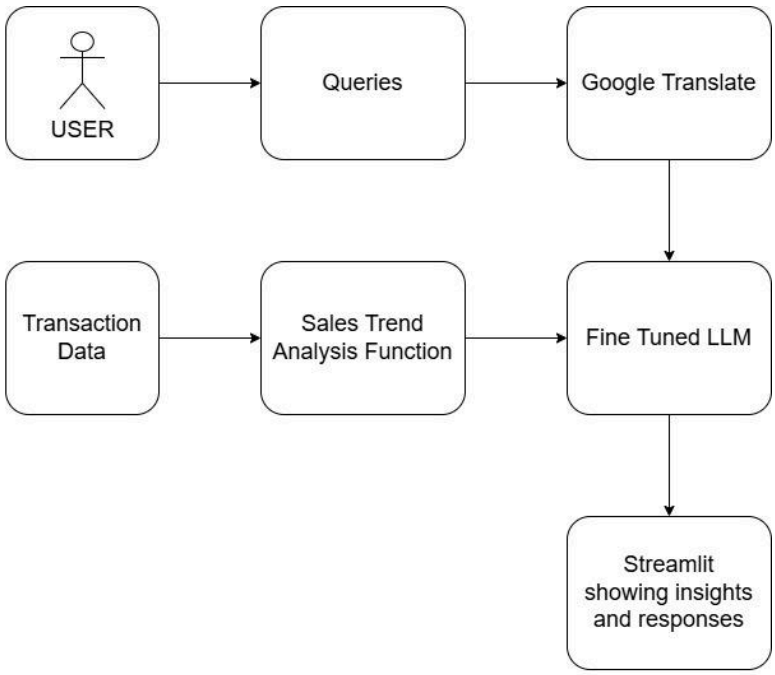
Solution Architecture



Data Utilization

To identify the sales trend and opportunities, we merged all the transaction data and developed a custom function to extract information from the merged data such as total sales, total orders, and average order value. The information then is displayed on the webpage and also fed into the LLM for context-aware responses. We also have a simple if-else condition to trigger alerts for critical issues from the extracted information for merchants.

Fine Tuning LLM

Fine-tuning an LLM customizes its behavior, enhances domain knowledge, and optimizes performance for specific tasks. To tailor the LLM specifically for Grab's merchant-partners, we created a sample synthetic dataset simulating merchant-related queries such as sales inquiries, trend interpretations, and customer insights. This dataset is then used to fine tune the "Llama-3.2-1B-bnb-4bit" model via the Unsloth framework.

Unsloth Framework

The reason we choose unsloth framework for fine tuning LLM is because it offers a cost-effective and time-efficient approach to develop a better performance model for specific tasks. Unsloth enables 2 times faster fine tuning with limited resources. Since we are fine tuning LLM with a laptop with a mid range GPU (Nvidia GeForce RTX 3070 laptop GPU), the Unsloth framework is the perfect solution for this.

Model Selection

The reason we choose Llama-3.2-1B-bnb-4bit as our base model is because the 1B model parameter size makes it lightweight enough to run on limited resources, which is ideal for small-scale deployments. Another reason is because this model enable bitsandbytes 4-bit quantization, which reduces memory usage and improves inference speed and performance. Besides, although this model is primarily trained in English, it is compatible with translation tools like googletrans to support multilingual queries.

Deployment

The fine-tuned model is deployed using Streamlit, which provides a simple, fast, and interactive frontend.