Pop Quiz! Comparing Educational Assessment Methods for Evaluating Context Utilization vs. Pre-Trained Knowledge in Large Language Models

Anura Deshpande, Suvaditya Mukherjee, Benjamin Nye, Ella Park, Russell Harral
Institute for Creative Technologies, University of Southern California
Los Angeles, CA, USA
{anurades, suvadity}@usc.edu
nye@ict.usc.edu

Abstract

Large Language Models (LLMs) are increasingly asked questions on documents, yet it remains unclear whether their responses reflect information from the provided source or their pretraining. Existing document question-answering benchmarks primarily evaluate answer accuracy. In this work, we adapt educational assessment techniques to evaluate LLMs and their use of context through document-based question-answering experiments. Over multiple iterations across a set of frontier models, we compare three different approaches that estimate a model's internal knowledge (as a form of LLM metacognition). First, multiple-choice questions are generated with the document to test models. Second, models must identify when no correct answer is present and generate a missing correct answer. Third, the models answers open-ended questions. Each approach tested models with and without document context. Altogether, we intend to evaluate LLM recognition, error detection, and reasoning, just as these assessments would for learners. Experiments are conducted across multiple models on a corpus of technical and nontechnical Wikipedia articles sampled across revision periods relative to model training cutoffs. While access to document context consistently improved performance, the magnitude of improvement depended strongly on assessment strategy. Open-response evaluation produced the largest absolute context gains compared with standard multiple-choice questions, indicating that recognition-based benchmarks may underestimate the contribution of external documents to natural language generation related tasks. We further find that document characteristics, including domain and revision period, influence context utilization, while model pairings have comparatively little effect. These findings provide a framework for evaluating document-grounded language generation and offer practical guidance for developing more reliable retrieval-augmented language systems.

1 Introduction

LLMs are increasingly becoming prevalent as educational tutor, enterprise search, and retrieval augmented generation (RAG) systems, where users input one or more source documents for models to retrieve and reason over their contents (Bhayana, 2024; Modran et al., 2024; Pasupuleti et al., 2024). Although access to specific domain-specific documents should naively improve performance, it remains unclear as to whether responses reflect reasoning over the provided context or parametric knowledge acquired during pretraining. Distinguishing between these behaviors is especially important in specialized or developing domains, where pretrained knowledge may be outdated or inconsistent with the supplied document. Existing evaluations of document-based question answering primarily measure answer precision (Chen and Mueller, 2024; Zhao et al., 2026). Furthermore, standard benchmarks rarely examine model behavior under alternative assessment settings, such as identifying missing information and recognizing flawed questions. For inspiration, we look towards the education domain, wherein assessments have historically employed multiple formats in order to measure understanding beyond factual recognition.

In this work, we propose a document-based evaluation framework for measuring context utilization and document usage. We evaluate models across question-answering tasks, beginning with standard multiple-choice questions (MCQ) answered with and without document context, followed by experiments requiring models to identify incorrect answer sets, generate missing answers, and produce open-ended responses. Together, they assess free-form reasoning and demonstrate whether models are able to answer factually with a dependence on context or not. In addition to answer precision, we evaluate explanation quality subjectively through a LLM-as-a-judge pipeline. Experiments

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 960–971
October 17-21, 2026 ©2026 Association for Computational Linguistics

are conducted across ChatGPT and Gemini models using a balanced corpus of technical and nontechnical Wikipedia articles sampled from revision periods spanning model training cutoffs. Our work is guided by these research questions:

RQ1: How does model performance differ with and without access to the source document?

RQ2: Which assessment strategy most effectively differentiates document-grounded reasoning from parametric knowledge?

RQ3: How do performance gains from document context vary across domains and time periods?

These research questions examine when and to what extent LLMs use documents over their parametric knowledge. Rather than evaluating end-to-end RAG, we isolate the document ingestion stage. This allows us to observe performance independent of retrieval quality. RQ1 investigates the extent to which document access may improve performance, RQ2 asks which assessment strategy creates the most clear distinction between document-grounded and parametric knowledge, and RQ3 examines how this varies across document characteristics. These experiments aim to characterize and provide a new framework for the evaluation of context-utilization. These findings could inform the evaluation of a wide range of downstream tasks, that require models to reason over provided context.

