Enhancing Observability with Generative AI and Large Language Models: Centralizing APIs and Documentation for Improved Support Team Responsiveness

CS4991 Capstone Report, 2024

Ali Houssain Sareini
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ahs8du@virginia.edu

ABSTRACT

The complex IT infrastructure of Fidelity Investments was characterized by thousands of APIs and a large volume of documentation and logs. To tackle the challenges of managing and enhancing observability, we proposed and developed a solution leveraging Large Language Models (LLMs) and Generative AI (GenAI) technologies. The core of our approach involved design and implementation of an intelligent system that centralizes and processes the myriad of APIs, documentation, and logs, subsequently presenting them in a coherent, accessible manner to the support team. We utilized advanced AI and machine learning techniques, requiring a comprehensive set of programming skills, tools, and understanding of AI-based text processing and data integration methods. The preliminary results of our implementation demonstrated significant improvements in the efficiency and effectiveness of the support team's operations by providing enhanced search capabilities and actionable insights into system performance issues. Further work is needed to refine the system's accuracy in document and log interpretation, expand its API coverage, and improve its user interface. Future efforts will also focus on extensive testing and evaluation to identify and rectify any bugs or glitches, as well as explore broader applications of this solution across different departments within Fidelity Investments to maximize its organizational value.

1. INTRODUCTION

In the era of digital transformation, the management and observability of complex IT infrastructures present a formidable challenge for corporations, particularly those with a vast array of application programming interfaces (APIs) and extensive documentation. Fidelity Investments, a titan in the financial services industry, is no exception. With thousands of APIs, an overwhelming volume of documentation and logs, and an offshore team, the task of ensuring efficient and effective system support and maintenance has become increasingly arduous. An acute need exists for a novel approach to enhance observability within such complex systems.

Leveraging the cutting-edge capabilities of GenAI and LLMs, we propose a groundbreaking solution aimed at centralizing APIs and documentation, thereby significantly improving support team responsiveness. The significance of this endeavor cannot be overstated. Rapid response times and system reliability are paramount, and the ability to swiftly navigate and interpret vast data landscapes is a competitive advantage. This introduction sets the stage for a comprehensive exploration of how advanced AI technologies can be harnessed to address these challenges, transforming the way support teams interact with and manage IT infrastructure, implementing a multi-disciplinary approach.

2. RELATED WORKS

The integration of LLMs and Generative AI in automating customer service is an evolving area of research that builds upon the convergence of AI, machine learning, and natural language processing technologies. The literature in this domain reflects a broad spectrum of approaches aimed at enhancing the efficiency, accuracy, and personalization of customer service interactions through automated systems.

Følstad and Skjuve (2019) delve into the user experience and motivational aspects of using chatbots for customer service, highlighting the importance of conversational interfaces in modern customer service environments. Their work provides a foundational understanding of the way chatbots, powered by advanced AI algorithms, can transform customer interactions by offering timely and relevant responses, thereby improving user engagement and satisfaction.

In a similar vein, Kim, et al. (2023) address the challenges associated with deploying resource-intensive neural models for information retrieval tasks, crucial for effective customer service automation. Their work on EmbedDistill presents a method for distilling complex dual-encoder and cross-encoder models into more efficient formats without significantly compromising performance. This research underscores the potential of model optimization techniques in making advanced AI models more accessible and applicable in real-world customer service applications.

Furthermore, Bonifacio, et al. (2022) explore the transformative impact of large pretrained transformer models on the field of Information Retrieval (IR), which is central to automating customer service. They emphasize the role of the MS MARCO dataset in facilitating zero-shot transfer learning, thereby enhancing the model's ability to understand and respond to a wide range of customer queries across different tasks and domains.

Our review extends these discussions by focusing on the practical implementation of these technologies within the framework of customer service automation. Specifically, we examine the development and deployment of custom LLMs, such as LangChain, for creating highly responsive and context-aware customer service bots. This approach leverages the latest advancements in AI and machine learning to provide a scalable, open-source solution capable of transforming traditional customer support mechanisms, such as FAQs, into more dynamic, interactive, and personalized customer service experiences.

By synthesizing these diverse strands of research, our literature review underscores the rapidly evolving landscape of customer service automation. It highlights the critical role of LLMs and Generative AI in driving innovations that promise to redefine the paradigms of customer interactions in the digital age.

3. PROJECT DESIGN

The endeavor to enhance observability and support responsiveness at Fidelity Investments necessitated the design and development of a sophisticated system. This section explains the intricate process design, highlighting the technologies and frameworks employed to create a seamless, efficient architecture capable of managing complex infrastructures. To this end, the team suggested a balanced approach to handling the document processing. The solution is built upon a synergy of advanced technologies including LangChain, PineconeDB, OpenAI embeddings, and leverages Python and React for backend and frontend development, respectively. Hosted on AWS with specific use of AWS SageMaker, our system represents a hybrid and secure way to implement practical uses of LLMs.

