Proposta para Projeto em Informática

A Self-Hosted ChatBot Platform

Número de alunos: 5/6

Orientadores: Sérgio Matos (aleixomatos@ua.pt)
Tiago Almeida (tiagomeloalmeida@ua.pt)

Keywords

Artificial Intelligence; Generative Language Models; Large Language Models; ChatBot Platform; Personal Assistance.

Context and objectives

Revolutionary advancements in AI, driven by Deep Learning and Large Language Models, have notably enhanced natural language processing. ChatGPT stands out, offering human-like interactions that prove essential across fields such as customer service, education, and personal assistance. Despite its widespread adoption, ChatGPT's closed-source framework and subscription-based model restrict customization and transparency. This is particularly problematic when handling sensitive information, raising significant privacy concerns. This situation suggests the necessity for a more controllable, on-premises alternative.

In response to these challenges, we propose harnessing the potential of Open-Source Large Language Models (LLMs) to build a locally deployable chatbot platform. The primary goal is to offer sophisticated chatting capabilities akin to those of ChatGPT, but with the added benefits of on-premises deployment and customization based on user preferences and feedback. With this platform, developers and organisations can leverage the advances in AI chatbot technology while maintaining full control over their data.

The primary objective of this project is to develop a comprehensive full-stack chatbot platform designed for straightforward on-premises deployment (run one command). Specifically, the platform should include a web-based frontend, backend, user management, and feedback management systems. A special emphasis will be placed on retrieving and utilising user feedback data, requiring students to devise and implement innovative methods for gathering valuable user feedback. Such feedback data will enable us to align the current LLM with users' needs in the future, enhancing its usability and effectiveness.

Key components of the project include:

Frontend: A chat web interface equipped with authentication and authorization mechanisms, and a feedback system to record and process user feedback. Refer to alternative solutions [5, 6].

Authentication and Authorization: Comprehensive systems to control access and determine who is allowed to interact with the chatbot.

Feedback Management: Mechanisms that prompt users to provide feedback on their interaction quality with the chatbot.

LLM Chatbot Integration: Here, students can integrate already available high-level services for interacting with LLMs via REST, such as Ollama [4] and LocalAI [3], or implement a low-level service using llama.cpp [1] or llama-cpp-python [2].

Students have total freedom regarding the technology selection for implementing this project, encouraging innovation and customization.

More information

[1] https://github.com/ggerganov/llama.cpp
[2] https://github.com/abetlen/llama-cpp-python
[3] https://github.com/mudler/LocalAI
[4] https://github.com/ollama/ollama
[5] https://github.com/ollama-webui/ollama-webui
[6] https://github.com/FlowiseAI/Flowise