Paper IC07

Revolutionizing Informed Consent Form Analysis with Generative AI: Enhancing Efficiency in Drug Development and Data Re-use

Hangyu Liu, Biogen, Cambridge, US
Andrew Borgman, Biogen, Cambridge, US
Jake Gagnon, Biogen, Cambridge, US
Dan Boisvert, Biogen, Cambridge, US
Haleh Valian, Biogen, Cambridge, US

ABSTRACT

The pharmaceutical industry faces challenges in re-using samples, data, and images for secondary research. Complex guidelines governing data re-use are often buried in Informed Consent Forms (ICFs) that are manually reviewed by subject matter experts (SMEs) on a case-by-case basis. This process is time-consuming and hampers efficiency. We developed a novel Artificial Intelligence (AI)-driven workflow leveraging Large Language Models (LLMs) for computer-assisted analysis of ICFs. This cutting-edge method substantially simplifies our process by using generative AI to summarize and identify allowable data usage promptly and precisely. This adaptable framework is applicable to any company requiring comprehension and analysis of ICFs or data use agreements. Through a domain-specific validation framework, combining manual SME review and quantitative measurements, we achieved 96% accuracy in Biogen study ICFs tested to date. The paper delves into the current workflow design, associated cost estimation, a multi-layered strategy to address potential hallucinations in LLMs, considerations for safeguarding data privacy within a secure computing environment, the design and development of an interactive chatbot powered by LLMs, and other pertinent technical details. Our overarching goal is to enhance data reusability by improving the efficiency of ICF document analysis and expediting the dissemination of therapeutic insights to the patient community.

Keywords: Large Language Models, Generative AI, Informed Consent Forms, Data sharing

INTRODUCTION

Informed Consent Forms (ICFs) hold a pivotal role in clinical trials, safeguarding the rights, safety, and wellbeing of trial participants. They serve as a vital document that researchers and clinicians often need to verify to ensure a patient's agreement for the utilization of specific samples in biomarker, translational science, and other scientific research areas. Currently, we have a dedicated data sharing team at Biogen to help scientists to review and verify the allowable data re-use guideline in ICFs. However, as the complexity of clinical trials grows, compounded by variations in ICFs across different countries, the data sharing team find themselves immersed in a labor-intensive task of manually reviewing numerous documents to extract necessary information. This challenge is exacerbated by the fact that multi-national clinical trials may have multiple country-specific versions of the ICF, making the review process even more cumbersome and time-consuming. Researchers often encounter difficulties in swiftly locating and extracting relevant data, leading to delays in data analysis and potentially hindering the progress of biomarker, translational science, and other scientific research.

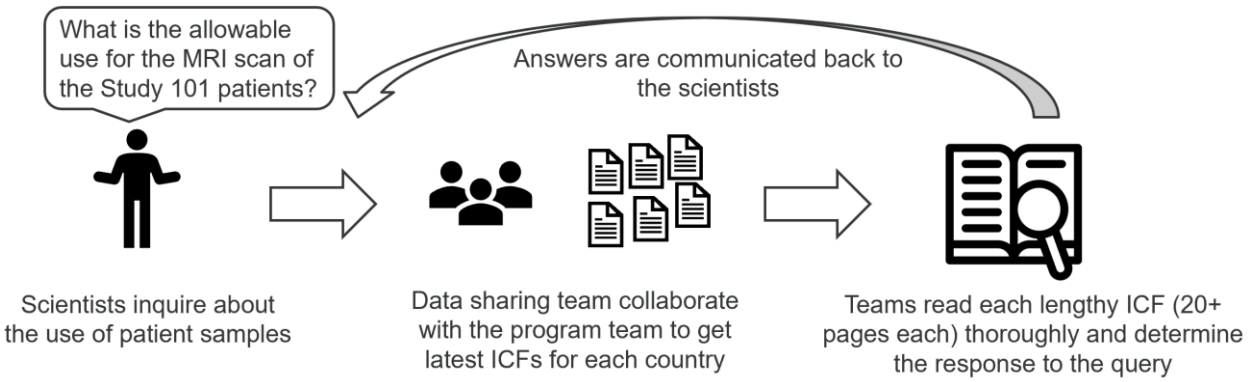

Figure 1: Overview of the Data Re-Use Enquiry Response Process

To address these challenges, our paper introduces an innovative solution—a novel Artificial Intelligence (AI)-driven workflow leveraging Large Language Models (LLMs) for the computer-assisted analysis of ICFs. The proposed solution aims to revolutionize the traditional paradigm by employing advanced AI techniques, specifically the LLM-powered Retrieval-Augmented Generation (RAG) framework, to streamline the extraction of pertinent information from ICFs.

LLMs, a recent advancement in artificial intelligence field, have been making waves in the field of natural language understanding and generation. These complex algorithms, such as GPT-3.5 (OpenAI, 2022) GPT-4 (OpenAI, 2023), and LLaMA (Touvron, H. et al., 2023), are revolutionizing computer's ability to interact with and understand text data. LLMs are trained on massive datasets of text and programming code, enabling them to grasp intricate linguistic patterns and generate human-quality text with remarkable fluency and coherence (Brown et al., 2022). This capability extends beyond simple text generation, it also allows LLMs to perform diverse Natural Language Processing (NLP) tasks, including question answering, summarization, translation, and even more complex tasks. The pharmaceutical industry has recently started to embrace the transformative capabilities of LLMs as a cutting-edge tool to navigate the complexities of drug development, regulatory compliance, and information retrieval. Researchers in the pharmaceutical industry leverage LLMs to extract valuable insights from scientific literature, patents, and clinical trial reports, accelerating the process of identifying potential drug candidates and understanding their safety profiles. The capacity of LLMs to understand context, generate human-like text, and interpret complex scientific jargon positions them as invaluable assets in the pharmaceutical landscape. As we explore in this paper, the integration of LLMs in the analysis of ICFs presents a novel application of this technology, offering a paradigm shift in the approach to mining information within ICFs and improving the efficiency of patient's data re-use. And unlike traditional machine learning models that require explicit training for each specific task, these LLMs, pre-trained on vast datasets, possess an unparalleled capacity to understand and generate human-like text in various domains. LLMs can generalize from the broad linguistic context they were provided with and generate human-like text that is contextually relevant. In the context of ICFs, these models can understand and interpret the content of ICFs, providing quick and accurate responses to specific queries, such as "How long will samples/Images be stored for as mentioned in the ICF?", '"How will samples be used as mentioned in the ICF?", etc.

