FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation

Tu Vu[1] **Mohit Iyyer**[2] **Xuezhi Wang**[1] **Noah Constant**[1] **Jerry Wei**[1]
Jason Wei[3*] **Chris Tar**[1] **Yun-Hsuan Sung**[1] **Denny Zhou**[1] **Quoc Le**[1] **Thang Luong**[1]

Google[1] University of Massachusetts Amherst[2] OpenAI[3]
freshllms@google.com

Abstract

Since most large language models (LLMs) are trained once and never updated, they struggle to dynamically adapt to our ever-changing world. In this work, we present FreshQA, a *dynamic* QA benchmark that tests a model's ability to answer questions that may require reasoning over up-to-date world knowledge. We develop a two-mode human evaluation procedure to measure both correctness and hallucination, which we use to benchmark both closed and open-source LLMs by collecting >50K human judgments. We observe that all LLMs struggle to answer questions that require *fast-changing* world knowledge as well as questions with *false premises* that need to be debunked. In response, we develop FreshPrompt, a few-shot prompting method that curates and organizes relevant information from a search engine into an LLM's prompt. Our experiments show that FreshPrompt outperforms both competing search engine-augmented prompting methods such as Self-Ask (Press et al., 2022) as well as commercial systems such as Perplexity.AI. To facilitate future work, we additionally develop FreshEval, a reliable autorater for quick evaluation and comparison on FreshQA. Our latest results with FreshEval suggest that open-source LLMs such as Mixtral (Jiang et al., 2024), when combined with FreshPrompt, are competitive with closed-source and commercial systems on search-augmented QA.

1 Introduction

Despite their impressive capabilities, modern LLMs often "hallucinate" plausible but factually incorrect information (Maynez et al., 2020; Liu et al., 2023), which reduces their trustworthiness especially in settings where accurate and up-to-date information is critical. This behavior can be partially attributed to the presence of outdated knowledge encoded in their parameters. While additional training using human feedback (Ouyang et al., 2022) or knowledge-enhanced tasks can mitigate this issue, it is not easily scalable for real-time knowledge updates (e.g., stock prices). In-context learning (Brown et al., 2020) is an appealing alternative by which real-time knowledge can be injected into an LLM's prompt. While recent work has begun to explore augmenting LLM prompts with web search results (Lazaridou et al., 2022; Press et al., 2022), it is unclear how to take full advantage of search engine outputs to increase LLM factuality.

In this work, we collect FreshQA, a novel benchmark to evaluate the factuality of LLM generations. FreshQA consists of 600 natural questions that are broadly divided into the *four* main categories shown in Figure 1. FreshQA's questions span a diverse set of topics with diverse difficulty levels (requiring single-hop and multi-hop reasoning), and require a model to "understand" up-to-date world knowledge to be able to answer correctly. Additionally, FreshQA is *dynamic* in nature: some of the ground-truth answers may change over time, and a question classified under a specific category may undergo reclassification at some later point in time (e.g., the current *false-premise* question *"How long has Elon Musk been married to his current spouse?"* will fall into the *fast-changing* category if Elon Musk gets married again in the future).

We benchmark a diverse range of both closed and open-source LLMs under a two-mode evaluation procedure: Relaxed, which measures only whether the main answer is correct; and Strict, which measures whether all of the claims in the response are factual and up-to-date (i.e., no hallucination). Through an extensive human evaluation (> 50K judgements), we shed light on limitations of these models and demonstrate significant room for improvement: for example, all models (regardless of model size) struggle on questions that involve fast-changing knowledge and false premises.

[*] Work done while at Google.

Findings of the Association for Computational Linguistics: ACL 2024, pages 13697–13720
August 11-16, 2024

Type	Question	Answer (as of this writing)
never-changing	*Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?*	***Yes**, Virginia Woolf's 1927 novel To the Lighthouse entered the public domain in 2023.*
never-changing	*What breed of dog was Queen Elizabeth II of England famous for keeping?*	***Pembroke Welsh Corgi** dogs.*
slow-changing	*How many vehicle models does Tesla offer?*	*Tesla offers **six** vehicle models: Model S, Model X, Model 3, Model Y, Tesla Semi, and Cybertruck.*
slow-changing	*Which team holds the record for largest deficit overcome to win an NFL game?*	*The record for the largest NFL comeback is held by the **Minnesota Vikings**.*
fast-changing	*Which game won the Spiel des Jahres award most recently?*	***Dorfromantik** won the 2023 Spiel des Jahres.*
fast-changing	*What is Brad Pitt's most recent movie as an actor*	*Brad Pitt is credited as Keith in **IF**.*
false-premise	*What was the text of Donald Trump's first tweet in 2022, made after his unbanning from Twitter by Elon Musk?*	*He **did not tweet** in 2022.*
false-premise	*In which round did Novak Djokovic lose at the 2022 Australian Open?*	*He **was not allowed to play** at the tournament due to his vaccination status.*

Figure 1: FRESHQA exemplars. Our questions are broadly divided into *four* main categories: *never-changing*, in which the answer almost never changes; *slow-changing*, in which the answer typically changes over the course of several years; *fast-changing*, in which the answer typically changes within a year or less; and *false-premise*, which includes questions whose premises are factually incorrect and thus have to be rebutted.

Motivated by these findings, we develop FRESHPROMPT, a few-shot prompting strategy that takes full advantage of a search engine by integrating up-to-date and relevant information into the prompt, including knowledge from related questions asked by other search users. FRESHPROMPT significantly boosts LLM factuality: for example, our best GPT-4 + FRESHPROMPT variant yields an improvement of 32.6% and 49.0% accuracy over the vanilla GPT-4 (OpenAI, 2023) on FRESHQA under RELAXED and STRICT, respectively. Further analysis of FRESHPROMPT reveals that both the number of retrieved evidences and their order are key factors behind the correctness of LLM-generated answers.

We make FRESHQA freely available at github.com/freshllms/freshqa and commit to updating the ground-truth answers *weekly* to encourage exploration of methods to improve LLM factuality. To facilitate future work, we develop FRESHEVAL, an LLM-based autorater that reliably replicates human judgments of model responses (> 96% average agreement with human annotators). Intriguingly, our latest evaluation with FRESHEVAL reveals that open-source LLMs such as MIXTRAL, when paired with FRESHPROMPT, rival the performance of closed-source and commercial systems on search-augmented QA (MIXTRAL 8X7B + FRESHPROMPT obtains a 11% absolute accuracy improvement over PERPLEXITY.AI's ONLINE LLM 70B), highlighting the potential of accessible AI solutions.

2 FreshQA

To address LLM factuality assessment, we build FRESHQA, a *dynamic* benchmark with 600 questions covering diverse question and answer types. We collected FRESHQA by recruiting both NLP researchers[1] and online freelancers[2] to write questions of varying difficulty and topics (arts, music, politics, government, religion, science and technology, environment, transportation, sports, etc.), focusing on questions whose answers evolve over time. Annotators were shown exemplars of the four broad question types (see Figure 1) and asked to write questions at two difficulty levels: *one-hop*, where the question directly states all information needed for the answer (e.g., *"Who is the CEO of Tesla"*); and *multi-hop*, where the question requires additional reasoning steps to find the answer (e.g., *"What country does the Wimbledon women's champion play for?"*). Annotators were encouraged to write questions that involve *fresh* knowledge and appear *natural* as search engine queries. For false-premise questions, we requested a brief explanation elucidating why the question is flawed.[3]

Quality control: Upon obtaining the initial dataset, we performed rigorous data cleaning and quality checks. This included manual review for

[1] including the authors and their colleagues

[2] We use UPWORK (https://www.upwork.com) with a compensation rate of $2 per example.

[3] Additionally, annotators were asked to include the year the answer last changed and a supporting URL.

well-formed questions, removal of duplicates and invalid questions (e.g., too easy or controversial), and verification of answers and supporting URLs.[4]

Data splits: FRESHQA is divided into a *test* set with 500 examples (125 per question type), a *development* set with 100 examples (25 per question type), and a 15-example *demonstration* set for few-shot learning. The development set is reserved for future use and was not used in this paper.[5]

2.1 Evaluation

Model responses were evaluated by two authors in a two-mode evaluation procedure: RELAXED, which assesses the correctness of the main answer; and STRICT, which additionally examines whether *all* of the facts in the answer are accurate (i.e., no hallucination). This approach provides both ends of the spectrum for evaluating factuality with the difference between a model's STRICT and RELAXED performance indicating the degree of hallucination. In both modes, the primary answer must be correct and either definitively stated or obviously inferable. Any additional information must not contradict the primary answer. For false-premise questions, the model must identify false premises to get credit. Our protocol also considers many other edge cases (e.g., approximate numbers, ungrammatical answers), which are fully detailed in Appendix A. Here, we provide an illustrative example:

> **Question:** Who won the biggest single-tournament payday in tennis history?
> **Gold answer:** Novak Djokovic
> **Model response:** The biggest single-tournament payday in tennis history was won by Novak Djokovic at the 2021 Australian Open.

The model receives credit under RELAXED for a correct primary answer; however, the tournament in question was actually the 2022 ATP Finals, so the answer is judged as incorrect under STRICT.

Inter-rater agreement: Two authors independently evaluated a subset of 100 model responses in both modes and had an agreement of 99% for RELAXED and 96% for STRICT, validating the reliability of our evaluation protocol.

[4]For each question, we also manually collected additional valid answers (e.g., different names of the same person) and included the expected next review date. To facilitate future answer updates, we excluded questions whose answers are likely to change more frequently than once per week.

[5]Although our data splits are initially balanced across question types, the distribution may change over time due to reclassification of questions from one category to another.

