VannGuard AI

CASE STUDY

SCALABLE ON-PREMISES DEPLOYMENT FOR HIGH-VOLUME LLM WORKLOADS

Overview

Querifai is an AI-powered platform that specializes in querying, extraction, and analysis of large datasets. By leveraging machine learning and natural language processing (NLP), Querifai simplifies data retrieval from unstructured sources, enabling businesses to quickly access actionable insights to improve decision-making and operational efficiency.

Customer

Querifi

Country: Germany

Customer Size: 1000+

Industry: B2B & B2C

Publish Date: 25/07/2024

Problem Statement

Querifai had integrated OpenAI's large language models (LLMs) to allow users to query vast datasets using natural language. However, as their dataset size grew significantly, the team encountered frequent rate limit issues despite increasing their quotas. This bottleneck affected throughput and hampered the user experience, limiting the platform's ability to efficiently scale.

Technical Solution

To resolve these challenges, we proposed and implemented an on-premise deployment of LLMs to eliminate rate limitations and improve performance. Specifically, we deployed the Mistral 7B model using vLLM on Microsoft Azure, ensuring high availability. Additionally, SkyPilot was integrated to handle autoscaling, dynamically adjusting the number of instances based on workload demands. This robust infrastructure utilized Microsoft Azure's cloud infrastructure paired with Docker and GitHub for version control and deployment automation. SkyPilot was instrumental in balancing the request load across multiple LLM instances, ensuring seamless scaling and improved responsiveness.

Results

Our solution provided a scalable on-premise infrastructure that resolved the performance issues Querifai had been facing. By eliminating rate-limiting bottlenecks, we significantly improved their platform's operational efficiency and user experience. With SkyPilot autoscaling in place, Querifai could dynamically manage system load, enhancing response times and ensuring a smooth, uninterrupted service for end-users. The overall outcome was a more reliable and scalable system, which enabled better workflows, greater user satisfaction, and an enhanced ability to handle high-volume queries from business clients.

Technologies

- vLLM
- SkyPilot
- MS Azure
- EC2 instances
- Python
- Mistral
- HuggingFace
- Docker
- Github

Domains

- MLOps
- Generative AI