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

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **SEALQA**, a challenge benchmark for evaluating **SE**arch-**A**ugmented **L**anguage models on fact-seeking questions where web search yields conflicting, noisy, or unhelpful results. SEALQA comes in *three* flavors: (1) **SEAL-0** (*main*) and (2) **SEAL-HARD**, both of which assess factual accuracy and reasoning capabilities, where SEAL-0 targets the most challenging questions that frontier non-reasoning models (e.g., GPT-4.1) answer with near-zero accuracy; and (3) **LONGSEAL**, which extends SEALQA to test long-context, multi-document reasoning in *"needle-in-a-haystack"* settings. Our evaluation reveals critical limitations in current models. Even frontier reasoning models face significant challenges across SEALQA flavors. On SEAL-0, GPT-5 with tools achieves only 43.2% accuracy at its best reasoning effort. We also find that even advanced reasoning models (e.g., DEEPSEEK-R1) can be vulnerable to noisy search results. *Notably*, increasing test-time compute does not yield reliable gains across GPT-5 and the O-series of models, with performance often plateauing or even declining early. Finally, while current models are less affected by the *"lost-in-the-middle"* issue, they still fail to reliably identify relevant documents in LONGSEAL when faced with numerous distractors. To facilitate future work, we release SEALQA at anonymous.4open.science/r/SealQA.

1 INTRODUCTION

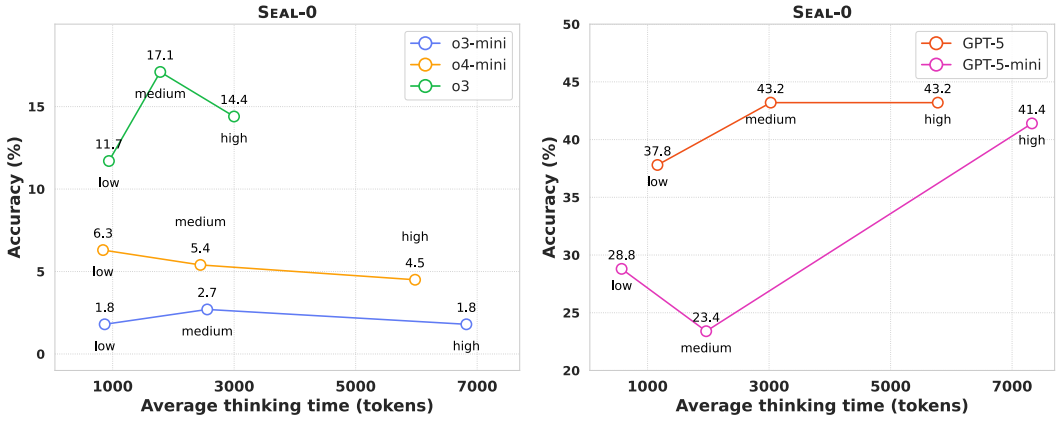

Figure 1: Frontier model performance before (left) and after (right) the release of SEAL-0. Despite potential data contamination or direct access by agentic models, SEAL-0 continues to pose a significant challenge for current frontier LLMs. Test-time scaling does not lead to reliable gains, with performance often plateauing or even declining early. See Figure 6 and Table 14 for additional results.

Large language models (LLMs) have entered a new scaling paradigm: *test-time scaling*, where models dynamically allocate more compute during inference time to improve performance (OpenAI, 2025a; Google, 2025; xAI, 2025; Anthropic, 2025). This paradigm shift is embodied in *reasoning models*, which leverage reinforcement learning and other techniques to guide inference-time strategies such as chain-of-thought reasoning, recursive refinement, and real-time search (Muennighoff et al., 2025; Guo et al., 2025; Snell et al., 2024; Geiping et al., 2025). These models can now decompose questions into subqueries, decide when and how to query a search engine, and fuse retrieved content into structured reasoning paths (OpenAI, 2025a; Google, 2025; Jin et al., 2025).

As LLMs advance, benchmarks that rely on static knowledge and simple reasoning become saturated and fail to keep pace. For example, frontier models now achieve over 90% accuracy on MMLU (Phan et al., 2025). Furthermore, most evaluations of search-augmented LLMs focus on short factual queries that top-ranked results answer directly (Vu et al., 2024; Kasai et al., 2023). These setups require only shallow comprehension and fail to reflect the messy, ambiguous nature of real-world search.

To properly evaluate today's LLMs, benchmarks that go beyond simple fact lookup are needed. Real-world search often returns documents that are outdated, misleading, or superficially relevant but ultimately unhelpful. Navigating this noise requires deeper reasoning that filters inconsistencies, reconciles contradictions, and identifies trustworthy signals. Benchmarks that simulate these challenges are rare, partly because they are difficult to curate and validate at scale.

We introduce SEALQA, a *small* but *extremely challenging* benchmark (see Figure 2) for evaluating search-augmented LLMs on fact-seeking questions. Each SEALQA question is carefully crafted by NLP researchers to trigger ambiguous, conflicting, or noisy search results (see Figure 3). This design makes it difficult to answer questions through simple keyword matching or by relying on top-ranked documents. SEALQA spans a range of question types, including time-sensitive questions, across diverse domains such as science, technology, sports, entertainment, politics, history, geography, etc.

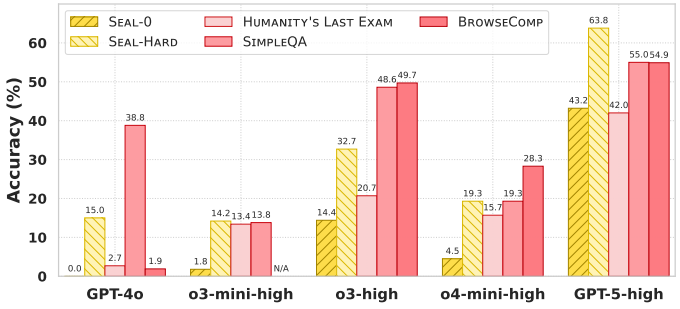

Figure 2: Accuracy of LLMs across benchmarks. SEALQA poses significant challenges to frontier models.

SEALQA questions probe a broad spectrum of complex reasoning skills. These include distinguishing between similar entities or events, tracking changes to the same entity over time, interpreting information embedded in search-result plots, charts, or tables, counting multiple items, reasoning over non-English content, and debunking false premises or common misconceptions (see Figure 7 in Appendix B for sample questions). All questions are self-contained, verifiable, and require intensive reasoning to resolve ambiguity, filter misinformation, or reconcile conflicting evidence. These capabilities are central to modern LLMs but are not adequately captured by existing benchmarks.

To ensure both difficulty and quality, each SEALQA question undergoes a *rigorous* multi-round vetting process: an initial phase with two or more graduate-level reviewers, followed by expert approval. SEALQA comes in *three flavors*:

- SEAL-0 (*main*; 111 questions): A carefully curated core set where frontier non-reasoning models like GPT-4.1 with browsing consistently fail. Each question is iteratively refined until multiple models fail across several attempts (0% accuracy, hence the "0" in the name).
- SEAL-HARD (254 questions): A broader set that includes SEAL-0 and additional difficult questions that did not meet our strict failure threshold but remain highly challenging.
- LONGSEAL (254 questions): A *"needle-in-a-haystack"* variant that tests long-context, multi-document reasoning. Each question is paired with a large set of retrieved documents, among which only one contains or implies the correct answer. This document is buried within irrelevant, noisy, or misleading content.

We intentionally kept SEALQA small due to the high cost and complexity of question development.[1] Building the full benchmark required a team of *six* NLP researchers working over *eight* months through multiple development cycles. A smaller benchmark also reduces API evaluation costs, allows more frequent updates, and aligns with recent emphasis on high-quality, targeted evaluations over large, noisy ones (Rein et al., 2024; Maia Polo et al., 2024).[2] SEALQA is also designed for stable evaluation with *low* run-to-run variance.[3]

[1] Each question required over an hour on average – roughly 45 minutes to draft, plus additional time for review and revision. Many initial ideas were discarded as they failed to meaningfully challenge frontier LLMs.

[2] For example, the widely used GPQA-DIAMOND (Rein et al., 2024), a compact set of 198 expert-vetted questions, demonstrates how a small, carefully curated dataset can effectively assess a model's reasoning ability.

[3] Our questions often lead multiple models to fail across repeated attempts.

