



LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA

汇报人：吴晨帆

时间：2024年11月18日



研究背景

研究领域

在超长上下文窗口的语言模型（例如GLM-4-9B-1M和Gemini 1.5）应用于长文档问答场景中，面临两大挑战：一是文档长度过长导致用户难以验证模型提供信息的准确性；二是模型输出常常偏离原文，出现幻觉现象，严重影响了模型输出的可信度。

现有方案

通过RAG或者后处理的方式，让大模型在回复中加入引用信息。这种方式在开放域问答和智能搜索引擎（例如 New Bing, Perplexity AI）中被广泛应用。

潜在问题

但 RAG 在长文本问答中，往往会丢失文本信息，使得模型回复的正确性下降。而后处理方式会让pipeline变得特别复杂，用户需要等待很长时间。



目录

DIRECTORY

一、

能力评测：LongBench-Cite

二、

数据构造：CoF

三、

模型训练：LongCite





NO.1

能力评测：LongBench-Cite



问题定义 (LQAC)



Question: What legislation governs wastewater treatment in Canada?

(a) Chunk-level citations



Context (divided into 128-token chunks): (...)

Chunk [6]: (...)

Chunk [7]: Water pollution control legislation. (...) To accomplish its water quality objectives the MOE produced legislation in the form of the Ontario Water Resources Act (OWRA). This legislation in conjunction with various regulations

Chunk [8]: made under the OWRA set out legal requirements (...)



Answer with chunk-level citations: (...)

<statement>- The Ontario Water Resources Act (OWRA): This is provincial legislation in Ontario that sets out legal requirements for managing environmental issues related to water.<cite>[7]</cite></statement> (...)



Incomplete sentences; Need further pinpointing; Bad user experience.

(b) Sentence-level citations



Context (divided into sentences): <C0> The water pollution (...) <C22>To accomplish its water quality objectives the MOE produced legislation in the form of the Ontario Water Resources Act (OWRA). <C23>This legislation in conjunction with various regulations made under the OWRA set out legal requirements for managing environmental issues. <C24>The City of Guelph, (...)



Answer with sentence-level citations: (...)

<statement>- The Ontario Water Resources Act (OWRA): This is provincial legislation in Ontario that sets out legal requirements for managing environmental issues related to water.<cite>[22-23]</cite></statement> (...)



Complete sentences; Accurate locating; User-friendly.

定义了一个带引文的长上下文问题解答任务 (LQAC) :

块级引用: 将上下文分割成固定128个token的语义块, 每个引文形式为[k]

句子级引用: 使用NLTK将上下文划分为句子, 每个引文形式为[k]或者[a-b]

数据来源

Figure 1: Comparison between chunk-level and sentence-level citations.

Dataset	Task	Source	Avg Len	Language	#data
MultiFieldQA-en	Single-Doc QA	Multi-field	4,559	English	150
MultiFieldQA-zh	Single-Doc QA	Multi-field	6,701	Chinese	200
HotpotQA	Multi-Doc QA	Wikipedia	9,151	English	200
Dureader	Multi-Doc QA	Baidu Search	15,768	Chinese	200
GovReport	Summarization	Government Report	8,734	English	200
LongBench-Chat	Multi-task	Real-world Query	35,571	English/Chinese	50

评测数据主要来源于现有的双语长文本benchmarks-longbench和longbench-chat



评价指标



LongBench-Cite从两个维度上，由GPT-4o进行自动判断：

正确性：

Correctness: 回答是否正确，与标准答案契合；

Correctness Ratio: 与普通长文本问答相比，加入引用后Correctness是否受损。

引用质量：

Citation Recall: 回答中的每个事实性陈述 (statement) 是否被对应的citation所支持；

Citation Precision: 每个citation是否包含了对应的statement的信息，不是无关的；

Citation F1: $2 * (R * P) / (R + P)$ ，综合考虑recall和precision；

Citation Length: 每个citation对应文本的长度(token数)。长度越短，说明粒度越细、定位越精准。





评价指标

从上图可以看出

开源模型citation f1很低，经常生成错误的或是不符合格式的引用；

闭源模型citation length普遍很高，引用粒度甚至比chunk-level citation还粗，需要进一步精准定位。例如，GPT-4o平均每个引用包含了原文中的6句话；

