AutoDataset：用于持续发现和搜索数据集的轻量级系统

Junzhe Yang[1,2], Xinghao Chen[1,3], Yunuo Liu[1],Zhijing Sun[1],Wenjin Guo[1],Xiaoyu Shen [1]

[1] 数字孪生研究院、东部理工学院，宁波[2] 上海交通大学 [3] 香港理工大学 yang_j@sjtu.edu.cn, xyshen@eitech.edu.cn

arXiv:2603.07271v1 [cs.IR] 7 Mar 2026

摘要

特定任务数据集的持续扩展已成为推动机器学习发展的重要动力。然而，新发布数据集的发现仍然困难，因为现有平台主要依赖人工整理或社区提交，导致覆盖范围有限且延迟显著。为应对这一挑战，我们推出 AUTODATASET，这是一款用于实时数据集发现和检索的轻量级自动化系统。AUTODATASET 采用论文优先的方法，持续监控 arXiv，直接从新发表的研究论文中检测并索引数据集。该系统通过低开销的多阶段流程运行。首先，轻量级分类器快速筛选标题和摘要，以识别发布数据集的论文，其 F1分数达到 0.94，推理延迟为 11 毫秒。对于识别出的论文，我们使用 GROBID 解析 PDF文件，并通过句子级提取器提取数据集描述。系统从论文正文中提取数据集网址，必要时自动回退到 LaTeX 源码分析。最后，系统使用稠密语义检索器对结构化记录进行索引，从而支持低延迟的自然语言搜索。我们将 AUTODATASET 部署为一个实时系统，持续摄取新论文并提供最新的数据集发现服务。实践表明，该系统显著缩短了研究人员查找新发布数据集所需的时间，数据集发现效率最高提升了 80%。[1]

1 引言

近年来，机器学习取得了快速进展，这在很大程度上得益于训练数据和模型规模的持续扩展（Kaplan等人，2020年；Zhao等人，2023年；Chung等人，2024年；Su等人，2023年，2024年；DeepSeek-AI，2025）。在许多子领域，进展如今已与高质量的特定任务数据集的可用性紧密相关（Haddow等人，2022年；Shen等人，2022年；Zhu等人，2023a年、b年）。这种依赖推动了数据集创建工作的爆发式增长，具体表现为发布新数据集和基准的论文数量稳步上升（Lhoest等人，2021年；Liu等人，2024年）。随着数据集发布速度不断加快，数据集发现变得日益困难。尽管数据集搜索已发展成为一个独立且重要的研究领域（Paton等人，2024年；Li等人，2025），但如表1所示，新论文的数量和发布速度之大，如今已使人工整理和临时跟踪难以应对。这些趋势呼唤能够在新数据集发表后立即对其进行跟踪、结构化和索引的自动化系统，而不是等待人工整理。

早期阶段（2004–2014）：逐渐兴起						
Year	2004	2006	2008	2010	2012	2014
论文	∼5	∼10	∼5	∼15	∼40	∼90
近期阶段（2020年–2025）：爆发式增长						
Year	2020	2021	2022	2023	2024	2025
论文	2,865	2,479	4,746	7,347	7,655	**8,566**
同比增长	–	−13%	**+91%**	**+55%**	+4%	+12%
对比2014	32×	27×	53×	82×	85×	**95**×

表1：我们的BERT-Gate检测器将论文判定为引入新数据集的逐年数量，依据为六个计算机科学类别中的arXiv投稿：信息检索、数据库、人工智能、计算语言学、计算机视觉和多智能体系统。这些数量从2015年前的基线到近年急剧增长，到2025年增长了 ∼95倍。

为应对这一挑战，我们推出 AUTO-DATASET，这是一个以论文为先的自动化系统，能够**低延迟地发现和检索**机器学习数据集。尽管 Hugging Face（Lhoest 等人，2021）和 OpenML（Vanschoren 等人，2013）等社区驱动的托管平台提供了宝贵的基础设施，但它们依赖用户自愿上传，往往导致文档记录不一致，

[1] **系统可用性：**源代码已公开发布于 https://github.com/EIT-NLP/AutoDataset。展示完整工作流程的屏幕录制视频可在https://youtu.be/_QSxHKyIYns获取。

维护情况参差不齐，去重问题也十分严重，同一数据集常常以不同名称、版本或格式出现。与此相反，我们基于一个简单却有力的观察：高质量机器学习数据集通常通过经同行评审的论文或预印本研究论文介绍。这些论文提供丰富且标准化的信息，涵盖数据收集、任务定义、实验验证和预期用途，能有力表明数据集的质量与相关性。此外，由于一个数据集通常对应一篇规范性介绍论文，以论文为锚点开展发现，自然可以避免冗余，无需依赖脆弱的启发式去重方法。

