从需求到投标：
数据驱动的智能技术方案生成 Agent 实战

演讲人：王志宏

商汤科技 / 大装置事业群 研发总监

AiCon
全球人工智能开发与应用大会

需求爆炸增长，产研陷入“评估挤压”

在**项目激增**与**场景分化**的双重压力下，**需求评估**已成为产研体系的**首要约束**。

需求激增

市场 AI 落地需求持续上升，售前团队并行推进大量项目机会，且每个项目都要求快速响应、快速出方案

高度定制

应用场景碎片化，不同行业、客户、部门给出完全不同诉求，可复现、可归纳的需求比例极低

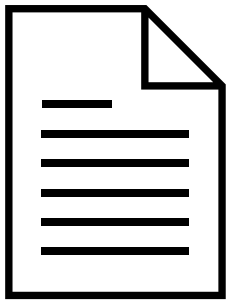
产研压力

产研一边保持标准产品持续迭代，同时必须协助售前评估复杂场景，反复理解和验证场景需求

核心瓶颈

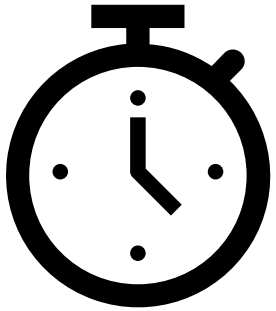
售前推进速度与产研评估速度严重不匹配，评估成本居高不下，大量需求无法及时响应

我们的困境



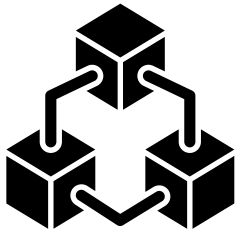
海量数据

产品说明书、需求文档、法律模板、过往案例…… 信息淹没决策。



时间紧张

响应周期以天计算，而非周。每一次评估都是一场与时间的赛跑。



知识孤岛

关键技术知识、过往案例、客户需求分散在不同团队和文档中，难以整合。

魔鬼细节

从一个参数错误到一项合规遗漏，任何微小失误都可能导致评估失败。

我们的答案

我们的答案：解决方案Agent

构造一个由多个专家Agent组成的协同系统，分析杂乱无章的数据并得到我们需要的结果。它不仅仅是自动化工具，更是一 个能够理解、推理、生成和验证的专家级合作伙伴。

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

数据简介

产品清单

能提供的标准产品

- 产品OnePage
- 产品功能清单
- 产品白皮书
- 产品使用手册
- 产品标准API
- 产品价格说明
- ……

技术能力

能提供的核心技术

- 工程开发能力
- 算法优化能力
- 接口集成能力
- 安全治理能力
- 监控日志能力
- 效率工具能力
- ……

案例成果

我们曾经做过的案例

- 过往项目案例
- 开源项目介绍
- 发表论文成果
- 公开课程合集
- 技术分享演讲
- 行业洞察报告
- ……

需求拆解

客户场景的真实需求

- 业务场景描述
- 功能诉求要点
- 项目范围界定
- 预算资源限制
- 交付质量标准
- 风险合规要求
- ……

产品经理视角下的需求分析

01

目标与定位

产品定位：自动化完成项目评估、方案生成、标书输出的智能解决方案助手。

核心目标：提升售前效率，提升方案质量，降低人员成本

02

用户与场景

使用者：售前顾问、项目经理、交付负责人、咨询顾问

典型场景：初次获取客户需求时，快速判断需求可行性，并输出结构化的解决方案

03

输入与输出

输入：我方的产品手册、技术能力和案例，客户的需求描述及项目上下文信息

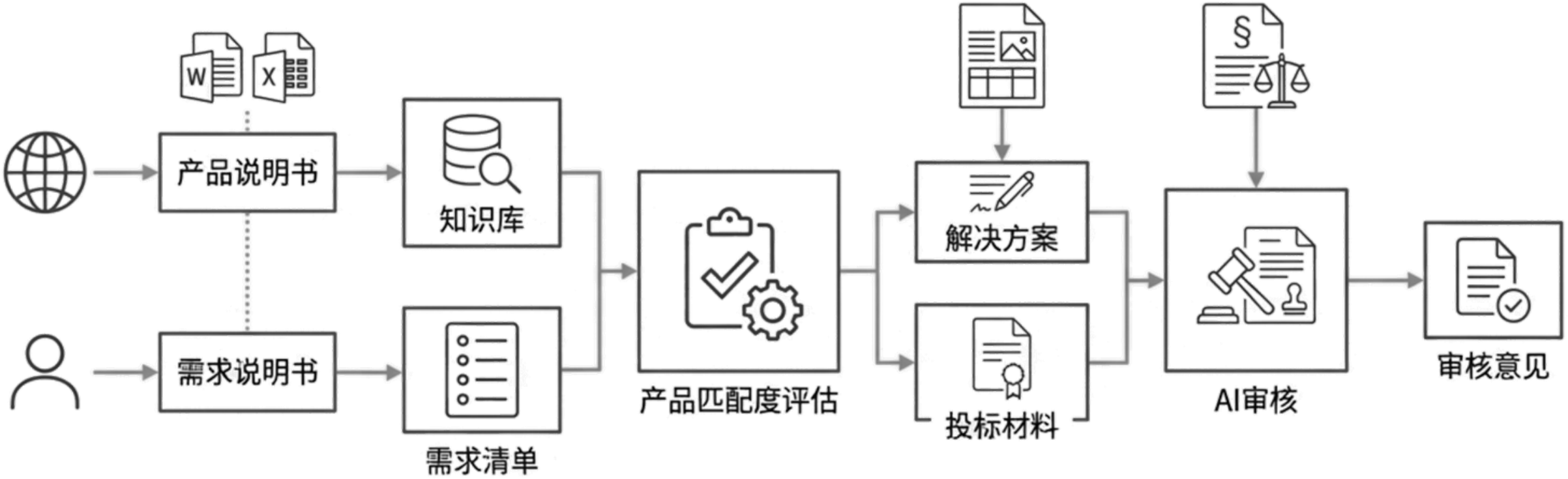
输出：项目需求解析，解决方案文档（对内）和投标材料（对外）

04

功能与模块

- 需求理解（输入解析）
- 可行性评估（决策判断）
- 方案生成（结构化方案）
- 风险提示（边界说明）
- 文档输出（可交付材料）

整体方案设计



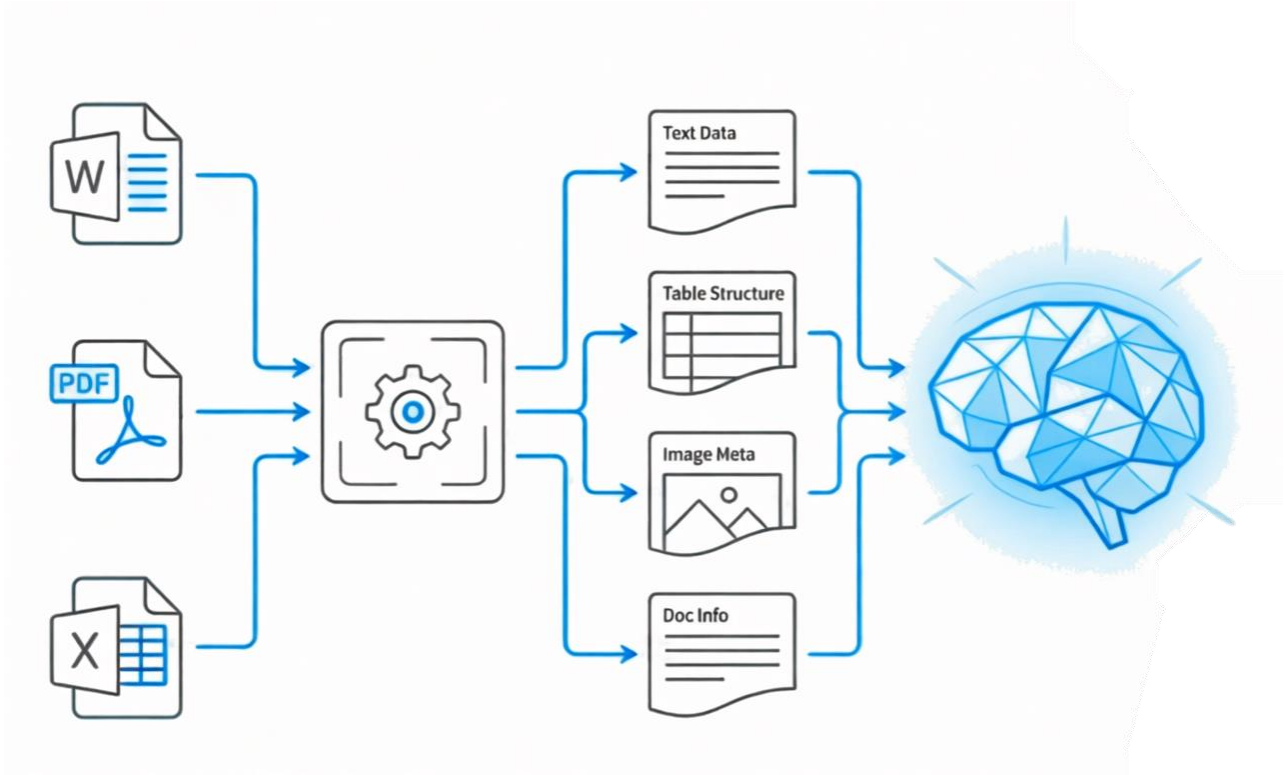
我们的期待很简单：

让 Agent 读取所有文档数据，自动分析需求，匹配产品数据，然后撰写解决方案和标书。

AiCon
全球人工智能开发与应用大会

目录

阶段1：产品和需求文档分析



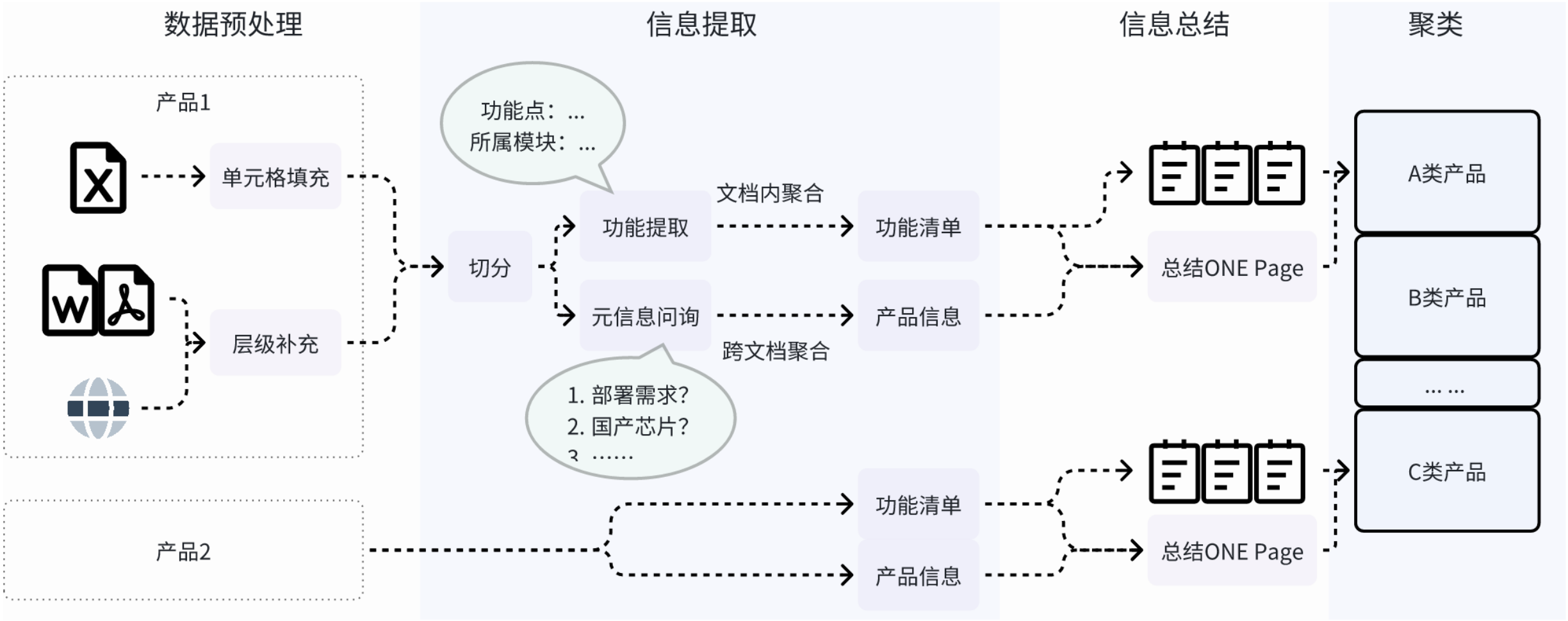
多格式解析与语义对齐：针对不同格式的文档数据，设计专门的信息抽取逻辑，并对齐到同一粒度

