
# 大模型时代下的企业智能合规：风险预防、检测与应对

何家旋　阿里巴巴

# 科技生态圈峰会 + 深度研习

## ——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

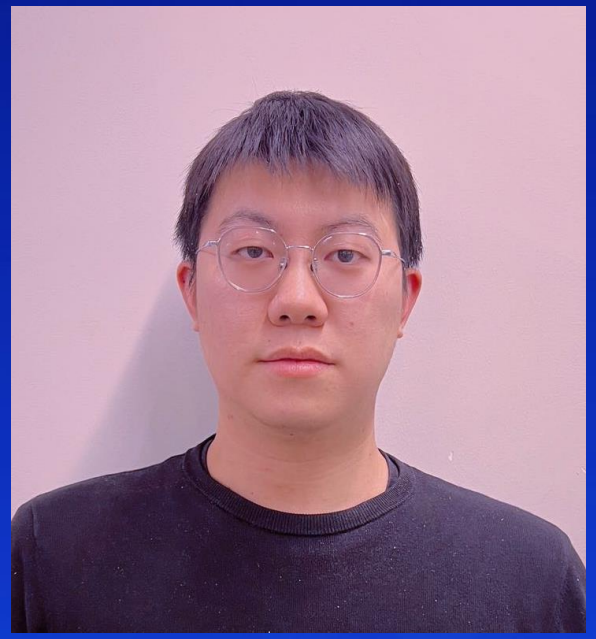
AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

# 演讲嘉宾





## 何家旋

阿里巴巴集团 算法工程师

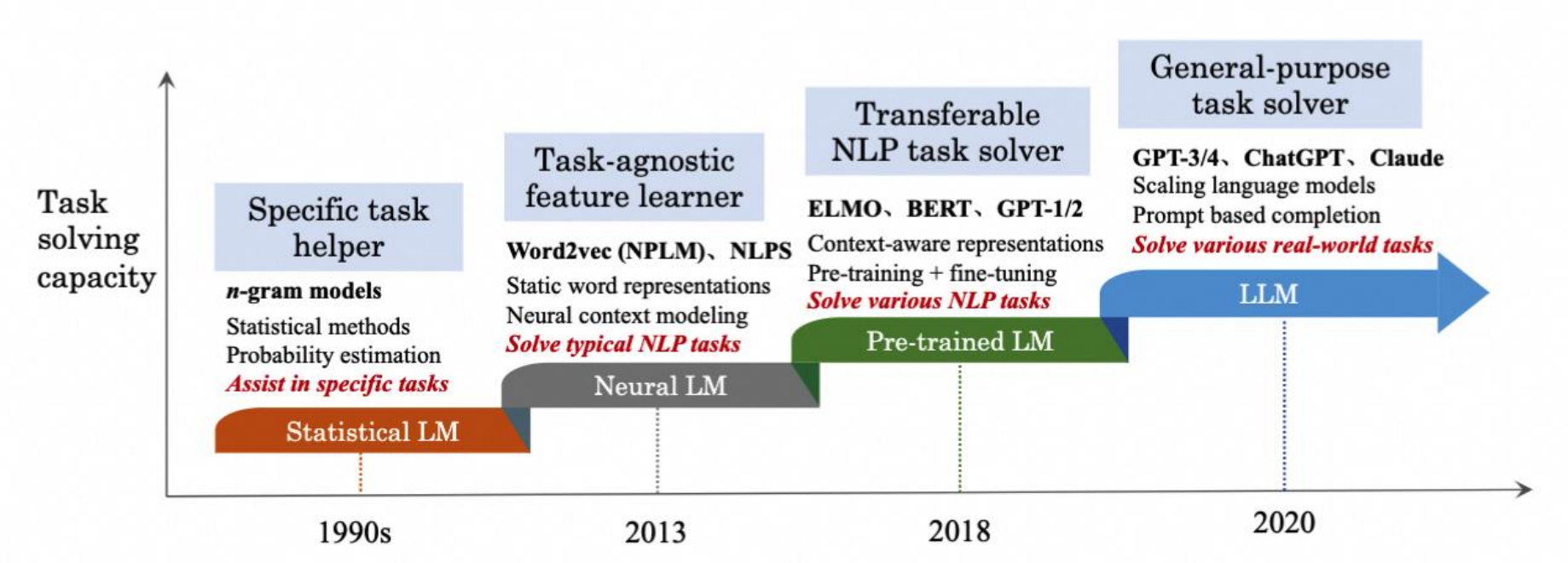
自然语言处理算法工程师，研究方向为多模态信息融合，企业智能合规，跨模态内容检索。目前负责阿里巴巴企业智能算法研发工作，负责企业智能法务合规业务，涉及审查、制裁、考试、自动收案等相关内容。发表过多篇多模态相关的SCI论文，在跨模态检索领域以及合规领域都有相关技术研究和内容沉淀。

# 目录
# CONTENTS

# PART 01
# 业务相关技术背景

# 1.1 多任务大语言模型（大语言/多模态大模型）

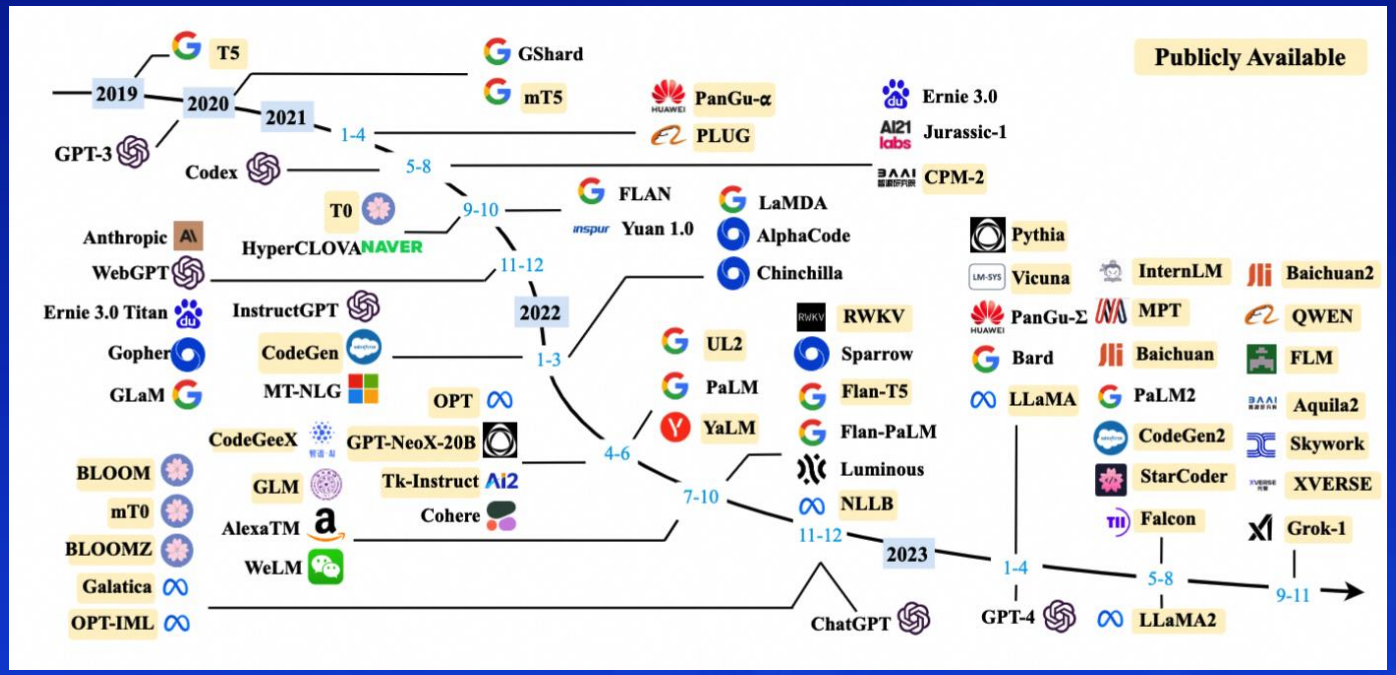


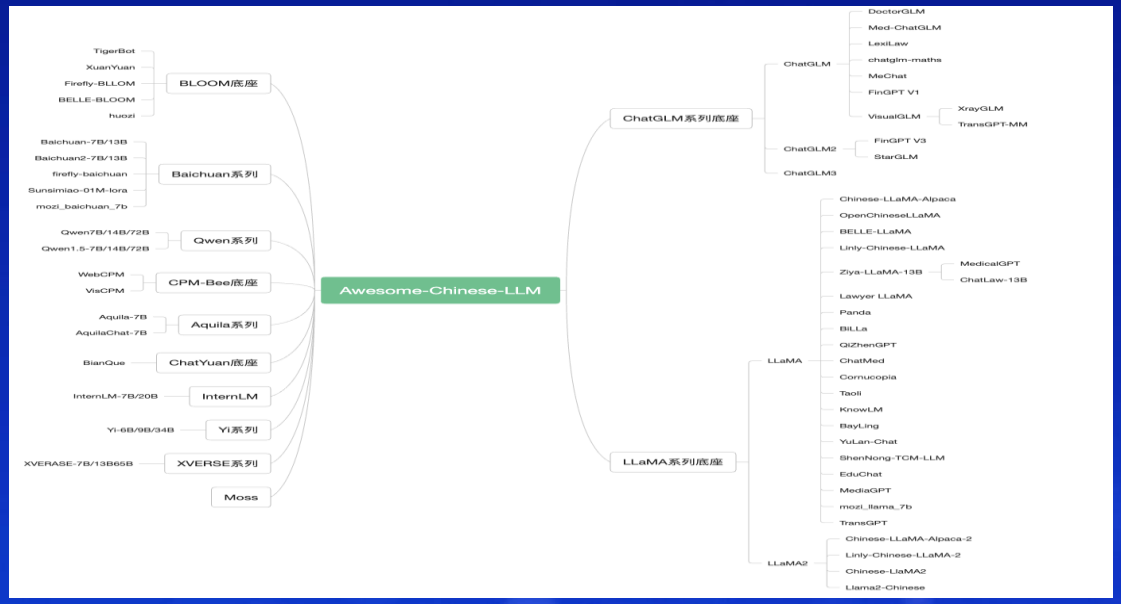



- 统计语言模型（SLM）：基于马尔科夫假设建立词预测模型，当词较多时，独热编码的维度会非常大，使得模型难以训练
- 神经语言模型（NLM）：通过引入分布式词向量来表征单词。word2vec通过一个浅层的神经网络（分为CBOW和Skip-gram两种方式）来学习分布式单词的表示。从而在向量空间中，可以通过语义的叠加和删减从一个单词得到另外一个单词（比如：皇后-女性=国王）
- 预训练模型（PLM）：transformer架构出现，ELMO，BERT的预训练-微调范式（在下游的分类头或其他适配器上进行参数训练）极大的提高了模型在各类NLP任务中的性能。
- 大语言模型（LLM）：基于decoder的GPT-3.5涌现能力。上下文学习ICL和思维链COT的特点，大模型在更多复杂任务表现出色

Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

# 1.1 多任务大语言模型（大语言/多模态大模型）




Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

https://github.com/HqWu-HITCS/Awesome-Chinese-LLM

# 1.1 多任务大语言模型（大语言/多模态大模型）



| 底座 | 包含模型 | 模型参数大小 | 训练token数 | 训练最大长度 | 是否可商用 |
|---|---|---|---|---|---|
| ChatGLM | ChatGLM/2/3 Base&Chat | 6B | 1T/1.4 | 2K/32K | 可商用 |
| LLaMA | LLaMA/2 Base&Chat | 7B/13B/33B/65B | 1T/2T | 2k/4k | 部分可商用 |
| Baichuan | Baichuan/2 Base&Chat | 7B/13B | 1.2T/1.4T | 4k | 可商用 |
| Qwen | Qwen/1.5 Base&Chat | 7B/14B/72B | 2.2T/3T | 8k/32k | 可商用 |
| BLOOM | BLOOM | 1B/7B/176B-MT | 1.5T | 2k | 可商用 |
| Aquila | Aquila/2 Base/Chat | 7B/34B | - | 2k | 可商用 |
| InternLM | InternLM/2 Base/Chat/Code | 7B/20B | - | 200k | 可商用 |
| Mixtral | Base&Chat | 8x7B | - | 32k | 可商用 |
| Yi | Base&Chat | 6B/9B/34B | 3T | 200k | 可商用 |
| DeepSeek | Base&Chat | 1.3B/7B/33B/67B | - | 4k | 可商用 |