在实际运行中，AutoDataset持续监控 arXiv [2]——现代机器学习研究领域首屈一指的预印本平台，以收录介绍数据集的论文。与一次性的文档理解任务或离线整理不同，我们的系统必须在严格的延迟和吞吐量限制下处理每一篇新提交的论文。这一要求使得简单依赖大语言模型（LLM）智能体并不实际，因为它们计算成本高、运维开销大（Guo 等，2024年；Shen 等，2025；Belcak 等，2025；Ding 等，2026）。因此，AUTODATASET 采用了刻意设计的轻量、模块化流水线，以实现可扩展的数据摄取，满足长期部署需求。具体而言，该系统采用两阶段基于BERT的架构（Devlin 等，2019）。第一阶段 BERT-GATE 是一个高速过滤器，仅依据论文的标题和摘要进行分类。我们采用这一设计，是因为标题和摘要通常包含足够的语义信号来识别数据集贡献，使系统能够准确筛选数千篇论文，同时将每篇论文的处理时间控制在约 11 毫秒。

对于已确认的论文，第二阶段 BERT-Desc会解析完整的PDF文件，以提取简洁、人类可读的数据集描述。同时，为确保成功提取主数据集URL，系统会直接从PDF中提取链接，并在必要时自动查阅LaTeX源文件。最后，这些结构化记录会被索引，采用最先进的稠密语义检索器（Karpukhin等人，2020年；Li等人，2023年），并通过直观的自然语言搜索界面加以呈现。我们部署将其部署为可配置的Web应用，支持实时爬取、参数调优以及即时的数据集探索。通过直接呈现已验证的数据集链接及简洁描述，AUTODATASET将查找新发布数据集所需的时间估计减少80%，把以往手动且易出错的工作流程转变为精简的检索体验。

总而言之，我们的工作有三项关键贡献：(1) 我们推出 AUTODATASET，这是一个完全自动化、以论文为先的系统，持续监控 arXiv，以发现并编入索引介绍数据集的论文。(2) 我们设计了一个高效的两阶段 BERT 流水线，用于快速筛选和描述提取，并结合 URL 模块，确保直接从 PDF文件中成功抽取链接，必要时还会查阅 LaTeX 源文件。(3) 我们构建了一个集中式可搜索索引，并发布了一个轻量、交互式的 Web演示，以促进高效的实时数据集检索。

2 相关工作

数据集发现平台 现有的数据集发现平台，例如 Google Dataset Search（Noy 等人，2019）、Hugging Face Datasets（Lhoest 等人，2021）、Kaggle（Kaggle）、OpenML（Vanschoren 等人，2013）和 Papers with Code（Papers withCode），在很大程度上依赖人工上传和社区整理。正如表 2所总结的，这些平台面临若干共同局限，包括元数据不一致、数据集质量参差不齐，以及论文发布与数据集在平台上可用之间存在明显延迟。除了目录式存储库之外，近期工作如 DataScout（Lin 等人，2025）探索了交互式、由人工智能辅助的数据集发现界面，但这些界面仍建立在现有的人工整理存储库之上，而非直接从文献中挖掘数据集。

自动化数据集提及抽取 早期的自动化数据集发现工作主要将提及抽取视为序列标注问题。这包括先对句子分类、再识别提及的两阶段流水线（Kumar 等人，2021年），以及应用于全文的 BiLSTM-CRF模型（Zeng 和 Acuna，2024年）。为支持该领域的稳健评估，DMDD（Pan 等人，2023年）推出了一个用于数据集提及检测的大规模基准语料库。近来，大语言模型和合成数据已成为主要工具

[2]https://arxiv.org/

平台	整理方法	基于文献	更新延迟	元数据一致性
Google Dataset Search	自动网络爬取	混合	中等	异构
Hugging Face Datasets	社区上传	部分	中等至高	中等
Kaggle	手动 / 社区	很少	High	Low
OpenML	手动上传	部分	High	High
Papers with Code	社区整理	Yes	低至中等	中等
AUTODATASET（我们的）	**自动化论文挖掘**	**Yes**	**极低**	**High**

表2：数据集发现平台的比较。与依赖人工整理或一般网络爬取的现有存储库不同，AUTODATASET直接自动挖掘文献，实现近乎零延迟。

，用于大规模识别和规范化数据集引用（Marini 等人，2025；Datta 等人，2024年；Solatorio 等人，2025）。与我们的设定最接近的是 ChatPD（Xu 等人，2025），它利用由大语言模型驱动的流水线，从文献中构建论文—数据集网络。相比之下，AUTODATASET 优先考虑摄取速度和效率，以轻量且高度针对性的 BERT 流水线取代繁重的大语言模型提取，从而实现对预印本服务器的持续实时监测。

3 AUTODATASET流水线

为了持续、低延迟地摄取新发布的数据集，AUTODATASET 被设计为一条高效的端到端处理流水线。如图1所示，系统包含四个核心模块：（1）**数据集检测**模块，充当高速过滤器，根据论文的标题和摘要识别介绍数据集的论文；（2）**描述提取**模块，处理这些被判定为正例的论文全文，生成简洁的数据集摘要；（3）**链接抽取**模块，与此同时运行，以可靠地提取主数据集网址；（4）**索引与检索**模块，汇总提取出的元数据，为直观的自然语言搜索界面提供支持。