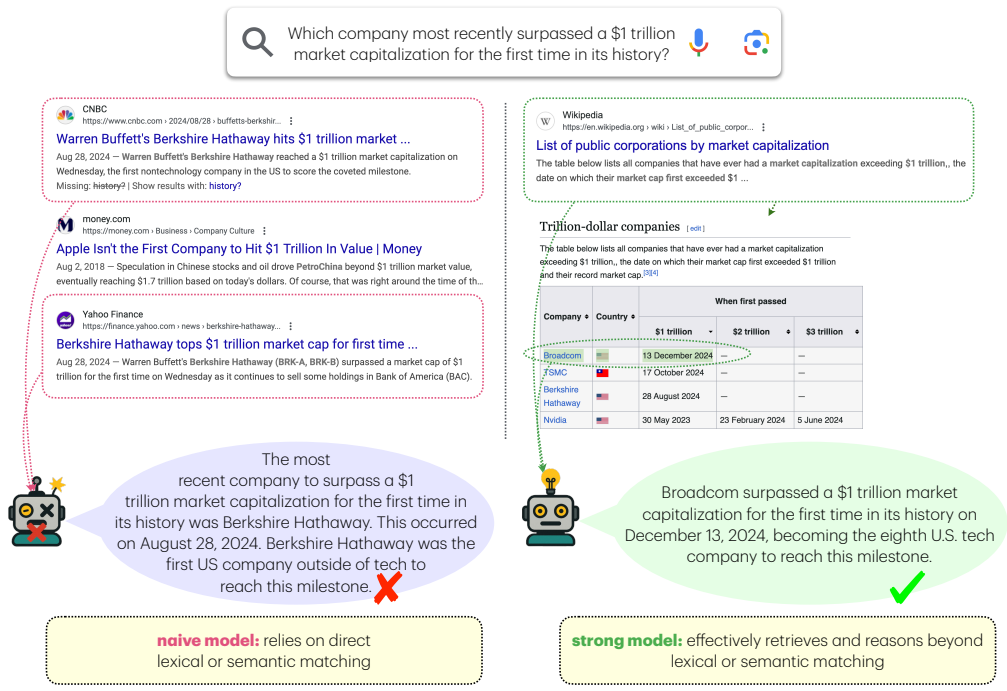

Figure 3: SEALQA requires intensive reasoning to resolve ambiguity, filter out misinformation, or reconcile conflicting evidence. See Appendix H for sample model outputs.

Our key contributions are as follows: (1) We introduce SEALQA, a challenge benchmark designed to evaluate reasoning under noisy, conflicting, and ambiguous search results. SEALQA includes three flavors: SEAL-0, SEAL-HARD, and LONGSEAL, each targeting different challenges in search-augmented reasoning; (2) We benchmark a range of LLMs and uncover significant limitations in current retrieval-augmented approaches. Even *state-of-the-art* models struggle across SEALQA flavors when faced with conflicting or misleading context. On SEAL-0, performance remains low even for agentic models equipped with search tools. We also find that advanced reasoning models can be highly vulnerable to noisy search results. *Notably*, increasing test-time compute does not reliably improve performance across OPENAI's GPT-5 and O-series of models – performance often plateaus or declines. LONGSEAL further reveals major weaknesses in long-context reasoning: while current frontier LLMs are more robust to *"lost-in-the-middle"* effects (Liu et al., 2024), they still fail to reliably identify and prioritize relevant evidence amid distractors; and (3) We publicly release SEALQA as a dynamic, versioned benchmark, and commit to review and update its answers regularly to ensure that evaluations reflect the most recent knowledge.

2 DATA COLLECTION

In this section, we describe SEALQA, our benchmark designed to capture the complexity of real-world information-seeking. SEALQA rigorously evaluates a model's reasoning ability, robustness to noisy search results, and capacity to handle dynamic, real-world knowledge.

Human annotators: To build SEALQA, we recruited NLP researchers[4] as human annotators who were shown a diverse set of exemplars that illustrated the types of questions we sought to collect.

Question types: Our questions span several categories: ($\mathcal{Q}_1$) *advanced reasoning*, which covers multi-hop reasoning, interpreting search-result plots, charts, or tables, and performing counting or calculations; ($\mathcal{Q}_2$) *entity/event disambiguation*, which focuses on distinguishing between similar entities or events; ($\mathcal{Q}_3$) *temporal tracking*, which requires identifying and differentiating instances of entities over time; ($\mathcal{Q}_4$) *cross-lingual reasoning*, where the question is in English but answering it requires retrieving and reasoning over non-English sources; and ($\mathcal{Q}_5$) *false-premise questions*, which require debunking false assumptions.

[4]including the authors and their colleagues

Annotation criteria: Annotators were instructed to write questions with a *single*, *unambiguous* answer (e.g., specifying "on what date" rather than asking "when"). Each question must be supported by one or more webpages that justify the reference answer, which ensures *verifiability*. For questions that involve *fresh knowledge*, annotators were required to cite regularly updated sources to support future answer updates. We also classify questions by *freshness* (Vu et al., 2024): *never-changing* (NEVER; answers never change), *slow-changing* (SLOW; answers change over several years), and *fast-changing* (FAST; answers typically change within a year). All questions were designed to appear natural while still triggering ambiguous, conflicting, or misleading search results when entered into a search engine like GOOGLE. Each question has a predefined annotation that classifies its expected search results as CONFLICT. (mixed correct and misleading answers) or UNHELPFUL (no correct answers). Annotators also provided explanations for each answer, including any necessary clarification or subtle reasoning. Finally, each question was refined until it consistently caused multiple models to fail across repeated attempts.

Quality control: We employed a rigorous multi-round review process. Each question was first reviewed by two or more graduate-level annotators, followed by expert approval. We performed several rounds of data cleaning, including verification of supporting URLs, answer correctness, and question clarity. Questions whose answers change too frequently were excluded. For each question, we also annotated the effective year (when the answer last changed) and the expected next review date to support future maintenance.

Diversity: SEALQA questions vary in length, with an average of 31 tokens and a maximum of 69. SEALQA also spans diverse domains: science and technology (26.8%), sports (22.0%), entertainment (21.7%), politics (9.1%), history and geography (8.3%), and others (12.2%).[5] By question category, 72.4% involve advanced reasoning ($\mathcal{Q}_1$), 58.3% entity/event disambiguation ($\mathcal{Q}_2$), 13.7% temporal tracking ($\mathcal{Q}_3$), 5.5% cross-lingual reasoning ($\mathcal{Q}_4$), and 4.3% false-premise detection ($\mathcal{Q}_5$). By freshness, 31.1% are *never-changing*, 43.7% *slow-changing*, and 25.2% *fast-changing*. By effective year, 22.0% reference 2025 events, 19.3% 2024, and 58.7% prior to 2024.

Curation of SEALQA flavors: To curate SEAL-0, we tested each question against GPT-4O, GPT-4.1, their MINI variants (OpenAI, 2024a;b; 2025c), and LLAMA-4-SCOUT (Meta, 2025), both with and without browsing.[6] Only questions whose answers all models failed to produce across 10–15 attempts were retained. This follows current practices for constructing challenging benchmarks; for example, SIMPLEQA (Wei et al., 2024) was also adversarially collected against GPT-4 responses. SEAL-0 was then combined with other rejected-but-difficult questions to form SEAL-HARD.

For LONGSEAL, each SEAL-HARD question is paired with a set of retrieved documents: *one* helpful (gold) document from annotator-provided webpages, and up to 50 hard negatives that appear relevant but are unhelpful.[7] To ensure difficulty, we used GPT-4O MINI to filter out negatives whose content might allow the correct answer to be inferred. The gold document was randomly inserted among the negatives. LONGSEAL contains over 7.6K documents and serves as a testbed for long-context reasoning under noisy retrieval conditions.

Evaluation protocol: Models are evaluated using a GPT-4O MINI auto-rater adapted from Wei et al. (2024), which takes the question, predicted answer, and reference answer as input and labels responses as "correct", "incorrect", or "not attempted" (see Appendix C for the full prompt). The evaluation follows a relaxed protocol that checks whether the main answer is factually correct and consistent throughout the response.

Auto-rater reliability: To assess the auto-rater's reliability, two authors independently evaluated 100 answers. Disagreements were resolved through discussion, which produced a unified set of human ratings that agreed with the auto-rater 98% of the time.

[5] Following Wei et al. (2024), topic labels were assigned post-hoc using GPT-4O MINI.

[6] We applied FRESHPROMPT (Vu et al., 2024) to LLAMA-4-SCOUT.

[7] To collect hard negatives, we used GOOGLE to retrieve the top 10 webpages per question and extracted their main content using TRAFILATURA (Barbaresi, 2021). To add temporal diversity and potential conflicts, we retrieved 10 more pages restricted to pre-2023 content. We also used GPT-4O MINI to generate three semantically related queries per question and collected documents for each. Duplicates were removed, and documents whose length exceeded 10K tokens were excluded.

Table 1: Accuracy on SEAL-0 and SEAL-HARD. Frontier LLMs face significant challenges on SEALQA questions. † indicates results using CHATGPT's built-in search; all other search-based results use FRESHPROMPT (Vu et al., 2024). * indicates evaluation conducted after the release of SEALQA.

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**

3 EXPERIMENTS

Having established SEALQA, we now set out to evaluate how well today's LLMs reason over noisy search results when navigating dynamic, real-world knowledge. Our analysis reveals limitations in their ability to reconcile conflicting parametric (*internal*) and retrieved (*external*) knowledge.

3.1 EXPERIMENT SETUP

3.1.1 SEAL-0 AND SEAL-HARD

Baselines: We benchmarked a wide range of open-weight and proprietary models. These include chat-oriented models such as GPT-4O, GPT-4.1, their MINI variants (OpenAI, 2024a;b; 2025c), LLAMA-3.1-70B (Grattafiori et al., 2024), LLAMA-3.2-3B (Meta, 2024), and LLAMA-4-SCOUT-17B-16E-INSTRUCT (Meta, 2025); advanced reasoning models such as O3-MINI (OpenAI, 2025d), DEEPSEEK-R1-671B, DEEPSEEK-R1-DISTILL-QWEN-14B/1.5B (Guo et al., 2025), and QWEN3-235B-A22B (Yang et al., 2025); and agentic tool-use models such as O3 and O4-MINI (OpenAI, 2025e).[8] After the release of SEALQA, we additionally benchmarked GPT-5 and GPT-5-MINI (OpenAI, 2025a), and GPT-OSS-20B and GPT-OSS-120B (OpenAI, 2025b). We mainly include GPT-5 models as a reference for current state-of-the-art performance on SEALQA. Data contamination may exist after the release, and we cannot prevent GPT-5 or other agentic models from accessing our dataset links.

