Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6 n. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

# Reliable Question Answering for Learning: When Should Educational AI Answer or Abstain?

**Ludovico Boratto [1], Francesco Congiu [1,2], Gianni Fenu [1], Mirko Marras [1,*] and Giacomo Medda [1]**

[1] University of Cagliari; ludovico.boratto@acm.org, francesco.congiu6@unica.it, gianni.fenu@unica.it, mirko.marras@unica.it, giacomo.medda@unica.it
[2] University of Macerata; f.congiu@unimc.it
* Correspondence: mirko.marras@unica.it

**Abstract:** Retrieval-augmented generation (RAG) is increasingly relevant to educational applications in which learners use conversational artificial intelligence (AI) to access course materials, institutional documents, textbooks, and other instructor-selected resources. In these settings, a useful response should not only be plausible or factually correct: it should also be supported by the materials that define the learning context. Conversely, when those materials do not provide sufficient evidence, refraining from answering may be more appropriate than generating a fluent but unsupported explanation. In this study, we examine this relationship through an evidence-aware evaluation framework that distinguishes the availability of supporting information from the subsequent behavior of the language model. Using three question-answering benchmarks as a controlled testing environment, we evaluate six open-weight language models under a common retrieval and abstention-aware generation protocol. Results show that providing more retrieved material has asymmetric effects. When supporting evidence is available, larger contexts generally reduce unnecessary abstention; when support is absent, additional context can instead increase unsupported and hallucinated responses. Within evidence-supported contexts, unnecessary abstention is also associated with lower evidence redundancy, while the position of sparse evidence affects several models and datasets. These findings suggest that the design of educational RAG systems should not focus exclusively on retrieving more information or maximizing answer coverage. More importantly, retrieval and generation should be coordinated so that learner-facing responses remain aligned with the resources made available to the system.

**Keywords:** Retrieval-Augmented Generation; Educational Artificial Intelligence; Question Answering; Large Language Models; Learning Support.

**Copyright:** © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Large language models are becoming conversational interfaces through which learners ask questions, request explanations, and navigate educational information. In a learning environment, these interactions may concern lecture notes, textbooks, syllabi, assessment instructions, institutional regulations, or other resources selected by instructors and educational institutions. This creates an important distinction between general-purpose question answering and educationally situated question answering. In the latter, a response should not only be plausible or factually correct in the abstract; it should also remain consistent with the materials that define the learning activity and the environment in which the learner is expected to operate.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Retrieval-augmented generation (RAG) provides a natural architecture for this purpose. A specific retrieval component selects potentially relevant passages from a document collection, and a language model uses those passages when producing a response (Fan et al., 2024). This separation is particularly attractive for educational applications because the information required by learners may be course-specific, institution-specific, frequently updated, private, or intentionally restricted to a curated set of resources. Rather than relying exclusively on knowledge encoded in the model parameters, a RAG-based assistant can be given the same materials that instructors expect students to consult. In this way, the system can function less as an unrestricted source of information and more as an interface to a knowledge space.

The presence of a retrieval component, however, does not guarantee that the resulting support is grounded in those resources. Existing evaluation frameworks commonly consider retrieval relevance, answer correctness, answer relevance, and faithfulness as complementary dimensions (Es et al., 2024; Saad-Falcon et al., 2024). These measures are important, but they do not fully capture whether the model's decision to answer is appropriate to the evidence actually available in the retrieved context. A model may produce a correct response even when no supporting passage has been retrieved, potentially relying on parametric knowledge. Conversely, it may abstain despite having received sufficient information to answer. Previous work on the interaction between parametric and contextual knowledge shows that language models can combine, prioritize, or disregard these information sources in different ways (Farahani & Johansson, 2024; Jin et al., 2024; Xu et al., 2024).

