

A Dataset Construction Details

We introduce the details of constructing the data in this section: The overall idea is: Firstly, we sample attributes for each person, allowing the benchmark to be generated on the fly and remain independent of any parametric knowledge. Second, we use these attributes to generate coherent biographical text for each individual. Finally, we concatenate multiple bios to construct a full context, where the configuration can be freely adjusted to control the task setup, thus making the framework highly extensible and interpretable.

Attribute Sampling To separate the parametric knowledge within LLMs and enable on-the-fly generation and inspired by the bioS dataset [19], we sample attribute values and corresponding sentence templates uniformly from a collected pool. Specifically, each biography includes seven attributes: *full name*, *birthdate*, *birthplace*, *hobby*, *graduated university*, *major*, and *working city*. We generate 100 unique first, middle, and last names independently using LLaMA-3.1-8B-Instruct and ensure that the resulting full names are unique. Birthdates are sampled uniformly from 1950-01-01 to 2001-12-31. For the other attributes, we extract values from datasets on Kaggle³, selecting the top 500 most common universities and 300 most common working cities.

Bio construction To generate a coherent bio for each individual, we manually write a clear and straightforward description template for each attribute. These templates are used to construct the biography as a sequence of six sentences (excluding the full name) in the bios construction stage. In the standard setting, all biographies use the same sentence templates, and the attribute order is fixed for consistency.

To support evaluation under more semantically diverse conditions, we also provide various paraphrases for each template generated from LLaMA-3.1-8B-Instruct. Each paraphrase is manually reviewed to ensure clarity and eliminate ambiguity.

To maintain control over content structure and quality, paraphrasing is applied at the sentence level only, and we retain the original attribute order since variations in order showed a negligible effect on the model performance.

Context Synthesis We use a controllable context construction based on three key configurations: *key information number*, *key information position*, and *distractor density*. Key information number refers to the number of information required to answer the question, and key information position means the position of the question information. Moreover, we introduce an exclusive feature distractor density, which represents the density of the same attribution appears within the context. Our experiments show that the knowledge density can be another strong bottleneck for the long-context tasks. Given those configs, we construct the context in a needle-insertion manner. Specifically, we first construct the haystack by keep concatenating the bios until it reaches a context threshold and then inserting the questioned bios. We use a specific config to control the position where the question bios are inserted. Then we got a sample context and its corresponding question-answer pair.

B Data Statistics

We provide the statistics across the average number of biographies in each task in Table 2 and the average token length for all biographies in Table 3.

C Task Description

This subsection will outline our motivation and explain how we developed all the current tasks in the proposed bench.

The tasks are split into three categories: understanding, reasoning, and trustworthiness, representing the core capabilities required to solve the tasks. An overview of the benchmark is presented in Table 4.

³<https://www.kaggle.com/datasets>

Table 2: The average number of biographies in a randomly generated Longbiobench dataset.

Task/Length	2	8	16	32	64	128
standard	14.93	73.77	152.19	308.83	622.18	1248.75
paraphrase	12.88	62.36	128.01	263.86	528.61	1060.80
pronoun	15.07	75.21	156.65	316.81	636.10	1276.81
multi_standard	12.05	70.87	149.28	305.35	617.94	1246.20
calculation	12.84	76.34	161.12	330.70	668.71	1348.31
rank	12.86	75.77	160.62	329.90	667.42	1345.70
multihop	12.06	70.88	149.27	305.82	619.39	1246.22
twodiff	12.85	76.31	161.42	330.84	670.33	1347.75

Table 3: The average number and standard deviation of tokens within each biography tokenized by Qwen2.5-7b-Instruct.

