# DataParasite 实现可扩展且可重复利用的在线数据整理

Mengyi Sun[1*]西蒙斯定量生物学中心、冷泉港实验室，冷泉港纽约州 11724 *通讯作者：msun@cshl.edu

**摘要**

计算社会科学中的许多问题都依赖于从异质的在线来源汇集而成的数据集，而这一过程往往耗费大量人力、成本高昂且难以复现。大语言模型的最新进展使得智能体搜索和从网络中进行结构化提取成为可能，但现有系统往往不透明、不灵活，或不适用于科学数据整理。在此，我们介绍 DataParasite，这是一种开源、模块化的处理流程，可用于大规模在线数据收集。DataParasite 将表格数据整理任务拆分为彼此独立的实体级搜索，这些搜索通过轻量级配置文件定义，并由一个通用的 Python 脚本执行。重要的是，同一处理流程还可通过自然语言指令改用于新任务，包括那些没有预定义实体列表的任务。我们在计算社会科学中的多项典型任务上评估了该处理流程，包括教师招聘历史、精英人物死亡事件和政治生涯轨迹。对于各项任务，DataParasite 均实现了较高的准确率，同时将数据收集成本降低了一个数量级，相较于人工整理尤为如此。通过降低在线数据汇编所需的技术和人力门槛，DataParasite 为计算社会科学及其他领域中可扩展、透明且可复用的数据整理提供了切实可行的基础。

## 引言

数据是科学进步的基础。在许多学科中，范式转变并非仅仅源于新理论的提出，也得益于新型数据的出现以及大规模分析这些数据的能力。从基因组学[1] 和天文学[2] 到神经科学[3] 和经济学[4], ，大量系统收集的数据集不断涌现，改变了研究人员能够提出的问题，也改变了他们可以采用的方法。

随着计算社会科学的兴起，社会科学也开始经历类似的转变[5]。在大规模数字痕迹、行政记录和在线数据的推动下，这一领域开辟了研究不平等、社会结构、集体行为和制度动态的新途径[6]。许多支撑这项研究的数据集并非以可直接使用的形式发布，而是必须从组织网站等在线来源[7] ，、公共记录[8] ，和半结构化文档[9]中汇集而来。实际上，收集此类数据往往依赖人工整理或临时搭建的网页抓取处理流程，因此耗费大量人力和成本，也难以扩展或复现。

大语言模型（LLM）的最新进展为自动化在线数据收集提供了可能的途径。现代基于大语言模型的智能体能够搜索网络[10], 、推理异构来源[11], ，并通过上下文学习提取结构化信息[12]。尤其是，支持多轮查询和迭代优化的新兴智能体搜索形式，相较于单次检索取得了显著进步[13]。尽管前景可期，但这些能力与其在计算社会科学中的实际应用之间仍存在差距。现有系统往往是闭源、不透明的，或被设计成难以针对新任务进行调整、审计或重新利用的单体式长周期智能体。

在此，我们介绍 DataParasite：一个开源、透明的处理流程，可实现可扩展且可复用的在线数据整理。DataParasite 源于一个简单的观察：计算社会科学中许多感兴趣的数据集天然呈表格形式，每行对应一个实体（例如个人、机构或事件），每列则表示一项待整理的属性。这种结构使在线数据收集能够分解为针对各个实体、彼此独立且可并行执行的搜索。

DataParasite 通过模块化设计实现了这一理念（图 1）。图 1A 展示了核心处理流程：任务专属的 YAML 文件定义模式、提示词和模型设置，而 CSV 文件则指定待整理的实体。这些输入会传递给一个独立且与任务无关的执行 Python 脚本，该脚本会启动针对实体的并行数据整理任务，并将生成的结构化输出汇总为统一的整理后数据库。图 1B 展示了如何通过智能体驱动的配置复用该处理流程。用户向现成的编码代理发送自然语言请求，以启动一项新任务；该代理由一份简明的代码库操作手册提供指导。根据是否提供实体列表，代理会直接生成任务专属的 YAML 配置，或者先识别相关在线来源并自动构建实体列表。在这两种情况下，生成的配置文件都可直接接入图 1A 所示的同一执行引擎。

