GLOBAL RESEARCH IMMERSION PROGRAM FOR YOUNG SCIENTISTS

CerebroCipher:
Your Private AI Assistant in the Box

Authors: Grigory FIlimónov (Polytechnic University of Catalonia) & Xinyu Li (National University of Singapore)

Background

Hosting Large Language Models(LLM) on traditional cloud servers causes signifcant memory and computational burdens, along with challenges of latency and privacy breach. This research addresses these challenges by using **edge computing solutions to deploy Llama3-7b on a micro server, providing LLM services for small and medium-sized enterprises (SMEs).**

Reseach Objectives

- Address privacy concerns by processing data locally.
- Deploy quantized Llama3-7b on edge servers which reduces memory and computational burdens.
- Utilize edge computing to improve performance, reduce latency, and lower data transfer costs for LLM services
- Provide small and medium-sized companies with a private and secure AI solutions.

Methods

Micro server Setup
-Configuration of the SOPHGO microserver SE7 32-EA4.

LLM Deployment
-Implementation of quantized Llama3-7b on the microserver.
-Optimization for efficient performance and resource usage.

Web Application Development
-Creation of a user-friendly UI for easy interaction with the LLM.
-Secure communication protocols to protect data privacy.
-Implementation of useful features.

Testing and evaluation
-Measuring response times and efficiency of the LLM.
-Evaluating the number of conversation rounds the system can manage**.**