This distinction has particular educational significance. When an artificial intelligence (AI) assistant mediates access to instructor-selected or institutionally controlled resources, a correct but unsupported answer is not necessarily equivalent to a correct answer grounded in those materials. The former may introduce explanations, terminology, or interpretations that are external to the intended learning context, while offering the learner little indication of where the information originates. The opposite failure is also relevant: if the required information is already available in the retrieved material but the system nevertheless refuses to answer, the learner is unnecessarily denied access to knowledge the educational environment was designed to provide. From this perspective, abstention cannot be interpreted as inherently desirable or undesirable. Its appropriateness depends on whether the available evidence is sufficient to support a response. Recent work similarly emphasizes the importance of distinguishing warranted uncertainty from unnecessary refusal (Madhusudhan et al., 2025; Peng et al., 2025; Zhang et al., 2024).

A further design challenge concerns how much material should be presented to the model. Increasing retrieval depth may appear beneficial because it raises the probability that relevant information is included in the context. At the same time, larger contexts introduce additional passages that may dilute, compete with, or obscure the evidence needed to answer a question. Research on long-context language models has shown that model behavior can be sensitive both to distracting information and to the position at which relevant evidence appears (An et al., 2024; He et al., 2024; Liu et al., 2024). In an educational setting, this means that decisions about retrieval depth and context construction are not mere technical configuration choices. They can influence whether knowledge contained in the learning resources is actually surfaced to the learner through the conversational interface. This problem becomes even more relevant when the retrieved context contains only limited

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

support. A single relevant passage may be harder for the generator to exploit than several mutually reinforcing passages, while evidence appearing later in a long context may be less effectively used than evidence presented earlier. In this sense, successful retrieval is only an intermediate condition for educational support. What ultimately matters is whether the retrieved information is sufficiently accessible to the generator to become accessible, in turn, to the learner. The distinction between retrieved information and usable information is therefore central to the design.

Against this background, in this study, we adopt an evidence-aware perspective on retrieval-augmented question answering. Rather than evaluating generation independently of retrieval, we first determine whether the retrieved context contains information capable of supporting an acceptable answer and then interpret the model's behavior relative to that evidence state. This makes it possible to distinguish supported correct answers, incorrect answers despite available support, unnecessary abstentions, justified abstentions, and answers produced without retrieved support. The analysis subsequently examines how these behaviors change as more material is retrieved and, when supporting information is available, how its redundancy and position are associated with the likelihood that the model makes effective use of it.

The evaluation uses well-known open-domain question-answering benchmarks as a controlled experimental setting, rather than as a direct simulation of classroom interaction. This choice makes it possible to manipulate retrieval depth and characterize evidence availability in a reproducible way while keeping the underlying question-answering task stable. The objective is therefore methodological and pedagogical rather than instructional-effectiveness oriented: the study does not test whether the systems improve learning outcomes, but investigates behaviors that are prerequisites for trustworthy educational use. More broadly, the study contributes to educational technology by framing RAG reliability as a relationship among the learning resources made available to the system, the way those resources are retrieved and organized, and the answering behavior presented to the learner. This perspective is relevant whenever conversational AI is expected to support learners in navigating instructor-selected, course-specific, or institutionally controlled knowledge. In such settings, a reliable assistant is the one whose willingness to answer remains aligned with what the educational resources can actually support.