FRESHEVAL: Human evaluation of LLM-generated answers can be extremely time-consuming, especially for long responses in the STRICT setting. To facilitate future evaluation and comparison, we develop FRESHEVAL, a simple autorater that uses few-shot in-context learning to teach an LLM to evaluate correctness and hallucination in LLM-generated responses given the questions and their valid answers, achieving an average agreement of 96.5% with human judgments for RELAXED and 96% for STRICT. We use FRESHEVAL in Section 5 to evaluate the latest LLMs; all other results in this paper are from human judgments. See Appendix B for more details about FRESHEVAL.

3 Offline LLMs struggle on FRESHQA

We first use FRESHQA to benchmark LLMs without access to real-time data or web browsing capabilities.[6] We simply feed individual questions into each model and use greedy decoding. While all LLMs (regardless of size) predictably struggle on questions requiring up-to-date knowledge, they also underperform on false-premise questions.

Baselines: We evaluate a series of models: **T5** (Raffel et al., 2020; Lester et al., 2021), **PALM** and **PALMCHILLA** (Chowdhery et al., 2022) (varying in size from 770M to 540B parameters), optionally using **FEW-SHOT** prompting (Brown et al., 2020) and Chain-of-Thought (**COT**) prompting (Wei et al., 2022); **FLAN-T5** and **FLAN-PALM** (Chung et al., 2022; Longpre et al., 2023), **GPT-3.5** (Ouyang et al., 2022), **CODEX** (Chen et al., 2021a), **CHATGPT**, and **GPT-4**.[7] See Appendix C for more details.

3.1 Results and Discussion

FRESHQA presents a challenge for LLMs: Figure 2 shows the accuracy of different LLMs on FRESHQA (see Appendix E for concrete numbers). All models struggle on FRESHQA, with overall accuracy ranging from 0.8% to 32.0% under STRICT, and

[6]Note that even without access to up-to-date information, a model can still give accurate answers to some current questions by making random guesses or using past valid responses. For example, for the question *"Which drama series won the most recent Primetime Emmy Award for Outstanding Drama Series?"*, a model trained in 2020 might correctly answer *"Succession"* (as of this writing) since it won in 2020 and the two most recent years.

[7]All models were benchmarked in 2023. Both CHATGPT (built upon GPT-3.5) and GPT-4 had access to the current date. We note that CHATGPT now has browsing capabilities.

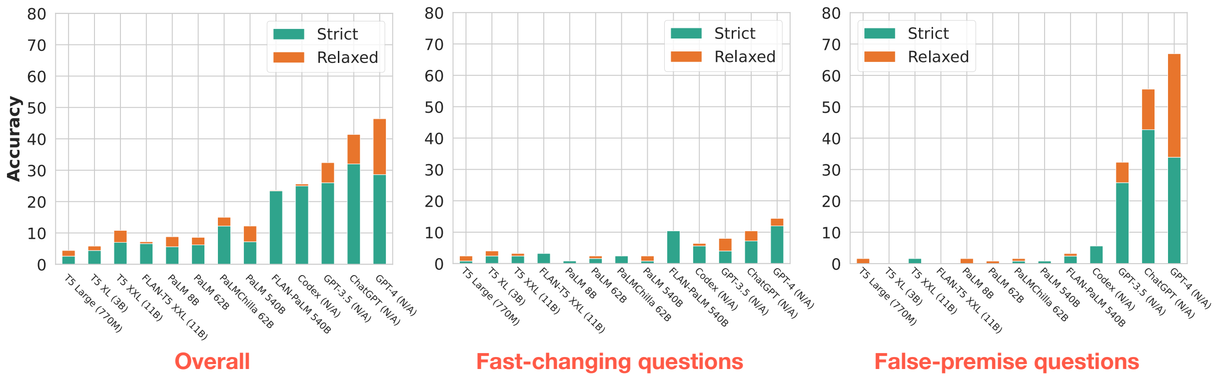

Figure 2: Accuracy of different LLMs on FRESHQA under RELAXED and STRICT (no hallucination) evaluations. *All* models (regardless of model size) fall short on *fast-changing* and *false-premise* questions.

0.8% to 46.4% under RELAXED. There is a marked decrease in accuracy for CHATGPT and GPT-4 when switching from RELAXED to STRICT due to the lack of access to up-to-date information, often resulting in "outdated" answers (starting with *"As of my knowledge cutoff date"*) or refusal to answer (e.g., *"I cannot provide real-time information"*). PALM's accuracy also drops significantly under STRICT, primarily due to ill-formed responses (conversation-like responses with unexpected end-of-turn `[eot]` tokens) and hallucination. In contrast, FLAN-PALM and CODEX show minimal hallucination due to their concise and direct answers.

LLMs struggle with questions involving up-to-date information and false premises: Outdated knowledge drastically lowers model accuracies on questions about fast-changing or recent knowledge. While GPT-4 generally obtains the highest accuracy on these questions, it never exceeds 15% in both evaluation modes. Our evaluation shows that CHATGPT and GPT-4 have been exposed to data containing information beyond their knowledge cutoff (see Appendix F). GPT-4 is more reluctant to answer fast-changing questions (refusing to answer 60% of the time) compared to CHATGPT (16%).

Questions with false premises also pose a hurdle for LLMs. Larger models do not improve accuracy for T5 and PALM ("flat scaling"), with performance ranging from 0.0% to 1.6%. However, GPT-3.5, CHATGPT, and GPT-4 outperform other models significantly, achieving accuracies between 25.8% to 42.7% under STRICT and 32.3% to 66.9% under RELAXED. Our findings suggest that these models may have been trained to handle false-premise queries.

COT increases hallucination and multi-hop reasoning is challenging for LLMs: Overall, FEW-SHOT and COT prompting confer benefits to large and moderately-sized models on questions with valid premises about never-changing or old knowledge. Under STRICT, FEW-SHOT and COT yields +36.1% and +26.9% accuracy improvement, respectively, over zero-shot prompting with PALM 540B on questions about pre-2022 knowledge (+21.9% and +29.7% under RELAXED). COT generally outperforms FEW-SHOT under RELAXED, whereas FEW-SHOT performs better under STRICT, as COT introduces more room for hallucination.

Most LLMs have difficulty with multi-hop questions. T5 LARGE and XL are incapable of dealing with this type of questions, while FLAN-PALM 540B, CODEX, and GPT-3.5 suffer the most when switching from one-hop to multi-hop questions. GPT-4, on the other hand, maintains stability across both types of questions, with less than a 2% difference in accuracy. See Appendix E for details.

4 Prompting Search Engine-Augmented Language Models

The low accuracies in Section 3 are expected, as none of the evaluated models had access to real-time information. Here, we evaluate the impact of *search engine augmentation* to LLMs on FRESHQA. We present FRESHPROMPT, a simple few-shot prompting method that substantially boosts LLM factuality by incorporating up-to-date information from a search engine into the prompt.

4.1 FRESHPROMPT

FRESHPROMPT uses a text prompt to (1) incorporate contextually relevant and current information (including answers to relevant questions) from a search engine into a model, and (2) teach the model to reason over retrieved evidences. More specifi-

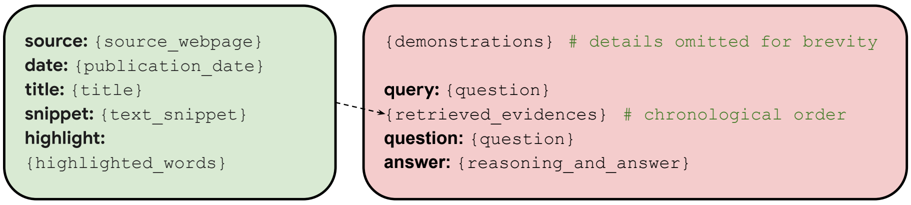

Figure 3: FRESHPROMPT's format. We standardize all retrieved evidences into a unified format with useful information: source webpage, date, title, text snippet, and highlighted words (left). The prompt begins with few-shot demonstrations, each presenting an example question along with a list of retrieved evidences, followed by reasoning to determine the most relevant and current answer (right).

cally, given a *question* $\boldsymbol{q}$, we first use $\boldsymbol{q}$ verbatim to query a search engine, i.e., GOOGLE SEARCH[8], and retrieve all search results, including the *answer box*, *organic results*, and other useful information, such as the *knowledge graph*, *questions and answers* from crowdsourced QA platforms, and *related questions* from search users (see Figure 9 in Appendix G). For each result, we extract the associated *text snippet* $\boldsymbol{x}$ along with details, such as *source* $\boldsymbol{s}$ (e.g., WIKIPEDIA), *date* $\boldsymbol{d}$, *title* $\boldsymbol{t}$, *highlighted words* $\boldsymbol{h}$. These snippets are standardized and then organized into a list of $\boldsymbol{k}$ retrieved evidences $\boldsymbol{E} = \{(\boldsymbol{s}, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{h})\}$ (Figure 3, left). To prioritize recent evidences, we arrange the evidences $\boldsymbol{E}$ in the prompt from oldest to newest.

To guide the model in learning the task, we provide a few input-output exemplars at the start of the prompt. Each demonstration includes an example question and a list of retrieved evidences, followed by a chain-of-thought reasoning to derive the most relevant and current answer (Figure 3, right). While we include some examples with false premises, we also test an explicit false premise check in the prompt: *"Please check if the question contains a valid premise before answering"*. Figure 10 in Appendix H shows a realistic prompt.

