive medical-context needle-in-a-haystack experiment, using heatmaps to illustrate the performance of LLMs at different lengths and depths. We selected three representative models: GPT-4o, Claude 3 Haiku, and InternLM2-chat-20b, and the complete experimental results are shown in Appendix Table A2 and Figure A2.

Advanced LLMs, such as the GPT-4 series, perform well on the naive needle-in-a-haystack task, even with new facts in the inserted needle. In contrast, other competitive LLMs see degraded performance as context length increases. Most open-source models got zero scores due to their inability to format outputs correctly for lengthy texts, especially the two foundational models. To address this, we relaxed the evaluation standard by removing formatting and using the subset string matching (SSM) algorithm, with results shown in Appendix Table A3 and Figure A3. Additionally, our error analysis showed that within the medical context, LLMs are more likely to make mistakes when addressing general "needles" compared to medical-specific "needles", with the error ratio being approximately 6:5.

Counting Results and Analysis. We present the performance of LLMs on four types of different Counting tasks in detail in Table 3 and an intuitive bar chart in Figure A1. This task is quite difficult with its fictional, counter-intuitive setting, even when using state-of-the-art LLMs. There is an interesting phenomenon where advanced LLMs can perform increasing counting tasks, likely due to their ability to capture this incremental pattern from the training corpus. However, this ability fades with disorganized counting. Most LLMs struggle with repeated counting and counting with corrections, highlighting their diminished reasoning ability, similar to a student confused by similar answer choices. Additionally, it reveals their vulnerability to self-

Table 3: The main experiment result of the En.Counting and Zh.Counting tasks.

Models	En.Counting				All	Zh.Counting				All
	Rep.	Inc.	Shuf.	Cor.		Rep.	Inc.	Shuf.	Cor.	
GPT-4	0	5	1	1	7/28	0	6	2	0	8/28
GPT-4o	1	5	3	0	**9/28**	1	6	3	0	**10/28**
Claude 3 Haiku	0	7	1	0	8/28	0	6	1	0	7/28
Claude 3 Sonnet	1	6	1	0	8/28	0	3	1	0	4/28
Moonshot-v1	0	5	1	0	6/28	0	6	1	0	7/28
ChatGLM3-6b-128k	0	1	0	0	1/28	0	0	0	0	0/28
InternLM2-chat-7b	0	1	1	0	2/28	0	2	0	0	2/28
InternLM2-chat-20b	0	2	0	0	2/28	0	3	0	0	3/28
Yi-6b-200k	0	0	0	0	0/28	0	0	0	0	0/28
Yarn-Mistral-7b-128k	0	0	0	0	0/28	0	0	0	0	0/28



(a) The results of En.Counting in medical contexts. (b) The results of En.Counting in generic contexts. (c) The results of En.Counting in maximum medical contexts.

(d) The results of Zh.Counting in medical contexts. (e) The results of Zh.Counting in generic contexts. (c) The results of Zh.Counting in maximum medical contexts.

Figure 6: Comparison of GPT-4o and Moonshot-v1 on Counting tasks in different context settings.

doubt, akin to a student who becomes skeptical when all answer options are identical.

Medical-related Tasks Results and Analysis. The overall performance of medical-related tasks is displayed in Table 2, and we also provide a loose version of the results using SSM in Table A4. The current state-of-the-art GPT-4o model performs well in terms of answer quality and format adherence, but is still not entirely reliable. Notably, the model's performance exhibits an overall decline as the context length increases, as shown in Figure 5. The open-source LLMs are almost impossible to accomplish the task, especially two base models, which lose the ability to output in format (marked with a green background). In particular, Moonshot-v1 has a good performance if only the content of the answer is considered for evaluation.

Analysis of Different Context Setting. We used the Counting task to experiment with different context settings: medical long context (MIC), generic long context (MIC), and maximum medical context length. The ablation results are shown in Figure 6. The experimental results support our proposed "MIC" principle. It is easy to observe that the performance is affected by different contexts whether the length is different or the domain is different, so we prefer to sacrifice an evaluation of extreme context length in exchange for sharing the same contextual texts between different LLMs. Due to different training corpus and training strategies, the degree of impact varies.

Error analysis. The errors observed primarily involved repetition, question forgetting, and reasoning flaws. While more advanced models like GPT-4o reduce the likelihood of question forgetting, the risk of repetition remains. Reasoning accuracy, however, is largely contingent on the LLMs' capabilities as reflected in Figure 1.

5 Conclusion

We take a step forward by building the first medical long-context evaluation benchmark, **MedOdyssey**, to facilitate the study of LLMs in long-context scenarios. Our benchmarks include medical-context needle-in-a-haystack tasks and several medical-related long-context tasks, totally build ten evaluation datasets. Additionally, we propose three effective principles to enhance the fairness and reliability of evaluations. We evaluated on ten state-of-the-art LLMs, providing performance results and analyses in various formats. Additionally, we provide examples of the impact of different contexts.

6 Limitations

Medical long-context evaluation is challenging, and our work faces some dilemmas. We sacrificed evaluating limit lengths to ensure different models share the same contextual cues, resulting in a restricted length being assessed. Effective open-ended QA is lacking due to difficulty in finding appropriate evaluation methods. Additionally, we took efforts to eliminate the effects of randomness (by fixing temperature and format constraints) and prevent data leakage, but these issues are unavoidable. We will continuously explore ways to improve our benchmark, as mentioned in Section 3.3.

7 Ethical Considerations

This paper proposes a new medical-domain long-context evaluation benchmark **MedOdyssey** for LLMs. All of the datasets in MedOdyssey are adhere to ethical guidelines and respect copyright laws. The entire data collection process is free of issues of copyright and issues of privacy, and there are three types of data sources, including license applications, the open source community, and public file cleaning and organizing. Meanwhile, the manual participation part in the dataset construction process was all done by the authors of this paper without any ethical issues.

References

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

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*.

Anthropic. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

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

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*.

Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. 2023. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Gabriel Kamradt. 2024. Llmtest: Needle in a haystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *Preprint*, arXiv:2001.04451.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 52430–52452. Curran Associates, Inc.

MoonshotAI. 2023. Moonshot.

OpenAI. 2023. New models and developer products announced at devday. Technical report, OpenAI.

OpenAI. 2024. Hello gpt-4o. Technical report, OpenAI.

Petros Papadopoulos, Mario Soflano, Yaelle Chaudy, Wilson Adejo, and Thomas M Connolly. 2022. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology*, 12(4):713–727.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Nicholas R. Rydzewski, Deepak Dinakaran, Shuang G. Zhao, Eytan Ruppin, Baris Turkbey, Deborah E. Citrin, and Krishnan R. Patel. 2024. Comparative evaluation of llms in clinical oncology. *NEJM AI*, 1(5):AIoa2300151.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

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.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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. *Preprint*, arXiv:2403.11802.