Our analysis demonstrates that the impact of external document context depends strongly on assessment type, with open-response producing the largest absolute separation between document-grounded reasoning and parametric knowledge. Additionally, we show that document domain and temporal recency influence context utilization, while these trends remain consistent across model pairings.

2 Related Work

2.1 Educational Evaluation Format

Educational assessment has long recognized that no single examination format captures all aspects of student understanding (Simkin and Kuechler, 2005; Yılmaz and Toker, 2022). MCQ primarily evaluate recognition, whereas constructed-response questions require learners to recall, synthesize, and explain information without cues (Bennett et al., 1991; Frary, 1985). Different assessment formats place different cognitive demands on learners, which can produce variable estimates of ability even when evaluating the same underlying material.

MCQ are widely used to benchmark LLMs, not only for their efficiency and simplicity, but also for enabling evaluation across a wide range of subjects (Hendrycks et al., 2021; Li et al., 2024). However, this format has important limitations, including sensitivity to option ordering, reliance on recognition cues, and an inability to fully test a model's knowledge (Balepur et al., 2025; Molfese et al., 2025; Zhang et al., 2024). In contrast, long-form generation has no fixed answer options, and avoids the order sensitivity and recognition cues that affect MCQ, though constructed responses are more complex to score reliably than multiple-choice selections (Balepur et al., 2025).

Recent work has examined LLM capabilities that go beyond standard MCQ accuracy. Kadavath et al. (2022) introduce P(IK) ("Probability I Know"), a learned prediction of whether a model can answer a given question, and find that this self-assessed knowledge increases when a relevant document is added to the context. They further observe that replacing an answer option with "None of the Above" degrades both accuracy and calibration, as models are biased against selecting it. This suggests that recognizing the absence of a correct answer is a distinct reasoning capability rather than a byproduct of answer selection. Madhusudhan et al. (2025) build on this with Abstain-QA, a benchmark whose unanswerable items remove the correct option and substitute a plausible distractor, so that selecting "None of the Above" is correct. This study found that strong models abstain reliably on simple factual questions, but degrade on complex reasoning tasks. However, these works each examine a single format in isolation, and do not ask whether the choice of format changes how much document reliance we observe. To our knowledge, no prior work has systematically compared recognition, abstention, and generation formats on the same documents to measure how the estimated benefit of context depends on the assessment format itself.

2.2 Parametric and Contextual Knowledge

When a language model answers a question about a document, its response may originate from two distinct sources. The first is the retrieved document provided in the prompt, commonly referred to as *contextual knowledge* (Longpre et al., 2021;

Neeman et al., 2023). The second is information encoded during pretraining and stored within the model parameters, referred to as *parametric knowledge* (Neeman et al., 2023; Petroni et al., 2019; Roberts et al., 2020). When the two sources disagree, models often disregard the provided passage in favor of parametric memory (Chen et al., 2022; Longpre et al., 2021), and whether they do so depends on factors as brittle as the coherence of the presented evidence and the model's confidence in its prior answer (Xie et al., 2024; Xu et al., 2024). Distinguishing whether a response is grounded in the document or recalled from memory therefore remains a central challenge for evaluating retrieval-augmented systems, however has been explored in prior work. Mallen et al. (2023) introduced PopQA, a benchmark designed to study when language models require retrieval versus relying on memorized knowledge. They demonstrate that retrieval provides the greatest benefit for long-tail entities and less frequently observed facts, while larger models primarily improve performance on more popular knowledge. Consistent with this, Kandpal et al. (2023) show that how often a fact appears in pretraining data strongly predicts whether a model can recall it. These findings suggest that the value of external documents depends on the availability of corresponding parametric knowledge, motivating evaluations that explicitly distinguish between the two sources.

Closely related is the problem of factual grounding. Ji et al. (2023) distinguish hallucinations that contradict the source document from those that introduce unsupported information, emphasizing the importance of determining whether generated content is grounded in retrieved evidence. Growing mitigation literature already treats over-reliance on parametric memory as a problem to engineer around, through context-faithful prompting (Zhou et al., 2023) and context-aware decoding (Shi et al., 2024); these methods, however, modify model behavior rather than measure it. Rather than hallucinations, our work examines how the need for document grounding varies across assessment formats.

