知识引擎原子能力

产品简介

腾讯云

文档目录

产品简介
产品概述

最近更新时间：2025-01-03 22:37:12

知识引擎原子能力（LLM Knowledge Engine Atomic Power）基于知识引擎研发的知识问答全链路能力，面向企业及开发者，提供灵活组建及开发模型应用的能力。您可通过多款原子能力组建您专属的模型服务，调用文档解析、拆分、embedding、多轮改写等服务进行组装，定制企业专属 AI 业务。
知识引擎原子能力（LLM Knowledge Engine Atomic Power）目前已提供多项检索增强生成（RAG）框架所需配套 API，以下为各个原子能力的简要介绍，可作为业务接入时选择的参考依据。

原子能力

提供 RAG 链路中解耦的各个能力，包括解析、拆分、embedding、多轮改写、重排序等。

原子能力名称	能力和特征	相关接口
文档解析（同步）	支持将多种格式文件转换成 Markdown 格式文件，可解析包括表格、公式、图片、标题、段落、页眉、页脚等内容元素，并将内容智能转换成阅读顺序。适用于对耗时要求较高的解析场景，如实时文档问答，支持的文件较小，耗时较短。	ReconstructDocument 文档解析 ReconstructDocumentSSE 文档解析 SSE
文档解析（异步）	支持将多种格式文件转换成 Markdown 格式文件，可解析包括表格、公式、图片、标题、段落、页眉、页脚等内容元素，并将内容智能转换成阅读顺序。适用于知识库问答等对耗时没有严格要求的场景，支持更大的文件。	CreateReconstructDocumentFlow 创建文档解析任务 GetReconstructDocumentResult 查询文档解析任务结果
文档解析拆分	支持将多种格式文件转换成 Markdown 格式文件并进行多级语义拆分，返回文件拆分后的结果。可用于后续的检索片段召回和阅读理解等。使用拆分模型后的相比传统正则切分方式，回答完整性提升20%。	CreateSplitDocumentFlow 创建文档拆分任务 GetSplitDocumentResult 查询文档拆分任务结果
embedding	支持调用文本表示模型，将文本转化为用数值表示的向量形式，可用于文本检索、信息推荐、知识挖掘等场景。	GetEmbedding 获取特征向量
多轮改写	该接口主要用于多轮对话中，进行指代消解和省略补全。使用本接口，无需输入 prompt 描述，根据对话历史即可生成更	QueryRewrite 多轮改写

	精确的用户查询语句。在应用场景上，本接口可应用于智能问答、对话式搜索等多种场景。	
重排序	重排序服务（ranker）提供 query 和切片片段之间的相关性排序服务，在 RAG 及搜索场景中，可通过排序服务找到相关性更高的内容并依次返回，引入排序服务可有效提升检索及大模型生成的准确率。	RunReRank 重排序

