Enhancing Drug Safety Documentation Search Capabilities with Large Language Models: A User-Centric Approach

Jeffery L. Painter
GSK
Durham, NC, USA
jeffery.l.painter@gsk.com

Olivia Mahaux
GSK
Wavre, Belgium
olivia.x.mahaux@gsk.com

Marco Vanini
GSK
Wavre, Belgium
marco.x.vanini@gsk.com

Vijay Kara
GSK
London, UK
vijay.x.kara@gsk.com

Christie Roshan
GSK
London, UK
christie.x.roshan@gsk.com

Marcin Karwowski
GSK
Poznan, Poland
marcin.x.karwowski@gsk.com

Venkateswara Rao Chalamalasetti
GSK
Durham, NC, USA
venkateswararao.x.chalamalasetti@gsk.com

Andrew Bate
GSK
London, UK
andrew.x.bate@gsk.com

***Abstract*—Integrating Large Language Models (LLMs) to enhance complex business document retrieval represents an emerging field known as retrieval-augmented generation (RAG). In highly regulated domains like drug safety (pharmacovigilance), its application has remained largely unexplored. This technology brings numerous advantages, including expedited staff onboarding, enhanced comprehension of contextual queries, and swift information retrieval through natural language inquiries, surpassing conventional keyword searches.**

This study delves into various operational tasks, such as locating regulatory process guidance, navigating intricate scenarios for advice, and ensuring the LLM's competence in recognizing uncertainties to prevent misinformation.

LLMs empower users to engage with documentation using natural language, markedly improving search efficiency. The case study underscores LLM's effectiveness in delivering prompt guidance within pharmacovigilance and adverse event processing and reporting, offering a user-centric solution that streamlines the search for intricate business documentation.

***Index Terms*—large language models, LLM, retrieval-augmented generation, drug safety, pharmacovigilance**

Submission Type: Full Research Paper for CSCI-RTAI

I. Introduction

Large language models (LLMs) have captured significant attention due to their versatile applications, particularly within the field of pharmacovigilance (PV). Pharmacovigilance involves the systematic evaluation of medication and vaccine safety in routine healthcare delivery [1]. Despite the extensive training of LLMs on public knowledge, their accessibility remains confined to publicly available information. Consequently, they often lack awareness of data hidden behind corporate firewalls, private sources, or specific contextual intricacies.

The aim of PV is to actively monitor, evaluate, prevent, and manage adverse drug reactions resulting from medication and vaccine use. The core of this process revolves around the collection of Adverse Event (AE) reports. These reports are meticulously stored in dedicated databases designed for PV, featuring intricate data validation processes to ensure data interoperability [2] [3]. Effective staff training for data ingestion and the use of user guides are essential for accurate data entry.

This research endeavors to explore the potential of LLMs in retrieving information accurately from procedural documents, thereby supporting PV activities. The specific focus is on maintaining consistent data entry into the GSK Global Safety Database (GSD). The procedural documents pertinent to this task originate from various interrelated sources, encompassing the GSD user manuals (comprising 11 individual documents) and country-specific guidelines.

The use of generative AI (GenAI) applications presents a unique set of challenges for assessment, including issues related to uncertainty, bias, and the complexity of explaining their outputs. LLMs, due to their intricate nature and relative opacity, are susceptible to generating responses that sound convincing but may lack accuracy, a phenomenon known as hallucination [4]. Notably, there is limited existing literature on assessing GenAI applications, underscoring the need for innovative evaluation methods. This research presents a novel approach to assess the reliability of context-constrained LLMs in generating user-specific responses, highlighting the importance of consistently accurate outcomes and strategies for mitigating suboptimal responses [5].

LLM performance is influenced by several factors, including the specific model, version, training data, application, output constraints, and input nature. This assessment narrows its focus to a specific LLM and version, concentrating on the specialized task of information retrieval from user manuals. It also evaluates the LLM's responsiveness to diverse prompts tailored to both novice and experienced users.

The methods in this research adhere to a systematic and

structured approach, expanding the array of inquiries to encompass a broad spectrum of tasks that LLMs can manage. This approach ensures a quantitative evaluation of performance, enhancing the assessment's rigor and comprehensiveness.

The primary goal of this study is to evaluate an LLM as a search interface using similarity embeddings extracted from the GSK GSD user manual. The central focus lies in assessing the accuracy and usability of the retrieved information concerning the user manual's content.

II. Background

In the realm of advanced search methodologies, LLMs offer a promising solution for the complex challenge of information retrieval. This study focuses on the application of LLMs within highly regulated environments, such as pharmacovigilance (PV). Prominent models like GPT-3 and GPT-4 are renowned for their natural language understanding and generation capabilities, facilitating various applications, including text summarization and advanced chat-bot systems. However, implementing LLMs in the context of critical business documentation poses intricate challenges, including data security, regulatory compliance, and user experience optimization.