4.2 Experiment setup

We closely follow the setup in Section 3, except for cases where we lack control over the model's decoding via an API (e.g., PERPLEXITY.AI).[9] In addition to GPT-3.5 and GPT-4, we evaluate **GOOGLE SEARCH**[10]; **PERPLEXITY.AI (PPLX.AI)**, which combines an LLM and a search engine to respond to users' queries;[11] and **SELF-ASK** (Press et al., 2022), which uses few-shot in-context learning to teach an LLM to decompose a question into simpler sub-questions that are answered via GOOGLE SEARCH.[12]

FRESHPROMPT setup: We employ FRESHPROMPT for both GPT-3.5 and GPT-4, sequentially adding the following retrieved evidences to the input prompt: $\boldsymbol{o}$ organic search results, $\boldsymbol{r}$ related questions from search users, $\boldsymbol{a}$ questions and answers from crowdsourced platforms, and the snippets from the knowledge graph and answer box (if available). Due to context limits, we retain the top $\boldsymbol{n}$ evidences (closest to the end of the prompt) sorted by date. Default values are $(\boldsymbol{o}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{n}, \boldsymbol{m}) = (10, 2, 2, 5)$ for GPT-3.5, and $(\boldsymbol{o}, \boldsymbol{r}, \boldsymbol{a}, \boldsymbol{n}, \boldsymbol{m}) = (10, 3, 3, 10)$ for GPT-4. Additionally, we include $\boldsymbol{m} = 5$ question-answer demonstrations at the start of the prompt.

4.3 Results and Discussion

FRESHPROMPT significantly boosts FRESHQA accuracy: Table 1 shows our results under STRICT (see Appendix I for RELAXED). FRESHPROMPT offers large improvements over vanilla GPT-3.5 and GPT-4 across the board. GPT-4 + FRESHPROMPT achieves absolute accuracy improvements of 47% and 31.4% over GPT-4 under STRICT and RELAXED, respectively. The absolute accuracy gap between STRICT and RELAXED diminishes substantially with FRESHPROMPT (from 17.8% to 2.2%), indicating a significant reduction in outdated and hallucinated answers. The most significant improvements for both models occur in categories related to recent knowledge, including fast-changing and slow-changing questions. However, even questions pertaining to older

[8]We scrape the results from GOOGLE SEARCH using SERPAPI (https://serpapi.com).

[9]All models were benchmarked in 2023. We note that some of the evaluated models may have evolved, posing a challenge to result reproducibility.

[10]We simply query GOOGLE SEARCH and use the answer in the answer box (if any) or the text snippet of the top-1 search result.

[11]https://www.perplexity.ai. At the time of evaluation, PPLX.AI was a combination of GPT-3.5 and BING SEARCH, and was able to provide both concise and detailed answers. We evaluated its concise answers.

[12]We used the few-shot prompt provided by SELF-ASK's authors and applied it to both GPT-3.5 and GPT-4. For simplicity, we evaluated solely the final answer from SELF-ASK, disregarding intermediate answers.

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	$<$ 2022	$\geq$ 2022	1-hop	m-hop	all	$<$ 2022
comparison against baselines												
GOOGLE SEARCH (N/A)	UTD	39.6	48.9	32.0	46.4	68.3	67.4	37.9	55.6	32.4	11.3	9.7
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + SELF-ASK (N/A)	UTD	41.6	51.1	36.8	43.2	73.0	73.8	37.4	52.2	48.1	12.9	17.2
GPT-3.5 + FRESHPROMPT	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
PPLX.AI (N/A)	UTD	52.2	57.2	38.4	53.6	79.4	73.0	47.7	63.8	40.7	37.1	38.7
GPT-4 (N/A)	2021$^+$	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 + SELF-ASK (N/A)	UTD	47.8	47.1	39.2	46.4	55.6	51.8	44.3	43.7	55.6	50.0	61.3
GPT-4 + FRESHPROMPT	UTD	**75.6**	**77.1**	**59.2**	**77.6**	**94.4**	**88.7**	**70.2**	**81.3**	**66.7**	**71.0**	**77.4**
sensitivity and ablation studies												
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + FRESHPROMPT	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
w/ PREMISE CHECK	UTD	35.2	27.1	14.4	28.0	38.9	36.2	21.7	31.0	17.6	59.7	67.7
GPT-4 (N/A)	2021$^+$	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 w/ SNIPPETS ONLY & SEARCH ORDER	UTD	74.0	75.5	56.8	75.2	94.4	87.9	68.1	79.9	64.8	69.4	77.4
GPT-4 w/ SNIPPETS ONLY & TIME ORDER	UTD	74.8	75.5	58.4	74.4	93.7	87.9	68.1	79.9	64.8	72.6	**82.8**
GPT-4 w/ SNIPPETS ONLY & RANDOM ORDER	UTD	72.4	73.7	56.8	69.6	94.4	87.9	65.1	78.4	62.0	68.5	76.3
GPT-4 + FRESHPROMPT	UTD	75.6	77.1	59.2	77.6	94.4	**88.7**	70.2	81.3	66.7	71.0	77.4
w/ PREMISE CHECK	UTD	75.0	74.2	56.8	76.0	89.7	85.1	67.7	79.5	61.1	**77.4**	79.6
w/o ANSWER BOX	UTD	74.2	74.7	57.6	74.4	92.1	**88.7**	66.4	79.1	63.9	72.6	78.5
w/o ANSWER BOX & RELEVANT INFO	UTD	72.4	72.9	54.4	71.2	92.9	87.2	64.3	78.0	60.2	71.0	78.5
w/ 1 EVIDENCE	UTD	61.4	60.9	40.0	55.2	87.3	79.4	49.8	66.8	46.3	62.9	75.3
w/ 5 EVIDENCES	UTD	70.6	72.1	56.0	69.6	90.5	81.6	66.4	78.0	57.4	66.1	73.1
w/ 15 EVIDENCES	UTD	**77.6**	**78.5**	**60.8**	**78.4**	**96.0**	**88.7**	**72.3**	**81.7**	**70.4**	75.0	80.6
w/ 15 DEMONSTRATIONS	UTD	74.6	75.5	56.8	76.0	93.7	87.9	68.1	79.9	64.8	71.8	76.3
w/ LONG DEMONSTRATION ANSWERS	UTD	73.0	72.6	55.2	71.2	91.3	83.7	66.0	77.6	60.2	74.2	81.7

Table 1: Accuracy of different search engine-augmented LLMs on FRESHQA under STRICT (no hallucination) evaluation. Accuracy reported across various question categories: *fast-changing* (*fast*), *slow-changing* (*slow*), *never-changing* (*never*), false-premise, questions about pre-2022 knowledge ($<$ 2022) and post-2022 knowledge ($\geq$ 2022), one-hop (*1-hop*) and multi-hop (*m-hop*) questions. $^+$ indicates a model with access to the current date. UTD stands for "up-to-date".

knowledge benefit from FRESHPROMPT (+30.5% and +9.9% improvements for GPT-4 under STRICT and RELAXED, respectively, on questions with valid premises about pre-2022 knowledge). Furthermore, FRESHPROMPT yields notable accuracy gains on false-premise questions (+37.1% and +8.1% respective improvements under STRICT and RELAXED).

FRESHPROMPT outperforms other search-augmented methods by a large margin: GPT-4 + FRESHPROMPT surpasses all other methods by a substantial margin, with its best variant (15 retrieved evidences per question) achieving 77.6% and 79.0% overall accuracies under STRICT and RELAXED, respectively. Compared to PPLX.AI and SELF-ASK (all built on top of GPT-3.5), GPT-3.5 + FRESHPROMPT demonstrates a respective increase of +3.8% and +14.4% in overall accuracy under STRICT. However, under RELAXED, PPLX.AI outperforms GPT-3.5 + FRESHPROMPT by +4.2%, mainly due to its higher accuracy on false-premise questions (58.1% vs. 41.1%). The significant 14.0% accuracy gap between STRICT and RELAXED for PPLX.AI indicates a considerable amount of hallucination in its outputs. Overall, all search-engine augmented approaches (SELF-ASK, PPLX.AI, and FRESHPROMPT) provide substantial improvements over vanilla GPT-3.5 and GPT-4.

Premise check improves accuracy on false-premise questions but can hurt accuracy on valid premise questions: Our findings indicate that GPT-3.5, GPT-4, and PPLX.AI, are likely trained to address false-premise queries. Additionally, we empirically find that several LLMs can debunk false-premise questions if explicitly prompted: *"Please check if the question contains a valid premise before answering"*, resulting in significant accuracy improvements on false-premise questions. For example, adding this premise check boosts accuracy by +23.4% and +6.4% for GPT-3.5 and GPT-4, respectively, under STRICT (+22.6% and +11.3% under RELAXED). However, this is detrimental for GPT-3.5 across other question types, reducing overall accuracy by 20.8% and 21% under STRICT and RELAXED, respectively. Conversely, GPT-4 experiences minimal impact, with only a 0.6% decrease under STRICT and a 1.2% increase under RELAXED.

Including more relevant and up-to-date evidences at the end of the input context is bene-

Model	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2024	≥ 2024	1-hop	m-hop	all	< 2024
without access to a search engine												
GPT-4	2023/04$^+$	63.0	61.2	26.0	65.6	92.7	65.7	30.6	64.6	51.0	68.5	70.2
MIXTRAL (8X7B)	2023/12$^+$	39.2	37.0	17.3	31.2	62.9	40.7	12.2	37.1	36.5	46.0	47.1
LLAMA-2 (70B)	2023/07$^+$	36.8	35.6	12.6	32.0	62.9	39.8	8.2	40.7	20.8	40.3	41.3
with access to a search engine												
GPT-4 + FRESHPROMPT	Online	80.6	80.9	67.7	79.2	96.0	84.1	59.2	85.7	66.7	79.8	81.0
MIXTRAL (8X7B) + FRESHPROMPT	Online	73.8	73.4	59.1	70.4	91.1	76.5	53.1	78.6	58.3	75.0	76.0
LLAMA-2 (70B) + FRESHPROMPT	Online	53.8	46.3	25.2	47.2	66.9	51.4	12.2	53.2	26.0	76.6	76.9
PPLX.AI'S ONLINE LLM (70B)	Online	62.8	72.6	56.7	75.2	86.3	76.8	44.9	78.6	55.2	33.1	33.1
YOU.COM'S WEB LLM	Online	53.4	56.9	27.6	60.0	83.9	64.5	6.1	63.9	36.5	42.7	43.8

Table 2: Accuracy of recent LLMs on FRESHQA under RELAXED evaluations with FRESHEVAL as of February 2024. We use greedy decoding for all models. For FRESHPROMPT, we use 5 retrieved evidences per question for LLAMA (due to a maximum context length of 4097 tokens) and 15 for other models.

ficial: We also analyze how the order of the evidences in the prompt impacts GPT-4's accuracy. Our results show that using the order returned by GOOGLE SEARCH (SEARCH ORDER, top search results at the end of the input context) or sorting the evidences by their associated date information (TIME ORDER, more recent results at the end) generally results in better accuracy compared to using a random order (RANDOM ORDER), with up to a +2.2% higher overall accuracy in both evaluation modes. However, using only the text snippet for each evidence without additional information (e.g., source, date) as in GPT-4 + FRESHPROMPT slightly decreases accuracy, with less than 1% in both evaluation modes.

