LLM OS report

第一期轮转汇报

柯宇斌

北京大学信息科学技术学院

2024 年 3 月 17 日

郭耀

- 两个轮转课题（新型操作系统）（面向大模型的泛在操作系统）
- 轮转内容
  - 各课题粗读五篇论文，确定具体轮转调研方向 (Week 1-2)
  - 每周搜集 1-2 篇论文精读，调研，完成报告 (Week 3-5)
  - 进度快的可进行初步实验，可在组会上汇报
- 轮转支持
  - 每周与老师一对一沟通一次
  - 可选参与组会

LLM+OS

- 大模型是近期极度火爆的概念，如今的大模型已经逐渐具有智能
- 人们对电脑的期望是函数式的（给定输入得到输出）
- LLM 符合这一特性
- LLM 已经走进应用，然后并未真正发挥智能，突破预定逻辑流
- 人们希望探索如何将 LLM 插入 OS，同时对 LLM 进行针对性的改造（严谨性等）
- 本文讨论的是 LLM+OS 的系统架构的粗浅设计

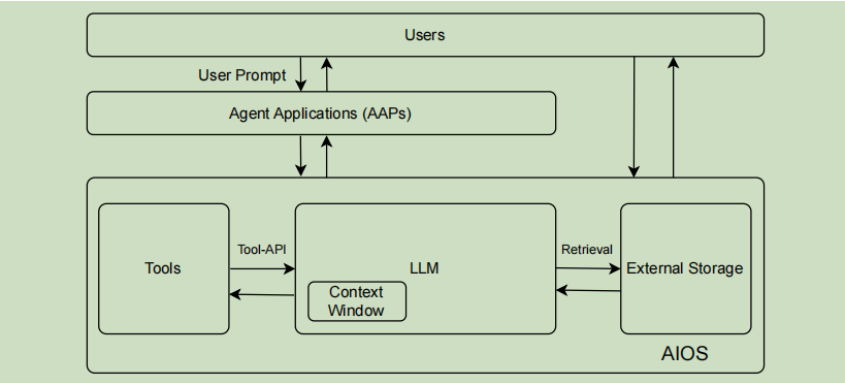
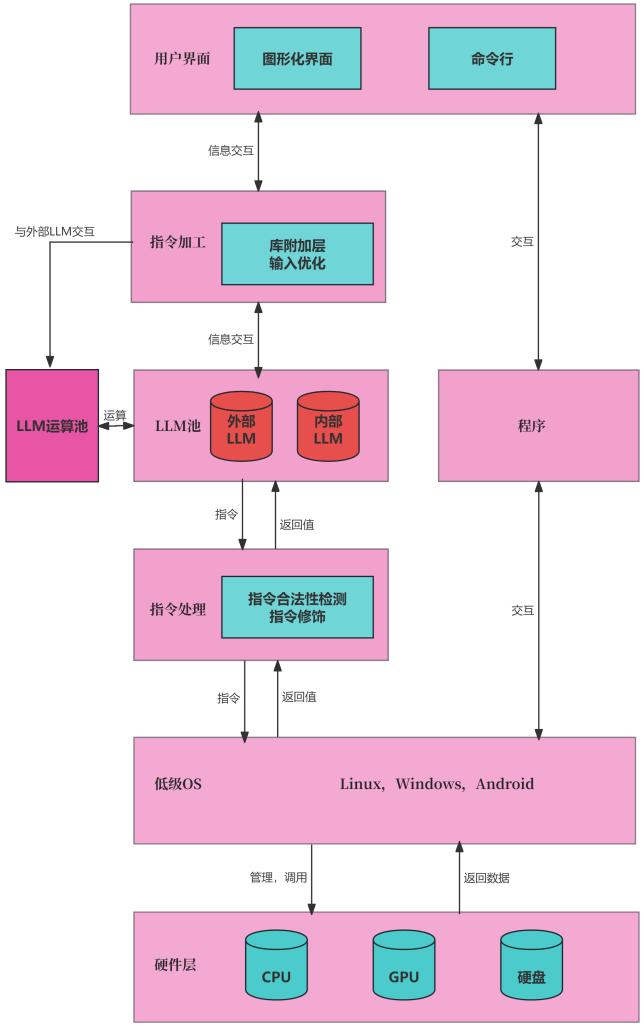