数据集检测 为持续监测文献，AUTODATASET 会自动摄取六个以数据集发表率高而著称的基础计算机科学领域的新 arXiv 投稿：信息检索（IR）、数据库（DB）、人工智能（AI）、计算语言学（CL）、计算机视觉（CV）和多智能体系统（MA）。在此初始摄取阶段，对论文元数据进行稳健解析时使用 BeautifulSoup（Richardson）以处理复杂的 HTML 和 XML 文档结构。由于每日新增预印本数量庞大，依赖重型生成模型进行初始筛选在计算上代价过高。

为解决这一问题，我们采用 BERT-GATE（一种基于 bert-base-uncased 的轻量级二元分类器）作为高速初始过滤器。BERT-GATE 仅处理每篇论文拼接后的标题和摘要；我们发现，这些信息提供了足够的语义信号，能够可靠地判断是否有新数据集被引入。该模型使用经过筛选的 7,789 篇标注论文进行训练 [3]。在留出测试集上的评估结果表明，BERT-GATE 具有很强的可分性，准确率达到 0.97，F1分数达到 0.94。尤为重要的是，这种轻量级架构使每篇论文的平均处理延迟约为 11 毫秒。预测分数超过 0.5 阈值的论文会被判定为正例，并立即传递至描述提取模块。[4]

描述提取 一旦 BERT-GATE将某篇论文识别为数据集引介论文，系统便会启动深度内容分析，以提取简洁、机器可读的数据集描述。为确保文本提取稳健可靠，系统会下载标准PDF，并使用 GROBID（Lopez，2009）进行处理，生成结构化的句子级XML表示。

为确定哪些句子确实描述了数据集，我们采用了经过微调的句子级分类器[4]。由于数据集描述通常跨越多个句子，孤立地对句子进行分类会严重限制性能。为解决这一问题，我们开发了一种类条件滑动窗口方法。对于每个目标句子，模型会接收一段以该句子为中心、对称分布的上下文文本，其长度不超过最大词元上限。为在高效降低冗余负跨度权重的同时提供丰富的话语语境，窗口在正标签附近以较小步幅滑动，而在负标签附近则采用较大步幅。

[3]通义千问-Plus生成的初始弱标签经过了人工核验。 [4]附录 A 包含详细的训练超参数。

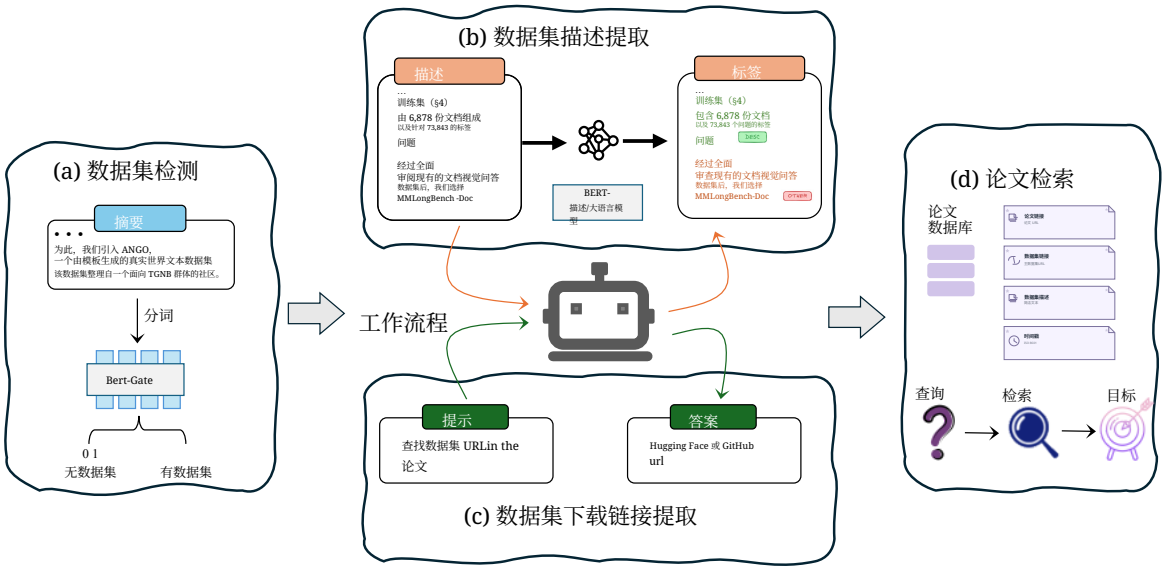


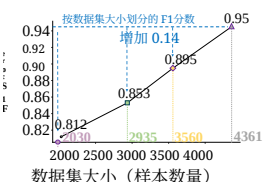
图1：AutoDataset 的系统架构

模型	Acc.	精确率	Rec.	F1	时间（毫秒）
Longformer	0.923	0.678	0.831	0.747	265.20
BigBird	0.919	0.679	0.774	0.723	257.98
通义千问-Plus（API）	0.809	0.636	**0.933**	0.757	1039.75
BERT-Desc	**0.949** ↑2.8%	**0.800** ↑17.8%	0.924 ↓1.0%	**0.858** ↑13.3%	**9.39** ↓ 96.5%
- 无窗口	0.828	0.544	0.654	0.594	9.46

