软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(8):3223−3228 [doi: 10.13328/j.cnki.jos.007730] [CSTR: 32375.14.jos.007730]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大语言模型能力上界、应用边界和可信模型系统*

何 玥[1], 初 旭[2], 汪云海[3], 魏哲巍[4], 杜小勇[1,3], 梅 宏[2]

[1](中国人民大学 信息学院, 北京 100872)

[2](北京大学 计算机学院, 北京 100871)

[3](数据工程与知识工程教育部重点实验室 (中国人民大学), 北京 100872)

[4](中国人民大学 高瓴人工智能学院, 北京 100872)

通信作者: 杜小勇, E-mail: duyong@ruc.edu.cn

摘 要: 生成式大语言模型 (以下简称大模型) 通过建模海量语料中的统计规律, 在语言理解、内容生成和交互式问题求解中展现出突出能力, 但其能力形成机制也限定了适用范围和可靠性条件. 大模型能力受到 3 类机制性条件约束: 训练数据决定知识覆盖和时效范围、有限条件下 Transformer 架构的结构性表达边界以及概率生成目标导致的原生性幻觉. 这些能力上界的约束会进一步在应用中显现为任务范畴、风险后果和数据形态的边界: 生成式模型与判别式任务的错位、安全攸关场景对黑箱幻觉的低容忍以及序列化表征与结构化数据的不适配. 面对单体模型在知识、计算、验证和责任方面的局限, 关键突破路径在于向模型系统演进. 智能体式工作流是模型系统的重要形态, 通过引入外部记忆、检索与工具调用等构件突破单体模型的能力限制. 同时, 该系统需依托生成与符号验证闭环确立信任锚点, 落实“生成归模型, 验证归系统”; 并嵌入人在回路机制确保最终决策的责任归属. 大模型的发展需要从参数规模扩张走向系统可信建设, 转向构建可解释、可验证、可追责的模型系统, 使其真正成为辅助人类解决问题的可靠工具.

关键词: 大语言模型; 能力上界; 应用边界; 可信模型系统

中图法分类号: TP18

中文引用格式: 何玥, 初旭, 汪云海, 魏哲巍, 杜小勇, 梅宏. 大语言模型能力上界、应用边界和可信模型系统. 软件学报, 2026, 37(8): 3223–3228. http://www.jos.org.cn/1000-9825/7730.htm

英文引用格式: He Y, Chu X, Wang YH, Wei ZW, Du XY, Mei H. Capability Upper Bounds, Application Boundaries, and Trustworthy Model Systems of Large Language Models. Ruan Jian Xue Bao/Journal of Software, 2026, 37(8): 3223–3228 (in Chinese). http://www.jos.org.cn/1000-9825/7730.htm

Capability Upper Bounds, Application Boundaries, and Trustworthy Model Systems of Large Language Models

HE Yue[1], CHU Xu[2], WANG Yun-Hai[3], WEI Zhe-Wei[4], DU Xiao-Yong[1,3], MEI Hong[2]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Computer Science, Peking University, Beijing 100871, China)

[3](Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), Ministry of Education, Beijing 100872, China)

[4](Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China)

Abstract: Generative large language models (hereinafter referred to as LLMs) have demonstrated outstanding capabilities in language understanding, content generation, and interactive problem-solving by modeling statistical regularities in massive corpora. However, the

