华东师范大学本科生毕业论文（设计）开题报告

<table>
<tr><td rowspan="2">论文题目</td><td rowspan="2">面向机器人端侧大模型的量化技术性能评估及工具实现</td><td colspan="2">数据科学与工程学院</td><td colspan="2">数据科学与大数据技术专业</td></tr>
<tr><td>学生姓名</td><td>姚凯文</td><td>学号</td><td>10224507041</td></tr>
</table>

一、选题的背景与意义

近年大语言模型以其卓越的语义表征与生成能力，已在自然语言理解、机器翻译、对话系统、代码生成及多模态融合等多个方向取得显著进展。这类模型通过大规模自监督预训练学习到的通用语言知识，使得它们在零样本或少样本情形下仍能展现较强的泛化能力。与此同时，机器人学界开始探索将 LLM 的高度抽象推理与语言理解能力，嵌入到机器人系统的高层控制与认知模块中，从而实现更自然的人机交互、更灵活的任务规划以及更鲁棒的指令执行。代表性研究（如将语言模型与机器人能力库或感知模块结合的工作）已展示了“语言→动作”桥接在复杂任务分解、语义指令理解与任务规划中的潜力，表明 LLM 能显著提升机器人在不确定、多变环境下的适应性与任务完成率。

然而，从理论可行性过渡到工程部署仍面临多重挑战。主流 LLM 通常具有数十亿至上千亿的模型参数，导致其推理阶段对算力、内存（显存）与能耗的需求极高。相比之下，多数移动机器人与嵌入式控制单元在计算资源与电源供给方面受限，无法直接承载原始大模型的运行负荷。此外，机器人应用对实时性与可靠性的需求远高于传统离线 NLP 任务：延迟、抖动或偶发的语义误判均可能导致任务失败甚至安全风险。因此，如何在严格的工程约束下保持模型在指令解析、动作生成与环境问答等任务上的有效性，成为落地应用的关键瓶颈。

量化技术作为一种降低模型表示精度以减少存储与计算开销的通用手段，因其实现成本相对低、工程性强而成为工业界与研究界的首选方案之一。近年来出现的多种量化方法（包括后训练量化、量化感知训练、混合精度及层/通道级别的自适应量化策略）在理论与实践上均表现出各自特点：一些方法注重最小化权重逼近误差，另一些方法通过激活校准或微调以修复量化带来的性能损失。但需要注意的是，不同方法在不同模型架构、不同任务类型以及不同硬件后端上的表现存在显著差异；此外，量化引入的数值误差在序列生成、语义稳定性以及任务依赖的后处理上可能呈现非线性影响，这在机器人任务中特别值得关注。

当前研究主体多集中在 NLP 通用基准上评估量化效果，而针对机器人任务的系统性研究相对稀缺。机器人任务通常具有多模态感知输入、状态-动作闭环、以及对错误类型（例如指令歧义、动作顺序错误）的敏感性，这些特性可能放大量化带来的影响。因此，有必要从工程视角出发，对量化方法在典型机器人场景中的适用性、鲁棒性与工程成本进行系统评估，明确在实际部署中应采取的折中策略与操作规范。

本课题从理论与工程双重视角切入：一方面对主流量化方法进行系统考察，分析其适用条件与数值机制；另一方面构建可复现的实验评估平台，将量化方法在典型机器人任务上进行对比验证与工程化测试。通过这一工作，期望达到两类具体贡献：一是面向机器人端侧场景对大语言模型量化技术进行系统性的性能评估，构建统一的评价协议与实验基准，并通过对比实验分析不同量化方法在典型机器人任务上的行为差异及其产生原因；二是实现一套工程化的量化—部署工具与文档，以便在受限资源的机器人平台上快速复现与验证结果。综上，本课题既有重要的理论研究价值，也具备明确的工程应用意义，能够为后续将 LLM 广泛部署于移动机器人、服务机器人及边缘智能设备提供实践参考与方法指导

二、研究的主要内容和预期目标

（一）主要研究内容

本课题的核心目标是系统研究不同 LLM 量化方法在机器人任务中的表现及其工程可行性，回答如下关键问题：

- 不同量化策略（后训练量化、量化感知训练、混合量化等）在机器人任务上对推理行为与任务性能的影响有何差异？
- 在资源受限的机器人工程环境中，如何构建一条可复现的量化与部署流程，以便权衡性能、延迟与资源消耗？
- 是否存在针对机器人任务的量化–部署建议或通用经验（例如对特定层分配不同精度、校准策略、微调步骤）？

