

教育部工程研究中心年度报告

(2025年1月——2025年12月)

工程中心名称：新一代智能搜索与推荐

所属技术领域：信息与电子工程

工程中心主任：文继荣

工程中心联系人/联系电话：王墨/18810775171

依托单位名称：中国人民大学

2026 年 5 月 11 日填报

一、技术攻关与创新情况

2025年，中心积极响应国家面向搜索和推荐“信息技术新基建”的战略需求，继续围绕新模型、新任务、新构架、可信可控和工程化实践五个方向开展深入的有组织科研。年度内，中心依托高瓴人工智能学院师生发表高水平论文77篇（含CCF-A类53篇、Nature子刊1篇）。此外，中心新增国家级项目9项、省部级项目9项，批准总经费达2270.60万元，新增横向课题71项，合同经费共计3899.25万元。同年，中心实际到账经费总计4181.79万元（纵向2026.69万元、横向2155.10万元），为核心技术攻关提供了坚实支撑。中心固定人员69人（含研究人员57人，技术人员1人，行政人员11人），人工智能方向专任教师25人，该方向国家级高层次人才12人，形成了由学术带头人、青年骨干、工程化人员和交叉学科团队共同组成的研究队伍。年度内，中心副主任魏哲巍教授入选国家“万人计划”领军人才，中心骨干李崇轩副教授获批国家自然科学基金青年科学基金项目（B类）。主要在以下五个关键方向进行技术攻关与创新：

1. 面向搜索的新模型研究：突破生成式检索底层范式

该方向由技术带头人为徐君教授。2025年，团队重点攻关扩散语言模型（DLM）的基础理论与关键技术，成功发布“玉兰-LLaDA”扩散语言大模型。该模型打破了传统“文字接龙”式的自回归范式，转向随机扩散生成，在性能上全面对标同规模代表性模型。目前系列模型下载量已超400万次，GitHub星标破3500，相关成果获ICLR 2025生成模型研讨会唯一最佳长文奖，并入选NeurIPS

2025 Oral，有力推动我国在基础模型底层架构上的原创性探索。

2. 智能搜索的新任务探索：支撑社会模拟等重要任务

该方向的技术带头人为孙浩教授。构建了“玉兰-万象”多智能体社会模拟迭代体系，填补了国内在复杂社会系统推演基础设施方面的空白。团队通过技术攻关，形成了由 YuLan-RecAgent（推荐智能体）、YuLan-GenSim（生成式仿真）到 YuLan-OneSim（通用社会模拟）的完整迭代路径，构建起一个具备自我演化能力的“AI社会科学家”科研闭环系统。目前，该平台已具备极高的工程化成熟度与跨领域适配性，支持八大核心领域，精细化模拟场景达 50 个。实现了十万级超大规模智能体的实时并发模拟与复杂交互，展现了卓越的分布式计算效率与逻辑一致性，为观察复杂社会现象、推演政策影响提供了高保真的“数字孪生”环境。相关核心技术已深度支撑社区治理智能体“亿心通”的建设与部署。

3. 搜索系统新构架研究：从RAG走向深度搜索智能体

该方向的技术带头人为窦志成教授。团队致力于推动信息检索从简单的检索增强（RAG）向深度搜索智能体演进：研制开源框架 FlashRAG，以模块化方式集成了23种RAG方法，并深度适配华为 MindSpore、百度PaddlePaddle等国产开发框架，GitHub星标突破2000次。此外，构建了具备“边思考、边搜索、边写作”能力的深度研究智能体WebThinker。此外，针对金融投研场景研发的“玉兰-融观”系统，在AFAC2025金融智能创新大赛中从1289支队伍中脱颖而出，摘得挑战组冠军。

4. 可信可控搜索和推荐技术：兼顾公平性与司法实务应用

该方向的技术带头人为丁晓东教授。针对互联网平台长期存在

的“马太效应”及分发偏差痛点，创新性地提出了一种受税收机制启发的公平检索模型。核心技术已成功转化为实践成果，深度赋能中心研发的“慧审”可信模拟法庭智能体系统。通过引入公平检索与可信推理框架，该系统实现了司法判例的精准调取与审理逻辑的闭环模拟，显著提升了法律实训的透明度与专业化水平。凭借卓越的技术前瞻性与社会应用价值，该项目在首届人工智能与智慧治理大学生创新创业大赛中脱颖而出，摘得全场一等奖。

5. 新一代智能搜索与推荐工程化实践探索：赋能涉外法治

该方向的技术带头人为赵鑫教授。重点推进了“涉外法治大模型”的建设与应用，通过深度数据治理与算法创新，构建了全球领先的涉外法治知识中枢。该大模型目前已成功汇聚涵盖全球24个主权国家、共计32,624部法律法规文件的海量数据资产，包含逾149.6万条原始法条及30万条高标准预设微调指令，夯实了模型在跨法域应用中的知识广度与专业深度。该系统目前已深度服务于跨国合规审查、跨法域法律比对等复杂实战场景，为我国涉外法治的智能化建设、提升涉外法律服务的全球竞争力提供了关键的技术底座与辅助决策支撑。