https://github.com/HqWu-HITCS/Awesome-Chinese-LLM

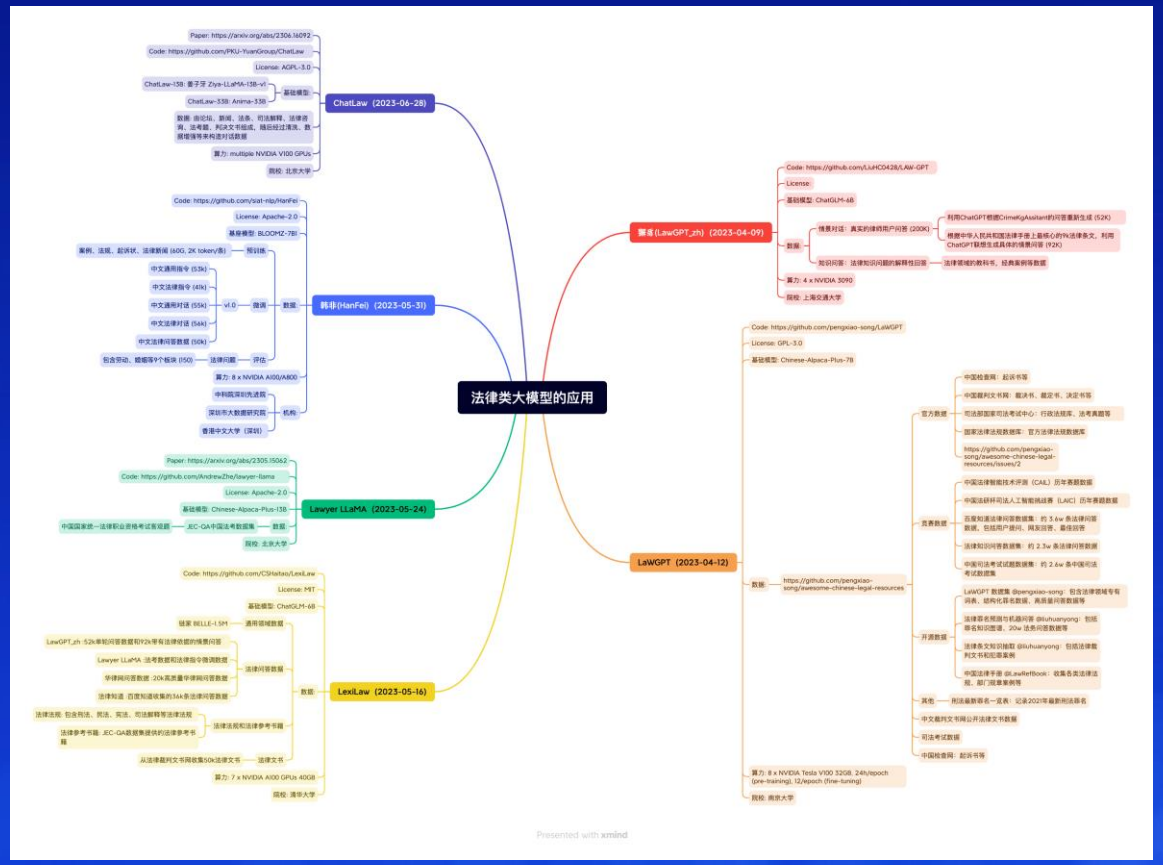
# 1.1 多任务大语言模型（大语言/多模态大模型）

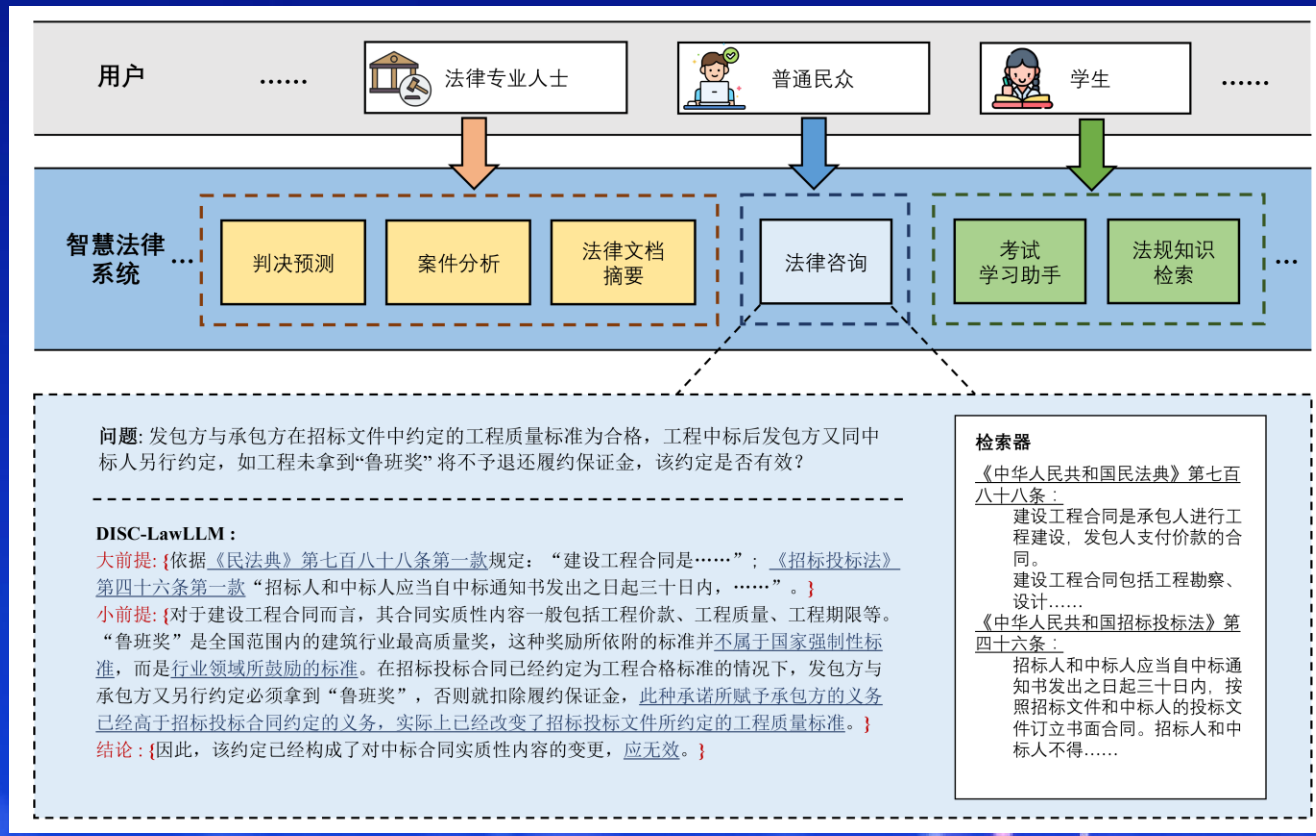







# 1.1 多任务大语言模型（法务大模型）

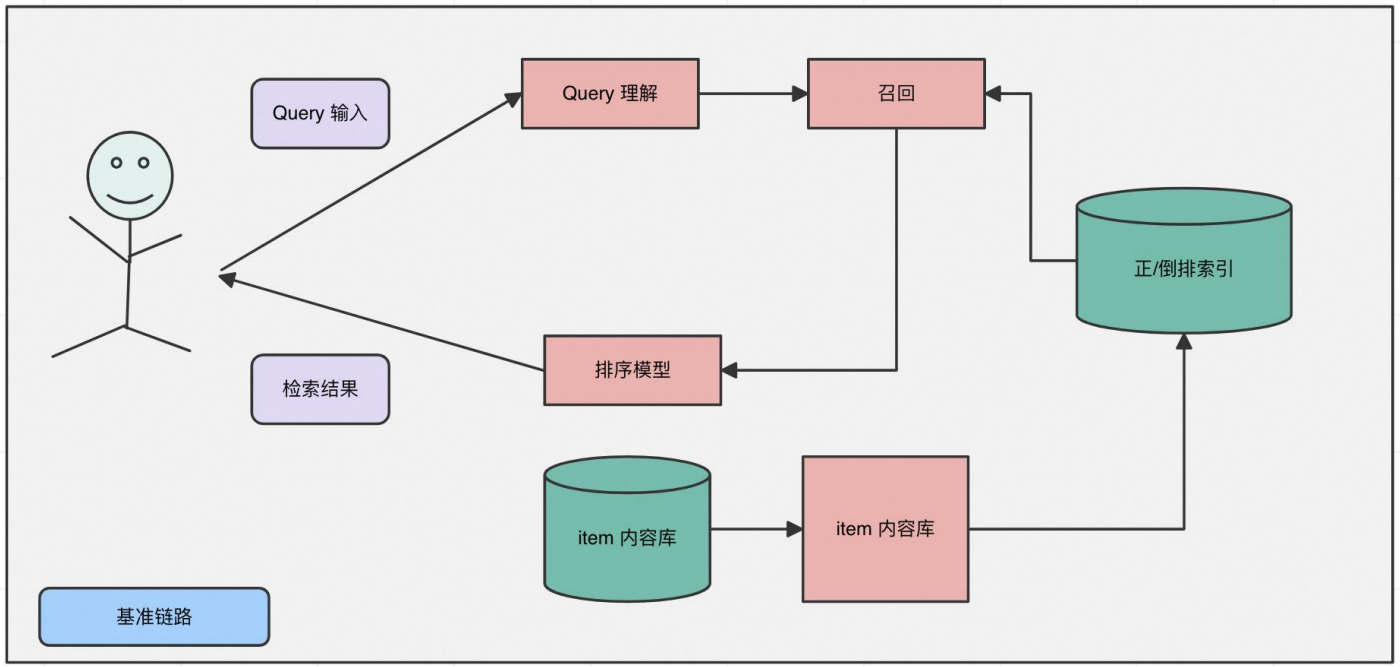



https://github.com/HqWu-HITCS/Awesome-Chinese-LLM

https://github.com/FudanDISC/DISC-LawLLM

# 1.2 检索系统相关技术




1. **Query 理解**：主观题作答需结合考生自我认知，无固定的参考答案，这类题目答案的修改主要依赖于考官个人判断，打分具有主观性
2. **召回**：参考答案可能有缩写、冗余内容，这些内容对考官无压力，而算法模型无法识别与判断
3. **排序**：问题和参考答案来自不同语言，阅卷需要考虑不同语言回答时的语境问题

# 1.3 法律文书要素(关系)抽取相关内容



## 民事起诉状

原告×××，男，19××年××月×日生，×族，住所地东莞市××区××路××号××室。电话：139××××××××

被告×××，女，19××年××月×日生，×族，住所地东莞市××区××路××号××室。电话：136*****

### 诉讼请求

1、请求法院判令被告支付原告欠款人民币***元（大写：***）；

2、请求法院判令被告支付原告以上欠款的利息***；

3、诉讼费用全部由被告承担。

### 事实与理由

原告与被告系（朋友/同事/××）关系。截至××年××月××日止，被告累计向原告借款金额为人民币***元。现还款期限已届满，被告拒不还款。

原告认为，原被告之间的借款合同关系合法有效，双方都应当诚实信用地履行合同义务，但现被告无正当理由拒不归还欠款，依法应当承担违约责任。据此，为维护原告的合法权益，依照《中华人民共和国合同法》与《中华人民共和国民事诉讼法》相关规定诉至贵院，请依法支持原告的诉请。

此致

XX市人民法院

起诉人：***

××年××月××日

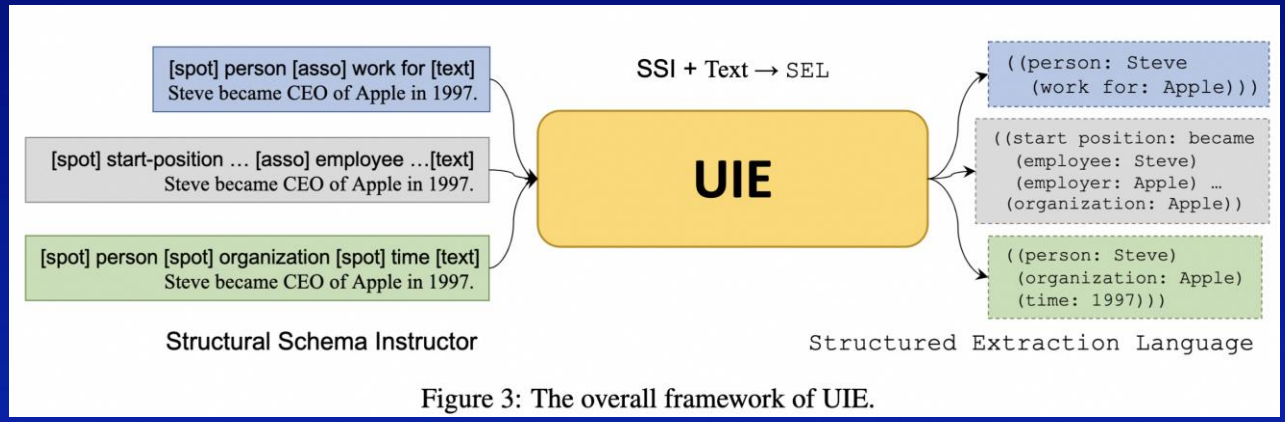



Figure 3: The overall framework of UIE.

## 通过微调UIE抽取起诉状中信息，提升法务收案智能化

**起诉状信息抽取**

原告张三，男，19××年××月×日生，×族，住所地东莞市××区××路××号××室。电话：139××××××××

被告李四，女，19××年××月×日生，×族，住所地东莞市××区××路××号××室。电话：136*****

诉讼请求

1、请求法院判令被告支付原告欠款人民币***元（大写：***）；

2、请求法院判令被告支付原告以上欠款的利息***；

3、诉讼费用全部由被告承担。

事实与理由

原告与被告系（朋友/同事/××）关系。截至××年××月××日止，被告累计向原告借款金额为人民币***元。现还款期限已届满，被告拒不还款。

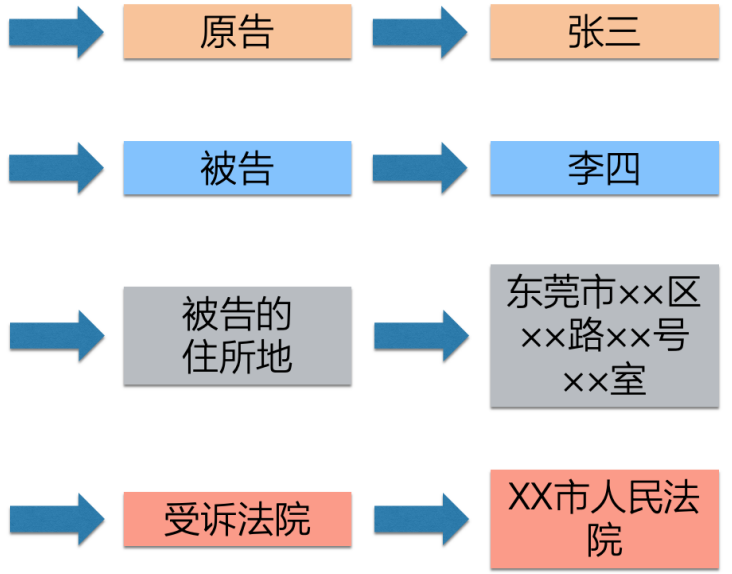
原告认为，原被告之间的借款合同关系合法有效，双方都应当诚实信用地履行合同义务，但现被告无正当理由拒不归还欠款，依法应当承担违约责任。据此，为维护原告的合法权益，依照《中华人民共和国合同法》与《中华人民共和国民事诉讼法》相关规定诉至贵院，请依法支持原告的诉请。

此致

XX市人民法院

起诉人：***

××年××月××日

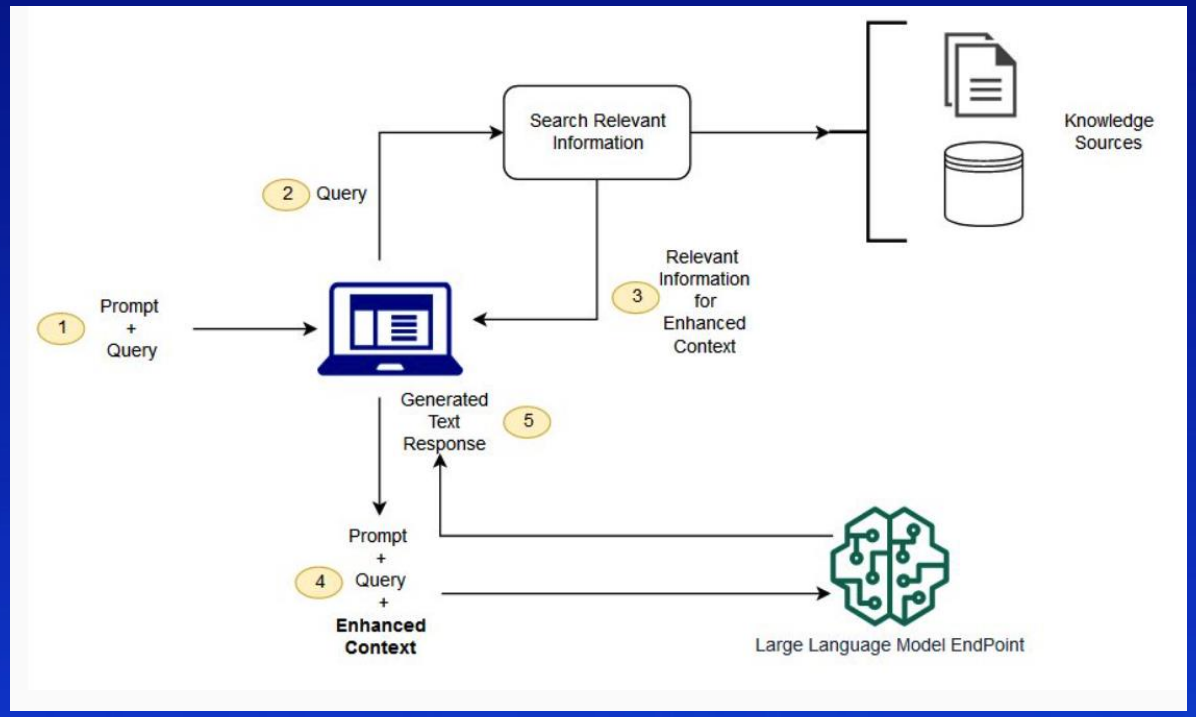


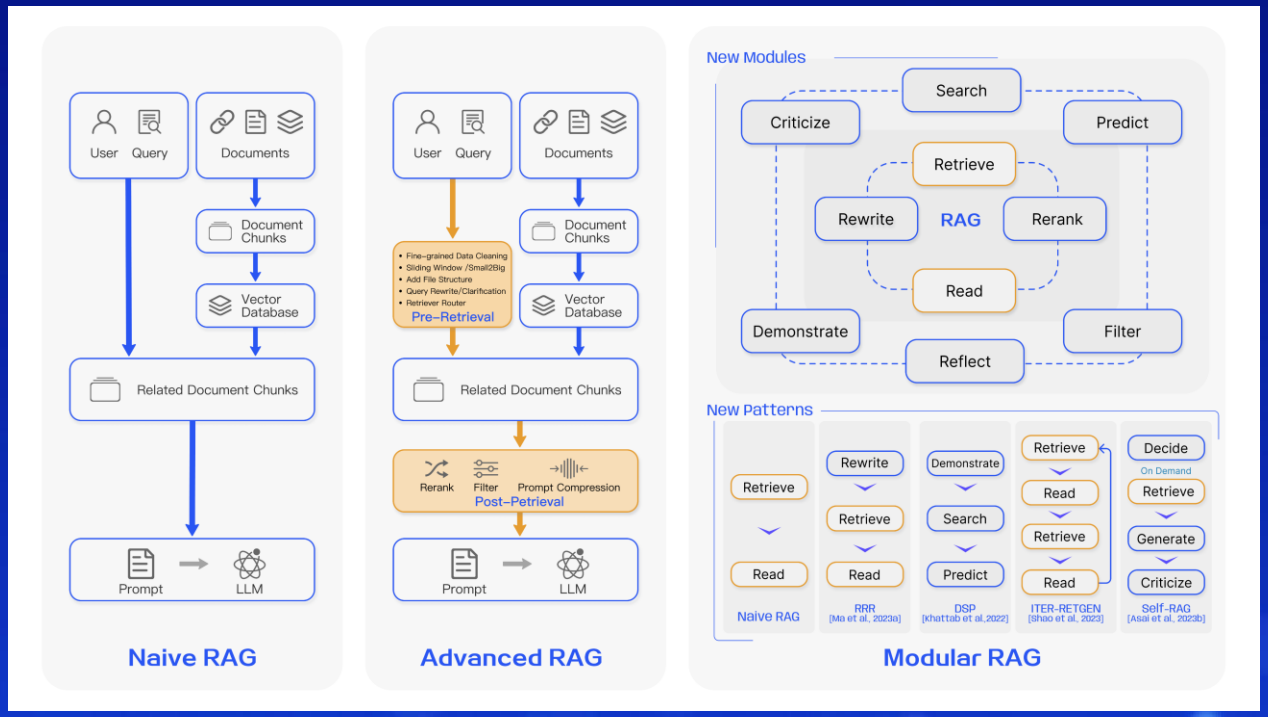

Lu Y, Liu Q, Dai D, et al. Unified structure generation for universal information extraction[J]. arXiv preprint arXiv:2203.12277, 2022.

