Llama 2 and Its Use Cases in Mobile Apps

Introduction

Llama 2 is an advanced open-access large language model (LLM) developed and released by Meta in 2023 (Schmid et al., 2023). Compared to its predecessor, the new model provides a number of advantages for better generative AI tasks. The new models offer greater context length that is double of the previous version, greater accessibility with generous licensing models open to any type of organisation, and more robust and extensive knowledge base and contextual understanding (Bergmann, 2023). The following report will explore 5 possible use cases of Llama 2 in mobile applications in enabling intelligent automation and improving user experience.

1. AI-Powered Keyboards

With its ability to “replicate linguistic and logical patterns in the training data” (Bergmann, 2023), Llama 2 can be integrated into mobile keyboard apps to enhance text prediction, autocorrection, and context-aware suggestions. Unlike conventional predictive text models, it can be personalised by training it with user-specific data such as demographics, preferences and habits so that the language predictions and suggestions become more natural to each user.

2. Intelligent Note-Taking and Summary

Llama 2 models can be implemented in mobile note-taking applications to provide AI-driven transcription, summarisation and organisation of notes. The models can train on the language patterns and preferences of each user, and develop a better alternative or suggestion on the note content so that it can be better suited for each user and make it more efficient and effective in using the note applications.

3. Real-Time Speech Translation

Llama 2 can enhance real-time speech translation that could be used in a number of mobile apps, such as for audio conversation (phone call, voice chat), video interpretation and live subtitle translation. While Llama 2 models do not support audio data processing natively, it can be integrated with speech-to-text technologies so that Llama 2 can obtain the text-based input to process translation.

4. Smart Assistant

Llama 2 offers specific chat models that are “fine-tuned for dialogue-driven use cases” (Bergmann, 2023). These chat models can be used for a number of virtual assistant types of functionalities and mobile applications. These include generating responses to user requests based on the trained data, and further enhancing the response quality based on the insights it has gathered from previous user data and interactions.

5. Code Generator

Llama 2 also offers Code Llama that is “fine-tuned for generating code from both code-based and natural language-based prompts” with support for a number of popular programming languages (Bergmann, 2023). At a basic level, this can be integrated in mobile applications to generate code suggestions or solutions based on the text-based input. Furthermore, the applications can further integrate image processing technologies to enable users to take pictures of code blocks and generate responses based on the image inputs.

References

Bergmann, D. (2023, December 18). What is Llama 2?. IBM. https://www.ibm.com/think/topics/llama-2

Schmid, P., Sanseviero O., Cuenca P., Tunstall L. (2023, July 18). Llama 2 is here - get it on Hugging Face. Hugging Face. https://huggingface.co/blog/llama2