AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·北京　08.18-19

AI+
software
Development
Digital
summit

大语言模型下的文本数据治理

于政　海致星图

科技生态圈峰会 + 深度研习 ——1000 + 技术团队的共同选择

K+峰会

2023K+
全球软件研发行业创新峰会
上海站
会议时间 | 06.09-10

2023K+
全球软件研发行业创新峰会
北京站
会议时间 | 07.21-22

2024K+
全球软件研发行业创新峰会
深圳站
会议时间 | 05.17-18

K+峰会详情

AiDD峰会

会议时间 | 08.18-19
AiDD AI+软件研发数字峰会
北京站

会议时间 | 11.17-18
AiDD AI+软件研发数字峰会
深圳站

AiDD峰会详情

演讲嘉宾

于政

海致星图研发副总裁

知识平台研发中心负责人，产业大模型研发与解决方案架构师，主要研究方向为自然语言处理、知识图谱、深度学习等，并在金融、能源电力和工业制造等领域拥有丰富的产品研发、解决方案咨询设计与项目落地经验，主持、参与了多项人工智能国家专项项目，近年来在IJCAI、TKDE等顶级期刊会议上发表多篇文章，申请专利十余项，科学出版社“新一代人工智能平台建设及其关键技术”丛书《人机协同》主要作者之一。

目录
CONTENTS

1. 大模型视角下的数据治理价值
2. 大模型文本数据治理的关键问题与挑战
3. 大模型文本数据治理的核心技术
4. 海致大模型应用落地案例

PART 01

大模型视角下的数据治理价值

数据将是未来 AI 大模型竞争的关键要素

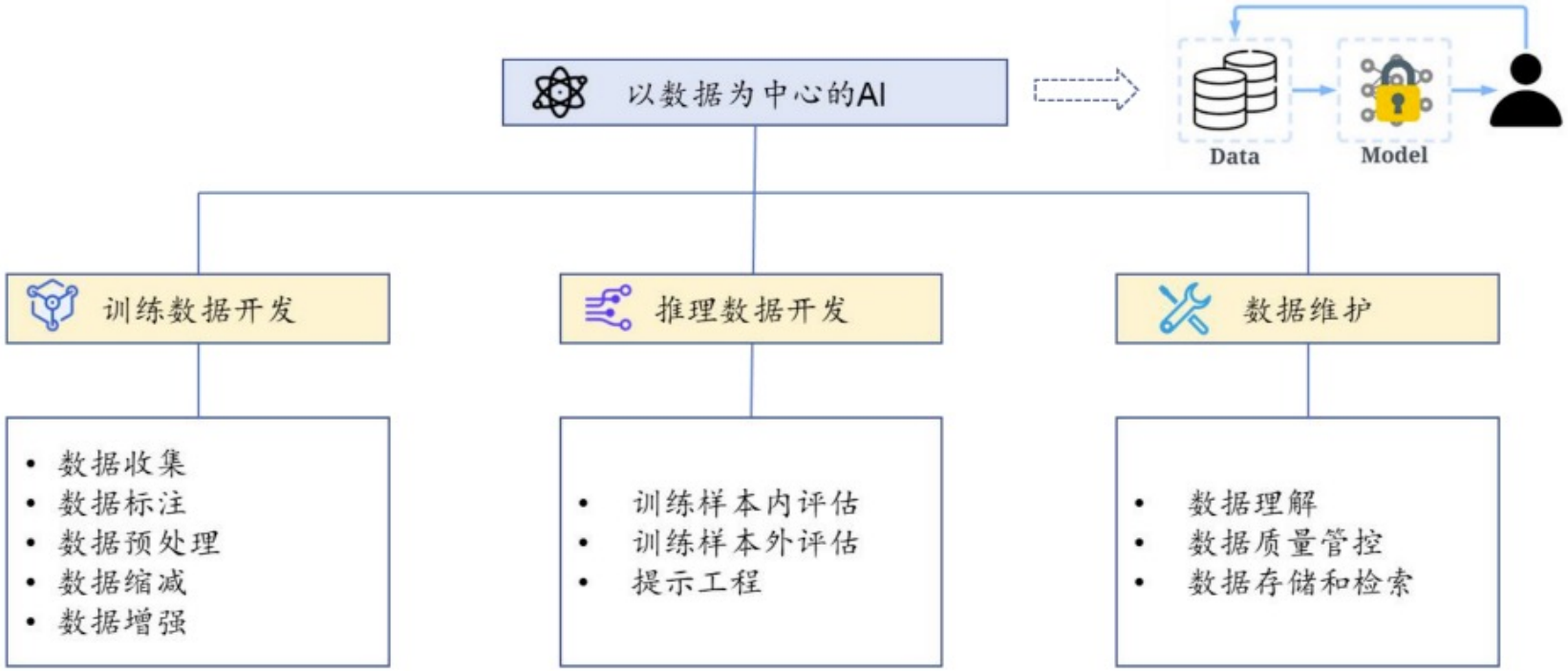
- 数据可以视为机器学习模型训练过程中的燃料，没有高质量的数据，就很难训练出高质量的模型。
- 模型训练对数据的依赖性主要体现在以下几个方面：

1. **数据质量**：模型的训练和预测表现在很大程度上取决于数据质量。如训练数据中缺乏高质量的写作数据、对话数据等，就不可能产生高质量的文章和高质量的对话。
2. **数据量**：对于深度学习模型，需要*大量的训练数据才能从中学习到复杂模式*。如果数据量不足，模型可能无法从中捕获到有用的信息，从而对未见过的新数据做出准确预测，限制模型对复杂长尾模式的学习能力。 OpenAI 在《Scaling Laws for Neural Language Models》中提出 *LLM 模型所遵循的“伸缩法则”（scaling law）*，即独立增加训练数据量、模型参数规模或者延长模型训练时间，预训练模型的效果会越来越好。
3. **数据代表性**：训练数据必须真实反映真实世界的情况，在*所有可能的输入空间中都有良好的覆盖*。否则，模型可能会在面对未见过的情况时表现不佳。
4. **数据多样性**：*训练集的数据应该具有高度多样性，模型可以学习到训练数据中的一般特征*，而不是特定于某些特例的特征。

资料来源：夕小瑶科技说《成为Llama 2唯一中国合作伙伴，TA凭什么？》

数据将是未来 AI 大模型竞争的关键要素

- 人工智能领域权威学者吴承恩发起了“**以数据为中心的 AI**”运动：
- 即**在模型相对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果**。提升数据集质量的方法主要有：添加数据标记、清洗和转换数据、数据缩减、增加数据多样性、持续监测和维护数据等。未来数据成本在大模型开发中的成本占比或将提升，主要包括数据采集，清洗，标注等成本。



资料来源：Daochen Zha et al. “Data-centric Artificial Intelligence: A Survey”2023, 华泰研究

