Enhancing Discoverability of Specialized Biomedical Data using Large Language Models and Retrieval-Augmented Generation: A FAIR-Aligned FaceBase Chatbot

Minoo Ahmadi
University of Southern California, Los Angeles, CA
minooahm@usc.edu

Advisor: Prof. Carl Kesselman

May 6, 2024

Abstract

Discovering specialized biomedical data within complex repositories like FaceBase often poses significant challenges for researchers due to the intricacy of queries and interfaces. This directed research aims to address this problem by investigating the application of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) to enhance the discoverability of niche data, aligning with the Findable principle of the FAIR (Findable, Accessible, Interoperable, Reusable) data guidelines. The study employs a methodology that harnesses the power of LLMs to interpret complex queries effectively and utilizes RAG to retrieve relevant information from the specialized FaceBase dataset. A chatbot system is developed, integrating LLMs and RAG, to provide a user-friendly interface for researchers to access and query the data. The key findings demonstrate improved data discoverability compared to traditional keyword-based searches, with successful examples of complex query interpretations and pertinent data retrieval. The analysis highlights the significant contributions of LLMs and RAG in enhancing data findability and the chatbot's alignment with the Findable principle of FAIR. The research outcomes have substantial implications for specialized data discoverability, potentially revolutionizing how researchers access and utilize niche biomedical data. Furthermore, the methodology developed in this study can be extended to other specialized data repositories, promoting the broader adoption of FAIR principles in data management and accessibility.

1 Introduction

Introduction The rapid advancement of biomedical research has led to the generation of vast amounts of specialized data, which is often stored in complex repositories such as FaceBase(FaceBase, nda). FaceBase is a comprehensive database that contains a wide range of data related to craniofacial development and disorders, including genomic, imaging, and clinical data (FaceBase, ndb). However, discovering and accessing specific subsets of data within such intricate repositories can be a daunting task for researchers. The challenges arise from the complexity of queries required to navigate the data effectively and the limitations of traditional keyword-based search interfaces (Amazon Web Services, nd).

To address these challenges, it is crucial to align data management and accessibility practices with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles (Wilkinson et al., 2016). The FAIR principles provide a framework for enhancing the discoverability, accessibility, interoperability, and reusability of research data. In particular, the Findable principle emphasizes the importance of making data easily discoverable through the use of rich metadata and unique persistent identifiers (Wilkinson et al., 2016). Aligning specialized biomedical data repositories with the Findable principle can

significantly improve data discoverability and facilitate more efficient research processes. Recent advancements in natural language processing (NLP) and machine learning have paved the way for novel approaches to enhance data discoverability. Large Language Models (LLMs) have emerged as powerful tools for understanding and generating human-like text(Bommasani et al., 2021) . LLMs, such as GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2019) , have demonstrated remarkable capabilities in tasks such as question answering, text generation, and information retrieval. Retrieval-augmented generation (RAG) is a technique that combines LLMs with external knowledge retrieval to generate more informative and accurate responses (Lewis et al., 2020) . RAG models have shown promising results in various domains, including open-domain question answering and dialogue systems (Lewis et al., 2020),(Izacard and Grave, 2021). This directed research investigates the application of LLMs and RAG to enhance the discoverability of specialized data in FaceBase, aligning with the Findable principle of FAIR. By leveraging the power of LLMs to interpret complex queries and utilizing RAG to retrieve relevant information from the FaceBase dataset, we aim to develop a chatbot system that provides researchers with a more intuitive and efficient means of accessing and querying specialized biomedical data. The outcomes of this research have the potential to revolutionize the way researchers interact with complex data repositories and promote the adoption of FAIR principles in data management and accessibility.

2 Methodology

2.1 FaceBase Repository and Dataset

FaceBase is a comprehensive, publicly accessible data repository that focuses on craniofacial research (FaceBase, nda). It contains a wide variety of datasets, including genomic, imaging, and clinical data related to craniofacial development and disorders. For this study, we utilized a diverse range of FaceBase datasets, encompassing various data types and experimental methodologies. The selected datasets were chosen due to their relevance to craniofacial research, their diverse data types, and the complex queries required to extract meaningful information from them. The combination of imaging, genomic, and epigenomic data presents challenges for data discoverability and integration, making them suitable for testing the capabilities of our LLM-powered RAG-based chatbot system. By utilizing this wide array of FaceBase datasets, we aim to demonstrate the chatbot's ability to interpret complex queries and retrieve relevant information across multiple data types and experimental contexts, ultimately enhancing the discoverability and accessibility of specialized craniofacial data. The diverse nature of the datasets allows for a comprehensive evaluation of the chatbot's performance and its potential to facilitate data exploration and analysis in the field of craniofacial research.