* 基金项目: 国家自然科学基金 (L2524018); 中国科学院学部学科战略类项目 (4-XKC2025019)
本文由《软件学报》执行主编金芝教授和责任编委崔斌教授推荐.
收稿时间: 2026-06-23; 修改时间: 2026-06-26; 采用时间: 2026-07-02; jos 在线出版时间: 2026-07-06
CNKI 网络首发时间: 2026-07-07

mechanisms underlying these capabilities also constrain their applicability and reliability. Specifically, LLMs' capabilities are constrained by three types of mechanistic conditions: training data that determines knowledge coverage and temporal scope; the structural expressive boundaries of the Transformer architecture under finite conditions; and inherent hallucinations stemming from probabilistic generation objectives. These upper-bound constraints on model capability further manifest in practical applications as boundaries across task categories, the severity of potential consequences, and data forms: the misalignment between generative models and discriminative tasks; the low tolerance of safety-critical scenarios for black-box hallucinations; and the incompatibility between serialized representations and structured data. To address the limitations of monolithic models regarding knowledge, computation, verification, and accountability, the critical evolutionary path lies in transitioning toward model systems. Agentic workflows constitute an important form of model systems, overcoming the capability limits of monolithic models by integrating components such as external memory, retrieval, and tool calling. At the same time, these systems must establish anchors of trust through a closed loop of generation and symbolic verification, following the principle that "generation is handled by the model while verification is handled by the system." Furthermore, a human-in-the-loop mechanism must be embedded to ensure the accountability of final decisions. The development of LLMs needs to shift from scaling up model parameters to constructing trustworthy systems, moving toward building explainable, verifiable, and accountable model systems, so that they truly become reliable tools that assist humans in solving problems.

Key words: large language model (LLM); capability upper bound; application boundary; trustworthy model system

1 引 言

从逻辑推理、专家系统到机器学习和深度学习, 人工智能始终围绕一个基本问题展开: 如何将人类问题转化为机器可处理的表示, 再将处理结果反馈到人的实践活动中, 以扩展人解决问题的能力. 当前, 在数据、架构、训练方法和算力条件的共同推动下, 生成式大语言模型展现出跨任务的语言生成和交互能力, 推动人工智能取得突破性进展. 从能力来源看, 大模型主要通过学习文本、代码、公式等语料中的统计规律形成表示和生成能力. 这种机制赋予大模型强大的语义表达能力, 同时也限定了其知识覆盖面、推算精确度和事实可靠性. 厘清大模型的能力上界和应用边界, 有助于避免对模型能力产生过高期望; 同时, 明确系统化发展路径, 有助于推动大模型在真实场景中稳定、审慎、可追责的应用.

2 能力上界: 数据、架构与生成目标的机制性约束

大模型能力持续扩展, 但其亦由客观机制形成和限制. 理解大模型的能力边界, 需回到能力形成机制本身: 经验数据分布、Transformer 架构和概率生成目标. 大模型不是脱离数据、结构与训练目标而存在的通用智能体, 这些条件共同限定模型的有效能力区间. 只有认识到这些基础性限制, 才能理解为什么模型能力难以稳定外推到任意领域, 为什么参数规模增长仍需配合数据、工具、验证和流程机制.

2.1 训练数据限定大模型的知识覆盖与时效范围

大模型的能力高度依赖训练数据. 它通过对海量文本、代码、公式等语料进行统计学习, 形成对语言、知识和模式的高维表示. 数据的领域覆盖、时间范围、分布密度和质量水平, 直接影响模型能力的覆盖范围[1]. 一方面, 高质量公开数据的增量供给正面临瓶颈. 互联网上可用于训练的高质量文本并不是无限资源, 尤其是经过人工组织、具有知识密度和逻辑结构的数据更加稀缺[2]. 另一方面, 生成训练数据虽然可以补充真实数据供给, 但在缺少真实数据锚定、质量过滤和分布控制时, 也可能产生分布偏畸、尾部信息丢失甚至模型坍塌风险[3]. 在缺少实验和人类反馈时, 大模型难以自发地产生可验证的新知识. 它可以重组已有知识, 生成可供进一步检验的方案与解释, 但其知识基础仍主要来自训练数据及后续检索与交互环境. 当任务明显偏离训练分布, 或者涉及高度专业、实时变化、事实稀缺的长尾问题时, 模型能力就会显著下降[4]. 数据覆盖、数据质量和数据时效构成大模型能力的基础上界.