二、成果转化与行业贡献

（一）总体情况

2025年，中心围绕新一代智能搜索、推荐、大模型与智能体等方向持续推进成果转移转化，新增横向课题71项，合同经费共计3899.25万元。合作对象覆盖华为、快手、腾讯、阿里、抖音、OPPO、高途、美团、苹果、淘宝、小米等头部企业和机构，形成了

“联合实验室+联合项目+开源社区+竞赛验证”的转化路径。年度内中心新增三个校企联合实验室：高途-中国人民大学AI智慧教育联合共建实验室、中国人民大学-OPPO人工智能终端技术联合实验室以及中国人民大学-百融智能智能体联合共建实验室，进一步把智能信息分发、未来媒体智能、多模态交互和智慧教育场景纳入常态化联合攻关。通过LLaDA、FlashRAG、WebThinker、玉兰-万象、涉外法治大模型等成果，中心在开源工具、行业应用、智慧治理和人才培养方面持续扩大影响力，对人工智能底层软件生态、区域数字治理和产业智能化升级形成了明显支撑。

（二）工程化案例

2025年度，工程研究中心通过高强度的有组织科研，成功将大批前沿基础理论转化为能够解决实际行业痛点、具备显著产业推广价值的系统级工程化产品与平台。以下重点详述五项最具代表性、影响力最广的典型工程化案例。

（1）典型工程化案例一：高校自主研发的玉兰（YuLan）基础大模型矩阵与YuLan-Mini高效预训练系统

①工程化背景与技术痛点

大语言模型（LLM）已成为新一代人工智能的核心操作系统。然而，大模型的研发是一项极度复杂的系统工程，涵盖海量数据清洗、架构设计、分布式预训练及后训练（含指令微调与价值观对齐）等无数底层细节。长期以来，工业界将高质量的训练数据配比及稳定训练策略视为核心机密。受限于算力规模与系统级工程经验壁垒，国内高校及科研机构的研究往往只能依赖于对已有开源工业模

型的二次微调，极少具备从零开始进行预训练到后训练全链条自主研发的能力。这导致学术界和广大中小企业面临“炼丹门槛极高、底层不可控、开源模型性能上限低”的困局。如何在极度有限的计算资源与数据量级下，研发出性能比肩工业界同规模模型的高效轻量化基座模型，是迫切需要解决的底层工程痛点。

②核心工程化技术与突破

工程研究中心团队打破了对工业界开源基座的依赖，依托自有算力集群从零代码起步，依靠高校自主科研力量完成了中英双语“玉兰（YuLan）”系列大模型的研发。该系列是国内乃至国际上为数不多的、由高校自主完成预训练和后训练全流程的基础大模型矩阵。从YuLan-1.0逐步迭代至参数规模达120亿的YuLan-3.0，中心构建了“多阶段课程学习”的指令微调机制与价值观高效对齐方法。

在此全链条自主研发经验的基础上，中心集中攻关，历时5个月，仅调用56张GPU算力卡，在经历1700余次实验迭代与2次系统重启后，成功推出了24.2亿参数的轻量化端侧基座模型YuLan-Mini-2.4B。该模型仅在1.08万亿（1.08T）词元上完成预训练，使用了高效的预训练方法，实现了三项重大底层工程突破：

精心设计的数据流程：将数据清理与动态数据调度策略深度结合，构建了涵盖去重清洗、规则筛选、污染检测等分层过滤系统；在预训练进程中动态调整各垂直领域数据配比，并规模化引入了基于Lean定理证明器的形式化数学增强数据及自主合成的高质量数据。

强大的优化方法：针对算力受限下需大幅提高学习率导致的训

训练不稳定问题，团队深入剖析了隐藏状态逐层递增及RMSNorm表示坍塌等导致训练崩溃的原因。通过工程化引入重参数化方法与 μP (Maximal Update Parametrization) 参数初始化策略等手段，有效缓解了训练不稳定性，实现了高学习率下的极速稳定收敛。

有效的退火方法：在预训练后期，创新性地结合了有针对性的目标数据选择与长文本 (Long-context) 训练进行退火，进一步释放并巩固了模型在长上下文理解与复杂逻辑任务中的潜力。

③工程化实施与应用效益

在基础模型底层创新方面，中心通过数据工程与算法优化的深度协同，成功研发并发布了轻量化基座模型 YuLan-Mini。该模型展现出卓越的训练能效，在仅消耗 1.08T 词元的预训练约束下，其综合性能表现已与 Qwen2.5-1.5B 等业界顶尖同规模工业模型持平。尤其在“数学推理”与“代码编写”两个反映逻辑思维能力的核心领域，YuLan-Mini 展现出显著的性能优势。这一成果不仅验证了高校团队在算力受限背景下实现高数据利用率的底层创新能力，更为大模型的能效优化提供了技术标杆。