功能分级建模：将产品功能信息构建为“功能概述 → 子功能 → 参数指标”的结构化数据，方便对比分析

技术要素提取：深入挖掘技术选型、软硬件依赖、国产化兼容性等关键要素作为元数据信息

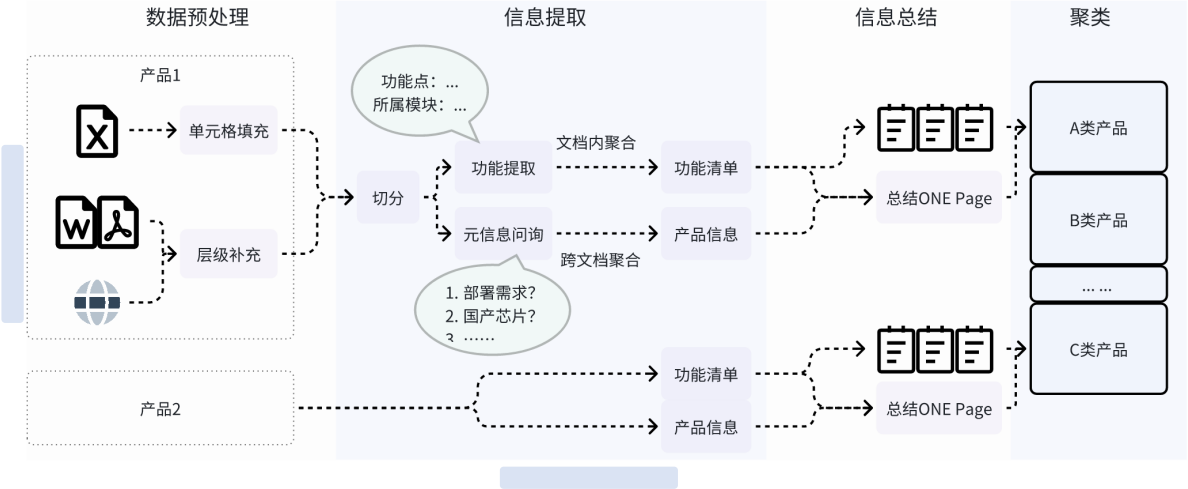
知识压缩与索引：对功能、要素、场景进行多维度信息提取并转换为向量化数据，以实现数据的精准匹配

阶段1：产品文档分析和结构化数据提取



阶段1：关键点

产品手册需要解析目录



对产品进行分层和聚类

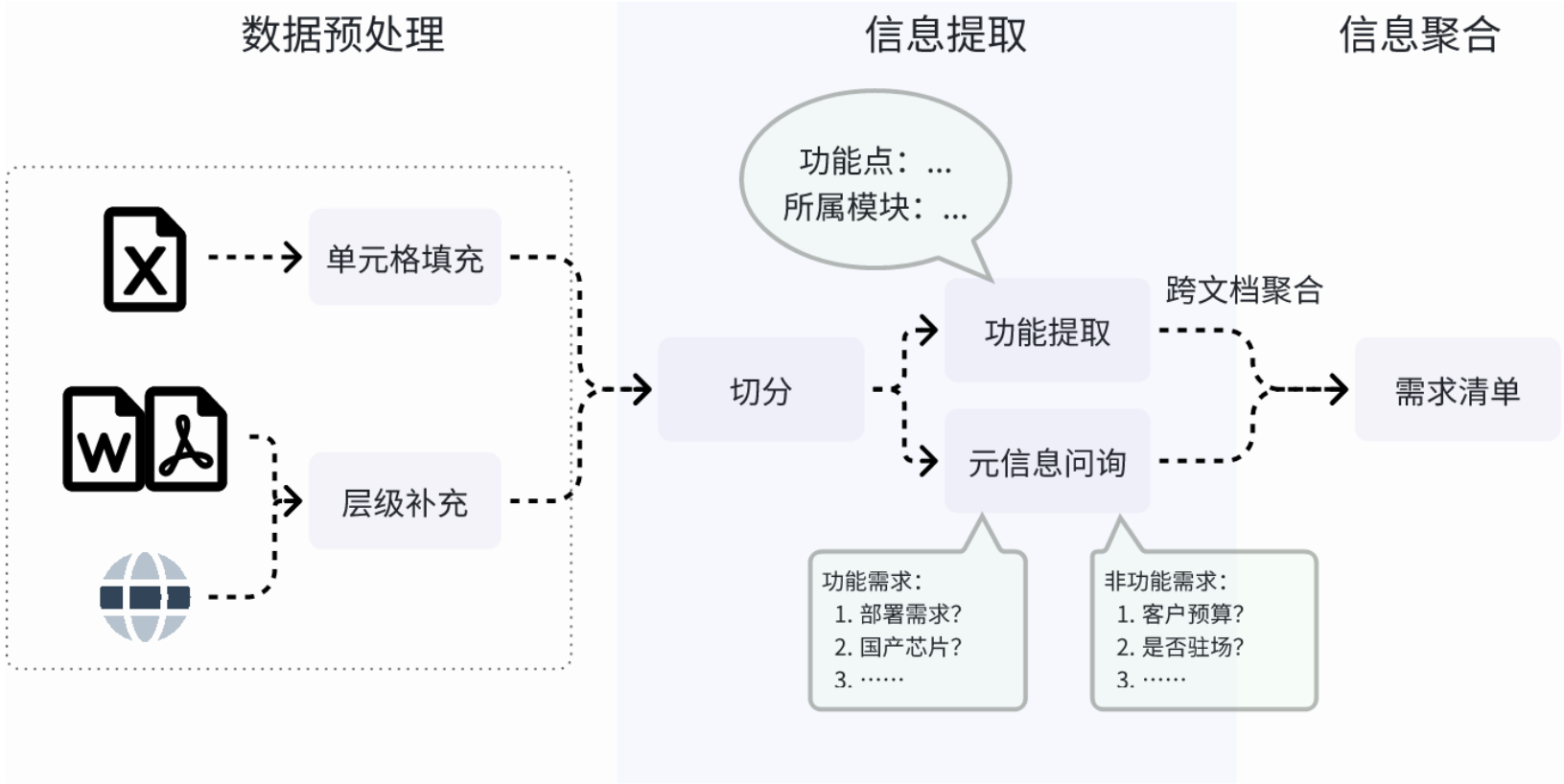
依赖和约束作为元信息

时间和费用

对于一篇50页左右的产品文档或需求说明：
约需要花费 30w Token的费用，耗时约15分钟

复用知识库持久化管理

阶段1补充：需求文档的结构化数据提取



灵魂N问：

1. 项目需求大概是什么？
2. 项目时间周期是多久？
3. 项目预算是多少？
4. 对硬件的有什么要求？
5. 对信创有什么要求？
6. 对模型有什么要求？
7. 是否需要支持小语种？
8. 对稳定性有什么要求？
9. 对权限有什么期待？
10. 是否接入旧系统？
11. 是否需要开放源码？
12. 外网能否访问环境？
13. 是否需要驻场开发？
14. … …

阶段1：输出展示

商汤 sensetime | 大装置 sensecore

需求解析


```
"metadatas": {
    演练经验，且至少1人持有相关认证；根据现场服务响应等级要求提供上门服务；要求提供现场巡检服务，每年4次，每次巡检需现场执行并提交书面报告；对于特殊日期和无法采用非现场方式解决的问题，中标人需委派技术人员抵达现场进行技术支持或维护；中标人需为招标人提供特殊防护期、重要节假日等现场值守服务，具体值守时间由招标人按实际需求提出；要求中标人安排技术人员在接到通知后规定时间内到场处理，具体时长需在合同中明确，驻场人员需具备相关技术资质和经验；现场技术支持人员应是参与过本项目技术服务过程的人员或资深系统设计师软件开发工程师及运维工程师；驻场实施总人数要求，具体人数未明确，但需在投标文件中响应并提供证明材料；驻场实施总人数 > 9人:2分；7人 < 驻场实施总人数 ≤ 9人:1分；驻场实施总人数 ≤ 7人:0分；驻场实施总人月数 > 20人月:2分；18人月<驻场实施总人月数 ≤ 20人月:1分；驻场实施总人月数 ≤ 18人月:0分；是，要求项目主要人员（如项目经理、技术经理）全职驻场服务，驻场周期为项目实施全过程；驻场实施人员不少于指定人数，总驻场人月数不少于18人月。",
    "售后要求": "售后要求：要求提供不少于1年的免费售后服务，包括系统使用咨询、问题解答、远程技术支持及定期回访。服务期内若系统更新或功能调整，需提供相应培训支持。本项目应提供3年的维保期。维保期结束后，中标人有义务在本项目的维保、运行管理和开发方面继续给予技术协作和咨询。需提供7×24小时电话、传真、电子邮件等方式的技术支持和咨询服务，需提供1名广州本地合格维护人员的直接联系方式。需提供持续的技术支持与维护服务，服务期限至少为合同签订后三年，包含系统升级、问题修复、定期巡检等。",
    "工期要求": "项目实施周期最长不超过12个月，自项目启动至招标人签署《系统开发项目上线验收报告》。系统软件上线后需稳定运行3个月且无重大生产问题方可终验，招标人可在3个月内根据运行情况提前终验。中标人需在投标时提交项目总体计划并分析其可行性。",
    "兼容性要求": "需与现有ERP系统、OA系统、财务系统等实现数据互通与业务协同，支持标准接口协议（如RESTful API、WebService）；需与招标人已有的云平台服务（TDSQL、CRedis、CLB）兼容，支持与现有系统通过域名方式调用，系统架构需与现有系统调用关系拓扑兼容，支持IPv4/IPv6双协议栈；需兼容现有信创云平台CVM环境、传统物理服务器环境、麒麟v10操作系统、国产数据库及中间件，支持与现有业务系统进行数据交互与接口对接。",
    "需求名称": "大模型协同计算平台项目招标文件.docx"
},
"function_list": [
    {
        "功能名称": "系统性能调优",
        "所属模块": "运维管理 -> 系统性能优化 -> 系统性能调优",
        "所属层级": "应用层",
        "功能描述": "系统上线后，中标人需对系统运行性能进行持续监控与分析，识别性能瓶颈，优化数据库查询、接口响应、资源调度等关键环节，确保系统在高负载场景下仍能保持稳定、快速响应。调优工作需覆盖系统架构、应用代码、数据库配置、缓存策略等多个方面，直至系统性能指标（如响应时间、吞吐量、并发能力等）达到合同约定要求。",
        "功能优先级": "",
        "功能验收要求": "系统性能调优需通过压力测试和真实业务场景验证，响应时间不超过1秒，系统吞吐量达到设计指标的120%，CPU和内存使用率在正常范围内，无崩溃或严重卡顿现象。验收标准包括性能测试报告、调优日志、优化前后对比数据，由需求方组织专家评审并签字确认。"
    },
    {
        "功能名称": "本地化部署",
        "所属模块": "模型运行 -> 本地化部署 -> 本地化部署",
        "所属层级": "IaaS层",
        "功能描述": "支持在本地环境中部署chatglm系列、qwen系列（包括qwen、qwen1.5及qwen2）、baichuan系列等大模型，确保模型可在本地服务器上独立运行，不依赖外部云服务，保障数据安全与系统可控性。部署过程应支持自动化或半自动化流程，具备良好的兼容性与稳定性。",
        "功能优先级": "",
        "功能验收要求": "通过实际部署验证各系列大模型可在目标环境中正常启动与运行；测试模型推理响应时间、内存占用等性能指标；提供部署日志与错误处理机制，确保部署过程可追溯、可复现。"
    },
```


