同行专家业内评价意见书编号：20250854329

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：项卓怡

学号：22260228

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03月17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)

人工智能药学是人工智能（AI）与药学的交叉学科，涉及深度学习、计算机视觉、自然语言处理、计算化学、分子模拟等多个领域。我系统掌握了人工智能的基础理论，包括神经网络、概率图模型、强化学习及大数据分析方法，并深入学习了其在药物研发中的应用，如基于深度学习的分子生成模型、药物-
靶点相互作用预测、虚拟筛选与分子对接，以及药物研发知识图谱构建等。在专业技术方面，我熟练使用TensorFlow、PyTorch等人工智能框架，并掌握生物医药数据的处理技术。通过理论与实践结合，我不仅掌握了人工智能技术的核心方法，还能够将其应用于新药研发，推动制药工业的智能化发展。

2.工程实践的经历(不少于200字)

在工程实践方面，我主要参与并主导了一个面向生化领域的大模型构建工作。该项目的核心目标是利用人工智能技术解决神生化领域的复杂问题。在项目初期，我深入研究了现有的生物医学预训练模型（如BioBERT、PubMedBERT、ChemBERTa），并结合药物化学、分子生物学等领域的特点，制定了针对性的训练策略。我负责数据收集和清洗，整理了来自PubMed、DrugBank、UniProt等多个数据库的大规模药学文本数据，构建了高质量的领域知识库。
在模型构建阶段，我选择qwen1.5-
7b作为基座模型。训练过程中，我优化了超参数调整策略，并结合高性能计算集群加速模型训练。在应用落地方面，我参与了模型的评测和优化，通过实际生化领域相关人物问答验证其效果。最终，该模型成功提升了生化领域复杂任务的处理能力。整个项目的实施不仅加深了我对人工智能药学的理解，也让我在实际工程应用中积累了丰富的经验，为未来生化领域研发提供了重要的技术支持。

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）

在实际工程实践中，我主导并参与了一个面向生化领域的大模型构建工作，该项目的核心目标是利用人工智能技术解决生化研究中的数据整合难、知识获取慢、推理能力有限等复杂问题。生化领域的数据来源广泛，包括研究论文、药物说明书、基因组数据、分子结构信息等，且数据格式各异，存在非结构化、半结构化和结构化数据共存的问题。为了解决这一挑战，我首先整合了PubMed、DrugBank、UniProt等多个权威数据库的生物医学数据，并运用自然语言处理技术进行数据清洗、去重、格式转换和结构化处理，建立了一个高质量的生化知识库。
在模型构建过程中，我选择了qwen1.5-
7b作为基座模型，并采用指令微调和对齐微调等先进技术，使模型能够更好地理解生化领域的任务。指令微调的核心在于构建高质量的领域数据集，因此，我设计了一个完整的数据收集和处理流程，从生化文献中提取专业知识，并结合化学和生物数据集，如PubChem、USPTO、UniProtKB等，生成高质量的指令数据集。为了确保数据的多样性和专业性，我提出了一种指令进化策略，利用GPT-4-
Turbo对初始指令进行扩展，生成更丰富、更复杂的任务指令，使模型能够处理从基础概念解释到复杂推理计算等不同难度层次的任务。在这之后我们还进行了数据的筛选，通过制定一套1-
5分的评分细则，使用gpt进行打分，筛选掉低于5分的指令。最终通过过滤获取了142万条训

练指令数据。
在模型训练过程中，我采用混合精度计算技术（FP16）优化计算效率和DeepSpeed库中实现的Zero Redundancy Optimizer（ZeRO）技术中的ZeRO-2策略等，以确保大规模数据训练的稳定性和高效性。同时，为了避免过拟合并提高模型泛化能力，我结合了早停（Early Stopping）、学习率衰减（Learning Rate Decay）等策略。训练完成后，我对模型进行了严格的评测，包括构建了一个基准评测数据集和利用其他生化领域标准基准数据集（如BC5CDR、GENIA）进行测试，并设计了一系列生化任务，如蛋白质功能预测、实验方案设计等，以验证模型的实际应用能力。实验结果表明，该模型在多个生化任务上均优于现有基线模型，尤其是在蛋白质功能预测任务中，性能提升显著。
该项目的成功不仅显著提升了生化领域复杂任务的自动化处理能力，也为生物医药研究提供了高效的智能工具，大幅降低了科研人员在文献检索、数据分析和实验设计中的工作负担。此外，这次工程实践使我积累了丰富的经验，使我更加深入理解了如何利用人工智能优化专业领域的大模型，并掌握了从数据整合、模型训练到实际应用落地的全流程技术。通过该项目，我不仅提升了在生化数据处理、大模型微调、计算优化等方面的能力，同时也进一步探索了人工智能在生化领域的应用前景，为未来在智能药物发现、精准医疗等方向的研究奠定了坚实的基础。