Task/Length	2	8	16	32	64	128
standard	105.65 _{8.35}	105.42 _{8.40}	105.45 _{8.41}	105.52 _{8.39}	105.54 _{8.26}	105.57 _{8.31}
paraphrase	126.85 _{12.68}	125.16 _{12.15}	125.59 _{11.60}	123.48 _{11.45}	124.14 _{11.23}	124.14 _{11.21}
pronoun	103.13 _{5.89}	103.25 _{7.33}	102.37 _{7.84}	102.90 _{7.73}	103.25 _{7.60}	103.28 _{7.46}
calculation	97.69 _{8.41}	97.58 _{8.32}	97.61 _{8.43}	97.62 _{8.42}	97.78 _{8.46}	97.61 _{8.41}
multihop	106.32 _{8.56}	105.74 _{8.47}	105.65 _{8.44}	105.65 _{8.36}	105.57 _{8.45}	105.56 _{8.32}
multi_standard	105.72 _{8.31}	105.66 _{8.52}	105.56 _{8.52}	105.77 _{8.38}	105.80 _{8.43}	105.56 _{8.31}
twodiff	97.63 _{8.29}	97.63 _{8.32}	97.44 _{8.33}	97.57 _{8.39}	97.55 _{8.36}	97.64 _{8.41}
rank	97.45 _{8.46}	97.65 _{8.41}	97.60 _{8.42}	97.70 _{8.46}	97.90 _{8.60}	97.76 _{8.43}

Standard Information Retrieval (Standard). We start with the simplest retrieval settings as the *Standard* version. To allow for increments in task difficulty, we ensure that all statements are expressed using the simplest and most direct sentences, such as “The hobby of *{person}* is”. This also avoids ambiguity for models, which establishes a robust baseline for subsequent, more challenging tasks. The model will be asked to retrieve a specific attribute for a person.

Multi Information Retrieval (Multi_standard). To further challenge the model to simultaneously retrieve information across different context locations, we upgrade the single retrieval task to a multi-retrieval task by asking models to retrieve *n* attributes from *n* people instead of one, where *n* can be modified when constructing the dataset. Here we let *n* equal 2, 5, 10 by default.

Retrieval on Paraphrased Bios (Paraphrase). To demand stronger contextual understanding, we paraphrase the expression of attributes within the bios. This prevents models from relying on exact matches between questions and sentences to locate answers. As a result, we can control for other confounding factors and more accurately assess the models’ true comprehension capabilities by examining the performance gap relative to the *Standard* version.

Retrieval on Bios stated with Pronoun (Pronoun). This task is an extension of the paraphrasing task. Based on the paraphrase setting, each bio is rewritten as a self-introduction. All sentences that describe a person’s attributes are expressed in the first person, with the individual’s name appearing only at the beginning of the bio. This design builds upon sentence-level understanding in paraphrasing and further challenges the LLM’s ability to understand the paragraph-level semantics, which is the hardest task in the understanding category.

Calculating the Ages (Calculation). For the reasoning level, we require the LLM to reason on the retrieved information. The calculation task asked LLM to calculate the subtraction of the ages of two people. We use subtraction here instead of the summation to make this task expandable to the TwoDiff task later. Besides, to prevent the ages of people from changing over time, we note that all birthdate attributes are replaced by the specific ages under this setting.

Ranking the Ages (Rank). We extend the Calculation setting to ranking the ages of different people so that we can freely define the number of retrieved information by specifying the number of

Table 4: Task Overview for Retrieval Benchmark

Task	Description	Metric	Example
Understanding			
Standard	Retrieve a specific attribute of one person.	Acc	Attribute: The hobby of {P1} is dandyism. Question: What’s the hobby of {P1}?
Multi_standard	Retrieve multiple attributes of different people.	All-or-Nothing Acc	Attribute: The hobby of {P1} is dandyism. {P2} is mycology. Question: What’s the hobby of {P1} and {P2}?
Paraphrase	Attribute expressions are paraphrased.	Acc	Attribute: {P1} worked in Dhaka. Question: Which city did {P1} work in?
Pronoun	Bio written from first-person view.	Acc	Attribute: I was born on 1993-06-26. Question: What is the birthday of {P1}?
Reasoning			
Calculation	Compute age difference between two people.	Acc	Attribute: {P1} is 61, {P2} is 43. Question: What’s their age difference?
Rank	Rank people by age.	Acc	Attribute: {P1} is 61, {P2} is 43. Question: Rank from youngest to oldest.
Multihop	Retrieve an attribute via cross-person reference.	Acc	Attribute: {P1} born in Santa Paula. {P2} born same place as {P1}. Question: Birthplace of {P2}?
Twodiff	Identify two people with specific age difference.	Acc	Attribute: {P1} is 61, {P2} is 43. Question: Who has 18 years age difference?
Trustworthy			
Citation	Answer plus source citation.	Citation Acc	Attribute: Bio [1]: {P1} born in Santa Paula. Question: Which university did Isabel graduate from?
IDK	No-answer case detection.	Refuse while Answer Acc	Attribute: Attribute removed. Question: What’s the hobby of {P1}?