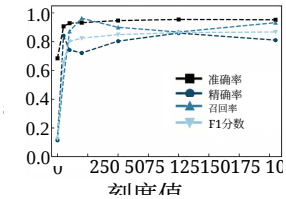
表3：数据集描述抽取任务上的性能。所提出的**BERT-Desc** 在总体F1值和效率上均显著优于基线方法。彩色百分比表示相对于各指标中表现最佳基线的相对提升（↑）或下降（↓）。最后一行给出了消融实验。

如表3所详述，这种上下文化方法取得了较高的准确率和F1分数，显著优于缺乏上下文窗口的基线方法。此外，BERT-DESC 的推理延迟极低。与Longformer（Beltagy等，2020年）和BigBird（Zaheer等，2020年）等计算开销较大的长上下文文档转换器相比，它在预测准确率和处理速度方面均具有显著优势。同样，它完全避免了与基于API的大语言模型相关的严重推理瓶颈，例如Qwen-Plus（Bai等，2023年）。

被分类为正向的句子按文档顺序汇总，构成最终的数据集描述。如果 BERT-DESC 在某篇文档中未找到任何正向句子，该论文就会被重新分类为负向。这样一来，该模块便能充当高效的第二阶段过滤器，减少初始检测阶段产生的误报。在生产环境中，这条完整的抽取流水线涵盖 PDF下载、解析和推理，处理



(a) BERT-GATE 的 F1值随数据集大小的变化。

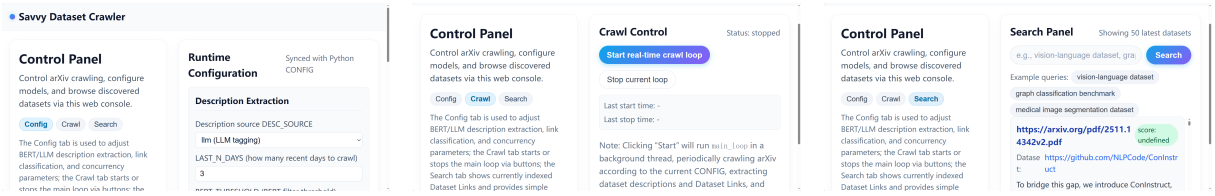


(b) BERT-DESC 指标随数据集大小的变化。

图2：数据规模消融实验。两个模型都展现出很强的数据效率，无需大量标注语料库便能迅速趋于饱和（达到最佳性能）。

文档的速度足够快，能够支持持续的实时摄取。

BERT-GATE 和 BERT-DESC 都展现出卓越的数据效率。如图2所示，我们通过评估模型在不同训练集规模下的性能，开展了数据规模消融实验。两种模型都能以极少的标注样本迅速接近性能饱和状态——



(a) 配置面板 (b) 爬取控制面板 (c) 搜索与检索面板

图 3：AUTODATASET Web界面。用户可以动态配置抽取参数(a)、实时监控 arXiv 摄取循环 (b)，并通过稠密语义搜索 (c) 浏览生成的数据集索引。

仅需极少的标注样本便能达到性能饱和。这种出色的数据效率验证了我们的设计选择：通过针对高度结构化的文本组件，AUTODATASET 无需付出大规模人工标注所带来的高昂成本，即可完成充分优化。

数据集链接抽取 在抽取描述的同时，系统会定位新引入数据集的托管主URL。仅依赖解析后的PDF进行链接抽取往往会失败，因为URL经常脱离其描述上下文，例如被 relegated 到脚注或参考文献中。这种错位破坏了准确识别所必需的语义关联。

为确保成功抽取链接，AUTODATASET 采用基于 LaTeX 源代码分析的稳健混合方法。由于目标 arXiv 类别中的绝大多数论文都提供可下载的 LaTeX 源文件，系统会获取这些文件并抽取所有标准超链接。随后，系统将每个候选URL与其周围的句子关联起来，以保留语义上下文。

完成候选项抽取后，我们采用一套全面的基于规则的评分系统（见附录 B）来过滤明显不相关的 URL，例如通用学术门户或根级存储库链接。对于剩余候选项，可选的轻量级语言模型验证步骤将作为最终分类器，以高置信度选出唯一的主数据集URL。如果禁用该验证模块或其返回不确定的结果，系统会无缝回退至基于规则的启发式方法中得分最高的候选项。最终输出是一个结构化 JSON 记录，其中包含论文链接、主数据集链接和抽取的描述，可直接用于后续索引。

数据集检索与评估 为了将提取的数据集记录与用户搜索意图衔接起来，AutoDataset 采用密集语义检索架构（Karpukhin等人，2020年）。用户的自然语言查询和流水线提取的数据集描述都会被映射到共享的高维向量空间中。对于这一嵌入层，我们采用`Alibaba-NLP/gte-large-en-v1.5`编码器（Li等人，2023年），因其强大的语义搜索能力而选用。为确保查询时的低延迟，所有数据集描述都会预先进行离线编码。当用户提交搜索请求时，查询会实时进行嵌入，并使用余弦相似度与离线索引进行匹配，从而返回最相关的最近邻。