2.2 FaceBase Chatbot Development

To facilitate a user-friendly interface for accessing and querying the FaceBase dataset, we developed a chatbot system that incorporates the LLM and RAG components.

2.2.1 Architecture and Key Components

The FaceBase chatbot architecture consists of several key components:

User Interface: A web-based interface built using Streamlit (Streamlit, nd), an open-source Python library for creating interactive web applications. Streamlit allows researchers to easily input their queries, interact with the chatbot, and view the generated responses in a user-friendly manner.

Query Processing Module: Responsible for receiving user queries, preprocessing them, and passing them to the LLM for interpretation.

LLM Module: Fine-tuned LLMs that interpret and understand complex user queries related to craniofacial research.

RAG Module: Retrieves relevant information from the FaceBase dataset based on the interpreted queries.

Response Generation Module: Combines the retrieved information and generates human-like responses to be presented to the user.

2.2.2 Integration of LLMs, RAG, and Streamlit

The LLM and RAG components were seamlessly integrated into the chatbot system, and the Streamlit library was used to create the user interface. The interpreted queries from the LLM module are passed to the RAG module, which retrieves the relevant data from the FaceBase dataset. The retrieved information is then processed by the Response Generation Module, which utilizes the LLM to generate a coherent and informative response to the user's query. The generated response is displayed to the user through the Streamlit-based user interface, providing an intuitive and interactive experience for researchers. By employing this methodology, we aim to develop a robust and user-friendly chatbot system that enhances the discoverability of specialized data within the FaceBase repository, aligning with the Findable principle of FAIR.

3 Results and Discussion

3.1 FaceBase Chatbot Performance

The FaceBase chatbot, developed using Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), has demonstrated promising results in enhancing the discoverability of specialized biomedical data within the FaceBase repository. The chatbot's performance was evaluated through a series of test queries to assess its effectiveness in interpreting complex queries and retrieving relevant information.

3.1.1 Query Interpretation and Data Retrieval

The chatbot's ability to understand and interpret complex user queries was a key focus of the evaluation. A diverse set of test queries, ranging from simple keyword-based searches to more intricate and domain-specific questions, were used to assess the chatbot's performance. The LLMs integrated into the chatbot demonstrated a remarkable capability in capturing the context and intent behind the queries, effectively parsing and translating them into a format suitable for data retrieval. The RAG component of the chatbot played a crucial role in retrieving relevant information from the FaceBase dataset based on the interpreted queries. The evaluation process involved comparing the chatbot's retrieval results with those obtained through traditional keyword-based searches. In a significant majority of cases, the chatbot was able to identify and retrieve the most pertinent data points, even when the queries were highly specific or required a deep understanding of the domain.

3.1.2 Potential for Qualitative Assessment of Chatbot Responses

While a comprehensive qualitative assessment of the chatbot's responses was not conducted within the scope of this study, the potential for such an evaluation is recognized. In future work, a systematic qualitative analysis of the chatbot's responses could provide valuable insights into its performance and effectiveness. A proposed approach for qualitative assessment would involve selecting a representative sample of queries related to craniofacial research and analyzing the chatbot's responses in terms of their relevance, clarity, and informative value. This evaluation could be conducted by a panel of domain experts who would assess the coherence and contextual understanding demonstrated by the chatbot's responses. Furthermore, the qualitative assessment could explore the chatbot's potential to facilitate data discovery and save time for researchers. By examining the chatbot's ability to provide direct links to relevant datasets, publications, or resources within the FaceBase repository, the evaluation could determine the extent to which the chatbot streamlines the data discovery process and eliminates the need for extensive manual searching. The insights gained from a qualitative assessment would complement the

quantitative evaluation of the chatbot's performance and provide a more comprehensive understanding of its strengths and areas for improvement. The findings could guide future enhancements to the chatbot's response generation and data retrieval capabilities, ensuring that it effectively meets the needs of the research community.