While LLMs like GPT-3.5 showcase exceptional language understanding and generation capabilities, the presence of low-quality and biased data in their training sets, along with their unfiltered and uncontrollable access to a vast expanse of knowledge introduces a potential drawback in the ICF use case. Simply asking a question to a generic LLM may yield responses based on a broad spectrum of information it has encountered, potentially including details not explicitly present in the ICF document. But in the context of ICF analysis, the need for evidence-based responses is imperative. To address this specificity requirement and ensure that responses are anchored in the evidence provided within the ICF, we leverage the Retrieval-Augmented Generation (RAG) framework, which is a synergy between the benefits of generative and retrieval-based models (Lewis et al., 2020). RAG retrieves relevant contents from documents and then uses them as context for generating responses. It leverages LLM's language understanding and reasoning capabilities to search through vast document corpora, yielding an advanced information retrieval system with impressive accuracy. By combining the strength of RAG to pinpoint relevant information, alongside the capability of LLMs and a carefully engineered prompt, our system ensures that answers are not only contextually accurate but are grounded in the evidence contained within the ICF documents, thereby aligning with the rigorous standards of evidence-based research in the pharmaceutical industry.

To facilitate seamless interaction with our AI-driven workflow, we have developed a user-friendly web-based application using the Python Streamlit framework. This chat-based platform allows users to interact with ICF documents in natural language and receive precise and prompt answers. In essence, it offers an intelligent, more intuitive way of dealing with ICF documents, alleviating the need for endless scrolling and searching. Instead, users can interact with their ICF documents as if they are having a conversation, ushering in a new era in the pharmaceutical industry where AI enhances productivity and efficiency.

In this paper, we will discuss the details of our novel AI-driven workflow for computer-assisted analysis of ICFs, providing insights into the technology, its implementation, and the benefits it offers to the pharmaceutical industry. We believe this represents a significant effort in addressing the practical challenges faced in clinical and scientific research, paving the way for an improved, efficient, and more accurate ICF analysis and data re-use process.

METHODOLOGIES

In this section, we delve into the technical details of our AI-driven workflow that is aimed at addressing challenges associated with the reuse of samples, data, and images for secondary research. Several key components will be explored in the subsequent discussions: (1) The design of the overall technical framework; (2) The strategic deployment and utilization of LLMs such as GPT-3.5, GPT-4, LLaMA, LLaMA2, and FLAN-T5, all within a secure and private environment; (3) The creation of a customized knowledge base with a Chroma vector database; (4) The validation of results through the utilization of a domain-specific framework; and (5) The development of an intuitive

and user-friendly web-based application using the Python Streamlit framework. Each of these components contributes collectively to the goal of enhancing the efficiency and effectiveness of secondary research endeavors.

OVERALL TECHNICAL FRAMEWORK

LLMs are designed for generalized tasks, so they lack knowledge of our company's private data, particularly regarding our ICF documents. Furthermore, these models are static with a cutoff date of their training data; for instance, ChatGPT's data is current only up to September 2021. Also, there is a risk of hallucination where inaccurate or misleading information can be generated by LLMs. Due to these limitations, we opt for the RAG framework to address these constraints. RAG ensures access to real-time and domain-specific data from an external knowledge base, minimizing the likelihood of hallucinations. In our ICF use case, RAG provides additional context and factual information to our Generative AI (GenAI) workflow's LLM during the generation process. This approach not only resolves issues related to recency and domain specificity but also enables GenAI applications to cite their sources, enhancing transparency and auditability, which is crucial in the pharmaceutical industry. RAG facilitates a more traceable way to understand and validate GenAI applications.

At a high level, there are three major components to setting up a RAG framework over our internal ICF data: (1) Ingestion of ICF documents as a knowledge base; (2) Context-based semantic search; and (3) LLM-powered question answering. The below diagram shows a brief overview of the technical framework:

