

INTSIG 合合信息™

大模型文档应用加速器

——TextIn·通用文档解析



目录

CONTEXT

01

产品基本信息

02

主要应用场景

03

产品性能

04

客户案例

01. 产品基本信息

- 1-1 产品背景
- 1-2 产品定位
- 1-3 标准服务模式

文档识别现状
场景少、内容少、处在价值链底部



识别

文字识别、表格识别、
手写字符识别



处理

水印去除、笔迹擦除



还原

版面还原、表格还原、
文字样式还原



比对

文档验真、合同比对



用户还需要完成的是更高价值的内容



分析

文字识别、表格识别、
手写字符识别



总结

版面还原、表格还原、
文字样式还原



输出

文档验真、合同比对



行动

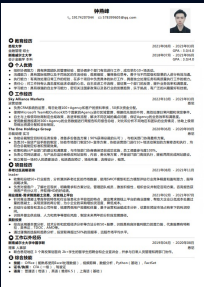
水印去除、笔迹擦除

LLM出现前的典型文档应用场景

识别 → 处理 → 还原 → 比对 → 分析 → 总结 → 输出 → 行动

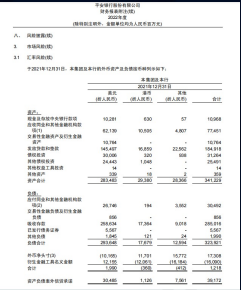
现有流程

招聘场景



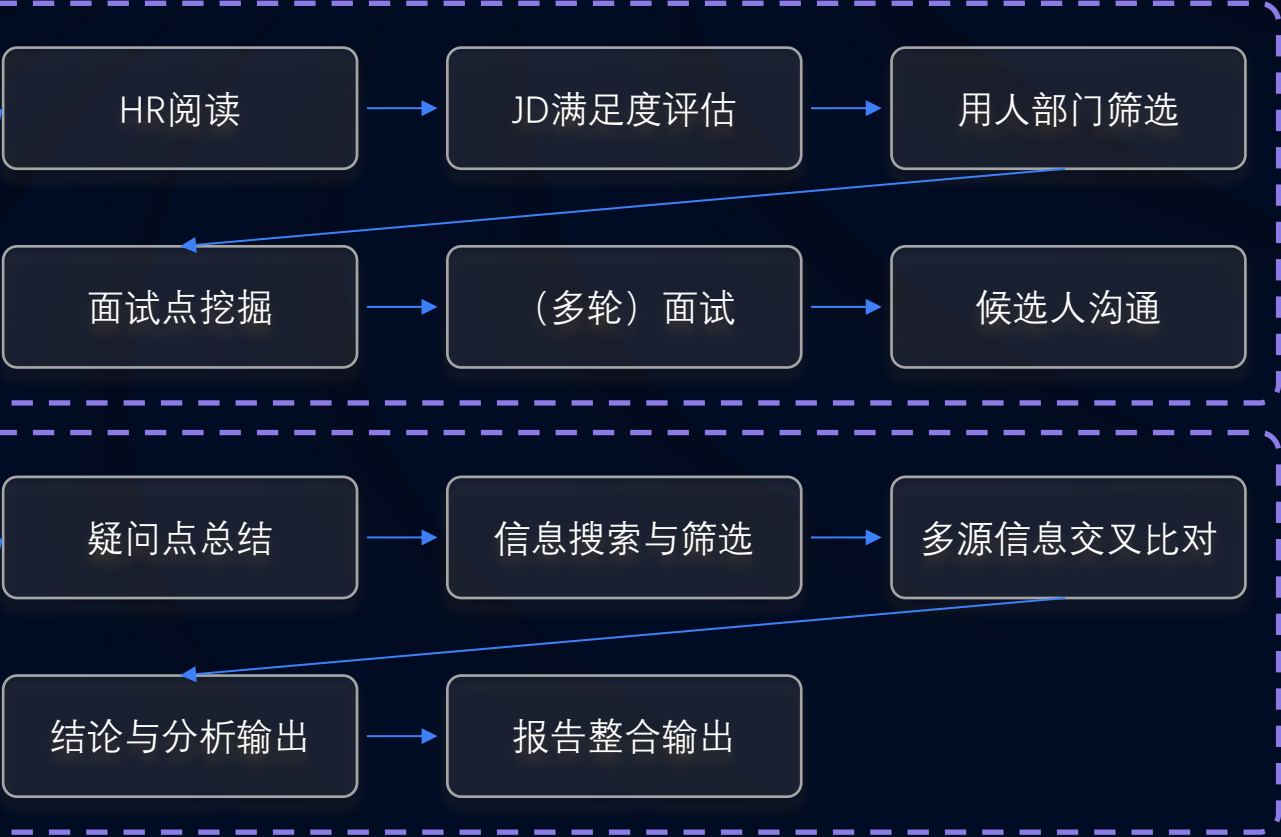
电子档简历解析 → 信息提取与入库

审计场景



财报与合同解析 → 信息抽取与比对

用户后续行动

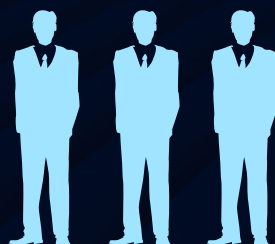


场景举例	场景介绍	典型应用
知识问答	利用LLM从海量、复杂、非结构化的文档中提取信息，回答用户问题	财报问答、产品说明书问答
内容生成	从文档中提取信息之后，快速高效地按特定格式或风格要求，生成各类内容	讲稿生成，营销方案生成
风控审核	会计审计场景下，要针对多源输入校验特定内容的一致性，传统做法纯靠人力，效率低下	财务审计
大数据清洗	利用LLM从文档中提取信息，将非结构文档结构化，提升大数据工程的效率	数据清洗、去重、提取、交叉比对验真
流程自动化	办公场景下，在业务系统中，识别屏幕或网页内容，自动进行下一步操作	RPA，Agent

