Trallie：使用大语言模型将自由格式文本转换为结构化数据

Vijayasri Iyer[1], Cristiano De Nobili[1,3],Mubashir Hussain Shah[1],Sébastien Bratières[1,2][1]Pi School, [2]Translated, [3]Critiqality

arXiv:2611.30773v1 [cs.CL] 23 Nov 2026

摘要

TRALLIE 是一个开源框架，利用大语言模型（LLM）从自然语言来源中执行信息抽取（IE），可选择是否接受用户指导。不同于依赖大量标注数据集或领域特定训练的传统方法，TRALLIE 能够在仅需极少或无需代表性示例的情况下，将非结构化数据转换为结构化格式。该框架有三个主要目标：（i）利用现代大语言模型基于提示的零样本和少样本能力，减少人工标注；（ii）灵活解析多种文档格式（即 HTML、PDF、TXT 和 JSON）；（iii）通过从文档集合中自动推导模式，并据此完善抽取流程，提高非专业用户的可访问性。我们在 FDA 510(k) 文件以及 SWDE 中的电影和大学网站上，使用开放权重大语言模型评估了 TRALLIE 。当为文本抽取提供目标属性时，其性能优于此前报告的系统；当必须自动发现模式所需的属性时，其表现也具有竞争力。项目地址：https://github.com/PiSchool/trallie。

1 引言

医疗保健、法律、金融和科学出版等领域的大量有价值信息，仍然封存在非结构化或半结构化文档中。这类文档难以进行高效搜索、查询和自动化分析，从而为知识发现和决策制造障碍。将这些内容转换为结构化、可查询格式，是推动数据驱动应用发展的基础。

传统的信息抽取（IE）方法面临两大主要局限。首先，它们依赖大量手动模式设计和经过繁复工程化处理的处理流程，因此开发成本高昂，开发起来既耗时又难以适应新领域。其次，它们依赖固定的标注模式；当需要自定义模式时，则必须进行成本高昂的重新标注和重新训练。即使采用最先进的模型，其迁移学习收益仍然有限。在资源的传统自然语言描述中，这些局限尤为明显，因为结构化信息仍嵌入其中

叙事文本中。现有的基于大语言模型的工具也需要使用者具备一定专业知识：用户必须为每项新任务编写提示、设计目标模式或提供带标注的示例。许多工具还属于闭源工具，或仅限于使用专有模型。

为应对这些挑战，我们推出了 TRALLIE。这是一种轻量且模块化的开源框架，利用大语言模型从文档集合中自动推导模式，并使用提取出的属性值填充这些模式。我们的方法只需极少的人工标注即可简化信息提取，同时兼容多种文档格式，并让非专业用户也能使用。TRALLIE 遵循三项核心原则：(1) 通过经过良好训练的大语言模型进行少样本学习，最大限度减少人工标注；(2) 支持 HTML、PDF、TXT 和 JSON，具备格式与语言灵活性；(3) 从文档集合中自动推断模式，而非要求用户具备领域专业知识，从而方便非专业用户使用。

我们的主要贡献如下。首先，我们提出了一个用于从非结构化文档集合中进行联合模式生成与抽取的统一框架，减少了对手动模式设计的需求。其次，我们引入了一种与提供商无关的架构，无需修改代码即可在商业大语言模型 API（例如 OpenAI 和 Groq）、自定义端点（例如 Hugging Face）和本地模型（例如 Ollama）之间切换；同时，通过重叠分块与基于大语言模型的整合，实现了长文档处理，

并支持 JSON 输出验证和多格式文档处理。第三，我们提供了完整的开源实现，包括 Python 软件包、评估脚本，以及用于模式发现和数据抽取的交互式 Streamlit 界面。

2 TRALLIE

TRALLIE 通过两阶段流水线运行：先进行模式生成，再进行数据抽取。**模式生成器** 从非结构化文档集合中识别相关属性，并生成模式。随后，**数据提取器**利用该模式从所有文档中提取属性值，并输出结构化 JSON 对象。通过将这两个阶段解耦，TRALLIE 支持灵活采用其中任一组件或同时采用两者，从而能够以极少的人工工作轻松适应新的领域。

2.1 模式生成

模式生成器通过三个步骤识别能够捕捉文档集合结构的属性：抽样、属性提取和聚合。下面的代码展示了使用 TRALLIE 框架进行模式生成的简单流程。

```
来自trallie 导入 模式生成器

记录 = ["数据/测试_数据_1.txt",
        "数据/测试_数据_2。PDF"]

模式_生成器 = 模式生成器（
    提供商 ="OpenAI"，
    模型_名称 ="gpt-4o"）

模式 = 模式_生成器。
    发现_模式（
    描述 ="美国食品药品监督管理局 510(k)许可
        文件"，
    记录 =记录）
```

步骤 1：抽样。 为降低计算成本，模式发现在集合中的一个小型子集上运行，而不是针对每个文档运行。默认使用前 10 个文档。

步骤 2：属性提取。 对于每个抽样文档，大语言模型会提出一组候选属性，用于捕捉文档的信息结构（例如，对于一篇科学论文，候选属性包括“标题”“作者”“摘要”“方法论”和“结论”）。

步骤 3：模式聚合。系统会聚合所有抽样文件中的候选属性，并选出出现频率最高的前 k 个属性（k = 20默认为）。这样即可生成一个紧凑的模式，反映文档集合的主要结构。

2.2 数据抽取

数据提取器接收一个文档集合和一个模式（由模式生成器生成或由用户提供），并输出结构化的 JSON，每个文档对应一个 JSON 对象。以下代码展示了使用 TRALLIE 框架进行数据提取的简单流程。

```
来自 trallie 导入 DataExtractor

记录 = ["data/test_data_1.txt",
        "数据/测试_数据_2.PDF"，]

模式 = ["标题" ，"作者" ，"日期"]

数据_提取器 = DataExtractor(
                提供商 ="OpenAI" ，
                模型_名称 ="gpt-4o")

针对 记录 在 记录:
    抽取_JSON = 数据_提取器。
        抽取_数据（模式，记录）
```

步骤 1：文档解析。系统会遍历每份文档，并根据其格式（即 PDF、HTML、TXT 或 JSON）将其转换为纯文本。

步骤 2：属性值提取。大语言模型接收文档文本及模式，并返回一个 JSON对象，其中包含每个属性的值（例如，给定一个包含“标题”“作者”和“出版_日期”属性的模式，系统会在文档中定位这些字段并抽取其值）。值会原样复制自文档，而文档中不存在的属性则返回空值。响应会被解析为 JSON；如果输出不是有效的 JSON，则会重试请求。

2.3 数据处理

TRALLIE 支持多种文档格式，包括 HTML、PDF、TXT 和 JSON。对于 PDF 等二进制格式，TRALLIE 内置了基于 PyPDF2 的解析器，用于文本抽取。

2.4 长文档处理

对于长度超过可配置限制的文件，TRALLIE 会将文本拆分为重叠文本块

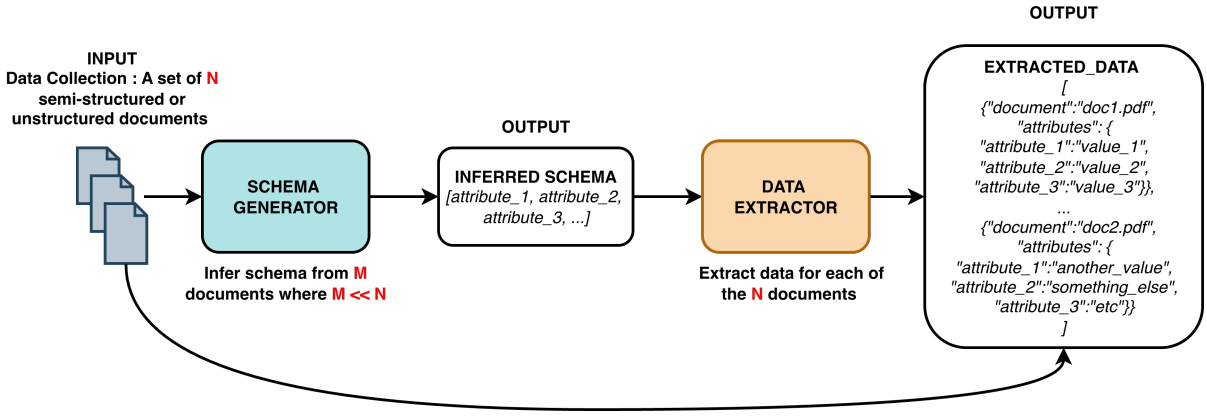


图 1：用于模式推断和数据抽取的两阶段流水线。

（默认每块 10 万个字符，重叠 1 万个字符，并尽可能在句末断开），从每个文本块中提取信息，然后通过最后一次大语言模型调用汇总各块的结果。

2.5 多种大语言模型后端支持

TRALLIE 的一项关键设计原则是大语言模型提供商独立性。该框架实现了统一的大语言模型接口，对不同提供商进行抽象，使用户只需通过一个配置参数，即可在商业服务（OpenAI、Groq）、自定义部署（Hugging Face Inference Endpoints）和本地模型（Ollama）之间切换。这一设计考虑了实际部署场景中用户可能面临的不同限制：API 费用上限、数据隐私要求、互联网连接状况或机构政策。

此外，TRALLIE 会将每个模型响应解析为 JSON，并在输出格式错误时重试请求，因此无论底层使用哪家提供商，下游应用都能收到格式良好的 JSON。

2.6 实现与易用性

TRALLIE 以开源 Python 软件包的形式在 Apache 2.0 许可证下发行。该实现提供了多种访问方式：用于集成流水线的编程接口、用于通过浏览器交互的轻量级 Streamlit 网页界面、展示常见工作流程的示例 Jupyter 笔记本，以及可复现的评估脚本。

3 评估

我们在两种互补的信息抽取设置下评估 TRALLIE：开放信息抽取和 ClosedIE。这些任务共同评估系统从非结构化文档中发现模式，以及在提供模式时抽取结构化属性值的能力。在 ClosedIE 设置中，目标模式预先定义并提供给模型，模型必须仅抽取这些属性对应的值，不得引入新字段或改写表述。相比之下，开放信息抽取设置不提供模式，因此要求 TRALLIE 推断出一组合适的属性，并直接从文档中填充这些属性。

3.1 评估方法

我们使用 Arora 等人 (2023年) 发布的评估指标和代码。ClosedIE 使用基于词元的 Text F1 进行评分（Rajpurkar 等人，2016年），先对每个属性的各文档得分取平均，再对属性取平均。开放信息抽取使用 (文档、属性、值) 三元组的 PairF1 进行评分，取排名前 k 个发现的属性，其中 k 为标准属性的数量。

3.2 评估数据集

我们在 EVAPORATE 基准（Arora 等人，2023年）上进行评估：美国食品药品监督管理局 FDA 510(k) 数据，以及来自 SWDE 电影和大学领域的 13 个网站（Hao 等人，2011年），共计 1,400 份文档。数据集可在 https://huggingface.co/datasets/hazyresearch/evaporate 获取。

监管领域。 FDA 510(k) 医疗器械许可文件；共使用全部 100 份文件。

SWDE 电影领域。 AllMovie、AMC TV、好莱坞、iHeartMovies、IMDb、Metacritic、Rotten Tomatoes 和 Yahoo Movies。每个网站

约有 2,000 个页面；我们按文件名选取前 100 个。

SWDE 大学领域。 CollegeProwler、eCampusTours、Embark、MatchCollege 和 US News，每个网站有 1,000–2,000 个页面；我们按文件名选取前 100 个。

4 结果

表 1比较了 TRALLIE 与 Arora 等人 (2023年) 在 EVAPORATE 基准上报告的系统。我们报告 ClosedIE 的结果：其待抽取的属性已给定，采用基于词元的文本 F1 评分（Rajpurkar 等人，2016年）；以及开放信息抽取的结果：系统还必须发现属性，并采用 Pair F1 评分。我们使用 gpt-oss- 120b（OpenAI，2025年）和 Qwen3.8-27B（Qwen 团队，2026年）运行 TRALLIE 。

TRALLIE 在使用两种大语言模型时，所有数据集上的 ClosedIE 得分均为最高（表 1），包括在提示开发期间留出的 SWDE 数据集。EVAPORATE-Direct（Arora 等人，2023年）每个属性只提示一次，而 TRALLIE 只需一次大语言模型调用，就能抽取文档的所有属性。开放信息抽取的表现则不太一致。TRALLIE 在 FDA 510(k) 上优于 EVAPORATE-Code+ ，使用 Qwen3.8-27B 时，在 SWDE 大学网站上也表现更佳；但在电影网站上则不及后者和 DOM-LM（Deng 等人，2022年）。各网站的开放信息抽取得分介于 20.2 至 80.1 之间（表 3），下文分析表明，这种差距主要源于属性命名，而非值提取。

开放信息抽取错误分析。 只有当值和属性名称都与金标准标注一致时，Pair F1 才会将预测判定为正确。TRALLIE 在 SWDE 上的大多数开放信息抽取错误属于命名错误，而非抽取错误：发现的模式改写了网站的标签（例如，导演 ，而金标准属性为导演），因此正确抽取的值也会被判定为错误。为区分命名错误与抽取错误，我们将每个金标准属性映射到与其值匹配程度最高的预测属性，并使用文本 F1 对结果进行评分。在这一映射下，TRALLIE 在电影网站上的得分分别为 78.0（gpt-oss-120b）和 80.2（Qwen3.8-27B），在大学网站上分别为 66.5 和 82.9。这仅用于诊断，不能与已发布的 Pair F1 数值直接比较。差距最大的是 Qwen3.8-27B 在烂番茄上的表现：值的抽取效果很好（映射后为 92.1），但模式使用了诸如电影标题 和发行这样的名称 年份 ，并包含近似重复的属性，使 Pair F1 仅为 20.2。

第二个误差来源是模式规模：发现的模式包含 20 个属性，因此对于拥有更多标准属性的网站（例如 Hollywood 上的 278 个和 MatchCollege 上的 66 个），其余属性无法匹配。由于 Pair 召回率是基于（文档、属性、值）三元组计算的，并且许多标准属性较为稀疏，因此召回率仍可能超过模式所覆盖的标准属性比例。Arora 等人 （2023年）还将属性名称与 SWDE 的一份有效替代名称列表进行匹配，但该列表尚未发布；因此，相较于已发表的结果，我们的严格匹配可能低估了 TRALLIE 的开放信息抽取得分。

提示设计的影响。 表2 展示了在大语言模型和文件保持不变的情况下，TRALLIE 修改后的提示对 FDA 510(k) 的影响。影响最大的是开放信息抽取，其中 Pair F1 从 25.3 提升至 68.6（gpt-oss-120b），以及从 22.8 提升至 80.2（Qwen3.8-27B）。大部分改进来自模式发现：指示模型在叶级复用文件自身的字段标签，使发现的模式中恢复出的标准属性名称数量从 16 个中的 6 个提升至 13 个（gpt-oss-120b），以及从 16 个中的 5 个提升至 14 个（Qwen3.8-27B）。ClosedIE 中给定了属性名称，因此变化较小（91.5 到 95.9，以及 97.0 到 97.2），主要是因为抽取现在会逐字复制值，并在字段缺失时返回空值。由于这些提示是在 FDA 510(k) 上开发的，FDA 的数值可能偏于乐观；SWDE 在开发期间被留作测试集，而两个大语言模型在该数据集上的 ClosedIE 高分表明，这些提示并非专门针对 FDA 510(k)。

5 系统演示

我们的演示通过交互式 Jupyter 笔记本和命令行界面展示 TRALLIE，涵盖以下内容：

- 在涵盖科学论文、监管申报文件和网络数据的示例文档集合上运行 liveschema 发现和数据抽取
- 通过更改单个提供商参数，比较不同大语言模型提供商（例如 GPT-4o，或 gpt-oss、Qwen 和 Llama 等本地模型）的抽取结果
- 观察系统如何通过重叠分块和基于大语言模型的整合来处理长文档

系统	LLM	FDA 510(k)		SWDE 电影		SWDE 大学	
		ClosedIE	开放信息抽取	ClosedIE	开放信息抽取	ClosedIE	开放信息抽取
DOM-LM†	–	–	–	71.9	54.1	68.0	55.2
EVAPORATE-Direct†	text-davinci-003	78.0[a]	45.5	84.4	45.2	72.6	53.8
EVAPORATE-Code+†	text-davinci-003	80.1	62.8	79.5	**56.8**	73.7	59.0
TRALLIE（我们的）	gpt-oss-120b	95.9	68.6	93.2[b]	49.8[b]	89.9[b]	56.3[b]
TRALLIE（我们的）	Qwen3.8-27B	**97.2**	**80.2**	**93.4**[b]	52.7[b]	**91.5**[b]	**64.4**[b]

表1：EVAPORATE 基准上的 ClosedIE（文本 F1）和开放信息抽取（Pair F1）。SWDE 分数为各网站的平均值（8 个电影类网站、5 个大学类网站）。 †数据报告自 Arora 等人（2023年）；未重新运行。 [a]按属性直接提示（其表6）。 [b]每个网站的前 100 个页面（EVAPORATE-Code+ 在所有页面上进行评估）；页面级 95% 自助法置信区间均在各平均值的 ±0.4 以内。TRALLIE 的提示是在 FDA 510(k) 数据上开发的；开发过程中未使用 SWDE。各系统使用不同的大语言模型，因此差异同时反映了方法和模型的影响。

LLM	ClosedIE		开放信息抽取	
	之前	之后	之前	之后
gpt-oss-120b	91.5	95.9	25.3	68.6
Qwen3.8-27B	97.0	97.2	22.8	80.2

表 2：TRALLIE 修订后的提示对 FDA 510(k) 的影响（100 份文件，使用相同的大语言模型和文件）。之前：原始提示；之后：模式发现复用文件自身在叶级的字段标签，抽取逐字复制值（上限约为 300 个单词），缺失字段则使用空值。发现的模式中与标准一致的字段名称，从 6/16 提升至 13/16（gpt-oss- 120b），从 5/16 提升至 14/16（Qwen3.8-27B）。

安装后（即 pip install trallie），用户只需几行代码即可从自己的文档中抽取结构化数据。示例文档和笔记本可在项目代码库中找到，[1]，在线演示则作为 Colab 笔记本提供。[2]

6 相关工作

早期的开放信息抽取（OpenIE）系统在没有预定义模式的情况下从文本中抽取关系元组，依赖自监督和语言结构（Banko 等人，2007年；Angeli 等人，2015年）。许多后续针对非结构化文本的方法使用依存句法分析器等语言工具（Covington，2025年；Tulpan 和 Tsur，2023年），先诱导出结构，再对所得结构应用启发式规则以抽取信息。对于 PDF 文件等较长的半结构化文档，这类处理流程表现不佳（Zhang 等人，2024年）。其他方法则假设整个流程中有人工标注员参与，用于标注数据并编写抽取规则（Schroeder 等人，2025年；Ding 等人，2025年），或者假设能够获取领域标注训练文档（Kim 等人，2022年；Meisenbacher 和 Norlander，2023年）。对于半结构化网页，SWDE 基准（Hao 等人，2011年）已被用于研究包装器诱导，以及从未见过的网站中进行零样本抽取（Lockard 等人，2020年）。

借助大语言模型（大语言模型），信息抽取正 increasingly 被表述为一种生成任务，在该任务中，实体、关系和属性值会直接以文本形式生成（Xu 等人，2023年）。ChatIE（Wei 等人，2023年）表明，零样本抽取可以被构建为与聊天模型进行的多轮问答对话。与我们的设置最为接近的是 EVAPORATE（Arora 等人，2023年），它利用大语言模型从半结构化文档集合（例如 HTML 页面和 PDF 文件）中生成结构化表格，无需领域专用训练或人工干预。它通过提示大语言模型处理少量文档样本并根据频率对提出的属性进行排序，从而确定一个模式；随后，它可以通过直接抽取或代码合成来填充该模式：在直接抽取中，大语言模型逐个处理文档；在代码合成中，大语言模型编写抽取函数，并将其大规模应用。其扩展版本 EVAPORATE-Code+通过弱监督组合多个合成的函数，从而降低计算成本。

其他研究则提供了用于模式引导抽取的专用模型和库。NuExtract（NuMind，2024年）是 NuMind 推出的一系列轻量级模型，基于合成数据进行微调，可从输入内容中填充用户提供的 JSON模板，既可用于零样本场景，也可进一步微调；NuMind 报告称，其抽取性能可与规模大得多的通用模型相媲美。LangExtract（Google，2025年）是 Google 推出的开源Python库，可根据任务描述和少量示例抽取结构化信息，并将每个抽取出的值与其在源文本中的位置对应起来；它通过分块和多轮抽取处理长文档，同时支持云端托管模型和本地模型。它还能够将每个抽取出的值定位到源文本中的

[1]https://github.com/PiSchool/trallie[2]https://colab.research.google.com/drive/124ECC8cQKD9zWSbPT9eB5Gcs-3UoB2vm

		gpt-oss-120b				Qwen3.8-27B			
		ClosedIE	开放信息抽取			ClosedIE	开放信息抽取		
网站	#属性。	F1	P	R	F1	F1	P	R	F1
SWDE 电影									
AllMovie	59	94.0	88.0	60.1	71.5	94.0	90.2	65.7	76.0
AMC TV	19	97.5	66.9	70.8	68.8	97.5	76.9	80.2	78.6
好莱坞	278	99.2	42.8	40.7	41.7	99.6	50.7	46.8	48.7
iHeartMovies	12	89.5	46.2	39.2	42.4	89.6	46.9	47.2	47.1
IMDb	39	88.9	50.1	26.4	34.5	91.6	48.6	24.8	32.8
Metacritic	20	86.1	46.1	45.6	45.9	77.8	70.4	40.4	51.3
烂番茄	13	97.3	39.4	44.3	41.7	97.5	18.7	21.9	20.2
Yahoo Movies	13	93.5	49.4	55.4	52.2	99.5	63.4	70.2	66.6
Mean		93.2			49.8	93.4			52.7
SWDE 大学									
CollegeProwler	23	96.6	41.5	45.1	43.2	97.0	56.9	60.7	58.7
eCampusTours	15	80.6	69.2	60.0	64.3	80.7	73.3	73.3	73.3
Embark	65	98.8	84.6	75.9	80.0	99.8	84.7	75.9	80.1
MatchCollege	66	82.7	47.0	19.6	27.7	85.9	65.0	26.7	37.8
US News	26	90.7	73.2	60.2	66.1	94.1	79.8	65.6	72.0
Mean		89.9			56.3	91.5			64.4

表 3：在每个网站的前 100 个页面上，TRALLIE 的各网站 SWDE 结果（两个大语言模型均没有失败页面）。#Attr. 表示标准属性的数量，这也 k 适用于开放信息抽取；发现的模式包含 20 个属性，因此对于标准属性更多的网站，其余属性无法匹配。平均值是按网站计算的，与表 1相同。

位置；在解码层面，受约束生成方法（Willard 和 Louf，2023年）和由提供商实现的 JSON 模式可确保生成语法有效的结构化输出，同时将模式的定义留给用户。

最后，NuNER（Bogdanov 等，2024年）、GLiNER（Zaratiana 等，2024年）、UniversalNER（Zhou 等，2024年）和 GoLLIE（Sainz 等，2024年）等通用命名实体识别模型，将抽取扩展到了训练期间未见过的实体类型或标注指南。这些模型针对语篇中的实体跨度进行处理，而文档级属性值抽取还需要判断哪些属性描述的是整个文档。

7 结论

我们介绍了 TRALLIE，这是一个轻量、模块化的开源框架，可将非结构化文档转换为结构化、可查询的格式。TRALLIE 利用大语言模型从文档集合中推导模式，并提取该模式对应的属性值，从而减少对手动模式设计和领域专用工程的需求。它可以处理 PDF、HTML、TXT 和 JSON 文档，并通过单一配置参数支持商业、定制及本地大语言模型提供商。在 EVAPORATE 基准上，TRALLIE 在使用两种大语言模型时，均在三个领域取得了最佳 ClosedIE 得分；在 FDA 510(k) 上取得了最佳开放信息抽取得分，并且使用 Qwen3.8-27B 时，在 SWDE 大学网站上也取得了最佳开放信息抽取得分。在 SWDE 电影网站上，开放信息抽取的表现仍低于 EVAPORATE-Code+，主要是因为发现的属性名称只是对网站标签的转述。这表明，让模式名称进一步贴合文档自身标签，是未来最有前景的研究方向。

局限性

与 EVAPORATE 的可比性。EVAPORATE 的结果取自 Arora 等人（2023年），因为他们使用的 LLM（text-davinci-003）已停止服务；因此，差异同时反映了方法和模型的影响。我们使用 EVAPORATE 发布的评估代码对 TRALLIE 进行评分。在 SWDE 上，出于成本考虑，我们使用每个网站的前 100 个页面，而不是全部页面。对页面进行自助法抽样（200 个样本；（Efron 和 Tibshirani，1993））得到的自助法置信区间，对于两个大语言模型而言都在各自平均值的 ±0.4 以内，但这一做法假设这些页面具有代表性，并忽略了用于模式发现的页面选择所带来的变化。EVAPORATE 针对 SWDE 列出的备选属性名称并未公开，因此我们严格的名称匹配可能低估了 TRALLIE 的开放信息抽取得分。

提示开发。 TRALLIE的提示在 FDA 510(k) 上进行了修订，因此这些结果可能偏于乐观。SWDE 被留作未参与训练的数据集；在该数据集上，ClosedIE 在两种大语言模型上都保持优势，但 OpenIE 仅在+ （Arora 等人，2023年）使用

大学网站上使用 Qwen3.8-27B 时表现更好，而在其他方面则落后于它，因为发现的属性名称是对网站标签的改写。

系统范围。 TRALLIE 每份文档调用一次大语言模型，因此不同于 EVAPORATE-Code+（后者每个集合仅合成一次抽取代码），其成本会随集合规模线性增长。发现的模式仅保留少量文档样本（在我们的实验中为 10 份）中最常见的属性，因此会遗漏长尾属性，而模式质量取决于该样本的代表性。TRALLIE 抽取扁平属性–值记录；它不对关系、时间信息或表格等嵌套结构进行建模。文档在抽取前会转换为纯文本，这会丢失多栏 PDF 文件和脚本密集型 HTML 中的版式信息。输出仅经过格式良好的 JSON 检查，不检查值类型或格式，因此高风险用途仍需要人工审查。

致谢

我们谨向所有为本项工作作出贡献的同事和合作伙伴表示感谢。本研究由欧盟委员会 NGI（下一代互联网）Search 倡议的第 4 次公开征集资助。

参考文献

Gabor Angeli、Melvin Jose Johnson Premkumar 和 Christopher D. Manning。2015年。利用语言结构进行开放域信息抽取。载于计算语言学协会第*53*届年会论文集（*ACL*），第344–354页。

Simran Arora、Brandon Yang、Sabri Eyuboglu、Avanika Narayan、Andrew Hojel、Immanuel Trummer 和 Christopher Ré。2023年。语言模型助力构建简单系统，以生成异构数据湖的结构化视图。*VLDB*基金会会议录，17(2):92–105。

Michele Banko、Michael J. Cafarella、Stephen Soderland、Matt Broadhead 和 Oren Etzioni。2007年。从网络中进行开放信息抽取。载于第二十届国际人工智能联合会议（*IJCAI*）论文集，第2670–2676页。

Sergei Bogdanov、Alexandre Constantin、Timothée Bernard、Benoit Crabbé 和 Etienne Bernard。2024年。NuNER：通过大语言模型标注数据进行实体识别编码器预训练。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。

Michael A. Covington。2025年。依存句法分析的一个基本算法（含修正）。*arXiv* 预印本 *arXiv:2510.19996*。

Xiang Deng、Prashant Shiralkar、Colin Lockard、Binxuan Huang 和 Huan Sun。2022年。DOM-LM：学习适用于 HTML 文档的通用表示。*arXiv* 预印本 *arXiv:2201.10608*。

Yiping Ding、Xiaorui Jiang 和 Opeoluwa Akinseloyin。2025年。通过人类与大语言模型协作对技术辅助医学证据综合进行计算综述：Cochrane 案例研究。*medRxiv* 预印本 *2025.11.08.25339805*。

Bradley Efron 和 Robert J. Tibshirani。1993年。自助法简介。Chapman and Hall/CRC，纽约。

Google。2025年。LangExtract：一个使用大语言模型从非结构化文本中抽取结构化信息的 Python 库。https://github.com/google/langextract。Qiang Hao、Rui Cai、Yanwei Pang 和 Lei Zhang。2011年。从一棵树到一片森林：结构化网页数据抽取的统一解决方案。载于第*34*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，第775–784页。Minsang Kim、Sang-hyun Je 和 Eunjoo Park。2022年。OASYS：从非结构化文本构建知识库的领域无关自动化系统。*arXiv* 预印本 *arXiv:2207.07597*。Colin Lockard、Prashant Shiralkar、Xin Luna Dong 和 Hannaneh Hajishirzi。2020年。ZeroShotCeres：从半结构化网页进行零样本关系抽取。载于第*58*届计算语言学协会年会（*ACL*）论文集。Stephen Meisenbacher 和 Peter Norlander。2023年。利用上下文规则辅助机器学习（CRAML）将非结构化文本转换为数据。*arXiv* 预印本 *arXiv:2301.08549*。NuMind。2024年。NuExtract 1.5：多语言、无限上下文，依然小巧，且优于 GPT-4o！
h
t
t
p
s://
w
w
w.
n
u
m
i
n
d.
a
i/
b
l
o
g/
n
u
e
x
t
r
a
c
t-1-5---multilingual-infinite-context-still-small-and-better-than-gpt-4o。OpenAI。2025年。gpt-oss-120b 和 gpt-oss-20b 模型卡。*arXiv* 预印本 *arXiv:2508.10925*。Qwen 团队。2026年。Qwen3.8-Max：编程与协作的新标杆。https://qwen.ai/blog?id=qwen3.8。所用模型：https://huggingface.co/Qwen/Qwen3.8-27B。Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。2016年。SQuAD：用于机器阅读理解的100000+个文本问题。载于*2016*年自然语言处理经验方法会议论文集，第2383–2392页，得克萨斯州奥斯汀。计算语言学协会。

Oscar Sainz、Iker García-Ferrero、Rodrigo Agerri、Oier Lopez de Lacalle、German Rigau 和 Eneko Agirre。2024年。GoLLIE：标注指南提升零样本信息抽取效果。载于第十二届国际学习表征会议（*ICLR*）。

Hope Schroeder、Deb Roy 和 JadKabbara。2025年。只需让人类参与其中？探究大语言模型辅助标注在主观任务中的应用。 *arXiv* 预印本 *arXiv:2507.15821*。

Yoav Tulpan 和 Oren Tsur。2023年。一种更深入的（自回归）非收敛话语解析方法。载于《*2023*年自然语言处理实证方法会议论文集》，页面 12883–12895。计算语言学协会。

Xiang Wei、Xingyu Cui、Ning Cheng、Xiaobin Wang、Xin Zhang、Shen Huang、Pengjun Xie、Jinan Xu、Yufeng Chen、Meishan Zhang、Yong Ji ang 和 Wenjuan Han。2023年。通过与 ChatGPT 对话实现零样本信息抽取。*arXiv* 预印本 *arXiv:2302.10205*。

Brandon T. Willard 和 Rémi Louf。2023年。面向大语言模型的高效引导式生成。*arXiv* 预印本 *arXiv:2307.09702*。

Derong Xu、Wei Chen、Wenjun Peng、Chao Zhang、Tong Xu、Xiangyu Zhao、Xian Wu、Yefeng Zheng、Yang Wang 和 Enhong Chen。2023年。用于生成式信息抽取的大语言模型：一项综述。*arXiv* 预印本 *arXiv:2312.17617*。

Urchade Zaratiana、Nadi Tomeh、Pierre Holat 和 Thierry Charnois。2024年。GLiNER：使用双向 Transformer 的通用命名实体识别模型。载于北美计算语言学协会分会 *2024*年会议论文集：人类语言技术（*NAACL*）。计算语言学协会。

Qintong Zhang、BinWang、Victor Shea-Jay Huang、Jun-yuan Zhang、Zhengren Wang、Hao Liang、Conghui He 和 Wentao Zhang。2024年。文档解析揭秘：结构化信息抽取的技术、挑战与前景。*arXiv* 预印本 *arXiv:2410.21169*。

Wenxuan Zhou、Sheng Zhang、Yu Gu、Muhao Chen 和 Hoifung Poon。2024年。UniversalNER：从大语言模型中进行面向目标的知识蒸馏，用于开放式命名实体识别。发表于第十二届国际学习表征会议（*ICLR*）。