基于上述问题，研究将围绕实现可复现的量化实现、构建评估平台并开展对比实验来展开，最终形成一套工程可行的技术路线与经验总结。

（二）预期目标

1. 实现多种量化算法在主流 LLM（如 LLaMA 或 Mistral）上的可复现部署；
2. 构建一个基于量化 LLM 的机器人任务实验平台；
3. 系统评估不同量化策略在机器人任务中的性能变化；
4. 提出适用于机器人平台的 LLM 量化优化方案，实现延迟降低与资源节约；
5. 形成一份具有实际工程价值的研究报告，为后续 LLM 在机器人中的部署提供参考。

三、拟采用的研究方法、步骤

（一）研究方法

1. 文献综述与方法学整理
    1. 系统梳理大语言模型的量化方法，包括：权重量化、激活量化、混合量化、后训练量化与量化感知训练。
    2. 总结各类方法的数学原理、实现策略、优劣与适用场景。
    3. 调研 LLM 在机器人领域的应用研究，识别任务特性对量化敏感性的潜在影响。
2. 机器人任务建模与实验平台设计
    1. 设计若干典型机器人语言任务作为实验场景（例如：自然语言指令理解与动作序列生成、关于环境状态的问答、序列任务规划与矫正）。
    2. 构建可复现的实验平台，包括模型推理接口、任务评估脚本、日志采集与可视化模块，确保对比实验的数据与运行环境一致。
3. 量化方法实现与工程化部署
    1. 实现并复现主流量化方法在目标模型上的流水线（包括模型转换、校准、导出与推理接口）。
    2. 维护所有实验脚本与配置，使其可通过版本控制复现。
4. 实验评估与结果分析
    1. 在保持实验可复现性的前提下，比较量化前后模型在任务精度、推理行为、资源消耗与工程稳定性方面的差异。
    2. 采用适当的统计分析方法评价结果稳定性，并通过可视化呈现性能—精度的权衡。
    3. 在分析中重点考察量化对机器人任务关键失败模式的影响（例如指令解析错误、动作序列不一致等），并提出工程化的改进建议（如保留关键层较高精度、结合微调策略等）。

（二）研究步骤

总体技术路线

1. 阶段化推进：从小规模验证到完整实验，先实现量化流程与评估脚本，再在若干任务上开展规模化对比，最后总结工程实践建议。
2. 可复现优先：所有代码、配置信息、校准与测试数据的生成脚本、运行日志与随机种子将按统一规范保存，并拟定发布计划以便复现。

3. 软硬件与工具链（示例性说明，具体版本在实验记录中注明）
4. 开发语言与框架：Python、PyTorch、Hugging Face Transformers 等主流工具链。
5. 量化实现与工具：参考并使用当前主流的量化实现与库（如 GPTQ 类实现、AWQ/SmoothQuant 思路、bitsandbytes 等社区工具），并在实验中注明具体实现来源与版本。
6. 推理与部署：采用常见的推理方式（如 Transformers 的推理流水线、ONNX/TensorRT 或框架原生加载方式），并对比不同推理后端的工程性差异。
7. 硬件平台：实验将覆盖具有代表性的桌面/服务器 GPU 环境与可得的边缘或嵌入式平台（在报告中给出设备类别与获得渠道说明），若可得硬件有限，将给出替代性评估手段（如受限资源的仿真或容器化内存限制测试）。

实验流程

1. 环境准备：记录 Python 版本、依赖项与库版本。
2. 数据与任务准备：设计并生成任务校准集、开发集与测试集，注明生成规则与随机种子。校准集用于 PTQ 的 activation calibration，测试集用于性能评估。
3. 基线测量：在原始（未量化）模型上运行基线测试，记录任务行为与资源占用，作为后续比较基线。
4. 量化流程：对选定模型分别应用不同量化方案，记录每种方法的步骤、参数与任何额外处理（如 layer-wise tuning、校准样本大小）。
5. 部署与测试：将量化后模型加载到统一推理框架下运行任务测试，记录推理延迟分布、内存占用、模型加载时间、任务评价指标与错误案例。
6. 重复与统计：多次重复关键实验以衡量结果波动，求取均值与置信区间；对关键比较采用配对统计检验以验证差异显著性。
7. 分析与可视化：绘制性能—精度曲线、资源消耗对比柱状图、错误类型分布等，分析量化方法的实际工程影响。

