



第三十二届 CHINA CONFERENCE
ON INFORMATION RETRIEVAL



CCIR 2026 全国信息检索学术会议

会议手册

主办单位:中国中文信息学会信息检索专委会
承办单位:吉林大学

2026年8月14日-16日 中国·长春





目录

▶ 关于CCIR	01
▶ 会议组织	02
▶ 简要日程	03
▶ 交通路线	04
▶ 会议程序	05
▶ 大会主旨报告	07
▶ 大会主题报告	11
▶ 工业论坛报告	14
▶ 青年论坛报告	16
▶ 讲习班	24
▶ 录用论文列表	26
▶ 赞助单位	28

关于CCIR

信息检索旨在满足人类在互联网上快速准确地获取信息与知识的需求,研究成果将支撑国家战略决策,推动互联网和人工智能领域的发展,提升整个社会的生产效率,并对社会生活各个领域产生重大影响。全国信息检索学术会议(CCIR)由中国中文信息学会(CIPS)举办,一路伴随着中国互联网产业的成长,是信息检索领域的旗舰会议。

CCIR2026

第三十二届全国信息检索学术会议(The 32nd China Conference on Information Retrieval, CCIR 2026)将于2026年8月14-16日在吉林长春举行,此次会议由中国中文信息学会信息检索专委会主办,吉林大学承办。

本次大会的主题是“AI for Science: 新范式与新未来”,聚焦科学信息检索与前沿科学研究的交叉融合,致力于构建海量知识检索、数据驱动、机理融合与智能发现的科研新范式。会议包含一系列学术活动,除传统的海内外知名学者的大会报告、会议论文报告外,还将组织青年学者论坛以及面向热点研究问题的讲习班等。CCIR已成为全国范围内信息检索领域新学术和技术工作的主要交流平台。



会议组织

大会主席:常 毅(吉林大学)

郭嘉丰(中科院计算所)

徐 君(中国人民大学)

程序主席:艾清遥(清华大学)

范意兴(中科院计算所)

王 鑫(吉林大学)

组委会主席:王 琪(吉林大学)

宣传主席:白 婷(北京邮电大学)

邬 渊(吉林大学)

出版主席:俞蔚捷(对外经济贸易大学)

龙 婷(吉林大学)

网络主席:詹靖涛(清华大学)

青年论坛主席:陈 旭(中国人民大学)

吕子钰(中山大学)

评测主席:朱小飞(重庆理工大学)

刘伟东(中国移动通信研究院)

赞助主席:马为之(清华大学)

刘子韬(暨南大学)

财务主席:孙培杰(南京邮电大学)



简要日程

时间	时长	2026/8/14 假日宴会中心二楼紫气东来厅
14:00—15:30	90分钟	专题讲习班1：以构象为核心的生成式大分子建模 许洪腾
15:30—17:00	90分钟	专题讲习班2：NLP到AI4Science的初探——基因组语言模型 倪仕文
19:00—20:00	60分钟	专委会工作会议
时间	时长	2026/8/15上午 假日宴会中心一楼锦绣中华厅
08:30—09:00	30分钟	大会开幕式
09:00—09:10	10分钟	合影
09:10—10:10	60分钟	大会主旨报告1：黎曼图深度学习: 图基础模型的新视角 Philip S.Yu
10:10—10:30	20分钟	茶歇
10:30—11:30	60分钟	大会主旨报告2：The Path to Trustworthy AI Agents 胡 侠
11:30—13:00	90分钟	午休
时间	时长	2026/8/15下午 假日宴会中心二楼紫气东来厅
13:00—13:30	30分钟	工业报告1：AgentVerse多智能体系统~驱动全行业生产力跃迁 丁来平（面壁智能）
13:30—14:00	30分钟	工业报告2：深度学习平台助力新材料智能化发现 刘旭韦（百度）
14:00—14:30	30分钟	AI for Science专题报告1：AI for Drug Discovery—新范式下的方法与挑战 兰艳艳
14:30—15:00	30分钟	AI for Science专题报告2：知识增强的AI4S科学智能体 陈华钧
15:00—15:30	30分钟	AI for Science专题报告3：从AI4S到AI科学家 李 勇
15:30—15:50	20分钟	茶歇
15:50—16:50	60分钟	大会主题圆桌论坛
16:50—18:35	105分钟	候选最佳论文报告及海报展示
18:35—20:35	120分钟	晚宴（一楼锦绣中华厅）
时间	时长	2026/8/16 假日宴会中心二楼紫气东来厅
08:30—09:30	60分钟	大会主旨报告3：AI赋能信息网络的研究与思考 张宏科院士
09:30—10:30	60分钟	大会主旨报告4：人工智能的进化路径与社会影响 刘奕群
10:30—10:50	20分钟	茶歇
10:50—12:50	120分钟	青年学者论坛
12:50—14:00	70分钟	午休
14:00—16:10	130分钟	新委员论坛
16:10—16:30	20分钟	茶歇
16:30—17:30	60分钟	评测报告
17:30—18:00	30分钟	大会闭幕式

交通路线

会议地点:长春净月假日酒店(长春市南关区净月开发区净月大街3070号)。

会议住宿:长春净月假日酒店:双人间/大床房450元/晚

长春净月智选假日酒店 :双人间/大床房350元/晚

地址:长春市南关区净月开发区净月大街3070号

会议交通:

长春站(北出口)

A公共交通:步行至地铁1号线(约5分钟),乘坐1号线至卫星广场,站内换乘3号线,乘坐3号线至金河街站,出站后步行至会场,全程票价4元,时长约1小时。

B网约车直达:全程约18公里,费用预计25-35元,约30分钟。

长春站(南出口)

A公共交通:直接乘坐轻轨3号线至金河街站,出站后步行至会场,票价5元,时长约1小时。

B网约车直达:全程约19公里,费用预计25-35元,约30分钟。

长春西站

A公共交通:步行至地铁2号线(约5分钟),乘坐2号线至解放桥,站内换乘3号线,乘坐3号线至金河街站,出站后步行至会场,全程票价5元,时长约1小时10分钟。

B网约车直达:全程约24公里,费用预计30-40元,约40分钟。

长春龙嘉机场

A公共交通:步行至龙嘉机场巴士(约5分钟),乘坐巴士2号线至会展中心,转乘3号线至金河街,出站后步行至会场。巴士票价25元,3号线2元,全程时长约1.5小时。

B网约车直达:全程34.5公里,费用预计80-100元,约45分钟。



大会主旨报告

报告一:黎曼图深度学习:图基础模型的新视角

讲者:Philip S. Yu 伊利诺伊大学芝加哥分校

时间:8月15日 9:10—10:10



个人介绍:Philip S. Yu 博士是美国计算机协会 (ACM)、电气和电子工程师协会 (IEEE) 以及美国科学促进会 (AAAS) 的会士 (Fellow)。Yu 博士长期致力于数据挖掘、知识发现、大数据分析 with 人工智能领域的研究,并因其在大数据挖掘、数据融合、数据匿名化以及大规模数据管理等方面的突出贡献,获得了 ACM SIGKDD 2016 创新奖、IEEE Computer Society 2013 技术成就奖以及 IEEE ICDM 2003 研究贡献奖等重要荣誉。他已发表超过 1,900 篇经同行评审的国际会议和期刊论文,论文总引用次数超过 269,000 次,H-index 达 222,并拥有超过 300 项专利申请。Yu 博士曾担任 ACM Transactions on Knowledge Discovery from Data (TKDD) 主编 (2011–2017) 以及 IEEE Transactions on Knowledge and Data Engineering (TKDE) 主编 (2001–2004),在数据挖掘与人工智能领域具有广泛影响力。

