代表性成果登记表

申请人姓名：肖仰华　　所在院系：计算与智能创新学院

研究类别：[　]基础创新类　　[√]应用创新类

成果名称：

大模型数据科学理论与方法

成果支撑材料清单（分点列出组成本项成果的高水平论文、专著、授权专利、政策咨询报告等）：

一、高水平论文

1.Nianqi Li, Siyu Yuan, Jiangjie Chen, Jiaqing Liang, Feng Wei, Zujie Liang, Deqing Yang, **Yanghua Xiao**（通讯作者）. Past Meets Present: Creating Historical Analogy with Large Language Models. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025): 3942–3957（**ACL 2025 杰出论文奖**）

2.Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, **Yanghua Xiao**（通讯作者）. Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024): 18099-18107（该成果被 AINLP 评为最具参考价值的中文大模型评测基准之一）

3.Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, **Yanghua Xiao**（通讯作者）. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): 1840-1873（该成果是角色扮演智能体人格评测方向被引最高的基准之一）

二、授权专利

1.面向政务 Text2SQL 的多约束融合六元组数据集构建方法及设备，专利号： 2026106445834

2.面向问答系统鲁棒性的知识丰富型问题生成方法及装置，专利号： 2022112396663

3.大规模语言模型生成答案可靠性检测方法，专利号： 2023111954525

三、政策咨询报告

1.《生成式语言模型与通用人工智能：内涵、路径与启示》（《人民论坛·学术前沿》2023 年 7 月下）

2.《人工智能大模型发展的新形势及其省思》（《人民论坛·学术前沿》2024 年第 13 期）

3.《人工智能发展数据瓶颈与突破路径》（《人民论坛·学术前沿》2026 年第 11 期）

成果简介（200-300 字）:大模型数据科学是通用人工智能发展的重要基础，大模型发展面临知识难评估、推理易失真、人格化交互难度量三重挑战。为此，提出融合事实核验与自反思的历史事件类比推理方法，抑制推理幻觉与刻板倾向，获 ACL 2025 杰出论文奖；构建全学科、可持续更新的大模型知识评测基准 Xiezhi，获多家机构认可并用于中文大模型评测；提出基于心理访谈的角色人格保真度评测基准 InCharacter，获业界关注及 Nature 子刊论文引用与评价。针对产业需求，提出大模型与通用人工智能产业发展咨政建议，构建训练数据与核心能力标签体系，形成数据课程优化方案并获华为火花奖。相关技术用于华为盘古、阿里千问多个基座模型训练，提升了数据治理与模型能力建设成效，应用于大模型出海、航运大模型等应用领域，推动行业的高质量发展。

附说明：

一、一、科学贡献

大模型发展正面临高质量数据供给不足、数据利用效率偏低，进而制约模型能力迭代提升的现实瓶颈，仅靠扩大数据规模和经验试错，难以实现高效训练与高阶智能。发展大模型数据科学，旨在揭示数据驱动智能形成的规律，推动模型训练从经验工程走向可量化、可解释、可控制的科学体系，为大模型工业化奠定基础。当前，大模型研究仍面临三个基础问题：其一，模型知识覆盖广泛且快速演化，传统静态基准难以全面揭示其跨学科知识结构与能力边界；其二，模型虽然积累了海量历史知识，却难以突破表层相似性，将历史经验可靠地迁移到现实问题；其三，角色智能体能够模仿人物的语言和经历，但其深层人格是否与目标角色保持一致仍缺乏有效测量。围绕上述问题，本项目从领域知识评测、历史类比推理和人格化智能交互三个方向展开研究，形成如下三个主题的科学发现：

1.建立了大模型全域知识能力的动态评测与边界刻画方法。面向传统评测覆盖范围有限、数据易过时以及难以区分模型能力差异的问题，构建可持续更新的全领域知识评测体系 Xiezhi，揭示了模型规模增长与各领域知识能力的内在关系。

2.建立了大模型跨时空知识迁移的历史类比推理机制。面向大模型“拥有历史知识但不会有效借鉴历史”的问题，提出历史类比获取任务，系统性地评测了大模型的类比发现能力，提高了大模型历史知识迁移能力。

3.建立了大模型角色智能的人格一致性测量与可信交互评价方法。面向角色智能体评测停留于知识复现和语言模仿、难以触及内在认知与行为倾向的问题，提出基于心理访谈的角色人格评测方法，将角色智能研究由“语言像不像”推进到“人格一致、行为可信”的深层评价。