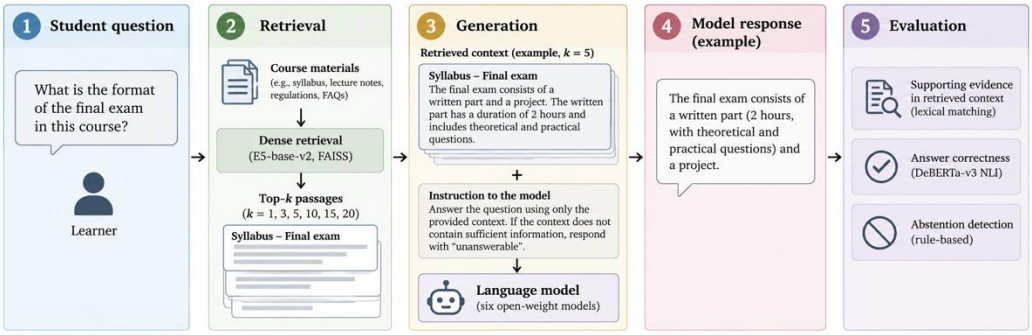
## 2. Materials and Methods

### *2.1. Study Design*

This study investigates a basic requirement for the use of retrieval-augmented generation in educational settings: whether an AI assistant provides information when the resources available to it support an answer and refrains from unsupported generation when they do not. The evaluation is computational and does not involve students, instructors, or classroom interventions. Established question-answering benchmarks are instead used as a controlled environment in which the evidence availability and system behavior can be examined under reproducible conditions. This distinction is important for the interpretation of the study. The experiments do not measure learning gains, teaching effectiveness, or student experience. Rather, they examine behaviors that are relevant before a RAG system can be considered appropriate for learner-facing use. In particular, an educational assistant that is expected to mediate access to course or institutional resources should be able to

exploit relevant information when it is available while recognizing when retrieved material is insufficient. To isolate these behaviors, the same retrieval process, prompt structure, generation conditions, and evaluation procedure are applied across datasets and language models. The resulting pipeline is shown in Figure 1.

**Figure 1. Evidence-aware RAG pipeline for educational question answering**. A learner question is matched against course materials through dense retrieval, and the top-k passages are provided to the language model together with an abstention-aware instruction. The generated response is then evaluated with respect to the presence of supporting evidence in the retrieved context, answer correctness, and abstention behavior. The example illustrates how the pipeline could mediate access to course-specific information while distinguishing supported answers from unsupported or unnecessary responses.




### *2.2. Experimental Pipeline*

For each question, the system first retrieves a ranked set of passages from a reference collection. The top-k passages are then provided to a language model together with a fixed instruction asking it to answer when the retrieved context contains sufficient information and to abstain otherwise. The generated response is subsequently examined along three dimensions: whether the retrieved context contains supporting evidence, whether the response is correct, and whether the model abstains. Figure 1 illustrates this process through an educational example in which a learner asks about the format of a final examination. In a real educational deployment, the collection consists of syllabi, lecture notes, regulations, frequently asked questions, or other resources selected by instructors or institutions. In the controlled evaluation, the same logic is reproduced using open-domain question-answering benchmarks and a common reference corpus.

The pipeline is intentionally kept simple. Query rewriting, reranking, iterative retrieval, and additional reasoning modules are not included. This design reduces the number of intermediate components that could influence the observed behavior and makes the relationship between retrieved evidence and generated responses easier to interpret. The procedure is repeated across retrieval depths and generator models.

### *2.3. Question-Answering Benchmarks and Retrieval Setting*

The adopted evaluation uses three established open-domain question-answering datasets: *Natural Questions* (Kwiatkowski et al., 2019), *TriviaQA* (Joshi et al., 2017), and *WebQuestions* (Berant et al., 2013). The experimental subsets contain 3,610 Natural Questions queries, 11,313 TriviaQA queries, and 2,032 WebQuestions queries. Each dataset provides open-ended questions together with one or more acceptable answers. These benchmarks provide reference answers that make it possible to determine, in a consistent manner, whether retrieved passages contain observable support for a question and whether a generated response can be

considered acceptable. Using three datasets also allows the analysis to test whether the observed patterns are confined to a single question-answering collection. Retrieval is performed over the *wiki18_100w* collection distributed by Jin et al., 2024, which contains approximately 21 million Wikipedia passages segmented into chunks of 100 words. Passages are represented using *E5-base-v2* embeddings and searched through a flat *FAISS* index. The ranking returned by the retriever is used directly. To examine how the amount of available material affects generation, retrieval depth is varied across k=1,3,5,10,15,20. The retrieved passages are concatenated in their original ranking order. This produces a controlled progression from relatively focused contexts to broader contexts containing a larger amount of potentially useful, but also potentially distracting, information.

