一站式 AI 智能体开发平台

产品白皮书

目 录

1. 引言

本文档对软件一部数字应用产品部自研产品中国电信一站式 AI 智能体开发平台的能力视图、主要功能模块以及应用场景进行了描述，方便相关同事对该产品形成系统性理解，方便后续二次开发，方案包装以及销售。

2. 概述

2.1. 背景

2.1.1. 全球人工智能发展整体态势

1. 大模型推动智能“涌现”，打开 AI 技术发展上限

人工智能大模型，是指通过在海量数据上依托强大算力资源进行训练后能完成大量不同下游任务的模型。在技术层面上，大模型的实现采用“预训练+指令微调+人类反馈的强化学习”的训练范式。首先通过预训练技术将深度学习网络在海量数据上进行自监督训练，然后利用指令数据进行有监督指令微调，提升模型对人类指令的追随能力，最后，基于由人类价值标注数据训练得到的奖励模型所提供的奖励信息进行强化学习，控制大模型的输入符合人类价值判断。在大模型使用时，通过设计提示进行即时学习可以进一步提升大模型完成各类任务的能力。规模化是使大模型强大的重要原因，研究表明当模型规模足够大的时候，会“涌现”智能能力，具备处理新的、更高层次的特征和模式的能力，能够为一系列下游任务带来更好的任务效果。大模型不断扩大的规模由“量变”引发“质变”，模型通用认知能力不断提升。大模型能力的迅速发展不仅有助于人类完成“规定动作”，还可能帮助人类去研究和发现未知领域，突破人类过去没有突破过的极限。

大模型的技术变革呈现数据巨量化、模型通用化、应用模式中心化的特点。整个发展历程可划分为三个阶段 2013—2018 年的深度学习阶段，主要还是基于传统的“针对特定任务的专用模型+大量标注数据”方式，在监督学习的机制下训练得到一个专用小模型，但是在词向量的自监督学习中，使用大规模数据进行预训练的方法已初见端倪。2017 年 Transformer 的提出为基础架构带来了规模化构建和规模化运算的潜力 Transformer 解决了 RNN 和 LSTM 的并行化训练和长距离依赖问题，解决了 CNN 的局部归纳偏差问题，能够容纳更多的参数规模，并且具备更强的语义特征提取能力、长距离特征捕获能力、综合特征提取能力。2018 年—2022 年的预训练阶段，基于“海量无标注数据”，在自监督学习机制下获得预训练大模型，通过少量标注数据微调后得到领域专用模型。自监督学习机制的成功使得可利用的数据愈发“巨量化”，从标

注数据拓展到无标注数据。Bert 将可利用的预训练数据量扩大 3 到 5 倍，成为自然语言理解任务中的基准模型。自此，“预训练+微调”的学习范式成为主流。在通用大模型上通过少量标注数据微调，即可适用于一系列下游任务 2022 年下半年以来的大语言模型阶段，预训练大模型的通用能力愈发强大，引入指令监督训练使得模型能更好地追随人类指令完成各种任务，并提升了在下游任务上的泛化能力，通过人类反馈学习让机器与人类价值对齐成为可能。

2. 大模型变革内容生产和技术服务模式，“无限生产”推动生产效率颠覆式提升

内容生产方面，生成式大模型率先在内容创作、图像生成、数字人、游戏等娱乐媒体领域广泛应用，内容生产效率和质量显著提升，内容生产模式从辅助人到“替代”人演变。据 Gartner 预测，至 2023 年底，将有 20%的内容被生成式大模型所创建；至 2025 年底，生成式大模型产生的数据将占所有数据的 10%。技术服务方面，大模型的“无限生产”能力重塑企业生产引擎。随着大模型能力的不断提升，AIAgent 成为重要发展趋势。未来，大模型将不仅仅是一种生产工具，更多是作为企业“合作者”，持续为企业注入生产动能。

3. 大模型作为新的“根”基础设施，驱动 AI 范式变革

大模型实现模型生产从“作坊式”到“流水线”的升级。大模型出现以前，AI 模型是“定制化、场景化”的开发方式，针对特定应用场景需求训练一个个小模型，模型难以复用和积累，导致 AI 落地的高门槛、高成本与低效率。大模型实现基础模型底座的标准化开发和泛在化应用，解决成本困境。

通用大模型通过从海量的、多场景、多领域的数据中学习共性知识，成为具有通用性和泛化能力的模型底座。基于通用大模型底座可搭建各行业的垂类大模型，可以有效缩减垂类大模型训练所需要的算力和数据量，缩短模型的开发周期，提升垂直领域的应用开发效率。openAI 以 GPT4 通用大模型为底座，通过快速增量训练和个性化微调的方式，允许普通用户通过简易对话界面自定义定制 GPT，支持开发者采用私有数据对 GPT 进行个性化微调，使大模型更易于访问和开发，产品形态更加丰富，以满足更广泛的市场需求。