此外大部分模型再加入引文后使得模型原本的长文本回答能力受损。

Model	Avg		Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport		
	F1	CL	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
<i>Proprietary models</i>																	
GPT-4o	65.6	220	46.7	53.5	46.7	79.0	87.9	<u>80.6</u>	55.7	62.3	53.4	65.6	74.2	67.4	73.4	90.4	79.8
Claude-3-sonnet	67.2	132	52.0	67.8	55.1	64.7	85.8	<u>71.3</u>	46.4	65.8	49.9	67.7	89.2	75.5	77.4	93.9	<u>84.1</u>
GLM-4	65.4	169	47.6	53.9	47.1	72.3	80.1	73.6	47.0	50.1	44.4	73.4	82.3	<u>75.0</u>	82.8	<u>93.4</u>	87.1
<i>Open-source models</i>																	
GLM-4-9B-chat	27.2	96	25.9	20.5	16.7	51.1	60.6	52.0	22.9	28.8	20.1	45.4	48.3	40.9	5.7	8.2	6.3
Llama-3.1-8B-Instruct	19.7	100	14.1	19.5	12.4	29.8	44.3	31.6	20.2	30.9	20.9	22.0	25.1	17.0	16.2	25.3	16.8
Llama-3.1-70B-Instruct	40.4	174	25.8	32.0	23.2	53.2	65.2	53.9	29.6	37.3	28.6	38.2	46.0	35.4	53.4	77.5	60.7
Mistral-Large-Instruct	51.5	132	19.8	23.9	19.0	71.8	80.7	73.8	34.5	40.9	32.1	58.3	67.0	60.1	67.9	79.6	72.5
<i>Our trained models</i>																	
LongCite-8B	72.0	85	62.0	79.7	67.4	<u>74.7</u>	93.0	80.8	<u>59.2</u>	<u>72.1</u>	<u>60.3</u>	<u>68.3</u>	85.6	73.1	74.0	86.6	78.5
LongCite-9B	<u>69.2</u>	<u>91</u>	<u>57.6</u>	<u>78.1</u>	<u>63.6</u>	67.3	<u>91.0</u>	74.8	61.8	78.8	64.8	67.6	89.2	74.4	63.4	76.5	68.2

Table 2: Citation recall (R), citation precision (P), citation F1 (F1), and citation length (CL) of different models on LongBench-Cite using LAC-S strategy. The best and second results are bolded and underlined, respectively.

Model	Avg			Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport		
	C	C _{LQA}	CR	C	C _{LQA}	CR	C	C _{LQA}	CR	C	C _{LQA}	CR	C	C _{LQA}	CR	C	C _{LQA}	CR
<i>Proprietary models</i>																		
GPT-4o	69.4	78.2	88%	61.6	77.4	80%	84.0	88.3	95%	74.5	80.8	92%	81.0	83.3	97%	46.0	61.3	75%
Claude-3-sonnet	77.6	78.3	99%	73.8	77.8	95%	88.6	88.1	101%	81.3	75.3	108%	75.8	80.3	94%	68.4	70.1	98%
GLM-4	73.7	77.2	95%	69.4	79.8	87%	87.6	88.1	99%	76.3	76.5	100%	<u>76.0</u>	<u>75.8</u>	<u>100%</u>	59.4	65.9	90%
<i>Open-source models</i>																		
GLM-4-9B-chat	62.3	70.8	88%	60.4	67.8	89%	74.2	84.9	87%	68.5	71.5	96%	49.3	68.1	72%	59.3	61.6	96%
Llama-3.1-8B-Instruct	52.1	60.2	86%	53.2	61.6	86%	63.9	73.3	87%	64.0	64.5	99%	29.8	39.4	76%	49.6	62.1	80%
Llama-3.1-70B-Instruct	62.0	65.5	95%	60.8	64.6	94%	78.4	78.3	100%	71.3	75.3	95%	43.3	42.5	102%	56.3	66.9	84%
Mistral-Large-Instruct	73.6	76.4	96%	63.8	67.8	94%	88.0	85.3	103%	77.0	77.3	100%	79.0	83.3	95%	60.4	68.3	88%
<i>Our trained models</i>																		
LongCite-8B	<u>71.7</u>	67.6	107%	69.0	68.6	101%	87.0	83.6	104%	70.8	69.0	103%	68.5	62.3	110%	63.0	54.4	116%
LongCite-9B	70.4	65.6	109%	67.6	64.6	105%	84.1	83.3	101%	71.8	67.5	106%	69.0	66.3	104%	59.6	46.4	128%

Table 3: Correctness in LQAC setting (C) using LAC-S strategy, correctness in vanilla long-context QA setting (C_{LQA}), and correctness ratio (CR) of different models on LongBench-Cite. We mark the cases where adding citations improves/hurts correctness (i.e., CR > 1 / CR < 1) in green/red.



