arXiv:2603.07050v1 [cs.IR] 7 Mar 2026

利用大语言模型自动化、规模化构建开放科学数据库

Nikita Gautam[1], Doina Caragea[1], Ignacio Ciampitti[2], 和 Federico Gomez[2]

[1] 计算机科学系，堪萨斯州立大学，堪萨斯州曼哈顿 66502，美国[2] 农学系，普渡大学，印第安纳州西拉斐特 47907，美国

摘要。 随着在线科学文献呈指数级增长，识别可靠的特定领域数据变得越来越重要，但也极具挑战性。针对特定领域科学文献进行手动数据收集和筛选不仅耗时耗力，还容易产生错误和不一致。为促进自动化数据收集，本文介绍了一种基于网络的工具，该工具利用大语言模型（LLM）自动化、规模化地构建开放科学数据库。更具体地说，该工具基于自动化统一框架，将基于关键词的查询、基于 API 的数据检索以及由大语言模型驱动的文本分类相结合，以构建特定领域的科学数据库。该工具采用并行查询技术，从多个可靠的数据源和搜索引擎收集数据，构建组合式统一数据集。随后，使用针对每个基于关键词的查询定制的提示词调用大语言模型，对数据集进行筛选，从中提取与目标科学查询相关的数据。研究针对一组面向不同特定领域任务的可变基于关键词的搜索进行了测试，这些任务与农业和作物产量相关。结果与分析显示，该方法与小型领域专家整理的数据库有 90% 的重合度，表明所提出的工具能够显著减少手动工作量。此外，所提出的框架具有良好的可扩展性且与领域无关，可应用于不同领域，以构建规模化的开放科学数据库。

关键词： 开放科学数据库 · 数据收集与筛选 · 文献挖掘 · 文本分类 · 大语言模型

1 引言

在现代研究中，获取精准且有条理的科学信息仍是一项重大挑战。尽管每年都有新知识以越来越快的速度产生，研究人员往往仍难以从大量分散且通常并非完全开放获取的来源中检索和汇集信息

以回答具体的科学问题。科学信息的碎片化特性，使得针对特定研究需求构建数据库需要耗费大量人力。研究人员必须手动从多个来源搜索、提取并整理信息，这不仅耗时，也容易出现错误和不一致。农业领域的既往研究指出，从现有文献中检索和汇集数据一直困难重重，表明各个科学领域在数据可获取性和标准化方面存在不足 [10,4,11]。例如，应用植物科学和农业领域的系统综述与荟萃分析常发现信息报告不一致或不完整，给综合得出可靠结论带来困难 [10,4,11]。文献分散，以及土壤有机碳（SOC） [13,1]和养分管理等领域覆盖有限等持续存在的挑战，凸显了对可扩展工具的需求，以协助汇集数据。更具体地说，我们迫切需要实用的自动化工具，从分散的来源中检索信息，并组建一致、全面且针对特定查询的数据集，从而解答有针对性的科学问题。

为满足这一需求，本文介绍了一种自动化网络工具，以实现可扩展的文献检索、分类和整合。该工具将基于关键词的查询与大语言模型（LLM）支持的零样本文本分类相结合，从多个分布式来源高效识别并筛选相关出版物。通过自动化以往通常需要大量人工工作量的任务，例如筛选相关文章、合并来自多个来源的数据以及删除无关的重复记录，该工具大幅降低了构建一致且针对特定查询的科学数据集所需的工作量。这种方法有助于研究人员高效构建自定义数据库，以回答不同领域中有针对性的科学问题。尽管该方法具有普适性，我们仍以农业相关任务进行评估，以突出其效用和领域挑战。

该工具整合了 Scopus [8],、Web of Science [5], 、Science Direct [7], 和 Google Scholar [12]等主要科学数据库。它使用与特定研究目标相关的预设关键词，指导在这些来源中进行检索。随后合并检索到的记录，并通过匹配数字对象标识符（DOI）识别和删除重复记录。此外，当数字对象标识符缺失或不一致时，还会将文章标题作为额外的核查依据。仅保留英文出版物以供后续处理。

该工具利用大语言模型自动进行相关性分类，减少人工筛选所带来的偏见和工作量。较小的大语言模型，例如 BERT [6]模型，可用于针对特定领域的文章标题进行训练，从而将出版物分类为相关或不相关。尽管基于 BERT 的分类器在其训练所针对的特定主题上效果良好，但要处理新主题，就需要重新训练，这限制了其普遍适用性。此外，如果无法获取与某些关键词相关的数据，就完全无法训练此类模型。为克服这些局限，我们使用了更灵活的大语言模型，例如 ChatGPT [2], LLaMA[19],和 Gemini [16] ，通过提示进行零样本分类，使该工具能够适应多样化的主题，无需重新训练——

