Paper ML16

A GenAI Agent for Statistical Output Review

Lizhong Liu, CIMS Global, New Jersey, USA
Peng Zhang, CIMS Global, New Jersey, USA
Tai Xie, CIMS Global, New Jersey, USA

1. ABSTRACT

The integration of advanced AI technologies is revolutionizing the healthcare industry, particularly in how medical information is processed, understood, and queried. In this paper, we introduce an innovative system that leverages state-of-art AI technology to transform clinical reporting processes. By automating the processing and enhancement of medical documents through natural language interaction and real-time response, this system significantly improves the efficiency of decision support systems within healthcare settings.

2. INTRODUCTION

The exponential growth of clinical data has created significant challenges for healthcare professionals, particularly in the context of clinical trials and medical decision-making. Reviewing and analyzing vast volumes of documents is a time-intensive process that often delays critical decisions. For example, **Data Monitoring Committee (DMC) [1]**, responsible for monitoring patient safety and treatment efficacy during clinical trials, must evaluate extensive datasets under tight deadlines. Traditional tools and methods struggle to keep up with these demands, leading to inefficiencies and potential errors.

Advances in artificial intelligence (AI) offer a promising solution to these challenges. AI technologies, particularly in natural language processing (NLP), have demonstrated their ability to automate document analysis and enhance decision-making processes. Among these innovations, Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) stand out for their ability to understand context, retrieve relevant information, and generate accurate, human-like responses. However, leveraging these technologies in a practical, user-friendly manner remains a complex task.

In this paper, we present DARIA (Data Analysis and Review Intelligent Assistant), an AI-powered assistant designed to transform the analysis and review of clinical documents. By integrating advanced AI tools and modular design, DARIA offers healthcare professionals a dynamic, efficient, and accurate system for querying and processing clinical data through natural language interaction. This paper details DARIA's methodology, applications, and potential future developments, highlighting its transformative impact on healthcare.

2.1. Development of AI

The integration of AI into healthcare is transforming how medical professionals manage data, make decisions, and interact with technology. From diagnostics to personalized treatments, AI has demonstrated its potential to improve efficiency and accuracy. One critical area of impact is natural language processing (NLP), which focuses on enabling machines to understand, process, and generate human language. In clinical reporting, NLP has opened up new possibilities for automating tedious tasks, such as reviewing extensive medical documents, summarizing key findings, and assisting in decision-making processes.

2.2. Related Work

2.2.1. Word Embeddings: A Foundation for NLP

Word embedding is a fundamental step in natural language processing (NLP). Word embeddings are vector representations of words in a continuous vector space, where words with similar meanings are positioned closer together. This proximity is calculated by optimizing the embedding vectors to minimize distances between words that frequently co-occur in text while maximizing distances between unrelated words. Embeddings capture semantic relationships between words, reduce the dimensionality of textual data, and make it computationally feasible for machines to process language. Once words or subwords are converted into numerical representations (embeddings), these vectors serve as input tokens for Large Language Models (LLMs). Word embeddings are created by constructing a large word dictionary matrix and projecting words into this vector space. Common word embedding algorithms include **Word2Vec [2]** and **BERT [3]**.

2.2.2. Large Language Models (LLMs)

Large Language Models, such as GPT-4, are designed to process and generate human-like text. These models have been trained on vast amounts of text data, allowing them to:

- Understand complex language structures and medical terminology.
- Generate high-quality text that mimics human writing styles.
- Answer questions in natural language, making them accessible even to non-technical users.

LLMs are predictive models where natural language input is first transformed into vectors using word embedding algorithms. These vectors are then treated as tokens and fed into the LLM. Early models, such as **Recurrent Neural Networks (RNNs) [4] and Long Short-Term Memory (LSTM) [5]** networks, processed text sequentially to predict the next word or understand context. However, these models had limitations, such as difficulty handling long-range dependencies and slower processing due to their sequential nature.

Modern LLMs, based on the **Transformer [6]** architecture, overcome these limitations by using a self-attention mechanism. This mechanism assigns different weights to tokens, simulating how humans pay attention to language with both local and global memory. Unlike RNNs and LSTMs, Transformers process all tokens in a sequence simultaneously, enabling them to capture long-range dependencies and context more effectively. This innovation has significantly improved the performance and scalability of language models.

During the text generation process, LLMs predict the next token sequentially. At each step, the model calculates the probability distribution of possible next tokens, selects the most likely one (or samples from the distribution), and appends it to the sequence. This newly generated token is then used as input for the next prediction step, creating a recurrent loop during the generation process. This sequential token-by-token prediction ensures that the output text remains coherent and contextually relevant. After generating each token in the vector space, it is transformed back into the word space to produce human-readable text which enables LLMs to create fluent, natural language outputs that closely resemble human writing.

2.2.3. Retrieval-Augmented Generation (RAG):

Retrieval-Augmented Generation (RAG) [7] is a cutting-edge approach that combines information retrieval systems with generative language models to enhance the accuracy and relevance of generated outputs. RAG addresses one of the key challenges of generative models: the potential for outputs to be factually incorrect or ungrounded, particularly when handling domain-specific information such as medical or clinical data. This issue, commonly referred to as "hallucination," arises because LLMs generate tokens based on probabilistic predictions, which may be reasonable in context but not factually accurate.

RAG operates through two tightly integrated stages. First, during the **retrieval** phase, the system queries relevant information from external sources such as knowledge bases, databases, or document repositories. In healthcare, for example, knowledge bases like PubMed, clinical trial databases, or electronic health record systems are commonly used. This ensures that the retrieved information is both factual and contextually relevant to the query. Second, in the **generation** phase, the retrieved data is passed to a generative language model, such as a Transformer-based LLM. Using this context, the LLM produces coherent, accurate natural language responses. The retrieved information serves as a grounding mechanism, guiding the model to ensure that its outputs are factually aligned with the query's context.

One significant advantage of RAG is its ability to provide **factual accuracy** by grounding responses in retrieved data, reducing the likelihood of hallucinations. Additionally, RAG offers **domain-specific adaptability**, making it highly effective in precision-demanding fields like healthcare. By leveraging domain-relevant repositories, it ensures outputs are aligned with current medical knowledge. Finally, RAG systems are capable of **dynamic updates**, meaning they can incorporate the latest information from external sources, unlike static language models that rely solely on their training data, which may become outdated over time.

2.2.4. AI Agents in Healthcare:

AI agents are autonomous software entities designed to perform specific tasks by interacting with users and systems in an intelligent manner. They utilize advanced technologies like machine learning, natural language processing (NLP), and data retrieval to understand user inputs, process complex information, and provide relevant outputs in real time. These agents are capable of adapting to user needs, learning from interactions, and executing tasks with minimal human intervention.

In healthcare, AI agents act as intelligent intermediaries, leveraging technologies such as Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) to provide dynamic, context-aware support across a range of medical applications. Unlike traditional software systems, AI agents can understand and respond to conversational prompts, offering a more intuitive user experience. They can be employed to automate administrative processes, support clinical decision-making, and enhance patient care.

3. Methodology

With these technologies, we have built an AI agent called DARIA (Data Analysis and Review Intelligent Assistant). DARIA leverages advanced techniques such as Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), and vector-based databases to assist users in reading and processing extensive clinical documents. By interacting with natural human language, DARIA delivers accurate and actionable results in a short time frame without requiring any programming skills from the user.

The workflow structure of the DARIA system is presented in Figure 1, showcasing its two main components: the User Side and the Server Side. Below, we describe the methodology behind DARIA's design and implementation in detail.

