当挖掘器成为生成器：重新思考大语言模型驱动的数据挖掘中的分离公理

Xi Chen*CSSE*，深圳大学中国深圳
2024150062@mails.szu.edu.cn

Mengke Li**CSSE*，
深圳大学中国深圳
mengkeli@szu.edu.cn

arXiv:2611.32207v1 [cs.DB] 24 Nov 2026

摘要——**数据挖掘的理论基础隐含地依赖一个未明言的前提，即分离公理（SA）：挖掘工具与被分析的数据在认识论上彼此独立。然而，随着大语言模型驱动的人工智能代理开始积极参与数据生成，这一前提正受到越来越大的挑战。我们认为，这一转变带来了一种新的认识论失败模式，我们称之为本体论自指：发现的模式可能反映的是工具的归纳偏置，而非数据的内在结构。这引发了对知识发现可信度以及自我强化的认识论循环风险的根本性担忧；在这种循环中，推断出的模式可能反映模型内部的偏见，而不是数据的内在结构。因此，我们呼吁重新思考大语言模型时代的数据挖掘范式，以充分考虑分离公理为理论基础，并将其拓展至系统层面的认知保障措施。我们认为，这种转变对于在概念上重整这一学科至关重要，可防止未来的知识发现陷入自我论证的算法回音室。**

索引词—**BlueSky、分离公理、生成式人工智能智能体、数据挖掘认识论**

I. 引言

A. 从外部数据到自生成数据：重新思考数据挖掘系统中的独立性

大语言模型和基于代理的系统的最新进展，正在从根本上重塑现代数据挖掘系统中的数据生成流水线。数据越来越不再是从外部环境中被动收集而来，而是由作为自主代理部署的生成模型主动合成、转换并记录。这一新兴范式挑战了数据挖掘领域长期存在但往往未被明确提出的假设，即生成数据的过程与分析数据的过程在认识论上相互独立。我们将这一假设称为**分离公理（SA）** [1]。在 SA 下，挖掘算法 M 被期望从数据中提取模式 D ，这些模式反映底层生成过程的属性 G_D，而非模型自身的归纳偏置。我们将这一直觉形式化为与二者相关的信念结构之间的弱独立条件： I（Ω_M；Ω_{G_D}）≈ 0，其中Ω_M表示编码于挖掘系统中的隐式归纳偏置，而Ω_{G_D}表示数据生成过程的潜在结构。从历史上看，由于数据采集与算法分析之间在物理和组织层面的分离，SA 一直被视为合理的假设 [2]。然而，在现代原生人工智能系统中，这种分离正变得日益脆弱，因为同一基础模型同时参与数据生成与数据分析。重要的是，这一转变不仅仅是模型能力方面的变化，更是数据生态系统中的结构性变化。同一模型家族如今可能既生成数据，又从数据中提取模式，从而导致一种新兴的认识论纠缠。

B. 代理式数据系统中的认识论纠缠

从通过强化学习激发的 DeepSeek-R1 深度推理能力 [3]，到 Manus 的完全自主任务执行范式 [4]，大语言模型的每一次突破都意味着生成式人工智能进一步渗透到人类数据生产流程的更多阶段 [5]。自主代理的出现 [6]–[8] 进一步推动了这一轨迹：从一次性调用交互发展到持续的自主数据生成循环，在这一过程中，代理在较长时间内与外部系统交互，形成闭环数据生命周期。这对分离公理（SA）构成了系统性挑战，而且这一挑战由数据生态系统的两端共同驱动。在**数据端**，基于大语言模型的代理越来越多地应用于企业 [9]、医疗保健 [10] 和研究场景 [11]，用于生成报告、摘要和日志 [12]，，随后这些内容被存储为原始数据，供下游挖掘使用。在**挖掘端**，现代数据挖掘系统越来越依赖大语言模型作为核心推理引擎，用于检索 [13], [14]、解释 [15], 和报告生成 [16]等任务。因此，数据生成和模式提取都越来越受到同一模型家族的中介作用影响。

当同一个大型语言模型 θ 同时参与这两个过程时，即D =智能体$_\theta$(上下文）和 $P_\theta = M_\theta(D)$，由此产生的模式可能反映模型的内在先验，而非底层数据结构。这造成了一种经典数据挖掘假设未明确涉及的认识论混淆。这种认识论纠缠引入了系统性偏差，并使其贯穿整个知识发现流水线，从而损害了经典数据挖掘的可靠性。在关联规则挖掘中，从代理生成语料中提取的模式反映了潜在的共现结构