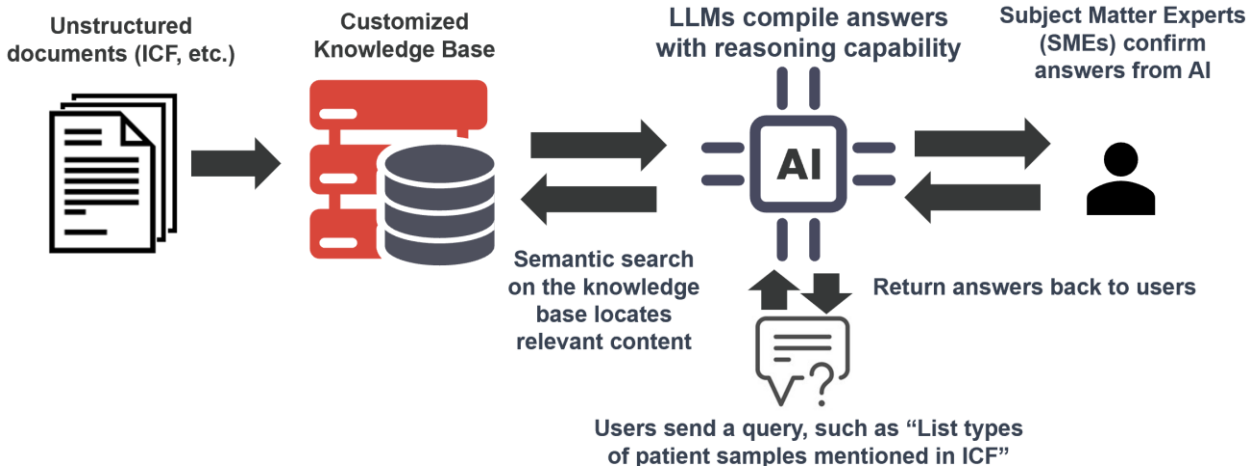

Figure 2: Brief Overview of the technical framework

CREATION OF KNOWLEDGE BASE

As computers cannot directly comprehend ICF documents, it becomes necessary to transform them into a machine-readable format. This is achieved through the application of NLP techniques, specifically utilizing an embedding model. An embedding model aims to capture the semantic meaning of the text by converting them into vectors, and the most important thing to understand is that embedding vectors aim to maintain similarity for sentences conveying similar ideas, even when they are expressed through different wording. In our use case, we use embedding model to convert ICF documents into numerical vectors, then establishing a vector database to store the document content as a knowledge base.

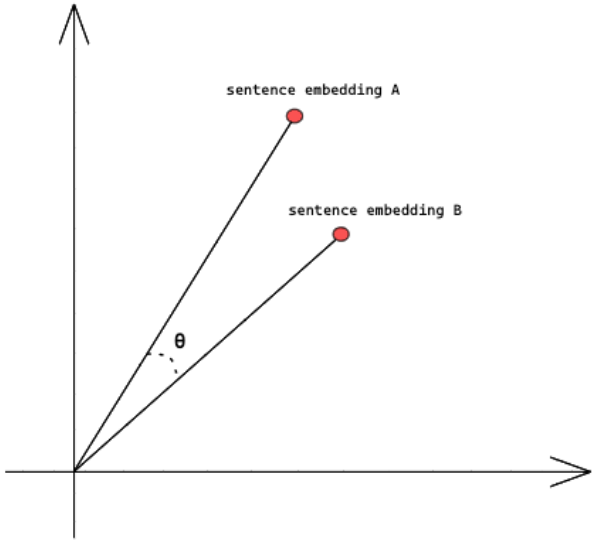

Figure 3: Example of Embedding Vectors of Two Similar Sentences

In our study, we assessed different embedding models to identify the optimal representation for textual data. Explored models included OpenAI's text-embedding-ada-002, all-mpnet-base-v2, and bge-large-en, selected for their advancements in capturing semantic relationships and proven efficacy in diverse NLP tasks. Following comparative analysis, OpenAI's text-embedding-ada-002 demonstrated superior performance, exhibiting robustness and versatility across various parameters.

Another important caveat is that embedding models are constrained by a token limit, indicating the maximum number of tokens they can process in a single sequence. In the context of text-embedding-ada-002, the model has a maximum token [1]limit of 8191. This limitation necessitates careful consideration during the embedding process, as entire ICF documents (usually 20+ pages) are very likely to exceed this token limit. Consequently, a chunking approach is needed to handle long documents to ensure effective utilization of the embedding model while adhering to its token limitations. Another important consideration in the process of embedding is the context window of LLMs. The context window represents the model's capability to ingest input data, and LLMs typically have a context window ranging from around 4,000 tokens to sometimes extending to 32,000 tokens. Embedding the entire ICF document without chunking and feeding the results as context to LLMs may surpass the model's context window limit, leading to potential challenges during LLM-powered question answering. Hence, the strategic chunking of documents is necessary to address both the token limit of the embedding model and the context window constraints of LLMs, ensuring effective and accurate utilization of the embedding model in the ICF analysis workflow. While various chunking methods were investigated, it is pertinent to note that the detailed exploration of these methods falls outside the primary scope of this work. As such, an in-depth discussion on chunking methodologies is deferred for future discussion.

Below, we present a schematic diagram illustrating our processes of document chunking into smaller units, subsequent conversion into embedding vectors, and the eventual creation of an ICF knowledge base in the form of a vector database. We used LangChain framework to facilitate PDF document processing, chunking, and interaction with the OpenAI API, ensuring a seamless integration of these crucial components. LangChain is an open-source framework designed to simplify the development of applications powered by LLMs. It offers a variety of tools and APIs that make it easy to connect LLMs to other data sources, interact with their environment, and build complex applications. Furthermore, Chroma served as the designated vector database for storing the resultant embedding vectors. The vector database is a type of database optimized for the storage and retrieval of embedding vectors generated by the embedding models. Chroma is an open-source vector database provider, which provides a structured and efficient storage solution, enabling quick access and retrieval of embedding vectors for downstream applications. It enhances the overall performance of our AI-driven workflow by facilitating rapid querying and retrieval of pertinent information encapsulated within the ICF knowledge base.

[1] A token is a fundamental unit of text, representing individual words or subwords extracted through tokenization for analysis and processing. Tokenization is a fundamental step in NLP tasks, enabling algorithms to better understand and process language by treating words or subwords as discrete elements.