3.1 Review of System Architecture

At the heart of our system's architecture is LangChain, a framework that allows for the seamless integration of Large Language Models (LLMs) into applications, enabling the intelligent processing and understanding of vast amounts of textual data. LangChain serves as the cornerstone for building our application, facilitating the creation of custom logic and workflows that leverage LLMs to interpret, summarize, and generate insights from the centralized data repositories. This approach ensures that the support team can access not just raw data, but

also contextualized information and recommendations generated by AI, thereby improving the speed and quality of decision-making processes.

PineconeDB plays a critical role in our architecture as the vector database responsible for storing the high-dimensional vector representations of text data. These representations are created using OpenAI embeddings, which convert text into vectors that encapsulate semantic meaning, allowing for highly efficient and accurate search capabilities across the system.

The development of our system was orchestrated using Python, a choice driven by its robust ecosystem for data science and machine learning, as well as its effectiveness in handling backend logic and AI model integration. For the frontend, we chose React due to its component-based architecture, which allows for the creation of a dynamic and user-friendly interface. React's ability to manage stateful interactions and update the UI in real-time has been instrumental in providing support team members with an intuitive and responsive platform to access the centralized information.

Hosting our solution on AWS offers the scalability, reliability, and security required for a system of this magnitude. AWS SageMaker, in particular, has been utilized for training and deploying our machine learning models, including those for generating embeddings and processing textual data with LLMs. SageMaker provides a managed environment that simplifies the machine learning workflow, from model development to deployment, enabling our team to focus on innovation and enhancing the system's capabilities.

The integration of these technologies into a cohesive system architecture has been a complex process, requiring careful consideration of security, scalability, and performance. By leveraging LangChain for LLM integration, PineconeDB for efficient data storage and retrieval, OpenAI embeddings for text representation, and AWS for hosting and model management, we have constructed a powerful solution capable of transforming how support teams interact with and manage IT infrastructure.

3.2 System Process

The process of embedding documents using an LLM and then retrieving them with semantic search involves a blend of natural language processing, machine learning, and information retrieval techniques. This process is central to our system's capability to efficiently manage and access the extensive documentation within Fidelity Investments technical ecosystem. The journey begins with the preprocessing of documents, which includes converting them from their original formats (such as PDF or TXT files) into a standardized, machine-readable format. This step often involves cleaning the data by removing any irrelevant information, such as headers, footers, or any non-textual content, to ensure that the focus remains on the meaningful content of the documents. Once the documents are preprocessed, they are fed into an LLM, OpenAI in this case, to generate embeddings. An embedding is a low-dimensional, dense vector representation of the document's content that captures its semantic meaning. This is achieved by the LLM analyzing the context and the intricacies of the language within the document, transforming it into a vector that represents the document's essence in a high-dimensional space.

The LLM models, such as those provided by OpenAI, are capable of understanding and capturing the nuances of human language, making them exceptionally suited for this task. The generated embeddings are then stored in a vector database, such as PineconeDB. Vector databases are designed to handle high-dimensional data efficiently, making them ideal for storing and managing the embeddings. They maintain the semantic relationships between documents, allowing for highly efficient and relevant retrieval based on the content's meaning rather than just keyword matching. When a query is received, the same LLM is used to convert the query text into an embedding, following a similar process as with the documents. This query embedding is then compared against the document embeddings stored in the vector database, using vector similarity measures. The database retrieves the documents whose embeddings are most similar

to the query embedding, thus ensuring that the search results are semantically related to the query. This process allows for a form of search that understands the intent and the meaning behind the query, rather than relying solely on the presence of specific keywords, effectively turning a knowledge-base into a search engine. The documents retrieved through semantic search are then presented to the user. Given the semantic nature of the search, the results are highly relevant to the user's query, often providing insights and information that traditional keyword search methods might miss.

This entire process, from embedding documents with an LLM to retrieving them with semantic search, enables a highly efficient, accurate, and user-friendly way to manage and access vast amounts of documentation. It represents a significant advancement over traditional search techniques, offering a more nuanced and intelligent approach to information retrieval that can dramatically enhance support team responsiveness and overall operational efficiency.

4. RESULTS

The deployment of the advanced platform within Fidelity Investments has notably enhanced the efficiency and responsiveness of support operations, a crucial improvement that has profoundly impacted offshore teams. Offshore teams, often grappling with the challenges posed by geographical dispersion, time zone differences, and varied access to centralized information, have historically encountered obstacles in accessing content swiftly and effectively. The introduction of our system, has significantly mitigated these challenges, fostering a more cohesive and efficient operational framework based on internal assessments. This section elaborates on the efficacy of the platform in transforming the operational dynamics for offshore teams, emphasizing its contribution to rapid content access and the overarching implications for global support strategies. While the recorded performance of this system is unknown in this context, The streamlined access to information and improved collaboration have led to a perceivable increase in operational efficiency. This efficiency translates into faster resolution times for support queries and a reduction in operational costs, as the reliance on manual, time-consuming processes decreases.

