상상을 구현하는
풀스택 AI 개발자
이한울 입니다.

CONTACT.
010-8423-4825
narehasedevel@gmail.com

상상을 구현하는 풀스택 AI 개발자 이한울 입니다.

Experience

2023 . 02 . 03 - 2024 . 01 . 03 선문대학교 컴퓨터 공학부

2024 . 02 . 03 - 0000 . 00 . 00 국립 공주대학교 인공지능 학부

Awards

0000 . 00 . 00 전국 고교 고려대학교 자율주행 경진대회

0000 . 00 . 00 SW중심대학 기술 창업 경진대회 우수상

0000 . 00 . 00 공학교육혁신협회 공학설계 캠프 장려상

Skill

- Python
- kotlin
- C++
- swift
- JS
- pytorch
- keras
- flask

01

2024 Portfolio

projects

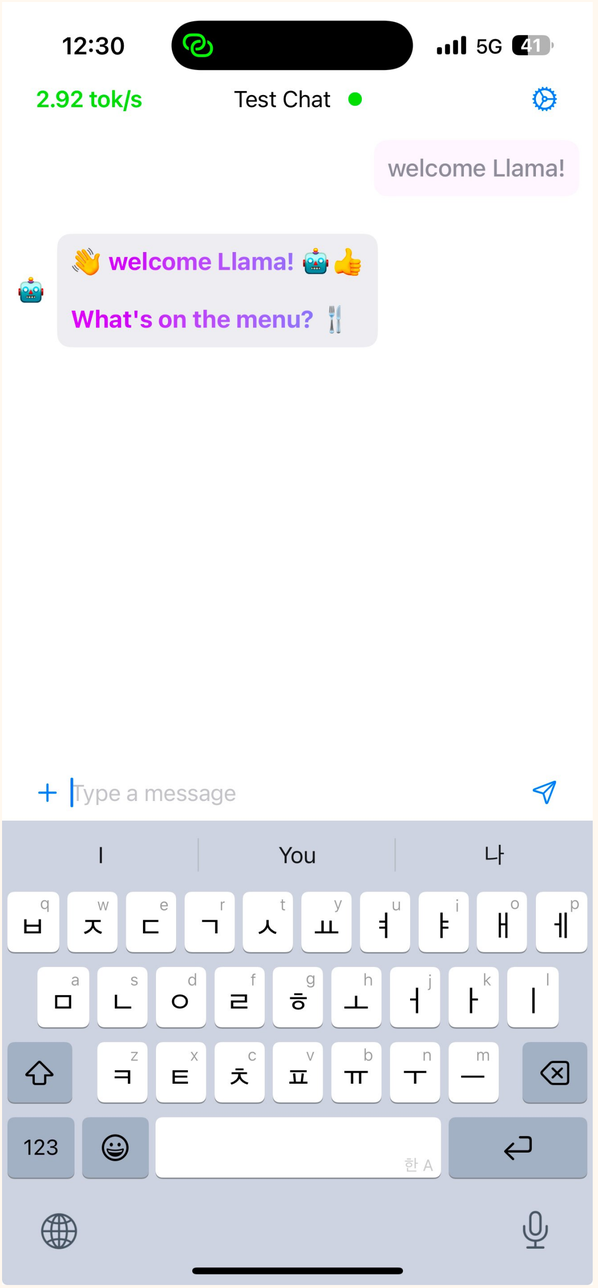
- Llama with iPhone
- Mail-Llama

Local-Llama Mobile

개요

OpenAI의 ChatGPT가 일상에 자리 잡으면서 개인정보 유출 문제가 부각되고 있다.
사용자 데이터가 서버로 전송되어 학습 데이터로 활용될 가능성이 이 문제의 핵심이다.
이를 해결하기 위해 On-Device AI 기술을 기반으로 한 **Local-Llama with Mobile** 솔루션을 제안한다.