上述相关贡献累计发表中国计算机学会（CCF）认定的 A 类期刊或会议论文 110 篇，授权专利 11 项，资政论文 3 篇。成果多次受工业界邀请作主旨演讲。大模型的课程学习方案荣获华为火花奖，类比推理评测研究获得 ACL 2025 杰出论文奖（见图 1）。获批华为全球范围内首个先进技术研究中心，专注于大模型数据科学方向的自然探索项目。获得华为 openPangu 计划支持，在国际上率先开设了《大模型数据工程》课程，并作为负责人承担建设了《openPangu 模型与应用》教育部虚拟教研室。成果应用于华为盘古基础模型训练以及阿里千问模型的多模态训练数据建设，应用于上海亿通国际的航贸大模型训练，支撑了华为翻译中心的跨文化跨语言语料建设，以及华为数据通信等业务部门的专业语料建设。

二、代表性论著评价

申请人提出的面向历史事件的类比推理相关工作荣获 ACL 2025 杰出论文奖（论文1），组委会认为该工作“创新性地提出自反思机制以抑制大模型推理过程中的幻觉与刻板倾向”。ACM、IEEE、AAAS、AAAI Fellow Nitesh V Chawla 多次引用本论文，并使用独立段落介绍本文工作，高度认可了本文提出的自我反思框架，以及自动多维相似度指标与人类评估的高度一致性（“Their automatic multi-dimensional similarity metric demonstrates high alignment with human evaluations”）。

申请人构建了一个全学科语言模型评估基准数据集 Xiezhi（论文2），包含 516 个领域维度，25 万条领域评估数据，同时提出 50 选 1 的测评机制，较好地验证了语言模型真正掌握知识的情况（见图 2）。该评测数据集被 AINLP 评为最具参考价值的中文大模型评测基准之一，是与 C-Eval 等并列的“学科知识评测”代表，并被 OpenDataLab 等平台收录。ACM、IEEE Fellow Xing Xie 在论文中认为 Xiezhi 能够有效评估大语言模型在不同学科领域的知识水平，通过 Xiezhi 进行的评测，能让研究者认识到这些模型的内在局限，并加深对其在不同领域能力的理解（“Xiezhi enables researchers to comprehend the notable limitations inherent in these models and facilitates a deeper comprehension of their capabilities in diverse fields”）。

申请人构建的大模型角色扮演人格保真度评测基准 InCharacter（论文3），是角色扮演智能体人格评测方向被引最高的基准之一（Semantic Scholar 约 240 次）。IEEE、ACM、AAAS Fellow、美国国家工程院院士 Maja J. Matarić 在 Nature Machine Intelligence（Nature 子刊）发表的文章中，将本论文与 CharacterEval 并列作为代表性“角色人格心理测量框架”加以引用。相关研究者在论文中认为，InCharacter 通过植根于心理学特质、以访谈式探查的方式评测角色人格保真度，为衡量角色扮演智能体是否真正“成为”其所扮演的角色提供了严谨的心理测量学依据。（InCharacter evaluates personality fidelity through interview-style probing grounded in psychological traits.）

申请人在国际上率先开设《大模型数据工程》课程，从训练数据治理、认知能力评测、领域模型优化等关键环节展开系统研究，形成了涵盖语料分级分类、数据合成、配比优化、质量评估、能力评测全流程的大模型数据工程体系。该课程受到华为 openPangu 专项资金、复旦大学 AI-BEST 课程改革资金资助，获批教育部“智能基础”虚拟教研室（openPangu 模型与应用）（见图 3）。

三篇资政文章均发表在由人民日报社主管、人民论坛杂志社主办的高端学术期刊《人民论坛·学术前沿》“二十四个重大问题研究”专栏，该期刊是中文社会科学引文索引（CSSCI）来源期刊、全国中文核心期刊、中国人文社会科学期刊 AMI 综合评价 A 刊核心期刊，也是国家社科基金成果指定发布期刊（见图 4）。

