arXiv:2603.01341v1 [cs.SI] 2 Mar 2026

大语言模型中的结构性幻觉：基于网络的知识组织与引文完整性评估

Moses Boudourides

西北大学专业研究学院

Moses.Boudourides@northwestern.edu

摘要

大语言模型（LLM）日益成为获取学术信息的中介，但其输出通常是在单个陈述层面而非知识结构层面进行评估。本文提出结构性幻觉：对概念组织、关系架构和书目依据的系统性扭曲，而这种扭曲对句子级准确率指标而言仍然不可见。为检测此类扭曲，我们开发了一种基于网络的幻觉压力测试，其基础包括知识图谱提取、图相似性分析、中心性比较和引文完整性验证。

该方案应用于三个代表学术知识核心形式的结构化领域：罗杰的同义词词典 （1911年）作为词汇本体；维基数据哲学家作为传记知识图谱；以及从Dimensions.ai 数据库检索的书目引文记录。在所有领域中均观察到显著的结构偏离。在词汇基准中，宏平均 F1 分数低于 0.05；在传记基准中，幻觉率超过 93%；在文献计量基准中，引文遗漏达到 91.9%。罗杰词典重构中的网络层面比较进一步显示，节点集杰卡德相似度为 0.028，捏造率高于 94%。

这些发现表明，不能仅凭局部流畅性推断结构保真度。所提出的压力测试为知识组织和信息质量研究中的大语言模型生成的知识表示的结构完整性评估提供了一种可复现工具。

关键词： 大语言模型；结构性幻觉；知识组织；知识图谱；网络分析；信息质量；文献计量学；引文分析

1 引言

大语言模型（LLM）日益融入各类工作流程，帮助人们获取学术信息。它们被用于概述文献、重构概念分类体系、生成传记数据以及编制参考书目。然而，对这些系统的评估仍主要局限于句子层面的准确性和语义合理性。对于通过本体、权威记录和引文结构来组织知识的领域而言，这种评估并不充分。因此，核心问题并非大语言模型输出在局部是否流畅，而是它们能否保留所模拟知识系统的结构完整性。

本文通过将概念分析与系统性实证检验相结合，探讨这一问题。我们认为，大语言模型在学术语境中最具影响的失误，并非孤立的事实错误，而是结构性幻觉：对概念组织、关系架构和书目依据的扭曲，而这些问题在局部（句子层面）评估中难以察觉。模型可能复现本体或网络的表面形式，却大幅改变其节点构成、边结构和中心性分布。仅凭传统事实核查无法检测出这类偏离。

为考察这一现象，我们提出一种基于网络的幻觉压力测试，该测试以知识图谱提取、图相似性分析和中心性诊断为基础。我们不将大语言模型的可靠性视为伦理合规或风格正确性问题，而是将其操作化为大语言模型生成的表征与权威真实值知识系统之间的结构可比性问题。

在实证方面，本文将这一流程应用于代表学术知识核心形态的三个领域：词汇本体（罗杰同义词词典，1911年版）、传记规范数据（维基数据哲学家）以及文献计量学研究（从Dimensions.ai检索的引文）。对于每个基准，我们采用多层评估流程，结合经典信息检索指标（精确率、召回率、F1值）、语义相似性分析、机器学习幻觉分类，以及对基于真实值数据和大语言模型生成数据构建的知识图谱进行网络层面比较。在所有三个领域中，我们都观察到显著的结构偏离，包括节点集与边集的极端不匹配，以及大规模引文遗漏。

本文作出三项贡献。首先，本文对结构性幻觉进行了正式定义并提出其类型学，将其与现有文献中的内在幻觉和Huang 等人，2025年；Ji 等人，2023年）所讨论的外在幻觉类别区分开来。其次，本文引入了一种可复现的网络分析压力测试，用于评估大语言模型生成的知识表示的结构保真度，并以知识组织理论（Hjørland，2003年；Zeng，2008年）和网络科学为基础

（Barabási，2016年；Newman，2010）。第三，本文提供了一项系统的跨领域实证评估，表明即使局部语义合理性得到保留，仍可在图层级检测到结构性失效。

本文其余部分安排如下。第2节考察基于流畅度的评估的局限性。第3节定义结构性幻觉，并指出其主要机制：概念重新聚焦、书目失真和逻辑错误结构化。第4节将大语言模型治理重新界定为方法论验证。第5节提出基于知识图谱的压力测试方案。第6节聚焦引文完整性这一文献计量案例。第7节概述其对学术实践的影响。第8节介绍实证研究。第9节结合信息质量与知识组织解读研究结果。第10节作出结论。

2 The 流畅性、错误与准确性的认识论问题

2.1 流畅性的认识论意义

大语言模型融入学术工作流程的范围如今已涵盖文献综述、假设形成、稿件撰写和学生评估（Bommasani 等人，2021年；Lund 等人，2023年；van Dis 等人，2023年）。这一转变的认识论意义不在于采用规模，而在于大语言模型输出的一项独特属性：语言流畅度与语言流畅度和认知可靠性之间的脱钩。

大语言模型在其Bender 等人、2021年；Bommasani 等人、2021年）。它们的设计目标是在统计意义上，依据训练数据的分布进行下一个词元预测（）。它们旨在生成看似合理的续写，而不是检索经验证的知识。因此，输出内容可能语法正确、风格得当、表面连贯，却在事实、语境或关系层面存在错误。在学术话语中，论断的有效性取决于概念与参考文献的精确组织，因此这种似是而非可能掩盖实质性的扭曲。

Bender 等人 （2021年）将此描述为语言形式与扎根的意义之间的错配。语言模型操纵词元序列，却无法直接获取这些词元所表示的实体和关系。在学术场景中，这种局限不仅限于孤立的事实错误，还会延伸至对概念结构与文献结构的错误呈现。

2.2 幻觉：流畅性失效的实证表征

这一局限的一个有充分文献记载的表现是幻觉：生成虽流畅却属捏造的陈述（Huang 等人，2025年；Ji 等人，2023年；Tonmoy 等人，2024）。幻觉出现在各个领域和不同规模的模型中，且很少自我标示（Huang 等人，2025年；Ji 等人，2023年）。错误的数字对象标识符、捏造的传记细节以及虚构的引文，可能以与准确信息相同的文体权威出现（Alkaissi 和 McFarlane，2023年；Weidinger 等人，2021年）。

现有分类法通过主要关注命题层面虚假的类别来区分内在幻觉与外在幻觉（Ji 等人，2023年）。本文将分析扩展至关系组织层面的失效。

2.3 句子层面的准确性与结构完整性

我们区分句子层面的准确性与结构完整性。句子层面的准确性关注单个陈述是否属实。结构完整性关注是否忠实呈现某个领域的关系架构。

一段文本中的单个陈述可能都正确，却错误呈现思想依存关系、引文模式或概念层级。孤立的事实核验无法发现这类扭曲。要发现它们，必须评估所呈现知识的整体组织（Pan 等人，2024；Sansford 等人，2024）。

标准大语言模型基准测试强调问答准确性和事实回忆（Chen 等人，2023年；Tonmoy 等人，2024）。它们并不评估模型在生成较长学术文本时，能否保留本体、规范档或引文记录的结构。这一方法论上的空白，正是本文所提出框架的研究动因。

这种区分类似于命题性知识与结构性知识之间的区别（Bommasani 等人，2021年；Wang 等人，2023年）。学术专长取决于能否将论断置于关系框架中。与孤立的错误不同，结构性扭曲会改变知识的底层架构，因此带来不同等级的风险。

2.4 事实核查的局限

事实核查关注句子层面的正确性，却没有考察结构组织。核实 LLM 生成文本中的个别主张，并不能保证最终形成的

表述保留了某一字段的智力结构或文献结构。

结构性扭曲是关系配置的属性，而非孤立陈述的属性。为了检测这类扭曲，我们在此提出一些方法，从生成的文本中提取知识图谱，并将其与权威的真实基准表征进行比较。此类方法必须处理的结构性失效模式分类体系将在下一节展开。

3 结构性幻觉：定义与机制

3.1 定义结构性幻觉

在自然语言生成中，幻觉通常被定义为不忠实于源文档或缺乏输入上下文支持的输出（Ji 等人，2023年；Maynez 等人，2020；）。对于开放式大语言模型生成，这一术语已扩展至涵盖看似合理却缺乏事实依据的内容（Alkaissi 和 McFarlane，2023年；Huang 等人，2025年；Rawte 等人，2023年）。尽管这些定义不可或缺，但它们主要针对单条陈述层面的失误。

我们引入**结构性幻觉**这一概念，以描述一种独特的失效模式。结构性幻觉指大语言模型生成文本中的陈述单独看来可能合理，甚至符合事实，但整体上却歪曲了某一知识领域的关系架构。扭曲不在于孤立的错误陈述，而在于文本构建或暗示的关系配置——概念依赖关系、思想谱系、引文网络或论证层级。这类扭曲无法仅通过逐句核验检测出来。

这一类别扩展了摘要研究文献中提出的内在／外在区分（Maynez 等人，2020）。内在幻觉与来源相矛盾；外在幻觉则引入缺乏依据的内容。相比之下，结构性幻觉指的是对隐含编码的知识结构层面的错误表征。它可能与局部正确的陈述并存，同时改变实体构成、关系结构或显著性。

学术知识本质上具有关系性和层级性（Chen 等人，2023年）。一项主张的重要性取决于它在由概念、方法和参考文献构成的有序网络中的位置。因此，结构性幻觉与孤立的事实错误有着本质不同：它影响的是知识的组织方式，而非某个单独陈述是否为真。

3.2 结构性幻觉的机制

结构性幻觉通过三种反复出现的机制表现出来。

3.2.1 概念重新聚焦

概念重新聚焦是指边缘概念或在统计上显著的概念被呈现为某一领域的核心。大语言模型通过共现模式而非认识论加权来学习关联（Bender 等人，2021年；Bommasani 等人，2021年）。因此，在大众语境或频繁重复的语境中突出的概念，可能会在生成内容中被赋予不成比例的重要性。通过考察相对于权威表征的中心性或社群结构变化，网络层面比较能够揭示此类重新聚焦。

3.2.2 书目失真

书目失真是指伪造引文、错误归属，或对思想谱系的误导性压缩。已有实证研究记录了格式风格正确的生成式伪造参考文献（Alkaissi 和 McFarlane，2023年；Orduña-Malea 和 Cabezas-Clavijo，2023年）。除直接捏造外，选择性引用也可能造成失真，改变某一领域中影响结构和认识论声望的表象。由于引文记录构成组织学术知识的基础设施，这一层面的失真影响的不只是个别论断，还会改变关系记录本身。

3.2.3 逻辑错误结构化

逻辑错误结构化 指的是尽管局部内容连贯，但论证关系却配置错误的情况。例如，模型可能准确描述两篇出版物，却暗示存在实际上并未发生的直接思想回应或依赖关系。每个陈述可能单独来看都是正确的，但它们之间构建的关系却并非如此。只有将推断出的关系结构与既有的知识表征进行比较时，才能检测出此类错误（Rawte 等人，2023年）。

3.3 结构性幻觉标准

上述机制彼此交互。概念重新聚焦可能会影响引文选择，而引文选择又会塑造论证结构。由此产生的文本可能内部一致，却在结构上偏离其所依据的知识领域。

因此，评估的核心问题在于结构：LLM生成文本是否保留了其所表述领域的关系配置？要回答这一问题，需要采用相应方法，从生成文本中提取知识图谱，并将其与权威真值系统进行比较。接下来的分部将运用网络分析技术，详细阐述这套方法。

4 治理作为方法论基础设施

4.1 从规范性治理到方法论验证

在关于人工智能的讨论中，治理一词通常被置于伦理、规范或政策导向的框架中（Lindebaum等人，2025年；Papagiannidis等人，2025年）。然而，在学术研究语境中，首要问题在于方法论：如何以可复现的方式评估大语言模型生成内容的结构可靠性？

我们在一种范围有限、可操作的方法论意义上使用“治理”一词。此处，治理指一种结构化的验证审查，使LLM输出能够接受检查，并与权威知识系统进行比较。这一概念并不抽象地着眼于披露或监督，而是强调能够生成结构保真度可测量指标的诊断程序。

这种重新界定与信息科学的方法论传统相契合；在这一领域，引文分析、文献计量映射和知识图谱构建都是分析知识组织的标准工具。本文提出的框架将这套方法应用于对LLM生成的学术表征的评估。

4.2 验证组件

所提出的验证工作流由三个组件构成。

知识图谱提取与比较。 从 LLM生成文本中提取其隐含的实体和关系，并将其表示为图。然后，将该图与根据精选数据库或权威文献计量学来源构建的真值知识图谱进行比较。节点存在性、边配置和关系方向方面的差异，构成可量化的结构偏离指标。

网络分析诊断。 构建图后，分别计算生成图和真实图的标准图指标——度中心性、中介中心性、模块度和社群结构。对这些指标进行定量比较，可以识别概念重新聚焦、异常社群形成或其他形式的关系失真。

