

自动化QA数据集生成管道：技术实现概述

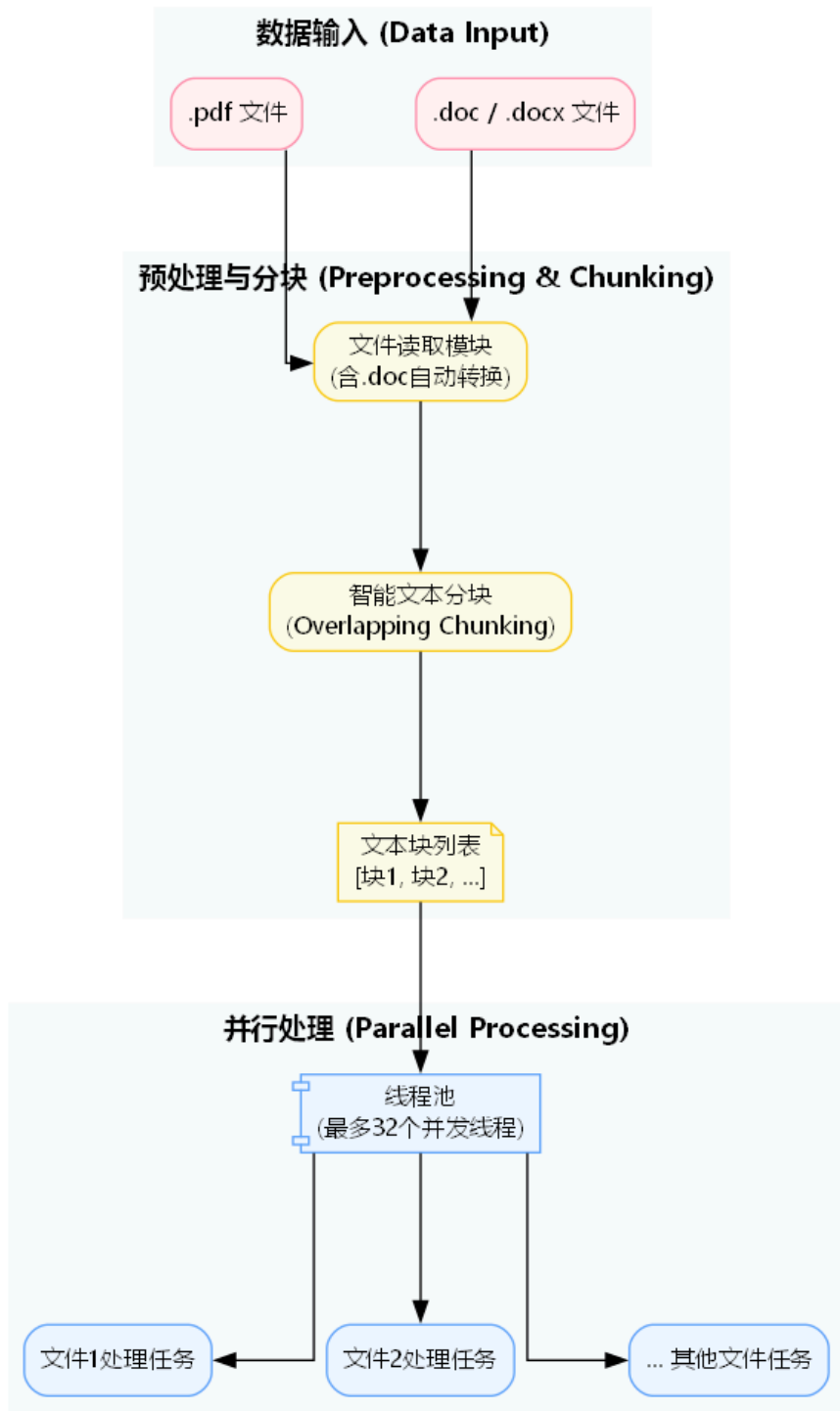
为满足从多种格式的非结构化文档中，批量、高效地生成高质量、结构化问答(QA)数据集的需求，我设计并实现了一套自动化的数据处理与生成管道。该管道的核心是利用大型语言模型（LLM）的自然语言理解与生成能力，并通过一系列工程技术确保流程的稳定性、可扩展性和鲁棒性。