我们通过一系列源自计算社会科学的在线整理任务，展示了 DataParasite 的实用性。我们表明，该处理流程能够准确且经济高效地构建常用于研究不平等和分层的教师招聘数据集，同时大幅减少数据收集所需的人力。我们进一步证明，仅凭自然语言指令，同一处理流程即可重新用于有预定义实体列表和无预定义实体列表的不同任务，并在两种情况下都达到完美准确率。总之，这些结果说明，DataParasite 降低了开展可扩展在线数据收集的门槛，并为计算社会科学中的数据驱动研究提供了切实可行的基础。

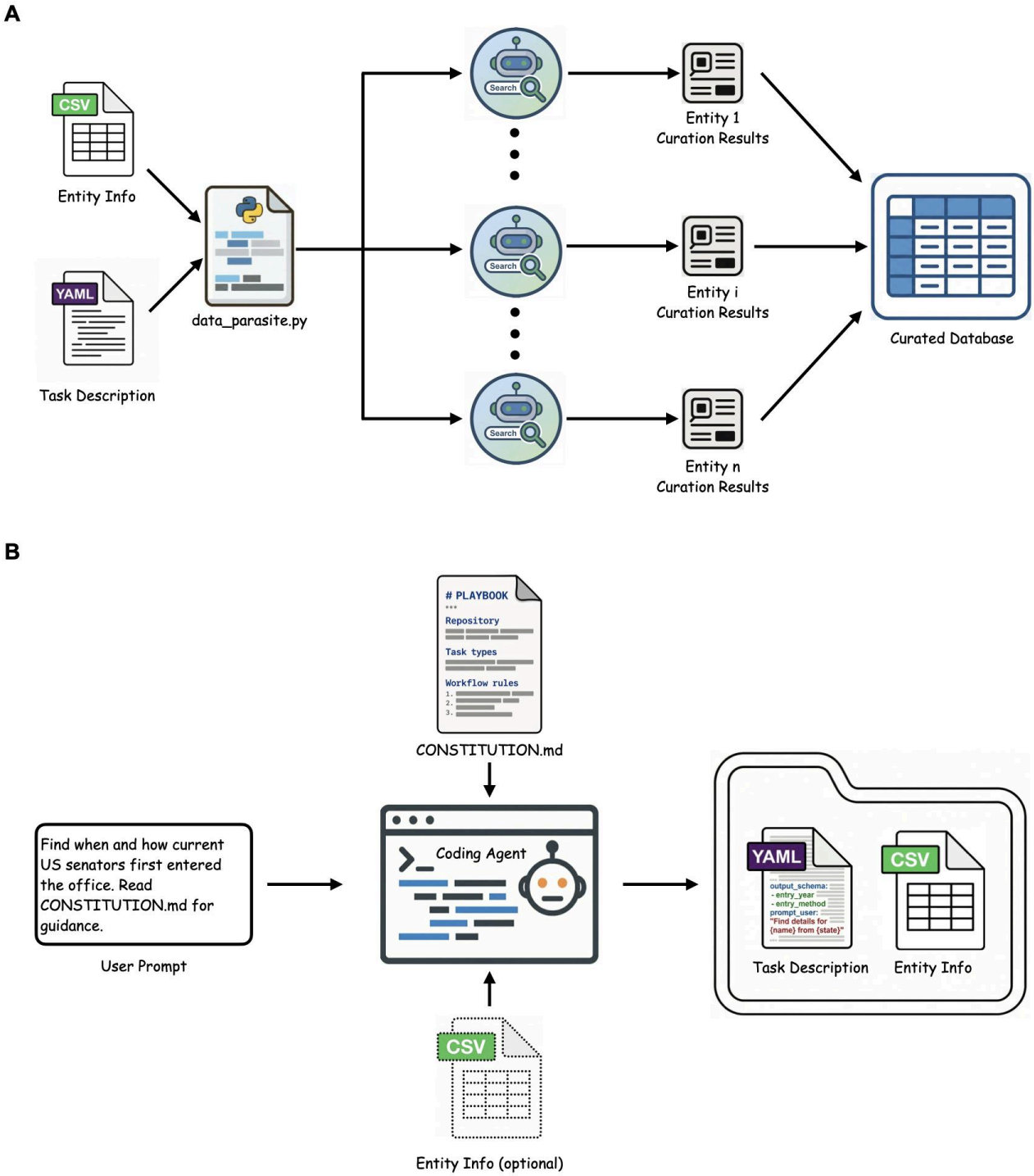



**图 1. DataParasite 用于可扩展且可复用的智能体辅助数据收集的工作流程。** **(A)** 核心处理流程。任务专用的 YAML 文件（“任务描述”）定义模式、提示词和模型设置，CSV 文件（“实体信息”）则列出待整理的实体。这些输入会传递给 data_parasite.py，由其启动并行的大语言模型驱动的数据整理任务——每个实体对应一个任务——并将生成的结构化输出汇编为统一的整理后数据库。**(B)**通过智能体驱动的配置实现可复用性。用户可以向编码代理发出自然语言请求，并指引其遵循代码库操作手册（CONSTITUTION.md），从而启动新任务。该代理可以通过**两种方式**复用处理流程：*(i)*提供实体 CSV 文件时，生成相应的 YAML 配置；或*(ii)*未提供实体列表时，定位相关在线来源，自动构建实体 CSV 文件，然后生成 YAML 描述文件。在这两种情况下，生成的轻量级配置文件都可以直接接入 (A) 所示的同一引擎，从而支持在异构数据收集任务之间快速复用该处理流程。

## 结果

我们在一组反映计算社会科学中常见瓶颈的在线数据整理任务上，评估了 DataParasite 处理流程的性能、成本和灵活性。这些任务要求从异构网络来源中大规模提取结构化信息，而相关信息可能不完整、报告方式不一致，或难以通过自动化方式定位。