报告摘要:图基础模型 (Graph Foundation Models, GFMs) 有望推动图学习领域的变革,但如何构建类似于大语言模型 (Large Language Models, LLMs) 的通用图基础模型,仍然是当前研究领域广泛讨论的重要问题。传统图神经网络 (Graph Neural Networks, GNNs) 在长期信息保持、理论驱动的可解释性以及跨领域适应能力方面仍存在局限;与此同时,将大语言模型直接应用于图数据也面临挑战,因为图序列化方法将复杂图结构转换为线性化标记序列,难以充分表达图中丰富的拓扑与语义信息。相比之下,黎曼几何 (Riemannian Geometry) 为图结构建模提供了一种自然且具有数学理论基础的框架,能够有效结合语义图学习与大语言模型技术。在本报告中,我们提出“对于图而言,几何比文字更能表达其本质”的观点,并介绍一种面向图基础模型的新范式——黎曼图基础模型 (Riemannian Graph Foundations)。该方法以图的内在几何结构为核心,使模型具备结构推理与结构生成的内生能力,从而突破传统表示空间变换的局限。报告将介绍近期在黎曼几何空间中构建图基础模型的研究进展,并展示该方法如何为下一代图智能的发展奠定理论与技术基础。

报告二:The Path to Trustworthy AI Agents

讲者:胡侠 浦江实验室

时间:8月15日 10:30—11:30



个人介绍:胡侠,国家级人才,现任浦江实验室主任助理、领军科学家。他曾任美国莱斯大学教授、莱斯大学数据科学中心主任,作为联合创始人兼首席科学家创立 AIPOW LLC;长期担任多本国际知名期刊副主编、编委及学术会议主席,并以项目负责人身份主持由 DARPA、NSF、NIH 资助的多项重大科研项目。胡侠长期深耕机器学习与人工智能领域,在 ICLR、NeurIPS、KDD 等国际顶级会议与期刊发表论文 200 余篇,论文总引用量超 45000 次,先后斩获 NSF CAREER Award、SIGKDD Rising Star Award 等 20 余项重要学术奖项,累计 12 次获得国际顶会最佳论文奖或提名,在国际学术界具备广泛影响力。其主导研发的开源系统 Auto-Keras 已是行业标杆项目,被集成至 TensorFlow 等主流工业框架并收录于专业教材;同时团队与 Meta、三星、Visa 等全球头部企业开展深度合作,持续推动人工智能前沿技术在关键场景落地应用。

报告摘要:AI models are actively assisting humans with an increasing number of daily tasks by leveraging external tools. While these tool-empowered agent systems can effectively carry out real-world tasks, they also pose risks of harming the physical world if they make unsafe decisions or errors. This talk begins by introducing representative agent frameworks, illustrating how the integration of different tools reduces the controllability of these agent systems. We then elaborate on our recent efforts to ensure AI safety and develop safe AI for controllable agent systems. Specifically, our progress spans all four core steps in the agent workflow: user input scanning, inherent model safety, tool usage monitoring, and output verification. Collectively, these advancements drive us toward our proposed AI safety development paradigm, which governs the co-evolution of AI systems' safety and performance. The talk concludes with a discussion of the existing challenges to this objective and our strategies to address them.



报告三:AI赋能信息网络的研究与思考

讲者:张宏科 北京交通大学

时间:8月16日 8:30—9:30



个人介绍:张宏科,中国工程院院士,北京交通大学学术委员会主任,IEEE Fellow。曾任两期国家973计划首席科学家。长期从事专用通信网络理论与工程技术研究,为解决国家和行业专网工程急需做出贡献。获国家技术发明奖、全国创新争先奖,享受国务院政府特殊津贴,是首批全国高校黄大年式教师团队带头人。

报告摘要:车联网、低空网络等典型应用场景,对网络的实时性、移动性等性能指标提出了更高要求。本

报告聚焦新型网络体系架构,阐述专网工程中智能网络的技术演进路径,并展望其未来应用前景。



报告四:人工智能的进化路径与社会影响

讲者:刘奕群 清华大学

时间:8月16日 9:30—10:30



个人介绍:刘奕群,清华大学计算机系教授,现任清华大学科研院院长、技术转移院院长、互联网司法研究院院长。国家级高层次人才,中国人工智能学会会士,ACM和CCF杰出会员,ACM SIGIR Academy成员。曾获中国青年科技奖、北京市科学技术一等奖、钱伟长中文信息处理科学技术一等奖等奖励。主要研究方向包括人工智能、信息检索、自然语言处理及其交叉应用,在人工智能赋能社会治理、司法实践和社会科学研究等方面具有丰富的研究积累与实践经验。

报告摘要:人工智能正在从专用任务系统快速迈向通用智能,其能力边界与社会影响也在持续扩展。本报告将回顾大模型的发展历程,结合规模定律、认知发育与生物智能讨论机器智能可能的进化路径及其极限,并以高等教育为例分析人工智能对学习方式与人才培养的影响。在此基础上,进一步通过职业能力需求图谱,探讨人工智能发展将如何重塑未来社会的能力结构与职业分工。

大会主题报告

报告一: AI for Drug Discovery——新范式下的方法与挑战

讲者: 兰艳艳 清华大学

时间: 8月15日 14:00—14:30



个人介绍: 兰艳艳, 清华大学智能产业研究院 (AIR) 副院长/首席研究员/教授, 人工智能学院博士生导师, 国家级高层次青年人才。研究方向为机器学习、信息检索和AI for Science。在Nature、Science等期刊和人工智能领域顶级国际会议发表论文80余篇, Google Scholar引用超过13000次。获得SIGIR最佳学生论文奖, CIKM最佳论文Runner-Up奖, 国际蛋白质结构比赛CAMEO冠军, 中文信息学会钱伟长中文信息处理科学技术奖一等奖、青年创新奖等。担任AI领域多项国际重要会议的PC Chair和Area Chair, 国际期刊AI Journal副主编。