产品聚类


```
{
  "cluster_result": {
    "应用层_产品类别_1": {
      "类别简介": {
        "品类介绍": "本类别产品是一套基于AI视觉与边缘计算技术的智能感知与管理解决方案，融合AI、IoT、5G、数字孪生等前沿技术，构建端边云协同的智能生态体系。产品可实现对人脸、人体、车辆等目标的高精度检测与识别，支持戴口罩人脸识别、本地事件判定与结构化数据实时分析，具备高抓拍率、高准确率与低误报率，显著降低对云端算力的依赖，保障数据隐私与安全。通过RTSP协议灵活接入视频流，实现从边缘侧到云端的端到端智能分析，广泛应用于智慧园区、智慧景区、智慧校园、智慧社区、智慧商业等场景。在智慧园区中，提供智慧通行、综合安防、设施管理、能效优化、智慧乘梯与车船管理等一体化服务，提升运营效率与管理智能化水平；在智慧景区中，实现刷脸入园、智能测温、客流统计与安防预警，助力景区数字化升级与游客体验优化；在智慧校园中，构建覆盖全场景的综合安防体系，支持智慧签到、异常行为分析与校园设施物联化管理，保障师生安全与教学秩序。整体方案具备高度可扩展性与场景适配能力，已在全国百余城市落地超300个场景，全面赋能城市关键空间的智能化转型，实现安全可控、高效便捷、可持续发展的智慧运营新范式。"
      },
      "产品列表": [
        "edge_cube",
        "商汤星云智慧园区",
        "商汤星云智慧景区",
        "商汤星云智慧校园"
      ]
    },
    "应用层_产品类别_2": {
      "类别简介": {
        "品类介绍": "该类别产品是一类面向大模型应用开发的综合性低代码平台，致力于帮助企业与开发者高效构建、部署和管理智能化应用。平台通过可视化拖拽方式，支持用户快速搭建复杂的工作流，集成多种基础模型、功能模块与控制流组件，实现从应用创建、数据管理、知识库构建到模型微调与推理部署的全流程闭环管理。其核心功能涵盖智能问答、多轮对话、文档解析、内容生成、文书审核、模板生成、批量处理、权限控制等，全面覆盖智能客服、企业知识管理、法律文书处理、金融报告生成、合同分析、市场洞察、智能办公等多样化业务场景。平台具备强大的多模态文件处理能力，支持PDF、DOC、PPTX、DOCX、Excel、图像等多种格式的上传与解析，可无缝对接本地文件、知识库、数据库及外部联网检索，实现基于多源信息的精准问答与内容生成。无需复杂Prompt工程或模型微调，即可实现零门槛调用AI能力，显著降低技术使用门槛。通过构建贴合企业实际业务场景的专业知识库，平台能够将大模型能力深度融入企业工作流，提升信息处理效率与决策智能化水平。无论是自动化数据处理、智能内容创作，还是高精度文档审核与分析，该类产品均能提供稳定、可扩展、安全可靠的AI应用支撑，助力企业实现数字化转型与智能化升级。"
      },
      "产品列表": [
        "LazyCraft",
        "商汤万象"
      ]
    },
    "应用层_产品类别_3": {
      "类别简介": {
        "品类介绍": "身份验证终端设备是一类集成了多种生物识别与证件核验技术的高安全性智能硬件，专为需要严格身份认证的场景而设计。该类产品通过融合身份证芯片读取、OCR文字识别、人脸活体检测、人脸比对及指纹识别等多重技术手段，实现对用户身份的多维度、高精度验证。在身份核验过程中，设备首先通过读取身份证芯片信息或对身份证进行OCR识别，自动提取持证人姓名、身份证号、照片等关键信息，并与证件原件进行真伪比对，有效防范伪造证件带来的安全风险。同时，结合先进的活体检测技术，设备可精准识别是否存在照片、视频或面具等欺骗行为，确保验证
```

阶段1：工程化

引入知识库以复用产品文档

通过将功能清单与 metadata 拆分为独立 node-group进行入库，实现产品手册内容的结构化复用与精细化管理。

需求做为临时文件按会话缓存

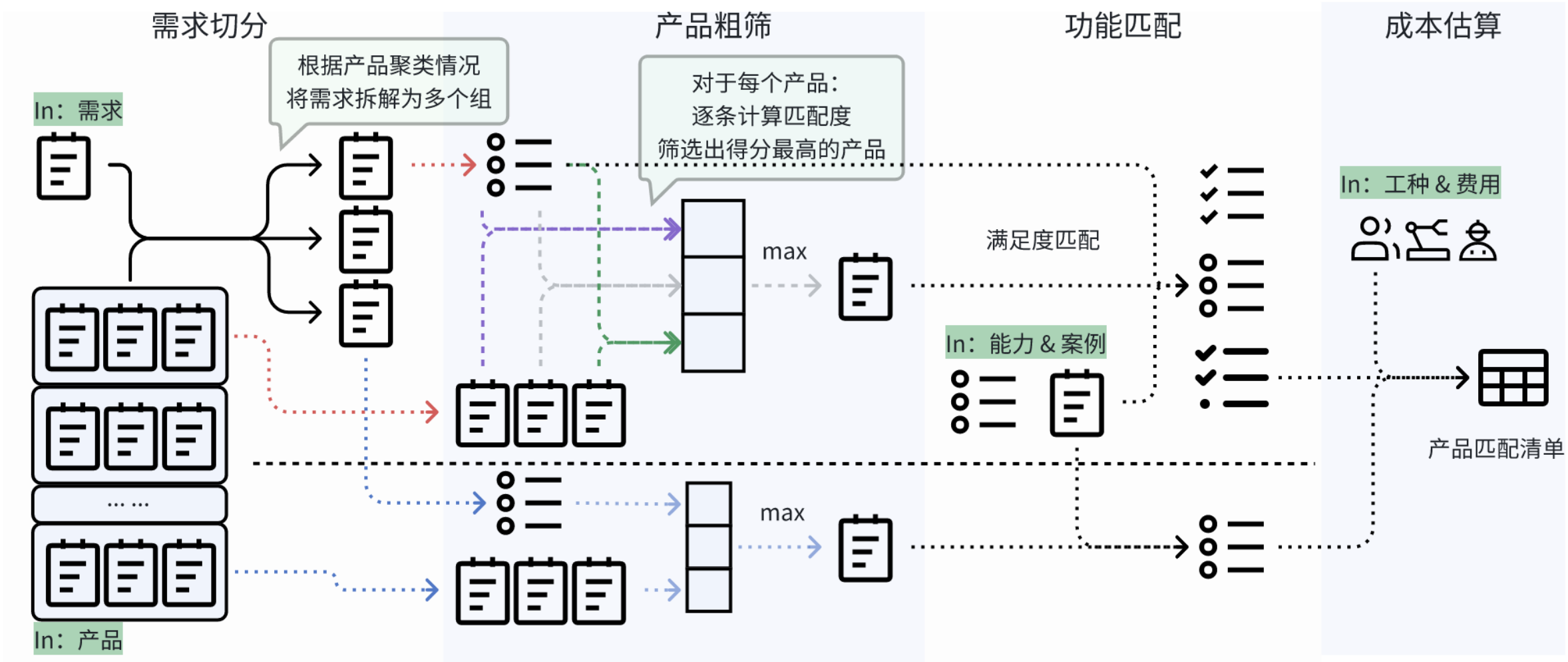
需求文档作为临时输入，不写入知识库，但通过短期缓存机制，加速相似需求的重复解析。

阶段1：代码片段

```
1.  from lazyllm.tools.rag import Document, JsonExtractor, MineruPDFReader, SentenceSplitter, ...

2.  doc = Document(dataset_path=dataset_path, embed=embed, manager=False,
                   store_conf = {'type': 'milvus','kwargs': {'uri': milvus_path}})

3.  doc.add_reader(["*.pdf", "*.docx", "*.doc"], MineruPDFReader(url="http://10.119.30.80:20234", upload_mode=True, llm=llm))
4.  doc.add_reader(["*.csv", "*.xlsx"], CustomPandasCSVReader())

5.  metadata_extractor = JsonExtractor(llm,
      schema='{"产品简介": "包括..., 二百字以内", "国产硬件支持": "可以支持的国产硬件...", ...',
      extra_requirements="所有输出内容必须严格来源于本资料片段，...")
6.  function_extractor = JsonExtractor(llm,
      schema='[{"功能名称": "", "所属模块": "功能所属的模块，格式为..."}, ...]',
      extra_requirements="所有输出内容必须严格...")

7.  doc.create_node_group(name="big_chunks", transform= SentenceSplitter, chunk_size=20480, chunk_overlap=100)
8.  doc.create_node_group(name="small_chunks", parent="big_chunks", transform= SentenceSplitter, chunk_size=1024, chunk_overlap=100)
9.  doc.create_node_group(name="metadata", transform=JsonTransform(metadata_extractor), parent="big_chunks")
10. doc.create_node_group(name="function_list", parent="small_chunks",
                          transform=JsonTransform(function_extractor, JsonLikeFormatter("[功能名称, 功能描述]")))
```

