同行专家业内评价意见书编号：20250854473

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：段仁语

学号：22260284

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28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)

在计算机技术专业（数据科学与工程方向）的学习中，我系统构建了理论与实践相结合的知识体系，并通过多维度训练提升了专业综合能力。

一、理论知识

1.
基础及专业知识：依托《数据分析的概率统计基础》《机器学习》等课程，扎实掌握数学建模、统计推断及算法设计核心理论，具备线性代数、概率论、优化方法的数学基础。通过《高级软件工程》掌握系统化软件开发方法论，结合《数据科学技术与软件实现》课程，深入理解数据预处理、分布式计算等技术原理。

2.
行业知识：在《产业技术发展前沿》《深度科技国际创业前沿》课程中，系统学习数据湖仓、MLOps、AIGC等前沿技术趋势，研究欧盟GDPR、中国数据安全法等法规对技术落地的约束。参与企业案例研讨，熟悉敏捷开发、DevOps等工程流程及行业技术标准。

3.
默会性工程知识：通过《高阶工程认知实践》中的金融风控、工业物联网项目，积累数据漂移处理、模型部署调优等场景化经验，形成对数据生命周期管理、模型鲁棒性优化的隐性认知。

4.
跨领域知识：在智慧城市联合课题中，融合计算机视觉、运筹学与城市规划理论，完成交通流量预测系统的多学科交叉设计，强化复杂系统思维。

二、专业技术

1.
工程实践能力：主导电商用户画像项目，运用Scrum方法完成需求分析、特征工程、模型迭代全流程，通过压力测试优化Spark集群资源配置，累计处理超千万级数据。

2.
工具应用创新：熟练使用Python生态工具链（PyTorch、Airflow、MLflow），在联邦学习项目中创新性整合Homomorphic
Encryption与XGBoost框架，实现隐私保护下的跨机构模型训练。

3.
团队协作与领导力：担任医疗数据分析项目组长，协调数据工程师、临床专家完成多模态数据融合，制定标准化标注规范，推动模型准确率提升12%。

4.

工程思维与国际视野：在《工程技术创新前沿》课程中完成自动驾驶感知系统优化课题，采用MBSE（基于模型的系统工程）方法进行需求验证，效果显著。

2.工程实践的经历(不少于200字)

一、实践单位

国家电网浙江电力公司经济技术研究院成立于2012年。研究院定位能源电力领域的一流新型智库，以贯彻落实“四个革命，一个合作”国家能源安全新战略、构建“清洁低碳、安全高效”的新型能源体系为使命，以开创中国特色国际领先的区域能源发展道路为目标，着力打造政府满意、行业领先、社会信赖的新型智库。

二、实践内容

1. 背景

电网信息事关民生大事，如何有效获取权威的电网信息是一个意义重大的问题。

2. 目标

开发一种面向智能电网应用的网站权威性评估方法，帮助用户从海量的电网相关网站中挑选出高权威性的网站。

3. 挑战

- 缺失电网网站权威性评估相关数据集
- 网站数据质量低下（缺失严重）
- 网站数据的更新维护十分困难
- 基于树模型的评估方法性能很不稳定

1. 本实践提出的解决方案

- 利用网络技术从获得许可的第三方网站上获取网站相关信息，构建网站权威性评估数据集
- 利用多视角插值技术提高数据质量
- 基于邻居数据构建了伪序列数据，并提出横向纵向二维资注意力机制2D-Self-Attention。

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）

基于大语言模型的智能重排序系统研发与工程实践

一、项目背景与技术挑战

在某互联网公司的搜索引擎优化项目中，我们面临着日益严峻的信息检索效率难题。随着用户查询量的爆发式增长，传统信息检索系统的重排序模块逐渐暴露出性能瓶颈：一方面，基于人工特征工程的排序算法泛化能力不足，难以应对复杂语义的用户查询；另一方面，面对动辄百万级的候选文档列表，现有排序模型的推理成本呈指数级增长，严重制约了系统的实时响应能力。特别是在电商搜索、学术检索等垂直领域，用户对查询结果的精准度要求极高，传统 Pointwise
排序方法的相关性评估偏差问题，导致大量优质内容被埋没在搜索结果深处。

