AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

AI驱动研发变革 促进企业降本增效

北京站 08/16-17

文档解析技术加速大模型训练与应用

常扬 合合信息

演讲嘉宾

常扬

合合信息 智能创新事业部研发总监 / 复旦大学 博士

合合信息智能创新事业部研发总监，复旦博士，复旦大学机器人智能实验室成员，国家级大学生赛事评审专家，发表多篇SCI核心期刊学术论文，多个学术会议讲师与技术社区AI专家博主，负责合合智能文档处理业务线的产品、技术、云服务平台研发工作。任职期间，先后主导AI数据清洗平台、信息抽取产品、智能文档处理云服务平台、智能文档场景落地产品，为金融、制造、物流等行业提供智能文档处理产品与解决方案，在人工智能领域具备丰富的技术落地经验和行业场景洞察力。

目录
CONTENTS

PART 01

当前大模型训练与应用中的挑战

研究背景

当前大模型训练与应用过程的关键环节面临的问题

训练Token耗尽

训练语料质量要求高

LLM RAG应用中
文档解析不精准

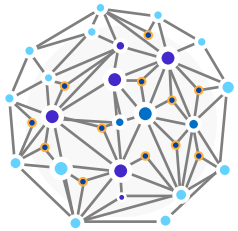
更多、更高质量的训练语料的需求 – 大模型训练

高质量预训练数据跟不上模型的进化

如何高效获取更多高质量数据?

互联网

CommonCrawl, C4, Github,
Wikipedia, StackExchange,
Huggingface数据集……



LLAMA2: **2T** Tokens
GPT4: **13T** Tokens

书籍、论文等
PDF/扫描件

? ?

Markdown

核心诉求

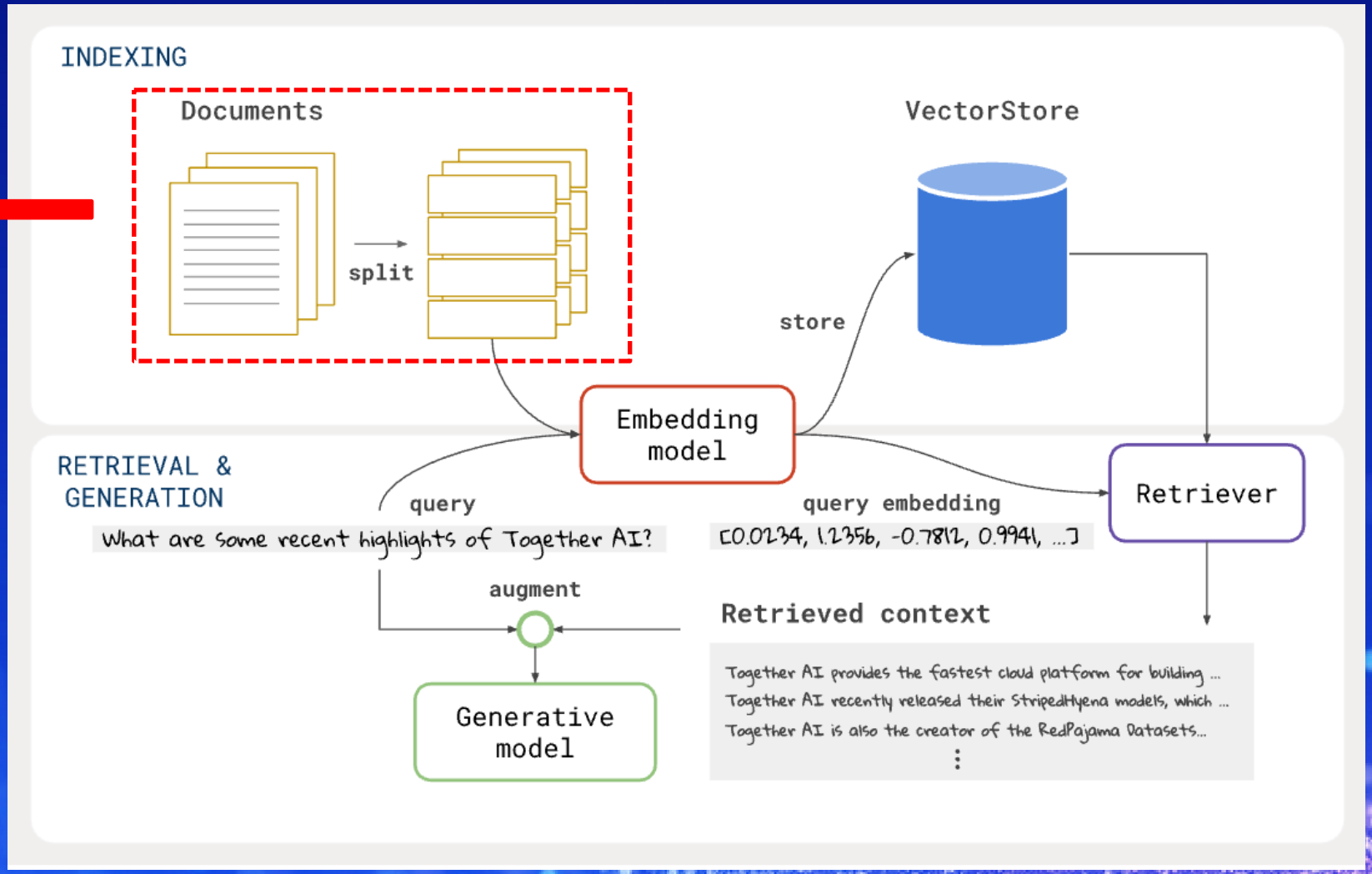
- **文档元素识别**，表格、段落、公式、标题
- **版面正确解析**，双栏、三栏、文表混合
- **转化速度**快，上百页PDF
- **阅读顺序**还原，避免混乱语序

更高精准、效率的文档解析的需求 – 大模型应用

大语言模型（LLM）驱动的检索增强生成（RAG）技术中确保能够从源文件中精准地提取内容，对于提高最终输出的质量至关重要。

在实际工作场景中，非结构化数据远比结构化数据丰富。但如果这些海量数据不能被解析，其巨大价值将无法发掘，其中 PDF 文档尤为突出。

RAG技术流程



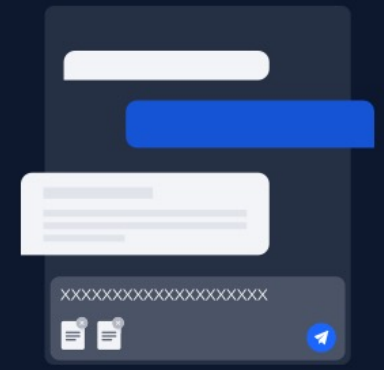
研究方向：多版式、高精度、高性能的文档解析技术

(书籍、论文、企业文档...)

各种格式、版式文档

训练

应用



文档问答、知识库问答

TextIn
文档解析技术

TextIn
文档解析技术

PART 02

文档解析技术发展与研究内容

文档介绍：计算机视角下两种类型的文档

有标记文档

Word文档

MarkDown文档

HTML文档

计算机视角下有标记的文档：

```
# 有标记文档MarkDown示例
## 第一部分
### 子标题
| 表格列1 | 表格列2 | 表格列3 |
|-------|-------|-------|
正文：有标记的文档指的是可以直接用计算机处理，结构化的文档
```

可以将文本组织成段落、单元格、表格

✅ 机器可以直接读取

无标记文档

扫描文档图像

PDF文档

计算机视角下无标记的文档：

```
%PDF-1.0
4 0 obj <</Length 65>>
stream 1. 0. 0. 1. 50. 700. cm
BT /F0 36. Tf (Hello, World!) Tj
ET endstream
endobj
```

没有储存任何结构信息，如表格或段落

❌ 机器无法直接读取

PDF文件格式

PDF文件：一系列显示打印指令的集合，非数据结构化格式。

Hello World

Hello World.pdf

Users > yang_chang > Documents > Hello World.pdf

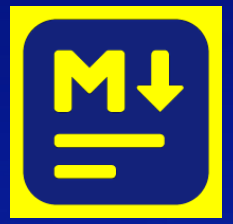
```
%PDF-1.7
%
1 0 obj
<</Names<</Dests 4 0 R >>/Outlines 5 0 R /Pages 2 0 R /
endobj
2 0 obj
<</Count 1/Kids[ 6 0 R ]/Type/Pages>>
endobj
3 0 obj
<</Author()/Comments()/Company()/CreationDate(D:20240724
endobj
4 0 obj
<</Names []>>
endobj
5 0 obj
<<>>
endobj
6 0 obj
<</Contents 14 0 R /MediaBox[ 0 0 595.276 841.89]/Parent
endobj
7 0 obj
<</Filter /FlateDecode /Length 29>>
stream
```

显示不受设备、软件或系统的影响 PDF(Portable Document Format 便携式文档格式)， 独立于应用程序、硬件和操作系统呈现文档的文件格式，能够完全保留原文档的格式。

非结构化文档、不具备可编辑性 为了极致的显示一致性，PDF 会将文本的位置、字体、间距、缩放比例、页边距等所有属性在文件格式中限定死，让软件没有自由发挥的空间。

解析 PDF 文档的挑战、让计算机可以获得PDF信息 准确提取整个页面的布局，并将所有内容（包括表格、标题、文本段落和图像）转化为结构化数据形式。

MarkDown文件格式



MarkDown文件：关注内容而非打印格式，可以表示多种文档元素。

```
# Hello World
## 二号标题
### 三号标题

**粗体**
***又斜又粗***

### 表格
|商品|数量|单价|
|---|---:|:---:|
|苹果苹果苹果|10|\$1|
|电脑|1|\$1999|

### 无线表
+ 一层
    - 二层
    - 二层
        * 三层
            + 四层

## 数学公式

支持 **LaTeX** 编辑显示支持，例如：$\sum_{i=1}^n a_i=0$， 访问 [MathJax][2] 参考更多使用方法。
```

“优雅、简约、统一”表达多种形式的数据

被互联网世界接受，充斥在各种数据中

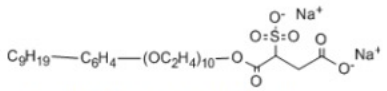
可以被大模型所理解

文档介绍：文档多版式示例

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二钠
CAS 号码：9040-38-4
分子式：$C_{39}H_{66}O_{17}Na_2S$



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，存贮时，应防止产品低温凝胶。

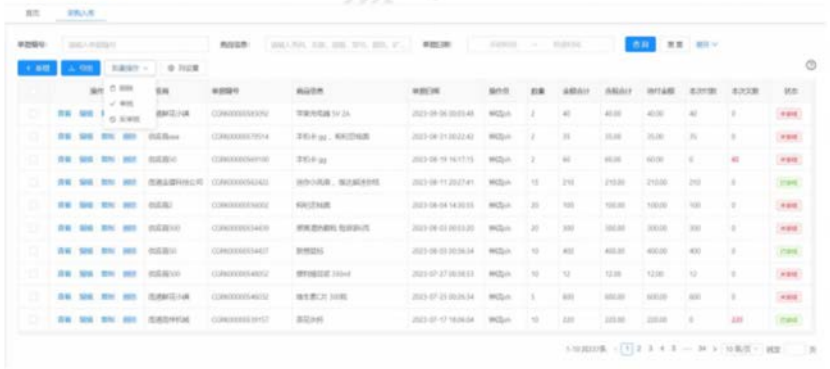
说明：本说明书仅供试验参考。用户应充分验证和测试本文中任何信息或产品，以决定是否适合他们的用途。本公司不保证对特殊用途的适用性。

用人单位（甲方）：

银行股份有限公司
财务报表附注(续)
2022年度
(除特别注明外，金额单位均为人民币百万元)

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。采购入库单据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。



项(2)	26,746	194	3,552	30,492
交易性金融负债及衍生金融负债	856	-	-	856
吸收存款	258,634	17,364	9,018	285,016
已发行债务证券	5,567	-	-	5,567
其他负债	1,845	121	24	1,990
负债合计	293,648	17,679	12,594	323,921
外币净头寸(3)	(10,165)	11,701	15,772	17,308
衍生金融工具名义金额	12,155	(12,061)	(16,184)	(16,090)
合计	1,990	(360)	(412)	1,218
资产负债表外信贷承诺	30,485	1,126	7,561	39,172

文档介绍：文档解析典型技术难点

单行公式与行内公式

1 原理

图 1 为串联谐振基本原理图。串联回路电流，

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

式中，R 为电阻；ω 为角频率；L 为电感；C 为电容。

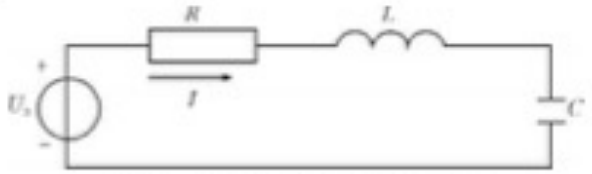


图 1 串联谐振基本原理图

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：

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

表格内公式

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 veries 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，电源仅需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 $1/Q$。

文档解析研究的过往发展

阶段	主要特点	主要方法、进展	主要应用或事件
概念阶段: 1920年代	纯光学技术	•光学模板匹配	首个OCR专利
第一阶段: 1950-70年代	字符识别方法探索与应用	•相关匹配 •统计模式识别(Chow 1957) •聚焦单字识别，假设规则版面	•商用OCR机器(IBM、日立、东芝、NEC等) •激光扫描仪出现 • IJCPR, IAPR
第二阶段: 1980-2000	简单结构文档分析与识别	•手写字符识别:特征匹配、结构匹配、统计分类、神经网络、多分类器 •词识别、字符串识别:过切分/候选切分、HMM、HMM+NN •版面分析:自上而下，自下而上	•印刷文档OCR，票据、支票识别，邮政编码和地址 •数据集:ETL, CENPARMI, CEDAR •会议:IWFHR (1990), ICDAR(1991), DAS (1994)
第三阶段: 2001-2013	复杂结构文档分析与识别	•文本行识别:英文、中文 •手写文档版面分析,联机手写文档 •自然场景/视频文本检测:边缘分析、连通成分分析	•历史文档数字化，谷歌BookProject, PhotoOCR •数据集:IAM,CASIA-HWDB等 •竞赛:ICDAR Robust Reading等
第四阶段 2014-至今	文档复杂内容识别新突破, 深度学习主导 -> 大模型主导	•手写字符和文本识别:CNN、CRNN •场景文本检测与识别:边界、分割、E2E •复杂文档版面分析:FCN、GNN •结构化图形符号:表格、公式	•自由手写文档 •网络图像、视频文本信息检索 •自由格式表格、表单

刘成林,金连文,白翔,et al.文档智能分析与识别前沿:回顾与展望[J].中国图象图形学报, 2023, 28(8):2223-2252.

文档解析研究的技术问题

图像处理

·文档预分类
-文档/非文档
·图像增强
-对比度、去噪
·图像校正
-光照/视角/变形
·二值化
·框线/装饰去除
·文档生成
·文档鉴伪

版面分析

·区域分割
·区域分类
·文本定位
·文本行分割
·手写/印刷区分
·表格分析
·逻辑版面
·签名/图标/印章

内容识别

·文本识别
-字符切分
-特征提取
-分类器设计
-序列模型
-上下文处理
-表示学习
-整页识别

·图形/符号识别
-流程、工程图
-符号、公式

·风格鉴定
-字体鉴别
-语种判别
-书写人鉴别
-签名验证

刘成林,金连文,白翔,et al.文档智能分析与识别前沿:回顾与展望[J].中国图象图形学报, 2023, 28(8):2223-2252.

文档解析技术成果应用中存在的问题?

基于规则的开源库

- pyPDF2
- PyMuPDF
- pdfminer
- pdfplumber
- papermage

基于深度学习/大模型的开源库

- Unstructured
- Layout-parser
- PP-StructureV2
- PDF-Extract-Kit
- pix2text
- MinerU
- marker
- Gptpdf

问题无法全部解决

- PDF扫描件不支持
- 无法支持全部版式
- 文档多页可用性低
- 阅读顺序无法还原
- 文档解析精度较低
- 速度慢不满足需求

TextIn 文档解析

解析更稳、识别更准、性能更快

电子档、扫描件

文档图像预处理

物理版面分析

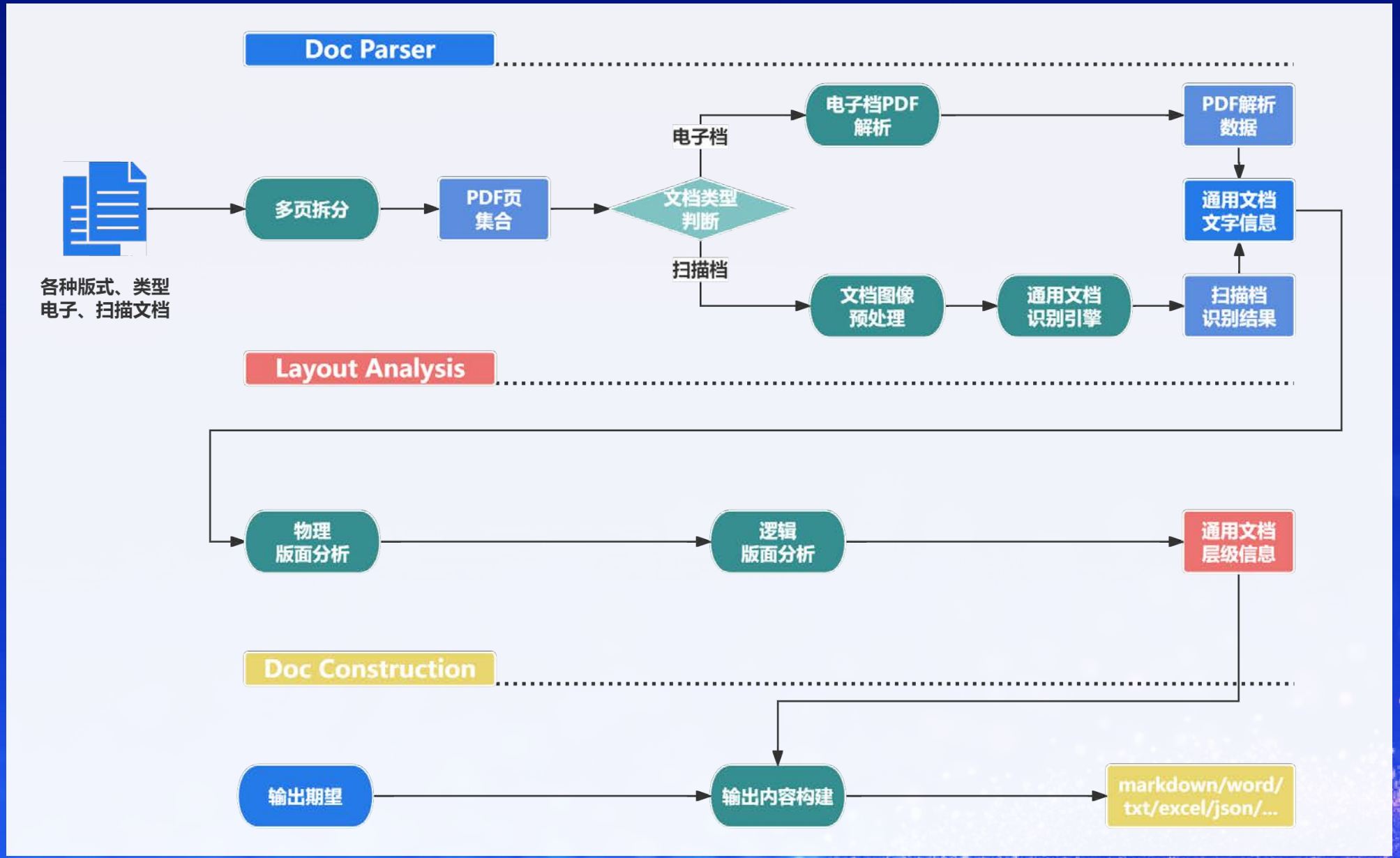
逻辑版面分析

文字识别

PART 03

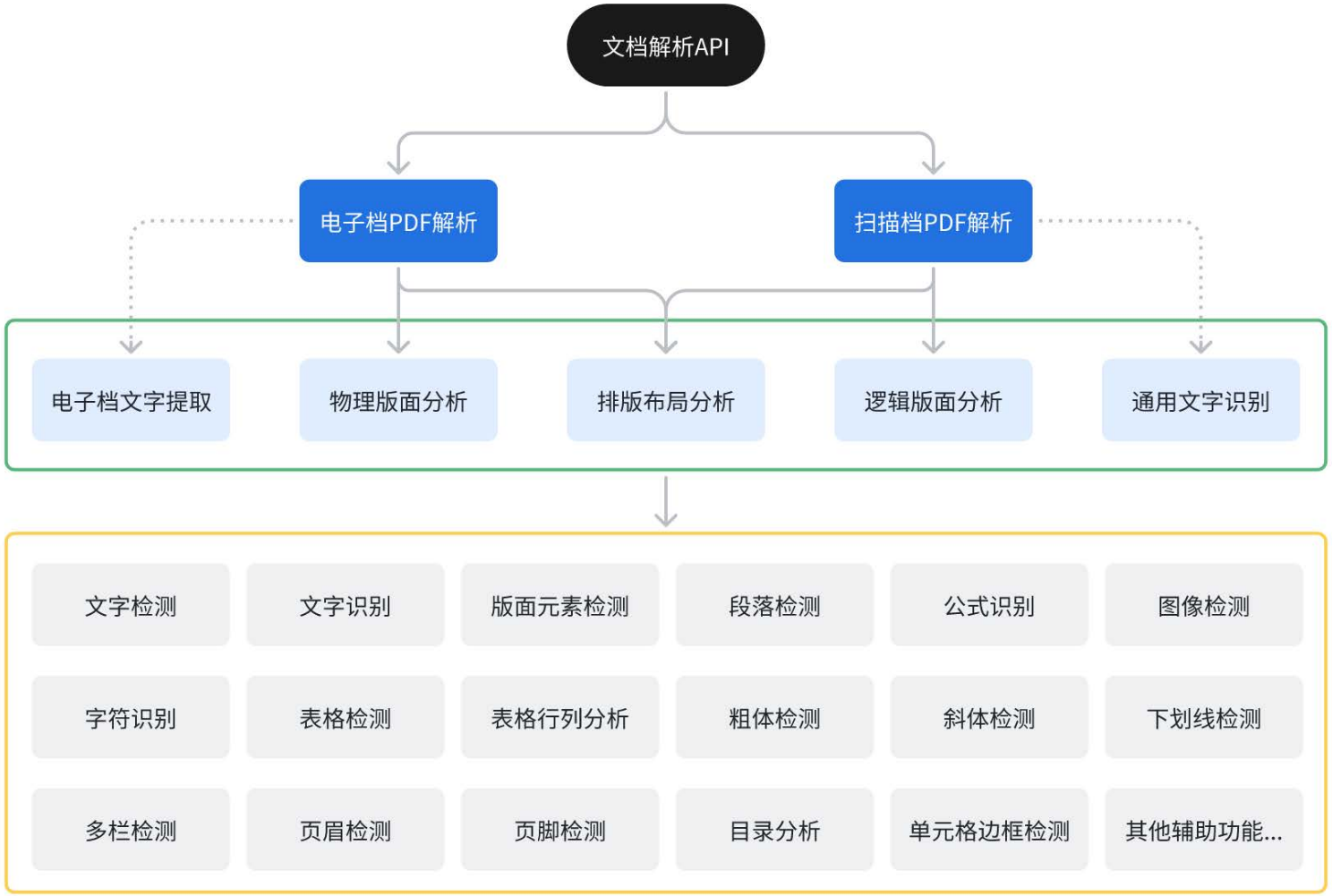
TextIn文档解析技术算法框架

TextIn文档解析 算法框架Pipeline



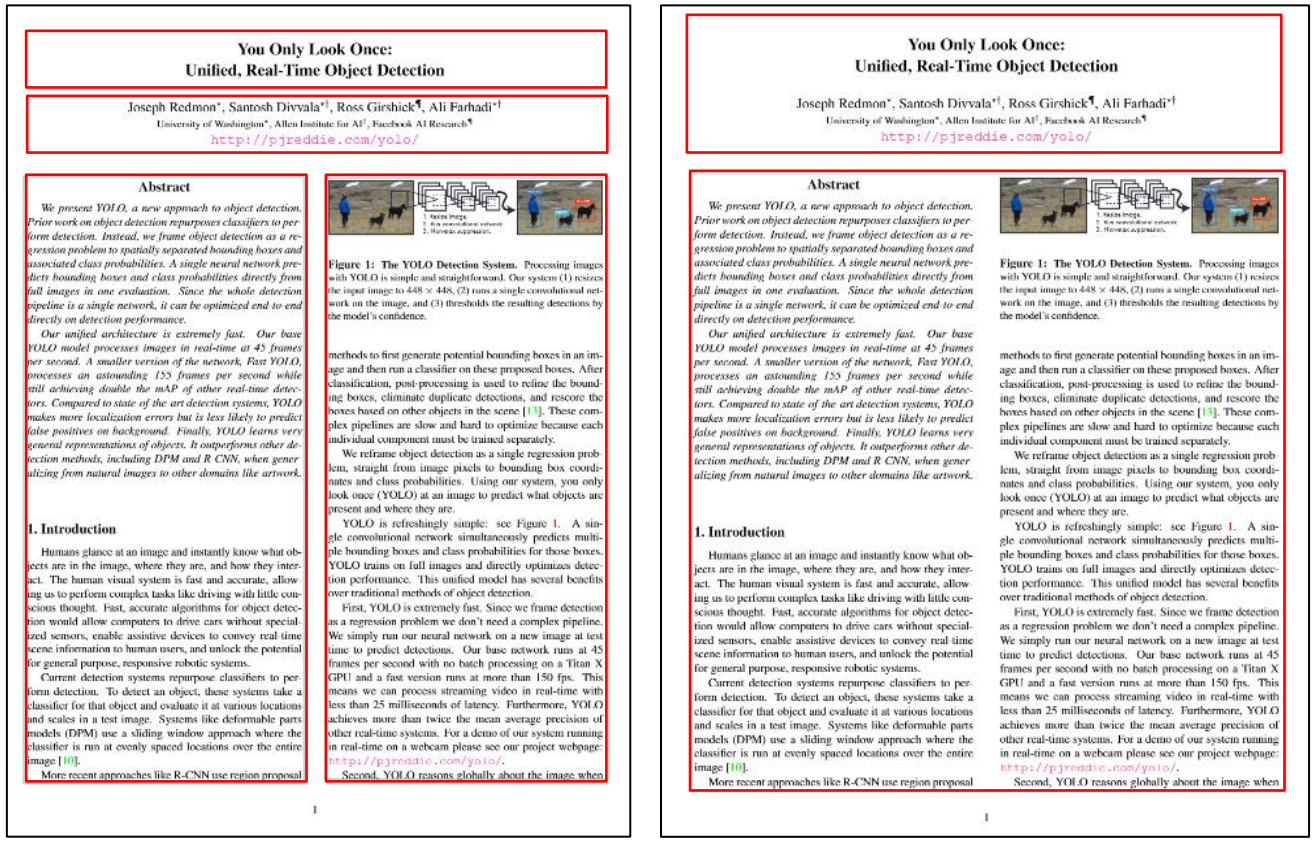
版面分析算法框架

版面分析算法框架



版面分析典型输出

版面分析算法 – 物理版面分析与逻辑版面分析



检测模型可视化：column区域（左图） vs section区域（右图）

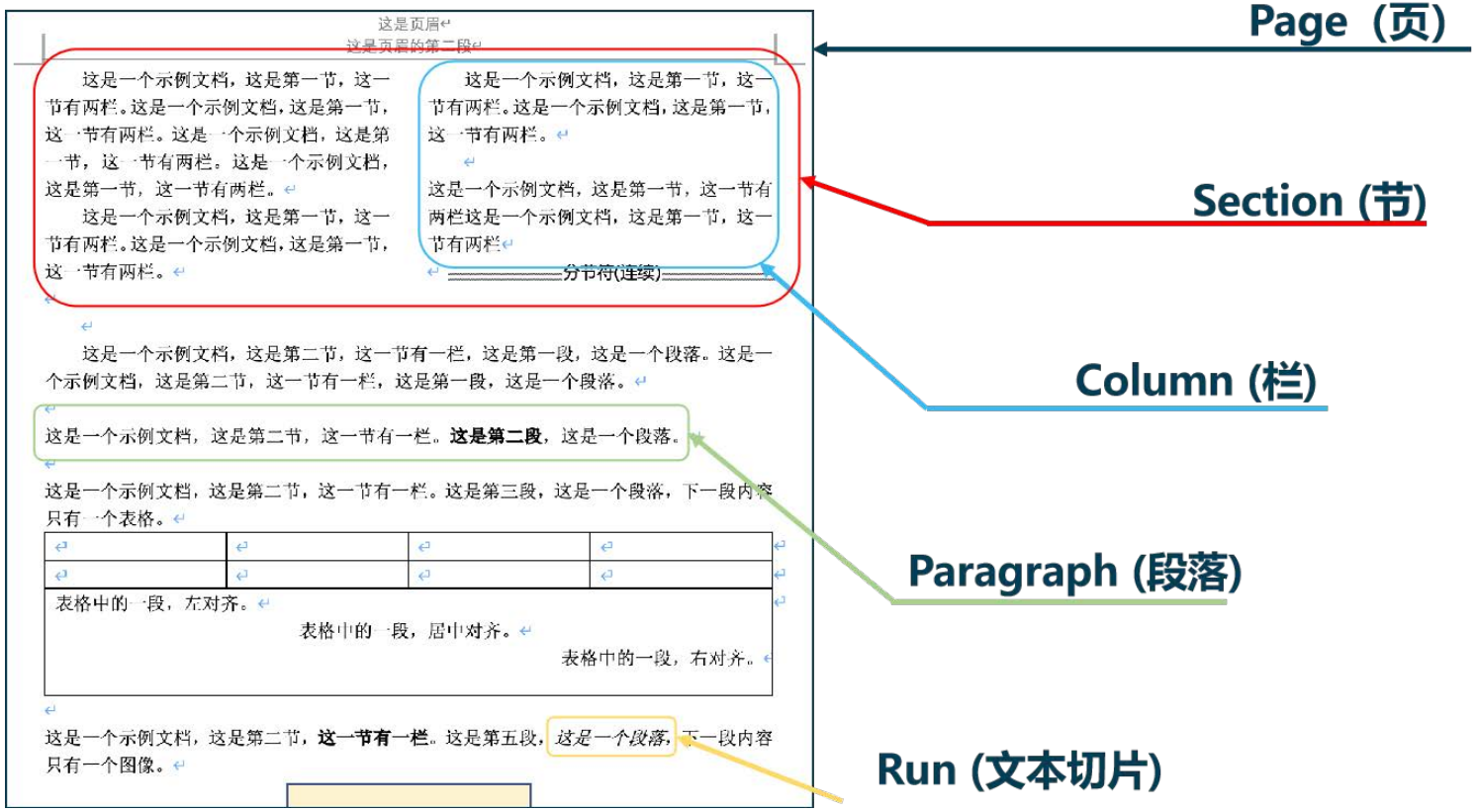
检测模型的发展：FasterRCNN/YOLO->(transformer)DETR/DINO

在产业落地时，综合考虑任务难度和推理速度，我们选用：

单阶段的检测模型，更多关注数据和模型小规模调优

- **物理版面分析-聚合**：侧重于**视觉特征**。主要任务是把相关性高的文字聚合到一个区域，比如一个段落，一个表格等等。
- **物理版面分析-布局**：选用目标检测任务进行建模，使用**基于回归的单阶段检测模型**进行拟合，从而获得文档中各种各样的布局方式。
- **逻辑版面分析**：侧重于**语义特征**。主要任务是把不同的文字块根据语义建模，比如通过语义的层次关系形成一个树状结构。

版面分析算法 – 物理版面分析



层级	概念
page层级	• 页 （page）
section层级	• 节（section） • 栏（column）
paragraph层级	• 段落（paragraph） • 列表（list） • 表格（table） • 图片（image）
run层级	• 切片（run）

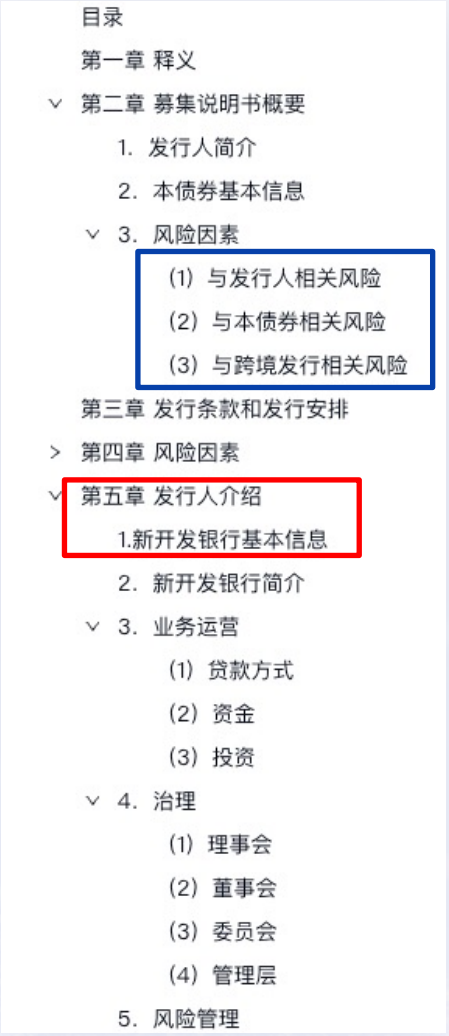
通过检测获得各个布局要素之后，我们可以建立文档的布局关系。例如，一个双栏的节（section）通常包括两个栏（column）。

版面分析算法 – 逻辑版面分析

输入文档

文档树引擎

输出目录树-树状结构

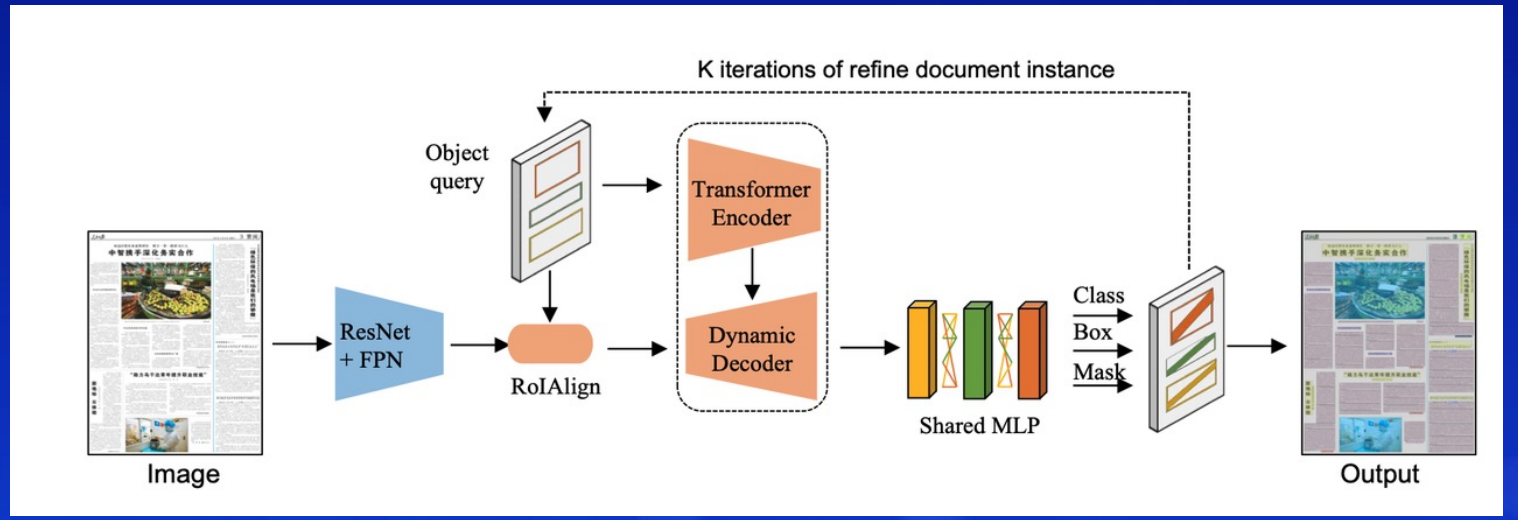


算法核心：通过Transformer架构，预测旁系类型与父子类型

预测每个段落和上一个段落的关系，分为子标题、子段落、合并、旁系、主标题、表格标题

如果是旁系类型，则再往上找父节点，并判断其层级关系，直到找到最终的父节点

最新研究方向 – 真实世界中更丰富布局类型的版面分析



Cheng H, Zhang P, Wu S, et al. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15138-15147. 合合信息华南理工联合实验室

文字识别算法逻辑

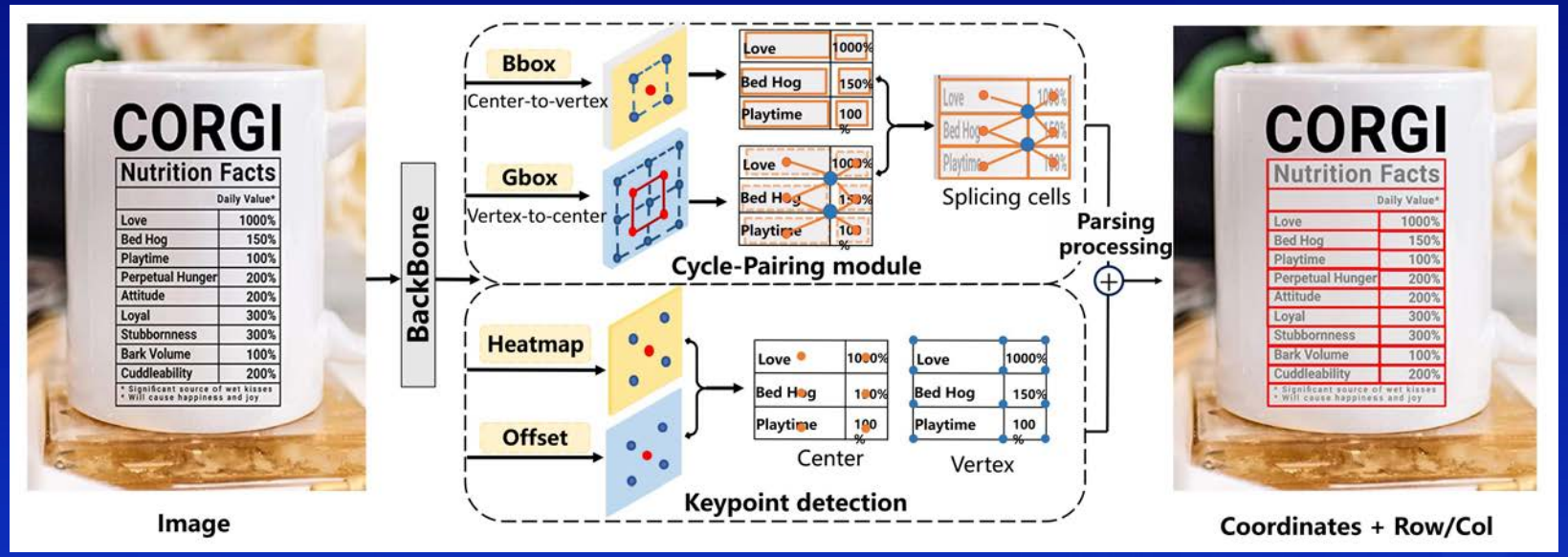
技术方案

- 多方向文字检测
- 多语言文字识别
- 其他辅助类模型

文字检测	Segmentation-based Model
文字识别	CTC-based Model

表格识别算法逻辑

- 表格类型判断
- 单元格角点模型
- 表格线段模型
- 单元格检测模型
- 后处理逻辑



[1] Parsing Table Structures in the Wild. ICCV 2021.

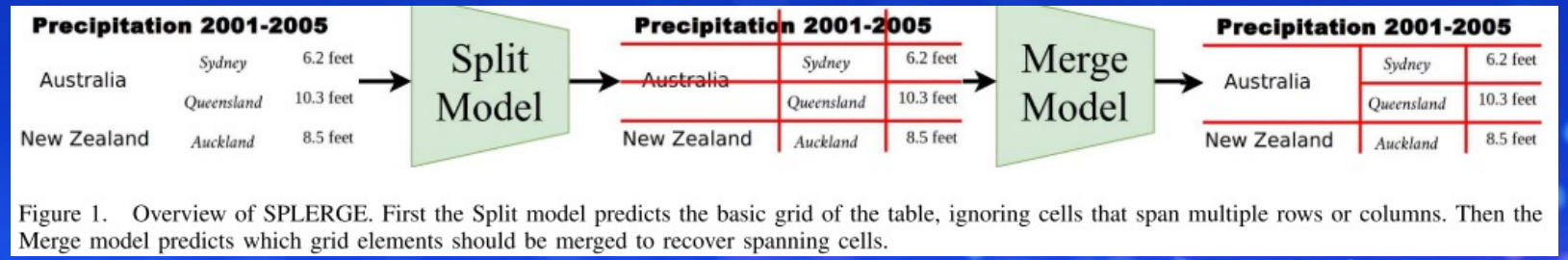
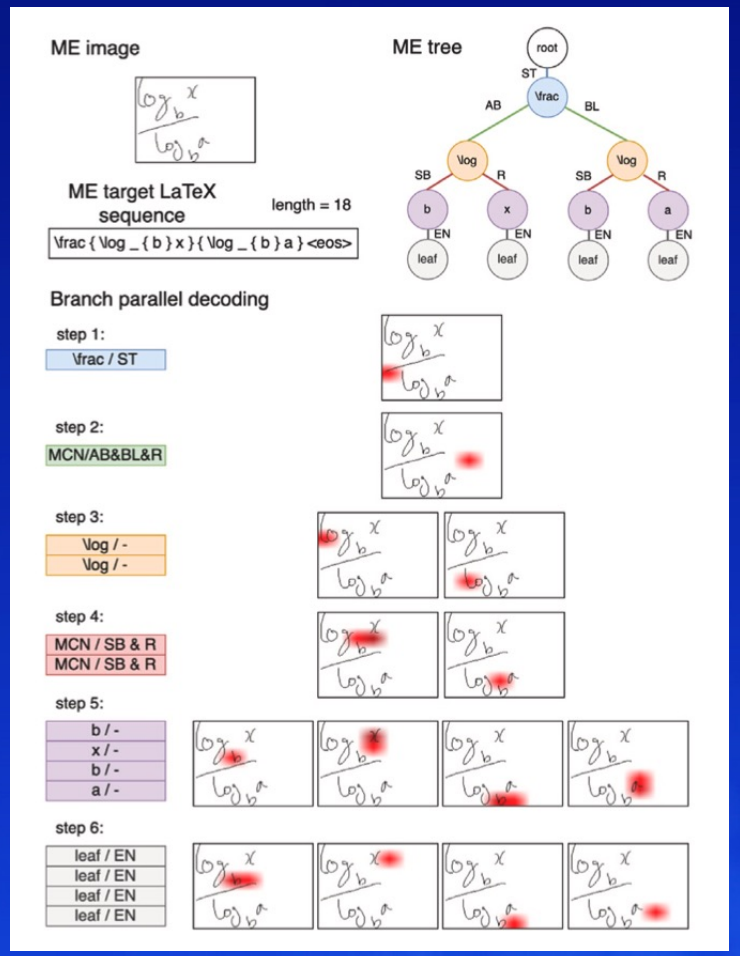
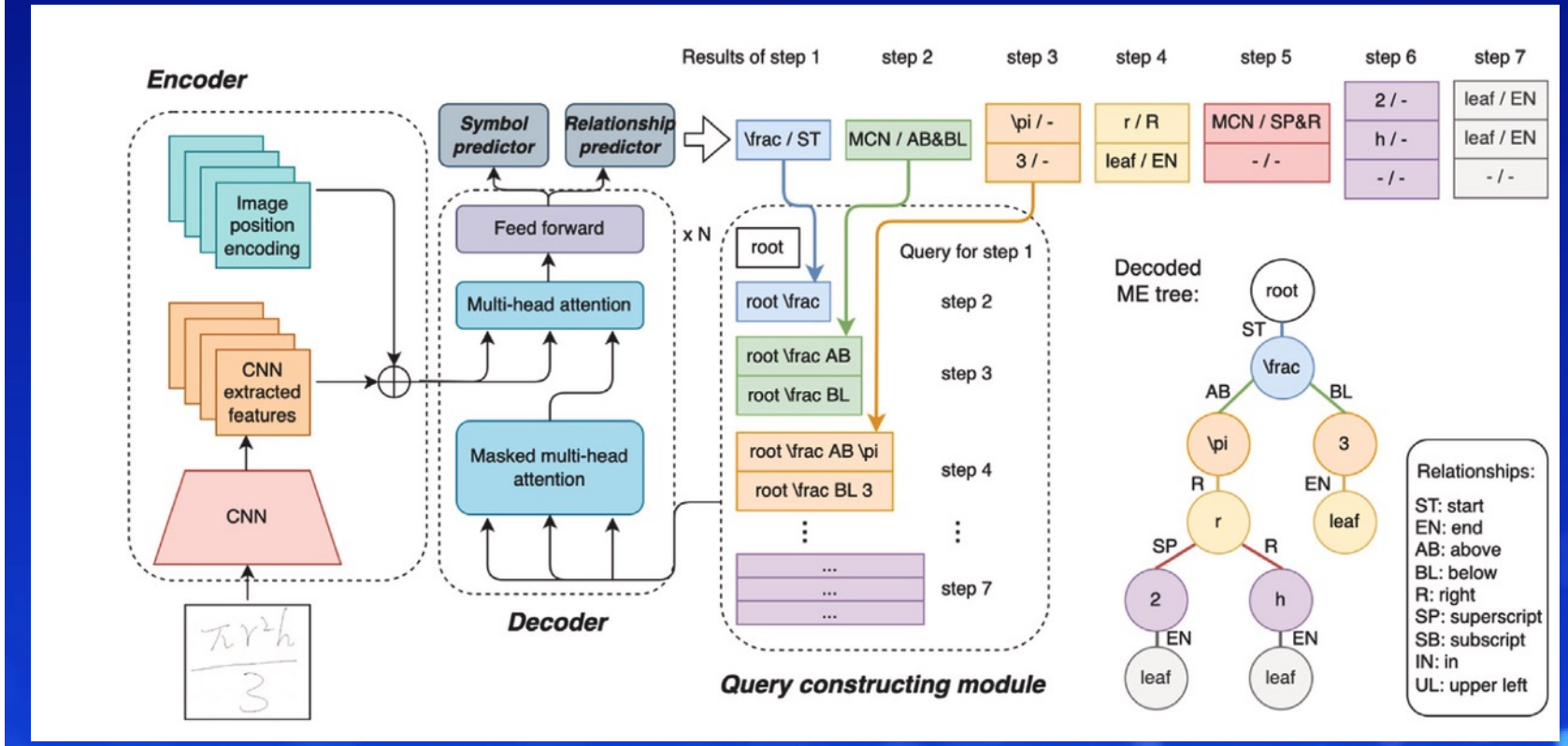


Figure 1. Overview of SPLERGE. First the Split model predicts the basic grid of the table, ignoring cells that span multiple rows or columns. Then the Merge model predicts which grid elements should be merged to recover spanning cells.

[2] Deep Splitting and Merging for Table Structure Decomposition. ICDAR 2019.

公式识别算法逻辑





Li Z, Yang W, Qi H, et al. A tree-based model with branch parallel decoding for handwritten mathematical expression recognition[J]. Pattern Recognition, 2024, 149: 110220. 合合信息华南理工联合实验室

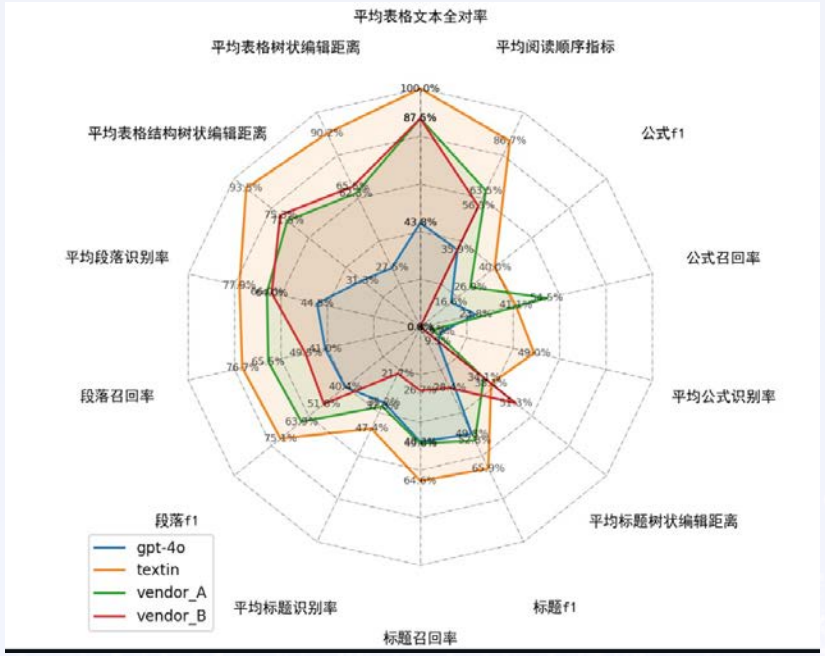
TextIn开源文档解析效果测试基准及工具

Markdown Tester

该测评脚本用于评价markdown文档相似性，从段落、标题、表格和公式四个维度进行评价：

指标	说明
段落识别率	段落匹配的个数（段落编辑距离小于0.8） / 预测出的总段落数
段落召回率	段落匹配的个数（段落编辑距离小于0.8） / 总的段落数
段落f1	2 * (段落识别率 * 段落召回率) / (段落识别率 + 段落召回率)
标题识别率	标题匹配的个数（标题编辑距离小于0.8） / 预测出的总标题数
标题召回率	标题匹配的个数（标题编辑距离小于0.8） / 总的标题数
标题f1	2 * (标题识别率 * 标题召回率) / (标题识别率 + 标题召回率)
标题树状编辑距离	所有标题树编辑距离分数之和（pred，包含文字） / 总标题数量（gt）
表格文本全对率	文本全对的表格个数（pred） / 总表格个数（gt）
表格树状编辑距离	所有表格树编辑距离分数之和（pred，包含文字） / 总表格数量（gt）
表格结构树状编辑距离	所有表格树编辑距离分数之和（pred，不包含文字） / 总表格数量（gt）
公式识别率	公式匹配的个数（公式编辑距离小于0.8） / 预测出的总公式数
公式召回率	公式匹配的个数（公式编辑距离小于0.8） / 总的公式数
公式f1	2 * (公式识别率 * 公式召回率) / (公式识别率 + 公式召回率)
阅读顺序指标	计算预测值和真值中，所有匹配段落的编辑距离

	gpt-4o	textin	vendor_A	vendor_B
平均表格文本全对率	0.4375	1	0.875	0.875
平均表格树状编辑距离	0.275310815	0.902359406	0.622942693	0.654674135
平均表格结构树状编辑距离	0.3126221	0.935267857	0.715132282	0.752590765
平均段落识别率	0.444770411	0.779088739	0.660429844	0.640144558
段落召回率	0.4097558	0.766987408	0.654844046	0.494996637
段落f1	0.404023194	0.751275893	0.638685022	0.518186347
平均标题识别率	0.352083333	0.473604827	0.369937496	0.21703854
标题召回率	0.481818182	0.646060606	0.48969697	0.266666667
标题f1	0.497619048	0.658869912	0.528264961	0.284125158
平均标题树状编辑距离	0.095151515	0.380606061	0.341375291	0.513333333
平均公式识别率	0.091836735	0.490384615	0.054791726	0
公式召回率	0.238181818	0.410909091	0.545454545	0
公式f1	0.166	0.399585921	0.268612475	0
平均阅读顺序指标	0.359378802	0.866915579	0.634898932	0.562797014



TextIn开源文档解析测试基准及工具 https://github.com/intsig/markdown_tester

TextIn 文档解析效果

PDF TO MARKDOWN

目录

费用管理办法（修订）

为了加强公司费用管理，贯彻落实党中央的“厉行节约、反对浪费”的精神，参照集团《集团总部费用开支管理办法（修订）》等规定，按照国家税务规定，结合本公司实际，特制定本办法。

一、 费用管理的原则

(一)预算管理：

根据公司全面预算管理的方针和目标，结合公司当年经营计划及目标，以“节约、务实、高效”为原则，参考上年度费用决算情况，在业务部、职能部门费用预算基础上，由财务部汇总各部门预算申报并结合项目预算，综合绩效增减、物价变动等因素，编制公司年度费用预算，经预算委员会核准后实施。

费用预算年度内原则上不得突破，确因特殊原因突破预算的，应按成本费用预算项目管理的有关规定，在征得分管领导同意后，向预算委员会提出追加预算申请并核准后方可，应严格实行费用预算管理。

（二）归口管理：公司各部门根据工作职责，对费用归口管理，对各项费用进行日常控制。

1. 业务部门对业务经营费用进行日常控制。

CH单位费用管理办法（修订）》.pdf

下载markdown　下载json　下载图片

重新上传

1/24

▶TextIn 文档解析+大模型演示 更高的文档问答精度

未使用文档解析 — 100% +

武汉理工大学硕士学位论文

参加活动中多结交朋友	人数	比例
非常希望	42	27.4%
比较希望	71	45.2%
一般	41	26.1%
不希望	2	1.2%

表 2.1 老年人对人际交往的态度

（根据《城市老年人精神需求的调查》改编绘制）

交往与归属需求：心理学表明每个人都害怕孤独和寂寞，老人们退休独自在家更是如此，他们渴望有人陪伴与人交流与外界交流，当今网络发达，一些老年人也开始玩起微信，加朋友圈，是老年人与外界交流、与人交往需求的充分体现。从实际调查（表 2.1）也可以看出大部分老年人希望在活动的参与中结交朋友，在交往中找到归属感，弥补精神上的孤独与失落。

4）自卑和抑郁感

老年人容易产生自卑情绪，他们感叹今非昔比，自己的社会地位、经济收入、身体状况不如从前，又跟不上现代生活节奏，而自信心丧失常自我否定，变得郁郁寡欢，严重者还出现心理疾病，患上老年焦虑症等。

受尊重与自我实现需求：让老年人摆脱自卑与抑郁的心理最好的办法就是让他们感受到尊重，做一些力所能及的事情发挥他们的余热，满足他们实现自身价值的需求。应多鼓励老年人参加社会活动、发展自身兴趣爱好，去老年大学中学习交流，陶冶情操，展示自己。学习知识不是老年人最终的目的，而是让老年人觉得自己在进步，有追求有特长，自身价值得以体现。

综上，老年人的心理特征体现在缺乏安全感、依恋和怀旧感、孤独和失落感、自卑和抑郁感；按美国心理学家马斯洛(A.H.Maslow)[1]对人类心理需求层级由低到高的划

图 2.1 老年人心理需求层次图（自绘）

[1] 马洛斯(1908~1970) 美国心理学家，主要从事社会心理学的研究，著作有《存在心理学导言》、《动机与人格》、《人类动机理论》等。

18

中国知网 https://www.cnki.net

AI Assistant

文档阅读完成 展开

表2.1讲了什么?

表2.1列出了参与活动的老年人们对人际交往的态度情况：非常希望、比较希望、一般以及不希望的比例分布。这些数据显示大部分老年人都希望通过参与各种活动找到归属感并弥补精神上的孤独与失落。

使用TextIn文档解析引擎 — 100% +

武汉理工大学硕士学位论文

参加活动中多结交朋友	人数	比例
非常希望	42	27.4%
比较希望	71	45.2%
一般	41	26.1%
不希望	2	1.2%

表 2.1 老年人对人际交往的态度

（根据《城市老年人精神需求的调查》改编绘制）

交往与归属需求：心理学表明每个人都害怕孤独和寂寞，老人们退休独自在家更是如此，他们渴望有人陪伴与人交流与外界交流，当今网络发达，一些老年人也开始玩起微信，加朋友圈，是老年人与外界交流、与人交往需求的充分体现。从实际调查（表 2.1）也可以看出大部分老年人希望在活动的参与中结交朋友，在交往中找到归属感，弥补精神上的孤独与失落。

4）自卑和抑郁感

老年人容易产生自卑情绪，他们感叹今非昔比，自己的社会地位、经济收入、身体状况不如从前，又跟不上现代生活节奏，而自信心丧失常自我否定，变得郁郁寡欢，严重者还出现心理疾病，患上老年焦虑症等。

受尊重与自我实现需求：让老年人摆脱自卑与抑郁的心理最好的办法就是让他们感受到尊重，做一些力所能及的事情发挥他们的余热，满足他们实现自身价值的需求。应多鼓励老年人参加社会活动、发展自身兴趣爱好，去老年大学中学习交流，陶冶情操，展示自己。学习知识不是老年人最终的目的，而是让老年人觉得自己在进步，有追求有特长，自身价值得以体现。

综上，老年人的心理特征体现在缺乏安全感、依恋和怀旧感、孤独和失落感、自卑和抑郁感；按美国心理学家马斯洛(A.H.Maslow)[1]对人类心理需求层级由低到高的划

受尊重与自我实现需求
交往与归属需求
便利与舒适需求
安全健康需求

图 2.1 老年人心理需求层次图（自绘）

[1] 马洛斯(1908~1970) 美国心理学家，主要从事社会心理学的研究，著作有《存在心理学导言》、《动机与人格》、《人类动机理论》等。

18

中国知网 https://www.cnki.net

AI Assistant

文档阅读完成 展开

表2.1讲了什么?

表2.1是一个统计表格，它展示了针对特定人群进行的人际关系倾向性研究的结果。具体来说：

- 表格标题是 "老年人对人际交往的态度"
- 数据来源于一项名为《城市老年人精神需求的调查》，但没有给出具体的来源链接。

该表格分为两栏：

第一列列出的是不同选项描述的行为或者心态，比如“参加活动中多结交朋友”。

第二列表示对应于每个行为的心态选择的比例。例如，“非常希望”这一类别的占比是27.4%。

总的来说，这张表格试图揭示受访对象——即老年人群——在面对人际关系时的具体偏好和期望程度。

TextIn 文档解析技术 工程性能

指标	指标说明	TextIn 表现	其他产品表现
单页耗时	单张PDF的耗时，从调用请求到收到请求	通常300ms~900ms	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(1.5s)	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	-
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	-

TextIn 文档解析技术 效果测试

测试集：年报全元素测试-单页-432张

	textin	Textin-v2	其他产品A	其他产品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：textin为纯OCR方案，textin-v2为综合方案

测试集：年报全元素测试-单页-540张

	textin	textin-v2	其他产品A	其他产品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：其他产品在解析部分样本时存在乱码，因此分数偏低

PART 04

基于文档解析技术的大模型应用探索

一、开放域多模态信息抽取

【信息抽取任务】

从以下资讯文本/多文档（票据）中抽取出关键信息

翔鹭钨业(9.500,0.12,1.28%)：股东众达投资854.68万股股份解除质押　来源：每日经济新闻

每经AI快讯，

公司股东潮州

海通证券(12.

押。公司表示

众达投资质押

生变更。

2019年年报显

占营收比例为

记者：曾剑"，

无境外永久居

龄32岁，硕士

传统方法

单证	字段	公司/版式	训练样本数量	测试数量
发票	买方 货品名 城市 船名 实体名称	1.1206NACHI-DG 2.1458LAYMHMI 3.1499FFSJIN 4.1620LDST--HB 5.1900CESANY 6.C LSCL 7.ICHIHIRO 8.MARUBENI SH 9.TRL CHINA 10.ZEON TRADING	每家公司20个sample	每家公司5个sample
海运提单	发货人 港口 船名 货品名 船公司	1.COMPASS 2.EVERGREEN 3.FOSHAN FOHANG 4.GAIA 5.JCT 6.MAERSK 7.ONE 8.SITC 9.TATSUMI 10.WAN HAI 11.YUSEN	1.18 2.16 3.19 4.20 5.20 6.23 7.22 8.20 9.20 10.20 11.20	每家公司5个sample
保单	保险公司 被保险人 赔付地 港口 保险代理公司	1.1628-TOKIO MARINE 2.AIG 3.C IRISS 4.INSURANCE-MITSUI 5.INSURNACE-TOKIO MARINE 6.INSURNACE-中国人民 7.INSURNACE-中国平安 8.INSURNACE-华泰保险 9.SOMPO 10.TOKIO-PANASONIC 11.日本财产MARUBENIITO	每家公司20个sample	每家公司5个sample
原产地证明	出口商 收货人 国别城市 出局机构	1.1628GDFURUP 2.1900CESANY 3.ARIAKE 4.ASEAN 5.C IRIS 6.DAIDOKOGYO 7.EBARA YT 8.TOSOH TRADE 9.WX YKK SNAP	每家公司20个sample	每家公司5个sample

总表　POC测试内容　计划安排与执行

大语言模型

建立模型

标注数百或数千份训练样本

模型训练和调优

后处理和补缺

- 需要开发人员有**丰富的算法经验**
- 新样本如语句变化则将**难以确保效果**

Prompt提示词

请从如下文本中判断出事件类型和相应的事件要素，结果按照results_style的形式进行呈现：

翔鹭钨业(9.500,0.12,1.28%)：股东众达投资854.68万股股份解除质押　来源：每日经济新闻

每经AI快讯，翔鹭钨业2，...

- **普通员工**会写提示词prompt即可
- 模型对语句变化后的**自适应性强**

开放域多模态信息抽取

三种抽取模式

- **智能解读业务文件，完成非结构化的关键信息提取，提高阅读效率，挖掘文档价值**

字段抽取

列表抽取

元素抽取

创建文档类型	设置提取字段	上传文件	AI提取返回结果	后续应用
• 直接创建文档类型 • 内置多种文档类型及预设字段，开箱即用	• 内置丰富语料信息	• 支持PDF、word、图片等主流文件格式	• 自动抽取文档中的关键信息	• 抽取结果可用于后续业务审核、一致性比对或导入其他系统

二、分析师问答产品 — 提高机构分析师信息检索效率

金融文档知识库

- 企业财报（年报、季报、半年报）
- 行业/公司研报（行业分析、个股深度分析）
- 其他公告（招股说明书、业绩预告、股权变动等）

个人知识库

- 非公开文档（专家调研报告、会议纪要等）

知识库信息检索

通过自然语言问答，精准检索知识库中相关内容；

多文档问答

支持多源信息检索及对比，洞察潜在趋势；

信息来源可靠

有效规避大模型幻觉，完整展示真实可靠来源；

关键内容总结

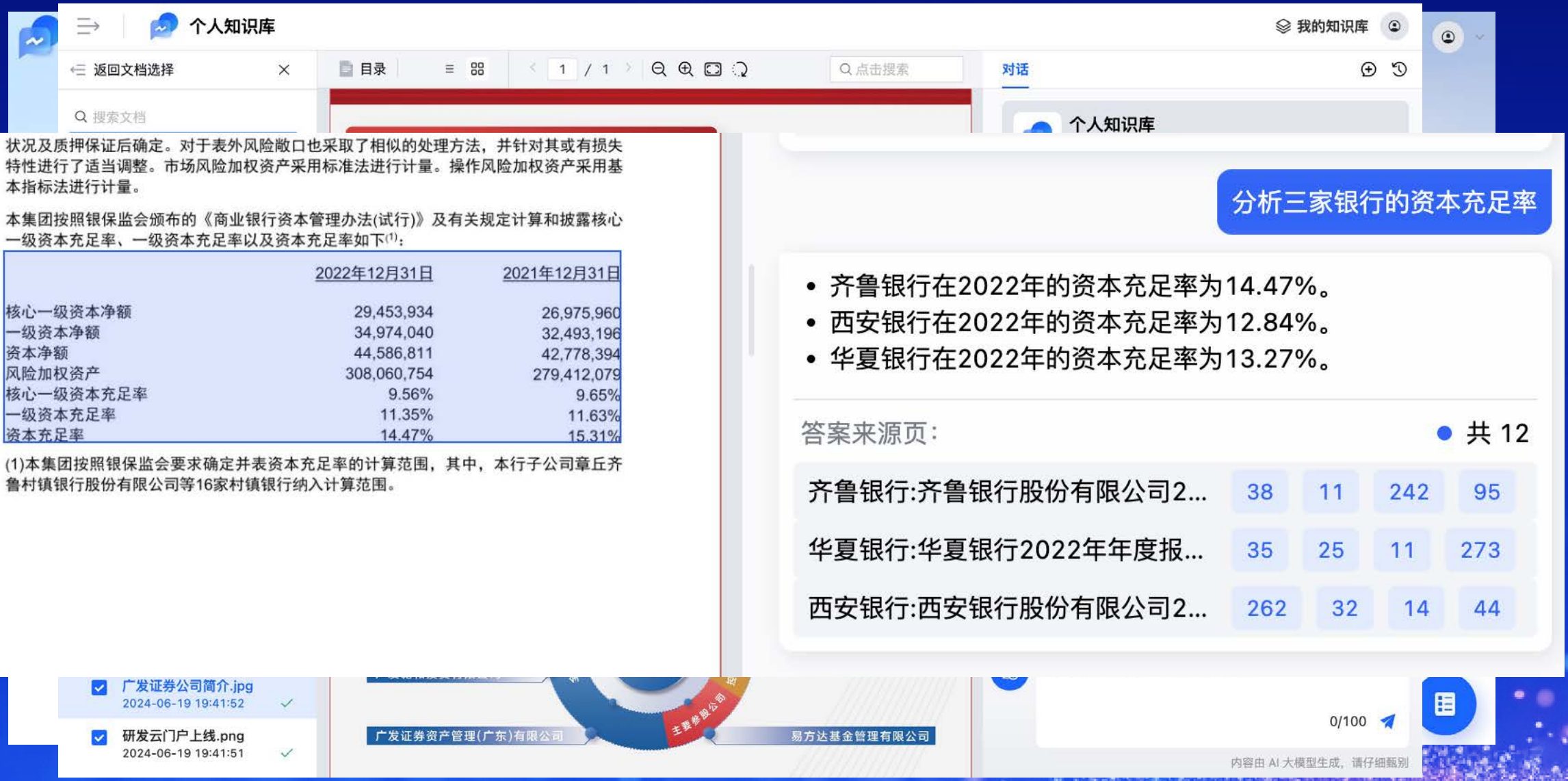
提炼文档重点内容，提高信息筛选效率；

投研知识管理

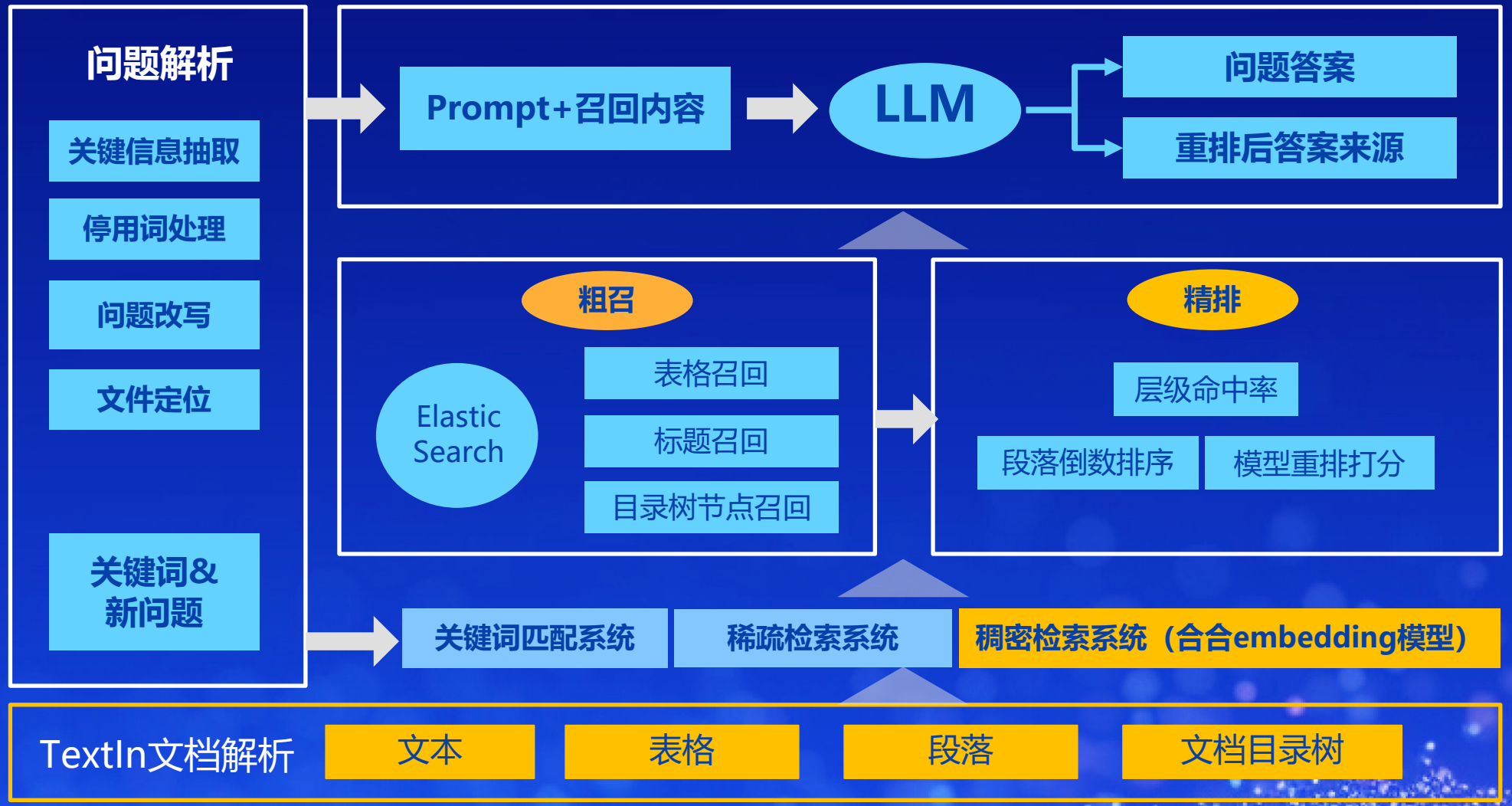
重点内容问询、标记、收藏，构建投研知识库；

- **专注有效信息阅读**
- **提高案头分析效率**
- **分析师个人投研助手**

分析师知识问答产品效果展示



分析师问答产品RAG技术架构



PART 05

总结与展望

合合TextIn智能文档处理平台矩阵 textin.com

综合产品

文档类应用

- 财报机器人
- 合同机器人
- 票据机器人
- 训练平台

体验端

- TextIn · 海外版
- TextIn · 小程序
- TextIn · Tools

基座产品

图像文档识别与处理

- 通用识别产品
- 个人证件识别产品
- 票据识别产品
- 公司证照识别产品
- 车辆识别产品
- 图像处理产品

文件转换

- 文档转换
- 图片转换

大模型加速器

- **通用文档解析**
- 文本向量模型

平台基建

- 个人账号体系
- 计费体系
- API上下架管理
- 多租户管理
- 调试模式
- 体验模式

TextIn文档解析，加速大模型训练与应用

TextIn文档解析核心特性

更多版式，更高精度

不漏检、不错检、识别准确

无线表、跨页表格、页眉、页脚、公式、图像、印章、流程图、目录树等

更高性能

速度快、服务稳定

100页PDF最快1.46s、云服务集群

更多格式

更多版式

更高精度

更高性能

THANKS