大模型数据集分析

类别	类别	名称	数据来源	发布方
大语言模型数据集	维基百科	Identifying Machine-Paraphrased Plagiarism	维基媒体基金会	德国伍珀塔尔大学、布尔诺孟德尔大学
		Benchmark for Neural Paraphrase Detection	维基媒体基金会	德国伍珀塔尔大学
		Quoref	维基媒体基金会	艾伦人工智能研究所、华盛顿大学
		QuAC (Question Answering in Context)	-	艾伦人工智能研究所、华盛顿大学、斯坦福大学、马萨诸塞大学阿默斯特分校
		TriviaQA	维基媒体基金会	华盛顿大学、艾伦人工智能研究所
		WikiQA	维基媒体基金会	微软研究院
	书籍	BookCorpus	Smashwords	多伦多大学、麻省理工学院
		Project Gutenberg	古腾堡文学档案基金会	古腾堡文学档案基金会
	期刊	ArxivPapers	arXiv	Facebook、伦敦大学学院、DeepMind
		MedQuAD	美国国家卫生研究院	美国国家卫生研究院
		Pubmed	PubMed	马里兰大学
		PubMed Paper Reading Dataset	PubMed	伊利诺伊大学厄巴纳香槟分校、滴滴实验室、伦斯勒理工学院、北卡罗来纳大学教堂山分校、华盛顿大学
		PubMed RCT (PubMed 200k RCT)	PubMed	Adobe Research、麻省理工学院
		MedHop	PubMed	伦敦大学学院、Bloomsbury AI
		unarXive	arXiv	Karlsruhe Institute of Technology
		arXiv Summarization Dataset	arXiv	Georgetown University、Adobe Research
		SCICAP	arXiv	宾夕法尼亚州立大学
	Reddit 链接	OpenWebText	Reddit	华盛顿大学、Facebook AI Research
	Commom Crawl	C4 (Colossal Clean Crawled Corpus)	Common Crawl	Google Research
		Common Crawl	Common Crawl	法国国家信息与自动化研究所、索邦大学
	综合	The Pile	-	EleutherAI
多模态数据集		Conceptual Captions	网络	Google Research
		YFCC100M	Flickr	卡内基梅隆大学、雅虎研究院、International Computer Science Institute
		AVA	-	Google Research
		LAION-400M	Common Crawl	慕尼黑工业大学、EleutherAI、LAION
		COCO	微软	微软
		LAION-5B	Common Crawl	LAION
		Language-Table	-	-

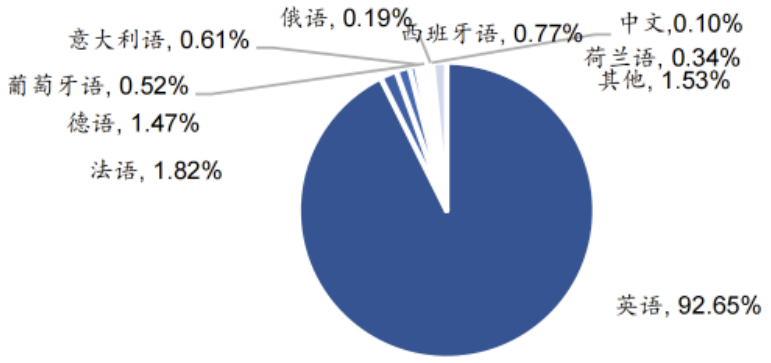
资料来源：OpenDataLab，CSDN，华泰研究

大模型数据集分析

大模型	维基百科	书籍	期刊	Reddit链接	Common Crawl	其他	合计
GPT-1		4.6					4.6
GPT-2				**40**			40
GPT-3	11.4	21	101	50	**570**		753
The Pile v1	**6**	**118**	**244**	**63**	**227**	**167**	825
Megatron-11B	**11.4**	4.6		**38**	**107**		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	**12.5**	**2100**	164.4		3450	4823	10550
LLaMA	**83**	**85**	**92**		**4162.2**	**406**	**4828.2**

注：以 GB 为单位，公开的数据以粗体表示，仅原始训练数据集大小

资料来源：Alan D. Thompson "What's in My AI" 2023，Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models" 2023，华泰研究



资料来源：中国网信网，华泰研究

- **尽管国内数据资源丰富，但由于数据挖掘不足，数据无法自由在市场上流通等现状，优质中文优质数据集仍然稀缺。**
- 如 ChatGPT 训练数据中中文资料比重不足千分之一，为 0.0991%，而英文资料占比超过 92.6%。据加利福尼亚大学和 Google 研究机构发现，机器学习和自然语言处理模型使用的数据集 50%由 12 家 Top 机构提供， 其中 10 家为美国机构，1 家为德国机构，仅 1 家机构来自中国，为香港中文大学。

典型中文开源数据集

典型中文数据集	数据集描述
DuReader 数据集	于 2018 年由百度发布。DuReader 是一个大规模的**开放域中文机器阅读理解数据集**。该数据集由 200K 问题、420K 答案和 1M 文档组成，是迄今为止最大的中文 MRC 数据集。问题和文档基于百度搜索和百度知道，答案是手动生成的。
WuDaoCorpora 数据集	于 2021 年由北京人工智能研究院、清华大学和循环智能联合发布。WuDaoCorpora 是北京智源研究院最新构建的高质量数据集，**悟道文本数据集采用20多种规则从 100TB 原始网页数据中清洗得出最终数据集**，注重隐私数据信息的去除，从源头上避免 GPT-3 存在的隐私泄露风险；**包含教育、科技等 50+个行业数据标签**，可以支持多领域预训练模型的训练。
CLUECorpus2020 数据集	于 2020 年由 CLUE 发布。CLUECorpus2020 是一个可以直接用于语言模型预训练或语言生成等自监督学习的大型语料库，它有 **100G 的原始语料库，包含 350 亿个汉字，这些语料库来自 Common crawl。**
CAIL2018 数据集	于 2018 年由清华大学、北京大学、中国科学院软件研究所和中国司法大数据研究院联合发布。CAIL2018 是第一个用于判决预测的**大规模中国法律数据集，收录了中国最高人民法院公布 260 万件刑事案件**，是现有判决预测工作中其他数据集的数倍。对判断结果的注释也更加详细和丰富。它由适用的法律条款、指控和刑期组成，根据案件的事实描述而定。
Ape210K 数据集	于 2020 年由猿辅导 AI Lab 和西北大学联合发布。**Ape210K 是一个新的大规模和模板丰富的数学单词问题数据集，包含 210K 个中国小学水平的数学问题**，是Math23K 的 9 倍。每个问题都包含黄金答案和得出答案所需的方程式，有 56K 个模板，是Math23K 的 25 倍。
WuDaoMM 数据集	于 2022 年由清华大学和北京智源人工智能研究院联合发布。WuDaoMM 是北京智源人工智能研究院 WuDaoCorpora 开源数据集的一部分。**WuDaoMM是图像和文本的多模态预训练数据，完整的数据集包含 6.5 亿对图像和文本，包含几千万对的强相关数据和 6 亿对弱相关数据，包含 19 大类**，分别是: 能源、表情、产业、医疗、景观、动物、新闻、花卉、教育、艺术、人物、科学、海洋、树木、汽车、社会、科技、体育等。
Zero 数据集	于 2022 年由北京航空航天大学、清华大学、奇虎 360 人工智能研究所联合发布。**Zero 是一种大规模的中文跨模态基准测试**，其中包含目前最大的公共预训练数据集ZERO-Corpus 和五个用于下游任务的人工注释微调数据集。**ZERO-Corpus 包含 2.5 亿张图片和 7.5 亿篇文字描述**，另外五个微调数据集中的两个也是目前中国跨模式下游任务中最大的数据集。