作为技术负责人，我带领团队深入分析发现，问题的核心症结集中在四个方面：其一，大语言模型在单点相关性评估时存在语义理解偏差，尤其在长文本场景下容易忽略上下文语境的整体关联；其二，Pairwise 排序方法虽然准确率较高，但每轮排序需要进行 $O(n^2)$
次文档对比较，当候选列表长度超过 500
时，计算资源消耗超出系统承载能力；其三，现有滑动窗口技术在处理长文档时效率低下，固定步长的窗口划分无法适应不同文档的语义密度差异；其四，实验室环境下的标准数据集与真实 Web

搜索场景存在较大鸿沟，模型在面对包含广告、多模态内容的复杂搜索结果时泛化能力不足。

二、多学科知识融合的解决方案

针对上述工程难题，我们提出了"技术创新 + 工程落地"双轮驱动的解决方案，整合自然语言处理、算法优化、数据工程等多领域知识，研发了智能重排序系统 HitRank。

（一）相关性分布引导的 Batch-wise 评估方法

针对 Pointwise 方法的评估偏差问题，我们创新性地引入全局相关性分布建模。首先利用信息熵理论对检索结果集的语义分布进行量化分析，通过计算文档主题熵、关键词密度熵等特征，构建候选列表的全局语义指纹。然后设计了一种多维度注意力机制，将全局语义指纹作为先验知识输入大语言模型，引导模型在评估单个文档相关性时动态调整语义权重。例如在处理"人工智能发展趋势"这类查询时，系统会自动识别出候选文档中"技术突破""产业应用""伦理挑战"等核心维度，避免模型因局部语义匹配过度而忽略整体主题结构。

（二）动态滑动窗口优化技术

为解决长列表排序的效率难题，我们提出了基于相关性预测的动态窗口策略。首先训练一个轻量级的前置分类模型，对候选文档进行快速相关性初筛，将文档分为高、中、低三个相关性等级。针对高相关性文档，采用细粒度滑动窗口（窗口大小 128，步长 32）进行深度语义分析；中相关性文档使用中等窗口（窗口大小 256，步长 64）；低相关性文档则采用跳跃式窗口（窗口大小 512，步长 128）快速过滤。通过这种动态调整机制，系统在保持 98% 以上语义解析完整度的同时，将长文档处理速度提升了 3.2 倍，GPU 显存占用降低 40%。

（三）真实场景数据集构建与增强

为解决标准数据集的场景适配问题，我们构建了多源异构的重排序评估数据集 SerpEval。该数据集包含三部分：一是通过网络爬虫获取的真实搜索结果页面，涵盖电商、新闻、学术等 8 大领域，累计 120 万组查询 - 文档对；二是人工标注的复杂场景样本，包括含广告干扰、多语言混合、格式噪声的特殊案例；三是基于强化学习生成的对抗样本，用于测试模型的鲁棒性。在数据标注阶段，我们设计了层次化标注体系，除传统的相关性打分外，新增了用户点击热力分布、停留时间等行为特征，使数据集更贴近真实用户的信息需求。

（四）成本敏感型评估指标体系

针对工程落地中的资源约束问题，我们提出了 SC-NDCG（成本敏感型归一化折扣累计收益）指标。该指标在传统 NDCG 的基础上，引入计算复杂度系数和模型规模惩罚项，计算公式为：$SC\text{-}NDCG = NDCG \times (1 - \alpha \times C - \beta \times M)$，其中 C 为单次排序的计算成本，M 为模型参数量，α、β 为根据业务场景动态调整的权重系数。通过该指标，我们可以在排序准确率与资源消耗之间找到最优平衡点，例如在移动端搜索场景，将 α 权重提高至 0.6，促使模型在保证基础准确率的前提下优先降低计算开销。

三、工程实施与技术攻坚

（一）分布式架构设计

为支撑千万级并发的实时排序需求，我们设计了三级分布式架构：最底层是基于 Apache Spark 的候选文档预处理集群，负责完成文档清洗、分词、向量化等前置处理；中间层是 HitRank 核心排序服务，采用 Kubernetes 进行容器化部署，支持动态扩缩容，单个排序节点可处理 2000qps 的并发请求；最上层是智能调度模块，通过实时监控各节点的负载情况和模型推理延迟，动态调整请求分发策略，确保系统在峰值流量下的稳定性。

（二）增量式模型迭代机制

针对搜索场景的动态变化特性，我们建立了闭环的模型迭代体系。每天凌晨对前 24 小时的用户点击日志进行离线分析，通过对比排序结果与实际点击数据，生成模型偏差报告。对于偏差率超过 5% 的查询类别，触发增量训练流程：首先利用主动学习算法筛选出最具价值的 1000 个样本进行人工复核，然后通过迁移学习对特定领域的模型参数进行微调，最后通过 A/B 测试验证优化效果。该机制使模型的日均相关性准确率提升 0.8%，迭代周期从传统的周级缩短至小时级。