从而提升其在各科学领域的可扩展性。

由此产生的工具可用于编制结构化、针对特定查询的数据库。每个生成的数据库都包含论文标题、作者、出版年份、DOI、URL 等关键信息。为展示其应用，我们介绍了一项案例研究，重点关注一个开放数据库，该数据库涉及塞内加尔主要作物对养分施用的产量响应。通过系统地识别、收集和整理相关出版物，该工具支持生成产量响应曲线，并评估养分对作物产量影响的预测因子。初步测试表明，该工具可显著减少人工工作量。该框架具有可扩展性且领域无关，因此适用于构建不同研究领域的开放科学数据库。总而言之，本研究的主要贡献如下：

– 我们设计了一套流程，以最少监督的方式收集和筛选大规模科学数据。– 我们提出了一种抽象筛选工具，集成基于大语言模型的分类功能，可自动去除噪声数据。– 我们对所提出的方法进行了全面评估，以量化其性能。

2 背景

我们回顾了与所提出的方法和评估相关的两个主要研究方向，具体而言，是农业领域构建特定主题数据库所面临的挑战，以及利用语言模型帮助构建此类数据库。

2.1 汇编农业 Scientific Data 的挑战

农业领域的近期研究凸显了土壤有机碳（SOC）在气候与农业研究中的重要性日益提升 [11]。全球和区域层面的土壤有机碳数据大型汇编已陆续建立 [13,1]。然而，这些努力仍显示，可获取且全面的土壤有机碳数据集存在显著缺口，尤其是在更精细的区域尺度上。类似地，近期综述通过将作物对养分施用的产量响应数据汇编到开放获取数据库中，指出了制定特定地点养分管理策略的必要性 [10]。此类汇编旨在提高主要大田作物的数据可用性，并支持农业管理中的分析与决策。尽管这些针对特定领域的举措展现了开放获取数据库推动研究发展的潜力 [3]，但也凸显了创建和维护全面、及时更新的数据集所面临的持续挑战。在许多领域，建立稳健的开放数据基础设施仍进展有限 [9]，而数据所有权、隐私和利益相关者利益等实际障碍，也使协作式数据共享变得复杂 [22]。尽管某些主题已有结构化数据库 [17,21,15]，但这些数据库往往涵盖范围狭窄、需要大量专家投入，而且难以扩展或适应新的研究问题。在

此外，当前许多方法缺乏自动化和标准化，导致从多种文献来源收集和合并数据、消除重复记录并确保一致性都十分耗时。这一局限阻碍了开放科学数据库的广泛采用，并拖慢了跨学科研究的发展。

为克服手动数据收集中的这些挑战，并简化特定查询数据集的构建，近期自然语言处理（NLP）和大语言模型的发展为自动化文献筛选与整合提供了很有前景的解决方案。基于这些进展，本文提出了一种自动化、可扩展的工具，将基于关键词的搜索、大语言模型执行的相关性分类以及标准化数据整合结合起来。该方法旨在减少研究人员从分散的文献来源构建定制科学数据集所需的人工工作量，使其能够在各种科学领域生成最新的任务专属数据库。

2.2 用于自动化文献筛选的自然语言处理与大语言模型

自然语言处理技术使我们能够以结构化且有意义的方式处理、交互和理解自然语言。在自然语言处理的各种应用中，文本筛选等分类任务对于整理和评估大量科学文献尤为重要。这些技术可根据目标关键词，帮助识别并提取研究论文中的相关内容。自然语言处理领域的一项重要进展，是开发出了基于 Transformer 的语言模型，例如双向 Transformer 编码器表征（BERT） [6], ，这类模型在包括文本分类在内的多种自然语言处理任务中展现出强大的性能。在 Transformer 的基础上，大语言模型的最新进展进一步提升了基于自然语言处理的分类能力。OpenAI 的 ChatGPT [2]、Meta 的 LLaMA[19] 、Google 的 Gemma 2 [18], 、 Microsoft 的 Phi 2 [14], 等模型，在各种特定领域任务上均显著提升了性能，且无需进行任务专属微调。由于大语言模型是在大量且多样的文本语料库上预训练的，因此能够捕捉复杂的上下文关系，并以极少的额外训练执行文献筛选等复杂任务。本研究评估大语言模型在文献筛选过程中用于文本分类的适用性，以支持开放获取科学数据库的自动化构建，并评估其在这一场景中的有效性。

3 方法

本节概述了一套模块化且可扩展的流程，用于从多个网络资源构建科学数据集。该流程包含三个主要阶段：数据收集、数据筛选和相关性分类。整体架构如图 1 所示。为支持可复现性并促进更广泛的使用，我们已在 GithHub 上发布所提出工具的完整流程。