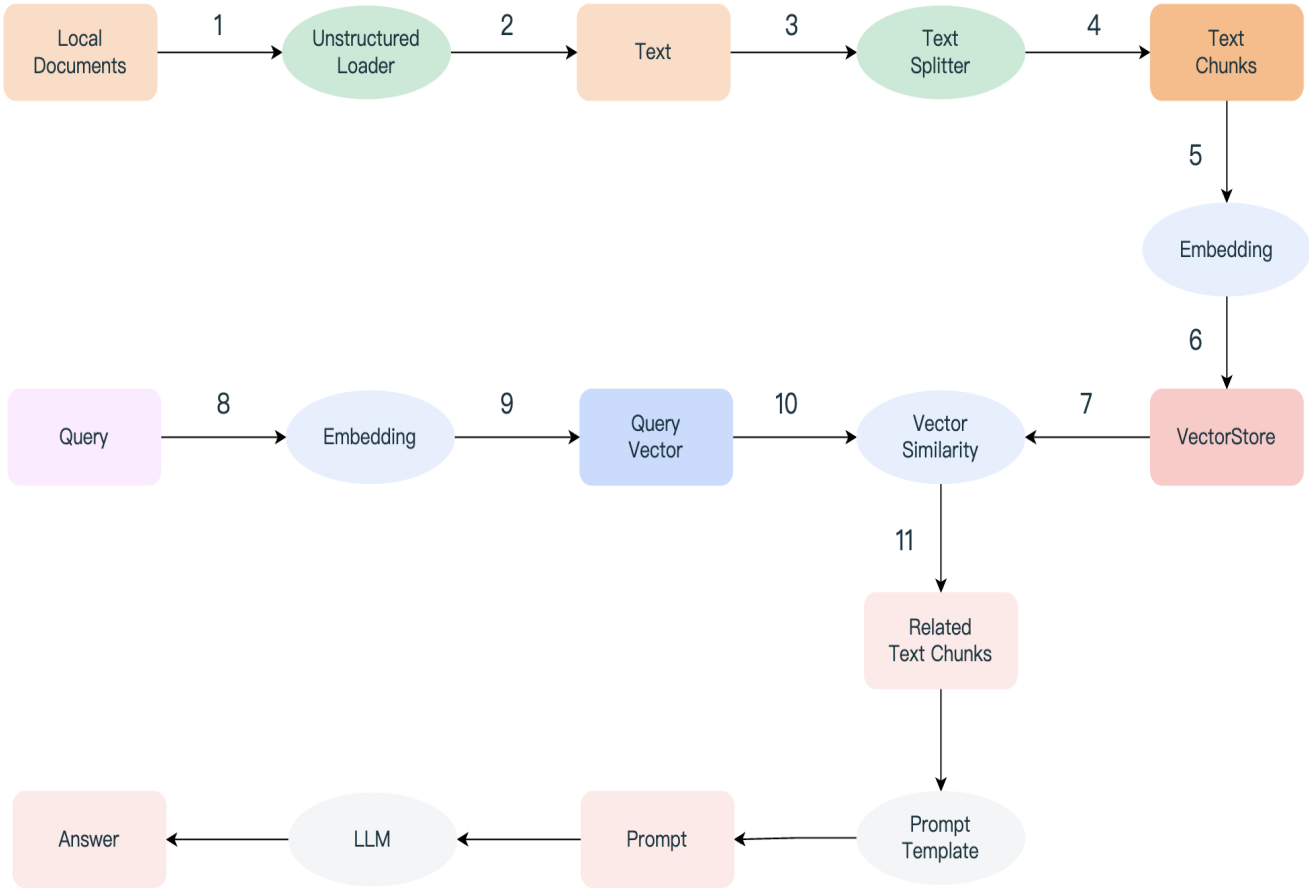
RAG 综合能力

检索增强生成 RAG（Retrieval-Augmented Generation）技术结合了检索系统的强大能力和生成模型的灵活性，可以解决更复杂的语言理解问题。知识引擎原子能力（LLM Knowledge Engine Atomic Power）基于大模型知识引擎 研发的知识问答全链路能力，面向企业及开发者，提供灵活组建及开发模型应用的组件化服务，可以基于此快速搭建一套效果更佳的 RAG 链路。

RAG 流程图

RAG 综合能力套件

提供 RAG 链路中的文件上传-解析-拆分- embedding -检索- rerank 的一站式全链路综合能力，您可综合使用各个接口获得 RAG 的检索结果，通过结合检索和生成模型的优势，实现高效的文档内容检索和精确的问答生成。详情请查看 操作指南 。


Local Documents
1
Unstructured Loader
2
Text
3
Text Splitter
4
Text Chunks
5
Embedding
6
VectorStore
7
Vector Similarity
10
Query Vector
9
Embedding
8
Query
11
Related Text Chunks
Prompt Template
Prompt
LLM
Answer