引文完整性验证。 通过权威书目数据库验证生成的参考文献，以确认其是否存在及元数据准确性。除逐条验证外，还会评估生成的参考文献目录相对于既有学术记录是否存在遗漏、选择性强调或关系失真(Alkaissi 和 McFarlane，2023年；Orduña-Malea 和 Cabezas-Clavijo，2023年)。

这些组成部分共同将结构评估从定性判断转变为可复现的分析流程。

4.3 方法论架构

这三个组成部分在互补的层面上发挥作用。引文验证关注参考文献层面的书目正确性。知识图谱比较评估关系配置。网络诊断则对结构属性进行量化汇总，以便在不同模型、提示词和领域之间进行比较。

每个步骤都会生成可解释的输出——缺失的节点/实体、多余的边/关系、中心性变化、引文遗漏率——这些输出可与其他程序参数一并报告。从这个意义上说，方法论治理被视为将既有验证实践扩展至一类新型生成式工具。

后续分部中介绍的实证基准将在词汇、传记和文献计量领域应用这一架构，展示如何在实践中实现结构验证。

本治理框架的应用层面——即其在学术评估、稿件撰写和教学大纲设计中的实际运用——将在第7节中探讨。

5 知识图谱与网络分析治理

5.1 知识图谱作为知识组织系统 s

从信息科学的角度来看，知识图谱是知识组织系统（KOS）的一种计算化实现：它以结构化方式呈现某一领域中的概念、实体及其关系（Hjørland，2003年，2021年；Zeng，2008年）。知识组织研究长期以来一直关注如何忠实地呈现领域结构，以服务于检索、推理和交流。知识图谱将这种结构呈现为带标签的（有向或无向）网络，其中节点代表实体或概念，边代表类型化关系（Hogan等，2021年）。

此处提出的框架将知识图谱构建视为评估结构保真度的主要工具。句子级验证关注单个论断的真实性；图级比较则关注生成文本的关系架构是否与权威领域表征的关系架构相匹配（Arazy 和 Kopak，2011年；Stvilia 等人，2007年）。这一转变将大语言模型评估置于既有的信息质量维度——准确率、完整性和一致性——之中，并将这些维度应用于关系结构，而非孤立记录。

通过将本研究提出的结构性幻觉指标对应到既有的信息质量（信息质量）维度，可以明确体现这种对应关系。在广泛使用的信息质量框架中，准确率指数据正确表征其所描述现实世界实体的程度；完整性指所有必要数据均已具备的程度；一致性指不同表征之间不存在内部矛盾；相关性则指数据对预期用途的适用程度（Lee 等人，2002年；Pipino 等人，2002年；Strong 等人，1997；Wang 和 Strong，1996年）。

根据这一映射，虚构率 直接体现了关系准确性：生成图中存在而权威参考图中缺失的节点，代表不准确的实体级断言。同样，错误生成 DOI 和凭空捏造传记字段，均违反了属性准确性。相比之下，书目基准中观察到的引文遗漏率 体现了结构完整性：书目中缺失的引文表明对该领域知识结构的呈现并不完整。当超过 90% 的规范引文链接缺失时，生成的图便在系统层面未能满足完整性标准。

中心性排名相关性 和向上流动 诊断将信息质量中的一致性概念延伸至关系架构。如果参考图中的规范节点中心性下降，而虚构节点获得了结构影响力，那么生成的表示在架构上

与权威领域模型不一致。这种不一致并非记录层面的矛盾，而是拓扑层面的失真。

最后，模块度比较和图相似度指数体现了表征连贯性与语境完整性，这些维度在信息质量文献中常与内在数据质量和语境数据质量相关联（Arazy 和 Kopak，2011年；Stvilia 等人，2007年）。图在局部层面上可能在语义上看似合理，但在依据外部结构基准进行评估时，却未能满足内在信息质量标准。

通过将伪造检测、引文遗漏和中心性偏差置于正式的信息质量维度——准确率、完整性和一致性——之中，该框架将结构性幻觉纳入既有的信息质量理论。因此，这项贡献不仅具有诊断意义，也具有概念意义：它将信息质量评估从属性级元数据扩展至网络级认知架构。

这一结构视角对数字图书馆和自动化索引系统的认知可靠性具有直接影响。当代发现平台日益采用知识图谱增强、自动主题标引、实体链接和大语言模型辅助元数据生成，以支持检索、推荐和语义导航。如果大语言模型生成的结构呈现出以下实证基准中记录的关系偏差——虚构实体、中心性层级扭曲以及规范节点遗漏——那么这些扭曲可能会传播至索引层、推荐系统和跨数据库链接基础设施。其结果不仅是局部元数据错误，还会系统性削弱数字知识环境中的检索精确率、引文追踪和领域表征。从信息质量的角度来看（Lee 等人，2002年；Wang 和 Strong，1996年），这类结构性偏差会威胁图书馆和书目系统中规范控制、主题一致性和语义互操作性的可靠性（Borgman，1999年；Tennis，2012年）。从这个意义上说，结构性幻觉并不局限于个别文档；它对数字图书馆赖以运作的认知基础设施构成系统性风险。

5.2 从文本到网络：提取与比较

结构评估始于将非结构化学术文本转换为明确的图表示。

概念与关系抽取。使用依存句法分析、共现分析或基于语义相似度的链接等标准自然语言处理技术，识别概念和关系

。由此生成的图谱编码了文本中隐含的概念架构。对于治理而言，基于依赖关系的提取和共现提取足以捕捉领域层面的结构。

文献计量网络构建。 利用书目内容构建引文网络或合著网络，以呈现文本的智力背景。这些网络可直接与根据文献计量数据库或经过整理的语料库构建的参考结构进行比较（Jaradeh 等人，2019）。

比较分析。 将大语言模型生成的图与基于权威来源构建的参考图进行比较。结构偏离——包括缺失节点、伪造边、关系倒置或社群结构改变——构成了结构性幻觉的操作性特征。

5.3 网络分析诊断

构建图后，会针对生成的图和参考图计算标准网络度量（Barabási，2016年；Newman，2010；Paulheim，2017年）。

中心性。 度、中介中心性、特征向量中心性和 PageRank 可识别结构上重要的节点（Bonacich，1987；Freeman，1977；Page 等人，1999年）。生成图与参考图中的中心性得分之间的秩相关，可为概念重新聚焦提供定量度量。

模块度与社群。 模块度和社群检测（例如 Louvain 算法）（Blondel 等人，2008年；Newman 和 Girvan，2004）用于评估生成图中的主题簇是否对应于公认的领域子结构。社区组成的偏离表明发生了结构重组。

图相似度。 杰卡德相似度等基于集合的度量用于量化节点集和边集之间的重合程度。相似度低表明存在大规模结构偏离。

5.4 幻觉压力测试协议

幻觉压力测试通过三个连续步骤将结构评估付诸实施。

步骤 1：伪造检测。将大语言模型生成的图中存在、但参考图中缺失的节点标记为虚构内容。此类节点所占的比例即为大语言模型捏造率。

步骤 2：中心性比较。 针对共同节点，计算其在多种指标下的中心性排名相关性。相关性较低表明概念重新聚焦。

步骤 3：向上流动分析。 若某个节点在大语言模型生成的图中的中心性排名相较于参考图大幅上升，则将其识别为向上流动节点。当此类节点同时属于虚构内容时，便代表严重的结构偏离。

这些步骤共同生成结构保真度的多层次诊断概况。由此得到的指标——虚构率、中心性排名相关性、模块度比较和图相似度——均为定量、可复现的指标，并且可在既有信息质量框架内加以解释。

与 BLEU 或 ROUGE 等标准自然语言处理评估指标不同，这些指标可以衡量生成文本与参考文本之间表层的 n 元语法重叠（Lin，2004；Papineni 等人，2002年）——幻觉压力测试考察的是关系结构层面。BLEU 和 ROUGE 是为机器翻译和摘要生成任务设计的，在这些任务中，词汇相似度近似反映语义充分性。然而，它们无法敏锐捕捉网络拓扑、层级重组和引文遗漏：生成文本即使达到中等程度的词汇重叠，也可能彻底改变中心性结构或遗漏关键关系边。相比之下，中心性排名相关性评估实体的相对结构重要性是否得到保留，而模块度比较则评估社群结构是否在不同图中得到维持（Blondel 等人，2008年；Newman，2010）。这些诊断能够捕捉词汇指标无法检测到的拓扑层面偏离。从信息科学的视角来看，知识组织依赖于关系架构的忠实保留，而非仅仅依赖文本相似度，因此网络分析指标提供了更适切且具有理论依据的评估框架。

从业者清单：实施幻觉压力测试

对于信息科学家而言，可以将该压力测试作为结构化验证工作流，集成到编目、元数据生成或知识图谱增强流程中。以下清单以操作性术语概述了该流程：

1. **定义身份验证权威参考图**。选择一个可信的知识组织系统

（受控词表、人工整理的本体、文献计量数据库或机构知识图谱），以表示目标领域。

2. **提取大语言模型生成的图**。 解析大语言模型生成内容，识别实体（节点）和类型化关系（边），并构建有向图表示。

3. **计算虚构率**。 比较节点集。虚构率为 $= \frac{|V_{LLM} \setminus V_{ref}|}{|V_{LLM}|}$，其中 V_{LLM}、 V_{ref} 分别表示大语言模型生成的图和参考图的节点集。

4. **计算结构对齐指标**。 对于共享节点：

- 计算中心性度量（度、中介中心性、PageRank）。
- 计算大语言模型与参考排名之间的斯皮尔曼等级相关。

5. **评估社群结构**。 对两张图应用社区检测（例如 Louvain）。比较模块度和社群组成。

6. **识别向上流动**。 标记排名显著上升的节点。将具有高中心性的虚构节点列为高风险失真。

7. **报告诊断概况**。 记录：

- 虚构率
- 秩相关
- 模块度比较
- 节点/边杰卡德相似度

从信息质量维度（准确率、完整性、一致性）解读结果。

伪代码摘要

输入：生成图 G_llm，参考图 G_ref

V_llm =节点(G_llm) V_
ref =节点(G_ref)

捏造_率 = |V_大语言模型 - V_参考 |/|V_大语言模型 |

V_共有 = 交集(V_大语言模型, V_参考)

对于中心性{degree, betweenness, pagerank}：rank_llm = centrality(G_llm, V_shared) rank_ref =centrality(G_ref, V_shared) rho =spearman(rank_llm, rank_ref)

模块度_llm = louvain(G_llm) 模块度_ref = louvain(G_ref)

向上_节点 ={v in V_shared : rank_llm(v) >> rank_ref(v)}

输出：
虚构率_秩相关_模块度
_比较向上_节点

这份操作性摘要将治理框架转化为一种可复现的验证工具，适用于集成到数字图书馆工作流、自动化索引系统和元数据质量审计中。

6 引文完整性与文献计量依据

引文完整性构成了结构评估中的文献组成部分。如果知识图谱呈现概念组织，引文网络则呈现学术谱系和证据基础。文献层面的结构性幻觉表现为伪造的参考文献、错误归属、选择性省略或扭曲的引文模式。

6.1 引文验证

第一步是系统地核实大语言模型生成或推荐的所有参考文献。每条引文都要与权威书目数据库（如 Web of Science、Scopus、

Dimensions、OpenAlex 等）进行核对，以确认其是否存在以及元数据准确性。这为书目可靠性确立了最低基准。

实证研究记录了LLM生成文本中存在伪造的参考文献和错误的元数据（Alkaissi 和 McFarlane，2023年；Orduña-Malea andCabezas-Clavijo，2023年）。除了核查是否存在之外，验证还必须评估语境相关性：被引用的研究是否实质性地支持其所要佐证的论点。存在性核验可以自动完成；语境相关性则需要专家审查。

6.2 文献计量基准测试

除了对单项内容进行验证，结构评估还在网络层面开展。将LLM生成文本所形成的引文图（或任何其他类型的文献计量图），与基于文献计量数据库构建的参考引文网络进行比较。

比较可以在多个层面进行：

节点层面： iden识别缺失的经典著作，或边缘来源被过度呈现的情况 s.