图 1: 架构 1

- paper:LLM as OS, Agents as Apps– Envisioning AIOS
- 提出了一个完整的 LLM 架构
- 完全新的一个系统，跨度大，可能需要一些中间过渡阶段

系统架构图


用户界面
图形化界面
命令行
信息交互
与外部LLM交互
指令加工
库附加层
输入优化
交互
信息交互
LLM运算池
运算
LLM池
外部
LLM
内部
LLM
程序
指令
返回值
指令处理
指令合法性检测
指令修饰
交互
指令
返回值
低级OS
Linux, Windows, Android
管理，调用
返回数据
硬件层
CPU
GPU
硬盘

指令处理模块

- 如何解决 LLM 输出的幻视现象，提高其正确率。(另一个调研方向)
- 如何训练 LLM 使用 API (THU Toolllm–Facilitating large language models to master 16000+ real-world apis)
-

LLM 池

- 如何实现 LLM 的通用性？
- 在 LLM 头部尾部加装一个转换器，每次输出时也输出一份标准化输入。可以以一个公认开源模型为标准
- 实现 LLM 快速切换？共享硬件数据？隔离问题？
- LLM 的调度管理系统（Fast Distributed Inference Serving for Large Language Models）
- LLM 的上下文窗口不足的问题（MemGPT– Towards LLMs as Operating Systems）

指令加工

- 自然语言修饰（大多数大模型已经提供这一功能）
- (ERNIE–Enhanced Language Representation with Informative Entities)
- 库附加层及调用

库附加层

- 类似 C 语言有很多小型库，用于实现部分功能
- LLM 的库（嵌入库）指的是经过微调后的特训库，例如医学 gpt 等等
- 我们希望实现库的可叠加可拆卸
- 特定知识层，知识目录页，特定附加输入
- 库附加层及调用

库附加层

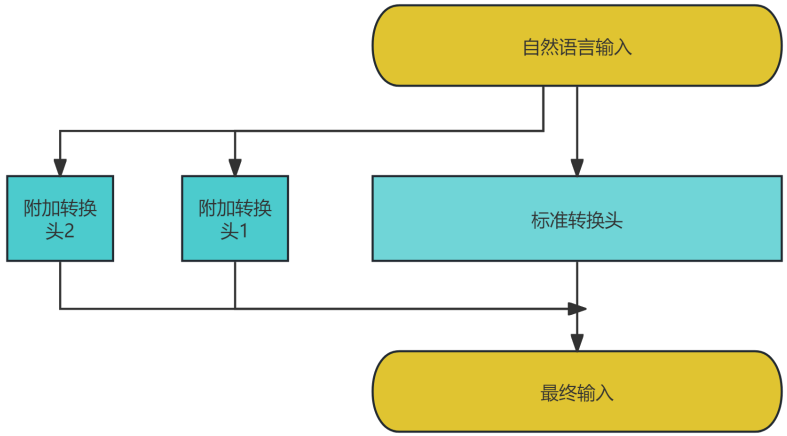


图 3: Enter Caption

库附加层需要训练的内容

- 训练模型到一个指定位置读取附加信息列表（这与训练模型读取磁盘信息等价，可以通过 API 训练实现）
- 训练模型的标准转换头，用于将自然语言转换为矩阵输入
- 在标准转换头的辅助下，训练针对于特定语言库的附加转换头
- （可选）如果无法简单地叠加，可能需要在最终输入前再增加一个整合器。整合器的训练应该保证任意的附近转换头与标准转换头都可自由拼加

库附加层

- 部分内嵌库研究
- (Knowledge Transfer from High-Resource to Low-Resource Programming Languages for Code LLMs)
- SEEKING NEURAL NUGGETS–KNOWLEDGE

Thanks!