基于引文验证文献语境的跨学科数据集发现

Zhiyin Tan*L3S*研究中心 汉诺威莱布尼茨大学汉诺威，德国
zhiyin.tan@l3s.de

Changxu Duan达姆施塔特工业大学 ¨达姆施塔特达姆施塔特，德国
duan@linglit.tu-darmstadt.de

arXiv:2601.05099v1 [cs.DL] 8 Jan 2026

摘要—**针对某一研究问题找到合适的数据集仍然颇具挑战，因为现有数据集搜索引擎高度依赖元数据质量和关键词重叠，而这些往往无法捕捉科学研究的语义意图。我们提出一种文献驱动框架，从科学论文的引文上下文中发现数据集，使检索结果立足于实际研究用途，而非元数据的可用性。我们的方法结合大规模引文语境提取、借助大语言模型进行模式引导的数据集识别，以及保留溯源信息的实体解析。我们使用八个基于调查得出的计算机科学查询对该系统进行评估，发现其召回率显著高于 Google Dataset Search 和 DataCite Commons，归一化召回率的平均值为 47.47%，最高达到 81.82%。除了找回金标准数据集外，该方法还发现了调查中未记录的额外数据集。针对科学领域五个顶级类别开展的专家评估表明，相当一部分额外数据集具有较高实用性，其中一些还被专家认为对所选特定主题而言具有新颖性。这些发现表明，引文语境挖掘是数据集发现的一种有效且可泛化的范式，尤其适用于数据集缺乏充足或可靠元数据的情形。为支持研究的可复现性并推动未来扩展，我们在 GitHub 上发布了代码、评估数据集和结果：https://github.com/Fireblossom/citation-context-dataset-discovery**。

索引词—**数据集发现、引文语境挖掘、科学文献分析、跨学科研究、实体消歧、开放科学**

I. 引言

不同学科的研究人员日益依赖公开可用的数据集来推动知识发现、确保可复现性，并促进后续的数据集成与复用 [1]。开放数据集的数量持续快速增长。DataCite [2] 单独索引的已注册研究对象就超过 3000 万个，而 Google Dataset Search [3] 覆盖的数量还要多出数百万。

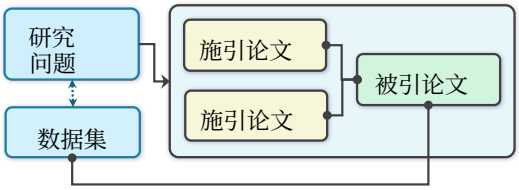


图 1：通过引文语境将研究问题与数据集联系起来。

然而，对于给定的研究问题，发现合适的数据集仍然十分困难。现有的数据集搜索引擎主要由元数据驱动。Google Dataset Search 会将查询与 schema.org 标记进行匹配，而 DataCite 则从结构化元数据字段中检索结果。当研究人员知道确切的数据集名称或精准术语时，这些系统能够很好地发挥作用；但当元数据不完整、不一致，或过于笼统而无法体现细致的适用性时，它们便会受到限制。先前的研究证实，不同知识库之间不一致的元数据标准会导致发现结果碎片化，并造成不可预测的搜索结果 [4]。

在实践中，研究人员很少一开始就知道精确的数据集名称。相反，他们会提出研究问题，这些问题通常源自调查主题、既有研究或新兴的跨学科问题，并寻找能够将这些问题付诸实践的数据集。对于此类任务，元数据关键词匹配并不足够。缺少的是上下文信息：为何选择某个数据集、如何使用该数据集，以及它帮助回答了哪些研究问题。

我们通过一个由文献驱动的数据集发现框架来弥补这一空白。我们的核心洞见是，科学论文已经包含了有关数据集使用情况的丰富上下文信息。引文上下文不仅描述使用了哪个数据集，还说明了该数据集如何支持研究设计，以及为何选择它。将这些上下文视为研究问题与数据集之间的语义桥梁（见图 1），我们便能超越静态元数据，使数据集发现立足于真实世界的研究实践。这一视角自然契合

Zhiyin Tan 获得了“HybrInt——通过可解释人工智能实现机器感知与交互中的混合智能”项目的资助（Zukunft Nds，Nieders¨achsisches Ministerium f¨ur Wissenschaft，资助编号：ZN4219）。Changxu Duan 获得了 InsightsNet 项目的资助（该项目由德国联邦教育与研究部（BMBF）通过资助编号 01UG2130A 提供资金支持）。我们衷心感谢 hessian.AIService Center 的支持（该中心由德国联邦研究、技术与航天部（BMFTR）通过资助编号 16IS22091 提供资金支持），以及 hessian.AI 创新实验室的支持（该实验室由黑森州数字战略与创新部通过资助编号 S- DIW04/0013/003 提供资金支持）。我们衷心感谢资源分配委员会批准的计算时间，并感谢 NHR 基础设施的一部分——NHR-Nord @G¨ottingen 超级计算机提供计算资源。本研究的计算使用了项目 nhr he starter 25563 所获批的计算资源。我们感谢 SIGIR 学生差旅资助。

研究者自身的工作方式：在寻找数据时，他们首先会查找与自身目标相似的论文，然后考察这些论文使用了哪些数据集。

在方法上，我们借鉴了科学信息提取中的技术。这些技术最初用于科学文献中的细粒度实体识别 [5], [6], ，并将其重新用于跨领域数据集发现。这种迁移使我们能够绕过对元数据的依赖，同时以科学文献中记录的真实使用模式为推荐提供依据。

我们的贡献如下：

- 我们提出了一种由文献驱动的数据集发现框架，通过利用科学论文中记录的使用情境，绕过对元数据的依赖。
- 我们开发了一个三阶段计算流程，将神经语言模型与引文图分析相结合，以实现可扩展的跨领域数据集提取。
- 我们建立了一套严谨且可扩展的评估方案，覆盖科学与技术领域（FOS）的六个一级领域[1]。对于计算机科学，我们使用源自调查论文的金标准数据集进行基于召回率的自动评估。为确保通用性，我们通过领域专家对相关性、实用性和新颖性进行评估，从而评估其余五个 FOS 领域。
- 我们证明，与领先的数据集搜索引擎相比，我们的方法显著提高了召回率；此外，它发现的额外数据集常被领域专家评定为具有较高实用性，并且在许多情况下对其研究领域而言具有新颖性。

本研究满足了一项实际需求，使研究人员能够在不熟悉的领域中高效发现相关数据集。同时，本研究也推动了科学知识组织智能化这一更广泛的愿景。

II. 相关工作

我们的工作通过系统分析科学文献来开展数据集发现。我们回顾了三个相关领域：现有的数据集搜索系统、基于文献的信息抽取，以及引文上下文分析。

A. 数据集搜索引擎与知识库

现有的数据集发现系统包括通用搜索引擎，例如 Google Dataset Search [3],，它依赖 schema.org 标记；以及 DataCite [2], ，它通过 DOI 注册确保元数据质量，但仅索引正式发表的数据集。Zenodo、Kaggle 和 Mendeley Data 等通用知识库接受来自各个学科的投稿，而蛋白质数据库 (PDB)、基因表达综合数据库 (GEO)、大学间政治与社会研究联合会 (ICPSR) 和英国国家语料库 (BNC) 等领域专用知识库，则在特定社群内提供经过整理的高质量资源，但仍各自为政，限制了跨学科发现。