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

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

THUDM. 2023. Chatglm3: Open bilingual language model. `https://github.com/THUDM/ChatGLM3`.

Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, et al. 2020. A study of entity-linking methods for normalizing chinese diagnosis and procedure terms to icd codes. *Journal of biomedical informatics*, 105:103418.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, et al. 2024. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *arXiv preprint arXiv:2402.05136*.

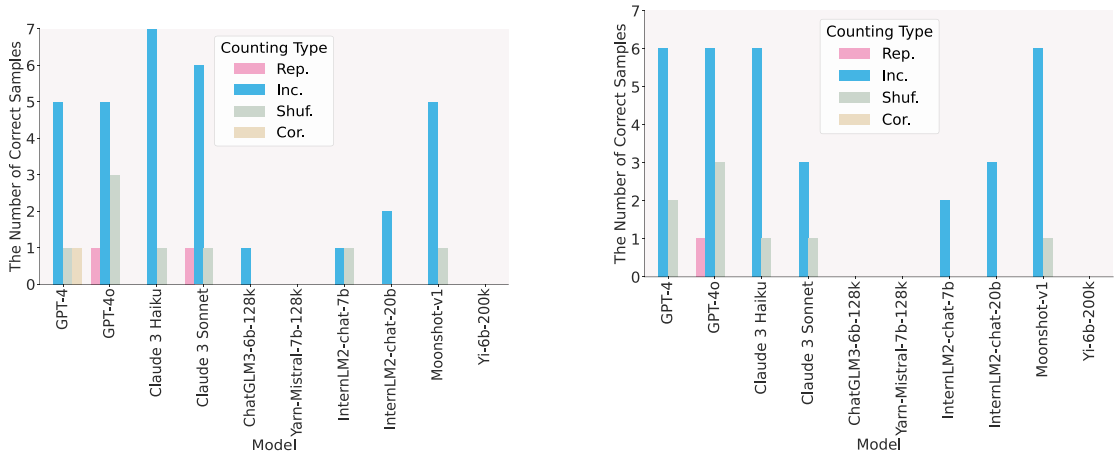
Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024. Infinitebench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. Promptcblue: A chinese prompt tuning benchmark for the medical domain. *arXiv preprint arXiv:2310.14151*.

A Full experiment results.

Table A1: The token-to-character conversion ratios of advanced long-context LLMs.

Models	En.NIAH	Zh.NIAH	En.KG	Zh.KG	En.Term	Zh.Term	Zh.Case	Zh.Table
GPT-4	0.281	**1.402**	0.267	**1.473**	0.267	**1.446−1.676**	**1.316**	**1.178**
GPT-4o	0.275	1.005	0.252	1.029	0.253	0.991−1.124	0.904	0.802
Claude 3 Haiku/Sonnet	0.289	1.342	0.275	1.330	0.264	1.291−1.483	1.191	1.072
Moonshot-v1	0.286	0.924	0.266	0.737	0.265	0.732−0.780	0.712	0.600
ChatGLM3-6b-128k	0.342	0.924	0.313	0.750	0.302	0.760−0.827	0.746	0.630
InternLM2-chat-7b/20b	0.299	0.899	0.292	0.739	0.289	0.750−0.797	0.725	0.608
Yi-6b-200k	0.342	0.992	0.301	0.812	0.293	0.791−0.883	0.773	0.659
Yarn-Mistral-7b-128k	**0.355**	1.394	**0.331**	1.430	**0.324**	1.362−1.607	1.286	1.139



(a) En.Counting Result (b) Zh.Counting Result

Figure A1: Histogram of Counting task results.

Table A2: The main experiment results of NIAH.