PART 02

# 基于大语言模型的自然语言处理任务

# 2.1 基于大模型的检索增强生成RAG



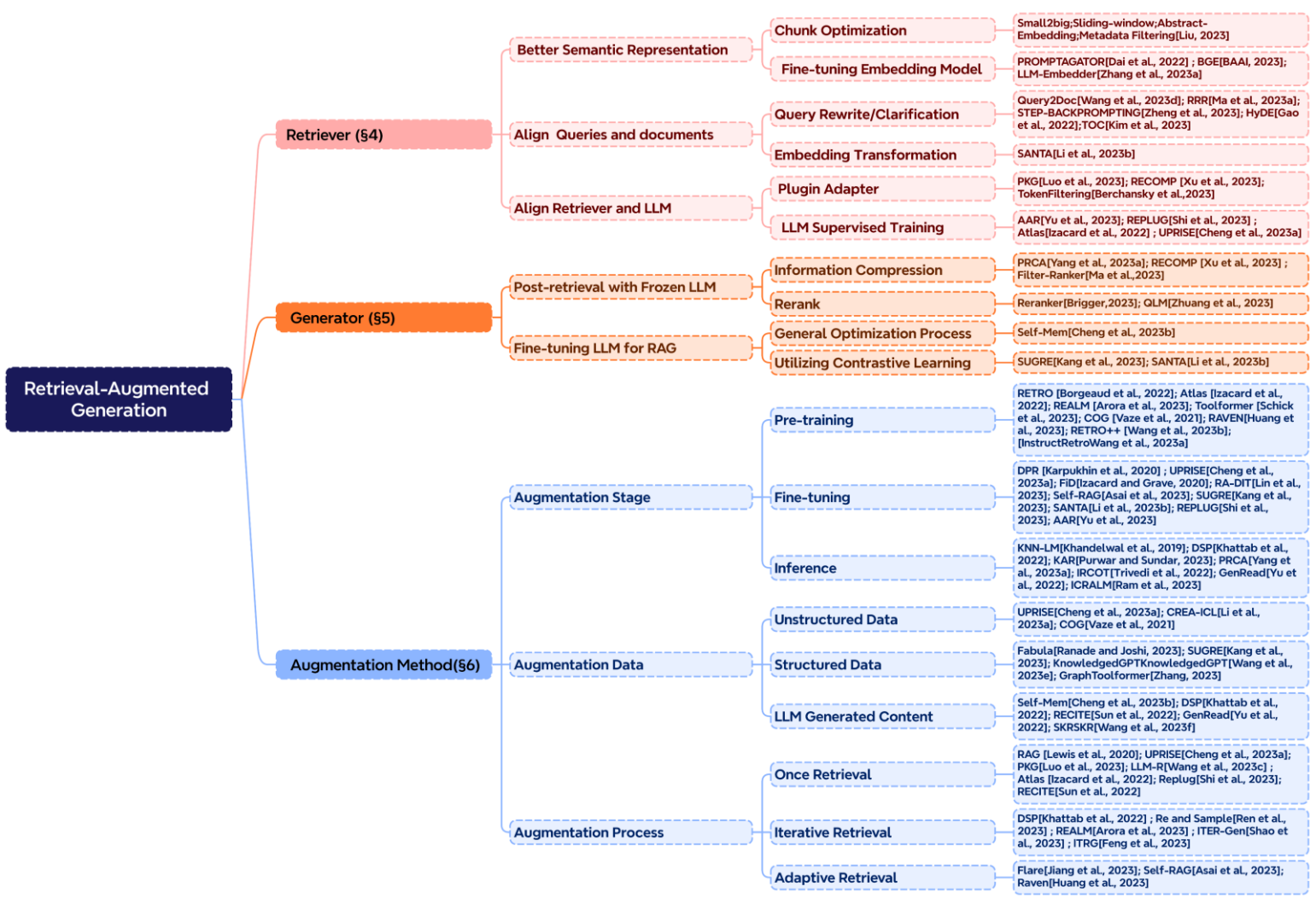




RAG-检索增强生成： 模型在回答问题或生成文本时，首先从广阔的文档库中寻找相关信息。然后，模型使用这些找到的信息来生成回答或文本，从而提高其预测的准确度。

1. 使用编码模型（如 BM25、DPR、ColBERT 等）根据问题找到相关的文档。
2. 生成阶段：以找到的上下文作为基础，系统生成文本

https://arxiv.org/abs/2312.10997
https://aws.amazon.com/cn/what-is/retrieval-augmented-generation/

# 2.1 基于大模型的检索增强生成RAG



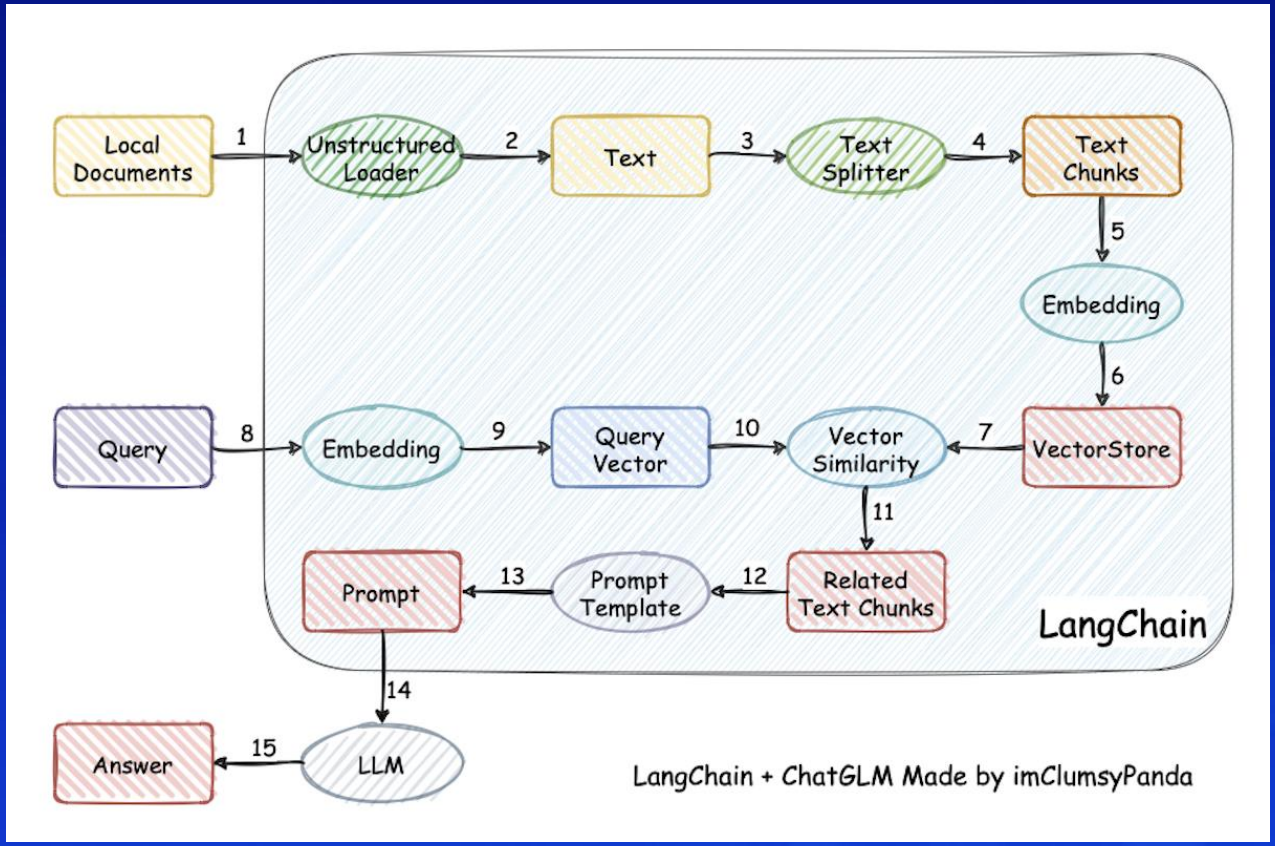

RAG-检索增强生成：

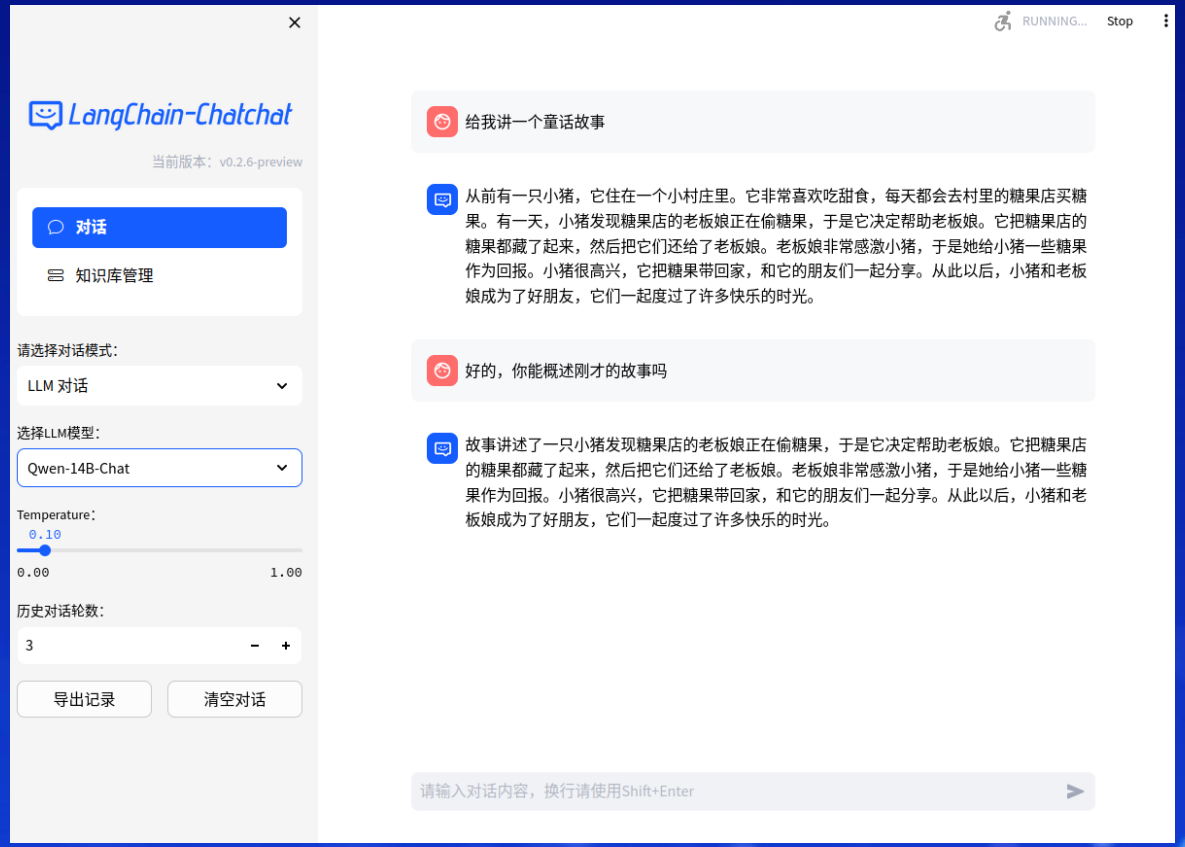
1. 检索: RAG首先通过一个检索系统（例如Dense Passage Retrieval, DPR）从一个大规模的文档数据库中检索出与输入查询最相关的文档。
2. 上下文选择: 模型选取检索到的文档中的相关片段，并将其与原始查询一起作为上下文送入到下一步骤。
3. 序列生成: 使用LLM结合原始输入和检索到的上下文，生成答案或者续写文本。
4. 边缘概率计算: 对于生成的每个词，计算其在所有检索到的文档片段上的边缘概率。
5. 后处理和输出: 选择概率最高的生成序列作为最终输出，完成整个生成任务。

https://arxiv.org/abs/2312.10997

# 2.1 基于大模型的检索增强生成RAG






https://github.com/chatchat-space/Langchain-Chatchat

# 2.2 大模型结合命名实体识别



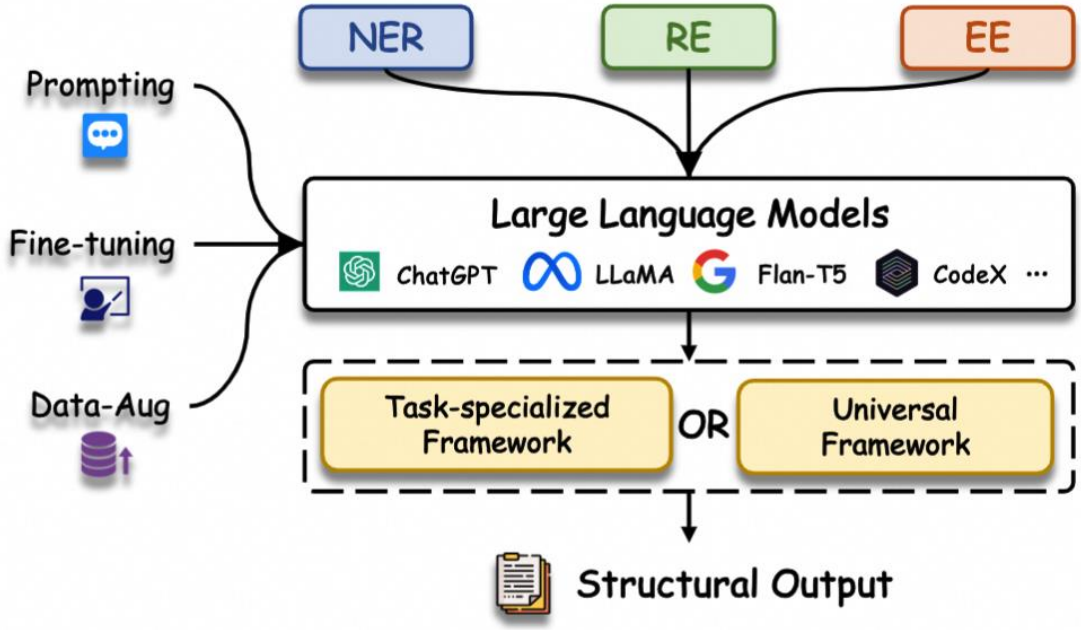
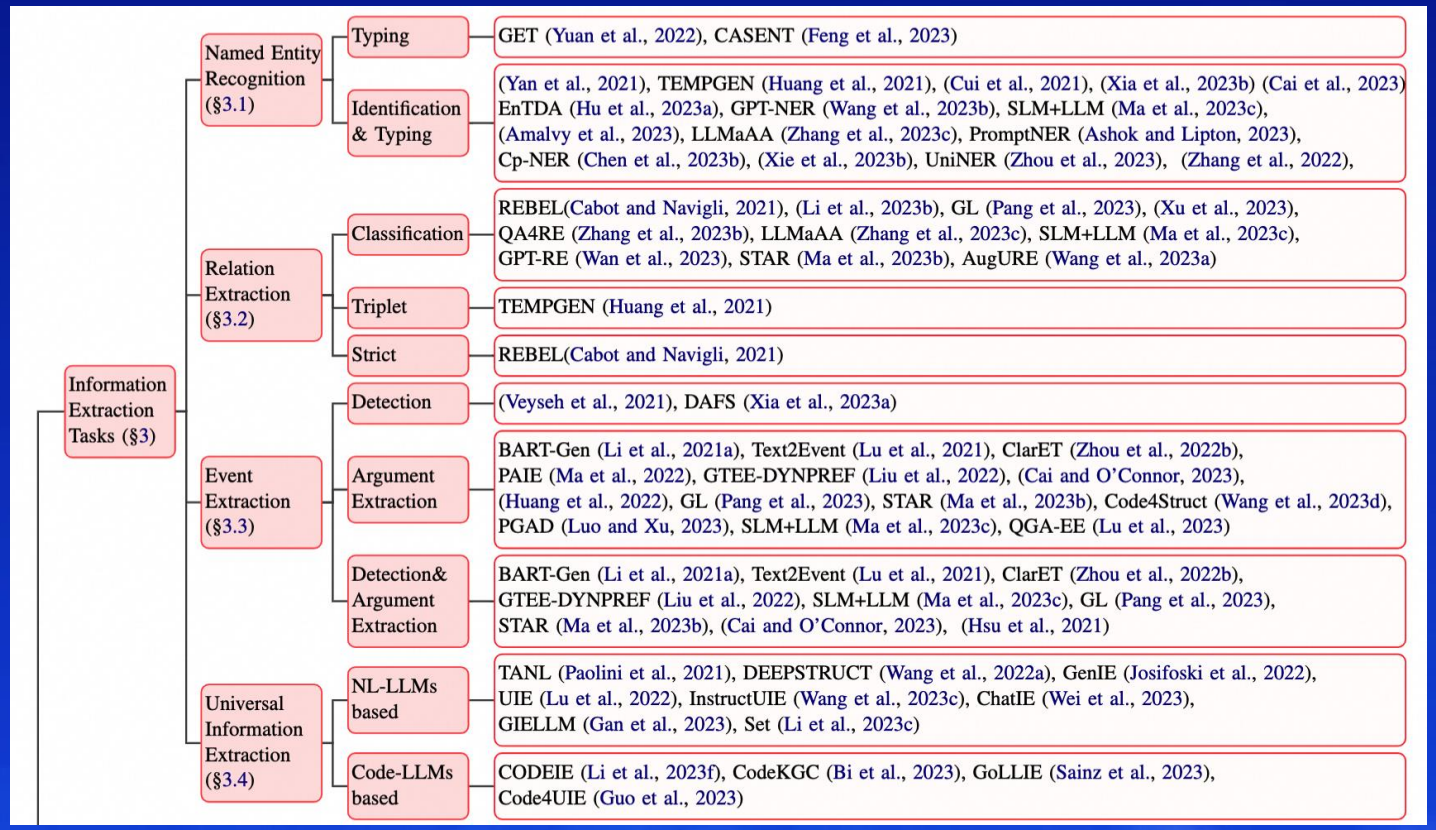



Figure 1: LLMs have been extensively explored for generative IE. These studies encompass various learning paradigms, diverse LLM architectures, and specialized frameworks designed for a single subtask, as well as universal frameworks capable of addressing multiple subtasks simultaneously.