这些系统存在一些根本性局限：严重依赖基于元数据的关键词匹配，缺乏对研究语境的语义理解，以及不同学科之间的术语不匹配 [7], [8]。很少有系统能够在元数据稀少或不一致时，将抽象的研究问题与数据集联系起来。

B. 基于文献的信息抽取

科学论文记录了资源使用的上下文细节，而这些细节无法仅凭元数据获取 [9]。以往工作利用这一信号构建了数据集提及抽取基准：SciERC [5] 专注于计算机科学论文，SciREX [6]将抽取扩展到文档级别，而 DMDD [10]则提供了大规模的多领域覆盖。

这些工作证明了大规模挖掘数据集提及的可行性，但其主要目标是构建目录或分析资源使用，而非引导研究人员找到与新研究问题相关的数据集 [11]。对全文进行抽取也可能计算开销很大，使实时搜索和推荐面临挑战 [12]–[14]。此外，这些方法还受限于对高质量结构化输入的依赖，以及缺乏可靠的跨学科评估。

我们在这项工作基础上，将抽取出的提及项视为连接研究问题与数据集的语义桥梁，而非目录编制的终点。为此，我们利用引文上下文实现多领域适用性，并支持高效、由文献驱动的数据集发现。

C. 引文上下文分析

引文上下文分析揭示了资源与应用之间的语义关系。Zhao 等人 [15] 开发了用于分类引文功能的框架，而 F¨arber 等人 [16]区分了实际使用的数据集与仅被提及的数据集。近期进展表明，神经语言模型擅长科学信息抽取 [13],，而 SOFT 等框架 [17] 则能将引文意图与内容类型区分开来。

然而，现有方法将上下文视为辅助信息，而非用于发现的主要语义信号。我们扩展了这些技术，将引文语境视为连接跨学科研究问题与数据集的语义桥梁，并利用科学文献作为涵盖数据集使用模式的综合知识库，构建了一个无需特定领域训练即可运行的通用框架 [1], [18]。

III. 方法

我们将数据集发现表述为科学文献中的语境信息抽取。核心思想是利用引文上下文作为语义证据，将自然语言研究问题与数据集联系起来。我们的系统采用三阶段流程，将高效的语料库访问与神经抽取和稳健的实体消歧相结合：(1) 可扩展的引文上下文检索，(2) 结合引文感知质量信号的神经数据集提及抽取，以及 (3) 数据集实体消歧与整合。这种设计直接解决了仅依赖元数据的系统长期存在的三个不足：语义不匹配、语境歧义和覆盖不完整。

[1]FOS：自然科学、工程与技术、医学与健康科学、农业科学、社会科学和人文学科。

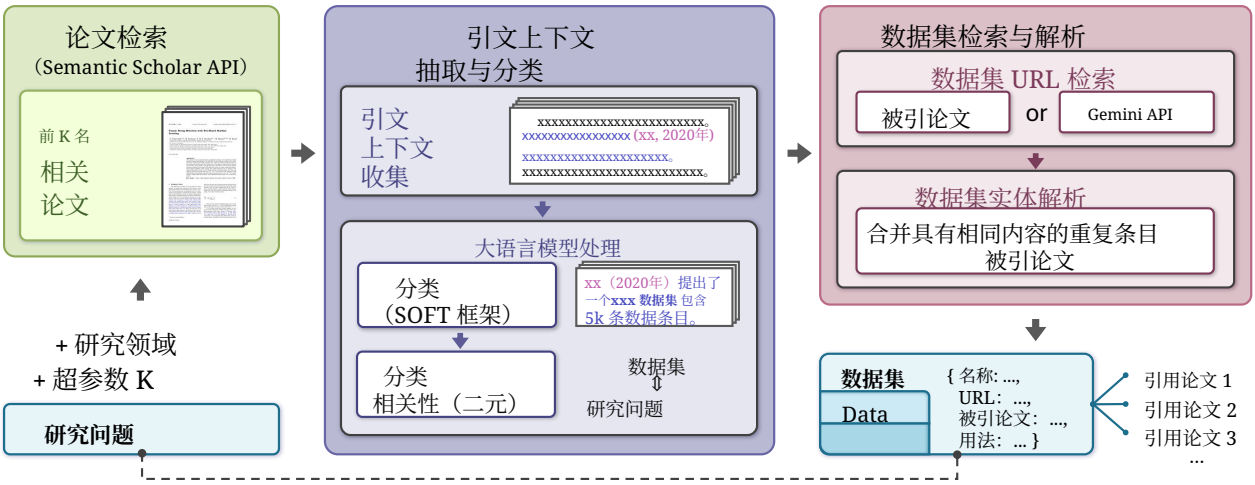


图 2：所提出的流水线通过引文上下文分析和元数据链接，将研究问题映射到相关数据集。

A. 问题形式化

给定研究查询 $\mathcal{Q}$ （自由文本）和可选的字段约束 $\mathcal{F}$ ，任务是从文献语料库中检索并排序数据集实体 $\mathcal{D} = \{d_1, \ldots, d_n\}$。在本研究中，我们将数据集 d 操作性地定义为任何在论文中被明确提及为使用、修改或用于评估的具名研究资源（例如语料库、基准数据集和数据库）。该定义不包括软件工具或通用方法的提及。

每个实体 d_i 都是一个结构化元组

$$d_i = \langle \text{name}_i,\ \mathcal{C}_i,\ \mathcal{M}_i \rangle, \tag{1}$$

其中，名称$_i$ 是规范标识符， $\mathcal{C}_i$ 是一组支持性引文上下文，$\mathcal{M}_i$ 包含增强元数据。

B. 三阶段流程

1) 阶段 *1*：可扩展的引文上下文检索：我们基于 Semantic Scholar Academic Graph (S2AG) [19]采用混合策略：使用预计算索引检索引文窗口，并按需筛选以实现查询专门化。给定 ($\mathcal{Q}$, $\mathcal{F}$)，我们检索候选论文，提取引文标记周围的句子级窗口，并保留 (论文，目标，窗口) 三元组，以及下游处理所需的最少元数据。为在处理数百万篇论文时维持吞吐量并控制延迟，我们结合了 (1) 预构建的上下文表和 (2) 轻量级的查询时连接；此外，在检索后阅读设置中使用基于大语言模型的相关性过滤器来进一步降低噪声 [13]。这种方法兼顾了准确性与计算成本（对全文进行处理的成本很高，不适用于实时推荐）。