2.2 单体 Transformer 架构存在结构性表达边界

当前大模型主要建立在 Transformer 架构之上. Transformer 的自注意力机制增强了模型对上下文关系、长程依赖和复杂模式的表示能力, 是大模型取得突破的重要基础. 但这并不意味着单体 Transformer 在现实约束下可以

模拟任意计算过程. 从计算复杂性角度看, 在有限深度、对数精度和受限计算资源等条件下, Transformer 的表达能力受到电路复杂度[5]①、信息传递路径和结构表示方式的约束[6]. 由此可见, Transformer 不能简单等同于任意计算过程, 参数规模增长仍需要上下文管理、外部记忆和执行机制配合. 从信息传递角度看, 自注意力能够建立全局关联, 但在长上下文、多层组合和复杂结构任务中仍会面临信息压缩、注意力稀释和过度聚合等问题[7]. 对于需要严格递归、精确符号操作、复杂函数复合或离散代数推理的任务, 模型通常表现出不稳定性. 从生成方式来看, 大模型通常采用自回归解码, 即逐 token 生成结果. 这种方式有利于生成局部语义连贯的连续文本, 但对于需要全局一致性、严格约束满足和可验证计算过程的任务, 单次生成容易出现局部合理而整体错误的问题. 因此, Transformer 架构本身虽具有强大的表达能力, 但仍然存在结构性能力边界.

2.3 概率生成目标决定幻觉具有原生性

大模型输出是在上下文条件下生成的概率结果. 它根据上下文预测最可能出现的后续内容, 这种生成过程并不等同于访问一个绝对真实、完备、一致的知识库. 因此, 大模型幻觉不能简单归为偶发系统漏洞, 它与概率生成机制、开放域知识不完备和训练目标密切相关[8]. 在开放域任务中, 尤其是面对长尾事实、实时信息、专业细节和不完整上下文时, 模型通常在证据不足时仍需生成连贯回答. 由于训练目标通常更直接奖励语言合理性和上下文一致性, 对事实真实性的约束需要额外机制补充, 因而模型可能生成看似合理但实际错误的内容[9]. 模型能够进行类比、补全和扩展, 其原因恰恰在于模型会在概率空间中进行重组和外推, 输出结果不等同于原文检索. 但这种外推能力缺乏事实锚定和验证机制时, 就可能转化为错误生成. 因此, 幻觉难以仅靠模型架构本身被彻底消除[10]. 更现实的治理方向在于建立可控机制, 使错误能够被发现、被追溯、被纠正. 也就是说, 可信大模型系统的关键在于使错误可发现、来源可追溯、结果可纠正、责任可界定.

3 应用边界: 判别式任务、安全攸关场景与结构化数据中的多重失配

大模型能力上界的约束, 在真实世界应用中具体表现为任务范畴、风险后果和数据形态上的边界. 当基于序列化表征的生成式模型, 面对需要清晰决策边界、极高可靠性以及结构化数据的现实应用时, 其局限性会更加集中地凸显. 讨论大模型的应用边界, 核心在于识别生成式模型与判别式任务的错位, 安全攸关场景对黑箱幻觉的低容忍度, 以及序列化表征与结构化数据的不适配.

3.1 生成式模型与判别式任务存在错位

生成式模型以数据重构生成为导向, 其输出通常追求语言连贯、语义相关和上下文一致; 而判别式任务面向特定目标变量进行判断或预测, 核心要求是清晰决策边界、置信度校准与可溯源判据, 并且通常需要低延迟响应. 例如, 在电力负荷预测或临床疾病诊断等场景中, 任务目标不是生成一个“看起来合理”的答案, 而是要求高置信度、精准量化的决策结果. 然而, 生成式模型直接处理此类任务时存在错位. 一方面, 生成式模型主要学习数据中的统计共现关系, 容易将高频共现特征混淆为判别依据. 仅依赖其输出, 可能带来校准偏差、越界判断和判据不明等问题[11]. 另一方面, 生成式模型支撑开放式生成需要较大的参数规模和推理开销, 对判别式任务而言也存在过度参数化、计算效率低下的问题. 因此, 生成式模型可以辅助判别式任务, 但不应直接替代判别式模型. 可以在通用层面探索具有跨任务泛化能力的判别式架构, 同时形成生成辅助和判别校准的协同流程, 依赖独立验证机制确保边界和归因的确定性.

