Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Learning with Caution: Assessing LLM Performance on Answerable and Unanswerable Questions

Aditya Raj[a], Gokul S Krishnan[b], Sanjay Bankapur[c], Abhilash C B[d], Manjunath K Vanahalli[a*]

[a]*Department of Data Science and Artificial Intelligence, Indian Institute of Information Technology Dharwad, India*

[b]*Centre for Responsible AI (CeRAI), Indian Institute of Technology Madras*

[c]*Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal, India*

[d]*Department of Computer Science and Engineering, JSS Academy of Technical Education Bengaluru, India*

*manjunath.k.vanahalli@gmail.com

Abstract: In the age of Artificial Intelligence, Large Language Models (LLMs) have become mighty question-answering (QA) tools, which increasingly shape the way students find and use information. Despite their outstanding performance on a wide range of domains, their propensity to "hallucinate" or make assertive but wrong answers turns their dependability in an educational setting into a point of worry. This research examines whether students can rely on LLMs when asking academic queries. We focus on the SQuAD 2.0 dataset, which incorporates both answerable and explicitly unanswerable queries, to assess the capability of state-of-the-art open-source LLMs in distinguishing correct answers from instances where no valid response is available. Particularly, experiments with various state-of-the-art 7-8 billion parameter models on representative validation samples from the SQuAD 2.0 dataset show strengths as well as limitations in state-of-the-art practices. Our results underscore the need for ethical and interpretable AI in learning, where avoiding dissemination of erroneous information is as vital as furnishing accurate responses. This effort helps toward developing guidelines that support safe LLM deployment within student learning environments.

Keywords: Large Language Models (LLMs), Explainable AI, Education, Learning, AI Safety, SQuAD-2.0, SHAP

1. Introduction

Large Language Models (LLMs) have emerged quickly to revolutionize natural language processing, with use extending from dialogue agents to educational tutoring systems. Specifically, their capacity to produce coherent, context-sensitive responses to student questions makes them desirable learning digital companions. Students of all subjects are increasingly using AI models like ChatGPT for educational purposes, with the share of U.S. teens using ChatGPT for schoolwork doubling from 13% in 2023 to 26% in 2024 (Pew Research Center, 2025), alongside other AI tools like Mistral and open-source equivalents to get rapid explanations, overviews, and clarifications. However, the trend also raises critical concerns: to what extent should learners believe such models, and how can we ensure that they don't spread false knowledge? This concern about factual reliability, or hallucination, is a key component of the broader challenge of LLM trustworthiness, which also includes aspects like toxicity, fairness, privacy, and ethical alignment (Mo et al., 2024).

Conventional QA systems used to depend on retrieval-based approaches or slender-domain knowledge bases in which the lack of an answer would be marked as such. LLMs, however, are inherently generative; when confronted with unanswerable or unclear questions, they tend to generate responses that are persuasively articulate but factually incorrect. This behavior, referred to as hallucination, introduces distinct danger in learning environments

wherein students might not have sufficient know-how to scrutinize the validity of an answer. This risk is amplified because the authoritative appearance of GenAI text can mislead young learners who lack the prior knowledge to recognize inaccuracies (UNESCO, 2023).

To address the fundamental challenge of ensuring reliable educational question-answering, our research focuses on two critical dimensions: evaluating how effectively models handle both answerable and unanswerable queries and understanding the underlying reasoning processes that drive their responses. In educational contexts, it is insufficient to simply measure accuracy as we must also verify that correct answers emerge from appropriate engagement with source material rather than from potentially unreliable memorized associations.

Our approach combines systematic performance evaluation with explainability analysis to examine whether models appropriately ground their responses in provided evidence or instead rely on fabricated information when faced with queries that lack sufficient contextual support. We employ explainability techniques to analyze which input elements most strongly influence model decisions, enabling us to distinguish between responses derived from genuine contextual reasoning versus those generated through potentially problematic pattern matching from training data.
Overall, our contributions are as follows:

- The study offer empirical observations on how well the model handles unanswerable questions in a QA task setting and understand the discrepancy between confidence and correctness.
- This study rigorously tests how an open-source LLM responds to student-like queries from SQuAD 2.0 (Rajpurkar et al., 2018), highlighting whether it can steer clear of misinformation.
- We use TokenSHAP (Goldshmidt et al., 2024), to quantify and explain how individual tokens in the input prompt influence the model's final response, enabling us to determine whether the model grounds its answers in the provided contextual evidence or instead relies on memorized patterns from its training data.
- Results are positioned in the context of ethical AI, given emphasis on the dangers of unbridled dependency on generative models within classrooms and the need for transparent, interpretable systems.
- Considering the recent state-of-the-art, the study aims to provide a few initial recommendations to assist in safely deploying LLMs on the part of educators and system designers in a manner that will not undermine trust but rather maximize learning.