资料来源：华泰证券《AI大模型需要什么样的数据》

领域数据对大模型的重要性

- 未来专业及垂类内容有望成为优质中文数据集的重要来源：

  1）**专业内容平台**：

  ✓ 知乎，问答知识平台，拥有超过 4300 万创作者产生的超过 3.5 亿条优质中文问答内容，涉及政治，经济，文化，历史，科技等几乎全部门类。

  ✓ 微信公众号，内容分享平台，2022 年公众号产出超 3.9 亿篇文章，其中既有专业领域内容分析，也有时事热点分析，这些内容对语言模型的训练迭代有重要作用。

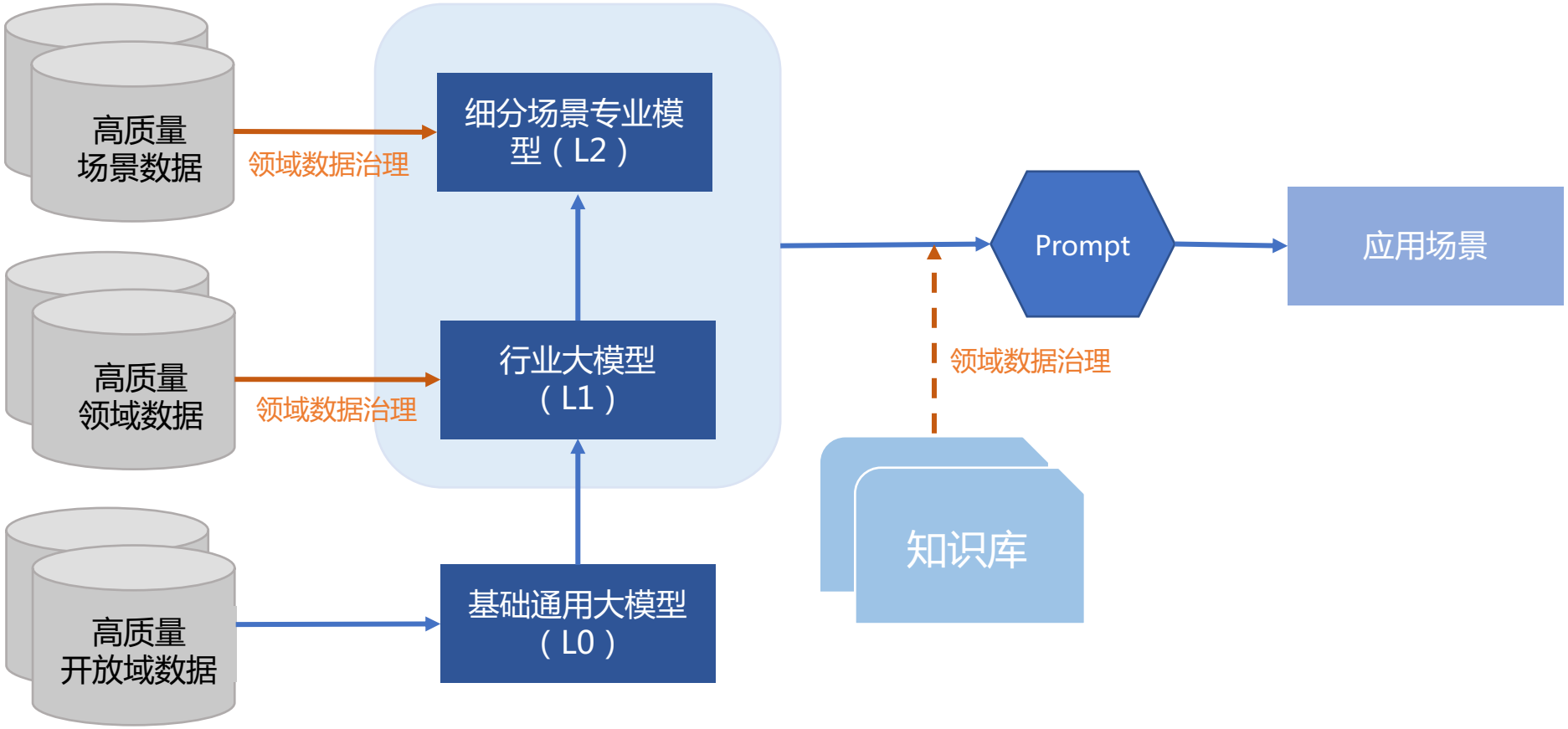
  2）**垂类内容平台**：

  ✓ 彭博BloombergGPT（基于金融垂类数据），国内在金融，医疗，生物等行业公司的数据可以作为细分领域大模型精确训练的基础。

- **领域数据具有格式多样、复杂，业务专业性强等特点，在数据治理方面面临诸多问题与挑战。**

行业大模型与领域数据治理

- 基于大模型的通用能力，经垂直领域数据的持续训练练就垂域认知能力，是行业大模型落地与应用的一条更为可行的路径。



PART 02

文本数据治理的关键问题与挑战

领域文本数据的特点

1、版面复杂多样

- 涵盖各类pdf、扫描件、图片等多种复杂文件类型；
- 需针对页眉页脚、分栏、列表、跨页等多样化信息进行版面解析；
- 需结合OCR能力。

2、内容复杂、组织多样

- 段落识别；
- 列表关系及内容识别。

3、多因素影响内容召回效果

- 文档内容相似度高（专业文档细分领域、版本迭代等）；
- 需要结合文档各级标题信息和段落内容进行匹配；
- 文档段落内容较长，影响与query核心信息匹配；
- **通用的向量相似度**算法效果不好（**问题与问题匹配 VS 问题与答案匹配**）；
- 召回率受文档库增大而降低。

4、有监督样本构造困难

- 需要较强的业务知识；
- 内网数据无法利用GPT接口构建。

典型问题和难点-法律场景

中华人民共和国行政诉讼法(2017).txt

（三）当事人所在社区、单位以及有关社会团体推荐的公民。
第三十二条 代理诉讼的律师，有权按照规定查阅、复制本案有关材料，有权向有关法院规定。
秘密、商业秘密和个人隐私的材料，应当依照法律规定保密。
当事人和其他诉讼代理人有权按照规定查阅、复制本案庭审材料，但涉及国家秘密、
第五章 证 据
第三十三条 证据包括：
（一）书证；
（二）物证；
（三）视听资料；
（四）电子数据；
（五）证人证言；
（六）当事人的陈述；
（七）鉴定意见；
（八）勘验笔录、现场笔录。
以上证据经法庭审查属实，才能作为认定案件事实的根据。
第三十四条 被告对作出的行政行为负有举证责任，应当提供作出该行政行为的证据
被告不提供或者无正当理由逾期提供证据，视为没有相应证据。但是，被诉行政行为
第三十五条 在诉讼过程中，被告及其诉讼代理人不得自行向原告、第三人和证人收
第三十六条 被告在作出行政行为时已经收集了证据，但因不可抗力等正当事由不能

中华人民共和国民事诉讼法(2017).txt

第六十一条 代理诉讼的律师和其他诉讼代理人有权调查收集证据，可以查阅本案有关材料。查阅本案有关材料的范围和办
第六十二条 离婚案件有诉讼代理人的，本人除不能表达意思的以外，仍应出庭；确因特殊情况无法出庭的，必须向人民法
第六章 证 据
第六十三条 证据包括：
（一）当事人的陈述；
（二）书证；
（三）物证；
（四）视听资料；
（五）电子数据；
（六）证人证言；
（七）鉴定意见；
（八）勘验笔录。
证据必须查证属实，才能作为认定事实的根据。
第六十四条 当事人对自己提出的主张，有责任提供证据。
当事人及其诉讼代理人因客观原因不能自行收集的证据，或者人民法院认为审理案件需要的证据，人民法院应当调查收集。
人民法院应当按照法定程序，全面地、客观地审查核实证据。
第六十五条 当事人对自己提出的主张应当及时提供证据。
人民法院根据当事人的主张和案件审理情况，确定当事人应当提供的证据及其期限。当事人在该期限内提供证据确有困难

query：民事诉讼的证据有哪些？

难点：关键信息在文件名上

典型问题和难点-规章制度场景1

- 附件3.北京市服务“专精特新”中小企业的公共服务示范平台服务方案编制要点.docx
- 附件3.财政部、工业和信息化部联合印发《关...“专精特新”中小企业高质量发展的通知》.pdf
- 附件3.财政部工业和信息化部联合印发《关于...“专精特新”中小企业高质量发展的通知》.pdf
- 附件3/第三批专精特新“小巨人”企业申请书.doc
- 附件4-第四批专精特新“小巨人”企业申请书.wps
- 附件4-第五批专精特新“小巨人”企业申请书.wps
- 附件4.北京市国家级专精特新“小巨人”企业名单.docx
- 附件4.北京市推荐国家第三批专精特新“小巨人”企业名单.doc
- 附件4.北京市专精特新中小企业认定标准.doc
- 附件4.申报示范平台服务“专精特新”中小企业绩效目标表.xlsx
- 附件4.XXXX服务“专精特新”中小企业绩效目标表.xlsx
- 附件5-第二批专精特新“小巨人”企业名单.docx

query：申请专精特新需要哪些申请材料

难点：文件版本管理

典型问题和难点-规章制度场景2

（一）作品实物照片及采用特有、珍贵、稀有原材料制作的证明；
（二）作品创作思想和艺术价值的说明；
（三）参加国内外有影响的展览会取得的荣誉证书；
（四）评审委员会认为应当提交的其他材料。