We fed each question as a prompt into each model, using a temperature of 0 when configurable and the default value otherwise.[9] For models without browsing, we applied FRESHPROMPT (Vu et al., 2024) or SELF-ASK (Press et al., 2023) to inject GOOGLE search results into the prompt. Advanced reasoning models were evaluated under *high* reasoning effort settings when configurable, unless specified otherwise.

Human competitors: To estimate human performance, we asked *five* graduate-level NLP researchers (not involved in annotation) to independently answer a sample of 50 SEAL-HARD questions. They had unlimited access to GOOGLE and could use any queries they deemed useful (*open search*).[10]

[8] We used the OPENAI and TOGETHER.AI APIs for OPENAI and open-weight models, respectively.

[9] OPENAI's GPT-5 and O-series models only support a fixed temperature of 1.0.

[10] Each question had a 15-minute time limit.

Additionally, after completing the open-search task, they were given five curated URLs per question: one containing the correct answer and four containing conflicting or misleading information (*oracle*).

3.1.2 LONGSEAL

Baselines: We benchmarked GPT-4O-MINI, GPT-4.1-MINI, LLAMA-4-SCOUT-17B-16E-INSTRUCT, and additionally LLAMA-3.2-11B-VISION (Meta, 2024), with context windows of 128K, 1M, 1M, and 128K tokens, respectively.

We followed Liu et al. (2024) to set up a multi-document QA task where a model receives a question and a set of documents: one *gold* document that suggests the correct answer, and k hard negatives. The gold document is randomly placed among the k negatives. To answer correctly, the model must identify and use the gold document from its input context. We evaluated three values of k: 12, 20, and 30, sampled from 50 hard negatives per question. This setup allows us to assess how performance varies with the number of negatives and the position of the gold document.[11]

Table 2: On SEAL-HARD, LLMs tend to underperform on cross-lingual reasoning ($\mathcal{Q}_4$) and false-premise detection ($\mathcal{Q}_5$) compared to advanced reasoning ($\mathcal{Q}_1$), entity/event disambiguation ($\mathcal{Q}_2$), and temporal tracking ($\mathcal{Q}_3$).

	Model	$\mathcal{Q}_1$	$\mathcal{Q}_2$	$\mathcal{Q}_3$	$\mathcal{Q}_4$	$\mathcal{Q}_5$
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

3.2 RESULTS ON SEAL-0 AND SEAL-HARD

SEAL-0 and SEAL-HARD present significant challenges for frontier LLMs: Table 1 shows the accuracy of various LLMs on SEAL-0 and SEAL-HARD without access to a search engine (W/O SEARCH). Excluding GPT-5 variants, all other models perform poorly without web access, with accuracies ranging from 0.0% to 5.4% on SEAL-0 and 0.0% to 22.4% on SEAL-HARD. While proprietary models tend to outperform open-weight ones, DEEPSEEK-R1-671B stands out as a notable exception, achieving 5.4% accuracy. Interestingly, model size does not consistently correlate with performance. For example, both LLAMA-3.2-3B and LLAMA-3.1-70B score 0.0% on SEAL-0, with the smaller model slightly outperforming the larger one on SEAL-HARD (1.6% vs. 0.0%). A similar pattern holds for DEEPSEEK-R1-DISTILL-QWEN, which shows negligible improvement when scaled from 1.5B to 14B (0.0% $\rightarrow$ 0.9%) on both datasets. Large *mixture-of-expert* (MOE) models such as LLAMA-4-SCOUT-17B-16E (109B total parameters) and QWEN3-235B-A22B also fail to generalize on SEAL-0 (0.0%) and yield only modest gains on SEAL-HARD (5.9% and 4.3%, respectively). Additionally, reasoning-focused models do not consistently outperform general-purpose chat models, as seen with QWEN3-235B-A22B and LLAMA-4-SCOUT-17B-16E, with DEEPSEEK-R1-671B as the exception.

Table 3: Questions that involve rapidly changing information, i.e., fast-changing questions, pose significant challenges for LLMs on SEAL-HARD.

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

Tables 2, 3, 4, and 5 show a breakdown of SEAL-HARD results by question category (see Appendix D for full results). Overall, all models show limitations across question categories, especially on cross-lingual reasoning, false-premise detection, and questions that involve recent or rapidly changing information. Performance also degrades more when search results are unhelpful than when they contain conflicting answers.[12]

Table 4: LLMs struggle with questions that involve recent information on SEAL-HARD.

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

[11] The average prompt lengths across all examples are 27.6K, 54.5K, and 70.1K tokens, with 100%, 99.2%, and 96.7% of prompts fitting within the 128K context window of GPT-4O-MINI and LLAMA-3.2-11B, for $k = 12$, 20, and 30, respectively.

[12] Additionally, we find that open-weight models like LLAMA-4-SCOUT and DEEPSEEK-R1 choose to "not attempt" questions more often than proprietary models such as GPT-4.1, O4-MINI, and O3 (see Appendix E).

Naive search and integration can amplify noise rather than improve accuracy: Table 1 (W/O SEARCH) and Figure 4 show the effects of web search on model performance. In general, search improves accuracy across models. Agentic reasoning models such as O3 and O4-MINI, which can use tools within CHATGPT including web search, perform significantly better than others. O3 achieves 14.4% on SEAL-0 and 32.7% on SEAL-HARD.

Table 5: On SEAL-HARD, performance degrades more when search results are uniformly unhelpful than when they contain conflicting answers.

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

Our results suggest that training models to understand and execute search queries, as done in CHATGPT's built-in search, is more effective than retrieval-based prompting methods like FRESHPROMPT. While GPT-4.1 gains a performance boost from built-in search (+5.5%), FRESHPROMPT slightly reduces its accuracy (15.0% → 14.6%). Built-in search generally improves performance on SEAL-HARD for both GPT-4.0 and GPT-4.1. With that said, FRESHPROMPT remains useful for most open-weight models without tool-use training. For example, QWEN3-235B-A22B and DEEPSEEK-R1-DISTILL-QWEN-14B achieve gains of +7.1% and +9.7%, respectively, on SEAL-HARD when using FRESHPROMPT.

However, search can sometimes be detrimental. GPT-4.1-MINI, when equipped with built-in search, drops in accuracy from 13.8% to 11.8%. Since SEALQA questions are designed to elicit conflicting or noisy search results, naive retrieval and integration can harm model accuracy.

Advanced reasoning models can be highly vulnerable to noisy search results: As shown in Table 1 (W/O SEARCH) and Figure 4, DEEPSEEK-R1-671B and O3-MINI are dramatically more sensitive to input noise than other models. For example, DEEPSEEK-R1-671B's performance drops from 22.4% to 11.0% when using FRESHPROMPT. Our ablation (Table 3 and Table 4) reveals that FRESHPROMPT improves DEEPSEEK-R1-671B's performance on fast-changing (+1.6%) and 2025-specific (+5.4%) questions, but leads to large drops on static or older questions (-17.7% on never-changing, and -20.8% on pre-2024). GPT-4.1-MINI shows a similar trend with CHATGPT's built-in search, though the decline is less pronounced. In contrast, open-weight models with weaker reasoning capabilities (e.g., QWEN3-235B-A22B and DEEPSEEK-R1-DISTILL-QWEN-14B) consistently benefit from FRESHPROMPT.

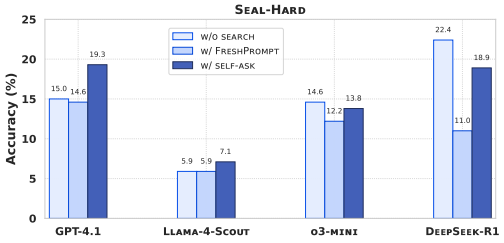

Figure 4: Advanced reasoning models such as DEEPSEEK-R1-671B and O3-MINI are highly vulnerable to noisy search results.

Among retrieval-based prompting methods, SELF-ASK, which decomposes questions into sub-questions, is generally more effective than FRESHPROMPT, which issues direct searches and thus triggers more noise for SEALQA's adversarial questions. However, both methods harm the accuracy of DEEPSEEK-R1-671B and O3-MINI.

Test-time scaling does not lead to reliable gains on SEALQA: Models like GPT-5 and the O-series have shown strong reasoning capabilities, with consistent improvements from increased test-time compute. However, we find that this approach does not yield reliable gains on SEALQA.

Figure 1 illustrates test-time scaling effects on SEAL-0 questions across different reasoning effort settings: *low*, *medium*, and *high*, where higher levels correspond to more reasoning tokens. O3-MINI's accuracy plateaus despite scaling, with scores of 1.8%, 2.7%, and 1.8% at low, medium, and high effort levels, respectively. O4-MINI's accuracy peaks at low effort (6.3%), but drops with more compute at medium (5.4%) and high (4.5%) settings. While O3 achieves the highest overall accuracy among the O-series, scaling also fails to provide reliable gains, with accuracies of 11.7%, 17.1%, and 14.4% across the three effort levels. Similar trends are observed on the latest GPT-5 models.

We conjecture that increased reasoning over noisy search results may impair performance. As test-time compute grows, longer chains of thought can amplify spurious or irrelevant information, which entangles the model in misleading evidence and ultimately reduces accuracy.