三、研究成果应用与影响

在领域知识数据集构建与优化方面，提出了高质量数据集构建的思路，围绕大模型认知能力提升，从基础评测、推理认知、情感认知、文化认知等多个角度，构建了 40 多个大规模、高质量、数据分布更合理的数据集，促进了学术界理论方法研究和工业界算法应用选型。面向旅行规划的模型评估基准研究工作《TravelPlanner：A Benchmark for Real-World Planning with Language Agents》被图灵奖获得者 Yann LeCun 在社交网络上转发（转发量达到 10W+），并入选 ICML 2024 亮点论文。另一项关于大语言模型与人类认知的工作《Why Did Apple Fall: Evaluating Curiosity in Large Language Models》，在海外社交平台 X 成功出圈，获海外头部 AI 科技资讯平台主动转发推介，影响力巨大（见图 5）。

面向大模型训练数据繁杂、管理困难、配比混乱、训练低效的核心难题，申请人从理论出发为大模型数据治理提出了一套训练数据与核心能力标签体系建设思路及数据课程优化方案，荣获华为火花奖（见图 6）。另外，开发的大规模多模态数据科学分类体系与概念理解科学评测体系，被应用于阿里大模型多模态任务训练中，并在实际应用中取得了显著效果，为阿里提供了高效的多模态训练语料数据治理支持（见图 7）；训练的航贸领域大模型有效赋能了航贸数据产品，并在 HSCode 归类等任务上取得良好效果（见图 8）。

申请人得到华为翻译中心连续三项研发项目资助（见图 9），面向大模型社会认知与跨文化交互方向，建立起完整的评测与数据增强研究体系。研究先后构建多项权威数据集：基于认知心理学构建文化能力评测数据集（AAAI 2026）；研发跨文化安全评测数据集，包括覆盖 77 个国家和地区、包含 2020 余条禁忌样例的 CulShield 文化安全基准，以及包含 16 种语言、3.68 万条专家精修样本的数据集（ACL 2026）；搭建大模型跨文化情境能力评测数据集（EMNLP 2026）。相关成果支撑华为建成当前规模最大的跨文化跨语言评测数据集 GaoYao（ACL 2026），数据集覆盖 26 种语言、51 个国家和地区，共计 18.2 万条样本。上述研究有效缓解了大模型存在的文化认知与实际行为脱节、跨语言能力不均衡、翻译数据本地化程度不足等痛点；通过模型能力诊断、数据优化与筛选等技术路径，提升了大模型多语言文化适配水平，为国产大模型安全可靠出海提供评测依据与数据底座。

申请人在国际上首次提出“人考 AI” 训练数据构建方法。依托 “逆向评测” 思路，由学生命题对大语言模型开展考核，评分机制与模型失效率正向关联。实践结果表明，多数学生可有效识别模型的推理盲区，部分测试案例实现模型作答零分的突破。本次课程评价改革，重在引导学生建立对 AI 系统能力边界的批判性认知，推动评测范式由知识记忆考核转向人机协同决策能力评估。“人考 AI” 评测模式引发国内外广泛关注：人民日报、新华社、环球时报、中国青年报等官方微信公众号均发布 10W+ 报道，德国《法兰克福邮报》亦刊发专题评述，成为全社会热点议题（见图 10）。班贝格大学认知系统领域专家 Ute Schmid（施

密德）对该改革作出评价：“不仅能够帮助学生了解人工智能模型的局限，也有助于培养学生批判性判断内容质量的能力。” 目前，该教学改革已获得阿里巴巴支持，正组织开展覆盖 100 门课程的 “人考 AI” 竞赛，用于规模化采集高质量人机对抗测试题目。

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>图 1 知识增强的类比推理模型荣获 ACL 2025 杰出论文奖</td>
<td>图 2 全学科语言模型评估基准 Xiezhi</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>图 3 openPangu 模型与应用虚拟教研室入选教育部高等教育司公布的第四批虚拟教研室建设试点</td>
<td>图 4 《人民论坛·学术前沿》收录 3 篇有影响力的资政论文</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>图 5 海外科技媒体 AIDB 关于成果论文《Why Did Apple Fall: Evaluating Curiosity in Large Language Models》的专题报道，阅读量 6 万次</td>
<td>图 6 训练数据与核心能力标签体系建设以及数据课程优化方案获华为火花奖</td>
</tr>
</table>