1. **Named Entity Recognition**：主观题作答需结合考生自我认知，无固定的参考答案，这类题目答案的修改主要依赖于考官个人判断，打分具有主观性
2. **Relation Extraction**：参考答案可能有缩写、冗余内容，这些内容对考官无压力，而算法模型无法识别与判断
3. **Event Extraction**：问题和参考答案来自不同语言，阅卷需要考虑不同语言回答时的语境问题
4. **Universal Information Extraction**：如合规、业务、个人成长相关，并且部分题目有明显的垂类特征，对考官评分也有一定要求

# 2.2 大模型结合命名实体识别





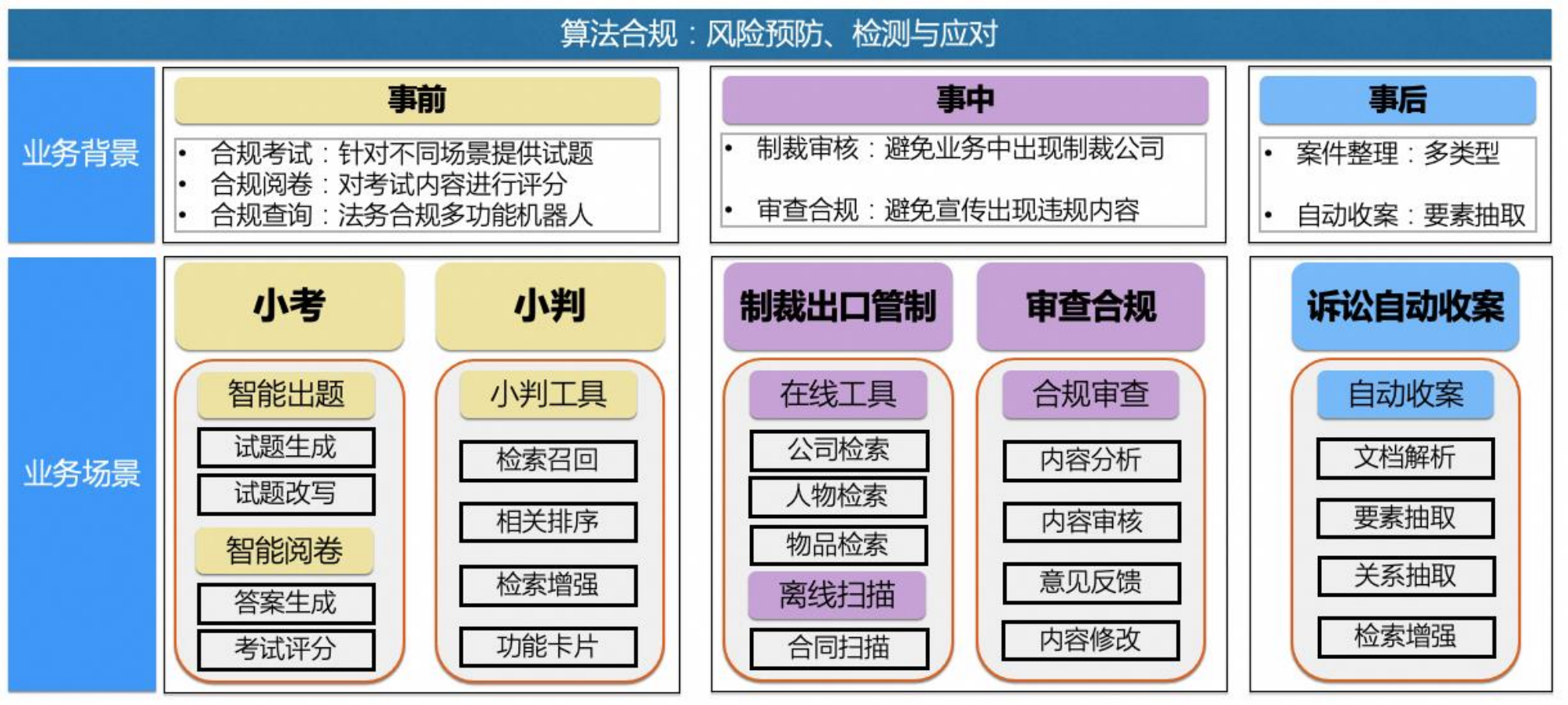

1. **Named Entity Recognition**：主观题作答需结合考生自我认知，无固定的参考答案，这类题目答案的修改主要依赖于考官个人判断，打分具有主观性
2. **Relation Extraction**：参考答案可能有缩写、冗余内容，这些内容对考官无压力，而算法模型无法识别与判断
3. **Event Extraction**：问题和参考答案来自不同语言，阅卷需要考虑不同语言回答时的语境问题
4. **Universal Information Extraction**：如合规、业务、个人成长相关，并且部分题目有明显的垂类特征，对考官评分也有一定要求

Large Language Models for Generative Information Extraction: A Survey

# PART 03
# 企业智能法务合规实践

# 3 框架大图






1. **事前**：合规考试、合规阅卷、查阅
2. **事中**：制裁管制、合规审核
3. **事后**：案件整理、内容收案

本质

1. **提示工程与小模型**：**合规考试、合规审核**
2. **内容检索与大模型**：**制裁管制、合规阅卷**
3. **信息抽取与大模型**：**案件整理、内容收案**

# PART 3-1
# 提示工程与模型组合

# 3.1 合规考试与审核




提升合规意识


验证学习内容



企业合规培训至关重要，确保员工明确法律界限。培训核心包括：

- **合规学习：简化材料，提取知识点，总结复习。**
- **合规测试：考题定制，评估员工合规学习情况。**
- **合规审核：营销合规，广告宣传是否符合法律。**

小考/考呀考试系统支持企业内外合规测试，广泛服务集团员工和合作商家。企业智能算法团队已在小考的应用场景中实现智能化探索，包括试题智能生成、改写和主观题自动评分，与小考团队协作推动自动化和智能化进程。

# 3.1.1 合规学习与评测：知识点提取与智能出题

## 文档标题：基础生物学概念复习

### 生物学定义

生物学是自然科学的一个分支，专注于研究生命和生物体的结构、功能、生长、起源、进化和分布。

### 细胞理论

细胞理论指出所有生命体均由一个或多个细胞构成，细胞是生命的基本单位，以及细胞只能由其他细胞分裂产生。

### 遗传学

遗传学是研究基因、遗传变异和生物体特征遗传的科学。门捷列夫通过实验揭示了遗传的基本规律，其中包括显性和隐性遗传。

### 生态系统

生态系统是由相互作用的生物群落和它们的非生物环境组成的系统。生态系统的核心概念包括食物链、能量流动和生物的循环。

## 主要知识点提取

1. 生物学是研究生命和生物体的自然科学分支。
2. 细胞理论包括所有生物体由细胞构成，细胞是生命的基本单位，细胞由细胞分裂产生。
3. 遗传学关注基因、遗传变异和特征的遗传，其中包括显性和隐性遗传规律。
4. 生态系统由生物群落和非生物环境组成，涉及食物链、能量流动和循环。

## 例题生成

### 多选题

问题：以下哪些陈述属于遗传学的基本组成部分？（多选）
A) 遗传变异可以导致种群多样性
B) 基因是遗传的基本单元
C) 生物体特征发生变化是随机的
D) 显性和隐性遗传是遗传的两个基本规律

正确答案：A, B, D

### 判断题

问题：生态系统只由生物群落构成，并不包括非生物环境。（对/错）

正确答案：错

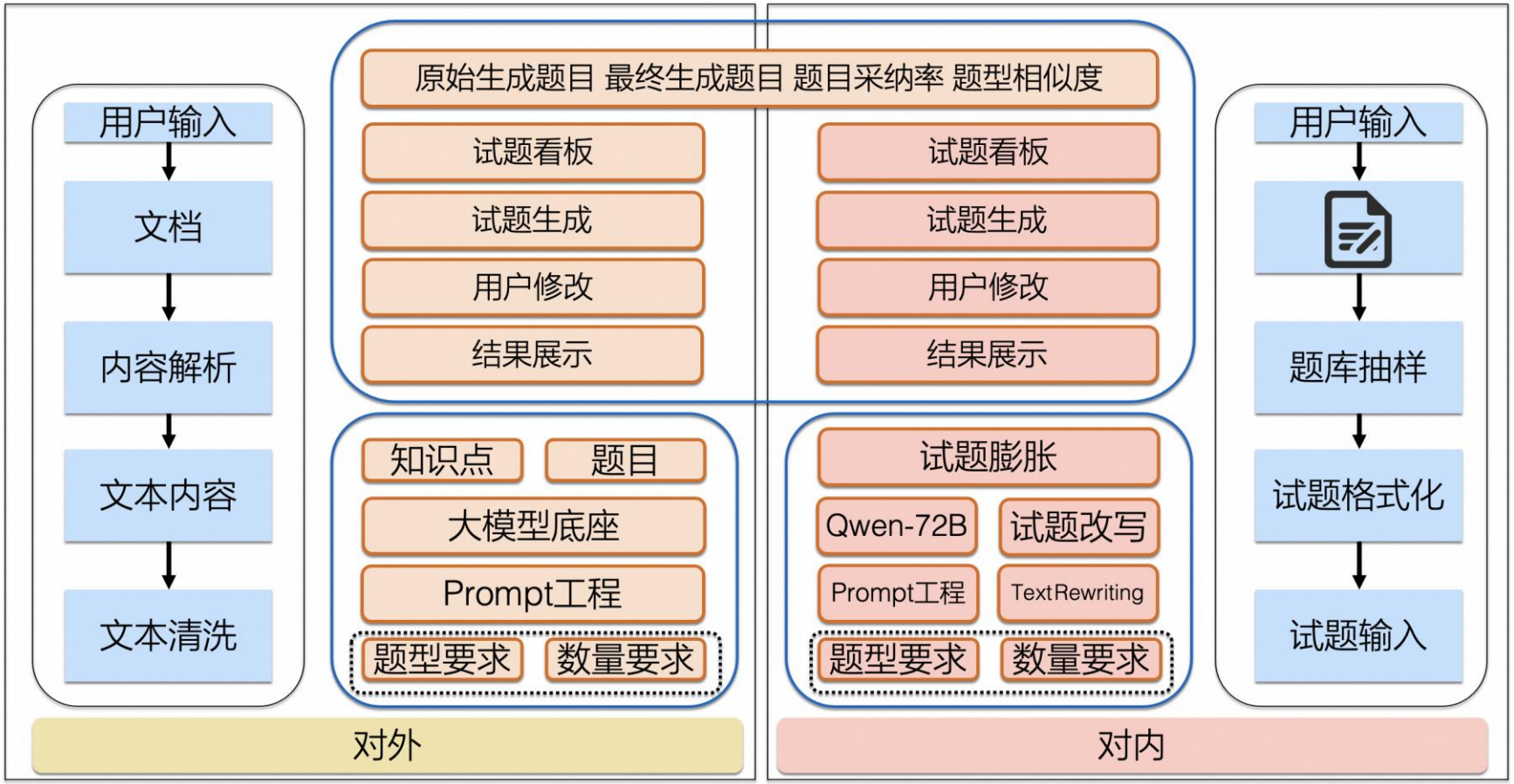
### 填空题

问题：生物学是自然科学的一个分支，专注于研究_____和生物体的结构、功能、生长、起源、进化和分布。

正确答案：生命

以上是一个示例，生成特定考试文档所需的知识点提取和习题创建依赖于具体的科目和内容。针对不同科目和专业领域，可以采用类似的框架来构建有针对性的文档和辅导习题。

# 3.1.1 合规学习与评测：知识点提取与智能出题



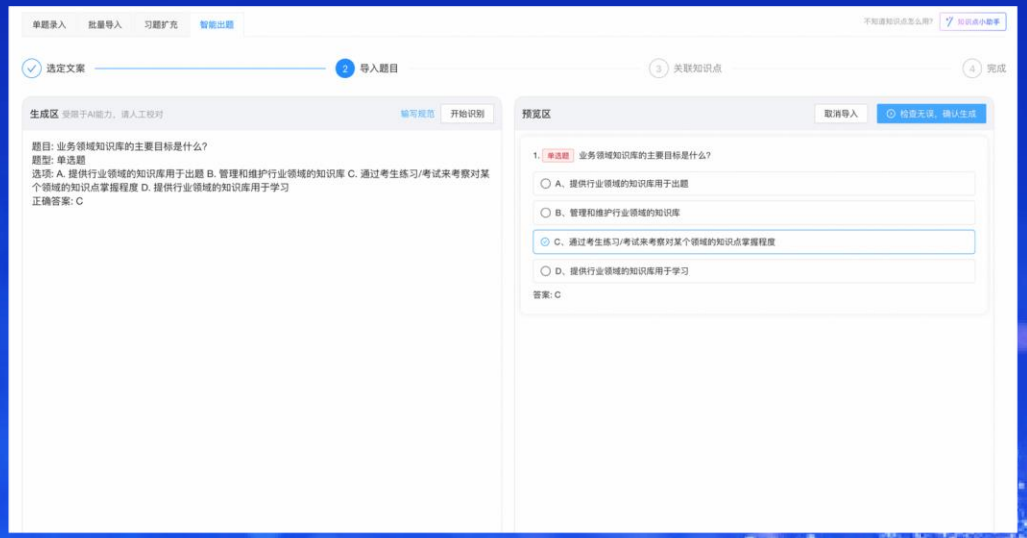

1. **特定领域出题**：用户提供关键词，系统检索现有文档库并提取top-k相关文档，生成相关问题。算法任务融合了检索、摘要以及文本生成的步骤。
2. **结合材料出题**：系统对用户提交的材料执行知识点提取，类似内容摘要。基于知识点，系统会进一步生成相应的问题。

# 3.1.1.1 知识点抽取与智能出题



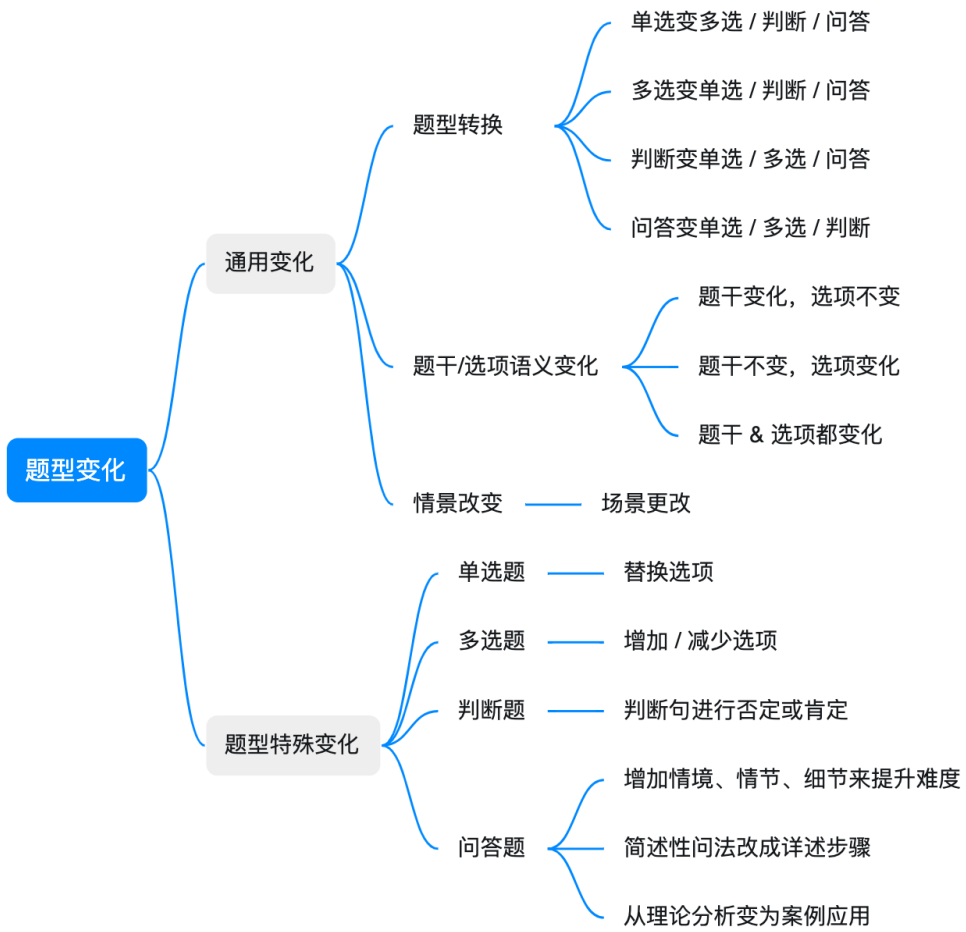

方式一：粘贴文本-快速出题


方式二：选择课件-领域出题




生成试题-人工修改

# 3.1.1.2 试题膨胀




现有问题：

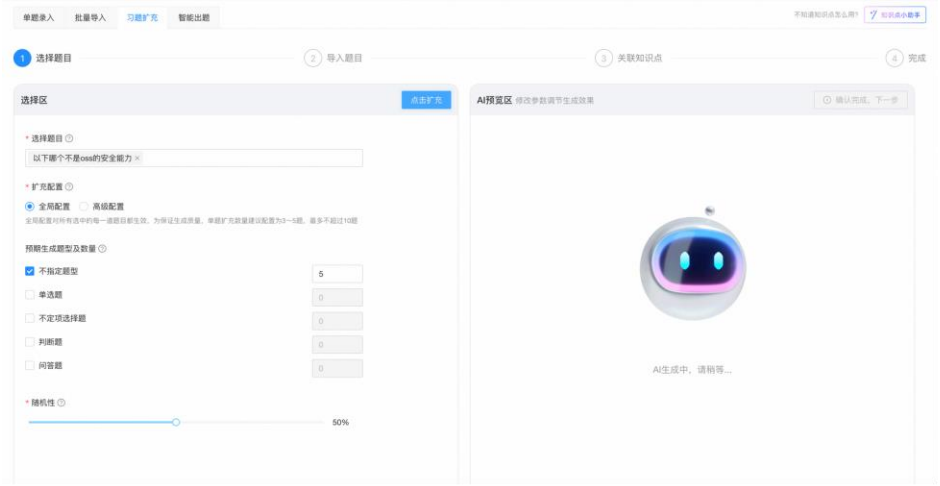
- 题库扩展：特定场景下单个知识点生成题目有限，无法满足无限训练的需求。
- 题型丰富："小考"目前主要包括单选题、多选题、判断题和问答题四种类型。通常，不是所有题型都涵盖了所有知识点，这限制了题型多样性。