900 people to be ranked. Here we let n equal 2 and 5 by default since we observe that 5 retrieval ranking
901 task is challenging enough for most models.

902 **Retrieve Two People Satisfying the Age Difference (Twodiff).** In this task, we give LLM an age
903 difference and ask LLM to retrieve two people whose age difference satisfies the age difference. This
904 demands that LLMs plan on retrieving the target instead of directly retrieving it based on the given
905 information. We design this task as a naive simulation of the scenario where LLMs are asked to do
906 some constrained retrieval (e.g in pairs trading, traders look for two stocks whose price difference
907 equals a predetermined target).

908 **Multi-hop Retrieval (Multihop).** Multi-hop question answering is a popular setting in document
909 question answering. In our benchmark, we replicate this setting by randomly changing the expression
910 of an attribute into “The {attribute} of {person 1 name} is the same as {person 2 name}” where
911 we ensure that person 2 appeared after person 1 in the context. This forces LLM to understand the
912 expression and retrieve sequentially across different positions in the context, which is an extended,
913 harder version of multi-retrieval.

914 **Citation (Cite).** Built upon the *Standard* setting, we index the bios and ask the model to retrieve
915 answers while referring to the bio presenting the target attribute with its index. Therefore, for this task,
916 we not only evaluate the final accuracy but also the precision of the model’s citation. Generating with
917 citations has long been an essential ability of trustworthy LLMs. We test their capabilities on citing
918 the correct bios after their answer in this task. To make the citation trackable, we add a number before
919 each bio and ask LCLM to generate both the answer and its corresponding number and measure the
920 accuracy of the citation in the end. As an extended version of *Standard/Multi_standard*, we set the
921 number of information pieces to 1 and 2 by default.

922 **I don’t know (IDK).** Expressing uncertainty is a critical aspect of trustworthy behavior in LLMs
923 [39]. To evaluate this, we simulate a controlled setting in which the target information is deliberately
924 removed, and the LLM is prompted to respond with “The answer is not explicitly stated.” Observing
925 that weaker LLMs tend to refuse all questions when the task becomes more difficult (e.g. with longer
926 context or a harder version), we evaluate models based on a combination of standard retrieval and
927 uncertainty expression. Specifically, a model is considered to have successfully passed a question only

if it (1) correctly retrieves the attribute when the relevant information is present, and (2) appropriately refuses to answer when the attribute sentence is removed.

D Analysis: Density and Needle position v.s Performance

To further investigate the factors influencing the performance of LCLMs, we conduct a stress test on Qwen2.5-Instruct-7B-1M by controlling the **position of the answer information** and the **density of distractors** as the variables while keeping the context length and task fixed. Specifically, we adjust the percentage of the haystack depth to insert the needle for controlling the position and set the probability of generating the same attribute as the needle attribute to control the distractor density. We evaluate the model on two representative tasks: the simplest reasoning task, *calculation*, and the most challenging understanding task, *pronoun*, as their baseline performances lie in a moderate range—neither too high nor too low. The results are visualized in the heatmap shown in Fig. 8.

Our key observations from the figure are as follows. We first observe a strong negative correlation between distractor density and model performance, suggesting that beyond context length, higher distractor density is a key factor contributing to the difficulty LCLMs face with long-context tasks. Second, we observe the lost-in-the-middle [28] phenomenon with our proposed synthetic task *calculation*, where performance declines when the needle appears in the middle of the context. Interestingly, this trend is less evident in the *pronoun* task. We conjecture that this is because the model already performs relatively well on the *pronoun*. Finally, both of these effects—performance decay with density and positional sensitivity—are more pronounced in the reasoning task than in the understanding task. This suggests that certain failure patterns emerge only under sufficiently challenging conditions, reinforcing the need to continue developing more difficult long-context benchmarks.

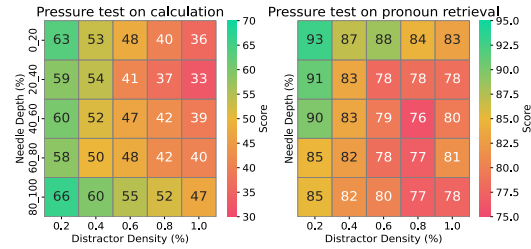


Figure 8: The performance on calculation (left) task and pronoun retrieval (right) task corresponding with the answer depth and distractor density. Y-axis shows the percentage of insertion depth in the context and x-axis shows the percentage of the distracted information appeared in the context.