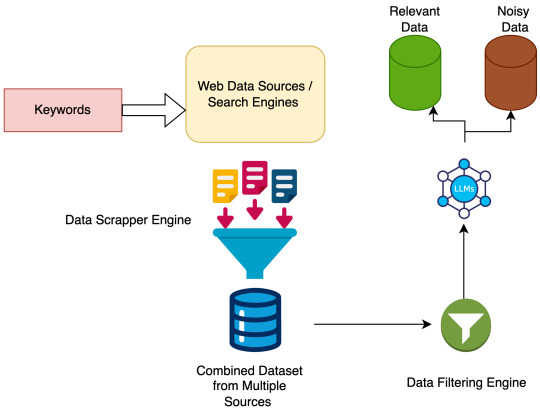

图1. 该图展示了如何通过网络文献来源端到端地构建领域特定数据集。流程从一组相关领域的关键词开始，并以此查询网络数据源或搜索引擎。数据抓取引擎收集检索到的文档，将摘要和元数据汇总为来自多个来源的综合数据集。随后，原始数据集会经过数据过滤引擎处理，该引擎利用大语言模型（LLM）将内容分类为相关数据和噪声数据。此流程可自动化并扩展筛选过程，从而高效构建干净且针对特定任务的数据集。

3.1 数据收集

数据收集过程始于一个自动化数据收集系统，该系统旨在针对各种领域特定查询开展大规模文献收集。该系统可查询多个学术数据库和搜索引擎，并支持在多个平台上进行关键词驱动的搜索，包括 Scopus、Web of Science、ScienceDirect 和 Google Scholar。针对每个来源，我们使用了官方 API（如有）或自动化抓取工具：

– **Scopus 和 ScienceDirect**：通过爱思唯尔 Scopus 应用程序接口[1] 访问 Scopus，该接口可通过其应用程序接口界面结构化地获取文章级元数据。ScienceDirect 也通过爱思唯尔 ScienceDirect 应用程序接口[2] ,访问，该接口支持检索文章级元数据。两个应用程序接口均通过 Pyscopus（基于 Python 的库）集成。

– **Web of Science (WOS)**：使用 Clarivate Web of Science API[3]，对 WOS 进行查询，该应用程序接口支持以编程方式检索搜索结果和元数据。WOSClient[4] 是一个官方客户端库，可配合来自 WOS 的 API 密钥获取搜索结果和文章级元数据。

[1] https://dev.elsevier.com/scopus.html[2]
https://dev.elsevier.com/sciencedirect.html[3]
https://developer.clarivate.com/apis/wos[4]
https://wos.readthedocs.io/en/latest/documentation/client.html

– **Google Scholar**：使用 Scholarly[5] 软件包开发了一个定制的网络抓取工具，用于从 Google Scholar 提取公开可用的元数据。

数据收集工具采用 Flask Web 框架实现，并集成了一个数据抓取引擎；该引擎在可用时通过API调用检索元数据、题名、摘要、文献标识符及其他信息。该方法的一个关键优势在于能够同时从所有来源提取数据，从而在时间和资源方面显著提高效率。该系统被用于在多个国家收集数据，并特别聚焦塞内加尔，以支持关于作物对养分施用响应的研究。例如，其中一项查询针对养分投入对塞内加尔主要作物可收获器官的影响，其表述如下：塞内加尔 *AND (*养分 *OR* 施肥 *OR* 肥料 *OR* 施用率 *OR* 施用量 *OR* 氮 *OR* 磷 *OR* 钾*) AND* 产量.同行评审文章检索自 Scopus、Web of Science 和 ScienceDirect，而 Google Scholar 则用于收集学术出版物和灰色文献，包括机构报告和大学学位论文。通过这一自动化方法，针对塞内加尔的检索共收集了 15,825 篇文章。其他国家的更多查询示例汇总于表1（Google Scholar 的数据除外）。这一可推广的框架能够在多种农业与环境领域实现高效、可扩展且可重复的数据收集。