技术探索：

- 数据增强：利用EDA（Exploratory Data Analysis）技术进行题目改写，但这种方法可能会改变题干意义，降低题目质量，并引入歧义。
- 文本复述：尝试通过微调中文文本复述模型来改写题目，但受限于模型性能，常出现编码错误、无关文本或控制不佳的复述长度，导致不理想的结果。
- 大语言模型SFT：在现有题库中，同一知识点下不同题型的关联性不强，缺乏多样的转换示例，这限制了基于SFT（Supervised Fine-Tuning）的应用范围。

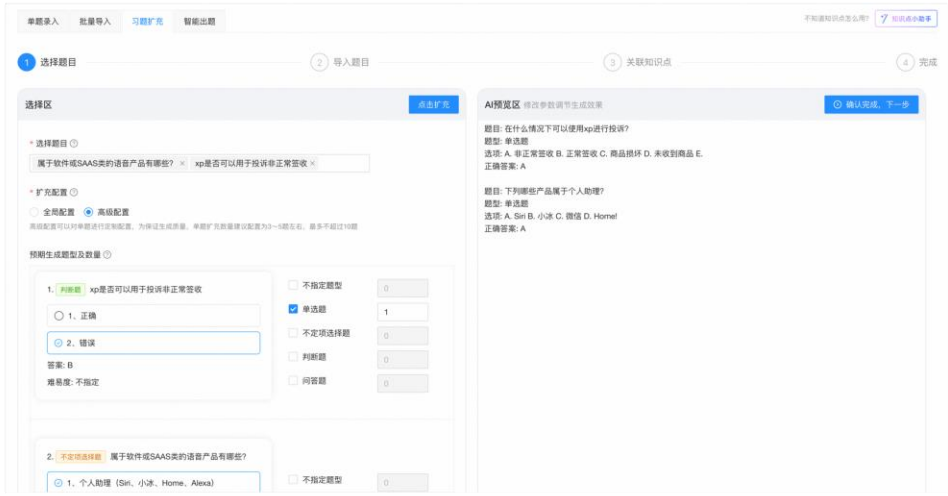
文本复述模型和数据增强工具主要用于选择题改写，对于更复杂的题目，我们结合Qwen-72B，重点聚焦在prompt工程，通过对长文档类型的题目输入进行改写，规范输出格式，进而实现四个题型间的多题型和多数量的题目转换。

# 3.1.1.2 试题膨胀

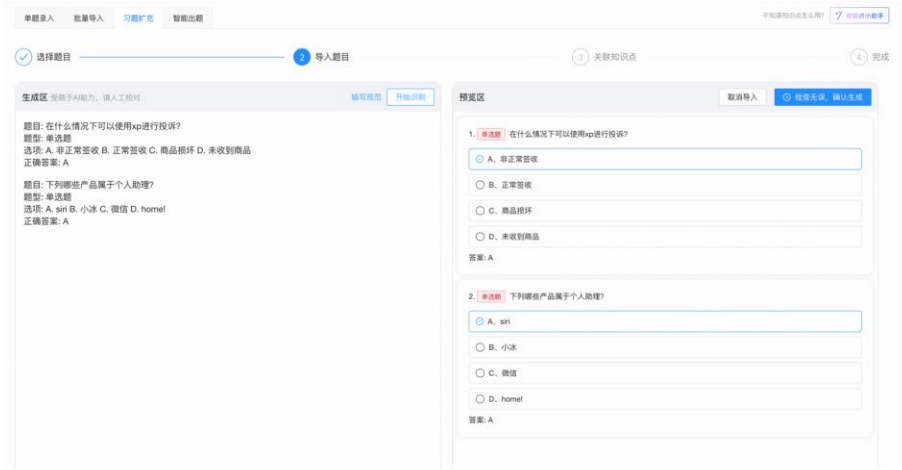



1.试题生成界面


2.多题目输入全局配置

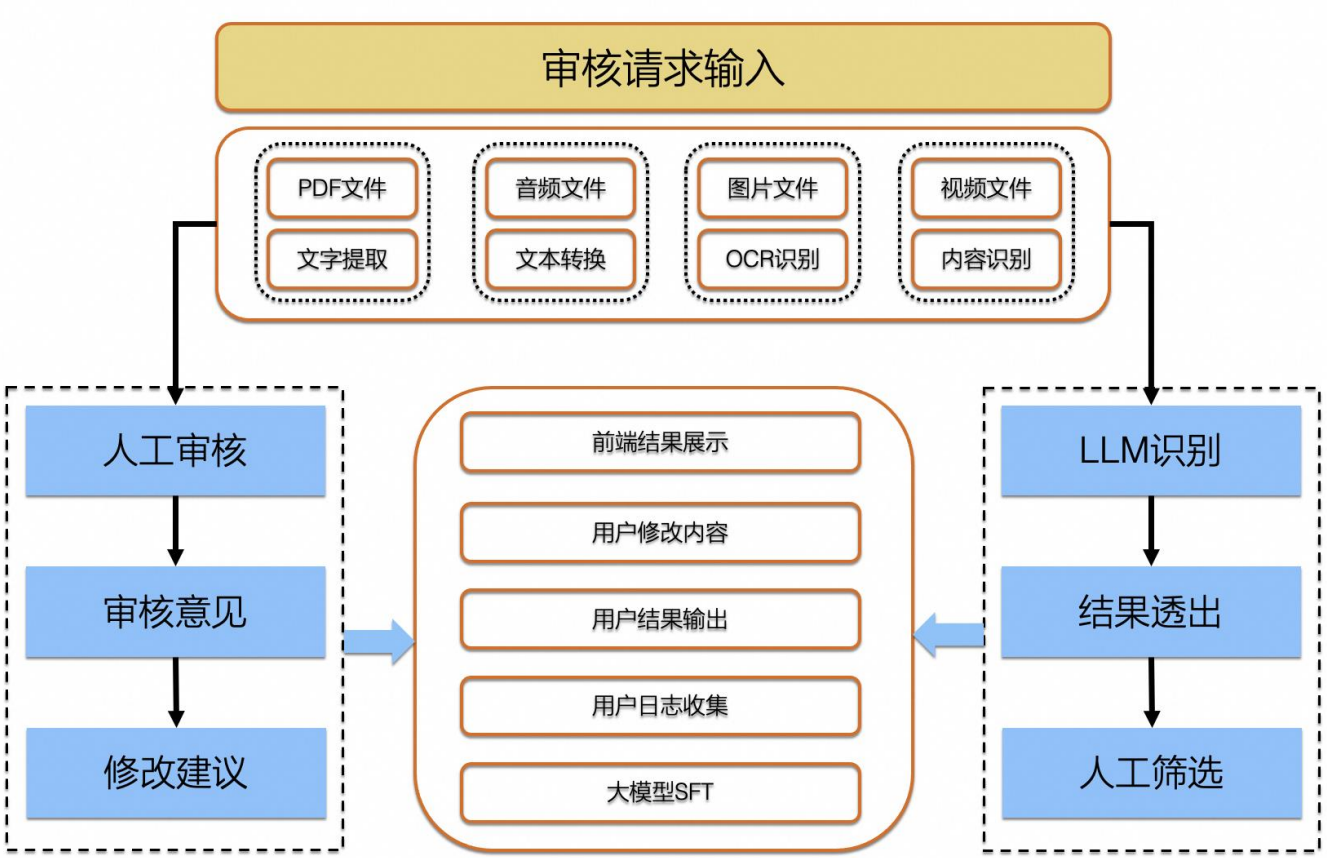



3、单试题配置



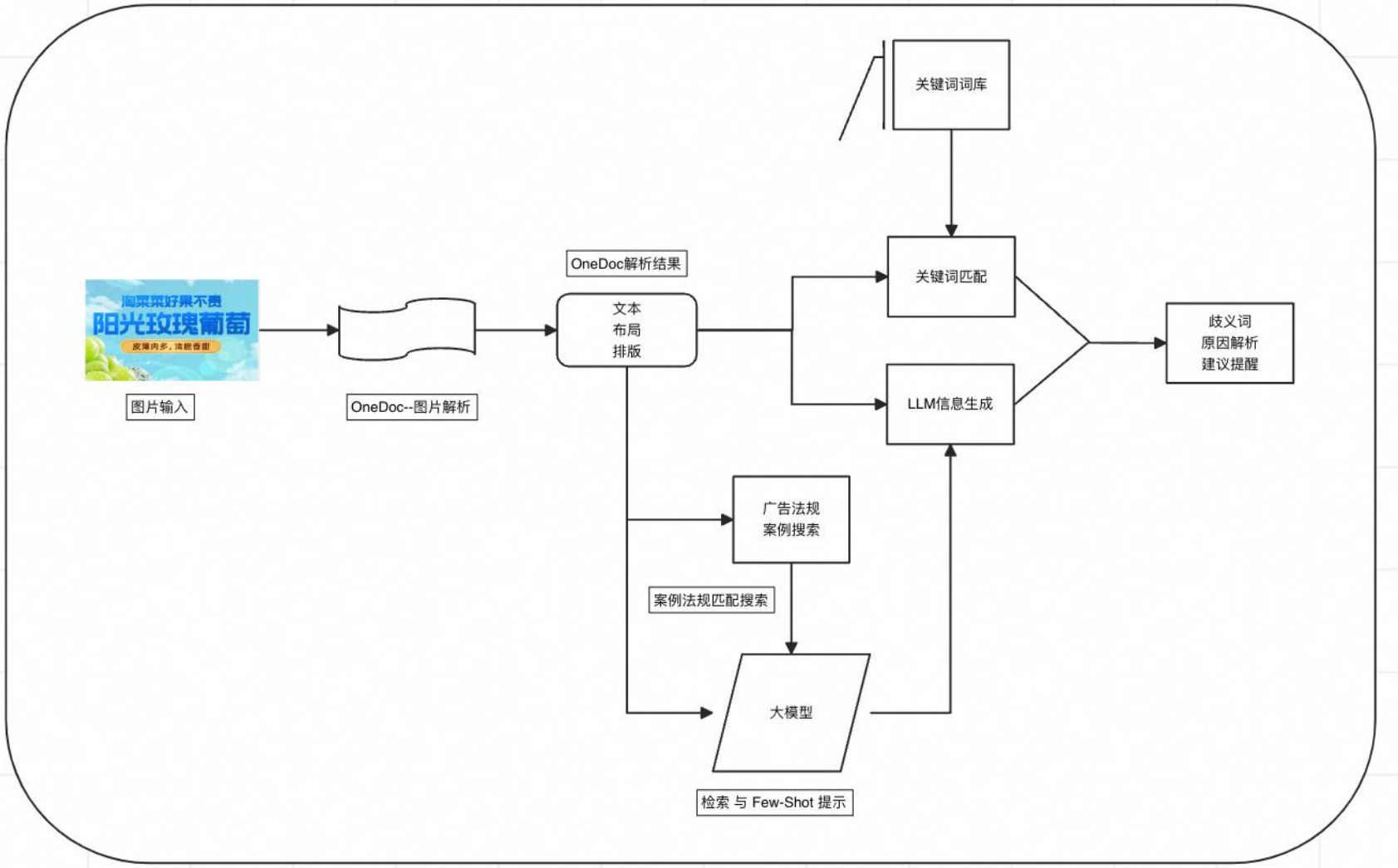

4、支持生成试题后进行人工修改

# 3.1.2 合规审核宣传审查合规




1. 一方面，采用关键词库的形式，命中文本中的关键词并标记，得到相应的修改意见（关键词和对应的词库都维护在词库中并持续更新）。
2. 另一方面，将OneDoc生成的文本输入到LLM（实验时分别使用了Qwen、ChatGLM、GPT-3.5、GPT-4、若干法务大模型等等，最终决定先采用Qwen）中，通过设计prompt工程，要求LLM输出审核后发现的问题以及相关的修改意见