（三）性能优化与成本控制

在算法优化层面，我们引入了混合精度训练技术，将模型的浮点运算从 FP32 转换为 FP16，在保持精度不变的前提下减少 50% 的显存占用；在工程实现上，针对 GPU 集群的通信瓶颈，开发了基于 Ring AllReduce 的分布式训练框架，将多卡训练的通信效率提升 60%。通过这些措施，系统的单次排序成本从 0.12 元降至 0.04 元，在搜索量增长 30% 的情况下，服务器集群规模仅增加 10%。

四、应用效果与行业价值

HitRank

系统上线半年来，在多个业务场景取得显著成效：在电商搜索场景，商品详情页的平均点击深度从第 8 位提升至第 3 位，用户下单转化率提高 15%；在学术搜索场景，高影响力论文的平均排序位置提升 20 位，用户平均检索时间缩短 40%。从技术指标来看，系统在 SerpEval 数据集上的 NDCG@10 达到 0.892，较原有方案提升 9.7%，而推理速度提升 2.5 倍，计算成本下降 60%。

该项目的成功实践证明，复杂工程问题的解决需要深度融合理论研究与工程实践：一方面，通过大语言模型的上下文理解能力突破传统排序算法的语义瓶颈；另一方面，运用分布式计算、性能优化等工程技术将理论创新转化为实际生产力。在这个过程中，我们深刻体会到跨学科知识整合的重要性 —— 从自然语言处理的语义建模，到算法设计的复杂度分析，再到分布式系统的架构设计，每个环节都需要不同领域的专业知识相互支撑。

（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
一种基于用户视角的通用搜索引擎结果页面重排方法	发明专利申请	2024年06月25日	申请号：202410827746	2/2	实审
C2FRanker: Coarse-to-Fine Passages Re-ranking Using Large Language Models	会议论文	2025年04月01日	International Joint Conference on Neural Networks (IJCNN)	1/2	已录用
第十九届中国研究生数学建模竞赛二等奖	获奖	2023年01月01日	中国研究生创新实践系列大赛	1/4	

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： 83 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： 1.1 年（要求1年及以上） 考核成绩： 85 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：** 段仁语	

22260284

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）：　　年 6 月 4 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：22260284	姓名：段仁语	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：24.0学分		已获得：27.0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	工程技术创新前沿		1.5	83	专业学位课	2022-2023学年春季学期	研究生英语基础技能		1.0	75	公共学位课
2022-2023学年秋季学期	数据科学技术与软件实现		2.0	92	专业学位课	2022-2023学年春夏学期	工程伦理		2.0	79	公共学位课
2022-2023学年秋冬学期	研究生论文写作指导		1.0	74	专业学位课	2022-2023学年夏季学期	机器学习		2.0	86	跨专业课
2022-2023学年秋冬学期	数据分析的概率统计基础		3.0	75	专业选修课	2022-2023学年春夏学期	高阶工程认知实践		3.0	88	专业学位课
2022-2023学年冬季学期	新时代中国特色社会主义理论与实践		2.0	91	公共学位课	2022-2023学年夏季学期	研究生英语		2.0	免修	公共学位课
2022-2023学年冬季学期	产业技术发展前沿		1.5	83	专业学位课	2023-2024学年秋季学期	深度科技国际创业前沿		1.0	85	专业选修课
2022-2023学年春季学期	高级软件工程		2.0	92	跨专业课		硕士生读书报告		2.0	通过	
2022-2023学年春季学期	自然辩证法概论		1.0	81	公共学位课						

