# （一）移动端大模型部署

## 部署背景:

随着人工智能技术的迅速发展，大规模语言模型（Large Language Models, LLMs）在自然语言处理、计算机视觉等领域取得了显著成果。然而，传统的大模型通常需要高性能服务器和大量的计算资源进行推理，这对用户端尤其是移动端的实时应用场景提出了挑战。为了提升用户体验、降低延迟以及减少对云计算资源的依赖，将大模型部署到移动端成为一个重要的研究方向和应用需求。  
 本项目旨在探索如何在移动设备上部署和优化大模型，利用边缘计算的能力，在保证模型推理精度的同时，提升推理速度和节约设备资源。该项目将以常见的移动平台（如Android或iOS）为基础，完成大模型的剪枝、量化等优化，并在真实设备上测试模型的性能。

## 主要功能

**问题回答**：实现一个基于大语言模型的智能问答功能，用户可以通过文本输入问题，模型将根据其语义理解能力进行自然语言生成，提供准确、连贯的回答。例如，用户可以输入关于天气、历史、技术等问题，模型将在移动端提供即时的反馈。

**语音输入**：集成语音识别功能，用户可以通过语音与应用进行交互。系统将语音转换为文本，并将该文本传递给大模型进行进一步处理。此功能适用于语音助手等应用场景，提升用户体验。

**语音朗读**：通过集成文本转语音（Text-to-Speech, TTS）技术，应用可以将大模型生成的文本内容朗读出来。用户无需阅读屏幕即可获取回答或信息。

**模型更新与动态加载：**支持模型的动态更新与加载，用户能够根据需求更新或切换不同的模型，从而提升系统的灵活性。

## 开发环境

**操作系统**：macOS，Ubuntu（用于开发和调试）

**移动设备**：Android（推荐Android 9及以上），iOS（推荐iOS 14及以上）

**开发工具**：

* + Android Studio（用于Android端开发）
  + Xcode（用于iOS端开发）
  + MLC-LLM（用于模型转换和优化）

## 基本架构模块

文字/语音输入模块

模型推理模块

前端界面

生成的回答

用户的输入

语音朗读模块

本地存储模块

**前端架构**前端采用跨平台框架或原生应用开发，负责与用户的交互界面。主要功能包括语音输入、文字输入、语音朗读、对话展示等。前端通过API接口与本地推理引擎进行数据交互，保证用户操作的流畅性。

**模型推理模块**  
模型推理模块是项目的核心组件之一，部署经过优化的大语言模型，并通过TVM等工具进行模型编译优化和推理加速。模块还需支持硬件加速，如GPU或NPU，以提高推理速度和节省电池消耗。

**本地存储模块**应用将用户与模型的所有对话记录保留在本地存储中，供用户随时访问、管理，并支持重新生成和修改回答的操作。通过本地存储的方式，系统可以在离线环境下依然提供历史记录访问功能，同时避免了将数据上传到云端，从而保护用户隐私。

**语音输入模块**

语音输入模块提供语音识别功能，允许用户通过语音与应用进行交互。语音输入模块使用 ASR 技术将用户的语音实时转换为文本，供本地推理进行后续处理。

**语音朗读模块**

语音朗读模块将大模型生成的文本通过 TTS 技术转换为语音反馈给用户，提升应用的交互体验。