Project Proposal – Open Source LLM application

Introduction

Since the beginning of the year 2023, when OpenAI released ChatGPT and the respective large language models (LLMs) GPT3.5-turbo and GPT-4, it has been clear that artificial intelligence is on a level that it will create a shift to how people will do work from now on. There has been numerous releases of high-quality open source LLMs since, that companies can utilize without being vendor-locked to proprietary providers of LLMs.

For businesses and end-users to utilize AI efficiently, relevant data sources must be connected to the LLM model to derive unique business value. The aim of this project is to create such AI enabled application that create unique value to an end-user with selected data sources.

Project goals

Main goal of the project is to generate a fully-fledged application that demonstrates features of ConfidentialMind cloud stack. The solution should combine open source LLM together with relevant data sources that create a benefit for certain use-case. Our team will help the student group to decide what kind of application they wish to develop.

Examples of AI enabled applications that could be:

- Personal Assistant
    - Data sources could be personal files on end-user's computer, their google drive, email and calendar
    - The application would help a person to stay on top of daily tasks and improve the efficiency of searching data and documents, plus many other tasks
- Job function assistant for human resources
    - Application would help recruiters to automatically generate job postings, go through applicants' CVs and cover letters, and search from existing employee data base whether there are ideal candidates from inside the organisation

The open source LLM and backend microservices are hosted on ConfidentialMind cloud stack that uses Kubernetes to provide scalable multi cloud support for running LLMs and related AI applications.

Technologies

Depending on the chosen end-user application, the used technologies will vary, especially on the data connector side. Here is the tech stack that we recommend and support

- Application side database, authentication and simple edge functions **Supabase**
    - PostgreSQL with good client-side libraries and real-time support (similar to firebase)
    - Edge functions with Deno support for typescript functions
    - Vector database for embeddings can be hosted on Postgres using pg-vector addon
- AI microservices hosted on ConfidentialMind cloud
    - FastAPI python servers that can utilize ConfidentialMind helper pip-package
    - Option also to use Deno containers, but most of the AI-libraries might require python environment
    - Option to have both python and Deno containers mixed as they are microservices
- Client-side application and interface
    - React Next.js project hosted on Vercel or Python Streamlit (backend that generates simple UIs) or mixture of these

Our team is familiar with all the used technologies listed here and provide help getting started and solving possible issues

Requirements for the students

The team should have some experience from the following technologies and concepts

- Full-stack projects (web application plus microservice/serverless backend)
- Some SQL based database
- Typescript, React, Next.js, Python, FastAPI, Node/Deno

In addition, it is beneficial to have expertise or interest in the following technologies and concepts

- Using large language models through APIs
- Understanding how text-embeddings and vector databases work
- Creating wrappers for accessing different data source (SQL, no-SQL databases, parsing PDFs or CSV files, accessing cloud-based email inboxes or calendar via API, etc)
- Supabase or Firebase experience

Project difficulty varies from moderate to demanding depending on the chosen end-user application and chosen data-sources

Legal Issues

All IPRs to all Results will be transferred to the Client. The client will share some confidential information with the students. We aim to open source at least some parts of the created application as part of our documentation and example templates.

Client

ConfidentialMind is a startup based in Otaniemi Espoo. The company was founded in the spring 2023 and has total of 6 employees. Our core business is to help companies to use AI in a secure way. We focus on open source LLMs.

The whole team has experience in building cloud infrastructure and full stack applications. In addition, we have people with long background in machine learning and AI. We are dedicated to guide the team, and we have participated in this course 5 times before with our earlier companies. The team's Scrum Master will be Esko Vähämäki who is also a founder at ConfidentialMind so he has insight into the market and technologies used.

ConfidentialMind will provide the necessary resources for hosting the developed microservices and host the open source LLM model used in the project. In addition, we have meeting rooms and co-working space available at our Otakaari 7 office on-agree-basis.

Client representative(s)

Project PO from our side

- Software Developer (AI applications and AI inference)
- Henri Dahl
- e-mail: henri@confidentialmind.com
- +358 50 330 5976
- Otakaari 7 B, 02150 Espoo

Scrum Master

- Founder & Chief Architect
- Esko Vähämäki
- e-mail: esko@confidentialmind.com

- +358 40 817 9553

Preselected Student Team Members

- Scrum Master Esko Vähämäki