2) 第二阶段：基于大语言模型的数据集提及抽取与引文感知质量：我们将抽取任务定义为在引文窗口中进行数据集提及抽取 [14] ，并结合引文功能线索。在科学信息检索 [6], [20], 的基础上，我们调整了 Qwen2.5-72B-Instruct [21],这一大型指令微调模型，以识别数据集提及，包括缩写或隐含的引用。输出结构化记录，包含（1）被引内容类型和（2）引文意图（遵循 SOFT [17] 对意图与内容类型的区分）。这种结构化抽取能够捕捉数据集的具体应用上下文，例如它主要是用作资源、经过修改以用于新任务，还是作为评估基准。这种语义洞察提供了元数据驱动的搜索系统根本无法实现的细粒度信息，并使研究人员能够根据资源在既有研究中的功能进行筛选。我们强制采用 JSON 模式，以确保下游一致性，并通过多层验证器（模式检查、语义检查、领域规则）评估抽取质量。我们还计算引用增强信号，例如近期性加权使用和上下文显著性，以补充模型置信度。算法 1 给出了高层流程。

算法 1 基于大语言模型的数据集提及抽取

输入： 引文窗口 $\mathcal{C}$，查询/主题提示 $\mathcal{T}$
输出： 包含置信度和证据的结构化抽取结果 $\mathcal{E}$ 证据

$\mathcal{E} \leftarrow \varnothing$
对 $c \in \mathcal{C}$ **执行**
  $y \leftarrow$ 大语言模型抽取(c, $\mathcal{T}$) {数据集、意图、内容类型}
  $z \leftarrow$验证(y) {模式/语义/领域}
  $r \leftarrow$ 相关性过滤(z, $\mathcal{T}$) {第二轮大语言模型处理}
  如果 r 有效**则**
    $\mathcal{E} \leftarrow \mathcal{E} \cup$ {记录(r.条目 r.置信度 r.证据)}
  结束条件判断
结束循环
返回 $\mathcal{E}$

3) 第三阶段：数据集实体消歧与整合：我们采用确定性且保留证据的整合

算法 2 确定性实体合并

要求： 提及项 $\mathcal{E}$（表面名称，关系，属性，来源信息）
确保： 合并后的实体 $\hat{\mathcal{E}}$

$\mathcal{E}' \leftarrow$按关系局部合并($\mathcal{E}$)
对 $m \in \mathcal{E}'$ **执行**
  m.key $\leftarrow$ 规范化(m.名称)
结束循环
$G \leftarrow$按键分组($\mathcal{E}'$)
对 $g \in G$ **执行**
  $display \leftarrow$选择显示名称(g.名称)
  $aliases \leftarrow$ 去重(g.名称)
  $evidence \leftarrow$收集来源信息(g)
  $links \leftarrow$ 优先使用持久标识符(收集链接(g))
  $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \{$实体$(display, aliases, evidence, links)\}$
结束循环
返回 $\hat{\mathcal{E}}$

流水线，而非图聚类。首先，将关联到同一书目关系的提及项在局部进行整合：选取最可靠的提及项，并合并兼容的属性（例如角色、简要描述），同时记录合并的提及项数量以供诊断。其次，我们依据严格的规范名称进行全局聚合：通过去除引号和括号标记、删除括号中的限定语和通用类型词（例如 dataset/corpus/benchmark 或 train/test/validation 标记）、合并空白字符、去除标点并转换为小写，对表层形式进行规范化。规范名称相同的提及项会被归为一组；我们保留所有来源信息，按频次选取便于阅读的显示名称，并将所有表层变体保留为别名。呈现引用时，优先使用持久标识符而非 URL。该流程总结于算法 2。

将各阶段整合起来，便形成了从查询到排序后的数据集的端到端流程。该架构在设计上具备可扩展性（预索引上下文和轻量级查询时过滤），能够保持准确性（基于大语言模型的抽取和引用感知信号），并提供可靠的输出、实体消歧结果和来源信息。

C. 评估框架与质量指标

我们依据 FAIR [1], 信息检索评估 [22],和数据质量标准 [23]，从抽取质量、证据质量和实际可访问性三个方面进行评估。我们针对基于调查得出的计算机科学领域金标准计算自动化指标，同时采用专家评估，判断其他 FOS 领域中不属于这些金标准集合的额外数据集的实用性和新颖性。

1）抽取质量：我们的评估采用多层次方法，设置三种匹配粒度，其设计借鉴了实体消歧评估实践 [24]。该方法衡量抽取效果和证据质量，遵循科学数据评估标准 [25]。

- 精确召回率：空白规范化处理后的精确字符串匹配。
- 规范化召回率：规范化处理（转为小写、移除标点、合并空白）。
- 模糊召回率：采用序列相似度聚类，阈值为 $\tau = 0.9$ [26]。

主要评估指标是规范化召回率，它在精确率与稳健性之间实现了最佳平衡，适用于实际应用，并遵循科学数据集评估的最佳实践 [25]。我们还报告

$$\textbf{FuzzyGain} = \textit{Fuzzy Recall} - \textit{Norm Recall} \tag{2}$$

，以量化对名称变体的稳健性。

2）证据质量：

- 可信来源（%）：匹配实体中，其证据可追溯至权威来源（PID 落地页，或官方目录、注册库、维护者代码库等可信主机）的比例。
- 具有数字对象标识符（%）：具有持久标识符（数字对象标识符/HANDLE/ARK）的实体比例；在表格中记为“具有数字对象标识符（%）”。

3）实际可访问性与效率：我们通过测量冗余度来评估实体合并的效果；分数越低，表示成功合并了指向同一数据集的重复提及：

$$\text{Redundancy} = \frac{|\text{Mentions}| - |\text{Entities Norm}|}{\max(1, |\text{Entities Norm}|)} \tag{3}$$

黄金标准是由专家整理的领域专属参考集，并采用标准化协议确保不同领域间的一致性；我们直接将系统的召回率与这些真实标注进行比较。

D. 可复现性

所有代码、配置文件和评估工作簿均已发布（第 IV 节）；附录 A–C 记录了影响复现的实现细节。

E. 流水线工作流可视化

图 2 展示了我们三阶段数据集发现流水线的完整工作流。系统以研究查询和字段约束为输入，依次进行引文上下文提取、结合引文增强质量评估的神经网络数据集提取以及实体消歧，最终生成排序后的数据集表。可选的大语言模型预筛选阶段可进行领域相关性评估，从而减少大规模检索场景中的噪声。

IV. 实验

A. 实验设置

我们在两种互补的设置下评估由文献驱动的数据集发现框架。第一种是在计算机科学领域开展自动化、大规模的基准测试，其中综述论文为数据集查询提供了自然的切入点。第二种是跨学科专家研究，旨在检验我们的方法在计算机科学以外的实际实用价值。综合这两种设置，我们探讨三个研究问题：（1）我们的流水线在覆盖范围和质量方面与现有方法相比如何？（2）

领域专家如何评价检索到的数据集的实用性？（3）各个组件对性能有何贡献?