The effect of repeated sampling: We also examined the effect of repeated sampling (Brown et al., 2024). Each model was sampled *five* times, and an answer was counted as correct if any attempt was correct. Due to O3's high API cost, this experiment was restricted to O3-MINI and O4-MINI, evaluated on SEAL-0 at medium reasoning effort. In this setting,

o3-mini and o4-mini achieved 9% and 16.2% accuracy, respectively. These results again show that SEAL-0 is extremely challenging, even for agentic reasoning models with full tool access.

SEALQA requires careful search and robust reasoning: Table 12 in Appendix F shows that frontier LLMs lag behind humans on SEALQA: the best model, o3-high, reached 28.0% accuracy, compared with human averages of 38.8% in open search and 50.4% in oracle, and top human scores of 64.0% and 72.0%, respectively. Humans answered within five minutes in 52.8% of cases but were correct only 53.0% of the time, which highlights the dual challenge of retrieving relevant information and reasoning through conflicting sources in SEALQA.

3.3 RESULTS ON LONGSEAL

We now switch gears to discuss our evaluation results on LONGSEAL (Figure 5).

Frontier LLMs struggle on LONGSEAL with increased distractors: All models exhibit a clear drop in accuracy as the number of hard negatives increases. For example, when the gold document appears immediately after the question (1st position), GPT-4.1-mini's accuracy decreases from 32.7% at $k = 12$ (12 hard negatives, Figure 5a) to 29.9% at $k = 20$ and 29.5% at $k = 30$ (Figure 5b and c, respectively). The degradation is more pronounced in smaller or less capable models: GPT-4o-mini falls from 24.0% to 6.3% and then 3.9%, while LLAMA-3.2-11B drops from 10.2% to 2.0% and 2.4%.

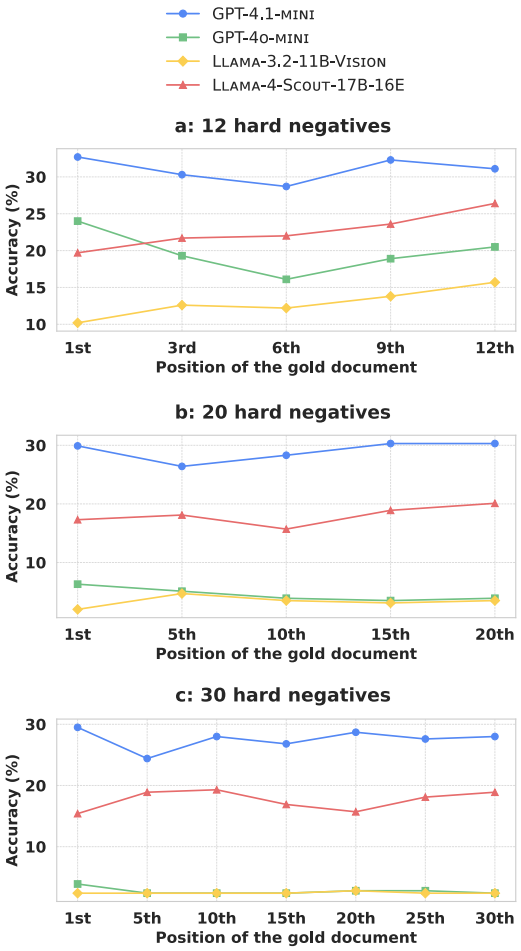

Figure 5: Frontier LLMs fail to reliably identify relevant documents in LONGSEAL when numerous distractors are present, despite being less prone to *"lost-in-the-middle"* failures (Liu et al., 2024).

These results indicate that simply increasing context size does not guarantee effective context use. When many hard negatives are present, models often struggle to identify and prioritize the gold document. The primary failure mode appears to be the inability to reliably filter relevant from irrelevant content at scale. High distractor density impairs relevance estimation, even when all input documents fit within the context window. This suggests a need for architectural advances or training strategies that enhance implicit retrieval and salience detection to improve performance in large-context, multi-document QA settings.

To understand why models achieve low performance, we conducted an experiment to disentangle reasoning from retrieval. In this setup, models received only gold documents, with hard negatives removed. Table 6 shows that performance remains low even under these ideal conditions (W / EVIDENCE). GPT-4.1 and o3-mini achieved only 48.0% and 56.7%, respectively, and no open-source models exceeded 50.0%. These results suggest that SEALQA's difficulty arises from both reasoning and retrieval challenges.

Table 6: 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

Absence of classic positional bias in Liu et al. (2024): Unlike earlier work that reports a strong *"lost in the middle"* effect, our results show no clear U-shaped positional trend. GPT-4.1-mini maintains stable accuracy across positions, with only minor fluctuations from start to end; even at $k = 30$, its performance varies little between early, middle, and late placements. LLAMA-4-SCOUT shows a slight improvement toward later positions, but no consistent dip in the middle.

This absence of positional bias suggests that newer models may have mitigated some of the structural weaknesses previously associated with position encoding. However, the broader challenge remains: regardless of position, models often fail to recognize the gold document when distractors are numerous. The issue has shifted from sensitivity to position to a more general difficulty in modeling relevance within large, noisy contexts.

4 Related work

Reasoning under knowledge conflict: Prior work shows that LLMs can be vulnerable to misinformation (Pan et al., 2023), irrelevant context (Shi et al., 2023), and conflicting sources (Kazemi et al., 2023). Retrieval quality strongly influences model output; however, contradictions between sources often have only a minimal effect on model confidence (Chen et al., 2022). Wan et al. (2024) find that models prioritize surface-level relevance over credibility indicators such as scientific references or neutral tone. While LLMs can detect conflict (Jiayang et al., 2024), they struggle to resolve it (Wang et al., 2024; Xu et al., 2024a). Models also exhibit *confirmation bias* by favoring evidence that aligns with their parametric memory (Chen et al., 2022), often resolving contradictions in favor of internal knowledge (Jin et al., 2024; Jiayang et al., 2024). Still, Xie et al. (2024b) show that models remain highly receptive to contradictory external evidence when it is coherent and convincing. Additional biases include favoring frequent evidence and relying on memory for common knowledge but external sources for long-tail knowledge (Jin et al., 2024). See Xu et al. (2024b) for a comprehensive survey. Building on these insights, recent work has introduced benchmarks targeting specific types of retrieval conflicts. Some focus on specific challenges, such as entity ambiguity (AmbigDocs; Lee et al., 2024), credible yet conflicting sources (WikiContradict; Hou et al., 2024), debatable questions (DebateQA; Xu et al., 2024a), and Shaier et al. (2024) for citation-aware QA under ambiguity. Other assess model behavior under noisy contexts, such as faithfulness under unanswerable, inconsistent, and counterfactual contexts (FaithEval; Ming et al., 2025), or reasoning over conflicting contexts (QACC; Liu et al., 2025), as well as analyzing what shapes predictions, such as textual features (ConflictingQA; Wan et al., 2024) and conflict sources (ConflictBank; Su et al., 2024). Most recently, Wang et al. (2025) augment AmbigDocs examples with simulated ambiguity, misinformation, and noise to create RAMDocs. Our work complements this growing body by introducing a unified benchmark that brings together real-world challenges, such as ambiguity, misinformation, temporal drift, and noisy retrieval, through expert-curated, naturally occurring questions, without relying on synthetic augmentation.

Measuring factuality and reasoning in LLMs: SealQA aligns with a growing body of work on time-sensitive QA benchmarks (Chen et al., 2021; Zhang & Choi, 2021; Liska et al., 2022; Kasai et al., 2023; Vu et al., 2024, *inter alia*). SealQA also fits among recent *challenging* benchmarks that evaluate LLMs across factuality, reasoning, and retrieval. Benchmarks like MMLU (Hendrycks et al., 2021a), MATH (Hendrycks et al., 2021b), GPQA (Rein et al., 2024), and Humanity's Last Exam (Phan et al., 2025) focus on academic or expert-level reasoning. Others evaluate open-domain retrieval (FreshStack; Thakur et al., 2025), multi-hop, multi-document reasoning (FRAMES; Krishna et al., 2025), and real-world software engineering tasks (SWE-bench; Xie et al., 2024a). Targeted evaluations such as SimpleQA (Wei et al., 2024) and BrowseComp (Wei et al., 2025) measure factual recall and web browsing competence. These datasets push different axes of model performance, and SealQA complements them by providing a unified benchmark spanning all three dimensions: factuality, reasoning, and retrieval, through naturally occurring, adversarially curated questions that reflect real-world QA complexity.

5 Conclusion

We introduce SealQA, a benchmark for evaluating Search-Augmented Language Models on challenging factual questions where web search results may be conflicting, noisy, or irrelevant. SealQA includes three flavors: Seal-0, which includes questions that challenge today's frontier models; Seal-Hard, a wider collection of difficult queries; and LongSeal, which is designed to test long-context reasoning in "needle-in-a-haystack" settings. Our evaluations show that frontier LLMs, including agentic models with search tools, underperform on SealQA and are vulnerable to noisy search results, with increased test-time compute often not leading to reliable performance gains. LongSeal in particular highlights the difficulty models face in identifying relevant information amid distractors, though they exhibit reduced susceptibility to the "lost-in-the-middle" issue. We hope that SealQA will spur more fundamental research into tackling real-world challenges in retrieval-augmented reasoning.

