LongMagpie: A Self-synthesis Method for Generating Large-scale Long-context Instructions

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Chaochen Gao

Motivation

- Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of tasks, with recent advancements significantly extending their context lengths.
- However, *fine-tuning LLMs to leverage long contexts requires access to high-quality long-context instruction data*. Existing methods for creating open-source instruction data face substantial limitations when extended to long contexts.
  - (1) *Human labor costs are prohibitively high for creating diverse*, high-quality long-context instruction data. The annotation difficulty is substantially greater than for short-context data, requiring individuals to read documents spanning thousands of tokens before formulating instructions—a demonstrably challenging task.
  - (2) Existing synthetic approaches, often relying on predefined templates or seed questions , *do not guarantee the diversity needed for effective long*- context instruction. While existing projects attempt to broaden seed data diversity, creating large-scale long-context instructions with high quality and diversity remains an expensive and time-consuming process.

Motivation

Key Insight: Auto-Regressive Document-Query Generation

- The foundation of LongMagpie is a key observation about aligned long-context LLMs: *when provided with a document followed by tokens that typically precede a user query (without the query itself), these models generate contextually relevant queries about that document.*

- Formally, for an aligned LLM $\mathcal{M}$ with vocabulary $\mathcal{V}$, we define the document-query generation process as follows: given a document $D = \{d_1, d_2, ..., d_n\} \in \mathcal{V}^n$ and pre-query template $T_{pre} = \{t_1, t_2, ..., t_m\} \in \mathcal{V}^m$ (containing tokens indicating a user or query role, e.g., <|im_start|>user), we provide input $X = D \oplus T_{pre}$, where $\oplus$ denotes sequence concatenation. The model then generates a sequence $Q = \{q_1, q_2, ..., q_k\} \in \mathcal{V}^k$ representing a query related to document D. This process can be described as:

$$p_{\mathcal{M}}(Q \mid D, T_{\text{pre}}) = \prod_{i=1}^{k} p_{\mathcal{M}}(q_i \mid D, T_{\text{pre}}, q_{<i}) \,, \tag{1}$$

Method

Document Preparation

We collect diverse long documents from multiple domains to create a rich dataset for long-context modeling.

Query Generation

We generate contextually relevant user queries for each document by prompting an aligned LLM with document text and instruction templates.

Response Generation

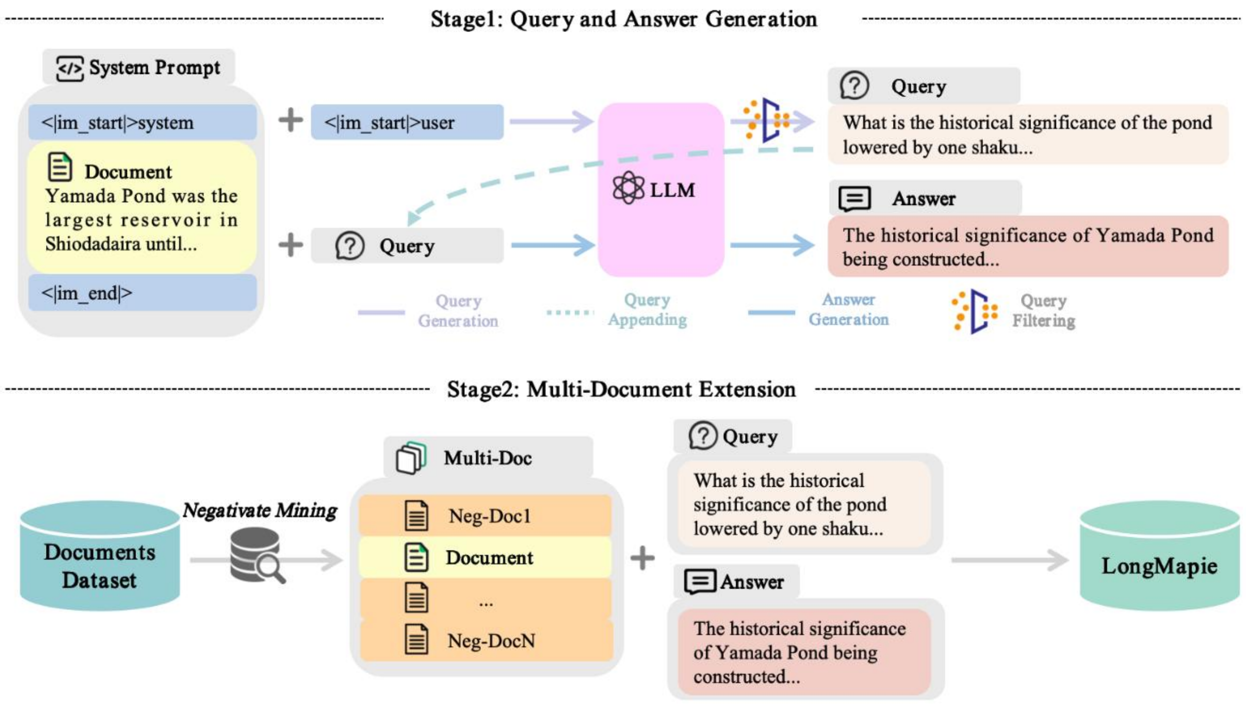
We produce assistant responses for each document-query pair.