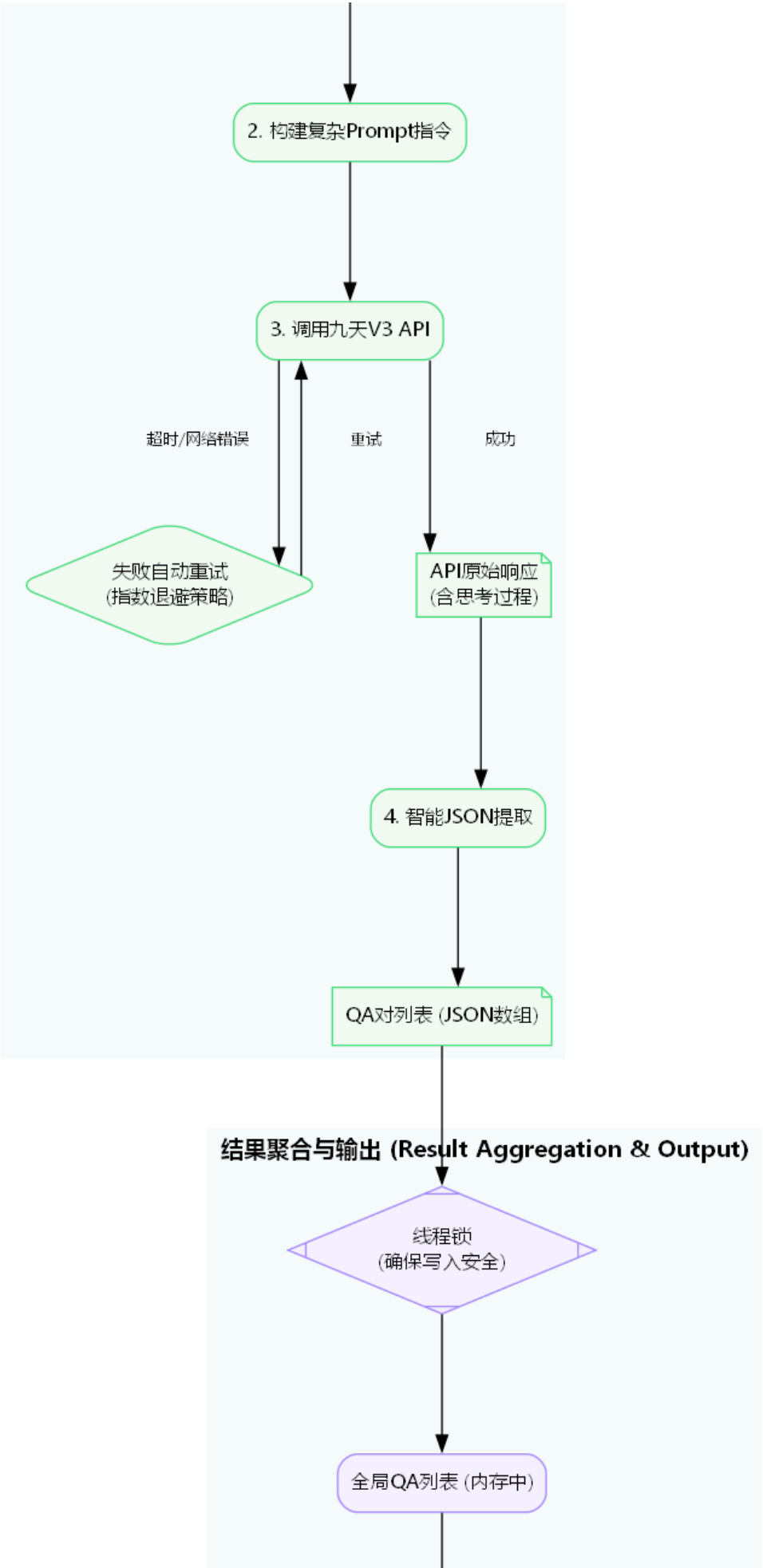
核心技术栈与组件：

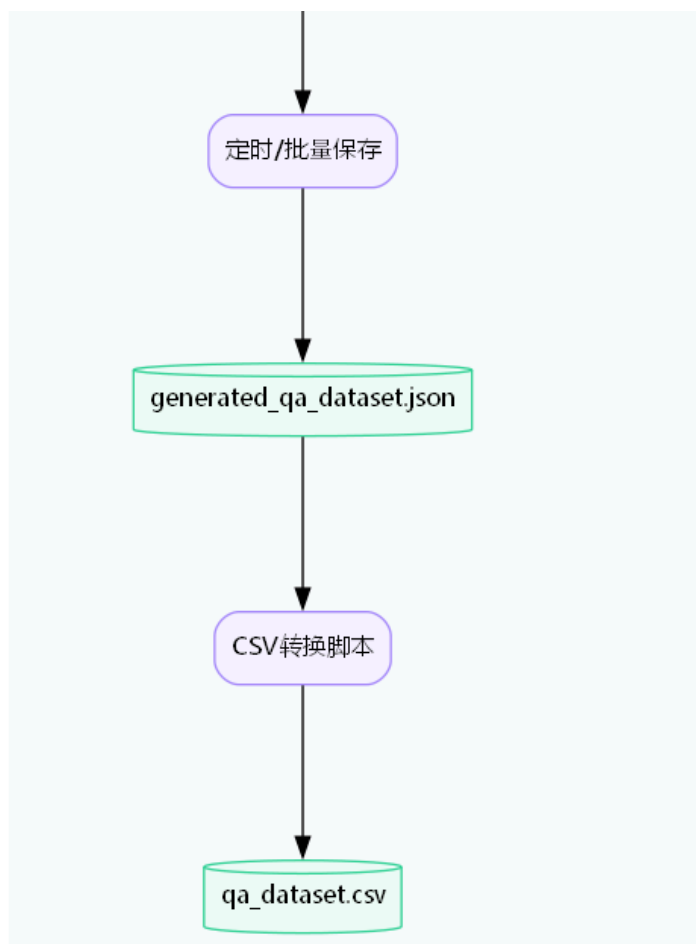
- **编程语言与环境:** Python 3.x
- **API交互:** `requests` 库用于执行与九天大模型V3 API (`/api/v3/chat/completions`)的HTTP POST通信。
- **安全认证:** 采用 `PyJWT` 库实现符合官方规范的JWT (JSON Web Token)动态令牌生成。通过HS256算法，将API Key的ID和Secret部分与时间戳和有效期结合，为每次API请求生成一个有时效性的Bearer Token，确保了通信的安全性。
- **多格式文档解析:**
 - **PDF处理:** 利用 `PyPDF2` 库从 `.pdf` 文件中提取纯文本内容。
 - **Word文档处理:**
 - 使用 `python-docx` 库直接解析现代 `.docx` 文件。
 - 为兼容老旧的 `.doc` 格式，通过 `pywin32` 库调用Windows COM接口，在后台启动Microsoft Word应用程序，以编程方式将 `.doc` 文件无损转换为 `.docx` 格式，从而实现对多种Word文档的无缝处理。
- **并行处理与并发控制:**