为打破大模型训练过程中的“黑盒”壁垒，中心坚定践行“全面开源、极致透明”的科学路线，旨在为全球研究者在退火数据策略 (Data Annealing) 及课程学习 (Curriculum Learning) 等前沿领域提供实证支撑。中心完整公开了 YuLan-Mini 预训练全生命周期的技术细节，开创性地实现了三项“首次”：1) 数据配比全透明：首次公布了涵盖各训练阶段的详细数据组成及动态演化方案；数据源全公开：2) 全面披露了中心自研的高质量合成数据与开源数据的融合路径；3) 训练状态全共享：史无前例地开放了预

训练过程中所有的中间检查点（Checkpoints）及对应的优化器状态。这一举措有效填补了学术界在高性能大模型全生命周期训练透明化方面的空白，有力推动了人工智能领域的开放科学进程。

在成果转化与社会服务层面，YuLan-Mini 及其核心调优技术已产生广泛的行业辐射效应。该成果在工程实践上的先进性已获得工业界高度认可，相关底层优化技术已被百川智能、蚂蚁集团等领军企业吸收，并深度应用于其自研模型的底座迭代进程中。同时，YuLan-Mini 已成功在中国人民大学高考招生智能问答系统等真实场景中完成工程化部署，在服务社会民生与提升高校治理智能化水平方面发挥了重要作用。该项成果的落地，显著彰显了工程研究中心作为国产基础大模型底层技术自主创新“核心骨干”的战略引领作用。

（2）典型工程化案例二：校企合作探索大模型新架构的扩散大语言模型LLaDA系列

①工程化背景与技术痛点

大语言模型（LLM）已成为新一代人工智能的核心操作系统。然而，当前主流大模型几乎全部依赖自回归生成机制，其生成过程基于逐词预测，虽然推动了模型能力的快速提升，但也面临推理速度受限、全局双向建模不足、复杂推理能力受生成顺序约束等瓶颈。特别是在逆序续写、反向推理以及多模态统一生成与理解等任务中，自回归架构的局限性日益显现。与此同时，扩散模型在图像、视频等连续模态生成中已取得显著进展，但在离散语言空间中长期面临训练目标复杂、稳定训练困难、尺度扩展规律不清晰、采样效率偏低等关键问题，尚未形成成熟可用的大模型技术路线。如何突

破自回归单一路径约束，构建兼具理论创新性、工程可实现性和产业可扩展性的新一代大模型生成架构，是迫切需要解决的重要底层工程痛点。

②核心工程化技术与突破

工程研究中心团队围绕扩散语言模型这一新型架构方向持续攻关，联合企业合作伙伴形成了从基础理论、规模扩展到模型研发、后训练优化和多模态拓展的完整技术链条，成功推出了LLaDA系列模型。该系列工作率先打破了自然语言生成长期由自回归模型主导的技术路径，是高校在新一代大模型生成架构方向上取得的系统性原创突破之一。

在此持续研发基础上，团队围绕扩散语言模型训练与实现中的关键瓶颈，实现了多项重大底层工程突破：

统一的扩散建模方法：针对离散扩散目标复杂、时间条件依赖强、训练与采样实现繁琐等问题，团队提出了RADD方法，从理论上揭示了吸收式离散扩散与干净数据条件分布之间的内在联系，将原有建模目标重参数化为无需显式时间条件的统一形式，显著降低了模型训练与采样复杂度，为扩散语言模型的大规模稳定实现奠定了基础。

清晰的规模扩展规律：针对扩散语言模型长期缺乏可复用训练范式、难以支撑大规模工程实现的问题，团队系统研究了文本masked diffusion的性能演化行为，总结并验证了扩散语言模型的scaling law，揭示了其在参数规模、数据规模和计算规模增长过程中的扩展特征，为后续模型设计、算力配置和训练组织提供了关键依据。

完整的系列模型研发与迭代：在上述理论和方法积累基础上，团队于2025年2月发布可对话扩散大语言模型LLaDA 8B，系统验证了扩散机制能够在开放式语言生成场景下取得与同规模自回归模型相竞争的性能。LLaDA 8B基于掩码扩散建模，从全局双向上下文出发进行语言生成，在数学推理、代码生成、问答、多语言等近20项标准任务中整体表现媲美同规模自回归模型LLaMA3 8B，并在逆序续写、反向推理等挑战性任务上甚至超越GPT-4o等领先闭源模型。此后，团队进一步推出LLaDA 1.5，提出面向扩散语言模型的方差约简偏好优化方法，显著提升了模型在数学、代码、问答和指令跟随等任务上的综合表现。随后，团队构建了LLaDA-V，形成纯扩散式多模态大模型方案，将扩散语言建模能力拓展到视觉语言统一建模场景，在图文问答与多轮对话等任务中展现出优异性能。在此基础上，团队与蚂蚁集团进一步合作发布LLaDA 2.0，持续推动扩散大语言模型向更大参数规模拓展，进一步验证了该技术路线在大规模训练、系统优化和工程落地中的可扩展潜力。