E Model Details

We provide the details of all models evaluated in Table 5.

F Full Results

The full results are shown in Fig. 9.

G Prompts

The prompt we used for each tasks are shown as follows:

Standard/Paraphrase/Pronoun:

(System): Your task is to answer the user’s question based on a long context, which consists of many bios. Output the answer only. Don’t explain or output other things.

(User): "Context: {given_context}"

Question: {question}"

(Assistant): Based on the provided context, {question_prefix}"

Table 5: Details of all evaluated Long-Context Language Models

Models	Release Date	Size	Support Context Length
gpt-4.1-nano-2025-04-14	2025-04	-	128,000
gpt-4o-2024-11-20	2024-11	-	128,000
gpt-4o-mini-2024-07-18	2024-07	-	128,000
internlm3-8b-instruct	2025-01	8B	131,072
Qwen2.5-7B-Instruct-1M	2025-01	7B	1,010,000
Qwen2.5-14B-Instruct-1M	2025-01	14B	1,010,000
Llama-3.3-70B-Instruct	2024-12	70B	131,072
Llama-3-8B-ProLong-512k-Instruct	2024-10	8B	524,288
Qwen2.5-7B-Instruct	2024-09	7B	131,072
Qwen2.5-72B-Instruct	2024-09	72B	131,072
Llama-3.2-1B-Instruct	2024-09	1B	131,072
Llama-3.2-3B-Instruct	2024-09	3B	131,072
Phi-3.5-mini-instruct	2024-08	4B	131,072
Llama-3.1-8B-Instruct	2024-07	8B	131,072
Llama-3.1-70B-Instruct	2024-07	70B	131,072
Mistral-Nemo-Instruct-2407	2024-07	12B	131,072
glm-4-9b-chat-1m	2024-06	9B	1,048,576
Phi-3-medium-128k-instruct	2024-05	14B	131,072

Multi_Standard:

(System): Your task is to answer all the user’s questions based on a long context, which consists of many bios. Output only the answers for each question sequentially. Don’t explain or output other things.

(User): Context: {given_context}

The Questions are as follows:

question

Answer each question in sequence.

(Assistant): Based on the provided context, the answer is

965

Rank:

(System): Following the format of the examples, your task is to rank the users based on their bios in a long context.

(User): Context: {given_context}

examples_with_cot

Question: {question}

(Assistant): Based on the provided context,

966

Calculation:

(System): Your task is to calculate the age difference of the given people based on the given instruction from a long context containing multiple bios.