### 教师招聘历史的可扩展在线整理

在计算社会科学领域，对机会、资源和地位获取不平等的关注贯穿始终。教师招聘已成为研究这些动态的典型实证视角，因为博士培养和学术职位安置的模式，为专业劳动力市场中的分层与流动提供了具体且广为确立的实例[7]。与此同时，整理教师招聘数据集仍然是一个持续存在的数据瓶颈，通常需要跨异构的院系网站和个人网站进行人工整理或定制化网页抓取。

为评估 DataParasite 能否缓解这一瓶颈，我们为随机抽取的 100 名教师整理了教师招聘历史，从在线来源中提取三个属性：博士院校、博士毕业年份和首个终身教职轨道招聘院校（图 2A）。我们在两种规则下评估准确性：将“未找到”响应视为错误，以及仅将成功返回“已找到”结果的条目纳入评估。

对于这三项属性，即使将“未找到”的回答视为错误，DataParasite 仍取得了较高的准确率；若分析仅限于已找到的条目，准确率则进一步提高（图 2A，左）。博士院校和博士毕业年份的准确率最高，首次聘用院校的准确率也保持在较高水平。这种区分体现了一种切实可行的人机协作模式：自动整理能够以较高可靠性处理大多数条目，而少量尚未解决的案例则可交由人工针对性核验，从而进一步提高端到端准确率。

除准确率外，DataParasite 还大幅降低了边际数据收集成本。对于每条教师记录（即整理全部三项属性的成本），成本分布集中在典型人工劳动费率远低于的水平（图 2A，右）。每条教师记录的平均成本约为 0.09 美元，而类似在线整理任务中具有代表性的研究助理成本约为每条记录 0.66 美元（估算详情见**方法**）。除了节省成本，该处理流程还支持对大量记录进行并行处理（受 OpenAI 速率限制约束），因此完成整项整理任务的速度可远快于研究助理。总之，这些结果表明，DataParasite 能够准确且经济高效地构建教师职位安置数据集，而此类数据集过去通常需要投入大量人工。

