

同行专家业内评价意见书编号：20250854479

附件1

浙江工程师学院（浙江大学工程师学院） 同行专家业内评价意见书

姓名：郑明辉

学号：22260470

申报工程师职称专业类别（领域）：电子信息

浙江工程师学院（浙江大学工程师学院）制

2025年05月31日

填表说明

一、本报告中相关的技术或数据如涉及知识产权保护、军工项目保密等内容，请作脱密处理。

二、请用宋体小四字号撰写本报告，可另行附页或增加页数，A4纸双面打印。

三、表中所涉及的签名都必须用蓝、黑色墨水笔，亲笔签名或签字章，不可以打印代替。

四、同行专家业内评价意见书编号由工程师学院填写，编号规则为：年份4位+申报工程师职称专业类别(领域)4位+流水号3位，共11位。

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1. 对本专业基础理论知识和专业技术知识掌握情况(不少于200字)

作为计算机科学技术专业人工智能方向的研究者，我系统构建了扎实的理论与技术知识体系。在基础理论层面，深入掌握概率统计、线性代数、优化算法等数学基础，能够熟练运用贝叶斯推断、马尔可夫链蒙特卡洛等概率方法解决不确定性建模问题，掌握矩阵分解、特征值分析等线性代数工具在神经网络中的应用。精通机器学习、深度学习等核心理论，包括监督学习、无监督学习、强化学习三大范式，对梯度下降、反向传播等优化算法有深入理解，尤其对Transformer架构、注意力机制、自监督学习等大模型底层技术原理有深刻理解，能够分析不同注意力变体（如稀疏注意力、线性注意力）的计算复杂度和适用场景。

在专业实践中，熟练运用PyTorch、TensorFlow等主流深度学习框架，具备从模型设计到部署的全流程开发能力。掌握分布式训练技术，包括数据并行、模型并行及混合并行策略，能够优化多机多卡环境下的通信效率。深入理解模型量化方法（如INT8/FP16量化、动态量化），熟悉知识蒸馏技术（如基于logits匹配、特征匹配的蒸馏策略），并成功应用于模型压缩任务。此外，具备丰富的模型调优经验，包括超参数搜索、损失函数设计、正则化策略选择等，能够针对不同任务场景定制高效的训练方案。这些理论知识与工程实践的结合，使我能够快速定位和解决复杂AI系统中的技术难题。

2. 工程实践的经历(不少于200字)

在vivo通讯有限公司担任NLP算法工程师期间，我深度参与了多项核心技术研发。针对代码翻译垂直领域优化，基于DeepSeek基座模型开展迁移学习研究，通过GPT4自动生成跨语言对齐语料，设计迭代式数据增强策略（GPT3.5重写编译失败样本），构建了包含20万对Python-Java的领域数据集。采用动态课程学习策略进行SFT微调，在HumanEval-X基准测试中实现pass@5指标4%的显著提升，验证了数据质量优化对代码语义保持的有效性。在端侧1B大模型优化方向，系统研究了RoPE、ALiBi等位置编码机制，创新性地将StreamingLLM的窗口注意力机制与动态缓存策略相结合，使对话上下文长度扩展至8K tokens时推理速度提升30%。针对模型小型化需求，深入分析LLM-QAT量化感知训练方法，在Blue-LLM平台完成INT8量化全流程验证，通过梯度补偿算法将文本摘要任务的性能损失控制在5%以内，形成可复用的轻量化部署方案。这些实践深化了我对Transformer架构优化、跨模态迁移学习及边缘计算优化的系统性认知，积累了从算法设计到工程落地的完整闭环经验。