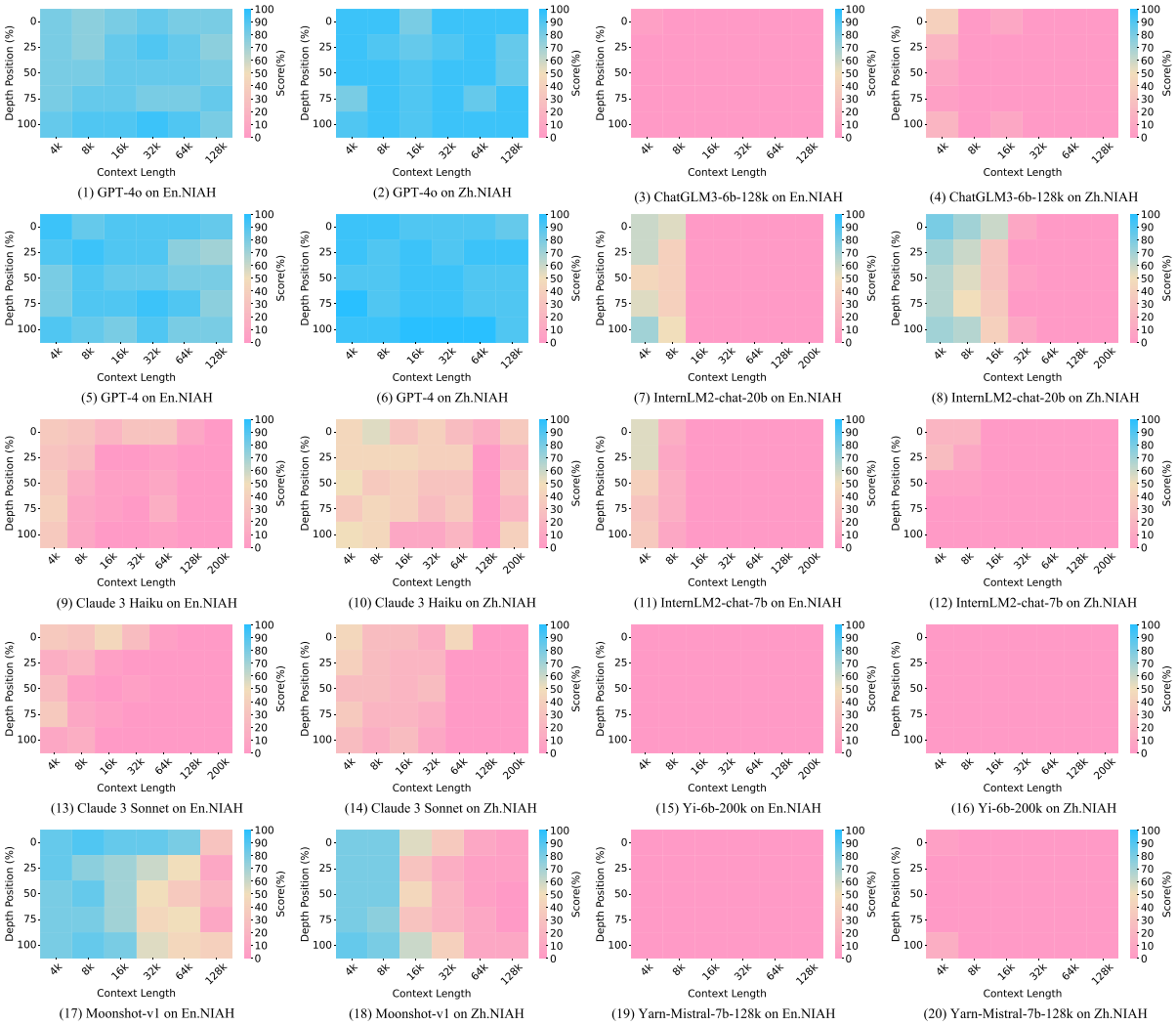
Models	Ablation	En.NIAH							ALL	Zh.NIAH							ALL
		4k	8k	16k	32k	64k	128k	200k		4k	8k	16k	32k	64k	128k	200k	
GPT-4	0%	19	17	18	18	18	17	–	107/120	19	19	18	18	18	17	–	109/120
	25%	18	19	18	18	15	14	–	102/120	19	18	19	18	19	19	–	112/120
	50%	16	18	17	17	16	16	–	100/120	18	18	19	19	18	18	–	110/120
	75%	16	18	18	19	18	15	–	104/120	20	18	19	19	18	18	–	112/120
	100%	18	17	16	18	16	16	–	101/120	19	19	20	20	20	18	–	116/120
	ALL	87/100	89/100	87/100	90/100	83/100	78/100	–	514/600	95/100	92/100	95/100	94/100	93/100	90/100	–	559/600
GPT-4o	0%	16	15	16	17	16	16	–	96/120	19	19	16	19	19	19	–	111/120
	25%	16	15	17	18	17	15	–	98/120	19	18	17	18	19	17	–	108/120
	50%	16	16	17	17	17	16	–	99/120	19	19	18	19	19	17	–	111/120
	75%	16	17	17	16	16	17	–	99/120	16	19	18	19	17	19	–	108/120
	100%	17	18	18	19	18	16	–	106/120	18	19	18	19	19	19	–	112/120
	ALL	81/100	81/100	85/100	87/100	84/100	80/100	–	498/600	91/100	94/100	87/100	94/100	93/100	91/100	–	550/600
Claude 3 Haiku	0%	7	6	4	6	6	2	0	31/140	9	11	6	8	5	3	7	49/140
	25%	6	5	0	0	1	0	0	12/140	9	9	9	8	8	0	4	47/140
	50%	7	3	1	1	2	0	0	14/140	10	7	8	6	6	0	6	43/140
	75%	8	2	1	0	3	0	0	14/140	7	9	8	5	7	0	4	40/140
	100%	7	2	0	0	1	0	0	10/140	10	9	2	2	4	0	8	35/140
	ALL	35/100	18/100	6/100	7/100	13/100	2/100	0/100	81/700	45/100	45/100	33/100	29/100	30/100	3/100	29/100	214/700
Claude 3 Sonnet	0%	7	6	9	5	1	0	0	28/140	9	5	5	3	9	0	0	31/140
	25%	3	4	1	0	0	0	0	8/140	8	5	4	4	0	0	0	21/140
	50%	5	1	0	1	0	0	0	7/140	5	5	4	5	0	0	0	19/140
	75%	7	2	1	0	0	0	0	10/140	7	4	4	3	0	0	0	18/140
	100%	2	3	0	0	0	0	0	5/140	5	3	5	2	0	0	0	15/140
	ALL	24/100	16/100	11/100	6/100	1/100	0/100	0/100	58/700	34/100	22/100	22/100	17/100	9/100	0/100	0/100	104/700
Moonshot-v1	0%	17	18	17	17	16	6	–	91/120	16	16	11	7	2	1	–	53/120
	25%	17	15	14	12	10	2	–	70/120	16	16	6	3	1	1	–	43/120
	50%	16	17	14	10	7	4	–	68/120	16	16	9	4	1	0	–	46/120
	75%	16	16	14	9	10	2	–	67/120	16	15	6	4	2	0	–	43/120
	100%	16	17	16	11	9	8	–	77/120	17	16	12	8	2	2	–	57/120
	ALL	82/100	83/100	75/100	59/100	52/100	22/100	–	373/600	81/100	79/100	44/100	26/100	8/100	4/100	–	242/600
ChatGLM3-6b-128k	0%	1	0	0	0	0	0	–	1/120	8	0	2	0	0	0	–	10/120
	25%	0	0	0	0	0	0	–	0/120	4	0	0	0	0	0	–	4/120
	50%	0	0	0	0	0	0	–	0/120	2	0	0	0	0	0	–	2/120
	75%	0	0	0	0	0	0	–	0/120	1	0	0	0	0	0	–	1/120
	100%	0	0	0	0	0	0	–	0/120	4	0	2	0	0	0	–	6/120
	ALL	1/100	0/100	0/100	0/100	0/100	0/100	–	1/600	19/100	0/100	4/100	0/100	0/100	0/100	–	23/600
InternLM2-chat-7b	0%	11	3	0	0	0	0	0	14/140	4	4	0	0	0	0	0	8/140
	25%	11	2	0	0	0	0	0	13/140	5	2	0	0	0	0	0	7/140
	50%	8	3	0	0	0	0	0	11/140	1	1	0	0	0	0	0	2/140
	75%	6	3	0	0	0	0	0	9/140	0	0	0	0	0	0	0	0/140
	100%	7	2	0	0	0	0	0	9/140	0	0	0	0	0	0	0	0/140
	ALL	43/100	13/100	0/100	0/100	0/100	0/100	0/100	56/700	10/100	7/100	0/100	0/100	0/100	0/100	0/100	17/700
InternLM2-chat-20b	0%	12	11	0	0	0	0	0	23/140	16	14	12	2	0	0	0	44/140
	25%	12	8	0	0	0	0	0	20/140	14	12	6	0	0	0	0	32/140
	50%	9	8	0	0	0	0	0	17/140	13	11	6	1	0	0	0	31/140
	75%	11	8	0	0	0	0	0	19/140	13	10	7	0	0	0	0	30/140
	100%	14	10	0	0	0	0	0	24/140	14	13	8	2	0	0	0	37/140
	ALL	58/100	45/100	0/100	0/100	0/100	0/100	0/100	103/700	70/100	60/100	39/100	5/100	0/100	0/100	0/100	174/700
Yarn-Mistral-7b-128k	0%	0	0	0	0	0	0	–	0/120	1	0	0	0	0	0	–	1/120
	25%	0	0	0	0	0	0	–	0/120	0	0	0	0	0	0	–	0/120
	50%	0	0	0	0	0	0	–	0/120	0	0	0	0	0	0	–	0/120
	75%	0	0	0	0	0	0	–	0/120	0	0	0	0	0	0	–	0/120
	100%	0	0	0	0	0	0	–	0/120	3	0	0	0	0	0	–	3/120
	ALL	0/100	0/100	0/100	0/100	0/100	0/100	–	0/600	4/100	0/100	0/100	0/100	0/100	0/100	–	4/600
Yi-6b-200k	0%	0	0	0	0	0	0	0	0/140	0	0	0	0	0	0	0	0/140
	25%	0	0	0	0	0	0	0	0/140	0	0	0	0	0	0	0	0/140
	50%	0	0	0	0	0	0	0	0/140	0	0	0	0	0	0	0	0/140
	75%	0	0	0	0	0	0	0	0/140	0	0	0	0	0	0	0	0/140
	100%	0	0	0	0	0	0	0	0/140	0	0	0	0	0	0	0	0/140
	ALL	0/100	0/100	0/100	0/100	0/100	0/100	0/100	0/700	0/100	0/100	0/100	0/100	0/100	0/100	0/100	0/700

Table A3: The main experiment result of NIAH based on subset string matching.