### *2.4. Generator Models and Answering Conditions*

Six open-weight instruction-tuned language models are evaluated: *Qwen-2.5 7B*, *Llama-3.1 8B*, *Qwen-2.5 32B*, *Gemma-4 31B*, *Mixtral 8×7B*, and *Llama-3.3 70B*. The models were selected to provide variation in model family and operational scale while remaining compatible with a common local inference workflow. The same retrieved passages and substantive instruction are provided to every model. Only the formatting required by the corresponding chat template is adapted. The prompt asks the model to answer when the retrieved context contains sufficient information and to return an unanswerable response when it does not. Generation is limited to 128 tokens. The use of multiple models is not intended to establish a performance ranking. Instead, it allows the analysis to determine whether the relationship between evidence availability and answering behavior recurs across different model families and scales or depends strongly on an individual generator.

### *2.5. Evidence Availability, Correctness, and Abstention*

For every question and retrieval depth, the retrieved context is first examined to determine whether it contains supporting information. Evidence availability is operationalized through a normalized exact matching. Retrieved passages and acceptable answers are normalized, and a passage is considered supporting when it contains at least one acceptable answer. This operational definition was chosen because evidence availability is central to the analysis and should therefore remain directly inspectable. When a context is classified as containing support, the corresponding answer string can be located in at least one retrieved passage. At the same time, the criterion is intentionally conservative: passages that express an answer only implicitly or through paraphrase may not be detected. Generated-response correctness is instead assessed separately from evidence availability. Since model outputs can be longer and more elaborated than the short reference answers, correctness is evaluated using a *DeBERTa-v3-large* natural-language-inference model. The question, generated answer, and reference answer are used to determine whether the response entails an acceptable answer under the calibrated evaluation criterion. Finally, abstention is identified before answer correctness is assessed. Responses are matched against a predefined set of refusal expressions, including the requested unanswerable response and equivalent formulations indicating that the available context is insufficient. When one of these patterns is detected, the output is classified as an abstention; all remaining responses are evaluated as (in)correct.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

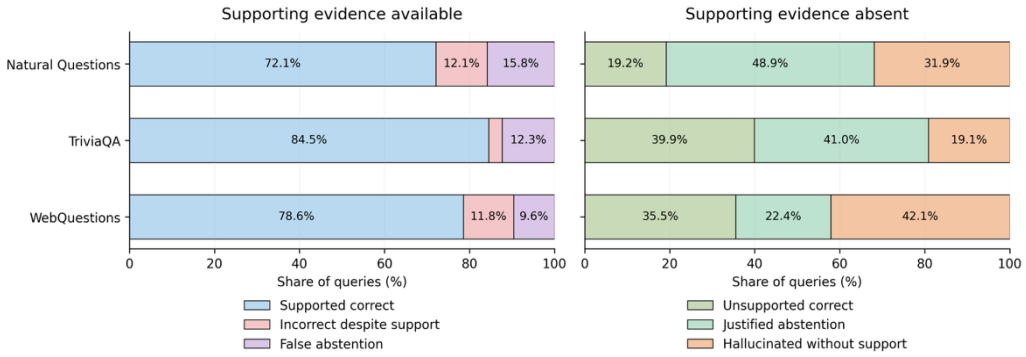
Combining these dimensions provides the evidence-aware interpretation used throughout the study. When supporting evidence is available, the model may produce a supported correct answer, an incorrect answer despite available support, or a false abstention. When evidence is absent, the model may produce a correct but unsupported answer, a hallucinated answer, or a justified abstention. We use the term unsupported correct answer rather than the more-common parametric answer because the experiment establishes only that retrieved support is absent; it does not directly identify which internal knowledge produced the response. This distinction is relevant for educational use. If an AI assistant is expected to help navigate a defined body of material, a correct answer that is unsupported by the retrieved resources is not equivalent to a correct answer that is traced back to those resources. Likewise, abstention is appropriate only when the available material is genuinely insufficient.

## 3. Experimental Results

The results show that increasing the amount of retrieved material does not translate into uniformly more reliable learner-facing behavior. At the largest retrieval depth (k=20), most queries for which supporting evidence is available are answered correctly, as illustrated in Figure 2. Supported correct answers account for 72.1% of Natural Questions, 84.5% of TriviaQA, and 78.6% of WebQuestions, averaged across the six evaluated generators. False abstention nevertheless remains present in all three datasets, affecting 15.8%, 12.3%, and 9.6% of evidence-supported cases, respectively. Thus, even when the retrieval stage successfully provides information that can support an answer, that information is not always converted into assistance.