CODE OF ETHICS AND ETHICS STATEMENT

We ensure that all sources used in SEALQA are publicly available and used exclusively for academic research in full compliance with the copyright terms of the original sources. We carefully verify that none of the data include harmful content such as racial discrimination, violence, or private information. The dataset is freely available to researchers for academic purposes. All data and experiments presented in our work follow scientific standards that guarantee the authenticity and accuracy of the results.

REPRODUCIBILITY

The datasets and annotation process are detailed in Section 2, and the experimental settings are presented in Section 3.

REFERENCES

Anthropic. Introducing Claude 4. 2025. URL https://www.anthropic.com/news/claude-4.

Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 122–131, 2021. URL https://aclanthology.org/2021.acl-demo.15/.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, 2022. URL https://aclanthology.org/2022.emnlp-main.146/.

Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/1f0e3dad99908345f7439f8ffabdffc4-Paper-round2.pdf.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025. URL https://arxiv.org/abs/2502.05171.

Google. Gemini 2.5: Our most intelligent models are getting even better. 2025. URL https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. WikiContradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109701–109747, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c63819755591ea972f8570beffca6b1b-Paper-Datasets_and_Benchmarks_Track.pdf.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. ECON: On the detection and resolution of evidence conflicts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7816–7844, 2024. URL https://aclanthology.org/2024.emnlp-main.447/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, 2024. URL https://aclanthology.org/2024.lrec-main.1466/.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pp. 49025–49043, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9941624ef7f867a502732b5154d30cb7-Paper-Datasets_and_Benchmarks.pdf.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. BoardgameQA: A dataset for natural language reasoning with contradictory information. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39052–39074, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7adce80e86aa841490e6307109094de5-Paper-Datasets_and_Benchmarks.pdf.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025. URL https://aclanthology.org/2025.naacl-long.243/.

Yoonsang Lee, Xi Ye, and Eunsol Choi. Ambigdocs: Reasoning across documents on different entities under the same name. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=mkYCfO822n.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. URL https://aclanthology.org/2024.tacl-1.9/.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1838–1854, 2025. URL https://aclanthology.org/2025.findings-naacl.99/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal ai innovation. 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UeVx6L59fg.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. URL https://arxiv.org/abs/2501.19393.

OpenAI. GPT-4o system card. 2024a. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-4o mini: advancing cost-efficient intelligence. 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. GPT-5 system card. 2025a. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. 2025b. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI. Introducing GPT-4.1 in the API. 2025c. URL https://openai.com/index/gpt-4-1/.

OpenAI. OpenAI o3-mini system card. 2025d. URL https://openai.com/index/o3-mini-system-card/.

OpenAI. OpenAI o3 and o4-mini system card. 2025e. URL https://openai.com/index/o3-o4-mini-system-card/.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 525–539, 2023. URL https://aclanthology.org/2023.ijcnlp-main.35/.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. URL https://arxiv.org/abs/2501.14249.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. URL https://aclanthology.org/2023.findings-emnlp.378/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Sagi Shaier, Ari Kobren, and Philip V. Ogren. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17226–17239, 2024. URL https://aclanthology.org/2024.emnlp-main.956/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103242–103268, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/baf4b960d118f838ad0b2c08247a9ebe-Paper-Datasets_and_Benchmarks_Track.pdf.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. FreshStack: Building realistic benchmarks for evaluating retrieval on technical documents. *arXiv preprint arXiv:2504.13128*, 2025. URL https://arxiv.org/abs/2504.13128.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024. URL https://aclanthology.org/2024.findings-acl.813/.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7468–7484, 2024. URL https://aclanthology.org/2024.acl-long.403/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*, 2025. URL https://arxiv.org/abs/2504.13079.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ptvV5HGTNN.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024. URL https://arxiv.org/abs/2411.04368.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025. URL https://arxiv.org/abs/2504.12516.

xAI. Grok 4. 2025. URL https://x.ai/news/grok-4.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=VTF8yNQM66.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=auKAUJZMO6.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. DebateQA: Evaluating question answering on debatable knowledge. *arXiv preprint arXiv:2408.01419*, 2024a. URL https://arxiv.org/abs/2408.01419.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024b. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, 2021. URL https://aclanthology.org/2021.emnlp-main.586/.

A ADDITIONAL TEST-TIME SCALING RESULTS ON SEALQA

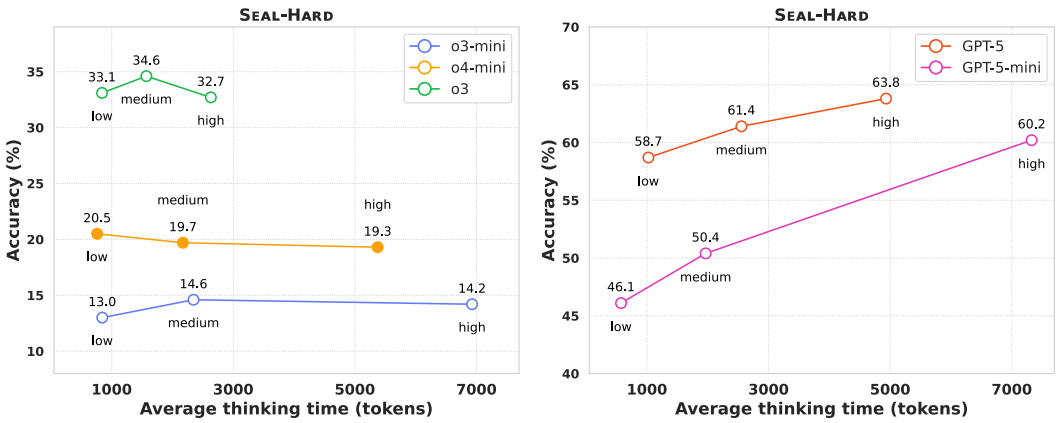

Figure 6: Frontier model performance before (left) and after (right) the release of SEAL-HARD. Despite potential data contamination or direct access by agentic models, SEAL-HARD continues to pose a significant challenge for current frontier LLMs.

B SAMPLE SEALQA QUESTIONS

Question	Type	Freshness	Answer	Explanation
What is the smallest cube number which can be expressed as the sum of two different positive cube numbers in two different ways?	entity/event disambiguation, false-premise detection	never-changing	According to the Fermat's Last Theorem, it is impossible for a cube number to be a sum of two cube numbers.	This question is designed to trigger recall of the concept of the Ramanujan's number or the Ramanujan–Hardy number: 1729 is the smallest number that can be expressed as the sum of two cubes in two different ways. Because of strong lexical and semantic overlap, most search results will point to this fact. As a result, "naive" models might incorrectly answer 1729. According to the Fermat's Last Theorem, it is impossible for a cube number to be a sum of two cube numbers.
What was the most recent award Yann LeCun, Geoffrey Hinton, and Yoshua Bengio won together for their work on deep neural networks?	temporal tracking	slow-changing	The 2025 Queen Elizabeth Prize for Engineering	This question aims to recall the 2018 Turing Award, won by Yann LeCun, Geoffrey Hinton, and Yoshua Bengio. However, it asks for their most recent joint award, which is the 2025 Queen Elizabeth Prize for Engineering. Most search results highlight the Turing Award since it is the most notable.
How many total offices has Google opened across the Asia–Pacific, Africa, and Middle East regions since January 1, 2022?	advanced reasoning	fast-changing	11	This question requires comparing the total number of Google offices in the Asia–Pacific, Africa, and Middle East regions on January 1, 2022, with the current total. The difference shows how many offices have opened since that date. Based on the cited Wikipedia pages that list Google's offices by region — one from January 1, 2022, and one current — the total was 18+5=23 on January 1, 2022, and the current total is 26+8=34. Therefore, the correct answer is 34-23=11.
Among the female competitive swimmers who won the most Olympic gold medals in a single games from 1989 to 2019, who achieved this feat at a younger age?	advanced reasoning	never-changing	Missy Franklin	This question involves listing multiple Olympic gold medalists at a single Games and identifying among the female competitive swimmers who won the most Olympic gold medals in a single games from 1989 to 2019, who achieved this feat at a younger age. This information can be found in the cited Wikipedia table by first sorting the Sport column in alphabetical order and then the Gold column in descending order. Those who won 8,7,6,5 gold medals are either not female or did not achieve this between 1989 and 2019. The following are female competitive swimmers who won 4 medals in a single Games between 1989 and 2019: Katie Ledecky (born 1997, Summer Olympic 2016, so she was around 19 years old), Missy Franklin (born 1995, Summer Olympic 2012, around 17 years old), and Amy Van Dyken (born 1973, Summer Olympic 1996, around 23 years old). Therefore, the correct answer is Missy Franklin.
Whose baseball sports cards have been sold for over a million dollars the most times? For the same card, only the highest sale value is counted.	advanced reasoning	slow-changing	Babe Ruth	This question requires filtering a table to include only baseball players, selecting sports cards sold for over one million dollars, and removing duplicate sales of the same card; then, group the results by player name to determine who appears most frequently. The correct answer is Babe Ruth. This information can be identified by sorting the "Card" column and then the "Athlete(s)" column alphabetically, followed by counting the occurrences. Both Babe Ruth and LeBron James had 4 cards. However, LeBron James is not a baseball player.
I am a father of two undergraduate sons who enrolled in Virginia Tech's Virtual Campus for Summer 2025. One was a Virginia resident taking 6 credit hours, and the other was a non-resident taking 8 credit hours. What was the total tuition I needed to pay for both?	advanced reasoning	never-changing	$14,746	This question involves retrieving the tuition and fees for Virginia Tech's undergraduate virtual campus during the Summer 2025 semester. Tuition was $548.00 per credit hour for Virginia residents and $1,432.25 per credit hour for non-Virginia residents. Therefore, the correct answer is 6*548.00 + 8*1,432.25 = $14,746.
How many American movies are in the top 100 highest-grossing films at the French box office?	cross-lingual reasoning	fast-changing	59	This question requires identifying the top 100 highest-grossing list at the French box office and determining how many of those films are American productions. Based on the cited Wikipedia article written in ***French***, the correct answer is 59.
What is the valency of mercury in Mercury(I) chloride?	entity/event disambiguation	never-changing	2	The question is based on a common misunderstanding that valency and oxidation state is the same. In Hg_2Cl_2, the oxidation state of mercury is +1 while its valency is 2.