4. 中美是大模型技术领域的主要“玩家”，大模型市场竞争持续深入

2023 年 5 月发布的《中国人工智能大模型地图研究报告》指出，美国和中国发布的通用大模型总数已占全球发布量的 80%。美国方面，形成了。penAI+微软 Meta、谷歌等多个“阵营”，openA 重点围绕 GPT4 底座模型完善上层开发者生态，Meta 通过开源 LLaMa 等大模型，引领了全球大模型开源浪潮。

中国工程院院士郑纬民指出，美国作为全球科技霸主一直引领人工智能领域发展前沿，整

个大模型的产业布局全面领先，在研发能力、人才储备、硬件设施及融资环境方面占据优势。相较而言，中国占据海量数据资源和应用场景优势，但顶尖的 AI 人才缺乏，在基础理论、原创模型等颠覆型、阶跃型技术方面仍缺乏引领能力。产业基础层的整体实力较弱，高质量数据积累不足，在高端芯片、关键基础软件等领域受制于美国。

2.1.2. 国内外大模型行业应用概述

1. 国外大模型行业应用情况

美国大模型商业化应用进展全球领先，商业化进展迅速。一是网络、存储等基础设施建设完备，技术发展成熟，为大模型广泛应用打下良好基础。二是具备充足的用于大模型训练推理的高端芯片，算力充沛。三是大模型技术领先，以 OpenAI 为代表的大模型公司对美国大模型在全球取得领先地位和广泛落地起到重要推动作用。openAI 作为全球人工智能顶尖研究机构，以 GPT4 为底座，为个人、开发者和企业大模型应用持续赋能，其近期推出的 GPTStore 为大模型应用带来爆发式增长。据不完全统计，美国大模型应用已经覆盖医疗、金融、房地产、媒体、军事、气候预测等领域，如微软将 GPT4 能力集成到 office 等办公软件中，提高办公效率和用户体验；摩根士丹利也接入 GPT4 能力，优化财富管理咨询流程；房地产服务商 Realtor.com 的大模型工具可根据用户提示自动生成房屋图像以及进行房源匹配；报纸出版商 Gannett 将大模型集成到出版系统中，简化运营，帮助记者摆脱日常繁琐任务、解放生产力欧盟、英国、加拿大、新加坡、日本、印度等国家和地区的大模型应用尚处于前期尝试阶段，仅个别头部企业开始应用。在英国，会计、法律等行业的国际知名企业在部署大模型，如普华永道已在英国员工测试使用尽职调查、识别合规问题、审批交易等功能，未来将面向全球推广；英国最大律师事务所之一麦克法兰宣布，与法律领域生成式 AI 企业 Harvey 达成技术合作，将在法律咨询、法律生成、查询，客户服务等领域全面应用生成式 AI。在日本，7-11 连锁便利店将大模型用于产品创意和规划，提升产品研发效率；本田汽车将大模型用于汽车设计。在印度，教育科技企业 PhysicsWallah 宣布引入 AlakhAI 平台，该平台将协助学生进行小组学习、解决学术和生活问题、提供支持和鼓励，甚至创建个性化的学习计划。

2. 我国大模型行业应用情况

大模型行业发展迅猛，我国政府积极制定相关政策加速大模型产业发展。2023 年 7 月，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布《生成式人工智能服务管理暂行办法》，鼓励和规范生成式人工智能创新发展。

同时，各省市地区也在积极出台大模型产业发展措施加速大模型落地。据不完全统计，截

至 2023 年 11 月，北京上海、广东、安徽等地均发布大模型相关政策，从算力支持场景开放、技术突破、产品生态等多方面鼓励大模型应用落地。2023 年 5 月，北京市人民政府印发《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023—2025 年）》，强调构建高效协同的大模型技术产业生态，建设大模型算法及工具开源开放平台，构建完整大模型技术创新体系，积极争取成为国家人工智能开放生态技术创新中心。北京市人民政府办公厅印发《北京市促进通用人工智能创新发展的若干措施》，强调开展大模型创新算法及关键技术研究加强大模型训练数据采集及治理工具研发，推动大模型在政务、医疗、科学、金融等领域的示范应用。上海市印发《上海市推动人工智能大模型创新发展若干措施（2023—2025 年）》，推进大模型创新应用，支持大模型在智能制造、生物医药、集成电路、智能化教育教学等领域构建示范应用场景打造标杆性大模型产品和服务。