（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
一种基于最大差异竞赛实现大语言模型样本的评估方法和装置	发明专利申请	2024年08月30日	申请号：202410530635.6	1/5	
MedIE-Instruct: A Comprehensive Instruction Dataset for Medical Information Extraction	会议论文	2025年12月12日	ICKG	1/6	

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： 85 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： 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：** 顶卓怡	

2226028

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）：2023年3月18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：22260228	姓名：项卓怡	性别：女	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：28.0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	新时代中国特色社会主义理论与实践		2.0	92	专业学位课	2022-2023学年秋冬学期	研究生论文写作指导		1.0	88	专业选修课
2022-2023学年秋季学期	研究生英语基础技能		1.0	免修	公共学位课	2022-2023学年冬季学期	新药发现理论与实践		2.0	90	专业学位课
2022-2023学年秋季学期	研究生英语能力提升		1.0	免修	跨专业课	2022-2023学年冬季学期	知识图谱导论		2.0	90	跨专业课
2022-2023学年秋季学期	工程技术创新前沿		1.5	87	专业学位课	2022-2023学年秋冬学期	高阶工程认知实践		3.0	90	专业学位课
2022-2023学年秋季学期	研究生英语		2.0	免修	专业学位课	2022-2023学年春季学期	自然辩证法概论		1.0	81	专业学位课
2022-2023学年秋冬学期	科技创新案例探讨与实战		2.0	85	专业选修课	2022-2023学年春季学期	数学建模		2.0	70	专业选修课
2022-2023学年秋冬学期	工程伦理		2.0	96	专业学位课	2022-2023学年夏季学期	药品创制工程实例		2.0	88	专业学位课
2022-2023学年冬季学期	产业技术发展前沿		1.5	83	专业学位课		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：2025-03-20

(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118569213 A
(43)申请公布日 2024.08.30

(21)申请号 202410530635.6

(22)申请日 2024.04.29

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 项卓怡 冯科华 丁科炎 张强 陈华钧

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
专利代理师 曹兆霞

(51)Int.Cl.
G06F 40/16(2020.01)
G06F 40/30(2020.01)
G06N 3/045(2023.01)

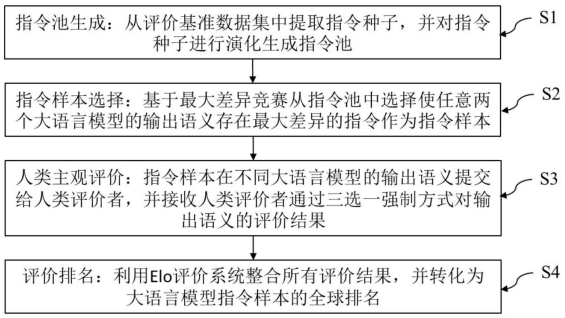
权利要求书2页 说明书8页 附图1页

(54)发明名称

一种基于最大差异竞赛实现大语言模型样本的评估方法和装置

(57)摘要

本发明公开了一种基于最大差异竞赛实现大语言模型样本的评估方法和装置,包括:指令池生成:从评价基准数据集中提取指令种子,并对指令种子进行演化生成指令池;指令样本选择:基于最大差异竞赛从指令池中选择使任意两个大语言模型的输出语义存在最大差异的指令作为指令样本;人类主观评价:指令样本在不同大语言模型的输出语义提交给人类评价者,并接收人类评价者通过三选一强制方式对输出语义的评价结果;评价排名:利用Elo评价系统整合所有评价结果,并转化为大语言模型指令样本的全球排名,这样可以克服机器评价偏见的同时,提升人类评估的效率和效果。



CN 118569213 A

MedIE-Instruct: A Comprehensive Instruction Dataset for Medical Information Extraction