# 3.1.2 合规内容之宣传审查合规



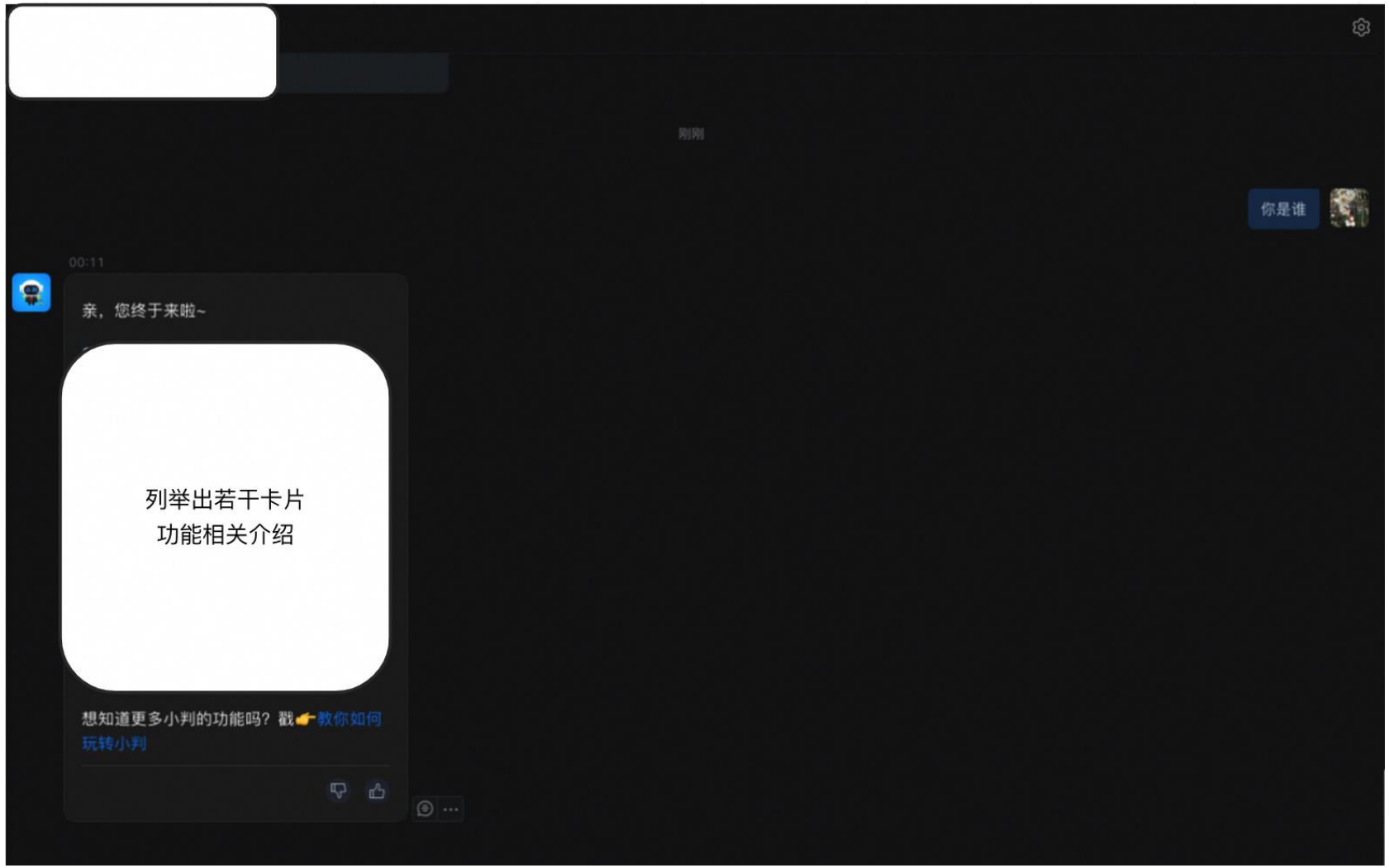

1. 内容处理：识别相关的文本内容
2. 关键词词库：内容筛选匹配。
3. 大模型检索：检索案例，给出原因，提出建议。

# PART 3-2
# 内容检索与大模型

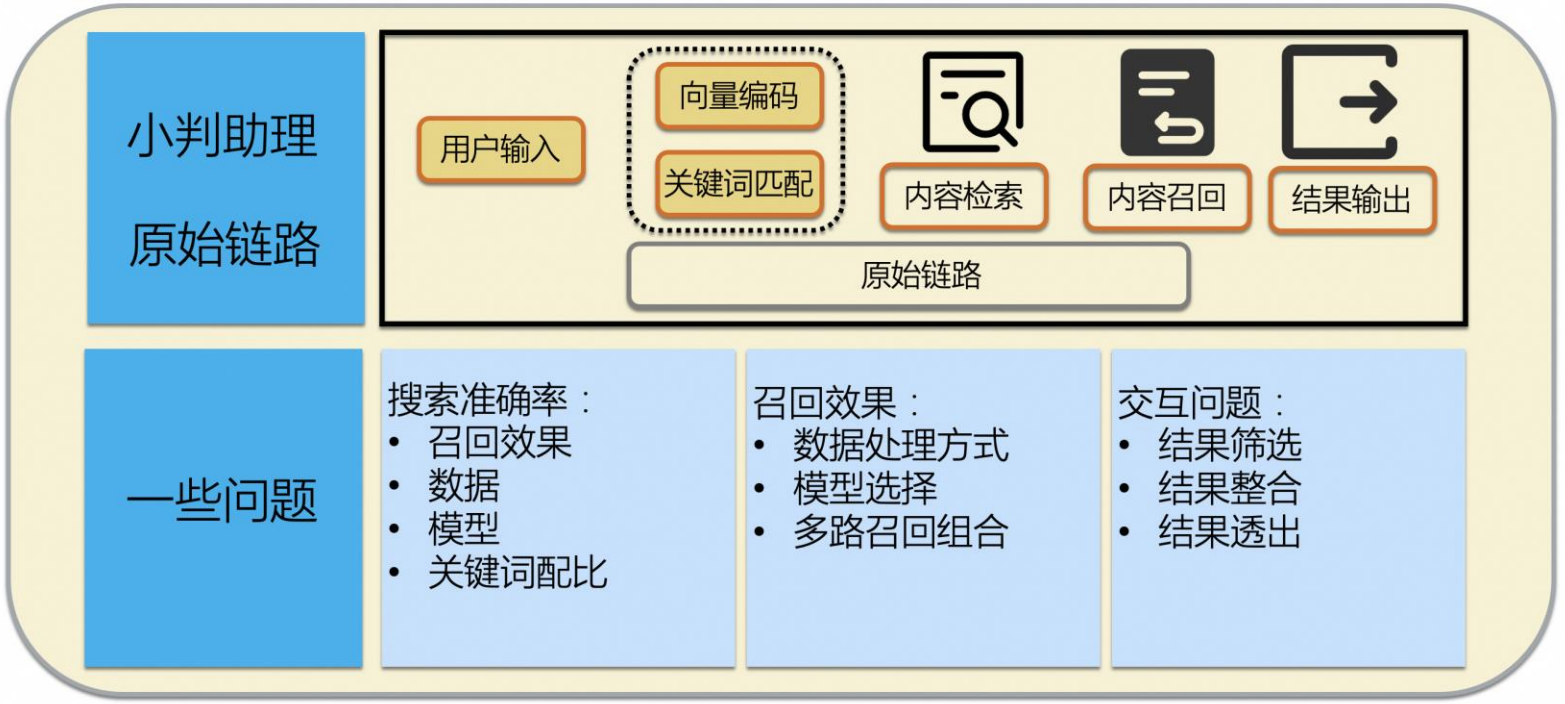
# 3.2.1 企业智能小判检索优化



# 3.2.1 企业智能小判检索优化



# 3.2.1 企业智能小判检索优化




1. **链路简单**：原始链路在接收用户输入后，仅有向量编码及关键词正则匹配进行检索召回得到结果。
2. **交互简单**：原始链路在检索到相关内容后，通过小判透出结果，交互简单，上下文理解能力有限，在需要进行多个候选文档内容整合时，知识整合能力不足。随着LLM的火热，人机交互掀起潮流，这对小判的优化是个启发。
3. **模型简单**：使用的向量编码模型较为简单，需要扩充多种场景内容和服务优化。

# 3.2.1 企业智能小判链路优化




数据处理:

- 针对现有业务数据集进行清洗和整合，对业务数据分布不均衡问题，我们尝试结合大模型和文本改写策略增强数据。
- 长文本分割，结合段落/章节、滑动窗口处理、主体分割（如文本聚类）、摘要抽取等来切分文本，得到完整语义的片段信息。

内容召回:

- query召回内容，采用更高维度双塔模型编码向量。
- 构造难负样本，多路召回，实现了更精确的内容召回

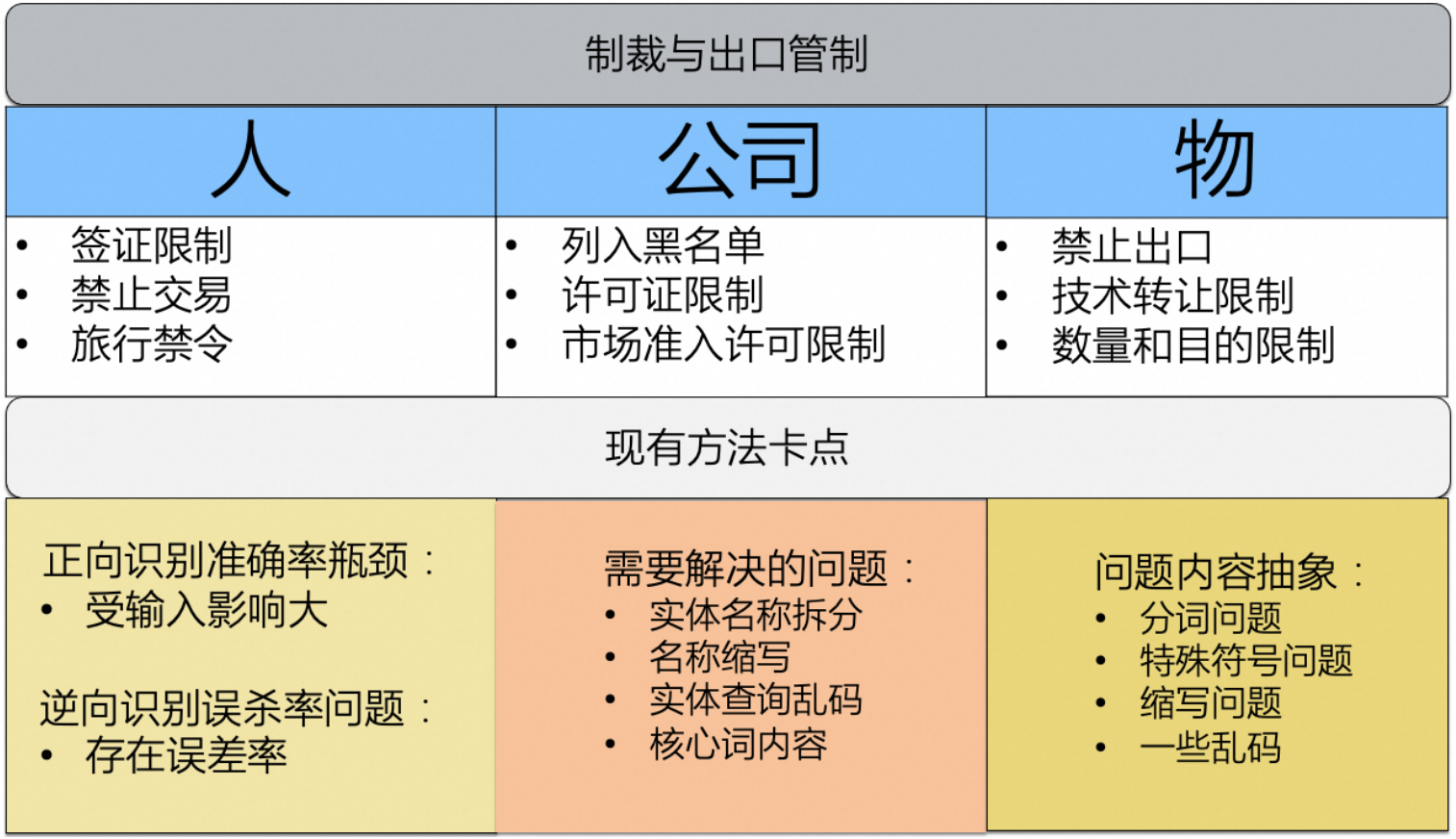
相关性与重排:

- 结合业务场景构造短-长文本(QA)匹配数据集，并对业务数据和输入query内容进行改写，构造难负样本，平衡样本分布。
- 法务专属trick，长文档文本分割与整合

大模型交互:

- 通过加入大语言模型（LLM），人机交互能力和上下文理解能力。
- 监督式微调（SFT）来提升Qwen模型的性能，确保了输出质量
- 设计微调指令，通过GPT-3.5/4和人工评分，确保输出的高质量。

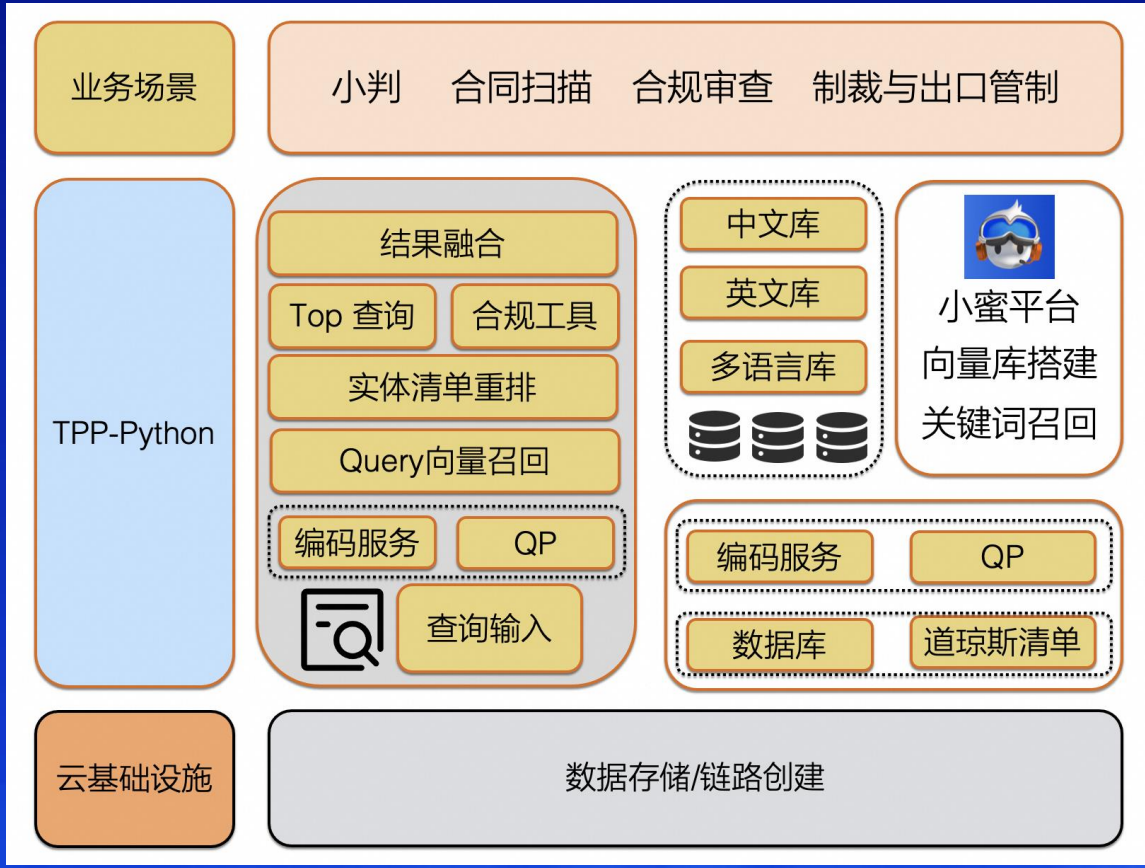
# 3.2.2 合规制裁审核




1. 多语言实体处理，确保对包含多个别名和语言版本的实体名单进行精确匹配。
2. 用户查询适配，提高对简称、误输或非标准查询的处理能力，保证结果的准确性。
3. 名单库的动态管理和维护，应对规模庞大且持续更新的数据集，实现实时、准确的实体识别和链接。

# 3.2.2 合规检索工具与制裁出口管制






**数据处理流程优化:**

- 设立多源数据整合管道，从道琼斯、阿里云等渠道更新制裁名单。
- 将名单按照语言类别分区，专门处理中英文和其他语言的多路查询，提高查询匹配效率。

**查询处理（Query Processing, QP）优化:**

- 实现高效的查询预处理，包括特殊字符处理、错别字纠正、乱码清理。
- 引入文本清洗和标准化流程，减少输入错误对检索结果的影响。

**向量化召回机制:**

- 中英文文本编码模型进行向量化，构建高效的向量索引。使用翻译接口，多语言召回结果。
- 内部算法平台上实现召回服务，结合开源faiss索引为备份策略，保证服务的稳定性。

**相关性排序与模型微调:**

- 微调模型以改善同语言匹配的相关性评分。对不同语言，分别收集相关语言开源数据和实体名单内容，形成三类数据(中、英、其他)。

# 3.2.3 小考考试主观题内容阅卷



## 主观题智能批改是什么?

2023-07-28 16:38

本章节将讲解"主观题智能批改"这个功能。

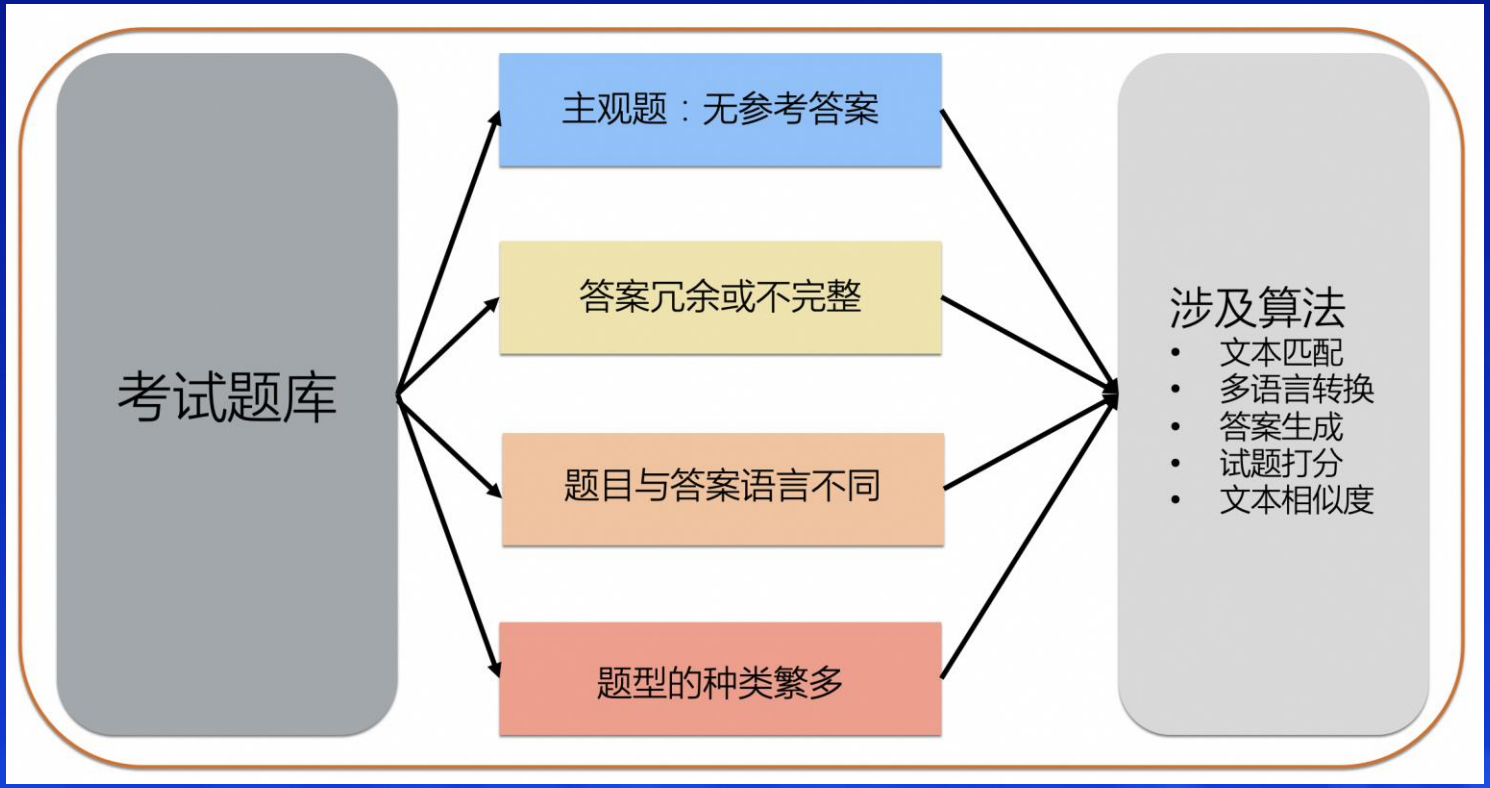
什么是主观题智能批改?