第九条　申请北京工艺美术大师认定的，申请人应当向评审委员会提交申请报告和以下材料：
（一）从事工艺美术专业的经历及相关证明；
（二）代表作品照片及说明；
（三）代表作品在国内外评比中的获奖证书；
（四）专业技艺论述或者论文。
申请北京民间工艺大师认定的，申请人应当向评审委员会提交申请报告及前款规定的（一）、（二）、（三）项材料。

第十条　评审委员会应当将申请人提交评审的传统工艺美术品种、技艺和工艺美术珍品目录及拟评工艺美术大师、民间工艺大师名单预先公告，公告期为30日。公告期内，任何单位和个人有权提出异议。

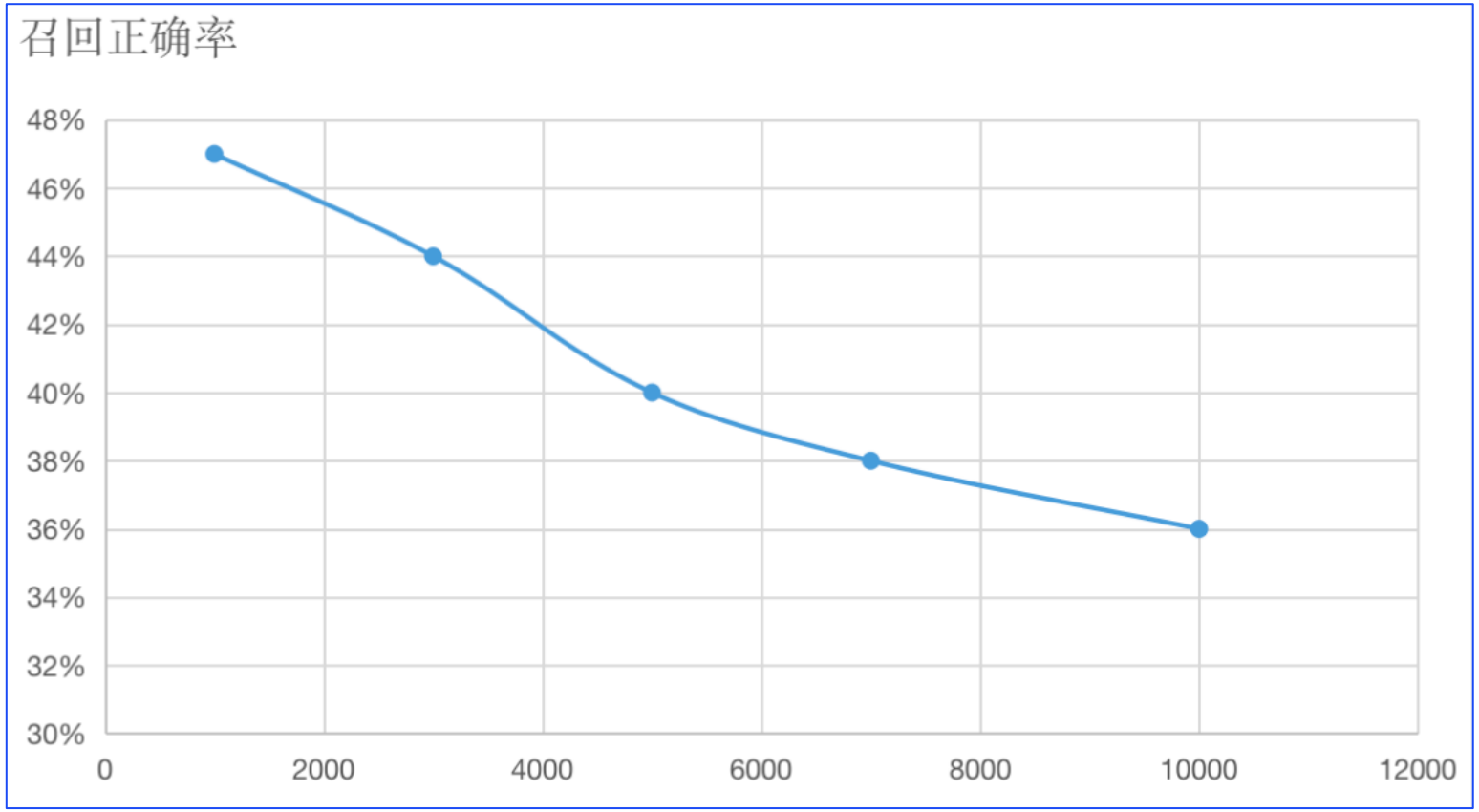
评审委员会应当在公告期满后30日内，对异议进行处理；不能及时处理完毕的申请材料，不列入评审范围，并由评审委员会向申请人说明。

Query：我有国内比赛的获奖证书，可以用来申请工艺大师吗？

难点：段落内容过长，Query与段落相似度低

典型问题和难点-召回率受文档数据量增大而降低

- 电力领域某专业多文档问答场景



- 纵坐标为TOP1召回准确率
- 横坐标为库内文档数量

领域数据治理需解决的核心问题

1. 文档解析
2. 文本召回优化
   - 文档的版本识别和管理；
   - 文档标题等信息的有效利用；
   - **文档内容的信息压缩；**
   - **适用于文档内容召回的向量训练。**
3. 领域高质量QA样本构建

PART 03

大模型文本数据治理的核心技术

大模型文本数据治理的核心技术

1. **文档解析**
2. 文本召回
   - 文档内容的信息压缩
   - 适用于文档内容召回的向量训练
3. 领域高质量QA样本构建

文档解析核心技术

1、文档周边信息识别

解决方案：版面识别模型，识别页眉、页脚、和上下左右周边附加信息坐标，凡在此类坐标范围内的信息解析后打上相应的标签

2、分栏识别段落、表格、图片等内容在上下文中的顺序

解决方案：版面识别模型，识别文档中的分栏情况。根据分栏坐标和上下左右的先后顺序，对分栏进行排序，再在各个分栏中根据从上到下原则，对段落、表格、图片等内容进行排序，从而确定整篇文档中的每个元素的先后顺序。

3、段落、表格、图片的精准识别

解决方案：版面识别模型先识别出段落、表格、图片元素信息和坐标信息。再基于具体解析方案（OCR解析、文本解析），对每个元素内的信息进行解析，对跨分栏、跨页等被分割的内容进行段落合并。

