大模型推理基础设施

研究生课程导论

张书豪

华中科技大学计算机学院

课程定位

核心结论

本课程面向研究生系统讲授大模型推理基础设施的关键问题、核心机制与研究方法。

- 关注在线服务系统，而非训练过程。
- 关注调度、状态管理、执行路径与异构适配。
- 结合论文精读、系统案例、实验训练与课程项目。

为什么值得独立成课

核心结论

大模型推理已经从“把模型跑起来”演进为“把服务稳定运行起来”的系统问题。

- 长上下文、MoE、RAG、智能体工作流重塑了工作负载。
- TTFT、P99、goodput、成本与稳定性成为核心目标。
- 真实收益取决于全链路而非局部算子。

统一视角：状态化系统

核心结论

推理系统的主要瓶颈，往往来自共享状态的访问、保留、搬运和复用。

- 模型参数之外，还有 KV Cache、知识缓存、会话记忆与调度元数据。
- 状态跨时间、跨请求、跨设备持续传播。
- 调度、缓存与执行路径因此连续耦合。

课程主线

核心结论

课程围绕三条主线展开：请求组织与调度、状态管理与复用、执行路径优化与验证。

1. 工作负载与评价指标
2. 推理执行路径与系统架构
3. 调度机制与 KV 管理
4. 异构平台适配与论文精读
5. 开源系统实践与课程项目

课程模块

- 模块 1：导论与问题地图
- 模块 2：工作负载与评价指标
- 模块 3：一次请求的完整链路
- 模块 4：请求调度与服务编排
- 模块 5：KV Cache 与状态管理
- 模块 6：推理系统架构
- 模块 7：执行优化与异构路径
- 模块 8：论文精读与研究训练

课程能力目标

核心结论

课程目标不是记住名词，而是形成系统研究与实验验证能力。

- 能分解一次推理请求的生命周期。
- 能分析瓶颈对象、控制点与收益链路。
- 能阅读系统代码并设计可信实验。
- 能提出具备实现路径的研究选题。

课程方法

核心结论

机制讲解、论文精读、代码导读、实验训练和课程项目共同构成完整教学闭环。

- 不把论文讲成摘要复述。
- 不把系统课讲成纯部署教程。
- 强调问题定义、机制边界与实验可信性。

课程总结

核心结论

真正的大模型推理基础设施研究，是让系统在复杂负载、持续状态和异构约束下稳定地跑得对、跑得快、跑得住。

- 这是研究问题，也是工程问题。
- 这也是本课程希望训练的核心能力。