(User): Context: {given_context}

examples_with_cot

Question: {question}

(Assistant): Answer: Based on the provided context,

967

	internlm3-8b-instruct						glm-4-9b-chat-1m						Qwen2.5-7B-Instruct-1M						Qwen2.5-7B-Instruct						
standard	99.4	98.8	99.2	97.1	94.6	87.9	98.6	98.2	97.1	97.6	92.5	82.9	99.9	99.5	99.6	98.9	98.6	94.9	99.8	82.9	72.2	71.9	56.9	26.1	
multi_standard_2	80.6	91.5	97.4	96.2	89.1	71.2	96.2	95.1	93.0	89.5	81.1	38.1	99.4	99.0	98.6	98.0	95.5	91.1	98.9	68.5	55.4	45.4	22.4	5.6	
multi_standard_5	-	88.6	76.5	70.5	58.6	36.6	-	94.2	91.1	89.1	76.5	54.8	-	98.2	97.5	96.1	92.5	82.4	-	38.6	21.0	12.0	3.4	0.2	
multi_standard_10	-	93.2	86.4	78.8	59.9	19.2	-	87.4	87.1	82.8	68.4	34.2	-	95.0	94.1	93.8	85.9	69.2	-	17.5	4.0	1.4	0.8	0.0	
paraphrase	95.4	97.6	98.6	97.6	94.4	84.9	96.8	97.5	92.8	90.2	80.0	64.1	99.4	98.8	99.2	98.5	95.6	91.0	98.2	83.5	69.1	65.2	49.1	22.0	
pronoun	95.0	96.8	94.8	91.8	77.9	59.2	93.0	87.6	83.1	70.6	54.6	33.0	95.8	98.0	97.4	95.0	90.6	81.4	97.8	72.6	57.2	41.6	29.4	11.8	
calculation	99.5	98.8	97.8	95.5	89.4	74.4	99.5	98.0	98.2	97.6	97.1	92.8	100.0	99.9	99.1	98.2	96.4	90.2	100.0	74.6	76.2	75.1	50.7	21.6	
rank_2	84.2	72.0	67.1	64.4	60.0	56.1	77.5	69.9	76.2	85.0	60.4	55.4	78.6	70.1	69.2	65.4	64.2	63.4	78.1	63.6	58.6	58.6	49.0	48.1	
rank_5	-	4.5	3.0	3.0	2.5	1.5	-	69.8	84.9	61.6	48.8	23.5	-	7.1	7.0	6.9	4.1	2.1	-	5.8	4.7	3.4	2.6	1.0	
multihop_2	92.0	60.0	57.8	41.8	20.4	8.2	71.8	43.6	24.4	19.6	17.2	7.5	89.8	76.4	57.9	51.5	33.8	25.9	95.4	37.9	39.0	24.0	9.8	1.5	
multihop_5	-	3.6	1.5	1.4	0.8	0.1	-	4.6	3.4	0.8	0.5	0.0	-	1.1	0.8	0.2	0.5	0.0	-	0.4	1.2	0.9	0.4	0.4	
twodiff	40.1	45.7	36.6	20.2	11.7	7.3	-	1.1	0.9	0.6	0.6	0.8	1.1	51.5	36.9	21.1	14.6	6.7	3.7	46.3	32.9	29.1	17.6	8.9	4.8
citation	100.0	99.6	98.5	91.7	83.9	58.1	99.9	98.3	94.5	85.3	64.1	38.0	99.9	98.9	98.4	84.4	73.5	76.6	99.7	59.2	41.5	30.1	18.5	6.6	
citation_2	99.8	97.0	93.9	88.6	66.1	30.7	99.5	98.0	94.1	83.9	52.1	20.2	100.0	98.9	86.5	61.2	42.1	39.7	99.6	42.3	17.6	9.5	1.7	0.0	
IDK	99.4	98.6	99.2	95.4	80.0	86.5	99.4	98.6	99.2	95.4	80.0	86.5	99.4	98.6	99.2	95.4	80.0	86.5	99.4	98.6	99.2	95.4	80.0	86.5	
	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	
	Qwen2.5-72B-Instruct						Qwen2.5-14B-Instruct-1M						Phi-3.5-mini-instruct						Phi-3-medium-128k-instruct						
standard	99.4	94.0	88.8	84.2	72.6	56.2	99.6	100.0	99.6	99.4	98.4	97.1	99.8	98.2	97.2	96.8	90.0	43.2	90.2	89.4	94.4	90.1	83.9	21.2	
multi_standard_2	99.6	99.1	98.2	92.6	51.8	32.0	99.5	98.9	99.5	99.1	96.9	94.1	99.2	94.1	91.1	82.6	56.1	10.0	99.2	96.4	89.8	75.0	52.1	0.4	
multi_standard_5	-	97.6	95.5	81.4	23.4	11.0	-	98.5	98.1	96.0	93.6	85.1	-	62.9	55.2	46.8	20.8	0.4	-	87.1	72.8	51.9	28.1	0.0	
multi_standard_10	-	95.5	92.4	31.9	5.8	4.0	-	96.8	96.2	94.8	89.1	74.4	-	33.8	29.9	23.9	10.1	0.0	-	73.4	58.9	38.8	10.0	0.0	
paraphrase	99.4	92.1	85.4	79.8	59.0	40.5	99.6	99.1	99.5	99.5	97.9	96.0	98.2	97.8	96.6	93.2	81.2	35.8	88.4	85.0	92.9	89.1	79.4	33.2	
pronoun	98.5	85.6	69.8	56.0	41.5	22.1	99.6	98.9	98.0	98.5	95.8	89.1	98.8	90.0	90.9	76.0	51.4	18.6	92.6	86.4	84.9	71.8	51.1	13.1	
calculation	100.0	95.8	86.8	84.2	74.2	48.8	100.0	100.0	99.8	99.6	98.0	97.5	100.0	94.5	92.2	89.8	67.1	8.1	100.0	99.1	97.5	91.1	80.9	3.0	
rank_2	97.5	85.6	80.2	86.4	96.8	99.8	96.4	90.6	91.0	90.4	95.5	99.5	99.9	99.9	98.8	99.6	97.4	98.2	84.2	59.8	54.9	53.4	52.6	46.2	
rank_5	-	28.6	16.0	14.2	19.5	34.8	-	37.1	30.4	34.4	69.2	85.1	-	53.0	48.4	51.5	71.0	60.1	-	3.1	1.9	1.1	1.8	0.6	