WWDC24에서 Apple이 CoreML의 LLM 지원을 공식화함에 따라,
CoreML을 활용하여 Meta의 Llama 모델을 변환 및 최적화하고,
Apple Neural Engine(ANE)을 통해 iOS 기기에서 구동하였다.

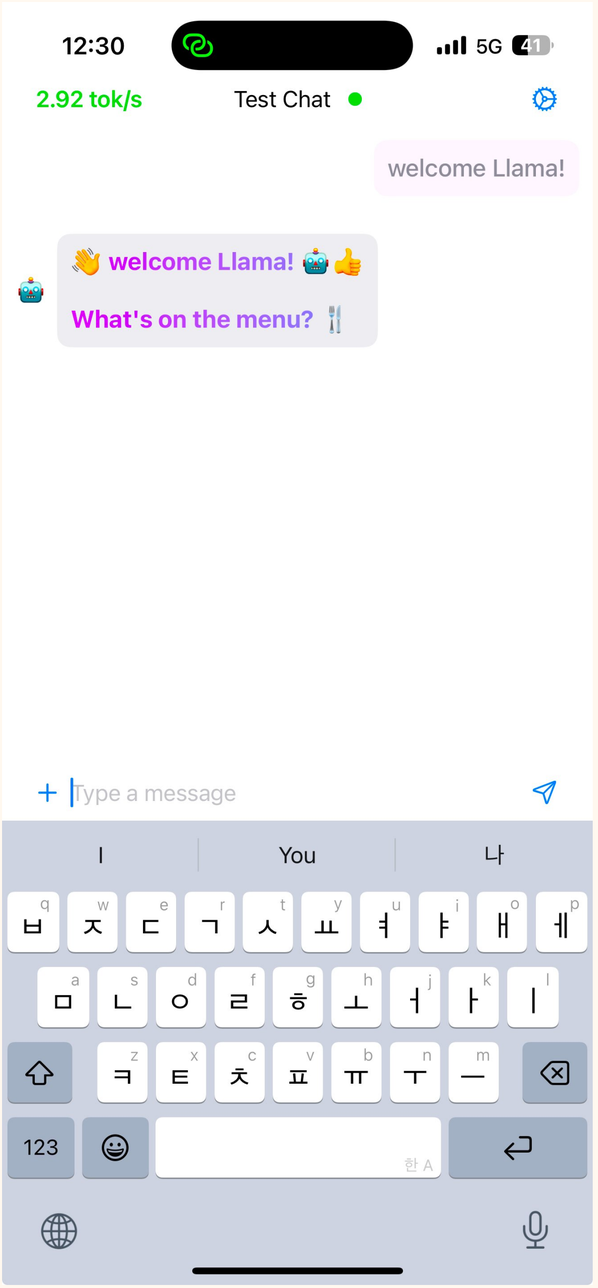


Local-Llama Mobile

문제 정의

최근 OpenAI의 ChatGPT와 같은 대형 언어 모델(LLM)이 일상생활의 일부로 자리 잡으면서 **개인정보 유출** 문제가 점차 부각되고 있다. 이러한 문제는 사용자의 대화 기록이 OpenAI 서버로 전송되고, 잠재적으로 학습 데이터로 활용될 가능성에서 기인한다.
이는 민감한 데이터의 외부 유출과 프라이버시 침해 우려를 야기하며, 사용자 신뢰를 약화시키는 요인으로 작용하고 있다.

이를 해결하기 위해 **On-Device AI** 기술을 활용한 **Local-Llama with Mobile** 솔루션을 구상하였다. 이 접근법은 사용자의 데이터를 장치 내에서만 처리하도록 하여, 서버와의 데이터 교환 없이 높은 수준의 프라이버시를 보장하고자 한다.