 - 为了最大化处理效率，引入了 `concurrent.futures.ThreadPoolExecutor`，构建了一个可配置（最多32个）工作线程的线程池，实现了对多个源文件的并行处理。
 - 为确保在多线程环境下对共享数据（最终的QA列表和输出文件）的写入操作是原子性的，使用了 `threading.Lock` 机制，有效防止了竞态条件和数据损坏。
- **数据处理与解析:**
 - **智能文本分块 (Overlapping Chunking):** 为解决因 `prompt` 过长导致的API网关超时（504 Gateway Time-out）问题，设计并实现了带有重叠区域的文本分块策略。将长文档切分为固定大小（如2500字符）的文本块，并让相邻块之间有部分内容重叠（如200字符），这在保证每次API请求轻量化的同时，最大程度地维持了上下文的连续性，保障了生成质量。
 - **健壮的JSON提取:** 针对LLM可能返回非纯净JSON（例如，包含其“思考过程”的文本）的情况，利用正则表达式(`re`)和字符串查找(`find / rfind`)方法，实现了一个智能解析器。该解析器能从混杂的文本中精确地提取出由 `[...]` 或 `{ ... }` 包裹的核心JSON内容，极大地增强了系统的容错能力。
- **错误处理与稳定性:**
 - **自动重试机制:** 在API调用函数中集成了带有指数退避策略的重试循环。当遇到网络波动、服务器临时超时等可恢复性错误时，脚本会自动等待一个逐渐加长的时间后重试（最多4次），显