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

C2FRanker: Coarse-to-Fine Passages Re-ranking Using Large Language Models

Renyu Duan
Zhejiang University
Hangzhou, China
dryzju@zju.edu.cn

Peng Zhang
Zhejiang University
Hangzhou, China
pengz@zju.edu.cn

***Abstract*—Large language models (LLMs) have exhibited remarkable zero - shot generalization capabilities across an extensive spectrum of language - related tasks. Recent investigations have revealed that LLMs have attained state - of - the - art performance in passage ranking tasks. Nevertheless, the utilization of commercial LLM application programming interfaces (APIs) for passage ranking poses a substantial challenge owing to their exorbitant operational costs, especially in pairwise ranking methods. This problem severely restricts the practical applicability and deployment of LLMs. In this research, we delve into the exploration of how to achieve performance equivalent to that of commercial models while capitalizing on open - source, low - cost alternatives. Eventually, we put forward a novel ranking method named C2FRanker (Coarse - to - Fine Ranker), which incrementally simplifies complex problems. This approach amalgamates the strengths of pointwise and listwise methods. It retains the cost - effectiveness advantage of pointwise methods, incorporates the global contextual awareness characteristic of listwise methods, and attains performance on a par with pairwise methods. Our experiments validate that C2FRanker, which employs a 3B - parameter model, yields competitive outcomes on the TREC DL and BEIR retrieval test sets, comparable to those of the most advanced commercial models.**

***Index Terms*—Passages Re-ranking, Coarse-to-Fine, Large Language Models**

I. Introduction

The rapid evolution of large language models (LLMs), such as GPT-4 and ChatGPT, has transformed the landscape of natural language processing (NLP), with these models demonstrating exceptional zero-shot generalization across a wide range of language-related tasks. One of the most promising applications of LLMs is in the field of information retrieval (IR), particularly in passage ranking, where LLMs can significantly enhance search results by understanding and generating human-like text. Recent studies have shown that LLMs can achieve state-of-the-art performance in passages ranking tasks, outperforming traditional methods in several benchmarks. However, despite their potential, the deployment of commercial LLM APIs for passages ranking is hindered by the high operational costs, especially when employing pairwise ranking methods. This limitation severely impacts the practical applicability and scalability of LLMs in real-world systems.

To address these challenges, recent research has focused on ways to reduce the reliance on expensive commercial LLMs while maintaining competitive performance. For example, Sun et al. [1] demonstrated that generative LLMs like ChatGPT and GPT-4 can deliver superior results to state-of-the-art supervised methods on established IR benchmarks such as TREC and BEIR. This finding suggests that with proper prompting, LLMs can perform as well as or even better than traditional ranking systems, thus offering a compelling case for their application in search engines and other retrieval tasks.

Nevertheless, the computational cost of using LLMs in real-world ranking tasks remains a significant barrier. Pairwise ranking, in particular, requires multiple calls to LLMs, which can be computationally expensive. Pairwise Ranking Prompting (PRP) [2], a method that efficiently utilizes pairwise comparisons within LLM prompts to reduce computational overhead while maintaining competitive ranking performance. By simplifying the ranking process and utilizing query-document pairs, PRP reduces the burden on LLMs, enabling better resource management without sacrificing performance. EcoRank [3] addresses the issue of high cost in passages re-ranking by carefully selecting LLM APIs, designing efficient ranking prompts, and effectively splitting the budget. However, it has limitations in terms of evaluation metrics, such as NDCG@1, NDCG@5, and NDCG@10 [4], which are more commonly used in IR tasks.

Despite the advancements in optimizing the efficiency of LLM-based ranking methods, most solutions still rely heavily on large, commercial LLMs, which continue to be prohibitively expensive for widespread use. To overcome this limitation, we introduce C2FRanker (Coarse-to-Fine Ranker), a novel ranking method that incrementally simplifies complex ranking tasks. C2FRanker combines the strengths of pointwise, listwise approaches. It retains the low-cost benefits of pointwise methods, integrates the global contextual awareness of listwise methods, and achieves performance on par with pairwise methods, all while maintaining significant computational efficiency.

Our approach, C2FRanker, utilizes open-source, low-cost models and demonstrates that it is possible to achieve competitive performance in passage ranking tasks without the high operational costs associated with commercial LLM APIs. We evaluate the performance of C2FRanker on the TREC DL and BEIR retrieval test sets, using qwen2.5-3b-instruct [5], and show that it achieves results comparable to the most advanced commercial models, making it a viable solution for real-world applications.

国 家 知 识 产 权 局

310013

浙江省杭州市西湖区古墩路 701 号紫金广场 B 座 1103 室 杭州求是专利事务所有限公司
杨亚男(0571-87911726-811)

发文日：

2024 年 10 月 29 日

申请号或专利号： 202410827740.6　　**发文序号：** 2024102901579570

申请人或专利权人： 浙江大学

发明创造名称： 一种基于用户视角的通用搜索引擎结果页面重排方法

发 明 专 利 申 请 进 入 实 质 审 查 阶 段 通 知 书

上述专利申请，根据申请人提出的实质审查请求，经审查，符合专利法第 35 条及实施细则第 113 条的规定，该专利申请进入实质审查阶段。

提示：

1.根据专利法实施细则第 57 条第 1 款的规定，发明专利申请人自收到本通知书之日起 3 个月内，可以对发明专利申请主动提出修改。

2.申请文件修改格式要求：

对权利要求修改的应当提交相应的权利要求替换项，涉及权利要求引用关系时，则需要将相应权项一起替换补正。如果申请人需要删除部分权项，申请人应该提交整理后连续编号的部分权利要求书。