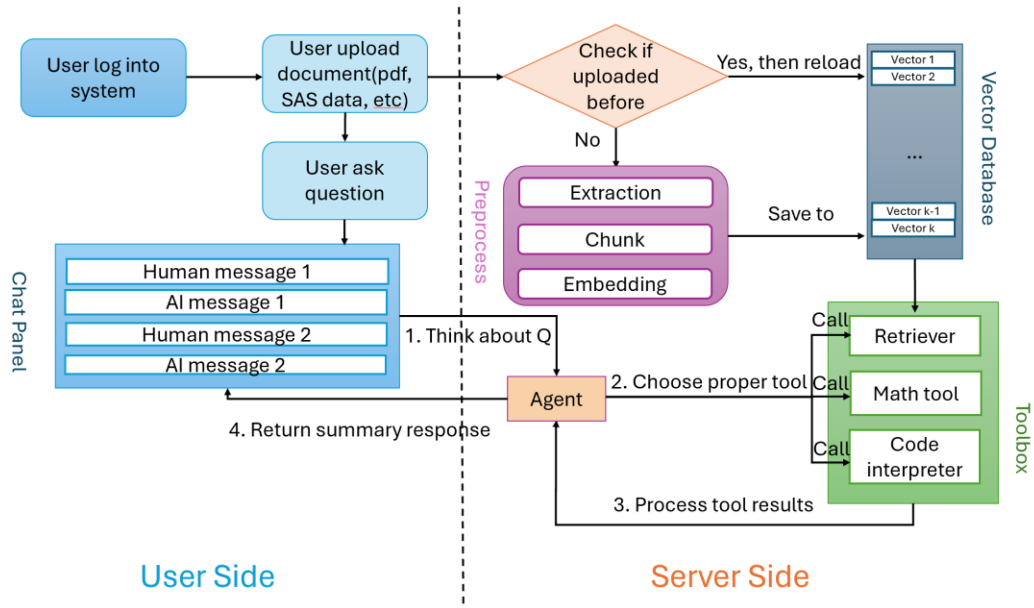

Figure 1: workflow of DARIA system

The DARIA system comprises a streamlined workflow that facilitates the interaction between the user and the underlying AI technologies to process clinical data efficiently. For example, during a Data Monitoring Committee (DMC) meeting, DARIA can assist by extracting and summarizing key safety and efficacy data from large clinical trial documents, generating visual graphs or tables, enabling committee members to focus on critical decision-making without the need to manually review extensive files or extra data processing. This practical application highlights how DARIA enhances both efficiency and accuracy in demanding healthcare environments

3.1 User Side

Users begin by logging into the system and uploading clinical documents, such as PDFs or SAS data. The system automatically checks whether the document has already been uploaded, ensuring efficiency and preventing duplication. Once the documents are in the system, users can interact via a conversational chat panel, asking questions in natural language. The intuitive chat interface displays both user queries and AI-generated responses, facilitating seamless communication and enabling users to obtain actionable insights.

3.2. Server Side

On the server side, the system processes the uploaded documents through several backend steps. For new documents, preprocessing includes text extraction, chunking, and embedding. Text extraction identifies and extracts relevant information from the document, while chunking splits the content into smaller, manageable segments for efficient processing. These segments are embedded into vector representations and stored in a vector database, allowing the system to efficiently retrieve and analyze content in response to user queries. Previously processed documents are directly retrieved from the database, bypassing preprocessing steps.

When a query is submitted, the AI agent interprets the user's question and selects the appropriate tools to address it:

- **Retriever**: Identifies key terms using the LLM's attention mechanism and searches the vector database using similarity-based algorithms, such as cosine similarity or approximate nearest neighbor methods, to fetch the most relevant information.

- **Math Tool**: Processes natural language input, converts it into mathematical expressions, and performs the required computations to deliver accurate results.

- **Code Interpreter**: Generates and executes Python code in a secure sandbox environment for tasks involving visualization or data transformation, returning charts or transformed data to the user.The outputs from these tools are integrated, and the AI agent generates a summarized response, which is sent back to the user via the chat panel. This structured workflow ensures accurate, efficient, and user-friendly interactions.

The system processes user questions in four steps: the agent interprets the question, selects the appropriate tool (e.g., retriever for document information, math tool for computations, or code interpreter for visualizations), gathers tool results, and generates a summarized response for the user. The system also includes private and public chat modes. In private chat sessions, the system temporarily stores chat history in memory to maintain context during the session, discarding it upon closure. Alternatively, users can save encrypted session histories in the database to resume interactions later. Tools are modular Python functions, enabling developers to add functionalities without altering the system structure. The system can also adapt to new tasks by simply modifying the agent's prompts, ensuring flexibility and scalability.

4. CONCLUSION

DARIA streamlines the process of analyzing and reviewing clinical documents, saving users significant time and effort. Its combination of natural language interaction, advanced retrieval mechanisms, and dynamic tool integration makes it an invaluable resource in healthcare and clinical research.

In healthcare, AI agents act as intelligent intermediaries to provide dynamic, context-aware support across a range of medical applications. Unlike traditional software systems, AI agents can understand and respond to conversational prompts, offering a more intuitive user experience. They can be employed to automate administrative processes, support clinical decision-making, and enhance patient care. For instance, an AI agent integrated into an electronic health record (EHR) system could assist physicians by summarizing patient histories, flagging potential drug interactions, and generating suggestions for treatment plans based on the latest clinical guidelines. Similarly, during clinical trials, AI agents can streamline the workflow by organizing datasets, providing real-time insights, and generating reports tailored to specific research objectives.