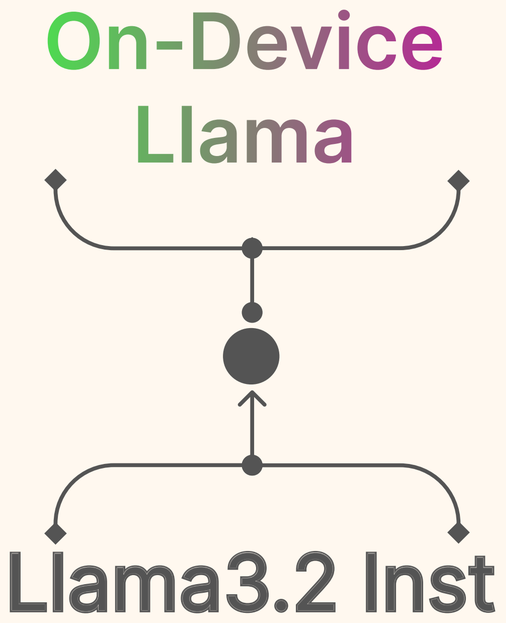


Local-Llama Mobile

접근 방법

WWDC24에서 Apple은 CoreML을 활용한 대형 언어 모델(LLM) 지원을 공식화하며 관련 문서를 공개하였다. 이로 인해 CoreML은 On-Device AI 구현의 강력한 도구로 자리 잡았다. 이를 기반으로, CoreML을 활용하여 Meta의 Llama 모델을 변환하고 최적화된 형태로 모바일 환경에서 구동하는 방식을 채택하였다.

변환된 Llama 모델은 Apple Neural Engine(ANE)과 같은 하드웨어 가속을 적극 활용하여 성능과 에너지 효율성을 극대화하며, 모델이 모든 데이터를 로컬에서 처리하도록 설계되었다. 이 접근법은 기존의 서버 기반 LLM 솔루션 대비 프라이버시를 강화하고, 사용자가 안전하게 AI를 활용할 수 있는 환경을 제공한다.