Additional retrieved information beyond organic search results provides further gains: Incorporating additional retrieved evidences beyond *organic search results*, such as *answer boxes* and *related questions* from search users, enhances performance. Removing *answer boxes* decreases GPT-4 + FRESHPROMPT's overall accuracy by 1.4% under STRICT (1.6% under RELAXED). Removing both *answer boxes* and other relevant information (including *related questions*) reduces GPT-4 + FRESHPROMPT's overall accuracy by 3.2% under STRICT (3.0% under RELAXED).

Increasing the number of retrieved evidences enhances FRESHPROMPT's effectiveness: We explore the effect of the number of retrieved evidences per question as well as the number of demonstrations by varying these numbers for GPT-4 + FRESHPROMPT. By default, we use 10 retrieved evidences for each question and 5 demonstrations. Our results reveal that the number of retrieved evidences per question is the most important ingredient for achieving highest accuracy. Under STRICT, increasing this number from 1 to 5, 10, and 15 results in respective accuracy improvements of of +9.2%, +14.2%, and +16.2%. This indicates GPT-4's adeptness in accommodating an increasing number of retrieved evidences, including conflicting answers, to provide responses grounded in the most factual and current information. Conversely, increasing the number of demonstrations from 5 to 15 marginally decreases accuracy in both evaluation modes, with a 1% overall decrease under STRICT.

Verbose demonstrations help with complex questions but also amplify hallucinations: To evaluate the effect of the writing style of the answer (including the reasoning) in each demonstration, we manually rewrite these answers into a more verbose version (LONG DEMONSTRATION ANSWERS). Our manual inspection reveals that while verbose answers may aid in tackling intricate questions, they can also be detrimental by allowing for hallucination, leading to a 2.6% decrease in overall accuracy under STRICT.

5 An updated evaluation in 2024

While our previous experiments were conducted in 2023, we use FRESHEVAL to evaluate new LLMs on an updated version of FRESHQA (February 5, 2024). We evaluate GPT-4, MIXTRAL 8X7B, LLAMA-2 (Touvron et al., 2021), PPLX.AI's ONLINE LLM, and YOU.COM's WEB LLM on the same date of February 5, 2024. See Appendix D for more details.

As shown in Table 2, even recent LLMs, with or without search engine augmentation, struggle with questions requiring up-to-date knowledge (with accuracies ranging from 6.1% to 44.9% on questions about post-2024 knowledge) as well as questions with false premises (accuracies ranging from 33.1% to 68.5%). Strikingly, we discover that open-source LLMs such as MIXTRAL 8X7B, when paired with FRESHPROMPT, are competitive with closed-source and commercial systems on FRESHQA. For example, MIXTRAL 8X7B + FRESHPROMPT obtains a 11% absolute accuracy improvement over PPLX.AI's ONLINE LLM 70B. FRESHPROMPT narrows the performance gap between GPT-4 and MIXTRAL 8X7B to under 9% in all question categories, with the largest gap on slow-changing questions decreasing from 34.4% to 8.8%.

6 Related Work

Knowledge augmented LLMs: Many prior works study semi-parametric knowledge augmentation in LLMs via additional fine-tuning (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022), while others advocate for knowledge generation instead of retrieval (Yu et al., 2023a; Sun et al., 2023). FRESHPROMPT aligns with the recent trend of retrieval-augmented generation (Nakano et al., 2021; Lazaridou et al., 2022; Menick et al., 2022; Yao et al., 2022; Press et al., 2022; Khattab et al., 2022; Schick et al., 2023; Luo et al., 2023). Similar to our method, Lazaridou et al. (2022) employ a few-shot prompting approach that inserts documents from GOOGLE SEARCH into LLM prompts. We refrain from comparing to their method due to its expensive inference cost, as they chunk retrieved documents into evidence paragraphs and perform $k = 50$ inference calls to the LLM to generate k answers followed by LLM reranking. In contrast, FRESHPROMPT only performs a single inference call to the LLM. SELF-ASK (Press et al., 2022) also uses few-shot in-context learning to teach an LLM to ask itself follow-up questions before answering the initial question, although it focuses more on decomposition.

Time-sensitive QA: FRESHQA fits into a growing body of work benchmarking LLMs' temporal reasoning capabilities (Chen et al., 2021b; Zhang and Choi, 2021; Liska et al., 2022; Kasai et al., 2022). Chen et al. (2021b) created TIMEQA by extracting evolving facts from WIKIDATA and then synthesizing timestamped question-answer pairs. Zhang and Choi (2021) constructed SITUATEDQA by annotating realistic questions from existing open-domain QA datasets with temporal context (i.e., timestamps). STREAMINGQA (Liska et al., 2022) consists of both LLM-generated and human-written questions, all answerable using a corpus of timestamped news articles. Also related is the dynamic REALTIMEQA benchmark (Kasai et al., 2022), which evaluates models weekly on multiple-choice questions about new events extracted from news websites. In contrast, FRESHQA contains a fixed set of human-written open-ended questions whose answers by nature can change based on new developments in the world and thus offers a complementary generative evaluation of time-sensitive QA.

QA over questionable or counterfactual premises: Recent work has also introduced QA benchmarks with questionable premises (Yu et al., 2023c; Kim et al., 2023) or counterfactual premises (Yu et al., 2023b). CREPE (Yu et al., 2023c) includes Reddit questions with false premises annotated by human workers. Kim et al. (2023) constructed $(QA)^2$ using frequently searched queries annotated by expert annotators and crowdworkers, distinguishing between those with and without questionable premises. Consistent with these efforts, we find that current LLMs struggle with false premise questions; additionally, several LLMs can debunk a false-premise question if explicitly asked to check for the premise's validity. Similar to above, these benchmarks are complementary and combining them is a promising direction for future work.

7 Conclusion

Our work offers a fine-grained and exhaustive evaluation of the capabilities of modern LLMs to adapt to ever-changing world knowledge with and without search engine augmentation. In the process, we develop a new dataset—FRESHQA—of 600 questions that test a broad range of reasoning abilities, from the incorporation of fast-changing knowledge to identification of questions with false premises. Our two-mode evaluation also provides a way to measure both correctness and hallucination. Additionally, we propose a simple few-shot in-context learning algorithm called FRESHPROMPT that incorporates relevant evidences retrieved from a search engine into an LLM's prompt. FRESHPROMPT significantly improves performance over competing search engine-augmented approaches on FRESHQA, and an ablation reveals that factors such as the number of incorporated evidences and their order impact the correctness of LLM-generated answers. We release FRESHQA and commit to updating its answers regularly to facilitate future research. Additionally, we develop FRESHEVAL, a reliable autorater for quick evaluation and comparison on FRESHQA.

8 Limitations and Future Work

One obvious challenge with FRESHQA is the need for regular answer updating by the maintainers; in the interim period between updates, the answers to some questions might become stale. This could be addressed by support from the open-source community. On the method side, FRESHPROMPT only performs one search query per question, and thus it can be further improved via question decomposition and multiple search queries (Khattab et al., 2022). Since FRESHQA consists of relatively simple questions, it is also unclear how well FRESHPROMPT performs in the context of long-form QA (Fan et al., 2019). Finally, FRESHPROMPT relies on in-context learning and thus may underperform approaches that fine-tune the base LLM on new knowledge.

9 Acknowledgements

We thank Colin Raffel, Hamed Zamani, and Subhransu Maji for helpful discussion and feedback. We would also like to thank Chengrun Yang, Xinyun Chen for their insightful comments on this manuscript. Finally, we are grateful to the following people for their contributions to creating our FRESHQA dataset: Marzena Karpinska, Dustin Tran, Daniel Cer, Sam Fullerton, Elizabeth Clark, Nishant Raj, Xiaoyu Song, Yapei Chang, Yixiao Song, Nader Akoury, Ankita Gupta, Bill Ray, Chau Pham, Wenlong Zhao, Maximilian Mozes, Simeng Sun, Ronan Salz, Kalpesh Krishna, Katherine Thai, Kanishka Misra, Salaheddin Alzu'bi, Erica Cai, Thibault Sellam, Jiao Sun, Dhruv Agarwal, Tessa Masis, Andrew Drozdov, Brian Lester, George Wei, Naveen Jafer Nizar, Shufan Wang, Youngwoo Kim, and Shib Sankar Dasgupta. This project was partially supported by award IIS-2046248 from the National Science Foundation (NSF), as well as NSF's CLOUDBANK program.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research (PMLR)*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021b. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, volume 1.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3558–3567.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research (PMLR)*, pages 3929–3938. PMLR.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.0408*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8466–8487.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

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. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research (PMLR)*, pages 13604–13622. PMLR.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.0877*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2021. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023b. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv preprint arXiv:2305.14010*.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023c. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10457–10480.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7387.

Appendix

A Evaluation protocol

Figure 4 shows specific examples of each evaluation criteria.