3.2 安全攸关场景难以容忍黑箱幻觉

在一般信息服务场景中, 大模型生成错误可能只是带来信息偏差或使用不便, 但安全攸关场景中的事实性错误可能直接造成严重后果[12], 高可靠需求无法接受不可解释的黑箱幻觉[13]. 因此, 大模型的应用边界必须根据后

① 电路复杂度以阈值门电路的深度与规模来衡量计算特定函数所需的资源, 旨在刻画完成计算的最小资源下界. 理论证明, 在常数深度与对数精度约束下, Transformer 的计算能力上限受限, 被包含于常数深度、多项式规模的阈值电路族 (即 TC^0 类) 中, 这从结构上界定了其无法高效表达被广泛猜想超出该复杂度类的函数.

果可逆性、人工可复核性和责任强度进行分层刻画. 第 1 类是有限风险的专业信息辅助场景 (如医学文献汇编、合规条款检索等). 在此类场景中, 大模型可以承担检索和解释任务, 但输出必须绑定原始证据和不确定性提示, 只能作为从业者的可溯源参考, 不能直接转化为操作指令. 第 2 类是高风险辅助决策场景 (如辅助诊断建议、金融交易预警), 大模型只能生成候选方案或风险提示, 系统应引入独立判别模型或外部知识库进行交叉校验, 并由具备资质的人类专家复核后方可进入业务流程[14]. 第 3 类是安全红线或涉及不可逆行动的场景 (如司法最终裁决、安全应急响应等). 在此类场景中, 大模型不应直接触发最终行动, 即便作为信息汇总或预案生成模块接入, 也必须前置权限隔离、人工确认和责任归属机制. 风险等级越高, 大模型越应从“决策者”退回为“证据组织者”和“候选方案生成器”, 系统验证与人类控制权越应成为不可绕过的硬约束. 相应地, 评价标准也需要从平均正确率转向最坏情况下的安全下界、证据可追溯性和人工兜底能力.

3.3 序列化表征与结构化数据不适配

科学计算、图分析与时序预测等结构化数据任务依赖数值关系、拓扑关系、时序动态和约束条件, 这些任务的目标超出语言模式匹配范畴. 大模型在语言语义建模上具有优势, 但在处理需要精确数值泛化和严格逻辑约束的结构化数据任务时, 输出稳定性仍然不足, 连续变量被离散化为 token 会引入量化误差, 并在多步计算推演中发生级联放大[15]. 同时, 分子、电路与社交网络等图结构数据包含复杂的拓扑依赖, 其空间度量基于非欧氏距离. 将高维图结构展平为一维 token 序列, 可能削弱节点间拓扑距离和代数关系的表达信息[16]. 此外, 表格与图等结构化数据具有置换不变性, 即特征列或节点的排列不改变数据本质, 而自回归模型的序列化生成天然对顺序敏感. 序列化处理会引入额外顺序信息, 模型需要额外学习与任务本质无关的顺序模式, 从而影响对同构排列的稳定泛化[16].

4 系统支撑: 从单体模型向可信模型系统演进

既然单体模型在知识、计算、验证和责任方面存在边界, 那么继续单纯依赖模型规模扩张就不是唯一发展路径. 回顾软件工程的发展历史, 软件能力的跃升不是依赖单体程序的无限膨胀, 而是源自架构演进、模块协同、流程控制和系统治理. 面对大模型的局限性, 研究视角也需要从“单体模型”扩展到“模型系统”. 一个模型难以独立承担所有认知、推理、验证和责任任务, 可信应用需要工程组织、工具调用、外部知识、形式化验证和人在回路共同支撑, 构建可解释、可验证、可追责的可信模型系统.

