ClaraNP: Generative Interface for Nursing Education with Hallucination Mitigation

Henry Salkever[1], Maria F. D. Rondon[1], Ixchel Y. Peralta-Martinez [1], Ahmet A. Kiziltunc[2], Gulustan Dogan[1], Julie Hinkle[1], Crystal H. Dodson[1]

[1]University of North Carolina Wilmington
601 College Rd
Wilmington, NC 28403 USA
[2]Gazi University
Emniyet, Gazi Üniversitesi Rektörlüğü, Bandırma Cad. No:6/1
06560 Yenimahalle/Ankara, Türkiye

Abstract

Education holds great potential for improvement from generative AI integration. Efforts to employ AI tools in curriculum creation and generation have been by tried by companies, but recent reports of hallucination rates among commonly used LLMs demonstrate a rate of erroneous output far too high for reliable deployment in medical education. We propose ClaraNP, a generative LLM fine-tuned from SBERT open-source code for nursing education with internal hallucination and bias mitigation. We demonstrate our preliminary progress, articulate our plan for achieving further domain specificity, and explain our plans for in-class testing and integration.

Introduction

Medical education involves great depth of study and high consequences of failure. It is estimated that 20-40 percent of tasks performed by K-12 educators could be automated with current technology (Bryant et al. 2023). Initial automation efforts have integrated general-purpose language models in medical education (Safranek et al. 2023), yet common models do not have the domain knowledge or accuracy to create medical course materials. The average hallucination rate of the top 10 language models is still 6 percent (Hughes, Connelly, and Ashimine 2023). Even the top LLMs in hallucination mitigation fail to return repeated correct information during iterative querying(Megahed et al. 2023).

We propose ClaraNP: a generative nursing education language model fine-tuned from an falcon-7b- base. Aiming to eliminate contextual and accuracy concerns, the model features internal hallucination and bias mitigation, as well as pdf input capability at the scale of large textbooks.

Methodology

Language Model Preparation and Querying

To begin, we imported libraries for PDF processing, text manipulation, and machine learning model deployment. This includes the langchain package for document parsing, text segmentation, and embedding, the transformers library for accessing transformer models, and standard Python libraries for regular expressions, JSON handling, and file system interactions. We established constants representing the specific pre-trained models Instructor XL, SBERT MPNet base, and FLAN T5 base.

The class architecture consists of functionalities pertinent to question-answering using PDF documents. This class encompasses methods for model and embedding initialization, vector database construction from PDFs, retrieval QA chain configuration, and output cleaning from language models, establishing the operational blueprint. A pre-prompt was designated, and a dedicated method is implemented for refining the language model's output, removing extraneous tags and spaces, and enhancing answer clarity and readability. A loop is defined to perform iterative querying. Ten different outputs of the same query are passed to the accuracy module.

Hallucination Mitigation Model Preparation and Implementation

PDF handling and tokenizing libraries such as pdfminer, scikit-learn, and TensorFlow were installed. Functions were defined for reading, tokenizing, and truncating pdf text. An except function is defined using an OCR library of pytesseract to enable extraction of text from photocopied pages. Tokenized pdf text and queries are concatenated into a 'question' against which the set of answers are compared. Functions of the keras and tensorflow libraries were used to initialize a Siamese neural network (SNN). For tokenizing and encoding the text, we utilized the fine-tuned version of the AllenAI's Longformer Encoder-Decoder (LED)(Beltagy, Peters, and Cohan 2020), specifically the 'led-large-16384-pubmed' checkpoint available on Hugging Face (von Platen 2021). A semantic similarity score is calculated by comparing the cosine angle between the resulting semantic vectors. A keyword ranking model is defined using similar pdf handling protocol and a Jaccard similarity algorithm. The SNN and keyword similarity scores are used to calculate a weighted average that is insensitive to comparative answer length, sensitive to semantic content, and semi-sensitive to question/answer keyword parity. Let S_{SNN} represent the SNN score and let $S_{keyword}$ represent the Jaccard keyword

similarity. The weighted average $S_{weighted}$ is given by the equation:

$$S_{\text{weighted}} = \frac{w_{\text{SNN}} \cdot S_{\text{SNN}} + w_{\text{keyword}} \cdot S_{\text{keyword}}}{w_{\text{SNN}} + w_{\text{keyword}}}.$$

Given the weights w_{SNN} = 0.6 and $w_{keyword}$ = 0.4, the answer representing the highest $S_{weighted}$ score is returned as model output.

Domain-Specific Considerations

For textbook capability, truncation and batching of the model were adjusted to allow for PDF input of indefinite size. An optical character recognition-based extraction method was also employed as a backup during text extraction to ensure functionality on scans or photocopies. The pre-prompt employs the "persona" method among others and asks the model to follow instructions as if they were a graduate nursing instructor. An open-source model was consciously chosen to create a copyright risk-averse resource for higher education institutions. The Longformer Encoder-Decoder (LED) (Beltagy, Peters, and Cohan 2020) model was used for tokenization and embedding in the accuracy algorithm due to its hybrid global/local attention mechanism, a feature that allows for more accurate representation of longer context input. The additional supervised fine tuning of the 'led-large-16384-pubmed' checkpoint (von Platen, 2021) introduces domain-specific knowledge and representation tendencies. The selection of a SNN model was conscious due to its reliance on semantic vector comparison instead of character length vectors. Each generated by one of the twin models, the resulting input (context and prompt) and response(answer) vectors are magnitude-blind, allowing for accurate comparison between sizable inputs and short answers.

