Llama 2, an open-source large language model (LLM) by Meta, was released in July 2023. LLMs use transformers, a type of neural network, at their core, with the mechanism "Attention is all you need," focusing on the significant meaning of sentences by adjusting relevant weights. The pre-trained version of Llama 2 uses Grouped Query Attention (QGA) and other techniques such as Supervised Fine-Tuning (SFT), Reinforcement Learning with Human Feedback (RLHF), and Ghost Attention (GAtt).

Annotators provide prompts to different models, choose the better responses from those, and feed them back to the model as a reward. Based on the rewards, the model adjusts the weights and generates better results next time; this is called helpfulness RLHF. Safety RLHF is a variant of RLHF which prevents harmful responses from both inputs and outputs using similar techniques. Ghost Attention is more similar to the GPTs, remembering specific instruction terms like "act as." For example, when a user requests the model to act as a recruiter at the beginning and then gives practical recruitment responses afterwards. GAtt works by concatenating an instruction with all user prompts in a conversation and generating instruction-specific responses.

The use cases and applications of Llama 2 include content generation, customer support, information retrieval, financial analysis, and healthcare use cases. Content generation can be used more for commercial marketing purposes, such as generating social media posts and advertising content. Information retrieval can better understand user intent and provide accurate information, improving productivity and efficiency.

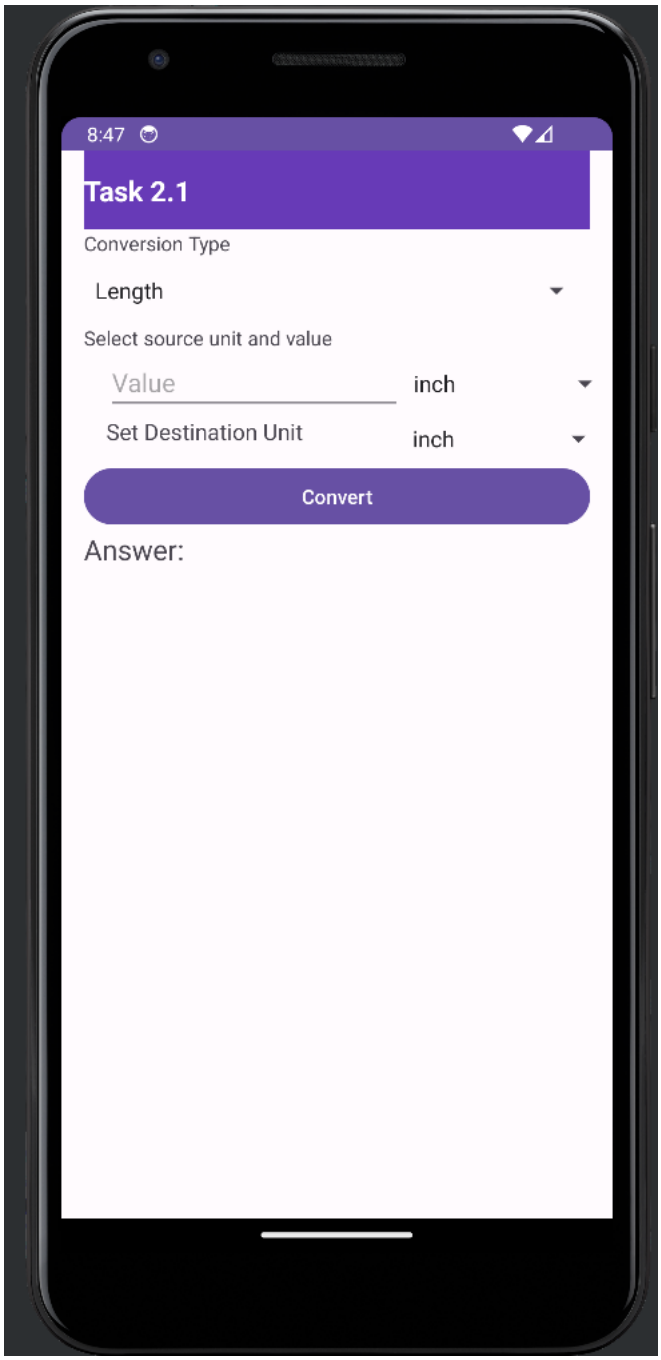
The results of the model evaluation indicate that Llama 2 excels in mathematical reasoning. As a result, financial organizations can develop efficient virtual assistants for finance, assisting customers with financial analysis and making informed decisions. Financial analysis provides customers with virtual financial assistance. This helps those who lack financial knowledge but face problems in this field, enabling them to make decisions using natural language.

In healthcare, Llama 2 can analyze medical reports containing several terminologies not well understood by laypersons. The model can convert complex terms into simpler ones. Additionally, patients can provide their symptoms to the pre-trained and fine-tuned model, which can then provide a rough diagnosis.

Customer support is the most concerning use case. Nowadays, e-commerce platforms equip chatbots on the side. They help not just by answering customers, but also by assisting with services connected to the backend, such as booking appointments or collecting points from physical receipts at Woolworth's market. This helps reduce the workload of customer support by automating repetitive tasks, allowing problems to be solved around the clock. This represents the latest trend in commercial websites, not just showcasing the company itself but also equipped with hotline customer services. This could overtake traditional websites in a new way, as there's unnecessary content on some websites today, causing customers to spend more time just to find specific information. LLMs-based chatbot provides quick and accurate responses by talking to customers, using natural languages, summarizing the contents, and retrieving the important contents.

GitHub Link: https://github.com/Yee1955/SIT305

Screenshot of the main app screen:



Demo Video Link: https://github.com/Yee1955/SIT305/blob/master/Demo%20Video.mov