Effectively leveraging LLMs for precise information retrieval relies on meticulous prompt formulation and secure, sand-boxed environments. A sand-boxed environment is a controlled and isolated computing environment that restricts the execution of unverified or potentially malicious code, ensuring the safety and security of the system it operates within. This study explores these essential aspects, demonstrating how LLMs can enhance the accessibility of complex business documentation while ensuring data integrity and security in a dynamically regulated landscape. Opportunities highlighted by Bate and Hobbiger suggest AI systems' potential integration into PV processes to enhance patient safety [6].

AI encompasses a broad field within computer science, focusing on systems and machines capable of human-like intelligent tasks. AI includes diverse techniques, such as machine learning, natural language processing, and computer vision. These systems address complex problems, make informed decisions, recognize patterns, and learn from data. While AI usage in PV is increasing, its routine application in such environments has been more limited [7] [8] [9].

LLMs represent a specific AI category, specializing in comprehending and generating human language. They excel in natural language understanding and generation, performing tasks like text completion, translation, question answering, and text generation. In summary, AI encompasses a wide range of technologies emulating human-like intelligence, with LLMs serving as a specialized subset excelling in language-related tasks, particularly in natural language understanding and generation.

A. Additional Motivation

Standard operating procedures (SOPs) in the pharmaceutical industry are a set of written instructions that describe how to perform a specific task or process. SOPs are prepared by employees of the pharmaceutical company and are important because they help to ensure that all tasks are performed consistently, to a high standard and meet regulatory requirements, which is important for the safety and quality of pharmaceutical products [10]. Pharmacovigilance-related processes can be interdisciplinary and span multiple areas, including clinical trials, manufacturing, and regulatory compliance, increasing the number of documents a PV scientist needs to be trained on. The skills and talents required to be an effective PV scientist are quite diverse, and the day-to-day activities are both complex and highly dependent on making medically informed decisions to process a safety report [2].

SOPs are a regulatory requirement in the pharmaceutical industry and are commonly inspected by national health authorities both at the process and individual level, ensuring that processes comply with regulatory requirements and individuals are adequately qualified to perform their expected tasks.

A typical pharmaceutical company has an average of 1,200 to 1,300 SOPs [11]. An industry survey conducted by Schmidt et al. in 2013 found that an average of 11 days (range 10-20 days) per year is devoted to training [12]. Even though extensive training is undertaken, many respondents (81%) to the survey criticized their existing SOP system citing complexity and lack of clarity of individual documents/SOP systems, which made it more difficult for users to rapidly seek and find the relevant sections/instructions required for day-to-day work. In many cases, instructions concerning parts of processes are spread out among different sections of an SOP or even among a number of different documents (SOPs, instructions, appendices, etc.).

B. System Design and Implementation

LLMs are renowned for their query-answering abilities, but face challenges when providing information from specific contextual or business documents. To enable informed, contextually relevant responses, LLMs require priming or training with document-specific knowledge. Research from Brown et al. [13] and Xie et al. [14], demonstrate the adaptability of LLMs through fine-tuning [15].

Our approach grounds the LLM by creating a retrieval-augmented generation (RAG) framework using vectorization methods [16] [17]. RAG enhances LLMs with current, verifiable data, and enriches prompting with vector-based data which are mathematical representations of data [18].

Figure 1 outlines our system, involving pre-processing and user interaction in a LLM application. We use the Faiss library [19] for vector store management. Document chunks, stored as embeddings, prime LLMs for effective interaction with complex business documents via natural language queries. A generative LLM interprets results and generates responses based on these embeddings within a secure, sand-boxed environment, preserving data integrity and meeting user needs. This integrated framework connects data-driven documents to the LLM within a sand-boxed environment, preserving sensitive information while meeting our user requirements.