边层面：考察引文邻域，以评估被引作品是否处于连贯的学术共同体之中。

网络层面：比较引文图的度分布、聚类系数、中心性特征和社群结构，使用第 5 节所述的相同网络分析工具。

大语言模型生成的引文网络与参考引文网络之间的系统性偏差，构成文献计量结构性幻觉：对学术交流实证结构的歪曲呈现。

6.3 作者与发表场所分布

其他诊断还会考察被引作者和发表场所的分布。过度集中于少数作者、忽视奠基性贡献，或场所代表性存在系统性偏差，都可能表明模型对领域结构的近似存在扭曲。

这些分布分析将大语言模型生成的引文行为置于既有文献计量基线中进行考察，从而能够评估表征平衡，而不只是事实正确性。

6.4 解释性含义

引文验证和文献计量基准测试补充了前文提出的知识图谱诊断。两者共同确保对大语言模型生成的学术表征进行评估时，不仅考察其局部准确率，也考察其与实证观察到的学科组织结构之间的结构对齐程度。

本文后文提出的实证基准将这些程序付诸实践，并量化实践中观察到的引文（或共同署名等）遗漏、捏造和网络偏离程度。第9节实证基准中记录的引文遗漏规模——其中91.9%的遗漏率被置于科学引文社会学的更广阔背景下考察——促使我们提出更有力的解释性主张。从科学社会学和学术交流的视角来看（Garfield，1955年；Merton，1973），这种偏离会扰乱认可、影响力追踪和累积性知识形成机制。因此，问题不仅在于数据质量，更在于认知连续性：引文在结构上的遗漏削弱了思想传承的可见链条，而这些链条正是维系学科连贯性的基础。

7 应用：评估、论文写作与教学大纲设计

前述分部所构建的结构验证框架，其适用范围已超越模型基准测试，延伸至更广泛的学术实践。其核心贡献在于将评估的重心从表面流畅性转向关系完整性与文献计量完整性。

7.1 评估情境

在教育场景中，结构诊断可用于评估提交的作品——无论其以何种方式创作——是否准确反映某一领域的概念组织和文献组织。知识图谱比较和引文验证使人们能够评估表征的一致性，而不仅仅是文风润色。

7.2 手稿撰写

在研究写作中，该框架能够对文献综述和概念分部进行系统验证。通过将草稿文本和可信的参考语料库都转换为可比较的知识网络和文献计量网络，作者可以发现被遗漏的奠基性研究、遭到扭曲的影响模式，或概念重新聚焦等问题；这些问题无法通过句子级审查发现（Chen 等人，2023年）。

7.3 课程设计

应用于课程设计时，结构比较能够揭示大语言模型生成内容中的缺漏、失衡或过度集中，例如阅读清单和主题序列。将教学大纲建模为知识网络和引文网络（或任何类型的文献计量共现网络），为评估主题覆盖度和学科连贯性提供了一种形式化方法。

在这些领域中，无论学术表征是否由大语言模型生成，该框架都可作为一种通用方法，用于评估其结构保真度。

7.4 伦理问题：知识可见性与认知公平

前文呈现的实证发现引出了超越技术可靠性的伦理考量。如果将大语言模型用作文献综述或领域映射工具，却遗漏了大量经典或专业研究（正如引文重建基准测试中所观察到的那样），其结果就不仅是学术研究不完整，更会系统性地扭曲知识可见性。

数字图书馆和索引系统历来作为认知基础设施发挥作用：它们决定哪些内容能够被发现，从而决定哪些内容能够被引用、传授并获得机构认可（Garfield，1955年；Hjørland，2003年）。当大语言模型生成的综述遗漏了被引用文献中的大量内容时，就可能造成算法中介的知识孤岛。如果某些研究传统在训练数据中代表性不足，或难以从参数化记忆中进行结构性重建，它们可能会彻底变得隐而不见。

这一问题与社会信息学领域长期以来关于技术系统与学术共同体之间互动的研究相呼应（Kling，2000）。知识组织研究表明，分类和索引实践从来都不是中立的；它们会塑造一个字段内的注意力分配和权威（Bowker and Star，1999；Olson，2002）。如果大语言模型辅助的文献综述系统性地偏重代表性充分或当代的材料，同时忽略

边缘、少数群体或历史久远的学术成果，那么由此产生的扭曲可能会加剧既有的认知不平等。

从信息质量的视角来看，此类遗漏构成了完整性与表征公平性的失败（Wang 和 Strong，1996年）。因此，伦理问题是结构性的，而非文体性的：若大语言模型在生成流畅摘要的同时遗漏现有参考书目的大量内容，便可能收窄学术话语，并以牺牲边缘化声音为代价放大主流叙事。

在书目基准中观察到的91.9%引文遗漏率，其影响超出了不完整的检索本身。被引文献编码了或可称为学术谱系的内容，即某一研究主题的理论基础、方法论创新与经典论争的历史序列；这些序列将当前工作置于持续的学术对话之中（Garfield，1955年；Merton，1973）。当大语言模型生成的文献综述遗漏这些联系中的大多数时，它并非仅仅排除个别论文，而是扰乱了可见的思想传承链条。由此形成的表征将研究呈现为脱离语境或非历史的，实际上削弱或抹去了当代学术所依赖的基础结构。在认可、引用与理论连续性对学科认同至关重要的领域，此类结构性遗漏有可能割裂集体记忆，并削弱学术影响的可追溯性。

本文提出的方法论治理框架直接应对了这些风险。通过要求明确进行引文网络重构、伪造检测，并与权威语料库进行结构比较，幻觉压力测试不仅是一种方法论保障，也是一种认识论问责工具。从这个意义上说，结构验证也是数字知识基础设施中社会责任的一种体现。

8 实证基准：结构性幻觉的三项压力测试

第4至第7节中构建的比较框架要求以系统化的网络分析方法，将大语言模型输出与权威的外部参考结构进行评估。本节通过三项基准实验将该框架付诸实践，分别针对学术知识组织的不同领域：词汇结构（罗杰同义词词典）、传记知识（维基数据哲学家）和书目引文记录（从研究数据平台 Dimensions.ai 检索）。三者共同构成了一项跨领域的实证压力测试，旨在检验

第3节提出的结构性幻觉假说。

共享评估流程

这三项基准测试均采用统一的三阶段评估流程。

阶段1：数据获取。 对于每个基准，均从权威的结构化来源（罗杰同义词词典、维基数据，以及 Dimensions.ai 等研究数据平台）构建基准真值数据集。抽取部分条目，并向大语言模型（gpt-4.1-mini）发起查询，要求其根据记忆重建每个条目的结构化字段。大语言模型的响应被限制为 JSON格式，最多包含 1,024 个词元，并解析为符合基准真值模式的结构化记录。所有大语言模型调用均采用指数退避重试逻辑（延迟时间为 1、2、4、8 秒），并在首次成功调用后缓存到磁盘，以确保结果可复现。

阶段2：传统评估。 使用基于词元集比率的模糊字符串匹配过程，将大语言模型生成的每个字段值与基准真值进行比较（阈值为 ≥ 80，通过`rapidfuzz` 库实现）。使用第三方判定器（DBpedia（Auer 等人，2007年））来区分真正的额外知识（经 DBpedia 确认的假阳性）与幻觉（基准真值和 DBpedia 中均不存在的假阳性）。分别针对每个字段计算精确率、召回率、F1值和准确率。

第 3 阶段：机器学习增强评估。 应用了两种机器学习技术。 语义相似度评分 使用预训练的句子转换器 `all-MiniLM-L6-v2` (Reimers 和 Gurevych，2019）计算真实值与大语言模型字段值之间的余弦相似度。该模型将句子映射到一个 384 维的稠密向量空间，该空间针对语义文本相似度进行了优化。余弦相似度的计算公式为：

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}, \tag{1}$$

其中，$\mathbf{a}$ 和$\mathbf{b}$ 分别是真实值和大语言模型字段值的句子嵌入。得分为 1.0 表示语义内容完全相同；得分为 0.0 表示两者正交。

幻觉分类使用第 2 阶段生成的二元幻觉标签训练逻辑回归模型。特征包括词元级重叠（词元集合的杰卡德相似度）、字符级归一化编辑距离（莱文斯坦距离除以最大字符串长度），

以及句子转换器得出的余弦相似度分数。分类器使用 F1值、精确率、召回率、准确率、ROC-AUC、对数损失和马修斯相关系数（MCC）进行评估。

8.1 基准测试1：罗杰同义词词典的词汇重建

第一个基准使用罗杰的1911年版《英语词汇与短语分类词典》（罗杰，1911年），取自古腾堡计划（1.5 MB纯文本，27,031行）。罗杰同义词词典是一种按层级组织的语义本体，将英语词汇映射为由类别、分部和词头构成的分类体系。在最高层级，类别将宽泛的概念领域归为一组（例如抽象关系、空间、物质、智力、意志、情感）。每个类别再细分为分部，将这些领域细化为更具体的主题分组，每个分部包含多个代表不同语义概念的词头。每个条目是一个语义概念，并配有经过整理的名词、动词、形容词、副词列表，以及指向相关词头的交叉引用。1911年版包含1,000个词头，是英语词典编纂学史上组织最为严谨的知识结构之一。

我们编写了一个自定义解析器，从纯文本中提取全部 997 个有效词头及其关联术语列表。该解析器对原始术语字符串执行以下预处理步骤：按逗号和分号拆分；去除括号注释（例如，(inquire) 461）；去除方括号中的拉丁语释义（例如，［Lat.］）；去除 &c. 标记和独立数字；并统一转换为小写。随后抽取 $n = 30$ 个词头作为随机样本（随机种子为 = 42），并向 gpt-4.1-mini 发出查询，要求其为每个词头重建五个结构化字段（名词列表、动词列表、形容词列表、副词列表、交叉引用）。

8.1.1 传统评估

传统评估使用词元集合模糊匹配，将大语言模型生成的术语列表与罗杰词典真实列表进行比较。结果表明，所有字段的重建保真度都极低（表1）。

表1：罗杰同义词词典重建（$n = 30$词头）的传统评估结果。所有分数均按词头进行宏平均。

字段	精确率	召回率	F1	准确率
noun_list	0.024	0.020	0.022	0.00
verb_list	0.019	0.020	0.020	0.00
形容词_列表	0.017	0.019	0.018	0.00
副词 _list	0.000	0.000	0.000	0.00
交叉–参考文献	0.038	0.053	0.044	0.00

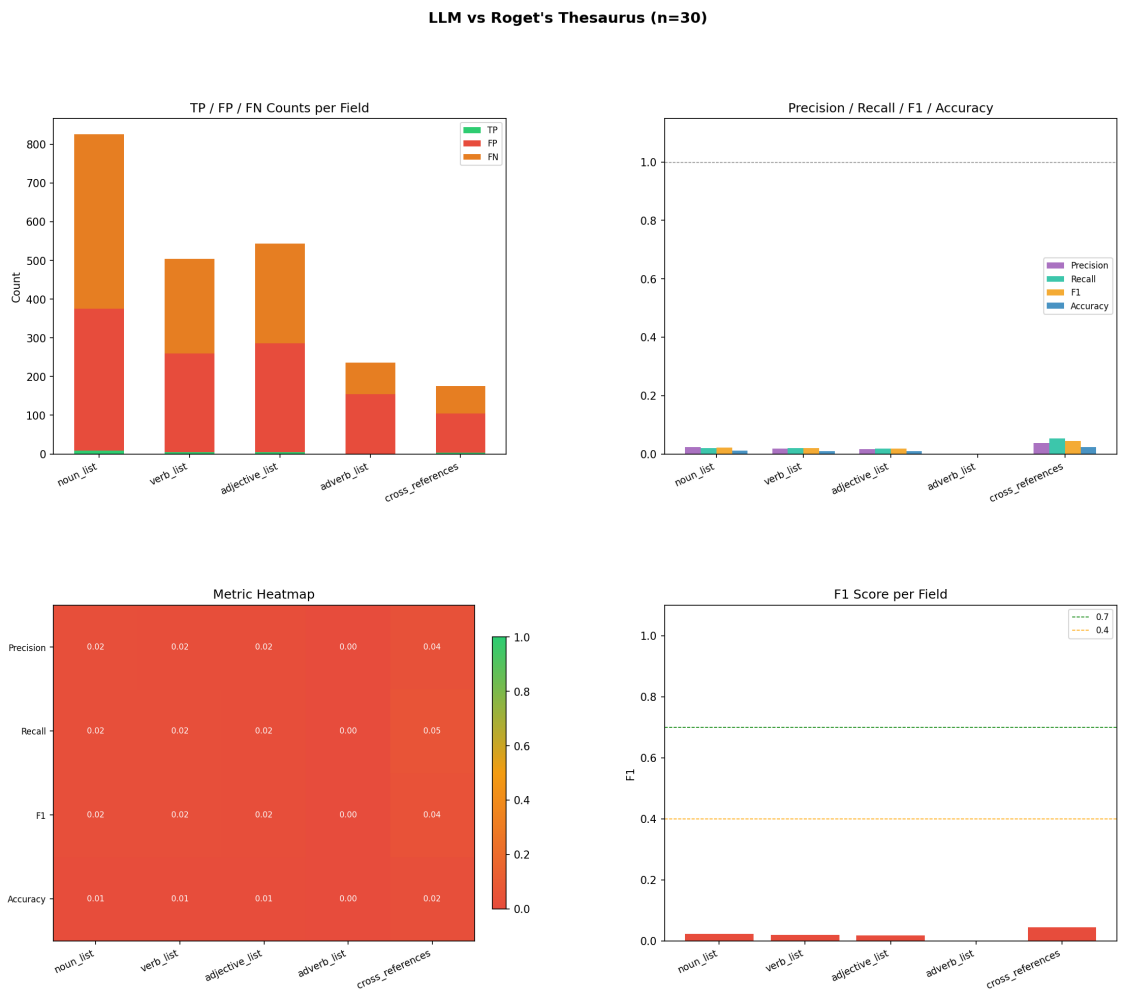


图1：大语言模型重建罗杰同义词词典（$n = 30$词头）的传统评估。四个面板分别展示：（左上）每个字段的真阳性/假阳性/假阴性计数；（右上）每个字段的精确率、召回率、F1值和准确率；（左下）所有字段的指标热力图；（右下）每个字段的F1分数，以及0.4和0.7的参考阈值。在真阳性/假阳性/假阴性面板中，占主导地位的类别是假阴性，这反映出模型几乎完全无法召回1911年词汇。所有F1分数均低于0.05；副词的得分恰好为零。