Prior efforts to measure this distinction have largely relied on knowledge-conflict constructions: editing the document by substituting entities or generating counterfactual passages that contradict memory, then observing which source the model favors (Chen et al., 2022; Longpre et al., 2021; Neeman et al., 2023). While informative, such evaluations require modified documents that deployed systems rarely encounter, leaving open whether context reliance can be measured on unmodified, naturally occurring documents. We address this by manipulating document availability and answer options rather than document content, and by stratifying our corpus by revision date relative to model training cutoffs (Dhingra et al., 2022; Kasai et al., 2023), so that whether content could have been memorized is controlled by construction rather than assumed. For RAG systems operating over fresh or specialized documents, this is the practical distinction: an answer that sounds correct may be recited from stale parametric memory rather than read from the source.

2.3 Document Question Answering

LLMs are increasingly deployed in document question answering pipelines, where retrieved documents are incorporated into the prompt to supplement a model's knowledge (Chen et al., 2017; Guu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020; Ram et al., 2023). Retrieval itself, however, can fail even on simple entity-centric queries (Sciavolino et al., 2021). These systems have consistently demonstrated improvements on knowledge-intensive tasks by grounding responses in external documents rather than relying solely on pretrained information. Despite these advances, evaluation remains largely centered on answer correctness. Reading-comprehension benchmarks score systems by exact match or F1 against gold answers (Kwiatkowski et al., 2019; Rajpurkar et al., 2018), and RAG benchmarks and automated evaluation frameworks similarly assess retrieval by measuring accuracy improvements once documents are provided (Es et al., 2024; Mohammadi et al., 2025; Ru et al., 2024; Saad-Falcon et al., 2024). Correctness alone, however, is an incomplete signal: Adlakha et al. (2024) show that traditional lexical metrics misjudge instruction-following models, and that a correct answer is not necessarily faithful to the provided document. Although extractive benchmarks introduced unanswerable questions precisely to penalize guessing (Rajpurkar et al., 2018), this scrutiny of the assessment format itself has largely not carried over to document-grounded evaluation of generative LLMs. Existing metrics quantify whether retrieval improves performance, but provide little insight into when models actually use retrieved information and whether different evaluation formats reveal different levels of document

reliance.

Our work addresses this limitation by adapting educational assessment principles to document question answering. We compare different assessment strategies—including recognition-based, abstention-based, and open-response evaluations—to measure document utilization. We show that the estimated benefit of retrieval depends strongly on the assessment format, implying that some assessments (e.g., self-generated multiple choice) may over-estimate pretrained knowledge.

3 Methods

3.1 LLM Selection

To separate document generation from document comprehension, we distinguish between *generator* and *candidate* models. Generator models are responsible for constructing assessment questions and producing reference answers or evaluation criteria, while candidate models are evaluated on the resulting assessments.

We evaluate two lightweight language models in both generator and candidate roles: GPT-5.4-mini (training knowledge cutoff: August 2025) and Gemini-3.1-Flash-Lite (training knowledge cutoff: January 2025) (OpenAI, 2025; Google DeepMind, 2025). Using both models (from different families) as generators and candidates produces four generator-candidate pairings, allowing us to measure not only differences in answering ability but also the influence of question generation.

3.2 Dataset

The document corpus consists of 120 Wikipedia articles sampled uniformly across six temporal buckets (with 20 documents per bucket). Articles were selected to span different stages relative to model training cutoffs in order to vary the amount of parametric knowledge likely available to the models. The six temporal buckets are:

- **Year Ahead:** June 2026 (current), Jan 2026
- **Month Ahead:** Sept 2025, Feb 2025
- **Year Behind:** Aug 2024, Jan 2024

Each time bucket contains an equal number of technical and nontechnical Wikipedia articles, enabling evaluation across both scientific and popular knowledge domains. Representative topics include the Google Willow Processor and the 2025 NBA All-Star Game, respectively.

3.3 Experimental Design

For every document, a generator model first produces the assessment materials. The same generated questions are cached and answered by every candidate model, ensuring that all candidate models are evaluated on an identical question set. Unless otherwise noted, every candidate model completes each experiment twice: once without access to the document and once with the complete document provided as context. This paired evaluation isolates the contribution of external document grounding over parametric knowledge. An example of the experiment workflow can be seen in Figure 1. Example questions are displayed in 2.