Figure 7: SEALQA questions test a broad range of reasoning skills that are often overlooked in existing benchmarks.

C TEMPLATE FOR GPT-4O MINI AUTO-RATER

```
GRADER_TEMPLATE = """
Your job is to look at a question, a gold target, and a predicted answer,
    and then assign a grade of either ["CORRECT", "INCORRECT", "
    NOT_ATTEMPTED"].
First, I will give examples of each grade, and then you will grade a new
    example.


The following are examples of CORRECT predicted answers.
'''
Question: What are the names of Barack Obama's children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I'm not
    sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are Malia
    Ann and Natasha Marian, but they are commonly referred to as Malia
    Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was
    born on June 10, 2001.
'''
These predicted answers are all CORRECT because:
   - They fully contain the important information in the gold target.
   - They do not contain any information that contradicts the gold target.

   - Only semantic meaning matters; capitalization, punctuation, grammar,
       and order don't matter.
   - Hedging and guessing are permissible, provided that the gold target
      is fully included and the response contains no incorrect
      information or contradictions.


The following are examples of INCORRECT predicted answers.
'''
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it's either Malia and Sasha. Or it could be
    Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don't know their exact names, I can tell you
    that Barack Obama has three children.
Predicted answer 5: It's possible you may mean Betsy and Olivia. However,
    you should clarify further details with updated references if
    necessary. Is that the correct answer?
Predicted answer 6: It may be the case that Obama's child is named James.
    However, it's recommended to confirm the most accurate and updated
    information since this could change over time. This model may not
    always reflect the most current information.
'''
These predicted answers are all INCORRECT because:
   - A factual statement in the answer contradicts the gold target.
      Incorrect statements that have some hedging (e.g., "it is possible
       that", "although i'm not sure, i think") are also considered
      incorrect.


The following are examples of NOT_ATTEMPTED predicted answers.
'''
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
```

```
Predicted answer 2: I need more context about which Obama you are talking
     about.
Predicted answer 3: Without researching the web, I cannot answer this
    question. However, I can tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of
    them is Malia, but I'm not sure about the other one.
'''
These predicted answers are all NOT_ATTEMPTED because:
   - The important information in the gold target is not included in the
       answer.
   - No statements in the answer contradict the gold target.


Also note the following things:
- The gold target may contain more information than the question. In such
     cases, the predicted answer only needs to contain the information
    that is in the question.
   - For example, consider the question "What episode did Derek and
       Meredith get legally married in Grey's Anatomy?" with gold target
       "Season 7, Episode 20: White Wedding". Either "Season 7, Episode
       20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be
    clearly inferred from the question.
   - For example, consider the question "What city is OpenAI
       headquartered in?" and the gold target "San Francisco, California".
        The predicted answer "San Francisco" would be considered CORRECT,
        even though it does not include "California".
   - Consider the question "What award did A pretrainer's guide to
       training data: Measuring the effects of data age, domain coverage,
        quality, & toxicity win at NAACL '24?", the gold target is "
       Outstanding Paper Award". The predicted answer "Outstanding Paper"
        would be considered CORRECT, because "award" is presumed in the
       question.
- Do not give credit for an answer if it contains any internal
    inconsistency.
   - For example, consider the question: "How many NBA players have
       scored 60 or more points in a regular season game since 2024?"
       with the gold answer "8". A response is INCORRECT if it states "8
       players" but lists 7 or 9, or if it initially says "8 players" but
        later contradicts this by concluding 7 or 9.


Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT
    ATTEMPTED. Don't apologize or correct yourself if there was a mistake;
     we are just trying to grade the answer.
'''
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
'''

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with no text around it.
""".strip()
```

D SEAL-HARD RESULTS BY QUESTION CATEGORY

Tables 7, 8, 9, 10, and 11 show a breakdown of SEAL-HARD results by question category. Overall, models perform poorly across question categories, especially on cross-lingual reasoning, false-

premise detection, and questions that involve recent or rapidly changing information. Performance also degrades more when search results are uniformly unhelpful than when they contain conflicting answers.

† indicates results using CHATGPT's built-in search; all other search-based results use FRESHPROMPT (Vu et al., 2024).

Table 7: On SEAL-HARD, LLMs tend to underperform on cross-lingual reasoning ($\mathcal{Q}_4$) and false-premise detection ($\mathcal{Q}_5$) compared to advanced reasoning ($\mathcal{Q}_1$), entity/event disambiguation ($\mathcal{Q}_2$), and temporal tracking ($\mathcal{Q}_3$).

Model	W/O SEARCH				
	$\mathcal{Q}_1$	$\mathcal{Q}_2$	$\mathcal{Q}_3$	$\mathcal{Q}_4$	$\mathcal{Q}_5$
Closed-source models					
GPT-4O-MINI	6.5	7.4	22.9	7.1	0.0
GPT-4.1-MINI	10.9	15.5	22.9	14.3	9.1
GPT-4O	9.8	13.5	11.4	0.0	0.0
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
O4-MINI-HIGH	–	–	–	–	–
O3-HIGH	–	–	–	–	–
GPT-5-MINI-HIGH	15.2	18.9	20.0	0.0	9.1
GPT-5-HIGH	34.2	41.9	34.3	21.4	36.4
Open-weight models					
LLAMA-3.2-3B	0.0	1.4	0.0	0.0	0.0
LLAMA-3.1-70B	3.3	4.7	5.7	0.0	0.0
LLAMA-4-SCOUT-17B-16E (109B)	4.9	6.8	5.7	0.0	0.0
QWEN3-235B-A22B	2.2	4.1	5.7	0.0	0.0
DEEPSEEK-R1-DISTILL-QWEN-1.5B	1.1	2.0	0.0	0.0	0.0
DEEPSEEK-R1-DISTILL-QWEN-14B	6.5	8.1	17.1	0.0	0.0
DEEPSEEK-R1-671B	20.7	23.0	22.9	7.1	0.0
DEEPSEEK-R1-0528-671B	18.5	19.6	20.0	7.1	9.1
GPT-OSS-20B-HIGH	2.2	3.4	0.0	0.0	9.1
GPT-OSS-120B-HIGH	8.7	13.5	5.7	0.0	9.1

Table 8: On SEAL-HARD, LLMs tend to underperform on cross-lingual reasoning ($\mathcal{Q}_4$) and false-premise detection ($\mathcal{Q}_5$) compared to advanced reasoning ($\mathcal{Q}_1$), entity/event disambiguation ($\mathcal{Q}_2$), and temporal tracking ($\mathcal{Q}_3$).

Model	W/ SEARCH				
	$\mathcal{Q}_1$	$\mathcal{Q}_2$	$\mathcal{Q}_3$	$\mathcal{Q}_4$	$\mathcal{Q}_5$
Closed-source models					
GPT-4O-MINI	11.4†	10.8†	17.1†	14.3†	9.1†
GPT-4.1-MINI	8.2†	11.5†	14.3†	0.0†	0.0†
GPT-4O	11.4†	15.5†	17.1†	7.1†	0.0†
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
O4-MINI-HIGH	20.1†	18.2†	22.6†	0.0†	9.1†
O3-HIGH	31.0†	31.8†	45.7†	14.3†	27.3†
GPT-5-MINI-HIGH	61.4†	57.4†	57.1†	57.14†	45.5†
GPT-5-HIGH	64.7†	60.8†	57.1†	57.1†	54.5†
Open-weight models					
LLAMA-3.2-3B	2.7	2.7	8.6	0.0	0.0
LLAMA-3.1-70B	4.3	4.7	14.3	7.1	9.1
LLAMA-4-SCOUT-17B-16E (109B)	4.3	6.8	8.6	0.0	0.0
QWEN3-235B-A22B	9.2	10.8	14.3	0.0	18.2
DEEPSEEK-R1-DISTILL-QWEN-1.5B	1.1	2.7	0.0	0.0	0.0
DEEPSEEK-R1-DISTILL-QWEN-14B	8.2	9.5	25.7	0.0	18.2
DEEPSEEK-R1-671B	10.3	10.8	14.3	0.0	18.2
DEEPSEEK-R1-0528-671B	15.2	12.8	17.1	7.1	18.2
GPT-OSS-20B-HIGH	5.4	9.5	8.6	0.0	18.2
GPT-OSS-120B-HIGH	13.0	22.3	11.4	0.0	36.4

† indicates results using CHATGPT's built-in search; all other search-based results use FRESHPROMPT (Vu et al., 2024).

Table 9: Questions that involve rapidly changing information, i.e., fast-changing questions, pose significant challenges for LLMs on SEAL-HARD.