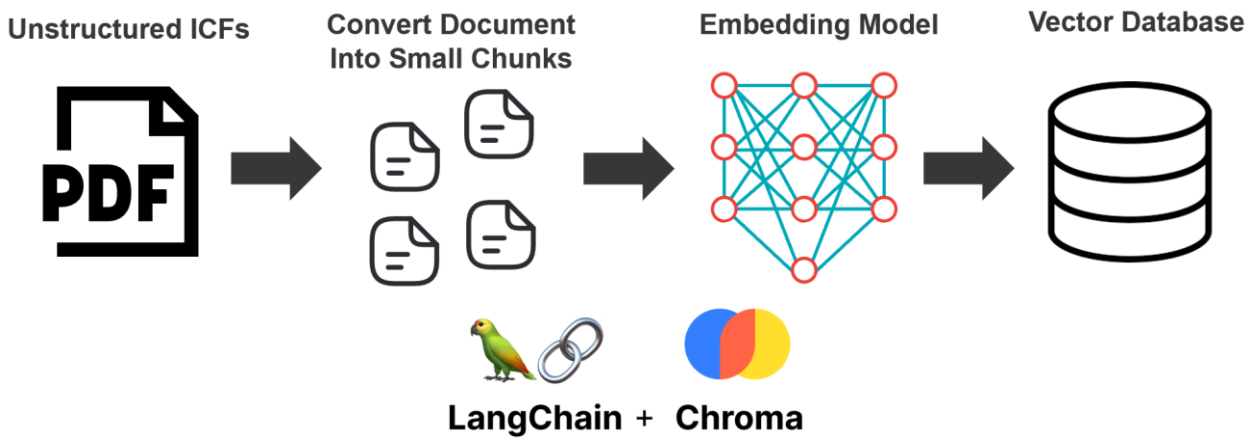

Figure 4: Steps To Create Knowledge Base From ICF Documents

LLM-POWERED QUESTION ANSWERING WORKFLOW

Following the conversion of ICF documents into a vector database, now our vector database contains the numerical representations of our ICF documents. This repository of embedding vectors then serves as the foundation for constructing an LLM-powered question-answering workflow, facilitating advanced and context-aware interactions with the ICF content encoded in the vector database.

Upon receiving a user query, such as ‘List all type of samples mentioned in the ICF’, our workflow first employs an embedding model to translate the natural language search terms into embedding vectors. These vectors are then dispatched to the vector database, initiating a "nearest neighbor" search mechanism. This process, known as semantic search, goes beyond traditional keyword matching by searching vectors that most closely align with the user's intended meaning. The emphasis lies in comprehending the true essence of the user's inquiry and retrieving relevant information, distinguishing itself from a mere keyword-based approach. Consequently, this refined approach enhances the precision and relevance of search results, aligning more closely with the user's intent.

Moreover, considering the context window size limitations of LLMs, our approach entails returning only the top k most similar document chunks from the semantic search, where k is a hyperparameter subject to fine-tuning. A larger value of k may encompass more contextual information but introduces potential noise. Further insights into the optimization strategy for the hyperparameter k will be discussed in the subsequent validation framework section. Once the top k relevant results are obtained, our workflow incorporates them into the LLM as context, alongside the user's query. This action prompts the LLM to undertake its generative task, utilizing its access to the most relevant and foundational information extracted from ICFs within our vector database. The integration of the RAG framework within this workflow serves to mitigate the likelihood of hallucination, ensuring the provision of accurate responses to the user's query. Below is a diagram to demonstrate our LLM-Powered question answering workflow:

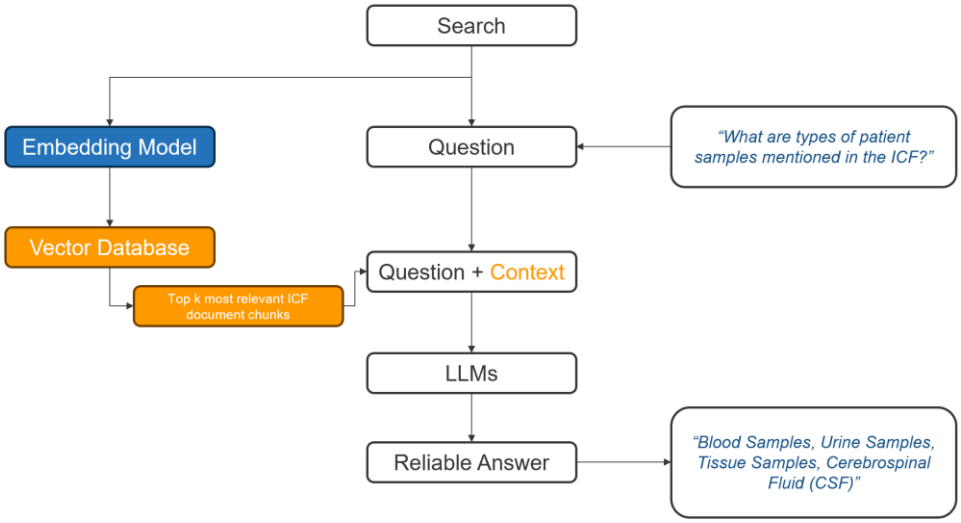

Figure 5: LLM-Powered Question Answering Workflow

DEPLOYMENT AND UTILIZATION OF LLMs

Due to the confidential nature of ICF documents, it is imperative to utilize LLMs with a commitment to privacy. In our study, we conducted comprehensive experiments involving a diverse array of LLMs, including both paid and open-source models. For the deployment of paid models, such as OpenAI's GPT-3.5, GPT-4, and text-embedding-ada-002, we opted for the Azure Cloud platform. To uphold data integrity and privacy, we established a data sharing agreement with Microsoft, ensuring that all data remained securely stored within our subscription, encrypted, and scheduled for auto-deletion after 30 days.