③工程化实施与应用效益

通过基础理论创新、训练方法设计与系统工程实现的深度协同，LLaDA系列成果实现了从底层建模方法、规模扩展规律到基座模型、对齐版本和多模态版本的全链条贯通，展现出高校与企业围绕新一代大模型架构开展协同创新的系统研发能力。LLaDA 8B作为全球最早的可对话扩散大语言模型之一，显著拓展了扩散生成机制在语言场景中的适用边界；LLaDA 1.5、LLaDA-V和LLaDA 2.0的持续推出，则进一步验证了该技术路线在后训练优化、多模态建模和大规模扩展方向上的工程延展性。

为了推动扩散语言模型技术路线的广泛传播和后续研究，团队坚持开源开放路线，持续开放LLaDA系列模型、训练方法和相关技术实现。相关研究获ICLR 2025生成模型研讨会长文最佳论文奖，开源项目GitHub星标超过3500枚，系列模型累计下载量超过400万，已成为国际扩散语言模型方向最具代表性的开源成果之一。该系列工作的核心思想和关键技术已对产业界探索新一代大模型生成架构产生重要影响，也为高校与企业围绕底层架构创新开展联合攻关提供了成功范例，充分彰显了工程研究中心在国产大模型底层架构创新中的骨干引领作用。

（3）典型工程化案例三：从检索增强基础架构到深度研究智能体的全链路工程体系

①工程化背景与技术痛点

在大语言模型时代，检索增强生成（RAG）已成为解决模型事实幻觉、接入外部知识的核心工程方案。然而，该领域在向复杂工业应用演进的过程中面临两大核心痛点：其一，底层研发基础设施高度碎片化。市场上现有的商用框架（如LangChain、LlamaIndex等）结构过于庞杂臃肿、底层封装过深，极度限制了科研机构与企业开发者对算法组件的灵活定制与标准化公平对比；其二，面对需要长时间跨度、多步逻辑推导、海量多源信息采集与交叉验证的“深度研究（Deep Research）”任务，传统的单次检索增强系统往往无能为力，极易产生逻辑断裂，亟需从“简单信息检索”向具备长链路规划能力的“深度研究智能体”跨越的系统级工程化解决方案。

②核心工程化技术与突破

工程研究中心团队秉承模块化与标准化的顶级软件工程理念，系统性构建了从底层RAG支撑库到顶层自主深度研究智能体的全栈工程体系：

大模型检索增强底层基础设施FlashRAG：专为学术前沿探索与轻量化工业部署设计。框架将复杂的RAG系统高度解耦重构，拆解为检索器、重排器、大模型生成器、精炼器、评测器五大可自由插拔的核心组件。平台高度格式化地预处理了36个国际公认的基准RAG数据集，并内嵌复现了23种最先进的RAG算法（如Self-RAG、ActiveRAG等）。开发者可通过极简配置一键调用底层加速库，为上层复杂智能体的开发提供了坚实、标准化的信息获取底座。

深度搜索与研究智能体引擎WebThinker与Search-ol：在标准化的RAG基座之上，团队进一步攻克了智能体的长链路规划与自主探索技术。中心研发了Search-ol深度搜索智能体与WebThinker智能研究系统，在模型中引入了复杂任务自主拆解、外部多工具协同动态调度与多信源交叉验证机制。这赋予了智能体深度探索与长链路信息分析能力，使其能够自主应对极度开放和复杂的长周期信息搜集与综合分析任务。

③工程化实施与应用效益

该全链路工程体系在开源生态构建、通用能力评测及垂直产业赋能三个维度均取得了卓越的效益：

在底层开源生态建设上，FlashRAG发布后引发国际开源社区热烈反响，短短数月内GitHub星标突破2000次，数据集下载量破万。它比同类商用框架更轻量且贴近学术前沿，已成为国内外众多高校实验室及中小科技企业快速构建、评估本地RAG系统，以及进行算

法商业化概念验证（POC）的首选标准脚手架平台。

在通用智能体效能提升上，依托WebThinker与Search-ol系统，大模型将原本需人工耗时数小时的复杂多源信息采集、深度分析与交叉验证任务压缩至数分钟内完成。在衡量AI解决真实世界复杂问题能力的通用智能体权威评测基准GAIA中，实现了整体性能较同类基线模型提升21.9%的显著跨越。

在垂直产业商业化应用中，中心将该深度研究智能体引擎面向高价值的金融投研场景进行工程化定制，衍生开发了“玉兰·融观（Yulan-FinSight）”多模态金融研报自动生成系统。该系统实现了从多源异构海量金融数据自动采集、深度专业逻辑分析到最终图文并茂的专业级金融研究报告生成的全流程自动化。在AFAC2025金融智能创新大赛中，该系统从1289支顶尖参赛队伍中脱颖而出，一举夺得挑战组总冠军，全面展现了该工程方案赋能复杂产业深度分析的巨大商业价值与应用潜力。