1）基于综述论文的查询（计算机科学基准数据集）：为进行大规模自动化评估，我们聚焦于计算机科学领域，因为精选综述反映了活跃的研究问题。我们从自然语言处理领域（ACL、EMNLP、NAACL）和计算机视觉领域（CVPR、ICCV、TPAMI）的顶级会议中选取了八篇具有代表性的综述论文。每篇综述的标题都被用作自然语言查询，以确保评估查询代表真实且由社区定义的研究问题，而非人为构造的提示。由此得到八项基准任务，详见表 II 及其结果。我们从相应的综述论文中手动提取每项任务的金标准，作为自动化评估的真实值。

2）专家提供的查询（跨学科评估）：为验证跨领域适用性，我们进一步使用直接取自不同领域专家的研究问题进行评估。根据 FOS分类法 [27], ，我们排除计算机科学（我们自身的领域）以避免偏差，并根据其余一级类别选取研究问题：工程与技术、医学与健康科学、农业科学以及社会科学与人文学科。我们邀请领域专家（博士学位持有者、博士研究生，以及在博士导师指导下的高年级硕士生）提出一个在其学科内确有研究价值的问题。这些查询构成了我们的专家评估协议（第 IV-C 节）的基础，确保评估既能体现框架的通用性，也能反映其在现实世界跨学科发现中的实际价值。

B. 基线系统

Google Dataset Search [3]是最全面的网络规模数据集搜索引擎，由 Google 维护，并通过 schema.org 标记索引数百万个数据集。作为数据集发现的事实标准，它覆盖了广泛的领域，但高度依赖元数据的质量。

DataCite Commons [2]是研究数据 DOI 的权威注册机构，由一个由各类机构和图书馆组成的全球联盟维护。它已注册超过 3,000 万个对象，代表了正式发布数据集的金标准，但其覆盖范围偏向经过策展的资源。

基线系统针对元数据搜索，而我们的模型则可推广至数据集和应用推理。

C. 专家评估协议

自动化指标无法衡量上下文适切性或现实世界中的实用性 [7]。因此，我们基于来自非计算机科学领域、由专家提供的查询开展双盲专家研究 [22], 。

1) 参与者招募：我们邀请了十位领域专家：三位博士学位持有者、四位博士研究生，以及三位在博士生导师指导下的高年级硕士生。涉及的学科包括工程与技术（食品与饮料）、医学与健康科学（临床医学）、农业科学（农业、林业与渔业）、社会科学（教育科学）和人文（艺术）。图 3 展示了各系统在这些由专家提供的查询上的对比表现，为我们的自动化基准测试提供了定性补充。

2) 匿名化策略：为避免偏见，三个系统的结果以匿名方式呈现为“系统 A、B 和 C”。专家不了解其背后的方法。

3) 评估维度：专家从以下六个 5 分制量表对候选数据集进行评分，这些数据集包括调查中记录的数据集以及各系统发现的额外数据集：

- 相关性：与所述研究问题的一致程度
- 实用性：在研究中实际使用的可能性
- 可访问性：描述的清晰度和获取的便捷程度
- 可信度：对可靠性和来源质量的信心
- 新颖性：发现此前未知的数据集
- 总体满意度：对系统进行整体评估

4) 数据收集与分析：通过匿名的 Google 表单收集响应。我们计算每位专家–系统组合内的平均评分，然后在各位专家之间进行汇总。鉴于每个查询的评分者数量较少且分布不均衡，我们未进行正式的显著性检验。我们报告各维度的平均值以及专家之间的最佳投票偏好，这些结果以基于领域的定性证据补充了自动化基准数据集。尽管样本量限制了统计功效，但结果通过提供基于领域的数据集实用性视角，补充了自动化基准数据集。

D. 结果与分析

我们展示了全面的实验结果，将我们的神经数据集发现方法与两个成熟的基线系统进行比较：Google Dataset Search 和 DataCite Commons。我们的评估涵盖八个多样化的研究领域，分析了总计 1,358 个数据集提及。

1) 总体自动化性能：表 I 总结了针对计算机科学基准数据集的汇总自动化评估结果。我们的引文上下文方法提取出 1,330 个去重的数据集实体，并实现了显著高于 Google Dataset Search（2.70%）和 DataCite（0.00%）的平均归一化召回率（47.47%），同时保持较高的证据质量和较低的冗余。表 II 报告了各查询的结果，在单个基于调查的任务中，我们系统的召回率最高达到 81.82%。

2) 专家评估结果：专家从六个 5 点李克特量表维度对数据集进行了评分：相关性、实用性、可访问性、可信度、新颖性和总体满意度。图 3 展示了各研究问题的评分，表 III 则给出了汇总结果。附录中的表 IV 报告了各研究查询的均值 ± 标准差评分。专家一致认可我们的方法在相关性 和实用性方面的表现，证实基于上下文的检索能够识别出真正对研究有用的数据集。在

表 I：所有计算机科学任务的基于统计的性能评估。 ↑ 表示数值越高，性能越好； ↓表示数值越低，性能越好。

方法	总计实体	FuzzyGain (%) ↑	可信来源 (%) ↑	含数字对象标识符 (%) ↑	冗余 ↓	证据质量
Ours	1,330	**1.71**	**14.23**	68.52	**0.26**	**High**
Google Dataset Search	79	0.00	1.63	32.81	0.29	中等
DataCite Commons	67	0.00	0.00	**87.50**	7.76	**High**

表 II：对于八个计算机科学查询中的每一个，我们报告金标准数据集（根据综述论文构建）的数量、匹配数据集的数量以及相应的宏平均召回率（%）。

研究问题	Gold	已匹配			召回率 (%)		
		Ours	Google	DataCite	Ours	Google	DataCite
多模态知识图谱推理 [28]	11	9	0	0	81.82	0.00	0.00
一体化图像复原 [29]	30	10	0	0	33.33	0.00	0.00
大语言模型的规划能力 [30]	38	21	0	0	55.26	0.00	0.00
基于事件的立体深度估计 [31]	17	9	0	0	52.94	0.00	0.00
自然语言处理中的专利分类 [32]	7	3	0	0	42.86	0.00	0.00
文档级事件抽取 [33]	23	9	3	0	39.13	13.04	0.00
历史文档文本行分割 [34]	43	13	1	0	30.23	2.33	0.00
个性化文本生成 [35]	16	1	1	0	6.25	6.25	0.00
平均召回率					**47.47**	2.70	0.00

表 III：专家平均评分汇总（1–5）。

维度	Ours	Google	DataCite
相关性	**4.33**	3.07	2.60
实用性	**4.09**	2.46	2.66
可访问性	**3.80**	2.87	3.17
可信度	**4.13**	2.49	2.84
新颖性	**3.64**	2.92	3.50
总体	**4.07**	2.60	2.48

可信度 和 总体满意度方面也取得了一致的提升。在新颖性方面，专家表示，我们的方法发现的额外（非调查）数据集中，有相当一部分对他们而言是新的，并且仍被认为在其领域内具有实用价值。