2. Literature Review

Large Language Models such as GPT-4 and GPT-5 are already changing how education is practiced. These models are able to automate teaching tasks (e.g. grading, generating feedback) and facilitate more tailored learning experiences for students through their new Study Mode (OpenAI, 2024) in ChatGPT designed to guide students through step-by-step learning rather than providing direct answers. Indeed, LLMs have reached near-student level performance on some standardized tests and can even be used as on-demand tutors or writing aids for students. Empirical research suggests that including tools such as ChatGPT in coursework can enhance student performance – for instance, a study f (Wang, S et al., 2024) found that students using ChatGPT to write or edit responses did better than average in certain university courses. The benefits could include enhanced student engagement and one-to-one support, as well as less teacher workload through AI-generated content. However, significant concerns remain, particularly regarding academic dishonesty (such as plagiarism or excessive reliance on AI-generated content) and the reliability and potential bias of AI-provided information (Kasneci, E et al., 2023). In addition, there are issues of unfair access (as not all students are likely to have equal access to these resources) and the necessity of educating both students and teachers in AI literacy. In brief, LLMs hold

revolutionary promise in education, but they need to be implemented and monitored with caution to tackle trust, fairness, and pedagogical concerns (Guizani, S et al., 2025).

In the conventional question-answering (QA) benchmarks, it was believed that any question had an answer in the text given. Realistic use cases shatter this presumption – users tend to ask unanswerable questions or ones answerable only via external knowledge. This problem came to the forefront with the release of datasets such as SQuAD 2.0 (2018), which merged the original SQuAD reading comprehension data with more than 50,000 adversarially composed unanswerable questions (Rajpurkar, P et al., 2018), (Reyes-Montesinos et al., 2025).

Hallucination is a key problem eroding confidence in LLMs. Hallucination involves the model creating information that sounds plausible but is untrue or not supported by the input (Huang, L. et al., 2025). For example, an LLM could provide a historical "fact" or a quotation that it completely fabricated. This has been a common phenomenon: LLMs learned from huge web data can generate coherent responses that appear to be right to the average reader but include fictions (Xu, Z, et al., 2025) . This many-faceted endeavour is essential if LLMs are to be safely implemented in education and other high-stakes applications. The risk is that the stylistic plausibility of AI-generated text can hide underlying falsehoods, eroding trust in established knowledge and hindering the development of critical thinking skills in students (Elsayed, 2024). Scientists in NLP and AI safety are working closely on these issues in the recent literature, and we are sure to witness fast-moving advances in methods that make LLMs more dependable, honest, and user-centered in the near future.

3. Methodology

Our research addresses the critical need to evaluate LLM reliability in educational question-answering scenarios through a comprehensive approach that combines performance assessment with explainability as to why the model behaves the way it does. Given the characteristics of educational environments, where learners increasingly rely on LLM-generated answers for learning, it is essential to understand not only when models fail but also the reasons behind their failures. We concentrate on the combined issue of assessing hallucination propensities while also exploring the fundamental decision-making mechanisms that result in both correct answers and false responses.

For this study, we presented prompts containing the context plus the answer and provided a couple of cases of few-shot prompt instructions to the model that explicitly demonstrated answerable and unanswerable cases to ensure that large language models grounded their responses in the given passage rather than relying on memorised knowledge from their training data. In the prompt fed into the LLM, it was explicitly instructed for the model to not use its memory to answer any of the questions and only consider the context provided to it within the prompt. In unanswerable situations, the correct response was asked to be shown as an empty string, similar to the unanswerable field in the SQuAD dataset, encouraging the model to abstain when evidence was absent.

The LLM Responses were further constrained to a strict JSON format, minimising the risk of free-form or hallucinated outputs. Using the few-shot prompting strategy allowed us to measure behaviour against the ground truth answers in the SQuAD dataset.