### 什么是主观题智能批改?

当所有题目都设置好试题分析（类似参考答案或者解题思路或者得分点）后，将可以打开主观题智能批改功能。当打开本功能后，在一份试卷被提交上来后，除了完成客观题的全部自动批改，系统还会尝试根据试题分析去对主观题进行批改，并将期望得分预填写至评分区域，考官可以根据期望得分去进行最后的批改。

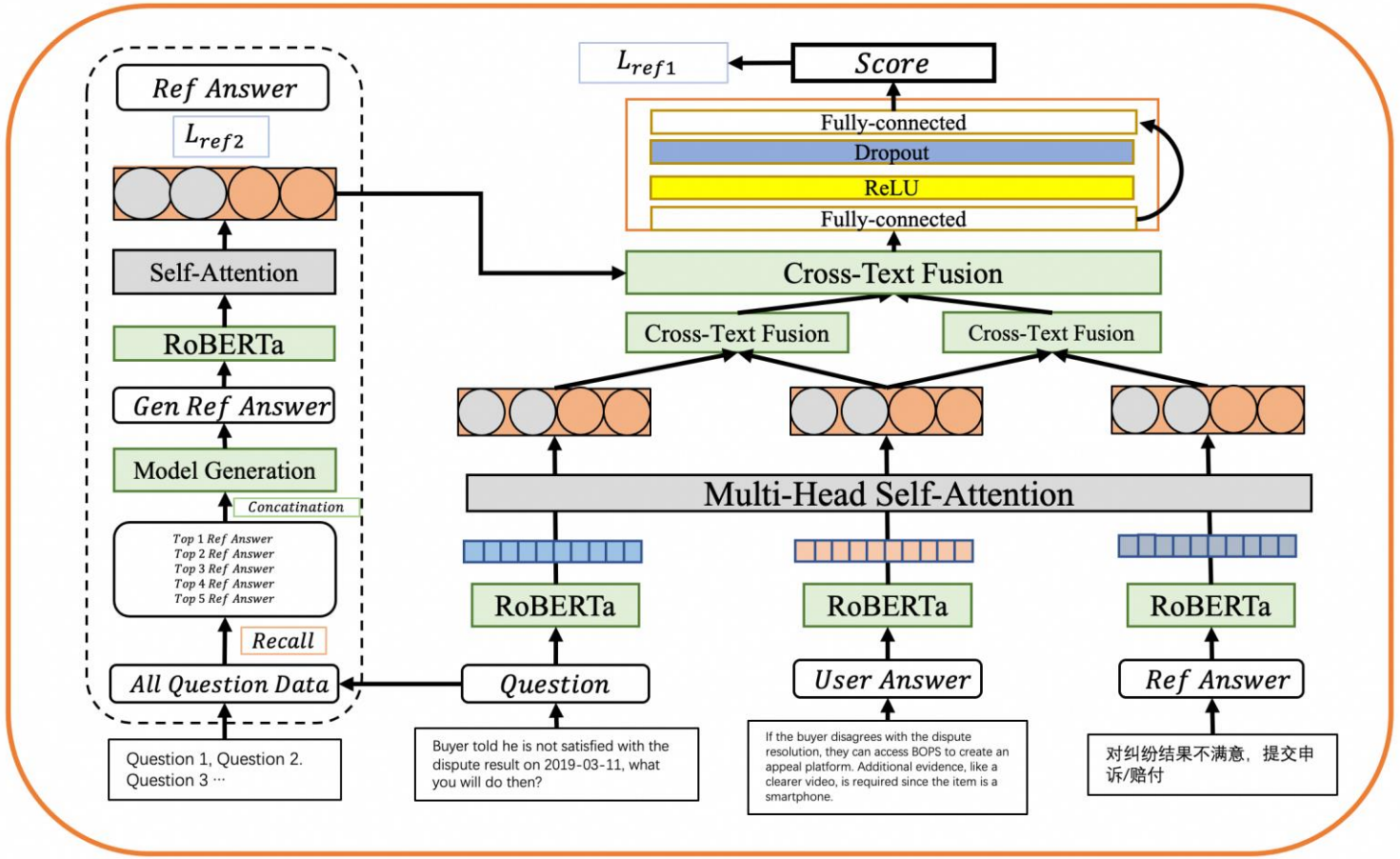
# 3.2.3 小考考试主观题内容阅卷






1. **无参考答案**：主观题作答需结合考生自我认知，无固定的参考答案，这类题目答案的修改主要依赖于考官个人判断，打分具有主观性
2. **答案缺失**：参考答案可能有缩写、冗余内容，这些内容对考官无压力，而算法模型无法识别与判断
3. **语言问题**：问题和参考答案来自不同语言，阅卷需要考虑不同语言回答时的语境问题
4. **题型繁多**：如合规、业务、个人成长相关，并且部分题目有明显的垂类特征，对考官评分也有一定要求
5. **题目重复**：考官使用相同题目制作试题，题目会有相似的同类型问题，并且带有考生的答案和对应的分数

# 3.2.3 小考考试主观题内容阅卷




1. 多任务主观题评分架构，首选低参数量的文本生成模型进行答案生成，并将现有题库用作辅助知识库，即使在缺少标准答案的情况下也能评分。
2. 答案召回与生成：当主观题缺少标准答案时，我们利用题库中相似题目的答案进行召回，得到top-5个用户答案/参考答案，选取人工评分较高的参考答案，并结合生成模型制定新参考答案。
3. 多维度评分系统：采用一种综合性评分方法，不仅比对参考答案和考生答案，还综合考虑题目内容、相关性以及答案相似度，从多个角度优化评分准确性和用户体验。
4. 经典算法辅助：为增强评分系统的健壮性，我们整合了N-Gram等传统技术作为补充，确保评分不仅基于语义匹配而忽略实际内容差异。

# PART 3-3
# 信息抽取与大模型

# 3.3 合规数据分析与内容整理



中国审判流程信息公开网
China Judicial Process Information Online

首页　公众服务

您所在位置：首页 > 公众服务 > **诉讼指南**

## 民事起诉状（样本）

发布单位：江油市人民法院　所属省份：四川　发布时间：2021-11-10　浏览量：1844

### 民事起诉状

原告×××，男（女），××××年××月××日出生，汉族，住××市×××路××号，身份证号码××××××××××××××××，电话：××××××××。

或原告×××有限公司，住所地×××路××号××室，电话：××××××××。

法定代表人：×××。

被告×××，男（女），××××年××月××日出生，汉族，住江油市×××路××号，身份证号码××××××××××××××××，电话：××××××××。

或被告×××有限公司，住所地×××路××号××室，电话：××××××××。

法定代表人：×××。

诉讼请求：

一、（写明具体的诉讼请求、标的。）

二、诉讼费由被告负担。

事实与理由：

（按具体情况陈述事实与理由。）

原告提交的证据：原告身份证明(个人为原告身份证复印件、法人为营业执照复印件，法定代表人身份证明书)；被告身份证明工商资料(个人为原告身份证复印件、法人为营业执照复印件，法定代表人身份证明书)；……（其他证据依次列明）。

此呈

江油市人民法院

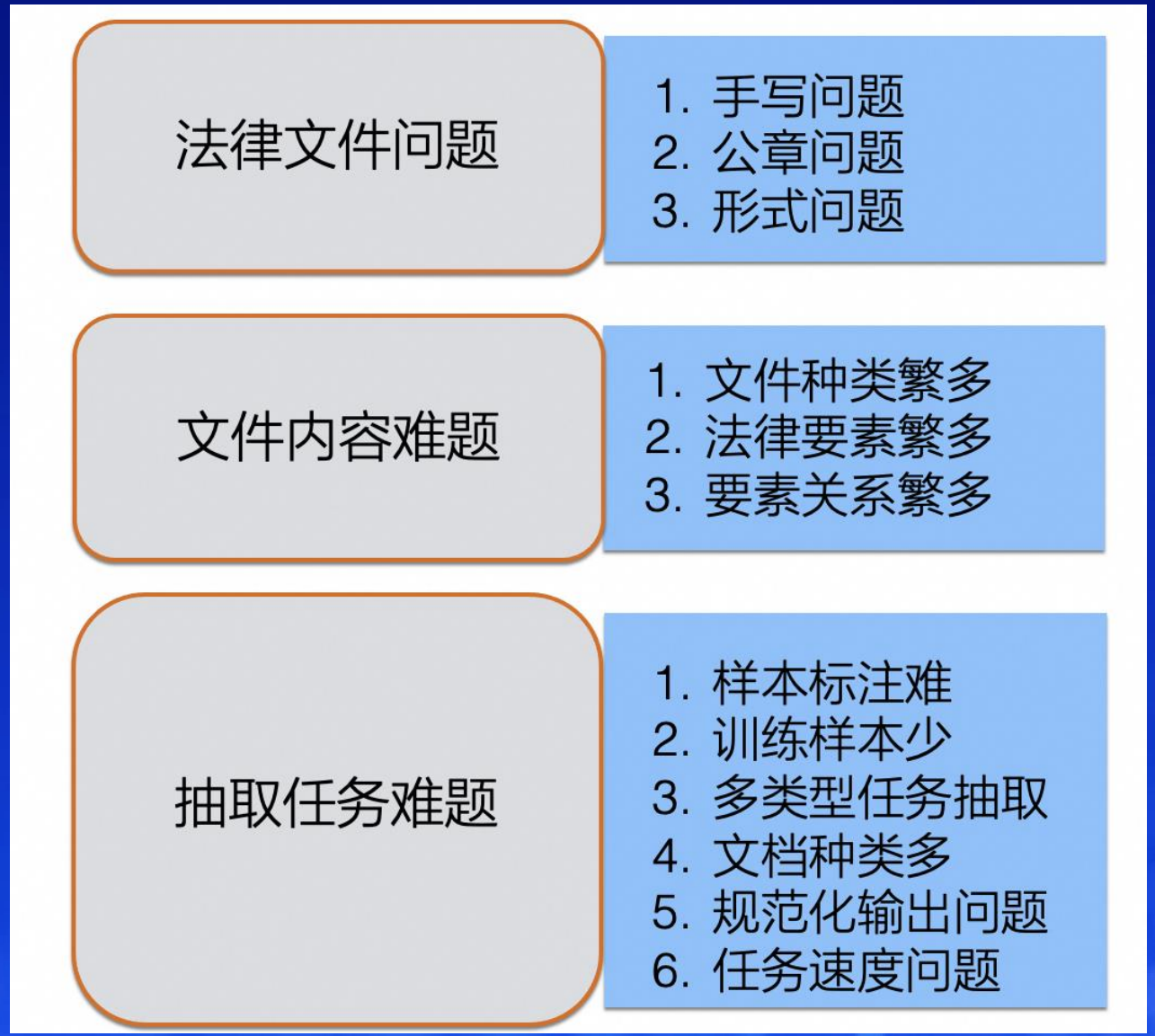
起诉人：（签名）

年　月　日

**起诉状范本**

# 3.3 合规数据分析与内容整理






1. 法律文件质量问题：法律文档通常通过拍照或扫描成PDF格式，多样化的文书格式和扫描质量问题，如手写内容和模糊字迹等，可能影响文字分析。
2. 文档内容泛化困难：当前系统处理的文书类型有限，泛化能力不强，若要支持更多文书种类，则需对服务架构进行大幅调整以包含更多字段。
3. 实体抽取难度：法律文书的处理要求参与者具备深厚的法律知识背景，导致专业标注资源稀缺，进而影响训练数据的质量和模型的性能。产品需求对文档类型的识别、标准化输出及处理时间均有明确期望。

# 3.3 合规数据分析与内容整理



## 民事起诉状

（民间借贷纠纷）

**说明：**

为了方便您参加诉讼，保护您的合法权利，请填写本表。

1. 起诉时需向人民法院提交证明您身份的材料，如身份证复印件、营业执照复印件等。
2. 本表所列内容是您提起诉讼以及人民法院查明案件事实所需，请务必如实填写。
3. 本表所涉内容系针对民间借贷纠纷案件，有些内容可能与您的案件无关，您认为与案件无关的项目可以填“无”或不填；对于本表中勾选项可以在对应项打“√”；您认为另有重要内容需要列明的，可以在本表尾部或者另附页填写。

★特别提示★

《中华人民共和国民事诉讼法》第十三条第一款规定：“民事诉讼应当遵循诚信原则。”

如果诉讼参加人违反上述规定，进行虚假诉讼、恶意诉讼，人民法院将视违法情形依法追究责任。

**当事人信息**

| | |
|---|---|
| 原告（自然人） | 姓名：<br>性别：男□ 女□<br>出生日期： 年 月 日 民族：<br>工作单位： 职务： 联系电话：<br>住所地（户籍所在地）：<br>经常居住地：<br>证件类型：<br>证件号码： |
| 原告（法人、非法人组织） | 名称：<br>住所地（主要办事机构所在地）：<br>注册地/登记地：<br>法定代表人/主要负责人： 职务： 联系电话：<br>统一社会信用代码：<br>类型：有限责任公司□ 股份有限公司□ 上市公司□ 其他企业法人□<br>事业单位□ 社会团体□ 基金会□ 社会服务机构□<br>机关法人□ 农村集体经济组织法人□ 城镇农村的合作经济组织法人□ 基层群众性自治组织法人□<br>个人独资企业□ 合伙企业□ 不具有法人资格的专业服务机构□<br>国有□ （控股□参股□）民营□ |
| 委托诉讼代理人 | 有□<br>姓名：<br>单位： 职务： 联系电话：<br>代理权限：一般授权□ 特别授权 □<br>无□ |
| 送达地址（所填信息除书面特别声明更改外，适用于案件一审、二审、再审所有后续程序）及收件人、电话 | 地址：<br>收件人：<br>电话： |

## 民事答辩状

（民间借贷纠纷）

**说明：**

为了方便您参加诉讼，保护您的合法权利，请填写本表。

1. 应诉时需向人民法院提交证明您身份的材料，如身份证复印件、营业执照复印件等。
2. 本表所列内容是您参加诉讼以及人民法院查明案件事实所需，请务必如实填写。
3. 本表所涉内容系针对一般民间借贷纠纷案件，有些内容可能与您的案件无关，您认为与案件无关的项目可以填“无”或不填；对于本表中勾选项可以在对应项打“√”；您认为另有重要内容需要列明的，可以在本表尾部或者另附页填写。

★特别提示★

《中华人民共和国民事诉讼法》第十三条第一款规定：“民事诉讼应当遵循诚信原则。”

如果诉讼参加人违反上述规定，进行虚假诉讼、恶意诉讼，人民法院将视违法情形依法追究责任。

| 案号 | | 案由 | |
|---|---|---|---|

**当事人信息**

| | |
|---|---|
| 答辩人（自然人） | 姓名：<br>性别：男□ 女□<br>出生日期： 年 月 日 民族：<br>工作单位： 职务： 联系电话：<br>住所地（户籍所在地）：<br>经常居住地： |
| 答辩人（法人、非法人组织） | 名称：<br>住所地（主要办事机构所在地）：<br>注册地/登记地：<br>法定代表人/主要负责人： 职务： 联系电话：<br>统一社会信用代码：<br>类型：有限责任公司□ 股份有限公司□ 上市公司□ 其他企业法人□<br>事业单位□ 社会团体□ 基金会□ 社会服务机构□<br>机关法人□ 农村集体经济组织法人□ 城镇农村的合作经济组织法人□ 基层群众性自治组织法人□<br>个人独资企业□ 合伙企业□ 不具有法人资格的专业服务机构□<br>国有□（控股□参股□） 民营□ |
| 委托诉讼代理人 | 有□<br>姓名：<br>单位： 职务： 联系电话：<br>代理权限：一般授权□ 特别授权 □<br>无□ |
| 送达地址（所填信息除书面特别声明更改外，适用于案件一审、二审、再审所有后续程序）及收件人、联系电话 | 地址：<br>收件人：<br>联系电话： |