在所有由专家评分的候选项中，我们的文献驱动方法生成了 105 个数据集，这些数据集均具有完整的 实用性 和 新颖性 评分。其中，45 个（42.9%）在这两个维度上的评分均至少为 4；相比之下，Google Dataset Search 的 31 个数据集中有 4 个（12.9%），DataCite Commons 的 6 个数据集中有 2 个（33.3%）。这些数据以定量方式表明，我们的方法发现的额外数据集中有相当一部分被领域专家认为具有较高的实用性，其中也有不少数据集被认为对特定研究主题具有新颖性。

3) 质量与可信度分析： 除原始召回率指标外，我们的评估还揭示了不同系统提取的数据集在质量与可信度方面存在显著差异。

a) 证据质量： 我们的系统实现了 14.23% 的可信证据支持的召回率，显著优于 Google Dataset Search（ 1.63%）和 DataCite（0.00%）。这表明，我们的神经抽取方法不仅发现了更多数据集，还能识别出学术来源信息可验证、质量更高的资源。

b) 持久标识符覆盖率： 由于专注于带有数字对象标识符的正式发布的数据集，DataCite 的 PID 比率最高（87.50%）。然而，这也导致其召回率极低（0.00%），表明正式发表要求严重限制了数据集发现范围。我们的系统在保持较高 PID 覆盖率（68.52%）的同时，实现了显著更高的召回率，既能识别正式发布的数据集，也能识别新兴数据集。

c) 冗余与效率： 我们的数据集实体消解流水线实现了最低的冗余率（0.26），体现出有效的重复检测与整合能力。DataCite 的冗余情况令人担忧（7.76），这表明可能存在索引不一致；而 Google 的效率则较为合理（0.29）。

4) 系统可扩展性与覆盖率：我们的方法展现出更强的可扩展性，提取出 1,330 个去重数据集实体，而 Google 和 DataCite 分别仅提取出 79 个和 67 个。这一覆盖率提升了 16.8×，反映出我们的系统能够识别嵌入引文上下文中的数据集，而非仅依赖显式元数据标记或正式注册。

a) 领域适应性：不同领域的表现差异显著，在计算机科学技术领域的召回率最高（多模态知识图谱推理: 81.82%），在跨学科领域的表现则较低（个性化文本生成：6.25%）。这反映了研究数据实践的根本差异：计算机科学领域已形成成熟的数据集共享文化，因此挖掘引文语境十分有效；而跨学科领域更注重新颖的数据收集，而非重复利用。这验证了我们的双重评估方法，因为高召回率未必能反映所有领域中的实际研究价值。

b) FuzzyGain 鲁棒性：我们的系统表现出正向 FuzzyGain（1.71%），表明其能够有效处理数据集

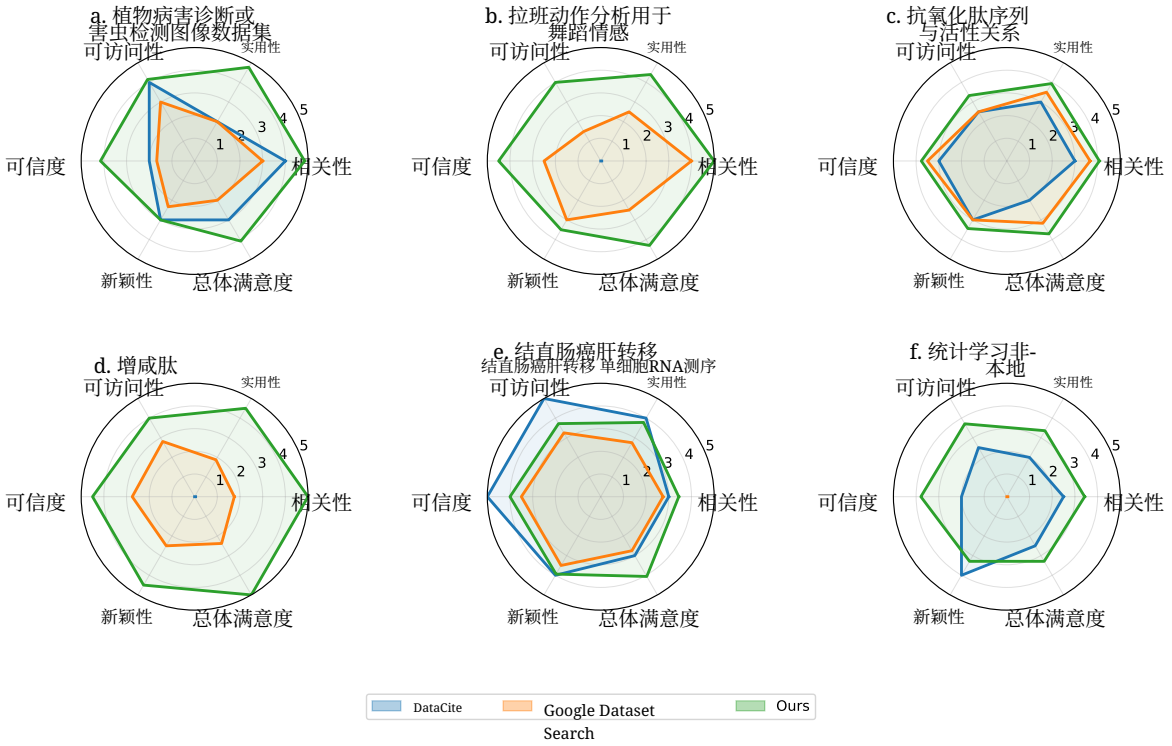


图 3：六个跨学科研究问题的专家评估评分。每个雷达图展示单个查询的用户平均评分，对比我们的系统与基线系统在六个质量维度上的表现。详细数值结果见表 IV。

名称变化带来的问题，并通过模糊匹配加以应对。基线系统的 FuzzyGain 均为零，说明它们对科学文献中常见的命名不一致缺乏鲁棒性。

5) 总结： 在自动评估和专家评估中，以引文上下文为基础的数据集发现所获得的资源，相较于元数据驱动的基线方法，都被认为具有更高的相关性、实用性和可信度。这些发现验证了我们的核心假设：基于文献的语义桥梁能够连接研究问题与数据集，从而在不同领域实现更有效且更具泛化能力的数据集发现，扩展可操作的数据集空间，而不仅仅是增加结果数量。基线系统的糟糕表现证实，直接进行元数据映射不足以捕捉复杂研究问题的意图，因此有必要采用本文提出的文献驱动方法。

E. 案例研究

1) 数据集族解析： 针对研究问题：文档级事件抽取。我们的系统识别出三个 ACE 变体（ACE、ACE 2005、ACE 2005 (zh)），它们指向同一目录（LDC2006T06），但在范围和语言上存在差异。传统方法要么过度合并这些变体（丢失语义区别），要么使用论文 DOI 作为键（导致实体错位）。