（4）典型工程化案例四：具有公平性约束与可解释、可溯源的智慧司法系统及涉外法治大模型

①工程化背景与技术痛点

“同案同判”是维护司法公正与权威的重要基石。随着智慧法院建设的推进，各地沉淀了海量数字卷宗，但传统基于深度匹配的AI类案推荐系统具有严重的“黑盒”属性，法官无法理解算法的推荐依据（缺乏可解释性与法理溯源性）；且深度学习模型极易放大历史司法数据中因地域或当事人属性带来的长尾统计偏差，导致推送的不公平性隐性歧视。此外，在国家统筹推进国内法治和涉外法治建设背景下，跨境法律需求呈爆发式增长，而国内亟缺能够准确

理解多国复杂法律体系、支持跨法域逻辑推理的高水平智能化工具。

②核心工程化技术与突破

工程研究中心深度整合法学院与人工智能学院的交叉学科力量，联合中国司法大数据研究院，依托国家重点研发计划，系统性地研发了司法专属的检索与生成模型体系，完成了两大政务级平台建设：

可解释、可溯源的公平性约束类案推送系统：针对系统公平性与解释性痛点，工程团队引入了“语言智能+经济学”工具变量方法量化提取司法裁判要素。在推送逻辑中，引入强化学习中的探索-利用机制动态平衡推送特征，在跨任务迁移学习中植入公平性惩罚因子，有效调平了干扰变量的分布偏差，确保了长尾案例的公平曝光与匹配。在解释溯源上，利用逆向最优传输、神经-符号机及异构信息网络技术，将家事纠纷、金融借贷等复杂案件特征解析为多源语义图谱，通过图神经网络（GNN）执行显式逻辑推理，将黑盒推荐结果还原为清晰的可视化法理溯源路径。

涉外法治大模型研发：团队系统性汇聚、清洗了覆盖全球24个国家、3.2万余部法律文件、149.6万条法条及超4810亿词的高质量多语种法律专业语料库。基于该庞大数据底座进行深度预训练，并利用定制的高质量涉外法治指令与价值观数据进行对齐微调。工程上自研了多通路、多粒度的专业检索增强架构（RAG），使大模型能够实时精准动态接入外部各国法律知识库，支持跨法域法律规则严密比对与高可信度的慢思考多语言推理。

③工程化实施与应用效益

该系列工程化成果取得了显著的政务与社会效益。依托国家重点研发计划转化为标准化政务SaaS产品的“具有公平性约束和可解释、可溯源的类案分析与推送系统”，已成功与地方法院的审判办案内网系统实现无缝接口对接。该系统目前在安徽省淮北市中级人民法院、亳州市中级人民法院及六安市中级人民法院等多地完成了长期的规模化示范应用。系统不仅实现了高精度的在办案件自动类案推送，更以逻辑图谱直观展示了判例相似依据，获得了法院方面出具的高度认可的用户使用报告，显著提升了法官的裁判效率与审判尺度的统一。

同时，“涉外法治大模型”已在十余家涉外法律服务机构及专业研究智库单位上线试运行。在涉外合规审查、跨国贸易争端应对与国际规则研究等高价值业务场景中发挥了“数字智囊”的关键作用，有力提升了我国涉外法律服务的智能化水准与国际纠纷应对能力。

（5）典型工程化案例五：“玉兰-万象”超大规模多智能体社会模拟与智慧治理平台

①工程化背景与技术痛点

在宏观社会科学研究、公共管理政策制定与商业平台复杂规则设计中，传统的“以真人为基础”或简单动力学模型的实验范式受限于伦理、高昂成本且样本量极小，难以进行大规模的反事实推演与长期预判。同时，在人工智能赋能人文社会科学（“AI+社科”）的转型浪潮中，传统文科师生面临极高的技术应用壁垒。如何在虚拟数字空间中，利用大语言模型构建具有高逼真度、个体异质性强、可自主演化的大规模人类行为模拟环境（即“计算社会科学

”仿真引擎），将社会科学的前沿议题系统性地转化为可计算的智能应用，从而准确预判公共政策或平台算法在真实人群中引发的宏观级联效应，是一项极具挑战的交叉学科系统工程。

②核心工程化技术与突破

中心基于强大的玉兰（YuLan）大语言模型底座与智能体（Agent）技术，在深刻理解社会科学机理与文化语境的基础上，研发了具有泛化协同能力的用户行为模拟平台RecAgent及全面升级版的“玉兰-万象”多智能体社会模拟计算平台，为社会科学研究提供了坚实的技术支点：

高并发智能体沙盒底层引擎：平台打破了传统仿真算力限制，开发了支持十万至百万级大语言模型智能体高效并行仿真的分布式计算引擎。工程团队将强化学习与认知心理学机制深度融入智能体架构，使每个被实例化的智能体均具备独立的异质化人格画像、记忆流模块（包括短时感觉记忆与长时经验记忆）及反思规划逻辑。智能体能够在高度拟真映射的虚拟物理与社交空间中自主进行信息消费、观点传播、经济交易与社会博弈演化。

全流程自然语言驱动与社科场景沉淀：为打破社科研究的代码壁垒，平台实现了“自然语言需求录入 -> 自动生成底层仿真代码 -> 分布式场景执行 -> 自动生成量化分析报告”的全流程零代码闭环，显著降低了人文社科学者的技术使用门槛。平台底层预置了覆盖社会学、经济学、公共管理、新闻传播等8大领域的50个典型复杂宏观仿真场景，形成了高度匹配社会科学问题特性的技术框架。

