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

Kazem Jahanbakhsh, Mahdi Hajiabadi, Vipul Gagrani, Jennifer Louie, Saurabh Khanwalkar
Course Hero

Abstract

Large language models such as GPT-4 have demonstrated performance comparable to human on various academic tasks. However, GPT models can generate incorrect information. They also lack providing custom academic references for their outputs.

This paper discusses how Course Hero leverages GPTs to increase answer coverage by 40% compared to a retrieval-based system. We also show how augmenting internal answers with explanations generated by GPTs leads to a 75% lift in users' approval ratings.

Lastly, we discuss a reference system, providing evidence to verify GPT responses. Through human evaluations, we show that we can achieve P=84% and R=69% when providing reference documents for GPT outputs.

Introduction

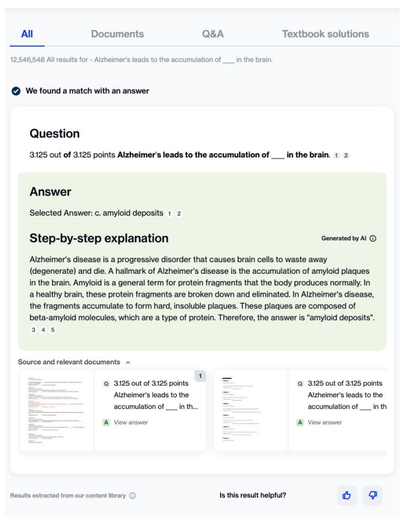
Course Hero has hundred of millions of academic Q&A's and documents for a wide range of subjects such as Finance, Nursing, and Computer Science. The website's search handles millions of queries weekly where a big percentage of them are questions (e.g. "solve $x^2+x-2=0$?").

We historically use semantic search to provide answers and explanations to users' questions. We run a semantic search using a vector database powered by Sentence-BERT to pull the best answer and explanation. By leveraging GPTs, we can bring the benefits of one-on-one tutoring to all students while addressing GPTs limitations such as hallucination.

Course Hero users can give a thumbs-up or thumbs-down to rate the presented Answer & Explanation based on its quality and format. From large scale 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%). Therefore, we want to employ GPTs to generate comprehensive explanations.

Generative AI Question Answering & Explanation

We added a few safeguards to our GPT prompts. Specifically, we instruct GPT models to only answer a question if it is ***rooted in truth***. Do not answer if a question is ***nonsense***, ***tricky***, or ***incomplete***. We also instruct GPTs to give an **accurate answer** and **step-by-step explanation**.



We ran a few A/B tests with GPT-3.5 and GPT-4 vs semantic search with internal Q&A. We designed a system to identify question type (MCQ vs FRQ) and route each question to its corresponding prompt.

Through A/B experiments, we measured the impact of GPTs on answer coverage compared to the semantic search system. We also measured the thumbs-up rate for each variant. Table 1 summarizes the A/B results.

Method	Thumbs-up rate lift	Answer % lift
Prompts safeguards	10% vs baseline prompt	40%
GPT-3.5	Similar to internal answers	40%
GPT-4	12% vs GPT-3.5	40%
MCQ question identification	5% vs wo question identification	NA

Table 1. Gen AI Answer & Explanation A/B Experiments Summary

Explanation Generation with GPTs

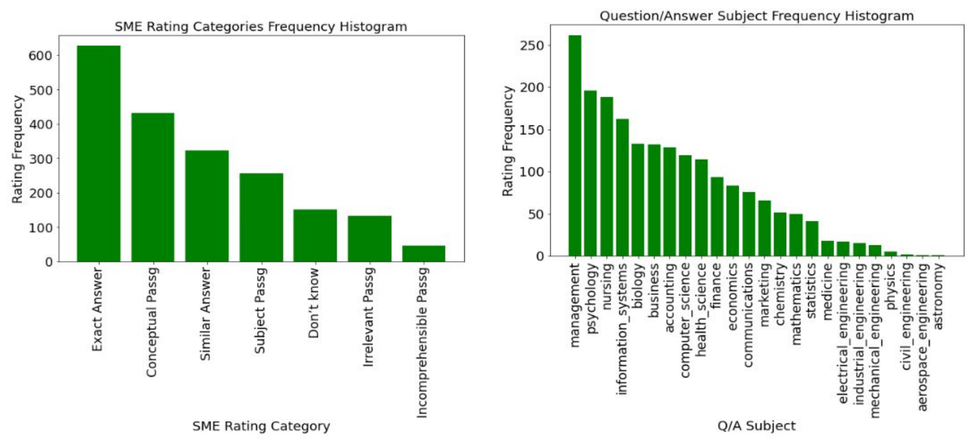
We have a large number of internal Q&A's without comprehensive explanations. We hypothesize employing GPTs to generate step-by-step explanations can enhance the quality of internal answers while mitigating the hallucination risk.

We devised two prompts for MCQ and FRQ questions. We append the matched question from the semantic search and its internal answer to the prompt. We instruct the GPT model to generate a step-by-step explanation for the provided question and answer.

For A/B tests, 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%.

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

Table 1: SME Reviews Stats



(a) Distribution of SME Rating Categories (b) Q&A Subject Distribution

Figure 1. Human Expert Reviews of Referenced Passages

Generative AI Verifiability with References

We developed a reference system to provide document evidence for users to verify the accuracy of GPT responses. We leverage a large index of study passages (2.5 billion). Given a GPT Answer & Explanation, we conduct a lexical search to retrieve top relevant passages. We rank the retrieved passages based on their semantic similarity to the GPT response. We select the top 3 ranked passages from study documents and add document links to the GPT response as references.

We evaluated the quality of references by sampling 700 questions spanning 24 subjects. We used 43 subject matter experts (SMEs) for evaluation. Each question and its Gen AI Answer & Explanation were accompanied by 3 reference passages identified by the reference system for review.

SMEs evaluated the quality of references according to 7 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.

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**. Figure 1 shows that 70% of all reference passages are relevant. The distribution of SME reviews across rating categories and Q&A subject distribution is depicted in Figure 1.

Conclusions

- A/B test results reveal a 40% increase in answer coverage.
- Augmenting internal answers with GPT explanations leads to a 75% lift in users' approval ratings.
- Reference system achieves P=84% and R=69% for providing academic documents for GPT outputs, compared favorably with STOA.

Contact

Kazem Jahanbakhsh, kazem.jahanbakhsh@coursehero.com
Vipul Gagrani, vipul.gagrani@coursehero.com
Mahdi Hajiabadi, mahdi.hajiabadi@coursehero.com
Jennifer Louie, jennifer.louie@coursehero.com
Saurabh Khanwalkar, saurabh.khanwalkar@coursehero.com

Course Hero, 2000 Seaport Blvd, Redwood City, California, USA

References

1. Best practices for prompt engineering with Open AI API (https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api).
2. GPT-4 technical report (https://arxiv.org/abs/2303.08774).
3. Open AI GPT models (https://platform.openai.com/docs/models).
4. 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.
5. Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines, 2023.
6. Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. Mathbert: A pre-trained model for mathematical formula understanding, 2021.
7. 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.