我们的解决方案将数据集族级标识符（目录 ID、持久标识符）用作主键，同时将变体元数据保留为属性。这使得能够在概念层面准确计算召回率，同时保留细粒度信息，避免召回率虚高，并确保比较结果可复现。

在概念层面，同时保留细粒度信息，避免召回率虚高，并确保比较结果可复现。

五、结论与讨论

本研究提出了一种文献驱动的数据集发现框架，将科学出版物视为连接研究问题与数据资源的语义桥梁。该三阶段流程由引文语境提取、数据集识别和数据集级实体消歧组成，结果表明，相较于仅依赖元数据，以学术使用为基础开展发现能够产生引文验证的数据集；这些数据集不仅具有更高的相关性和可信度，还拥有更丰富的语义语境。这些数据集直接从其文本与引文环境中提取，传达了有关其用途、方法及领域相关性的信号，而这些意义维度通常在基于元数据的知识库中缺失。

该方法的一项突出优势在于，每个发现的数据集都已隐含地经过同行验证。每个数据集都曾出现在已发表的研究中，并在学术工作流程中得到描述、评估和应用，体现了其可访问性、可用性和可靠性。这一特性使发现过程天然具备质量意识，从而减轻人工审查的负担。此外，通过保留数据集提及项的语言和概念上下文，我们的框架揭示了采用数据集背后的理由，让研究人员能够更有效、更全面地理解数据格局。

实证结果表明，该框架在多个不同领域中始终优于依赖元数据的系统。这些发现确立了引文语境挖掘作为一种可扩展且可泛化的发现范式，能够找回已广泛使用的数据集，以及此前未记录但颇具价值的数据集。通过将静态书目参考文献转化为研究问题、数据资源及其已记录应用之间可操作的语义链接，该框架推动了科学搜索进入一种新的上下文感知模式。

a) 局限性与挑战： 尽管该框架实现了稳健且有意义的发现，其设计自然也带来了一些权衡。(1) 依赖引文丰富的文献，会优先发现那些已通过先前发表和复用而获得关注的数据集。这确保所识别的数据集可信，并经过研究界验证，但也可能限制对尚未被引用的新发布或小众数据集的覆盖。同样，出版周期固有的时间滞后，会使新兴数据集直到出现在后续作品中才被纳入。(2) 依赖 S2AG 等大规模开放获取语料库，会造成结构性偏差，倾向于开放数据实践成熟的学科，同时低估以付费墙论文为主的领域。这些并非方法本身的缺陷，而是更广泛的数据生态系统的体现。未来整合多源语料库和自适应检索策略，有助于逐步缓解这些不平衡。

b) 伦理考量： 依赖英语文献和基于引文的索引，可能会无意中低估来自非英语研究群体的数据集，或来自数据引用规范化程度较低学科的数据集。这些偏差并非源于方法本身，而是可获取学术记录的构成所致。要解决这一问题，就需要扩大对多语言和区域语料库的覆盖范围，并调整抽取模型，使其能够识别各学科不同的数据集致谢惯例。此类扩展将推动发现过程更加均衡，并更具全球代表性。

更广泛的影响与未来方向

这项工作的更广泛意义在于，将引文语境挖掘确立为一种可推广的范式，用于基于文献的资源发现。尽管我们目前聚焦于数据集，但其底层原理自然也适用于其他学术实体。例如，同一方法可用于描绘计算软件在不同学科间的传播、追踪实验方案的演变，以及考察理论框架如何在实证研究中得到操作化。通过收录这些多样化的成果，未来的工作可以构建一个全面的学术知识图谱，将数据、方法和工具联系起来，从而加深对该领域的理解。将这样的图谱与大语言模型相结合，将能够以自然语言查询复杂的研究关系，进而推动该框架从一种特定的发现工具发展为统一的学术知识互联模型，为透明且可复现的科学生态系统提供支持。

参考文献

[1] M. D. Wilkinson 等，“科学数据管理与 stewardship 的 FAIR 指导原则”，Scientific Data，第3卷，第1期，2016年。[2] J. Brase，“DataCite——研究数据的全球注册机构”，SSRN Electronic Journal，2010年。[3] D. Brickley、M. Burgess 和 N. Noy，“Google Dataset Search：在开放网络生态系统中构建数据集搜索引擎”，载于万维网大会，2019年，第1365–1375页。[4] A. Zimmerman，“仅凭元数据还不够：利用多种知识形式定位可再利用的数据”，International Journal on Digital Libraries，第7卷，第1–2期，第5–16页，2007年5月。[5] Y. Luan、L. He、M. Ostendorf 和 H. Hajishirzi，“用于科学知识图谱构建的实体、关系和共指多任务识别”，载于《2018年自然语言处理实证方法会议论文集》。计算语言学协会，2018年。[6] S. Jain、M. van Zuylen、H. Hajishirzi 和 I. Beltagy，“SciREX：用于文档级信息抽取的挑战数据集”，载于《第58届计算语言学协会年会论文集》。计算语言学协会，2020年。[7] C. L. Borgman，《大数据、小数据、无数据：网络世界中的学术研究》。麻省理工学院出版社，2015年1月。[8] I. V. Pasquetto、B. M. Randles 和 C. L. Borgman，“科学数据的再利用”，Data Science Journal，第16卷，第8页，2017年3月。[9] X. Chen、S. Jia 和 Y. Xiang，“综述：知识图谱上的知识推理”，Expert Systems with Applications，第141卷，第112948页，2020年3月。[10] H. Pan、Q. Zhang、E. Dragut、C. Caragea 和 L. J. Latecki，“DMDD：用于数据集提及检测的大规模数据集”，计算语言学协会会刊，第11卷，第1132–1146页，2023年。[11] J. Heddes、P. Meerdink、M. Pieters 和 M. Marx，“科学文章中数据集名称的自动检测”，Data，第6卷，第8期，第84页，2021年。[12] A. Xu、R. Ding 和 L. Wang，“ChatPD：由大语言模型驱动的论文—数据集关联系统”，载于《第31届ACM SIGKDD知识发现与数据挖掘大会论文集》第2卷，KDD ’25系列。美国纽约州纽约市：计算机协会，2025年。[13] P. Marini、A. Santos、N. Contaxis 和 J. Freire，“Data Gatherer：利用大语言模型从科学文献中抽取数据集参考文献”，载于《第五届学术文档处理研讨会（SDP 2025）论文集》。计算语言学协会，2025年。[14] P. Datta、S. Datta 和 D. Roy，“直面文献：利用大语言模型抽取数据集提及项”，载于《第24届ACM/IEEE数字图书馆联合会议论文集》，JCDL ’24系列。美国纽约州纽约市：计算机协会，2025年。[15] H. Zhao、Z. Luo、C. Feng、A. Zheng 和 X. Liu，“用于建模科学文献中在线资源引文作用与功能的基于上下文的框架”，载于《2019年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集》。计算语言学协会，2019年。[16] M. F¨arber、A. Albers 和 F. Sch¨uber，“识别科学出版物中使用的方法和数据集”，2021年。[17] C. Duan 和 Z. Tan，“用于引文分类的语义正交框架：分离意图与内容”，载于《连接数字图书馆理论与实践》，W.-T. Balke、K. Golub、Y. Manolopoulos、K. Stefanidis 和 Z. Zhang 编。Cham：Springer Nature Switzerland，2026年，第183–206页。[18] A. Cohan、S. Feldman、I. Beltagy、D. Downey 和 D. Weld，“SPECTER：利用引文信息转换器进行文档级表示学习”，载于《第58届计算语言学协会年会论文集》。计算语言学协会，2020年，第2270–2282页。[19] A. I. for Artificial Intelligence，“Semantic Scholar开放数据平台”，2025年。[20] Q. Zhang、Z. Chen、H. Pan、C. Caragea、L. J. Latecki 和 E. Dragut，“SciER：用于科学文档中数据集、方法和任务的实体与关系抽取数据集”，载于《2024年自然语言处理实证方法会议论文集》。计算语言学协会，2024年，第13083–13100页。