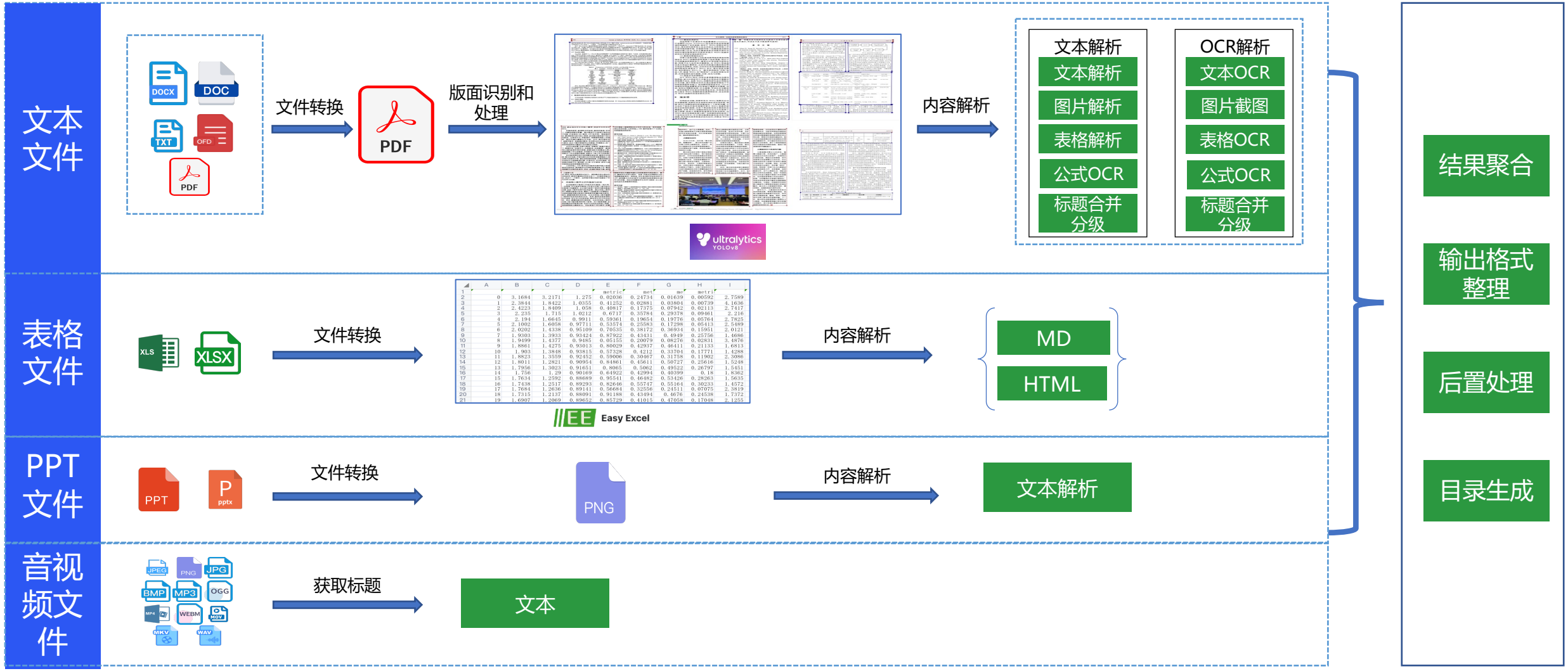
4、列表完整性识别

解决方案：版面识别模型，识别列表元素。解析过程中，对跨分栏、跨页等被分割的列表进行合并，对上下相连的列表，判断是否具有相同的列表特征再进行合并。

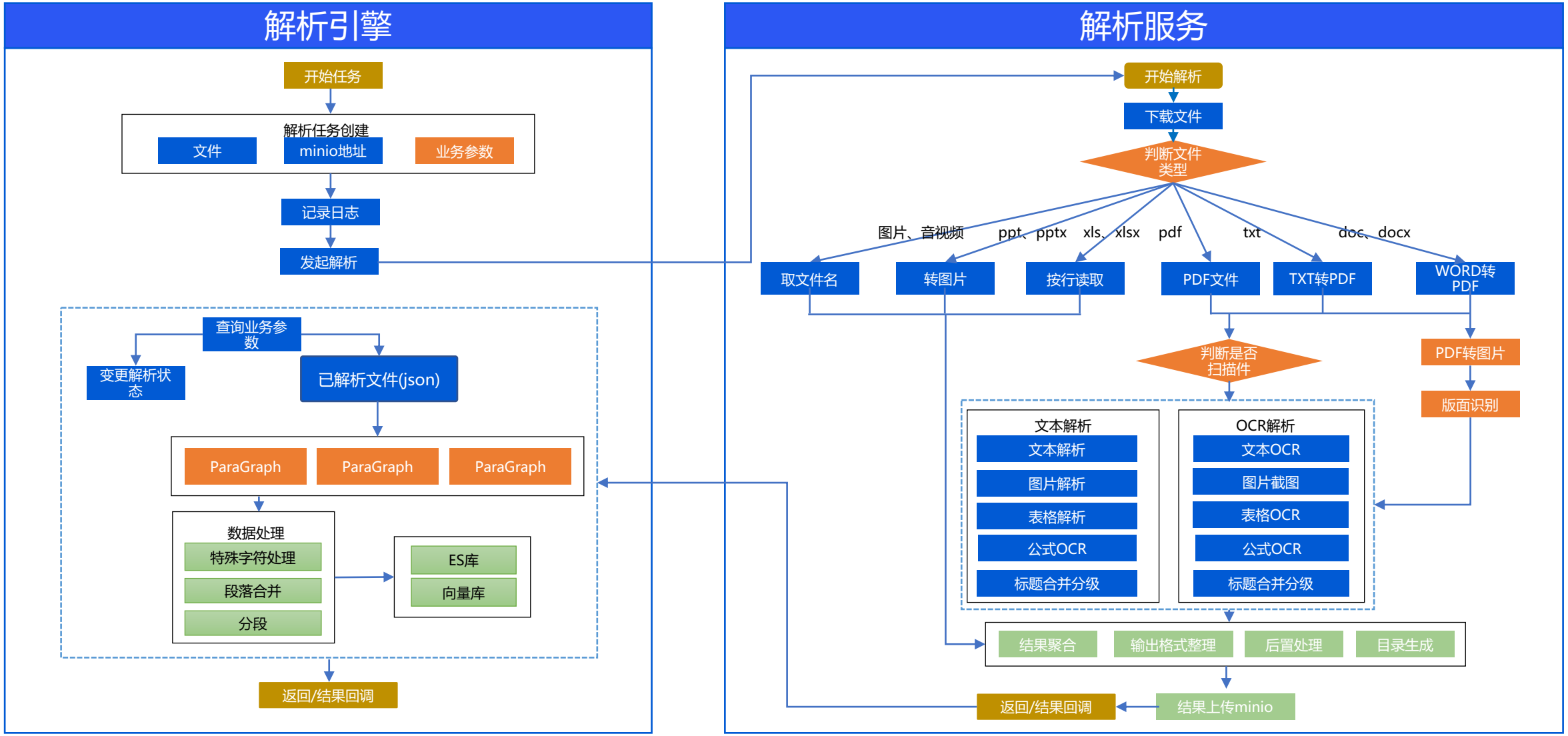
5、标题层级结构判断

解决方案：版面识别出的标题，各自截图，按顺序排列。应用标题识别优化的版面识别模型对标题的层级进行识别：结合标题的字体、加粗、数字等标题识别策略，判断标题层级。

文档解析引擎技术路径


文本文件
DOCX
DOC
TXT
OFD
PDF
文件转换
PDF
版面识别和处理
ultralytics YOLOv8
内容解析
文本解析
文本解析
图片解析
表格解析
公式OCR
标题合并分级
OCR解析
文本OCR
图片截图
表格OCR
公式OCR
标题合并分级
表格文件
XLS
XLSX
文件转换
Easy Excel
内容解析
MD
HTML
PPT文件
PPT
pptx
文件转换
PNG
内容解析
文本解析
音视频文件
JPEG
PNG
JPG
BMP
MP3
OGG
MP4
WEBM
MOV
MKV
WAV
获取标题
文本
结果聚合
输出格式整理
后置处理
目录生成