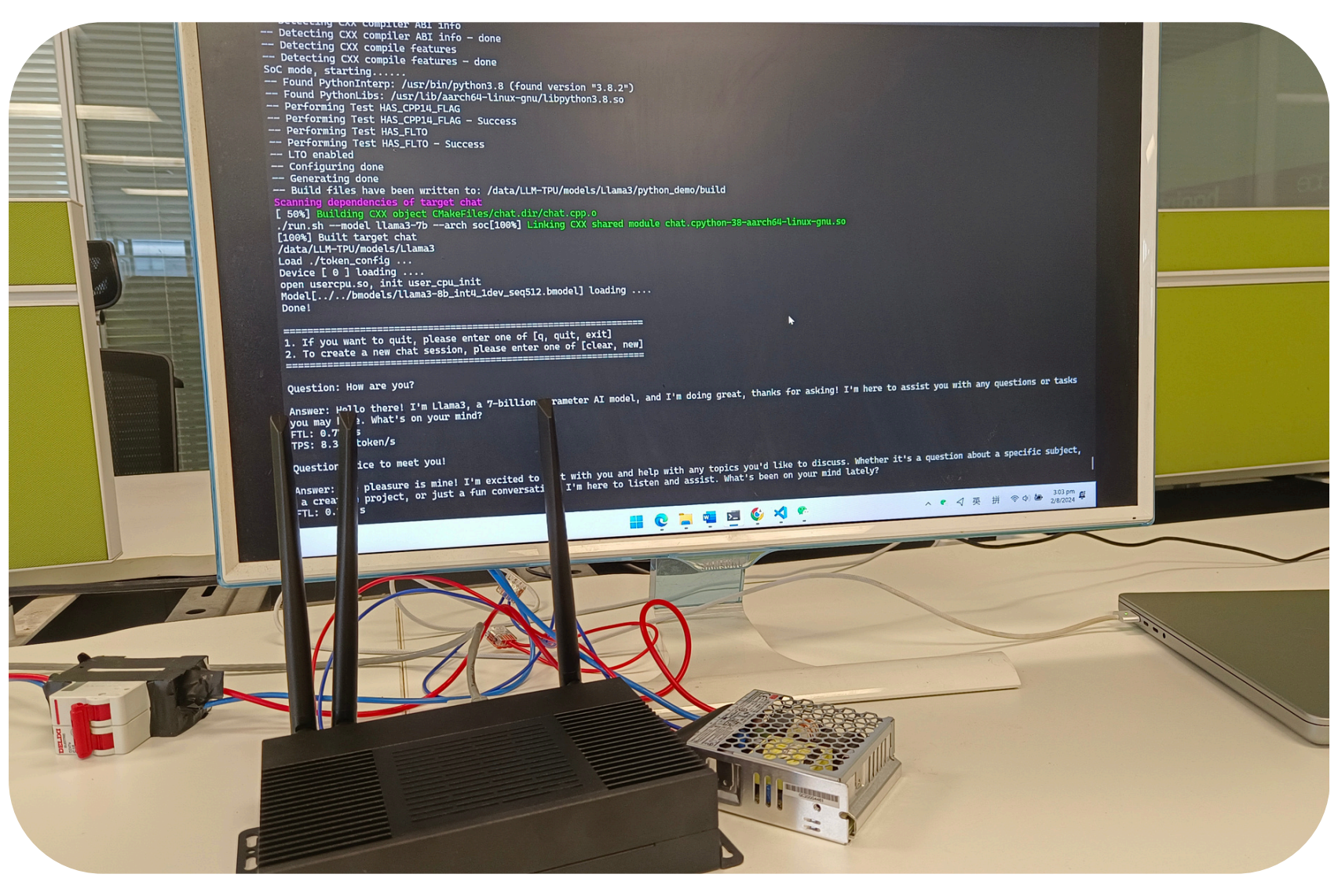
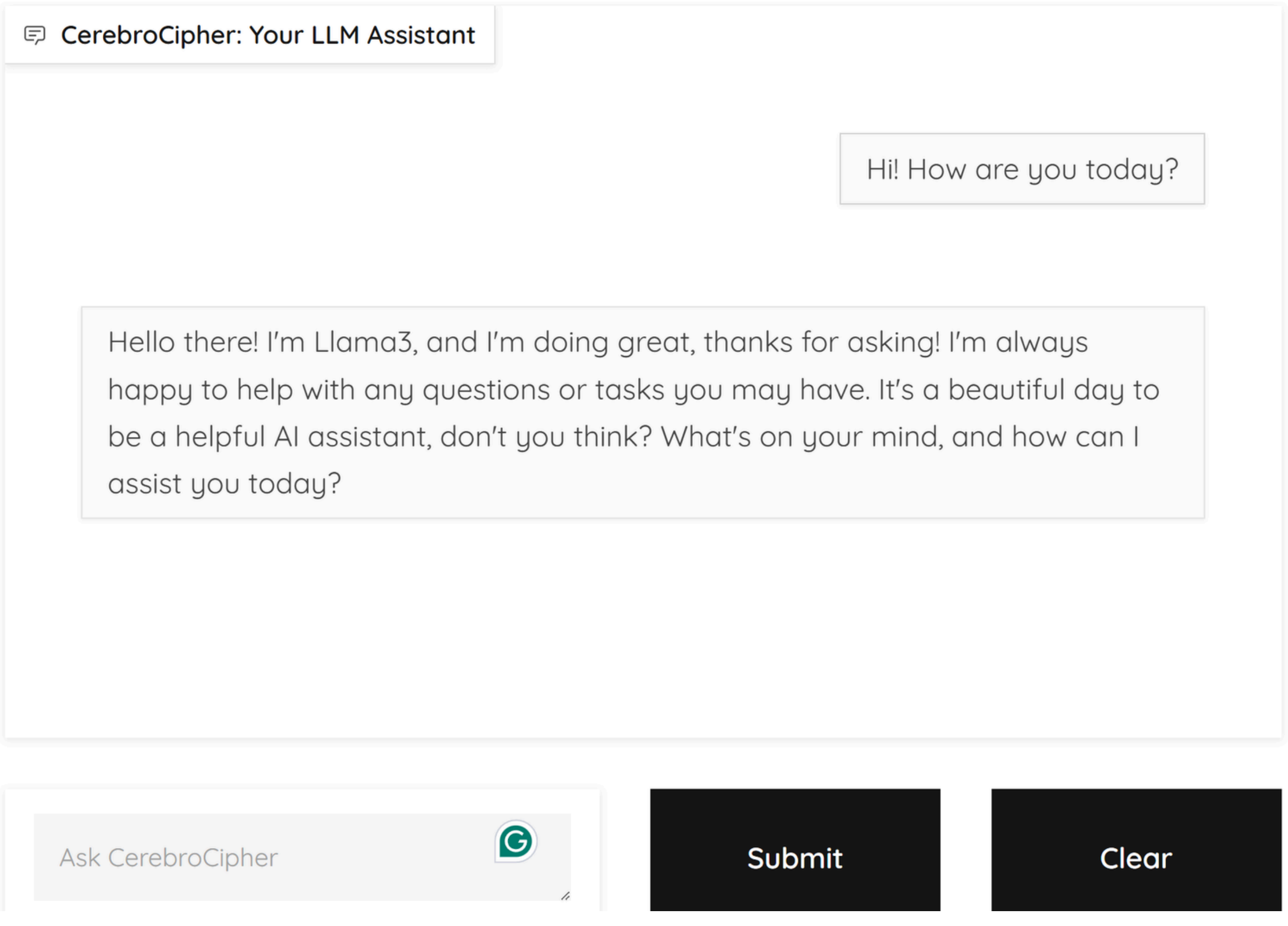


Figure 1: Interaction with quantized Llama3-7b on SE7 32-EA4 equipped with the 4th generation deep learning SoC BM1684X