新的产品体验带来了大量潜在新用户

LLM核心价值在于降低了专业门槛，让普通人也对复杂、专业、动辄上百页的内容有了需求

以法律和财务专业问答为例：

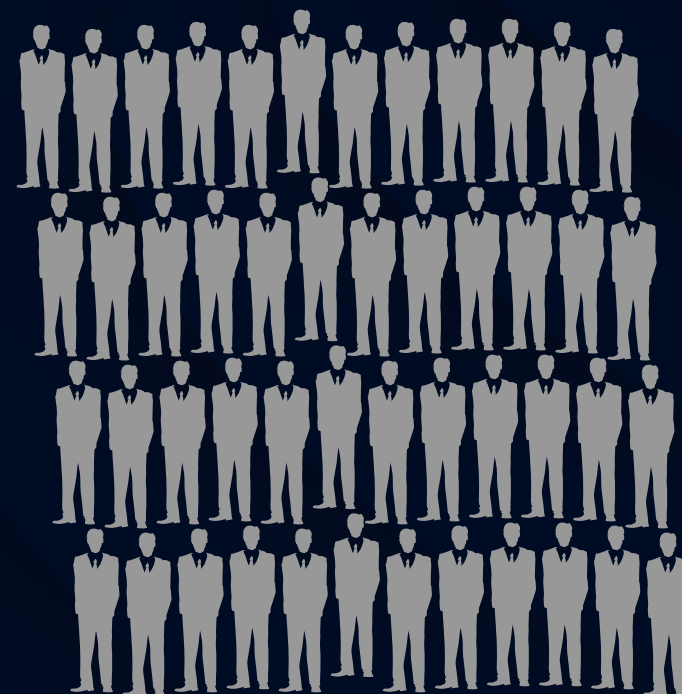


执业律师：65万



注册会计师：33.5万

约100 ~ 1000倍



劳动人口：87905万

白领：约20000万

更多的文档类型也被挖掘出来

以单页/简单文档为主

票据	合同
卡证	银行流水
试卷	清单



以复杂版式和长文档为主

论文	财报
函件	屏幕截图
合同	产品说明书
法律条文	判例说明
书籍	工作报告
研报	

Manuel Ziegler and Thomas Rednar / Energy Procedia 78 (2015) 693 – 698

695



表面活性剂
A-103

产品描述

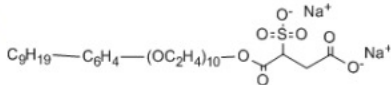
离子类型：阴离子

化学名称：壬基酚聚氧乙烯醚琥珀酸单酯磺酸钠

CAS 号码：9040-38-4

分子式：C₃₉H₆₆O₁₇ Na₂S

简述：KD A-103是用于丙烯酸、苯丙和乙烯苯烯酸类乳液的极佳的主乳化剂，尤其是光泽体系。



物化性能

有效成分 (%)	比重 25℃	CMC,%	耐钙性 CaCO ₃ @0.5%	表面张力 at CMC	溶剂	外观
24-26	1.09	0.01-0.02	2250ppm	36	水	无色透明液体

产品应用

- 聚合物领域：用于纯丙、醋丙和 EVA 乳液，尤其是高光体系，提供具有良好膜透明的小颗粒乳液。能显著提高上述乳液体系中自交联单体的用量，而不改变乳液性能。
- 非聚合物领域：具有杀菌力的清洗剂、化妆品和洗发液；发泡水泥、墙板和粘合剂；对阳离子表面活性剂和多价阳离子具有良好的承受力。对中等 HLB 值的树脂或颜料体系是良好的增溶剂和分散剂。

产品使用

用法用量：作为主乳化剂建议用量 0.5-2.0%。

为获得最佳性能，建议稀释后添加，使用量在很大程度上取决于应用体系，用户在使用前应通过实验确定最佳添加量。

为避免发生凝胶，产品应在 15℃ 以上储存。如果出现凝胶现象，应加温或加入适量水并搅拌使之全部溶解。

包装规格

25kg/桶或 220kg/桶

注意事项

本品不含易挥发组分，存储时，应防止产品低温凝胶。

说明：本说明书仅供试验参考。用户应充分验证和测试本文中任何信息或产品，以决定是否适合他们的用途。本公司不保证对特殊用途的适用性。

用人单位（甲方）：

所（盖章）

平安银行股份有限公司
财务报表附注(续)

2. 采购入库

采购入库单可以由采购订单转过来，也可以单独创建。采购入库单中的仓库列表只显示当前用户有权限的仓库。采购入库单可以使用多账户付款。

结论：LLM的出现，给文档解析带来了增量市场

LLM越强大，
下游受众会越多，文档解析的市场机会也就越大



产品定位：大模型场景下最好用的文档解析工具

INTSIG 合合信息™

核心特性

表格效果好

不漏检、不错检、内容准确

无线表、跨页表格、合并单元格、密集表格、表格中有手写字符、有公式等

综合体验佳

速度快、有层级、元素检测准

支持多种样式的文档、支持多种元素识别、识别速度快、
服务稳定、开发者上手轻松

产品功能



- 将PDF（电子档+扫描档）转换成Markdown
- 输出json，用以信息回溯与精准定位

产品形态



公有云API



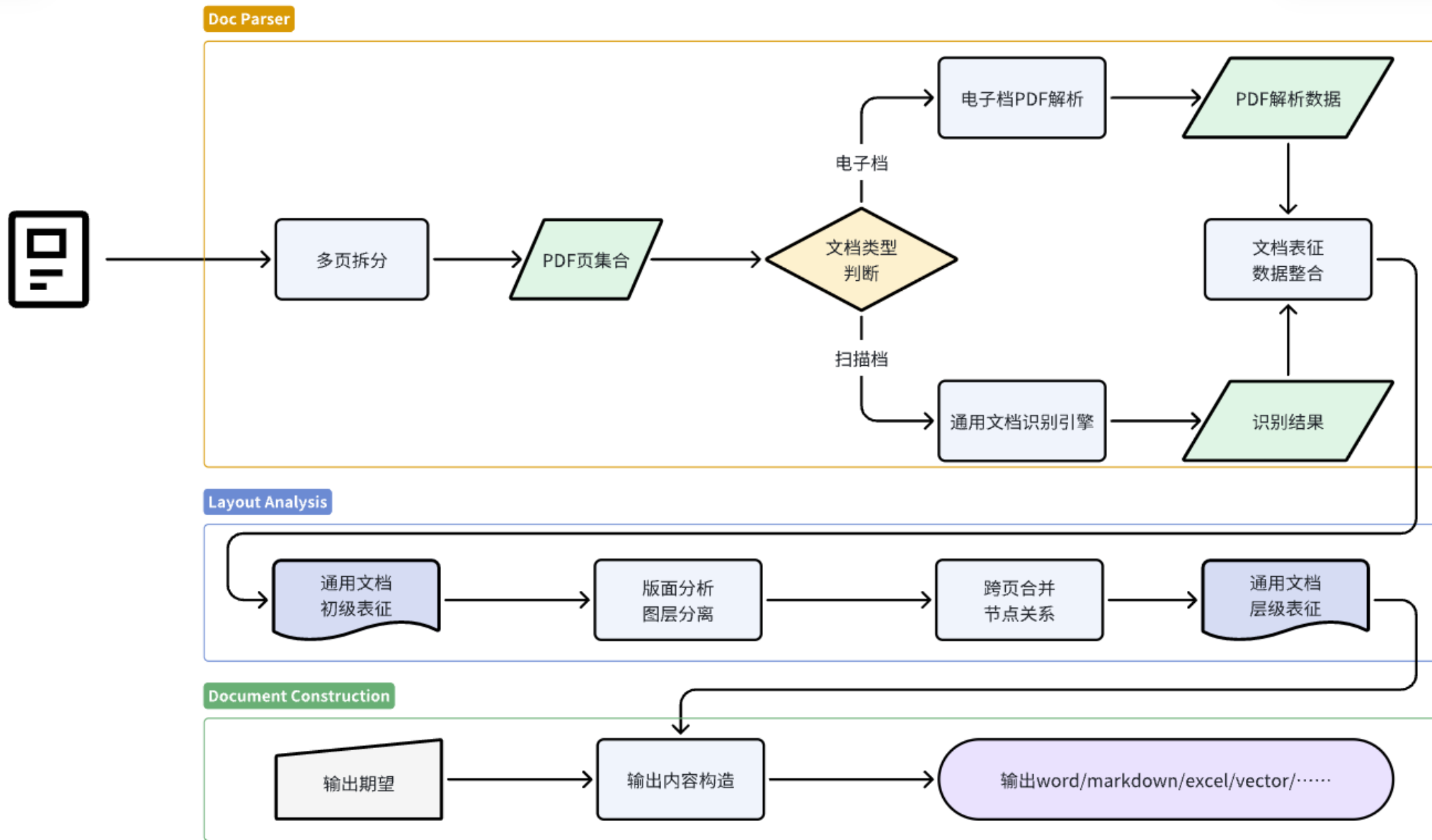
私有云标品镜像



入口网页

独特卖点

- **最全量的扫描件识别能力**，既支持单张的卡、证、票，也支持数百页的扫描版文档
- **最精准的表格还原能力**，能应对合并单元格、跨页表格
- **清晰的目录结构**，扫描版长文档也能轻松支持



