Beyond Hallucination: Building a Reliable Question Answering & Explanation System with GPTs

Kazem Jahanbakhsh
Course Hero, USA
kazem.jahanbakhsh@coursehero.com

Mahdi Hajiabadi
Course Hero, Canada

Vipul Gagrani
Course Hero, USA

Jennifer Louie
Course Hero, USA

Saurabh Khanwalkar
Course Hero, USA

Abstract

Large language models such as GPT-4 have demonstrated performance comparable to human on various academic assessments, including the Uniform Bar Exam and LSAT. This opens up unprecedented opportunities for advancement of online learning through generative AI. However, there are a number of challenges using GPT models for educational use cases. For example, GPT models can generate incorrect information. They also lack providing custom academic references for their outputs. This paper discusses the design and implementation of a GPT-powered question answering/explanation system at Course Hero. We present A/B test results revealing a notable 40% increase in answering coverage compared to a retrieval-based question answering system. Moreover, we describe how augmenting our internal questions' answers with step-by-step explanations generated by GPTs lead to a 75% lift in users' approval ratings. Lastly, we outline the design for a production-ready reference system, providing evidence for users to verify GPT responses. Through human evaluations, we show that we can achieve $Precision = 84\%$ and $Recall = 69\%$ when providing reference documents for GPT outputs.

1 Introduction

As one of the largest online learning platforms, Course Hero [1] has hundred of millions of questions & answers (Q&A) and study documents for a wide range of academic subjects (e.g. Finance, Nursing, Management, and Computer Science). Course Hero search handles millions of queries weekly, where users utilize the website's search to get instant help with their questions or find study materials for their courses. A big percentage of the search queries are questions for which users need help to get answers and step-by-step explanations (e.g. "solve $x^2 + x - 2 = 0$?").

We historically use semantic and lexical search to provide answers and explanations to users' questions. Behind the scene, when a user types a query, we predict the query's intent using a machine learning model. If the query's intent is "asking a question", we run a semantic search using a vector database powered by Sentence-BERT to pull the best answer and explanation (i.e. A&E) and presents them to the user [12].

By leveraging recent advances in GPTs (Generative Pre-Training Transformer), we can bring the benefits of one-on-one tutoring to all students while maintaining academic integrity and addressing GPTs limitations such as hallucination (i.e. producing content that is nonsensical or not factual). With our internal Q&A's [2], we can answer & explain a substantial percentage of search questions

[1] https://www.coursehero.com
[2] The internal vector database contains hundred of millions of Q&A's.

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

while ensuring a minimized rate of hallucinations. Having said that, the supply gap opens up an opportunity to leverage GPTs [4] for answering & explaining the remaining search questions for which we do not have internal A&E's.

When we show an answer and explanation to a user, they can give a thumbs-up or thumbs-down to rate the A&E based on its quality in the product. From product users ratings, we know that the main reason for thumbs-up is "thorough explanation" (20%). More interestingly, one of the leading reasons for thumbs-down is "needs more explanation" (15%). We have a large number of Q&A's in our vector database that do not have step-by-step explanations. Thus, we employ GPTs to generate comprehensive explanations. Figure 1 illustrates an example where our internal response is complemented by an explanation generated by a GPT model. The paper contributions can be summarized as follows:

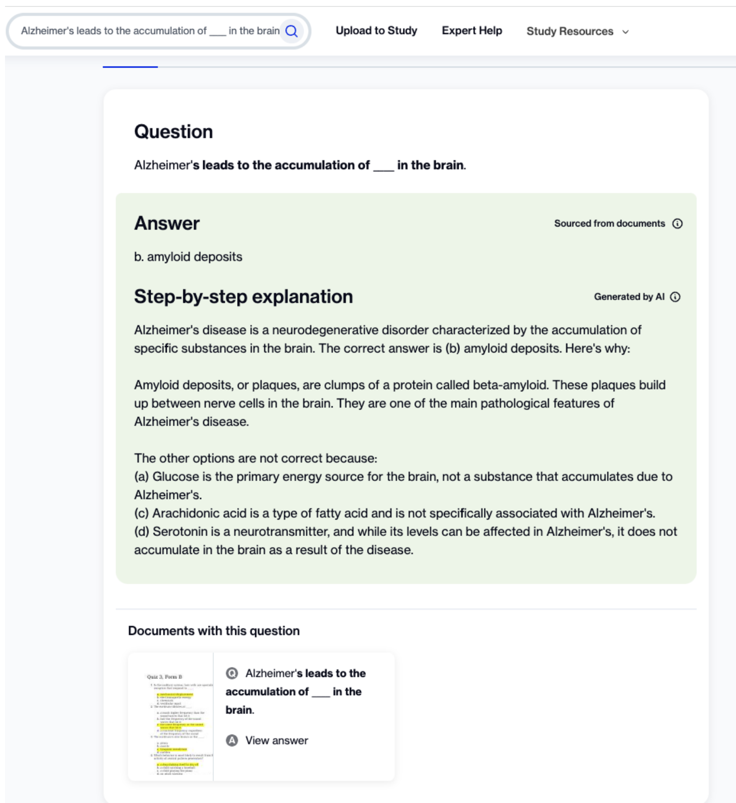


Figure 1: Providing Answer and Explanation for Question Queries in Search

- Implement a few novel prompts to enhance GPT response style and address hallucination for question answering and explanation.
- Devise an algorithm for question type identification helping GPT models produce higher quality outputs.
- Show how GPT models can provide step-by-step explanations, enhancing student comprehension of new concepts.
- Implement a reference system to provide evidence for generative AI's answers and explanations.

2 Generative AI Question Answering & Explanation: Opportunities & Challenges

Leveraging large language models (i.e. LLMs) for question answering and explanation comes with a number of challenges. We need to ensure that the generated answers & explanations are factually correct. We know that GPT models are not 100% reliable and can suffer from hallucination [3]. For example, if a user types a partial question, GPT models could complete the question and answer it. Another issue with GPT models is that they are not connected to a search engine therefore they can not provide evidentiary support for their responses independently. This is especially critical for students to verify a generative AI answer and explanation. In the following sections, we present a few novel solutions that we developed to address GPTs shortcomings.

2.1 GPT Prompting & Safeguards

We ran a number of experiments to improve the quality of GPT-3.5 and GPT-4 responses when answering/explaining users' questions [4]. In the prompts, we instruct the GPT model to act as an *academic tutor*. Based on search query logs, we know that users might type partial questions in the search bar. Sending such questions to GPTs could lead to hallucinated responses. We added a few safeguards to prompts commanding GPTs to not answer incomplete questions. Specifically, we instruct GPT models to only answer a question if it is *rooted in truth*. And, do not answer if a question is *nonsense*, *tricky*, or *incomplete*.

Table 1 shows two examples of incomplete questions. Our prompts' safeguards for handling incomplete questions reduced the hallucination rate by 10%. This consequently improved the thumbs-up rate for generative AI answers & explanations. Finally, in prompts we instruct GPT models to give an accurate answer and step-by-step explanation. This also improved the quality of GPT responses. Figure 1 demonstrates how we show the "answer & explanation" on top of search results when a user asks a question using the search bar.

Which of the following does not occur during Sprint execution?
Which of the following is not a common reason management might seek assistance from OBM specialists?

Table 1: Examples of Incomplete Search Question Queries

2.2 GPT-3.5 vs GPT-4: Hallucination Comparison

We ran a number of A/B tests to benchmark the quality of GPT-3.5 and GPT-4 models for question answering in search. In the control bucket we used semantic search with internal Q&A; however, the variant bucket was using both semantic search and GPT models while prioritizing internal Q&A for answering. The control bucket had $147k$ users. The GPT-3.5 and GPT-4 buckets had around $147k$ users respectively. Both GPT models showed around 40% lift on question answering coverage compared to the baseline system (i.e. semantic search with internal Q&A). We use thumbs-up and thumbs-down feedback inside the product to measure users' satisfaction with the presented answer/explanation. The A/B tests showed that both GPT-3.5 and GPT-4 models had a thumbs-up rate comparable with the internal answers and explanations. Through A/B tests, we also learned that the GPT-4 model has around 12% higher thumbs-up rate than GPT-3.5. This directionally aligns with GPT-4 and GPT-3.5 performance reported in [3].

2.3 Hallucination Mitigation by Question Identification

We receive two types of question in search: free response questions (i.e. FRQ) and multiple choice questions (i.e. MCQ). From our first A/B tests, we discovered that we get lower ratings for MCQs compared to FRQs. For the GPT-3.5 model, the thumbs-up rate for MCQs was 21% lower than FRQs. Through manual reviews, we noticed that GPT responses for MCQs were quite short. Often GPT models just picked the option that it thought to be the right answer without any explanations. This leads to a poor user experience since students prefer a step-by-step explanation to fully understand an answer.