Zhuoyi Xiang[1], Xinda Wang[2], Xiaodong Yan[3], Deng Zhao[3], Keyan Ding[4], Qiang Zhang [*4 5]
[1]*Polytechnic Institute, Zhejiang University*
[2]*School of Software Technology, Zhejiang University*
[3]*Machine Intelligence Department, Ant Group*
[4]*ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University*
[5]*ZJU-UIUC Institute, Zhejiang University*
{xiangzhuoyi, wangxinda, dingkeyan, qiang.zhang.cs}@zju.edu.cn
{qiqing.yxd, zhaodeng.zd}@antgroup.com

2024 IEEE International Conference on Knowledge Graph (ICKG) | 979-8-3315-0882-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICKG63256.2024.00059

***Abstract*—Medical information extraction (IE) tasks, including named entity recognition (NER), relation extraction (RE), and event extraction (EE), are crucial for constructing medical knowledge graphs from unstructured text. However, existing IE datasets in the medical domain are often limited, fragmented, and lack uniformity. To address these issues, we present MedIE-Instruct, a comprehensive bilingual (English and Chinese) medical IE instruction corpus comprising 11 datasets with over 100,000 instructions. This corpus was constructed using schema-based instruction generation to create a large-scale, diverse IE dataset to support large language models (LLMs) in medical IE tasks. Experimental results show that fine-tuning state-of-the-art LLMs with MedIE-Instruct significantly enhances model performance, especially in zero-shot scenarios. We hope this dataset and research findings will provide valuable resources and insights for understanding and processing medical content.**

***Index Terms*—Medical Information Extraction, Instruction Generation, Zero-shot/Few-shot Learning**

I. Introduction

In natural language processing (NLP), information extraction (IE) is a core technology for converting unstructured text into structured knowledge graphs. Recent advancements in deep learning have significantly improved various IE subtasks, including named entity recognition (NER), relation extraction (RE), and event extraction (EE) [1], [2]. Concurrently, large language models (LLMs), such as GPT-4 [3] and BERT [4], [5], have revolutionized information extraction tasks. These models, pre-trained on vast text corpora, exhibit outstanding abilities in understanding and generating human-like text [6]. They excel in scenarios where traditional models falter, such as interpreting diverse linguistic structures, comprehending context, and identifying entities and relationships within complex sentences [7]. By integrating large language models into the IE pipeline, the ability of LLMs to capture nuanced linguistic patterns and semantic relationships can be leveraged to enhance NER, RE, and EE tasks. This integration increases the flexibility and adaptability of information extraction systems [8].

The success of IE models relies on high-quality instruction datasets, which are essential for improving generalization and ensuring robust performance in real-world applications [9], [10]. However, current medical IE datasets are often limited in scale and coverage, failing to meet the diverse and complex demands of practical scenarios [11], [12]. While datasets like CoNLL-2003 and OntoNotes are valuable for NER tasks across various domains [13], the medical field urgently requires comprehensive datasets that include NER, RE, and EE. Moreover, inconsistencies in schema naming conventions and variations in annotation quality further constrain the potential for model performance improvement [14]. Additionally, these datasets are not designed as instructions that LLMs can readily understand and execute. The lack of such comprehensive instruction datasets not only limits the broader application of LLMs in medical information extraction but also hampers their development in critical areas like clinical text analysis and medical knowledge graph construction.

In response to these limitations, we developed MedIE-Instruct [1], a large-scale, bilingual (English and Chinese) medical extraction instruction dataset aimed at advancing information extraction in medical NLP applications. MedIE-Instruct consists of 11 datasets and over 100,000 instructions, created through schema-based instruction techniques to support LLMs in various medical IE tasks. In the generative IE setting [15]–[17], we employed the Structural Schema Instructor (SSI), as illustrated in Figure 1, which incorporates the schema into the model input. The entire predefined tag set of the dataset is utilized as SSI to guide the model's output during inference.

Fine-tuning LLMs on the MedIE-Instruct dataset led to decent performance improvements compared to baseline models in the medical domain, especially in zero-shot scenarios. Even in Few-shot scenary, our analysis revealed that the SFT models showed performance declines com-

*Corresponding author.

[1]Available at https://github.com/HICAI-ZJU/MedIE-Instruct.

979-8-3315-0882-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICKG63256.2024.00059