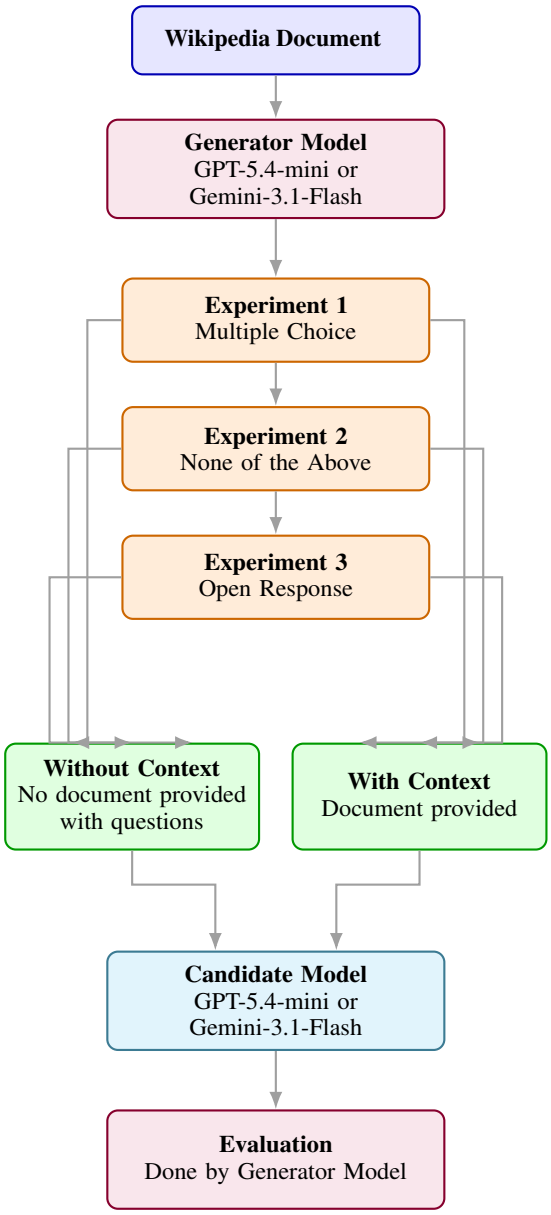


Figure 1: Questions generated from Wikipedia documents are evaluated without and with document context.

Example document: *Cockatiel Color Genetics, Technical, September 2025* Generator: GPT-5.4-mini

Experiment 1: MCQ

Q: What causes White-faced cockatiels to have no psittacofulvin pigments?

A. Pearl gene increases gray pigments
B. **Blue gene deactivates yellow/orange pigments**
C. Cinnamon oxidizes eumelanin
D. Pied removes wing markings

Experiment 2: None of the Above

Q: Which mutation has no visual effect on pigment, but instead changes pigment distribution?

A. White-faced
B. Lutino
C. Cinnamon
D. Pied
E. **None of the Above**

Experiment 3: Open Response

Q: Explain how Cinnamon and Pearl mutations differ in their effects on coloration and inheritance.

Reference answer:
Pearl changes pigment distribution whereas Cinnamon alters eumelanin pigmentation. Both are sex-linked recessive, and male Pearls lose their pearled appearance after the first molt.

Figure 2: Example questions generated from a single document illustrating the three assessment strategies.

3.3.1 Experiment 1: Standard MCQ

Experiment 1 measures standard document question answering performance. For each document, the generator model produces five MCQ, each containing one correct answer and three plausible distractors. The generator also provides the correct option and a reference explanation.

Each candidate model answers every question under both the no-context and with-context conditions by selecting a single answer option and providing a brief explanation.

3.3.2 Experiment 2: None of the Above

Experiment 2 evaluates error detection by asking whether a model can recognize that none of the presented answers are correct. For each document, the generator produces MCQ in which all four options (A-D) are intentionally incorrect but remain plausible distractors derived from the document. A fifth option, *None of the Above*, is appended and is the correct answer for every question.

Candidate models answer each question with and without document context, selecting one option and providing a short explanation. Unlike standard multiple-choice evaluation, success requires rejecting every presented answer rather than selecting the most plausible distractor, providing a direct measure of error detection.