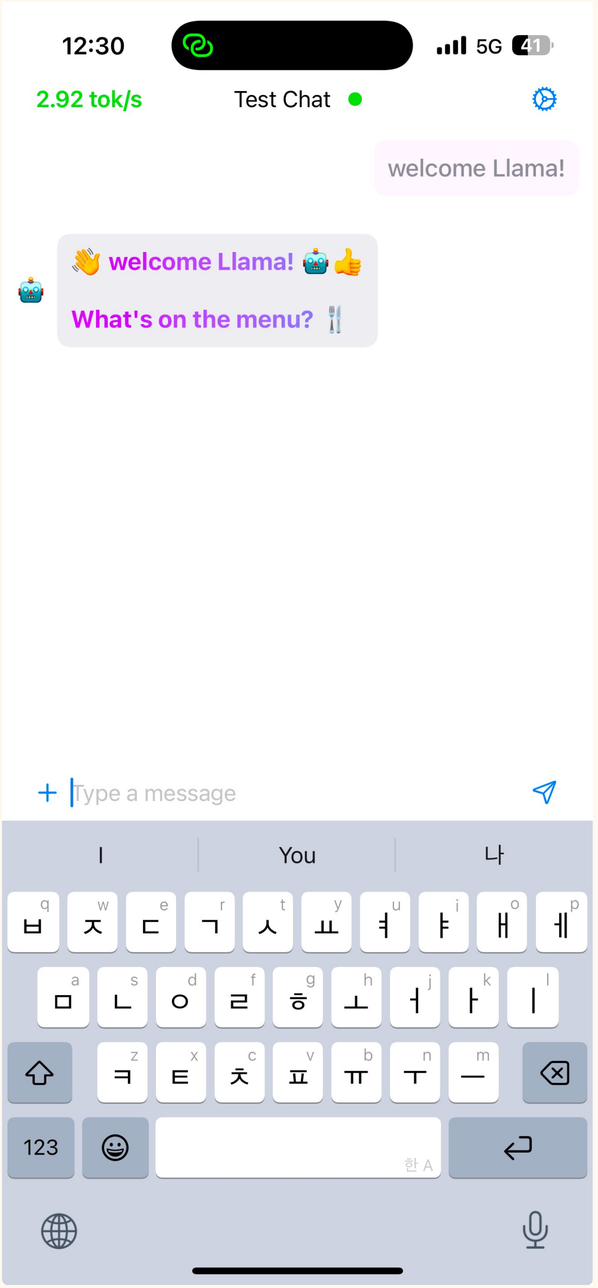


Local-Llama Mobile

결과

CoreML을 통해 변환된 Llama 모델은 iOS 기기에서 성공적으로 구동되었으며,
On-Device 환경에서 대형 언어 모델(LLM)의 실시간 응답성을 확인할 수 있었다.
로컬 처리 방식 덕분에 서버와의 데이터 교환 없이 모든 연산이 기기 내에서 이루어졌고,
이를 통해 **사용자 개인정보 보호**와 **데이터 보안성**을 크게 강화할 수 있었다.

또한, Apple Neural Engine(ANE)을 활용한 최적화 덕분에 모델의 응답 속도와
에너지 효율성도 기대 이상의 성능을 발휘하였다. 이는 모바일 환경에서 대형 언어 모델의
상용 가능성을 입증하며, 개인정보 유출에 대한 우려를 실질적으로
해소할 수 있는 기반을 마련하였다.



Local-Llama Mobile

문제 정의

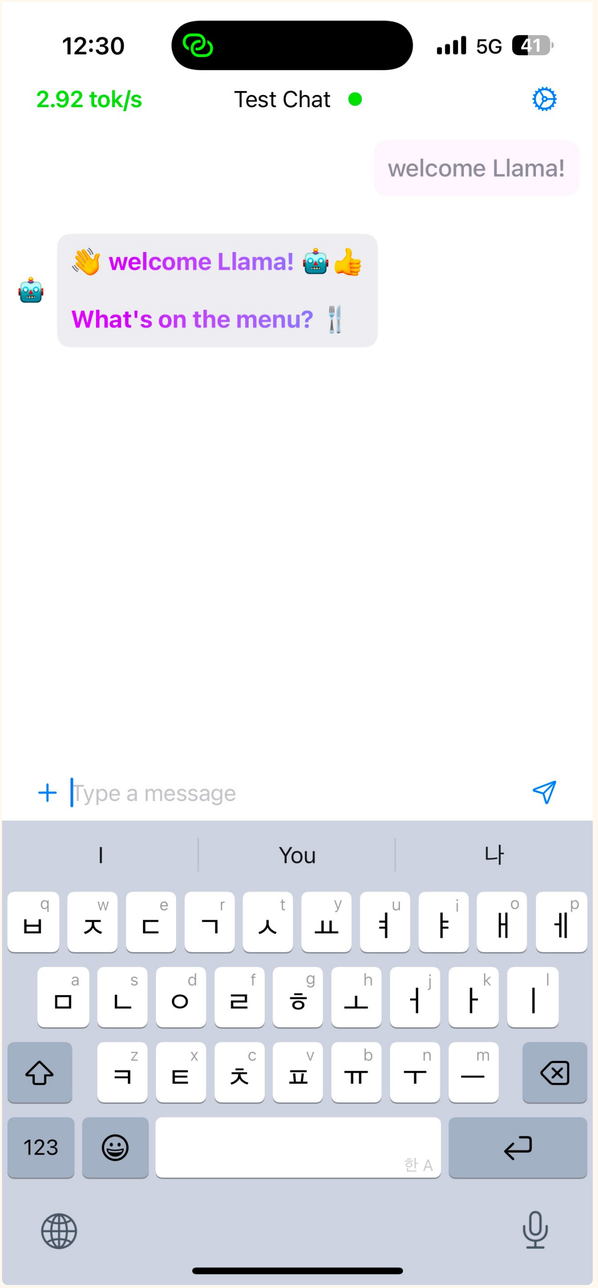
Meta등 다양한 AI기업에서 Open-SLM을 출시하면서 OpenAI의
Chat GPT와 달리개인정보 보호를 위한 personal LM의 필요성을 느끼고 있다.
이에 기존의 Nvidia-GPU가 아닌 mobile device에서 구동할 수 있지 않을까 라는
생각에 프로젝트를 구상하였다.

