前　言

在当前科技革命与产业变革的浪潮中，深度学习技术尤其是以生成式人工智能（generative artificial intelligence）为代表的新型信息技术，正以极快的速度发展，这一进程得益于海量数据与强大计算能力的支撑. 在过去几十年里，人工智能（artificial intelligence, AI）领域的研究者和实践者主要聚焦于模型结构与算法优化，这一研究思路被称为“模型中心范式”. 然而，数据的获取与选择同样显著影响模型的有效性（如性能）与效率（如标注成本或训练开销）. 尤其在模型规模呈指数式扩张的背景下，训练数据需求急剧增长. 人工智能领域知名学者吴恩达曾强调，高质量数据的重要性甚至超过复杂模型的改进，这一观点推动了“数据中心范式”的兴起. 从模型范式向数据范式的转变在深度学习领域迅速引起广泛关注.

这一新范式的核心不在于单纯扩大数据规模，而在于聚焦数据构建与处理、数据治理与安全、数据合成与评估、数据表征与挖掘以及数据应用等全流程环节. 以数据为中心的人工智能致力于突破性能瓶颈，克服传统“数据堆砌”问题，并构建更高效能的AI数据生态系统. 为进一步推动我国学者在以数据为中心的AI领域的融合研究，及时报道相关最新成果，我们组织了本期“数据为中心的AI前沿进展及应用”专题. 自2025年6月5日截稿以来，经领域专家严格评审与多轮修订，最终收录8篇论文，汇集了科学数据特征生成、表格数据聚类、个性化序列数据学习推荐、多模态数据鲁棒表征、文本-图像数据检索、半监督人群计数、视觉语言数据安全鲁棒性及时序数据补全等方向的最新研究成果，充分展现了数据为中心的人工智能在方法创新与应用实践中的前沿进展，为同行学者提供参考与启发.

肖濛等作者的研究“基于强化学习的科学数据特征生成算法”聚焦于科学数据特征生成中面临的高阶特征组合构建难、领域知识依赖强及搜索空间指数扩张等问题，提出了一种多智能体协作构建数学变换算式来实现特征自动化生成的框架，同时引入了大语言模型（LLMs）以评估生成特征的可解释性，展示了数据为中心的AI在数据优化生成与可解释性研究中的广阔前景.

李培文等作者的研究“基于合成数据预训练基础模型的表格数据聚类方法”针对表格数据特征异构性、高维稀疏性及弱局部相关性导致的分组结构挖掘困难，提出了基于合成数据预训练表格数据基础模型的“预训练-迭代推理”两阶段聚类框架，显著超越传统具有代表性的聚类方法，为数据结构化分析提供了新动能.

陈泽君等作者的研究“面向作答序列数据的情境自适应学习推荐方法”针对现有学习推荐方法忽视作答序列情境信息及知识迁移困难，提出了情境自适应学习推荐方法CAE-RRS，融合知识概念、习题情境感知、时序大语言模型和知识图谱动态关联，实现对学习者高价值、稀疏时序行为数据的高效挖掘和建模，推动数据为中心的AI在学习推荐领域的应用.

尹小龙等作者的研究“置信引导的多模态数据测试时适应框架”针对多模态测试时低质量视图引入有害梯度、跨模态置信度失衡和时序特征漂移等问题，研究了置信度引导的在线自适应机制，利用视图不确定性评估、视图裁剪与时序对比自监督模块，提高了模型在噪声干扰与分布偏移环境下的鲁棒性，为多模态模型测试时的持续表征更新与数据质量评估提供创新方案.

胡冰玉等作者的研究“面向复杂噪声数据的鲁棒文本-图像行人检索方法”针对文本-图像行人检索中存在的噪声关联与多对多语义对应问题，提出语义感知噪声关联学习框架SNCL，通过语义感知的噪声识别函数与噪声鲁棒互补学习策略实现噪声样本精准辨识与鲁棒训练，有效缓解假正/假负样本对问题，提升模型在高噪声场景下的语义对齐能力与泛化性能，为数据为中心的AI提供了可靠的数据优化与鲁棒表征范式.

余鹰等作者的研究“渐进式认知引导的双域半监督人群计数”针对半监督人群计数中伪标签质量受认知不确定性与任意不确定性干扰的问题，提出渐进式认知引导的双域半监督人群计数网络PCDNet，通过认知评估模块实现多尺度伪标签筛选、并结合频域与空间域的联合约束，有效抑制噪声干扰并提升模型的鲁棒性与计数精度，推动数据为中心的AI在高噪声环境下的数据质量优化与表征学习发展.

曾诚等作者的研究“多模态视觉语言表征学习模型及其对抗样本攻防技术综述”系统梳理了多模态视觉语言表征学习的发展脉络，从细粒度对齐、轻量高效化与多任务通用等方面总结了近年来的模型方法，归纳了模态联合攻击、通用扰动与补丁攻击等多类对抗攻击技术及相应防御策略，并对未来在安全鲁棒性与可解释性方面的研究趋势进行了展望，揭示了数据安全与鲁棒性在多模态智能中的核心地位，为数据为中心的AI在安全可信方向的发展提供了重要支撑.

周立等作者的论文“基于深度学习的不完整时序数据补全方法综述”系统分析了时序数据中缺失值带来的挑战，从基于RNN、CNN、注意力机制及大模型的补全方法进行了归纳总结，并通过实验评估不同方法在多种缺失场景下的性能差异，最后对大规模数据、复杂缺失场景、多模态融合及轻量化部署等未来研究方向进行了展望，凸显了高质量数据预处理和表征对保障下游智能分析与决策效果的关键作用，强化了数据为中心的AI在复杂时序场景下的实践价值.

在当今人工智能模型快速迭代与数据规模指数增长的时代，亟需数据高效获取、智能处理与质量优化等关键技术的突破，以应对数据异构性、噪声干扰及复杂场景下不同下游任务性能受限等挑战.希望本专题能够抛砖引玉，促进我国在相关领域研究的进一步发展.受篇幅限制，本专题无法全面覆盖该领域的所有研究进展，难免存在不足之处，敬请学界同仁批评指正！我们衷心感谢期刊编委会的指导与支持，感谢审稿专家的辛勤付出以及作者们的积极参与！希望本专题能够为研究人员、工程师及从业者提供有益的启发与参考，进一步推动数据为中心的AI在理论创新与实践应用中的发展.我们期待更多学者持续投入，共同促进该领域的繁荣与进步！

白翔 教授（华中科技大学）
蒋杰 副总裁（腾讯集团）
尹义龙 教授（山东大学）
王皓 副教授（中国科学技术大学）

2025年8月