表1. 针对各类领域特定关键词收集的文章总数

关键词	总计
（多光谱机载影像 OR 无人机 OR 无人机 OR 无人机系统 OR 无人驾驶航空器 OR 远程驾驶航空器系统）AND 氮 AND 产量	77489
玉米 OR 玉米 AND (谷物品质 OR 谷物成分) AND (氮肥施用 OR 水分胁迫 OR 干旱胁迫)	63056
产量 AND (固氮 OR 固氮 OR 大气来源氮 OR 大气来源氮比例 OR 氮-吸收 OR 氮吸收 OR 种子氮 OR 氮收获指数 OR 氮收获指数) AND (鹰嘴豆 OR 菜豆 OR 豇豆 OR 蚕豆 OR 豌豆 OR 花生 OR 小扁豆 OR 羽扇豆)	4413
氮稀释曲线 AND 氮营养指数 AND 临界氮浓度 AND (一年生黑麦草 OR 黍 OR 棉花 OR 饲料甜菜 OR 杂交黑麦草 OR 玉米 OR 燕麦 OR 多年生黑麦草 OR 马铃薯 OR 救荒草 OR 水稻 OR 高粱 OR 甘蔗 OR 向日葵 OR 甘薯 OR 高羊茅 OR 梯牧草 OR 小麦 OR 白甘蓝)	121
尼日利亚 AND (养分 OR 施肥 OR 肥料 OR 施用率 OR 施用量 OR 氮 OR 磷 OR 钾 OR 硫 OR 硫) AND 产量	5853
利比里亚 AND (养分 OR 施肥 OR 肥料 OR 施用率 OR 施用量 OR 氮 OR 磷 或 钾 或 硫 或 硫）与 产量	685
尼日尔 与（养分 或 施肥 或 肥料 或 施用率 或 施用量 或 氮 或 磷-磷 或 钾 或 硫 或 硫）与 产量	5774
加纳 与（养分 或 施肥 或 肥料 或 施用率 或 施用量 或 氮 或 磷-磷 或 钾 或 硫 或 硫）与 产量	5527
马里 与（养分 或 施肥 或 肥料 或 施用率 或 施用量 或 氮 或 磷-磷 OR 钾 OR 硫 OR 硫) AND 产量	5136

[5] https://github.com/scholarly-python-package/scholarly

3.2 数据筛选

在汇总来自多个来源的数据后，实施了数据清理步骤，以消除不一致之处，并确保所收集的文献数据集无误。此阶段的目标是识别并移除不相关、冗余和重复条目，从而创建统一的高质量数据集。最初，通过比较文档网址、数字对象标识符（DOI）和文章标题来识别重复文章。对于 Scopus 和 Science Direct 等来源，还通过比较唯一的$Scopus_id$增加了一层筛选，以消除重复记录。尽管检索到的大部分内容为英文，但通过 API 查询也纳入了一些非英文文章。因此，实施了语言筛选，以检测并移除非英文条目，确保语言一致性。

表 2 展示了我们从 Science Direct、Web of Science 和 Scopus 提取数据时的数据整合与筛选流程示例。我们从 Scopus 检索到 4,999 篇文章，从 ScienceDirect 检索到 670 篇，从 Web of Science 检索到 638 篇。合并 Scopus 和 ScienceDirect 的数据后，我们依据 Scopus_id 删除了 496 条重复条目，表明这两个来源之间存在大量重叠。合并来自 Web of Science 的数据后，我们采用基于 DOI 的去重筛选，发现有 277 篇文章具有相同的 DOI。此外，我们发现，由于数据集中存在 DOI 缺失或不一致的情况，仅依赖 DOI 可能存在局限性。因此，我们增加了一层基于标题的筛选，最终又删除了 7 条重复记录，证明了在其他标识符不完整时，这一方法仍然有效。

表 2.加纳相关关键词合并/去重过程统计

筛选阶段	**详情**
初始记录数	
Scopus	4,999
ScienceDirect	670
Web of Science	638
合并 Scopus 和 ScienceDirect 后	
Scopus_id 去重前	5,669
Scopus_id 去重后	5,173
已删除重复记录（Scopus_id）	496
与 WOS 合并并按数字对象标识符去重后	
数字对象标识符去重前	5,811
按数字对象标识符去重后	5,534
已移除的重复记录（数字对象标识符）	277
按题名去重后	
标题去重前	5,534
标题去重后	5,527
已删除的重复记录（题名）	7
最终记录数	
唯一记录的最终数量	5,527

3.3 使用大语言模型进行数据分类

传统筛选方法主要依赖领域专家进行人工检查，通过目视分析文章的题名和摘要来识别相关变量。这一过程不仅耗时，而且由于依赖领域专业知识，成本也很高。为缓解这一问题，我们研究了大语言模型理解不同领域上下文信息的能力，因为大语言模型所需的监督最少，并且已展现出理解自然语言文本中文本线索的强大能力。与依赖领域专门微调或专家目视检查的传统方法不同，大语言模型为筛选与领域相关的文献提供了一种可扩展且具有泛化能力的替代方案。