We also performed a SHAP-based token attribution as shown in Figure 1., which was applied to a sample of the data, providing insight into which parts of the passage and question informed the model's decisions and output. This study was key for verifying that the generated answers arose from meaningful contextual cues rather than from previous training data memory or the consideration of irrelevant tokens, thereby measuring how trustworthy the model is when generating answers.

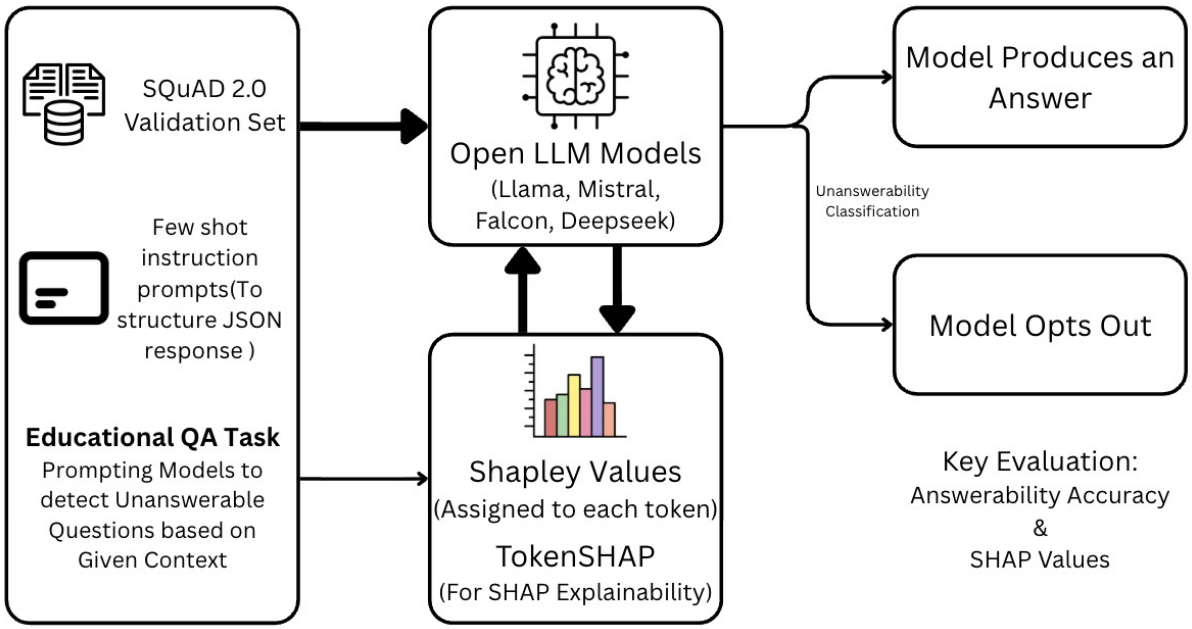

Figure 1. The research pipeline for evaluating LLM answerability on the SQuAD 2.0 dataset, utilizing few-shot prompting to structure responses and TokenSHAP for model explainability.

3.1 Dataset Brief

For evaluating open-source LLMs' reliability in handling both answerable and unanswerable queries from a given context, we utilized the Stanford Question Answering Dataset (SQuAD 2.0) validation set (11,873 samples). The dataset offers a perfect setting to quantify whether LLMs may abstain from answering as needed or mislead the user instead. We specifically chose SQuAD 2.0 for our study over the original SQuAD dataset as it introduced a subset of unanswerable questions from the given context, allowing us to test the tendency of models to hallucinate while giving out responses.

We selected the validation data to prevent any overlap with training data that could bias the models' performance through memorization (during training) rather than actual reasoning or comprehension.

3.2 Models Evaluated

We selected four state-of-the-art 7-8 billion parameter open-source models for our research (2 Instruction Tuned models and 2 Base Models):

- Mistral-7B-Instruct-v0.3
- Meta-LLaMA3-8B-Instruct
- DeepSeek-LLM-7B-Base
- Falcon3-7B-Base

These models were specifically chosen due to their easy accessibility to students, widespread adoption within the open-source community, balance between computational performance and efficiency. In this initial study we are evaluating only open-source models without any task-specific fine-tuning as foundational models are already trained on large internet corpus with instruction following capabilities like question answering. We chose this approach because it reflects a realistic scenario: as their availability enables developers to build conversational models and chatbots on top of them leading to students and general users often interacting with these models almost directly. Therefore, assessing their performance in this "vanilla" state provides a more authentic and essential baseline. We intend to explore the effects of hallucinations under fine-tuning and Retrival Augmented Generation (RAG) setting as part of future work.