### 有无输入实体列表时跨任务的复用能力

接下来，我们评估了 DataParasite 的灵活性：在不修改执行引擎的情况下，通过向编码代理发出单次自然语言指令，将同一核心处理流程改用于另外两项在线数据收集任务（图 2B）。这些任务是计算社会科学和政治学中的典型案例，在这类研究中，离散事件常被用作准外生冲击，以研究长期结果。在这两种情况下，复用都是通过让编码代理重写特定任务的 YAML 配置实现的；如果未提供实体列表，代理还会查找相关在线来源，构建相应的 CSV，然后生成 YAML 描述文件。

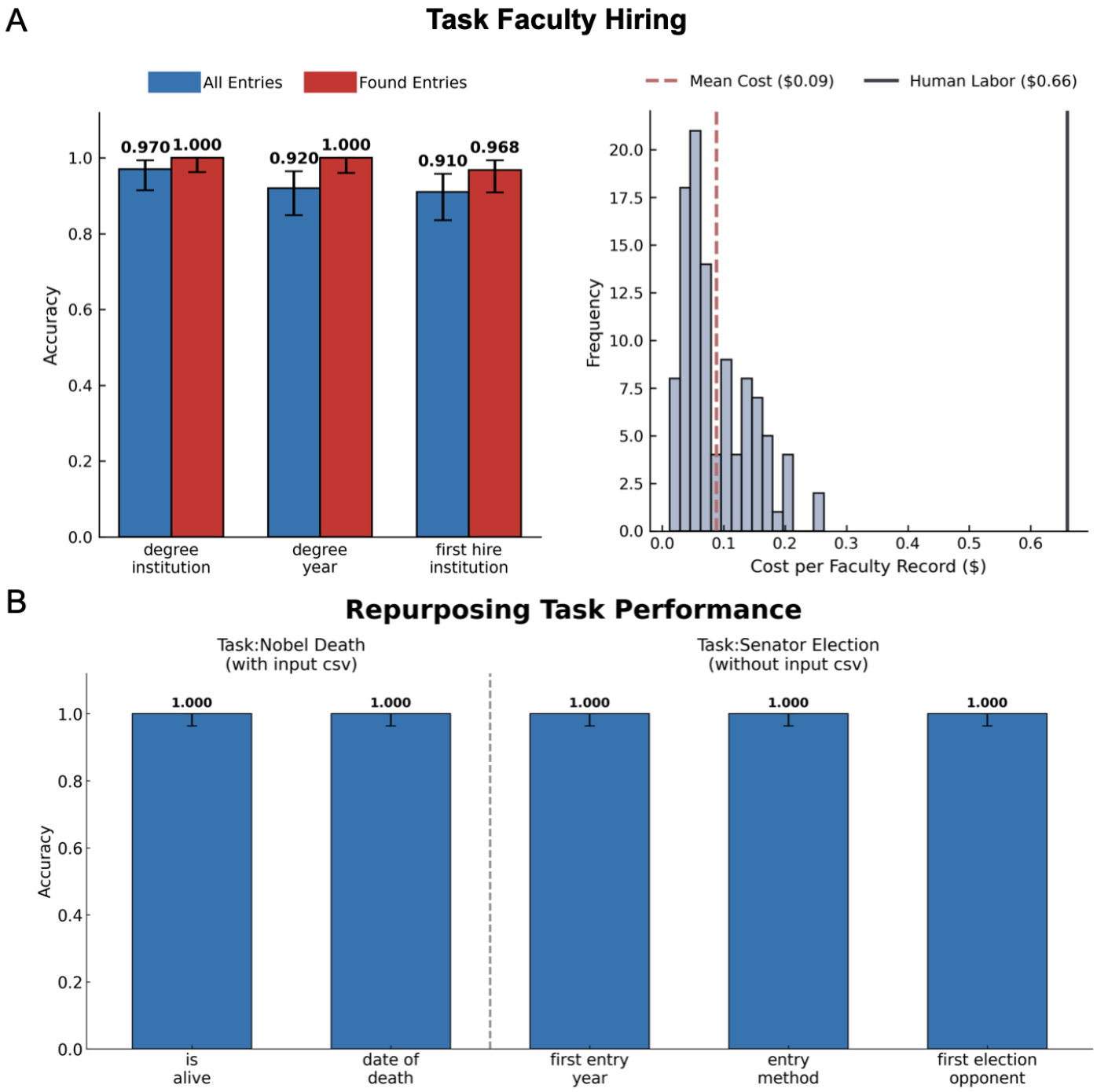



**图2. DataParasite 在教师招聘任务及再利用任务中的准确率和成本。（A）**针对100名教师的教师招聘任务。左侧面板展示了三项属性（学位授予院校、学位年份、首次聘用院校）在线整理的准确率，并对比了将“未找到”条目计为错误的评估结果（蓝色）与仅针对已找到条目的评估结果（红色）。柱形显示点估计值及其95% Clopper–Pearson 精确置信区间，柱形上方标注了数值。右侧面板展示了每条教师记录的成本分布，该成本定义为整理一名教师的全部三项属性所需的总成本。虚线表示 DataParasite 的平均成本（每条记录0.09美元），实线表示一名具有代表性的人工研究助理的成本（每条记录0.66美元）。**（B）**再利用任务表现（每项任务100个实体）。左侧面板（使用输入CSV文件）展示了整理诺贝尔奖得主是否在世，以及若已去世则整理其死亡日期的准确率。右侧面板（不使用输入CSV文件）展示了整理每位现任美国参议员的首次入职年份、入职方式和首次选举对手的准确率。柱形显示点估计值及其95% Clopper–Pearson 精确置信区间。

**任务1（带输入CSV文件）**聚焦于高影响力人物的死亡事件，这类事件被广泛用于研究“伟大人物”对同伴效应[14], 创新方向[15]以及制度轨迹[16]的因果影响。给定一份包含100位随机抽取的诺贝尔奖得主的输入CSV文件后，编码代理生成了一份 YAML 配置，用于在线整理每位个人是否在世；若已去世，则整理其死亡日期。

**任务2（不带输入CSV文件）**代表了一种更具挑战性的情境，因为在开展任何整理工作之前，必须先在线识别相关实体集合。首次进入立法职务是一项重要的政治事件，不仅会影响个人职业生涯[17] ，也会影响所代表的选民[18], ，因为早期选举结果可能通过现任优势，造成政治代表性方面持久的差异[18]。在此任务中，编码代理首先通过在线来源识别出美国现任参议员，并自动构建实体列表，随后生成 YAML 配置，用于整理每位参议员的首次入职年份、入职方式和首次选举对手。

尽管任务复杂度有所提升，DataParasite 在两项任务的所有查询属性上均实现了完美准确率（图 2B）。由于这两项任务涉及的事实在网上都有广泛记录，因此一个潜在问题是，高准确率可能源于模型记忆，而非真正的在线整理。为排除这一可能性，我们进行了一项基线评估：要求用于整理的同一模型（gpt-5-mini）仅使用其内部知识回答相同问题，并明确禁用网络搜索。在此设置下，模型性能显著下降。在诺贝尔奖得主死亡任务中，判断一位得主是否在世的准确率为 44%，而对于所有死亡日期查询，模型均返回未¯找到。在参议员选举任务中，对于所有查询属性，模型同样均返回未¯找到。这些结果表明，DataParasite 的出色表现源于主动进行在线数据收集与验证，而不是依赖语言模型中嵌入的记忆知识。

## 讨论

DataParasite 提供了一种实用工具，用于开展可扩展的在线数据整理，并优先考虑易用性、可复用性和低开销。在多项任务中，该处理流程能够从异构网络来源中准确且高效地提取结构化信息，无需编写特定于任务的代码或定制网页抓取处理流程。通过使用简单的配置文件描述数据收集任务，并通过共享处理流程执行这些任务，DataParasite 使研究人员能够快速从研究问题推进到可运行的数据收集工作流，并以最少的工作量将该工作流调整到新的任务中。