* 通讯作者。本研究部分得到国家自然科学基金委员会（NSFC）（62676254、62306181）、广东省基础与应用基础研究基金（2024A1515010163）以及雷达信号处理国家重点实验室（JKW202403）的资助。

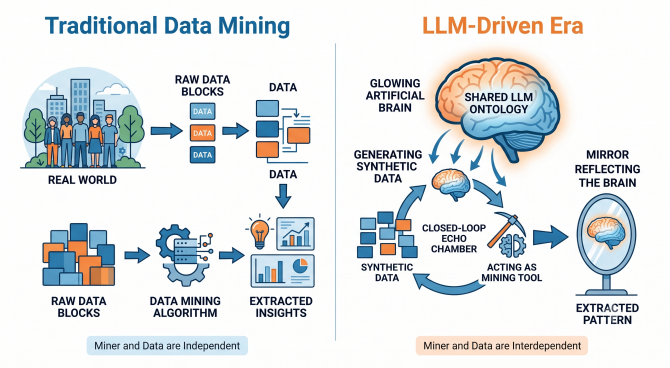

图 1. SA 的崩塌。经典数据挖掘假设数据生成与挖掘相互独立（左）。在人工智能代理驱动的系统中（右），这两个过程通过共享的 LLM 本体（Ω）相互耦合，形成认知闭环和本体论自指。

这些结构继承自预训练，导致统计上看似合理、实则具有误导性的统计依赖关系。图挖掘中也会出现类似现象：由代理介导的交互可能微妙地扭曲观测到的拓扑结构，并影响下游推理，例如社区检测。更广泛地说，当代理生成日志或编译器输出等系统级产物时，挖掘过程所发现的规律反映的是底层模型的生成偏差，而非环境本身的属性。这些观察表明，在共享模型基础上，自动化数据生成与分析系统中的本体论自指是一种普遍的失效模式，如图 1所示。

C. 我们对 *BlueSky* 的愿景

我们设想，在原生于人工智能的数据时代，重新思考数据挖掘认识论：数据生成系统与挖掘系统之间的交互应成为一项首要研究对象，而非一个隐含假设。我们提出两个紧密关联的方向：

- **具备 SA 意识的理论视角**。在现代数据生态系统中重新审视 SA，因为数据与模型日益耦合；重点关注这种耦合如何重塑我们对有意义模式的理解。
- **系统层面的认知保障措施**。将这一视角转化为适用于大规模原生于人工智能的数据系统的系统级设计原则，以减轻本体论自指在数据架构、多智能体生态系统和评估协议中的产生与传播。

II. SA 违背下经典数据挖掘的崩溃

在将 SA 违背确立为人工智能原生数据挖掘的一种新失效模式后，一个自然而然的问题随之而来：现有的可靠性范式是否仍能确保可信的知识发现？大多数现有方法 [17]–[19] 都是基于这样一种隐含假设发展起来的：数据生成与数据挖掘之间保持认知分离。因此，它们针对的都是与 SA 违背所引入的根本不同的不确定性来源；在 SA 违背情形下，挖掘系统与其分析的数据在认知层面发生纠缠 [20]。以下讨论将重新审视几种具有代表性的范式，以说明在存在 SA 违背时，其底层假设的适用边界。

A. 幻觉检测：输出层面的可靠性

幻觉检测 [21] 是一种输出层面的可靠性范式，侧重于通过检索增强、事实核查或其他验证机制，确保单条记录在事实层面准确无误。它隐含地假设，提高每条生成输出的保真度，就能为下游分析带来更可信的数据集。SA 违背带来的挑战处于不同的抽象层面。即使每条生成记录在事实上都准确无误，数据集仍可能继承共享基础模型的语义偏好和统计规律，而非源自底层现实。因此，幻觉检测验证的是单条观测的正确性，而 SA 违背关乎整个数据集的认知保真度。数据集即使完全符合事实，仍可能产生系统性误导知识发现的模式。