Simultaneously, a comprehensive evaluation of various open-source models, including LLaMA, FLAN-T5, Llama-2-13b-chat, Wizard LM 70B, all-mpnet-base-v2, and bge-large-en, was conducted. These models were deployed internally, and we tried both the Databricks hosting service and the vLLM hosting framework on High-Performance Computing (HPC) infrastructure, both deployment strategies ensured the restriction of ICF data within Biogen's secure internal network. During this evaluation, we carefully considered factors such as inference speed and hardware requirements, especially significant for larger models. While techniques, such as model quantization, offered a potential solution to hardware constraints, it came with trade-offs in terms of reduced capabilities and accuracy. The overall performance of these open-source models was also thoroughly assessed to ascertain their efficacy within our workflow.

Prompt engineering also plays a pivotal role in optimizing the utilization of LLMs, particularly in the context of our ICF use case. The tailored prompt employed in our experimentation serves as a strategic advantage in guiding the LLM to provide contextually accurate and document-grounded responses. The specified prompt, structured as a personal Bot assistant for answering ICF-related questions, includes explicit instructions emphasizing the importance of deriving answers solely from the provided documents, preventing the hallucination of the LLM. Below is the prompt template we created:

```
# template for prompt message sent to LLMs
prompt_template = """You are a personal Bot assistant for answering any questions
about documents of Informed Consent Forms.\nYou are given a question and a set
of documents. If the user's question requires you to provide specific
information from the documents, give your answer based only on the documents
provided below. DON'T generate an answer that is NOT written in the provided
documents. If you don't find the answer to the user's question with the documents
provided to you below, answer that you didn't find the answer in the documents.

QUESTION: {question}

DOCUMENTS:

{context}

ANSWER:

"""
```

Figure 6: Prompt Template for ICF Document Question Answering

The structured prompt introduces a set format including the user's question, the relevant documents, and a placeholder for the answer, ensuring a consistent and standardized approach in LLM interaction. This design is important in enforcing constraints necessary for reliable information retrieval from ICF documents, aligning with the rigorous nature of the biomedical research domain. The question component serves as the user's inquiry, while the documents section provides the contextual information available to the LLM for formulating responses. The explicit instruction to refrain from generating answers not in the documents willreduce the risk of introducing inaccuracies. If the LLM fails to find relevant information, it is instructed to respond honestly, stating that the answer was not found in the provided documents. Prompt engineering is a critical factor in refining the capabilities of LLMs for our ICF use case, ensuring that responses are both precise and grounded in the context of the documents. This well-crafted prompt template enhances the reliability of the LLM-powered question-answering workflow, highlighting the significance of prompt engineering in optimizing the performance of LLMs in domain-specific applications. In addition, it is important to note that our experimentation included various other prompts; however, as these explorations are not the central focus of this paper, specific details regarding them are not provided here.

To enhance the performance of our LLM-powered question-answering workflow, we incorporated text summarization as another key strategy. Recognizing the inherent non-deterministic nature prevalent in many LLMs like GPT-4 and

GPT-3.5, even under a temperature setting of 0 often associated with Mixture of Experts (MoE) approaches, we have identified a practical means to improve output stability and robustness. By iteratively summarizing the outputs of LLMs (e.g., 5 times) through the LLM itself, we observed a reduction in output variability and an enhancement in comprehensiveness. This technique proves particularly beneficial in our ICF analysis, especially when dealing with queries that require answers scattered across multiple sections of the ICF document.

After extensive experimentation, we determined that the optimal combination for our purposes consists of OpenAI's GPT-3.5, coupled with the above tailored prompt template and summarization approach, and the text-embedding-ada-002 as the chosen embedding model. This model setting demonstrated superior performance in capturing semantic relationships, aligning with the goals of our ICF analysis workflow. In the subsequent section, we delve into the validation framework and model selection process, providing a comprehensive overview of how we assessed the performance of these models and determined the optimal configuration tailored to our specific ICF use case.

VALIDATION FRAMEWORK

For the validation of our workflow, we leveraged pre-existing ICF language forms prepared by Subject Matter Experts (SMEs) to serve as our validation dataset. The ICF language form comprises question-and-answer pairs pertaining to various aspects of ICF content. For instance, a question such as 'Extract of ICF Text relating to confidentiality' corresponds to an SME-provided answer, serving as the gold standard for comparative evaluation. Notably, we sampled two ICF language forms from two studies, amounting to a validation set encompassing 150 question-answer pairs.

To assess the model's performance, we employed a dual-validation methodology. Firstly, we calculated the cosine similarity score between the model-generated answers and the human-provided answers, utilizing this metric to gauge the quality and accuracy of the model's responses. Additionally, SMEs manually reviewed the model's answers, providing valuable feedback on the nuances and context-specific elements. Throughout this validation process, multiple hyperparameters were fine-tuned, including document chunking strategies, LLM model selection, and the top 'k' for semantic search results.