Models	Ablation	En.NIAH							ALL	Zh.NIAH							ALL
		4k	8k	16k	32k	64k	128k	200k		4k	8k	16k	32k	64k	128k	200k	
GPT-4	0%	20	19	20	20	20	20	–	119/120	19	19	18	18	18	18	–	110/120
	25%	20	20	19	20	18	19	–	116/120	19	18	19	18	19	20	–	113/120
	50%	18	20	20	20	19	19	–	116/120	19	18	19	19	19	18	–	112/120
	75%	19	20	20	20	20	20	–	119/120	20	19	19	19	19	19	–	115/120
	100%	19	19	19	20	20	20	–	117/120	19	19	20	20	20	19	–	117/120
	ALL	96/100	98/100	98/100	100/100	97/100	98/100	–	587/600	96/100	93/100	95/100	94/100	95/100	94/100	–	567/600
GPT-4o	0%	19	19	19	19	20	20	–	116/120	20	20	18	20	20	20	–	118/120
	25%	18	18	20	20	20	18	–	114/120	20	20	19	20	20	20	–	119/120
	50%	19	19	19	20	19	20	–	116/120	20	20	19	20	20	20	–	119/120
	75%	18	20	19	19	20	20	–	116/120	19	20	20	20	20	20	–	119/120
	100%	20	19	20	20	20	17	–	116/120	20	20	20	20	20	20	–	120/120
	ALL	94/100	95/100	97/100	98/100	99/100	95/100	–	578/600	99/100	100/100	96/100	100/100	100/100	100/100	–	595/600
Claude 3 Haiku	0%	19	20	20	20	20	16	18	133/140	19	20	19	19	19	18	18	132/140
	25%	17	16	18	16	17	16	17	117/140	18	18	18	18	20	16	19	127/140
	50%	16	19	20	17	17	19	19	127/140	19	20	18	19	19	19	19	133/140
	75%	16	18	19	17	18	19	18	125/140	18	19	19	19	18	18	20	131/140
	100%	18	20	19	19	19	19	19	133/140	18	19	18	19	18	19	19	130/140
	ALL	86/100	93/100	96/100	89/100	91/100	89/100	91/100	635/700	92/100	96/100	92/100	94/100	94/100	90/100	95/100	653/700
Claude 3 Sonnet	0%	18	19	19	19	16	13	13	117/140	19	19	20	19	19	18	17	131/140
	25%	16	18	18	17	16	13	13	111/140	16	18	19	18	18	17	18	124/140
	50%	15	18	17	18	17	15	14	114/140	15	17	19	18	19	19	19	126/140
	75%	17	19	17	17	18	15	16	119/140	20	17	19	19	16	19	18	128/140
	100%	18	20	17	19	18	18	17	127/140	19	17	19	19	18	18	18	128/140
	ALL	84/100	94/100	88/100	90/100	85/100	74/100	73/100	588/700	89/100	88/100	96/100	93/100	90/100	91/100	90/100	637/700
Moonshot-v1	0%	19	20	19	19	19	17	–	113/120	20	20	20	20	20	20	–	120/120
	25%	19	19	18	19	19	18	–	112/120	20	20	20	20	18	19	–	117/120
	50%	18	19	18	18	18	18	–	109/120	20	20	20	20	19	19	–	118/120
	75%	18	18	18	19	19	19	–	111/120	20	19	19	20	19	20	–	117/120
	100%	18	18	18	17	17	18	–	106/120	19	19	19	19	18	18	–	112/120
	ALL	92/100	94/100	91/100	92/100	92/100	90/100	–	551/600	99/100	98/100	98/100	99/100	94/100	96/100	–	584/600
ChatGLM3-6b-128k	0%	17	18	17	17	18	16	–	103/120	20	19	19	18	18	15	–	109/120
	25%	17	17	18	18	16	14	–	100/120	18	18	19	17	15	14	–	101/120
	50%	17	17	17	18	15	15	–	99/120	18	19	17	19	15	16	–	104/120
	75%	17	15	18	17	17	19	–	103/120	17	18	17	17	18	14	–	101/120
	100%	15	16	14	16	15	16	–	92/120	18	19	18	19	17	15	–	106/120
	ALL	83/100	83/100	84/100	86/100	81/100	80/100	–	497/600	91/100	93/100	90/100	90/100	83/100	74/100	–	521/600
InternLM2-chat-7b	0%	20	19	19	17	17	12	1	105/140	19	19	19	19	16	13	5	110/140
	25%	20	19	19	17	16	11	7	109/140	19	19	17	19	17	13	5	109/140
	50%	20	19	19	17	14	8	12	109/140	19	19	18	17	13	10	6	102/140
	75%	20	20	17	17	14	15	13	116/140	19	19	19	17	15	13	11	113/140
	100%	20	20	19	18	19	18	10	124/140	19	19	19	19	19	19	15	129/140
	ALL	100/100	97/100	93/100	86/100	80/100	64/100	43/100	563/700	95/100	95/100	92/100	91/100	80/100	68/100	42/100	563/700
InternLM2-chat-20b	0%	20	19	19	16	14	8	4	100/140	19	19	18	18	14	9	8	105/140
	25%	20	19	19	19	19	12	9	117/140	19	17	17	16	9	7	9	94/140
	50%	20	19	19	19	15	17	16	125/140	18	18	18	18	12	7	8	99/140
	75%	19	20	19	19	17	17	13	124/140	18	18	17	18	17	12	4	104/140
	100%	19	19	19	20	19	18	16	130/140	18	18	18	18	19	17	16	124/140
	ALL	98/100	96/100	95/100	93/100	84/100	72/100	58/100	596/700	92/100	90/100	88/100	88/100	71/100	52/100	45/100	526/700
Yarn-Mistral-7b-128k	0%	13	12	9	9	7	0	–	50/120	15	10	8	6	6	0	–	45/120
	25%	13	14	6	5	3	0	–	41/120	9	9	6	4	2	1	–	31/120
	50%	12	13	6	7	2	0	–	40/120	8	10	5	5	2	2	–	32/120
	75%	14	15	11	6	2	0	–	48/120	14	9	6	8	2	1	–	40/120
	100%	12	13	15	13	13	0	–	66/120	16	14	15	12	12	10	–	79/120
	ALL	64/100	67/100	47/100	40/100	27/100	0/100	–	245/600	62/100	52/100	40/100	35/100	24/100	14/100	–	227/600
Yi-6b-200k	0%	2	2	2	2	2	1	2	13/140	19	18	19	18	16	15	14	119/140
	25%	2	2	2	2	2	2	2	14/140	18	18	15	13	14	13	11	102/140
	50%	2	2	2	2	2	3	2	15/140	17	17	15	17	14	14	13	107/140
	75%	2	2	2	3	2	3	2	16/140	19	17	16	17	15	15	14	113/140
	100%	2	2	2	2	2	2	2	14/140	19	18	19	16	17	16	16	121/140
	ALL	10/100	10/100	10/100	11/100	10/100	11/100	10/100	72/700	92/100	88/100	84/100	81/100	76/100	73/100	68/100	562/700