产品优势

最近更新时间：2025-01-03 22:37:12

业内领先的文档解析能力

基于大模型知识引擎的 OCR 大模型解析引擎，识别准确率提升30%，具有以下优势：

- 独创多模态文档解析大模型：在算法上，基于腾讯优图实验室自研新一代多模态文档解析大模型，通过粗粒度生成元素的位置及顺序，并辅以内容生成赋予上下文的语义感知，可以解决各种复杂排版的问题，并在图文表混排的场景下更具优势。
- 智能版面分析：与传统的 OCR 文字识别不同，文档解析产品能够快速抽取文档的关键属性，支持对多栏、内容混排等复杂版式的文档进行精准解析，如论文、报告、书籍等文档中的标题、段落、图片、表格、公式、页眉、页脚等多种版面元素，并按照阅读顺序提取内容。
- 表格结构识别：针对传统表格识别难题，全新的表格结构识别模型在常规、有线、无线、少线、多表格、跨页表格等复杂场景下能对表格精准检测和内容识别，并做结构化复原。
- 高精度文本识别：能够准确识别中英文、繁体字、生僻字等多种类字体，即使是图片和扫描的 PDF 文档，也能够进行高精度识别。
- Markdown 格式输出：支持将图片、PDF 文档转换为 Markdown 格式，这种轻量级的标记语言易于阅读和编写，非常适合大型模型训练和文档电子化。

业界首创基于 LLM 的多级语义切分模型

- 业界首创基于 LLM 的多级语义切分模型，通过语义理解的方式对文档进行切分，保障文档切分片段的语义完整性。
- 采用多级文档切分方式，将文档切分成适合检索和大模型问答的片段。
- 传统切分方式文档类型受限，缺乏通用性，且容易截断语义，语义模型的切分方式可有效解决该类型问题，端到端检索准确度大幅提升。

混合检索能力

- 支持向量检索，全文检索等多种混合检索策略，可根据业务场景灵活配置。
- 针对多行业（政务，汽车，文旅，教育，金融，制造等），多格式（PDF，Docx，Excel，MD，PPT等）多文档元素（普通文本，表格，图文，流程图等），端到端综合检索准确率达到90%。

基于 LLM 的 Embedding 模型

- 通过不同的 Instruction 区分Embedding 和生成任务，让LLM能同时在这两种任务上训练，从而得到一个同时具备文本表征和文本生成能力的模型。
- 借助LLM的多语言能力，同时支持多种语言的混合检索。

应用场景

最近更新时间：2025-01-03 22:37:12

知识引擎原子能力服务于RAG（Retrieval-Augmented Generation）检索增强生成框架，融合了检索系统的精准定位能力与生成模型的创新灵活性，针对复杂的语言理解挑战提供了强有力的解决方案。这一技术依托于先进的知识引擎原子能力——基于 大模型知识引擎 研发的知识问答全链路能力，旨在打造一套高效的知识问答全链路能力。
面向企业和开发者，我们提供了一套组件化的服务，使他们能够灵活地组建和开发模型应用。借助这套服务，用户可以迅速搭建起一套效果卓越的 RAG 链路，灵活接入 Embedding 及大语言模型，从而极大地提升了语言处理任务的性能和准确性。
知识引擎原子能力适用于以下场景：

智能客服

知识引擎原子能力能够快速准确地检索相关信息，结合生成模型的自然语言处理能力，为客户提供及时、准确且友好的咨询服务。

员工服务

通过知识引擎原子能力构建 RAG 框架，企业可以为员工提供个性化的信息服务，解答各种工作相关的问题，从而提高工作效率和员工满意度。

文档问答

针对大量文档资料，知识引擎原子能力能够迅速找到相关答案，为用户提供高效、便捷的文档查询体验。

查询助手

知识引擎原子能力可应用于各种查询场景，如搜索引擎优化、专业领域知识检索等，为用户提供更加精准、全面的查询结果。

车载助手

结合车载系统的特点，知识引擎原子能力能够实时响应驾驶员的需求，提供导航、娱乐、安全等方面的信息支持，提升驾驶体验。