Evaluating the Value of ChatGPT in Generating Customer Responses

Final Report

Alex Henigman

May 2024

Project sponsor: Dianne Babski

**Table of Contents**

Contents

[Abstract 3](#_Toc167804334)

[Objectives: 3](#_Toc167804335)

[Methods: 3](#_Toc167804336)

[Results: 3](#_Toc167804337)

[Conclusions: 3](#_Toc167804338)

[Introduction 3](#_Toc167804339)

[Methods 3](#_Toc167804340)

[Background and Scope 3](#_Toc167804341)

[NLM GenAI Pilot 3](#_Toc167804342)

[Background 4](#_Toc167804343)

[Model Development 4](#_Toc167804344)

[Evaluation 5](#_Toc167804345)

[Initial Testing 5](#_Toc167804346)

[Qualtrics Questionnaire 5](#_Toc167804347)

[Google Trial 6](#_Toc167804348)

[Results 6](#_Toc167804349)

[Recommendations 6](#_Toc167804350)

[Conclusion 7](#_Toc167804351)

[Resources 8](#_Toc167804352)

[Appendix A. 9](#_Toc167804353)

[Appendix B. 10](#_Toc167804354)

[Appendix C. 11](#_Toc167804355)

[Appendix D. 12](#_Toc167804357)

# Abstract

## Objectives:

The project aims to develop an experimental proof-of-concept Generative AI (GenAI) platform and evaluate the potential benefits of GenAI for customer service at the National Library of Medicine (NLM) for both the organization and its users.

## Methods:

Alex Henigman, Dianne Babski, and Nicole Sroka, and the NLM Reference and Web Services Team collaborated using an interdisciplinary approach to create, test, evaluate, and disseminate a GenAI solution for NLM customer service. Alex led a team comprising Microsoft engineers and representatives from the NIH Center for Information Technology to design a GenAI chatbot grounded with NLM products within the Microsoft Azure OpenAI Platform. Upon finalizing the proof-of-concept, she evaluated the GenAI outputs with an internal questionnaire. Additionally, Alex collaborated with a team from Google to compare results from two different technical products.

## Results:

The project successfully developed a GenAI chatbot, achieving a 70% satisfaction rate among 29 internal reviewers. The findings indicated that GenAI could provide significant value to NLM, justifying further development of this proof-of-concept towards full implementation. The chatbot and its evaluation results were disseminated through various formats and subsequently published at the 2024 Medical Library Association’s Annual Conference.

## Conclusions:

The project demonstrated that GenAI could be effectively utilized to enhance workflow efficiency. The proof-of-concept received positive ratings from reviewers, and further development is planned to explore additional frameworks for a GenAI solution. Preliminary testing of a Google prototype also showed a high satisfaction rate, suggesting that alternative frameworks should be considered.

# Introduction

The "Evaluating the Value of ChatGPT in Generating Customer Responses" project was part of a broader NLM Generative AI Pilot, wherein a selected cohort was granted access to the Microsoft Azure OpenAI (MS AOAI)1 platform to explore use cases that could enhance operational efficiency and support NLM’s mission. Driven by the high volume of customer inquiries and a commitment to innovation, an experimental proof-of-concept Customer Response Generative AI Use Case was developed. This project aimed to investigate the integration of GenAI to improve the efficiency of customer response handling and enhance customer satisfaction.

# Methods

## Background and Scope

### NLM GenAI Pilot

This project was part of a larger 6-month NLM GenAI Pilot, where a selected cohort developed working proof-of-concept prototypes for 10 GenAI use cases aimed at increasing workflow efficiencies through innovative solutions. Alex led the Customer Response Use Case, which was designed to evaluate whether GenAI implementation could benefit both NLM staff and users by improving customer response metrics, satisfaction, and the efficiency of handling inquiries. During this pilot, Alex participated in weekly stand-ups using Agile methodology, discussing achievements, plans, and challenges with fellow participants for feedback and collaboration. She also engaged in training, checkpoint presentations, and a final demonstration of the use case to an NLM-wide audience.

### Background

The NLM receives over 74,000 user inquiries annually, prompting the need for an innovative solution to enhance efficiency with GenAI. Alex collaborated with Dianne and Nicole to establish project expectations, scope, and desired outcomes, as well as to understand the types of questions submitted to the NLM. Through these discussions, the team determined the following objectives:

* Develop a customized GenAI platform to draft responses for Customer Service Representative (CSR) review.
* Focus the project on consumer health-related questions.
* Ensure the accuracy and relevancy of responses by grounding the platform in NLM products.
* Evaluate generated responses based on specific criteria.
* Ensure the system does not offer medical advice.

