QuantumBasel

Improving Cancer Treatment Options

2024 | Use Case | Fine-tunded LLMs for Precise Cancer Genomics Diagnostics

CHALLENGE

Cancer genomics is a critical element of precision oncology, which uses broad genomic profiling of tumor tissue or blood to **identify genomic biomarkers** of sensitivity or resistance to various cancer therapies. **The lab reports from these tests are routinely 20-50 pages long and impractical for integration into the workflow of a practicing oncologist**. Even for oncologists who are experts in the field, these machine-generated **reports can be too complicated, resulting in incorrect treatment recommendations**.

Qnomx

QuantumBasel

SOLUTION

To improve this process, the goal is to **develop an AI workflow that integrates retrieval (searching for relevant information) with generation (creating text based on retrieved information) to improve the quality of output and produce more informed and accurate results**.

For this case, several state-of-the-art open source components and a customized large language model (LLM) with local data storage and processing were used to improve the accuracy, explainability, and transparency of the data flow. An open source **LLM was fine-tuned through supervised and reinforcement learning based on input from a genomics expert**. Qnomx experts were able to provide the AI with the correct answers to typical genomic questions, which served as ground truth. The model was then tested in multiple iterations on new datasets to evaluate its ability to perform the task. **Through a Retrieval-Augmented-Generation (RAG), the user is able to upload a document (the cancer report) and chat with the document.**

Mistral Base Model

100% GPU power

Finetuning Process

Mistral Finetuned Model

-24% GPU power

RESULTS

The new solution dramatically **streamlines analysis, significantly enhances hardware efficiency, and ensures trustworthy outputs, all while providing users with seamless and visually intuitive access to relevant information**.

The key improvements include:

90% reduction in text reading time

24% increase in hardware efficiency through model quantization