3. 在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字）

一、端侧大模型长文本处理技术体系构建

在移动端大模型架构优化中，针对长文本外推与多轮对话的核心挑战，构建了融合位置编码创新与动态计算优化的技术框架。系统研究旋转位置嵌入（RoPE）、线性偏置注意力（ALiBi）等编码机制的内在特性，深入分析不同位置编码方案在短文本语义捕捉与长序列扩展能力上的权衡关系。基于Transformer架构的注意力计算模式，提出分层位置编码融合策略：在底层网络保留旋转位置编码对局部语义的精确建模能力，在高层网络引入可学习的线性衰减因子，渐进式增强模型对超长上下文的适应力。为解决多轮对话中的位置漂移问题，开发跨轮次位置索引同步模块，通过对话状态机记录历史位置偏移量，确保连续交互时的位置编码连贯性。

针对移动端内存限制，创新设计动态稀疏注意力计算范式。结合滑动窗口机制与关键token保留策略，构建基于注意力权重的缓存淘汰算法：实时监测各注意力头的激活分布，动态保留高信息熵的上下文片段，在内存阈值内实现最大化的语义完整性保持。为提升计算效率，研究混合精度KV缓存存储方案，将关键向量保留全精度特征，非关键区域采用量化压缩，平衡计算精度与资源消耗。

二、模型轻量化与推理加速技术深度探索

面向端侧部署的严苛约束，建立覆盖模型压缩、量化训练、编译优化的三维技术体系。在参数稀疏化方向，研究结构化剪枝与动态稀疏训练的协同机制：提出基于二阶导数敏感度的参数重要性评估方法，结合梯度引导的渐进式剪枝策略，实现模型骨架的精准压缩；开发运行时自适应稀疏引擎，根据设备算力状态动态调整计算密度，构建“静态架构优化+动态资源适配”的双重压缩范式。

量化感知训练领域，针对大模型低比特表征失真问题，构建分层补偿训练框架：在嵌入层采用知识蒸馏引导的表征对齐，在注意力模块设计残差量化校准机制，在输出层引入对抗训练增强分布鲁棒性。重点改进LLM-

QAT方法的梯度传播路径，提出量化误差反向传播修正算法，通过构建虚拟量化算子实现梯度流的精准控制。同步研究混合精度量化策略，对关键模块保留高精度计算，辅助结构采用激进量化，形成精度与效率的最优平衡。

在编译优化层面，深度定制计算图编译流程：针对移动端异构计算单元，开发算子自动切分与调度引擎，实现NPU/DSP/CPU的协同计算；重构注意力计算内核，设计缓存友好的分块计算策略，优化片上内存复用效率；研究量化算子融合规则，将线性层与量化操作合并为单一计算单元，减少中间结果传输开销。

三、跨技术域协同创新实践

在技术融合创新层面，探索将硬件特性与算法设计的深度结合：研究移动端NPU的矩阵加速指令集，定制化开发稀疏矩阵乘法的硬件加速内核；利用DSP的定点计算优势，设计面向量化模型的低功耗计算流水线；开发边缘设备集群协同推理框架，通过模型分片与流水线并行提升多设备协作效率。

同步推进软件栈生态建设：构建自动化模型转换工具链，支持PyTorch到TensorFlow Lite的一键式格式转换；开发移动端推理引擎插件体系，实现算子库的模块化扩展；建立模型安全加固机制，集成权重混淆、计算图加密等隐私保护技术。

通过上述技术体系的系统化建设，形成覆盖算法创新、工程实现、系统优化的完整能力矩阵，为移动端大模型的落地应用构建坚实技术基础。这些实践不仅深化了对Transformer架构、模型压缩原理、边缘计算等核心技术的理解，更培养了从理论到工程的全链条问题解决能力。

(二)取得的业绩(代表作)【限填3项,须提交证明原件(包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等)供核实,并提供复印件一份】

1. 公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利(含发明专利申请)、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
一种基于大模型的港口知识摘要生成方法	发明专利申请	2024年12月29日	申请号: 202411879818.5	1/3	

2. 其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

(三) 在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 81 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年(要求1年及以上) 考核成绩： 87 分
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！	
申报人签名： 郑明祥	