The platform not only streamlines access to critical information and facilitates superior collaboration across global teams, but it also significantly bolsters the relevance of document retrievals and the efficacy of semantic search functionalities. To achieve this, metadata has been meticulously integrated into documents. This enrichment process has been crucial in fine-tuning the semantic search capabilities, ensuring that the most pertinent documents are surfaced with precision, directly contributing to an accelerated and more accurate query resolution.

Moreover, the architecture incorporates a cloud-hosted database designed to augment API endpoints with essential internal details about their corresponding sub-systems. This aspect of the system is particularly instrumental in the management of API ownership. Each API is assigned an owner, typically deeply knowledgeable about the system's nuances, who also serves as an emergency contact in the event of a critical endpoint failure. This delineation of responsibility ensures that, should a high-priority issue arise, the system can proactively notify the relevant API owners, thereby mobilizing a swift response to potential system failures.

This proactive notification mechanism represents a pivotal enhancement in the realm of site reliability, offering a robust framework for addressing and mitigating potential disruptions in large-scale IT infrastructures. By integrating these advanced functionalities, the platform not only elevates the efficiency of global team collaboration but also significantly contributes to the cultivation of a culture of continuous learning and improvement. The inclusion of enriched data and the strategic deployment of a cloud-hosted database for enhanced API management exemplify a forward-thinking approach to maintaining system integrity and reliability, positioning the firm—and any future adopters of this system—at the forefront of innovative solutions for managing complex infrastructure challenges.

5. CONCLUSION

The integration of Large Language Models and Generative AI into Fidelity Investments' IT infrastructure management has proven to be a transformative solution, addressing the challenges posed by the extensive volume of APIs, documentation, and logs. By centralizing these resources and leveraging advanced AI techniques, the proposed system has significantly enhanced the efficiency and effectiveness of support operations. The ability to rapidly access relevant information and gain actionable insights has empowered the support team, particularly the offshore members, to deliver swift and accurate resolutions to queries.

This project not only showcases the immense potential of LLM applications in optimizing complex IT infrastructures but also highlights the importance of a multidisciplinary approach in developing innovative solutions. The knowledge gained from this endeavor extends beyond the industry of financial institutions, offering valuable insights for organizations across various industries grappling with similar challenges. As we continue to refine and expand the capabilities of this system, we anticipate even greater value for Fidelity Investments and its customers, solidifying its position as a pioneer in the application of cutting-edge technologies.

6. FUTURE WORK

Efforts should be directed towards refining the accuracy of the system in interpreting documents and logs. This will involve exploration of more advanced NLP techniques and the fine-tuning of the LLMs to better understand the domain-specific language and context. Additionally, expanding the coverage of APIs will be a key focus, ensuring that the system can provide comprehensive support across the entire IT infrastructure. As stated by Schwartz (2023), to have a higher adoption rate and belief in the automations granted by generative AI, we must increase trust in our systems. The technical nature of IT infrastructure is a perfect place to start building this trust. Improving the user interface is another critical aspect, aiming to enhance the user experience and facilitate seamless interaction with the system in an intuitive manner. Furthermore, exploration in the potential applications of this solution across different departments, such as customer service, risk management, and compliance will be conducted.

Looking beyond Fidelity Investments, the technology envisions the possibility of generalizing this solution to benefit other organizations facing similar challenges in managing complex IT infrastructures. This would involve developing a more flexible and customizable framework that can be easily integrated into various systems and adapted to different industry requirements. By sharing our findings and collaborating with the broader AI and IT communities, we aim to contribute to the advancement of AI-driven solutions for infrastructure management and inspire further innovation in this field.

REFERENCES

Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19). Association for Computing Machinery, New York, NY, USA, Article 1, 1–9. https://doi.org/10.1145/3342775.3342784

Bonifacio, L., Abonizio, H., Fadaee, M., & Nogueira, R. (2022). InPars: Unsupervised Dataset Generation for Information Retrieval. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval

Kim, S., Rawat, A. S., Zaheer, M., Jayasumana, S., Sadhanala, V., Jitkrittum, W., Menon, A. K., Fergus, R., & Kumar, S. (2023). EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval. ArXiv

Schwartz, S., Yaeli, A., & Shlomov, S. (2023). Enhancing trust in LLM-based AI automation agents: New considerations and future challenges. arXiv preprint arXiv:2308.05391.