ECNU

开题报告答辩

面向机器人端侧大模型的量化技术性能评估及工具实现

姓名：姚凯文
学号：10224507041
日期：2026年1月8日

研究背景与意义

- 背景：
  - 大语言模型（LLM）在多任务、多场景下表现优异。
  - 机器人探索将语言模型结合高层控制，提升交互与任务适应性。
- 意义：
  - LLM部署面临硬件算力和能耗限制。
  - 探讨量化技术降低模型负担，保障实时性和可靠性。

研究问题与目标

- 核心研究问题：
  1. LLM不同量化策略对机器人任务性能的影响？
  2. 如何构建可复现的量化部署流程？
  3. 针对机器人场景的量化优化经验？
- 预期目标：
  - 可复现量化部署。
  - 构建机器人任务测试平台。
  - 提出优化建议，降低延迟和消耗。

研究方法与技术路线

- 方法：
  1. 文献综述：总结量化方法和LLM机器人应用。
  2. 任务建模：设计典型机器人语言任务场景。
  3. 验证与优化：实验分析性能与工具工程性。
- 技术路线：
  - 阶段推进，优先可复现性。
  - 工具链：Python、PyTorch、Transformers等。

研究计划及进度

- 第1-3周：文献调研，确定方向。
- 第4-6周：搭建实验环境，选择模型数据。
- 第7-10周：模型量化与性能测试。
- 第11-13周：结果分析与优化。
- 第14-16周：完成论文与答辩。

创新点与预期贡献

- 创新点：
  - 从机器人场景出发，评估量化效果。
  - 提出面向资源受限平台的优化方案。
- 贡献：
  - 工具与实验基准，助力后续研究。
  - 工程化建议，应用于机器人场景。

致谢与提问

- 致谢：
  - 感谢导师、同学与研究团队的帮助与指导。
- 期待您的问题与建议。

THANKS