The validation efforts resulted in the identification of the best-performing model with the following parameters:

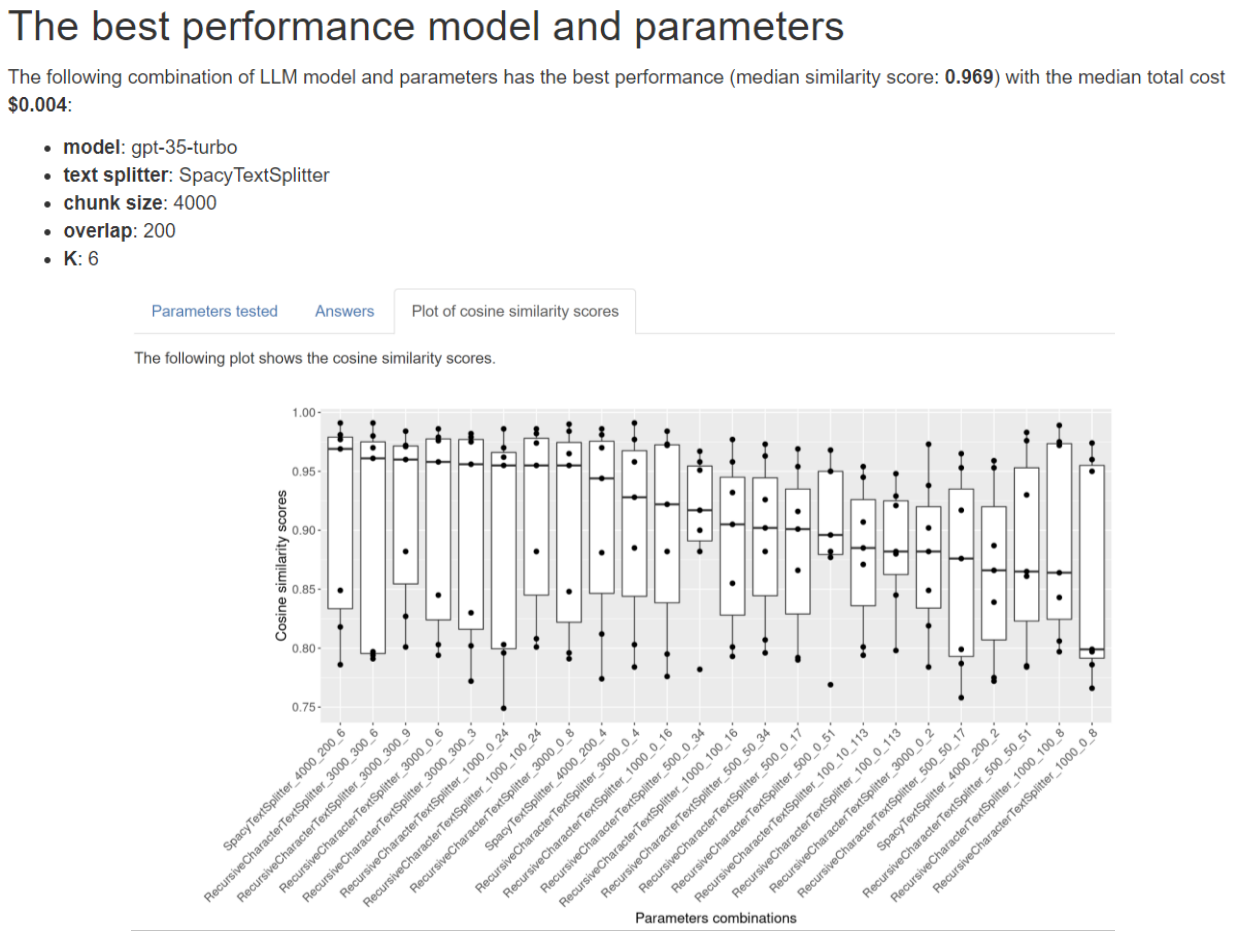

Figure 7: The Parameters and Cosine Similarity Scores of The Best Performance Model

The optimal model demonstrated exceptional performance, achieving an impressive 96.9% cosine similarity score on validation set when compared to the answers provided by SMEs. Upon independently evaluating the quality of answers, SMEs confirmed the model's responses for accuracy and readability with a very high level of confidence. Notably, this high level of accuracy was accompanied by a remarkably low median cost of $0.004 for the given validation set. This robust validation framework serves as a testament to the reliability of our proposed AI-driven workflow. Furthermore, it underscores the alignment between our model-generated responses and expert-vetted answers within ICF language forms, while also highlighting the cost-effectiveness inherent in this approach.

AI-POWERED PYTHON APPLICATION - ICF Q&A CHATBOT

Apart from a solid backend workflow, a well-designed UI is also very important, and it simplifies the interaction process, a user-friendly interface can eliminate the need for researchers to delve into complex command-line interfaces or navigate intricate systems manually. And therefore, to provide a smooth interaction experience between researchers and our AI-driven workflow, we developed an intuitive and user-friendly web-based application utilizing the Python Streamlit framework called “ICF Q&A Chatbot”. Streamlit is a powerful Python library designed for creating interactive and data-driven web applications with minimal effort. Its simplicity allows for the swift development of applications for proof-of-concept (POC), making it an ideal choice for our user interface.

Our ICF Q&A Chatbot offers a straightforward way for researchers to articulate queries and obtain accurate and prompt responses. The interface is designed to accommodate a diverse user base, regardless of their technical proficiency, thereby democratizing access to the advanced capabilities of our AI-driven system.