著提升了长时间、大批量任务的成功率。

- **全面的日志系统:** 使用 logging 模块，将详细的运行日志（包括线程信息、API响应、错误详情等）同时输出到控制台和本地日志文件，为监控、调试和问题追溯提供了坚实的基础。







```
2025-07-03 19:09:27,788 - QAGenerator - ERROR - Thread 37824 - 读取Word文件 政务系统问答数据集/origin_files/未转换.docx 失败: Package not found at '政务系统问答数据集/origin_files/未转换.docx'
2025-07-03 19:09:27,791 - QAGenerator - ERROR - Thread 35720 - 读取Word文件 政务系统问答数据集/origin_files/未转换.docx 失败: Package not found at '政务系统问答数据集/origin_files/未转换.docx'
2025-07-03 19:09:27,792 - QAGenerator - WARNING - Thread 34644 - 文件内容为空或读取失败: 政务系统问答数据集/origin_files/未转换.docx
2025-07-03 19:09:27,792 - QAGenerator - WARNING - Thread 40032 - 文件内容为空或读取失败: 政务系统问答数据集/origin_files/未转换.docx
2025-07-03 19:09:27,793 - QAGenerator - WARNING - Thread 37824 - 文件内容为空或读取失败: 政务系统问答数据集/origin_files/未转换.docx
2025-07-03 19:09:27,793 - QAGenerator - WARNING - Thread 35720 - 文件内容为空或读取失败: 政务系统问答数据集/origin_files/未转换.docx
2025-07-03 19:09:27,794 - QAGenerator - WARNING - Thread 41476 - 文件内容为空或读取失败: 政务系统问答数据集/origin_files/未转换.docx
2025-07-03 19:09:27,824 - QAGenerator - INFO - Thread 39292 - 成功读取Word文件: 政务系统问答数据集/origin_files/未转换/8.docx
2025-07-03 19:09:27,824 - QAGenerator - INFO - Thread 39292 - 文件 8.docx 已被切分成 1 个重叠的文本块。
2025-07-03 19:09:27,824 - QAGenerator - INFO - Thread 39292 - --- 正在为文件 '8.docx' 的第 1/1 个文本块生成QA对 ---
2025-07-03 19:09:27,829 - QAGenerator - INFO - Thread 32972 - 成功读取Word文件: 政务系统问答数据集/origin_files/未转换/9.docx
2025-07-03 19:09:27,830 - QAGenerator - INFO - Thread 32972 - 文件 9.docx 已被切分成 2 个重叠的文本块。
2025-07-03 19:09:27,830 - QAGenerator - INFO - Thread 32972 - --- 正在为文件 '9.docx' 的第 1/2 个文本块生成QA对 ---
2025-07-03 19:09:27,862 - QAGenerator - INFO - Thread 36904 - 成功读取Word文件: 政务系统问答数据集/origin_files/未转换/企业文件.docx
2025-07-03 19:09:27,862 - QAGenerator - INFO - Thread 36904 - 文件 企业文件.docx 已被切分成 15 个重叠的文本块。
2025-07-03 19:09:27,862 - QAGenerator - INFO - Thread 36904 - --- 正在为文件 '企业文件.docx' 的第 1/15 个文本块生成QA对 ---
2025-07-03 19:09:27,895 - QAGenerator - INFO - Thread 37720 - 成功读取Word文件: 政务系统问答数据集/origin_files/未转换/企业4.docx
2025-07-03 19:09:27,895 - QAGenerator - INFO - Thread 37720 - 文件 企业4.docx 已被切分成 7 个重叠的文本块。
2025-07-03 19:09:27,895 - QAGenerator - INFO - Thread 37720 - --- 正在为文件 '企业4.docx' 的第 1/7 个文本块生成QA对 ---
2025-07-03 19:10:24,130 - QAGenerator - INFO - Thread 36904 - 成功生成并添加了 10 个QA对。总数: 10
2025-07-03 19:10:24,132 - QAGenerator - INFO - Thread 36904 - 进度已保存, 当前总计 10 个。
2025-07-03 19:10:24,132 - QAGenerator - INFO - Thread 36904 - --- 正在为文件 '企业文件.docx' 的第 2/15 个文本块生成QA对 ---
```