Experiments and Preliminary Results

The preliminary structure of ClaraNP was recently completed, thus extensive tests have not been conducted. We obtained a textbook from the graduate course in which this model will be tested and generated several series of 10 outputs through prompting a general purpose model to make a given quiz 10% less specific with each iteration. We recorded the output and ran this series through the hallucination mitigation model, and it assigned higher rankings to the un-augmented answers, regardless of their length. The independent human ranking of these sets by two professors loosely reflected the model results but did not directly coincide.

Challenges and Limitations

As output results are generated common consensus around top outputs may be the subject of debate between staff members. Although the tiiuae falcon-7b-instruct base model hallucinates on 16.2% of outputs (Hughes 2023) , this fallacy rate is not high enough to cause observable disparity between the top 4-6 outputs and thus the top 2-3 human labels are likely to vary based on individual preference. A very small percentage of internal bias toward longer answers may be introduced by the keyword similarity algorithm. The Jaccard Index computes similarity by dividing the intersection of two the two vectors by their union (Jaccard, 1901). Longer answers, even if incorrect, are likely to achieve higher Jaccard scores because higher numbers of terms will be shared between the documents. However, keyword comparison is necessary for achieving terminology parity between textbook context and the output.

# of Fallacies	Cosine Similarity Index	SNN Similarity Index	Accuracy Index
0	1	0.50923615694046	0.85277085
1	0.929778804228801	0.542960226535797	0.81373323
2	0.860045393911641	0.542955815792083	0.76491852
3	0.720578573277321	0.51461923122406	0.65879077
4	0.581111752643	0.514575839042663	0.56115098
5	0.41840046190296	0.510691344738006	0.44608773
6	0.3486670515858	0.510729908943176	0.39728591
7	0.2324447010572	0.443137675523757	0.29565259
8	0.1162223505286	0.444635421037673	0.21474627
9	0	0.450810253620147	0.13524308

Figure 1: Accuracy Module Output for Incrementally Augmented Model Responses

Future Project Plan

In the next phase of the project, we will fine-tune the falcon-7b-instruct base and the various embedding models on a large corpus of nursing textbooks. We will employ supervised fine tuning with datasets such as pubmed and SQuAD, and perform direct preference optimization on a dataset we have hand labeled from medical queries of general purpose models. We will add designated use options to the generation interface, allowing for increased pre-prompt specificity, and we will vectorize the code in all iteration structures. An anti-bias toolkit such as IBM's AIF360 (Bellamy 2018), Fairlearn (Weerts et. al, 2023), or factored verification by the LED itself will be integrated into the accuracy model. Although racial and gender references in medical textbooks are rare and most information is purely empirical, these mechanisms will ensure that model outputs never feature references or overgeneralizations of race or gender. We are also in the process of developing a vector store with additional nursing domain textbook content. Later, a prototype interface will undergo in-class testing with Dr. Crystal Dodson and Dr. Julie Hinkle of the UNCW nursing department.

Conclusion

ClaraNP is a step in a long-overdue integration of generative tools into nursing education. Test cases and assessment materials are two of the biggest limiting factors on the quality of healthcare coursework, and ClaraNP aims to provide a reliable source of this material free from the hallucination concerns that surround general-purpose transformer models. Our architecture (Safranek et al. 2023) represents an easy-to-implement strategy that could be duplicated by other institutes of higher education seeking similar results in their interfaces. This paper discusses the challenge hallucinations

pose in higher education as well as the methods we are taking to address the problem. We also outlined the remaining steps to completion and discussed our plans for in-classroom integration.

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. (arXiv:2004.05150). ArXiv:2004.05150 [cs].

Bryant, J.; Heitz, C.; Sanghvi, S.; and Wagle, D. 2023. How artificial intelligence will impact K–12 teachers. https://www.mckinsey.com/industries/education/our-insights/how-artificial-intelligence-will-impact-k-12-teachers. Accessed: 2023-11-21.

Hughes, S.; Connelly, S.; and Ashimine, I. 2023. Hallucination Leaderboard. https://github.com/johnsmith/awesome-project. Accessed: 2023-11-22.

Megahed, F. M.; Chen, Y. J.; Ferris, J. A.; Knoth, S.; and Farmer, A. J. . 2023. How Generative AI Models Such as CHatGPT Can Be (Mis)used In SPC Practice, Education, and Research? An Exploratory Study. https://arxiv.org/abs/2302.10916. Accessed: 2023-11-21, arXiv:2302.10916.

Safranek, C. W.; Sidamon-Eristoff, A. E.; Gilson, A.; and Chartash, D. 2023. The Role of Large Language Models in Medical Education: Applications and Implications. *JMIR Medical Education*, 9.

von Platen, P. 2021.