SEALQA: RAISING THE BAR FOR REASONING IN SEARCH-AUGMENTED LANGUAGE MODELS

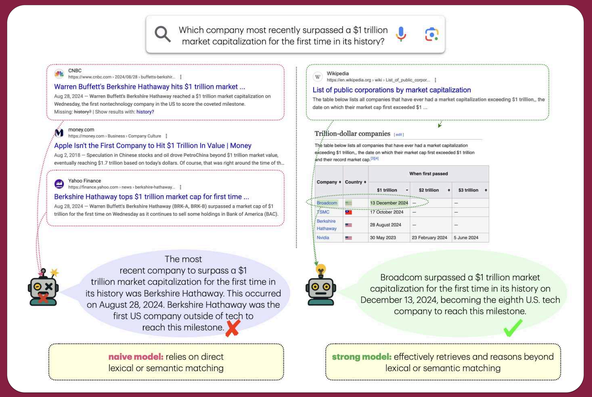
Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, Tu Vu

Our benchmark dataset has been used by Google's Gemini, DeepSeek, and Kimi

Paper

VIRGINIA TECH ICLR

A challenging benchmark for evaluating search-augmented LLMs that requires deep reasoning to handle noisy web searches.

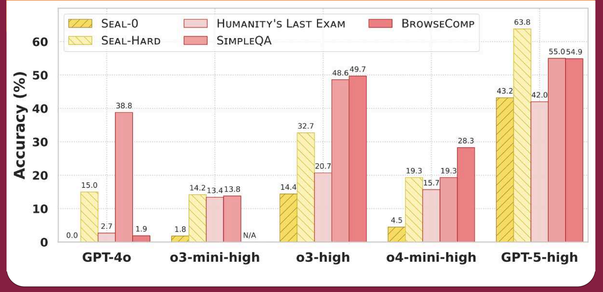


SEALQA comes in three flavors:

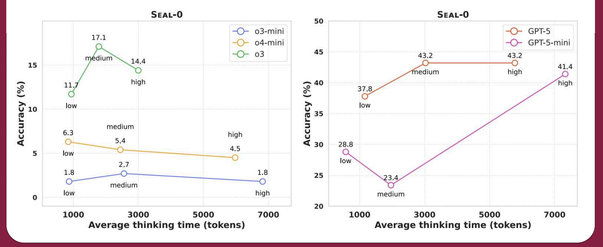
- SEAL-0 (main; 111 questions): A carefully curated core set where frontier non-reasoning models consistently fail.
- SEAL-HARD (254 questions): A broader set of SEAL-0 along with additional challenging questions.
- LongSeal (254 questions): A "needle-in-a-haystack" variant to test long-context multi-document reasoning.

Search Language Models | Challenging Benchmarks

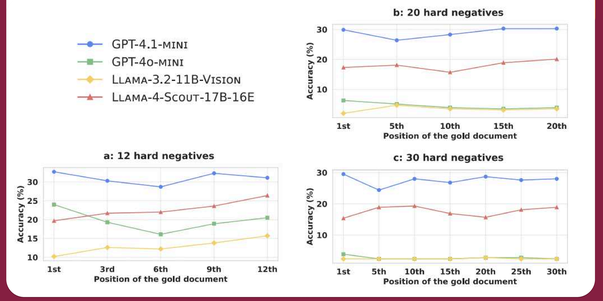
SEALQA poses significant challenges to frontier models



Test-time scaling does not lead to reliable gains



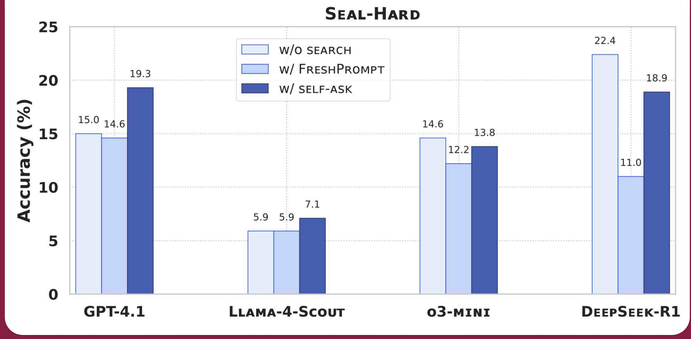
Frontier LLMS fail to reliably identify relevant documents in LONGSEAL when numerous distractors are present



Accuracy on SEAL-0 and SEAL-HARD. Frontier LLMs face significant challenges on the SEAL-0 and SEAL-HARD subsets, as SEALQA demands careful search and robust reasoning .