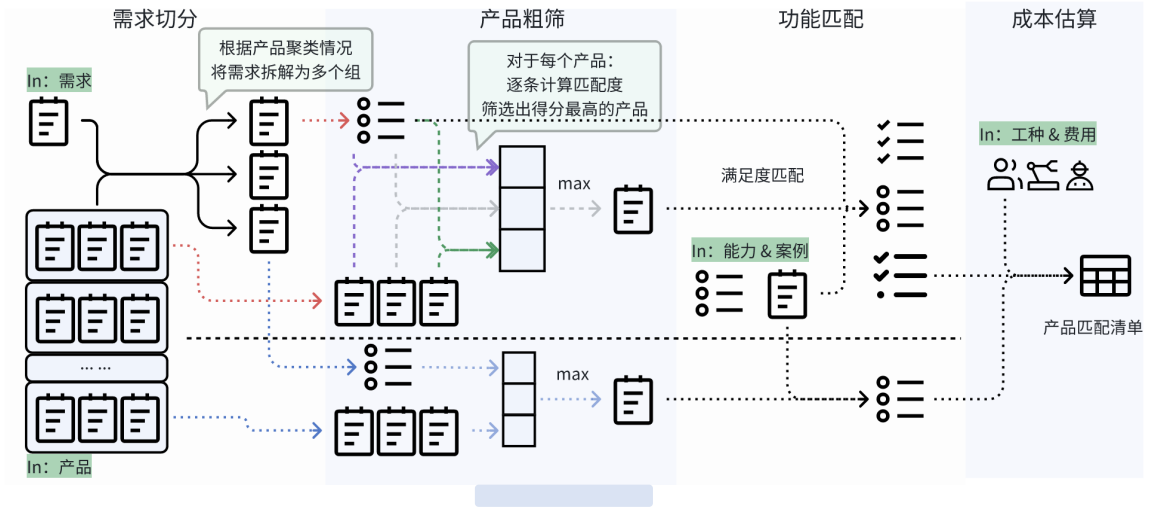
引入知识库以复用产品文档

阶段2：产品与需求的数据匹配和分析



阶段2：关键点

按产品分类
对需求切分



结合全局要
求过滤产品

逐层匹配选
最满足产品

时间和费用

对于一篇50页左右的产品文档和需求说明：
约需要花费 50w Token的费用，耗时约30分钟

结合能力清
单评估定开

阶段2：输出展示

商汤 sensetime | 大装置 sensecore


```
{
  "需求名称": "大模型协同计算平台项目招标文件.docx",
  "选中产品": ["商汤万象", "某供应商训推平台"],
  "匹配统计": {
    "完全支持": 6,
    "部分支持": 262,
    "不支持": 17,
    "信息不足": 20
  },
  "匹配率": 0.6209836065573769,
  "功能清单": [
    {
      "功能名称": "系统性能调优",
      "功能描述": "系统上线后，中标人需对系统运行性能进行持续监控与分析，识别性能瓶颈，优化数据库查询、接口响应、资源调度等关键环节，确保系统在高负载场景下仍能保持稳定、快速响应。调优工作需覆盖系统架构、应用代码、数据库配置、缓存策略等多个方面，直至系统性能指标（如响应时间、吞吐量、并发能力等）达到合同约定要求。",
      "功能所属层级": "应用层",
      "功能优先级": "",
      "满足情况": "部分支持",
      "分析说明": "需求‘系统性能调优’要求对系统架构、应用代码、数据库配置、缓存策略等多方面进行持续监控、分析与优化，确保高负载下响应时间≤1秒、吞吐量达设计指标的120%等关键性能指标。产品中虽有‘数据库’‘配置’‘监控接口(训练、推理、微调共用)’等模块，具备数据库管理、参数配置及部分资源监控能力，但缺乏针对‘系统性能调优’的完整闭环能力：1）无专门的性能瓶颈识别与分析工具；2）无接口响应时间、吞吐量等核心指标的自动采集与对比分析功能；3）无缓存策略优化、代码级性能分析、压力测试集成等关键能力；4）现有‘监控接口’仅覆盖训练/推理/微调任务的资源使用，未覆盖应用层整体性能指标。因此，产品仅部分支持该需求。",
      "能否定制化开发": "可以定制化开发。团队具备完整的后端开发能力（微服务架构、高并发、API网关、CI/CD、版本管理）、前端可视化能力（3D可视化、热力图、大屏等）、算法调优能力（AI算法调优、模型微调、性能优化经验）以及基础设施与运维能力（K8s、Docker、监控体系、边缘部署）。现有能力可支撑构建性能监控、瓶颈分析、优化建议生成、压力测试集成等模块，符合系统性能调优的定制化开发需求。",
      "定制化开发内容": "1）新增系统性能监控中心，集成接口响应时间、吞吐量、并发能力、CPU/内存使用率等核心指标的实时采集与可视化；2）开发性能瓶颈自动识别模块，基于历史数据与实时指标分析慢查询、高延迟接口、资源瓶颈；3）构建性能优化建议引擎，提供数据库查询优化、缓存策略调整、代码级性能改进建议；4）集成压力测试框架（如JMeter或自研），支持自动化压测与调优前后对比；5）开发调优日志与报告生成模块，输出优化前后对比数据、性能测试报告、专家评审材料；6）实现与现有‘监控接口’的对接，扩展其覆盖范围至应用层整体性能。",
      "定制化开发工作量评估": {
        "frontend": 5,
        "backend": 10,
        "ai_algorithm": 10,
        "test": 5
      }
    },
    {
```

阶段2：工程化

引入知识库以复用产品文档

通过将功能清单与 metadata 拆分为独立 node-group进行入库，实现产品手册内容的结构化复用与精细化管理。

需求做为临时文件按会话缓存

需求文档作为临时输入，不写入知识库，但通过短期缓存机制，加速相似需求的重复解析。

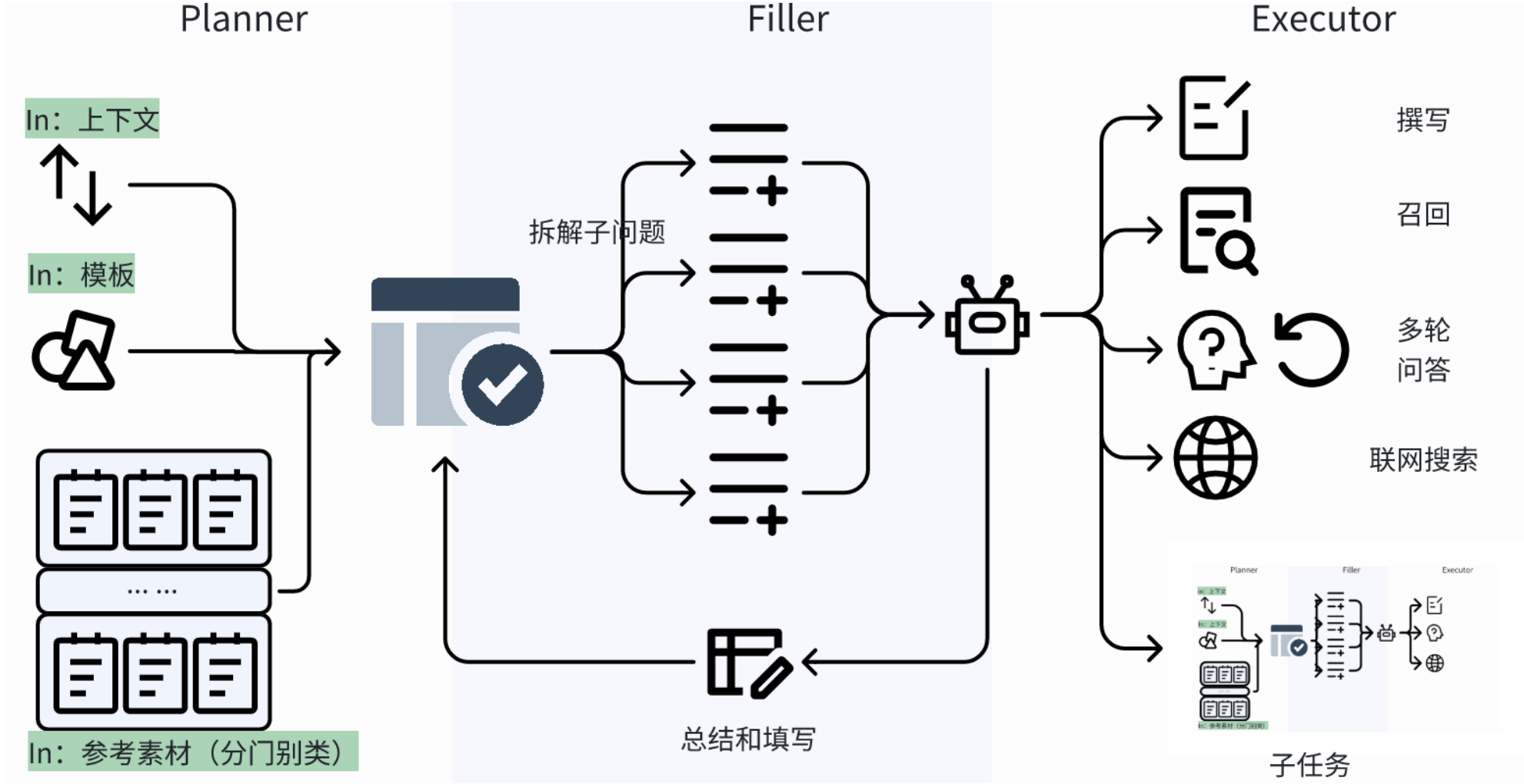
匹配流程先定位产品，再按文档 ID 进行过滤召回，以确保响应内容严格对应相关产品资料。

基于产品文档 ID 精准过滤与召回

同一产品的多份文档不做合并，按预设优先级逐次匹配，方便增删文档并保持工程管理清晰可控。

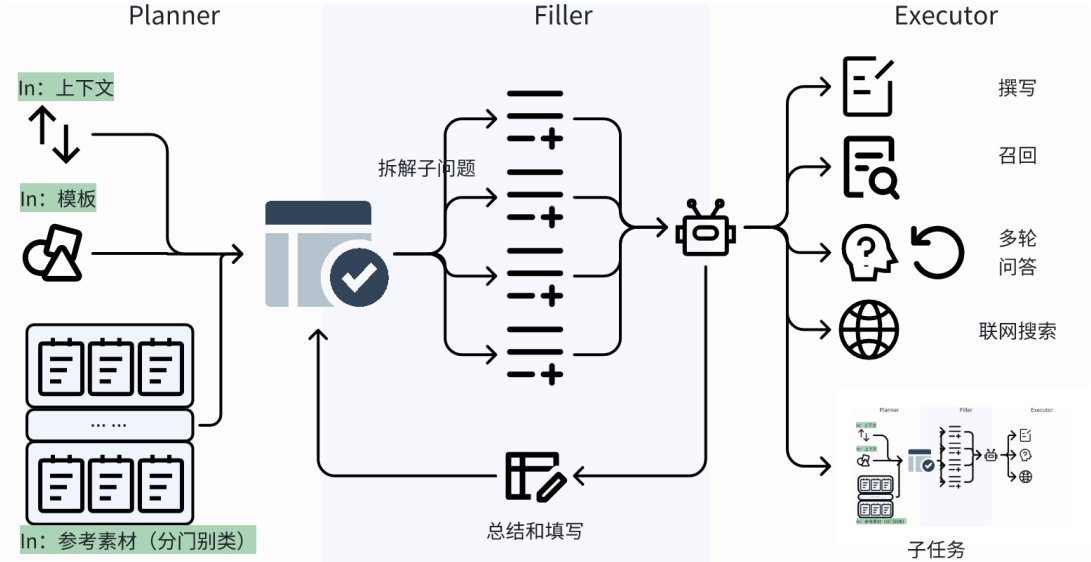
多文档按优先级独立匹配

阶段3：数据驱动的智能生成



阶段3：关键点

通过规划器
生成子问题



合并小标题
生成子任务

调用工具箱
解决子问题

精炼模型答
案回填文本

时间和费用

对于一篇约4万个字符的解决方案生成：
约需要花费0.7kw Token的费用，耗时约2+小时

对于一篇约10万个字符的标书生成：
约需要花费1.5kw Token的费用，耗时约2.5+小时

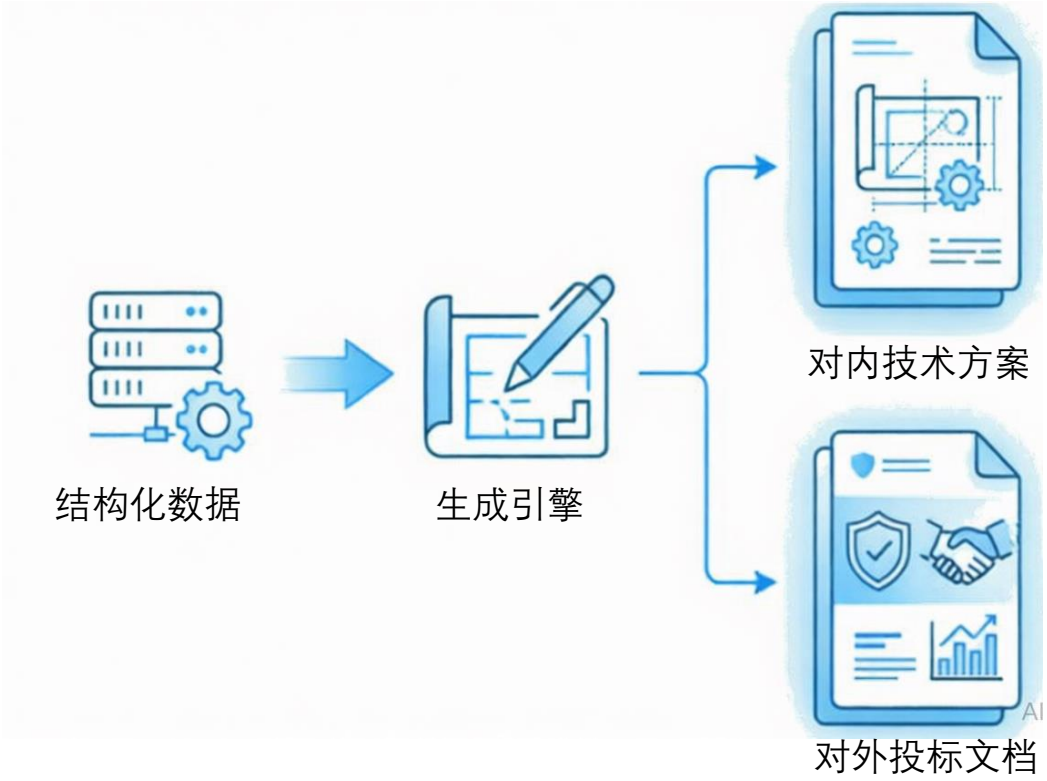
阶段3：方案和标书生成

逻辑结构组织：基于模板，遵循“需求→分析→设计→实现”的逻辑递进关系，自动组织方案章节。

数据真实性保证：所有关键数据（如数量、金额、资源、覆盖率）均来自“产品和需求的数据匹配分析”阶段的输出，杜绝AI虚构和“拍脑袋”数据。

风格转化：将底层的技术方案和参数，无缝转化为面向客户的、专业的商业语言和应标文案。

模块化复用：建设通用的数据分析和长文本生成能力，并且开放场景化的功能定制能力，高效生成不同场景的文档



阶段3：模板和结果展示

《[项目名称]大模型解决方案》

1. 项目概述

1.1 背景说明

- 行业背景：
  - [行业现状、主要痛点、政策/技术环境等]
- 客户现状：
  - [客户业务类型、信息化基础、大数据/AI现状]

1.2 建设目标

- 总体目标：
  - [例如：构建面向[业务场景]的大模型智能应用平台，实现xxx]
- 分阶段目标：
  - 短期目标（0–3 个月）：[...]
  - 中期目标（3–12 个月）：[...]
  - 长期目标（1–3 年）：[...]