Alex began by reviewing literature on the current use of GenAI in libraries and customer service. She established connections with the Microsoft FastTrack (MS FT) team and gained education and training on GenAI through LinkedIn Learning and Microsoft Azure via Microsoft Learn. She also worked with the NLM Reference and Web Services Section to gather a list of questions submitted to the NLM Help Desk and to understand the methods for answering consumer health questions. Terry Ahmed and Mabel Mendez from this section provided Alex with 50 user-submitted questions, which she de-identified for research use (see Appendix A).

At this stage, Alex conducted baseline testing on a ChatGPT-4 Playground within MS AOAI, hosted securely behind the NIH Firewall. Baseline testing involved submitting de-identified user inquiries to the Playground and recording the responses. It was noted that the responses often provided medical advice and hallucinated citations, which were outside the scope of the desired responses. Alex reported these findings to the Microsoft FastTrack team and led the development of a GenAI platform grounded in three NLM products: PubMed, MedlinePlus, and ClinicalTrials.gov.

## Model Development

After discussions with the Microsoft FastTrack (MS FT) team, it was determined that the MS AOAI Bot Service solution would best serve the needs of the end users (CSRs). This solution allows user inquiries to be submitted and generates outputs similar to current human-created answers. Alex then met with Zhiyong Lu (ClinicalTrials.gov), Amanda Sawyer (PubMed), and Michael Honch (MedlinePlus) to understand how data is stored in these products, identifying the best methods for integration into the GenAI platform. Through these meetings, Alex proposed utilizing the product APIs so that the chatbot could conduct real-time searches of user inquiries, generating up-to-date, relevant, and accurate answers based on NLM products (Appendix B).

In collaboration with Microsoft and the NIH Center for Information Technology (CIT), a Python script with PubMed, MedlinePlus, and ClinicalTrials.gov API add-ons was developed and connected to the MS AOAI Bot Service. The service utilizes a Semantic Kernel Bot3, which provides the chatbot architecture to determine the optimal keywords for searching within the relevant NLM product. This allows the bot to intelligently decide the best keywords based on the input (Appendix C).

The Bot Service also includes a system message (through prompt engineering) to set the tone and format the responses, providing CSRs with a ready-to-use template (Appendix D).

## Evaluation

### Initial Testing

Initial testing revealed success in grounding the model with NLM resources, which is essential to eliminate hallucinations2. However, further experimentation showed buggy performance, prompting Microsoft engineers to collaborate with NLM CIT to improve the code based on Alex’s documentation. Once the chatbot performed consistently, Alex tested it with 50 user inquiries. The initial examination showed that the chatbot provided accurate, up-to-date answers to consumer health questions. By referencing NLM products, the chatbot could cite resources used, generate answers based on content from the product websites, and provide hyperlinks to the relevant resources.

### Qualtrics Questionnaire

To gain better insight into the chatbot's success, a Qualtrics questionnaire was developed and sent to internal participants to evaluate the generated outputs. Alex developed an evaluation methodology by consulting with Kimberly Thomas to understand evaluation principles and drew inspiration from the NIST AI Management Framework4 to create an evaluation focused on both the utility and safety of the chatbot.

The questionnaire was based on nine themes shown in the table below.

Participants were also asked to rate their overall satisfaction with the GenAI outputs on a scale of 1-10. A selection of ten user inquiries and GenAI outputs from the chatbot, categorized into easy, medium, and hard inquiries as evaluated by Terry Ahmed and Mabel Mendez, were included in the Qualtrics questionnaire. Feedback from 29 participants revealed a 70% overall satisfaction rate with the generated responses. Some volunteers trialed the chatbot and their comments included:

* “I am having a good time with the chatbot. I'm asking questions about a health-related episode we had in my family this week and so far, it's giving excellent information.”
* “I think it does a great job answering this 'question' [where] the user isn't clearly asking anything.”
* “Tried a few questions and was very impressed by the responses.”