As a follow-up to Experiment 2, we evaluate whether models can reconstruct the missing correct answer rather than simply recognizing that none of the provided options are correct. Responses are evaluated by the generator model using the rubric score described below, where 0 indicates an incorrect answer and 3 indicates complete satisfaction. This provides a stricter test of document-grounded answer generation.

3.3.3 Experiment 3: Open Response

Experiment 3 evaluates unrestricted document-grounded reasoning. For each document, the generator model produces open-ended questions together with a reference answer and a rubric describing the required concepts.

Candidate models answer each question with and without document context using natural language. The generator model acts as an LLM judge and evaluates the factual accuracy of each response using a Likert-style satisfaction rubric (0-3), a commonly used approach for assessing graded levels of performance rather than binary correctness (Rokeman, 2024; Harpe, 2015; Lee et al., 2025). Higher scores indicate closer agreement with the generator model's accuracy criteria based on the document content. Responses are subsequently rescaled to the range $[0, 1]$ to enable comparison with the multiple-choice experiments' scores. Figure 3 shows an example evaluation rubric.

4 Results

Across all experiments, providing document context resulted in statistically significant improvements. For the remainder of the analysis, context gain refers to the within-document difference between rescaled performance $[0, 1]$ with and without access to the source document, and is numerically comparable across assessments. For all graphs, error bars indicate 95% confidence intervals.

4.1 RQ1: How does access to document context influence LLM performance?

Figure 4 summarizes model performance with and without access to the source document across the three assessment strategies.

For standard MCQ (Experiment 1), average accuracy increased from 0.848 without context to 0.976

Example document: *Cockatiel Color Genetics, Technical, September 2025*
Generator: GPT-5.4-mini
Candidate: Gemini-3.1-Flash-Lite

Question
Explain how Cinnamon and Pearl mutations differ in their effects on coloration and inheritance.

Required Accuracy Criteria

- Pigment alteration vs. distribution
- Sex-linked recessive inheritance
- Male Pearl first-molt effect

LLM Judge Accuracy Scale (0–3)

0	Does not satisfy the required concepts.
1	Partially satisfies the required concepts.
2	Mostly satisfies the required concepts with minor inaccuracies.
3	Fully satisfies the required concepts and agrees with the reference answer.

Score: Without Context = **1.6/3**,
With Context = **3.0/3**.

Judge rationale:
The without-context response incorrectly explains avian sex-linked inheritance and misuses X/Y and ZZ/ZW terminology. With document context, the response correctly identifies both mutations as sex-linked recessive and explains the male Pearl first-molt effect.

Figure 3: Example of an abridged LLM rubric evaluation used in Experiment 3. The judge assigns an accuracy score from 0-3. Higher scores indicate closer agreement with the reference answer.

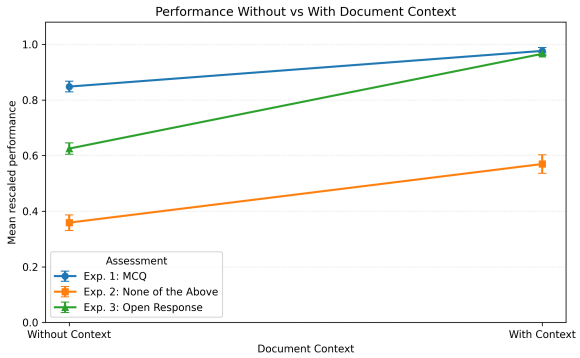


Figure 4: Performance with and without document context across the three assessment strategies.

with context, corresponding to a mean improvement of 0.128 (+15.1% relative gain; $t(479) = 14.83$, $p < 0.001$). While baseline performance was already high, document access nearly eliminated remaining errors.

The improvement was larger in the None of the Above task (Experiment 2). Accuracy increased from 0.359 to 0.570, representing a mean gain of 0.211 (+58.8% relative gain; $t(479) = 14.71$, $p < 0.001$). This suggests that identifying when every presented answer is incorrect benefits more from document grounding than the traditional multiple-choice evaluation. Despite generating the assessment questions itself, the model was substantially less effective at answering the None of the Above questions than standard multiple-choice questions. The low baseline performance on this task and the fact that it is well under other methods, even with context, suggests that alternate prompts might yield even larger context gains.