在政策指引下，全国各地大模型落地速度加快。北京大模型应用进展全国领先，各领域全面开花。目前，北京大模型已在政务、金融、医疗等领域实现落地，大模型应用氛围浓厚。如北京市计算中心为帮助政协更好适应新时代政协提案工作新要求，开发了大模型相关的政协提案应用。如今，政协提案系统能够准确地从多源信息中凝练和关联语义，实现根据工作重点和社会热点丰富提案线索和选题的功能。元保险基于大模型构建智能客服和智能理赔应用，保险智能客服相较传统机器人客服问题解决率提升 80%，智能理赔应用相较之前审核速度提升 70%。北京友谊医院引入云知声科技的山海大模型，在内分泌科试点门诊电子病历生成，在医生问诊的过程中，电子病历系统能自动过滤无关对话，医生和患者的口语交流自动转化为标准化的书面语言，并从非结构化自动梳理为结构化的表述方式，形成电子病历文书初稿，随后经医生审核形成正式的电子病历。

其他城市大模型落地以政务示范为主，当地领军企业尝试为辅。广州、深圳、厦门均积极推进大模型在政务领域的落地，如广州白云区城管局与华为云合作探索华为云盘古政务大模型在城市治理领域的创新应用，并成立政务大模型实验室，对占道经营、垃圾堆积、城中村治理等城市治理典型场景展开探索。深圳福田政府借助华为盘古大模型，已经在政府服务、城市治理、政府办公等方面实现政务创新。厦门人社发布了人社领域大语言模型，并应用于厦门 12333 智能客服，通过 AI 智能座席替代人工座席完成咨询服务工作在企业方面，四川领军企业发布长虹大模型智慧家电 AI 平台“长虹云帆”，通过云帆 AI 平台，用户可以通过语音指令一次性完成多个操作，同时电视还能够根据多维感知技术了解用户的身体信息，并提供相关健康建议。

2.2. 产品定位

一站式 AI 智能体开发平台是一款基于专家知识库数据能力、训推一体化平台模型能力，联合内外部生态构建丰富的提示词库、AI 算法仓、插件库、大模型基础共性组件，搭建的一站式 AI 智能体应用定制开发平台，实现智能体应用快速构建，提升智能体应用二次开发效率和交付能力。

2.3. 产品价值主张

AI 智能体应用研发的使能者，提供便捷、高效的 AI 大模型应用开发工具，降低智能体应用开发的技术门槛，提升应用二次开发效率，助力客户应用快速交付

2.4. 架构图

2.4.1. 功能架构

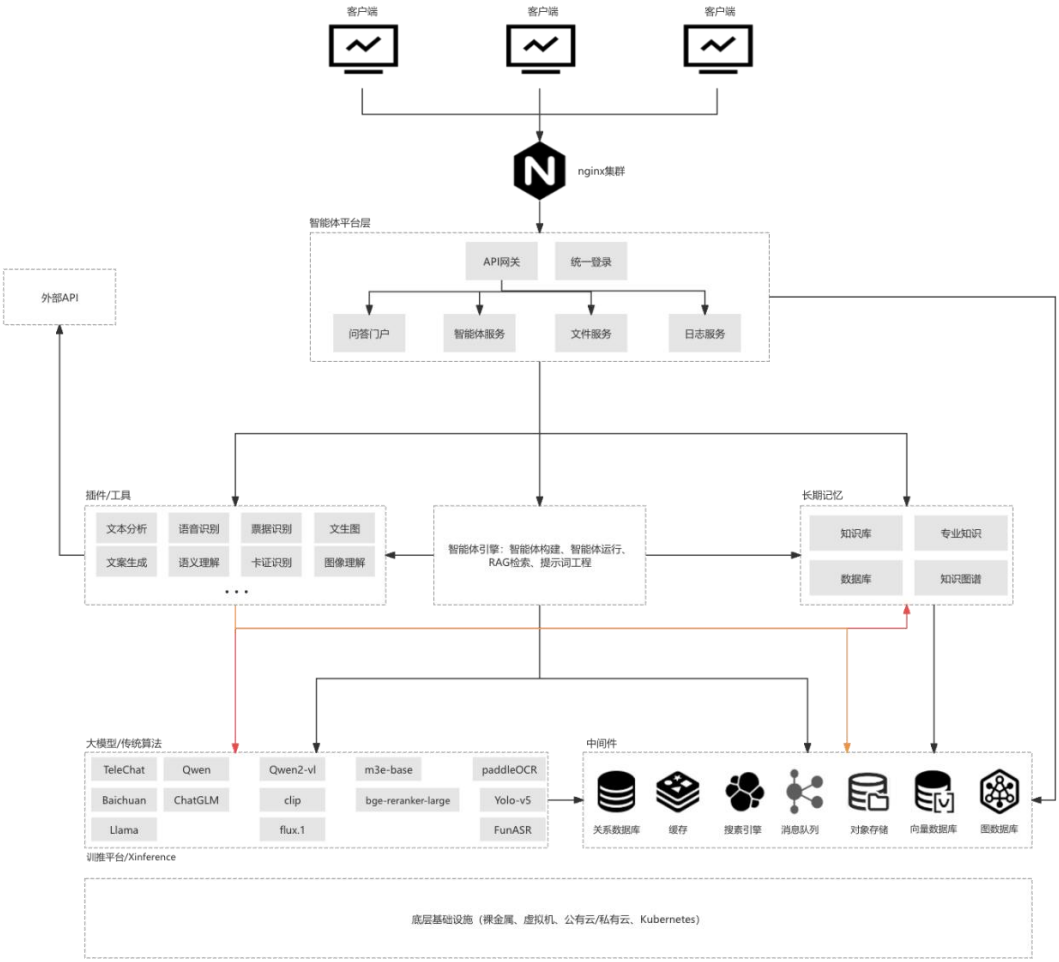
基于一站式 AI 智能体开发平台，可以 5 步新建智能体应用，结合提示词库、知识中心、插件库功能，为通用场景、业务场景提供能力支持。智能体平台功能主要包含基础服务、智能体开发、智能体及平台管理、智能体应用、平台及应用部署几部分。如下图所示：