B. 自动化启发式设计（*AHD*）系统：通过认知分离实现可靠性

近期的 AHD 系统，包括 FunSearch [22], EoH [23], ReEvo [24], 和 OPRO [25], ，通过在候选生成与评估之间保持认知分离，取得了卓越的性能。大语言模型提出启发式方法，而独立的外部评估器（例如可执行沙箱或客观优化基准）则负责判定其正确性 [26],[27]。因此，它们的可靠性不仅取决于模型能力，也取决于是否保持了 SA。

一般的数据挖掘很少能找到此类客观评估器。没有外部沙箱能够验证从大语言模型生成的医疗记录、科学文献或社会互动中挖掘出的模式是否忠实地反映现实。因此，AHD 系统的成功非但没有解决 SA 违背带来的挑战，反而凸显了这一挑战。它们的可靠性从根本上取决于能否通过外部评估流程保持认知分离，而这一条件在一般的数据挖掘中鲜少具备。其他范式中也存在类似的假设不匹配，包括概念漂移检测 [17],表演性预测 [28]，模型崩塌 [29]，和人工智能内容水印 [30]。尽管这些方法解决了关键的可靠性挑战，但它们所依赖的条件与 SA 违背下的条件并不相同。

其他研究记录了生成式人工智能时代数据质量的下降。例如，数据反馈循环可能放大训练集偏差 [31]，而对合成表格数据进行朴素推理则可能产生统计错误和假阳性 [32], [33]。大语言模型在完善自身输出时也表现出明显的自我评估偏差 [34]。这些发现揭示了模型生成内容带来的风险，但仍属于经验性或输出层面的观察。相比之下，SA 违背及其

所导致的本体论自指提供了一种认识论解释，统一说明了这些现象的根本原因。

III. 分离公理之外的根本挑战

SA 违背并非原生于人工智能的数据挖掘中的另一种误差来源。它消解了数据生成与挖掘之间的分离，进而动摇了发现生命周期中知识形成、传播和验证所依赖的假设。

A. 失去外部真实依据

可解释性假定，解释最终能够以独立于模型本身的观测为依据 [35]。然而，在 SA 违背的情况下，这一假设不再成立。在原生于人工智能的数据系统中，整个认识论链条可以表示为

$$D = \text{Agent}_\theta(\text{context}) \to P = M_\theta(D) \to E = \text{LLM}_\theta(P, D), \tag{1}$$

其中，数据生成、模式发现和解释都源自同一模型家族。因此，即使缺乏外部参照来验证所发现的模式反映的是现实还是仅仅反映了模型自身的先验，解释仍可能在内部保持一致。由此，可解释性成为整个知识发现流水线的属性，而非仅仅是模型的属性。这进而要求我们建立一种适用于认识论耦合情形的全新可解释性理论。

B. 集体认识论强化

现代原生人工智能系统越来越多地由多个相互交互的代理组成，这些代理通过共享上下文、检索记忆和生成输出持续交换信息 [36]。在 SA 违背的情况下，每个代理都会反复将其他代理的输出作为上下文，并生成新的观测结果，从而在共享数据生态系统内形成闭环反馈。由于这些代理共享相同的认识论基础，继承自底层模型的信念会被反复传播和强化。因此，最初存在偏差的信念可能逐渐积累统计支持，最终在缺乏外部依据的情况下成为看似可靠的共识。不同于模型崩塌，这种强化完全通过上下文交换发生，而非通过模型更新实现。在 SA 违背的情况下，仅凭共识已不能再被视为可靠性的证据。

C. 丧失独立评估

传统评估方法依赖这样一个假设：模型输出与观测数据之间的一致性反映了外部效度。然而，在 SA 违背的情况下，数据和习得模式都受到同一认识论基础的塑造。因此，评估不再是将预测结果与独立证据进行比较，而是衡量共享认识论系统内部的一致性。 Consequently，统计显著性、可复现性和泛化能力不再是外部效度的可靠指标，而转而成为认识论自洽性的衡量标准。

这三项挑战表明，SA 违背会导致数据挖掘系统发生统一的认识论崩溃。核心问题并非个别阶段出现故障，而是整个数据生成、模式发现和评估流程中的认识论独立性消失。这种崩溃凸显了这样一种需求：必须建立一种全新的理论，用于在认识论耦合系统下推理可信度。