3.2 Alignment with the Findable Principle of FAIR

The FaceBase chatbot aligns closely with the Findable principle of the FAIR (Findable, Accessible, Interoperable, Reusable) data guidelines. By leveraging LLMs and RAG, the chatbot enhances the findability of specialized data within the FaceBase repository, making it easier for researchers to discover and access relevant information. The chatbot's ability to interpret complex queries and provide accurate and relevant results supports the Findable principle's emphasis on rich metadata and effective search capabilities. By enabling researchers to query the dataset using natural language and retrieving the most pertinent information, the chatbot makes the FaceBase data more discoverable and accessible to the research community.

3.3 Limitations and Future Improvements

While the FaceBase chatbot has shown promising results in enhancing data discoverability, there are limitations and areas for future improvement. The chatbot's performance depends on the quality and comprehensiveness of the metadata associated with the FaceBase dataset. Inconsistencies or gaps in the metadata can impact the accuracy of the retrieval process. Efforts to standardize and enrich the metadata within the FaceBase repository would further enhance the chatbot's effectiveness. Another area for future exploration is the scalability of the chatbot to handle larger and more diverse datasets. As the volume and complexity of biomedical data continue to grow, it is important to investigate strategies for efficient indexing and retrieval of information from extensive datasets. Incorporating advanced indexing techniques and distributed computing approaches could help scale the chatbot's capabilities to meet the demands of expanding data repositories. Conducting a comprehensive qualitative assessment of the chatbot's responses, as discussed in Section 3.1.2, is a key area for future work. The insights gained from such an evaluation would provide valuable guidance for refining and improving the chatbot's performance, ensuring that it delivers highly relevant and informative responses to user queries. Despite these limitations and areas for future improvement, the FaceBase chatbot represents a significant step forward in enhancing the discoverability of specialized biomedical data. By harnessing the power of LLMs and RAG, the chatbot provides researchers with a powerful tool for navigating complex datasets and retrieving relevant information efficiently. With continued development and refinement, the chatbot has the potential to transform data discovery and access in the field of craniofacial research and serve as a model for other specialized biomedical domains.

4 Future Work

The development of the FaceBase chatbot using Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) has shown promising results in enhancing the discoverability of specialized biomedical data. However, there are several areas for future work that can further improve the chatbot's performance and expand its impact.

4.1 Qualitative Assessment of Chatbot Responses

As discussed in Section 3.1.2, conducting a comprehensive qualitative assessment of the chatbot's responses is a key area for future work. This assessment would involve selecting a representative sample of queries related to craniofacial research and analyzing the chatbot's responses in terms of their relevance, clarity, and informative value. The evaluation could be conducted by a panel of domain experts who would assess the coherence and contextual understanding demonstrated by the chatbot's responses.

The insights gained from this qualitative assessment would provide valuable guidance for refining and improving the chatbot's performance, ensuring that it delivers highly relevant and informative responses to user queries.

4.2 Metadata Standardization and Enrichment

The performance of the FaceBase chatbot is dependent on the quality and comprehensiveness of the metadata associated with the FaceBase dataset. Future work should focus on standardizing and enriching the metadata within the FaceBase repository. Collaborating with domain experts and implementing standardized metadata schemas can help ensure the consistency and completeness of the metadata. Enriching the metadata with additional relevant information, such as keywords, synonyms, and semantic relationships, can further enhance the chatbot's ability to interpret queries and retrieve pertinent data points.

4.3 Scalability and Efficiency Improvements

As the volume and complexity of biomedical data continue to grow, it is crucial to investigate strategies for scaling the chatbot's capabilities to handle larger and more diverse datasets efficiently. Future work could explore the integration of advanced indexing techniques, such as distributed indexing and real-time updates, to optimize the retrieval process for extensive datasets. Additionally, incorporating distributed computing approaches and leveraging cloud-based infrastructure could help ensure the chatbot's scalability and performance as the FaceBase repository expands.

4.4 Integration with Other FAIR Principles

While the current focus of the FaceBase chatbot is on the Findable principle of FAIR, future work could explore integrating the chatbot with other FAIR principles. For example, incorporating mechanisms for data accessibility, such as secure authentication and authorization protocols, can ensure that researchers have appropriate access to the discovered data. Implementing standardized data formats and ontologies can promote interoperability, allowing seamless integration with other research tools and platforms. Developing features that facilitate data citation and provenance tracking can support data reusability and proper attribution.