Results

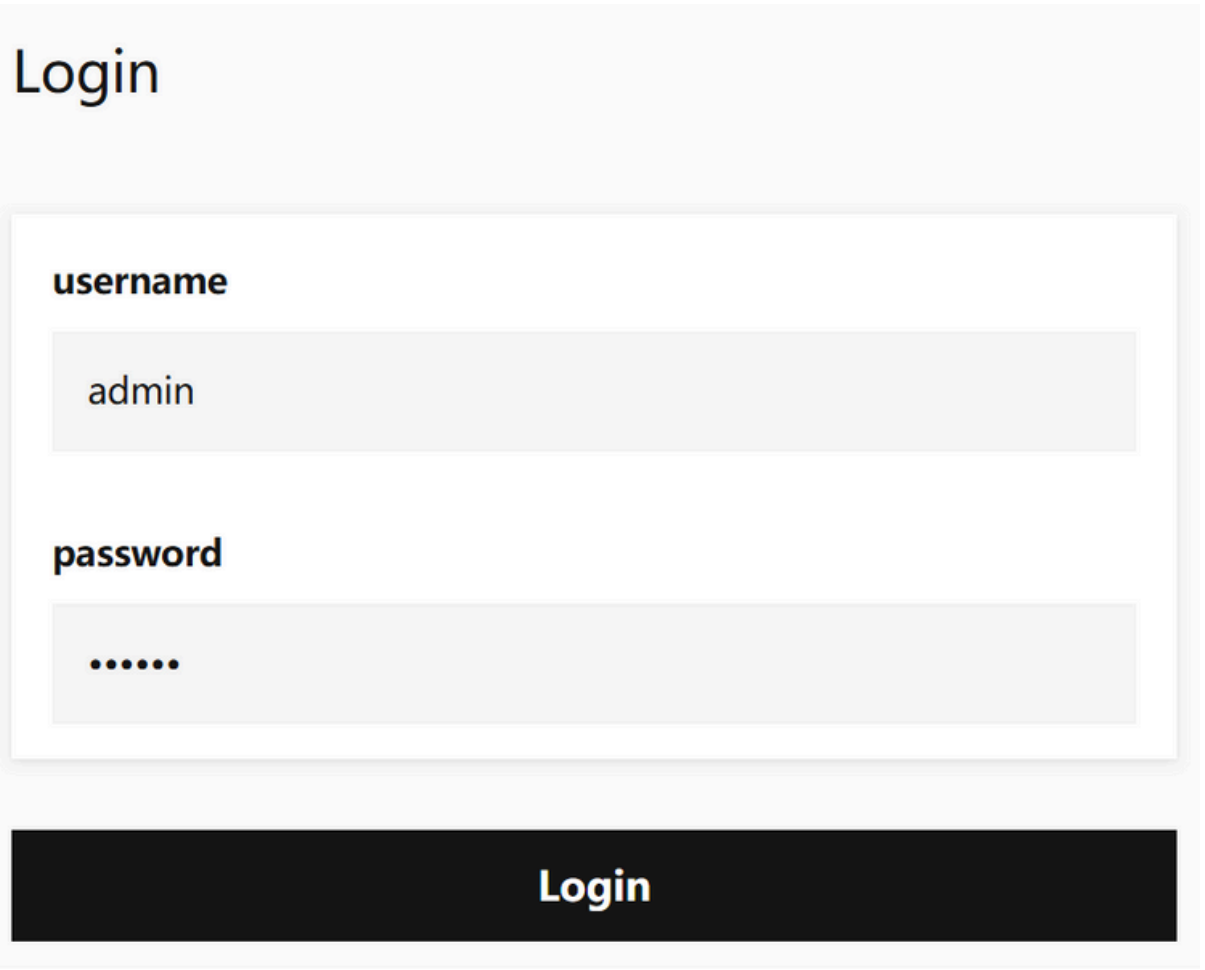
A web application was built with Gradio (Figure 2). Users can access it by simply entering the server address into their browser. The application supports text processing capabilities with short response times. An authentication page ensures that only permitted users can access the application (Figure 3). On average, it takes **0.773 seconds** for Llama3-7b to process and start generating responses, with an average processing rate of **8.078 tokens per second** (Figure 4). This has proven to be a stable and efficient AI solution with low latency.


CerebroCipher: Your LLM Assistant

Hi! How are you today?

Hello there! I'm Llama3, and I'm doing great, thanks for asking! I'm always happy to help with any questions or tasks you may have. It's a beautiful day to be a helpful AI assistant, don't you think? What's on your mind, and how can I assist you today?