模块	特性	能力边界
物理版面识别	版面	- 支持双栏、三栏等复杂版式 - 不支持倾斜文字 - 针对年报、研报场景有专门优化
	表格	- 支持有线表识别、无线表识别，且性能较好； - 能按html和md格式输出表格，html语法支持合并单元格展示 - 支持跨页表格合并 - 单元格中多行文字的处理
	公式	支持单行公式；不支持复杂化学式，不支持多行数学公式
逻辑版面识别	标题	- 标题识别同时支持电子档与扫描档，一般竞品的标题识别仅支持电子档 - 标题预测仅支持按序进行，不支持跨级预测
	段落	- 所有文字均以‘段落’对象输出，不支持注释、列表等细分格式 - 暂不支持跨页段落合并
工程性能	响应时间	单页处理时间约为0.5 ~ 1.5s；100页耗时并行处理最快2s以内
	稳定性	接口超时等待为600s，支持1000页的文档
	文档格式	支持扫描档和电子档PDF，支持图片，建议文件体积最大支持500MB

调用方法

```
1 curl --location --request POST 'https://api.textin.com/ai/service/v1/pdf_to_markdown' \
2 --header 'x-ti-app-id: c81f*****e9ff' \
3 --header 'x-ti-secret-code: 5508*****1c17' \
4 --data-binary '@/path/to/example.jpg'
5
```

参数名	数据类型	是否必填	描述
pdf_pwd	string	否	当pdf为加密文档时，需要提供密码。备注：对前端封装该接口时，需要自行对密码进行安全防护。
page_start	integer	否	当上传的是pdf时，page_start 表示从第几页开始转
page_cout	integer	否	当上传的是pdf时，page_count 表示要进行转换的pdf页数。不指定时，最大转换前24页
markdown_type	string	否	指定输出markdown 类型 (当前版本不可使用该参数) •regular(标准) •mathpix(mathpix-markdown-it)

输出示例

```
{
  "duration": 607
  "message": "Success"
  "result": {...}
  "metrics": [
    {
      "status": "Success"
      "duration": 505.61486816406
    }
  ]
  "code": 200
  "version": "3.0.36"
}
```

字段名	类型	描述
code	integer	错误码，详见“错误码说明”
message	string	错误信息
version	string	doc_restore 引擎版本号
result	object	
detail	object	每个段落对象的详情
markdown	string	markdown 正文字符串

02. 主要应用场景

2-1 for 大模型应用

2-2 for 数据清洗

2-3 for 数据录入



场景要点	说明
背景	所有需要LLM结合文档提供服务的应用，都需要先将文档中的信息转换成对LLM友好的格式
典型场景	- 文档问答 - 文档比对
客户需求	- 解析速度快，尤其对C端产品更为关键 - 信息不错、不漏 - 对容错率要求不高的场景，不要求完全正确解析排版信息，只要阅读顺序正确即可
我们的价值	作为LLM+文档类应用的前置工具，为客户提供稳定且高质量的文档输出结果，从而优化最终产品体验
客户画像	- LLM+文档类应用的开发商，文档问答为其C端或B端SaaS产品的主要功能之一 - 民营企业背景
代表目标客户	- 大模型厂商：百川、智谱、Kimi、通义千问等 - 大模型小应用：CS·AI简历项目、英飞流信息科技·RAG开源项目
服务模式	公有云API直接调用

场景要点	说明
背景	所有打算通过LLM做数据清洗、分类、筛选等工作的客户，都需要先将文档中的信息转换成对LLM友好的格式
典型场景	- 教育场景：题库清洗与打标签 - 大模型建设场景：预训练或微调语料处理（仅适合技术实力强的公司，可以用不够干净的语料进行训练）
客户需求	- 能短期处理至少百万级的数据 - 能提取正文内容，忽略非正文内容（期望我们能提取文章标题、作者、摘要等的客户，都不是当前典型客户）
我们的价值	- 将非结构化的文档正文按对LLM友好的格式还原出来，加速数据建设 - 能为多种非正文信息打上标签，供二次开发 - 支持按阅读顺序对结果进行还原
客户画像	- 民营企业背景，有C端或B端SaaS产品 - 主营业务中有文档库、知识库相关产品，提供搜索查询服务 - 数据建设最后一环，通常是人工审核和校正，因此客户通常有一定规模的人工数据清洗团队（10人以上）
代表目标客户	- C端：蜜蜂试卷·题库清洗，亿山睦赶考状元·题库清洗 - B端：百川·特定行业的大模型预训练语料处理
服务模式	公有云API直接调用

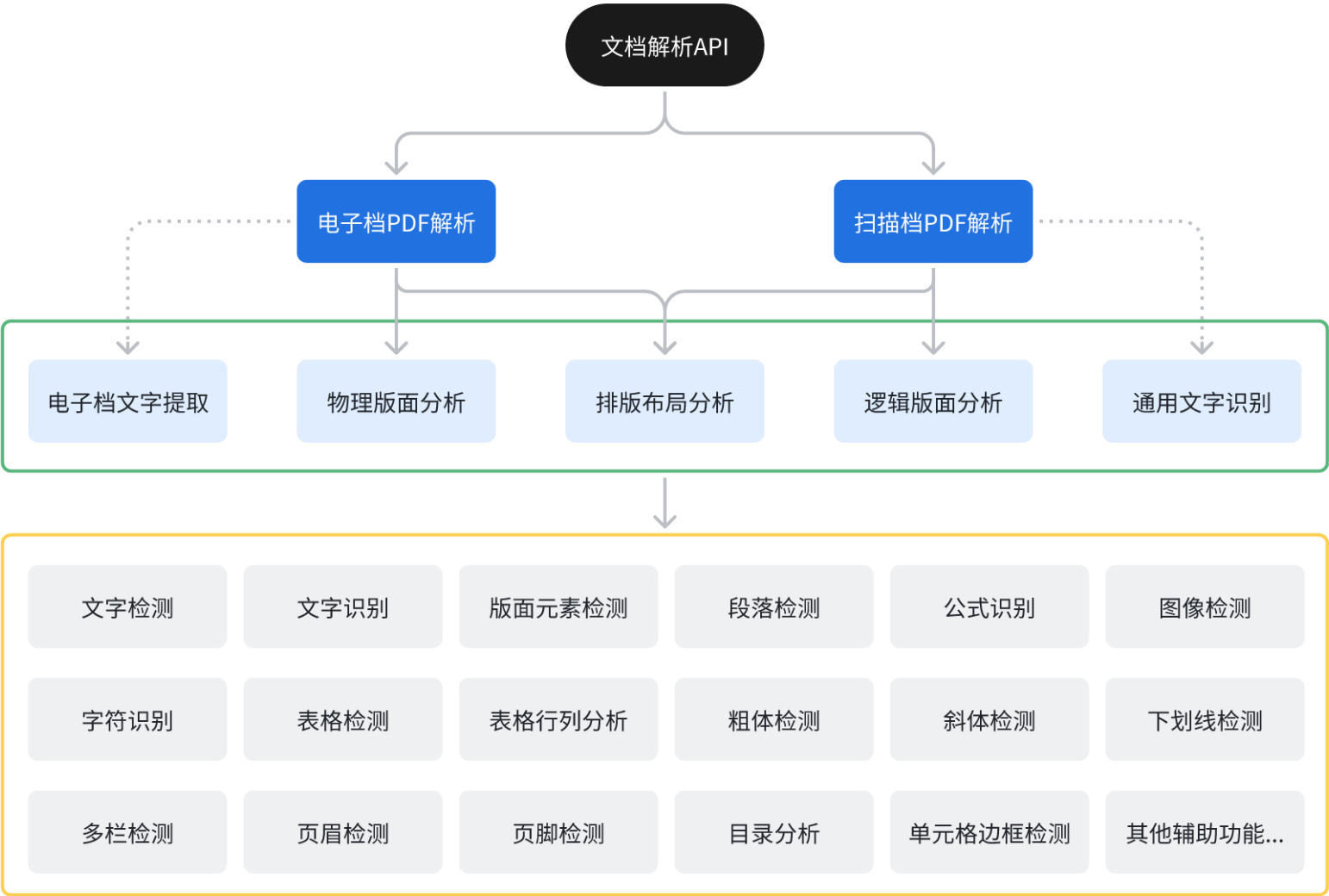
场景要点	说明
背景	- 传统数据录入需要洗数据，编写大量正则表达式去提取网页内容，面对文档类数据无从下手 - 有了LLM后，行业范式变为“数据 + 文档解析 + LLM + Prompt”的模式，写prompt比写正则相对更容易维护，也更低门槛 - 要将文档中的内容转成对LLM友好的格式，就需要文档解析工具
典型场景	各类大数据服务
客户需求	- 解析速度快（新的报告通常要在T+0天就上线） - 对各类表格结构的还原准确 - 对各类版式都要有一定兼容性，能较好的还原版式结构
我们的价值	- 为客户提供稳定且高质量的文档输出结果，提升数据录入的效率 - 能返回清晰的坐标信息，支持溯源 - 支持各类扫描类内容识别，如手写字符
客户画像	- 民营企业背景，提供大数据查询和检索服务，有C端或B端SaaS产品 - 数据录入最后一环，通常是人工审核和校正，因此客户至少有一定规模的人工数据清洗团队（10人以上）
代表目标客户	B端：DG·经信委项目·线下经济数据录入、法询科技·线上年报信息录入、启信宝·企业经营信息录入
服务模式	公有云API直接调用