B Inter-rater agreement and automatic evaluation

Two authors independently evaluated a randomly sampled subset of 100 answers across models (including 50 questions with valid premises and 50 questions with false premises) in both modes RELAXED and STRICT.

To facilitate future evaluations, we also develop FRESHEVAL, an autorater that uses few-shot in-context learning to teach an LLM to judge model responses. In each evaluation, the model is conditioned on a given question, a list of valid answers, and a model response, and is then expected to generate a comment on the correctness of the response, followed by a final judgement. At the start of each input prompt, we also provide an instruction of the evaluation task, and sample comments and evaluations of the examples in Figure 4 as demonstrations.[13] See Figure 5 and Figure 6 for FRESHEVAL's prompts for RELAXED and STRICT evaluations, and Figure 7 for FRESHEVAL's sample output for STRICT evaluation.

Table 3 reports the inter-rater agreement between the two human raters, and between FRESHEVAL and each human rater, in terms of exact accuracy. The two human raters had an agreement of 99% for RELAXED and 96% for STRICT, while FRESHEVAL achieved an average agreement of 96.5% with human evaluations for RELAXED and 96% for STRICT. Overall, the high accuracies demonstrate that our evaluation protocol is reproducible and reliable, and FRESHEVAL can be used in place of human evaluation on FRESHQA.

C Additional experiment setup details for Section 3

To increase reproducibility, we used greedy decoding (with a temperature of 0), which selects the most likely token at every decoding timestep, and a maximum number of 256 generated tokens for all models. Note that the API for some models is non-deterministic by default, even with a temperature of 0.

For non-chat models that were not pre-trained with a QA task, we fed them a text prompt of the format: "Q: <question>\nA: " ("\ n" is the new line character). For few-shot in-context learning with T5 and PALM, we used the 5-shot demonstrations for TRIVIAQA (Joshi et al., 2017) used in Sun et al. (2023).

For OPENAI models, we used the `2023-03-15-preview` API in AZURE OPENAI SERVICE with the model names `text-davinci-003`, `code-davinci-002`, `gpt-3.5-turbo`, and `gpt-4` for GPT-3.5, CODEX, CHATGPT, and GPT-4, respectively.

D Additional experiment setup details for Section 5

For the evaluated models, we used OPENAI API for GPT-4 (`gpt-4-0125-preview`), TOGETHER API for MIXTRAL (`Mixtral-8x7B-Instruct-v0.1`) and LLAMA-2 (`Llama-2-70b-chat-hf`), PPLX-API for PPLX.AI's ONLINE LLM (`pplx-70b-online`), and YOU API for YOU.COM's WEB LLM.

For the base LLM of FRESHEVAL, we used GPT-4 (`gpt-4-1106-preview`).

E Additional experiment results for Section 3

Table 4 and Table 5 show the accuracy of different LLMs on FRESHQA under STRICT (no hallucination) and RELAXED evaluations, respectively.

F CHATGPT/GPT-4's awareness of recent knowledge

Our manual evaluation suggests that CHATGPT and GPT-4 have been exposed to data containing information beyond their knowledge cutoff date (see Figure 8).

G GOOGLE SEARCH results

Figure 9 shows different types of search results from GOOGLE SEARCH for given a query.

H A realistic prompt for FreshPrompt

Figure 10 displays a realistic prompt for FRESHPROMPT.

[13] In our experiments, we found that using separate prompts for RELAXED and STRICT evaluations resulted in better performance compared to using a single, combined prompt for both evaluation modes. We also found that additionally incorporating retrieved evidences for the question into the prompt did not improve inter-rater agreement between FRESHEVAL and human raters.

I Additional experiment results for Section 4

Table 6 presents the accuracy of different search engine-augmented LLMs on FRESHQA under RELAXED.

ID	Question	Type	Primary answer (as of this writing)	Model response	Relaxed	Strict	Explanation
1	*How old is the world's oldest verified living person?*	fast-changing	*116 years old*	*The oldest verified living person is Maria Branyas Morera, who was born on March 4, 1907, making her 116 years old.*	✓	✓	*the primary answer is correct, and all the provided information is accurate and up-to-date*
2	*When did the UK adopt the Euro?*	false-premise	*The UK has never adopted the Euro.*	*The UK has never adopted the Euro as its official currency. The country has retained the British pound sterling (GBP) as its currency.*	✓	✓	*the response debunks the question's false premise, and all the provided information is accurate and up-to-date*
3	*How much longer will Brittney Griner spend in the Russian prison?*	false-premise	*She was released in 2022.*	*I'm sorry, but I have no information to suggest that Brittney Griner is currently in a Russian prison.*	✗	✗	*a confident and definitive answer is required*
4	*Which language has the third largest population by number of native speakers?*	slow-changing	*English*	*1. Mandarin 2. Spanish **3. English***	✓	✗	*though the correct answer can be obviously inferred from the response, a response without an explicit statement is less desirable*
5	*Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?*	never-changing	*Yes*	***1927's To the Lighthouse is in the public domain in the United States.** The final answer is To the Lighthouse.*	✗	✗	*the final answer when standing alone must be correct*
6	*Is Tesla's stock price above $200 a share?*	fast-changing	*Yes*	*Tesla's stock price is above $200 a share. The stock is trading at a price of $120.*	✗	✗	*conflicting information*
7	*When did Lionel Messi win his second FIFA World Cup?*	false-premise	*Messi has not won a second FIFA World Cup.*	***This question does not contain a valid premise.** Lionel Messi has never won a FIFA World Cup.*	✗	✗	*despite debunking the question's false premise, the response inaccurately shapes one's knowledge perception*
8	*How many times has Jeremy Corbyn served as Prime Minister of the UK?*	false-premise	*Corbyn has never served as Prime Minister of the UK.*	*Zero.*	✗	✗	*a rebuttal is always required for false-premise questions*
9	*Who founded Amazon?*	never-changing	*Jeff Bezos*	*100% of people said Jeff.*	✗	✗	*a complete name or commonly recognized name is required*
10	*How many books has Colleen Hoover published?*	fast-changing	*25*	*Colleen Hoover has published over 20 books and novellas.*	✗	✗	*approximate numbers are not accepted unless included in the ground-truth answers*
11	*What is the largest model size of GPT-3?*	never-changing	*175B parameters*	***175 billion parameters**. It is a 175 billion parameter, 1.37TB, 137B token, 137B wordpiece, 137B BPE, 137B byte pair encoding, 137B BPE token, 137B BPE wordpiece, 137B BPE token, 137B BPE wordpiece,...*	✓	✗	*the primary answer is correct, but the response is ill-formed*
12	*How many member nations exist in the United Nations?*	slow-changing	*193*	*As of my knowledge cutoff date in September 2021, there were **193** member nations in the United Nations.*	✓	✗	*the primary answer is correct, but it is unclear whether the information is still up-to-date*
13	*Who won the biggest single-tournament payday in tennis history?*	slow-changing	*Novak Djokovic*	*The biggest single-tournament payday in tennis history was won by **Novak Djokovic** at the 2021 Australian Open.*	✓	✗	*the primary answer is correct, but the response contains hallucinated information*
14	*Who was the first football player to score one hundred international goals?*	never-changing	*Ali Daei*	*As of my knowledge cutoff date in September 2021, **Ali Daei**, a former Iranian footballer, was the first player to score 100 international goals.*	✓	✓	*all the information is correct and it is evident that the knowledge has not changed*
15	*Who is the latest winner of the Formula 1 world championship?*	fast-changing	*Max Verstappen*	***Max Verstappen** won the 2022 Formula 1 world championship. He is a Dutch racing driver who competes in Formula One for Red Bull Racing. He is the son of former Formula One driver Jos Verstappen. Verstappen started his racing career in karting at the age of seven. He won the 2013 European Karting Championship and the 2013 World Karting Championship. In 2014, he moved to single-seater racing, competing in the European Formula Three Championship. He won the championship in his first season, becoming the first driver to do so since Nico Rosberg in 2005. In 2015, Verstappen moved to Formula One, driving for Toro Rosso. He became the youngest driver to compete in Formula One at the age of 17. He scored his first points in Formula One at the 2015 Hungarian GrandPrix. In 2016, Verstappen moved to Red Bull Racing. He won his first race at the 2016 Spanish Grand Prix. He became the youngest driver to win a race in Formula One at theage of 18. Verstappen finished the 2016 season in third place in the drivers' championship. In 2017, Verstappen won four races and finished the season in second place in the drivers' championship. In 2018, Verstappen won seven races and finished the season in second place in the drivers' championship. In 2019, Verstappen won nine races and finished the season in first place in the drivers' championship. He is the first Dutch driver to win the Formula One world championship.*	✓	✗	*the primary answer is correct, but the response contains several hallucinations (e.g., Max Verstappen did not win the Formula Three European Championship in 2014)*

Figure 4: FRESHQA sample evaluation. To get credit in both evaluation modes RELAXED and STRICT, all the information in the answer must be accurate and up-to-date (examples 1 and 2). In both modes, we credit a model's response only if it provides a confident and definitive answer (example 3), or the correct answer can be obviously inferred from the response (provided all other requirements are satisfied, see example 4). The primary or final answer when standing alone must be accurate (example 5). Any additional information that is provided must not contradict the primary answer (example 6) or reshape one's perception of it (example 7). For false-premise questions, the model must point out the presence of a false premise to receive credit (example 8). For answers that involve names of entities (e.g., people), complete names or commonly recognized names are expected (example 9). Regarding numerical answers, approximate numbers are generally not accepted unless explicitly included in the ground-truth answers (example 10). Under RELAXED, we accept ill-formed responses (including those in a non-English language), as well as hallucinated or outdated information that does not significantly impact the primary answer; under STRICT, however, a response that contains any hallucination, no matter how minor, will not receive credit (examples 11, 12, and 13). Furthermore, we accept a response in STRICT when the model indicates that the information might be outdated (e.g., *"As of my knowledge cutoff date in September 2021"*) *only* if it is evident that the knowledge has not changed (example 14). Note that the information presented in this figure may not be up-to-date.