3.3 Evaluation Metrics

We extended beyond standard SQuAD metrics such as exact match and F1 for evaluation. While these capture the quality of the answers generated on a surface-level, we also incorporated semantic similarity measures such as Cosine Similarity to account for meaning within answers produced and BLEU and ROUGE-L to gauge lexical overlap. Crucially, we measured unanswerable accuracy to assess the model's ability to recognise and abstain when there was no answer to be found in the given context.

- Exact Match (EM): Represents the proportion of predictions that exactly match the ground truth answer. This is a binary metric that requires complete string matching and provides the most stringent evaluation, making it particularly useful for factual QA tasks where precision is critical.
- F1 Score: Represents the harmonic mean of precision and recall calculated on token overlap with the ground truth (van Rijsbergen, 1979). This metric balances the trade-off between precision (what fraction of predicted tokens are correct) and recall (what fraction of ground truth tokens are captured), providing a more nuanced view than exact match alone.
- BLEU Score: Measures n-gram overlap between predicted and gold answers (on answerable questions). Originally developed by Papineni et al. (2002) for machine translation evaluation, BLEU computes modified precision for n-grams and includes a brevity penalty to discourage overly short outputs.
- ROUGE-L Score: Measures the longest common subsequence between predicted and ground truth answers (on answerable questions). ROUGE-L captures sentence-level structure similarity and is less sensitive to word reordering compared to n-gram based metrics, making it particularly effective for evaluating text generation quality (Lin, 2004).
- Cosine Similarity: Embedding-based semantic similarity metric applied to answerable predictions (on answerable questions). This approach captures semantic relatedness beyond lexical overlap by comparing vector representations of predicted and reference answers in a high-dimensional embedding space, providing insight into meaning preservation.
- Unanswerable Accuracy: Represents the proportion of unanswerable questions correctly identified as having no answer. This metric specifically evaluates a model's ability to abstain from answering when appropriate, which is crucial for robust question-answering systems that must handle queries outside their knowledge scope.

These metrics aim to capture not only the LLM's answer correctness but also semantic similarity and reliability in refusing to hallucinate.

3.4 Insights into Feature Explainability

In addition to the evaluation metrics defined, we used TokenSHAP, an interpretability method that explains the output of large language models (LLMs) by computing Shapley values for input tokens. It estimates how much each token contributes to the final model response.

This analysis highlights which specific words or phrases in the context influenced the model's decision to either provide an answer or abstain. We randomly sampled 10 queries from SQuAD (answerable and unanswerable) for each selected model. We extracted SHAP attribution scores for both tokens, which increased (Positive SHAP score) and decreased (Negative SHAP score) the likelihood of generating the answer given by the LLM.

This interpretability layer allows us to examine whether models genuinely anchor their responses in the provided context passages or instead rely excessively on their pre-trained knowledge. This differentiation is crucial in educational settings, where a model that generates fabricated answers while disregarding the assigned reading material could misinform students. Conversely, a model that authentically engages with the text can promote fundamental reading comprehension skills.

4. Results and Discussion

4.1 Quantitative Findings

Across the four evaluated models, performance varied. We saw a consistent trend emerge, i.e., while all the models demonstrated moderate success on answerable questions, their ability to correctly handle unanswerable queries was markedly limited.

Table 1. *Summary of the results*

	EM	F1	BLEU	ROUGE-L	Cosine Similarity	Unanswerable Accuracy
Mistral-7B-Instructv0.3	0.481	0.557	0.512	**0.807**	**0.872**	0.305
Meta-Llama3.2-8B-Instruct	**0.632**	**0.684**	**0.518**	0.786	0.833	0.584
Deepseek-LLM-7B-Base	0.478	0.515	0.332	0.530	0.599	0.504
Falcon3-7B-Base	0.549	0.576	0.267	0.406	0.439	**0.748**
Average	0.535	0.583	0.407	0.632	0.685	0.535

The results in Table 1 indicate that all four models achieve near comparable performance on EM and F1, averaging 53%. This suggests that while they capture the gist of correct answers, exact matches remain limited, and the models tend to wander off or bring other complexities of the language it has learnt during training into the mix. Unanswerable accuracy classification is also a concerning metric, averaging around 53% (only due to Falcon3 performing at 74.8% accuracy). It displays the persistent challenge of hallucination, and the difficulty models face in refraining from answering when no valid answer is present in the context provided.