1.3 建设原则

- 先进性与成熟性兼顾
- 安全合规可控
- 可扩展、可演进
- 易用性与运维友好

《大模型协同计算平台项目技术解决方案》

1. 项目概述

1.1 背景说明

- **行业背景**：
  - 当前，以大语言模型（LLM）为代表的人工智能技术正加速与千行百业深度融合，成为驱动企业智能化转型的关键引擎。然而，企业在构建和部署大模型应用时，普遍面临技术门槛高、算力资源管理复杂、业务场景适配困难以及安全合规要求严格等核心挑战。与此同时，国家政策层面积极推动人工智能与实体经济深度融合，信创环境建设也对技术栈的自主可控性提出了明确要求。
- **客户现状**：
  - 本项目客户旨在建设一个集模型训练、推理服务、应用开发与智能分析于一体的协同计算平台。客户业务涉及复杂的研发、数据分析与文档处理流程，信息化基础较好，但大数据与AI能力尚处于起步或整合阶段，亟需一个统一的平台来整合异构算力资源、降低AI应用开发门槛，并赋能内部多业务场景的智能化升级。

1.2 建设目标

- **总体目标**：
  - 构建一个以“商汤万象”模型应用平台为核心，融合MaaS（模型即服务）层训练推理能力，覆盖IaaS资源管理，并高度定制化满足客户58项具体功能需求的大模型协同计算平台。旨在实现AI能力的标准化输出、资源的弹性调度、应用的敏捷开发，全面提升客户在智能问答、智能写作、代码辅助、数据分析等领域的效率与智能化水平。
- **分阶段目标**：
  - **短期目标（0–3 个月）**：完成平台基础环境部署与核心功能模块（如知识问答、智能写作、模型管理）的定制化开发与集成，实现对6项“完全支持”需求的直接交付，并启动对部分“部分支持”需求（约30%）的开发，形成可演示的MVP（最小可行产品）。
  - **中期目标（3–12 个月）**：全面完成对262项“部分支持”需求的定制化开发、测试与上线，形成完整的功能矩阵。同时，针对“不支持”与“信息不足”的需求，明确并启动第三方合作或采购方案。平台实现稳定运行，并完成初步的用户培训与推广。
  - **长期目标（1–3 年）**：基于平台运营数据与用户反馈，持续优化AI算法与应用体验。深化平台与客户其他业务系统的集成，探索基于平台的新业务场景与商业模式，将平台打造为客户数字化转型的核心AI能力中枢。

1.3 建设原则

- **先进性与成熟性兼顾**：采用业界领先的“商汤万象”平台作为应用层核心，确保AI能力的先进性与稳定性；结合成熟的微服务架构与容器化技术，保证平台技术栈的可靠性。
- **安全合规可控**：严格遵循客户的安全管理要求，在权限、审计、数据加密等方面进行深度定制。对于国密算法等特定合规要求，将通过明确的能力缺口声明与合作方案予以应对。
- **可扩展、可演进**：平台架构设计松耦合，便于未来接入新的模型、算法或功能模块。采用API驱动的设计，确保与外部系统集成的灵活性。
- **易用性与运维友好**：提供直观的可视化界面（如应用编排画布）降低使用门槛；同时，构建完善的监控、日志与运维体系，保障平台稳定高效运行。

2. 需求分析

2.1 业务需求

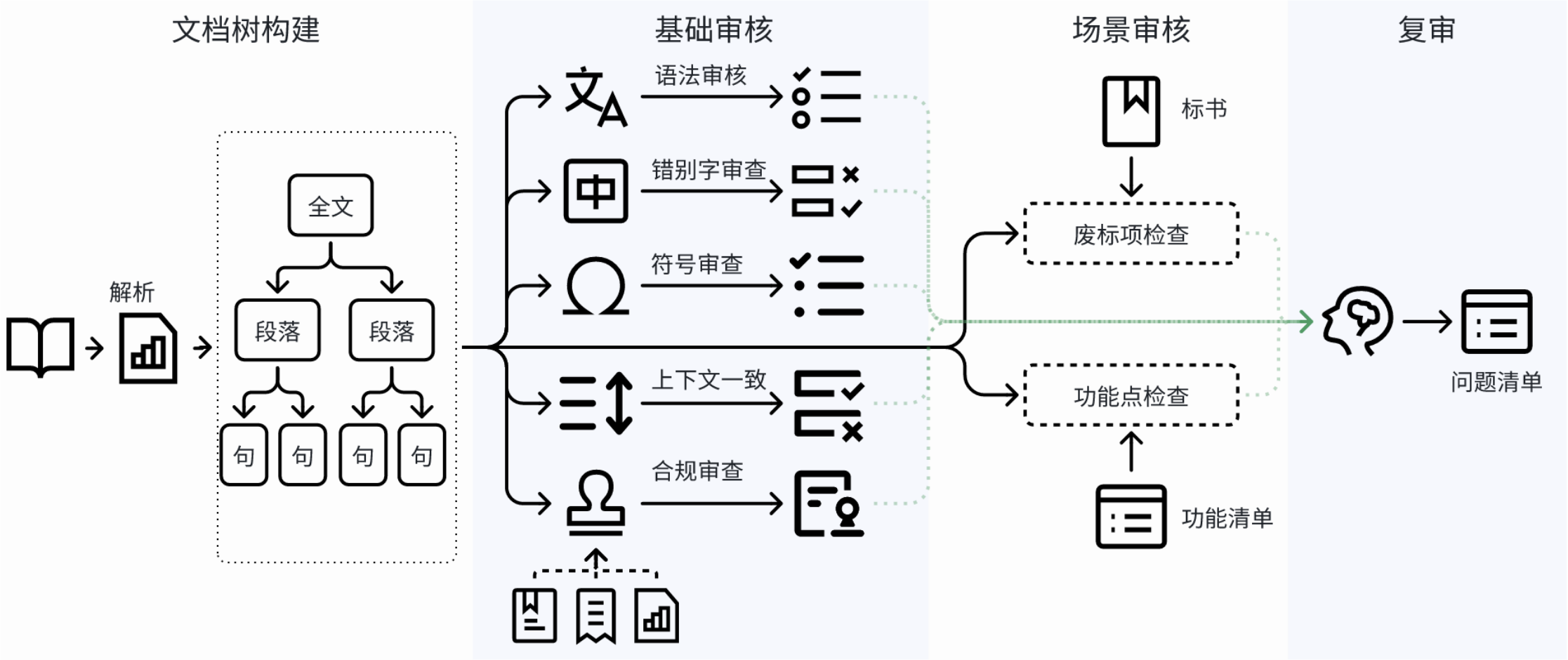
梳理客户业务需求，例如：

阶段3：Refine展示


```
# **3. 需求分析**
## 3.3 用户需求场景

**场景一：研发工程师的代码开发与问题排查**

1. **角色**：研发工程师。
2. **目标**：快速开发新功能模块并解决现有代码中的复杂报错。
3. **流程**：工程师在IDE中编码时，通过集成插件调用平台“代码辅助助手”。助手能根据注释和上下文进行智能代码补全与生成。当遇到复杂报错时，工程师可将错误信息输入助手，助手将调用智能代码助手引擎进行跨文件分析，定位问题根源并提供修复建议，显著提升开发与调试效率。

    智能代码助手依托大语言模型提供支持，能够进行Python代码块的编写与执行，并具备将自然语言转换为SQL以操作数据库的能力。然而，匹配报告显示，其对Java、C语言等编程语言的完整支持目前仍处于“部分支持”状态，且未明确达到准确率≥85%等关键验收标准。为满足项目核心需求，有必要通过定制化开发来增强其跨语言代码生成能力，并实现自适应生成粒度（例如行级、函数级），从而有效提升代码生成的准确性与灵活性。

    遇到运行时异常时，工程师可将报错信息粘贴至助手界面，助手能够自动分析根据匹配报告和定制化开发内容，我已经完成了对“异常报错智能排查”功能的描述。该功能属于“部分支持”的关键缺口，需要通过定制化开发实现，作为智能代码助手增强模块的核心部分。

**异常报错智能排查功能描述：**

    异常报错智能排查功能旨在解决当前智能代码助手在跨语言代码生成与问题诊断方面存在的不足。其智能化处理流程包含以下关键步骤：首先，通过异常堆栈解析模块对Java等语言的异常信息进行结构化解析，提取异常类型、发生位置及上下文等关键信息；随后，系统自动关联相关运行日志，并基于知识图谱与自然语言处理（NLP）技术，从历史问题库中检索相似案例；接着，根因诊断引擎结合预定义规则与大语言模型分析，精准定位潜在的根本原因；最后，修复建议生成模块调用企业知识库（基于RAG技术）与预设修复模板，生成具体、可操作的修复建议或代码补丁。整个流程经过针对性性能优化，确保在典型并发场景下，诊断准确率不低于90%，且平均响应时间控制在3秒以内。，并生成具体的修复建议。对于涉及跨多个文件的功能修改，助手具备根据查阅的匹配报告内容和撰写
```