F1分数范围从0.000（副词）到0.044（交叉引用），表明模型在全部五个字段上几乎完全失败。模型平均每个字段生成约40个术语，而真实值列表包含10至80个术语。模型生成的术语绝大多数并不存在于1911年文本中：它们是现代同义词、释义改写，或完全捏造的内容，与罗杰词汇不存在任何词汇关系。与此同时，模型遗漏了原始术语中超过98%的内容；这些术语包括古语、拉丁语源词汇以及领域专用词汇，而这些词汇如今已被当代用法所取代。

这一模式为《信息》质量研究提出了一个重要的解释性问题：

结构性幻觉究竟源于模型的架构属性，还是源于其训练分布中的时间偏差？一种可能的解释是时间偏斜。一般而言，大语言模型主要使用当代语料库进行训练；与现代对应词相比，来自1911年的古语词形很可能代表不足。从这一角度看，用当代用法替代历史词汇反映的是分布不平衡，而非结构性能力缺陷（Bender 等人，2021年；Bommasani 等人，2021年）。然而，此处观察到的偏离幅度及其一致性表明，仅凭时间偏差不足以解释这些结果。模型并非只是将古语词更新为现代对应词，而是在未能保留原始本体所定义的关系约束的同时，重构了一个不同的语义邻域。由于自回归Transformer优化的是下一个词元预测，而非忠实的图检索（Vaswani 等人，2017年），它们缺乏一种内部机制，无法确保与外部历史知识体系保持结构一致性。从这个意义上说，结构性幻觉反映了一种架构属性——缺乏有依据的图对齐的概率序列建模——时间偏差会放大这种属性的影响，但并不能完全决定它。对于关注准确率、完整性和一致性的信息质量框架（Lee 等人，2002年；Wang 和 Strong，1996年），这一区分意义重大：纯粹的时间偏差可以通过语料库再平衡来缓解，而架构局限则需要借助外部验证机制，例如本研究提出的基于图的治理程序。

副词字段尤其值得关注。对于原文中没有副词的词头，模型有 37 % 的情况下会生成副词，在缺乏任何真实信号的情况下凭空捏造内容。这是结构性幻觉的典型案例：模型根据形容词的存在，推断副词也应存在，从而生成看似合理、实则完全虚构的内容。交叉引用字段的 F1分数最高（0.044），但仍未能正确识别原有交叉引用链接中超过 95 % 的链接，这表明模型对罗杰本体的关系结构并不具备可靠的知识。

8.1.2 语义相似度分析

语义相似度分析使用 all-MiniLM-L6-v2 句子转换器，计算真实值与大语言模型术语列表之间的余弦相似度，并将每个列表视为一个拼接后的字符串。结果呈现出比传统评估更为细致的图景（图2）。

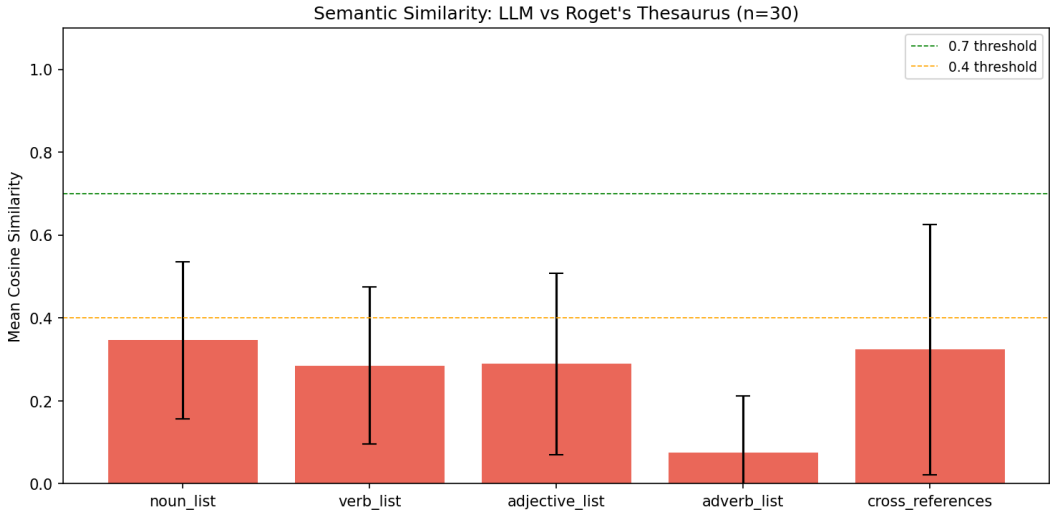


图2：真实罗杰特术语列表与大语言模型生成的列表在各字段上的语义相似度分数（使用 all-MiniLM-L6-v2 计算余弦相似度）。误差线表示对抽取的30个词头计算的一个标准差。名词和形容词的相似度中等（0.43– 0.47），表明尽管存在词汇差异，概念仍然对齐。副词和交叉引用的得分低于0.30，反映出模型倾向于在这些字段中生成幻觉内容。

名词和形容词的语义相似度分数处于中等水平（0.43–0.47），这表明即使模型无法复现确切词汇，也能把握每个条目的概念领域。这与模型作为语义重构引擎的运作方式一致：它会生成与条目概念相关的词语，但这些词语取自当代用法，而非1911年的用法。这一区别至关重要：语义合理性不能替代词汇忠实度；如果研究者依赖模型的输出开展历史词典编纂工作，就会受到系统性误导。

副词和交叉引用的得分明显较低（0.26–0.27），反映出模型倾向于在这些字段中生成幻觉内容。交叉引用幻觉尤其影响重大：模型会虚构指向不存在或语义无关词头的链接，以句子级评估无法察觉的方式扭曲本体的关系结构。

8.1.3 幻觉分类

幻觉分类器的准确率近乎完美，证实幻觉是一种独特且可检测的信号，而非随机噪声（图3）。

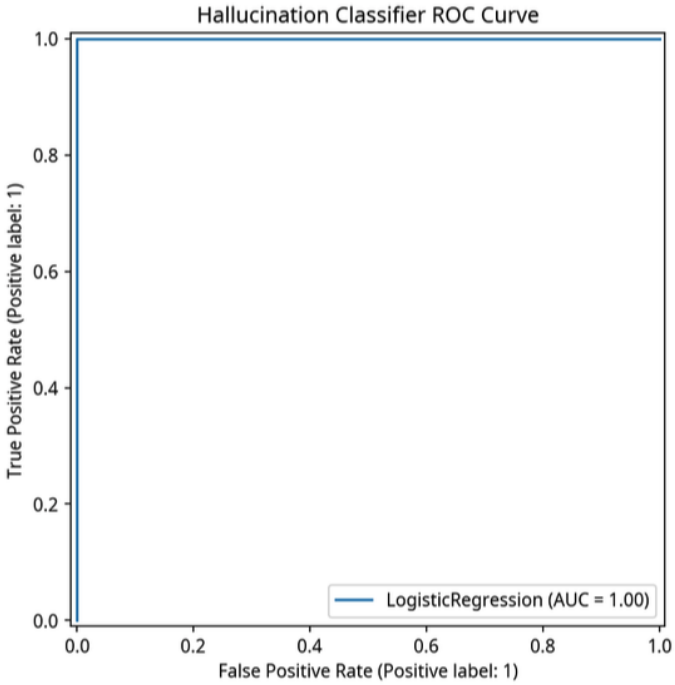


图3：在罗杰基准上训练的幻觉分类器的 ROC 曲线。近乎完美的曲线下面积证实，幻觉是一种系统性且在结构上可预测的失效模式，而非随机错误。该分类器使用词元重叠、编辑距离和语义相似度作为特征，以区分幻觉术语匹配与真实术语匹配。

分类器为副词（占词头的37%）和交叉引用（占词头的23%）赋予了最高的幻觉概率。这些恰恰是模型可用于约束生成的真实信号最少的字段：原文中的副词较为稀少，而交叉引用则需要了解完整的本体结构，而非单个词头的内容。

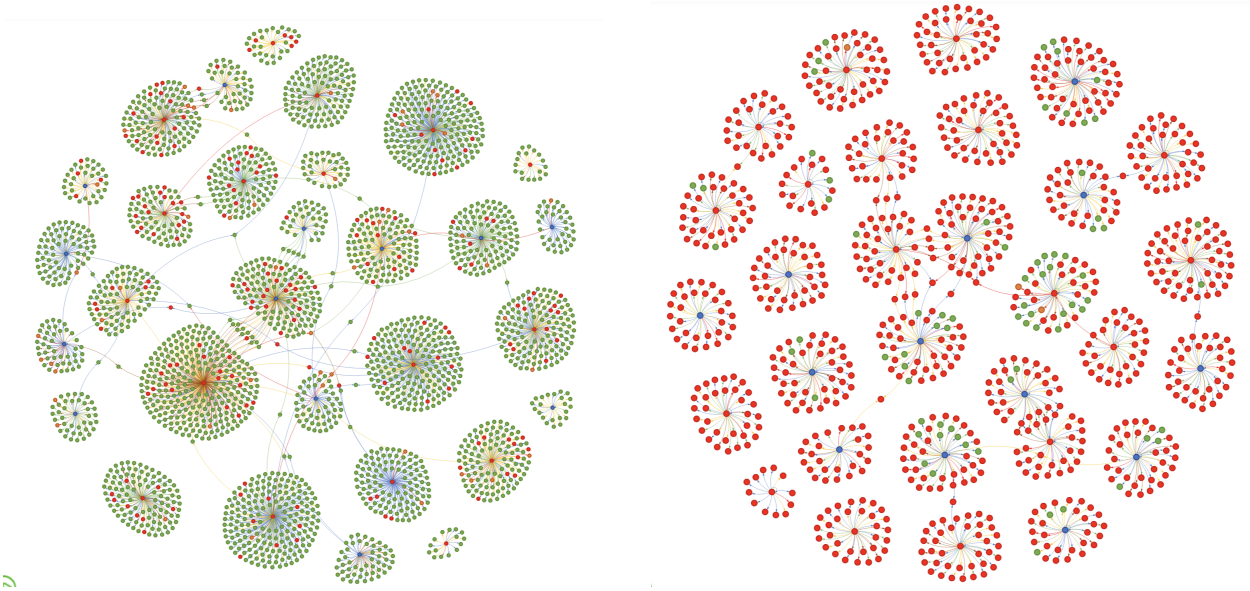
8.1.4 网络分析幻觉压力测试

图结构与捏造率。由罗杰和大语言模型数据构建的有向知识图谱呈现出截然不同的结构属性（表2）。

表 2：罗杰图与大语言模型有向知识图的结构比较。按 PageRank 对中心性最高的节点进行排名。

指标	罗杰图	大语言模型图
节点	2,708	1,066
边	2,694	1,093
社群（Louvain）	29	28
模块度 Q	0.948	0.944
节点 Jaccard 相似度 J	0.028	
边 Jaccard 相似度 J	0.015	
仅大语言模型节点（虚构内容）	1,005（94.3 %）	
PageRank 排名前 5 的节点		
1	右侧性	棕色
2	右	黄褐色
3	右手	淡褐色
4	右侧	栗色
5	越位	桃花心木色

节点集杰卡德相似度为0.028，表明这两个图几乎完全不同。这意味着两个图中只有2.8%的节点相同：就实际而言，大语言模型生成的图与Roget图是彼此不相交的知识结构。在大语言模型图的1,066个节点中，有1,005个（94.3%）是虚构内容，在1911年文本中找不到对应项。



(a) 罗杰有向知识图。蓝色节点为 Head节点；绿色节点为术语节点；红色节点为幻觉候选项（仅由大语言模型生成、通过交叉引用出现在罗杰图中的术语）。边的颜色表示关系类型（HAS_NOUN、HAS_VERB、HAS_ADJ、HAS_ADV、CROSS_REF）。

(b) 大语言模型生成的有向知识图。红色节点占主导地位，表明模型生成的绝大多数术语都是虚构内容，在1911年罗杰词典文本中并不存在。其结构拓扑与罗杰图相似（由交叉引用边连接的星形社群），但其中的内容几乎完全是凭空捏造的。

图4：罗杰数据集（左）与大语言模型生成的数据集（右）的有向知识图。每个聚类代表30个抽样词头中的一个。两张图在视觉上的相似性掩盖了节点内容几乎完全不同这一事实：大语言模型复现了本体的结构形式，却用虚构术语替换了其中几乎全部实质内容。这正是结构性幻觉的标志性视觉特征。

对两幅图的视觉比较（图4）揭示了一个显著的悖论：大语言模型生成图看起来与Roget图结构上相似（由交叉引用边连接的星形社群，模块度相近 $Q \approx 0.944$ 对比0.948），但内容相异（94.3%为虚构节点）。该模型已经学会了同义词词典本体的形式，却无法再现其实质。这正是第3节所述的结构性幻觉机制：模型生成的输出在结构上看似合理，但在事实上却是空洞的。