Table A4: The main experiment results of medical-related tasks based on subset string matching.

Models	**En.KG**	**Zh.KG**	**En.Term**	**Zh.Term**	**Zh.Case**	**Zh.Table**
GPT-4	51.00	60.00	38.00	47.00	72.00	63.00
GPT-4o	72.00	**80.00**	48.00	**61.00**	76.00	67.00
Claude 3 Haiku	57.00	50.00	37.00	33.00	79.00	77.00
Claude 3 Sonnet	**73.00**	67.00	38.00	41.00	83.00	78.00
Moonshot-v1	46.00	72.00	**52.00**	59.00	**92.00**	**85.71**
ChatGLM3-6b-128k	3.00	3.00	14.00	8.00	73.00	58.00
InternLM2-chat-7b	2.00	3.00	23.00	20.00	67.00	65.00
InternLM2-chat-20b	0.00	2.00	22.00	11.00	67.00	60.00
Yi-6b-200k	0.00	0.00	0.00	2.00	54.00	44.00
Yarn-Mistral-7b-128k	0.00	0.00	1.00	0.00	42.00	17.00



(1) GPT-4o on En.NIAH (2) GPT-4o on Zh.NIAH (3) ChatGLM3-6b-128k on En.NIAH (4) ChatGLM3-6b-128k on Zh.NIAH

(5) GPT-4 on En.NIAH (6) GPT-4 on Zh.NIAH (7) InternLM2-chat-20b on En.NIAH (8) InternLM2-chat-20b on Zh.NIAH

(9) Claude 3 Haiku on En.NIAH (10) Claude 3 Haiku on Zh.NIAH (11) InternLM2-chat-7b on En.NIAH (12) InternLM2-chat-7b on Zh.NIAH

(13) Claude 3 Sonnet on En.NIAH (14) Claude 3 Sonnet on Zh.NIAH (15) Yi-6b-200k on En.NIAH (16) Yi-6b-200k on Zh.NIAH

(17) Moonshot-v1 on En.NIAH (18) Moonshot-v1 on Zh.NIAH (19) Yarn-Mistral-7b-128k on En.NIAH (20) Yarn-Mistral-7b-128k on Zh.NIAH

Figure A2: Heatmaps of the performance of all LLMs on NIAH task.

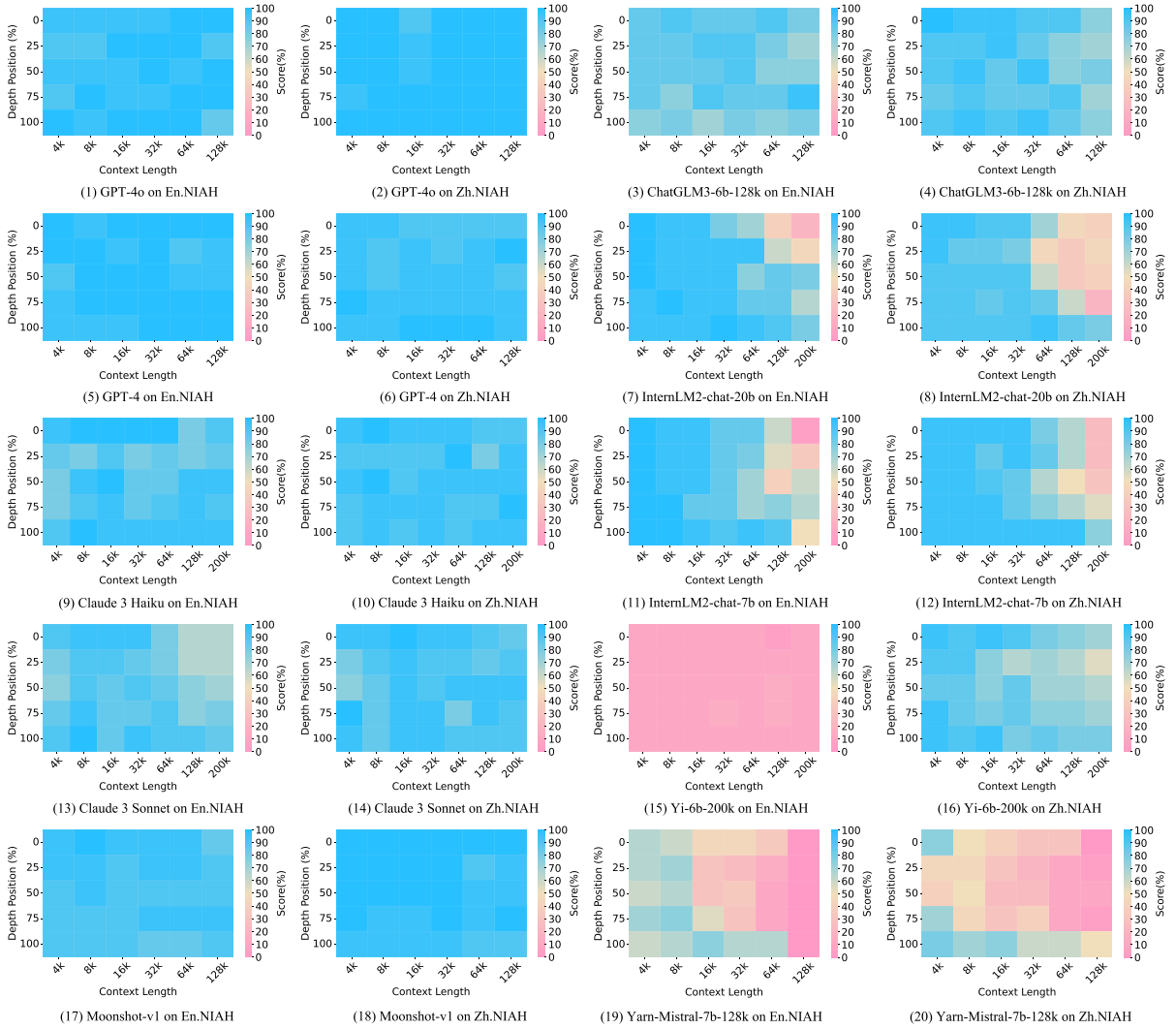


Figure A3: Heatmaps of the performance of all LLMs on NIAH task based on subset string matching.

