VannGuard AI

CASE STUDY

SCALABLE ON-PREMISES DEPLOYMENT FOR HIGH-VOLUME LLM WORKLOADS

Overview

Querifi faced challenges of rate-limiting and latency with the existing LLMs. In collaboration with Querifi we devised a scalable on-premises Generative AI system that successfully addressed the customer's challenges. This solution allowed high-volume, low-latency Large Language Model (LLM) workloads across environments that ensured seamless user experience, ensuring that the system dynamically adapted to varying traffic demands.

Customer

Querifi

Country: Germany

Industry: B2B & B2C

Customer Size: 1000+

Publish Date: 25/07/2024

Problem Statement

Querifi faced significant challenges in managing the high traffic demands of their AI-powered applications. Their existing infrastructure struggled with rate-limiting issues, leading to bottlenecks in performance and scalability. To meet the growing needs of concurrent users while optimizing operational costs, Querifi needed a robust, scalable solution capable of dynamically adapting to fluctuating traffic demands without compromising performance.

Technical Solution

We deployed HuggingFace's Mistral-7b model on Azure, leveraging vLLM's advanced memory management features, particularly PagedAttention, to optimize GPU utilization and efficiently handle high traffic volumes. This ensured that even during peak demand, the system maintained fast response times. To support the dynamic nature of traffic, we integrated SkyPilot for auto-scaling, which allowed the system to intelligently scale resources up or down based on real-time user traffic. This approach ensured that additional GPU instances were automatically deployed during traffic spikes, while resources were deallocated during idle periods, optimizing operational costs.

We also implemented distributed cloud inference on Azure to guarantee low-latency access for concurrent users, coupled with robust load balancing mechanisms to handle traffic surges. Our focus on building a resilient and scalable infrastructure ensured that Querifi's AI-powered applications could seamlessly adapt to fluctuating user demands while maintaining high performance and cost efficiency.

Results

Our scalable on-premises infrastructure addressed Querifi's challenges with their existing LLMs, eliminating rate-limiting issues and significantly enhancing operational efficiency. The system successfully handled high-volume workloads, delivering fast response times and uninterrupted service, leading to improved user satisfaction. By dynamically adjusting to traffic demands and optimizing resource usage, our solution also reduced operational costs, providing Querifi with a scalable, cost-effective Generative AI platform.

Technologies

- Python
- Tensorflow
- Keras
- Docker
- SimpleITK
- SnapITK

Domains

- Computer Vision
- Deep Learning
- Machine Learning