2.4.2. 技术架构

平台选择前后端分离的架构，其中前端采用 VUE3 + Element-UI + Vite 的技术选型。后端采用 Springboot 3.0 + SpringSecurity + MybatisPlus + MySQL + Redis + Minio 的技术选型。在此架构下，前后端可以并行工作，各自专注自己的领域，减少了相互等待的时间，提高了整体项目进度。

在基础设置层的基础上，平台设置有的中间件层，搭建有统一的关系型数据库，缓存，搜索引擎，对象存储和向量数据库等中间件，可供上层应用和大模型进行调用，方便管理。

另外平台使用训推平台或 Xinference 搭建统一的大模型/传统算法层，统一对外提供模型及算法能力。系统架构如下图所示：



2.4.3. 系统流程

平台的基本业务流程如下：

1) 用户登录智能体平台，创建新智能体，进入智能体编排页面。

2) 基座模型选择：在智能体编排页选择所需的 AI 模型，并调整对应参数。

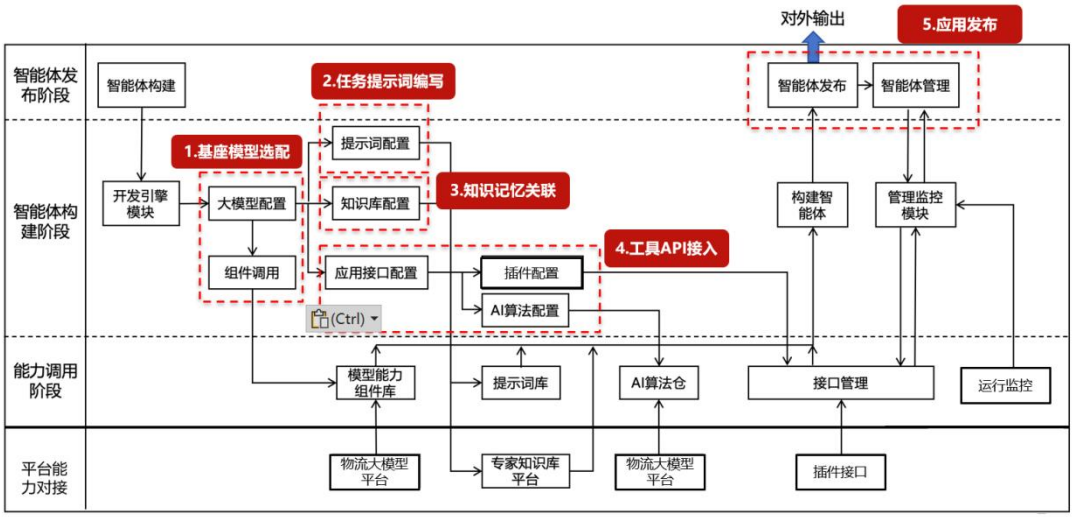
3) 任务提示词编写：根据需求编写提示词，定义智能体的功能。

4) 知识记忆关联：选择并优化私域知识库，提升智能体应用的专业性。

5) 工具库 API 接入：从工具库中选择所需工具，集成到智能体应用中。

6) 应用发布：一键发布智能体应用，完成智能体开发和部署。

访问智能体站点，通过网站直接访问，嵌入别的系统访问或 API 的方式将智能体嵌入到三方系统中。



业务流程图

3. 产品功能介绍

3.1. 系统概述

智能体平台是一站式 AI 智能体应用定制开发平台，可深度挖掘并整合业务数据，无缝对接大模型、算法、插件能力。平台集成了全面的模型服务工具箱、全链条定制化应用开发套件，以及多样化的应用发布渠道，通过精简的五个步骤——基座模型选择、任务提示词编写、知识记忆关联、工具库 API 接入、应用发布，即可迅速发布适配于各类系统的智能体应用。