The complementary evidence-absent condition reveals a substantially different pattern. When no retrieved passage contains supporting information, abstention is not consistently the dominant behavior. Hallucinated answers account for 31.9% of Natural Questions, 19.1% of TriviaQA, and 42.1% of WebQuestions, whereas justified abstention reaches 48.9%, 41.0%, and 22.4%, respectively. Correct answers without retrieved support are observed with non-negligible frequency. Although such answers may be acceptable in unrestricted question answering, they do not demonstrate that the response is grounded in the resources made available to the system. This distinction is relevant in educational settings where learners reasonably assume that an assistant is answering from the selected learning materials.

**Figure 2. Evidence-conditioned answering behavior at the maximum retrieval depth**. Distribution of generation outcomes at k=20, averaged across the six models. The left panel reports responses when at least one supporting passage is available in the retrieved context, distinguishing supported correct answers, incorrect answers despite available support, and false abstentions. The right panel reports responses when supporting evidence is absent, distinguishing unsupported correct answers, justified abstentions, and hallucinated answers.

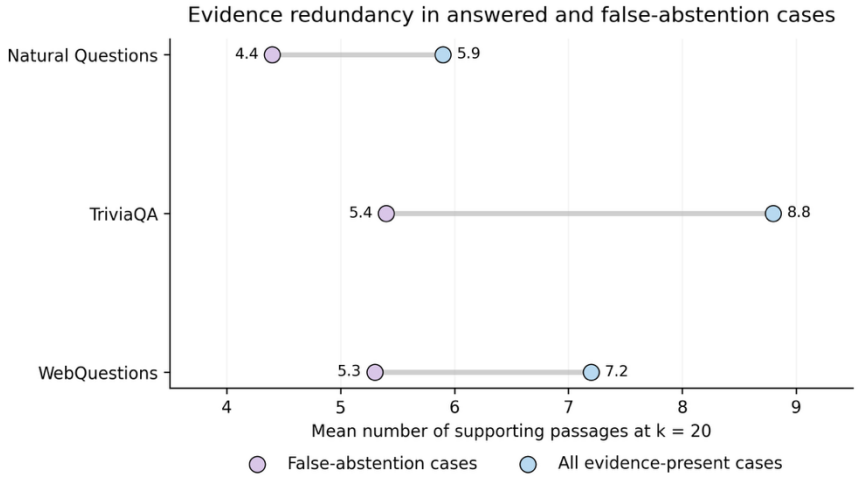


EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Considering the full range of retrieval depths helps explain why these two evidence conditions need to be interpreted together. As more passages are retrieved, false abstention generally decreases when supporting information is available, particularly for medium- and larger-scale generators. Additional context can therefore help the model exploit evidence that has entered the retrieved set. When supporting evidence remains absent, however, increasing retrieval depth is associated with more unsupported generation, with hallucination rising before generally stabilizing at larger context sizes. Increasing the amount of context can thus have opposite consequences depending on what the retrieved material actually contains: it can improve access to useful information while also increasing the likelihood that the system provides an unsupported answer. From an educational perspective, this asymmetry is important because retrieval depth determines more than how much information is technically made available to the model. It also influences the kind of support ultimately presented to the learner. A configuration that encourages answering can be beneficial when the retrieved learning resources contain sufficient evidence, but problematic when they do not. Retrieval depth should therefore be calibrated together with evidence availability and abstention behavior rather than treated as an isolated technical parameter.

A second pattern concerns the strength with which relevant information is represented in the retrieved context. Figure 3 compares the average number of supporting passages across all evidence-present queries with the corresponding average among cases resulting in false abstention. Across all datasets, unnecessary abstention is associated with lower evidence redundancy. In Natural Questions, the mean number of supporting passages decreases from approximately 5.9 across all evidence-present cases to 4.4 among false-abstention cases. TriviaQA shows the largest difference, from 8.8 to 5.4, while WebQuestions decreases from 7.2 to 5.3.