The below is the landing page of the chatbot, where users can upload an ICF document to trigger the Q&A workflow:

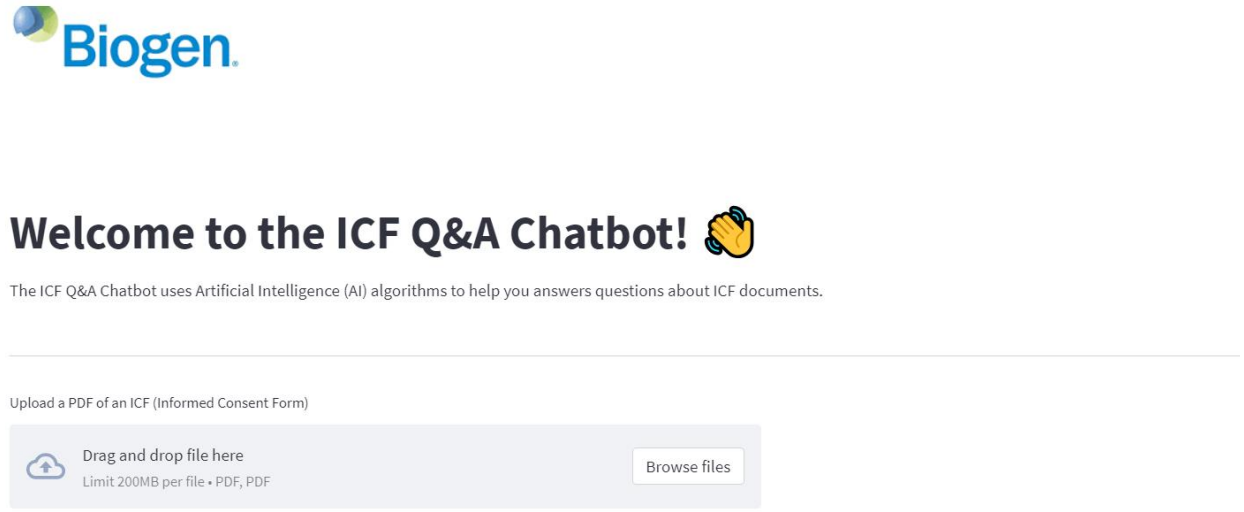

Figure 8: The Landing Page of the ICF Q&A Chatbot

Upon users uploading their ICF document, the underlying Python program initiates a process where the documents are chunked with optimized parameters, and embedding models are employed to transform them into embedding vectors. These vectors are then cached in a Chroma vector database, which functions as a customized knowledge base for subsequent question-and-answer (Q&A) sessions. Upon completion of the upload, users are seamlessly redirected to the following page, enabling them to pose inquiries regarding the ICF. Here, they can promptly receive answers generated by the underlying AI-driven workflow.

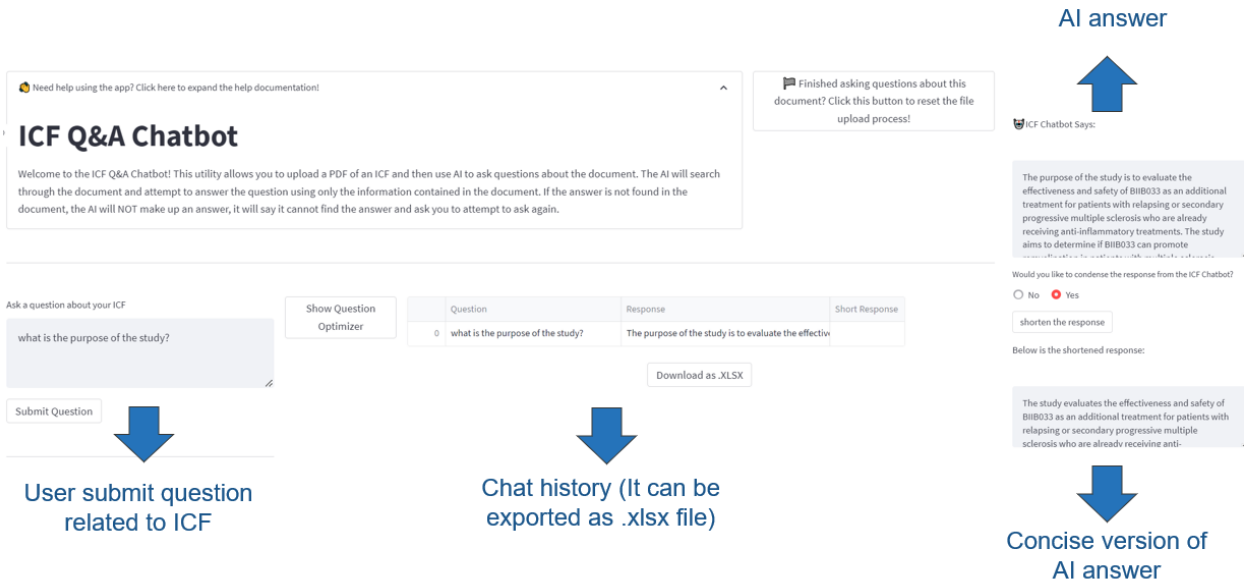

Figure 9: The Q&A Page of the ICF Q&A Chatbot

As illustrated in the above screenshot, the Q&A chatbot interface is intuitively designed, enabling users to easily pose queries and receive interactive responses. The entire chat history is systematically preserved and can be exported as an Excel file for comprehensive review. Additionally, we have introduced an extra feature that enables users to leverage the underlying LLM to condense the AI's responses. This enhancement stems from user feedback indicating a preference for more concise answers, addressing instances where responses were noted to be overly lengthy.

The chatbot solution simplifies the ICF analysis process for the data sharing team, allowing easy interaction and export of chat history. This user-friendly tool reduces the labor-intensive task of manually reviewing numerous ICF documents to extract necessary information and offering efficiency and practicality for the data sharing team to verify the allowable data re-use guidelines in ICFs.