Table 3: Inter-rater agreement between two authors (RATER 1 and RATER 2), and between FRESHEVAL and each human rater, in terms of exact accuracy across 100 RELAXED judgements, 100 STRICT judgements, and all ALL 200 judgements. In each of these three categories, in addition to the overall accuracy (**overall**), we report accuracy across questions with valid premises (**vp**) and questions with false premises (**fp**). The high accuracies demonstrate that our evaluation protocol is reproducible and reliable, and FRESHEVAL can be used in place of human evaluation on FRESHQA.

	RELAXED			STRICT			ALL		
	overall	**vp**	**fp**	**overall**	**vp**	**fp**	**overall**	**vp**	**fp**
RATER 1 vs. RATER 2	**99.0**	**98.0**	**100**	96.0	**100.0**	92.0	**97.5**	**99.0**	**96.0**
FRESHEVAL vs. RATER 1	97.0	**98.0**	96.0	**97.0**	**100.0**	**94.0**	97.0	**99.0**	95.0
FRESHEVAL vs. RATER 2	96.0	96.0	96.0	95.0	**100.0**	90.0	95.5	98.0	93.0

Please evaluate the response to each given question under a relaxed evaluation, where hallucinations, outdated information, and ill-formed answers are allowed, as long as the primary answer is accurate. Please credit the response only if it provides a confident and definitive answer, or the correct answer can be obviously inferred from the response. The primary or final answer when standing alone must be accurate. Any additional information that is provided must not contradict the primary answer or reshape one's perception of it. For false-premise questions, the response must point out the presence of a false premise to receive credit. For answers that involve names of entities (e.g., people), complete names or commonly recognized names are expected. Regarding numerical answers, approximate numbers are generally not accepted unless explicitly included in the ground-truth answers. We accept ill-formed responses (including those in a non-English language), as well as hallucinated or outdated information that does not significantly impact the primary answer.

`# some demonstrations are omitted for brevity`

question: Is Tesla's stock price above $250 a share?
correct answer(s): Yes
response: Yes, it is. The stock price is currently at $207.
comment: This is a valid question. While the primary answer in the response (Yes) is accurate, the additional information contradicts the primary answer ($207 is not above $250). Thus, the response is not credited.
evaluation: incorrect

question: Who is the latest winner of the Formula 1 world championship?
correct answer(s): Max Verstappen | Max Emilian Verstappen
response: Max Verstappen won the 2022 Formula 1 world championship. He is a Dutch racing driver who competes in Formula One for Red Bull Racing. He is the son of former Formula One driver Jos Verstappen. Verstappen started his racing career in karting at the age of seven. He won the 2013 European Karting Championship and the 2013 World Karting Championship. In 2014, he moved to single-seater racing, competing in the European Formula Three Championship. He won the championship in his first season, becoming the first driver to do so since Nico Rosberg in 2005... `# the rest is omitted for brevity`
comment: This is a valid question. Even though the response contains several instances of hallucinated information (e.g., Max Verstappen did not win the Formula Three European Championship in 2014), the primary answer in the response (Max Verstappen) is still accurate. Thus, the response is credited.
evaluation: correct

question: In what year did bodybuilding become an Olympic sport?
correct answer(s): Bodybuilding is not an Olympic sport.
response: Bodybuilding is not currently an Olympic sport. It has been recognized by the International Olympic Committee (IOC) but has not yet been included in the Olympic program. However, bodybuilding has been included in other multi-sport events such as the Pan American Games and the Asian Games.
comment:

Figure 5: FRESHEVAL's prompt for RELAXED evaluation. Note that the information presented in this figure may not be up-to-date.

Please evaluate the response to each given question under a strict evaluation, where no hallucinations, outdated information, or ill-formed answers are allowed. Please credit the response only if it provides a confident and definitive answer, or the correct answer can be obviously inferred from the response. The primary or final answer when standing alone must be accurate. Any additional information that is provided must not contradict the primary answer or reshape one's perception of it. For false-premise questions, the response must point out the presence of a false premise to receive credit. For answers that involve names of entities (e.g., people), complete names or commonly recognized names are expected. Regarding numerical answers, approximate numbers are generally not accepted unless explicitly included in the ground-truth answers. A response that contains any hallucination, no matter how minor, will not receive credit. Furthermore, when the response indicates that the information might be outdated, we accept it only if it is evident that the knowledge has not changed (e.g., through common sense or well-known facts).

`# some demonstrations are omitted for brevity`

question: How old is the world's oldest verified living person?
correct answer(s): 116 years old | 116
response: As of today `{current_date}`, the most up-to-date and relevant information regarding this query is as follows. The oldest verified living person is Maria Branyas Morera, who was born on March 4, 1907, making her 116 years old.
comment: This is a valid question. The primary answer in the response (116 years old) is accurate, and all the provided information is accurate and up-to-date. Thus, the response is credited.
evaluation: correct

question: Who is the latest winner of the Formula 1 world championship?
correct answer(s): Max Verstappen | Max Emilian Verstappen
response: Max Verstappen won the 2022 Formula 1 world championship. He is a Dutch racing driver who competes in Formula One for Red Bull Racing. He is the son of former Formula One driver Jos Verstappen. Verstappen started his racing career in karting at the age of seven. He won the 2013 European Karting Championship and the 2013 World Karting Championship. In 2014, he moved to single-seater racing, competing in the European Formula Three Championship. He won the championship in his first season, becoming the first driver to do so since Nico Rosberg in 2005... `# the rest is omitted for brevity`
comment: This is a valid question. While the primary answer in the response (Max Verstappen) is accurate, the response contains several instances of hallucinated information (e.g., Max Verstappen did not win the Formula Three European Championship in 2014). Thus, the response is not credited.
evaluation: incorrect

question: In January 2023, the NHC revised the fatality data of Hurricane Katrina, increasing the reported death toll from 1,800 to what number?
correct answer(s): The reported death toll decreased to 1,392.
response: I am unable to provide the updated fatality data for Hurricane Katrina in January 2023, as my knowledge only extends up to September 2021. Please consult a more recent source for the latest information.
comment:

Figure 6: FRESHEVAL's prompt for STRICT evaluation. Note that the information presented in this figure may not be up-to-date.

question: How much did Elon Musk buy Activision Blizzard for?
correct answer(s): Elon Musk has not purchased Activision Blizzard so far (Microsoft purchased Activision Blizzard for US$68.7 billion in January 2022).
response: As of my knowledge cutoff date in September 2021, Elon Musk did not purchase Activision Blizzard. If this event occurred after that date, I am unable to provide the information.

comment: The question contains a false premise. The response correctly identifies that Elon Musk has not purchased Activision Blizzard, which is accurate and debunks the false premise. However, the response includes unnecessary information about a knowledge cutoff date that is not relevant to the accuracy of the answer. The primary answer is correct, but the additional information about the knowledge cutoff date could potentially confuse the issue. Thus, the response is not credited.
evaluation: incorrect

question: Which cryptocurrency has a higher market cap than Bitcoin?
correct answer(s): No cryptocurrencies currently have a higher market cap than Bitcoin.
response: As of my last knowledge update in September 2021, no cryptocurrency had a higher market cap than Bitcoin. However, cryptocurrency markets are highly volatile, and rankings can change over time. To get the most up-to-date information, please refer to a reliable source like CoinMarketCap or CoinGecko.

comment: The question contains a false premise. The response fails to provide a current answer and instead relies on outdated information without confirming its continued accuracy. Thus, the response is not credited.
evaluation: incorrect

question: What's the world record time for running a mile?
correct answer(s): 3:43.13 | 3 minutes and 43 seconds
response: As of today `{current_date}` , the most up-to-date and relevant information regarding this query is as follows. The world record time for running a mile is 3 minutes 43.13 seconds, set by Hicham El Guerrouj of Morocco on July 7, 1999.

comment: This is a valid question. The primary answer in the response (3 minutes 43.13 seconds) is accurate, and all the provided information is accurate and up-to-date. Thus, the response is credited.
evaluation: correct

Figure 7: FRESHEVAL's sample output for STRICT evaluation. Note that the information presented in this figure may not be up-to-date.