IV. BlueSky倡议：迈向具备认识论意识的数据挖掘系统

防止认识论崩塌需要在数据、模型和评估之间建立协调机制，而不是在各个独立阶段采取彼此孤立的干预措施。

A. 第一阶段：认识论可观测性

首要要求是将认识论耦合从默认假设转变为可明确测量的对象。我们引入一种认识论分离信号，称为超额对齐比 $S(M,\ D \mid D_0)$。该信号以基线为参照，量化模型诱导先验与数据生成结构之间的对齐程度。不同于传统的分布指标，$S(M,\ D \mid D_0)$能够捕捉观测数据集是否由与挖掘模型本身相同的认识论基础塑造。从操作层面看，这一概念可以通过信念指纹探测加以实例化。我们构建一个标准化探测集$Q = \{q_1, ..., q_k\}$，覆盖目标领域中的关键命题。模型 M 和数据集 D 均被映射为该探测空间上的可比分布，从而得到信念指纹 F_M 和生成指纹F_D。随后将认识论分离定义为

$$S(M, D \mid D_0) = \min\left(1, \frac{\bar{W}(F_M, F_D)}{\bar{W}(F_M, F_{D_0}) + \epsilon}\right), \tag{2}$$

其中，$\bar{W}(\cdot)$表示归一化的分布距离，例如 Wasserstein 距离；D_0 表示一组基线客观观测，ϵ 是用于数值稳定性的小常数。S 的值较高表示模型与数据之间的认识论耦合较弱，而较低的值则表示由共享认识论先验诱导出的强对齐。该公式使认识论纠缠变得可明确测量，为后续的干预与稳定化机制奠定基础。

B. 第二阶段：认识论干预

一旦认识论耦合通过$S(M,\ D)$变得可观察，数据挖掘面临的挑战便从在固定数据集下估计模式，转变为在共享模型先验所诱导的可变认识论结构下进行推理。这一转变的核心影响之一是，有效性不再能仅以对观测数据的拟合程度来定义。相反，必须考量数据和模式如何随底层认识论假设 Ω_M的变化而共同变化。由此产生了一种基于稳定性的推理概念：如果某个模式在认识论扰动 [37]–[39] 下保持不变，就可认为它是可靠的。形式上，这可以通过对模型诱导的认识论结构进行受控变化来理解，例如修改提示词或模型配置，

并观察数据表征和提取出的模式 P所发生的变化。在此类扰动下表现出高度稳定性的模式，更有可能反映外部环境中的结构，而非共享先验造成的假象。与此同时，可以根据观测数据集中各个样本与模型生成倾向的一致性 $P_M(x)$，调整其影响力，从而部分缓解其中的认识论偏差。这并非是将结果校正至未知真实依据，而是一种重新平衡机制：它能减少与模型一致的观测结果被过度放大的情况，同时保留认识论多样性信号。由此确立了一项统一的干预原则：在 SA 违背的情况下，数据挖掘的目标不再是消除偏差，而是识别能够经受认识论变化的稳健模式。

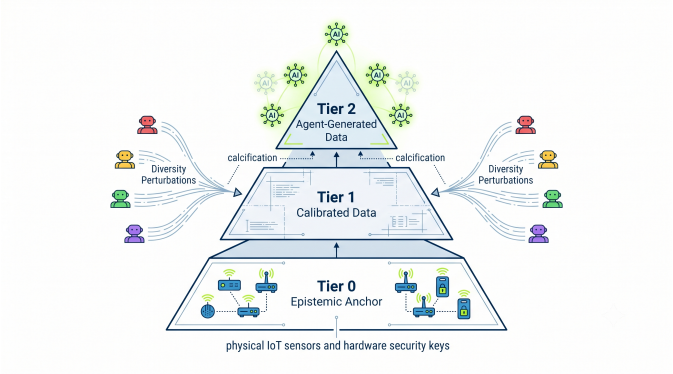

图 2. 三层认识论锚定（TEA）架构。用于表示和管理原生于人工智能的数据系统中认识论分层的系统级框架。

C. 第三阶段：认识论稳定化

上述理论构念需要在系统层面加以实例化，才能在现实世界的原生于人工智能的数据管道中运行。不同于传统的系统设计，其目标并非模块化增强，而是防止数据生成、交互和评估过程中出现认识论封闭。为此，我们引入三层认识论锚定（TEA）架构，使数据潜在的认识论结构能够得到明确表征。TEA 不再将数据集视为同质化输入，而是根据数据与模型诱导先验的纠缠程度对数据进行划分。第 0 层对应于近似外部现实的原始观测锚点；第 1 层包含经过模型部分转换、同时保留对第 0 层可追溯性的数据；第 2 层则对应于完全由模型生成的内容。在实践中，即使名义上由人类生成的数据，也可能由于模型辅助生产的普遍存在而属于更高层级，这使认识论污染成为一种结构性属性，而非例外情况。该架构如图 2所示。