접근 방법

문제를 해결하고자 방법으로 On-Device 기술을 활용하기로 하였다. On-Device를 통해
사용자의 Device에서 LM을 구동하는 것을 목표로 하였으며

결과

INT4, INT8 등으로 모델을 양자화 하여 모델을 개선하는 등,
여러 최적화를 통해서 iPhone 에서 모델을 구동할 수 있도록 하였다..



Local-Llama Mobile

접근 방법

Local-Llama Mobile

문제 정의

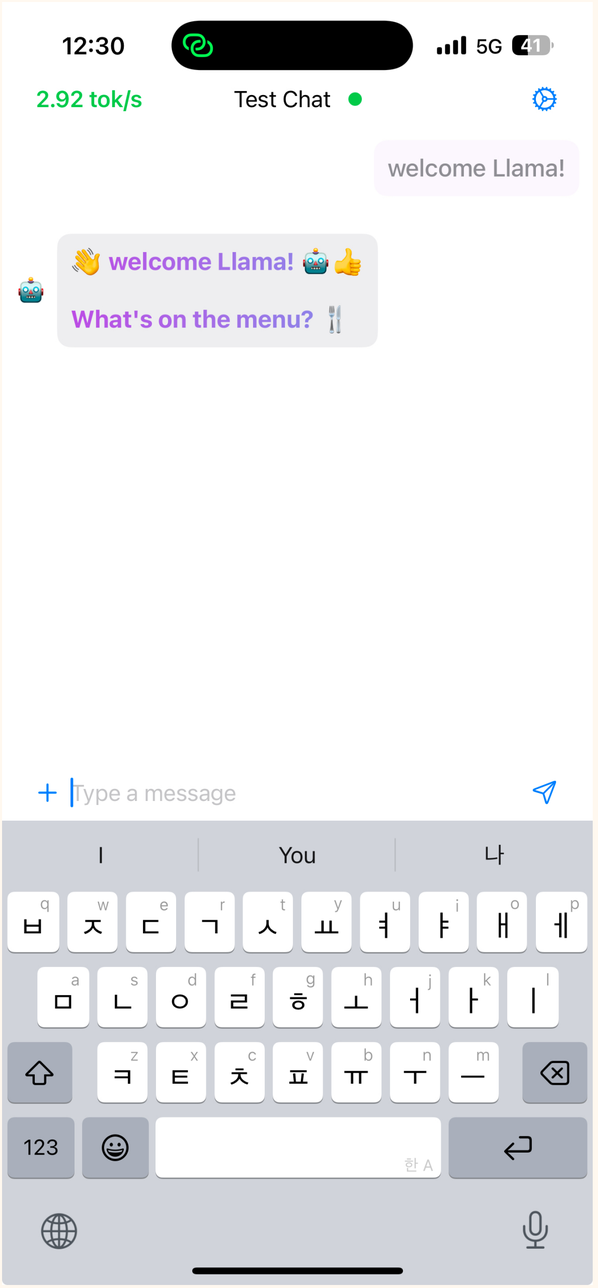
Meta등 다양한 AI기업에서 Open-SLM을 출시하면서 OpenAI의
Chat GPT와 달리개인정보 보호를 위한 personal LM의 필요성을 느끼고 있다.
이에 기존의 Nvidia-GPU가 아닌 mobile device에서 구동할 수 있지 않을까 라는
생각에 프로젝트를 구상하였다.

접근 방법

WWDC24를 통한 CoreML을 통한 LLM 모델 지원을
공식화 하며 여러 문서를 공개하였다. 이에 CoreML을 통해서
Llama를 변환하여 구동하기로 함.