报告摘要: 人工智能正重塑药物发现的底层范式, 从“实验试错”的假设驱动范式走向“计算预测与生成”的数据驱动范式。过去五年, 我们课题组围绕AI药物研发构建了一体化计算平台, 覆盖靶点发现、结构预测、药物虚拟筛选与设计、性质预测与优化等药物研发主要环节。本次报告将聚焦虚拟筛选、药物设计和性质预测这三个应用, 针对现有药物发现场景的问题与挑战, 重点探讨基于信息检索、生成式AI与基础大模型等技术的创新解决思路, 并展望从“单点计算”走向“闭环智能设计”的可行路径。

报告二: 知识增强的AI4S科学智能体

讲者: 陈华钧 浙江大学

时间: 8月15日 14:30—15:00



个人介绍: 陈华钧, 浙江大学求是特聘教授, 主要研究方向为人工智能、知识图谱、自然语言处理、AI for Science等, 牵头发起中文开放知识图谱OpenKG, 以一作或通讯作者在Nature Machine Intelligence、Nature Biotechnology、Nature Computational Science, Nature Communications、ICLR、ICML、NeurIPS、IJCAI、AAAI、ACL、EMNLP、WWW、SIGIR、KDD、VLDB、Proceeding of IEEE、ACM Computing Survey、National Science Review等国际顶会顶刊上发表多篇论文。担任浙江省数智科技研究会副会长, 中国人工智能学会知识工程专委会副主任, 中国中文信息学会语言与知识计算专委会副主任。入选浙江省有突出贡献中青年专家, 浙江省高层次人才特殊支持计划科技创新领军人才, 斯坦福全球前2%顶尖科学家终身榜单(人工智能领域)。作为项目负责人及首席科学家主持科技部创新2030“新一代人工智能”国家科技重大专项“知识增强的科学具身智能体平台及应用”重大项目, 并主持负责国家自然科学基金委重大研究计划重点项目、国家自然科学基金联合基金重点项目及多项重大企业合作项目。牵头或作为核心骨干获得省部及国家级科技奖5项, 及浙江大学首届优秀教材奖一等奖、中国工信传媒出版集团优秀出版物一等奖等教材专著类奖。

报告摘要: 面向科学的Agentic AI旨在超越被动式问答, 迈向能够运用科学知识和工具进行推理、规划、行动和学习的系统。本次报告中, 将首先介绍通过结合大语言模型(LLM)与知识图谱(KG)来构建知识增强智能体的基本思路。大语言模型提供灵活的语言理解和推理能力, 而知识图谱贡献结构化知识、显式关系、来源追溯和可控的语义锚定。二者的协同可以提升科学智能体的可靠性、可解释性和可扩展性。接着, 我将探讨科学具身智能体的新视角, 强调科学智能不仅应基于文本和数据进行推理, 还应以目标导向的方式与工具、环境和工作流进行交互。最后, 将介绍团队近期的几个项目, 包括InstructBioMol、SciToolAgent、SciGraph和SkillNet。这些工作分别探索了面向生物分子任务的指令遵循、工具使用的科学智能体、以图为中心的科学知识组织与推理, 以及智能体科学智能的技能习得。更广泛地看, 这些方法可应用于多个科学领域, 包括化学合成、合成生物学和药物发现, 支撑文献挖掘、假设生成、实验方案规划和迭代实验。

报告三：从AI4S到AI科学家

讲者：李勇 清华大学

时间：8月15日 15:00—15:30



个人介绍：李勇，清华大学电子工程系长聘教授、教育部长江学者。围绕人工智能、数据科学及交叉学科方向，承担国家重点研发计划项目、自然科学基金重点项目等20余项，研究成果发表于Nature（自然）、Nature Computational Science（自然·计算科学）、Nature Machine Intelligence（自然·机器智能）、Nature Human Behaviour（自然·人类行为）、Nature Cities（自然·城市）等综合性期刊，在ACM KDD、NeurIPS、ICLR、WWW等CCF A类会议期刊发表学术论文200余篇，引用42000次，授权专利100余项。先后入选全球“高被引科学家”、国家“万人计划”青年拔尖人才计划，获计算机领域重要国际会议ACL、WWW、SIGIR、UbiComp最佳/优秀论文奖7次，获教育部科技进步一等奖、湖北省技术发明一等奖、电子学会科技进步一等奖等。

报告摘要：大语言模型和智能体时代，AI4S智能科研范式正在走向AI科学家的新阶段。本报告将介绍支撑人机深度协同的通用智能科研平台，通过融合“科学、科研和学者”三个维度知识与流程的科学大模型与多智能体工作流，在理论提出与自动化实验闭环中有望实现突破性进展，标志着从“AI辅助”向“AI主导”的科研范式根本性跃迁，旨在重塑未来科学发现路径。

工业论坛报告

报告一：AgentVerse多智能体系统～驱动全行业生产力跃迁

讲者：丁来平 面壁智能

时间：8月15日 13:00-13:30



个人介绍：丁来平，面壁智能 Agent 解决方案架构师，东南大学硕士。拥有十余年 ICT 与人工智能领域工程与生态经验，技术轨迹横跨通信网络、AI 框架与智能体三大方向。2015 年至 2021 年任职于华为，负责无线 OSS 网管网络优化特性的方案设计与全球推广，相关方案覆盖 10+ 国家、20+ 运营商，具备从技术方案到规模商用的全链路落地能力。2021 年至 2026 年中担任昇腾 AI 框架布道师，主导昇腾 AI 框架生态建设与开发者赋能。完成 Qwen、智谱、DeepSeek 等主流大模型在昇腾 NPU 上的适配与性能优化；统筹 Torch-NPU 与 PyTorch 生态的 API 对齐工作，累计对齐 500+ API，向社区提交 100+ PR；同时负责 AI 框架技术课程体系设计与讲授及开源项目管理，持续推动昇腾框架生态的兼容性与开发者体验提升。2026 年中加入面壁智能，担任 Agent 解决方案架构师，聚焦企业级智能体系统的架构设计与行业落地。

报告摘要：以「AGI for Lives」为背景，研判大模型迈向 AGI 的全模态、高效化与自主化趋势，阐释从对话工具到专业智能体、再到多智能体协同的生产力变革，并介绍 AgentVerse 框架及其在教育、法律、软件等领域的落地，展望人机协同的群体智能。

报告二:深度学习平台助力新材料智能化发现

讲者:刘旭韦 百度

时间:8月15日 13:30-14:00



个人介绍:刘旭韦, 百度大模型前沿探索部高级研发工程师, 博士毕业于日本国立群马大学, 中科院过程工程所联合培养。研究方向关注AI for Materials, 科学智能, 微、介观数值模拟方法和应用, 发表多篇JCR 2区以上SCI论文。