Query Filtering

We filter out invalid queries using rule- and length-based heuristics to ensure quality and relevance.

Multi-Document Extension

We extend LongMagpie to multi-document settings by combining multiple documents into a single input to enable cross-document reasoning.



Result

□LongMagpie demonstrates better performance on average.

- ✓ As shown in Table 1, models trained solely on LongMagpie data already set a leading performance on long-context evaluation, topping HELMET (62.10), RULER (91.17), LongBench-v2 (34.4) and the LongAVG score (62.56) within the Long Instruction Data group

Table 1: Main experimental results comparing LongMagpie with other methods on long-context and short-context benchmarks. Best scores in each column are bolded. LongAVG is the average of HELMET, RULER, and Longbench v2, ShortAVG is the average of different short-context tasks.

Dataset	Long Evaluation				Short Evaluation
	HELMET	RULER	Longbench v2	LongAVG	ShortAVG
Short Instruction Data					
Tulu	61.93	87.92	28.4	59.42	63.90
Magpie	60.18	87.06	31.4	59.55	63.32
UltraChat	60.55	83.85	30.4	58.27	**64.43**
Long Instruction Data					
ChatQA	60.23	89.82	30.8	60.28	63.58
LongAlign	57.79	86.08	24.5	56.12	60.97
LongMagpie	**62.10**	**91.17**	**34.4**	**62.56**	62.37
***p*-Mix: Long + Short Instruction Data**					
ChatQA + UltraChat	60.80	87.42	31.4	59.87	64.38
LongAlign + UltraChat	60.98	89.49	30.6	60.36	64.17
LongMagpie + UltraChat	**62.11**	**89.70**	**33**	**61.60**	64.10

Result

□Impact of Different Multi-Document Settings.

- ✓ We observe that the multi-document strategy significantly improves performance on long-context tasks (from 60.19 to 62.56). As the value of n increases, the performance on long-context tasks improves and degrades, with the best performance observed when n = 10.
- ✓ *We hypothesize that this trend is due to an excessive number of documents increasing the task difficulty beyond the model's learning capacity, thereby leading to a drop in performance.*

n	HELMET	RULER	Longbench v2	LongAVG	ShortAVG
0	60.13	89.04	31.4	60.19	**63.20**
5	61.42	89.91	31.4	60.91	61.98
10	**62.10**	**91.17**	**34.4**	**62.56**	62.37
20	61.75	91.08	32.8	61.88	62.04
40	62.08	90.77	31.0	61.28	62.37
80	61.15	90.65	31.0	60.93	62.13

Result

□Impact of Different Data Size and Different Source Model Size.

- ✓ Table 4 demonstrates that *increasing the volume of high-quality long-context instruction data significantly enhances the model's ability*.
- ✓ This superior performance stems from larger models' enhanced ability to model long-context capabilities, which translates to better results when applied to the LongMagpie method.