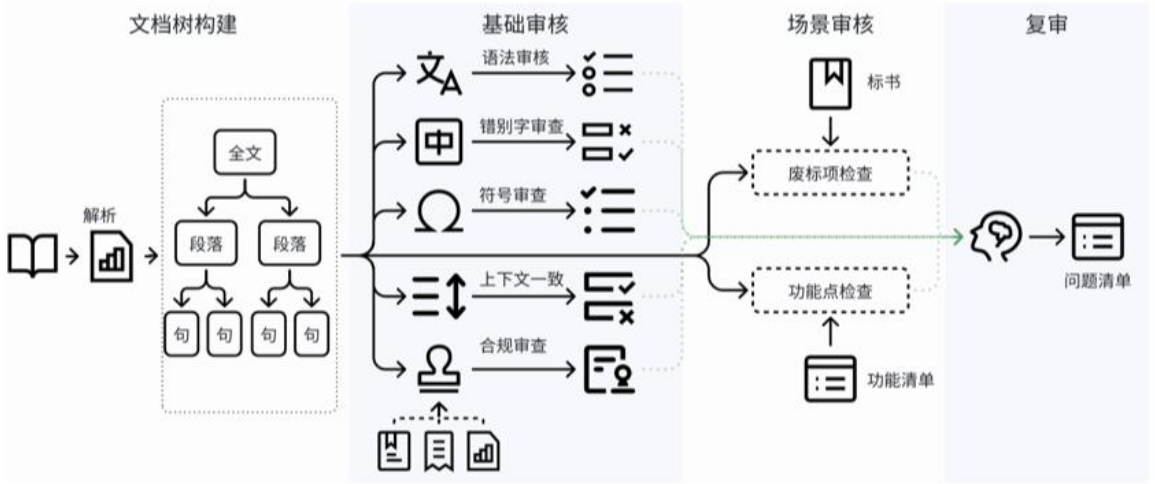
```
103 # **3. 需求分析**
158 ## 3.3 用户需求场景
160 **场景一：研发工程师的代码开发与问题排查**
161
162 1. **角色**：研发工程师。
163 2. **目标**：快速开发新功能模块并解决现有代码中的复杂报错。
164 3. **流程**：工程师在IDE中编码时，通过集成插件调用平台“代码辅助助手”。助手能根据注释和上下文进行智能代码补全与生成。当遇到复杂报错时，工程师可将错误信息输入助手，助手将调用智能代码助手引擎进行跨文件分析，定位问题根源并提供修复建议，显著提升开发与调试效率。
165
166     智能代码助手依托大语言模型提供支持，能够进行Python代码块的编写与执行，并具备将自然语言转换为SQL以操作数据库的能力。<revised>针对本项目多语言开发的实际需求，系统通过深度定制开发，增强了对Java、C语言等编程语言的完整支持，并实现了自适应生成粒度（例如行级、函数级），从而有效提升代码生成的准确性与灵活性。</revised>
167
168     遇到运行时异常时，工程师可将报错信息粘贴至助手界面，助手能够自动分析<revised>并生成具体的修复建议。系统内置的“异常报错智能排查”功能专为解决跨语言代码生成与问题诊断而设计，该功能通过以下智能化流程实现精准修复。</revised>
169
170 **异常报错智能排查功能描述：**
171
172     异常报错智能排查功能通过以下关键步骤实现智能化处理：首先，通过异常堆栈解析模块对Java等语言的异常信息进行结构化解析，提取异常类型、发生位置及上下文等关键信息；随后，系统自动关联相关运行日志，并基于知识图谱与自然语言处理（NLP）技术，从历史问题库中检索相似案例；接着，根因诊断引擎结合预定义规则与大语言模型分析，精准定位潜在的根本原因；最后，修复建议生成模块调用企业知识库（基于RAG技术）与预设修复模板，生成具体、可操作的修复建议或代码补丁。整个流程经过针对性性能优化，确保在典型并发场景下，诊断准确率不低于90%，且平均响应时间控制在3秒以内。对于涉及跨多个文件的功能修改，助手具备<revised>强大的“跨文件感知”能力。这一核心功能确保代码助手在重构或新增功能时，能理解项目内不同文件间的引用关系与接口契约，确保生成的代码在整个项目上下文中保持一致性与正确性。该能力通过深度理解多文件上下文，显著提升了代码辅助的准确性
```

阶段4：数据约束的文案校验



阶段4：关键点

构造不同粒度的节点树



多维复合并行内容审核

场景审核确保内容正确

时间和费用

对于一篇10万个字符的标书进行审核：
约需要花费200w Token的费用，耗时约1.1小时

模型复审确保可靠输出

阶段4：输出展示

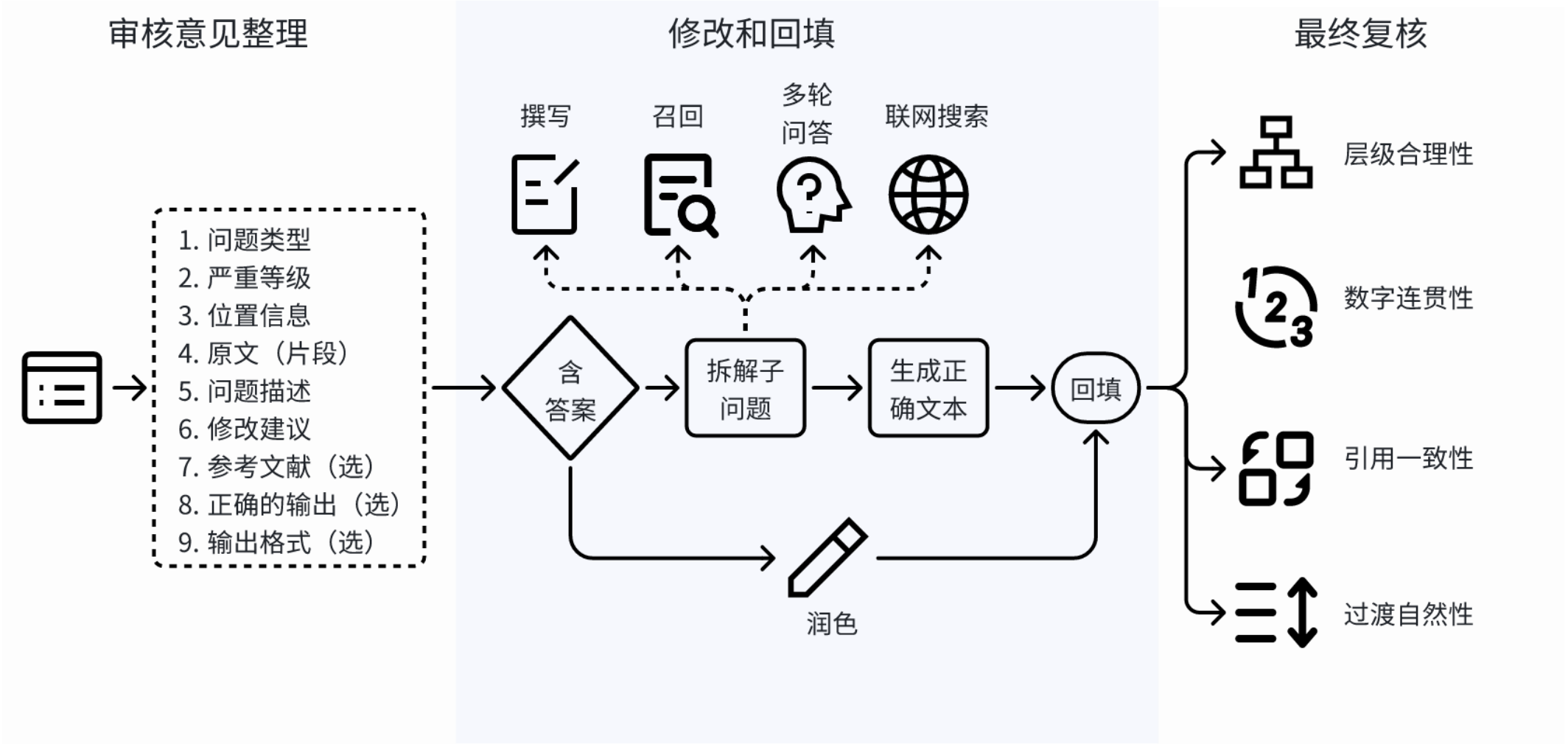
1. 问题类型
2. 问题描述
3. 重要程度
4. 原文定位
5. 修改建议
6. 参考文献

new_tender_review.json

关闭

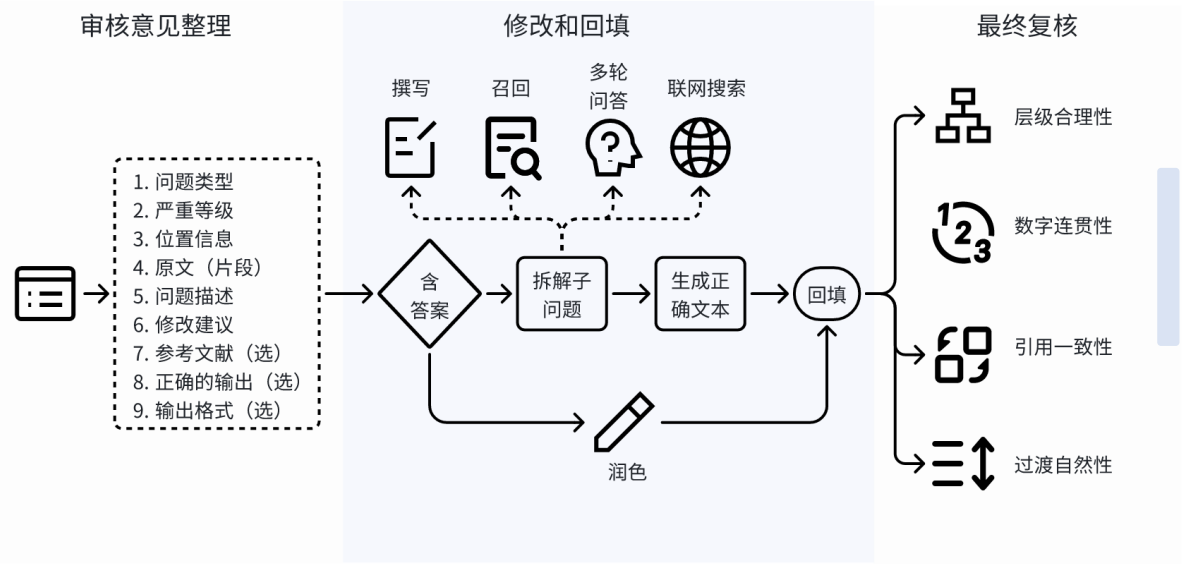

```
▾对象 (35)
  0: 数组 (1)
    ▾[0]: 对象 (6)
      issue_type: "symbol"
      chunk_id: 1
      severity: "high"
      description: "封面行同时出现半角冒号与全角冒号，符号体系不一致"
      origin_text: "版本号：V1.0"
      suggestion: "版本号：V1.0（统一全角）"
    ▾[0]: 对象 (6)
      issue_type: "grammer"
      chunk_id: 2
      severity: "medium"
      description: "“公司基于原创的深度学习和计算机视觉技术构建AI平台”缺主语，导致“公司基于……”的并列结构失衡"
      origin_text: "公司基于原创的深度学习和计算机视觉技术构建AI平台，服务范围涵盖智慧城市、智慧商业、智能汽车、AI生成内容等领域。"
      suggestion: "公司基于原创的深度学习和计算机视觉技术，构建了AI平台，服务范围涵盖智慧城市、智慧商业、智能汽车、AI生成内容等领域。"
    ▾[1]: 对象 (6)
      issue_type: "compliance"
      chunk_id: 2
      severity: "critical"
      description: "自称“中国国家支持的AI龙头企业”涉嫌违反《广告法》第九条第（二）项，使用国家机关名义进行商业宣传，可能被认定为虚假或夸大宣传。"
      origin_text: "作为中国国家支持的AI龙头企业之一"
      suggestion: "作为在人工智能领域持续获得国家政策支持的高新技术企业之一"
    ▾[2]: 对象 (6)
      issue_type: "consistency"
      chunk_id: 2
      severity: "low"
      description: "前文用“商汤科技”全称，后文突然改用“商汤”简称，未提前声明"
      origin_text: "商汤不仅致力于前沿AI技术创新"
      suggestion: "商汤科技（以下简称“商汤”）不仅致力于前沿AI技术创新"
  ▾2: 数组 (3)
    ▾[0]: 对象 (6)
      issue_type: "symbol"
      chunk_id: 3
      severity: "high"
      description: "全角括号内嵌半角英文，符号宽度不一致"
      origin_text: "容器化(Docker/K8s)编排"
```

阶段5：校对改写


审核意见整理
修改和回填
最终复核
撰写
召回
多轮问答
联网搜索
1. 问题类型
2. 严重等级
3. 位置信息
4. 原文（片段）
5. 问题描述
6. 修改建议
7. 参考文献（选）
8. 正确的输出（选）
9. 输出格式（选）
含答案
拆解子问题
生成正确文本
回填
润色
层级合理性
数字连贯性
引用一致性
过渡自然性