在本研究中，我们探讨了使用大语言模型自动将收集到的文章分类为相关（有关）和不相关（无关）类别，从而减少对专家主导的人工筛选的依赖。具体而言，我们在零样本设置下，将大语言模型——LLaMA2-7b [20], Phi-2 [14], 和 Gemma-2 [18]——集成到分类流程中。我们进行了提示设计、推理和评估，以提取并分析大语言模型的输出。

领域专家认为，文本中关键词出现次数可以为判断文本与这些关键词的相关性提供定量依据。为此，我们将关键词频率分析纳入用于借助大语言模型将文本分类为相关或不相关的零样本提示中。图2展示了我们使用的确切提示，说明了如何整合领域特定关键词和上下文线索来引导模型。

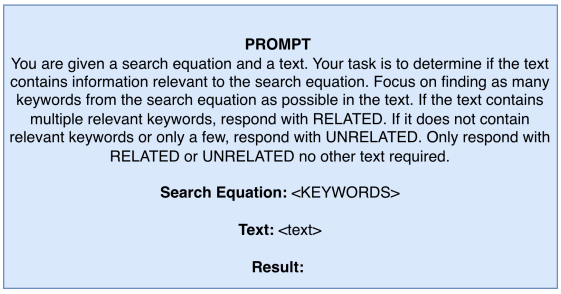

图 2。 使用大语言模型进行零样本分类的提示词

零样本学习期间，用于控制大语言模型输出的超参数如下：max_new_tokens 设置为 32，以限制模型响应的长度；eos_token_id 已定义，使模型在适当的时机停止生成文本；do_sample 设置为 True，以启用采样，从而生成更灵活、多样的输出；温度设置为 0.6，top_p 设置为 0.9。这些超参数在初步运行中产生了稳定的

预测，因此，为了确保公平比较，我们在所有模型中始终采用这些超参数。

4 结果

4.1 使用大语言模型进行零样本分类

我们的领域专家制定了四个农业相关搜索查询，并通过自动化关键词管道检索文章。表3展示了各种大语言模型的表现，并与领域专家手动筛选为相关的文章集进行比较；专家针对每个查询检索到的文章集要小得多。**人工判定相关** 列表示专家使用自己的方法从手动检索的文章中判定为相关的文章数量。**工具检索到的文献** 列表示我们的工具通过多个来源自动提取的文章数量。**模型**列表示用于分类的大语言模型。**模型判定相关** 指大语言模型从工具检索到的文献中分类为相关的文章数量。**人工判定相关 ∩ 工具检索到的文献** 列表示人工判定相关且同时包含在工具检索结果中的文章数量。**人工判定相关 – 工具检索到的文献** 列表示人工判定为相关但未被工具检索到的文章数量。**人工判定相关∩ 模型判定相关**列表示人工判定相关且同时被模型预测/分类为相关的文章数量。**重叠准确率** 列衡量模型在人工判定相关结果与工具检索结果的交集内正确识别相关文章的能力，计算方式为共享子集中模型正确判定为相关的条目所占百分比，即

$$\%\ \textbf{Overlap} = 100 \times \frac{\text{HumanRelevant} \cap \text{ModelRelevant}}{\text{HumanRelevant} \cap \text{ToolRetrieved}}.$$

评估每个基于搜索关键词的查询的重叠准确率尤为重要，因为它反映了模型在人工判定为相关且自动检索到的文章这一共同子集中的表现。如表3所示，所有大语言模型在所有查询中的准确率均稳定超过90%。Phi-2模型的整体准确率最高，其中4个查询有3个达到100%。这些结果证明了我们的方法能够有效地从摘要中提取有意义的信息，并准确识别相关内容。

分析由大语言模型标记为相关的文章后可以发现，模型判定相关文献的数据集明显大于仅通过人工检索通常能够得到的数据集。不过，也有一些文章被专家人工判定为相关，却未被模型判定为相关，即模型漏掉了一些文章（列**人工判定相关–工具检索到的文献**）。对这些文章进行检查后发现，许多未纳入重叠集的文章仍以其他形式存在于模型判定相关文献的数据集中，只是未通过数字对象标识符匹配或标题精确匹配识别出来。例如，两个未匹配的标题以不同方式表示氮（nitrogen

和化学符号 N）：“冬小麦作物临界 N 稀释曲线的确定”与“冬小麦作物临界氮稀释曲线的确定”。这些差异通常可归因于数据源不同、各平台上文章标题的格式差异以及数据收集的时间差异。此外，搜索引擎会频繁更新其索引算法，这可能导致搜索结果发生变化，例如删除过时条目或优先显示较新的出版物。这些因素可能共同造成了观察到的重叠集差距。尽管如此，重叠准确率仍是衡量模型能否准确识别相关文章的一项关键指标。结果表明，大语言模型有望有效复现并可能取代传统上用于数据收集和筛选的双盲人工方法。