合作成果证明

我公司与复旦大学计算机科学技术学院王怀华、李志旭老师团队于 2024 年 5 月至今在大模型数据治理领域展开了深度合作。双方就多模态大模型数据质量治理问题进行了深入研究，主要针对高质量图文对的标准化、数据噪声的剔除、语料分布不均衡以及评测集中核心问题进行攻关。基于上海市科技创新行动计划项目“基于大规模多模态知识库的概念理解与多模态数据质量治理研”（NO. 2251174 84700）中提出的理论框架与方法论，有效解决了数据质量评估的瓶颈，并提出了一系列创新解决方案。

合作过程中，项目组开发了大规模多模态数据科学分类体系与概念理解科学评测体系，将强大模型对理论评价的高度重视和多样性，这一体系已成功应用于阿里内部大模型的多模态治理中，并在实际应用中取得了显著的效果。相关成果已形成发明或技术部门成功上线，为基多模态训练治理提供了高效的数据治理支持。

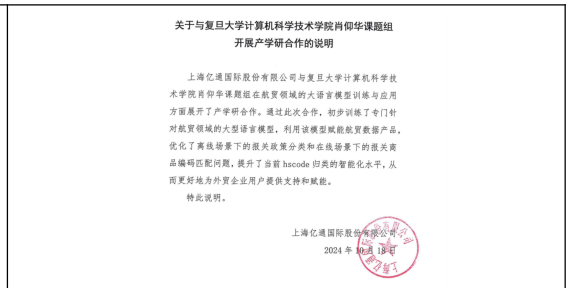
合作项目已顺利通过阿里巴巴集团审核验收，评价结果：项目已成功完成合同任务，达到预期目标，并在多个关键指标上超额完成，其中 DMLP 一项，AAAI 一项，为大模型数据治理领域树立了新的行业标准，具有广泛的社会与经济价值。

特此说明。

上海亿通国际股份有限公司

2024 年 10 月 18 日

图 7 阿里巴巴合作成果证明：大规模多模态数据科学分类体系与概念理解科学评测体系



关于与复旦大学计算机科学技术学院肖仰华课题组开展产学研合作的说明

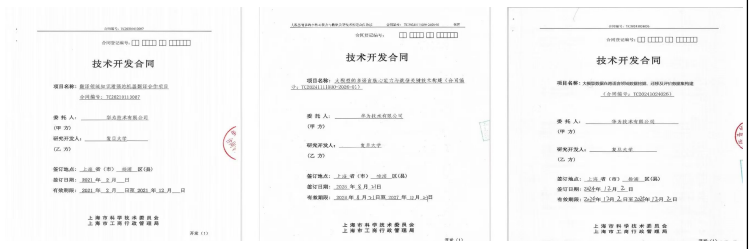
上海亿通国际股份有限公司与复旦大学计算机科学技术学院肖仰华课题组在航贸领域的大语言模型训练与应用方面展开了产学研合作。通过此次合作，初步训练了专门针对航贸领域的大型语言模型，利用该模型赋能航贸数据产品，优化了离线场景下的报关政策分类和在线场景下的报关商品编码匹配问题，提升了当前 hscode 归类的智能化水平，从而更好地为外贸企业用户提供支持和赋能。

特此说明。

上海亿通国际股份有限公司

2024 年 10 月 18 日

图 8 上海亿通国际合作证明：在航贸领域的大语言模型训练与应用

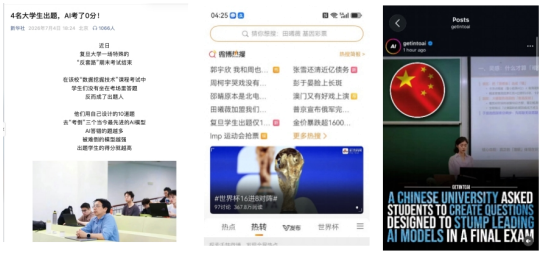


技术开发合同

技术开发合同

技术开发合同

图 9 华为翻译中心连续三项研发项目资助：大模型社会认知与跨文化交互方向



4名大学生出题，AI考了0分！

近日，复旦大学一场特殊的“反面教”期末考试结束

在该校“数据挖掘技术”课程考试中学生们没有去在考场答题反而成了出题人

他们用自己设计的10道题去“考倒”三个当今最先进的AI模型AI答错的概率多被推荐的模型成绩出题学生的得分就越高

图 10 数据挖掘课程 “人考 AI” 教学改革：登上微博热搜榜，获国内外多家权威媒体报道