NO.2

数据构造: COF





数据构造：CoF

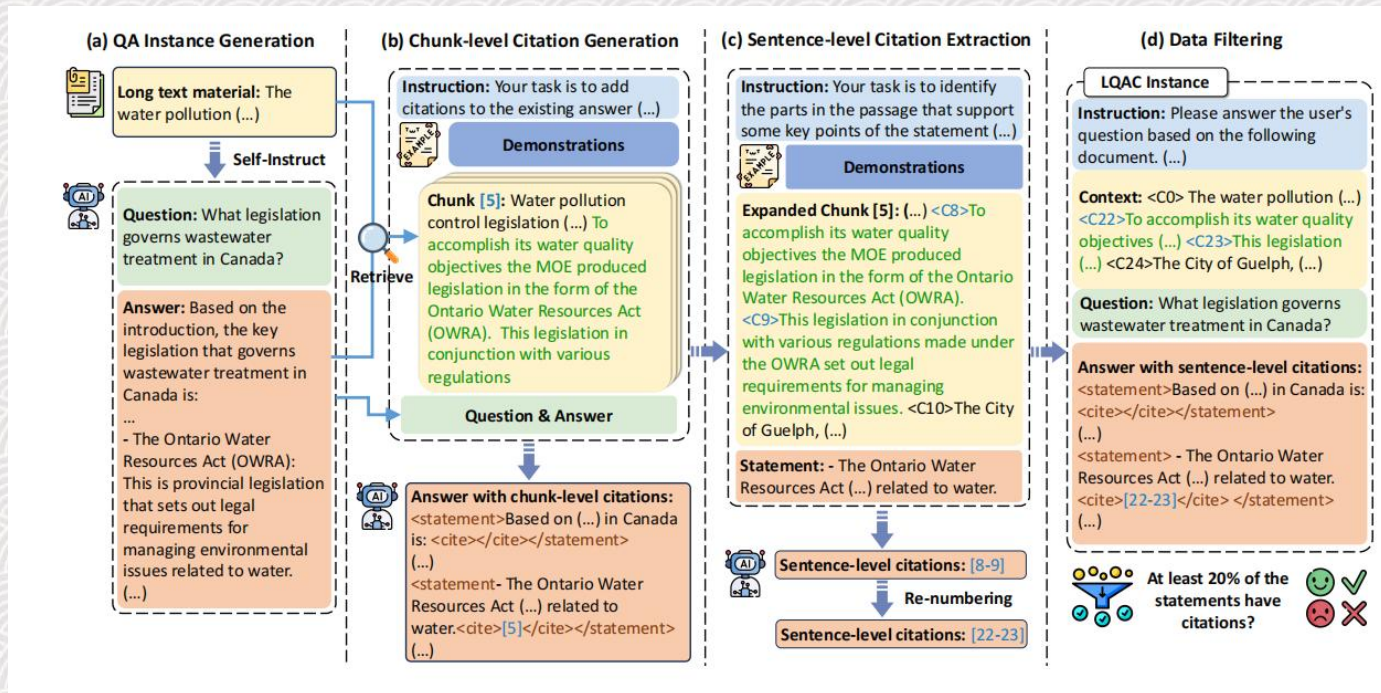
提出了 CoF (Coarse to Fine) 的方法，并利用现有长文本模型来自动构造数据。构造方法遵循两个原则：

原则一：充分利用现有模型的长文本能力，保证问答的质量。采用后处理的方式，先用普通的长文本问答得到答案，再往里加引用。

原则二：生成的citation质量要高，粒度也要细。先生成粗粒度的chunk-level citation，再从中抽取出细粒度的sentence-level citation。



构造流程



问题生成。使用self-instruction的方法，自问自答生成文本问答对。

生成chunk-level citation。使用答案中的句子从原文检索出相关片段。再通过 in-context learning 的方式让大模型把原有答案分成 statements 并加入 chunk-level citation。

抽取sentence-level citation。对于每个 statement，将其对应的 chunk-level citation 中的句子

数据筛选。将前几步的数据整理，得到最后的带有 sentence-level citation 的长文本问答数据。筛去引用过少的数据。这些数据可能没有忠于原文，包含有幻觉。

构造流程

You will receive a passage and a factual statement. Your task is to identify the parts in the passage (i.e., chunks $\langle C\{s1\} \rangle - \langle C\{e1\} \rangle$, $\langle C\{s2\} \rangle - \langle C\{e2\} \rangle$, ...) that support some key points of the statement, and output the chunk number in the format:

[s1-e1]

[s2-e2]

...

”

If the passage contains no key information relevant to the statement, you must output "No relevant information".

Here are some examples:

{Example 1}

{Example 2}

{Example 3}

Now get ready to process the following test case.