03. 产品性能

- 3-1 主要难点
- 3-2 我们的效果
- 3-3 产品研发计划

版面分析关键技术 Layout-engine

版面分析算法框架



INTSIG 合合信息™

版面分析典型输出

☒ 段落 ☒ 列表 ☒ 图像 ☒ 有线表格 ☒ 节 ☒ 栏 ☒ 页眉 ☒ 页脚

证券研究报告·公司点评报告·军工电子 II

振华科技（000733）

2022 年年报及 2023 年一季度点评：业绩增长稳健，看好军用电子龙头发展前景

买入（维持）

盈利预测与估值	2022A	2023E	2024E	2025E
营业收入（百万元）	7,267	9,294	11,333	13,707
同比	28%	28%	22%	21%
归属母公司净利润（百万元）	2,382	2,994	3,693	4,542
同比	60%	26%	23%	23%
每股收益-最新股本摊薄（元/股）	4.58	5.75	7.10	8.73
P/E（现价&最新股本摊薄）	19.63	15.62	12.66	10.30

关键词：#进口替代

事件：公司发布 2022 年年度报告和 2023 年一季度报。公司 2022 年实现营收 72.67 亿元，同比上升 28.48%；归母净利润 23.82 亿元，同比上升 59.79%。2023Q1 实现营收 21.02 亿元，同比上升 11.46%；归母净利润 7.35 亿元，同比上升 20.96%。

投资要点

- **2022 年净利润大增 59.79%，盈利能力稳健提升。**受益于十四五期间军工电子元器件需求旺盛和公司稳定的高新供应链产业链，2022 年公司实现营收 72.67 亿元，同比上升 28.48%；归母净利润 23.82 亿元，同比上升 59.79%；毛利率 62.72%，同比提升 1.90pct，归母净利率 32.79%，同比提升 6.32pct。公司 Q1 业绩仍呈稳步增长趋势，实现营收 21.02 亿元，同增 11.46%；归母净利润 7.35 亿元，同增 20.96%；毛利率 63.53%，归母净利率 34.95%，同比上年同期有所提升。
- **定增 25.18 亿元已获深交所审核通过，积极布局半导体分立器件。**公司定增申请已通过审核，投资项目聚焦半导体分立器件的产能提升和纵向延伸机电板相关产业链。公司作为国内电子元器件龙头，深耕相关技术研发超 20 年，高端产品处于国际领先水平，其中二极管产品市占率高达 60%，在研发投入与专利数量上均处于行业领先地位。随着定增投资项目的投产，将进一步提升核心竞争力。
- **存货高达 22.88 亿元，电子元器件库存同比下降 41.57%。**公司为满足订单交付提前备货，2022 年存货 22.88 亿元，同比增长 23.95%。公司主营产品库存量为 52044.26 万只，同比下降 41.57%；此外，公司绩效奖金增长，薪酬同比增加 2326 万元，体现公司相关产品在市场上需求和竞争力较高，公司销售策略和业绩取得非常大的突破。
- **盈利预测与投资评级：**考虑下游装备的放量节奏，我们预计 2023-2025 年归母净利润为 29.94（+0.20）/36.93（+0.70）/45.42 亿元，对应 PE 分别为 16/13/10 倍，维持“买入”评级。
- **风险提示：**1）下游需求及订单波动；2）原材料价格上涨；3）客户集中度较高风险。

中度较高风险。

东吴证券
SOGUCHOW SECURITIES

2023 年 05 月 03 日

证券分析师 苏立赞
执业证书：S0600521110001
sulz@dwzq.com.cn

证券分析师 钱佳兴
执业证书：S0600521120002
qianjx@dwzq.com.cn

研究助理 许敏
执业证书：S0600121120027
xumu@dwzq.com.cn

股价走势

市场数据

收盘价(元)	89.87
一年最低/最高价	81.33/142.83
市净率(倍)	4.46
流通 A 股市值(百万元)	46,741.49
总市值(百万元)	46,769.53

基础数据

每股净资产(元.LF)	20.16
资产负债率(% LF)	28.25
总股本(百万股)	520.41
流通 A 股(百万股)	520.10

相关研究

- 《振华科技(000733)：2022 年中报点评：营收利润增势迅猛，军用电子龙头持续发力》
2022-09-04
- 《振华科技(000733)：军用电子元器件大本营，平台化发展前景光明》
2022-02-29

典型技术难点——版面检测

元素遮盖重叠



元素本身有多多样性



Obesity's Suppression of Metabolic Enzyme Promotes Breast

新农村建设中农民素质瓶颈与化解路径研究

王展芹,李宝林

摘要 分析了新农村建设中面临的农民科学技术知识水平偏低,专业技能差,思想道德素质低等问题,指出农民素质瓶颈成为制约新农村建设的瓶颈。在此基础上提出了相应的可行路径,即:加快城镇化发展,巩固农村基础教育,加强农村职业教育,开展提高农民思想道德素质建设等。