Instruction-tuned models, particularly Meta-LLaMA-3.2-8B-Instruct, consistently outperform base models across most metrics, demonstrating being better at matching answers with the Ground Truth through the EM, BLEU, ROUGE-L metric and also having a higher semantic similarity through the cosine similarity metric. This also uncovers an advantage that reflects the effectiveness of instruction tuning in adapting models towards more precise and contextually faithful outputs. However, base models like Falcon3-7B show stronger performance in unanswerable detection, even if the overall quality of the answers it provides lags compared to the other Instruction-tuned models, indicating a trade-off between answer generation and uncertainty handling.

4.2 Attribution-Based Explainability Findings

We employed token-level SHAP analysis on a subset of ten samples (TokenSHAP is very computationally intensive and thus the reason for taking 10 samples) from the SQuAD dataset per model. This allowed us to uncover tokens in the context that influenced the decision to answer or abstain. We generalized this across 4 cases, i.e. when the LLM decides to answer an answerable question correctly, when the LLM answers an answerable question incorrectly, when the LLM decides to not answer an unanswerable question (correct), when the LLM answers an unanswerable question (incorrect).

Table 2. *Generalisation of the TokenSHAP findings across 4 cases*

	Correct Prediction	Incorrect Prediction
Answerable (Ground Truth)	When the passage contained a demarcated span that aligned closely with the question, SHAP strongly assigned those tokens with	In relatively denser contexts, SHAP highlighted irrelevant tokens relative to the Ground Truth (e.g. Numbers) with strong positive weights, causing it to

	high positive weights, while assigning negative weights to filler or background text. This clear overlap made the answer span differentiable, thus enabling the LLM to pick out the answer accurately.	get distracted, while underweighting the actual answer span. It showed the tendency of LLMs to be drawn towards contextually incorrect cues in longer contexts, thinking it was highlighting relevant facts when it wasn't.
Unanswerable (Ground Truth)	SHAP was able to detect and assign lower weights to domain-relevant but misleading tokens with respect to the question, reducing the risk of fabrication of an answer. Generally, little to no tokens received high positive weights showing that no suitable span was available within the context to serve as the answer.	SHAP assigned strong positive weights to question-related tokens (e.g. "What"), and to filler tokens (e.g. "Question: ") fed through the prompt. In turn, it under-weighted the actual critical span containing the answer, losing its confidence in being able to answer the question.

There were also edge cases where the model was able to produce a rephrased answer which was correct, e.g. '1st' instead of 'first'. Here we observe that the model did not just directly extract the token from the context but rather used its internal knowledge to summarize. This also shows that the model's predictions are not just solely context-dependent and rather is shaped by a blend of contextual cues and habits it has picked up from the training data it was exposed to.

4.3 Safety Implications on Educational Sector

The quantitative results reveal concerning patterns that have direct implications for educational applications. The accuracy for unanswerable questions is only 53.5%, indicating that when students ask questions that lack sufficient information in the provided materials, language models often generate plausible sounding but incorrect responses nearly half of the time. Along with that, our analysis of explainability uncovers that when models incorrectly answer unanswerable questions, they tend to focus on superficially relevant tokens instead of genuinely reasoning about whether adequate evidence exists. This pattern suggests that the models are employing shallow pattern matching instead of the deeper comprehension skills expected from reliable educational tools. This finding validates concerns raised by organizations like UNESCO, which worry that young learners, being less expert, might accept such superficial or inaccurate AI output without the necessary critical engagement (UNESCO, 2023).

A 47% risk of receiving fabricated information could significantly undermine learning outcomes. These findings highlight the need for a few necessary interventions to ensure safe educational deployment:

- Refusal/Abstention: Language models should be explicitly trained to prioritize abstaining from fabrication or hallucinating. The level of creativity needs to be curbed under the educational task setting as models can fabricate content in such settings,
- Uncertainty Indicators: Language Model based educational interfaces or systems should incorporate indicators of uncertainty or lack of confidence to prevent blind trust of uncertain outputs.
- Explainability: Explainability mechanisms should be integrated with Language Model based systems to help students and educators discern when responses are supported by evidence versus reliant on potentially unreliable memorized associations. Such mechanisms can also enable traceability of references, based on which the students or educators can make informed decisions about the systems' responses.