为了评估系统端到端的检索效果，我们整理了一个由自然语言查询构成的基准数据集，其中每条查询都对应一条特定的真实标签数据集记录。如搜索界面（图 3c）所示，检索器展现出极高的结果呈现准确率，能够持续将正确的数据集排在搜索结果的最前列。将检索窗口进一步略微扩大后，覆盖率几乎达到完美。这些指标证实，该系统成功弥合了非正式用户查询与自动抽取的描述之间的语义鸿沟，提供了响应迅速且精准的搜索体验。

4 InteractiveWeb演示

为了展示 Auto-Dataset 的实际应用价值，我们将该系统部署为一个轻量级、交互式的 Flask Web应用。如图 3 所示，界面分为三个不同的面板，用户可借此管理整个论文到数据集流水线：包括 **配置** 面板、**爬取**控制器，以及自然语言**搜索**界面。

系统配置与爬取配置面板（图3a）通过直观的表单呈现流水线的核心超参数。用户可以动态调整 BERT过滤

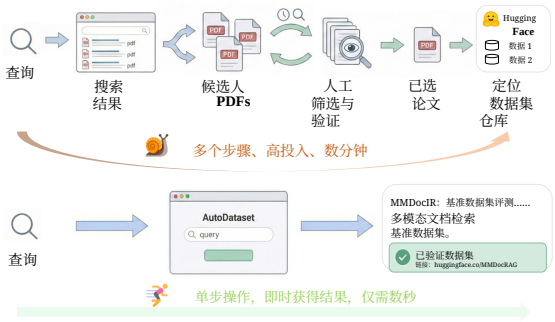


图4：传统人工工作流与AUTODATASET工作流程在发现多模态文档信息数据集方面的对比。

置信度阈值，指定要监控的 arXiv 论文时间窗口，并切换是否启用大语言模型回退机制进行链接验证。系统级参数（例如并行工作线程数和 PDF 下载的最大并发数）也可以在运行时调整，以适应不同的硬件限制。完成配置后，用户可以通过爬取面板（图3b）监控自动化摄取流程。系统轮询 arXiv，并通过两阶段 BERT 流水线处理新论文时，所有提取的数据集记录（包括描述和已验证的 URL）都会序列化为 JSON 对象，并持续追加到后端索引中。

5 案例研究

为了说明 AUTODATASET 如何支持贴近实际的研究工作流程，我们以一位对新发布的多模态文档信息检索数据集感兴趣的用户为例。一个自然的起点是在通用搜索引擎中输入“多模态文档信息”等网络查询，并手动搜索合适的基准数据集。

图4 对比了这种交互方式与传统的手动工作流程。没有 AUTODATASET 时，用户首先进行网络搜索，检查结果摘要，打开几个候选 arXiv PDF 文件，然后浏览正文或附录，查找数据集提及。对于每篇候选论文，用户还必须点击项目链接（例如 GitHub 或 Hugging Face），并手动确认哪个 URL 对应实际的基准数据集仓库。以 MMDocIR 为例，这需要从搜索引擎跳转到论文《长文档多模态检索基准测试》的 arXiv PDF，接着找到 MMDocIR 项目页面，最后找到数据集仓库。

使用 AUTODATASET 时，同样的信息需求可以更直接地得到满足。用户在搜索面板中提交查询“多模态文档检索”，系统便会返回一条以数据集为中心的 MMDocIR 基准数据集记录，其中包含指向来源论文的链接 [5] 、预测的主数据集 URL [6]，，以及简短的文字描述和时间戳，标明该记录最近一次在爬取过程中被发现的时间。我们手动核实了返回的 URL 是官方项目页面，其中概述了该基准数据集及其任务定义，并提供评测集和训练集的访问途径。

这种直接呈现已验证链接的方式显著改进了工作流程。根据一项小规模试点研究，三位研究人员搜索五个近期发布的数据集时，手动发现新推出的数据集通常每次查询需要 4–9 分钟，主要耗时在扫描 PDF 和验证链接上。相比之下，使用 AUTODATASET 后，工作流程缩短为检查一条排序结果并点击已验证的数据集网址，通常只需 0.5–2 分钟。据估计，这使获取首个有效链接所需的时间减少了 60–80%；对于访问尚未被社区平台收录的最新数据集，这一优势尤为重要。

6 结论与未来工作

本文介绍了 AUTODATASET，这是一个以论文为先、全自动化的系统，旨在持续监测预印本文献，并以近乎零延迟的方式为新发布的机器学习数据集编制索引。该系统以高效的两阶段 BERT 流水线和稳健的 LaTeX 辅助链接抽取，取代了庞大的生成模型，在不牺牲精确率的情况下实现了快速处理。配套的 Web 演示证明，这种自动化、基于文献的方法在技术上可行，也具有很强的实用性，能够加速数据集发现。

在未来的工作中，我们计划直接从文本中提取更细粒度的元数据（例如任务适用性、数据模态和语言标签），以增强系统。我们还计划加强跨论文去重，并与现有社区存储库（如 Hugging Face 和 Papers with Code ）实现更紧密的集成。最

[5] https://arxiv.org/pdf/2501.08828v3.pdf [6] https://mmdocrag.github.io/MMDocIR/

