页1：标题页

尊敬的各位老师，同学们，大家好！
我是姚凯文，今天我将向大家汇报我的毕业设计开题报告。我的论文题目是“面向机器人端侧大模型的量化技术性能评估及工具实现”。感谢各位在百忙之中拨冗参加，接下来我将围绕研究背景、核心问题、研究计划及预期贡献等方面展开介绍。

页2：研究背景与意义

近年来，大语言模型凭借其卓越的语义表达与泛化能力，已经成为诸多领域的核心技术。然而，这些模型的高算力、高能耗特性，使得它们在机器人中部署面临挑战。
相比传统NLP任务，机器人对指令解析的实时性和动作生成的鲁棒性要求更高，而机器人设备的资源有限，难以直接承载原始的大模型运行负荷。
因此，探索如何通过模型量化技术降低资源需求，同时确保任务性能，是一个重要且富有工程意义的方向。本课题旨在研究主流量化技术在机器人任务上的适用性，提出优化方案并实现工程化工具支持。

页3：研究问题与目标

针对以上背景，本课题着眼于以下三个核心问题：

1. 不同量化方法对机器人任务性能会产生怎样的影响？
2. 如何构建一个面向受限环境的量化与部署流程，以实现性能和效率的良好平衡？
3. 是否可提炼出针对机器人场景的通用量化优化经验？

因此，我的研究目标包括：复现与比较多种量化方法，搭建机器人任务实验平台，评估量化策略的工程成本和效果，并通过具体实验提出优化建议。

页4：研究方法与技术路线

本课题采用以下方法开展研究：

- 首先，通过文献调研系统梳理主流量化策略和LLM的机器人应用，明确量化方法的基本原理和适用场景。
- 接着，设计若干典型机器人语言任务作为主要实验场景，例如指令理解、动作生成和任务规划。
- 然后，复现多种量化方法并进行对比实验，评估其在模型性能、资源消耗和实际部署中的表现。
- 最终，综合分析实验数据，从算法、硬件和部署的维度提出优化策略与实践总结。

我的研究强调阶段化推进与可复现性，工具链包括Python、PyTorch、Hugging Face Transformers等，实验将在GPU和边缘设备等资源受限平台上进行。

页5：研究计划及进度

研究计划主要分为五个阶段：

1. 第1-3周，完成文献调研，确定研究方向与实验框架。
2. 第4-6周，搭建实验环境，选择模型和数据。
3. 第7-10周，进行量化实验，测试性能指标。

4. 第11-13周，分析实验结果，优化方案与验证。
5. 第14-16周，完成论文撰写与答辩准备。

通过这五个阶段，我计划逐步推进研究，确保成果的系统性和完整性。

页6：创新点与预期贡献

我的研究有以下几方面创新点：

- 关注以机器人为核心的应用场景，系统评估量化方法的工程适用性，这与传统NLP任务有显著不同。
- 从工程落地的角度优化量化部署流程，提出具体的性能与资源权衡解决方案。

预期贡献包括：

- 搭建一套开放且可复现的实验框架，为相关领域研究提供基准。
- 提出针对机器人平台的工程化部署建议，助力高效应用LLM于移动机器人等设备。

页7：致谢与提问

最后，特别感谢我的导师对本研究方向的悉心指导，也感谢研究团队和同学们给予的支持与帮助。
以上是我的开题报告，请各位老师提出宝贵意见，同时也欢迎大家的提问！