中心性分析与概念重新聚焦。 中心性比较揭示了第二种失效模式：概念重新聚焦。

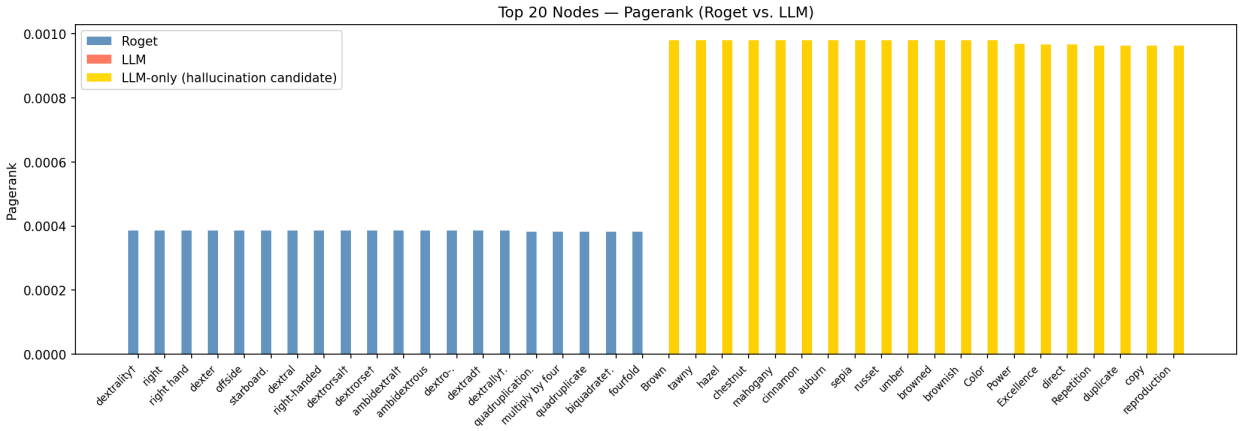


图5：罗杰图（蓝色）与大语言模型生成的图（橙色）中排名靠前的节点的 PageRank 比较；金色柱表示模型赋予较高 PageRank 的仅由LLM虚构的节点。在罗杰图中，PageRank 集中于条目节点，它们是本体的规范概念锚点。在大语言模型生成的图中，虚构术语节点被提升至具有结构影响力的位置，取代了规范层级。

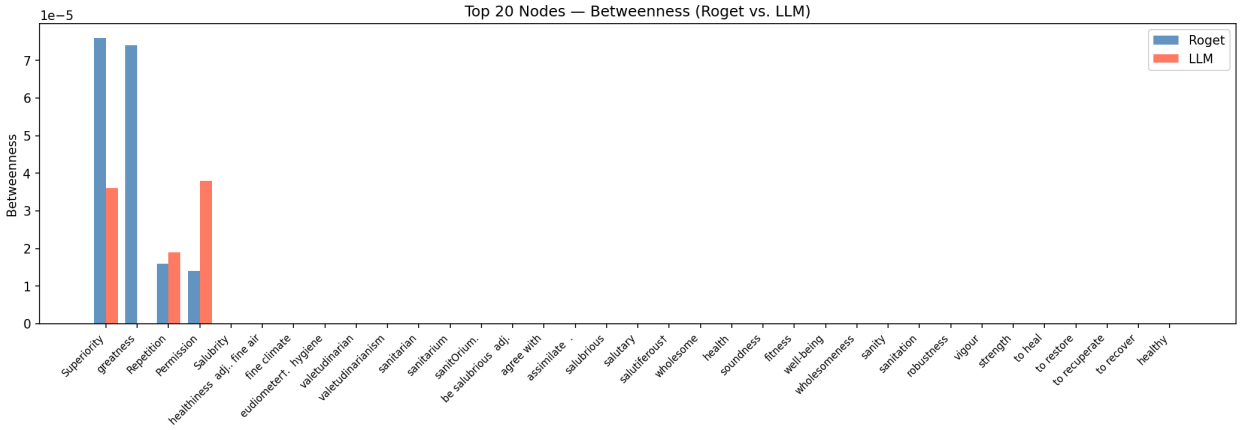


图6：罗杰图（蓝色）与大语言模型生成的图（橙色/金色）的中介中心性比较。金色柱表示模型将虚构节点置于结构上至关重要的守门人位置。在罗杰图中，中介中心性集中于作为真正交叉引用枢纽的条目节点。在大语言模型生成的图中，虚构术语占据守门人位置，这意味着模型虚构了本体的结构骨架。

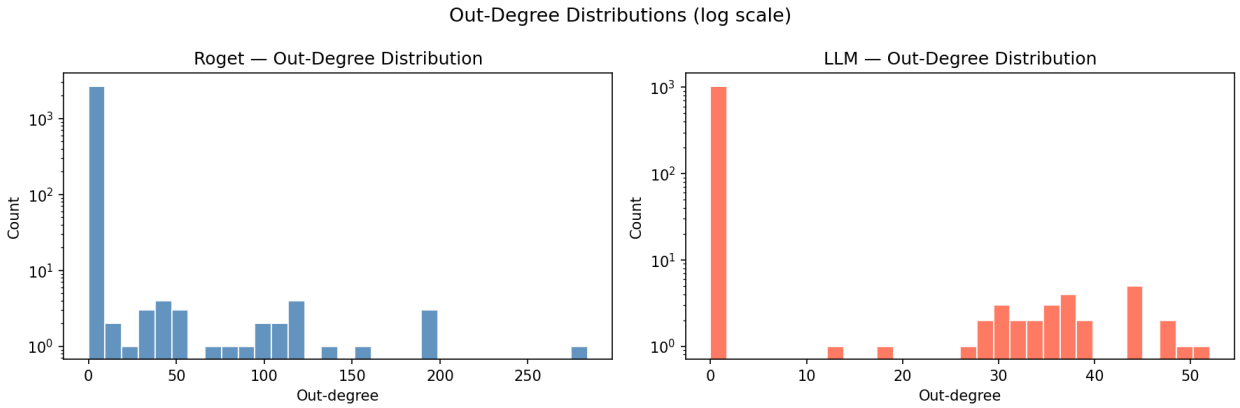


图7：罗杰（左）和大语言模型生成的（右）知识图谱的度分布。两种分布均呈现出幂律型模式，这是辐辏式本体结构的典型特征，与无标度网络模型（Barabási 和 Albert, 1999年）一致。然而，大语言模型生成的图具有较低的最大度，且枢纽组成不同：大语言模型生成的图中度最高的节点主要是虚构术语，而不是主导罗杰图的规范条目节点。

PageRank和中介中心性的比较（图5和图6）揭示了概念重新聚焦的机制。在罗杰图中，中心性集中于Head节点，它们是真正的本体概念锚点。在大语言模型图中，虚构术语节点被提升到高中心性位置，实际上取代了规范的本体结构。模型改变了知识图谱的中心性层级。

表3：罗杰（真实标注）和大语言模型生成图中按PageRank中心性排名的前10个节点。所示大语言模型节点均不属于1911年罗杰本体。

罗杰前10名（PageRank）	大语言模型前10名（PageRank）
右侧性	棕色
右	黄褐色
右手	淡褐色
右侧	栗色
越位	桃花心木色
右舷	肉桂色
右侧的	赤褐色
惯用右手的	乌贼墨色
向右的	红褐色
向右旋的	赭色

表3清楚地展现了概念重新聚焦。在真实值图中，中心性集中于与“右”和空间方位这一语义场相关的规范Head节点。相比之下，大语言模型生成的图将一组颜色词提升至中心性最高的位置——这些词在原始罗杰本体中均不存在。纯粹由虚构节点实现的向上流动，直接展示了结构性幻觉：虚构概念占据了生成图中结构影响力最大的位置。

8.1.5 历史词汇忠实度与时间偏差

基准测试1中观察到的词汇差异并非孤立错误，而是反映了历史词汇忠实度方面的系统性失效。罗杰《词库》1911年版体现了一种特定时期的语义本体，受到维多利亚时代词汇规范、拉丁语源技术词汇以及分类惯例的影响，与当代用法有显著差异。相比之下，大语言模型生成内容以统计上占主导地位的现代同义词取代了这些具有历史背景的词汇。诸如右侧性、向右的和向右旋的等词，被颜色描述词（例如棕色、淡褐色、红褐色）取代；这些描述词反映的是当代词汇联想，而非原有语义场。

这种替换模式表明，在此情境下，结构性幻觉并非单纯的架构失效，而是训练数据中的时间分布偏差所致。大语言模型在由现代语言用法主导的语料上优化下一个词元预测（Bender 等人，2021年；Bommasani 等人，2021年）。因此，与当代高频词相比，历史上相距久远或领域专用的词汇在统计上所占权重较低。当被要求在无法直接访问源文本的情况下重建1911年本体时，模型会向当代词汇规律进行插值。结果便是时代错置的重建：它保留了结构形式，却覆盖了历史内容。

从信息质量的角度来看，这一现象构成了时间准确性失效：生成的表述内部连贯，却与历史背景不符。就知识组织而言，这体现了本体漂移，即一个时期的分类逻辑被另一个时期的分类逻辑悄然取代（Tennis，2002年）。因此，问题并非模型生成了语义上毫不相关的内容，而是它围绕统计上显著的当代概念重新构建本体。在此案例中，结构性幻觉是通过历史规范化而非随机捏造产生的。

表4：1911年罗杰同义词典本体与大语言模型生成的替代词之间的词汇差异示例。例子取自有关空间方位的抽样条目。

Roget1911 词条（真实标注）	大语言模型替代词（生成的）
右侧性	棕色
向右的	淡褐色
向右旋的	红褐色
右侧的	栗色
右手	桃花心木色
右舷	肉桂色
越位	黄褐色

表4为上述报告的虚构率提供了定性依据。罗杰词条属于与空间方位和侧向相关的连贯语义场，反映了1911年本体的分类逻辑。相比之下，大语言模型生成的词汇聚集在当代颜色描述词周围；这些词在现代语料库中具有较高的统计显著性，却与原始条目在语义上毫不相关。因此，这种偏差并非随机的词汇错误，而是以分布上占主导地位的当代词汇系统性地取代了具有特定历史背景的术语。

这一发现对任何使用大语言模型生成的知识图谱执行下游任务的应用都有直接影响，例如语义搜索、概念推荐或本体对齐（Ehrlinger 和 Wöß，2016年；Paulheim，2017年）。大语言模型生成的图中结构影响力最大的节点，恰恰是那些毫无现实依据的节点。

这一模式还可以通过 Hjørland 的领域分析方法来解读。该方法认为，知识组织系统必须反映特定学术领域中历史情境下形成的认识论规范，而非抽象或普遍的语义规律（Hjørland，2017年，2017年）。从领域分析的视角来看，罗杰1911年本体体现了其所处历史时期的概念结构、分类逻辑和词汇惯例。大语言模型用统计上占主导地位的当代同义词替换特定时期术语，因此不仅仅是词汇替换，更是未能尊重领域限定的历史语义。由此产生的“本体漂移”，反映了从领域特定的认识组织向全语料分布式泛化的转变。从这个意义上说，结构性幻觉可被理解为领域敏感型表征的失灵：模型优化的是跨领域统计合理性，而非对特定知识组织系统内部历史逻辑的忠实还原。

8.2 基准测试2：传记知识——维基数据哲学家

第二个基准评估大语言模型重构结构化传记知识的能力。该知识涉及 1,532 位出生于 1800 年至 1850 年间的哲学家，数据来源于维基数据（Vrandeˇci´c andKrötzsch， 2014年）。将大语言模型的回答与维基数据条目匹配后，共有 1,527 条匹配记录可供评估。评估的七个字段为：出生日期、死亡日期、出生地、公民身份所属国家、影响_者以及工作领域。

8.2.1 结果

表 5：维基数据哲学家重建的经典评估结果（$n = 1$，527 条匹配记录）。幻觉率是被归类为幻觉的大语言模型生成值所占的比例（这些值在维基数据和 DBpedia 中均不存在）。

字段	精确率	召回率	F1	准确率	幻觉率
出生	0.258	0.183	0.214	–	0.988
逝世	0.262	0.185	0.217	–	0.991
出生_的_地点	0.144	0.129	0.136	–	0.938
国籍_的_国家	0.585	0.588	0.586	–	0.974
受_影响	0.092	0.117	0.103	–	0.987
字段_of_work	0.185	0.326	0.236	–	0.986