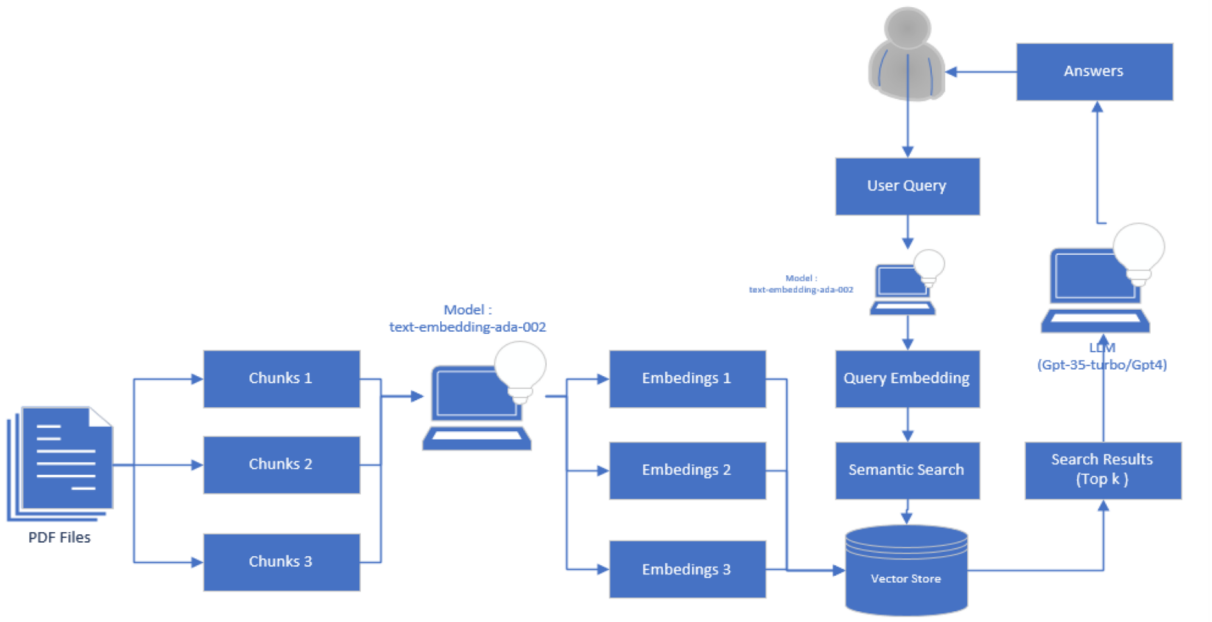

Fig. 1. Chat-Bot System Architecture

C. Importance of sand-boxed Environments

Sand-boxing is of paramount importance when integrating LLMs, creating a protective barrier between sensitive business documents and the LLM. This protection can be achieved through local GPU deployment or secure cloud services like OpenAI's Chat-GPT[1] models available via Azure[2].

These specialized LLMs do not share data with public versions or other accessible LLMs, ensuring secure corporate use. In regulated environments, such as those subject to HIPAA[3] regulations, this isolation is vital for data protection.

Sand-boxing also mitigates security threats, particularly prompt injection risks, which can arise when LLMs interact with enterprise systems. These models can bypass traditional security measures, making prompt generation within a sand-boxed context crucial for maintaining control over file access, library usage, credential management, and network connectivity. The sand-boxing approach safeguards against potential security breaches, reinforcing the separation between LLM interactions and the infrastructure, thus enhancing overall security [20] [21].

D. Understanding LLM Prompts and Streamlining User Interaction

An LLM prompt serves as a textual or command input to instruct LLMs like GPT-3 or GPT-4 for specific tasks. These prompts can vary from simple questions to detailed instructions. The clarity and quality of prompts significantly impact the relevance and accuracy of LLM responses. Crafting precise prompts is essential for meaningful results. For instance, if one intends to employ an LLM to generate a summary of a news article, a typical prompt might resemble, "Please summarize the following article: 'Title of the Article..'" In this context, the prompt serves as a directive to the LLM, instructing it to undertake a summarization task.

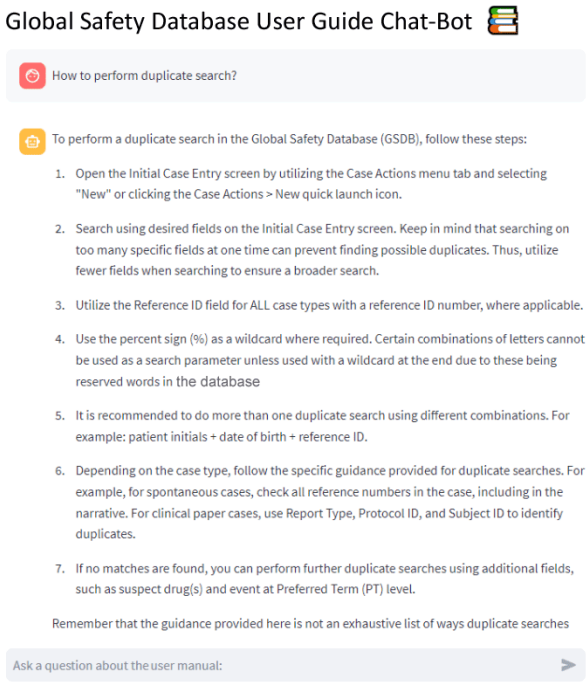

Fig. 2. Streamlit Chat-Bot Application

[1] https://openai.com/
[2] https://azure.microsoft.com/
[3] https://www.cdc.gov/phlp/publications/topic/hipaa.html

To cater to our internal users, we developed a dedicated chat-bot accessible through a Streamlit application, shown in Figure 2. This application facilitates real-time interaction, enabling users to ask questions and receive immediate Python Streamlit responses, enhancing the user experience [22].

