innovate
GenAI Use Case

enago
Life Sciences

Leveraging AI for Querying Research to Find Answers

THE CONTEXT

The importance of artificial intelligence (AI) in research and knowledge discovery cannot be overstated. It has transformed the speed and efficiency with which information is processed, analyzed, and synthesized, leading to faster and more accurate results. AI-driven insights are enabling breakthroughs that were once considered unfeasible in disciplines ranging from healthcare to environmental science.

In this rapidly evolving landscape, a variety of tools and technologies have emerged, revolutionizing the way we query and analyze research data. These tools range from advanced natural language processing (NLP) algorithms to sophisticated data mining and analysis software, each designed to facilitate different aspects of research.

Large language models (LLMs), which are the foundation of text-based Generative AI (Gen AI) models (e.g., ChatGPT and Bard), represent a ground-breaking shift in the field of artificial intelligence. They have the ability to create new content that is almost indistinguishable from human-generated content.

An LLM such as Open AI's GPT-4[1] and Google's Bard[2] is a specific application of Gen AI. LLMs offer significant benefits to researchers, institutions, and biotech, biomedical, and pharma companies. Some common uses in the field of medicine and healthcare are listed below:

- *Accelerating literature review and meta-analyses*: Gen AI tools can swiftly analyze and compile relevant literature from vast databases, significantly speeding up the process of literature review and meta-analyses.
- *Answering questions about medical research:* Using Gen AI and natural-language queries, users can query the content of research articles and quickly find contextually relevant answers.
- *Extracting insights*: Gen AI tools can process vast amounts of complex medical literature, translate dense medical jargon and statistics, and deliver comprehensible insights. This helps researchers and practitioners save valuable time to proceed to actionable research tasks.
- *Summarizing medical literature*: Gen AI can distill lengthy medical documents into concise summaries, allowing researchers and healthcare professionals to extract essential information faster or decide what needs deeper research/reading.

However, challenges remain in terms of data privacy, security, and transparency/explainability of AI models. These challenges become even more pronounced in specialized contexts or with information from particular domains because AI models are unable to fully factor in the intricacies and distinct nuances in these situations. Genuine-sounding but factually inaccurate responses generated by Gen AI models are referred to as *hallucinations*, and they are especially critical in bio-sciences and healthcare because of the potential consequences in decision-making.[3,4]

Remotely hosted LLMs like ChatGPT, Bing, and Bard present a security concern because of lack of transparency of data-security protocols and potential infringements on intellectual property. In contrast, locally deployed LLM technology are challenging in terms of high infrastructure and deployment costs and seamless integration into live workflows while ensuring a small computational/carbon footprint.

Thus, a system that helps researchers quickly understand medical literature, ensures accurate, reliable generation of answers and summaries, provides security for sensitive data, and allows for smooth deployment would be the optimal combination. To this end, we developed a generative-AI-based question and answering (Q&A) application, *Trinka AI Co-Pilot*, to help medical professionals gain a deeper understanding of research papers while spending less time.

OUR SOLUTION: *TRINKA AI CO-PILOT*

Trinka AI Co-Pilot is a versatile tool designed to make scientific literature more accessible and comprehensible for researchers, academic institutions, corporate entities, and stakeholders. Users can gain an in-depth understanding on and insights from a research paper within minutes. This feature fits seamlessly into any workflow, allows users to quickly grasp technical details and generate insights from research papers, and saves valuable human time and increases research productivity.

To ensure accuracy and relevance of results, we combined the functions of our Q&A application with an LLM from Open AI's GPT family. Our proprietary Q&A application was developed in-house and uses an enormous dataset of research content, including 122 million abstracts of research papers across 124,000 journals and conferences. Coupled with the Open AI GPT model, *Trinka AI Co-Pilot* provides accurate, focused, and relevant answers to queries on manuscripts.

Trinka AI Co-Pilot uses the generative method for answers and summaries and has the following features built-in.