**Figure 3. Evidence redundancy associated with false abstention**. Comparison of the mean number of supporting passages available at k=20 among all evidence-present queries and among the subset resulting in false abstention, averaged across the six evaluated models. The distance between the two markers highlights the reduction in evidence redundancy associated with unnecessary abstention; values next to the markers report the corresponding mean number of supporting passages.




The instance-level analysis is consistent with this pattern. False abstention is generally most frequent when only one supporting passage is available and tends to

decrease as the number of passages increases, with the clearest relationship observed for Natural Questions and TriviaQA. Redundant evidence should not, however, be interpreted as sufficient for successful answering: false abstentions remain possible even when several supporting passages are present. Rather, repeated support appears to make the relevant information easier for the generator to exploit.

The position of supporting information provides an additional qualification. When the analysis is restricted to queries containing exactly one supporting passage, hence separating position from redundancy, TriviaQA shows the clearest positional sensitivity: false abstention becomes more frequent once the relevant passage moves beyond the first five retrieved positions, particularly for smaller and medium-sized models. Natural Questions exhibits a similar, yet less regular, tendency. Specifically, WebQuestions behaves differently, showing no stable monotonic degradation as the supporting passage moves deeper. In many conditions, evidence at intermediate ranks performs comparably to or better than evidence presented earlier.

Taken together, these findings indicate that retrieving relevant information is not equivalent to making that information available through the interaction. Supporting evidence is more likely to result in an answer when it is represented more strongly in the retrieved context and, for some datasets, when sparse evidence appears earlier in the ranked sequence. For the educational RAG systems, this distinction is particularly important: a learning resource may be technically accessible to the system while still failing to become accessible to the learner. The design of learner-facing assistants should therefore consider not only whether relevant material has been retrieved, but also how prominently, repeatedly, and coherently that material is presented to and used by the generator.

## 4. Discussion

The present study examined RAG reliability from the perspective of whether the generator's behavior corresponds to the evidence retrieval has made available. This perspective is especially relevant for educational assistants that are expected to mediate access to defined learning resources rather than answer exclusively from general model knowledge. Three main considerations emerge from the findings.

The first takeaway is that educationally reliable answering requires evaluating the response together with the evidence state. Existing RAG evaluation approaches have made important progress by separating retrieval relevance, answer relevance, and faithfulness (Es et al., 2024; Saad-Falcon et al., 2024), while recent work has emphasized unanswerability and abstention (Madhusudhan et al., 2025; Peng et al., 2025). The present results show why these perspectives need to converge. An answer can be correct without being supported by the retrieved resources, while an abstention can be either responsible or unnecessary depending on whether those resources already contain sufficient information. In an educational application, this distinction is consequential. If a learner asks about a course-specific policy or concept, a fluent answer derived from external parametric knowledge may not be equivalent to an answer grounded in the instructor-selected material. Conversely, a system that repeatedly withholds information contained in the course resources introduces an avoidable barrier to access. Reliability should therefore be understood as alignment between available evidence and model behavior.

A second takeaway is that retrieval depth should be treated as part of the instructional configuration of an educational assistant rather than simply maximized.

Larger contexts improve the chance that relevant information is retrieved and can reduce false abstention when evidence is available. At the same time, the results show that deeper retrieval can also make unsupported answering more frequent when evidence remains absent. This complements previous research showing that language models do not use long contexts uniformly and may be sensitive to distractors and evidence position (An et al., 2024; He et al., 2024; Liu et al., 2024). For educational technology, the implication is that retrieval and generation should be configured jointly. Reranking, evidence selection, adaptive retrieval depth, or explicit source highlighting may help ensure that the additional material given to learners' assistants contributes relevant support rather than increasing context size.