[21] Qwen，“Qwen2.5技术报告”，2025年。[22] C. D. Manning、P. Raghavan 和 H. Sch¨utze，《信息检索导论》。剑桥大学出版社，2008年。[23] C. Batini、C. Cappiello、C. Francalanci 和 A. Maurino，“数据质量评估与改进的方法论”，*ACM*计算调查，第41卷，第3期，第1–52页，2009年7月。[24] P. Christen，《数据匹配：记录链接、实体消歧与重复检测的概念和技术》。施普林格·柏林·海德堡，2012年。[25] L. Koesten、E. Simperl、T. Blount、E. Kacprzak 和 J. Tennison，“关于数据集的一切：数据摘要研究”，《国际人机研究杂志》，第135卷，第102367页，2020年3月。[26] W. W. Cohen、P. Ravikumar 和 S. E. Fienberg，“名称匹配任务中字符串距离度量的比较”，载于《*2003*年国际万维网信息集成会议论文集》，IIWEB’03系列。AAAI出版社，2003年，第73–78页。[27] A. Sinha、Z. Shen、Y. Song、H. Ma、D. Eide、B.-J. P. Hsu 和 K. Wang，“微软学术服务（MAS）及其应用概述”，载于《第*24*届国际万维网大会论文集》，WWW ’15系列。ACM，2015年5月，第243–246页。[28] K. Liang、L. Meng、M. Liu、Y. Liu、W. Tu、S. Wang、S. Zhou、X. Liu、F. Sun 和 K. He，“图谱类型上的知识图谱推理综述：静态、动态与多模态”，*IEEE*模式分析与机器智能汇刊，第46卷，第12期，第9456–9478页，2024年12月。[29] J. Jiang、Z. Zuo、G. Wu、K. Jiang 和 X. Liu，“全能图像复原综述：分类体系、评估与未来趋势”，*IEEE*模式分析与机器智能汇刊，第47卷，第12期，第11892–11911页，2025年12月。[30] H. Wei、Z. Zhang、S. He、T. Xia、S. Pan 和 F. Liu，“PlanGenLLMs：大语言模型规划能力的现代综述”，载于《第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文）》。计算语言学协会，2025年，第19497–19521页。[31] S. Ghosh 和 G. Gallego，“基于事件的立体深度估计：综述”，*IEEE*模式分析与机器智能汇刊，第47卷，第10期，第9130–9149页，2025年10月。[32] H. H. Shomee、Z. Wang、S. N. Ravi 和 S. Medya，“专利分析综述：从自然语言处理到多模态人工智能”，载于《第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文）》。计算语言学协会，2025年。[33] H. Zheng、S. Wang 和 L. Huang，“文档级信息抽取综合综述”，载于《事件检测未来研讨会（*FuturED*）论文集》。计算语言学协会，2024年，第58–72页。[34] I. Rabaev 和 M. Litvak，“历史文档图像中的文本行分割与基线检测最新进展：系统综述”，《国际文档分析与识别杂志》，2025年5月。[35] Y. Xu、J. Zhang、A. Salemi、X. Hu、W. Wang、F. Feng、H. Zamani、X. He 和 T.-S. Chua，“大模型时代的个性化生成：综述”，载于《第63届计算语言学协会年会论文集（第1卷：长篇论文）》 style>。计算语言学协会，2025年，第24607–24649页。

附录

详细流水线方法

本附录详细介绍了我们文献驱动的数据集发现框架的端到端操作流程。该流程分为三个主要阶段。

A. 阶段 *1*：可扩展的语料库预处理与索引

此初始阶段构建一个聚焦于查询相关内容的科学文本语料库，以供分析。

a) 种子论文检索：给定一个自然语言研究问题，我们首先查询 Semantic ScholarAPI，以检索一组前 K 名相关种子论文（通常为 $200 \leq K \leq 400$）。这些论文构成了分析语料库中初始且高度相关的核心。

b) 引文图扩展与上下文收集：使用我们预先建立索引的 S2AG 语料库本地快照。我们收集所有引文上下文，其中种子论文要么引用了其他作品，要么被其他作品引用。这一扩展使相关文本池超越了初始的关键词匹配范围。

c) 离线索引：为大规模执行此阶段，我们的流水线基于本地 S2AG 快照运行。原始数据被转换为 Parquet 列式格式，并使用嵌入式 DuckDB 实例建立索引；该实例配置在一台高内存服务器上（128 GB RAM、32 线程并行处理、1TB SSD 存储）。我们为每条引文链接预先物化文本上下文，从而实现近乎即时的检索。此架构将查询延迟中位数从纯 API 方案下的数十秒缩短至约 0.1 秒，这对于交互式分析至关重要。

B. 第二阶段：基于大语言模型的抽取与相关性筛选

此阶段采用大语言模型执行多步骤流程，以识别数据集提及并评估其与初始研究问题的相关性。

提示词 1：数据集提及抽取

［系统］ 你是一名专业的科学标注员。从给定的引文上下文中抽取数据集

研究问题（研究问题）：“{RQ_TEXT}” 施引论

任务：1) 识别明确或隐含提及的数据集/基准数据集/语料库名称。2) 对每个数

"datasets": [{ "name": "...", "usage_role": "使用 |修改 |评估对象", "

a) 数据集提及抽取：我们使用大语言模型（Qwen2.5-72B-Instruct）处理每个收集到的引文上下文，以识别其中提及的数据集、基准数据集或语料库。根据 SOFT 分类框架，提示词会指示模型提取数据集的名称及其在上下文中所述的具体用途。

b) 数据集相关性筛选：随后会再次调用大语言模型，筛选提取出的数据集，判断其与原始研究问题的相关性。模型会获得研究问题、引文上下文，以及引用论文和被引论文的标题。模型根据数据集的用途与研究目标之间的语义对齐情况，做出二元判断（纳入/排除）。此步骤会筛除那些虽然有效、但与用户查询无关的数据集。

