Introduction: This android app aims to achieve the real-time speech switch to text and real-time translate and using Llama2 to enhance the accuracy. This app use Llama2 as a LLM to train to adapt dialect and accent unless the traditional tools which rely on phrase based and context-aware translations, it can be used in study and communication.

Objective: This application can solve the traditional translate tools can’t understand special dialect and account and the time delay by using Llama2 for dynamic language adapt it can also provide the results with low-latency and high-fidelity.

Proposed Solution: the core architecture is Phone input-Android audio record-Speech to text engine (API)-Llama2 enhance the translate and output-UI display to user, it can be considered to add a local data record to prove and cache the users habit to enhance the speed.

Users can open the app and input the voice or text then wait for the dynamic UI designed to output the result they need. Also users can ask questions to Llama2 to get the answer to enhance their learning for the language.

As for the real time speech to text I decide to use google cloud speech-to-text API (cloud based) it can balance the latency and accuracy.

As for Llama2 I use it to clean misunderstand and the adapt the difficult means with polish then output a better result. Such as the misunderstand that riverbank is a bank.

I want to integrate Llama2 based more by cloud to reduce the app’s format and reduce latency then it can also made the app more light just cache frequent used phrases such as the greeting words.

For the workflow Llama2 can preprocess the text and resolve the ambiguities then post-process the translated text to polish and improve the fluency.

Technology Stack: using Java in android studio for audio push and UI. Connect Google’s Speech-to-text API to handle the text transfer and use Llama2 to preprocess the raw text and polish the result.

Development Methodology: As level1 I decided to use some offline translate and audio to build my own database and prepare to train the model. In level2 integrate the real-time translate API and Llama2 context layer. Then in level3 optimize latency by training and model pruning. In the final level use dataset and samples to text and optimize the model.

Conclusion: This app aims to broken the barrier of language with AI powered engine and There also have some limits such as how to balance the latency and model size and achieve the high accuracy in complex language environment. When overcome these barriers it can push the real-time translate tools.