결과

INT4, INT8 등으로 모델을 양자화 하여 모델을 개선하는 등,
여러 최적화를 통해서 iPhone 에서 모델을 구동할 수 있도록 하였다..



Local-Llama Mobile

문제 정의

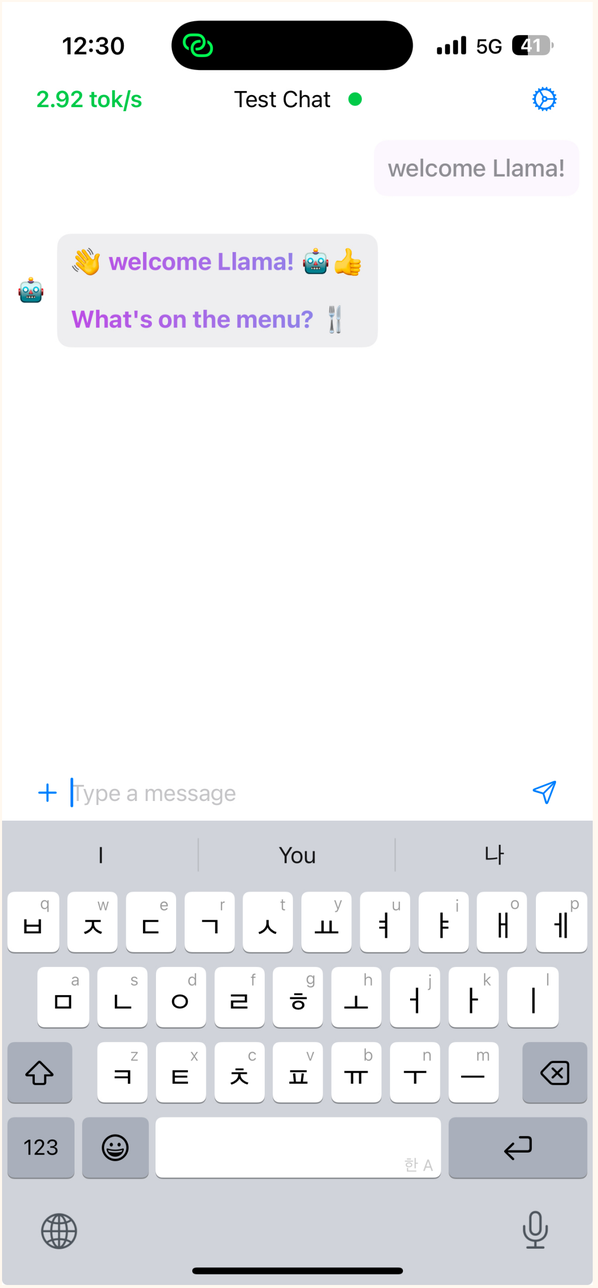
Meta등 다양한 AI기업에서 Open-SLM을 출시하면서 OpenAI의
Chat GPT와 달리개인정보 보호를 위한 personal LM의 필요성을 느끼고 있다.
이에 기존의 Nvidia-GPU가 아닌 mobile device에서 구동할 수 있지 않을까 라는
생각에 프로젝트를 구상하였다.

접근 방법

WWDC24를 통한 CoreML을 통한 LLM 모델 지원을
공식화 하며 여러 문서를 공개하였다. 이에 CoreML을 통해서
Llama를 변환하여 구동하기로 함.

결과

INT4, INT8 등으로 모델을 양자화 하여 모델을 개선하는 등,
여러 최적화를 통해서 iPhone 에서 모델을 구동할 수 있도록 하였다..