[Passage Start]

$\langle C0 \rangle \{Sentence\ 0\}$ $\langle C1 \rangle \{Sentence\ 1\}$ $\langle C2 \rangle \{Sentence\ 2\}$...

[Passage End]

[Statement]

{statement}

[output]

Figure 13: Prompt for sentence-level citation extraction in the CoF pipeline.

Your task is to add citations to the existing answer. Specifically, when a factual statement S in the answer uses information from context snippets 11, 12, ..., 1n, please add citations by appending these snippet numbers to S in the format " $\langle \text{statement} \rangle \{S\} \langle \text{cite} \rangle [\{11\}][\{12\}]...[\{1n\}] \langle \text{cite} \rangle \langle \text{statement} \rangle$ ". For other sentences such as introductory sentences, summarization sentences, reasoning, and inference, you still need to append " $\langle \text{cite} \rangle \langle \text{cite} \rangle$ " to them to indicate they need no citations. Except for adding citations, do not change the original content and format of the existing answer.

Here is an example:

{An Example}

Now get ready to add citations for the following test case.

[Contexts Start]

Snippet [1]

{Chunk 1}

Snippet [2]

{Chunk 2}

Snippet [3]

{Chunk 3}

...

[Context End]

[Question]

{Question}

[Existing Answer Start]

{Answer}

[Existing Answer End]

[Answer with Citations]

Figure 12: Prompt for chunk-level citation generation in the CoF pipeline.



NO.3

模型训练: Longcite



训练数据

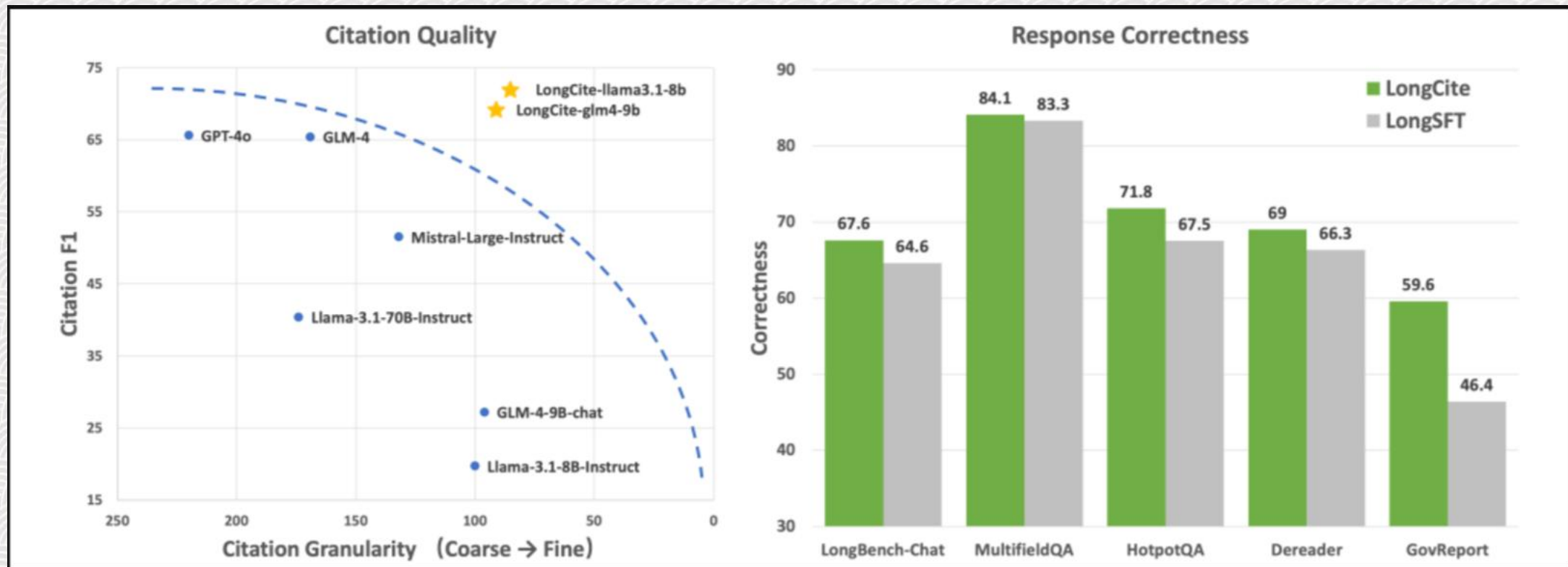
训练数据集来源与longcite-45k和shareGPT:


longcite-45k:从 GLM-4 的预训练语料库中收集了 50k篇长度 再128-256k之间的文档，应用 CoF，以 GLM-4 为骨干 LLM，以 Zhipu Embedding-v2 为检索器为每篇文档生成带有句子级引文的 QA 对，经过过滤阶段后得到 44,600 个高质量 LQAC 实例。

shareGPT: 76k通用的SFT实例,保证模型的通用能力。


同时使用 LongCite-45k 中的纯长语境 QA 对(不含任务指令和引文)训练了两个基础模型，并将训练后的模型命名为LongSFT

结果对比





结果对比



Query: Are both Duke Energy and Affiliated Managers Group based in Massachusetts?