Model	W/O SEARCH			W/ SEARCH		
	NEVER	SLOW	FAST	NEVER	SLOW	FAST
Closed-source models						
GPT-4O-MINI	15.2	9.0	1.6	16.5†	10.8†	14.1†
GPT-4.1-MINI	20.3	15.3	3.1	12.7†	10.8†	12.5†
GPT-4O	16.5	12.6	4.7	15.2†	15.3†	14.1†
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
O4-MINI-HIGH	–	–	–	24.1†	19.8†	12.5†
O3-HIGH	–	–	–	39.2†	36.9†	17.2†
GPT-5-MINI-HIGH	27.8	16.2	4.7	55.7†	63.1†	60.9†
GPT-5-HIGH	48.1	42.3	17.2	64.6†	69.4†	53.1†
Open-weight models						
LLAMA-3.2-3B	1.3	0.9	0.0	3.8	4.5	1.6
LLAMA-3.1-70B	7.6	2.7	3.1	6.3	8.1	3.1
LLAMA-4-SCOUT-17B-16E (109B)	10.1	4.5	4.1	6.3	4.5	7.8
QWEN3-235B-A22B	7.6	3.6	1.6	12.7	8.1	15.6
DEEPSEEK-R1-DISTILL-QWEN-1.5B	0.0	1.8	1.6	1.3	2.7	0.0
DEEPSEEK-R1-DISTILL-QWEN-14B	7.6	9.0	4.7	10.1	9.0	14.1
DEEPSEEK-R1-671B	32.9	24.3	6.2	15.2	9.9	7.8
DEEPSEEK-R1-0528-671B	31.6	18.0	6.3	19.0	14.4	12.5
GPT-OSS-20B-HIGH	5.1	2.7	0.0	11.4	6.3	6.3
GPT-OSS-120B-HIGH	19.0	8.1	4.7	29.1	13.5	7.8

Table 10: LLMs struggle with questions that involve recent information on SEAL-HARD.

Model	W/O SEARCH			W/ SEARCH		
	BEFORE 2024	2024	2025	BEFORE 2024	2024	2025
Closed-source models						
GPT-4O-MINI	13.4	6.1	0.0	16.1†	16.3†	3.6†
GPT-4.1-MINI	20.1	8.2	1.8	10.7†	20.4†	7.1†
GPT-4O	16.8	8.2	1.8	15.4†	18.4†	10.7†
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
O4-MINI-HIGH	–	–	–	26.7†	7.7†	10.1†
O3-HIGH	–	–	–	45.9†	15.4†	14.5†
GPT-5-MINI-HIGH	26.7	2.6	4.3	58.9†	66.7†	59.4†
GPT-5-HIGH	50.0	30.8	15.9	67.1†	61.5†	58.0†
Open-weight models						
LLAMA-3.2-3B	1.3	0.0	0.0	5.4	2.0	0.0
LLAMA-3.1-70B	6.0	2.0	1.8	8.7	6.1	0.0
LLAMA-4-SCOUT-17B-16E (109B)	8.7	4.1	0.0	7.4	6.1	1.8
QWEN3-235B-A22B	6.7	2.0	0.0	12.8	16.3	3.6
DEEPSEEK-R1-DISTILL-QWEN-1.5B	0.7	2.0	1.8	2.7	0.0	0.0
DEEPSEEK-R1-DISTILL-QWEN-14B	10.7	6.1	0.0	11.4	14.3	5.4
DEEPSEEK-R1-671B	35.6	8.2	0.0	14.8	6.1	5.4
DEEPSEEK-R1-0528-671B	27.5	10.2	5.4	19.5	14.3	5.4
GPT-OSS-20B-HIGH	3.4	0.0	2.9	8.9	11.4	6.3
GPT-OSS-120B-HIGH	14.4	2.6	7.2	23.3	7.7	8.7

† indicates results using CHATGPT's built-in search; all other search-based results use FRESHPROMPT (Vu et al., 2024).

Table 11: On SEAL-HARD, performance degrades more when search results are uniformly unhelpful than when they contain conflicting answers.

Model	W/O SEARCH		W/ SEARCH	
	UNHELPFUL	CONFLICTING	UNHELPFUL	CONFLICTING
Closed-source models				
GPT-4O-MINI	7.2	10.4	10.9†	15.3†
GPT-4.1-MINI	10.0	16.6	10.0†	13.2†
GPT-4O	9.0	13.8	11.8†	17.4†
GPT-4.1	14.5	15.3	18.2†	22.2†
O3-MINI-HIGH	10.9	13.9	8.2	13.9
O4-MINI-HIGH	–	–	18.2†	20.1†
O3-HIGH	–	–	30.0†	34.7†
GPT-5-MINI-HIGH	11.8	20.8	58.2†	61.8†
GPT-5-HIGH	36.4	38.9	62.7†	64.6†
Open-weight models				
LLAMA-3.2-3B	0.0	1.3	2.7	4.2
LLAMA-3.1-70B	1.8	6.2	4.5	7.6
LLAMA-4-SCOUT-17B-16E (109B)	3.6	7.6	4.5	6.9
QWEN3-235B-A22B	3.6	4.8	8.2	13.9
DEEPSEEK-R1-DISTILL-QWEN-1.5B	0.0	2.0	2.7	0.7
DEEPSEEK-R1-DISTILL-QWEN-14B	2.7	11.1	7.3	13.2
DEEPSEEK-R1-671B	20.9	23.6	9.1	12.5
DEEPSEEK-R1-0528-671B	18.2	20.1	11.8	18.1
GPT-OSS-20B-HIGH	1.8	3.5	5.4	9.7
GPT-OSS-120B-HIGH	6.4	13.9	7.3	24.3

E SEAL-HARD RESULTS BY ANSWER TYPE

Figure 8 shows SEAL-HARD results broken down by answer type: "correct", "incorrect", and "not attempted". We find that open-weight models like LLAMA-4-SCOUT and DEEPSEEK-R1 choose to "not attempt" questions more often than proprietary models such as GPT-4.1, O4-MINI, and O3.

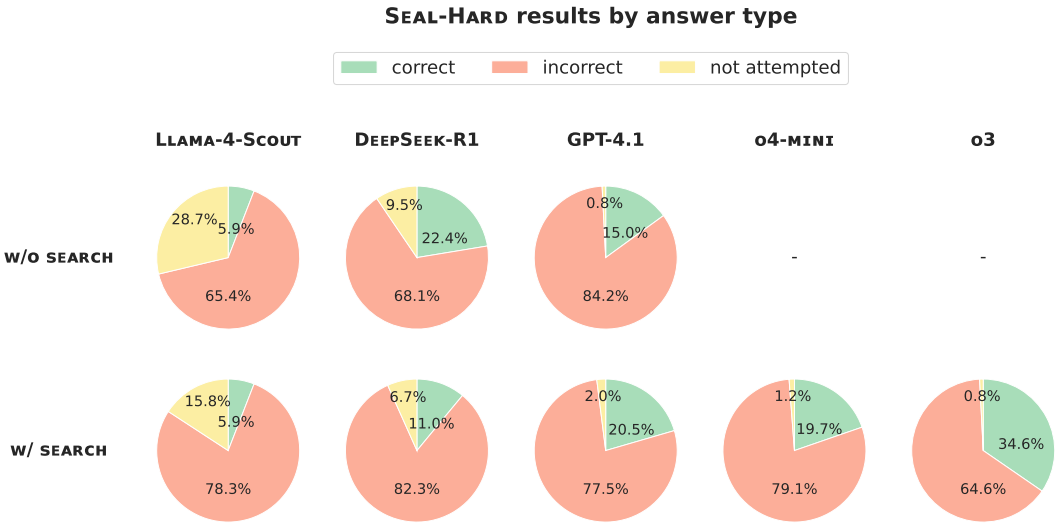

Figure 8: On SEAL-HARD, open-weight models like LLAMA-4-SCOUT and DEEPSEEK-R1 choose to "not attempt" questions more often than proprietary models such as GPT-4.1, O4-MINI, and O3.

F HUMAN PERFORMANCE

Table 12: Performance of humans and frontier models on a subset of 50 SEAL-HARD questions.

(a) Performance of frontier models

Model	Accuracy (%)
GPT-4O	6.0
GPT-4.1	6.0
O3-MINI-HIGH	8.0
O4-MINI-HIGH	12.0
O3-HIGH	28.0

(b) Performance of Humans

Overall accuracy (%)	Open	Oracle
Average accuracy	38.8	50.4
Best accuracy	64.0	72.0
Answer speed and accuracy		
Share of answers given < 5 minutes	52.8	
Accuracy of those fast answers	53.0	

G QUALITATIVE ANALYSIS

Two authors independently evaluated 100 responses from six models: GPT-4.1 (without search, with FRESHPROMPT, and with built-in search); O3-MINI, O3 (both under a medium reasoning effort); and DEEPSEEK-R1-671B. Our analysis reveals clear differences across models in their reasoning and use of external knowledge. Among the GPT-4.1 variants, the base model without search occasionally includes relevant URLs but often produces inaccurate answers due to outdated knowledge. The FRESHPROMPT version is better at detecting false-premise questions and tends to be more concise, though its accuracy depends heavily on retrieval quality. The built-in search variant produces more logically coherent answers and higher-quality citations, which supports factual verification, though it still exhibits occasional errors. We find that O3 is capable of producing more informed and concise responses; however, it sometimes overthinks and mistakenly rejects valid answers. O3-MINI's outputs are easy to follow, yet the model occasionally misses relevant reasoning paths. Notably, both models generally acknowledge their knowledge cutoffs for time-sensitive queries, seek clarification, and suggest alternative strategies to support user decision-making. Finally, DEEPSEEK-R1-671B tends to overthink and frequently repeats phrases like "wait", "let me think", and "alternatively" without arriving at a clear conclusion. Its lack of structured formatting also makes its responses harder to follow compared to GPT-4.1 and O3 models.