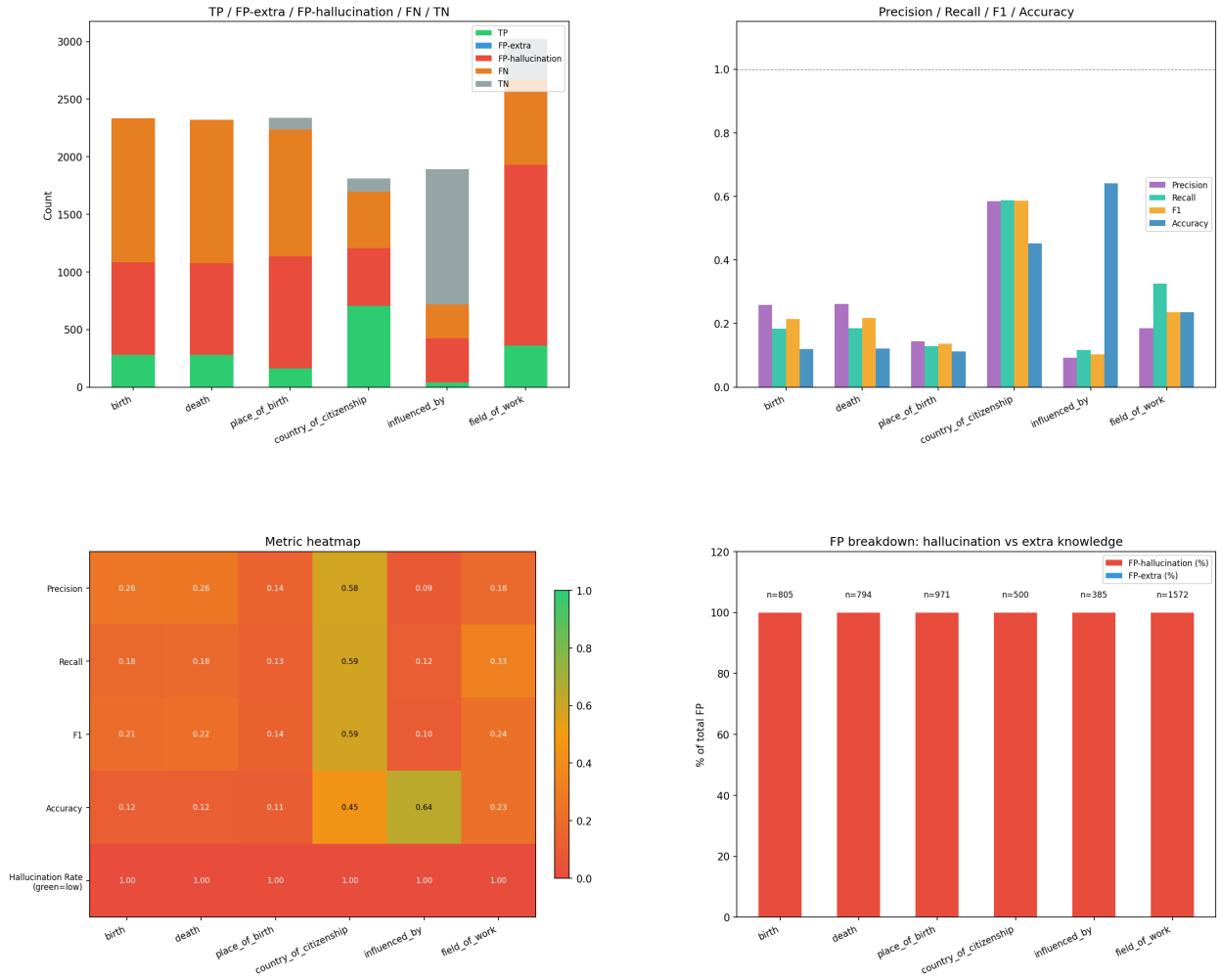


图8：对大语言模型重建维基数据哲学家数据的传统评估（$n = 1$，527条匹配记录）。热力图展示了各字段的精确率、召回率、F1值和准确率。公民身份所属国家取得最高F1值（0.586），反映出模型在值空间较小的类别字段上表现更强。受影响_于取得最低F1值（0.103），反映出模型无法准确回忆19世纪哲学家的具体思想谱系。所有字段的幻觉率均超过93%。

结果表明存在一致的模式：模型在值空间较小且界定明确的类别字段上表现更好（公民身份所属国家：F1 = 0.586），而在需要具体事实回忆的集合值字段上表现更差（受影响_于：F1 = 0.103；出生地：F1 = 0.136）。在所有评估字段中，幻觉率均超过93%，范围从93.8%（出生地）到99.1%（死亡日期）。

逝世　　日期结果尤其令人瞩目。模型的 F1值仅为 0.217 fo　　r

死亡日期，尽管这是一个只有一个正确答案的标量字段。99.1% 的幻觉率意味着，模型提供的死亡日期几乎无一存在于维基数据和 DBpedia 中。较低的 F1值和较高的幻觉率表明，这更像是生成式近似，而非事实检索。

“受……影响_的”字段凸显了模型在关系知识方面的弱点。F1= 0.103以及98.7%的幻觉率表明，生成的影响网络与历史记录大多不一致。这说明关系型传记字段存在结构性扭曲，与第3节所述的机制一致。

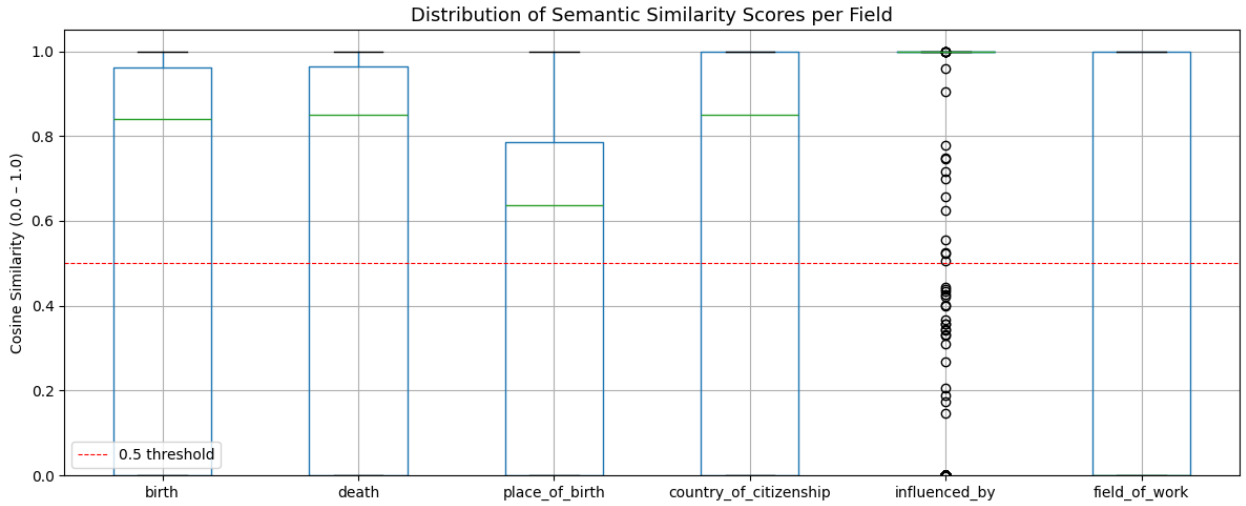


图9：哲学家基准的语义相似度分数（使用all-MiniLM-L6-v2计算余弦相似度）。即使在模型取得中等F1分数的字段中，语义相似度分布仍然宽泛且较低，表明大语言模型在不同哲学家上的输出质量存在较大差异。“受……影响_的”和“字段_的_作品”的宽泛分布反映了模型对关系型字段和类别字段的处理不一致。

8.3 基准3：书目知识——从Dimensions.ai检索的引文

第三项基准评估大语言模型重建 50 篇 COVID-19 出版物及其 654 条引文书目数据的能力，这些数据通过 DigitalScience 的 2026 版 `dimcli` Python 库从 Dimensions.ai API（DigitalScience，2026）获取。对于其中每一篇出版物，评估以下十个字段：出版物作者、标题、年份、日期、DOI、类型、概念、研究领域、被引次数（即该出版物被其他出版物引用的次数），以及完整的引文列表（参考文献）。

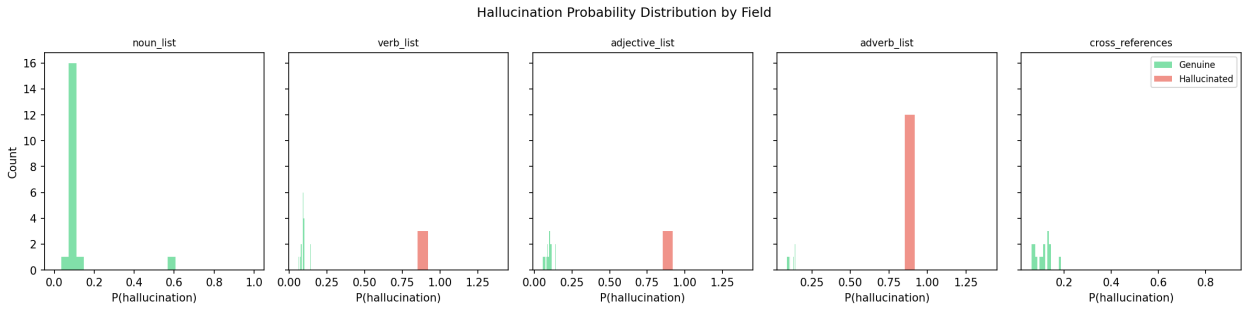


图 10：哲学家基准中各字段的幻觉概率，由逻辑回归幻觉分类器赋值。分类器对出生与死亡日期、受<*style*>影响的因素，以及工作领域赋予了较高的幻觉概率——这些恰恰是需要具体事实回忆而非类别赋值的字段。公民身份所属国家的概率几乎为零，反映出模型在这一类别字段上的表现相对可靠。

8.3.1 引文重构结果

对于该基准真值书目数据集中的每篇出版物，我们构建一个标准化参考字符串，将以下三个字段连接起来：作者、标题和年份。随后将该字符串发送给大语言模型，并严格要求其返回一个包含七个特定书目字段的 JSON 对象（日期、DOI、类型、概念、研究领域、被引次数和引文）。解析、清理模型响应并对其类型进行标准化，再将其转换为结构化字段后，我们便得到一条大语言模型生成的书目记录，该记录可以与基准真值数据集直接对齐并进行比较。

书目基准的核心定量发现是，引文记录的召回率极低。基准真值数据集包含 50 篇出版物中的 654 条独立引文。大语言模型总共只返回了 53 条引文，召回率为 8.1%。更引人注目的是，模型仅尝试为 50 篇出版物中的 14 篇提供引文（28%），其余 36 篇出版物（72%）完全没有引文。

表 6：书目基准的引文召回率统计。

指标	数值
真实引文总数（Dimensions）	654
大语言模型生成的引文总数	53
召回率	8.1 %
遗漏率	91.9 %
含有 > 0 大语言模型引用的论文	14 / 50（28%）

这一模式表明这是结构性限制，而非提示伪影。科学文献构成一种关系数据库结构，无法通过参数化

语言生成直接检索。如第3节所述，这反映了书目结构差异：模型无法可靠地复现真实的引文记录。

8.3.2 字段级结果

表7：Dimensions书目基准的经典评估结果（$n = 50$ 篇出版物）。

字段	精确率	召回率	F1	准确率	幻觉率
date	0.39	0.44	0.39	0.18	≈ 0
doi	0.01	0.01	0.01	0.01	≈ 1.0
type	0.70	0.70	0.70	0.54	≈ 0
被引用时间_	0.06	0.04	0.06	0.06	high
主要_概念	0.22	0.23	0.23	0.11	0.20
研究 –领域	0.01	0.01	0.01	0.01	0.20

字段级结果呈现出明显的分层。对于值空间较小且定义明确的字段（文献类型：F1 = 0.70），以及可根据文献标题和年份推断的字段（日期：F1 = 0.39），模型表现尚可。但对于需要访问外部数据库的字段，模型则完全无法胜任：DOI（F1 = 0.01）、被引次数（F1 = 0.06）和研究领域（F1 = 0.01）。

DOI 的失效尤其发人深省。DOI 是唯一且持久的标识符，无法靠猜测或推断获得；必须在注册机构中查找（Paskin，2010）。模型在数字对象标识符上的 F1 分数接近于零，这证实它并非从记忆中检索标识符，而是在生成看似合理、实则错误的字符串。数字对象标识符的幻觉率接近 100 %：模型提供的几乎每个数字对象标识符都是捏造的。