We also noticed that the thumbs-up rate for MCQs with options is 90% higher than MCQs without options. This is another example for GPTs hallucination when MCQs are incomplete. These observations encouraged us to improve the design of question answering system using generative AI. First, we devised two separate prompts for MCQ and FRQ questions. Specifically, for the MCQ prompt we instruct GPT models to not only provide the correct answer option but also a step-by-step explanation for the selected option.

To detect the type of a question, we implement a rule-based classifier which takes a question and detects if it is MCQ or not, by matching with patterns such as *"what is the correct answer"*, *"which of the"*, and *"none of the above"*. If a question is MCQ, we next run some regex-based rules to check if it has all options or not. If an MCQ has options, we send it to GPTs for answering. Otherwise, we do not send them to GPTs. This reduces the hallucination and improves the thumbs-up rate. We analyzed our generative AI results after shipping the MCQ classifier to production. The new question identification system leads to an absolute 5% improvement in thumbs-up rate which is significant.

2.4 Explanation Generation with GPTs

We have a large number of Q&A's in our internal databases without explanations. Our hypothesis suggests that employing a GPT model to generate step-by-step explanations can enhance the quality of our internal answers while mitigating the hallucination risk. We devised two prompts for MCQ and FRQ questions. We append the matched question from the vector database and its internal answer to the explanation prompt. We instruct the GPT model to generate a step-by-step explanation for the provided answer. For the A/B test, we had $210k$ users assigned to the control bucket and $96k$ users to the variant (i.e. internal answers supplemented by GPT explanations). The A/B results show that presenting generative AI explanations improves the thumbs-up rate by 75%. Table 2 shows an example of a generative AI explanation.

Question: A person bites her nails but wants to stop. The behavior she wants to change (biting her nails) is called the: **Internal Answer**: Target behavior **Generative AI Explanation**: In the field of psychology, particularly in behavior modification, the term "target behavior" is used to describe the behavior that a person wants to change. In this case, the person wants to stop biting her nails. Therefore, the behavior she wants to change, which is nail biting, is referred to as the "target behavior". This term is used to identify and specify exactly what behavior is to be changed, which is the first step in any behavior modification program.

Table 2: Example of a Generative AI Explanation by GPT-3.5

3 Generative AI Verifiability with References

When using GPTs for question answering, adding references to GPT responses is crucial for high quality learning [9]. A reference system can provide document evidence for users to verify the accuracy of GPT responses. We developed a system with a large index of study passages (over 2.5 billion) across thirty eight academic subjects. Given a GPT answer & explanation, we first conduct a lexical search using Opensearch to retrieve top relevant passages (e.g. 20). Next, we rank the retrieved passages based on their semantic similarity to the GPT response using a Sentence-BERT model. We select the top three ranked passages from study documents and add document links to the GPT response as references.

3.1 References: Subject Matter Expert Reviews

We evaluated the quality of references by sampling seven hundred questions spanning twenty-four academic subjects (e.g., accounting, business, biology). The questions are randomly sampled from recent Course Hero search queries which were answered by a GPT model. Each question and its generative AI A&E were accompanied by three reference passages/documents identified by the reference system explained in the previous section.

Forty three subject matter experts (SMEs) evaluated the quality of references (i.e. reference passages and documents) according to seven categories: (1) incomprehensible passage, (2) irrelevant passage, (3) subject passage (matching subjects only), (4) conceptual passage (matching concepts only), (5) explains a similar answer (supports a similar question/answer), (6) explains the exact answer (supports the exact question/answer), (7) I don't know. Table 3 shows an example of a GPT answer/explanation, one of its reference passages, and its SME review.