4.1 构建大模型系统, 补足单体模型能力短板

未来大模型的发展需要在参数规模提升之外, 加强模型系统建设. 当前的智能体式工作流体现了这一系统化趋势[17]. 在模型系统中, 大模型承担语言理解、任务分解和候选方案生成, 外部记忆、知识检索、数据库、计算工具、专业软件和流程编排机制为其提供支撑. 外部记忆[18]用于保存长期状态和历史信息, 检索系统[19]提供可追溯、可更新的外部知识, 工具调用[20]支持精确计算、代码执行和数据分析, 流程编排[21]将复杂任务拆解为可验证、可回滚、可审计的步骤. 通过这些系统机制, 大模型输出由单次文本生成转向可编排、可复核、可迭代的任务执行系统; 而在计算理论层面, 引入外部记忆与迭代控制, 也增强了模型系统处理长程任务和复杂控制流的能力 (完善了图灵完备性方面的计算能力)[22]. 相应地, 评价标准也需要覆盖系统级能力, 包括任务分解、工具调用、流程稳定、异常识别和错误恢复.

4.2 “连接主义”生成与“符号主义”验证结合: 以形式化校验建立信任锚点

大模型具有突出的生成能力和直觉式模式识别能力, 这种能力需要接入可审查、可验证的符号系统, 才能形成稳定可靠的知识沉淀. “连接主义”生成与“符号主义”验证结合, 是构建可信模型系统的重要方向[23]. 一种可行路径是构建“生成-验证”闭环, 形成“生成归模型, 验证归系统”的格局: 模型负责生成候选方案、说明思路和进行交互表达; 系统负责证据检索、规则校验、权限控制、结果验证和异常处理, 进而实现发散性生成与确定性验证的优势互补. 例如, 在数学证明任务中, 大模型首先将自然语言表述的数学命题转化为形式化语言, 接着生成证明策略或构造中间引理; 符号证明器或形式化验证工具随后逐步执行这些策略, 严格检查逻辑推导是否满足底层公理体系; 若中间某步推导验证失败, 系统会将具体的错误信息作为反馈返回给大模型, 大模型据此修正证明思路或补充

引理, 循环往复直至整个证明通过确定性验证. 需要强调的是, 连接主义与符号主义相结合并不是解决所有问题的通用答案, 属于可信系统设计原则, 适用于具有明确规则、类型或约束的任务. 尤其在涉及安全红线和高风险决策的场景中, 大模型输出必须通过系统验证和人工确认后才可执行.

4.3 人在回路嵌入最终责任机制

无论大模型能力如何提升, 其工具属性始终不应改变, 不能作为责任主体. 工具可以增强人的能力, 但不能替代人的判断权、价值选择和责任承担. 这一不可让渡的责任底线, 必须转化为系统架构中的刚性约束, 要求人机协同必须明确角色分工. 模型负责生成、检索和归纳, 系统负责约束、验证和审计, 人负责最终确认、价值判断和责任承担. 人应处于确认链、解释链和责任链之中, 掌握关键节点的否决权、纠偏权和最终决定权[24]. 同时, 人在回路需要提供有效的审核条件, 其核心在于形成实质性的责任闭环. 系统应当向人提供充分的依据、来源、风险提示和替代方案, 支持人能够进行有效判断, 并降低流畅表达造成的误导风险. 组织也应对流程设计、权限配置、审核机制和风险控制承担责任, 形成“人管理工具、组织把控流程、制度划定边界”的分层责任体系.