最终成果:

通过上述技术的综合应用，最终交付的是一个高效、稳定且可扩展的Python脚本。它能够：

1. 并行处理包含 .pdf ， .docx ， .doc 等多种格式的大批量文件。
2. 自动、安全地完成与九天V3 API的认证和通信。
3. 通过智能分块和上下文重叠技术，在保证生成质量和避免API超时的矛盾中取得了最佳平衡。

4. 利用精心设计的 prompt 工程，引导模型生成符合特定场景、特定格式（包括对流程图的文字描述）的高质量QA对。
5. 最终将所有生成的QA对聚合，并保存为结构化的 .json 和 .csv 文件，便于后续的数据分析、模型训练或知识库构建。

最终生成的CSV数据集文件示例

技术总结.md • qa_dataset.csv × 流程图.py pipeline_flowchart.png 算法优化.md def.h ...Graph Prim.c 笔记

政务系统问答数据集 > qa_dataset.csv

```
1 question, answer
2 作为一名新手家长，我想知道为什么培养孩子的阅读兴趣如此重要？，“阅读是孩子生存和发展的重要的能力。
3 培养孩子的阅读能力的关键是激发他们对阅读的兴趣。”
4 幼儿园在激发孩子的阅读兴趣方面可以做哪些工作呢？，“幼儿园应该为孩子创造阅读环境。
5 给孩子们指导阅读方法。
6 开展阅读活动。
7 让孩子们感受和喜爱书本和阅读。”
8 文章中提到了一些具体的阅读环境和方法，能详细说说吗？，“例如，创造像‘书香校园’这样的阅读氛围。
9 引导孩子进行‘亲子共读’。
10 利用‘故事大王’等活动形式。
11 鼓励孩子参与‘阅读分享’。”
12 对于小班、中班和大班的孩子，他们在阅读方面有什么不同的特点和需求吗？，“小班孩子可能更喜欢听故事，需要更多的亲子互动。
13 中班孩子可以开始尝试自主阅读简单的图画书。
14 大班孩子可以阅读更复杂的绘本，并进行简单的阅读理解。”
15 文章中提到的“早期阅读”是指什么？，“早期阅读是指在孩子入园前就开始进行的阅读启蒙和培养。
16 它侧重于激发孩子对书本和阅读的兴趣，培养良好的阅读习惯。”
17 在阅读活动中，如何让孩子感受到阅读的乐趣呢？，“可以通过表演故事、角色扮演等方式，增加阅读的趣味性。
18 鼓励孩子表达对故事的理解和感受。
19 创造轻松愉快的阅读氛围，避免强迫孩子阅读。”
20 父母在家中如何配合幼儿园的工作，共同培养孩子的阅读兴趣？，“家长可以在家中也为孩子创造阅读环境，如设置家庭图书角。
21 每天抽出时间与孩子进行亲子共读。
22 鼓励孩子在家中也进行阅读，并给予积极的反馈。”
23 文章中提到了“书香校园”，这具体指的是什么？，“书香校园是指营造一种浓厚的读书氛围，让书籍成为校园文化的重要组成部分。
24 这包括在教室、走廊等地方摆放书籍，鼓励孩子随时随地阅读。”
25 “亲子共读”具体应该如何进行呢？，“家长和孩子一起选择喜欢的绘本。
26 家长声情并茂地给孩子讲故事。
27 在阅读过程中，可以和孩子进行互动，提问并引导孩子思考。
28 鼓励孩子在阅读后表达自己的感受和理解。”
29 除了阅读活动，还有什么其他方法可以激发孩子的阅读兴趣？，“可以利用绘本剧表演，让孩子在角色扮演中感受故事的魅力。
30 组织读书分享会，让孩子交流阅读心得。
31 鼓励孩子创作自己的故事，激发他们的想象力和创造力。”
32 作为一名幼儿园老师，我应该如何激发孩子们对阅读的兴趣？，“作为一名幼儿园老师，您可以从以下几个方面着手激发孩子们对阅读的兴趣：
33 1. **创设阅读环境**：为孩子们创建一个充满书香的环境，例如设置图书角，摆放各种有趣的绘本，让孩子们随时都能接触到书本。
34 2. **提供阅读指导**：教给孩子们正确的阅读方法，例如如何翻页、如何指读，以及如何理解书中的图画和文字。
35 3. **开展阅读活动**：组织各种有趣的阅读活动，例如讲故事、角色扮演、阅读分享等，让孩子们在轻松愉快的氛围中体验阅读的乐趣。
```

技术总结.md • qa_dataset2.csv × 流程图.py pipeline_flowchart.png 算法优化.md def.h ...Graph Prim.c 笔记