终，AUTODATASET 为构建更丰富的实时发现服务奠定了坚实基础，使其能够跟上实证机器学习的快速发展。

参考文献

Jinze Bai 等。2023年。 通义千问技术报告。*arXiv* 预印本 *arXiv:2309.16609*。

Peter Belcak、Greg Heinrich、Shizhe Diao、Yonggan Fu、Xin Dong、Saurav Muralidharan、Yingyan Celine Lin 和 Pavlo Molchanov。2025。 小型语言模型是智能体人工智能的未来。 预印本，arXiv:2506.02153。

Iz Beltagy、Matthew E. Peters 和 Arman Cohan。2020年。Longformer：长文档 Transformer。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。2024年。规模化指令微调语言模型。机器学习研究期刊，25(70):1–53。

Priyangshu Datta、Suchana Datta 和 Dwaipayan Roy。2024年。与文献较量：利用大语言模型进行数据集提及抽取。载于第*24*届 *ACM/IEEE* 数字图书馆联合会议论文集（*JCDL2024*），第 29:1–29:12 页，中国香港。ACM。

DeepSeek-AI。2025。Deepseek-v3.2：推动开放式大语言模型的前沿发展。预印本，arXiv:2512.02556。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会会议论文集，第 4171–4186 页。

Longwei Ding、Anhao Zhao、Fanghua Ye、Ziyang Chen 和 Xiaoyu Shen。2026年。从大语言模型到大推理模型：重新思考推理中心型模型的剪枝。*arXiv* 预印本，*arXiv:2601.18091*。

Taicheng Guo、Xiuying Chen、Yaqi Wang、Ruidi Chang、Shichao Pei、Nitesh V Chawla、Olaf Wiest 和 Xiangliang Zhang。2024年。基于大语言模型的多智能体：进展与挑战综述。载于第三十三届国际人工智能联合会议论文集，第 8048–8057 页。

Barry Haddow、Rachel Bawden、AntonioValerio Miceli-Barone、Jindˇrich Helcl 和 Alexandra Birch。2022年。低资源机器翻译综述。计算语言学，48(3)：673–732。

Kaggle。Kaggle：机器学习与数据科学社区。https://www.kaggle.com。访问日期：2026-02-24。

Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B. Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。2020年。神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*。Vladimir Karpukhin、Barlas Oˇguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。面向开放域问答的密集段落检索。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第6769–6781页。计算语言学协会。

Sandeep Kumar、Tirthankar Ghosal 和 Asif Ekbal。2021年。Dataquest：一种从科学论文中自动提取数据集提及的方法。载于迈向开放可信数字社会：第 *23* 届亚太数字图书馆国际会议（*ICADL 2021*），第 13133 卷，收录于计算机科学讲义，第 43–53 页。Springer。

Quentin Lhoest、Albert Villanova del Moral、Yacine Jer- nite、Abhishek Thakur、Patrick von Platen、Suraj Patil、Julien Chaumond、Mariama Drame、Julien Plu、Lewis Tunstall、Joe Davison、Mario Šaško、Gunjan Chh- ablani、Bhavitvya Malik、Simon Brandeis、Teven Le Scao、Victor Sanh、Canwen Xu、Nicolas Patry 等 13 人。2021。《数据集：自然语言处理社区库》。载于《*2021* 年自然语言处理实证方法会议论文集：系统演示》。

Pengyue Li、Sheng Wang、Hua Dai、Zhiyu Chen、Zhifeng Bao 和 Brian D. Davison。2025。《大语言模型时代的开放数据集搜索综述：回顾与展望》。*arXiv*预印本 *arXiv:2509.00728*。

Zehan Li、Xin Zhang、Yanzhao Zhang、Dingkun Long、Pengjun Xie 和 Meishan Zhang。2023年。迈向采用多阶段对比学习的通用文本嵌入。*arXiv* 预印本 *arXiv:2308.03281*。

Rachel Lin、Bhavya Chopra、Wenjing Lin、Shreya Shankar、Madelon Hulsebos 和 Aditya G. Parameswaran。2025。重新思考 DataScout 的数据集发现。收录于第*38*届 *ACM* 用户界面软件与技术年度研讨会论文集（*UIST '25*），大韩民国釜山。计算机协会。

Yang Liu、Jiahuan Cao、Chongyu Liu、Kai Ding 和 Lianwen Jin。2024 年。大语言模型的数据集：一项全面综述。*arXiv* 预印本 *arXiv:2402.18041*。

Patrice Lopez。2009年。GROBID：将自动书目数据识别与术语提取相结合，用于学术出版物。载于第*13*届欧洲数字图书馆研究与先进技术会议（*ECDL*）论文集，计算机科学讲义第5714卷，计算机科学讲义，第473–474页。Springer。

Pietro Marini、Aécio Santos、Nicole Contaxis 和 Ju- liana Freire。2025 年。数据收集器：利用大语言模型从科学文献中提取数据集引用。载于第五届学术文档处理研讨会（SDP 2025）论文集，第114–123页，奥地利维也纳。计算语言学协会。