提示 2：相关性筛选

```
[系统]
仅根据提供的上下文和标题，判断候选数据集是否与
      研究问题相关，
并作出回答。
      且仅包含有效的 JSON。

[用户]
研究问题（研究问题）：
"{RQ_TEXT}"

候选数据集：
"name": "{CANDIDATE_DATASET_NAME}"

上下文：
[开始上下文]
{引文_上下文_文本}
[结束上下文]

施引论文标题："{施引_标题}"
引用论文摘要："{施引_摘要}"
被引论文标题："{CITED_TITLE}"
被引论文摘要："{CITED_ABSTRACT}"

判定规则：
- 如果上下文表明该数据集已被使用，则判为相关                    ，
      修改过        或经过评估
  旨在解决研究问题（或与之直接相关的目标）。
- 如果用途无关、纯属背景信息，
      或属于不同领域，则不相关。
- 应采取保守判断；要求上下文中有明确的
      证据。
仅输出 JSON：
{
  "is_相关": true/false,
  "置信度": 0.0-1.0,
  "reasoning": "简短且基于证据的说明"
}
```

c) 提示与基础设施：所有大语言模型任务均在一台由 vLLM 驱动的推理服务器上运行，该服务器配备四块 NVIDIA A100 80GB GPU。我们的提示工程策略将大语言模型设定为专家标注员，并提供严格的纳入／排除规则（例如，提取 MIMIC-III，忽略 BERT），同时要求所有输出均以所提供的文本为依据，以防止产生幻觉。所有输出均采用基于 SOFT 的结构化 JSON 模式，并将模型温度设为 0.0，以确保结果具有确定性。

C. 第三阶段：实体消歧与元数据增强

此阶段对原始筛选后的数据集提及进行清理和整合，生成规范列表。

a) 规范化与整合：我们采用确定性的规则流程，将不同的表层形式（例如 ACE 2005 (zh)、ACE-2005）统一为单一的规范实体。该流程先进行严格的词汇规范化（转换为小写、移除标点等），再将规范化形式相同的提及项归为一组。在每组中，根据出现频率选取易于阅读的显示名称，并将所有原始表层形式保留为别名。所有来源信息均予以保留，以确保整个过程可审计。

b) 多来源 *URL* 与 *PID* 检索：我们采用三级策略，为每个数据集查找 URL 或 PID。首先，检查大语言模型是否已从引文上下文中直接提取 URL。若没有，则查看 S2AG 中被引论文的元数据；如果它是资源论文，我们便沿用其数字对象标识符。最后，我们使用规范的数据集名称，以编程方式查询外部来源（Google Gemini 的搜索功能），查找其官方网站或代码库。

c) 最终排序与结果生成：整条流水线的结果是一个经过清理、去重并排序的列表，其中包含与原始研究问题相关的数据集。最终排名由引文数量决定：在我们检索到的语料库中，被发现用于特定数据集的去重论文数量。这种基于使用情况的频次可作为该数据集在研究问题背景下的普及程度和重要性的替代指标。最终列表中的每个条目都包含数据集的规范名称、使用上下文、引文数量，以及可验证的 URL 或 PID。

逐查询专家评分

表 IV：汇总所有跨学科研究问题的专家评估评分。每个区块显示某个查询的用户评分均值（1–5） ±标准差。

系统	相关性	实用性	可访问性	可信	新颖性	总体
农业科学：植物病害诊断或害虫检测图像数据集						
Ours	4.38 ± 0.52	4.00 ± 0.63	3.83 ± 0.75	4.17 ± 0.75	3.67 ± 0.82	4.00 ± 0.63
Google	2.88 ± 0.64	2.56 ± 0.73	2.88 ± 0.64	2.56 ± 0.73	2.88 ± 0.64	2.56 ± 0.73
DataCite	2.63 ± 0.74	2.69 ± 0.70	3.19 ± 0.54	2.81 ± 0.66	3.50 ± 0.60	2.56 ± 0.66
人文（艺术）：用于舞蹈情感的拉班动作分析						
我们的方法	4.67 ± 0.52	4.17 ± 0.75	3.67 ± 0.82	4.17 ± 0.75	3.50 ± 0.84	4.00 ± 0.89
Google	3.67 ± 0.82	2.67 ± 0.82	2.83 ± 0.75	2.50 ± 0.84	3.00 ± 0.89	2.50 ± 0.84
DataCite	2.50 ± 0.84	2.67 ± 0.52	3.17 ± 0.41	2.67 ± 0.52	3.50 ± 0.55	2.50 ± 0.55
工程与技术（食品）：抗氧化肽序列与活性关系						
Ours	4.75 ± 0.35	4.43 ± 0.53	3.94 ± 0.70	4.31 ± 0.60	3.63 ± 0.80	4.38 ± 0.56
Google	2.88 ± 0.64	2.63 ± 0.74	2.88 ± 0.64	2.69 ± 0.70	2.88 ± 0.64	2.69 ± 0.70
DataCite	2.44 ± 0.73	2.69 ± 0.70	3.13 ± 0.64	2.75 ± 0.71	3.50 ± 0.76	2.69 ± 0.74
工程与技术（食品）：增咸肽						
Ours	4.89 ± 0.20	4.78 ± 0.33	4.40 ± 0.52	4.67 ± 0.33	3.90 ± 0.57	4.78 ± 0.33
Google	2.10 ± 0.57	2.20 ± 0.42	2.50 ± 0.53	2.30 ± 0.48	2.50 ± 0.53	2.20 ± 0.42
DataCite	2.00 ± 0.47	2.30 ± 0.48	2.40 ± 0.52	2.50 ± 0.53	4.10 ± 0.32	2.30 ± 0.48
医学与健康科学（临床）：结直肠癌肝转移 结直肠癌肝转移 单细胞RNA测序						
Ours	4.33 ± 0.58	4.00 ± 0.63	3.83 ± 0.75	4.17 ± 0.75	3.67 ± 0.82	4.00 ± 0.63
Google	3.00 ± 0.63	2.67 ± 0.52	2.83 ± 0.75	2.50 ± 0.55	3.00 ± 0.63	2.50 ± 0.55
DataCite	2.67 ± 0.52	2.67 ± 0.52	3.17 ± 0.41	2.83 ± 0.41	3.50 ± 0.55	2.50 ± 0.55
社会科学（教育）：统计学习 非母语						
Ours	4.25 ± 0.96	3.75 ± 0.96	3.75 ± 0.96	4.00 ± 0.82	3.75 ± 0.96	3.75 ± 0.96
Google	3.00 ± 0.82	2.75 ± 0.96	2.75 ± 0.96	2.75 ± 0.96	3.00 ± 0.82	2.75 ± 0.96
DataCite	2.75 ± 0.96	2.75 ± 0.96	3.25 ± 0.96	3.00 ± 0.82	3.50 ± 0.58	2.75 ± 0.96