从第一届全国人民代表大会第三次会议《高级农业生产合作社示范章程》提出“建设社会主义新农村”奋斗目标,到1998年中共十五届三中全会通过的《中共中央关于农业和农村工作若干重大问题的决定》中明确“建设富裕文明农村的社会主义新农村”和“建设中国特色的社会主义新农村”的思想,再到党的十七大胡锦涛总书记对新农村建设的“生产发展,生活富裕,乡风文明,村容整洁,管理民主”概括,新农村建设的内涵不断深入。

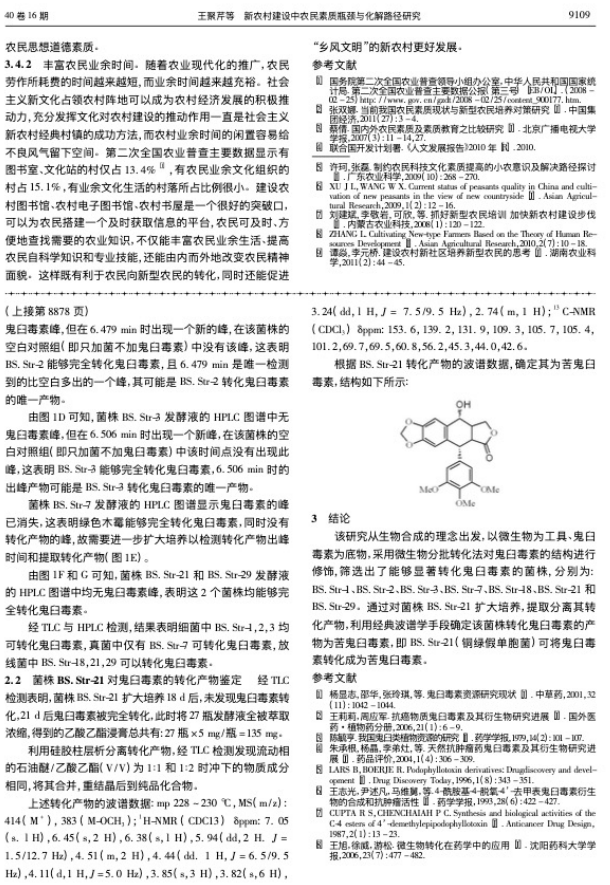
1 新农村建设中农民素质瓶颈

1.1 科学技术知识瓶颈 要使“农民”不再成为愚昧、无知,

页眉形式1

页眉形式2

复杂版式：双栏、跨页、三栏



INTSIG 合合信息™

城镇天然气管道泄漏原因分析

李 静 汪 彤 吕良涛

【摘要】 本文结合北京地区近年来发生的天然气管道泄漏事故,通过对事故原因的分析,提出关于燃气管道安全对策建议措施。

【关键词】 天然气管道; 泄漏; 安全; 分析

1 天然气管道的发展

1958年,1958年焦化厂开始供应人工煤气,70年代又增加了重油催化裂解气。随着石油工业的发展,1965年开始发展用瓶装方式供应的液化石油气。1985年,华北油田第一条进京输气管线建成,北京市管道供气又增加了天然气气源。

截至2004年年底,北京市天然气使用量已达到23.18亿立方米,燃气用户总数超过380万户。北京市天然气管网管线总长度已达到6827km。

随着天然气供应的不断扩大,城镇燃气供应过程中出现了许多安全问题。

2 天然气的性质

天然气主要成分甲烷,属于第2.1类易燃气体,熔点(℃):-182.5,沸点(℃):-161.5,燃烧热(kJ/mol):889.5,临界压力(MPa):4.59,溶解性:微溶于水,溶于醇、乙醚。燃烧(分解)产物:一氧化碳、二氧化碳。与空气混合

形成爆炸性混合物。遇热源和明火有燃烧爆炸的危险。泄漏应急处理,迅速撤离泄漏污染区人员至上风处,并进行隔离,严格限制出入,切断火源。建议应急处理人员戴自给正压式呼吸器,穿消防防护服。

3.1 近年来发生的天然气管道泄漏事故

2005年5月21日13时许,延庆县某人员驾驶农用车,行驶时,将天然气管道接口割断,导致天然气泄漏。延庆警方封锁现场,疏散过往的群众。一旦发生爆炸,后果不堪设想。经过40min,天然气公司

的抢修人员关闭了县医院的3个阀门,才停止了泄漏。

2005年4月12日下午4时20分,北京亦庄开发区天华北路西口在施工时地下1m多深的天然气管线被挖断,破口处喷出了数米高的气柱,附近小区的600户居民被紧急疏散。

近年发生了多起天然气管道泄漏事故,事故的发生及人民群众生命安全,造成严重的经济损失,对生态环境造成破坏,要避免事故发生,就要对事故进行原因分析。

3.2 发生事故的原因分析

管道燃气泄漏的原因很多,也很复杂,归纳起来主要有:设计和安装、使用、安全管理等几个方面的原因。

(1) 设计和安装存在的问题

燃气管道选址不合理,安全距离不够。

燃气管道设备质量不满足技术标准的要求,存在防护设施不全,设备老化,日常维护保养措施不利,出现管道腐蚀、易导致跑、冒、滴、漏等现象的发生。

燃气管道在安装施工过程中,存在安装技术水平不高、焊接质量差等问题。根据对燃气泄漏事故的统计,有很大一部分泄漏事故与管道的焊接质量有关。

(2) 使用过程中发生的问题

多栏的影响

多栏+插入表格的影响

《安全》2005年第4期

专项研究

城镇天然气管道泄漏原因分析

李静汪彤吕良涛

北京危险化学品应急技术中心

【摘要】 本文结合北京地区近来发生的天然气管道泄漏事故，通过对事故原因的分析，提出关于燃气管道安全对策建议措施。

【关键词】 天然气管道；泄漏；安全；分析

1 天然气管道的发展

北京的城市燃气事业起始于1958年。1958年焦化厂开始供应人工煤气，70年代又增加了重油催化裂解气。随着石油工业的发展，1965年开始发展天然气供应。1965年第一条进京输气管线——北京石油第一条进京输气管线，北京市管道供气又增加了天然气气源。

截至2004年年底，北京市天然气使用量已达到23.18亿立方米，燃气居民用户总数超过380万户。北京城市天然气管网管线总长度已达到6827km。

随着天然气供应的不断扩大，城镇燃气供应过程中出现了许多安全问题。

2 天然气的性质

天然气主要成分是甲烷，属于第2.1类易燃气体，熔点(℃)：-182.5，沸点(℃)：-161.5，燃烧热(kJ/mol)：889.5，临界压力(MPa)：4.59，溶解性：微溶于水，溶于醇、乙醚、氯仿。与空气混合能形成爆炸性混合物，遇热源和明火有燃烧爆炸的危险。泄漏应急处理：迅速撤离泄漏污染区人员至上风处，并进行隔离，严格限制出入，切断火源。建议应急处理人员戴自给正压式呼吸器，穿消防防护服。

3 天然气管道发生泄漏事故的原因分析

3.1 近来发生的天然气管道泄漏事故

2005年5月21日13时许，延庆县某人驾驶农用车时，将天然气管道接口处撞断，造成天然气泄漏。延庆警方封锁现场，疏散过往的群众。一旦发生爆炸，后果不堪设想。经过40min，天然气公司

4 城镇天然气管道发生泄漏事故的原因分析

尽可能切断泄漏源。合理通风，加速扩散。喷雾状水稀释、溶解。构筑围堤或挖坑收容产生的大量废水。如有可能，将漏出气用排风机送至空旷地方或装设水头烧掉。也可以将漏气的管道、空矿处，注意通风。漏气者要妥善处理，修复，检验后再用。