multihop_2	99.5	89.0	74.5	62.5	43.5	16.6	99.2	94.2	92.0	88.9	83.4	66.2	89.6	62.1	57.1	47.2	15.8	0.6	96.4	57.4	33.8	17.0	7.2	0.1	
multihop_5	-	32.9	18.1	7.4	2.0	0.9	-	22.8	26.1	18.9	10.9	3.1	-	1.2	1.4	0.1	0.4	0.1	-	1.6	0.9	0.6	0.0	0.0	
twodiff	5.2	2.8	5.5	6.2	5.2	3.0	35.4	39.6	32.1	18.0	13.7	7.4	28.9	34.5	33.9	27.1	24.6	26.1	1.5	1.6	1.0	0.9	0.9	0.9	
citation	100.0	90.3	75.9	47.9	25.4	9.1	100.0	97.5	91.5	96.2	91.3	88.0	97.8	26.6	22.9	15.7	5.1	1.6	99.8	93.9	77.0	61.3	43.4	5.2	
citation_2	100.0	85.7	62.5	26.4	8.9	1.5	100.0	97.9	90.3	87.9	81.2	68.5	97.9	9.6	4.2	1.5	0.1	0.0	100.0	85.3	56.8	24.7	9.4	0.0	
IDK	70.5	60.9	48.5	30.0	16.2	10.0	94.6	79.9	81.1	80.9	61.1	49.4	99.5	97.6	97.2	96.8	89.9	43.0	44.1	70.2	86.9	88.5	83.6	1.0	
	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	
	Mistral-Nemo-Instruct-2407						Llama-3.2-3B-Instruct						Llama-3.2-1B-Instruct						Llama-3.1-8B-Instruct						
standard	99.5	98.6	93.6	71.6	18.0	5.0	99.4	90.6	85.1	63.8	40.2	27.5	97.5	75.4	48.9	38.4	20.9	8.8	99.6	99.4	99.6	99.4	97.2	79.2	
multi_standard_2	99.1	99.6	89.8	37.4	0.0	0.0	97.4	81.5	65.1	32.1	12.0	6.8	89.8	38.1	9.6	3.5	1.8	0.8	41.2	97.6	98.1	97.0	93.0	53.9	
multi_standard_5	-	78.9	71.1	15.5	0.0	0.0	-	47.2	28.0	6.1	1.6	0.4	-	16.8	0.8	0.4	0.0	0.0	-	95.9	93.5	90.2	82.0	22.8	
multi_standard_10	-	95.6	64.8	4.0	0.0	0.0	-	19.2	9.9	1.6	0.1	0.2	-	4.6	0.2	0.1	0.0	0.0	-	91.9	90.4	82.6	68.8	8.0	
paraphrase	99.5	98.2	91.1	53.2	10.6	2.6	96.5	86.1	72.9	42.5	24.4	15.1	91.0	61.8	32.9	23.8	9.8	2.2	99.0	98.8	98.5	97.8	94.8	59.6	
pronoun	99.4	96.4	84.8	50.2	9.4	1.8	92.9	62.2	62.8	34.0	22.9	7.5	85.6	-	22.1	13.1	6.2	1.0	97.1	95.6	93.0	85.6	69.2	35.0	
calculation	100.0	99.9	95.0	61.4	4.0	3.2	99.9	93.1	82.5	49.9	39.9	20.0	100.0	75.8	46.9	46.6	15.0	5.2	100.0	99.9	99.1	98.8	88.6	31.9	
rank_2	97.5	85.9	68.4	66.6	57.9	48.5	80.5	69.4	63.1	65.5	65.4	54.5	49.4	71.6	54.0	27.6	84.8	92.8	88.5	81.4	78.8	76.4	66.8	60.1	
rank_5	-	15.9	1.9	1.1	1.1	0.6	-	18.0	8.2	35.1	49.6	37.6	-	0.5	1.0	2.0	3.1	5.2	1.4	28.1	15.4	11.2	5.5	1.4	
multihop_2	99.4	92.6	32.1	1.6	0.4	0.2	94.4	66.6	38.8	15.6	8.2	1.5	66.4	10.4	1.9	0.4	0.5	0.2	97.2	93.5	91.8	81.6	48.4	2.8	
multihop_5	-	22.5	3.0	0.4	0.2	0.1	-	11.9	3.6	2.5	0.6	0.6	-	1.2	0.4	0.2	0.4	0.2	-	29.2	20.1	9.0	1.4	0.1	
twodiff	33.6	30.7	2.0	2.0	0.9	0.9	0.8	0.9	0.4	0.6	2.0	1.6	0.8	0.9	1.5	4.4	3.6	12.1	7.2	7.6	8.0	9.6	10.9	3.2	
citation	94.1	98.9	87.1	44.6	14.7	0.0	89.5	94.6	87.8	60.8	23.1	9.5	47.8	1.1	1.6	0.0	0.0	0.0	100.0	99.7	97.5	91.1	88.4	50.1	
citation_2	0.0	0.0	0.0	0.0	0.0	0.0	63.5	89.0	66.7	35.5	4.8	0.7	42.1	1.6	0.6	0.5	0.0	0.0	97.2	99.5	97.5	90.5	77.6	28.1	
IDK	97.4	93.9	91.5	60.8	11.8	0.1	48.5	74.4	77.0	61.1	40.1	27.4	97.5	75.2	48.9	38.1	20.8	8.8	84.0	82.5	82.4	84.4	97.1	79.0	
	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	2k	8k	16k	32k	64k	128k	
	Llama-3.1-70B-Instruct						Llama-3.3-70B-Instruct						Llama-3-8B-ProLong-512k-Instruct						gpt-4o-mini-2024-07-18						
standard	99.6	99.1	99.6	99.2	98.0	60.9	100.0	99.4	99.6	98.5	95.2	19.8	99.6	98.8	98.5	96.5	90.5	85.5	99.9	99.8	99.5	98.6	97.5	88.5	
multi_standard_2	98.8	99.8	99.1	99.0	95.1	0.0	99.9	99.2	98.4	96.9	90.8	1.6	95.2	96.8	95.0	86.5	76.4	65.2	99.5	98.6	98.8	98.1	95.9	82.0	
multi_standard_5	-	98.5	98.4	96.2	85.6	1.4	-	98.5	98.0	94.5	77.2	0.1	-	90.1	83.4	60.8	47.2	32.8	-	98.4	95.9	95.6	92.5	69.6	
multi_standard_10	-	96.1	97.1	93.59																					