这种分层的数据表示会直接影响信息在多智能体系统中的传播方式。当代理共享同一个基础模型时，它们的输出会在迭代式多智能体交互中被反复重新引入后续上下文 [40], [41], ，从而围绕共享的认识论先验形成强化循环。这种反馈动力学往往会使系统趋向单一的认识论吸引子。通过纳入基于异质模型家族或训练语料库构建的代理，认识论多样性注入可以减轻这一效应。这些代理充当受控扰动，打断收敛动力学。因此，异质代理在评估相同模式时产生的分歧，便成为潜在认识论伪影的诊断信号，揭示了共识形成掩盖的不稳定性。

因此，评估必须置于同一认识论结构下加以解读。准确率、F1 和可复现性等标准评估指标的有效性，依赖于数据与模型之间的认识论独立性。一旦这一假设不再成立，评估衡量的便不再是外部效度，而只是共享认识论系统内的内部一致性。更恰当的标准是在认知扰动下保持稳定：即某个已发现的模式能否在数据溯源、模型家族或推理情境发生变化时依然保持不变。因此，评估从基于固定数据集的一致性转向在认识论重构下的不变性。人工智能原生数据挖掘中的认知有效性，由一种耦合控制机制决定；在这一机制中，可观测性、不变性和分层共同刻画推理保持外部根基、而非陷入自我指涉式一致性的条件。

D. 主要论点

在 SA 违背的情况下，数据挖掘的特征是数据生成与推断之间存在认识论上的不可识别性，外部结构与模型诱导结构在观测上变得无法区分。BlueSky 框架并不试图恢复经典的分离。相反，它通过可观测的分离信号将认识论纠缠形式化，以认识论扰动下的不变性取代有效性，并强制实施系统级分层，以防止系统坍缩为自我指涉的一致性。

V. 结论

在现代数据挖掘中，分离公理长期以来一直隐含其中；然而，随着原生于人工智能的数据系统日益普及，它已不能再被视为理所当然。我们认为，数据生成与挖掘之间的认识论耦合会造成一种新的失效模式，而现有关于分布偏移、可解释性或统计评估的论述都未能充分涵盖这一问题。我们不应将这一挑战视为又一个需要缓解的问题，而应重新审视数据挖掘的基础。本文提出的 BlueSky倡议提供了一种可能的方向：让认识论耦合变得可观测，从而能够进行面向稳定性的干预，并推动建立系统层面的认知保障措施。更广泛地说，我们希望这项工作能促使学界将认知分离视为未来人工智能原生数据挖掘系统的一项明确设计原则。随着智能体日益负责数据的生成、整理和解读，理解知识是如何产生的或许会变得与提升能够发现哪些模式同样重要。