报告摘要:传统材料研发面临诸多挑战, 包括“组分—结构—工艺—性能”(构-效)关系的高度复杂非线性、材料的多尺度空间特征、巨大的材料搜索空间以及高昂的实验成本。随着人工智能技术的持续进步, AI 正在加速破解这些瓶颈。然而, AI 驱动的材料研发仍然存在场景任务多样、模型演进迅速、缺乏面向材料研究者的统一且易用的平台, 以及扩散模型、多模态模型等新技术集成难度大等问题。为此, 本文提出一种基于“数据—机理”双驱范式的材料研发平台。该平台通过智能算法与先进仿真工具的深度融合, 为材料发现、化学合成以及合金设计等任务提供全流程的智能化支撑。平台核心组件 PaddleMaterials 采用统一的训练器(trainer)架构, 兼容回归模型、扩散模型与多模态模型;同时将训练器、采样器(sampler)与预测器(predictor)进行解耦设计, 实现高效的训练与推理, 以适配不断演进的研究需求与多样化的应用场景。其底层数据封装支持晶胞、分子图以及多种谱学图像等多种数据格式, 为终端用户提供便捷易用的工作流, 尤其适合在材料发现场景下集成各类仿真工具。此外, 平台基于飞桨 PaddlePaddle 深度学习框架构建, 兼容多样化硬件架构, 并提供分布式训练与推理能力。实验结果表明, 该一体化平台可显著提升新材料研发的效率与准确性, 为材料研发的智能化转型提供有力支撑, 推动行业技术创新。

青年论坛报告

报告一:当AI成为智能显微镜:理解智能机理的新科学范式

讲者:张宁豫 浙江大学

时间:8月16日 10:50-11:05



个人介绍:张宁豫, 博士, 副教授、启真优秀青年学者, 多次入选斯坦福全球前2%顶尖科学家榜单, 主要研究方向为自然语言处理、大模型、知识图谱、知识编辑等, 主持国家自然科学基金青年/面上、科技部科技创新2030新一代人工智能重大专项子课题、CCF-百度松果/腾讯犀牛鸟基金、CAAI华为MindSpore基金及头部互联网企业合作项目等, 研发了海洋大模型OceanGPT·沧渊、知识获取系统

DeepKE、知识编辑系统EasyEdit、轻量记忆系统LightMem、智能体技能图谱系统SkillNet、AI机理科学家系统Mechanist, 曾获得ACL SAC Highlight Award, 担任TMLR、Neural Networks、Machine Intelligence Research、ACM TALLIP、IP&M Associate Editor, ACL、EMNLP、NeurIPS、ICLR、KDD领域主席, ARR Action Editor, IJCAI 高级程序委员、多个Nature子刊审稿人。联合培养学生入选中国科协青年科技人才培养工程博士生专项计划、中国电子学会-腾讯博士生科研激励计划、中国人工智能学会潜力清源学者、中国中文信息学会“博士学位论文激励计划”、浙江省青苗人才计划、腾讯犀牛鸟精英人才计划。

报告摘要:本报告围绕“如何理解智能”这一科学前沿问题, 系统梳理从碳基智能到硅基智能的发展脉络, 回顾人类在哲学、认知科学、脑科学与人工智能等领域对智能本质的持续探索, 分析人工智能机理研究的发展历程及其代表性工作, 包括神经元分析、回路发现、行为调控等关键机制研究。在此基础上, 报告进一步讨论当前智能机理研究面临的核心挑战, 介绍AI研究AI的新范式, 探讨人工智能如何从被研究对象逐渐演变为理解智能的新型科学仪器。结合多学科知识库与统一分析框架, 展示基于AI辅助智能研究的新机理发现、定量复现结果及其科学价值。最后, 报告展望未来智能机制科学的发展方向, 讨论认知科学、脑科学与人工智能深度融合背景下, 构建可理解、可验证、可复现的智能理论体系的机遇与挑战, 探索从“创造智能”迈向“理解智能”的新科学范式。

报告二:迈向自主进化的智能体

讲者:曹艺馨 复旦大学

时间:8月16日 11:05-11:20



个人介绍:曹艺馨,上海创智学院全时导师,复旦大学可信具身智能研究院青年研究员、博士生导师。国家级青年人才计划入选者、上海市青年领军人才计划入选者。于清华大学获得博士学位,曾先后在新加坡国立大学、南洋理工大学和新加坡管理大学担任博士后、研究助理教授和助理教授职位。研究领域为大语言模型评估与自主进化,推理和检索增强,主持并参与各类国家与企业项目,收到包括国家自然科学基金委员会(面上)、国家科技重大专项课题、新加坡政府等资助。在国际知名会议和期刊发表论文80余篇,谷歌学术引用11,000余次,并多次被领域内国际顶级会议评为口头报告。研究成果获得两项国际会议的最佳论文及提名,曾获Lee Kong Chian Fellowship、Google South Asia & South-east Asia Awards和AI2000最具影响力学者奖的荣誉提名,Elsevier2024、2025全球前2%顶尖科学家。担任多个国际会议演示程序主席、领域主席和国际期刊审稿人,并在ACM上海分会担任秘书长、CSIG多媒体专委会担任副秘书长。

报告摘要:“自进化代表着大语言模型从“被训练的工具”走向“持续成长的智能体”的关键转变。其核心思想是让模型在与任务、环境以及自身输出的持续交互中,不断生成新的经验、发现新的问题、修正已有错误,并将这些反馈逐步内化为自身能力的增长。换言之,模型不再只是被动地学习人类提供的知识,而是能够主动探索自身能力边界,在尝试、反馈、反思和优化的循环中实现持续改进。这一过程使大模型具备类似“经验积累”的能力,从静态的知识载体逐步转向动态的学习系统。本报告将围绕大语言模型自进化这一主题展开,重点介绍其基本概念、研究动机和关键技术路径,包括评估、环境、harness以及基础模型等层次。通过这些方法,大语言模型有望在复杂推理、智能体任务和科学发现等场景中实现更强的适应性与自主提升能力。总体而言,自进化不仅是提升模型性能的一种技术路径,也为构建能够长期学习、持续适应和自主增强的人工智能系统提供了重要方向。

报告三:面向智能系统的神经交互接口

讲者:叶子逸 复旦大学

时间:8月16日 11:20-11:35



个人介绍:叶子逸,复旦大学可信具身智能研究院助理教授,分别于2025年和2020年在清华大学获得博士学位和学士学位。主要从事具身智能和脑机接口相关的科研工作,代表性工作包括基于过程记忆的具身操作规划,大模型的鲁棒可解释偏好对齐,国际上首个基于脑机接口的交互式搜索系统等。在Nature Comm. Biol.、ICLR、ICML、CVPR、SIGIR、Multimedia等国际一流期刊和会议上发表论文30余篇。相关研究内容与美团、字节跳动、五季医学、上海电气等行业头部企业建立深度联动。曾获中国中文信息学会优秀博士毕业论文,ACM上海分会优秀博士毕业论文,北京市优秀毕业生,CCIR2024最佳论文提名等荣誉。担任Neurips、ICLR、ICML、SIGIR、WWW、CVPR等多个权威国际会议的程序委员,以及Nature主刊、TOIS、TKDD等多个知名国际期刊的审稿人。