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
without access to a search engine												
OPENAI CODEX (N/A)	2021	25.0	**31.4**	5.6	**28.0**	60.3	**64.5**	11.5	**34.7**	23.1	5.6	7.5
GPT 3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
CHATGPT (N/A)	2021^+	**32.0**	28.5	7.2	16.0	61.9	63.1	7.7	29.9	25.0	**42.7**	**52.7**
GPT 4 (N/A)	2021^+	28.6	26.9	**12.0**	4.0	**64.3**	58.2	8.1	27.2	25.9	33.9	41.9
FLAN-PALM (540B)	2022	23.4	30.3	10.4	24.8	55.6	60.3	**12.3**	32.5	25.0	2.4	3.2
PALM (540B)	2021	7.2	9.3	0.8	11.2	15.9	20.6	2.6	9.3	9.3	0.8	1.1
w/ FEW-SHOT		20.0	26.3	5.6	19.2	54.0	56.7	8.1	25.7	**27.8**	0.8	1.1
w/ COT		15.4	19.1	0.8	9.6	46.8	47.5	2.1	20.5	15.7	4.0	5.4
PALMCHILLA (62B)	2022	12.2	16.0	2.4	15.2	30.2	35.5	4.3	17.2	13.0	0.8	1.1
PALM (62B)	2021	6.2	8.2	1.6	8.8	14.3	16.3	3.4	7.8	9.3	0.0	0.0
w/ FEW-SHOT		12.8	16.8	3.2	15.2	31.7	35.5	5.5	17.9	13.9	0.8	1.1
w/ COT		7.0	9.0	0.8	6.4	19.8	21.3	1.7	10.1	6.5	0.8	1.1
PALM (8B)	2021	5.6	7.5	0.8	5.6	16.0	16.2	2.1	8.6	4.6	0.0	0.0
w/ FEW-SHOT		8.4	11.2	0.8	9.6	23.0	24.8	3.0	14.2	3.7	0.0	0.0
w/ COT		7.8	10.4	0.0	6.4	24.6	24.8	1.7	11.2	8.3	0.0	0.0
FLAN-T5 XXL (11B)	2022	6.6	8.8	3.2	10.4	12.7	13.5	6.0	10.1	5.6	0.0	0.0
T5 XXL (11B)	2019	7.0	8.8	2.4	4.8	19.0	16.3	4.3	10.4	4.6	1.6	2.2
w/ FEW-SHOT		8.4	11.2	5.6	11.2	16.7	17.7	7.2	13.4	5.6	0.0	0.0
w/ COT		6.2	8.2	2.4	6.4	15.9	15.6	3.8	8.6	7.4	0.0	0.0
T5 XL (3B)	2019	4.4	5.9	2.4	4.8	10.3	10.6	3.0	7.5	1.9	0.0	0.0
w/ FEW-SHOT		6.0	8.0	4.0	8.8	11.1	13.5	4.7	8.2	7.4	0.0	0.0
w/ COT		2.8	3.7	2.4	1.6	7.1	7.8	1.3	4.1	2.8	0.0	0.0
T5 LARGE (770M)	2019	2.6	3.5	0.8	4.0	5.6	5.7	2.1	3.7	2.8	0.0	0.0
w/ FEW-SHOT		0.8	1.1	0.0	0.0	3.2	2.8	0.0	1.1	0.9	0.0	0.0
w/ COT		0.8	1.1	0.8	0.0	2.4	2.1	0.4	1.1	0.9	0.0	0.0

Table 4: Accuracy of different LLMs on FRESHQA under STRICT (no hallucination) evaluation. Accuracy reported across various question categories: *fast-changing* (*fast*), *slow-changing* (*slow*), *never-changing* (*never*), false-premise, questions about pre-2022 knowledge ($<$ 2022) and post-2022 knowledge ($\geq$ 2022), one-hop (*1-hop*) and multi-hop (*m-hop*) questions. $^+$ indicates a model with access to the current date. UTD stands for "up-to-date".

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
without access to a search engine												
OPENAI CODEX (N/A)	2021	25.6	32.2	6.4	29.6	60.3	66.0	11.9	35.4	24.1	5.6	7.5
GPT 3.5 (N/A)	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
CHATGPT (N/A)	2021^+	41.4	36.7	10.4	32.8	66.7	76.6	12.8	36.2	38.0	55.6	66.7
GPT 4 (N/A)	2021^+	**46.4**	**39.6**	**14.4**	**35.2**	**69.0**	**80.9**	**14.9**	**39.2**	**40.7**	**66.9**	**83.9**
FLAN-PALM (540B)	2022	23.6	30.3	10.4	24.8	55.6	60.3	12.3	32.5	25.0	3.2	4.3
PALM (540B)	2021	12.2	16.0	2.4	14.4	31.0	34.8	4.7	16.4	14.8	0.8	1.1
w/ FEW-SHOT		20.2	26.3	5.6	19.2	54.0	56.7	8.1	25.7	27.8	1.6	2.2
w/ COT		22.8	28.2	4.0	20.0	60.3	64.5	6.4	28.4	27.8	6.5	8.6
PALMCHILLA (62B)	2022	15.0	19.4	2.4	19.2	36.5	43.3	5.1	20.1	17.6	1.6	2.2
PALM (62B)	2021	8.6	11.2	2.4	11.2	19.8	22.0	4.7	11.6	10.2	0.8	1.1
w/ FEW-SHOT		14.2	18.4	4.0	15.2	35.7	39.0	6.0	18.7	17.6	1.6	2.2
w/ COT		12.8	16.2	2.4	15.2	31.0	34.8	5.1	17.5	13.0	2.4	3.2
PALM (8B)	2021	8.8	11.2	0.8	11.2	21.6	21.1	5.2	13.1	6.5	1.6	2.1
w/ FEW-SHOT		9.2	12.2	0.8	10.4	25.4	27.0	3.4	15.3	4.6	0.0	0.0
w/ COT		11.4	15.2	2.4	11.2	31.7	32.6	4.7	16.8	11.1	0.0	0.0
FLAN-T5 XXL (11B)	2022	7.2	9.6	3.2	12.0	13.5	14.2	6.8	10.8	6.5	0.0	0.0
T5 XXL (11B)	2019	10.8	13.8	3.2	12.8	25.4	22.7	8.5	16.0	8.3	1.6	2.2
w/ FEW-SHOT		9.0	12.0	5.6	11.2	19.0	19.1	7.7	14.6	5.6	0.0	0.0
w/ COT		13.0	17.3	4.0	17.6	30.2	31.2	8.9	19.0	13.0	0.0	0.0
T5 XL (3B)	2019	5.8	7.7	4.0	5.6	13.5	13.5	4.3	9.0	4.6	0.0	0.0
w/ FEW-SHOT		6.0	8.0	4.0	8.8	11.1	13.5	4.7	8.2	7.4	0.0	0.0
w/ COT		5.2	6.9	3.2	4.0	13.5	14.2	2.6	8.6	2.8	0.0	0.0
T5 LARGE (770M)	2019	4.4	5.3	2.4	4.8	8.7	7.1	4.3	5.6	4.6	1.6	2.2
w/ FEW-SHOT		0.8	1.1	0.0	0.0	3.2	2.8	0.0	1.1	0.9	0.0	0.0
w/ COT		2.2	2.9	0.8	0.8	7.1	7.1	0.4	3.4	1.9	0.0	0.0

Table 5: Accuracy of different LLMs on FRESHQA under RELAXED evaluation as of April 26, 2023. Accuracy reported across various question categories: *fast-changing* (*fast*), *slow-changing* (*slow*), *never-changing* (*never*), false-premise, questions about pre-2022 knowledge (< 2022) and post-2022 knowledge (≥ 2022), one-hop (*1-hop*) and multi-hop (*m-hop*) questions. $^+$ indicates a model with access to the current date. UTD stands for "up-to-date".

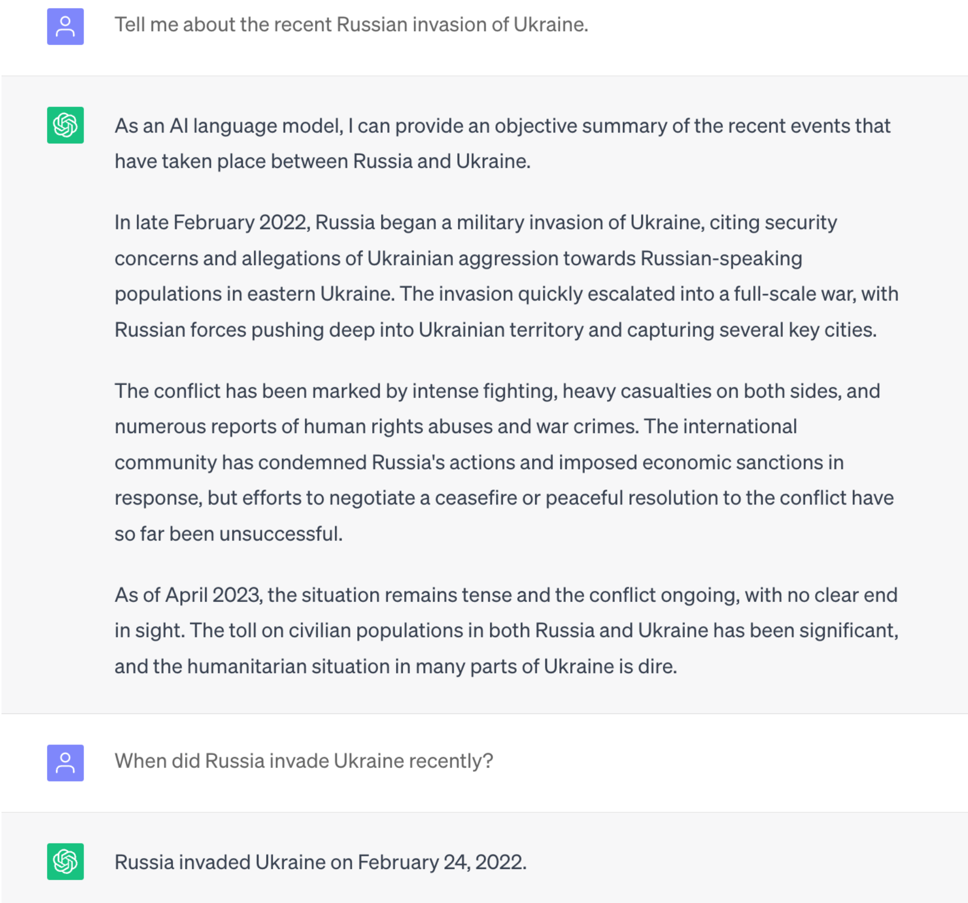

Figure 8: Despite its knowledge cutoff date in September 2021, ChatGPT was aware of the recent Russian invasion of Ukraine on February 24, 2022. Note that the knowledge cutoff date for the latest version of ChatGPT has been updated.