### Google Trial

During the evaluation phase, Alex collaborated with a team from Google to compare their product, MedLM-Large, a health domain-specific GenAI, with MS AOAI. The team developed a MedLM-Large and PubMed solution and produced results for ten user inquiries, which were included in the same Qualtrics questionnaire. Although a smaller cohort participated in this evaluation, the results for the Google solution approached an 80% satisfaction rate. These results were impressive, given the shorter development time, the solution being grounded in only one NLM product, and the lack of response formatting.

# Results

It was determined that a GenAI solution to generating customer responses is valuable to NLM. As a result of this proof-of-concept, development will continue and be expanded. This use case has displayed that a GenAI chatbot can generate accurate answers and provide citations to PubMed, MedlinePlus, and ClinicalTrials.gov. It also appears to have eliminated hallucinations as the chatbot will always provide an answer with a citation to a research article or NLM product, increasing accuracy of results as compared to baseline testing.

The use case was featured in a presentation to an NIH-wide AI Community of Practice meeting, a 2024 AI Summit, and accepted for publication as a program poster at the 2024 Medical Library Association Annual Meeting.

# Recommendations

Alex has provided the following recommendations for this project:

1. Utilize a More Robust Retrieval-Augmented Generation: Although this will require additional resources and time, development of a more robust RAG system is the best way to improve the accuracy, speed, and relevance of the generated responses. It will also be needed to craft a production-level product and move beyond the three NLM products.
2. Explore Other GenAI Products: Continue evaluating Google’s MedLM-Large and consider other solutions from IBM and Amazon Web Services. It may also be beneficial to develop an internal open-source solution.
3. Policy Versus GenAI Implementation: Determine which issues can be addressed through policy changes rather than relying solely on GenAI implementation. For example, a policy change to what may be considered medical advice as to allow CSRs to respond with a more empathetic tone and/or provide direct contextual answers (versus directing to the information). It may be beneficial to craft a customer-driven template to use as a draft.
4. **Consider NIH and HHS-Wide Implementation:** Leverage the wealth of resources, databases, and products available within NIH and HHS to train a future AI platform.
5. **Maintain Human Oversight:** Always keep a human-in-the-loop during the process and expand upon the evaluation by staying current with developments in AI safety, policy, and regulation.

# Conclusion

The "Evaluating the Value of ChatGPT in Generating Customer Responses" project has successfully demonstrated the potential of a GenAI platform to enhance customer service at NLM. Through a comprehensive proof-of-concept, the project showed that a GenAI chatbot could generate accurate, relevant, and well-cited responses to consumer health-related inquiries, effectively eliminating hallucinations, providing a solution to workflow efficiencies, and potentially improving customer satisfaction.

This project highlighted the significant value that GenAI can bring to NLM by increasing efficiency in handling the high volume of customer inquiries. The success of this use case may prove to be not only beneficial for CSRs but also pointed towards broader applications across NIH and HHS.

Moving forward, continued development and expansion of this GenAI platform are recommended. Emphasis should be placed on utilizing more robust retrieval-augmented generation techniques, exploring additional GenAI products, and addressing policy versus GenAI solutions.

A crucial recommendation is to always maintain human oversight in the process, ensuring adherence to the latest developments in AI safety, policy, and regulation. By leveraging the extensive resources within NIH and HHS, the NLM can continue to innovate and improve its customer service capabilities, ultimately advancing its mission to disseminate consumer health information efficiently and accurately.

# Resources

1. [Azure OpenAI Service Documentation](https://azure.microsoft.com/en-us/products/ai-services/openai-service/?ef_id=_k_EAIaIQobChMIpLLGi-bGgwMVjElHAR2sXAqPEAAYASAAEgJwHvD_BwE_k_&OCID=AIDcmm5edswduu_SEM__k_EAIaIQobChMIpLLGi-bGgwMVjElHAR2sXAqPEAAYASAAEgJwHvD_BwE_k_&gad_source=1&gclid=EAIaIQobChMIpLLGi-bGgwMVjElHAR2sXAqPEAAYASAAEgJwHvD_BwE)
2. “AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate” (https://www.ibm.com/topics/ai-hallucinations). The main concern for this project was hallucinations of articles.
3. [Semantic Kernel Bot-in-a-box](https://github.com/Azure/semantic-kernel-bot-in-a-box)
4. https://www.nist.gov/itl/ai-risk-management-framework

# Appendix A.



# Appendix B.









# Appendix C.



# Appendix D.

# Appendix D.

