课题组细分方向介绍与基础了解-2025

针对我们现在主要研究的两个方向：AI for System 和 System for AI，我们整理了一个细分方向介绍与基础了解。主要包含两个方面：

1. 相关论文阅读。

首先，我们在两个方向里面各选择了几篇相关的论文，你可以任选其中一个方向的其中一篇论文阅读，以此来了解我们的研究方向。论文阅读完成后，请你用你喜欢的方式告诉我：

- ✧ 研究背景是什么？
- ✧ 解决的问题是什么？
- ✧ 如何解决的问题？
- ✧ 针对这篇论文你认为可改进的地方和改进方案是什么？

2. 动手尝试

我们针对不同方向需要用到的能力，整理了一个动手尝试，“Keep your hands dirty”对于我们计算机的同学来说是很重要的。你可以借此体会一下相关的技术栈是否熟悉，当然最重要的是兴趣，技术栈不会的可以未来我们再和大家一起学习。不同方向的动手尝试的要求不同，详情见下面标准。同样，你也可以用你喜欢的方式告诉我你在动手尝试中做了什么、有什么心得和体会☺

根据自己兴趣可选择如下方向之一进行测试。

■ AI for System in IoT

1. 相关阅读列表

[1] MobiCom’25 - AutoIOT: LLM-Driven Automated Natural Language Programming for AIoT Applications. (AutoIOT: LLM-Driven Automated Natural Language Programming for AIoT Applications (polyu.edu.hk))

[2] SIGCOMM’24 - NetLLM: Adapting Large Language Models for Networking. (NetLLM: Adapting Large Language Models for Networking (acm.org))

2. 动手尝试

- ➢ 简要了解 WebAssembly。
- ➢ 将 PolyBench/C 的代码编译为 WebAssembly 代码，并使用至少 3 种 WebAssembly 运行时（Wasmer，Wasmtime，Wasmedge，WAVM，WAMR，Lucet，Wasm3 等）运行编译出的 WebAssembly。
- ➢ 运行并测试 WebAssembly 代码的执行时间，并将其与原生编译（也即直接将 PolyBench/C 用 gcc 编译为可执行文件）。
- ➢ 任何你想到的相关实现，都可以进一步尝试。

■ System for AI in IoT

1. 相关论文阅读

[1] ACL’25 - MobiLoRA: Accelerating LoRA-based LLM Inference on Mobile Devices via Context-aware KV Cache Optimization.
(https://www.liborui.cn/publication/27-acl25-mobilora/27-ACL25-MobiLoRA.pdf)

[2] DAC'25 - InfScaler: Enabling Efficient ML Inference Serving on Multi-Accelerator Edge Devices via Asymmetric Auto-Scaling. (https://www.liborui.cn/publication/23-dac25-infscaler/23-DAC25-InfScaler.pdf)

2. 动手尝试

（可选 1）

- 借助 Pytorch 实现一个 Transformer。
- 阅读 llama.cpp 的文档，尝试在本地（PC，手机均可）部署基于 LoRA 的 llama2-7b 模型（可以选择任何量化版本）。

（可选 2）

- 了解 Serverless 定义与架构。
- 使用 Knative/OpenFaaS 框架实现一个简单的 Serverless 应用案例。