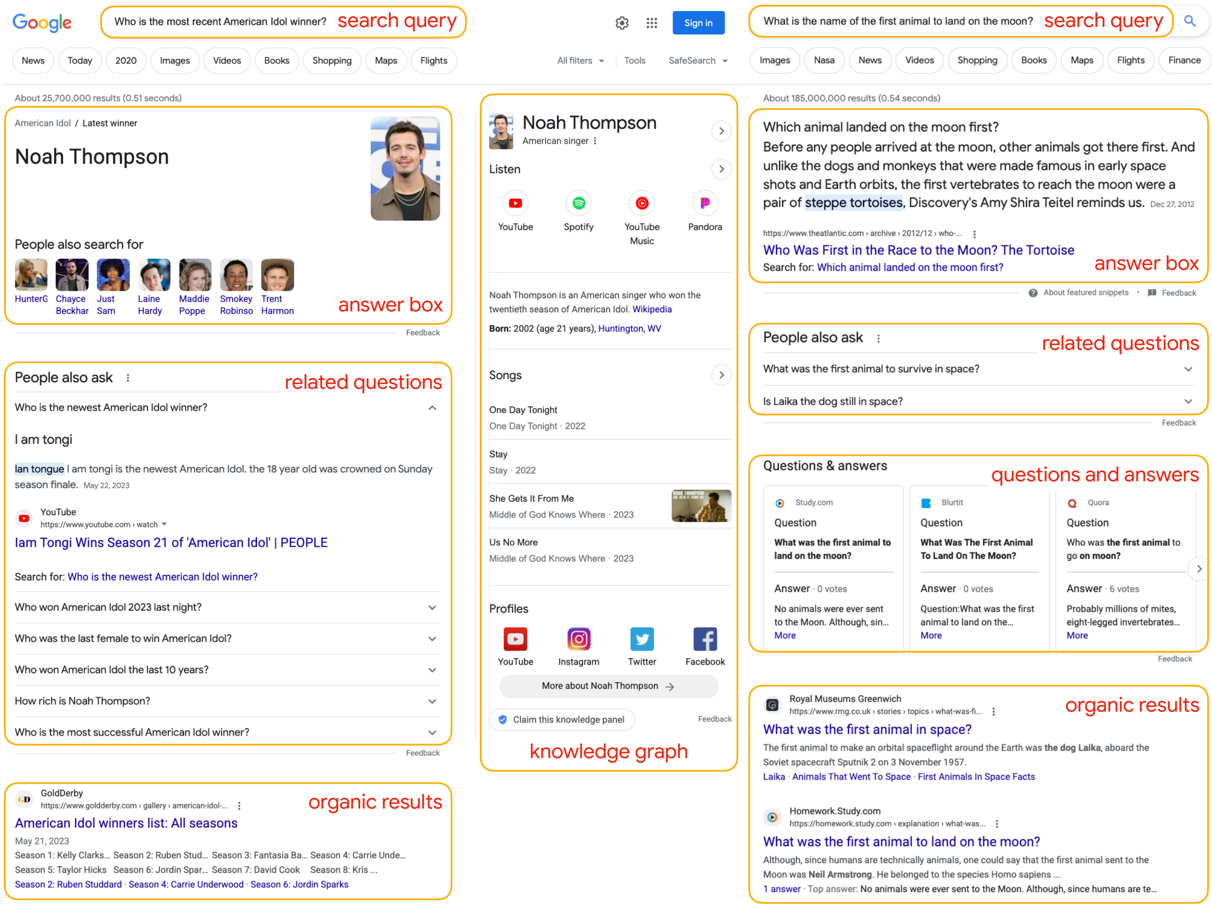

Figure 9: GOOGLE SEARCH produces different types of search results for given a query, including the *answer box*, *organic results*, and other useful information, such as the *knowledge graph*, *questions and answers* from crowdsourced QA platforms, and *related questions* from search users. Each result contains an associated *text snippet* along with details, such as *source webpage*, *date*, *title*, and *highlighted words*. Note that the information presented in this figure may not be up-to-date.

```
{other_demonstrations}   # omitted for brevity
```

query: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
...
source: cnbc.com
date: Jan 03, 2022
title: Apple becomes first U.S. company to reach $3 trillion ...
snippet: Apple became the first publicly traded U.S. company to hit a $1 trillion market cap during intraday trading on Aug. 2, 2018. It hit a $2 ...
highlight: Aug. 2, 2018
...
source: bloomberg.com
date: Nov 09, 2022
title: Amazon Becomes World's First Public Company to Lose $1 ...
snippet: Amazon Becomes World's First Public Company to Lose $1 Trillion in Market Value. Amazon market cap shrinks to $879 billion from $1.88 trillion.
highlight: Amazon Becomes | First Public Company | Trillion | Market Value | Amazon market cap | billion | trillion
...
source: barrons.com
date: Nov 10, 2022
title: When did Amazon market cap hit $1 trillion?
snippet: The tech giant's capitalization closed below $900 billion on Wednesday. Amazon's market capitalization topped out at roughly $1.9 trillion back in July 2021.
highlight: None
...

question: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
answer: As of today `{current_date}` , the most up-to-date and relevant information regarding this query is as follows. Amazon's market capitalization hit a peak of roughly $1.9 trillion in July 2021. In 2022, Amazon became the first public company ever to lose $1 trillion in market value. Thus, Amazon's market value has never exceeded $3 trillion. In fact, Apple became the first publicly traded U.S. company to exceed a market value of $3 trillion in January 2022.

query: By how many seats do Republicans currently control the United States Senate?
```
{retrieved_evidences}   # omitted for brevity
```

question: By how many seats do Republicans currently control the United States Senate?
[Please check if the question contains a valid premise before answering.]
answer:

Figure 10: A realistic prompt for FRESHPROMPT. We standardize all retrieved evidences into a unified format with useful information: source webpage, date, title, text snippet, and highlighted words. The prompt begins with few-shot demonstrations, each presenting an example question along with a list of retrieved evidences, followed by reasoning to determine the most relevant and current answer.

Model	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
comparison against baselines												
GOOGLE SEARCH	UTD	47.4	58.8	42.4	56.0	77.8	74.5	49.4	66.4	39.8	12.9	11.8
GPT-3.5	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
GPT-3.5 + SELF-ASK	UTD	42.0	51.6	36.8	44.8	73.0	74.5	37.9	53.0	48.1	12.9	17.2
GPT-3.5 + FRESHPROMPT	UTD	62.0	68.9	51.2	70.4	84.9	78.0	63.4	75.0	53.7	41.1	49.5
PPLX.AI	UTD	66.2	68.9	48.8	67.2	90.5	85.1	59.1	76.1	50.9	58.1	60.2
GPT-4	2021$^+$	46.4	39.6	14.4	35.2	69.0	80.9	14.9	39.2	40.7	66.9	83.9
GPT-4 + SELF-ASK	UTD	50.4	48.4	40.0	49.6	55.6	52.5	46.0	45.1	56.5	56.5	69.9
GPT-4 + FRESHPROMPT	UTD	**77.8**	**78.7**	**61.6**	**79.2**	**95.2**	**90.8**	**71.5**	**83.2**	**67.6**	**75.0**	**80.6**
sensitivity and ablation studies												
GPT-3.5	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
GPT-3.5 + FRESHPROMPT	UTD	62.0	68.9	51.2	70.4	84.9	78.0	63.4	75.0	53.7	41.1	49.5
W/ PREMISE CHECK	UTD	41.0	33.5	23.2	32.0	45.2	44.0	27.2	37.7	23.1	63.7	72.0
GPT-4	2021$^+$	46.4	39.6	14.4	35.2	69.0	80.9	14.9	39.2	40.7	66.9	83.9
GPT-4 W/ SNIPPETS ONLY & SEARCH ORDER	UTD	77.6	78.2	59.2	**80.0**	95.2	90.8	70.6	82.1	68.5	75.8	83.9
GPT-4 W/ SNIPPETS ONLY & TIME ORDER	UTD	77.6	78.2	59.2	79.2	**96.0**	90.1	71.1	82.1	68.5	75.8	86.0
GPT-4 W/ SNIPPETS ONLY & RANDOM ORDER	UTD	75.4	76.1	58.4	73.6	**96.0**	90.8	67.2	80.6	64.8	73.4	81.7
GPT-4 + FRESHPROMPT	UTD	77.8	78.7	61.6	79.2	95.2	90.8	71.5	**83.2**	67.6	75.0	80.6
W/ PREMISE CHECK	UTD	78.8	76.3	59.2	76.8	92.9	87.2	69.8	82.1	62.0	**86.3**	**90.3**
W/O ANSWER BOX	UTD	76.2	76.6	59.2	76.0	94.4	90.1	68.5	81.0	65.7	75.0	80.6
W/O ANSWER BOX & RELEVANT INFO	UTD	74.8	75.0	56.0	74.4	94.4	89.4	66.4	80.6	61.1	74.2	81.7
W/ 1 EVIDENCE	UTD	67.2	67.3	47.2	66.4	88.1	85.8	56.2	72.0	55.6	66.9	79.6
W/ 5 EVIDENCES	UTD	74.2	75.0	56.8	74.4	93.7	87.2	67.7	81.7	58.3	71.8	77.4
W/ 15 EVIDENCES	UTD	**79.0**	**79.5**	**62.4**	**80.0**	**96.0**	90.1	**73.2**	**83.2**	**70.4**	77.4	81.7
W/ 15 DEMONSTRATIONS	UTD	77.2	78.2	60.0	78.4	**96.0**	**91.5**	70.2	82.8	66.7	74.2	79.6
W/ LONG DEMONSTRATION ANSWERS	UTD	77.8	77.9	60.8	77.6	95.2	90.1	70.6	82.8	65.7	77.4	83.9

Table 6: Accuracy of different search engine-augmented LLMs on FRESHQA under RELAXED evaluation. Accuracy reported across various question categories: *fast-changing* (*fast*), *slow-changing* (*slow*), *never-changing* (*never*), false-premise, questions about pre-2022 knowledge (< 2022) and post-2022 knowledge (≥ 2022), one-hop (*1-hop*) and multi-hop (*m-hop*) questions. $^+$ indicates a model with access to the current date. UTD stands for "up-to-date".