文档解析引擎-数据处理流程图



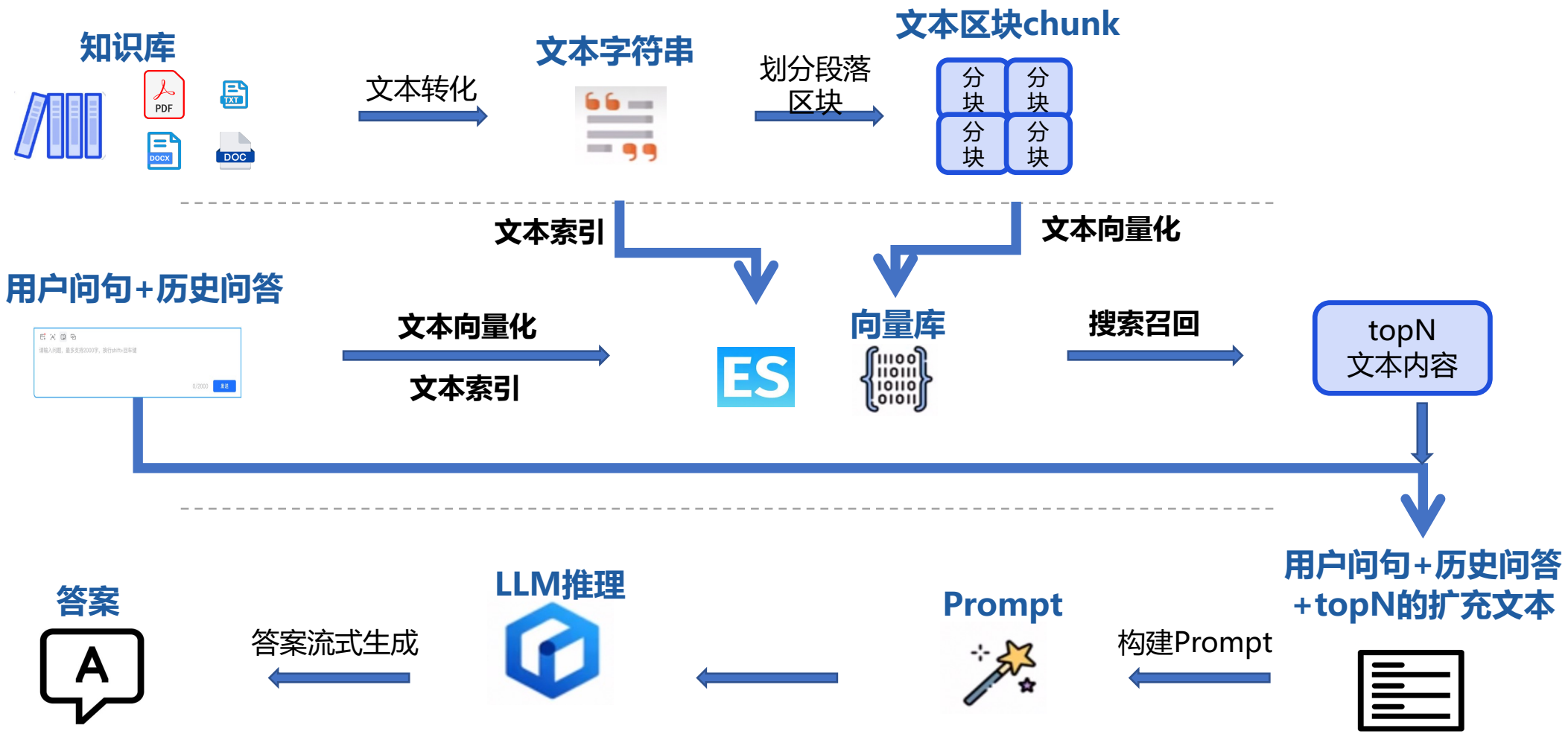
大模型文本数据治理的核心技术

1. 文档解析

2. **文本召回**
   - **文档内容的信息压缩**
   - **适用于文档内容召回的向量训练**

3. 领域高质量QA样本构建

一张图理解大模型外挂知识库的实现路径



文档内容的信息压缩——文本关键词

方案1、利用keybert关键词提取算法

方案2、利用prompt调用大模型生成对应关键词

方案3、对大模型进行关键词提取领域微调

方案4、对文档内容进行摘要

方案5、基于成分句法分析，提取名词短语，生成关键词

1）基于传统 NLP 的成分句法分析，提取名词短语；再通过短语间的依存关系，生成关键词列表

2）从完整语句的 Embedding，切换为关键词 Embedding：

- 知识库构建时，基于单知识点入库，入库时提取关键词列表进行 Embedding
- 查询时，对用户的问题提取关键词列表进行 Embedding 后，从本地知识库命中多条记录。

3）将单问句中的多知识点拆解后检索，将召回的多条记录交付给 LLM 整合。

方法优势：

- 相比传统 Embedding，大幅提升召回精准度。
- 支持单次交互，对多知识点进行聚合处理
- 使用传统 NLP 在专项问题处理上，相比 LLM 提供更好的精度和性能。
- 减少了对 LLM 的交互频次；提升了交付给 LLM 的有效信息密度；大大提升问答系统的交互速度。

基于句法关键词方案参考文章 https://zhuanlan.zhihu.com/p/641132245

向量召回模型优化

1. 向量召回模型

传统模型：基于Query与Query训练而成的相似度检测模型，而非Query与段落的相关性。

优化模型：训练目标优化为提升Query与段落的相关性，使得问题和相关段落的语义向量表示更接近，训练模型有sbert，cosent等。

2. 训练数据

构造 <问题，匹配文档段落，不匹配文档段落>的数据结构，使用语义相似度高但不是匹配文档段落的数据作为困难负样本。

3. 实验对比结果（某电力领域场景）

向量方案	top1	top3	top10
base-cosent	46.5%	57.7%	80.3%
cosent微调后	52.1%	74.6%	90.1%
BGE	56.3%	81.7%	91.5%
BGE微调后	69.0%	87.3%	97.1%

