Google **Retrieval Augmented Generation or Long-Context LLMs?** DeepMind **A Comprehensive Study and Hybrid Approach**



UNIVERSITY OF Mingyang Zhang¹ Qiaozhu Mei² Zhuowan Li¹ Michael Bendersky¹ Cheng Li¹ **MICHIGAN** ¹Google DeepMind ²University of Michigan

Summary of Contributions	A simple-yet-effective method: Self-Route
 A comprehensive comparison of RAG and LC A simple-yet-effective method: Self-Route Detailed analysis 	 Step-1 (RAG): "Answer the question based on the retrieved passages. If the question cannot be answered, output unanswerable" Step-2 (LC): if "unanswerable" in step-1, do LC.

Main Results

(a)			C RAG	Self-Route						Avg	Narr	Qasp	Mult	Hotp	2Wiki	Musi	Sum	En.QA	En.MC
	50 —		_		40.70			1-1	LC	49.70	32.76	47.83	52.33	61.85	62.96	40.22	20.73	43.08	85.57
		48.67	48.89		49.70	4	46.41	1-2	RAG	37.33	22.54	44.68	49.53	48.36	54.24	26.56	19.51	19.46	51.09
ince	40 —					27.22		1-3	Self-Route	46.41	28.32	45.23	51.47	55.18	62.68	40.66	19.77	37.51	76.86



Comparison: LC performs consistently better than RAG, but with a higher cost. Self-Route achieves comparable performance with LC with much lower cost.



- \succ For 63% queries, RAG/LC predictions are exactly identical. For 70% queries, RAG/LC predictions are very similar (score difference < 10).

- "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?"
- "What caused the shadow behind the spaceship?" for a long story without explicit D. mention of the shadow when the cause is revealed.



 \succ For 63% queries, RAG/LC predictions are exactly identical. For 70% queries, RAG/LC predictions are very similar (score difference < 10).

	without "based only on"	with "based only on"				
Avg	50.57	45.53				

	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 % token %	58.50 56.14	47.00 65.08	52.63 48.46	45.33 53.43			