表3。（Llama2-7b、Gemma-2、Phi-2）模型筛选与若干特定领域的基于关键词的查询相关文献的表现，并与人工标注的相关文献进行比较。**关键词**：用于检索文献的搜索查询。**人工判定相关**：人工识别为相关的文献数量，即专家使用自己的方法在规模小得多的文献子集中找到的数量。**工具检索到的文献**：自动化工具检索到的文献数量。**人工判定相关 ∩ 工具检索到的文献**：人工识别为相关且同时出现在工具检索结果集中的文献数量。**人工判定相关 − 工具检索到的文献**：人工识别为相关但未被工具检索到的文献数量。**模型**：用于预测的大语言模型。**模型判定相关**：模型从工具检索结果集中选为相关的文献数量。**人工判定相关 ∩ 模型判定相关**：人工识别为相关且也被模型预测为相关的文献数量。**(% 重叠 = 重叠准确率)**：$100 \times$ 人工判定相关∩模型判定相关 人工判定相关∩工具检索到的文献 。

关键词	人类 相关的检索结果	Tool ∩工具检索到的文献	人工判定相关 − 工具检索到的文献	人工判定相关	模型	模型 相关	人工判定相关 ∩模型判定相关	重叠 (%)
多光谱	41	77489	36	5	Llama2-7b	47666	33	91.67
					Gemma-2	68063	33	91.67
					Phi-2	78681	36	100
Corn	21	63056	13	8	Llama2-7b	48752	12	92.31
					Gemma-2	57569	12	92.31
					Phi-2	62460	13	100
固氮	83	4413	46	37	Llama2-7b	3644	43	93.48
					Gemma-2	3991	39	84.78
					Phi-2	4282	43	93.48
氮稀释	34	121	14	20	Llama2-7b	120	14	100
					Gemma-2	109	13	92.86
					Phi-2	121	14	100

多光谱：（多光谱机载影像或无人机或无人机或无人机系统或无人驾驶航空器或远程驾驶航空器系统）以及氮以及产量

玉米：玉米或玉米以及（谷物品质或谷物成分）以及（氮肥施用或水分胁迫或干旱胁迫）

固氮：产量 AND（固氮 OR 固氮 OR 大气来源氮 OR 大气来源氮比例 OR 氮吸收 OR 氮吸收 OR 种子氮 OR 氮收获指数 OR 氮收获指数）AND（鹰嘴豆 OR 菜豆 OR 豇豆 OR 蚕豆 OR 豌豆 OR 花生 OR 小扁豆 OR lu

pi͡ŋs **氮稀释**：氮稀释曲线以及氮营养指数以及临界氮浓度以及（一年生黑麦草或扫帚高粱或棉花或饲料甜菜或杂交黑麦草或玉米或燕麦或多年生黑麦草或马铃薯或救荒草或水稻或高粱或甘蔗或向日葵或甘薯或高羊茅或梯牧草或小麦或白甘蓝）

5 摘要筛选工具

为了实现可扩展的自动化数据收集与分类，我们使用Flask框架开发了一个基于网络的工具。该工具包含多个集成组件，其中包括一个数据抓取模块，可从Scopus、Web of Science、Science Direct和Google Scholar等学术来源检索文献。系统通过API调用从各个数据库或搜索引擎中提取关键元数据字段，例如题名、摘要、作者和文献标识符，以及其他特定于搜索引擎的信息。图3展示了我们基于Web的应用程序的仪表板，用户可借此使用特定关键词在多个来源中高效检索数据，或访问下载页面以获取先前收集的数据集。数据收集在多个来源上并行执行，从而加快检索速度并提高可访问性。

