**南京航空航天大学**

**本科毕业设计（论文）任务书**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学 院 | 计算机科学与技术学院/软件学院 | | | | |
| 专 业 | 物联网工程 | | | | |
| 题 目 | 大模型KVCache优化方法研究 | | | | |
| 学生姓名 | 曾旭阳 | 学号 | | 162140212 | |
| 起迄日期 | 2024年12月16日至2025年5月18日 | | | | |
| 毕设地点 | 南航计算机学院 | | | | |
| 指导教师 | 蔡淼 | | 职称 | |  |

2024年12月31日

|  |
| --- |
| 毕业设计（论文）课题任务书  学生姓名 曾旭阳 学号 162140212 学院 计算机科学与技术学院/软件学院  题 目 大模型KVCache优化方法研究  题目来源: 其它 题目性质: 工程设计类-技术研究  类 别: 🌕毕业设计 🌕毕业论文  团 队: |
| 1. **内容及要求：**   调研目前大模型KVCache的优化方法，提出并设计KVCache卸载机制，缓解GPU缓存压力 |
| 1. **主要指标：**   短序列文本推理延迟在1s内 |
| 1. **进度安排：**   2024.12-2025.1，完成文献调研，方案制定  2025.2-2025.3，完成方案实现、代码编写和实验测试  2025.3-2025.4，完成毕业论文撰写  2025.5，准备毕业论文答辩 |
| 1. **参考文献：**   PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU, SOSP 2024 |
| **指导教师签名：**  蔡淼 |
| **系审查意见：** |
| 同意 |
| **主任：**  2025年1月3日 |

说明： 本科毕业设计（论文）任务书由指导教师根据课题的具体情况填写，经系审查和主任签字后

生效。此任务书在毕业设计（论文）开始前一周内填写并下达至学生。