3.2. 智能体构建引擎

搭建智能体编排引擎，通过自主规划、工作流编排两种方式，实现 5 步快速构建智能体。

3.2.1. 自主规划应用

提供 5 步构建智能体编辑页面，实现模型选择、模型参数调整、提示词撰写、知识库选择、插件选择、对话开场白配置、智能体发布等功能。核心功能包括以下内容

- 模型选择：允许用户根据需求选择最合适的 AI 模型，以提升任务执行效果
- 模型配置：提供灵活的参数设置，帮助用户优化模型表现，包括但不限于 temperature，top-p 等常见参数。
- 提示词配置：让用户为智能体配置提示词，以精准引导生成内容
- 提示词优化：以 react 框架为基准，优化用户输入提示词
- 知识检索：支持用户配置知识库，为大模型提供知识检索增强（RAG）
- 插件配置：支持用户配置已有插件

- 对话开场白：支持用户配置开场白，使智能体问题更加人性化
- 推荐问题：支持用户提供推荐问题，提供问题示例
- 语音输入：支持语音输入
- 语音输出：支持大模型回答转语音
- 输入审查：支持对用户输入进行审查
- 输出审查：支持对大模型输出进行审查

3.2.2. 工作流应用

提供工作流编排方式快速构建智能体，包含大模型、知识库、插件等 9 类节点选择，开展智能体编排、调试及发布，实现 5 步构建智能体。

- 开始节点：开始节点支持配置开场白以及推荐问题，默认包含用户输入参数
- 结束节点：支持直接回答以及模型回答
- 大模型节点：支持模型选择、提示词配置
- 知识库检索节点：支持用户选择知识库
- 插件工具节点：支持选择用户配置或系统内置插件
- 条件分支器节点：支持创建 if else 分支节点
- 问题分类器节点：识别用户意图，进行问题分类
- 代码节点：支持用户代码撰写
- Http 请求节点：支持配置 http 请求节点
- 对话开场白：支持用户配置开场白
- 推荐问题：支持用户提供推荐问题
- 语音输入：支持语音输入
- 语音输出：支持大模型回答转语音
- 输入审查：支持对用户输入进行审查
- 输出审查：支持对大模型输出进行审查

3.2.3. 智能体调试

提供智能体调试功能，进行智能体功能效果测试，支持停止回答等功能。

- 停止回答：支持回复中停止回答
- 重新回答：支持重新回答
- 思维链输出：支持输出思维过程，可展示每步骤输出

- 流式输出：支持流式输出内容
- 参考来源：支持展示参考来源
- 文件上传：支持上传文件进行问答

3.2.4. 智能体应用发布

支持多种智能体状态发布，如仅自己可见、平台内可见等，支持通过 API、站点访问等多种发布访问形式。

- 私密发布：支持用户私密发布智能体，仅自己可见
- 公开发布：支持用户公开发布智能体，平台内可见
- 公开可配置发布：支持用户公开发布智能体，平台内可见，且可配置
- API 发布：支持以 API 形式发布智能体应用
- 嵌入式发布：支持以代码嵌入式形式发布智能体应用，包括 iframe 嵌入以及悬浮球嵌入
- 站点式发布：支持在平台内发布智能体应用，通过页面试用
- API 密钥申请：支持用户自行申请 API 密钥

3.2.5. 智能体调用分析

统计分析会话数、回复速度、互动数、token 消耗数等运营指标。

- 全部会话数：智能体应用与用户的每日会话总数，会话数即用户与智能体建立的对话数，包括智能体调试及正式运行阶段
- 平均会话互动数：正式投入运行的智能体应用，每日每个会话中用户与智能体沟通次数的平均值，用户与智能体一问一答算一次沟通
- 智能体回复速度：智能体回答结果的 Tokens 输出速度，包括智能体调试及正式运行阶段
- Tokens 消耗量：智能体应用每日消耗的 Tokens 总量，包括智能体调试及正式运行阶段，智能体思考以及回答消耗的全部 tokens

3.3. 智能体构建工具

提供提示词库、插件库、智能体评测三类工具支持智能体快速构建。

3.3.1. 提示词库

提供预置提示词模板，支持用户查看、搜索等功能，并可根据具体使用场景进行提示词优

化，自建提示词模板

- 预置提示词模板：平台预置的成熟提示词，可供用户调用
- 提示词查看：查看预置的提示词模板内的具体内容
- 提示词分类：预置提示词按照标签进行分类，包括角色扮演等
- 提示词搜索：用户可搜索自己想查看的提示词
- 提示词优化：提供提示词优化功能
- 自建提示词：提供自建提示词模板功能，用户自定义创建提示词

