Paper OS01

Integrating GenAI with Open-Source Toolkit for Clinical Insights Generation to Boost Data Science Efficiency

Vincent Shen, Roche, Mississauga, Canada

ABSTRACT

Roche data scientists began transitioning to a new computing environment (OCEAN) and adopting NEST tools (https://github.com/insightsengineering) as the default for reporting events in 2024. To alleviate the change management burden and enhance productivity and efficiency, we have started integrating the latest GenAI technologies into solutions that can be embedded into data scientists' daily workflows. This paper will detail our journey in developing our in-house GenAI solution for data scientists, the impact it has brought so far, and our vision for how GenAI can transform the way our data scientists work within the industry.

INTRODUCTION

About a year ago, Roche data scientists started a journey to transition into a new, internally-developed computing environment (OCEAN). During the journey, people are expected to learn new programming languages, as well as new processes and tools, including various R packages for SDTM, ADaM, TLG creation, and git for workflow. Learning so many things at the same time requires significant efforts and resources for change management and user support, and even with good documentation on new tools as well as training materials, the steep learning curve still exists and presents a significant business challenge.

To address such challenges, we chose to leverage the latest GenAI and LLM technologies. GenAI was (a year ago), and still is today, a fast-growing field and has generated a great level of interest globally. GenAI/LLM has demonstrated the ability to effectively extract information from domain-specific contexts and generate answers. Additionally, it offers a relatively low cost in terms of time and effort required to build a domain-specific application. Furthermore, GenAI/LLM can communicate in a human-like manner. All these capabilities have made GenAI/LLM a natural choice to solve our business challenges.

GENAI PRODUCTS & BUSINESS IMPACT

We built two GenAI products to address the aforementioned business challenges. In this section, we will briefly cover how these products were designed and the business impacts they have brought so far.

OCEAN ASSISTANT

The OCEAN Assistant's design architecture centers around a RAG (Retrieval Augmented Generation) based application powered by GPT-4o. This core is supported by an extensive knowledge base of over 20 sources of documentation, covering a wide array of systems, tools, and processes. To ensure that the information provided is always up-to-date, a semi-automated data pipeline continuously fetches the latest information from these diverse sources. Importantly, the OCEAN Assistant provides traceability of its answers and includes a self-evaluated confidence score, allowing users to assess the reliability of the information provided.

Ocean Assistant has reduced the workload on SMEs by serving as an automatic first-line user support. It has also lessened the learning curve for OCEAN and related tools, and helped users troubleshoot complex processes and production runs. Additionally, it has made the 20+ sources of documentation more digestible for users. We collected some metrics after launching OCEAN Assistant, and it has by far greatly exceeded our expectations. We have over 1000 users with 44000 questions asked and answered already, specifically, 4300 production runs on OCEAN have been debugged by OCEAN Assistant. Putting things into perspective, we believe it takes an average of at least 15 minutes to get any question answered by one data scientist alone, including time spent on documentation search and also reaching out to SMEs. Given that OCEAN Assistant gives the answer instantly with a great level of accuracy and details, we believe that we could save about 25 hours per 100 questions asked on OCEAN Assistant.

NEXPERT

NExpert's architecture is built on a foundation of GPT-4o and a fine-tuned open-source model, bolstered by comprehensive package documentation and curated code examples that serve as its knowledge base. The system incorporates a code review process and a semi-automated pipeline to ensure the quality of the retrieved information and optimize the efficiency of the retrieval process.

Since its launch, NExpert has lessened the learning curve for many data scientists on packages like {admiral}, {rtables}, {tern}, etc. and it specifically served as a first-line support to help users to address their ADaM & TLG's customization needs. In addition, NExpert is able to provide human-readable explanations to the recommended code and also help users to

troubleshoot common errors when using selected R packages. Some of these features rely on the underlying LLM, which showed great promise of how these models may help in the clinical insights generation process.

REFLECTIONS ON OUR JOURNEY SO FAR

As mentioned earlier, the field of GenAI/LLM evolves quickly and almost presents new developments and advances on a daily basis. Since launching the two products, we have made some key observations, which led to reflections on our product strategy as well as the future development directions for these products. We'd like to discuss three main areas of our reflections in this section.

REFLECTION ON KNOWLEDGE RETRIEVAL TASKS

The recent success of the OCEAN Assistant has underscored the growing need to integrate additional data sources into our system. However, relying on a single aggregated database for retrieval using RAG has proven inefficient, particularly as the number of data sources increases. This expansion has not only complicated the data retrieval pipeline but also made its maintenance more challenging. Despite these obstacles, our evaluation shows that developing a knowledge bot based on a single data source is relatively low in cost and effort, and there is a clear user preference for a unified interface that consolidates all inquiries, rather than having to navigate multiple domain-specific platforms.

