

# Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Michael Bendersky











#### RAG v.s. LC?

Understanding long-context is an important capability for LLMs.

On one side, LLMs are supporting longer and longer contexts, e.g. Gemini supports up to 2M tokens

On the other side, RAG (retrieval augmented generation) has been an effective & efficient way to process long contexts.

Which one should we use?

## Summary of contributions

- Comparison finding
  - LC > RAG when resourced sufficiently
- A simple-yet-effective method
  - Self-Route: try RAG, then LC
- Analysis
  - RAG failure reasons
  - Scaling
  - Synthetic data
  - LLM's internal knowledge
  - 0 ...

#### Compare RAG and LC

Published as a conference paper at ICLR 2024

# RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

Peng Xu<sup>†</sup>, Wei Ping<sup>†</sup>, Xianchao Wu, Lawrence McAfee Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina Mohammad Shoeybi, Bryan Catanzaro

**NVIDIA** 

Can Long-Context Language Models Subsume
Retrieval, RAG, SQL, and More?

Jinhyuk Lee\* Anthony Chen\* Zhuyun Dai\*
Dheeru Dua Devendra Singh Sachan Michael Boratko Yi Luan
Sébastien M. R. Arnold Vincent Perot Siddharth Dalmia Hexiang Hu
Xudong Lin Panupong Pasupat Aida Amini Jeremy R. Cole
Sebastian Riedel Iftekhar Naim Ming-Wei Chang Kelvin Guu
Google DeepMind

LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding

Yushi Bai<sup>12</sup>, Xin Lv<sup>2</sup>, Jiajie Zhang<sup>12</sup>, Hongchang Lyu<sup>3</sup>, Jiankai Tang<sup>1</sup>,
Zhidian Huang<sup>1</sup>, Zhengxiao Du<sup>12</sup>, Xiao Liu<sup>12</sup>, Aohan Zeng<sup>12</sup>,
Lei Hou<sup>1</sup>, Yuxiao Dong<sup>1†</sup>, Jie Tang<sup>1</sup>, Juanzi Li<sup>1†</sup>

<sup>1</sup>Tsinghua University <sup>2</sup>Zhipu.AI <sup>3</sup>Institute of Automation, Chinese Academy of Sciences

Google Cloud

# Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG

Bowen Jin<sup>1</sup> <sup>2</sup> \*, Jinsung Yoon<sup>1</sup>, Jiawei Han<sup>2</sup> and Sercan Ö. Arık<sup>1</sup> <sup>1</sup>Google Cloud AI Research, <sup>2</sup>University of Illinois at Urbana-Champaign

<sup>†{</sup>pengx, wping}@nvidia.com

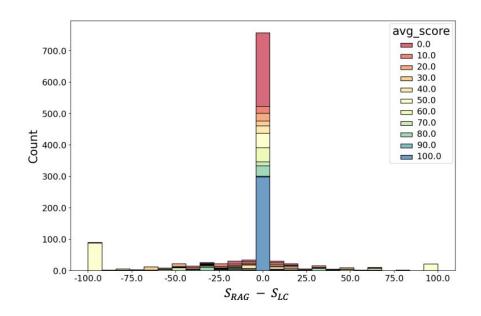
## Comparison: LC performs better than RAG

- 3 models:
  - Gemini-1.5-Pro (1M)
  - o GPT-40 (128k)
  - o GPT-3.5-Turbo (16k)
- 9 Datasets
  - o 7 datasets from LongBench
    - NarrativeQA,
    - Qasper,
    - MultiFieldQA,
    - HotpotQA,
    - 2WikiMultihopQA,
    - MuSiQue,
    - QMSum
  - 2 datasets from InfiniteBench
    - EN.QA,
    - EN.MC
  - Real dataset in English; Can do RAG



### RAG and LC agree for most queries

- For most queries, RAG and LC predictions are very similar
  - For 63% queries, RAG/LC predictions are exactly identical
  - For 70% queries, RAG/LC predictions are very similar (score difference < 10)</li>



#### Our method: Self-Route

#### Two steps:

- 1. RAG: prompt LLM to predict whether the query is answerable; if so, output the answer
- 2. If the query is not answerable using RAG, do LC

Use LLM itself as a router, to route different queries to RAG or LC.

### Self-Route performance is similar to LC, with reduced cost



Gemini-1.5-Pro	1-3	SELF-ROUTE	46.41	28.32	45.23	51.47	55.18	62.68	40.66
	1-4	answerable %	76.78	73.00	85.00	96.67	84.50	81.00	58.50
	1-5	token %	38.39	23.07	49.93	36.88	32.97	53.49	56.14
GPT-4O	2-1	LC	48.67	32.78	44.54	55.28	62.42	70.69	41.65
	2-2	RAG	32.60	18.05	46.02	50.74	36.86	50.21	16.09
	2-3	SELF-ROUTE	48.89	31.36	47.99	53.17	62.14	70.14	41.69
	2-4	answerable %	57.36	44.00	67.50	94.00	52.50	62.00	30.00
	2-5	token %	61.40	66.40	72.25	39.65	65.79	77.05	85.00
GPT-3.5-Turbo	3-1	LC	32.07	23.34	42.96	49.19	45.33	41.04	17.92
	3-2	RAG	30.33	18.22	38.15	49.21	37.84	35.16	16.41
	3-3	SELF-ROUTE	35.32	24.06	38.65	52.07	47.28	44.62	34.44
	3-4	answerable %	74.10	71.50	80.00	91.33	68.50	69.00	47.00
	3-5	token %	38.85	20.56	55.08	35.29	48.70	65.91	65.08

Caveat: "answerable rate" (different models show different answerable rates, due to different alignments).