政务系统问答数据集 > qa_dataset2.csv

```
12 8. 成立日期，公司正式成立的日期；
13 9. 统一社会信用代码，这是企业在全国范围内的唯一身份标识。这些信息都是公司非常重要的基础信息，需要准确无误地载明在营业执照上。”
14 我打算设立一家有限责任公司，想了解股东们认缴出资有什么具体的规定吗？，您好！关于有限责任公司股东认缴出资，文档里有详细说明。根据第五条，股东
15 如果我的有限责任公司想增加注册资本，应该怎么办呢？，您好！有限责任公司增加注册资本也是有明确规定的。根据第七条，公司增加注册资本后，股东认缴
16 除了用现金出资，股东还可以用哪些方式出资设立公司呢？，您好！出资方式有很多种，不限于现金。根据第六条，股东可以用货币出资，也可以用实物、知识
17 我有一家老公司，是2024年6月30日之前登记设立的有限责任公司，想了解下股东的出资期限有什么新的规定吗？，您好！针对2024年6月30日之前登记设立的
18 在什么情况下，公司登记机关会对公司的注册资本进行真实性、合理性的研判呢？，您好！公司登记机关会在一些特定情况下对公司注册资本的真实性、合理
19 1. 认缴出资期限三十年以上；
20 2. 注册资本十亿元人民币以上；
21 3. 其他明显不符合客观常识的情形。登记机关会结合公司的经营范围、经营状况以及股东的出资能力、主营业务、资产规模等进行综合研判，必要时还会组
22 如果我想委托别人帮我办理公司的登记或者备案事项，可以吗？有哪些需要注意的地方吗？，您好！当然可以委托别人代办公司登记、备案。根据第十六条，
23 1. 代理人需要诚实守信，依法履责。
24 2. 要标明代理身份，并提交授权委托书，明确委托事项和权限。
25 3. 不得提交虚假材料或者采取欺诈手段隐瞒重要事实。
26 4. 不得利用代理业务损害国家、社会公共利益或者他人合法权益。选择信誉良好的中介机构或代理人非常重要，要确保他们能够诚信合法地办理相关事宜。”
27 我想指定一位登记联络员来负责公司与登记机关之间的联络工作，有什么具体的要求吗？，您好！指定登记联络员是很有必要的。根据第十四条，公司设立登
28 如果我是一家新成立的小型企业管理者，想要给我的公司起名，根据这份文件，我应该遵循哪些基本原则呢？，“作为新成立的小型企业管理者，给公司起名时
29
30 1. **使用规范汉字**：企业名称必须使用规范的汉字，确保名称的标准化和易读性。
31
32 2. **构成要素**：企业名称通常由行政区划名称、字号、行业或者经营特点、组织形式这四个部分组成，并且需要按照这个顺序排列。例如：XX省XX市XX
33
34 3. **显著性**：企业名称中的字号需要具有显著性，由两个以上的汉字组成，可以是字、词或者它们的组合，确保您的企业名称能够区别于其他企业。
35
36 4. **行业特点**：企业名称中的行业或者经营特点要用语需要根据您的公司的主营业务和国民经济行业分类标准来确定。这样能够清晰地表明您的业务
37