The largest absolute improvement occurred in the open-response evaluation (Experiment 3). Performance increased from 0.625 to 0.966, corresponding to an average improvement of 0.341 (+54.5% relative gain; $t(479) = 34.69$, $p < 0.001$). The larger gain indicates that unrestricted generation benefits more from document access than recognition-based assessment.

Open-response evaluation produced the largest absolute context gain, while the None of the Above assessment exhibited the largest relative improvement over its no-context baseline. Collectively, these findings demonstrate that document context consistently improves model performance, with increasingly larger effects as assessment formats require greater reasoning and generation.

4.2 RQ2: Which assessment strategy best distinguishes document-grounded reasoning?

To quantify the influence of document context, we compared context gain across the three assessment strategies. As shown in Figure 5, context gain differed significantly across assessment formats (one-way ANOVA, $F = 91.27$, $p < 0.001$). Post-hoc Tukey comparisons indicated that all pairwise differences were significant ($p < 0.01$).

Standard MCQ exhibited the smallest context gain (0.128), followed by the None of the Above task (0.211), while open-response evaluation produced the largest context gain (0.341). Although the None of the Above assessment showed the largest proportional improvement over its lower no-context baseline (58.8%), open-response evaluation yielded the greatest absolute gain from document context. These results indicate that conventional MCQ are less sensitive to the contribution of external documents, whereas open-ended generation provided the clearest separation between document-grounded reasoning and responses generated from parametric knowledge.

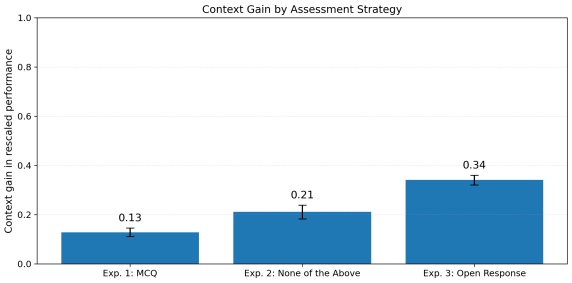


Figure 5: Mean context gain for each assessment strategy.

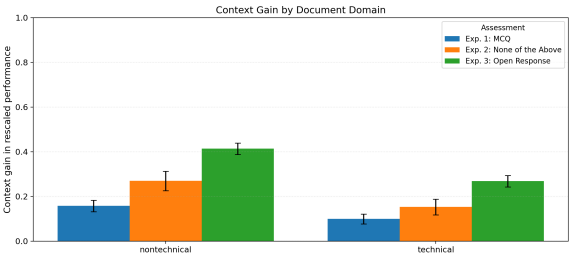


Figure 6: Context gain across technical and nontechnical documents.

4.3 RQ3: How do document characteristics (content, recency) influence context utilization?

We next examined whether document type and document recency influenced reliance on external context. Across every assessment strategy, nontechnical documents produced larger context gains than technical documents Figure 6. For open-response questions, context gain reached 0.413 for nontechnical documents compared with 0.268 for technical documents. Similar patterns were observed for standard MCQ (0.158 vs. 0.099) and None of the Above (0.269 vs. 0.153). These results suggest that models may need less context on technical topics, since the material can be more consistent over time.

Document modification period also affected context utilization Figure 7. Older articles generally exhibited larger context gains than more recent articles. For example, open-response gains were highest for August 2024 documents (0.444) and lowest for June 2026 documents (0.288). While the relationship is not strictly monotonic across all temporal buckets, this finding suggests that the recency of content affects not just direct knowledge but also context use. These results raise the question of whether stronger pretrained knowledge of a topic enables models to leverage newly provided document context more effectively.

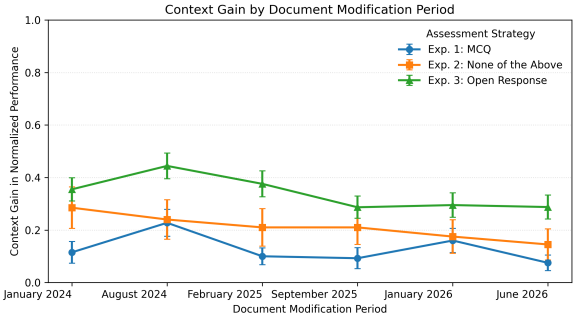


Figure 7: Context gain across document modification periods. Results are averaged across Gemini and ChatGPT and shown by month. Interpretation of the aggregated trend should account for the different model knowledge cutoffs (January 2025 and August 2025, respectively).