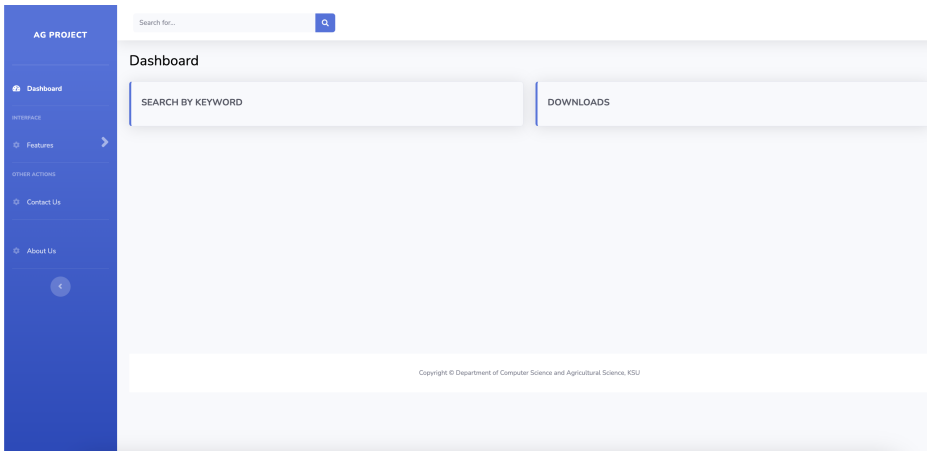

图 3. 摘要筛选工具仪表板

如图 4 所示，简易输入表单提供一站式界面，可启动跨多个数据源/搜索引擎的关键词搜索。用户可以为每次搜索指定自定义别名，之后系统会用该别名在下载页面标记相应的数据集。为确保唯一性，每个别名都必须与其他别名不同。对于 Scopus 和 Science Direct，用户可以设定要检索的文章数量，每个数据源最多可检索 5,000 篇论文。对于 Web of Science (WOS)，数据按页检索，用户可以输入 0 到 100 之间的数值，以确定每次 API请求要抓取的页数。这样可以提取这些页面中的所有可用文章。用户还可以选择将 Google Scholar 设为数据源。由于平台对自动抓取的限制，以及缺少摘要元数据会降低下游分类性能，此选项默认设为否。如果用户选择

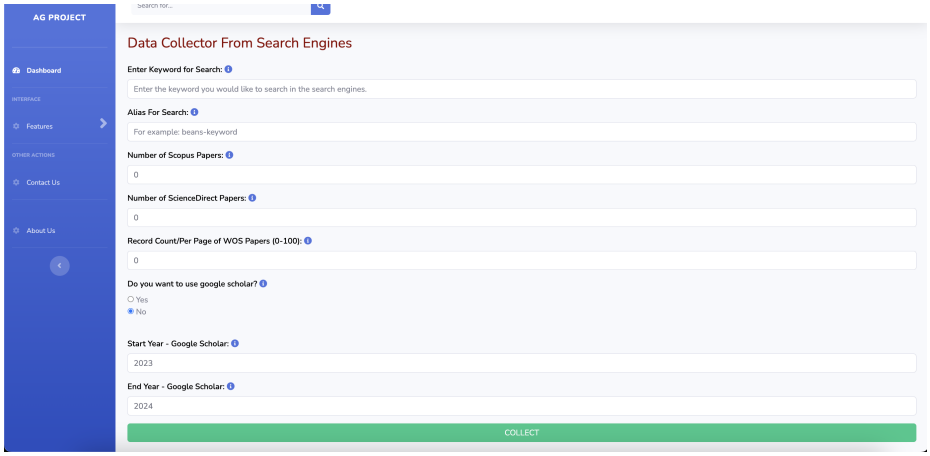

图4. 摘要筛选工具的数据采集页面

纳入 Google Scholar，工具将使用文章题名而非摘要进行分类。用户还可以指定出版年份范围，每年最多收集 1000 篇文章。用户点击收集按钮后，系统便会开始同时从所有选定的数据源检索数据，从而实现高效且可扩展的数据收集。

数据收集完成后，该工具会自动执行一系列预处理步骤，包括删除重复条目、筛除非英语内容以及验证元数据。随后，清理后的数据会在后台传递给由大语言模型驱动的分类层/模块，该模块使用与当前搜索关键词相关的领域特定提示词，将文章分类为相关或不相关。最后，用户可以下载以开放获取 CSV 格式提供的采集数据，从而提高透明度并便于重复使用。

该工具旨在利用并行处理，同时查询多个数据源，显著提升速度和效率。它还具备多项功能，可灵活适配不同研究领域。通过自动化原本需要领域专家投入大量人工工作量和时间的任务，该工具提供了一种可扩展且有效的解决方案，其适用范围远超最初的农业文献应用场景。尽管广泛的跨学科开放数据倡议难以在科学界获得广泛采用，但此类领域特定平台在促进有效数据共享和提升数据可及性方面更具潜力。

6 结论与未来工作

本文提出了一种可推广、可扩展的方法，利用大语言模型自动收集、分类和整理科学文献。所提出的方法能够高效地收集大量

来自多个数据源/搜索引擎的数据。该方法实现了出色的分类性能（即，领域专家选取的数据与大语言模型选取的数据重合度超过 90% 的准确率，且无需专家监督）。本文介绍的工具支持创建开放获取的特定领域数据库，并通过提供可下载的数据集促进可复现研究。

该工具的一个局限是，Google Scholar 等平台对数据抓取有所限制。此外，分类准确率可能因摘要的清晰度、搜索查询中所用关键词的相关性，以及用于零样本学习的大语言模型而异。未来的工作可以尝试采用少样本学习来提高分类准确率。尽管所提出的方法旨在做到领域无关，但将更多特定领域的数据源纳入搜索流程并实现动态集成，也是未来工作的重要方向。