III. Methods

This test was conducted using a sand-boxed environment of Chat-GPT 4 and the GSD user manuals as well as the following subset of country specific reference guides: France, UK, Belgium, Bulgaria, Germany, Italy, Portugal and Switzerland. The system architecture for ingesting the user manuals followed best practices of modern LLM development using Langchain, vector database and chunking of the user manuals to prime the LLM for search [23] [24]. In total, the 11 individual PDFs which compose the GSD user manuals include 566 pages, while the country specific reference guides average 3-5 pages each (total of 30 pages).

The system utilizes the user question to search for pertinent documents within the retriever. Subsequently, it passes these documents and the question to a question-answering chain to generate a response. The retriever uses a vector store, which is established by employing embeddings through the text-embedding-ada-002 model. The generator, as depicted in Figure 1, is powered by the GPT-4 LLM model. A character text splitter was employed with the following parameters (chunk size=1,000, overlap=300, separator='\n'). The retriever employed a cosine similarity search, returning the top 7 documents as context for the LLM (GPT-4) model with temperature setting of zero[4]. The parameters and prompt used were specific to the implementation and require future exploration for fine-tuning the model.

The built-in template for the prompt is as follows:

> **System:**
> Use the following pieces of context to answer the users question.
>
> If you don't know the answer, just say that you don't know, don't try to make up an answer.
>
> {context as retrieved from the retriever}
> **Human:** {question as asked to the chat-bot}

To accurately evaluate the LLM's true performance, assessment of both components is necessary. However, due to resource constraints of this experiment, the retriever will not be tested separately from the generator in this scenario, preventing us from determining if answer quality is constrained by the context window provided to the generator. In addition, for each question under evaluation, no history is preserved, and the questions are presented to the LLM without any prior history of a user interaction to determine performance metrics. Further testing may be required if the generated answer quality is found to be subpar.

[4]Temperature is a parameter passed to the LLM. Increasing the temperature leads to more varied and inventive output, whereas decreasing the temperature makes the output more predictable and concentrated. In this experiment, the temperature was set to the minimum threshold in order to mitigate potential hallucinations by the LLM.

A. Assessment of LLM Performance

To conduct the evaluation, a set of questions was designed to assess the LLM's proficiency in answering queries across a broad spectrum of complexities and tasks. The team aimed to create at least 20 questions (5 per category outlined below).

(A) **(Confirm Understanding, n=5):** Evaluating the LLM's ability to accurately identify the exact location of information within the GSD Manual and to accurately confirm the understanding of a specific process. Primarily designed to support expert users.

(B) **(Guidance and Advice, n=5):** Focusing on the LLM's capability to provide accurate and useful information when users ask open-ended questions, especially when uncertainty exists regarding actions in particular situations. This category is designed to cater to intermediate users.

(C) **(Describe and Summarize, n=5):** Assessing the LLM's capacity to respond to open-ended questions by describing or summarizing processes. This category is particularly tailored for novice users.

(D) **(Nonsensical or Out of Context, n=7):** Measuring the LLM's ability to respond appropriately, including the capability to ignore open-ended questions that are irrelevant to the user manuals. *Two additional out of context questions were devised to fully explore the LLM's ability to assert that it would not generate answers to out of scope topics.*

The questions were generated by subject matter experts based on prior experience with the GSD related to user rule interactions, with the exception of the out-of-context or nonsensical questions. Table III contains a brief list of example questions (one from each category), along with the responses generated by the LLM.

For the category of nonsensical questions, these were intentionally designed to evaluate the LLM's discernment regarding when to withhold an answer. Within this category, one question was crafted to address a topic outside the scope of the GSD user manual but within the broader domain of PV reporting. Two questions were formulated using keywords from the GSD user manual in conjunction with unrelated words. Another question was structured to be relevant to PV in a general sense but not specific to the GSD user manual, while a separate question was intentionally designed to be entirely unrelated to PV or the GSD user manual. Lastly, two questions were created to be contextually relevant within the GSD user manual but on fabricated topics.

To ensure a comprehensive assessment encompassed a wide range of complexities, the assessor formulated questions aimed at evaluating the model's proficiency in traversing multiple documents, thereby testing its "context-hopping" capabilities [25].

For each question, subject matter experts prepared a "gold standard" answer in advance for comparison with the re-

TABLE I
LIKERT ACCURACY SCALE

Value	Description
1	Completely incorrect
2	More incorrect than correct
3	Approximately equal correct and incorrect
4	More correct than incorrect
5	Nearly all correct
6	All correct

TABLE II
LIKERT COMPLETENESS SCALE

Value	Description
1	Incomplete
2	Adequate
3	Comprehensive

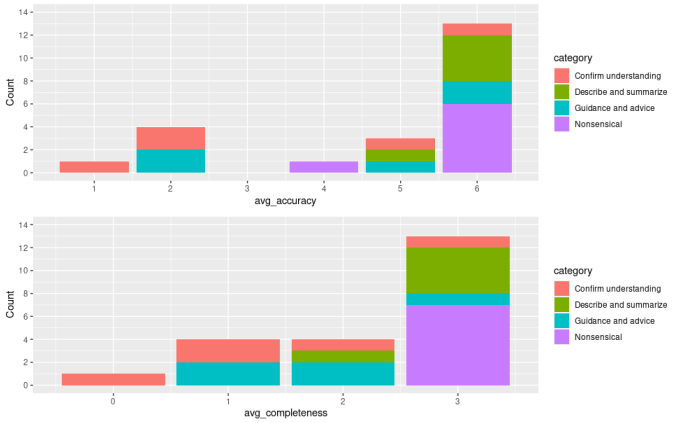

Fig. 3. Adjudicated Rank by Question Category

sponses generated by the LLM. These answers were meticulously crafted by experts who referred directly to the GSD user manual and country-specific reference guides.

The assessment of LLM responses for each question involved reviewers using a Likert scale. Likert rating scales are commonly employed to evaluate the performance of natural language processing models, including LLMs [26] [27] [28]. In this experiment, responses to questions were scored using the following methodology:

An accuracy score was determined using a six-point Likert scale, as outlined in Table I, while a completeness score was assigned using a three-point Likert scale, as defined in Table II.

These assessments explored the perceived accuracy and completeness of the LLM responses. Each question was prompted to the LLM twice, with no retention of the response history, and both generated responses were recorded for subsequent evaluation by two independent reviewers. In instances where reviewers held differing opinions, a third reviewer carried out an adjudication to reach a consensus.

IV. RESULTS

While our primary intention is to provide a comprehensive quantitative and qualitative analysis, it is worth noting that, overall, the LLM's responses, as evaluated by independent reviewers, were generally deemed acceptable. Moreover, when the LLM was uncertain or lacked an appropriate answer, it explicitly communicated this limitation, instilling confidence in the potential of this framework to assist users in navigating intricate business documents as a search interface.

The results displayed in Figure 3 showcase the adjudicated responses derived from independent reviews of the 22 questions posed to the LLM. The adjudication process involved either concurring with one of the two reviewers or, more frequently, taking an average between the two metrics assessed. In most cases, the adjudicator favored the more conservative reviewer's response. However, in instances where the LLM indicated it couldn't provide an answer, the adjudicator concurred with the independent review that this response was more appropriate than providing inaccurate or invalid information.

The graph demonstrates that, in terms of both accuracy and completeness, most questions were rated within the acceptable range by the reviewers. The mean score for average accuracy among these 22 questions was 4.750, with a median of 5.750. Similarly, the mean score for average completeness for the same set of questions was 2.227, with a median value of 2.500. Detailed summary statistics are available in Table IV.

An examination of the consistency of answers generated by the LLM when prompted twice with the same question reveals that 73% of the LLM's responses were consistent and received identical scores. In cases where they did not match (n=6), the variations were limited to one Likert scale point in both accuracy and completeness. To simplify reporting, the metrics were averaged between the completeness and accuracy of the LLM's outputs for both responses.

Further analysis delved into performance based on the categories of questions, as illustrated by the color-coding in Figure 3. Notably, when it came to describing and summarizing procedures, the LLM excelled, garnering the highest scores. A relatively balanced distribution was observed among guidance and advice questions, as well as those related to confirming understanding. It's worth mentioning that, given the context-sensitive nature of these questions, the expected responses hinged on specific country operating procedures or product-specific distinctions (e.g., vaccines versus general medicines). While some of the answers to these questions received lower ratings, the LLM consistently recommended that users seek more specific guidance from their managers when it was uncertain about the response. Impressively, the LLM demonstrated exceptional performance when confronted with nonsensical questions, providing assurance that it refrained from answering when it lacked a suitable response.

A. *Question Variation Analysis*

Of the 22 questions in our study, 34 additional variations were crafted to assess the impact of context modifications on LLM responses, particularly in the presence of formal

TABLE III
EXAMPLE OF QUESTIONS SUBMITTED TO THE CHAT-BOT

Category	Question	Chat-Bot Output
Confirm Understanding	In which part of the user guide is info on LPs?	Information on Licensing Partner (LP) Cases can be found in section 2.5 of the Global Safety Database User Manual - General Data Entry.
Guidance and advice	Can I search for duplicate cases using a batch number?	The provided information does not specifically mention searching for duplicates using a batch number. It is recommended to follow the guidance on duplicate searches as described in the document, which includes searching using reference numbers, key fields, and other relevant information. If you need to search using a batch number, it might be best to consult with your database administrator or trainer for further guidance.
Describe and summarize	Explain in simple words when to use No-PII classification?	In simple words, use the "No-PII" classification when the reporter (the person reporting the adverse event or health issue) does not wish to share their personal identifiable information (PII) or refuses to provide their contact details. This classification helps to maintain the reporter's anonymity while still allowing the case to be entered into the global safety database.
Nonsensical	Can you describe the steps for the best spaghetti Bolognese sauce?	I'm sorry, but I don't have any information on cooking or recipes in the provided context. My knowledge is limited to the information given in the context.

TABLE IV
SUMMARY STATISTICS OF ACCURACY AND COMPLETENESS

Accuracy	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	1.000	4.125	5.750	4.750	6.000	6.000
Completeness	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.000	1.625	2.500	2.227	3.000	3.000

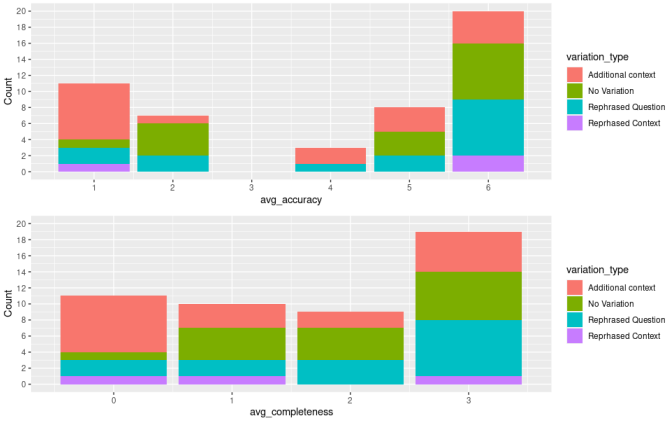

Fig. 4. Adjudicated Rank by Question Variation

prompts or specific contexts. The results, as shown in Figure 4, highlighted the significance of context changes in improving response accuracy and completeness. Variables like vaccines, drugs, and the source of adverse events, such as clinical trials vs. spontaneous reports, played a key role in question contexts.

In Figure 5, performance variations were categorized into increased accuracy or completeness, no change, or decreased performance. Notably, 64.7% of variations resulted in either no change or improvements, while 35.3% led to reduced accuracy and completeness. Generally, it was observed that succinct prompts and questions yielded better LLM responses.

Further examination of variations resulting in decreased performance revealed several valuable insights:

1) Three of the variations were primarily influenced by additional context, emphasizing the benefit of maintaining concise prompts and questions.
2) One variation highlighted the specificity of the GSD user manual concerning vaccines, which was otherwise implicit for drugs. This observation underscored potential clarity issues within the manual, contingent on the circumstances.
3) Another instance emphasized the importance of mirroring the wording found in the GSD user manual for improved results.
4) In one case, when the topic was not clearly articulated in the GSD user manual, using the original keyword search proved more effective than a complex question.
5) Two variations indicated that using key phrases such as "steps to do" and "in simple words" yielded better results compared to alternative approaches.

Interestingly, 41% of context-added variations decreased performance, while 35% improved it. Rephrased variants aligned with the manual's context, specifying circumstances for context-dependent queries, consistently improved performance by 67%.

Qualitative analysis favored concise and simple questions with single-document references. Nonsensical prompts were well-handled, with no attempt to generate a response.

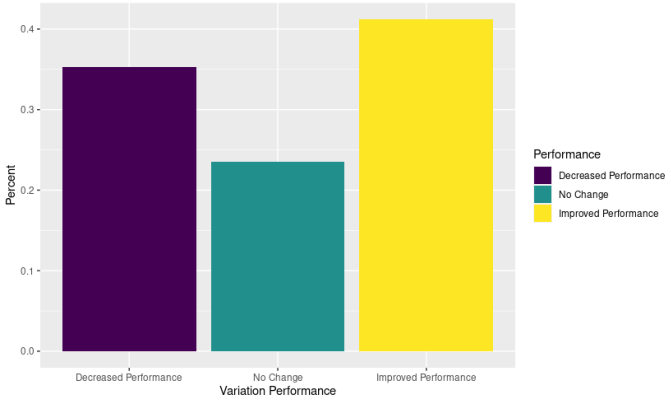

Fig. 5. Impact of Question Variation on Performance

In complex situations with information spanning documents or specific guidance needs, the output often lacked context or failed to cover varying scenarios, such as those between vaccines and drugs or AE source differences (e.g., clinical trials vs. spontaneous reports). When guidance related to specific chapters, accuracy suffered due to mixed titles and chapter numbers.

Low-scoring LLM answers (i.e., less than 3) struggled with hard-to-retrieve GSD user manual information, often due to incomplete references or data within unreadable tables or figures. Notably, no hallucination was observed despite intentionally crafted nonsensical questions.

B. Key Advantages and Limitations

The experiment underscores the substantial advantages of employing an LLM powered chat-bot, which include more efficient information retrieval, result consistency, reduced human error, and adaptability to users' specific needs. Users can swiftly access critical information within complex business documents, streamlining the process of content retrieval and saving valuable time. The system's proficiency in searching across multiple documents and sources enables users to access a comprehensive array of information, even from interconnected documents. Moreover, the system's adaptability to various prompts and contexts ensures it can cater to a broad spectrum of user needs and scenarios.

In our extensive evaluation of the LLM's ability to address inquiries from PV scientists regarding the GSD user manual, it is crucial to recognize inherent limitations as we strive to create trustworthy AI for PV [5]. Notably, the present iteration of our LLM application demonstrates some limitations in handling tabular data, which may have implications for response accuracy. Although the LLM (GPT-4) has the capacity to process images, diagrams, or screenshots as inputs, our current implementation is constrained, as it exclusively extracts information from PDF files as plain text.

Secondly, our assessment predominantly focuses on the LLM's ability to answer a specific set of questions within the context of our business-critical documents, it is not capable of seeking external fact checking or validation [29]. We do not delve into efficiency-related aspects, such as computational costs, model routing, or potential time-saving benefits for users. Furthermore, the model's capacity to address ambiguous or vague queries by seeking clarifications remains unexplored.

Lastly, it is crucial to acknowledge the constrained generalizability of this study, given its reliance on a relatively small set of questions purposefully generated for a specific LLM version. The field of document evaluation is still in its nascent stages [30], and, to the best of our knowledge, there exists a general lack of established protocols for evaluating GenAI applications without access to an objective, publicly available dataset suitable for such experiments with minor adjustments.

The study operates within the boundaries of a dataset primarily comprising the GSD user manual and select country-specific *Reporting Reference Guides*. These assessments are based on information extracted from these documents, possibly overlooking insights that subject matter experts with practical experience may possess or guidance that surpasses the provided materials. Notably, the employed LLM models lack the capability for learning or adaptation and cannot be trained on annotated datasets, a significant constraint to consider.

V. Conclusion

In this study, we've explored the intersection between LLMs and essential business documents within the highly regulated field of PV. The need to leverage LLM capabilities for efficient information retrieval from complex documents like the GSD user manual is evident. However, this endeavor poses challenges concerning security, user privacy, and data protection.

To tackle these challenges, we emphasize the significance of sand-boxed environments. These controlled virtual spaces act as protective barriers between LLMs and sensitive documents, ensuring data confidentiality. Implementing sand-boxed prompt contexts provides precise control over various aspects, effectively mitigating prompt injection and security risks.

Additionally, we've emphasized the adaptability of LLMs and their potential for in-context learning. Leveraging this adaptability, we've developed a framework to enhance information retrieval from complex documents, improving user experience and workflow efficiency.

In a landscape where companies increasingly harness LLMs, sand-boxed environments have become a requirement, ensuring both secure interactions and regulatory compliance, especially in highly regulated domains like PV. We anticipate the development of more comprehensive frameworks for test sets and guidelines for reporting AI experiments, fostering trust in the generated outputs.

In summary, integrating LLMs with business-critical documents offers numerous opportunities for enhanced information retrieval and user interaction. Ensuring these interactions occur within sand-boxed environments allows us to harness LLM capabilities while safeguarding data and user privacy. The journey to optimize LLMs for complex documents is ongoing, promising further advancements in natural language understanding and information retrieval.

VI. Acknowledgments

This research was supported in part by sponsors of the PV Systems operations teams at GSK. Michael Glaser who provided feedback and guidance, as well as support from Ray Kassekert, GSK head of PV Systems management team for Global Safety.

References

[1] World Health Organization, "What is pharmacovigilance?" https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance, 2018, [Online; accessed 04-May-2022].

[2] R. Bhangale, S. Vaity, and N. Kulkarni, "A day in the life of a pharmacovigilance case processor," *Perspectives in Clinical Research*, vol. 8, no. 4, p. 192, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5654220/

[3] ICH, "E2B(R3) Electronic transmission of individual case safety reports (ICSRs) - data elements and message specification. Step 4 version," https://ich.org/page/e2br3-individual-case-safety-report-icsr-specification-and-related-files, 2013, [Online; accessed 30-Oct-2023].

[4] S. Schwartz, A. Yaeli, and S. Shlomov, "Enhancing Trust in LLM-Based AI Automation Agents: New Considerations and Future Challenges," *arXiv preprint arXiv:2308.05391*, 2023. [Online]. Available: https://arxiv.org/pdf/2308.05391.pdf

[5] J.-U. Stegmann, R. Littlebury, M. Trengove, L. Goetz, A. Bate, and K. M. Branson, "Trustworthy AI for safe medicines," *Nature Reviews Drug Discovery*, pp. 1–2, 2023. [Online]. Available: https://www.nature.com/articles/s41573-023-00769-4

[6] A. Bate and S. F. Hobbiger, "Artificial Intelligence, Real-World Automation and the Safety of Medicines," *Drug Safety*, vol. 44, pp. 125–132, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s40264-020-01001-7

[7] J. L. Painter, R. Kassekert, and A. Bate, "An industry perspective on the use of machine learning in drug and vaccine safety," *Frontiers in Drug Safety and Regulation*, vol. 3, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdsfr.2023.1110498/full

[8] A. Bate and Y. Luo, "Artificial intelligence and machine learning for safe medicines," *Drug safety*, vol. 45, no. 5, pp. 403–405, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s40264-022-01177-0

[9] B. Kompa, J. B. Hakim, A. Palepu, K. G. Kompa, M. Smith, P. A. Bain, S. Woloszynek, J. L. Painter, A. Bate, and A. L. Beam, "Artificial intelligence based on machine learning in pharmacovigilance: a scoping review," *Drug Safety*, vol. 45, no. 5, pp. 477–491, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s40264-023-01273-9

[10] R. C. Nelson, B. Palsulich, and V. Gogolak, "Good pharmacovigilance practices: technology enabled," *Drug safety*, vol. 25, pp. 407–414, 2002. [Online]. Available: https://link.springer.com/article/10.2165/00002018-200225060-00004

[11] J. Bhattacharya, "Guidance for preparing standard operating procedures (sops)," *IOSR Journal of Pharmacy*, vol. 5, no. 1, pp. 29–36, 2015. [Online]. Available: https://www.academia.edu/11454777/Guidance_for_Preparing_Standard_Operating_Procedures_Sops_

[12] G. Schmidt, D. Baier, A. Hecht, and M. Herschel, "SOPs in clinical research," *Applied Clinical Trials*, vol. 22, no. 7/8, p. 40, 2013.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[14] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," *arXiv preprint arXiv:2111.02080*, 2021. [Online]. Available: https://arxiv.org/abs/2111.02080

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[17] O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing langchain: A primer on developing LLM apps fast," in *International Conference on Applied Engineering and Natural Sciences*, vol. 1, 2023, pp. 1050–1056. [Online]. Available: https://as-proceeding.com/index.php/icaens/article/view/1127

[18] IBM, "What is retrieval-augmented generation?" https://research.ibm.com/blog/retrieval-augmented-generation-RAG, 2023, [Online; accessed 31-Oct-2023].

[19] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8733051

[20] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," *arXiv preprint arXiv:2302.12173*, 2023. [Online]. Available: https://arxiv.org/abs/2302.12173

[21] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against LLM-integrated applications," *arXiv preprint arXiv:2306.05499*, 2023. [Online]. Available: https://arxiv.org/pdf/2306.05499.pdf

[22] M. Khorasani, M. Abdou, and J. Hernández Fernández, "Streamlit use cases," in *Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework*. Springer, 2022, pp. 309–361. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4842-8111-6_11

[23] K. Pandya and M. Holia, "Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations," *arXiv preprint arXiv:2310.05421*, 2023. [Online]. Available: https://arxiv.org/abs/2310.05421

[24] A. Pesaru, T. S. Gill, and A. R. Tangella, "AI assistant for document management Using Lang Chain and Pinecone," *International Research Journal of Modernization in Engineering Technology and Science*, 2023. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2023/42630/final/fin_irjmets1687886863.pdf

[25] C. M. Allwood and T. Kalén, "Evaluating and improving the usability of a user manual," *Behaviour & Information Technology*, vol. 16, no. 1, pp. 43–57, 1997. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/014492997120002

[26] F. Montastruc, W. Storck, C. de Canecaude, L. Victor, J. Li, C. Cesbron, Y. Zelmat, and R. Barus, "Will artificial intelligence chatbots replace clinical pharmacologists? an exploratory study in clinical practice," *European Journal of Clinical Pharmacology*, pp. 1–10, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00228-023-03547-8

[27] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir *et al.*, "Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model," *Research square*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10002821/

[28] C. Van Der Lee, A. Gatt, E. Van Miltenburg, S. Wubben, and E. Krahmer, "Best practices for the human evaluation of automatically generated text," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 355–368. [Online]. Available: https://aclanthology.org/W19-8643.pdf

[29] X. Shi, J. Liu, Y. Liu, Q. Cheng, and W. Lu, "Know Where to Go: Make LLM a Relevant, Responsible, and Trustworthy Searcher," *arXiv preprint arXiv:2310.12443*, 2023. [Online]. Available: https://arxiv.org/abs/2310.12443

[30] M. de Jong and P. J. Schellens, "Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods?" *IEEE Transactions on professional communication*, vol. 43, no. 3, pp. 242–260, 2000. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/867941