4.4 Missing Answer Generation

As a follow-up to the None of the Above experiment, we evaluated whether models could reconstruct the missing correct answer after recognizing that no listed option was correct. Performance was consistently high across both document domains, evaluated using the previously described rubric score out of 3. Mean scores between generated answers and the reference answers were 2.53 for both technical and nontechnical documents, with narrow confidence intervals and negligible differences between domains. These results indicate that once models recognize that the correct answer is absent, they are generally capable of reconstructing semantically equivalent answers directly from the document.

4.5 Model Pairings

We also compared performance across our generator-candidate model pairings to determine whether observed effects were consistent across model families. Averaged across all assessment strategies and generator models, candidate model performance was similar. Without document context, GPT-5.4-mini achieved a mean score of 0.586, compared with 0.635 for Gemini-3.1-Flash-Lite. With document context, models achieved mean scores of 0.825 and 0.850, respectively. Overall, trends remained consistent regardless of the LLM used for question generation and answering, suggesting that the observed improvements are primarily driven by assessment strategy and document context rather than specific model type.

5 Conclusion

The proposed assessment framework reveals complementary aspects of model behavior. Experiment 2, None of the Above, measures whether a model can recognize a case where all provided options are incorrect, while the follow-up missing-answer generation task evaluates whether it can reconstruct the correct information once recognition occurs. Together with Experiment 3, open-response evaluation, these assessments distinguish LLM recognition, error detection, answer generation, and free-form reasoning, providing a more comprehensive characterization of document utilization than standard MCQ (Experiment 1) accuracy alone. Although document context consistently improved performance across all experiments, the magnitude of improvement varied substantially with assessment strategy. Standard MCQ showed only modest gains and ceiling effects, while None of the Above and open-response evaluation produced larger improvements. Consequently, evaluations based solely on multiple-choice accuracy may underestimate the contribution of external documents.

Document characteristics also influenced context utilization. Nontechnical documents consistently produced larger context gains than technical documents, suggesting that models benefited more from explicit grounding when reasoning about general interest topics than specialized technical content, which could be because these subjects are associated with more diverse discourse and evolving contextual information, as compared to structured technical domains. In addition, context gains were generally smaller for more recently modified documents. Additional experiments would be needed to determine the factors driving this effect.

Finally, we observed relatively small differences across generator-candidate model pairings compared with the much larger effects of assessment strategy. This suggests that the proposed evaluation framework is largely model-agnostic, and captures properties of document-grounded reasoning rather than artifacts of a particular language model family.

6 Future Work

This framework provides a useful benchmark for evaluating document-grounded language generation and contributes toward more rigorous and interpretable assessment of retrieval-augmented language models. Our findings complement recent work on LLM self-assessment that seek to estimate a model's prior knowledge before document retrieval via metacognition (Deshpande et al., 2026; Iqbal, 2026; Zhou et al., 2024). Together, these directions move toward evaluation frameworks that can determine both what a model already knows and when external documents meaningfully contribute additional knowledge.

This work has implications beyond document question answering, including document-grounded classification, explanation generation, educational tutoring, and other RAG-based applications where distinguishing retrieved evidence from parametric knowledge is essential to interpret model behavior (Akbarova et al., 2026; Modran et al., 2024). The present framework intentionally adopts a more rigid assessment strategy, providing a high-level characterization of document utilization across multiple task formats. Future work could investigate adaptive evaluation strategies that iteratively probe model understanding to identify the boundary between parametric and contextual knowledge. Similar to adaptive testing or oral examinations, such approaches could progressively refine questions until the point at which retrieval becomes necessary, distinguishing background knowledge from information that depends on the provided document (Cheng et al., 2024; Zhuang et al., 2023).

Limitations

This work has several limitations. First, our evaluation corpus consists exclusively of Wikipedia articles. While Wikipedia provides high-quality and varied documents, it may not capture expert complexity or the nature of real-world retrieval settings, such as enterprise knowledge bases, medical records, legal documents, etc.

Second, our temporal analysis is based on each article's most recent revision date. Although documents were selected from a year before, a month after, and a year after model training cutoffs, information in later revisions may have existed in earlier versions or in other publicly available sources. Consequently, revision date serves only as a proxy for model familiarity rather than a definitive measure of whether particular knowledge was available during pretraining.