大模型文本数据治理的核心技术

1. 文档解析

2. 文本召回
   - 文档内容的信息压缩
   - 适用于文档内容召回的向量训练

3. **领域高质量QA样本构建**

领域QA样本构建（以电力场景为例）

1. 获取电力相关的题库，如注册电气工程师考试等，通过文档解析、正则抽取等步骤，获取题目和答案。

2. 获取电力领域的图书、教材、技术标准等文档，构建文档库。

3. 根据题目类型分别进行处理，从文档库中召回段落：

（1）填空题以及选择题，把答案回填到题干，作为检索文本，从文档库召回相关段落；

（2）问答题使用题目和答案作为检索文本，从文档库召回段落。

4. 根据检索文本和段落的相关度得分（如段落内容对检索文本的覆盖度等），对结果进行初筛，保留得分高的题目和段落组合。

5. 人工进行筛选，保留可信的题目和段落组合。

6. 把<问题，段落，答案>按照prompt模板，构造为文档问答的微调训练样本。

PART 04

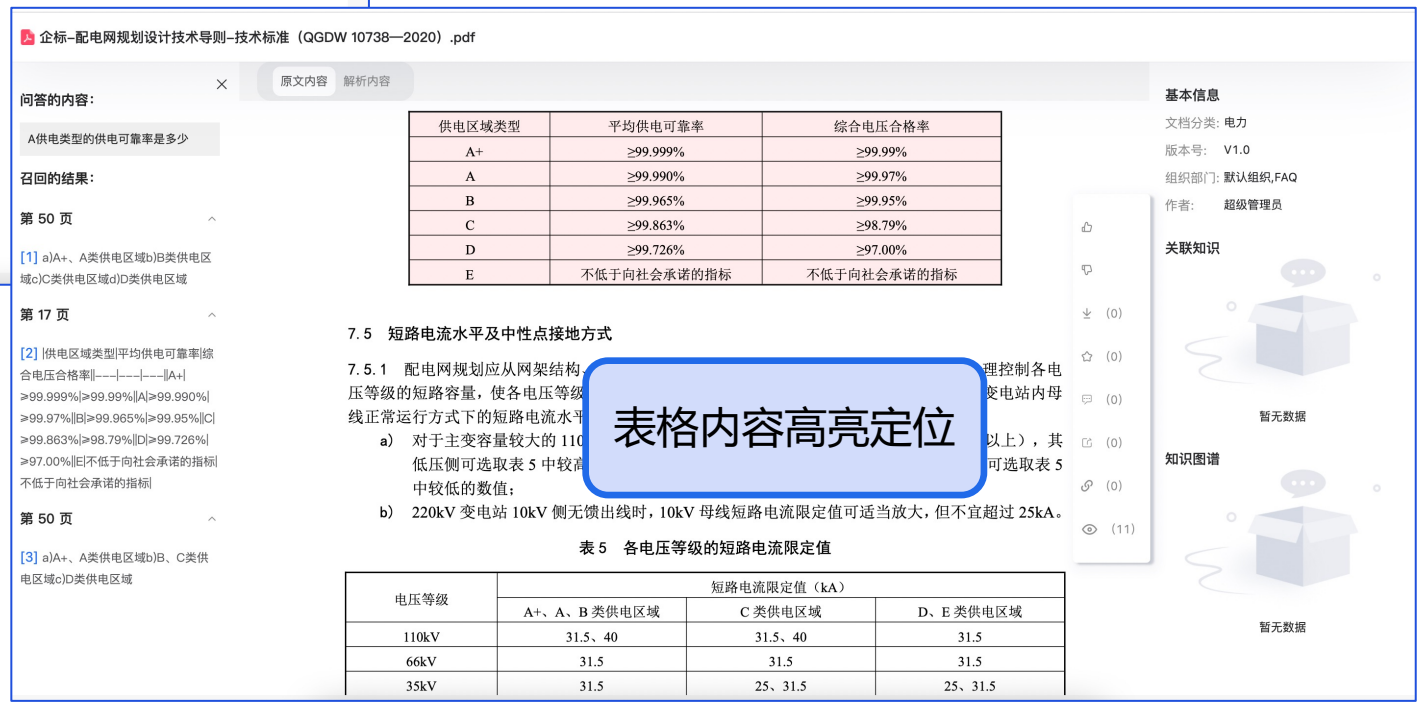
海致大模型应用落地案例

基于大模型外挂知识库的问答

- 支持pdf、word、txt等文档格式的解析，快速便捷构建知识库；
- 通过语义向量化、内容检索、对话管理等技术手段，实现了对文档知识问答、多轮对话、答案溯源等能力。

基于大模型外挂知识库的问答

- 基于统一解析引擎，对文档内部表格的内容识别和解析。
- 结合大模型的能力，可以实现表格答案的高亮溯源。



供电区域类型	平均供电可靠率	综合电压合格率
A+	≥99.999%	≥99.99%
A	≥99.990%	≥99.97%
B	≥99.965%	≥99.95%
C	≥99.863%	≥98.79%
D	≥99.726%	≥97.00%
E	不低于向社会承诺的指标	不低于向社会承诺的指标

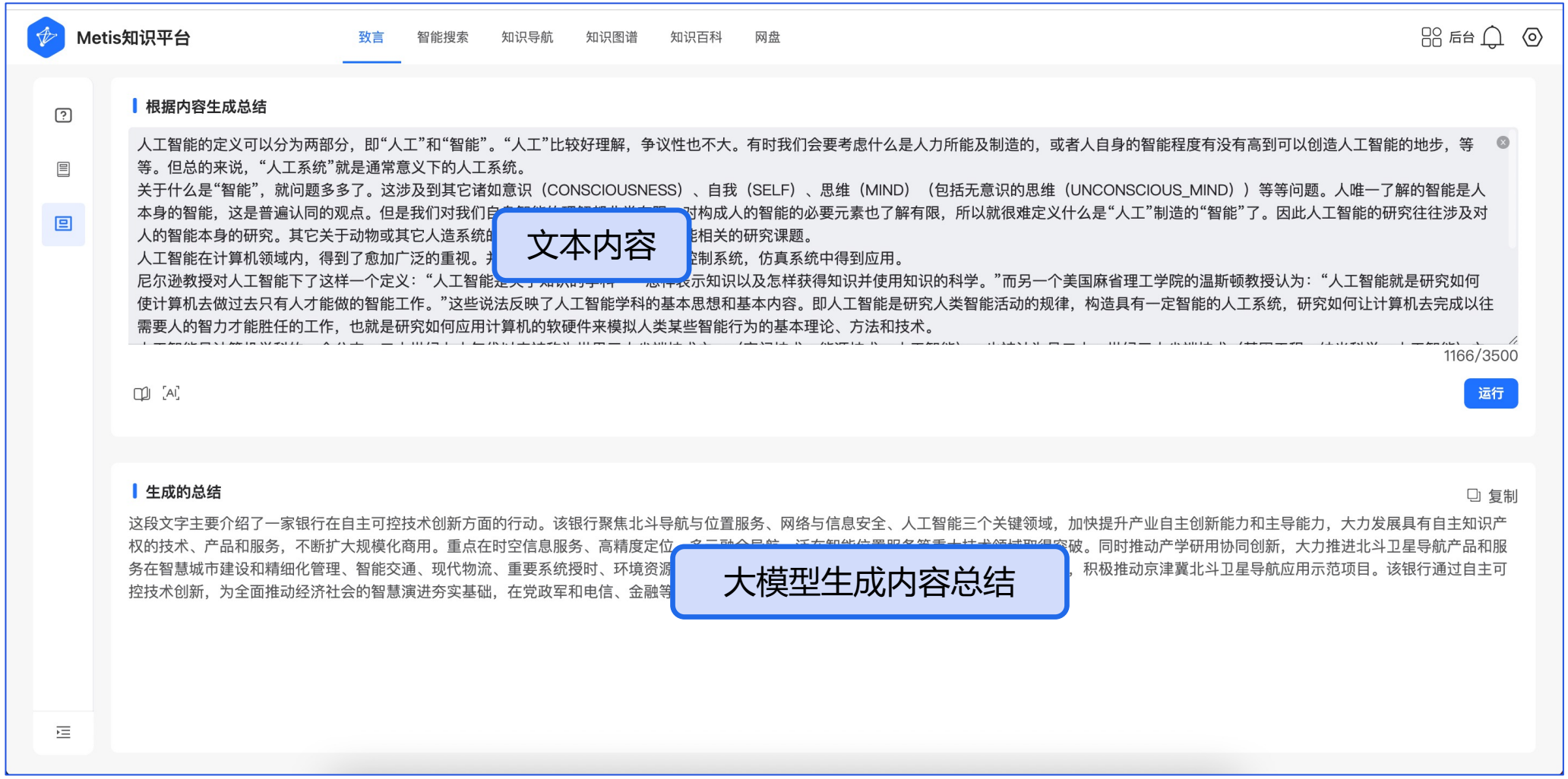
电压等级	短路电流限定值（kA）		
	A+、A、B 类供电区域	C 类供电区域	D、E 类供电区域
110kV	31.5、40	31.5、40	31.5
66kV	31.5	31.5	31.5
35kV	31.5	25、31.5	25、31.5

基于大模型外挂知识库的多轮问答

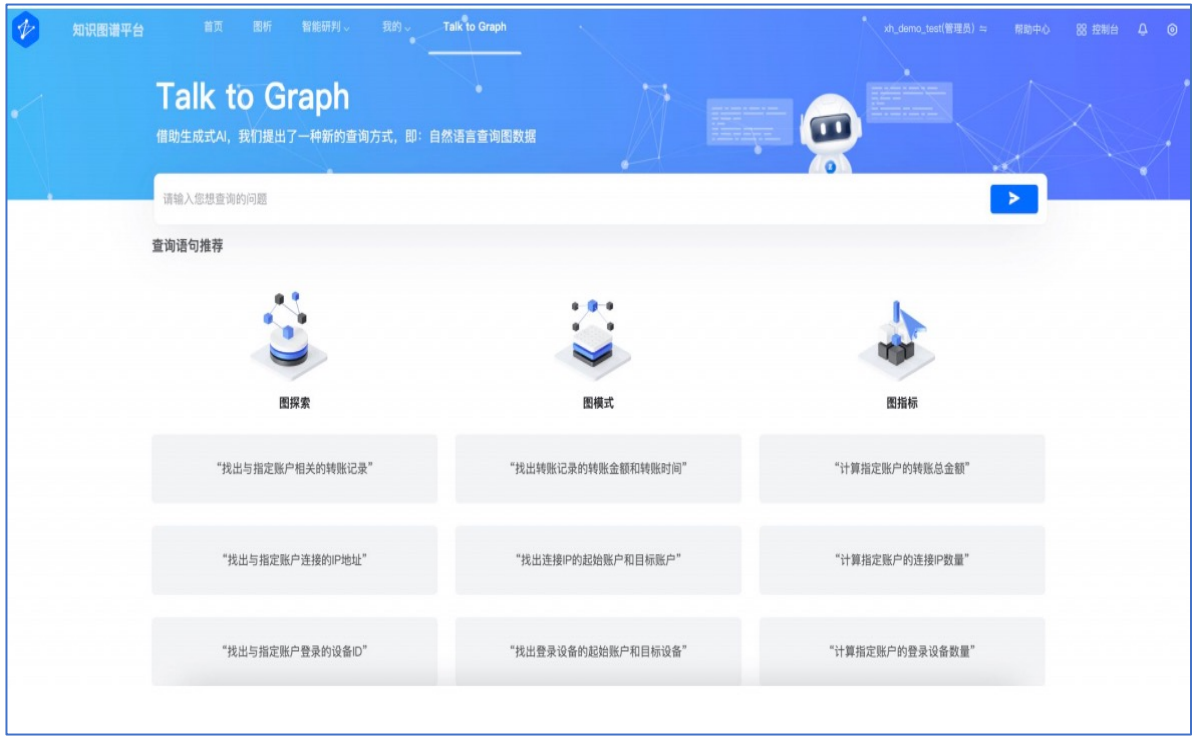
基于大模型的报告生成

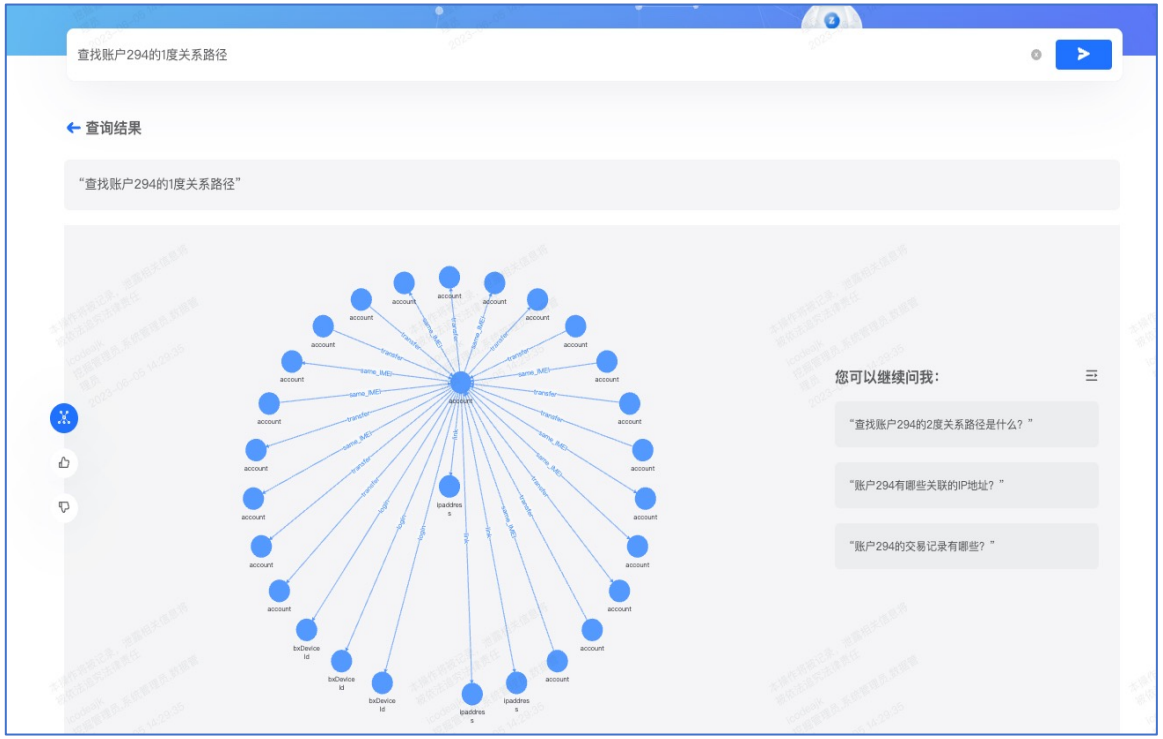
Metis知识平台
致言
智能搜索
知识导航
知识图谱
知识百科
网盘
后台
提纲
示例体验
1、什么是人工智能
2、人工智能的优势和劣势
2.1人工智能的优点
2.2人工智能的缺点
提纲
57/1000
说说你想写什么内容
人工智能
主题
4/1000
中断
生成的报告
报告：人工智能在能源领域的应用及发展趋势
大模型生成报告
一、什么是人工智能
人工智能是指通过计算机模拟人类智能思维和行为，实现智能化决策和自动化执行的技术。人工智能系统通过学习、推理、感知、自适应等方式，对数据进行分析、识别、处理和预测，从而实现自主决策和智能化服务。

基于大模型的摘要总结


Metis知识平台
致言
智能搜索
知识导航
知识图谱
知识百科
网盘
后台
根据内容生成总结
人工智能的定义可以分为两部分，即“人工”和“智能”。“人工”比较好理解，争议性也不大。有时我们会要考虑什么是人力所能及制造的，或者人自身的智能程度有没有高到可以创造人工智能的地步，等等。但总的来说，“人工系统”就是通常意义下的人工系统。
关于什么是“智能”，就问题多多了。这涉及到其它诸如意识（CONSCIOUSNESS）、自我（SELF）、思维（MIND）（包括无意识的思维（UNCONSCIOUS_MIND））等等问题。人唯一了解的智能是人本身的智能，这是普遍认同的观点。但是我们对我们
对构成人的智能的必要元素也了解有限，所以就很难定义什么是“人工”制造的“智能”了。因此人工智能的研究往往涉及对人的智能本身的研究。其它关于动物或其它人造系统
能相关的研究课题。
人工智能在计算机领域内，得到了愈加广泛的重视。
控制系统，仿真系统中得到应用。
尼尔逊教授对人工智能下了这样一个定义：“人工智能
示知识以及怎样获得知识并使用知识的科学。”而另一个美国麻省理工学院的温斯顿教授认为：“人工智能就是研究如何使计算机去做过去只有人才能做的智能工作。”这些说法反映了人工智能学科的基本思想和基本内容。即人工智能是研究人类智能活动的规律，构造具有一定智能的人工系统，研究如何让计算机去完成以往需要人的智力才能胜任的工作，也就是研究如何应用计算机的软硬件来模拟人类某些智能行为的基本理论、方法和技术。
文本内容
1166/3500
运行
生成的总结
复制
这段文字主要介绍了一家银行在自主可控技术创新方面的行动。该银行聚焦北斗导航与位置服务、网络与信息安全、人工智能三个关键领域，加快提升产业自主创新能力和主导能力，大力发展具有自主知识产权的技术、产品和服务，不断扩大规模化商用。重点在时空信息服务、高精度定位
破。同时推动产学研用协同创新，大力推进北斗卫星导航产品和服务在智慧城市建设和精细化管理、智能交通、现代物流、重要系统授时、环境资源
，积极推动京津冀北斗卫星导航应用示范项目。该银行通过自主可控技术创新，为全面推动经济社会的智慧演进夯实基础，在党政军和电信、金融等
大模型生成内容总结

基于大模型的图谱查询





通过大模型能力将自然语言转换为图查询语言；结合图谱信息进一步实现分析探查图路径的能力

AiDD AI+软件研发数字峰会
AI+ software Development Digital summit

感谢聆听