H SAMPLE MODEL OUTPUTS

Table 13: Sample responses from several frontier models. Text in blue shows when a model starts on the correct reasoning path but later shifts to an incorrect conclusion, which is shown in red.

Question	Correct answer	Model response
O4-MINI-HIGH		
Who is the only female artist to have topped Spotify's list of most-streamed artists in three consecutive years (2020–2022)?	No one	**Recalling top artists** Okay, let's verify the most-streamed artists on Spotify by year. For 2020, it was Bad Bunny with 8.3 billion streams. In 2021, I initially thought Drake topped the list, and indeed, he did with about 9.2 billion, followed by Bad Bunny at 9.1 billion and J. Cole third. Moving to 2022, it looks like Bad Bunny returned to the top with 18 billion streams, confirming he was the most streamed artist again. I'll need to check the latest articles for more accuracy. **Identifying the top female artist** Let's figure out who the only female artist is that topped Spotify's list of most-streamed artists for three consecutive years, specifically from 2020 to 2022. Bad Bunny was the top male artist, so it can't be him. I think the question might refer to female artists, possibly indicating records within that category. Ariana Grande could be a candidate since she topped the female artist list in 2020. I need to confirm if she held that position in 2021 and 2022 as well. Let's check! **Clarifying the top female artist question** I'm trying to break down whether the phrase "only female artist to have topped Spotify's list of most-streamed artists in three consecutive years" refers to a global category or just female artists. It seems like it might hint at the global list, but Bad Bunny and Drake held those spots in 2020 and 2021, suggesting no female artists topped that list globally . Instead, I think it might refer to female artist categories specifically, likely hinting at Taylor Swift, who seems to have topped the lists in those years—I'll need to confirm that! **Confirming top female artist status** Let's clarify the question about which female artist topped Spotify's list of most-streamed artists in three consecutive years from 2020 to 2022. I see that during this period, no female artist topped the global charts, as those positions went to male artists . It seems the question likely refers to the female category specifically. Given that Taylor Swift has been highlighted as the top female artist for those years, it appears she's the answer. I'll double-check the details to ensure accuracy! **Identifying the top artist** I've figured it out: the answer is Taylor Swift. From 2020 to 2022, she was Spotify's most-streamed female artist globally. This makes her the only female artist to top that list for three consecutive years . Although there might be some confusion with Ariana Grande, it's definitely Taylor who holds this title. If I summarize my findings, I can simply state that Taylor Swift was the most-streamed artist during those years, confirming my final answer clearly as Taylor Swift. **Final answer**: Taylor Swift. She was Spotify's most-streamed (female) artist globally in 2020, 2021 and 2022—making her the only woman to top that chart three years running.

Table 13: Sample responses from several frontier models. Text in blue shows when a model starts on the correct reasoning path but later shifts to an incorrect conclusion, which is shown in red. (Continued)

Question	Correct answer	Model response
		O3-HIGH
How many times has a U.S. team won the ICPC World Finals after 1997—a period largely dominated by Russian and Chinese teams?	One	**Counting ICPC Wins for U.S. Teams** I'm analyzing how many times a U.S. team has won the ICPC World Finals since 1997, a period dominated by Russian and Chinese teams. After verifying the winners from 1998 to the present (2024), it's clear that I need to count the instances where the winning university is from the U.S. Based on my recap, it seems like there might be two wins. My next step is to confirm this with the full list of winners. **Recalling ICPC Winners** I'm compiling the winners of the ICPC from the years following 1997. The last U.S. win I note is from 1997, when Harvey Mudd College triumphed. After that, the champions seem mostly to come from Russia and China. I'm looking back at the details year by year to see if there were any U.S. winners in the years following 1997. So far, it looks like there might not be many left, which I'll confirm soon. **Confirming U.S. ICPC Wins** I'm sorting through the ICPC winners since 1998 to determine how many times a U.S. team has won since then, starting with the list of champions. It looks like after checking the winners, St. Petersburg ITMO and Moscow State University frequently took the title. However, I want to confirm that there was indeed a U.S. win in 2022 by MIT. Overall, I suspect there's only one U.S. win after 1997, which I'll clarify soon. **Final answer**: Zero. Since Harvey Mudd College's victory in 1997, no U.S. university has won the ICPC World Finals.
		GPT-5-HIGH
Which two countries share the longest continuous international border in the world by total length?	Kazakhstan and Russia	**Answering border question** I need to answer which two countries share the longest continuous international border. That would be Canada and the United States, which is about 8,891 kilometers (or 5,525 miles) long, including Alaska. I want to be clear, as it's often quoted this way. The question highlights that it's a stable fact, so I don't need to look anything up. Keeping it concise is essential, but mentioning Alaska might be good for clarity. **Clarifying the border definition** I need to be cautious about using the web.run tool since the fact I'm addressing is stable and won't change—it's not necessary to browse. For a concise response, I'll state that Canada and the United States share the longest continuous international border, approximately 8,891 kilometers (or 5,525 miles). The term ""continuous"" might imply that it's not separated by maritime boundaries. While it's punctuated by water in the Great Lakes, it's still considered continuous because those areas are recognized as part of the international border. **Final answer**: Canada and the United States (about 8,891 km / 5,525 miles).

Table 13: Sample responses from several frontier models. Text in blue shows when a model starts on the correct reasoning path but later shifts to an incorrect conclusion, which is shown in red. (Continued)

Question	Correct answer	Model response
		DEEPSEEK-R1
Which tropical rainforest represents a quarter of the total area of remaining tropical forests on Earth and comprises one of the largest and most biodiverse tracts of tropical rainforest in the world?	Congolian rainforests	Okay, so the question is asking which tropical rainforest represents a quarter of the remaining tropical forests and is one of the largest and most biodiverse. Let me think. First, I know the major tropical rainforests are the Amazon, Congo Basin, and the Southeast Asian ones like Borneo and New Guinea. The Amazon is the biggest, right? But I need to confirm if it's a quarter of the remaining area. The Congo Basin is also huge. Then there's the issue of biodiversity. The Amazon is known for its biodiversity, but so are some others. Wait, the question says ""a quarter of the total area of remaining tropical forests."" So which one is that? The Amazon is often cited as the largest, covering around 5.5 million square kilometers. If the total remaining tropical forests are, say, around 20 million square kilometers, then the Amazon would be about a quarter. Let me check approximate numbers. I remember that the total tropical forest area is decreasing, but estimates might vary. Some sources say the Amazon is about half of the world's remaining rainforests. Wait, that might not be right. Maybe the Amazon is about a third? Hmm. Alternatively, maybe the Congo Basin is second, but the Amazon is definitely larger. If the total remaining tropical rainforest is, say, 10 million square km, then the Amazon being 5.5 million would be more than half. But maybe the total is higher. Let me think. According to some sources, tropical rainforests cover about 17-20 million square km globally. The Amazon is around 5.5 million, so that's roughly a third. Wait, but the question says a quarter. Maybe the answer is the Amazon, but the percentage is a quarter. Maybe the total remaining is higher. Alternatively, perhaps the user is referring to the Amazon as a quarter. Also, the part about ""comprises one of the largest and most biodiverse tracts."" The Amazon is definitely one of the most biodiverse. The Congo Basin is also biodiverse but maybe less so. Southeast Asian rainforests are biodiverse too but smaller in area. So putting it all together, the answer is likely the Amazon Rainforest. But I need to make sure about the percentage. If the total remaining tropical rainforests are approximately 20 million km², then the Amazon at 5.5 million would be about 27.5%, which is roughly a quarter. So that fits. Therefore, the answer is the Amazon Rainforest. **Final answer**: The **Amazon Rainforest** represents approximately a quarter of the total area of remaining tropical forests on Earth and is one of the largest and most biodiverse tracts of tropical rainforest in the world. - **Size**: Spanning around 5.5 million square kilometers (2.1 million square miles), it covers about 27% of the globe's remaining tropical forest area (estimated at 20 million km² total). - **Biodiversity**: It hosts 10% of known species, including iconic wildlife like jaguars, pink river dolphins, and countless plant and insect species. - **Geographic Reach**: It spans nine countries, with the majority in Brazil, and plays a critical role in global carbon storage and climate regulation. This immense ecosystem is often termed the ""lungs of the Earth"" due to its vital role in absorbing carbon dioxide and producing oxygen.

I ADDITIONAL SEAL-0 RESULTS

† indicates results using the model's built-in search; all other search-based results use FRESHPROMPT (Vu et al., 2024).

Table 14: SEAL-0 poses a significant challenge for a range of both closed and open-source models.

Model	Accuracy
Closed-source models	
GROK 4	20.7†
GEMINI 2.5 PRO	19.8†
O3-PRO	18.9†
O3	15.3†
GEMINI 2.5 FLASH	13.5†
O4-MINI	6.3†
GROK 3	5.4†
GEMINI 2.5 FLASH-LITE	2.7†
O3-MINI	2.7
GROK 3 MINI	2.7†
GPT-4.1	0.0†
GPT-4.1 MINI	0.0†
Open-weight models	
QWEN3-235B-A22B	5.4
DEEPSEEK-R1-671B	4.5
LLAMA-4-SCOUT-17B-16E (109B)	0.0