5 结语: 从规模扩张走向系统可信

总体来看, 大模型的“能”在于基于人类已有知识扩展人的知识重组与迁移、内容生成和方案探索能力; 其“不能”在于无法自动保证输出结果真实性、价值正当性与责任合法性. 从能力形成机制看, 大模型能力由训练数据、基础架构和概率生成目标所限定; 这些约束则进一步导致其在判别式任务、安全攸关场景和结构化数据处理中呈现明显局限. 面对单体模型的能力上界和应用边界, 单纯依赖模型规模扩张不是未来大模型发展的唯一出路. 更加务实且有效的发展路径是将模型自身能力提升与检索、工具、验证、审计和人在回路机制结合起来, 补足单体模型的能力短板, 构建可解释、可验证、可追责的可信模型系统, 推动大模型在真实场景中稳定审慎地运行. 归根结底, 人工智能的发展目标在于让机器成为扩展人类判断和行动能力的工具, 大模型亦是如此. 在这一过程中, 应坚持以人为中心、科技向善和责任可溯的基本原则. 由此, 大模型才能在清晰边界和责任框架内成为可靠的工具.

致谢 本文工作受到国家自然科学基金委员会-中国科学院前沿交叉研判联合项目“数据智能的发展路径研判”(项目编号: L2524018、4-XKC2025019) 资助, 本文观点受益于围绕“数据智能的发展路径研判”项目开展的系列专题研讨会, 以及 2026 年 5 月 15–17 日召开的第 41 届中国计算机学会秀湖会议“数据智能的能与不能: 理论上限与应用边界、模型系统与人机协同”, 谨向参与讨论的专家和同事表示感谢.

References

[1] Mei H. Some criticisms on current AI boom. Communications of the China Computer Federation, 2024, 20(9): 38–45 (in Chinese).

[2] Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In: Proc. of the 41st Int’l Conf. on Machine Learning. Vienna: JMLR.org, 2024. 49523–49544.

[3] Shumailov I, Shumaylov Z, Zhao YR, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature, 2024, 631(8022): 755–759. [doi: 10.1038/s41586-024-07566-y]

[4] Yang LY, Zhang SB, Qin LB, Li YF, Wang YD, Liu HM, Wang JD, Xie X, Zhang Y. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto: ACL, 2023. 12731–12750. [doi: 10.18653/v1/2023.findings-acl.806]

[5] Merrill W, Sabharwal A. The parallelism tradeoff: Limitations of log-precision Transformers. Trans. of the Association for Computational Linguistics, 2023, 11: 531–545. [doi: 10.1162/tacl_a_00562]

[6] Cui GY, Wei ZW, He K. Position: The turing-completeness of real-world autoregressive Transformers relies heavily on context management. In: Proc. of the 43rd Int’l Conf. on Machine Learning. Seoul: PMLR, 2026.

[7] Barbero F, Banino A, Kapturowski S, Kumaran D, Araújo JGM, Vitvitskyi A, Pascanu R, Veličković P. Transformers need glasses! Information over-squashing in language tasks. In: Proc. of the 38th Int’l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 98111–98142. [doi: 10.52202/079017-3114]

[8] Ji ZW, Lee N, Frieske R, Yu TZ, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 248. [doi: 10.1145/3571730]

[9] Kalai AT, Nachum O, Vempala SS, Zhang E. Evaluating large language models for accuracy incentivizes hallucinations. Nature, 2026, 653(8116): 1047–1051. [doi: 10.1038/s41586-026-10549-w]

[10] Kalai AT, Vempala SS. Calibrated language models must hallucinate. In: Proc. of the 56th Annual ACM Symp. on Theory of Computing. Vancouver: ACM, 2024. 160–171. [doi: 10.1145/3618260.3649777]

[11] Brown KE, Yan C, Li ZH, Zhang XM, Collins BX, Chen Y, Clayton EW, Kantarcioglu M, Vorobeychik Y, Malin BA. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. Journal of the American Medical Informatics Association, 2025, 32(5): 811–822. [doi: 10.1093/jamia/ocaf038]