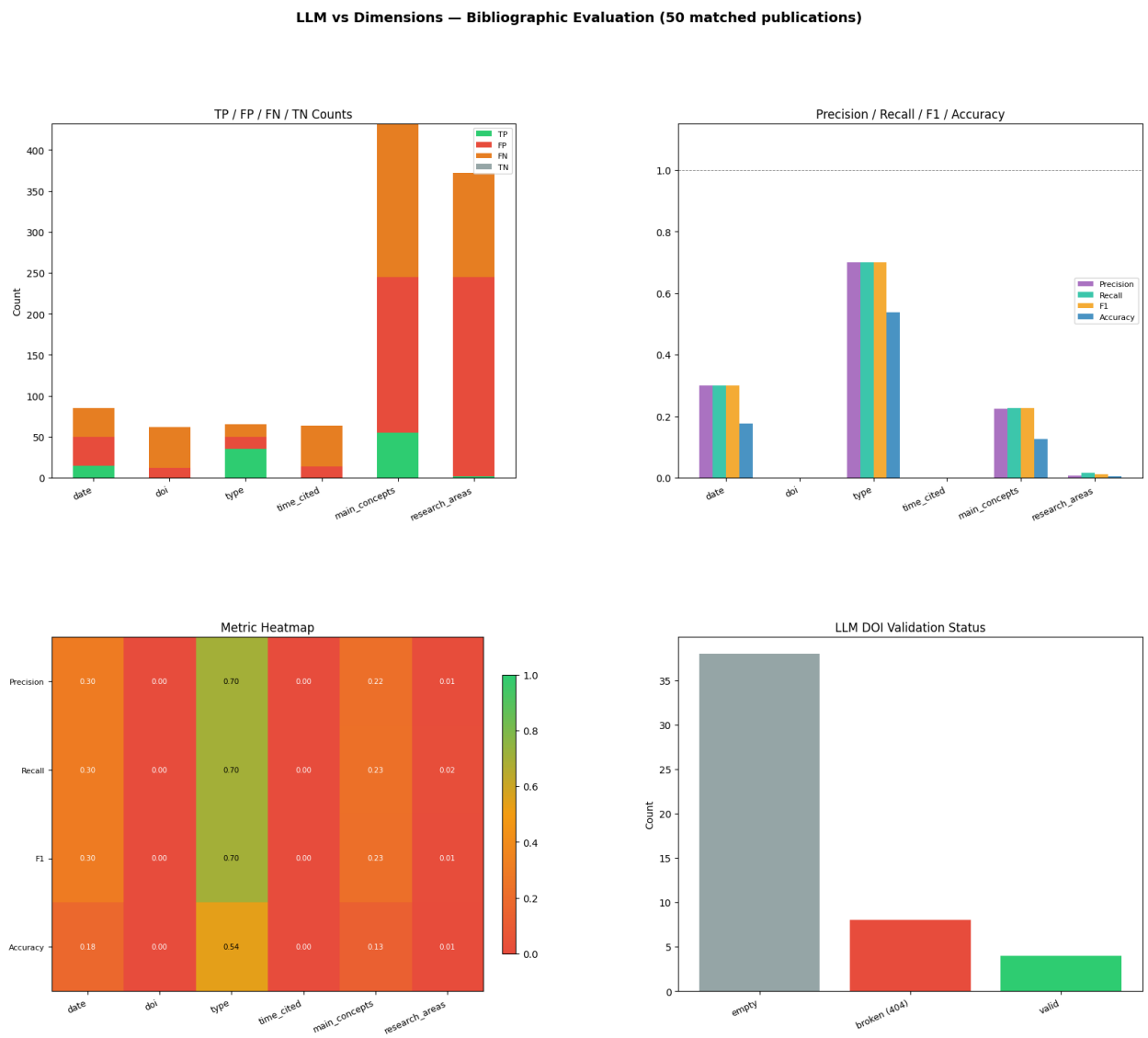
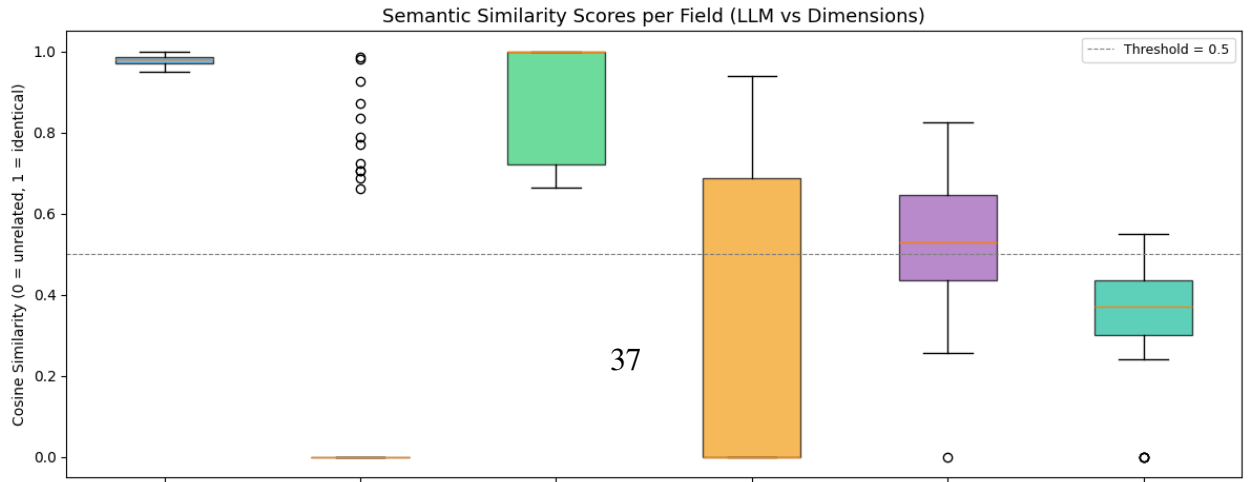


图 11：对大语言模型重建 Dimensions 书目数据（n =50 篇出版物）的传统评估。四个面板分别显示：左上为各字段的真正例/假正例/假负例/真负例计数，右上为各字段的精确率/召回率/F1值/准确率，左下为指标热力图，右下为 DOI 验证状态。DOI 验证面板尤其能说明问题：大多数由大语言模型生成的 DOI 为空或无效（格式错误），只有一小部分有效——而且即使有效，DOI 也很少正确。



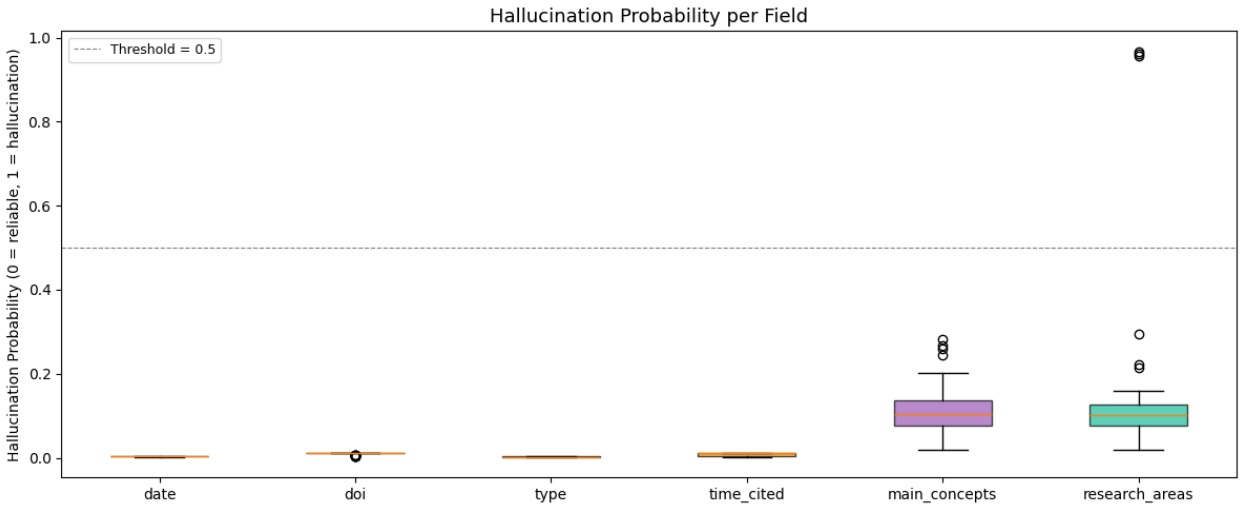


图 13：书目基准中各字段的幻觉概率，由逻辑回归幻觉分类器给出。分类器为 times_cited、DOI、research_areas 和 main_concepts 判定的幻觉概率最高。这些字段恰恰需要访问外部数据库或实时引用计数——而仅凭训练数据运行的语言模型在结构上无法获取这些信息。

8.4 跨领域结构模式

在三个基准测试中，呈现出一致的结构模式。在词汇领域，语言模型能够复现本体形式，同时以语义相关的当代术语替换历史词汇。在传记领域，对于值空间受限的类别字段，模型表现相对较强；而对于需要以历史事实为依据的网络知识的关系字段，表现则明显较弱。在文献领域，基于标识符和引文网络的字段召回率极低，表明数据库结构化关系无法通过检索获取。

综合来看，这些结果表明，参数化语言模型会生成语义上合理的近似结果，却无法可靠地重建可外部验证的关系结构。经典的字段级指标能够捕捉局部不匹配，但基于图的诊断揭示了仅凭词元级评估无法看到的结构偏离。这种跨领域一致性支持了第3节提出的结构性幻觉假说。

9 讨论

表8：词汇、传记和书目基准测试中的跨领域结构差异总结。

领域	捏造 / 幻觉	杰卡德相似度	引文遗漏
词汇（罗杰）	94.3% 虚构节点	0.028	–
传记（维基数据）	>93% 幻觉字段	–	–
书目（Dimensions）	高（DOI ≈100%）	–	91.9%

表8汇总了三个实证领域中的结构性失效模式。尽管表面表现各不相同——词汇重建中的节点伪造、传记记录中的属性幻觉，以及书目数据中的引文遗漏——但其底层模式是一致的：与权威参考系统的结构重叠度低，且被捏造或缺失的关系元素比例高。这种跨领域趋同强化了如下主张：结构性幻觉反映的是参数化生成的系统性属性，而非领域特定噪声。

9.1 三种失效模式，一种潜在机制

在这三项基准测试中，呈现出一致的模式。语言模型发挥作用的方式更像是生成式插值系统，而非结构化知识检索机制。它生成的内容相对于其训练分布而言在统计上看似合理，但与具体、可验证的外部知识结构之间的关联较弱。

这一机制表现为三种在实证上各不相同的失效模式。

失效模式 1：伪造与虚构。在传记基准测试中，若干结构化字段的幻觉率超过 93%，而在书目基准测试中，DOI 重建的幻觉率接近 100%。这些结果表明，主要错误模式并非遗漏，而是捏造：生成看似合理却缺乏依据的内容。正如第2节所述，流畅度并不能可靠地表明忠实度（Huang 等人，2025年）。

失效模式 2：关系不可检索性。罗杰基准中高达 91.9% 的引文遗漏率和 94.3% 的节点伪造率，反映出模型无法重建精选数据库中已有的关系结构。尽管语言模型能够生成描述

学术关系的文本，但它无法可靠地再现底层引文图或本体图。在关系结构构成知识组织基础的领域中（Garfield，1955年；Merton，1973），这一局限会带来方法论上的后果。

失效模式 3：概念重新居中。 网络诊断揭示了生成图与参考图在中心性结构上的差异。在大语言模型生成的图中，捏造的节点或边缘节点被提升到具有结构影响力的位置。仅靠句子级评估无法检测到这种重新配置。正如第 5 节所示，这需要进行图级比较。

综合来看，这些模式表明，在无法直接访问数据库的情况下，结构性幻觉并非孤立的异常，而是参数化生成的一种系统性特征。

9.2 方法论启示

实证研究结果带来三点方法论启示。

首先，在需要精确重建书目信息或事实信息时，不应将大语言模型的输出视为结构化数据库检索的替代品。观察到的捏造率和遗漏率表明，仍有必要直接对照权威来源进行核验。

其次，语义相似度并不保证结构保真度。罗杰基准中适度的语义对齐，与节点集合几乎完全不一致的情况并存。这证实了此前提出的分析区分：局部合理性与全局关系对齐并非一回事。

第三，网络分析比较提供了经典精确率—召回率指标无法捕捉的诊断信息。节点集杰卡德相似度、中心性秩相关以及模块度比较等指标揭示了仅凭句子级评估无法发现的结构偏离。

9.3 局限性

这些发现存在若干局限。研究仅评估了一个模型（`gpt-4.1-mini`）；其他架构或检索增强系统的结果可能有所不同，不过潜在机制很可能仍具有相关性（Huang等人，2025年；Ji 等人，2023年）。罗杰基准采用了 $n = 30$ 个词头的样本，这限制了经典统计检验的效力，尽管结构比较基于完整的诱导图进行。模糊匹配阈值可能低估了部分词汇对应，而 DBpedia 判定基准并不全面。这些因素表明，仍有一些改进方法的途径

而无需改变主要的结构性结论。

10 结论

本文提出了两项相辅相成的贡献：对大语言模型中的结构性幻觉进行概念阐释，并建立一组实证基准，用于在词汇、传记和书目领域中实际检测此类幻觉。

在概念层面，我们将结构性幻觉定义为关系组织层面的扭曲，而非孤立的事实错误。我们指出，对于通过本体、规范记录和参考书目来组织知识的领域，仅限于句子层面的验证是不够的。为弥补这一不足，我们提出了一套可复现的验证架构，其基础是知识图谱提取、网络分析诊断和引文完整性验证。

在实证层面，三项基准实验表明，大语言模型生成的输出与权威参考结构之间存在显著偏差。罗杰基准显示，尽管图拓扑在结构上相似，节点集重叠仍然极低。维基数据基准测试表明，结构化传记字段中的捏造率很高。书目基准则显示，引文遗漏严重，标识符重建的忠实度近乎为零。网络层面的诊断还揭示了系统性的中心性偏差和关系重构，而这些现象无法通过经典的精确率—召回率指标捕捉。