AI agents also exhibit continuous learning capabilities, enabling them to improve performance over time by incorporating feedback and adapting to new information. This adaptability is critical in dynamic environments like healthcare, where knowledge and best practices evolve rapidly. As a result, AI agents not only reduce the cognitive burden on healthcare professionals but also enhance the quality and efficiency of medical services.

5. Discussion

There are several areas where the DARIA system can be improved. First, while DARIA focuses on handling documents, its retriever component can only query parts of a document, which limits its ability to address broader questions, such as those requiring summarization or involving terms like "maximum," "minimum," or "largest." In the future, integrating capabilities that allow the system to interact directly with data—for example, by enabling the agent

to write and execute Python or SQL code—could provide more accurate and comprehensive answers to these types of queries.
Second, implementing a multi-agent architecture could make the system more organized and powerful. In this setup, individual agents would specialize in specific tasks, while an AI coordinator would manage task assignments, review the agents' outputs, and ensure the results are accurate and high-quality. This approach would enhance the system's scalability and reliability, ultimately improving its performance in complex and dynamic scenarios.
What's more, using the local LLM that implement in local server can ensure more privacy and confidentiality. And the smartness of the system keeps involves with LLM's performance.

6. Public Health Significance and Considerations

There is an increasing trend of involving open-source and AI technology in the pharmaceutical industry and healthcare and stakeholders have interests on how such technique can improve the workflow. During our research and based on our experience, we would like to comment the followings on how/when to adapt the AI into their considerations and plannings:

1) GenAI can help improve the workflow from various aspects. However, one might expect to see what kind of case can be the best fit for using GenAI, that is, something we must use AI to deal with rather than conventional approach or algorithm-based approach. One suggested way of thinking includes an almost infinite number of inputs or parameters that will take almost forever for algorithm-based approach to develop (e.g., if one wants to develop a unified solution but for unknown or various input types, structures, or the models include too many parameters for testing). If this is the case, people can leverage the efforts from writing algorithms, to writing test cases for GenAI applications, to improve the performance. Admittedly, both GenAI approach and algorithm-based approach will need the preparation of samples for their validations or robustness. Such test cases will be provided by ones or department's capability, based on their empirical practice.

2) There is a choice between using API and local model, while using API does not require a local/cloud to host the model. From another end, using API might need you push your proprietary data/materials to API side, which causes the data privacy concern. To utilize API properly, one might need to consider masking or similar approach to ensure the privacy issue can be addressed before sending the info out. Otherwise, a localized approach should be considered to guarantee such a concern. However, due to incoming both open-source model from deepseek [8] with similar performance, and lower price of API, the organization can have less limitations and worry less about resources. We might expect more models with better performance or less price in the future, from which case one can focus more on how to construct their applications.

3) While AI can help improve performance, it is not an algorithm-based generation or calculation. Although one can include more accurate and detailed prompts, or including different agents to improve the quality, there will be always a small chance to generate non-relavant results. When this is a concern during the procedure, we might suggest including an algorithm for the extreme case to avoid any unnecessary outcomes.

4) Validation as well as development both play an important role in the application or products. The concern arises when people start to challenge how I can trust your results. There is no unified solution across the industry, but a suggested approach of A Risk-Based Credibility Assessment Framework proposed by FDA Draft guidance [9] can be referred for the validation procedure, compared with previous complete-case testing.

7. REFERENCES

1. Fleming TR, DeMets DL, Roe MT, Wittes J, Calis KA, Vora AN, Meisel A, Bain RP, Konstam MA, Pencina MJ, Gordon DJ, Mahaffey KW, Hennekens CH, Neaton JD, Pearson GD, Andersson TL, Pfeffer MA, Ellenberg SS. Data monitoring committees: Promoting best practices to address emerging challenges. Clin Trials. 2017 Apr;14(2):115-123. doi: 10.1177/1740774516688915. Epub 2017 Feb 1. PMID: 28359194; PMCID: PMC5380168.
2. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
3. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
4. Elman J L. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179-211.
5. Hochreiter S. Long Short-term Memory[J]. Neural Computation MIT-Press, 1997.
6. Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

7. Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.
8. **https://www.deepseek.com/**
9. FDA Draft Guidance for Industry and Other Interested Parties: Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products, January 2025, https://www.fda.gov/media/184830/download

8. CONTACT INFORMATION

Author Name: Lizhong Liu
Company: CIMS Global
Address: 285 Davidson Ave, Somerset, NJ 08873
Work Phone: (844) 491-7480
Email: lliu@cims-global.com
Website: https://cims-global.com/