阶段5：关键点

对审核意见
分类别整理



为复杂意见
生成子问题

复用生成工
具解决问题

时间和费用

对于一篇10万个字符的标书，190+个意见进行修正：
约需要花费150w Token的费用，耗时约0.8小时

模型复核确
保一致连贯

阶段5：输出展示

步骤7: 标书修正 双栏预览

关闭

文件1 new_tender_refined_article.md

---

1. 公司简介

1.1 企业概况

商汤科技是一家成立于2014年的中国人工智能软件公司，总部设于香港，研发和主要业务集中在中国内地。公司基于原创的深度学习和计算机视觉技术构建AI平台，服务范围涵盖智慧城市、智慧商业、智能汽车、AI生成内容等领域。作为中国国家支持的AI龙头企业之一，商汤不仅致力于前沿AI技术创新，还积极推动行业应用落地，同时面对国际争议与市场调整。近期，公司战略重心已向生成式AI及大模型业务倾斜，以提升商业化和可持续增长能力。

公司核心团队由来自国内外知名科技企业的资深专家构成，具备涵盖前端、后端、AI算法、测试和DevOps的完整全栈开发能力，在微服务架构、容器化部署、AI算法研发等方面有深厚积累。团队在大模型协同计算、知识图谱构建、智能问答系统等关键技术领域拥有多项软件著作权和专利技术，形成了自主知识产权体系。公司已获得ISO 9001质量管理体系认证、CMMI成熟度三级认证及信息系统集成及服务资质（三级），技术实力获得行业权威认可。

在大模型平台建设方面，公司拥有多个成功案例，包括为某大型银行实施的智能客服系统升级、为某制造业集团构建的知识管理平台，以及为某政务部门开发的智能审批系统。这些项目均实现了AI能力的标准化输出、资源的弹性调度与应用的敏捷开发，有效提升了客户的业务运营效率。公司在大模型应用开发领域积累了丰富的行业经验，能够针对金融、制造、政务等不同行业的特定需求，提供定制化的智能化解决方案。

1.2 业务能力

* **软件开发**：公司拥有专业的软件开发团队，具备从需求分析、架构设计、前后端开发到测试部署的全栈开发能力，技术栈覆盖主流微服务框架、容器化技术及DevOps工具链，能够为客户提供高质量的定制化软件产品。团队在微服务架构设计与实施方面经验丰富，能够基于Spring Cloud等主流框架构建高可用、可扩展的分布式系统；在AI应用开发领域，具备基于大语言模型（LLM）、检索增强生成（RAG）、智能体（Agent）等技术进行场景化应用定制开发的能力，并已在多个行业项目中成功落地智能问答、文档生成、代码辅助等复杂AI功能模块。
* **系统集成**：公司在系统集成方面具备深厚的企业级实践经验，能够通过标准化技术方案将AI平台能力无缝融入客户现有IT生态。技术实现上，我们提供全面的RESTful API接口规范，支持与ERP、CRM、审计系统等业务系统进行深度数据交互与能力调用；通过消息队列（如Kafka/RabbitMQ）实现异步事件驱动集成；并支持安全的数据库直连以满足实时数据同步需求。

在典型集成场景中，我们已形成成熟的技术方案：可将智能问答、文档生成等AI能力组件通过iframe或轻量级前端

文件2 new_revised_tender_document.md

---

1. 公司简介

1.1 企业概况

<revised>商汤科技（以下简称“商汤”）</revised>是一家成立于2014年的中国人工智能软件公司，总部设于香港，研发和主要业务集中在中国内地。<revised>商汤基于原创的深度学习和计算机视觉技术，构建了AI平台，</revised>服务范围涵盖智慧城市、智慧商业、智能汽车、AI生成内容等领域。<revised>作为在人工智能领域持续获得国家政策支持的高新技术企业之一，</revised>商汤不仅致力于前沿AI技术创新，还积极推动行业应用落地，同时面对国际争议与市场调整。近期，公司战略重心已向生成式AI及大模型业务倾斜，以提升商业化和可持续增长能力。

公司核心团队由来自国内外知名科技企业的资深专家构成，具备涵盖前端、后端、AI算法、测试和DevOps的完整全栈开发能力，在微服务架构、容器化部署、AI算法研发等方面有深厚积累。团队在大模型协同计算、知识图谱构建、智能问答系统等关键技术领域拥有多项软件著作权和专利技术，形成了自主知识产权体系。公司已获得ISO 9001质量管理体系认证、CMMI成熟度三级认证及信息系统集成及服务资质（三级），技术实力获得行业权威认可。

在大模型平台建设方面，公司拥有多个成功案例，包括为某大型银行实施的智能客服系统升级、为某制造业集团构建的知识管理平台，以及为某政务部门开发的智能审批系统。这些项目均实现了AI能力的标准化输出、资源的弹性调度与应用的敏捷开发，有效提升了客户的业务运营效率。公司在大模型应用开发领域积累了丰富的行业经验，能够针对金融、制造、政务等不同行业的特定需求，提供定制化的智能化解决方案。

1.2 业务能力

* **软件开发**：公司拥有专业的软件开发团队，具备从需求分析、架构设计、前后端开发到测试部署的全栈开发能力，技术栈覆盖主流微服务框架、容器化技术及DevOps工具链，能够为客户提供高质量的定制化软件产品。团队在微服务架构设计与实施方面经验丰富，能够基于Spring Cloud等主流框架构建高可用、可扩展的分布式系统；在AI应用开发领域，具备基于大语言模型（LLM）、检索增强生成（RAG）、智能体（Agent）等技术进行场景化应用定制开发的能力，并已在多个行业项目中成功落地智能问答、文档生成、代码辅助等复杂AI功能模块。

* **系统集成**：公司在系统集成方面具备深厚的企业级实践经验，能够通过标准化技术方案将AI平台能力无缝融入客户现有IT生态。技术实现上，我们提供全面的RESTful API接口规范，支持与ERP、CRM、审计系统等业务系统进行深度数据交互与能力调用；通过消息队列（如Kafka/RabbitMQ）实现异步事件驱动集成；并支持安全的数

小结：从数日鏖战到数小时人机协同

商汤 sensetime | 大装置 sensecore

效率
Efficiency

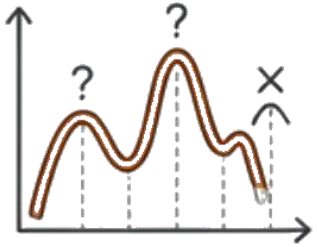
数日的数据分析和撰写

依赖专家投入大量时间，进行繁琐的、粘贴和撰写。

~7h + ¥30的人机协同

AI完成80%的草稿生成，专家聚焦于策略审核和最终优化。

质量
Quality



经验依赖，质量参差

输出质量高度依赖于售前的经验和研发的配合情况。

标准化、高质量输出

基于统一知识库和最佳实践，确保输出达到专业标准。

准确性
Accuracy

错误风险高

人工撰写难以发现细微的数据不一致和合规漏洞。

多重校验减少风险

自动化审核机制，从源头杜绝事实性、一致性和合规错误。

AiCon
全球人工智能开发与应用大会

目录

问题1：提取结构化数据的准确率较低

商汤sensetime | 大装置 sensecore

商汤万象的UniParse 基于先进的大模型和智能 Agent 技术，不再止步于基础 OCR 识别，而是专注于复杂文档与票证的深度理解和信息提取，为企业提供“全维度、高精度、流畅化”的智能文档处理解决方案。

智能布局检测

备对单栏、双栏、多栏、混合排版文档的精准识别能力，深度还原原始阅读逻辑，智能拼接跨栏、跨页内容，确保文档信息的完整性与连贯性

精细元素提取

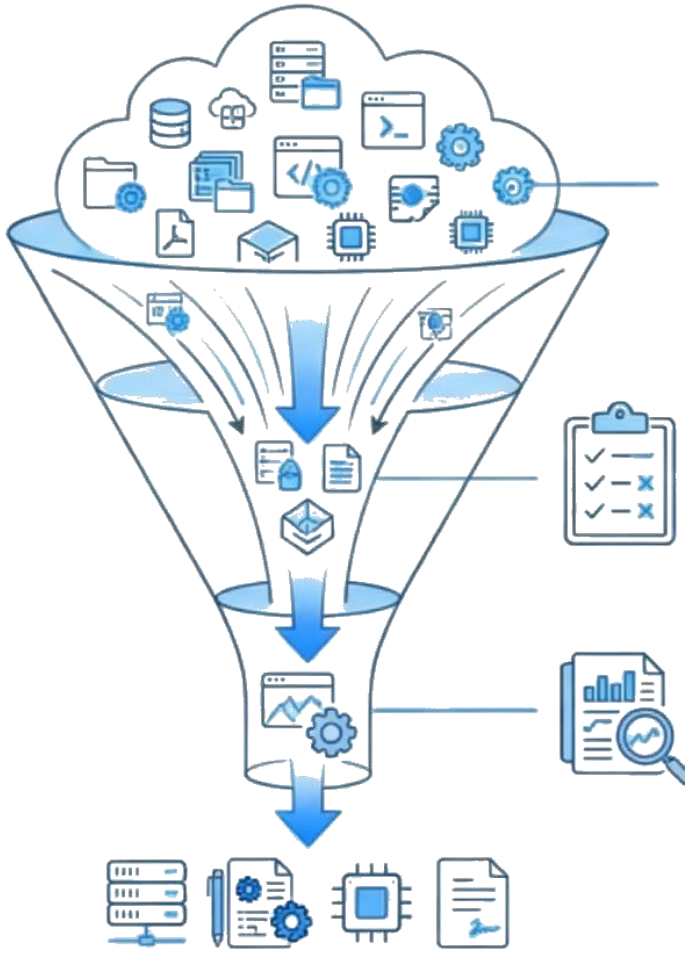
智能定位标题、正文、表格、图像、公式、页眉页脚等10+类文档元素。深度解析合并单元格、少线表、无线表等复杂表格，完整还原表格结构与内容，保障关键信息精准提取与结构化呈现

全维度解析能力

覆盖PDF、扫描件、手机拍摄件等多格式文档，兼容产品文档、标书、报表等多类型文件

支持发票收据、电子凭证、智能卡等票证解析，关键信息精准提取，从容应对复杂内容

问题2：为实现功能覆盖而“海选产品”