Narr

32.76

22.54

Qasp

47.83

44.68

Mult

52.33

49.53

Hotp

61.85

48.36

Avg

49.70

37.33

LC

**RAG** 

Particularly good for GPT-3.5 (suspectively due to shorter context length)

1-1

1-2

"token%": cost

2Wiki

62.96

54.24

Musi

40.22

26.56

Sum

20.73

19.51

19.77

93.50

17.96

21.92

19.97

21.31

92.00

20.26

19.61

18.94

19.88

93.50

16.40

En.QA

43.08

19.46

37.51

56.41

42.25

32.36

14.43

34.95

27.07

73.01

14.73

15.39

22.03

50.43

38.17

En.MC

85.57

51.09

76.86

62.45

32.84

76.42

41.05

77.29

47.16

53.21

34.50

43.67

44.54

95.63

4.50

#### Ablations of k

As k increases, RAG can answer more and more queries.

Self-Route consistently performs better than RAG.

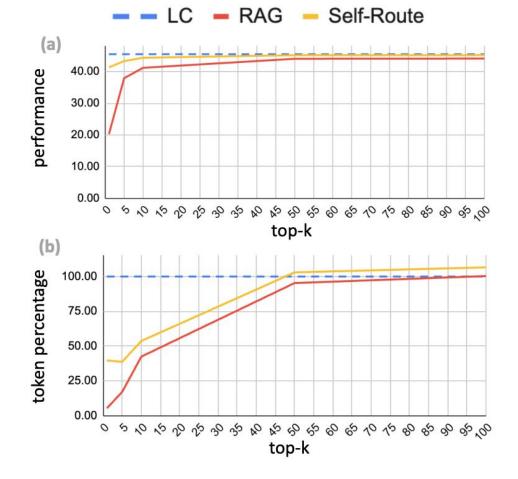


Figure 3: Trade-off curves between (a) model performance and (b) token percentage as a function of k.

#### Why does RAG fail?

Manually check the RAG failures where LC succeeds. 4 typical reasons:

- A. The query requires **multi-step reasoning** so the results of previous steps are needed to retrieve information for later steps,
  - a. What nationality is the performer of song You Can?
  - b. Where does the director of film Wine Of Morning work at?
  - c. What is another notable work made by the author of Miss Sara Sampson?
- B. The query is **general**, thus it is hard to retrieve relevant chunks.
  - a. What does the group think about XXX"
- C. The query is **long and complex**, which is challenging for the retriever to understand.
  - a. What did Julie Morgan elaborate on the online survey when talking about the evaluations on the legitimacy of the children's rights, protection and demands?
  - b. The Huskies football team were invited to the Alamo Bowl where they were defeated by a team coached by Art Briles and who played their home games at what stadium?
- D. The query is **implicit**, demanding a thorough understanding of the entire context o connect the dots and deduce the answer
  - a. "What caused the shadow behind the spaceship?" for a long story without explicit mention of the shadow when the cause is revealed.

#### Failure case distribution

Prompt Gemini to classify the failure cases:

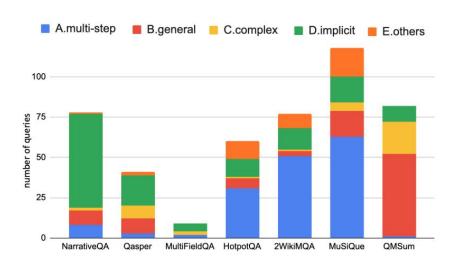


Figure 4: Distribution of typical RAG failure reasons.

Different distributions for each dataset.

## Why not synthetic dataset? A toy experiment

Needle-in-the-haystack style synthetic dataset may contain artifacts that affect the analysis results

"Passkey" dataset: a sentence with passkey (e.g. "the passkey is 123456") is hidden in large chunks of texts.

"What is the passkey?"	r	RAG	LC
"What is the special token	Original	80.34	65.25
hidden inside the texts?"	Variant-1: "special token"		69.32
"Which passkey is larger? First or Second?" (for 2 passkeys)	Variant-2: "which is larger"	47.63	64.24

### Exclusion of LLM's internal knowledge

Dataset leakage; LLM's world knowledge.

Many of the datasets are based on Wikipedia (HotpotQA, 2WikiMultihopQA, MuSiQue); books (NarrativeQA); or scientific papers (Qasper, MultiFieldQA)

LLM's internal knowledge should be excluded.

### How to exclude the internal knowledge?

#### Method-1:

use a simple prompt: "based only on the provided passage".

	without "based only on"	with "based only on"	
NarrativeQA	36.35	32.76	
Qasper	50.69	47.83	
MultiFieldQA	56.07	52.33	
HotpotQA	66.47	61.85	
2WikiMQA	68.97	62.96	
Musique	54.56	40.22	
QMSum	20.87	20.73	
En.QA	49.20	43.08	
En.MC	90.83	85.57	
Avg	50.57	45.53	

#### Method-2

Exclude questions that can be correctly answered without context.

Cons: different models will have different eval

	all qu	estions	w/o commonsense		
	Gemini	GPT-3.5	Gemini	GPT-3.5	
# questions	200	200	133	150	
LC	40.22	17.92	31.76	13.00	
RAG	26.56	16.41	15.51	13.05	
Self-Route	40.66	34.44	31.32	19.76	
answerable %	58.50	47.00	52.63	45.33	
token %	56.14	65.08	48.46	53.43	
			*		

# **Thanks! Questions?**