③工程化实施与应用效益

“玉兰-万象”平台开创了社会科学智能化研究的新模式，不仅在学术界成功复现了“信息茧房演化”、“网络舆情极化”、“从众行为”等复杂的宏观群体动力学现象，更在赋能“AI+社科”交叉创新、国家宏观智慧决策及商业验证中取得了卓越效益：

驱动“AI+社会科学”教育创新与交叉学科落地：该平台作为核心技术底座，全面支撑了第一届中国人民大学“百融杯”AI智能体创新应用大赛。大赛专设了“大模型驱动的社会模拟与复杂决策”等赛道，吸引了全校15个学院的63支师生队伍参赛。依托平台的底层赋能，多项“AI+社科”前沿应用成功孵化并扎实落地：例如荣获全场唯一“最佳创意奖”的“亿心通教学智能体”，将大模型技术与基层社区治理场景深度融合，直击传统社会工作与公共管理教学中“理论脱离实践、案例静态陈旧、决策训练缺失”的核心痛点，引导学生在与AI的互动仿真中亲身体验构建社会治理共同体的复杂过程；此外，平台还支撑研发了“基于AI的生成式量化研究方法课程”、“面向非遗影像的多模态结构化解析与智能检索系统”等多个获奖应用。这充分展现了该平台在赋能社会科学复杂应用孵化与科研范式转型中的关键支撑能力。

赋能国家与地方政府宏观智慧决策：在公共管理政务领域，团队利用该系统为特定城市建立了“公租房资源智能分配”的宏观模拟沙盘。平台高精度推演并预测了在不同分配政策（如摇号、积分等）实施后，不同收入及社会背景群体在地理分布与社交网络中的群体演化行为及潜在不满情绪，为政府部门优化公共资源配置、提前防范社会治理风险提供了强有力的“数据与机理双驱动”的辅助决策支撑。

支撑工业界商业验证与高水平学术产出：在工业界商业领域，该平台作为“零真实风险”的数字验证测试场，与华为、快手等合作企业深度对接，为推荐系统的流量分配机制验证、广告精准投放策略及舆情干预机制提供了低成本、高仿真的模拟环境。依托该平台生成的社科交叉科研成果荣获了国际顶级会议ICAIS的“最具社会影响力论文奖”，标志着我国在AI赋能宏观社会计算与政策推演领域的工程化水平全面迈入国际领先行列。

（三）行业服务情况

（1）本年度与企业的合作技术开发

2025年中心新增横向课题71项，合同经费共计3899.25万元。按合作单位看，华为7项，758.90万元；OPPO 4项，402.89万元；抖音9项，510万元；高途7项，250万元；快手5项，250万元；阿里巴巴5项，250万元；腾讯7项，235万元；其他单位27项，1242.46万元。合作内容覆盖下一代智能信息分发、未来媒体智能、大模型技术、多模态交互、智慧教育和智能搜索系统。

（2）提供技术咨询情况

中心依托联合实验室和开源项目，为合作企业和机构提供面向RAG系统构建、深度搜索智能体训练、推荐公平性、多模态内容理解、法律知识检索和智能治理场景落地的技术咨询。通过FlashRAG、WebThinker等项目维护文档、基准和示范流程，中心持续向学术界、产业界输出可复现、可迁移的技术方案。

（3）为企业开展技术培训情况

中心围绕人工智能、大模型、智能搜索与推荐、智慧教育等方

向开展技术培训和联合研讨。高途-中国人民大学AI智慧教育联合共建实验室等平台将科研成果与教育场景结合，推动面向教师、学生和企业研发人员的技术交流；中心师生也通过开源社区说明文档、专题报告和竞赛指导，提升合作单位对RAG、智能体、多模态模型和可信AI技术的应用能力。

（4）参加行业协会、联盟活动情况

2025年中心成员积极参加中国计算机学会、中国人工智能学会及相关学术与产业组织活动。窦志成获CCF ADL优秀学术主任，骨干黄文炳副教授团队获CCF产学研合作基金优秀项目案例，骨干林衍凯副教授获吴文俊人工智能科学技术奖青年科技奖。中心相关成果在众多国际会议和研讨会中获得奖项或重要展示，进一步提升了行业影响力。

三、学科发展与人才培养

（一）支撑学科发展情况

2025年，中心继续作为中国人民大学人工智能和计算机科学与技术学科建设的重要支撑平台。根据CSRankings统计，中国人民大学人工智能学科稳定位居世界前列，互联网与信息检索方向2024-2025年位居世界第1，计算机学科进入国内第一方阵并在新一轮学科评估中位列“强优”学科。中心以智能搜索、推荐、大模型、智能体和智慧治理为牵引，推动计算机科学与技术、法学、新闻传播学、公共管理、金融、教育等学科交叉融合，形成涉外法治大模型、明新新闻传播智能体、社区治理智能体“亿心通”、玉兰-融观金融研报系统等一批跨学科成果。年度内中心新增国家级项目9项