参考文献

[1] T. Cramer，《分离公理的定义》，Canadian Mathematical Bulletin，第17卷，第4期，第485–491页，1974年。[2] U. M. Fayyad、G. Piatetsky-Shapiro 和 P. Smyth，《从数据挖掘到数据库中的知识发现》，AI magazine，第17卷，第3期，第37–54页，1996年。[3] D. Guo、D. Yang、H. Zhang、J. Song、P. Wang、Q. Zhu、R. Xu、R. Zhang、S. Ma、X. Bi 等，《Deepseek-r1 通过强化学习激励大语言模型进行推理》，Nature，第645卷，第8081期，第633–638页，2025年。[4] M. Shen、Y. Li、L. Chen、Z. Fan、Y. Li 和 Q. Yang，《从思维到机器：Manus AI 作为完全自主数字代理的崛起》，arXiv 预印本 arXiv:2505.02024，2025年。[5] D. H. Hagos、R. Battle 和 D. B. Rawat，《生成式人工智能与大语言模型的最新进展：现状、挑战与展望》，IEEE 人工智能汇刊（TAI），第5卷，第12期，第5873–5893页，2024年。[6] J. S. Park 等，《生成式代理：人类行为的交互式模拟》，载于用户界面软件与技术研讨会（UIST）论文集，2023年。[7] X. Wang 等，《基于大语言模型的自主代理综述》，Frontiers of Computer Science，第18卷，第6期，文章编号186345，2024年。[8] P. Wei 等，《Agentbench：将大语言模型作为代理进行评估》，载于国际学习表征会议（ICLR）论文集，2024年。[9] G. Zeng、X. Chen、J. Hu、S. Qi、Y. Mao、Z. Wang、Y. Nie、S. Li、Q. Feng、P. Qiu 等，《Routine：面向企业大语言模型代理系统的结构化规划框架》，arXiv 预印本 arXiv:2507.14447，2025年。[10] W. Wang、Z. Ma、Z. Wang、C. Wu、J. Ji、W. Chen、X. Li 和 Y. Yuan，《医学领域基于大语言模型的代理综述：距离 Baymax 还有多远？》，计算语言学协会（ACL）论文集，2025年，第10 345–10 359页。[11] C. Liu、C. Wang、J. Cao、J. Ge、K. Wang、L. Zhang、M.-M. Cheng、P. Zhao、T. Li、X. Jia 等，《基于大语言模型代理的自动化研究愿景》，arXiv 预印本 arXiv:2504.18765，2025年。[12] J. Lu、B. Pan、J. Chen、Y. Feng、J. Hu、Y. Peng 和 W. Chen，《Agentlens：对基于大语言模型的自主系统中的代理行为进行可视化分析》，IEEE 可视化与计算机图形学汇刊（TVCG），2024年。[13] Y. Zhang、Z. Liu、Q. Wen、L. Pang、W. Liu 和 P. S. Yu，《用于信息检索的人工智能代理：生成与排序》，载于第33届 ACM 信息与知识管理国际会议论文集，2024年，第5605–5607页。[14] W. Zhang、J. Liao、N. Li、K. Du 和 J. Lin，《代理式信息检索》，arXiv 预印本 arXiv:2410.09713，2024年。[15] A. Bechar、A. Oulefki、A. Amira、F. Kurogollu 和 Y. Himeur，《利用视觉大语言模型与小波和三维递归表示，从建筑能耗数据中提取可操作的见解》，载于 IEEE 数据挖掘国际会议（ICDM）。电气电子工程师学会，2025年，第1040–1047页。[16] R. Li、H. Zhang、E. Gehringer、T. Xiao、J. Ding 和 H. Chen，《揭示大语言模型在科学论文自动生成评审中的优点与缺陷》，载于 IEEE 数据挖掘国际会议（ICDM），2025年，第1370–1379页。[17] J. Park、F. Chiang 和 M. Milani，《Andri：概念漂移情况下的自适应异常检测》，载于 ACM 信息与知识管理国际会议（CIKM）论文集，2025年。[18] F. Gr¨oger、S. Lionetti、P. Gottfrois、A. Gonzalez-Jimenez 等，《Cleanpatrick：图像数据清洗基准》，arXiv 预印本 arXiv:2505.11034，2025年。[19] S. J¨ager 和 F. Biessmann，《从数据插补到数据清洗——自动清洗表格数据可提升下游预测性能》，载于人工智能与统计国际会议（AISTATS）论文集，2024年。[20] J. Yang、Q. Zhang、X. Xu、Y. Wu、S. Zhang 和 Z. Ding，《Knobtunex：通过结构化推理增强大语言模型的自动数据库调优》，载于 IEEE 数据挖掘国际会议（ICDM）。电气电子工程师学会，2025年，第843–852页。[21] Z. Ji 等，《自然语言生成中的幻觉综述》，ACM 计算调查，第55卷，第12期，文章编号248，2023年。[22] B. Romera-Paredes 等，《利用大语言模型进行程序搜索所获的数学发现》，Nature，第625卷，475，2024年，468页。

[23] F. Liu 等，《启发式算法的演化：利用大语言模型实现高效自动算法设计》，载于《国际机器学习大会》（ICML）论文集，2024 年。[24] H. Ye 等，《ReEvo：将大语言模型用作具有反思式演化能力的超启发式》，载于《神经信息处理系统大会》（NeurIPS）论文集，2024 年。[25] C. Yang 等，《将大语言模型用作优化器》，载于《国际学习表征会议》（ICLR）论文集，2024 年。[26] G. Wang、Y. Xie、Y. Jiang、A. Mandlekar、C. Xiao、Y. Zhu、L. Fan 和 A. Anandkumar，《Voyager：一种由大语言模型驱动的开放式具身智能体》，《机器学习研究汇刊》（TMLR），2024 年。[27] J. Yang、C. Jimenez、A. Wettig、K. Lieret、S. Yao、K. Narasimhan 和 O. Press，《Swe-agent：智能体—计算机接口赋能自动化软件工程》，《神经信息处理系统进展》（NeurIPS），第 37 卷，第 50 528–50 652 页，2024 年。[28] J. C. Perdomo、T. Zrnic、C. Mendler-Dunner 和 M. Hardt，《表演式预测》，载于《国际机器学习大会》（ICML）论文集，2020 年。[29] I. Shumailov 等，《在递归生成数据上训练时，人工智能模型会发生崩溃》，《Nature》，第 631 卷，第 755–759 页，2024 年。[30] X. Zhao、P. V. Ananth、L. Li 和 Y. Wang，《面向人工智能生成文本的可证明鲁棒水印》，载于《国际学习表征会议》（ICLR），2024 年。[31] R. Taori 和 T. B. Hashimoto，《数据反馈循环：模型驱动的数据集偏差放大》，载于第 40 届《国际机器学习大会》（ICML）论文集，第 202 卷，PMLR，2023 年，第 33 883–33 920 页。[32] B. van Breugel、Z. Qian 和 M. van der Schaar，《合成数据，真实错误：如何（不）发布和使用合成数据》，载于第 40 届《国际机器学习大会》（ICML）论文集，第 202 卷，PMLR，2023 年，第 34 793–34 808 页。[33] A. Decruyenaere、H. Dehaene、P. Rabaey、C. Polet、J. Decruyenaere、S. Vansteelandt 和 T. Demeester，《人工吸引力背后的真相：表格合成数据的推断效用》，载于第 40 届《人工智能不确定性大会》（UAI）论文集，第 244 卷，PMLR，2024 年，第 966–996 页。[34] W. Xu、G. Zhu、X. Zhao、L. Pan、L. Li 和 W. Y. Wang，《傲慢与偏见：大语言模型在自我精炼中放大自我偏差》，载于第 62 届计算语言学协会年会论文集（第 1 卷：长篇论文），2024 年，第 15 474–15 492 页。[35] H. Zhang、C. Zhang、P. Hu、S. Qiu 和 J. Wang，《利用大语言模型定制检索增强生成，以在推荐系统遗忘中去偏》，载于 IEEE 数据挖掘国际会议（ICDM），2025 年，第 1702–1711 页。[36] Z. Zhang、Q. Dai、X. Bo、C. Ma、R. Li、X. Chen、J. Zhu、Z. Dong 和 J.-R. Wen，《基于大语言模型的代理记忆机制综述》，《ACM信息系统汇刊》，第 43 卷，第 6 期，第 1–47 页，2025 年。[37] M. Arjovsky、L. Bottou、I. Gulrajani 和 D. Lopez-Paz，《不变风险最小化》，arXiv 预印本 arXiv:1907.02893，2019 年。[38] K. Ahuja、K. Shanmugam、K. Varshney 和 A. Dhurandhar，《不变风险最小化博弈》，载于《国际机器学习大会》（ICML），2020 年，第 145–155 页。[39] Y. Wang、Z. Lai 和 T. Zhong，《基于全变差的不变风险最小化的分布外泛化》，载于第十三届《国际学习表征会议》（ICLR），2025 年。[40] Q. Wu、G. Bansal、J. Zhang、Y. Wu、B. Li、E. Zhu、L. Jiang、X. Zhang、S. Zhang、J. Liu 等，《Autogen：通过多智能体对话赋能新一代大语言模型应用》，arXiv 预印本 arXiv:2308.08155，2023 年。[41] S. Hong、M. Zhuge、J. Chen、X. Zheng、Y. Cheng、J. Wang、C. Zhang、S. Yau、Z. Lin、L. Zhou 等，《MetaGPT：面向多智能体协作框架的元编程》，载于《国际学习表征会议》（ICLR），第 2024 卷，2024 年，第 23 247–23 275 页。