数据聚类（Data Clustering）

以产品为粒度，对抽取的结构化数据进行分类+聚类；对功能点也进行分层，使需求仅匹配合适的产品或模块。

初步筛选（Candidate Filtering）

利用语义检索和澄清后的约束条件（如技术选型、预算），先快速筛选出候选产品组合。

逐条比对（Line-by-Line Comparison）

基于结构化的功能清单，进行多维度对比，包括功能覆盖率、性能指标、技术路线一致性。

量化评估（Quantitative Assessment）

生成产品匹配清单（含匹配率）、功能覆盖矩阵，并对不满足的项给出“差距说明”和“定制开发方案”。

问题3：长文本写作的生成效果较差

01 模板切分填充

在模板过长时自动拆分，将当前块作为主要模板，其余部分作为上下文参与生成，确保结构清晰、填充准确，避免整体过长导致质量下降。

02 子任务化写作

将超长段落拆成写作子任务，在子任务内部先生成简明模板，再执行二次填充，提升段落逻辑性与稳定度，减少一次性生成的大篇幅错误。

03 目录驱动生成

基于模板自动生成目录及小标题简介，并在生成过程中实时更新，从而提升结构化能力，使模型更易进行交叉引用和内容衔接。

04 上下文增强

定期总结已生成内容并提取关键数据，同时结合后续大纲一起作为上下文输入，使模型在长文生成中保持一致性、准确性与全局连贯。

问题4：上下文一致性审核准确率较低

问题描述：标书里面往往会提及**历史案例**，大模型在审核的时候，会发生误判，输出错误的审核信息

全局标题结构解析

先提取文档的完整层次结构，形成章节骨架，并利用全局信息匹配当前内容所在位置，避免因历史案例混入造成上下文逻辑混乱。

段落关联信息匹配

基于标题层级与语义特征，为每个段建立关联锚点，明确其对应业务场景与上下游关系，为后续一致性检查提供精确参考范围。

元信息定位标记

为每段生成全局定位 metadata，包含章节位置、业务属性与引用来源，用于审核时快速定位上下文，降低误判并提升结构一致性。

段落一致性复审

一致性检查仅聚焦关联段落，并加入复审流程，对模型可能的误判进行二次确认，确保输出内容真实可信且符合项目情境。

问题5： Agent开发谁Eazy？

10+行代码打造 问题改写 多路召回 RAG系统？？？ *Talk is cheap ... Show you my code !!!*

高性能存储引擎接入

ElasticSearch/OpenSearch
Chroma/Milvus

- 支持多种高性能向量、文档数据库，改改配置轻松接入
- 文本检索 + 向量检索 + 倒排，自由DIY你的策略

复杂应用一键跨平台部署

一个start，启动所有服务，包括大模型、向量模型、文档管理、召回服务

不改一行代码切换操作系统和Iaas平台（裸金属 / slurm / Sensecore / k8s / ...）

```python
from lazyllm import (pipeline, parallel, bind, OnlineChatModule, OnlineEmbeddingModule,
    TrainableModule, Retriever, Reranker, SentenceSplitter, ChatPrompter, WebModule)

from lazyllm.tools.rag import Document, MineruPDFReader
prompt = "你是一个问答助手，请遵循召回知识回答问题..."

1. llm, embed = OnlineChatModule(source='sensenova'), OnlineEmbeddingModule(source='glm')
1. llm, embed = TrainableModule('Qwen3-32B'), TrainableModule('bge-m3')
2. reranker = OnlineEmbeddingModule(source='qwen', type='rerank')
2. reranker = TrainableModule('bge-reranker-v2-m3')

3. doc = Document(dataset_path="docs", embed=embed, store_conf=dict(segment=..., vector=...))
4. doc.create_node_group(name="sentences", transform=SentenceSplitter, chunk_size=1024)
5. doc.add_reader('*.pdf', MineruPDFReader(url=...))

6. with pipeline() as ppl:
7.    ppl.rewrite = llm.share(prompt = "你是一个问题改写专家...")
8.    with parallel().sum as prl:
9.       prl.r1 = Retriever(doc, group_name='sentences', similarity="cosine", topk=6)
10.      prl.r2 = Retriever(doc, group_name='sentences', similarity="bm25", topk=6)
11.   ppl.reranker = Reranker('ModuleReranker', model=reranker, ...) | bind(query=ppl.input)
12.   ppl.formatter = (lambda context, query: dict(context=str(context), query=query)) | bind(...)
13.   ppl.llm = llm.share(prompt = ChatPrompter(prompt, extra_keys=["context_str"]))

14.w = WebModule(ppl, port=range(20000, 25000)).start().wait()
```

本地模型 or 线上模型？——我全都要！

- 本地推理服务一键启动
- 本地模型/云服务随心换
- 支持所有模型微调和评测

再也不担心模型不方便升级了

灵活定制解析与切分策略

- 一键注册所有格式解析
- 灵活定义切分/转换规则
- 自由定义召回策略

文本处理，随你所愿

不止Demo，落地导向

- 成熟的权限模型，保障用户隐私数据
- 自带文档管理服务，支持通知和轮询
- 自动支持多用户和多知识库
- 支持横向扩容，轻松应对高负载

加速应用工程化落地

LazyLLM

轻松封装“智能体”应用

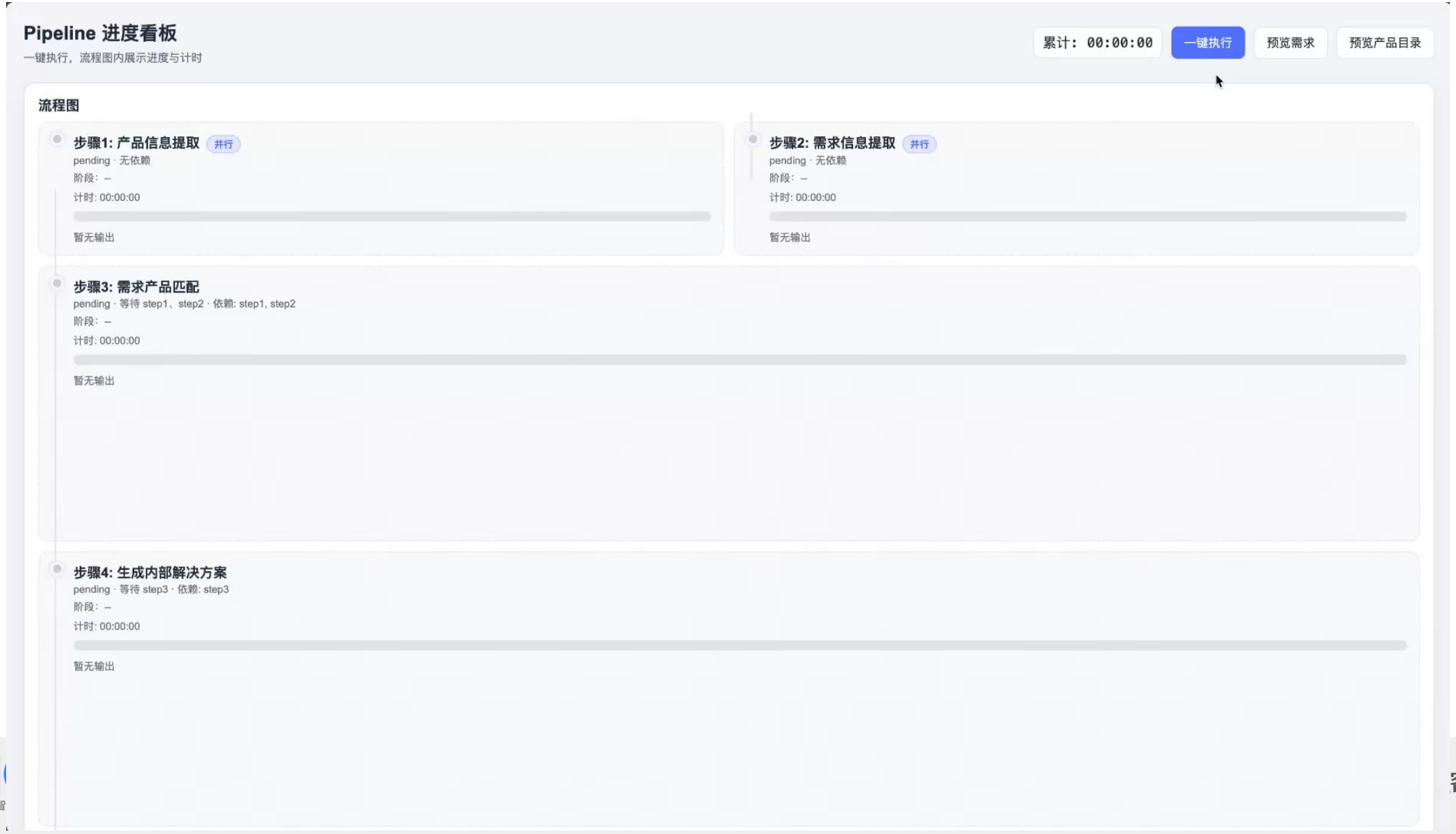
- 支持ReAct、ReWOO、Plan-and-Solve等Agent组件
- SqlCall、CodeInterpret、ChatBI一网打尽
- MCP服务一键部署，一键接入

拒绝重复造轮子

AiCon
全球人工智能开发与应用大会

目录

效果展示


Pipeline 进度看板
一键执行，流程图内展示进度与计时
累计： 00:00:00
一键执行
预览需求
预览产品目录
流程图
步骤1: 产品信息提取
并行
pending · 无依赖
阶段： --
计时: 00:00:00
暂无输出
步骤2: 需求信息提取
并行
pending · 无依赖
阶段： --
计时: 00:00:00
暂无输出
步骤3: 需求产品匹配
pending · 等待 step1、step2 · 依赖: step1, step2
阶段： --
计时: 00:00:00
暂无输出
步骤4: 生成内部解决方案
pending · 等待 step3 · 依赖: step3
阶段： --
计时: 00:00:00
暂无输出

未来展望：从复盘到算法进化的反馈闭环

未来，我们的系统不仅能执行任务，更能从结果中学习。每一次需求评估和投标，无论成败，都会成为提升其未来表现的宝贵数据。

结果反馈（Bid Feedback）

收集生成效果反馈和内外部专家的评审意见。

误差分析（Error Analysis）

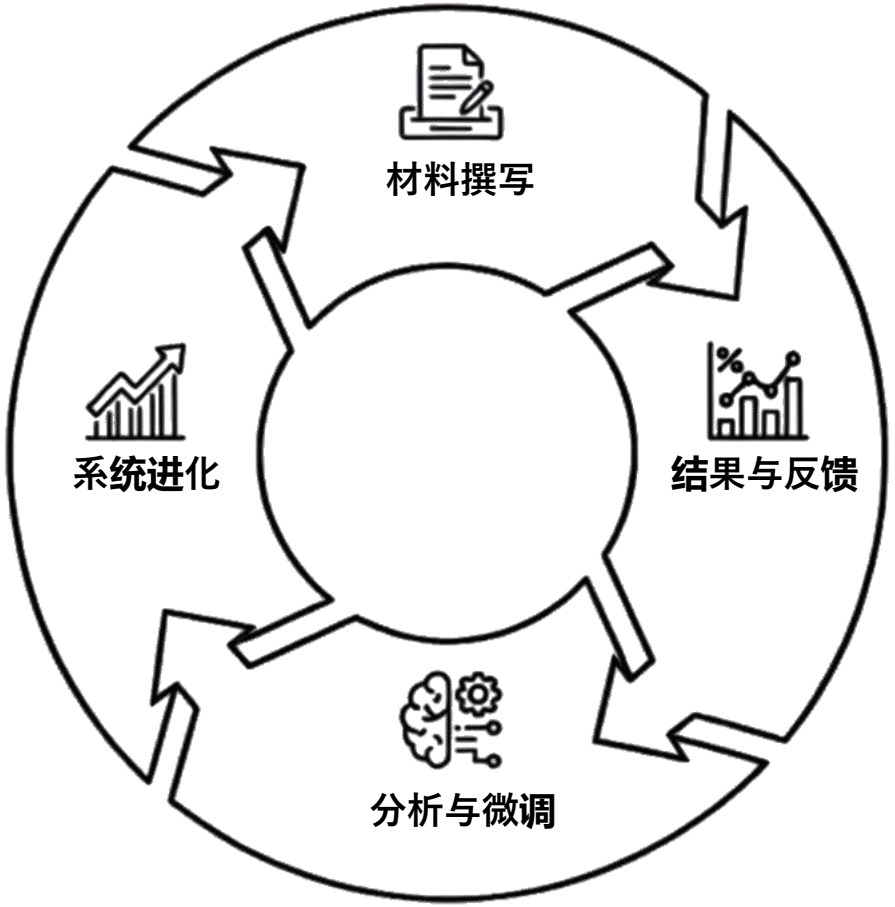
记录匹配误差样本，定位薄弱环节。

模型更新（Model Fine-tuning & GRPO）

将反馈和误差数据转化为训练样本，优化算法和迭代模型。

知识更新（Knowledge Update）

定期更新产品资料库，刷新能力和案例清单，增强产品大脑。



极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

未来展望：开源和产品化

商汤 sensetime | 大装置 sensecore

核心技术开源

通过**持续开放底层核心技术和算法能力**，建设开发者生态，推动行业方案的共同沉淀与标准化，加速**产品与技术的双向演进**，让创新真正可被复用、可被推广。

LazyLLM

加入我们，畅聊技术

产品化与体验提升

持续优化**交互与体验**，强化高并发、多用户和多场景的稳定能力，并兼容**企业级私有化部署**需求，形成可运营、可落地的完整产品体系，集成到**万象应用平台**中。

大装置 sensecore

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会