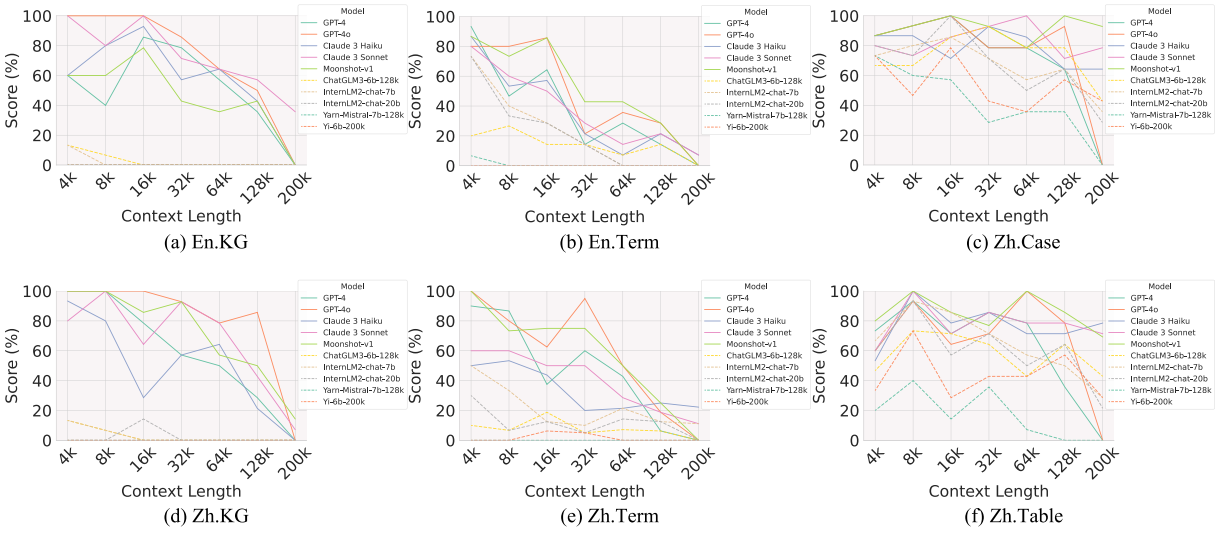


Figure A4: Trends in the performance variations of LLMs on medical-related tasks across different context lengths based on subset string matching.

B Details of our datasets.

En.NIAH

Needles: In Elio's script, there are three nameless guests who got off the Astral Express in Penacony: 1. Razarina Jane Estella, a ex-surveyor aboard the Astral Express and a young scholar, brimming with enthusiasm and curiosity. 2. Breukelen Tiernan, the former guard of the Astral Express and a outstanding gunslinger. 3. Mikhail Char Legwork, Former mechanic of the Astral Express, and the legendary big shot of Penacony, 'the Watchmaker'.

Q: Who is 'the Watchmaker' of Penacony?

A: Mikhail Char Legwork

Needles: The Lancet Diabetes & Endocrinology recently published a study comparing advanced hybrid closed-loop therapy and standard insulin therapy in pregnant women with type 1 diabetes. The study found that for pregnant women starting strict blood glucose control, advanced hybrid closed-loop therapy did not improve overall time in target range but improved overnight time in target range, reduced time below target range, and increased treatment satisfaction. These data suggest that MiniMed 780G (an advanced hybrid closed-loop therapy) can be safely used during pregnancy and offers some additional benefits compared to standard insulin therapy; however, it is important to improve the algorithm to better meet pregnancy requirements.

Q: In the study from The Lancet comparing advanced hybrid closed-loop therapy and standard insulin therapy in pregnant women with type 1 diabetes, which therapy improved the overnight time in target range?

A: Advanced hybrid closed-loop therapy

Zh.NIAH

Needles: 阿哈是执掌欢愉命途上的星神，祂有一群追随者叫做「假面愚者」，但也有一群反对者叫做「悲悼伶人」，祂曾经炸毁了阿基维利的星穹列车。

Q: 谁曾经炸毁了星穹列车?

A: 阿哈

Needles: 以下是一段2019年5月的医患聊天记录\n\n患者：\n刚抛妇产9天。今天上医院检查有点高血压。高压140，，低压96，还要母乳喂孩子，请问产后能吃什么。比较好（女，27）\n\n医生：\n拉贝洛尔是可以吃的。。。\n\n患者：\n吃饭。能吃什么\n\n医生：\n盐量要控制，不吃活血的动西，其他没有禁忌啊\n\n以下是一段2019年4月医患聊天记录\n\n患者：\n医生，我又来了，今天查了个尿常规，隐血1个加，要不要紧（女，25岁）\n\n医生：\n以前查过吗？末次月经什么时候\n\n患者：\n没有，就今天查的\n\n医生：\n嗯嗯\n\n患者：\n这个月15号，昨天还有一点点\n\n医生：\n哦，那考虑跟月经有关，没事的，半月后复查尿常规\n\n患者：\n哦哦，好的，谢谢医生了\n\n医生：\n嗯嗯，不客气。

Q: 请问在2019年4月的医患聊天记录里，患者做了什么检查?

A: 尿常规

Figure B1: Examples of NIAH task.

En.Counting

Repeat: The little star looked to a small area and counted 1 little penguin, The little star looked to a small area and counted 1 little penguin. ⋯, The little star looked to a small area and counted 1 little penguin.

Ground Truth: [1, 1, …, 1]

Increase: The little star looked to a small area and counted 1 little penguin, The little star looked to a small area and counted 2 little penguins, ⋯, The little star looked to a small area and counted 8 little penguins.

Ground Truth: [1, 2, 3, 4, 5, 6, 7, 8]

Zh.Counting

Shuffle: 小星星看向一小块区域，数了8只小企鹅，小星星看向一小块区域，数了5只小企鹅，小星星看向一小块区域，数了1只小企鹅，小星星看向一小块区域，数了3只小企鹅，小星星看向一小块区域，数了4只小企鹅，小星星看向一小块区域，数了6只小企鹅，小星星看向一小块区域，数了5只小企鹅，小星星看向一小块区域，数了11只小企鹅，小星星看向一小块区域，数了9只小企鹅，小星星看向一小块区域，数了7只小企鹅，小星星看向一小块区域，数了9只小企鹅，小星星看向一小块区域，数了6只小企鹅。

Ground Truth: [8, 5, 1, 3, 4, 6, 5, 11, 9, 7, 9, 6]

Correction: 小星星看向一小块区域数了5只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是7只小企鹅，小星星看向一小块区域数了1只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是2只小企鹅，小星星看向一小块区域数了6只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是5只小企鹅，小星星看向一小块区域数了9只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是10只小企鹅，小星星看向一小块区域数了2只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是1只小企鹅，小星星看向一小块区域数了3只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是4只小企鹅，小星星看向一小块区域数了6只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是7只小企鹅，小星星看向一小块区域数了7只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是6只小企鹅。

Ground Truth: [7, 2, 5, 10, 1, 4, 7, 6]

Figure B2: Examples of Counting task.