3.3.2. 插件库

提供预置插件，支持用户查看、搜索等功能，支持用户通过本地文件导入自建插件

- 插件管理：平台预置的成熟插件进行管理，可供用户调用
- 插件详情查看：查看预置插件的具体内容
- 插件分类：预置插件按照标签进行分类，包括助手等
- 插件搜索：用户可搜索自己想查看的插件
- 本地文件导入插件：支持自定义创建插件，通过 json 本地文件导入
- URL 和原始数据导入插件：支持自定义创建插件，通过 URL 和原始数据导入

3.3.3. 智能体评测

开展智能体效果测试，展示评测清单、评测集等功能

- 评测清单：展示已有评测任务列表，支持分页展示，支持关键字搜索
- 评测集：展示已有评测集，支持分页展示，支持关键字搜索
- 评测集上传与下载：支持用户上传评测集，下载评测集模板
- 评测任务管理：用户填写评测任务名称，选择需要评测的智能体应用，选择模型，选择评测集，创建评测任务

3.4. 知识管理中心

知识管理中心包含专家知识库、专业知识、数据库以及知识图谱等类型的知识管理系统，为智能体提供知识来源。

3.4.1. 知识管理中心架构

3.4.1.1. 功能架构

知识管理中心以知识库为核心，面向云边端协同场景提供多层次检索增强能力，通过复杂

文档解析和大模型技术深度优化 RAG 功能，集成知识审核、版本管理等全生命周期运营工具，并扩展知识图谱、文本纠错等辅助功能，既可作为智能体平台的知识中枢，也能以独立服务形式输出检索接口，最终构建具备多场景适配性和自我迭代能力的知识服务体系。