可复现性与开放性规范

1. 代码管理：所有实现与实验脚本放在版本控制系统中，并附带 README 与运行示例。
2. 实验记录：每次实验保存完整运行日志，重要结果以机器可读格式保存。
3. 结果复核：在论文中附上必要的复现说明与最小可复现示例，便于审核者或后续研究者复现关键结论。
4. 预期成果与论文结构
5. 实验报告：记录量化方法的实现细节、运行日志与比较结果，包含图表与讨论。
6. 开发产出：可复现的实验代码、配置与数据生成脚本。
7. 学术论文：按照本科论文格式整理研究背景、方法、实验、结果与结论。
8. 工程建议书（附录）：总结在机器人平台上部署 LLM 的工程注意事项与实用建议，供后续工程化移植参考。

四、研究的总体安排与进度

本课题的研究工作计划分为五个阶段，主要安排如下：

第一阶段（第 1—3 周）：文献调研与方案设计
查阅大语言模型量化及其在机器人应用中的相关研究，确定研究方向、实验模型、量化方法与性能指标，完成总体方案设计。

第二阶段（第 4—6 周）：实验环境与模型准备
搭建实验环境，配置量化工具和深度学习框架，选择合适的语言模型并准备机器人应用任务数据。

第三阶段（第 7—10 周）：模型量化与性能测试
对模型进行多种精度下的量化实验，测试在机器人语言任务中的推理速度、精度和资源占用等性能指标。

第四阶段（第 11—13 周）：结果分析与优化
整理实验数据，分析不同量化策略对性能的影响，尝试结合优化方法提升模型推理效率。

第五阶段（第 14—16 周）：论文撰写与答辩准备

完成论文撰写、修改与答辩准备，整理实验成果并形成最终论文。

五、参考文献

[1] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth and S. Han, "SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models," arXiv preprint arXiv:2211.10438, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.10438

[2] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan and S. Han, "AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration," arXiv preprint arXiv:2306.00978, Jun. 2023. [Online]. Available: https://arxiv.org/abs/2306.00978

[3] J. Frantar and T. Dettmers, "GPTQ: Accurate Post-Training Quantization for Generative Pretrained Transformers," arXiv preprint, 2023. [Online]. Available: https://github.com/IST-DASLab/gptq (implementation repository)

[4] T. Dettmers, "bitsandbytes: 8-bit optimizers and quantization tools," GitHub repository, 2022. [Online]. Available: https://github.com/TimDettmers/bitsandbytes

[5] T. Dettmers, L. Shleifer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint, 2023. [Online]. Available: https://github.com/artidoro/qlora

[6] A. Zeng et al., "PaLM-E: An Embodied Multimodal Language Model for Robotic Reasoning and Control," arXiv preprint, 2022. [Online]. Available: https://arxiv.org/abs/2212.08086

[7] M. Ahn et al., "SayCan: Connecting Language Models to Robot Capabilities for Task Execution," in Proc. Robotics: Science and Systems (RSS), 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[8] J. Lee, S. Park, J. Kwon, J. Oh and Y. Kwon, "A Comprehensive Evaluation of Quantized Instruction-Tuned Large Language Models: An Experimental Analysis up to 405B," arXiv preprint arXiv:2409.11055, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.11055

[9] Hugging Face, "Transformers: State-of-the-art Natural Language Processing for PyTorch and TensorFlow," GitHub repository. [Online]. Available: https://github.com/huggingface/transformers

[10] PyTorch Contributors, "PyTorch," 2024. [Online]. Available: https://pytorch.org

<table>
<tr><td rowspan="2">论文题目</td><td rowspan="2">面向机器人端侧大模型的量化技术性能评估及工具实现</td><td colspan="2">数据科学与工程学院</td><td colspan="2">数据科学与大数据技术专业</td></tr>
<tr><td>学生姓名</td><td>姚凯文</td><td>学号</td><td>10224507041</td></tr>
</table>

六、指导教师意见

本课题既有重要的理论研究价值，也具备明确的工程应用意义，能够为后续将 LLM 广泛部署于移动机器人、服务机器人及边缘智能设备提供实践参考与方法指导。同意开题。

	签字：		年	月	日	

七、开题答辩小组意见

	小组成员签字：		年	月	日	