报告摘要:神经信号作为人类认知活动的信息表征形式,为构建新型人机交互接口提供了新的可能。随着大模型和智能系统的发展,如何建立人与人工智能之间更加自然、直接的信息交互通路成为重要研究问题。本报告介绍面向智能系统的神经交互接口探索:一方面,研究脑信号作为输入,通过结合生成式大模型实现开放场景下的语言意图解码、信息检索增强以及机器人控制,使脑信号能够承载更加开放和高层次的人类意图。另一方面,探索脑信号作为反馈,利用人类神经响应刻画大模型输出的认知风险,分析人脑与模型内部表征之间的关联,并进一步实现面向幻觉风险的智能系统调控。



报告四:可信推理驱动的大模型智能体

讲者:陈科海 哈尔滨工业大学(深圳)

时间:8月16日 11:35-11:50



个人介绍:陈科海,哈尔滨工业大学(深圳)信息学部计算机学院/计算与智能研究院教授,博导,国家级青年人才。获2025年度“吴文俊人工智能科学技术奖”科技进步特等奖、2024年度“钱伟长中文信息处理”科技进步一等奖、2022年度北京市科技进步二等奖和2020年度中国中文信息学会“优博”论文奖,主要研究方向为自然语言处理、大模型智能体、多模态可信推理等,在自然语言处理、机器学习 and 人工智能国际会议和期刊上发表论文百余篇。多次担任自然语言处理国际会议(高级)领域主席和人工智能国际会议资深程序委员,担任Neural Networks期刊编委。

报告摘要:本次报告聚焦可信推理驱动的大模型智能体,揭示当前智能体在多模态感知、强化学习优化与自主任务执行中普遍存在证据理解不充分、推理过程不稳定、奖励反馈不可靠和工具执行易出错等问题。首先从多模态结构理解与证据校验出发,介绍如何提升模型对图结构、表格、文档和图文信息的可靠感知能力。然后从强化学习与对齐视角,介绍偏好学习、奖励建模和策略优化如何增强模型推理行为的稳定性与可控性。最后,以自主智能体任务为例,介绍状态感知推理、工具调用、动态回退和多智能体协作等方法,探讨如何构建能够可信感知、稳定推理、可靠执行并持续进化的大模型智能体。



报告五:信息检索在深度研究智能体中的作用

讲者:孟川 爱丁堡大学

时间:8月16日 11:50-12:05



个人介绍:孟川,英国爱丁堡大学博士后研究员,博士毕业于荷兰阿姆斯特丹大学,师从ACM Fellow、荷兰皇家艺术与科学院院士Maarten de Rijke教授。主要研究方向为智能体搜索、对话式搜索与搜索质量评估。截至目前,已在SIGIR、ACL、EMNLP、NAACL、CIKM、ACM TOIS等国际会议和期刊发表论文30余篇。曾任亚马逊(Amazon)应用科学家实习生,相关研究成果已应用于亚马逊智能购物助手。在智能体搜索、对话式搜索与搜索质量评估等方向牵头组织8次国际会议教程与专题研讨会。曾受邀在英国皇家学会、伦敦大学学院等高校和研究机构作学术报告。担任SIGIR、ACL、EMNLP、WWW、WSDM、CIKM、KDD等国际会议程序委员会委员,以及ACM TOIS、IP&M、Pattern Recognition等期刊审稿人。

报告摘要:深度研究旨在回答需要对开放互联网进行广泛探索的复杂问题。为解决该任务,现有深度研究智能体通常生成查询、调用网络搜索API并阅读检索到的文档,重复上述过程,直至收集足够证据并生成答案。然而,网络搜索API通常是黑盒,现有信息检索方法在深度研究场景中的表现尚未得到系统研究。针对该空白,本研究从三个方面探究现有信息检索方法在深度研究中的表现:1)文档级与段落级检索单元的影响;2)重排序方法的有效性;3)智能体生成查询与检索模型训练查询之间分布差异的影响。实验发现,智能体生成的查询往往呈现基于关键词的网络搜索风格,这使稀疏检索和多向量检索方法更具优势;在上下文窗口有限的情况下,段落级检索单元可提升深度研究的效用,并避免稀疏检索在文档检索中面临的文档长度归一化问题;重排序方法能够大幅提升深度研究效用,并减少搜索调用次数。为缓解智能体生成的关键词式查询与当前检索模型训练中常用的自然语言查询之间的分布差异,本研究提出关键词到自然语言的查询改写方法,显著缓解查询不匹配问题。

报告六:基于图结构的大模型知识增强研究

讲者:张庆刚 吉林大学

时间:8月16日 12:05-12:20



个人介绍:张庆刚,吉林大学人工智能学院教授、博士生导师,国家高层次青年人才。本科毕业于西北工业大学计算机学院,博士毕业于香港理工大学计算机系,2026年5月加入吉林大学人工智能学院。长期从事大语言模型、检索增强生成、智能体记忆系统与自进化机制等方向的研究。近三年,以第一作者或通讯作者发表CCF-A类论文20余篇,涵盖NeurIPS、ICLR、ICML、KDD、ACL、TKDE、TOIS等机器学习 and 信息检索领域顶会顶刊。曾获 ACL 2025 亮点论文奖、香港理工大学计算机系优秀博士论文奖等。

报告摘要:大语言模型正在从“语言生成模型”迈向“知识驱动的智能系统”。检索增强生成(RAG)作为连接大模型与外部知识的重要技术路径,近年来受到广泛关注。然而,面对复杂领域知识、多跳推理以及可信决策需求,传统RAG仍存在知识表示不足、检索效率有限和生成不可靠等问题。本报告将以图结构知识增强为切入点,介绍GraphRAG如何通过实体关系建模和结构化检索突破传统RAG局限,并围绕GraphRAG评测体系、高效索引图构建、上下文忠实生成等方向分享最新研究进展。同时结合未来LLM Agent的发展趋势,探讨知识增强技术在智能体系统中的应用前景。

报告七:大模型规模定律的机制研究

讲者:詹靖涛 清华大学

时间:8月16日 12:20-12:35



个人介绍:詹靖涛,清华大学深圳国际研究生院助理教授。詹靖涛分别于2020年和2025年在清华大学电子系和清华大学计算机系获得本科学位与博士学位,主要从事信息检索模型、大模型等相关研究。在论文发表方面,詹靖涛在SIGIR、ICLR、ACL、WWW、WSDM、CIKM等重要国际会议上发表多篇论文,研究成果受到学术界广泛关注。受邀担任TOIS、SIGIR、WWW、WSDM、CIKM等多个国际一流学术期刊和会议的审稿人,担任YOCSEF深圳分论坛AC委员。荣获SIGIR 2024 最佳论文奖、WSDM 2022最佳论文奖、国家奖学金、北京市优秀毕业生等荣誉。