> *Querying the paper*: The user can quickly get specific information about a particular section or the concepts from the paper. The context of the questions and their responses are retained in case the user has follow-up queries based on the previous responses.
>
> *Simplified responses*: *Trinka AI Co-Pilot* provides simplified explanations regarding unclear or difficult concepts.

Literature Summarization: The tool will generate a summary to help the user decide whether deeper, human reading is required.

Our technology team has relied on the following techniques to ensure that Trinka AI Copilot provides accurate (*hallucination*-free) responses and can be deployed locally to ensure data security.

AI guardrails: AI guardrails mitigate the risks associated with AI, such as bias, discrimination, security breaches, and potentially harmful LLM behavior. *Trinka AI Co-Pilot*'s guardrails include measures to protect confidential information and, if requested by a client or user, to safeguard private, sensitive data through masking or redaction before it is sent to the LLM.

Retrieval-augmented generation:[5] LLMs are primarily trained on non-specialized data; hence, it is essential to tap into the knowledge in specialized data to provide accurate responses and detailed contextual insights. *Trinka AI Co-Pilot* digests structured/unstructured data and uses vector databases with custom domain data, thus minimizing the risk of generating inaccurate information or falsehoods.

LLM Fine-tuning: With the variety of LLMs available, it is important to understand which LLM provides the best outcomes. Fine-tuning various open-source LLMs using domain-specific data helps uncover their full potential. We employed parameter-efficient fine-tuning[6] techniques like LoRA, QLoRA, and P-tuning to optimize the processes, reduce costs, and enable swift deployment. This exploration of different LLMs led us to select an Open AI GPT model for Trinka AI Co-Pilot.

LLM deployment: Model deployment requires LLM optimization through techniques such as quantization, pruning, and distillation. Using these techniques, we made the Trinka AI Co-Pilot tool as compact as possible so that it can be deployed on smaller infrastructure without compromising on accuracy and speed.

With the incorporation of all of the abovementioned methods, *Trinka AI Co-Pilot* represents a careful combination of an in-house Q&A application, an LLM, and well-designed services to make sure the users benefit from time reduction, costs savings, deep understanding, and fast decision-making with the responsible use of gen AI and LLMs.

CONCLUSION

In summary, we created Trinka AI Co-Pilot, a generative-AI-powered Q&A tool, designed to assist researchers in quickly grasping complex concepts in medical literature. Moreover, by utilizing our knowledge and practical experience in LLM technology, we have crafted solutions aimed at enabling biotech and pharmaceutical firms, along with academic institutions, to ethically leverage the power of gen AI and LLMs, thereby speeding up scientific discoveries and addressing the challenge of inaccuracy and security in LLM outputs.

REFERENCES

1. OpenAI. GPT-4 Technical Report. arXiv (Cornell University). Published online March 15, 2023. doi:https://doi.org/10.48550/arxiv.2303.08774

2. *Google Bard*. https://bard.google.com. (n.d.). https://bard.google.com/chat

3. Naddaf, M. (2023). *CHATGPT generates fake data set to support scientific hypothesis*. Nature News. https://www.nature.com/articles/d41586-023-03635-w

4. Pal, A., Umapathi, L.K. and Sankarasubbu, M. (2023) *Med-HALT: Medical Domain Hallucination Test for Large Language Models*. Available at: https://arxiv.org/pdf/2307.15343.pdf (Accessed: 17 November 2023).

5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D., Facebook, & Research, A. (n.d.). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. https://arxiv.org/pdf/2005.11401v4.pdf

6. Mangrulkar, S., & Paul, S. (2023, February 10). *Parameter-Efficient Fine-Tuning using 🤗 PEFT*. Hugging Face. Retrieved November 4, 2023, from https://huggingface.co/blog/peft

Visit Enago Life Sciences at https://lifesciences.enago.com/ or contact services@ls.enago.com to learn more.