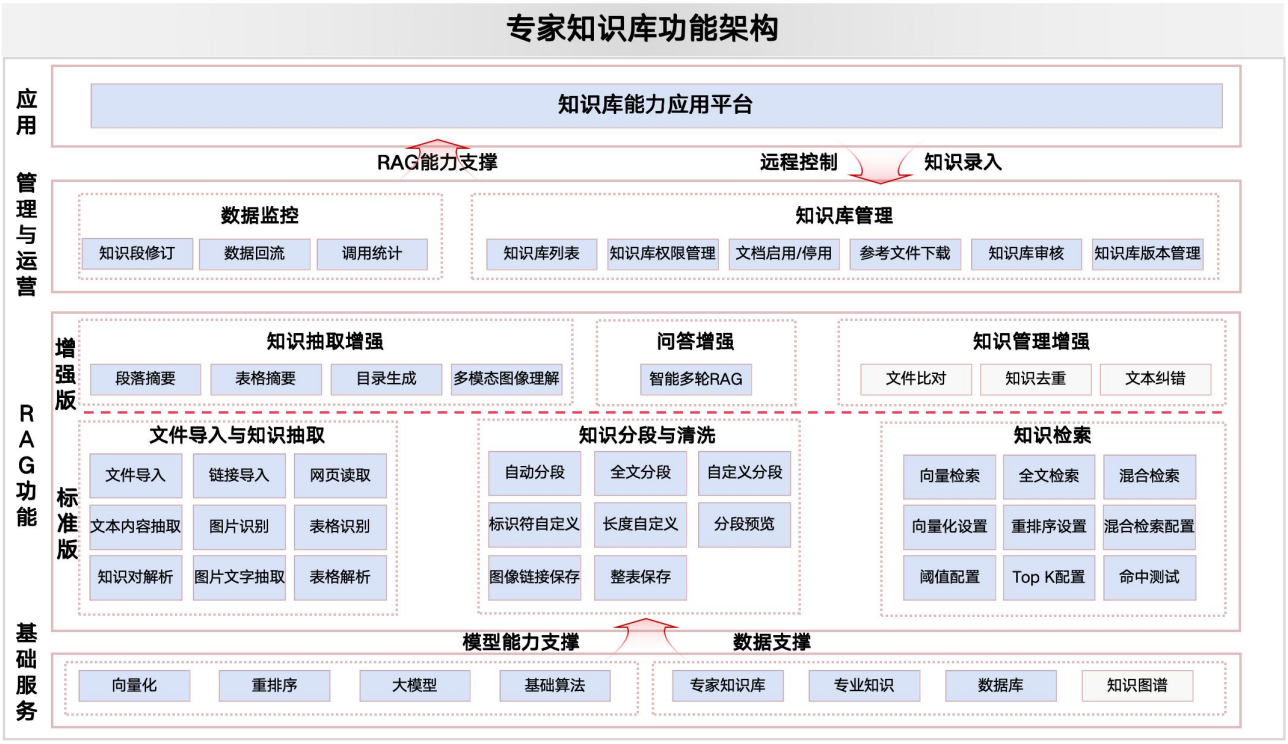


图 3：专家知识库功能架构图

3.4.1.2. 业务流程

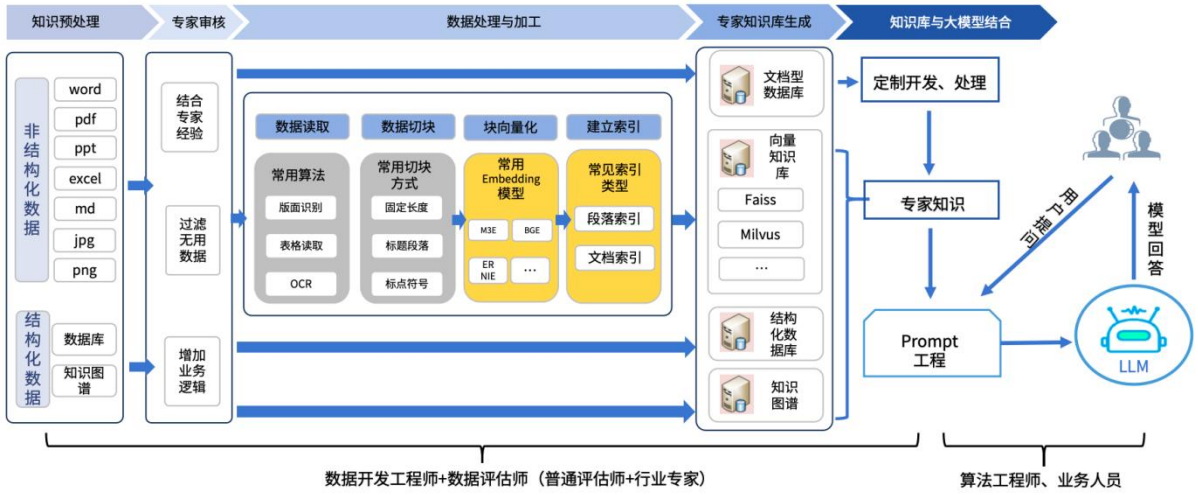


图 4：知识库业务流程图

专家知识库为构建一站式知识图谱、向量、结构化及非结构化知识库生成工具，为“智能

体应用”提供丰富的语料支持。

第一步，对知识进行预处理，通过人工筛选，找到需要存储到知识库的知识文件，如 PDF、WORD 文件等。经过专家审核或经过知识过滤，确保知识为有效的知识数据。

第二步，经过数据读取、数据切块、块向量化、建立索引四个步骤，完成对非结构化数据的知识索引构建。

第三步，把分段处理的知识分别存储在文档型数据库和向量知识库中，同时可接入数据库或知识图谱丰富知识检索能力。

第四步，通过知识库与大模型结合，帮助客户实现知识检索增强问答功能，为模型提供专业知识，保障问答的准确性。

3.4.2. 专家知识库

专家知识库包括知识库展示、导入数据、知识分段、知识索引、知识检索、知识片段预览、数据回流等功能

3.4.2.1. 知识库展示

知识库展示功能具备可视化展示能力，以卡片的方式展现存储在知识库中的所有知识内容。展示知识库描述以及包含的知识片段数量。知识库的正确描述有助于智能体在选择知识库时命中正确的知识库。

3.4.2.2. 导入数据

导入数据功能支持将外部数据源的知识内容批量导入知识库，能够自动识别并处理不同格式的数据（如 TXT、 MARKDOWN、 PDF、 HTML、 XLSX、 XLS、 DOCX、 CSV、JPG、PNG 等）。此功能可以将大量非结构化数据转化为向量化知识，并确保数据质量符合知识库标准。支持本地文档导入、文件服务器导入以及网页链接导入。

3.4.2.2.1. 本地文档导入

本地文档导入功能支持用户将存储在本地设备上的文档文件（如 Word、PDF、Excel、TXT 等格式）批量导入到专家知识库中。此功能使得用户可以轻松将日常文件转化为可在知识库中检索、共享和管理的知识内容，减少了手动输入和文档整理的工作量，提升了工作效率。

3.4.2.2.2. 文件服务器导入

文件服务器导入功能允许用户将企业内部文件服务器上的文档自动导入到专家知识库中。这些文件可以是存储在共享文件夹中的各种类型的文档，通过与文件服务器的接口对接，系统

可以定期或按需自动抓取文件并进行处理。此功能可以大大提升文件管理效率，使得知识库能够及时同步企业内部的文档更新。通过集成化操作，用户无需手动干预，即可确保知识库内容的最新和完整性。