报告摘要:规模定律为理解和规划大模型发展提供了重要的经验框架:随着模型参数量、训练数据量和计算量的增长,语言模型的损失往往呈现可预测的幂律关系。近年来,规模定律已经成为连接资源配置、模型能力和系统设计的核心方法。本报告将围绕规模定律这一主题展开,首先介绍规模定律的历史,尤其是规模定律在其他领域中的表现形式和内在机制。其次,本报告介绍大模型领域中的规模定律在预训练配置、下游任务适配和能力涌现机制等问题中的应用;最后,本报告探讨规模效应背后的可能机制,包括数据表征、模型架构、优化目标等因素如何共同影响规模定律的形式。报告希望通过经验规律与机制解释的结合,为更可预测、更高效的大模型研究提供新的视角。

报告八:大模型在司法场景的应用和思考

讲者:刘泽阳 山东大学

时间:8月16日 12:35-12:50



个人介绍:刘泽阳,现就职于山东大学浪潮人工智能学院,担任助理教授,硕士生导师。于2024年获得英国诺丁汉大学博士学位。其主要研究方向为信息检索、自然语言处理和应用机器学习,研究内容涵盖对话式搜索系统评估、大模型信息检索增强等,已在SIGIR、TOIS、CIKM等信息检索领域高水平会议和期刊发表多篇论文,曾担任SIGIR、WSDM、CIKM、WWW、KDD、MM等国际高水平会议的程序委员会委员。现担任中国中文信息学会信息检索专委会通讯委员、青工委委员、大模型与生成专委会正式委员等。

报告摘要:构建可信的智能司法系统,要求大语言模型不仅能够准确理解与检索法律知识,更需在真实场景中表现出与其设计目标一致的行为特征。然而,现有研究在这两个核心维度上均存在系统性偏差:在人格评估层面,模型依赖心理量表进行自陈报告,评估的是内省标签而非可观测行为,与司法场景对AI行为可预期性的现实需求脱节;在法律检索层面,主流方法将查询与法条之间的语义鸿沟视为表层现象加以弥合,忽视了大陆法系的深层结构,导致检索本质上退化为词句匹配而非法律推理重构。两类偏差虽表现形式不同,但都反映了以表层近似替代结构对齐,并由此制约着智能司法系统的可信性边界。本报告从行为对齐与结构化检索两个互补视角切入,分别介绍BehaviorBench与tSTARD两项研究的理论框架与实验发现,探讨在更深层结构意义上重建智能司法系统可信性的技术路径。

讲习班

报告一:以构象为核心的生成式大分子建模

讲者:许洪腾 中国人民大学

时间:8月14日 14:00-15:30



个人介绍:许洪腾,中国人民大学高瓴人工智能学院副教授,国家高层次青年人才,2017年于佐治亚理工学院获得博士学位,2018至2020年担任杜克大学客座研究员,研究方向聚焦最优传输驱动的机器学习、大模型技术与科学智能。在科学智能方向,带领团队研发全球首个聚合物构象生成模型PolyConf和基准测试集。研究工作曾获得ICML2025 AI for Math研讨会杰出论文提名奖,入选《Environmental Science & Technology》副封面文章等荣誉。主持国家自然科学基金重大研究计划培育项目、中央引导地方人工智能专项等项目,常年担任ICML、NeurIPS、ICLR等国际会议领域主席,曾在KDD2019、AAAI2022、IJCAI2023、AAAI2026成功举办讲习班,获得国内外同行的关注。

报告摘要:针对蛋白质、聚合物为代表的柔性大分子进行构象生成和设计已经成为药物和材料研发领域的前沿方向。目前,针对这一方向的生成式人工智能技术面临模型泛化能力弱、通用性不足、计算效率低等挑战。本次讲座将介绍面向柔性大分子设计的生成式建模技术的进展,针对分子基态构象预测、蛋白质骨架生成、聚合物构象生成、聚合物属性预测等任务,介绍一系列模型架构和学习算法方面的创新性工作,特别是最优传输、几何代数等数学工具在柔性大分子建模方面的新应用。此外,本次讲座也将针对柔性大分子建模的技术路线发展趋势提供新的思考和观点。

