Large Language Model (LLM) Limitations when Used on a Limited Dataset

Bert Noble, Jim Chen, Mark Smith
Department of Information Systems, SCSU, St. Cloud, MN, United States
bert.noble@go.stcloudstate.edu

Extended Abstract

Introduction

Since the launch of ChatGPT in 2022, the use of Artificial Intelligence (AI) has taken an exponential increase. Private industry, as well as public services, embraced the possibilities and has implemented countless AI-driven chatbots (Mohamad Suhaili, Salim & Jambli, 2021). Almost all these chatbots use the capabilities of Large Language Models (LLMs) to interact with the users using human understandable language. Although much research has been done to understand and solve common LLM issues (e.g., bias, provocative language, false information)(Rana et al., 2023; Passmore & Tee, 2023; Labs, 2024; Trandabăț & Gifu, 2023; Li, Xu & Fan, 2022), our research wasn't able to find research on the use of LLM-based AIs on a limited set of documents and the specific issues that come with this use.

Use-Case

The use-case for this research project was the creation of a FAQ-type chatbot that would only be sourced by internal information (fact sheets, emails, documents).

Research Project Objectives and Methodology

Our research project explored the options to create an AI-chatbot that would provide the students with answers to their questions about the Master of Science in Information Assurance program at Saint Cloud State University in Minnesota. In general, there were four activities involved: (1) exploring the available research and options, (2) determining the preferred implementation for our use-case, (3) creating the prototype, (4) evaluating the prototype as a solution for the use-case.

Based on earlier experience, existing research, current user feedback, and driven by limited time and funds, we opted for creating the prototype chatbot based on (a) the OpenAI GPT-3.5 Turbo engine and the OpenAI embeddings as a LLM and to contextualize the conversation, (b) Qdrant vector database for storing the information and executing similarity searches, (c) the Langchain framework and Python language for easier implementation, and (d) the Streamlit platform and GitHub repository for hosting and running the prototype. This prototype does not consider the confidentiality of the information. Due to the limited resources in the available hardware, running the prototype locally would have been too time-consuming given the time constraints.

Because of the possible security implications, during the evaluation activities, we focused on two known issues of using AI-chatbots: (1) hallucinating AI, and (2) misinterpreting of the question. These two issues can lead to the provisioning of false information. Within the scope of the prototype, this could potentially lead to students making the wrong choice for their academic career, but in a more extreme situation where a similar AI-chatbot would be used for operational military decision-making, could lead to loss of life.

Prototype

The general information flow within the prototype follows the following steps:

(i) Information is fed to the prototype, this information is split into chunks and stored in a vector database to allow for similarity searches to extract relevant information.
(ii) A user enters a question in the chatbot. The question is embedded into a vector that allows for a similarity search in the vector database to extract the relevant chunks of information.
(iii) The relevant chunks of information, together with the system prompt setting limitation to how the AI-model can behave, and the session chat history is fed into an LLM to construct an answer to the question.

Steps (ii) and (iii) are repeated for follow-up questions. During the research project, the model has been fed with variating amounts of data in step (i).

Findings

Security Issue #1 – Hallucinating AI

We were able to limit the amount of hallucination by the AI by using a very strict system prompt.

Security Issue #2 – Misinterpretation of the Question by the AI

During the research, we encountered some problems that showed the limitations of the use of a LLM on a limited amount of data. When presented with focused questions that are positively related to the stored information and requested in the correct context, the chatbot repeatedly answered the questions in a satisfactory manner. Unfortunately, that is not the purpose of an automated chatbot. We must start from the idea that questions are asked about information that is not commonly known or not put into the right context. And that is exactly where the prototype didn't function as hoped.

When the chatbot was presented with a random question that had no relation with the information in the database, it would answer that it was unable to answer the question in the current context. But when a random question was entered using words that were related to the information in the database, the chatbot would provide an answer that could be misleading or was even plain wrong.

We did notice that the prototype was less prone to this issue when provided with more data.

Future research

Our research isn't completed after the conclusion of this semester-long research project, and we are still in the progress of researching the following topics/questions:

1. *AI Model*: Is it possible to fine tune the model for a better contextual awareness? What is the influence on the results if we adapt the main parameters that are fed into the LLM (temperature, prompt)? Which models are better suited for small datasets?

2. *Embeddings/vectors*: What is the influence of changing the parameters of the vectorization of the data (vector dimension, chunk size, overlap)? Are some vector stores more suited to perform similarity searches on small datasets? Can we set a threshold to similarity search results to obtain better contextualized chunks of text?

3. *Answer/question model*: Can we approach the found issues on the question component instead of the answer component?

Directly related to the use-case but not yet considered in the current prototype research:

4. *Proprietary information*: Is it possible to implement a similar prototype on proprietary information stored in a local database? What are the additional limitations that are put on the used LLM, embeddings and vector store?

References

Labs, W. (2024, January 4). *Artificial Intelligence: Still needs fine tuning to succeed in manufacturing*. Food Engineering RSS. https://www.foodengineeringmag.com/articles/101774-artificial-intelligence-still-needs-fine-tuning-to-succeed-in-manufacturing

Li, A.-W., Xu, X.-K., & Fan, Y. (2022). Immunization strategies for false information spreading on signed social networks. Chaos, Solitons & Fractals, 162, 112489. https://doi.org/10.1016/j.chaos.2022.112489

Mohamad Suhaili, S., Salim, N., & Jambli, M. N. (2021). Service Chatbots: A systematic review. *Expert Systems with Applications, 184*, 115461. https://doi.org/10.1016/j.eswa.2021.115461

Passmore, J., & Tee, D. (2023). Can Chatbots replace human coaches? Issues and dilemmas for the coaching profession, coaching clients and for organisations. *Coaching Psychologist, 19(1)*, 47–54. https://doi.org/10.53841/bpstcp.2023.19.1.47

Rana, S. A., Azizul, Z. H., & Awan, A. A. (2023). A step toward building a unified framework for managing AI bias. *PeerJ Computer Science, 9*, e1630. https://doi.org/10.7717/peerj-cs.1630

Trandabăţ, D., & Gifu, D. (2023). Discriminating AI-generated Fake News. Procedia Computer Science, 225, 3822–3831. https://doi.org/10.1016/j.procs.2023.10.378