Third, generated-answer quality was evaluated using an LLM-as-a-judge framework. While this provides a scalable and consistent evaluation procedure, it may introduce systematic biases and may not perfectly align with human judgment of perfor-

mance. Furthermore, because the assessment questions were themselves generated by LLMs, they may be more closely aligned with the capabilities of language models than with human-authored assessments, as suggested by the high baseline performance in Experiment 1 without document context.

Finally, our experiments focus on two proprietary model families (GPT and Gemini). Future work should extend this evaluation framework to newer frontier models, additional commercial systems such as Claude, and open-source models including Llama and Qwen. Evaluating a broader range of architectures would help determine how well these findings generalize across different model families and deployment settings.

Ethical Considerations

All evaluation documents were collected from Wikipedia, whose content is distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license. No human subject data was used.

We used commercially available LLMs, accessed through their respective APIs. All evaluations were fully automated and no model outputs were used to make decisions affecting individuals.

Experiments were parallelized across multiple CPU processes on a high-performance computing cluster. Parallelization was used solely to reduce runtime and did not alter the evaluation protocol or model outputs. The complete evaluation framework, prompts, and analysis code will be made publicly available to support reproducibility.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Sabinakhon Akbarova, Felix Dobslaw, and Robert Feldt. 2026. Understanding on the edge: Llm-generated boundary test explanations. In *Proceedings of the 7th ACM/IEEE International Conference on Automation of Software Test*, pages 23–33.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418.

Randy Elliot Bennett, Donald A Rock, and Minhwei Wang. 1991. Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28(1):77–92.

Rajesh Bhayana. 2024. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1):e232756.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. 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*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200.

Cheng Cheng, GuanHao Zhao, Zhenya Huang, Yan Zhuang, Zhaoyuan Pan, Qi Liu, Xin Li, and Enhong Chen. 2024. Towards explainable computerized adaptive testing with large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2655–2672.

Anura Deshpande, Celine Cerezci, Benjamin Nye, Mark G Core, Suvaditya Mukherjee, and Joshua Shay Kricheli. 2026. Have (a) i seen this before? exploring llm metacognition using self-assessed rankings and scoring. In *The International FLAIRS Conference Proceedings*, volume 39.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Robert B Frary. 1985. Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement*, 22(1):21–31.

Google DeepMind. 2025. Gemini 3 flash model card. https://deepmind.google/models/model-cards/. Accessed: 2026-07-13.

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

Spencer E Harpe. 2015. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Sajid Iqbal. 2026. Do ai know what they know? exploring metacognition in llms. *Intelligent Data Analysis*, page 1088467X261436903.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 15696–15707, Honolulu, Hawaii, USA. PMLR.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809.

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*, volume 33, pages 9459–9474.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Horia Alexandru Modran, Ioana Corina Bogdan, Doru Ursuțiu, Cornel Samoilă, and Paul Livius Modran. 2024. Llm intelligent agent tutoring in higher education courses using a rag approach. In *International Conference on Interactive Collaborative Learning*, pages 589–599. Springer.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6129–6139.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18477–18494.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Rajesh Pasupuleti, Ravi Vadapalli, Christopher Mader, and Norris Timothy. 2024. Popular llm-large language models in enterprise applications.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. Association for Computational Linguistics.

Nor Roselidyawaty Mohd Rokeman. 2024. Likert measurement scale in education and social sciences: explored and explained. *EDUCATUM Journal of Social Sciences*, 10(1):77–88.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, and 1 others. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Mark G Simkin and William L Kuechler. 2005. Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1):73–98.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Erdi Okan Yılmaz and Türker Toker. 2022. Analysing the effects of assessment and evaluation applications and exam formats in distance education. *International Journal of Psychology and Educational Studies*, 9(1):165–176.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust LLM evaluators. *arXiv preprint arXiv:2405.11966*.

Min Zhao, Inez Y Oh, Aditi Gupta, Sally Cohen-Cutler, Kathryn M Harmoney, Albert M Lai, and Bryan A Sisk. 2026. Automating evaluation of llm-generated responses to patient questions about rare diseases. *JAMIA open*, 9(2):ooag054.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1453–1463.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and 1 others. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective.