# AI 项目站外数据获取调研

# 元典智库分析

## 元典智库

- **定义与定位**
  元典智库是华宇软件旗下、面向专业法律人士（法官、检察官、律师、企业法务、法学师生等）提供的“一站式法律知识服务与智能检索平台”
- **数据规模与内容覆盖**
  - 法律法规：收录超过 400 万条
  - 公开裁判文书：涵盖约 1.5 亿篇
- **核心功能亮点**
  - 法律知识图谱、搜索、智能推荐与图谱关联
  - 智能案例研判、知识管理、团队协作、量刑辅助、可视化大数据分析、风险画像、智能文书生成等模块
- **部署方式与使用场景**
  - 提供 SaaS 在线访问与单位内网本地部署两类方案
  - 支持 Web 端与小程序访问，满足移动多终端使用需求

## 元典问答

- **定位与功能**：作为基于大模型 AI 的智能法律问答与类案检索工具，整合元典智库的数据和全网资源，提供“一站式法律问题检索服务”。
- **核心模块功能**：
  - **智能问答**：融合 DeepSeek-R1 模型和元典自有海量案例 (1.6 亿)、法规 (460 万) 与全网权威页面，支持多角度观点生成并可追溯来源和推理路径。
  - **法律调研场景**：提供“综合研究”“全网观点”“法规研究”“案例洞察”四大视角，快速覆盖问题全貌。
  - **智能文书生成**：支持元素式起诉状、答辩状、律师函、法律咨询意见书等多种文书的一键生成、编辑与下载。
  - **文档分析与输出**：支持多文档阅读、智能提取关键内容，并导出调研报告或分享至微信移动端。

## 对比

| 维度 | 元典智库 | 元典问达 |
| --- | --- | --- |
| **核心功能** | 智能搜索、知识图谱、类案研判、大数据分析、团队协同 | 智能问答、调研视角、文书生成、来源可溯、报告输出 |
| **技术支撑** | 图谱、意图识别、推荐系统、大数据可视化 | 大模型（DeepSeek-R1）、LLM 推理、多角度摘要 |
| **目标场景** | 综合法律研究、案情分析、团队知识管理 | 快速解决法律问题、文书自动生成、实务问答与调研 |

| 维度 | 元典智库 | 元典问达 |
| --- | --- | --- |
| **数据支撑** | 自有法规、文书库与企业/律师信息 | 利用元典智库数据 + 全网观点 + 大模型理解能力 |
| **输出形式** | 检索结果、报告导出、笔记、协同 | 生成答案、文书文档、调研报告、分享链接 |
| **使用方式** | SaaS 或本地部署平台 | 在线问答工具，支持网页与小程序访问 |

**核心要点：海量数据，案例支撑**

# 一、项目背景

- 目标：将外部网页/站点转成 LLM 友好文本或结构化数据（Markdown/JSON），用于 RAG 检索、Agent 工具调用、摘要与监控。
- 核心难点：
  - 反爬/动态渲染网页处理
  - 数据清洗与结构化
  - 稳定性、速率和成本控制
  - 合规性（robots.txt、站点 ToS、PII 保护）

# 二、调研工具概览

| 工具 | 核心定位 | 适配 RAG/Agent | 接入难度 | 动态页面处理 | 成本/伸缩 | 最佳场景 |
| --- | --- | --- | --- | --- | --- | --- |
| Firecrawl | SaaS/API 抓取 + 搜索 + 全站爬行，输出 Markdown/JSON | 很强 | 低 | 平台处理 | 按量或套餐(自托管版本功能有限) | 快速上线 RAG/Agent，全站抓取 |
| **crawl4ai** | **开源 LLM 友好爬虫库，强调鲁棒解析与清洗** | **很强** | **中** | **可通过渲染管线处理** | **自托管成本（算力/带宽/维护）** | **高度可控、自托管深度集成** |

| 工具 | 核心定位 | 适配 RAG/Agent | 接入难度 | 动态页面处理 | 成本/伸缩 | 最佳场景 |
| --- | --- | --- | --- | --- | --- | --- |
| Jina AI Reader API | 极简 API 单页抽取（ReaderLM-v2 支持复杂页） | 中 | 极低 | 平台处理 | 按量计费 | 单页信息获取，轻量化快速接入 |
| **Scrapegraph-ai** | 图式流程抓取（抓取→解析→存储），可挂本地 LLM | 强 | 中-高 | 自定义节点/浏览器 | **自托管**成本（算力/代理/维护） | 复杂任务编排与企业级流水线 |

# 三、可行性与接入分析

## 1. Firecrawl

- **可行性**：支持搜索+抓取一体，直接输出 LLM-ready Markdown/JSON。
- **接入**：REST/SDK（Python、Node、Go），几行代码即可接入。
- **优势**：快速上线原型，直接适配 RAG/Agent。
- **限制**：大规模抓取成本较高，依赖第三方服务。

## 2. crawl4ai

- **可行性**：开源，自由度高，可将网页转成结构化文本。
- **接入**：Python 工程集成，自建代理、重试和分布式。
- **优势**：私有化部署，长期成本可控。
- **限制**：维护工作量大，需技术团队支持。

## 3. Jina AI Reader API

- **可行性**：快速单页抓取，支持 SERP 批量抓取。
- **接入**：极简 HTTP API 调用。
- **优势**：轻量化、快速接入。
- **限制**：大规模或全站抓取需要额外编排和速率控制。

### 4. Scrapegraph-ai

- **可行性**：可将抓取流程模块化，支持复杂任务。
- **接入**：Python 图式流程，需要设计节点和数据流。
- **优势**：复杂流程、企业内网与公网混合抓取。
- **限制**：工程复杂度高，需要监控和维护

## 四、接入成本分析

| 工具 | 免费额度 | 付费成本 | 主要成本因素 | 备注 |
| --- | --- | --- | --- | --- |
| Firecrawl | 每月 500 次请求 | $19–$399/月 | 套餐限制并发、请求频率 | SaaS 托管，快速集成 |
| **crawl4ai** | 开源免费 | 自托管成本（服务器、带宽、存储、维护） | 算力、带宽、运维 | 灵活可控，自建流水线 |
| Jina AI Reader API | 每月 1M 词元 | $50–$500/月 | 词元数 | 单页快速接入，轻量化 |
| Scrapegraph-ai | 开源免费 | 自托管成本 | 服务器、带宽、代理、维护 | 复杂任务编排，适合企业级 |

## 五. 架构建议

- **抓取层**：Firecrawl / Jina / **crawl4ai**(🌟) / Scrapegraph-ai
- **清洗/抽取层**：Markdown / JSON / 表格 / 列表抽取
- **存储层**：对象存储 + 文本仓 + 向量库
- **编排层**：任务队列、代理池、失败重试、速率控制
- **合规层**：robots.txt、ToS 白名单、PII 过滤、缓存与回源控制
- **消费层**：RAG 检索、Agent 工具调用