CONCLUSION / FUTURE PLANS

In conclusion, our POC work presents a novel AI-driven workflow leveraging LLMs for the computer-assisted analysis of ICFs. Through extensive experimentation, we identified and optimized the model setting, OpenAI's GPT3.5 with text-embedding-ada-002 as the embedding model achieved a remarkable 96.9% cosine similarity score when compared to SMEs' answers. Also, the development of an intuitive ICF Q&A Chatbot further enhances accessibility, allowing researchers to engage seamlessly with the AI-driven system. Our validation framework, utilizing pre-existing ICF language forms, not only attests to the reliability and accuracy of our proposed solution but also highlights its cost-effectiveness. Overall, this cutting-edge method substantially simplifies our process by using generative AI to summarize and identify allowable data usage from ICF documents promptly and precisely.

However, it is crucial to acknowledge the inherent risks associated with LLMs, particularly the potential for generating untruthful information. Therefore, embracing a human-in-the-loop approach becomes imperative to validate outputs, mitigate errors, and ensure ethical adherence in research practices. And we also acknowledge that there is further room for improvement in the chunking strategy employed in our document processing. With the evolution of LLMs, newer models like GPT-4-Turbo feature substantially longer context windows, reaching a token limit of 128,000, and therefore we can potentially get rid of the chunking and semantic search steps and feed the entire document as the context into the LLM-base Q&A workflow, but this approach may introduce additional noises. We intend to explore these advanced models to assess their performance and determine their potential to improve the efficiency and accuracy of our ICF analysis workflow. In addition, our research intends to investigate sophisticated methodologies for augmenting the efficacy of the RAG framework, including strategies like Hypothetical Document Embeddings (HyDE), reranking, among others.

To maximize on this innovative solution, the industry can gain valuable insights from the design and implementation of our workflow, seamlessly developing and integrating a similar workflow into their existing research practices, streamlining the analysis of ICF documents and expediting the decision-making process in data sharing and data re-use areas. And in adopting our solution, the industry stands to benefit from increased efficiency, reduced manual efforts, improved accuracy in ICF analysis, and mitigated challenges in determining the re-use of samples, data, and

images for secondary research. While the risks of inaccurate information generation persist, a thoughtful integration of human oversight ensures the responsible and ethical application of AI technologies, ultimately advancing the efficiency of data re-use and expediting the dissemination of invaluable therapeutic insights to our patient community.

In next steps, we aim to implement a batch processing pipeline to analyze a substantial volume of ICFs automatically. This future enhancement tackles the practical challenges associated with large-scale document processing, mitigating the limitations of our current interactive solution. The overarching objective is to streamline and expedite the extraction of information from a large amount of historical ICFs, thereby optimizing the workflow for enhanced scalability. Besides, we are dedicated to post-processing and structuring the output in a format suitable for Subject Matter Experts (SMEs) to discern and determine appropriate data re-uses. Additionally, we plan to expand this approach to include other type of clinical documents, such as clinical study reports (CSRs), clinical protocols, investigator brochures (IBs), and more. The continuous refinement of our system aligns with our dedication to staying at the forefront of advancements in AI-driven clinical document analysis.

In alignment with our commitment to fostering innovation and collaboration across the pharmaceutical industry, we are exploring the possibility of making our Python-based framework available on GitHub. This strategic move will enable pharmaceutical companies of all sizes to utilize, contribute to, and enhance the framework. By doing so, we aim to create a vibrant community of contributors who can collectively drive advancements and share best practices of using LLM in pharmaceutical industry, ensuring that our solution evolves to meet the industry's emerging needs and challenges.

REFERENCES

1. Touvron, H. et al. (2023) Llama: Open and efficient foundation language models, arXiv.org. Available at: https://arxiv.org/abs/2302.13971 (Accessed: 03 January 2024).
2. Brown, T. B., Mann, B., Ryder, N., Subramanian, A., Dhariwal, P., Kaplan, J., ... & Mishkin, A. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
3. Wikimedia Foundation. (2024, January 3). GPT-3. Wikipedia. https://en.wikipedia.org/wiki/GPT-3#GPT-3.5
4. GPT-4 OpenAI (2023). https://openai.com/research/gpt-4
5. Lewis, M., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 37th International Conference on Machine Learning (Vol. 119, pp. 5916-5927).
6. Proser, Z. (n.d.). Retrieval augmented generation (RAG): The solution to GenAI hallucinations. Pinecone. https://www.pinecone.io/learn/retrieval-augmented-generation/
7. Schwaber-Cohen, R. (n.d.). Chunking Strategies for LLM Applications. Pinecone. https://www.pinecone.io/learn/chunking-strategies/
8. Google/Flan-T5-XL · hugging face. google/flan-t5-xl · Hugging Face. (n.d.). https://huggingface.co/google/flan-t5-xl
9. Deploy private LLMs using Databricks model serving. Databricks. (n.d.). https://www.databricks.com/blog/announcing-gpu-and-llm-optimization-support-model-serving

ACKNOWLEDGMENTS

We would like to thank our Biogen colleagues Tiffany Johnson, Anuja Das, Carla Marchini Silva, Tiffany Colacchio, Jennifer Giannetti, Buthainah Al Rifaie, and our Pharmalex colleague Matthew Ryals, Theodoros Sofianos for their collaborative work to help validate the results and insight on the application development.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Hangyu Liu
Biogen inc.,
225 Binney Street
Cambridge, MA 02142, USA
hangyu.liu@biogen.com

Andrew Borgman
Biogen inc.,
225 Binney Street
Cambridge, MA 02142, USA

andrew.borgman@biogen.com

Jake Gagnon
Biogen inc.,
225 Binney Street
Cambridge, MA 02142, USA
jake.gagnon@biogen.com

Dan Boisvert
Biogen inc.,
225 Binney Street
Cambridge, MA 02142, USA
daniel.boisvert@biogen.com

Haleh Valian
Biogen inc.,
225 Binney Street
Cambridge, MA 02142, USA
haleh.valian@biogen.com