一些局限性为未来改进指明了方向。首先，尽管 DataParasite 完全开源，但当前实现依赖商业语言模型 API，这可能限制可访问性，并使中间推理轨迹变得不透明。支持开源模型[19,20], 包括可部署在边缘设备上的模型[20–22]将提高透明度并扩大使用范围。其次，网络搜索目前是成本的主要驱动因素（占成本的 >90%），构成了总支出的大部分。虽然内置搜索工具具有稳健性，但整合无头浏览等本地检索机制可以显著降低成本，不过速率限制和反爬虫措施带来的实际约束也必须考虑。第三，当前处理流程将实体视为相互独立的搜索单元，尽管相关信息通常分布在多个实体之间。利用这种结构可以提高效率和准确性，但需要采用新的方法来协调搜索和证据复用。最后，虽然我们的示例集中于计算社会科学，但许多科学领域也存在类似的在线整理挑战，包括生物学[23], 医学[24], 和化学[25], ，因此将 DataParasite 推广到更广泛的应用领域是顺理成章的下一步。

通过让在线数据整理更易于设置、运行成本更低且更易于复用，DataParasite 降低了数据驱动研究的门槛，使研究人员能够专注于实质性问题，而不是数据汇编。

## 方法

### Data

对于教师招聘任务，我们使用 CSRankings 数据集（https://github.com/emeryberger/CSrankings/tree/gh-pages），从美国、中国、英国和德国随机抽取 100 名计算机科学家，构建输入CSV文件。对于每位个人，我们从在线来源手动整理其真实的博士学位授予机构、毕业年份和首个聘用机构，并将这些信息用于评估 DataParasite 的整理结果。

对于诺贝尔奖得主死亡任务，诺贝尔奖得主名单及相关奖项信息取自 Kaggle 数据集（https://www.kaggle.com/datasets/imdevskp/nobel-prize）。我们随机抽取了 100 名诺贝尔奖得主，并从在线来源手动整理其生存状态（在世或已故），以及适用情况下的死亡日期。这些真实记录随后与 DataParasite 的整理结果进行了比较。

对于参议员选举任务，事先未提供输入名单。相反，编码代理（我们使用了 Cursor-Agent；使用 Codex 进行的测试得出了类似结果）会自动从在线来源检索美国参议员及其对应州的当前名单，并生成后续任务的 YAML 文件。我们根据参议院官方名录（https://www.senate.gov/general/resources/pdf/senators_phone_list.pdf）手动核实检索到的参议员名单，并将整理的职业信息与维基百科进行了交叉核对。

### DataParasite 的实现

DataParasite 使用 OpenAI Python SDK 实现。除正文中介绍的核心逻辑外，该处理流程还会自动记录遥测数据，包括 token 使用量、网页搜索工具调用次数及相关成本。本实现支持在不同的 OpenAI 模型之间切换；不过，本研究中的所有分析均使用 gpt-5-mini 完成。DataParasite 的源代码可在以下网址获取：https://github.com/mengysun/DataParasite。

### 人工劳动成本估算

人工劳动的平均成本估算基于一项近期且任务复杂度相近的研究，该研究的任务是确定给定作者的性别和获得博士学位的年份[26]。该研究报告称，研究助理每小时可整理约 30 条记录；这与我们在线手动整理真实标注数据时的经验一致。假设美国研究助理的典型时薪约为 20 美元，则估算每条记录的成本为 0.66 美元（20 美元/30）。

### 用于重新利用任务的提示词

诺贝尔奖得主死亡任务的提示词如下：

“我已在任务目录下的文件夹 `nobel_winner_death` 中整理了一份诺贝尔奖得主名单。你的任务是判断每位个人是否仍然在世；如果已经去世，请确定其逝世时间。在启动任务之前，请务必仔细阅读 `@CONSTITUTION.md` 文件。”

参议员选举任务的提示词如下：

“收集所有现任（2025年）美国参议员的数据，包括他们首次进入参议院的年份（经任命或选举），以及若经选举进入参议院，他们在首次成功当选时击败的最后一位对手。为这项任务命名 `senator_election`。请在启动任务前仔细阅读`@CONSTITUTION.md` 。”

**参考文献**

1. Gates, A. J., Gysi, D. M., Kellis, M. & Barabási, A.-L. 基于人类基因组计划取得的丰硕发现——数字一览。《自然》*Nature* **590**, 212–215 (2021). 2. Edwards, K. J. & Gaber, M. M. 天文学与大数据：一种用于识别不确定星系形态的数据聚类方法. (施普林格国际出版, 施普林格, 瑞士, 2014). 3. Helmstaedter, M. 突触分辨率连接组学：迈向大脑规模扩展与连接组筛查。《自然综述：神经科学》1–20 (2025). 4. Nagaraj, A. & Tranchero, M. 数据访问如何塑造科学？联邦统计研究数据中心对经济学研究的影响. https://www.nber.org/system/files/working_papers/w31372/w31372.pdf (2023). 5. Lazer, D. 等 计算社会科学。《科学》 **323**, 721–723 (2009). 6. Salganik, M. J. 一点一滴：数字时代的社会研究。(2019). 7. Clauset, A., Arbesman, S. & Larremore, D. B. 教师招聘网络中的系统性不平等与等级制度。*Sci. Adv.* **1**, e1400005 (2015). 8. Jin, C., Ma, Y. & Uzzi, B. 科学奖项与科学主题的非凡增长。《自然·通讯》**12**, 5619 (2021). 9. Gross, D. P. & Sampat, B. N. 美国的腾飞：第二次世界大战期间的研发与美国创新体系的崛起。《美国经济评论》 **113**, 3323–3356 (2023). 10. Nakano, R. 等 WebGPT：借助浏览器并结合人类反馈的问答。*arXiv [cs.CL]*(2021). 11. Wei, J. 等 思维链提示能够激发大语言模型的推理能力。*arXiv [cs.CL]* (2022). 12. Brown, T. B. 等 语言模型是少样本学习器。*arXiv [cs.CL]* 1877–1901 (2020). 13. Trivedi, H., Balasubramanian, N., Khot, T. & Sabharwal, A. 将检索与思维链推理交替结合，以应对知识密集型多步问题。*arXiv [cs.CL]* (2022). 14. Azoulay, P., Zivin, J. S. G. & Wang, J. 巨星消亡*. 《经济学季刊》 **125**, 549–589 (2010). 15. Azoulay, P., Fons-Rosen, C. & Zivin, J. S. G. 科学是否一次葬礼向前迈进一步？《美国经济评论》

**109**, 2889–2920（2019）。16. 琼斯，B. F. 和奥尔肯，B. A. 领导者重要吗？第二次世界大战以来的国家领导力与增长。《*SSRN* 电子期刊》（2004）doi:10.2139/ssrn.524042。17. 卡罗，R. A.《崛起之路：林登·约翰逊的岁月 II》。**2**，（2011）。18. 福勒，A. 和霍尔，A. B. 选举结果的长期影响。《英国政治学杂志》 **47**，351–372（2017）。19. 郭，D. 等。 DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》 **645**，633–638（2025）。20. 杨，A. 等。 Qwen3 技术报告。*arXiv [cs.CL]* （2025）。21. OpenAI 等。 gpt-oss-120b 和 gpt-oss-20b 模型卡。*arXiv [cs.CL]* （2025）。22. Gemma团队 等。 Gemma 3 技术报告。*arXiv [cs.CL]* （2025）。23. 阿什伯纳，M. 等。基因本体：实现生物学统一的工具。基因本体联盟。《自然遗传学》 **25**，25–29（2000）。24. Landrum, M. J. 等。 ClinVar：序列变异与人类表型之间关系的公共档案。《核酸研究》 **42**，D980–5（2014）。25. 本托，A. P. 等。使用 RDKit 的开源化学结构整理流程。《化学信息学杂志》 **12**，5（2020）。26. 李，L.、陈，Y.、莱文斯坦，M. C. 和维尔胡伯，L. 众包数字公共产品：关于元数据贡献的田野实验。社会科学研究网络（2024）doi:10.2139/ssrn.5008203。