A third takeaway is that information availability and information usability are distinct properties. We showed that false abstention is associated with a deficit in evidence redundancy and with less favorable evidence positions. A conventional retrieval metric would count all of these queries as successful because at least one supporting passage appears in the retrieved set. From the learner's perspective, however, a retrieved passage that the generator effectively ignores does not provide the intended access to information. This extends the long-context literature from a model-performance issue to an educational-design concern: when AI becomes the interface through which students navigate learning resources, document selection, ordering, repetition, and signaling can influence which information is actually surfaced. Could you please replace with "Some exceptions also caution against treating position bias as a universal phenomenon.". Context design should be empirically evaluated for the type of questions, sources, and models used.

These findings also clarify the role that abstention could play in educational assistants. A system that always answers may appear helpful but can provide unsupported explanations; a system optimized to avoid hallucination may become excessively conservative and withhold useful information. Appropriate abstention should instead be conditional on evidence sufficiency. In practice, an educational system might also make abstention pedagogically informative (for example, by indicating that the available course materials do not support an answer, identifying which sources were checked, or suggesting that the learner consult an instructor).

## 5. Conclusions

In this study, we introduced an evidence-aware perspective for evaluating retrieval-augmented question answering with relevance to AI-supported learning. Findings show that increasing retrieval depth does not uniformly improve reliability: larger contexts generally reduce unnecessary abstention when supporting evidence is available, but can simultaneously increase hallucinated answering when that evidence is absent. Within evidence-supported contexts, false abstention is associated with lower evidence redundancy, while the position of sparse evidence further affects several models. These results suggest that the design of educational RAG systems should move beyond the assumption that successful retrieval or larger contexts are sufficient. What ultimately matters is whether the generator's decision to answer or abstain is appropriate to the learning resources it has been given.

The limitations of the study directly motivate its next educational extensions. First, evidence availability is currently identified through exact answer matching; this makes the criterion transparent but can miss implicit or paraphrased support, motivating semantic evidence labeling that remains inspectable by educators.

Second, abstention is identified through rule-based language patterns and should be validated with human judgments, especially for responses that combine caveats with partial answers. Third, the present analyses are observational, so redundancy and position are associated with false abstention but are not established as its causal mechanisms; controlled manipulation of evidence number, order, and distractor content is therefore needed. Fourth, the evaluation uses open-domain QA benchmarks rather than fully-authentic educational queries. Future studies should evaluate course-grounded assistants using teacher-provided resources, naturally occurring student questions, and instructor judgments of whether an answer is appropriately supported. Fifth, the study assesses system behavior rather than learner outcomes. Classroom-oriented work should examine whether evidence-aware answering and pedagogically designed abstention improve trust, source use, and the learner's ability to recognize the boundaries of available information. Finally, different educational contexts may assign different costs to unsupported answers and unnecessary abstention; future human-in-the-loop evaluation should therefore examine how instructors wish to calibrate this trade-off for different activities.

## References

An, S., Ma, Z., Lin, Z., Zheng, N., Lou, J.-G., & Chen, W. (2024). Make your LLM fully utilize the context. Advances in Neural Information Processing Systems, 37, 62160–62188.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on Freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1533–1544).

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (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 (pp. 150–158). Association for Computational Linguistics.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6491–6501).

Farahani, M., & Johansson, R. (2024). Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 16966–16977).

He, J., Pan, K., Dong, X., Song, Z., Liu, Y., Sun, Q., Liang, Y., Wang, H., Zhang, E., & Zhang, J. (2024). Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp. 13628–13642).

Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., Li, Q., & Zhao, J. (2024). Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Proceedings of LREC-COLING 2024 (pp. 16867–16878).

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (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 (pp. 1601–1611).

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 453-466.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157–173.

Madhusudhan, N., Madhusudhan, S. T., Yadav, V., & Hashemi, M. (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 (pp. 9329–9345).

Peng, X., Choubey, P. K., Xiong, C., & Wu, C.-S. (2025). Unanswerability evaluation for retrieval augmented generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (pp. 8452–8472).

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In Proceedings of NAACL 2024 (pp. 338–354).

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., & Xu, W. (2024). Knowledge conflicts for LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8541–8565).

Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H., & Zhang, T. (2024). R-Tuning: Instructing large language models to say "I don't know." In Proceedings of NAACL 2024 (pp. 7113–7139).