Model	knowl. cutoff	type	SEAL-0		SEAL-HARD	
			w/o search	w/ search	w/o search	w/ search
Closed-source models						
GPT-4o-mini	Sep 30, 2023	Chat	0.0	0.0†	9.1	13.4†
GPT-4.1-mini	May 31, 2024	Chat	0.0	0.0†	13.8	11.8†
GPT-4o	Sep 30, 2023	Chat	0.0	0.0†	11.8	15.0†
GPT-4.1	May 31, 2024	Chat	0.0	0.0†	15.0	20.5†
o3-mini-high	Sep 30, 2023	Reason.	3.6	1.8	12.6	14.2
o4-mini-high	May 31, 2024	Agentic	-	4.5 †	-	19.3†
o3-high	May 31, 2024	Agentic	-	14.4†	-	32.7†
GPT-5-mini-high *	May 31, 2024	Reason.	6.3	41.4†	16.9	60.2†
GPT-5-high *	Sep 30, 2024	Reason.	**15.3**	**43.2**†	**37.8**	**63.8**†
Open-weight models						
Llama-3.2-3B	December 1, 2023	Chat	0.0	0.0	1.6	3.5
Llama-3.1-70B	December 2023	Chat	0.0	0.0	0.0	6.3
Llama-4-Scout-17B-16E (109B)	August 2024	Chat	0.0	0.0	5.9	5.9
Qwen3-235B-A22B	-	Reason.	0.0	5.4	4.3	11.4
DeepSeek-R1-Distill-Qwen-1.5B	-	Reason.	0.0	2.7	0.0	1.6
DeepSeek-R1-Distill-Qwen-14B	-	Reason.	0.9	3.6	0.9	10.6
DeepSeek-R1-671B	-	Reason.	**5.4**	1.8	**22.4**	11.0
gpt-oss-20b-high *	June, 2024	Reason.	0.9	4.5	2.7	7.8
gpt-oss-120b-high *	June, 2024	Reason.	0.9	**7.2**	10.6	**16.9**

Advanced reasoning models (DeepSeek, o3-mini) are highly vulnerable to noisy search results



Breakdown of SEAL-HARD results by question category

	Model	Q_1	Q_2	Q_3	Q_4	Q_5
w/o search	GPT-4.1	14.1	14.2	25.7	0.0	0.0
	o3-mini-high	10.9	14.9	14.3	0.0	0.0
	o3-high	–	–	–	–	–
	Llama-4-Scout	4.9	6.8	5.7	0.0	0.0
	DeepSeek-R1	20.7	23.0	22.9	7.1	0.0
w/ search	GPT-4.1	20.1†	17.6†	25.7†	21.4†	9.1†
	o3-mini-high	9.8	10.1	22.9	7.1	9.1
	o3-high	31.0†	31.8†	45.7†	14.3†	27.3†
	Llama-4-Scout	4.3	6.8	8.6	0.0	0.0
	DeepSeek-R1	10.3	10.8	14.3	0.0	18.2

Model	w/o search			w/ search		
	never	slow	fast	never	slow	fast
GPT-4.1	21.5	18.0	1.6	17.7†	24.3†	17.2†
o3-mini-high	20.3	12.6	3.1	12.7	10.8	10.9
o3-high	–	–	–	39.2†	36.9†	17.2†
Llama-4-Scout	10.1	4.5	4.1	6.3	4.5	7.8
DeepSeek-R1	32.9	24.3	6.2	15.2	9.9	7.8

Model	w/o search			w/ search		
	< 2024	2024	2025	< 2024	2024	2025
GPT-4.1	23.5	6.1	0.0	25.5†	20.4†	7.1†
o3-mini-high	20.5	2.6	1.4	14.4	12.8	4.3
o3-high	–	–	–	45.9†	15.4†	14.5†
Llama-4-Scout	8.7	4.1	0.0	7.4	6.1	1.8
DeepSeek-R1	35.6	8.2	0.0	14.8	6.1	5.4

Model	w/o search		w/ search	
	unhelpful	conflict.	unhelpful	conflict.
GPT-4.1	14.5	15.3	18.2†	22.2†
o3-mini-high	10.9	13.9	8.2	13.9
o3-high	–	–	30.0†	34.7†
Llama-4-Scout	3.6	7.6	4.5	6.9
DeepSeek-R1	20.9	23.6	9.1	12.5

Frontier models fail to extract correct answers even when no distractors are provided.

Models	w/ evidence	w/o search	w/ search
GPT-4.1	48.0	15.0	20.5
o3-mini	56.7	14.6	12.2
Llama-4-Scout	33.5	5.9	5.9
DeepSeek-R1	49.2	19.3	15.4