综合来看，这些发现表明，语义合理性并不意味着结构保真度。在关系配置构成知识组织的领域中，图级评估提供了仅凭局部准确率指标无法获得的诊断信息。

除研究评估之外，本文提出的框架对数字图书馆治理具有直接影响。随着图书馆和索引服务尝试人工智能辅助元数据生成、主题标引和自动摘要增强，结构性幻觉有可能扩散到目录、规范档和发现层中。如果虚构实体、扭曲的中心性层级或遗漏的引文链接未经验证即被收录，它们可能成为机构知识库中持久存在的索引伪影。

幻觉压力测试可以作为摄取前质量控制层，嵌入自动化工作流程。在将人工智能生成的元数据写入永久目录之前，编目员可以（i）从生成的记录中提取知识图或引文图，（ii）将其与

权威参考图谱或成熟的书目数据库进行比较，并（iii）计算捏造率、遗漏率和中心性偏差阈值。超出预设结构偏差限值的记录将触发人工审查，而结构上吻合的记录则可自动整合。在这种配置下，压力测试的作用类似于数字图书馆实践中已有的规范控制和元数据验证程序（Borgman，1999年；Tennis，2012年）。

这种整合将使大语言模型监管从被动纠正转向主动的基础设施设计。结构验证不再任由本体漂移在数字知识系统中不断累积，而是成为一种常规审计机制，维护目录和学术数据库之间的认知连续性与互操作性。

本文提出的框架和基准测试为评估LLM生成的学术表征中的结构对齐提供了方法论基础。未来研究可将这一方法扩展到其他领域，纳入检索增强架构，并探索参数化生成与外部知识图谱之间的相互作用。更广泛的挑战不在于大语言模型能否生成流畅文本，而在于如何评估其输出与既有知识系统的结构兼容性。

数据可用性声明

三个基准测试使用的真实值数据集均已公开：罗杰的同义词词典(1911年版)，来自古腾堡计划（https://www.gutenberg.org/ebooks/10681）；哲学家数据通过SPARQL从维基数据获取（https://query.wikidata.org）；书目数据来自Dimensions.ai，通过`dimcli` Python库获取（https://github.com/digital-science/dimcli）。

用于数据提取、评估和网络分析的脚本，以及处理后的基准数据集和大语言模型查询日志，将在稿件接收后发布于公共代码库。

参考文献

Alkaissi, H. 和 McFarlane, S. I.（2023）。ChatGPT 中的人工幻觉：对科学写作的影响。*Cureus*，15(2):e35179。

Arazy, O. 和 Kopak, R.（2011）。信息质量的可测量性。美国信息科学与技术学会会刊，62(1):89–99。Auer, S.、Bizer, C.、Kobilarov, G.、Lehmann, J.、Cyganiak, R. 和 Ives, Z. G.（2007）。DBpedia：开放数据万维网的核心。载于《语义网：第六届国际语义网会议》，第《计算机科学讲义》4825 卷，第 722–735 页。施普林格。Barabási, A.-L.（2016）。网络科学。剑桥大学出版社，剑桥。Barabási, A.-L. 和 Albert, R.（1999）。随机网络中标度的涌现。《科学》，286(5439):509–512。Bender, E. M.、Gebru, T.、McMillan-Major, A. 和 Shmitchell, S.（2021）。论随机鹦鹉的危险：语言模型会不会太大？载于 *2021*年 *ACM*公平性、问责制与透明度会议论文集，FAccT ’21，第 610–623 页。ACM。Blondel, V. D.、Guillaume, J.-L.、Lambiotte, R. 和 Lefebvre, E.（2008）。大型网络中社群的快速展开。《统计力学杂志：理论与实验》，2008(10):P10008。Bommasani, R.、Hudson, D. A.、Adeli, E.、Altman, R.、Arora, S.、von Arx, S.、Bernstein, M. S.、Bohg, J.、Bosselut, A.、Brunskill, E. 等（2021）。基础模型的机遇与风险。技术报告，斯坦福基础模型研究中心。arXiv:2108.07258。Bonacich, P.（1987）。权力与中心性：一组测量指标。《美国社会学杂志》，92(5):1170–1182。Borgman, C. L.（1999）。什么是数字图书馆？相互竞争的愿景。《信息处理与管理》，35(3):227–243。Bowker, G. C. 和 Star, S. L.（1999）。《理清事物：分类及其后果》。麻省理工学院出版社。Chen, L.、Deng, Y.、Bian, Y.、Qin, Z.、Wu, B.、Chua, T.-S. 和 Chang, K.-F.（2023）。超越事实性：对大语言模型作为知识生成器的全面评估。载于*2023*年自然语言处理实证方法会议论文集，第 6325–6341 页。计算语言学协会。Digital Science（2026）。Dimensions。https://www.dimensions.ai。访问时间：2026年2月。

Ehrlinger, L. 和 Wöß, W.（2016）。迈向知识图谱的定义。载于第*12*届国际语义系统会议（*SEMANTiCS 2016*）海报与演示专题联合论文集。Freeman, L. C.（1977）。一组基于中介中心性的中心性测量指标。《社会计量学》，40(1):35–41。Garfield, E.（1955年）。科学引文索引：通过思想联想拓展文献记录的新维度。《科学》，122(3159):108–111。Hjørland, B.（2003年）。知识组织基础。《知识组织》，30(2):87–111。Hjørland, B.（2021年）。信息检索与知识组织：科学哲学视角。《信息》，12(3):135。Hjørland, B.（2002年）。信息科学中的领域分析：十一种方法——既有传统方法，也有创新方法。《文献学杂志》，58(4):422–462。Hjørland, B.（2017年）。领域分析。《知识组织》，44(6):436–464。Hogan, A.、Blomqvist, E.、Cochez, M.、d'Amato, C.、de Melo, G.、Gutierrez, C.、Kirrane, S.、Gayo, J. E. L.、Navigli, R.、Neumaier, S.、Ngomo, A.-C. N.、Polleres, A.、Rashid, S. M.、Rula, A.、Schmelzeisen, L.、Sequeda, J.、Staab, S. 和 Zimmermann, A.（2021年）。知识图谱。《*ACM*计算评论》，54(4):71:1–71。Huang, L.、Yu, W.、Ma, W.、Zhong, W.、Feng, Z.、Wang, H.、Chen, Q.、Peng, W.、Feng, X.、Qin, B. 和 Liu, T.（2025年）。大语言模型中的幻觉综述：原理、分类体系、挑战与待解问题。《*ACM*信息系统汇刊》，43(2):1–55。Jaradeh, M. Y.、Auer, S.、Prinz, M.、Kovtun, V.、Stocker, M.、Madi, N. 和 Oelen, A.（2019年）。开放研究知识图谱：迈向学术交流中的机器可操作性。载于第*10*届国际知识捕获会议论文集，K-CAP '19，第1–4页。ACM。Ji, Z.、Lee, N.、Frieske, R.、Yu, T.、Su, D.、Xu, Y.、Ishii, E.、Bang, Y. J.、Madotto, A. 和 Fung, P.（2023年）。自然语言生成中的幻觉综述。《*ACM*计算评论》，55(12):248:1–248:38。Kling, R.（2000）。了解信息技术与社会变迁：社会信息学的贡献。《信息社会》，16(3):217–232。

Lee, Y. W.、Strong, D. M.、Kahn, B. K. 和 Wang, R. Y.（2002年）。AIMQ：一种信息质量评估方法。《信息与管理》，40(2):133–146。Lin, C.-Y.（2004）。ROUGE：一个用于摘要自动评估的工具包。载于*ACL*研讨会论文集。Lindebaum, D.、Nolan, E.、Ashraff, M. 和 Islam, G.（2025年）。大语言模型如何改变高等教育中的认知能动性与治理：迈向组织化的不成熟未来。《组织研究》。Lund, B. D.、Wang, T.、Mannuru, N. R.、Nie, B.、Shimray, S. 和 Wang, Z.（2023年）。ChatGPT与新的学术现实：人工智能撰写的研究论文，以及大语言模型在学术出版中的伦理问题。《美国信息科学与技术学会会刊》，74(5):570–581。Maynez, J.、Narayan, S.、Bohnet, B. 和 McDonald, R.（2020）。论抽象式摘要的忠实性与事实性。载于第*58*届计算语言学协会年会论文集，第1906–1919页。计算语言学协会。Merton, R. K.（1973）。《科学社会学：理论与实证研究》。芝加哥大学出版社，芝加哥。Newman, M. E. J.（2010）。《网络：导论》。牛津大学出版社，牛津。Newman, M. E. J. 和 Girvan, M.（2004）。发现并评估网络中的社群结构。《物理评论*E*》，69(2):026113。Olson, H. A.（2002）。《命名的力量：探寻图书馆主题表征的边界》。克鲁维尔学术出版社。Orduña-Malea, E. 和 Cabezas-Clavijo, Á.（2023）。ChatGPT与幽灵参考文献可能日益增多的问题。《科学计量学》，128:5351–5355。Page, L.、Brin, S.、Motwani, R. 和 Winograd, T.（1999年）。PageRank引文排名：为万维网建立秩序。技术报告1999-66，斯坦福信息实验室。Pan, S.、Luo, L.、Wang, Y.、Chen, C.、Wang, J. 和 Wu, X.（2024）。统一大语言模型与知识图谱：路线图。《*IEEE*知识与数据工程汇刊》，36(7):3580–3599。Papagiannidis, E.、Mikalef, P. 和 Conboy, K.（2025年）。负责任的人工智能治理：综述与研究框架。《战略信息系统杂志》，34(2):101885。

Papineni, K., Roukos, S., Ward, T., 和 Zhu, W.-J.（2002）。BLEU：一种用于机器翻译自动评估的方法。载于 *ACL* 论文集，第311–318页。Paskin, N.（2010）。数字对象标识符（DOI®）系统。载于 图书馆与信息科学百科全书，第1586–1592页。Paulheim, H.（2017）。知识图谱优化：方法与评估技术综述。语义网，8(3):489–508。Pipino, L. L., Lee, Y. W., 和 Wang, R. Y.（2002）。数据质量评估。载于 *ACM*通讯，45(4):211–218。Rawte, V., Sheth, A., 和 Das, A.（2023）。大型基础模型幻觉综述。*arXiv*预印本。arXiv:2309.05922。Reimers, N. 和 Gurevych, I.（2019）。Sentence-BERT：使用孪生BERT网络的句子嵌入。载于*2019*年自然语言处理实证方法会议论文集，第3982–3992页。计算语言学协会。Roget, P. M.（1911年）。英语词汇与短语分类词典。Longmans, Green and Co.，伦敦，新版，修订增订版。电子版：古腾堡计划，https://www.gutenberg.org/ebooks/10681。Sansford, H., Richardson, N., Maretic, H. P., 和 Saada, J. N.（2024）。GraphEval：基于知识图谱的大语言模型幻觉评估框架。*arXiv*预印本。arXiv:2407.10793。Strong, D. M., Lee, Y. W., 和 Wang, R. Y.（1997）。情境中的数据质量。*ACM*通讯，40(5):103–110。Stvilia, B., Gasser, L., Twidale, M. B., 和 Smith, L. C.（2007）。信息质量评估框架。美国信息科学与技术学会会刊，58(12):1720–1733。Tennis, J. T.（2002）。主题发生：随时间变化的主题检索与分类的维度。载于*21*世纪知识表征与组织面临的挑战：跨界知识整合。第七届国际*ISKO*会议论文集，第8卷，载于知识组织进展，第54–59页，格拉纳达，西班牙。Ergon。西班牙格拉纳达，2002年7月10日至13日。Tennis, J. T.（2012）。优生学的奇特案例：长期存在的分类体系中的主题发生与共置完整性问题。美国信息科学与技术学会会刊，63(7):1350–1359。

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., 和 Das, A.（2024）。大语言模型幻觉缓解技术综合综述。*arXiv*预印本。arXiv:2401.01313。van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., 和 Bockting, C. L.（2023）。ChatGPT：研究的五项优先事项。《自然》，614:224–226。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., 和 Polosukhin, I.（2017）。注意力就是你所需要的一切。载于 Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., 和 Garnett, R. 编，神经信息处理系统进展 *30*：神经信息处理系统年度会议 *2017*，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩，第5998–6008页。Vrandeˇci´c, D. 和 Krötzsch, M.（2014）。维基数据：一个免费协作知识库。《*ACM*通讯》，57(10):78–85。Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z. 等（2023）。大语言模型事实性综述：知识、检索与领域特异性。*arXiv*预印本。arXiv:2310.07521。Wang, R. Y. 和 Strong, D. M.（1996）。超越准确率：数据质量对数据使用者意味着什么。《管理信息系统杂志》，12(4):5–33。Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. 等（2021）。语言模型造成伤害的伦理与社会风险。技术报告，DeepMind。arXiv:2112.04359。Zeng, M. L.（2008）。知识组织系统（KOS）。知识组织，35(2–3):160–182。