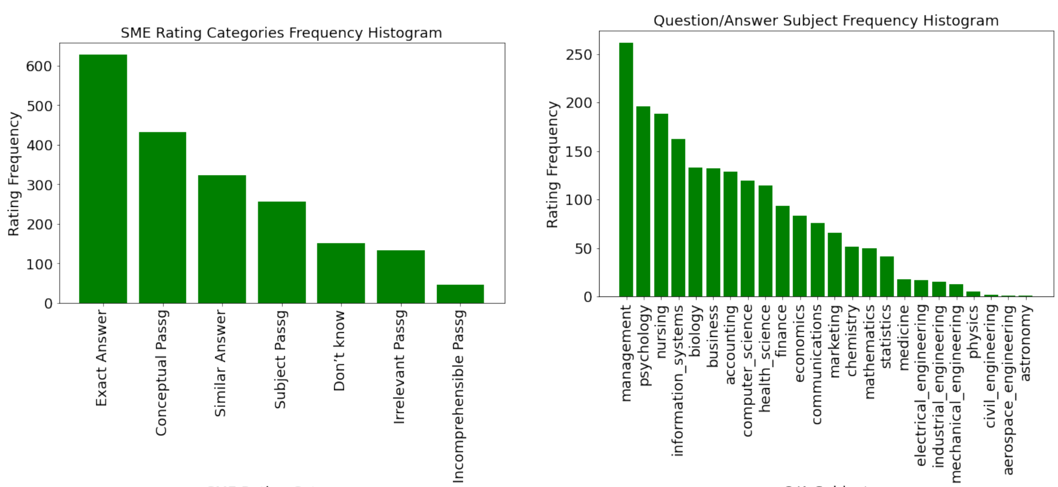
Question: Select all that apply. Which of these line items appear on a balance sheet? a) Assets b) Liabilities c) Expenses d) Shareholders' equity e) Revenues.
Gen AI Answer: The line items that appear on a balance sheet are: Assets, Liabilities, and Shareholders' equity. So, you should select the first, second, and fourth options.
Reference Passage: The balance sheet contains assets, liabilities, and owners' or shareholders' equity. Line items within the asset and liability classification are presented in their order of liquidity, so that the most liquid items are stated first. The assets include cash, property, inventory, and anything else owned by the company. Assets are listed on the left side of the balance sheet. Liabilities and equity are listed on the right side.
SME Review: Explains the exact answer

Table 3: An Example of Reference Passage and its SME Review

For the analysis, we excluded all rated passages marked as "I don't know". We categorized "Incomprehensible passage", "Irrelevant passage", and "Subject passage" as **"Irrelevant" passages**, while "Conceptual passage", "Explains a similar answer", and "Explains the exact answer" were considered as **"Relevant" passages**. Table 4 presents the high-level statistics. As you can see, 70% of all reference passages are relevant. Figure 2a shows the distribution of SME reviews across seven rating categories.

Total no of rated passages: 1960 [3]
No of Relevant passages: 1378 (70.3%)
No of Irrelevant passages: 433 (22.1%)
No of IGNORED passages: 149 (7.6%)

Table 4: SME Reviews Stats



(a) Distribution of SME Rating Categories

(b) Q&A Subject Distribution

Figure 2: Rating Categories and Subject Distribution for SME Reviews

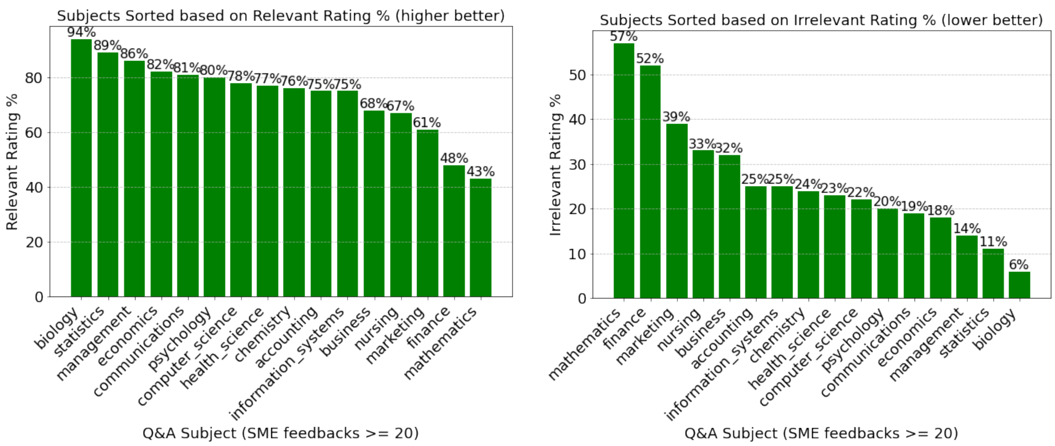
3.2 References: Subject Analysis & Precision/Recall Curve

Figure 2b displays SME reviews distribution across subjects. We filtered out subjects with less than twenty reviews ensuring sufficient samples for conclusive decisions. The subject distribution (i.e. Figure 2b) reflects Course Hero search distribution. We computed the percentages of "Relevant" and "Irrelevant" passages per subject. Figure 3a displays the subjects with the highest percentages of "Relevant" passages. Biology, statistics, management, economics, communications, and psychology

[3] We generated around three reference passages per Q/A/E for 700 questions.

have more than 80% "Relevant" passages (i.e. exact or similar answer). This can be attributed, in part, to the greater abundance of high-quality content available for these subjects in comparison to others (Figure 2b).