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☒优秀 ☐良好 ☐合格 ☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）： 2024年5月31日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过 ☐不通过（具体原因： ） 工程师学院教学管理办公室审核签字（公章）： 年 月 日

浙江大学研究生院
攻读硕士学位研究生成绩表

学号：22260470		姓名：郑明辉		性别：男		学院：工程师学院			专业：计算机技术				学制：2.5年		
毕业时最低应获：27.0学分				已获得：29.0学分						入学年月：2022-09			毕业年月：		
学位证书号：						毕业证书号：						授予学位：			
学习时间		课程名称		备注	学分	成绩	课程性质	学习时间		课程名称		备注	学分	成绩	课程性质
2022-2023学年秋季学期		创新设计方法			2.0	通过	专业选修课	2022-2023学年春季学期		港口设施规划与布局			2.0	76	专业学位课
2022-2023学年秋季学期		工程技术创新前沿			1.5	88	专业学位课	2022-2023学年春季学期		新时代中国特色社会主义思想理论与实践			2.0	80	公共学位课
2022-2023学年秋季学期		工程数值分析			2.0	87	专业选修课	2022-2023学年春季学期		研究生论文写作指导			1.0	70	专业学位课
2022-2023学年秋季学期		工程伦理			2.0	79	公共学位课	2022-2023学年春季学期		自然辩证法概论			1.0	70	公共学位课
2022-2023学年冬季学期		机器视觉及其应用			2.0	81	专业学位课	2022-2023学年春季学期		研究生英语基础技能			1.0	79	公共学位课
2022-2023学年冬季学期		海洋管理学			2.0	90	专业选修课	2022-2023学年春夏学期		高阶工程认知实践			3.0	82	专业学位课
2022-2023学年冬季学期		产业技术发展前沿			1.5	86	专业学位课	2022-2023学年夏季学期		数据科学前沿技术导论			2.0	84	专业学位课
2022-2023学年秋冬学期		研究生英语			2.0	88	公共学位课			硕士生读书报告			2.0	通过	

说明：1. 研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2. 备注中“*”表示重修课程。

学院成绩校核章：
成绩校核人：张梦依
打印日期：2025-06-03

浙江大学
(60)
成绩校核章



国家知识产权局

311400

浙江省杭州市富阳区银湖街道富闲路9号银湖创新中心11号第三层
315室 杭州泓呈祥专利代理事务所(普通合伙)
张婵婵(0571-88843902)

发文日:

2024年12月19日



申请号: 202411879818.5

发文序号: 2024121901242770

专利申请受理通知书

根据专利法第28条及其实施细则第43条、第44条的规定,申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日等信息通知如下:

申请号: 2024118798185

申请日: 2024年12月19日

申请人: 浙江大学

发明人: 郑明辉,陈根浪,高云君

发明创造名称: 一种基于大模型的港口知识摘要生成方法

经核实,国家知识产权局确认收到文件如下:

权利要求书 1份2页,权利要求项数: 8项

说明书 1份9页

说明书附图 1份2页

说明书摘要 1份1页

专利代理委托书 1份2页

发明专利请求书 1份4页

实质审查请求书 文件份数: 1份

提示:

1.申请人收到专利申请受理通知书之后,认为其记载的内容与申请人所提交的相应内容不一致时,可以向国家知识产权局请求更正。

2.申请人收到专利申请受理通知书之后,再向国家知识产权局办理各种手续时,均应当准确、清晰地写明申请号。

审查员: 自动受理

联系电话: 010-62356655

审查部门: 初审及流程管理部



200101
2023.03

纸件申请,回函请寄: 100088 北京市海淀区蓟门桥西土城路6号 国家知识产权局专利局受理处收
电子申请,应当通过专利业务办理系统以电子文件形式提交相关文件。除另有规定外,以纸件等其他形式提交的文件视为未提交。