、省部级项目9项，师生发表论文77篇，其中CCF-A类论文53篇、Nature子刊论文1篇，显著提升了学科的国际显示度和工程化支撑能力。

（二）人才培养情况

2025年，中心继续把重大科研任务、开源项目、校企联合实验室和国际交流作为人才培养主阵地，强化学生在真实问题中的工程能力和创新能力。中心支撑本科生科研领航项目，2020级本科生于兆洋、张佳钊参与Open Manus开源复刻版Manus并形成较大影响；夏文科、刘恒畅、张伟滔等学生获2025世界人形机器人运动会拆药分装技能竞技银牌。学生创新创业成果活跃，高一钊创立的智子引擎完成多轮融资并跻身国产多模态技术领域前列；学生团队获得AFAC2025金融智能创新大赛挑战组冠军、首届人工智能与智慧治理创新创业大赛一等奖、第十六届“京东杯”学生创新创业大赛最高奖、中国国际大学生创新大赛（2025）北京赛区一等奖等。国际交流方面，仅2025年上半年已有100余位师生赴外参与国际交流。

（三）研究队伍建设情况

2025年，中心固定研究人员57人，人工智能方向专任教师25人，该方向国家级高层次人才12人，形成了由学术带头人、青年骨干、工程化人员和交叉学科团队共同组成的研究队伍。年度内，魏哲巍入选国家“万人计划”领军人才，李崇轩获批国家自然科学基金青年科学基金项目（B类），张瑞、邹睿等青年教师获批国家自然科学基金青年科学基金项目（C类），毛佳昕、王子贺等教师承担国家自然科学基金面上项目或北京市科技项目。中心通过“预聘-

长聘制”双轨机制、青年科技创新团队支持、校企联合项目牵引和国际顶会交流，持续支持40岁以下青年教师在大模型、深度搜索、多智能体、可信推荐和AI+社会治理等方向成长。

四、开放与运行管理

（一）主管部门、依托单位支持情况

2025年，主管部门和依托单位继续为中心建设运行提供稳定支持。中国人民大学在中关村校区等区域为中心配置总计5100余平方米实体场地，涵盖立德楼教师办公区和研究生研发工位、信息楼数据中心机房、公共政策实验平台及相关学术交流空间，保障中心师生和校企联合团队开展有组织科研。学校持续支持高性能与大数据公共计算云平台建设和运维，新增自建机房投入110.13万元，主要用于配备高端GPU加速卡等关键设备。学校还在科研平台运行维护、青年科技创新团队建设、人才引进、研究生招生名额、国际交流和产学研协同等方面给予优先保障，为LLaDA、FlashRAG、WebThinker、玉兰-万象等项目提供了稳定条件。

（二）仪器设备开放共享情况

本年度未在材料中列示单台原值30万元以上新增大型仪器设备。中心继续依托学校科学技术保障中心（原大型科学仪器共享平台）和学院自建高性能计算集群开展开放共享，实行专人运维、在线预约、资源调度和数据安全管理工作，在优先保障中心重大科研任务的同时，面向校内团队和合作单位开放算力、存储与模型训练环境，提高设备使用效益。

中心新增30万以上大型仪器设备具体情况如下：

设备名称：无

型号：无

原值（万元）：无

存放地点：无

是否纳入“首都科技条件平台开放仪器设备”：无

备注：2025年度无新增。

（三）学风建设情况

2025年，中心把学风建设、科研伦理和负责任人工智能要求贯穿科研组织全过程。中心常态化举办“高屋建瓴AI公开课”“AI学术思辨会”及学术前沿讲座，引导师生围绕大模型、深度搜索、智能治理、法治科技等方向开展高质量研讨。同时，中心承办“人工智能自主创新发展论坛”，发布自主研发的“玉兰”大模型、涉外法治大模型等前沿成果，有效展示了中心在大模型领域的阶段性成效。此外，中心加强数据安全、知识产权和科技伦理教育，对涉外法治、新闻传播、社区治理等涉及真实场景的数据和模型研发任务，落实合规审查与分级管理。

（四）技术委员会工作情况

2025年度，新一代智能搜索与推荐教育部工程研究中心技术委员会第二次会议顺利召开。会议由技术委员会主任胡事民院士主持，共有13位委员通过线上线下相结合的方式参会，覆盖清华大学、中国科学院、中国人民大学等高校，以及华为、京东、快手、OPPO、中国联通等产业界单位。会议围绕中心年度工作进展、基础模型路线（如LLaDA）、深度搜索与智能体技术、RAG与智能推荐工

程化应用、校企联合实验室建设及下一年度工作计划等议题展开了深入研讨。委员们对中心在大模型新范式、开源工具平台、社会治理智能体及学科建设方面取得的进展给予高度评价，并就技术规模化验证、评估体系创新、开源生态建设、法律合规与出海业务支持、智能体与大模型边界演变等关键方向提出了建设性意见。会上中心还正式聘请胡事民院士担任技术委员会主任，并新增李晗（快手）、秦征（OPPO）、谭焜（华为）三位委员，进一步强化了产学研协同机制，为中心下一阶段的技术突破与成果转化提供了有力支撑。