Local-Llama Mobile

문제 정의

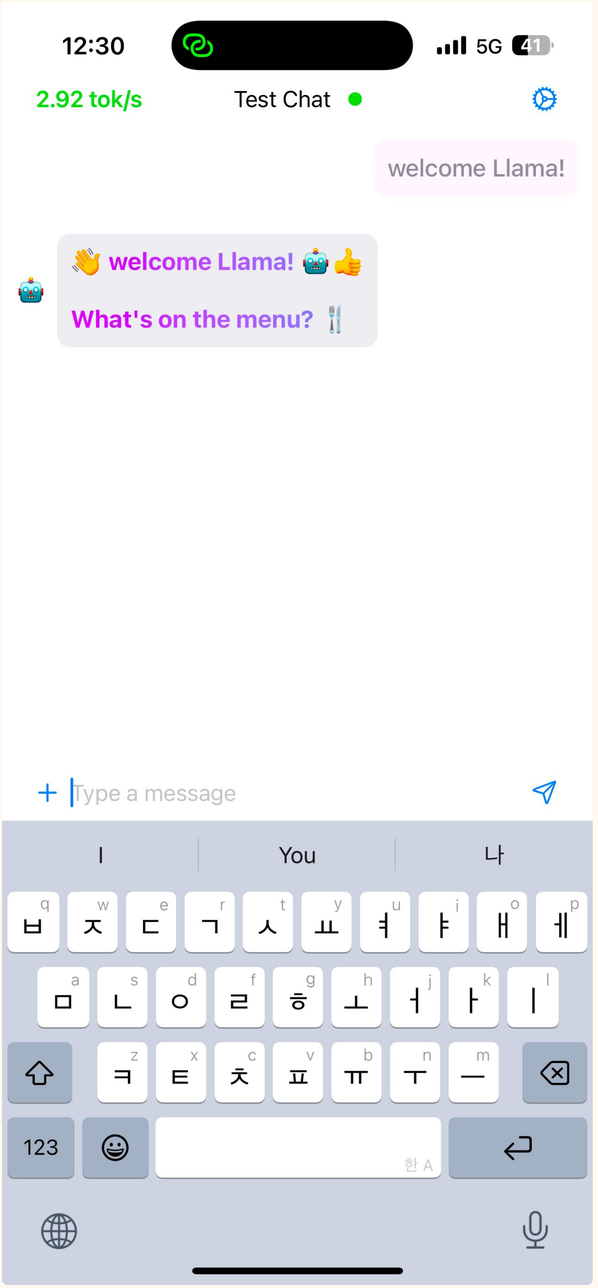
OpenAI의 ChatGPT가 일상에 한부분으로 자리잡기 시작하며 개인정보 유출이란 문제가 떠오르고 있으며 이는 사용자의 기록은 OpenAI에서 학습데이터로 취급하며 문제가 생기는 것 이다. 이를 해결하고자 On-Device를 활용한 Local-Llama with Mobile을 구상하게 되었다.

접근 방법

WWDC24를 통한 CoreML을 통한 LLM 모델 지원을
공식화 하며 여러 문서를 공개하였다. 이에 CoreML을 통해서
Llama를 변환하여 구동하기로 함.

결과

INT4, INT8 등으로 모델을 양자화 하여 모델을 개선하는 등,
여러 최적화를 통해서 iPhone 에서 모델을 구동할 수 있도록 하였다..



Local-Llama Mobile

문제 정의

OpenAI의 ChatGPT가 일상에 한부분으로 자리잡기 시작하며 개인정보 유출이란 문제가 떠오르고 있으며 이는 사용자의 기록은 OpenAI에서 학습데이터로 취급하며 문제가 생기는 것 이다. 이를 해결하고자 On-Device를 활용한 Local-Llama with Mobile을 구상하게 되었다.

접근 방법

WWDC24를 통한 CoreML을 통한 LLM 모델 지원을
공식화 하며 여러 문서를 공개하였다. 이에 CoreML을 통해서
Llama를 변환하여 구동하기로 함.

결과

INT4, INT8 등으로 모델을 양자화 하여 모델을 개선하는 등,
여러 최적화를 통해서 iPhone 에서 모델을 구동할 수 있도록 하였다..