甘肃法院网起诉状范本

# 3.3 合规数据分析与内容整理



## 最丰富的企业智能场景应用案例

针对企业服务业务场景化的最佳智能算法解决方案

智慧法务 | 智慧行政 | 智慧HR | 阿里内外

**智慧法务白皮书**
集团法务中台运营团队联合企业智能算法和产品开发团队在法律科技领域进行一些探索

**智能起草（问卷式）**
通过简单交互实现合同的自动起草，极大提升合同起草效率，减少潜在风险

**智能起草（协同式）**
利用NLP技术提供标准条款推荐，必备条款补全推荐等功能辅助法务实现智能起草

**文档树**
通过OneDoc解构保留文档重要结构信息，...按照树的形式组织起来，清晰的理清文档脉络

**合同抽取**
提供合同结构化分析模块，能够自动构建和保留合同的基本meta信息，以及23+个关键字...

**智能审查**
利用NLP技术实现合同的形式、要素、条款的智能审查，覆盖多个风险点和业务场景

**智能文档解析**
通过OneDoc解析，清晰展现文档结构，方便下游使用

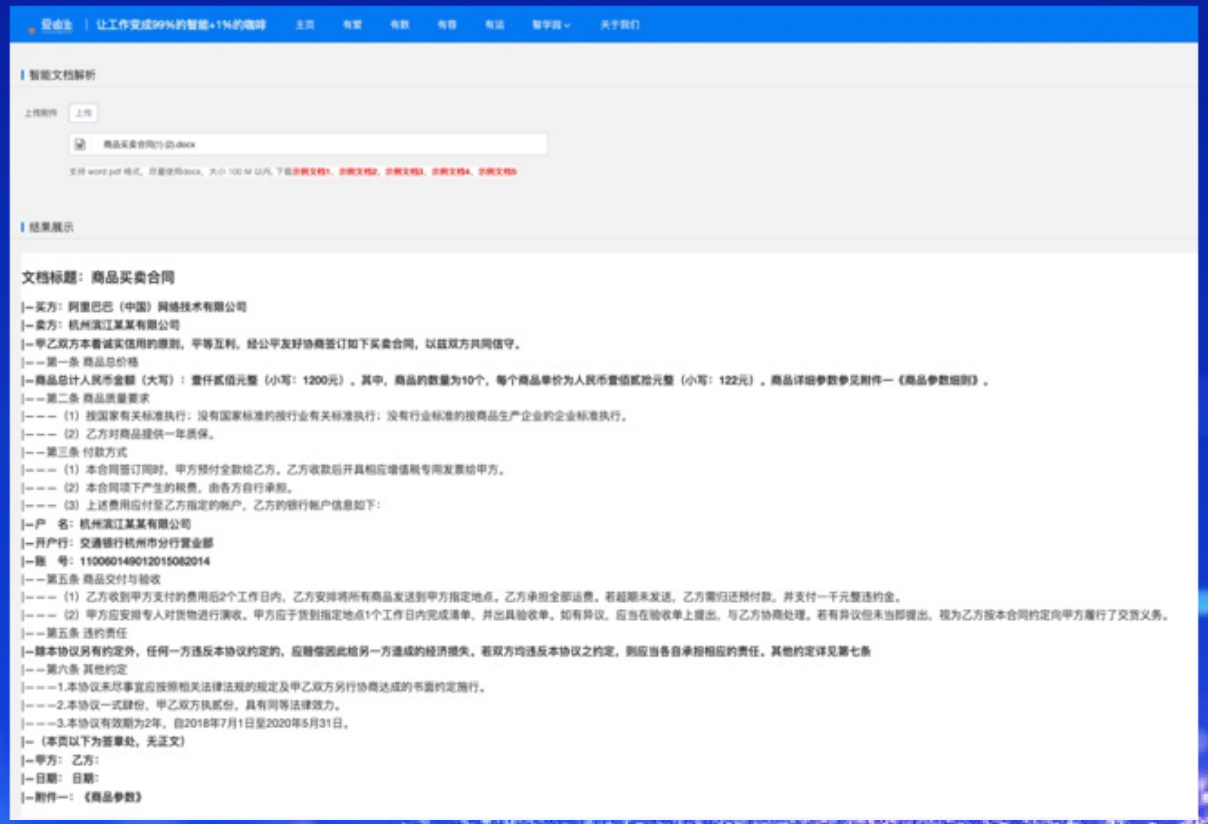
**智能履约**
针对不同履约场景生成对应的履约计划，协助法务人员顺利完成合同履约

**协议智能诊断**
利用人工智能辅助法务在起草、审批时对在线协议的内容进行负面关键词的审核把关

**诉讼智能收案**
将纸质传票转化为结构化的电子文档，准确度超过90%，召回度超过85%，效率提升300%

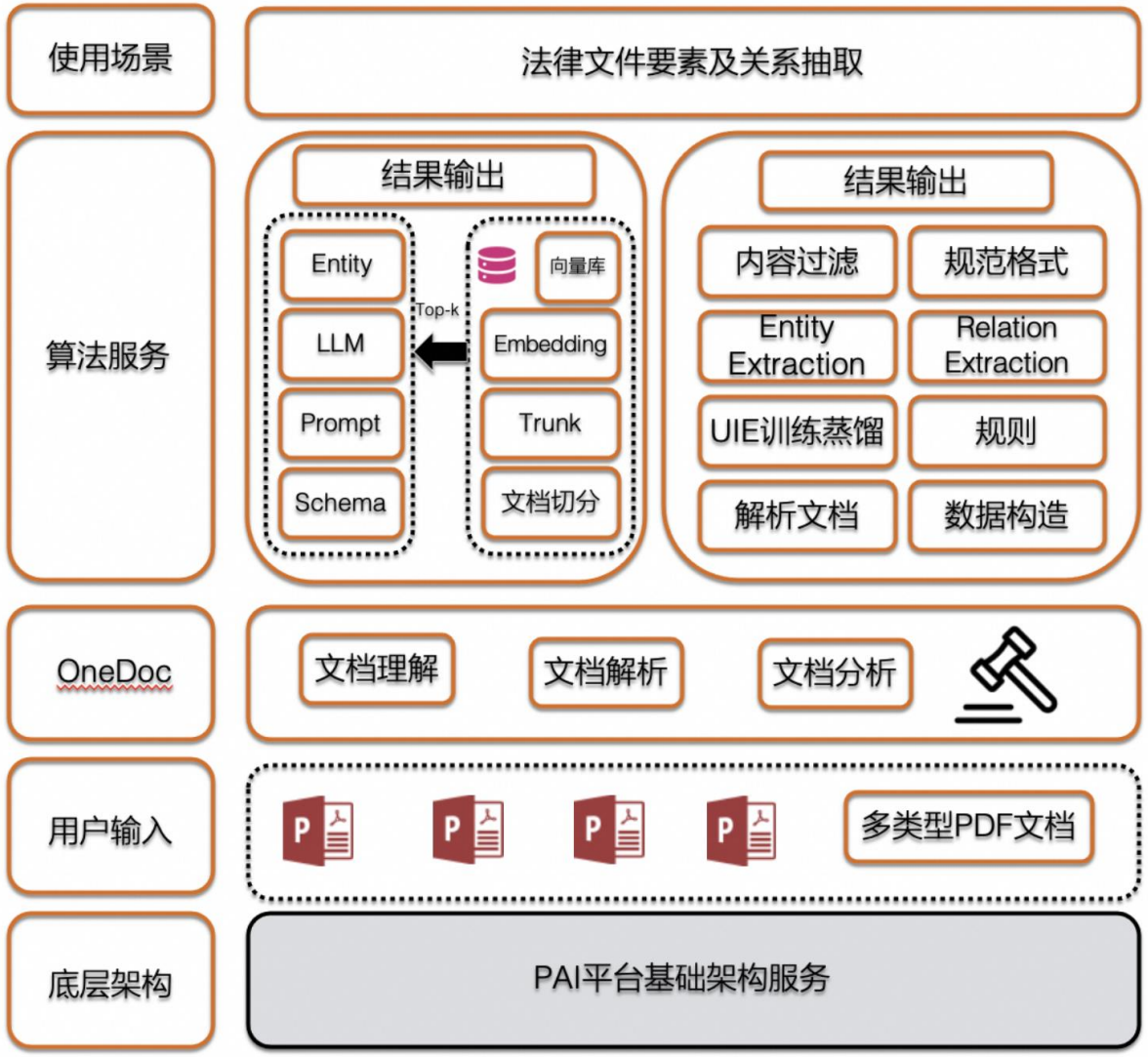
**条款库**
合同条款库应用，包括标准条款审查、条款...

**语雀知识管理**
支持将语雀文档进行知识生产、知识加工、...



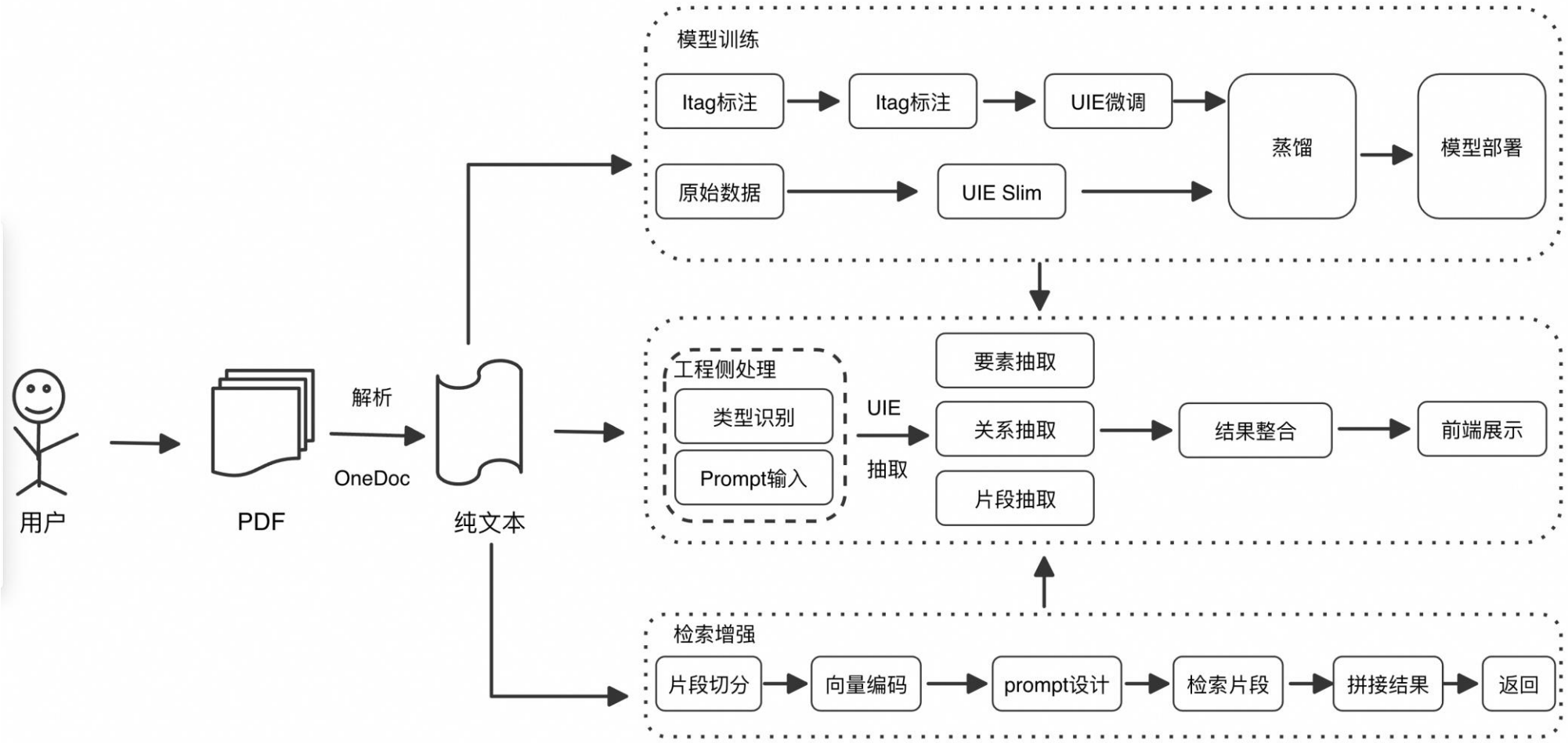

甘肃法院网起诉状范本

# 3.3.1 诉讼内容自动收案



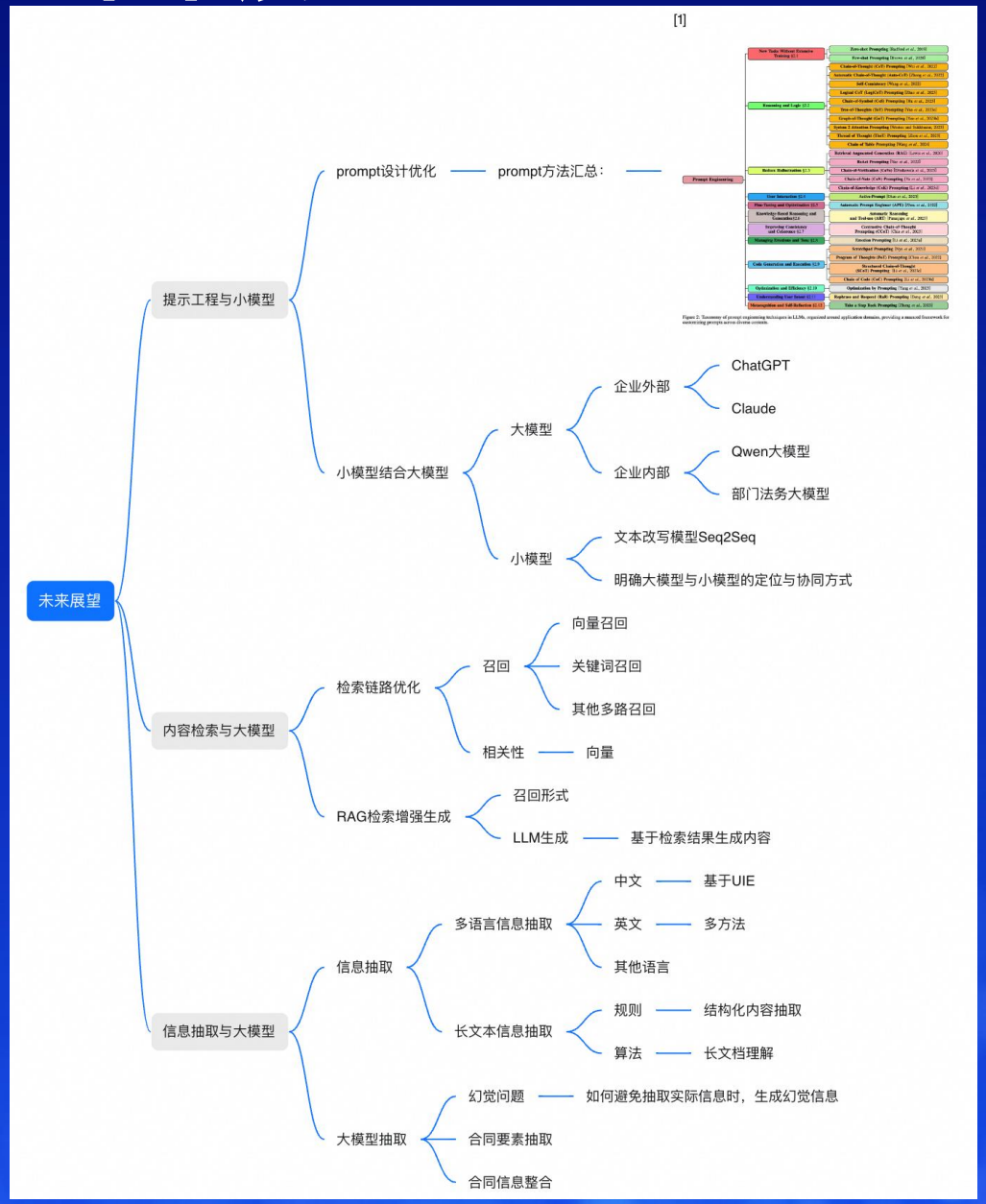

1. 数据标注: 十余种文档类型的业务数据进行了详细标注，由法务知识背景的专业人员完成。
2. 模型选择: UIE（Unified Information Extraction）模型进行要素及其关系的抽取。
3. 模型微调: 通过对UIE模型的少量样本微调。
4. 知识蒸馏: 解决长文本分析时间和资源消耗问题，利用更小的模型（UIE-small）进行了知识蒸馏，有助于减少模型推理时间并保持性能。
5. 模型格式转换: 将模型文件转换为ONNX格式以进一步减少处理时间。
6. 规则设计: 结合业务数据设计大量规则，以提高模型检测的准确率。
7. 长文本处理: 您采用了基于检索增强的大型语言模型（LLM）策略。

# 3.3.1 诉讼内容自动收案



# PART 04
# 未来展望

# 4 未来展望




1. 提示工程与小模型：提示工程在主要在输入层面影响大模型，在一些场景中如何结合小模型提升LLM的工作效率。
2. 内容检索与大模型：检索增强生成依赖于前链路检索能力与大模型检索能力。
3. 信息抽取与大模型：实体信息抽取必须依赖于原始内容，不能产生幻觉，并且要与原文内容相同。

[1] Sahoo et al. (2024) A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications.

# 科技生态圈峰会 + 深度研习

## ——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

# THANKS