五、下一年度工作计划

下一年度，中心将继续围绕新一代智能信息检索和基础大模型技术开展有组织科研。技术研发方面，重点推进LLaDA自主大模型架构的尺度扩展、解码采样优化、强化学习、原生多模态和高效推理库建设，在已有扩散语言模型基础上形成更大规模、更易部署、更具国际影响力的开源模型和工具；同时深化WebThinker等深度搜索智能体研究，提升复杂推理、跨源检索、自动报告生成和多智能体协作能力。

成果转化方面，中心将持续升级FlashRAG、WebThinker、玉兰-融观、玉兰-万象、涉外法治大模型、明新大模型等工程化平台，围绕金融投研、智慧治理、涉外法治、新闻传播、智慧教育和智能信息分发等场景形成可复制示范。中心将深化与华为、快手、联通、百川智能、腾讯、OPPO、高途等企业和机构的联合攻关，发挥新增校企联合实验室作用，推动开源成果、模型能力和行业数据场

景更紧密结合。

人才培养方面，中心将进一步完善本硕博贯通和交叉学科育人体系，继续支持本科生科研领航、研究生高水平论文与工程项目训练、学生创新创业和国际联合培养。团队建设方面，依托“预聘-长聘制”和青年科技创新团队机制，持续引进大模型、智能体、多模态、可信AI和AI+社会科学方向青年人才，强化专职工程化队伍与科研团队协同。制度优化方面，中心将完善数据合规、知识产权、开源社区维护、仪器设备开放共享和成果转化激励机制，进一步提高工程中心正式运行阶段的组织效能。

六、问题与建议

中心本年度运行总体平稳，无重大问题。面向下一阶段发展，建议依托单位继续在稳定算力供给、跨学科研究空间、工程化人员岗位和研究生招生名额方面给予支持，保障基础模型训练、深度搜索智能体和大规模社会模拟等高算力任务持续开展。

建议主管部门和教育部进一步支持教育部工程研究中心之间的开放协同，推动人工智能、法治、新闻传播、公共管理和产业应用场景的数据合规共享与联合评测机制建设；同时在成果转化、开源基础软件评价、校企联合实验室长期运行和青年工程技术人员评价方面给予政策指导，帮助中心把原创技术更快转化为服务国家战略和行业升级的新质生产力。

七、审核意见

工程中心负责人审核意见：

本年报的数据及内容真实有效，信息填报属实。

工程研究中心主任：

年 月 日

依托单位审核意见：

依托单位：

（单位公章）

年 月 日

八、年度运行情况统计表

研究方向	研究方向1	面向搜索的新模型研究		学术带头人		徐君
	研究方向2	智能搜索的新任务探索		学术带头人		孙浩
	研究方向3	搜索系统新构架研究		学术带头人		窦志成
	研究方向4	可信可控搜索和推荐技术		学术带头人		丁晓东
	研究方向5	新一代智能搜索与推荐工程化实践探索		学术带头人		赵鑫
工程中心面积	5113.64 m ²			当年新增面积		113.64 m ²
固定人员	69 人			流动人员		1 人
获奖情况	国家级科技奖励	一等奖	0项	二等奖	0项	
	省、部级科技奖励	一等奖	0项	二等奖	0项	
当年项目到账总经费	4181.79万元	纵向经费	2026.69万元	横向经费	2155.1万元	
当年知识产权与成果转化	专利等知识产权持有情况	有效专利	22项	其他知识产权	0项	
	参与标准与规范制定情况	国际/国家标准	0项	行业/地方标准	0项	
	以转让方式转化科技成果	合同项数	0项	其中专利转让	0项	
		合同金额	0.0万元	其中专利转让	0万元	
		当年到账金额	0.0万元	其中专利转让	0.0万元	
	以许可方式转化科技成果	合同项数	0项	其中专利许可	0项	
		合同金额	0.0万元	其中专利许可	0.0万元	

				当年到账金额	0.0万元	其中专利许可	0.0万元
		以作价投资方式 转化科技成果		合同项数	0项	其中专利作价	0项
				作价金额	0.0万元	其中专利作价	0.0万元
		产学研合作情况		技术开发、咨询 、服务项目合同 数	71项	技术开发、咨询 、服务项目合同 金额	3899.25万 元
当年服务情况		技术咨询		355次		培训服务	300人次
学科发 展与人才 培养	依托学科 (据实增删)	学科1	计算机科学技术	学科2	法学	学科3	
	研究生 培养	在读博士	246人		在读硕士		72人
		当年毕业博士	24人		当年毕业硕士		33人
	学科建设 (当年情况)	承担本 科课程	2118.4学时	承担研究生 课程	816学时	大专院校 教材	1部
研究队 伍建设	科技人才	教授	28人	副教授	21人	讲师	8人
	访问学者	国内		2人	国外	0人	
	博士后	本年度进站博士后		1人	本年度出站博士后		0人