Figure 3b shows the subjects with the highest percentages of "Irrelevant" passages. Mathematics and Finance are the top subjects with more than 50% "Irrelevant" ratings (i.e. "Incomprehensible passage" and "Irrelevant passage"). It seems that reviewers do not find reference passages to be really helpful, especially for Mathematics. We think this is partly because BERT and specifically Sentence-BERT models do not do well in recognizing math terms and formulas due to their training data and masking strategy [11]. Therefore, we decided to not support references for Math questions.



(a) Relevant Passages Percentage (higher better)

(b) Irrelevant Passages Percentage (lower better)

Figure 3: Relevant & Irrelevant Passages Percentages across Subjects

In the reference algorithm, we have the semantic similarity threshold by which we can tune how similar a reference passage should be to its generative AI A&E. Figure 4 shows the Precision/Recall curve for references as a function of semantic similarity threshold. As we increase the threshold, the reference precision increases however the recall drops. We picked 0.65 as the optimal threshold to achieve $Precision = 0.84$ and $Recall = 0.69$, compared favorably with STOA [9].

3.3 Related Work

The Transformer architecture, pioneered by Google in 2017 [13], led to the creation of advanced GPT models [4]. GPT-4 demonstrates human-level performance on academic benchmarks [3]. ChatGPT, built on GPT-3.5 or GPT-4, has gained popularity among students, despite facing challenges like hallucination, misinformation, and potential for enabling plagiarism [5, 2]. GPT-4 has the tendency to "hallucinate" when responding to a user's question, i.e. "produce content that is nonsensical or not factual" [8, 10]. This tendency is harmful especially for educational use cases. Despite these limitations, we believe in the potential of generative AI to accelerate student learning [6].

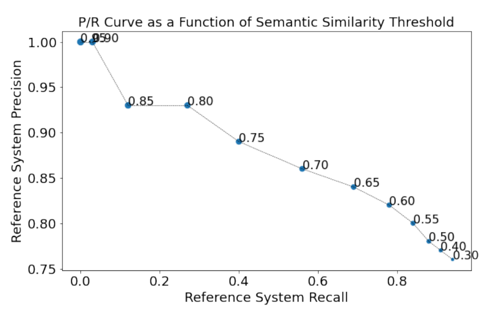


Figure 4: Reference Precision/Recall as a Function of Semantic Similarity Threshold

We utilize some of the best practices in prompt engineering to maximize the quality of question answering and explanation system [1]. Providing references is another technique to verify the quality of generative AI responses. Nelson et al. examined reference accuracy and recall for generative search engines, finding that only 74.5% of references support their statements (precision), and 51.5% of generated statements have references (recall) [9]. Lewis et al. introduced the retrieval-augmented generation (RAG) framework, combining parametric (e.g. BART) and non-parametric memory (e.g. a knowledge database) to enhance LLMs' factual responses [7].

3.4 Conclusion

In this paper, we discuss how our prompts and safeguards could mitigate the chance of GPT responding to incomplete, tricky, or non-academic questions. We describe a question identification system to ensure provision of both answers and step-by-step explanations for MCQs and FRQs. We show how this approach improves the quality of GPT responses especially when answering MCQs. We additionally demonstrate how GPT models could enrich internal Q&A's with comprehensive explanations empowering students to gain a deep understanding of the underlying concepts. We stress the importance of providing references for generative AI responses, enabling students to verify answers and access reference documents for further study. We have plans to imminently roll out the reference system for all generative AI A&E's. In the future, we also want to test the RAG framework and measure its impact on the question answering quality.

References

[1] Best practices for prompt engineering with openai api. Accessed on 2023-09-19.

[2] Educators battle plagiarism as 89chatgpt for homework. Accessed on 2023-09-19.

[3] Gpt-4 technical report. Accessed on 2023-09-19.

[4] Open ai gpt models. Accessed on 2023-09-19.

[5] Parents and students are optimistic about ai, but parents have a lotto learn to catch up to their kids – and want rules and ratings to help them. Accessed on 2023-09-19.

[6] Teaching with ai. Accessed on 2023-09-19.

[7] 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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[8] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines, 2023.

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

[11] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. Mathbert: A pre-trained model for mathematical formula understanding, 2021.

[12] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Acknowledgments and Disclosure of Funding

We would like to acknowledge several team members who contributed to this project: Galdino Rosas, Freddy Fei, Mohammad Yusuf, Krishna Yalamanchili, Cece Liu, Ishan Parmar, Isha Ghai, Andrew Shi, Rucha Patwardhan, Paresh Ravikumar, Josh Tyler, Leah Romm, Jacqueline Morales, Niamh Hughes, and Sean Michael Morris.