Table 4: Increasing the volume of training data improves performance on long-context benchmarks.

Source Model	Data Volume	HELMET	RULER	Longbench v2	LongAVG	ShortAVG
Qwen-2.5-70B	190k	61.29	90.65	32.6	61.51	62.30
Qwen-2.5-70B	450k	**62.10**	**91.17**	**34.4**	**62.56**	**62.37**

Table 5: Using the larger source model improves performance on long-context benchmarks..

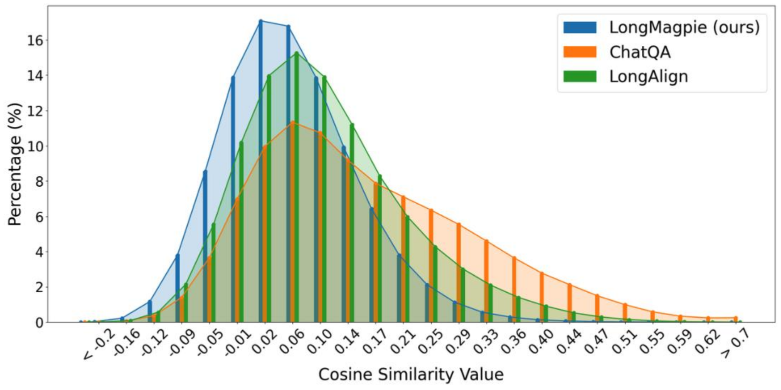
Source Model	Data Volume	HELMET	RULER	Longbench v2	LongAVG	ShortAVG
Qwen-2.5-7B	450k	59.28	86.95	32.6	59.61	62.18
Qwen-2.5-70B	450k	**62.10**	**91.17**	**34.4**	**62.56**	**62.37**

Result

◻Analysis of of LongMagpie Queries.

- ✓ *Higher Quality of LongMagpie Queries*: The overall data quality of LongMagpie is significantly higher than previous methods.
- ✓ *Better Diversity of LongMagpie Queries*: LongMagpie queries generally exhibit lower similarity among themselves, which also reflects their good diversity.

(a) Reward model scores for different datasets.

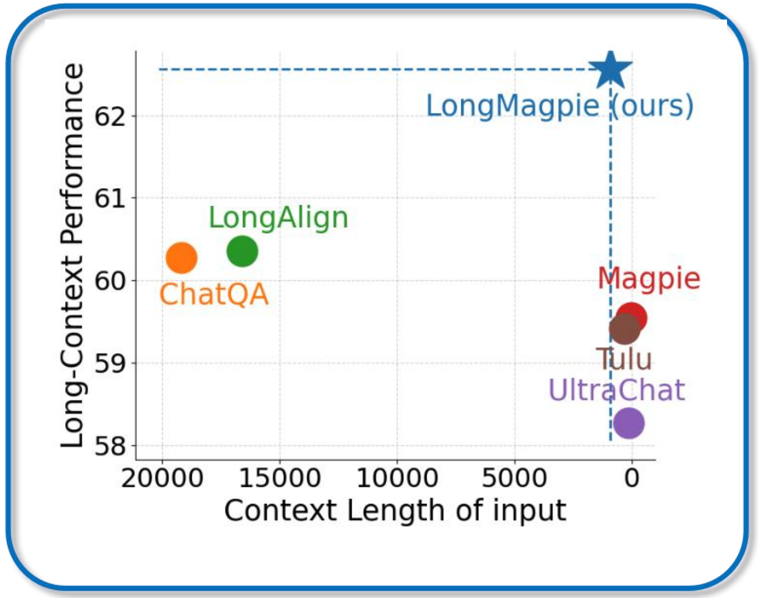


(b) Query similarities within different datasets.

Result

◻Sample Efficiency of LongMagpie.

- ✓ This efficiency stands in stark contrast to existing methods, which consume 10-13× more tokens per instruction during synthesis yet produce inferior performance outcomes. *LongMagpie's remarkable sample efficiency facilitates greater scalability and diversity.*



中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Thanks!