[12] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proc. of the 2021 ACM Conf. on Fairness, Accountability, and Transparency. ACM, 2021. 610–623. [doi: 10.1145/3442188.3445922]

[13] Zhou YJ, Yang JD, Huang Y, Guo KH, Emory Z, Ghosh B, Bedar A, Shekar S, Liang ZW, Chen PY, Gao T, Geyer W, Moniz N, Chawla NV, Zhang XL. Benchmarking large language models on safety risks in scientific laboratories. Nature Machine Intelligence, 2026, 8(1): 20–31. [doi: 10.1038/s42256-025-01152-1]

[14] Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, Charney AW, Freeman R, Kummer B, Glicksberg BS, Nadkarni GN, Klang E. Sociodemographic biases in medical decision making by large language models. Nature Medicine, 2025, 31(6): 1873–1881. [doi: 10.1038/s41591-025-03626-6]

[15] Yan JH, Zheng B, Xu HX, Zhu YH, Chen D, Sun JM, Wu J, Chen JT. Making pre-trained language models great on tabular prediction. In: Proc. of the 2024 Int'l Conf. on Learning Representations. OpenReview.net, 2024.

[16] Jin BW, Liu G, Han C, Jiang M, Ji H, Han JW. Large language models on graphs: A comprehensive survey. IEEE Trans. on Knowledge and Data Engineering, 2024, 36(12): 8622–8642. [doi: 10.1109/TKDE.2024.3469578]

[17] Zhang KC, Li J, Li G, Shi XJ, Jin Z. CodeAgent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Bangkok: ACL, 2024. 13643–13658. [doi: 10.18653/v1/2024.acl-long.737]

[18] Xu WJ, Liang ZJ, Mei K, Gao H, Tan JT, Zhang YF. A-MEM: Agentic memory for LLM agents. In: Proc. of the 39th Conf. on Neural Information Processing Systems (NeurIPS 2025). San Diego: NeurIPS Foundation, 2025.

[19] Asai A, Wu ZQ, Wang YZ, Sil A, Hajishirzi G. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: Proc. of the 2024 Int'l Conf. on Learning Representations. OpenReview.net, 2024.

[20] Chen GX, Zhang Z, Cong X, Guo FD, Wu YS, Lin YK, Feng WZ, Wang YS. Learning evolving tools for large language models. In: Proc. of the 2025 Int'l Conf. on Learning Representations. OpenReview.net, 2025.

[21] Wang YB, Xu ZX, Huang Y, Wang XQ, Song ZR, Gao L, Wang CX, Tang XR, Zhao Y, Cohan A, Zhang XL, Chen XY. DyFlow: Dynamic workflow framework for agentic reasoning. In: Proc. of the 39th Conf. on Neural Information Processing Systems (NeurIPS 2025). San Diego: NeurIPS Foundation, 2025.

[22] Schuurmans D. Memory augmented large language models are computationally universal. arXiv:2301.04589, 2023.

[23] Yang XW, Shao JJ, Guo LZ, Zhang BW, Zhou Z, Jia LH, Dai WZ, Li YF. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. In: Proc. of the 34th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI, 2025. 10770–10778. [doi: 10.24963/ijcai.2025/1195]

[24] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

附中文参考文献

[1] 梅宏. 对当前人工智能热潮的几点冷思考. 中国计算机学会通讯, 2024, 20(9): 38–45.

作者简介

何玥, 博士, 讲师, CCF 专业会员, 主要研究领域为因果机器学习, 可信人工智能.

初旭, 博士, 助理研究员, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 数据挖掘.

汪云海, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为大数据可视分析, 智能体交互.

魏哲巍, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为图计算与图学习, 数据流算法与学习.

杜小勇, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为数据库, 大数据管理.

梅宏, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为软件工程.