38 5. **组织形式明确**：您需要根据公司的组织结构或者责任形式，在企业名称中标明相应的组织形式用语，比如“有限责任公司”、“股份有限公司”等，确
39 我注意到文件中提到企业名称登记管理由国家市场监督管理总局负责，请问具体来说，国家市场监督管理总局在企业名称登记管理方面承担哪些主要职责呢？
40
41 1. **制定规则**：负责制定企业名称禁用规则、相同相近比对规则等企业名称登记管理的具体规范，为全国的企业名称登记工作提供指导和依据。
42
43 2. **系统建设和维护**：负责建立、管理和维护全国企业名称规范管理系统和国家市场监督管理总局企业名称申报系统，为全国的企业名称申报和管理提供
44
45 3. **监督管理**：负责监督全国的企业名称登记管理工作，确保各地企业登记机关按照统一的规则 and 标准进行企业名称登记，维护全国企业名称登记管理制
46 文件中提到了企业公共机构的名称，如果我计划为我公司设立分公司，那么分公司的名称应该如何规范呢？，“当您为公司设立分公司时，分公司的名称需要
```

```
政务系统问答数据集 > qa_dataset3.csv
1 question,answer
2 请问作为一家小规模纳税人，我们公司可以享受哪些增值税减免优惠？,根据文件，作为小规模纳税人，您可以享受以下增值税减免优惠：1. 增值税小规模纳
3 我们公司主要做出口业务，请问如何利用‘数智通’平台申请出口退税，有哪些便捷措施？,为了利用‘数智通’平台申请出口退税并享受便捷措施，您可以按照
4 我们公司近期资金周转有些困难，想了解一下什么是金融支持重点企业‘白名单’，以及入选‘白名单’的企业能获得哪些支持？,金融支持重点企业‘白名单’是
5 最近政府都在强调提高信贷比重，请问有哪些具体的措施来支持企业获得信贷？,为了提高信贷比重，政府采取了以下具体措施来支持企业获得信
6 我听说‘政银担’合作可以降低企业融资风险，请问这个机制是如何运作的？政府在其中扮演什么角色？,‘政银担’合作机制是一种风险分担机制，旨在降低企
7 我们公司经常运输农产品，请问整车合法装载运输鲜活农产品的车辆，高速公路通行费有哪些优惠政策？,对于整车合法装载运输全国统一《鲜活农产品品种目
8 请问政府在发展多式联运，推动‘公转水’‘公转铁’方面有哪些支持政策？,为了发展多式联运，推动‘公转水’‘公转铁’，政府出台了以下支持政策：1. **资
9 请问哪些主体可以享受‘六税两费’减半征收的优惠？,根据文件，以下主体可以享受‘六税两费’减半征收的优惠：1. **增值税小规模纳税人**；这是最主要
10 什么是‘政采贷’？我们公司经常参与政府采购，请问如何利用政府采购合同进行融资？,‘政采贷’是指中小微企业利用政府采购合同向金融机构申请融资的一
11 我们公司遇到了临时性的资金周转难题，请问有什么政策可以帮助我们缓解？,针对企业面临的临时性资金周转难题，政府出台了以下政策来帮助您缓解：1.
12 我们是一家湖北省的中小企业，最近想尝试参与东风汽车的供应链，想了解一下湖北省有没有相关的扶持政策可以帮助我们？,当然有！湖北省非常支持像你们