En.KG

Q: What is the relationship between 'Instillation site tenderness' and 'Instillation site pain'?

A: ["Instillation site tenderness|classifies|Instillation site pain"]

Zh.KG

Q: 鼻炎的相关药物是什么？？

A: ["鼻炎|相关药物|丙酸倍氯米松气雾剂", "鼻炎|相关药物|必畅", "鼻炎|相关药物|斯卫尔", "鼻炎|相关药物|信龙", "鼻炎|相关药物|鼻通滴鼻剂"]

En.Term

Medical phrase: double vision

Ground Truth: Diplopia

En.Term

Medical phrase: 神经精神问卷

Ground Truth: 神经精神量表

Medical phrase: 膀胱镜取石术

Ground Truth: 经尿道膀胱取石术

Medical phrase: 黑粪

Ground Truth: 黑便

Medical phrase: 胎儿双足内翻

Ground Truth: 先天性内翻足

Figure B3: Examples of KG QA and Terminology Normalization.

Zh.Table

Table: 心律失常的治疗

疾病状态	治疗*
窄QRS综合波心动过速	
多源性 房性早搏 房性早搏 室上性病灶（通常在心房）引起各种心律。诊断依靠心电图。无症状不需要治疗。 也可以参考 心律失常概述 异位室上性心律包括 房性早搏 房性心动过速 阅读更多	请放心，非二氢吡啶类钙通道阻滞剂或beta受体阻滞剂
心房颤动 心房颤动 心房颤动是一种快速、不规则的房性心律。症状包括心悸，有时疲乏，体力下降和晕厥先兆。可能形成心房栓子，有引起栓塞性脑卒中的明显危险性。诊断靠心电图。治疗包括用药物控制心率，用抗凝药物预防血栓栓塞，有时用药物或心脏转复的方法使心房颤动转复成窦性心律。... 阅读更多	抗凝 用于控制心率: beta受体阻断剂 维拉帕米 地尔硫䓬，或 地高辛 对于心律控制: 抗心律失常药物（例如伊布利特，胺碘酮，普罗帕酮，决奈达隆，索他洛尔，多非利特） 心脏电复律，或 射频消融 有时需要Maze手预
心房扑动 心房扑动 心房扑动是由心房折返引起的快速规则的心房节律症状包括心悸，有时虚弱感，体力耐受性差，呼吸困难，晕厥先兆。心房血栓可能导管栓塞。通过心电图进行诊断。治疗涉及以下方面：用药物控制心室率，用抗凝药预防血栓栓塞，常用药物或心脏电复律转复成窦性心律。... 阅读更多	抗凝 射频消融（往往是最好的治疗） 有时 DC 复律、地高辛、β 受体阻滞剂和/或维拉帕米
异位室上性心动过速 异位室上性心律 室上性病灶（通常在心房）引起各种心律。诊断依靠心电图。无症状不需要治疗。 也可以参考 心律失常概述 异位室上性心律包括 房性早搏 房性心动过速 阅读更多 (如房性心动过速)	有时，直流电复律、心率控制药物（地高辛除外）、抗心律失常药、超速起搏和/或消融
折返性室上性心动过速，如房室结折返性心动过速	加强迷走神经张力动作 房室结阻滞药物（例如，beta阻滞剂，维拉帕米） 消融（往往是最好的治疗）
宽QRS综合波心动过速	
室性心动过速 室性心动过速（VT） 室性心动过速是指心率≥120次/分、≥3个连续的室性搏动。症状取决于发作的时限而有不同，可表现为无症状、心悸、血流动力学紊乱甚至死亡。诊断依据心电图。比短暂略长的室性心动过速，其治疗是用心脏电复律还是抗心律失常药，取决于症状。如果必要，长期治疗可采用植入性心脏复律除颤器，... 阅读更多	立即药物治疗或直流电复律 胺碘酮、索他洛尔、普罗帕酮、利多卡因、美西律、氟卡尼、射频消融术 有时是植入性除颤器
尖端扭转型室性心动过速 尖端扭转型室速 尖端扭转型室性心动过速是一种发生于长QT间期患者的特殊类型的多形性室性心动过速。它的特点为快速、不规则的QRS波群，此QRS波群似乎围绕心电图（ECG）的基线扭转。这种心律失常可自发地终止或蜕变成心室颤动。它可引起明显的血流动力学损害，常致死亡。诊断靠心电图。治疗是静脉推注镁剂，缩短QT间期的处理措... 阅读更多	如果不稳定，立即进行直流电复律，镁和/或钾，有时会植入除颤器 持续进行必要的镁，钾，beta-受体阻滞剂，异丙肾上腺素或超速心脏起搏治疗 有时是植入性除颤器
心室颤动 心室颤动（VF） 室颤导致心室不协调的颤动，丧失有效收缩。它可在导致晕厥并且数分钟内死亡。治疗措施是心肺复苏，包括立即除颤。 也可以参考 心律失常概述 室颤（VF）是由多个子波折返的电活动所致，心电图（ECG）表现为非常快速的围绕基线的波动，时间上和形态上完全不规则。... 阅读更多	除颤 有时也可用药（如胺碘酮） 有时是植入性除颤器
Brugada综合征 Brugada综合征 Brugada综合征是一种遗传性心脏电生理异常，可导致晕厥和猝死的危险性增加。 也可以参考 心律失常概述 涉及几种不同的突变，大部分影响SCN5A基因编码电压依赖性心脏钠离子通道的alpha-亚基。通常情况下，患者无... 阅读更多	通常是直流电复律或植入性除颤器
*总是需要识别和纠正病因及加重因素（如电解质异常，低氧血症，药物）。	
AV = 房室；直流电 = 直流电。	

Q: 异位室上性心动过速的诊断依靠什么?

A: ["心电图"]

Zh.Case

Case: 产褥期抑郁症病历1例
基本信息: 女
主诉: 患者因“少言，焦虑一周”入院
现病史: 患者于一周前出现少言，焦虑，失眠等症状，在家未进行治疗，遂入我院。
既往史: 体健
查体: T：36.3℃，P：87次/分，R：18次/分，BP：105/70mmHg神志清楚 精神淡漠，体检配合，头部端正，甲状腺无肿大，胸部对称，心肺听诊无异常。
初步诊断: 产褥期抑郁症
诊治经过: 入院后给予指导家属对产妇要耐心，关心体贴患者，同时给予药物治疗，指导药物服用方法及注意事项，指导按时复诊。

Q: 主诉中提到患者由于什么入院?