LongSFT-9B:

Based on the information provided, **Duke Energy has offices in several locations including West Palm Beach, Florida; Prides Crossing, Massachusetts; Stamford, Connecticut; and London, United Kingdom.** Affiliated Managers Group has principal offices in West Palm Beach, Florida; Prides Crossing, Massachusetts; Stamford, Connecticut; and London, United Kingdom. So they both have an office in Massachusetts.

LongCite-9B:


<statement>Based on the information provided, **Duke Energy is headquartered in Charlotte, North Carolina,** while Affiliated Managers Group has principal offices in West Palm Beach, Florida; Prides Crossing, Massachusetts; Stamford, Connecticut; and London, United Kingdom.</statement> <statement>So only Affiliated Managers Group is based in Massachusetts.</statement>

Citations:

[25-25]: Duke Energy Corporation is an American electric power and natural gas holding company headquartered in Charlotte, North Carolina.

[219-219]: (NYSE: AMG) is a strategic partner to independent investment management firms globally, with equity stakes in a number of partner-owned traditional investment managers, hedge funds, and specialized private equity firms which it calls “Affiliates.” The company has principal offices in West Palm Beach, Florida; Prides Crossing, Massachusetts; Stamford, Connecticut; and London, United Kingdom.

Table 9: Case study. LongSFT-9B hallucinates the office location of Duke Energy, directly copying that of Affiliated Managers Group, while LongCite-9B gets the correct answer due to SFT with citations. We mark the wrong and correct statements in red and green, respectively.



结果对比

Query: Please write a one-page summary of the above government report.

LongSFT-9B:

The Department of Homeland Security (DHS) has established policies and processes for managing its multi-billion dollar acquisition investments. These policies and processes are primarily set forth in its Acquisition Management Directive 102-01 and Acquisition Management Instruction 102-01-001. The department uses these policies and processes to deliver systems that are intended to close critical capability gaps, helping enable DHS to execute its missions and achieve its goals.

The DHS acquisition management policy establishes that a major acquisition program's decision authority shall review the program at a series of predetermined acquisition decision events to assess whether the major program is ready to proceed through the acquisition life-cycle phases. Depending on the program, these events can occur within months of each other, or be spread over several years.

The report found that 10 of the 24 programs assessed were on track to meet their schedule and cost goals, while 14 programs were not on track. The report made two recommendations to DHS to improve its acquisition management policies and processes.

LongCite-9B:

<statement>The report assesses the Department of Homeland Security's (DHS) major acquisition programs and their progress towards meeting schedule and cost goals. It also evaluates DHS's policies and processes for acquisition management, resource allocation, and requirements development to determine if they reflect key portfolio management practices.</statement>

<statement>Key findings include:</statement>

<statement>- Of the 28 programs reviewed, only 10 were on track to meet their schedule and cost goals as of January 2018. The remaining 14 programs had slipped on schedule, cost, or both.</statement>

<statement>- The most common reasons for schedule slips were new requirements, technical challenges, and external factors like staffing shortages or funding cuts.</statement>

<statement>- Some programs had cost breaches due to scope increases, technical challenges, or contractor labor costs.</statement>

<statement>- DHS has taken steps to improve its policies and processes, but more is needed. For example, it should require components to submit affordability information when programs re-baseline due to breaches.</statement>


<statement>- The report makes two recommendations for DHS to improve its acquisition management: require components to submit affordability information when programs re-baseline and assess the results of post-implementation reviews to identify opportunities to improve performance across the portfolio.</statement>

<statement>- DHS concurred with the recommendations and has identified actions it will take to address them.</statement>



总结

综上所述，该论文的主要贡献体现在以下三个方面：

- 提供了针对带引文的长上下文问题解答任务的评测方法——Longcite-bench。
 - 提出了构建细粒度引文数据的数据构建流程，命名为COF。
 - 公开了一个专注于带引文的长上下文问答领域的数据集——Longcite-45k。
- 

挚终



谢谢观看

Thanks for watching

汇报人：吴晨帆

时间：2024年11月18日