Twodiff:

(System): Your task is to find the names of people based on the given instruction from a long context containing multiple bios. Follow the format provided in the examples closely and give the final answer.

(User): Context: {given_context}examples: {question}

(Assistant): Answer:

969

Cite (Standard):

(System): Your task is to answer the user's question with citation based on a long context, which consists of many bios. You must output the answer following with the citation number of the relevant bios strictly surrounded by square brackets such as [1]. Don't explain or output other things.

(User): Context: {given_context}

examples

Question: {question}

Answer:

(Assistant): Based on the provided context, {question_prefix}

Cite (Multi-Standard):

(System): Your task is to answer all the user's questions with citation based on a long context, which consists of many bios. Following the format of the examples, You must output the answer ending with the citation number of the relevant bios strictly surrounded by square brackets such as [1]. You should give the answer and citation for each question sequentially. Don't explain or output other things.

(User): Context: {given_context}

examples

question

Answers:

(Assistant): Based on the provided context,

970

IDK:

(System): Your task is to answer the user's question based on a long context, which consists of many bios. Output the answer only. If you don't know the answer or the answer is not explicitly stated, you should strictly output 'The answer is not explicitly stated'. Don't explain or output other things.

(User): Context: {given_context}

Question: {question}

(Assistant): Based on the provided context, {question_prefix}

971