致谢

本出版物所报告的研究得到了 USDA-NRCS 资助项目（NR233 A750011G001）的支持。该研究还部分得到了美国国立卫生研究院生物医学研究卓越中心（COBRE）下属的认知与神经生物学可塑性研究中心（CNAP）的支持，资助编号为 P20GM113109。本文内容仅由作者负责，并不一定代表美国国立卫生研究院和 USDA-NRCS 的官方观点。

参考文献

1. Beillouin, D., Demenois, J., Cardinael, R., Berre, D., Corbeels, M., Fallot, A., Boyer, A., Feder, F.：土地管理、土地利用变化和气候变化对土壤有机碳影响的全球数据库。Scientific Data **9**(1), 228 (2022) 2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee- lakantan, A., Shyam, P., Sastry, G., Askell, A., 等：语言模型是少样本学习者。神经信息处理系统进展 **33**, 1877–1901 (2020) 3. Chamorro-Padial, J., García, R., Gil, R.：农业开放数据系统综述。Computers and Electronics in Agriculture **219**, 108775 (2024) 4. Ciampitti, I., Van Versendaal, E., Rybecky, J.F., Lacasa, J., Fernandez, J., Makowski, D., Lemaire, G.：用于拟合主要作物物种氮稀释临界曲线的全球数据集。Scientific Data **9**(1), 277 (2022) 5. 科睿唯安：Web of Science：研究发现平台。https://www.webofscien ce.com (2025 12 6 2025 11), 访问日期：- -。Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向变换器预训练。载于：Burstein, J., Doran, C., Solorio, T. (编)《2019年北美计算语言学协会会议论文集：人类语言技术》，第1卷（长短篇论文）。第4171–4186页。计算语言学协会，明尼阿波利斯，明尼苏达州（2019年6月）

7. 爱思唯尔：ScienceDirect：科学、技术与医学研究。https://www.sciencedirect.com，访问日期：2025-11-12。8. 爱思唯尔：Scopus：文摘与引文数据库。https://www.scopus.com，访问日期：2025-11-12。9. Gamage, P., Rajapakse, C., Galpaya, H.：什么样的农业倡议才算开放？关于共享农业信息、写作权和不同结果的思考。开放发展的批判性视角，第227页 10. Gomez, F., Carcedo, A., Diatta, A., Djighaly, P., Nagarajan, L., Singh, U., Stewart, Z., Zingore, S., Majumdar, K., Prasad, P.V., 等：迈向特定地点养分管理策略：塞内加尔开放数据库。Science of The Total Environment **969**, 178926 (2025) 11. Gomez, F., Carcedo, A., Mean, C.M., Reyes, M., Hok, L., Tivet, F., Seng, V., Vara Prasad, P., Ciampitti, I.：东南亚地区农业系统土壤有机碳数据集。Scientific Data **11**(1), 374 (2024) 12. Google：Google Scholar。https://scholar.google.com，访问日期：2025-11-12。13. Ledo, A., Hillier, J., Smith, P., Aguilera, E., Blagodatskiy, S., Brearley, F.Q., Datta, A., Diaz-Pines, E., Don, A., Dondini, M., 等：多年生作物种植下土壤有机碳变化的全球经验性协调数据集。Scientific Data **6**(1), 57 (2019) 14. MicrosoftResearch：Phi-2：潜力巨大的小型语言模型。https://www.microsoft.com/en-us/research/blog/phi-2-a-small-language-model-with-big-potential (2023)，访问日期：2025-11-12。15. Nagarajan, L., Singh, U.：土壤联盟（2022）16. Pichai, S., Hassabis, D.：推出 Gemini：我们规模最大、能力最强的 AI 模型。Google。检索日期：2023年12月**8**（2023）17. Tamene, L.D., Erkossa, T., Tafesse, T., Abera, W., Schultz, S.：自愿者联盟：为埃塞俄比亚农业提供数据驱动型解决方案（2021）18. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., 等：Gemma 2：以实用规模改进开放语言模型。arXiv预印本 arXiv:2408.00118 (2024) 19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., 等：LLaMA：开放且高效的基础语言模型，2023 13971 20 2302 。网址 https://arxiv. org/abs/ 。Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bash- lykov, N., Batra, S., Bhargava, P., Bhosale, S., 等：Llama 2：开放式基础模型与微调聊天模型。arXiv预印本 arXiv:2307.09288 (2023) 21. WALL, A.G.G.：非洲开放协议（2021）22. Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J.：智慧农业中的大数据——综述。农业系统 **153**, 69–80 (2017)