录用论文列表

1. Jiayaqi Cheng, Zhongxiang Sun, Xiao Zhang, Jun Xu. 基于多智能体协作的检索增强生成可归因拒答方法
2. 张琼文, 张宇. FACT-SQL: 基于智能体的细粒度 Text-to-SQL 纠错方法
3. 田雨欣, 段增豪, 庞亮, 尹芷仪, 程学旗. SkillAtlas: 面向智能体技能漏洞的众包风险轨迹库
4. Xinyuan Cao, Qingyao Ai, Qingyi Pan, Bangde Du, Yiqun Liu, Min Zhang. 基于人口学特征参数化注入的人类价值观模拟方法研究
5. Jun Yang, Yixing Fan, Ruqing Zhang, Jiafeng Guo, Xueqi Cheng. 基于大语言模型的查询扩展再研究: 深度分析与策略增强
6. Bangde Du, Yueyue Wu, Qingyao Ai, Ziyi Ye, Wei Lin, Zhifeng Zheng, Min Zhang, Yiqun Liu. 面向大语言模型生成式问答的中文价值一致性评测基准
7. Chen Xu, Xiaopeng Ye, Baosong Yuan, Sunhao Dai, Teng Shi, Hengyi Jin, Weijie Yu, Jun Xu. 生成式人工智能信息分发的隐式不公平性研究与不确定性决策驱动的公平排序框架
8. Jingping Yan, Hengran Zhang, Jiafeng Guo, Keping Bi. 面向智能体检索的关键词查询: 稀疏检索与通用密集检索对比分析
9. 卜明仪, 张亚, 林原, 许侃, 王思翰, 杨亮, 林鸿飞. 基于三级递进提示推理与知识蒸馏的情感分析方法
10. Wenting He, Bing Li, Yixing Fan, Jiafeng Guo. 基于相对片段内注意力的领域模型词表扩展选词方法
11. Qiang Yan, Wenting He, Juan Chen, Jiafeng Guo. 面向金融监管问答的证据图检索增强方法
12. Yichen He, Xinyan Deng, Xiaorou Zheng, Shoubin Dong. 基于检索增强生成和患者级动态图推理的鲁棒临床预测
13. 徐泽西, 朱振方, 卢强. 一种基于语义先验增强长距离学习的视频摘要方法
14. Siqi Bao, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Dawei Yin, Xueqi Cheng. 面向检索增强问答的自适应证据选择与压缩方法
15. Haoxiang Li, Xiang Liu, Yu Bai. 基于语义锚点引导和近似对比学习的图文检索的方法
16. Hongjin Qian. 参数化知识激发能否替代长程推理? 以学术文献发现为例
17. Lin Chen, Zhen Li, Jie Liu, Cheng Chi. Planning-Guided Dynamic Knowledge Graph Exploration RAG for Multi-Hop Question Answering
18. Ruiqi Zhao, Haifeng Liu, Nan Zhao, Yanda Wang, Junsheng Zhou. Retrieval-Augmented Generation with Probabilistic Rewards and Dynamic
19. Zihui Liu, Yongbing Zhang, Shengxiang Gao, Zhengtao Yu. 基于实体锚点增强的低资源跨语言新闻推荐方法
20. 刘安琪, 郭文彬, 吕子钰. 面向检索增强生成的领域感知多智能体可信评估方法
21. Chenkai Qiao, Xing Gong, Guanglei Song, Xingyang Chang. TG-RAG: 基于意图验证实现多轮检索过程监督学习方法
22. Lichang Song, Yuxuan Liu, Ting Long, Yi Chang. PND: 一种专利新颖性检测基准数据集
23. Guohao Jian. 面向Top-K推荐的用户模拟器对比评估
24. Xuanang Chen, Ben He, Le Sun, Yingfei Sun. 从攻击到防御: 文本检索中对抗排序攻击的检测性防御研究
25. Yiwei Wang, Zhijing Wu, Yan Yuan, Xianling Mao. 面向外观专利审查的专利级多视图检索方法
26. Yatian Gao, Junyu Luo, Xueqin Ni, Xinyu Wang, Xinyan Shi, Piji Li. 大模型智能体复杂技术推理研究
27. 谢乐妍, 李逸聪, 臧天梓, 胥帅, 刘苧. FAIR-DECAF: 基于表征解耦与反事实监督的 LLM 推荐去偏方法
28. Xinyuan Yang, Shasha Guo, Liang Pang, Xinrui Tan, Xinran Liu. 基于图感知压缩增强的多轮对话意图理解框架

报告二: NLP到AI4Science的初探——基因组语言模型

讲者: 倪仕文 深圳理工大学

时间: 8月14日 15:30-17:00



个人介绍: 倪仕文, 深圳理工大学人工智能研究院副教授, 入选国家级海外引才计划。中国中文信息学会青工委委员, 中国计算机学会自然语言处理专委会委员, 斐陶斐荣誉学会荣誉会员。博士毕业于台湾成功大学计算机系, 曾任中国科学院深圳先进技术研究院博士后/助理研究员, 曾获深圳市优秀博士后, 台湾信息学会最佳博士论文佳作奖, 中国科学院深圳先进院“卓越员工奖”、“十大优秀博士后”奖、“院长博士后择优专项计划”、“优秀产业合作奖”、“优秀论文奖”等荣誉和奖励。近三年共主持10项国家/省/市/横向项目, 个人累计经费超500万。在自然语言处理、大模型等方向有着丰富的研究经历, 并取得了丰富的研究成果, 以(共)一作/通讯在ACL、ICLR、ICML、NeurIPS、AAAI、EMNLP、IEEE TASLP等顶级会议和期刊发表论文30余篇, 相关工作被DeepSeek、Qwen、KIMI、混元、豆包等头部大模型多次引用。担任NLP三大会议ACL, EMNLP, NAACL领域主席, 以及IEEE TASLP、ACL、EMNLP、AAAI、ICLR、NeurIPS等国际顶会和SCI期刊审稿人。此外, 申请人主导或参与研发的得理法律大模型、晓星教育&心理大模型、墨子知识产权大模型、顺丰财务审查大模型等多项技术均成功落地应用, 累计服务超10万人次, 经济效益超千万。



- 29.卢树乐, 王凌翔, 张沁楠, 温思嘉, 张海楠. FedDTRE:基于信任性评估的联邦对话生成模型
- 30.Yan Zhou, Yi Chang, Xin Wang, Yili Wang. 面向复杂专利匹配的多智能体动态推理方法
- 31.Haowen Li, Danyang Zong, Kaike Zhang, Du Su, Qi Cao, Fei Sun, Xueqi Cheng. 面向长尾交互的流行度去偏成员推断攻击
- 32.Hao Xiping, Chuanming Yu. 面向价值感知记忆的多智能体端到端科学发现
- 33.刘勘, 高欣怡, 任雨晴, 桂盼. LEAF:一种基于法律要素对齐的法条检索重排序框架
- 34.Kun Ai, Jiaxin Mao. 面向学科评估的长证据匿名全序排序:TourRank 适配与校准
- 35.Zhaoshuo Liang, Jiayu Sun, Zhao Cao. ThinkCal:同源错误检索驱动的智能体过程级持续校准
- 36.Min Hou, Tao Yu, Xin Liu, Hao Liu, Qiansong Zhang, Guangyi Xiao, Fei Liu, Le Wu. PRISM:基于共享-多分支低秩适应的个性化大模型构建方法
- 37.杜宜甫, 赵维纳, 马龙龙, 安波.《四部医典》关系约束与证据追溯问答方法研究
- 38.Wenxi Xiang, Jiancan Wu, Xiang Wang, Xiangnan He. Improving Training Data Efficiency for Tabular Foundation Models
- 39.Jiayu Cui, Tao Lian, Haocheng Dou. AdaPass: Adaptive Low-Pass and High-Pass Spectral Graph Filters for Complementary Recommendation
- 40.Weijen Liu, Jiarui Qin, Xu Huang, Xingshan Zeng, Yunjia Xi, Jianghao Lin, Chuhan Wu, Yasheng Wang, Lifeng Shang, Ruiming Tang, Defu Lian, Yong Yu and Weinan Zhang. Position: The Real Barrier to LLM Agent Usability is Agentic ROI
- 41.Boyang Yan, Yunqi Mi, Guoshuai Zhao, Chengxu Liu, Xueming Qian. Multi-view Embracing User Heterogeneity in Social Recommendation
- 42.Tengjiao Li, Yunqi Mi, Chengxu Liu, Xueming Qian, Guoshuai Zhao. MEHP: Multi-Grained Evidence and Hierarchical Priors for Multimodal Intent Recognition
- 43.Yizhe Li, Zipeng Fan, Xindian Ma, Peng Zhang. ETPDiff: Evolution Trajectory Prediction Diffusion Model for Influenza Virus
- 44.Yang Yang, Bo Xu, Xiaokun Zhang, Yuan Lin, Kan Xu, Hongfei Lin. FI-RAG: Failure-Induced Evidence-Use Control for Multi-Hop Retrieval-Augmented Generation
- 45.Haochuan Qian, Daoxuan Liu, Jiancan Wu, Junchao Gao, Zhe Lv, Yongbo Yao, Xiangnan He. Evaluating Large Language Models for Children's Story Generation and Rewriting
- 46.Kun Zhang, Jinglong Li, Le Wu, Hailin Zhou, Richang Hong. Bias-Amplification and Bias-Elimination Framework for Debaised User Modeling and Cognitive Diagnosis
- 47.Diru Han, Yingchen Zhang, Ruqing Zhang, Jiafeng Guo. Learning to Memorize Less: Scaling Generative Retrieval with Representative Documents
- 48.Yuyue Zhao, Mingjun Xie, Weijian Chen, Jiancan Wu. Attribute Bias in Collaborative Filtering: An Invariant Contrastive Learning Approach
- 49.Cheng Yang, Jinda Lu, Jinghan Li, Xiang Wang, Jiancan Wu. Beyond Static Rubrics: RubricRefiner for Discriminative Rewards in GRPO Training