Natasha Noy、Matthew Burgess 和 Dan Brickley。2019年。Google Dataset Search：为开放网络生态系统中的数据集构建搜索引擎。载于万维网大会（WWW）论文集。Huitong Pan、Qi Zhang、Eduard Dragut、Cornelia Caragea 和 Longin Jan Latecki。2023年。DMDD：用于数据集提及检测的大规模数据集。《计算语言学协会汇刊》，11：1132–1146。Papers with Code。Papers with Code。https:// paperswithcode.com。访问日期：2026-02-24。Norman W. Paton、Jiaoyan Chen 和 Zhenyu Wu。2024年。数据集发现与探索：一项综述。《ACM计算调查》，56(4)：102:1–102:37。Leonard Richardson。Beautiful Soup 文档。https://www.crummy.com/software/BeautifulSoup/。访问日期：2026-02-24。Minjie Shen、Yanshu Li、Lulu Chen 和 Qikai Yang。2025年。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。Xiaoyu Shen、Svitlana Vakulenko、Marco del Tredici、Gianni Barlacchi、Bill Byrne 和 Adrià de Gis- pert。2022年。开放域问答的低资源密集检索：全面综述。预印本，arXiv:2208.03197。

Aivin V. Solatorio、Rafael Macalaba 和 James Liounis。2025年。利用大语言模型和合成数据监测研究论文中的数据集提及。arXiv 预印本 arXiv:2502.10263。Hui Su、Zhi Tian、Xiaoyu Shen 和 Xunliang Cai。2024年。揭开缩放定律之谜：第一部分。arXiv 预印本 arXiv:2403.06563。Hui Su、Xiao Zhou、Houjin Yu、Xiaoyu Shen、Yuwen Chen、Zilin Zhu、Yang Yu 和 Jie Zhou。2023年。WELM：面向中文的博览群书式预训练语言模型。预印本，arXiv:2209.10372。Joaquin Vanschoren、Jan N. van Rijn、Bernd Bischl 和 Luis Torgo。2013年。OpenML：机器学习中的网络化科学。《SIGKDD探索》，15(2)：49–60。

Anjie Xu、Ruiqing Ding 和 Leye Wang。2025年。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5106–5116页，加拿大多伦多。计算机协会。应用数据科学方向。

Manzil Zaheer、Guru Guruganesh、Avinava Dubey、Joshua Ainslie、Chris Alberti、Santiago Ontanon、Philip Pham、Anirudh Ravula、Qifan Wang、Li Yang 和 Amr Ahmed。2020年。Big bird：面向更长序列的 Transformer。载于神经信息处理系统进展。

Tong Zeng 和 Daniel Acuna。2024年。使用双向长短期记忆网络-条件随机场模型抽取科学文章中的数据集提及。*arXiv* 预印本。

Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、RuiyangRen、Yifan Li、XinyuTang、Zikang Liu 等 3 人。2023年。大语言模型综述。*arXiv* 预印本 *arXiv:2303.18223*。

Dawei Zhu、Xiaoyu Shen、Michael Hedderich和Dietrich Klakow。2023a年。面向弱监督稳健学习的元自我精炼。载于第*17*届计算语言学协会欧洲分会会议，第1043–1058页。

Dawei Zhu、Xiaoyu Shen、Marius Mosbach、Andreas Stephan 和 Dietrich Klakow。2023b年。《弱于你的预期：对弱监督学习的批判性审视》。载于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第14229–14253页。

模型训练细节与规模分析

数据准备和超参数 BERT-GATE 和 BERT-DESC 模型均使用表格 4中详述的配置进行了微调。所有实验均在单个 NVIDIA RTX 3090（24GB）图形处理器上进行。

对于 BERT-GATE，分类器使用 7,789 个实例进行训练，并按 10:3:3 的比例划分为训练集、验证集和测试集，同时严格确保所有划分中的正负样本比例为 1:2.8。对于 BERT-DESC，模型基于由 179 篇论文和 21,297 个句子构成的精细标注语料库进行训练，并在包含 2,286 个句子的留出测试集上进行评估。此外，BERT-DESC 使用掩码均值池化，随后通过一个两层多层感知机（线性层–层归一化–丢弃 0.3–高斯误差线性单元–线性层）生成最终的句子级逻辑值，其交叉熵损失配置了 ignore_index $= -100$。

超参数	**BERT-GATE**	**BERT-DESC**
最大序列长度	512	512
优化器	AdamW	AdamW
学习率	3×10^{-5}	2×10^{-5}
权重衰减	0.1	0.01
批量大小	16	1（有效：4）
评估策略	每 100 步	基于轮次
模型选择指标	验证损失	验证召回率

表4：两个基于BERT的模块的训练配置和超参数。

上下文窗口优化（BERT-DESC） 为了实现描述提取器的类别条件滑动窗口，我们以目标句子为中心，使用包含 ± 2 个句子的对称窗口，并向外扩展，直到达到 512 个词元的上限。窗口在正标签附近以较小的步幅（$\lfloor$ 个窗口 $/3\rfloor$，最小为 1）滑动，在其他位置则以较大的步幅（$\lfloor$ 个窗口 $/2\rfloor$）滑动。这一具体的工程设计选择丰富了实际数据集描述的局部上下文，同时大幅降低了冗余负跨度的权重。