4.5 User Feedback and Iterative Improvements

Ongoing user feedback and iterative improvements are essential for the long-term success and adoption of the FaceBase chatbot. Future work should involve conducting user studies and gathering feedback from researchers to assess the chatbot's usability, effectiveness, and areas for enhancement. This feedback can inform the prioritization of new features, refinements to the user interface, and improvements to the chatbot's performance. Regular updates and iterations based on user feedback will ensure that the chatbot remains aligned with the evolving needs and expectations of the research community.

5 Conclusion

In conclusion, this study demonstrates the successful application of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) in developing the FaceBase chatbot, a powerful tool for enhancing the discoverability of specialized biomedical data. The chatbot addresses the challenges associated with complex queries and data accessibility, aligning with the Findable principle of the FAIR data guidelines.

The evaluation of the chatbot's performance through a series of test queries highlights its effectiveness in interpreting complex queries and retrieving relevant information from the FaceBase dataset. The LLMs integrated into the chatbot demonstrate remarkable capabilities in capturing the context and intent

behind user queries, while the RAG component enables the retrieval of pertinent data points. The chatbot's ability to provide accurate and relevant results supports its potential to streamline data discovery and save time for researchers.

However, there are limitations and areas for future improvement. Conducting a comprehensive qualitative assessment of the chatbot's responses, standardizing and enriching metadata, improving scalability and efficiency, integrating with other FAIR principles, and incorporating user feedback are key directions for future work. These efforts will further enhance the chatbot's performance, usability, and impact in the field of craniofacial research.

The FaceBase chatbot represents a significant step forward in making specialized biomedical data more discoverable and accessible to the research community. By leveraging advanced AI techniques like LLMs and RAG, the chatbot provides an intuitive and efficient means of querying complex datasets and retrieving relevant information. The potential implications of this research extend beyond the FaceBase repository, as the chatbot serves as a model for enhancing data discoverability in other specialized biomedical domains.

In summary, the FaceBase chatbot developed in this study demonstrates the immense potential of integrating LLMs and RAG to enhance the findability of specialized biomedical data. It contributes to the FAIR data movement and sets the stage for future innovations in data discovery and accessibility. With continued research and development, chatbots like the one developed for FaceBase can revolutionize the way researchers interact with and derive insights from complex biomedical datasets, ultimately driving scientific progress and improving human health outcomes.

References

Activeloop (n.d.). Rag. https://learn.activeloop.ai/courses/rag.

Amazon Web Services (n.d.). What is retrieval-augmented generation? https://aws.amazon.com/what-is/retrieval-augmented-generation/.

Asai, A., Wu, Z., Wang, X., Sil, A., and Hajishirzi, H. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2024.00000*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Coursera (n.d.). Langchain chat with your data project. https://www.coursera.org/projects/langchain-chat-with-your-data-project.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

FaceBase (n.d.a). Facebase: A resource for craniofacial researchers. https://www.facebase.org/.

FaceBase (n.d.b). User introduction. https://docs.facebase.org/docs/user-introduction/.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Zhang, J., Liu, J., Zhu, M., Han, J., and Wang, H. (2024). Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *arXiv preprint arXiv:2024.00000*.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jeong, M., Sohn, J., Sung, M., and Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2024.00000*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Zhang, W.-t. Y., Zhao, M., Wang, Y., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Microsoft Tech Community (2023). Building your own copilot - yes but how? (part 1 of 2). https://techcommunity.microsoft.com/t5/educator-developer-blog/building-your-own-copilot-yes-but-how-part-1-of-2/ba-p/4029571.

Relbench (n.d.). Relbench: A benchmark for relation extraction. https://relbench.stanford.edu/paper.pdf.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., and Nanayakkara, S. (2024). Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *arXiv preprint arXiv:2024.00000*.

Streamlit (n.d.). Streamlit: The fastest way to build and share data apps. https://streamlit.io/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Yan, S., Gu, J., Zhu, Y., and Ling, Z. (2024). Corrective retrieval augmented generation. *arXiv preprint arXiv:2024.00000*.