赞助单位



金牌赞助



银牌赞助





「高效」是面壁发展的第一性原理

面壁智能是以「高效」为第一性原理的大模型公司，专注高效大模型的全栈研发、产业应用与生态构建，致力于成为端侧智能技术引领者。面壁把同等参数、性能更高、成本更低的高效大模型，放到离用户最近的地方，面壁「小钢炮」MiniCPM 系列端侧模型规模化赋能汽车人机交互系统、智能家居、AI手机、AIPC等终端。面壁攻关普惠的通用人工智能，助力终端产业智能化升级与垂直行业提效革新，加速智能平权普惠进程。



李大海

面壁智能 联合创始人 & CEO
CCF 杰出会员、理事

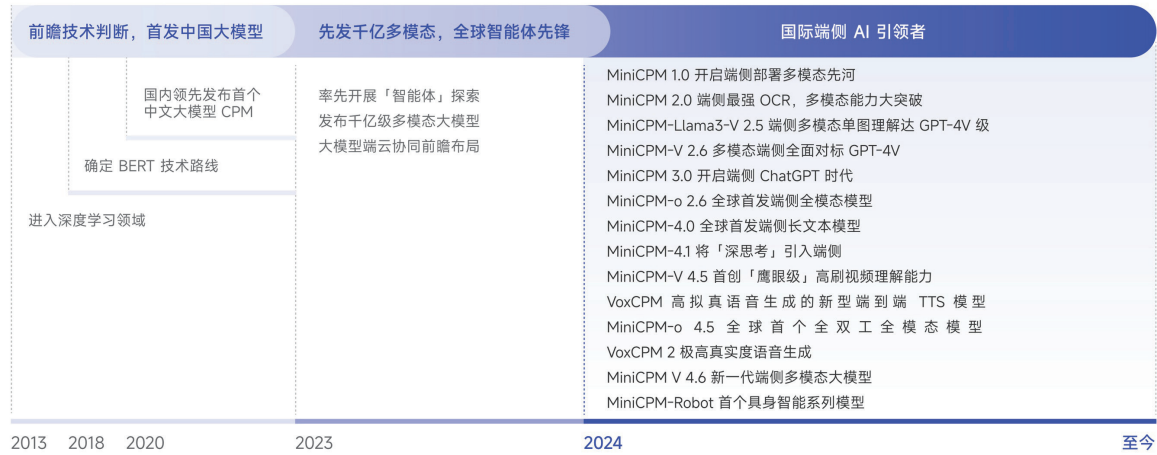


刘知远

面壁智能 首席科学家
清华大学 长聘教授

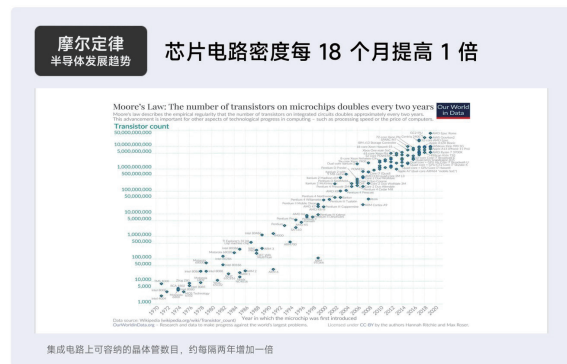
走先人一步的路，打以少胜多的仗

面壁智能研发团队具有鲜明的「先锋」特色。通过对大模型本质规律的深刻探索，团队一次次对人工智能技术发展方向进行精准预判，展现出惊人的战略判断力。



高效，来自对大模型科学化的不懈探索

面壁是大模型演进的「密度定律」的提出者，坚信模型知识密度与芯片算力密度的双向跃升，必将加速实现普惠的通用人工智能，最终实现“智周万物”的公司愿景。



面壁「小钢炮」端侧大模型系列，是全球领先的轻量高性能大模型，以「以小博大」的性能和卓越端侧表现，受到世界开源社区的广泛赞誉，接连登顶 GitHub，Hugging Face 热度榜，并成为 Hugging Face 2024 年度最受欢迎中国大模型，全平台下载量累计破 2400 万。面壁「小钢炮」携手行业标杆合作伙伴，成为 AIPC、AI Phone、智能座舱、具身机器人等行业创新发展中不可或缺的角色。



适配国内外主流芯片



携手全球合作伙伴





BAIDU

Founded in 2000, Baidu's mission is to make the complicated world simpler through technology. Baidu is a leading AI company with a strong Internet foundation. We are one of the few companies worldwide operating on a full-stack layout in AI. Driven by the company's deep learning platform PaddlePaddle and pre-trained model ERNIE with real-world applications across industries, Baidu stands at the forefront of the industry, achieving end-to-end optimization and significantly improving operational efficiency.

Over the past decade, we've invested nearly 190 billion RMB in research and development. This sustained investment has propelled us to the forefront of the AI landscape, with pioneering self-developed technologies and leading positions in AI patent applications.

Baidu has long been at the forefront of deep learning research. Its PaddlePaddle, open-sourced in 2016, is the first open-source deep learning framework in China. Its ERNIE 1.0, released in 2019, marks a significant step forward in LLM development. Based on these technological advancements, its AI chatbot "ERNIE Bot" was developed and publicly launched in 2023. Baidu's open-source reasoning model, ERNIE-4.5-21B-A3B-Thinking, topped Hugging Face's trending models list in 2025. In January 2026, Baidu officially released its foundation model ERNIE 5.0. In May 2026, Baidu launched ERNIE 5.1 to further enhance ERNIE's capabilities and performance.



文心大模型

BAIDU WENXIN

A series of knowledge-enhanced large models, from general-purpose to industry-specific, independently developed by Baidu, speed up AI innovations and empower the industry to upgrade.

ERNIE