数据扩展分析 为了评估两种架构的数据效率，我们比较了它们在规模不断增大的训练数据子集上的表现。

如图2(a)所示，BERT-GATE 的 F1分数随着数据集大小的增加而单调上升，从 0.812（对应 2,030 个样本）提升至 0.945（对应 4,361 个样本）。这表明，即使在较低的资源水平下，该检测器也具有很高的数据效率，并且完全能够满足我们将其用作轻量级过滤器的需求。

同样，如图2(b)所示，随着数据集规模从小到中等不断增大，BERT-DESC 的四项主要指标均迅速提升。随后，随着训练数据规模接近我们的完整训练规模，各曲线明显趋于平缓，这表明句子级分类器目前已处于近乎饱和的状态，增加标注数据带来的性能提升将微乎其微。

B 基于规则的数据集网址评分

给定候选URL u、其锚文本 a 以及局部上下文 c（超链接前后的各两个句子），我们通过对少量可解释特征求和，计算标量分数 score(u)。最终分数为

$$\text{score}(u) = \sum_i w_i f_i(u, a, c),$$

其中每个 $f_i \in \{0, 1, 2, \dots\}$ 都是手工设计的指标或计数特征，w_i 是一个较小的整数权重。表 5总结了主要特征组及其代表性条件。

B.1 主URL选择模式

为适应不同的计算约束和准确率要求，我们的系统支持三种不同的最终链接选择运行模式：

1. **纯规则模式**。 此模式完全依赖上述启发式评分系统。将候选项按分数降序排列：(u_1, s_1)，(u_2, s_2)…。我们采用以下选择逻辑：

• 高置信度接受：如果只有一个候选项，我们就返回该项。否则，当 $s_1 \geq \tau$ 较高，或（$s_1 \geq \tau$ 中等 $\wedge$ $s_1 - s_2 \geq \Delta$）时，我们接受 u_1，其中τ较高=22、τ中等=16，且 Δ=5。

• 首选托管平台平局决胜：如果仍无法确定，我们会检查排名前K 的候选项（K=5），并优先选择托管在以数据集为主的平台上的候选项（例如 Hugging Face、Zenodo）。如果仍然平局，则优先选择落地页而非直接文件链接。

•一般平局决胜与拒绝：如果找不到首选主机的 URL，我们会选择

分组	条件	权重
主机$^{+}$	Hugging Face数据集 位于主机+路径中	+10
	Zenodo记录 ，Kaggle数据集	+8–9
	Figshare ，Dataverse ，OSF.IO	+7–8
主机$^{-}$	ARXIV.ORG ，DOI.ORG ，ACM.ORG ，	-8 至 -10
	IEEEXPLORE.IEEE.ORG ，SCHOLAR.GOOGLE.*	
	RESEARCHGATE.NET ，MEDIUM.COM	-6
路径提示	路径包含 /DATASET ，/DATASETS	+3
	路径包含 /DATA ，/下载 ，	+2
	/文件 ，/记录 ，/发布版本	
文件扩展名	以 .CSV 结尾 ，.TSV，.JSON ，	+6
	Parquet	
	以 .ZIP、.TAR(.GZ)、.TGZ 结尾,	+5
	.XZ, .7Z	
	以 .RAR 结尾	+4
词汇$^{+}$	在 (A, C) 中：“数据集”、“我们的数据集”,	+2 每个
	“我们发布”、“可在.....获取”	(上限 +8)
词汇$^{-}$	在 (A, C) 中：“代码”、“源代码”、	-2 每个
	“实现”、“BibTeX”	（大写 -6）
特殊$^{-}$	在（A, C）中：“我们在.....上评估”和“数据集”共同出现	-3
GitHub 惩罚	不含 /RELEASES 或 /DATA 路径的 GitHub.com 仓库根目录	-4
	中的 /RELEASES 或 /DATA 路径	

表 5：用于候选数据集网址评分的可解释整数权重特征。

得分最高的候选项，并在得分相同时优先选择较短的 URL。如果最佳得分$s_1 < \tau_{\text{min}}$（其中 τ_{min}=15），我们会将该记录标记为“没有可靠的数据集链接”。

尽管这种纯粹基于规则的方案效率很高，但偶尔会对复杂情况分类错误。

2. **仅大语言模型模式**。 此模式完全绕过基于规则的评分，并将候选项选择交由大语言模型（LLM）完成。它优先考虑语义理解而非提取速度，依靠大语言模型理解每个链接的上下文。

3. **混合模式（规则 +大语言模型）**。这是我们的默认方案，也是最稳健的方法。它利用基于规则的系统作为第一阶段筛选器，根据负分排除明显不是数据集链接的链接（例如，通用学术门户）。随后，将剩余的高质量候选项及其精简的语义上下文传递给大语言模型。大语言模型作为最终分类器，用于可靠地确定主数据集URL。如果大语言模型被禁用或返回不确定的结果，系统会无缝回退到仅使用规则逻辑得出的最高分候选项。

实际上，基于规则组件的整数权重和阈值是通过人工检查 arXiv 论文开发集进行调整的。这种模块化设计确保整个流程易于调试，同时保持较高的抽取精确率。