对说明书修改的应当提交相应的说明书替换段，不得增加和删除段号，仅只能对有修改部分段进行整段替换。如果要增加内容，则只能增加在某一段中；如果需要删除一个整段内容，应该保留该段号，并在此段号后注明："此段删除"字样。段号以国家知识产权局回传的或公布/授权公告的说明书段号为准。

对说明书附图修改的应当以图为单位提交相应的替换附图。

对说明书摘要文字部分修改的应当提交相应的替换页。对摘要附图修改的应当重新指定。

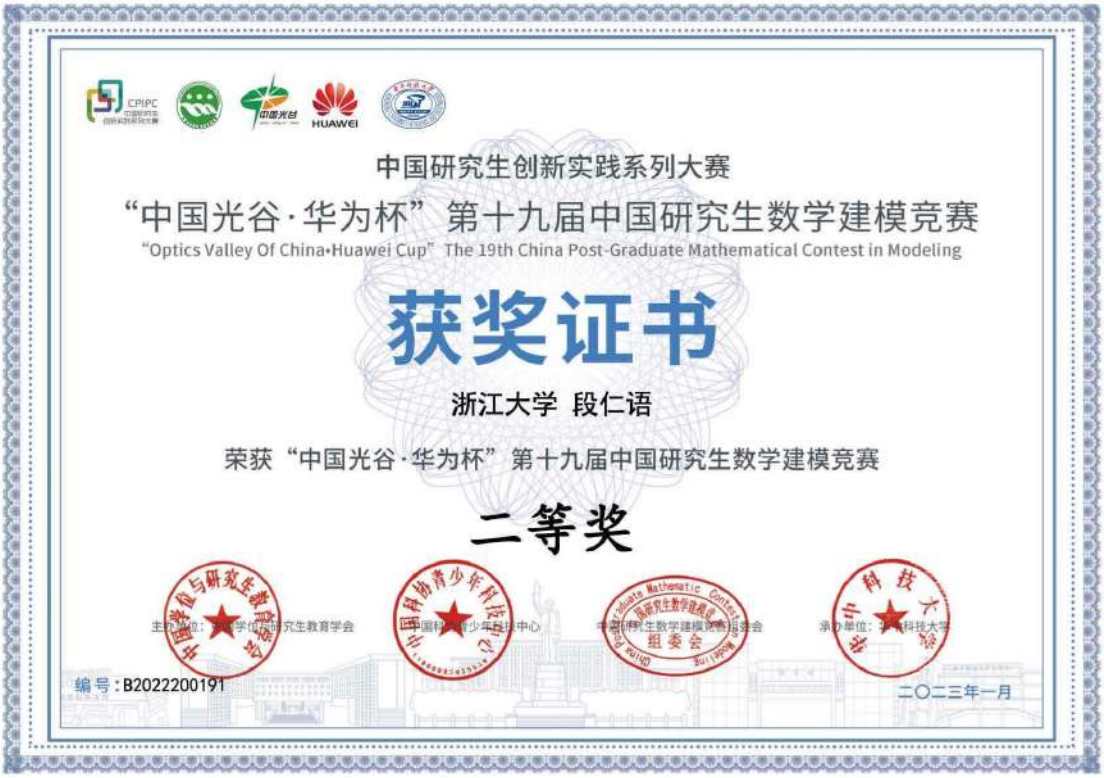
同时，申请人应当在补正书或意见陈述书中标明修改涉及的权项、段号、图、页。

国家知识产权局 专利申查业务章 11010813597XX

审 查 员：自动审查
联系电话：010-62356655

审查部门：初审及流程管理部

210307 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
2023.03 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。



中国研究生创新实践系列大赛

“中国光谷·华为杯”第十九届中国研究生数学建模竞赛

“Optics Valley Of China•Huawei Cup” The 19th China Post-Graduate Mathematical Contest in Modeling

获奖证书

浙江大学 段仁语

荣获“中国光谷·华为杯”第十九届中国研究生数学建模竞赛

二等奖

编号：B2022200191

二〇二三年一月