5. Conclusion

This study provides a systematic evaluation of open-source LLMs on the SQuAD 2.0 benchmark, with emphasis on the dual challenge of answering correctly when possible and abstaining responsibly when not. Our experiments demonstrate that while current 7–8B parameter models are reasonably effective at generating about accurate answers, their low unanswerable accuracy indicates that vulnerability to hallucination remains.

Our initial research work indicates that modern language models are not always reliable in their instruction following capability with their performance on unanswerable queries being a significant concern. The models correctly identified unanswerable questions only 53.5% of the time on average, meaning they opted to fabricate plausible but incorrect information in the remaining 46.5% of cases. Our explainability analysis shows this stems from a reliance on superficial pattern-matching rather than deep contextual reasoning. For educational applications, this failure rate of nearly 47% presents a serious risk of misleading students and undermining learning outcomes.Overall, the findings emphasise the challenge of optimising answer quality and reliability whilst aligning with the study's broader focus on trust and robustness in LLM-based question answering.

Future research should extend these evaluations to other datasets, models, proprietary systems and subject domains, explore domain-specific fine-tuning with balanced proportions of unanswerable queries, and test the integration of interpretability mechanisms into real classroom settings. Our preliminary findings on model unreliability highlight the need to foster students' critical thinking skills, specifically their ability to question and validate information provided by AI tools (Greyling & Cilliers, 2023). Additionally, future work could focus on system-level safeguards that bypass generative models for factual queries. For example, some educational systems are being designed to analyze a user's input and, if a factual question is detected, retrieve the answer from a curated database of authorized materials, thereby preventing the possibility of hallucination entirely (Jančařík & Dušek, 2024). Only by aligning the model's performance and priority towards generating correct, viable and trustworthy answers in line with the actual needs of the education domain can LLMs be responsibly embedded and trusted in student learning environments.

Acknowledgements

We would like to thank all the people who have helped prepare the dataset and models and kept it open source for us to use for our research.

References Elsayed, H. (2024). The Impact of Hallucinated Information in Large Language Models on Student Learning Outcomes: A Critical Examination of Misinformation Risks in AI-Assisted Education. *Northern Reviews on Applied Technology and Cross-Cultural Communications, 1*(1). https://northernreviews.com/index.php/NRATCC/article/view/2024-08-07

Goldshmidt, R., Neeman, T., Mizrahi, M., & Goldberg, Y. (2024). TokenSHAP: Interpreting large language models with Monte Carlo Shapley value estimation. *arXiv preprint arXiv:2407.10114*. https://arxiv.org/abs/2407.10114

Greyling, L. M., & Cilliers, L. (2023). Redefining the Role of Educators: The Impact of Artificial Intelligence on Curriculum Design in Higher Education. In *Proceedings of the 22nd European Conference on e-Learning - ECEL 2023* (pp. 136-144). Academic Conferences International Limited.

Guizani, S., Mazhar, T., Shahzad, T., Ahmad, W., Bibi, A., & Hamam, H. (2025). A systematic literature review to implement large language model in higher education: issues and solutions. *Discover Education, 4*(1), 1-25.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1-55.

Jančařík, A., & Dušek, O. (2024). The Problem of AI Hallucination and How to Solve It. In *Proceedings of the 23rd European Conference on e-Learning, ECEL 2024* (pp. 122-128). Academic Conferences International Limited.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, (pp. 74-81). Association for Computational Linguistics.

Mo, L., Wang, B., Chen, M., & Sun, H. (2024). How Trustworthy are Open-Source LLMs? An Assessment under Malicious Demonstrations Shows their Vulnerabilities. In *Proceedings of the*

2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2775-2792). Association for Computational Linguistics.

OpenAI. (2024). *ChatGPT study mode*. https://openai.com/index/chatgpt-study-mode/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Pew Research Center. (2025, January 15). *About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023*. Pew Research Center. https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/

Rajpurkar, P., Jia, R., & Liang, P. (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)* (pp. 784-789). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

Reyes-Montesinos, J., Rodrigo, Á., & Peñas, A. (2025). None of the above: comparing scenarios for answerability detection in question answering systems. *Applied Intelligence, 55*(12), 1-18.

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000386693

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworths.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., ... & Wen, Q. (2024). Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Xu, Z., Song, T., & Lee, Y. C. (2025). Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *International Journal of Human-Computer Studies, 197*, 103455.