Moving forward, it is recommended that we adopt a modular and scalable architecture to seamlessly integrate multiple data sources without overburdening the retrieval pipeline. Transitioning to a federated retrieval system could alleviate the inefficiencies associated with a single database approach and simplify maintenance. Additionally, focusing on the development of a centralized user interface will enhance user experience by providing a one-stop solution for all queries. The next steps should include prototyping the modular architecture, testing the new retrieval approach with diverse data inputs, and gathering comprehensive user feedback to refine the system further.

REFLECTION ON GENAI CODING ASSISTANCE

From a user perspective, the most effective approach for providing coding assistance is directly within an Integrated Development Environment (IDE), ensuring developers receive real-time support as they code. While public coding assistance tools that rely solely on large language models (LLMs) excel in this regard, it is important to note that Pharmaverse packages are relatively new and considered niche. Additionally, the R language has a smaller footprint compared to more popular languages like Python and JavaScript, which influences the adoption and community support available. Our evaluation suggests that while LLM-based tools offer broad assistance within the IDE, NExpert—leveraging both LLMs and curated domain knowledge—has an advantage in handling niche aspects associated with Pharmaverse packages and R-specific insights.

Looking ahead, our recommendation is to leverage existing coding assistance solutions such as GitHub Copilot and Cursor—known for their robust UI and seamless integration within popular IDEs—while embedding our own model enriched with domain-specific and Pharmaverse knowledge. This hybrid approach would combine the user-friendly features and widespread adoption of these platforms with our specialized insights, offering developers both a familiar coding environment and enhanced, targeted support for R and Pharmaverse-related queries.

We have recently built a prototype on this idea, albeit in its early stages and not yet subjected to user testing, our next steps will focus on iterative refinement and comprehensive evaluation. We plan to initiate controlled user testing in select environments to collect valuable feedback on usability, performance, and the effectiveness of our domain-specific enhancements. These insights will guide further iterations, ensuring that our integrated solution not only maintains the high-quality user experience of established tools but also delivers the specialized support needed by our internal data scientists.

REFLECTION ON E2E WORKFLOW

Our current implementations of OCEAN Assistant and NExpert are designed primarily to tackle two business challenges—efficient knowledge retrieval and coding assistance—while the broader insights generation workflow is considerably more complex, encompassing several additional components beyond coding and how-to queries. Moreover, we have yet to capitalize on some of the key strengths of large language models, such as advanced document comparison and automated writing, which are critical in numerous workflow steps. Evaluations suggest that integrating more GenAI use cases could bring significant efficiency gains, especially in areas where routine yet tedious tasks currently dominate and could benefit from a well-established automation pipeline.

Moving forward, we recommend expanding our toolset to incorporate a wider array of GenAI functionalities that address the full spectrum of our workflow. This should include embedding capabilities for document comparison, automated writing, and other essential tasks to streamline the entire process—from data ingestion to insights generation. Given that many routine tasks could be automated effectively, the next steps involve mapping out these workflow components, developing pilot projects for the new use cases, and initiating user testing to validate their impact on efficiency. This strategic expansion will ensure that our solution evolves to meet the complete set of challenges within our insights generation process.

VISION ON GENAI FOR INSIGHTS GENERATION

To fully harness the power of GenAI/LLM for clinical insights generation, we must make strategic investments in two key areas: Data & Domain Knowledge and an Agentic Approach. First, a robust foundation in data and domain knowledge is critical. This involves developing a well-designed database that encompasses all relevant data sources and metadata essential for insights generation. Equally important is maintaining an up-to-date and accurate domain knowledge base. Without precise, comprehensive mapping of the information landscape, even the most advanced LLMs can misguide users—much like a high-performance GPS that leads astray when its maps are outdated or incomplete.

In parallel, an agentic approach offers a promising pathway to streamline the inherently complex insights generation workflow. Rather than treating the process as a monolithic task, this approach envisions an AI Agent that intelligently decomposes complex challenges into manageable sub-tasks. Each sub-task can then be addressed by specialized AI bots, whether it's executing intricate coding assignments or triggering a straightforward automation pipeline. This modular structure not only enhances operational efficiency but also ensures that each component of the workflow receives the focused attention it requires.

Looking to the future, our long-term vision is to establish a human-driven, AI Agent-enabled workflow that accelerates and smooths out the clinical insights generation process. In this envisioned system, human experts will guide the AI Agent, making strategic decisions while individual AI solutions handle routine or specialized tasks. This human-in-the-loop approach is essential—not only for refining and validating the results but also for ensuring accountability in clinical reporting. By investing in these areas now, we lay the groundwork for a more efficient, reliable, and integrated clinical insights generation process, positioning ourselves at the forefront of GenAI-driven innovation in the clinical reporting domain.

ACKNOWLEDGMENTS

It has been a great team work to develop and launch the GenAI solutions internally at Roche. Here is a list of people who have made great contributions so far to the project and their work is greatly appreciated:

Adam Forys
Anzhou Zhang
Bo Ci
Dony Urnadi
Jian Dai
Mathieu Cayssol
Zoila Rossi
Zuoyu Wang

CONTACT INFORMATION

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

Author Name: Vincent Shen
Company: Roche
Email: vincent.shen.vs1@roche.com