13 我们是一家湖北省制造企业，最近想做一些减排的技术改造，想了解下政府有没有什么补贴政策？,有的，湖北省非常鼓励制造业企业进行碳减排技术改造
14 我们是湖北省一家小型科技企业，想利用湖北省重点实验室的设备做一些研发，但又担心费用太高，请问有什么好的办法吗？,这是一个好消息！湖北省为了
15 我们湖北省的企业一直很注重产品质量和品牌建设，最近听说有个“荆楚精品”的认证，请问申报条件是什么？,“荆楚精品”认证是湖北省为了打造高品质品牌
16 我是湖北省退役军人，最近准备创业，请问有什么特殊的政策可以支持我吗？,湖北省非常支持退役军人创业！在金融支持方面，你们可以申请的贴息贷款额度
17 我们公司想入驻湖北省的省级产业园，请问除了公示的政策外，还有什么隐藏的福利吗？,入驻省级产业园确实有一些非公示的政策福利！以武汉市光谷为例，
18 我们是湖北省一家二类医疗器械企业，最近想注册新产品，请问有什么加速通道吗？,有的！湖北省为鼓励生物医药产业发展，开通了二类医疗器械注册的加
19 我们是湖北省一家农业企业，想建设一些温室大棚，请问政府有什么支持政策吗？,湖北省非常支持设施农业发展！对于你们建设的智能连栋温室，政府会补
20 我们是湖北省一家民营企业，想参与军品订单，请问湖北有什么便利化措施吗？,湖北省大力支持民营企业参与军品订单！为了让你们更方便地获得相关认证，
21 请详细描述一下湖北省中小企业数据资源如何确认为资产的流程？,好的，根据湖北省的实施细则，中小企业数据资源确认为资产的流程如下：第一步是**数
22 作为一个普通市民，这份文件《公共企事业单位信息公开规定制定办法》到底是什么意思？它想做什么？,“这个文件就像是国家为了规范一些重要的公共服务
23
24 1. **建立健全信息公开制度：** 就是要让这些公共服务单位更加公开透明地运营，不能藏着掖着。
25
26 2. **深入推进信息公开：** 不只是说说，而是要真正把这些单位的信息公开展示出来，让老百姓知道。
27
28 3. **加强监督管理：** 通过信息公开，也能更好地监督这些单位，看他们是否合规合法地运营，服务质量如何。
29
30 4. **提升服务水平：** 信息公开也能倒逼这些单位提升服务质量和效率，更好地为老百姓服务。
31
32 5. **维护人民群众切身利益：** 最终目的是为了保护我们老百姓的权益，让我们能更方便地获取信息，更明白地消费服务。
33
34 6. **助力优化营商环境：** 信息公开能让营商环境更加透明公平，有利于企业的发展。”
35 这个办法里说的‘公共企事业单位’具体是指哪些单位呢？我平时生活里会接触到哪些？,“这个办法说的‘公共企事业单位’，主要指的是那些和我们日常生活紧
36
37 1. **教育：** 比如学校、幼儿园等等。
38
39 2. **医疗卫生：** 比如医院、社区卫生服务中心等等。
```

该方案不仅完成了预定任务，还在工程实践的多个层面（如并发、容错、数据解析）进行了优化，形成了一套可复用的、用于大规模语言模型数据工程的解决方案。