天然气管道，天然气的输送设施按照燃气设计压力分为7级，高压A(2.5MPa~4MPa)、高压B(1.6MPa~2.5MPa)、次高压A(0.8MPa~1.6MPa)、次高压B(0.4MPa~0.8MPa)和中压A(0.2MPa~0.4MPa)、中压B(0.01MPa~0.2MPa)和低压(小于0.01MPa)。

管道输送是燃气输配的主要方式，燃气虽然是一种洁净卫生、使用方便的能源，然而，它却是一种极其危险的气体。如果管理不善或使用不当，一旦泄漏，将会给人们带来灾难，使人中毒、甚至死亡。

5 城镇天然气管道发生泄漏事故的原因分析

3.2 发生事故的原因分析

管道燃气泄漏的原因很多，也很复杂，归纳起来主要有：设计和安装、使用、管理等方面的原因。

(1) 设计、施工存在的问题

燃气管道选址不合理，安全距离不够。

燃气管道设备质量不满足技术标准的要求，存在防护设施不全、设备老化，日常维护保养措施不利，出现管道腐蚀、易导致跑、冒、滴漏等现象的发生。

燃气管道在安装施工过程中，存在安装技术水平不高、焊接质量差等问题。根据对燃气泄漏事故的统计，有很大一部分泄漏事故与管道的焊接质量有关。

(2) 使用过程中发生的问题

《软科学》2011年8月·第25卷·第8期(总第140期)

科技与经济

城镇天然气管道泄漏原因分析

李静汪彤吕良涛

北京危险化学品应急技术中心

【摘要】 本文结合北京地区近来发生的天然气管道泄漏事故，通过对事故原因的分析，提出关于燃气管道安全对策建议措施。

【关键词】 天然气管道；泄漏；安全；分析

1 天然气管道的发展

北京的城市燃气事业起始于1958年。1958年焦化厂开始供应人工煤气，70年代又增加了重油催化裂解气。随着石油工业的发展，1965年开始发展天然气供应。1965年第一条进京输气管线——北京石油第一条进京输气管线，北京市管道供气又增加了天然气气源。

截至2004年年底，北京市天然气使用量已达到23.18亿立方米，燃气居民用户总数超过380万户。北京城市天然气管网管线总长度已达到6827km。

随着天然气供应的不断扩大，城镇燃气供应过程中出现了许多安全问题。

2 天然气的性质

天然气主要成分是甲烷，属于第2.1类易燃气体，熔点(℃)：-182.5，沸点(℃)：-161.5，燃烧热(kJ/mol)：889.5，临界压力(MPa)：4.59，溶解性：微溶于水，溶于醇、乙醚、氯仿。与空气混合能形成爆炸性混合物，遇热源和明火有燃烧爆炸的危险。泄漏应急处理：迅速撤离泄漏污染区人员至上风处，并进行隔离，严格限制出入，切断火源。建议应急处理人员戴自给正压式呼吸器，穿消防防护服。

3 天然气管道发生泄漏事故的原因分析

3.1 近来发生的天然气管道泄漏事故

2005年5月21日13时许，延庆县某人驾驶农用车时，将天然气管道接口处撞断，造成天然气泄漏。延庆警方封锁现场，疏散过往的群众。一旦发生爆炸，后果不堪设想。经过40min，天然气公司

4 城镇天然气管道发生泄漏事故的原因分析

尽可能切断泄漏源。合理通风，加速扩散。喷雾状水稀释、溶解。构筑围堤或挖坑收容产生的大量废水。如有可能，将漏出气用排风机送至空旷地方或装设水头烧掉。也可以将漏气的管道、空矿处，注意通风。漏气者要妥善处理，修复，检验后再用。

天然气管道，天然气的输送设施按照燃气设计压力分为7级，高压A(2.5MPa~4MPa)、高压B(1.6MPa~2.5MPa)、次高压A(0.8MPa~1.6MPa)、次高压B(0.4MPa~0.8MPa)和中压A(0.2MPa~0.4MPa)、中压B(0.01MPa~0.2MPa)和低压(小于0.01MPa)。

管道输送是燃气输配的主要方式，燃气虽然是一种洁净卫生、使用方便的能源，然而，它却是一种极其危险的气体。如果管理不善或使用不当，一旦泄漏，将会给人们带来灾难，使人中毒、甚至死亡。

5 城镇天然气管道发生泄漏事故的原因分析

3.2 发生事故的原因分析

管道燃气泄漏的原因很多，也很复杂，归纳起来主要有：设计和安装、使用、管理等方面的原因。

(1) 设计、施工存在的问题

燃气管道选址不合理，安全距离不够。

燃气管道设备质量不满足技术标准的要求，存在防护设施不全、设备老化，日常维护保养措施不利，出现管道腐蚀、易导致跑、冒、滴漏等现象的发生。

燃气管道在安装施工过程中，存在安装技术水平不高、焊接质量差等问题。根据对燃气泄漏事故的统计，有很大一部分泄漏事故与管道的焊接质量有关。

(2) 使用过程中发生的问题

表1 环境变量 Tobit 模型回归结果

环境变量	人均GDP	信息产业资产比例	人均受教育年限	制造业资产比例	常数C
软硬件投入松弛量拟合系数	0.052279	0.135194	0.271935	-0.147603	0.787023
P 值	0.799166	0.039079	0.796860	0.000000	0.193222
人员投资松弛量拟合系数	0.184183	0.145991	1.042824	-0.1248743	0.257231
P 值	0.312573	0.012274	0.279009	0.000000	0.631895

表2 调整前后信息技术投资效率

区域	省区	未剔除环境变量的效率值			剔除环境变量的效率值			区域效率平均值
		综合效率	技术效率	规模效率	综合效率	技术效率	规模效率	
东部地区	辽宁	0.386093	0.916312	0.421356	0.388709	1.000000	0.388709	CE _{it} = 0.613243 CE _{it} [*] = 0.626936 TE _{it} = 0.789749 TE _{it} [*] = 0.737325 SE _{it} [*] = 0.794542 SE _{it} [*] = 0.878472
	北京	0.622428	1.000000	0.622428	0.478657	0.517130	0.925603	
	天津	1.000000	1.000000	1.000000	0.454729	0.457057	0.994908	
	河北	0.352509	0.353958	0.995908	1.000000	1.000000	1.000000	
	山东	0.481939	0.736402	0.654451	0.626608	1.000000	0.626608	
	江苏	1.000000	1.000000	1.000000	0.659576	0.683123	0.965530	
	上海	0.515186	0.535363	0.9623	0.574660	0.578207	0.993867	
	浙江	0.514000	0.609465	0.843363	0.489895	0.505883	0.974172	
	福建	0.594048	0.786287	0.755510	0.671758	0.771461	0.870761	
	广东	0.585496	0.905090	0.615929	0.551707	0.597713	0.923030	
广西	0.693974	0.798865	0.868700	1.000000	1.000000	1.000000		
中部地区	黑龙江	0.556767	0.658163	0.845940	0.582434	0.605797	0.961434	CE _{it} = 0.544653 CE _{it} [*] = 0.646685 TE _{it} = 0.676105 TE _{it} [*] = 0.761614 SE _{it} = 0.805798 SE _{it} [*] = 0.865338
	吉林	0.551908	0.587395	0.939587	0.634020	0.651064	0.973822	
	内蒙古	0.487958	0.626734	0.778572	0.572948	0.602726	0.950595	
	安徽	0.724385	0.868697	0.833875	0.697211	0.873980	0.797743	
	河南	0.347514	0.570864	0.608751	0.557434	1.000000	0.557434	
	湖北	0.527059	0.600743	0.877345	0.637302	0.704478	0.904644	
	湖南	0.484277	0.759316	0.637781	0.634260	0.683494	0.927966	
	江西	0.557557	0.612899	0.809914	0.866252	0.938459	0.803380	
	山西	0.664450	0.800138	0.830420	0.638306	0.794525	0.803380	
	四川	0.674714	0.676996	0.996629	0.803135	0.829705	0.967976	
西部地区	重庆	0.774424	1.000000	0.774424	0.573835	0.708796	0.809591	CE _{it} = 0.606541 CE _{it} [*] = 0.705292 TE _{it} = 0.803526 TE _{it} [*] = 0.856767 SE _{it} = 0.777917 SE _{it} [*] = 0.837073
	陕西	0.456578	1.000000	0.456578	0.859143	1.000000	0.859143	
	甘肃	0.561324	0.729908	0.769033	0.623671	1.000000	0.623671	
	贵州	0.635954	0.806379	0.788654	0.781688	1.000000	0.781688	
	新疆	0.536251	0.607870	0.882181	0.590282	0.602103	0.980366	
	宁夏	0.536251	0.607870	0.882181	0.590282	0.602103	0.980366	
	青海	0.536251	0.607870	0.882181	0.590282	0.602103	0.980366	
	西藏	0.536251	0.607870	0.882181	0.590282	0.602103	0.980366	

注：CE、TE、SE 分别代表综合效率、技术效率和规模效率，带“*”表示调整后的效率值，字母 e、m 和 w 代表东部、中部和西部地区

从表1的回归结果可以看出，人均GDP和制造业资产比例对两个松弛量都有负向影响，人均GDP变量没有通过显著性检验，其影响是不显著的。根据对燃气泄漏事故的统计，有很大一部分泄漏事故与管道的焊接质量有关。

(2) 使用过程中发生的问题

无线表格识别

合并单元格识别

泰康人寿保险有限责任公司	
泰康鑫福年金保险（分红型）条款	
在本条款中，“您”指投保人，“我们”、“本公司”均指在保险单上签章的泰康人寿保险有限责任公司的分公司，“本合同”指您与我们之间订立的“泰康鑫福年金保险（分红型）保险合同”，“被保险人”指本合同的被保险人，投保人、被保险人的姓名在保险单上载明。	
1. 我们保多久、保什么	
1.1 保险期间	本合同的保险期间自本合同生效日零时开始，至被保险人年满105周岁后首个本合同的年生效对应日 ² 的24时止。
1.2 基本保险金额	本合同的基本保险金额由您在投保时与我们约定，并在保险单上载明。
1.3 养老保险金起领年龄	如果被保险人的投保年龄不超过55周岁 ³ （含55周岁），则养老保险金起领年龄为65周岁。 如果被保险人的投保年龄超过55周岁，则养老保险金起领年龄为70周岁。 （此页正文完）

表2 调整后信息投资效率							
区域	省区	未剔除环境变量的效率值			剔除环境变量的效率值		
		综合效率	技术效率	规模效率	综合效率	技术效率	区域效率平均值
东部地区	辽宁	0.386093	0.916312	0.421356	0.388709	1.000000	0.388709
	北京	0.622428	1.000000	0.622428	0.478657	0.517130	0.925603
	天津	1.000000	1.000000	1.000000	0.454729	0.457057	0.994908
	河北	0.352509	0.353958	0.995908	1.000000	1.000000	1.000000
	山东	0.481939	0.736402	0.654451	0.626608	1.000000	0.626608
	江苏	1.000000	1.000000	1.000000	0.659576	0.683123	0.965530
	上海	0.515186	0.535363	0.962312	0.574660	0.578207	0.993867
	浙江	0.514000	0.609465	0.843363	0.489895	0.505883	0.974172
	福建	0.594048	0.786287	0.755510	0.671758	0.771461	0.870761
	广东	0.585496	0.950590	0.615929	0.551707	0.597713	0.923030
中部地区	广西	0.693974	0.798865	0.868700	1.000000	1.000000	1.000000
	黑龙江	0.556767	0.658163	0.845940	0.582434	0.605797	0.961434
	吉林	0.551908	0.587395	0.939587	0.634020	0.651064	0.973822
	内蒙古	0.487958	0.626734	0.778572	0.572948	0.602726	0.950595
	安徽	0.724385	0.868697	0.833875	0.697211	0.873980	0.797743
	河南	0.347514	0.570864	0.608751	0.557434	1.000000	0.557434
	湖北	0.527059	0.600743	0.877345	0.637302	0.704478	0.904644
	湖南	0.484277	0.759316	0.637781	0.634260	0.683494	0.927966
	江西	0.557557	0.612899	0.899914	0.866252	0.938459	0.803380
	山西	0.664450	0.800138	0.830420	0.638306	0.794525	0.803380
西部地区	四川	0.674714	0.676996	0.996629	0.803135	0.829705	0.967976
	重庆	0.774424	1.000000	0.774424	0.573835	0.708796	0.809591
	陕西	0.456578	1.000000	0.456578	0.859143	1.000000	0.859143
	甘肃	0.561324	0.729908	0.769033	0.623671	1.000000	0.623671
	贵州	0.635954	0.806379	0.788654	0.781688	1.000000	0.781688
	新疆	0.536251	0.607870	0.882181	0.590282	0.602103	0.980366

单行公式与行内公式

表格内公式

1 原理

图 1 为串联谐振基本原理图。串联回路电流

$$I = U \sqrt{R^2 + [\omega L - 1/(\omega C)]^2}^{1/2} \quad (1)$$

式中， R 为电阻； ω 为角频率； L 为电感； C 为电容。

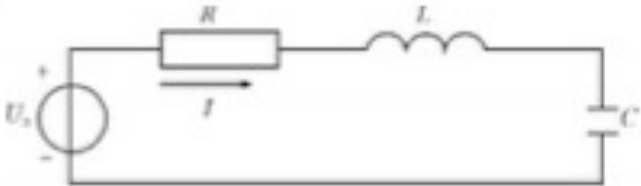


图 1 串联谐振基本原理图

Fig.1 Basic principle diagram of series resonance

对串联回路，流过各元件的电流相等。当 $\omega L = 1/(\omega C)$ 时，回路即处于串联谐振状态^[3]，此时的频率称为谐振频率：

$$f_0 = 1/(2\pi \sqrt{LC}) \quad (2)$$

表 6 变压器 2 采用变频串联谐振方法实测数据

Tab.6 Data of transformers 2 test via varies resonance with variational frequency

被试绕组/kV	C _x /pF	试验电压/kV	I _L /A	f ₀ /Hz
110	19 620	81<100	0.510<1	52.1
35	30 020	72<100	0.805<2	59.6
10	24 110	30<100	0.180<1	47.3

由串联谐振原理可知，该试验方法利用高压电抗器与电容谐振产生高电压和大电流，电源仅提供系统中有功消耗的部分，试验所需的电源功率仅有试验容量的 $1/Q$ ，所需电源容量大幅减小，试验设备的体积与重量亦大幅减小；在串联谐振状态，当试品的绝缘弱点被击穿时，电路立即脱谐，回路电流迅速下降为正常试验电流的 $1/Q$ 。

指标	指标说明	我们的水平	竞品最强水平
单页耗时	单张PDF的耗时，从调用请求到收到请求	通常300ms~1200ms	通常400 ~ 1200ms
多页耗时	整份多页PDF的耗时，从调用请求到收到完整返回	100页，P90<2s	100页，最快<2s
文件错误率	一定周期内，无法解析或解析失败的文件占总文件数的比值	约20份/万份	暂无数据
页面丢失率	一定周期内，解析失败的页面占总页面数的比值	约5页/万页	暂无数据

—— 将测试指标拆分为 元素对象、页面、数据集 3个层级

指标	模块	说明
整份文档编辑距离	总模块	整份文档当成一个字符串做编辑距离比较，意义不是特别大
整份文档BLEU	总模块	把所有内容合并在一起计算BLEU，物理意义是否有问题还不确定。整体上可能比编辑距离的意义更不大
表格召回率	元素检测模块+表格还原模块	表格匹配的个数（完全匹配或树状编辑距离小于阈值）/总表格数
公式召回率	元素检测模块+公式识别模块	检测到的公式，公式的内容都对公式个数/公式总数； 元素检测一行包含公式的文本时，会分为多个run，比如“一班比二班多1/2人”，会有“一班比二班多”，“1/2”，“人”这三个run
公式精度	元素检测模块+公式识别模块	检测到的公式，公式的内容都对公式个数/预测出的公式总数
段落文本召回率	元素检测模块+后处理模块	完全匹配的段落总数/总的段落数
段落文本精度	元素检测模块+后处理模块	完全匹配的段落总数/预测出的总段落数
标题召回率	总模块	完全匹配的标题总数/总的标题数
标题精度	总模块	完全匹配的标题总数/预测出的标题数
文档树的准确率	总模块	用树状编辑距离衡量
阅读顺序指标	总模块	平均阅读顺序指标： 将整份文档当作字符串并分成多个窗口，每个窗口尽量覆盖两个以上段落，计算每个窗口的编辑距离分数，窗口编辑距离分数超过阈值的数量/总窗口数量。

效果测试结果

——年报场景下，表格效果已优于竞品，目录识别差距明显

测试集：年报全元素测试-单页-432张

	intsig	intsig-V2	友商A	友商B
平均表格文本全对率	0.638	0.636	0.294	0.628
平均表格树状编辑距离	0.942	0.942	0.688	0.891
平均表格结构树状编辑距离	0.959	0.959	0.734	0.953
平均段落识别率	0.79	0.796	0.441	0.785
段落召回率	0.754	0.754	0.631	0.806
平均标题识别率	0.753	0.763	0.57	0.855
标题召回率	0.877	0.916	0.716	0.52
平均标题树状编辑距离	0.307	0.333	0.171	0.1
平均阅读顺序指标	0.853	0.845	0.652	0.841

注：intsig为纯OCR方案，intsig-V2为综合方案

测试集：年报全元素测试-单页-540张

	intsig	intsig-v2	友商A	友商B
平均表格文本全对率	0.603	0.603	0.16	0.587
平均表格树状编辑距离	0.918	0.918	0.535	0.874
平均表格结构树状编辑距离	0.94	0.94	0.621	0.921
平均段落识别率	0.692	0.692	0.222	0.563
段落召回率	0.775	0.775	0.672	0.716
段落f1	0.731	0.731	0.334	0.63
平均标题识别率	0.938	0.938	0.714	0.857
标题召回率	0.771	0.771	0.423	0.466
标题f1	0.846	0.846	0.531	0.604
平均标题树状编辑距离	0.482	0.484	0.288	0.193
平均阅读顺序指标	0.68	0.68	0.348	0.467

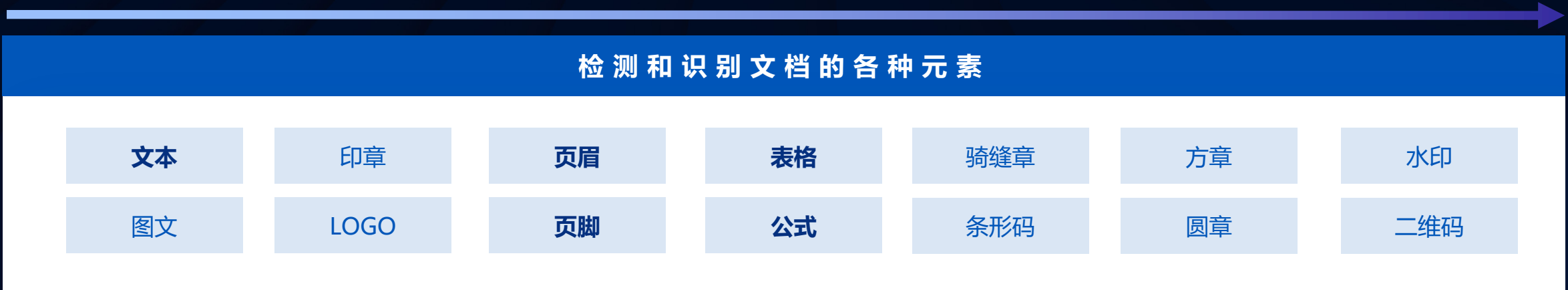
注：友商在解析部分样本时存在乱码，因此分数偏低

按场景和元素逐一迭代，依次做到分项最优

阶段一



阶段二



04. 客户案例

- 4-1 某大模型创业公司
- 4-2 某生物科技公司
- 4-3 某传媒科技公司

某大模型创业公司案例基本情况与项目节奏

客户情况	说明
所在行业	IT科技公司
主营业务	C端 · 提供通用文档问答产品 B端 · 提供基座大模型API
主要需求	其C端文档问答产品需要一款好用的前置文档解析工具，提升整体文档问答的产品体验，减少用户差评
对应产品场景	For 大模型应用
业务形态	用户上传文档后，若是PDF，则直接调用合合文档解析接口，获取解析结果后交由LLM处理
收费模式	后付费，标准定价
关键决策人	算法工程师、技术负责人

梳理测试集，准备对比测试

23.12月中

上线多线程测试版本，体验下来初步达标

24.1月中

流量切换前灰度测试

24.3月中

23.12月初

有C端用户反馈效果太差，于是开始调研好的文档解析工具

24.1月初

合合总体效果还不错，但速度太慢，不可用

24.3月初

线下交流，反馈我们的效果只高一些，但客户老板认为合合天花板更高，推动切换

24.4月初

开通后付费模式，全部线上流量切换到合合文档解析

客户情况	说明
所在行业	生物科技公司
主营业务	一站式大分子生物药CDMO服务
主要需求	生物或医药论文的数据清洗和抽取，预估调用量为180W页
对应产品场景	For 数据清洗
业务形态	客户调用文档解析处理论文，拿到解析结果后，再写规则清除无关信息，最后给到NLP模型构建知识图谱
收费模式	预付费
关键决策人	算法工程师、技术负责人

某传媒公司案例基本情况

客户情况	说明
所在行业	传媒公司
主营业务	专业视角的媒体报道和评论文章
主要需求	建立服务分析师或文章编辑们的知识库，加快文章写作效率
对应产品场景	For 大模型应用
业务形态	分析师或编辑在写作专业向内容时，向知识库应用提问知识点，或是让应用帮忙整合信息，从而加速文章写作
收费模式	预付费
关键决策人	算法工程师、技术负责人

INTSIG 合合信息™

Thanks



← 左转产品团队

右转技术交流群→