A: 少言，焦虑一周

Figure B4: Examples of Table QA and Case QA.

Input:
请根据接下来的内容回答后续的问题。请按照JSON格式要求直接输出答案，格式要求：{"答案": "xxx"}。要求答案来自所给内容，严禁要给出无关文本。
内容：
{heystack_prefix_part}{needle}{heystack_suffix_part}
问题：*{question}*

答案：

Input:
Please answer the question based on the context. Please output the answer directly according to the JSON format requirements. The format requirements is: {"answer": "xxx"}. The answer is required to come from the given content, and irrelevant text is strictly prohibited.
Context:
{heystack_prefix_part}{needle}{heystack_suffix_part}
Question: *{question}*

Answer:

Figure B5: Prompt of the NIAH Tasks.

Input:
在某个月光皎洁、云雾缭绕的夜晚，南极洲上空有一只小星星睁开了眼睛往下面看，它很无聊于是开始全神贯注地数地面一共有多少只小企鹅。请帮助小星星收集所数的小企鹅只数，按照如下格式：{"小星星": [x, x, x, ...]}，不要求和，[x, x, x, ...]中数字为小星星每次数小企鹅的只数，仅以JSON格式输出结果，不需要输出任何解释。
*{heystack_part1}*小星星看向一小块区域，数了*{number1}*只小企鹅。*{heystack_part2}*小星星看向一小块区域，数了*{number2}*只小企鹅。*{heystack_part3}...{heystack_partn}*

Input:
On a moonlit and misty night, a little star in the sky above Antarctica opened its eyes and looked down, it was bored and started to count the number of little penguins on the ground. Please help the little star collect the number of little penguins, for example: {"little_star": [x, x, x, ...]}. The summation is not required, and the numbers in [x, x, x, ...] represent the counted number of little penguins by the little star. Only output the results in JSON format without any explanation."
*{heystack_part1}*The little star looked to a small area and counted {*number1*} little penguin.*{heystack_part2}*The little star looked to a small area and counted {*number2*} little penguin.*{heystack_part3}...{heystack_partn}*

Figure B6: Prompt of the Counting Tasks (Type of Rep., Inc., and Shuf.).

Input:
在某个月光皎洁、云雾缭绕的夜晚，南极洲上空有一只小星星睁开了眼睛往下面看，它很无聊于是开始全神贯注地数地面一共有多少只小企鹅。请帮助小星星收集所数的正确小企鹅只数，按照如下格式：{"小星星": [x, x, x, ...]}，不要求和，[x, x, x, ...]中数字为小星星每次数小企鹅正确的只数，仅以JSON格式输出结果，不需要输出任何解释。
*{heystack_part1}*小星星看向一小块区域数了*{false_number1}*只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是*{true_number1}*只小企鹅。*{heystack_part2}*小星星看向一小块区域数了*{false_number2}*只小企鹅，但发现数错了，于是又数了一遍，这次数对了，是*{true_number2}*只小企鹅。*{heystack_part3}...{heystack_partn}*

Input:
On a moonlit and misty night, a little star in the sky above Antarctica opened its eyes and looked down, it was bored and started to count the number of little penguins on the ground. Please help the little star collect the correct number of little penguins, for example: {"little_star": [x, x, x,...]}. The summation is not required, and the numbers in [x, x, x,...] represent the correctly counted number of little penguins by the little star. Only output the results in JSON format without any explanation.
*{heystack_part1}*The little star looked to a small area and counted *{false_number1}* little penguins, but found that a mistake had been made, so the counting was done again, and this time *{true_number1}* little penguins was counted correctly. *{heystack_part2}*The little star looked to a small area and counted *{false_number2}* little penguins, but found that a mistake had been made, so the counting was done again, and this time *{true_number2}* little penguins was counted correctly. *{heystack_part3}...{heystack_partn}*

Figure B7: Prompt of the Counting Tasks (Type of Cor.).

Input:
请完成医疗术语标准化任务，从给定的术语库中选出输入医疗名词对应的标准术语，标准化结果按照下面的JSON格式输出：{"result": "xxx"}
医疗名词：*{medical_phrase}*
术语库：*{termbase}*

标准化结果：

Input:
Please complete the medical terminology normalization task by selecting the standard terminology that corresponds to the input medical noun from the given Termbase, and then output the normalized result in the following JSON format: {"result": "xxx"}
Medical Phrase: *{medical_phrase}*
Termbase: *{termbase}*

Normalization result:

Figure B8: Prompt of the Term Tasks.

Input:
请给定一些三元组，格式为 实体1|关系|实体2，请找出能回答所提供问题的三元组，回答按照下面的JSON格式。给出的答案三元组仍需保持提供的格式。仅限从提供的三元组中给出答案，严禁给出答案JSON以外的内容：{"result": ["xxx", "xxx", "…"]}
三元组：*{triplets}*
问题：*{question}*

答案：

Input:
Given some triplets in the format Entity1|Relation|Entity2, please find the triplets that can answer the provided question. The answer is in the JSON format below. The given answer triplets must still be in the format provided. Answers can only be given from the provided triplets, and answers other than JSON are strictly prohibited: {"result": ["xxx", "xxx", "…"]}
Triplets: *{triplets}*
Question: *{question}*

Answer:

Figure B9: Prompt of the KG Tasks.

Input:
给定一些markdown格式的表格，请根据表格给出后续问题的答案。给出的答案需符合下面的JSON格式。仅限从提供的表格中给出答案，严禁给出未提供的内容，严禁给出额外内容：{"result": ["xxx", "xxx", "…"]}
表格：*{tables}*
问题：*{question}*

答案：

Input:
Given some markdown tables, please give answers to the subsequent questions based on the tables. The answers given must conform to the following JSON format. Only answers from the provided tables are allowed. It is strictly forbidden to give answers that are not provided, and it is strictly forbidden to give additional content: {"result": ["xxx", "xxx", "…"]}
Tables: *{tables}*
Question: *{question}*

Answer:

Figure B10: Prompt of the Table Tasks.

Input:
给定一些病例，请根据病例给出后续问题的答案。给出的答案需符合下面的JSON格式。仅限从提供的病例中给出答案，严禁给出未提供的内容，严禁给出额外内容：{"result": ["xxx", "xxx", "…"]}
病例：*{medcases}*
问题：*{question}*

答案：

Input:
Given some medical cases, please give answers to the follow-up questions based on the cases. The answers given must conform to the following JSON format. Only answers based on the cases provided are allowed. It is strictly forbidden to give answers that are not provided or to give additional content:{"result": ["xxx", "xxx", "…"]}
Medcases: *{medcases}*
Question: *{question}*

Answer:

Figure B11: Prompt of the Case Tasks.