Figure 2: CerebroCipher web application


Login

username: admin

password: ••••••

Figure 3: Authentication page

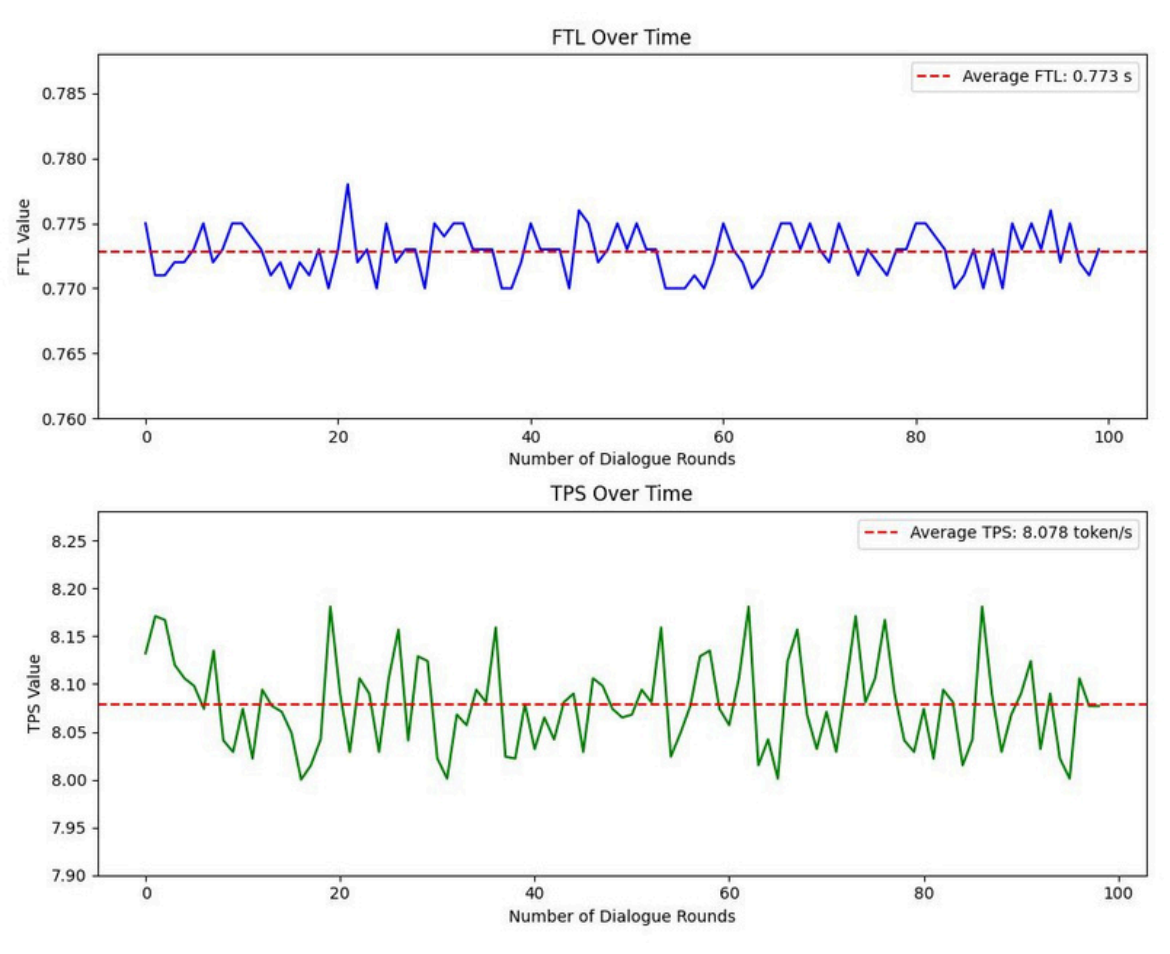


Figure 4: FTL and TPS changed over time

Discussion

In the current implementation, a constriant of limited memory from only one server was encountered, which impacts performance and scalability of our project. It restricts the maximum SEQLEN, feature complexity and necessitates single-thread usage, causing users to queue for responses. To address this issue, future improvements could include **deploying a stack of servers to handle more simultaneous users or exploring a device-server hybrid inference strategy to optimize performance and enhance user experience.**

Conclusions

A secure and efficient AI solution for small and medium-sized enterprises was successfully built by deploying Llama3-7b on SOPHGO SE7 microservers. The system supports fast text processing, providing robust capabilities while ensuring data privacy and control. Despite the current constraints on server memory and performance, the implementation demonstrates feasibility and potential of using edge servers to address privacy concerns associated with traditional cloud-based AI solutions. Future improvements will focus on enhancing scalability and performance to further meet the needs of SMEs.

Key References

Bang, J., Lee, J., Shim, K., Yang, S., & Chang, S. (2024, June 11). Crayon: Customized On-Device LLM via Instant Adapter Blending and Edge-Server Hybrid Inference. ArXiv.org. https://doi.org/10.48550/arXiv.2406.07007
Gradio. (n.d.). Gradio.app. https://www.gradio.app/
sophgo/LLM-TPU. (2024, July 31). GitHub. https://github.com/sophgo/LLM-TPU
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly. High-Confidence Computing, 4(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Contact Information

grigory.filimonov@estudiantat.upc.edu
xinyu.li26@u.nus.edu