3.4.2.3. 知识分段

知识分段功能具备对复杂信息进行自动拆解与分类的能力，可以将大量的知识内容依据主题、领域或其他标准切分为独立的知识单元。此功能能够实现知识的模块化管理，使得每个知识片段更加精准和易于更新。通过知识分段，用户能够快速定位所需内容，提升知识存取的效率，并便于后续的知识补充与迭代。

3.4.2.3.1. 自动切段

自动切段功能利用自然语言处理（NLP）和机器学习算法，能够根据知识内容的主题、结构和语义自动将文本划分为若干个逻辑清晰的段落或单元。系统会分析文档中的句子结构、关键字、上下文语境等因素，智能地识别并切割出每个知识片段，确保每个分段都具备独立性和完整性。此功能不仅提高了分段的效率，还减少了人工干预和操作，提高了处理大量文本时的自动化程度，使得大规模知识库的管理更加轻松和高效。

3.4.2.3.2. 全文切段

全文切段功能是一种基于文档整体结构和逻辑的切割方式，旨在将整个文档或知识内容按照预定规则或文本结构进行系统化切割。与自动切段和自定义切段相比，全文切段更多考虑文档的整体逻辑性和流畅性，能够根据文档的章节、标题、段落结构等，完整地划分出一系列符合语义和上下文的段落或部分。该功能能够自动识别文档中的标题、子标题、段落标志等结构性元素，精准切分并保持文本的连贯性和层次感。

全文切段功能能够：

1. 提升大规模文档处理效率：自动对大篇幅文档进行切段，减少人工干预，快速将文档内容转化为易于管理和检索的段落。

2. 保持文档的语义完整性：通过对文档的整体结构分析，确保分段后的内容逻辑通顺，便于后续的检索、学习和知识应用。

3. 提供一致性和标准化的分段格式：适用于需要对大量文档进行统一处理和分类的场景，保持企业或项目文档的标准化格式，便于知识的进一步扩展和应用。

3.4.2.3.3. 自定义切段

自定义切段功能允许用户根据具体需求手动调整或定义分段规则，对知识内容进行个性化切割。用户可以根据主题、章节、内容类别等不同标准，灵活设定切段的方式，确保每个知识片段更符合业务需求或用户习惯。此功能支持根据关键词、段落长度、内容逻辑等多个维度进行精细化分段，使得知识库中的内容在展示和管理时更加灵活和高效，用户可以根据不同场景对知识进行更高层次的定制和优化。

3.4.2.4. 知识索引

知识索引功能具备智能化的知识分类和索引能力，通过分析知识内容的关键词、语义和关系，自动生成多层次、全方位的索引体系。该功能能够实现知识的快速定位与检索，用户可以通过关键词、主题或关联性高的内容快速找到相关资料。知识索引的建立不仅提升了知识库的使用效率，还增强了系统在处理大规模知识时的查询性能，帮助用户高效应对复杂的信息检索需求。

3.4.2.4.1. 文件索引

文件索引功能能够对导入到知识库中的各类文档进行高效的索引，确保用户能够通过关键词或主题快速检索到相关的文件内容。系统通过对文件内容进行语义分析、关键词提取、文档结构解析等技术手段，自动生成多层次、结构化的索引。每个文件会被赋予多个标签和关键词，这些索引项支持用户基于主题、关键字、领域等维度进行精准检索。通过文件索引，用户可以快速找到文档中的关键信息，避免传统文档管理中的信息碎片化问题。

文件索引功能的关键能力：

1. 多维度索引：支持按内容、关键字、标签、日期等多维度对文档进行索引，提升检索灵活性。

2. 全文检索：支持全文检索功能，即使用户仅记得文档中的部分信息，也能快速通过索引找到相关文件。

3. 自动化更新：每当新的文件导入时，系统会自动更新索引，确保知识库中的文件始终可以通过最新的索引项被快速检索。

3.4.2.4.2. 知识对索引

知识对索引功能主要针对知识库中包含的知识对（FAQ、技术支持、常见问题等）进行索引管理。通过对每个问题进行语义分析和关键词提取，系统能够生成与问题内容相关的多个索

引项，并根据用户查询的意图，准确匹配到相关的答案。该功能还支持智能化的匹配机制，能够在用户提问时，依据历史问题和答案进行优化推荐，帮助用户高效找到所需解答。

3.4.2.5. 知识检索

知识检索功能具备强大的检索引擎能力，支持基于关键词、语义、标签等多维度进行查询。借助自然语言处理（NLP）技术，该功能能够理解用户的查询意图，提供精准的搜索结果，并支持模糊匹配和高相关性排序。知识检索功能通过高效的数据处理和深度学习算法，不仅能提供实时的搜索结果，还能根据用户的行为进行个性化推荐，提升知识查找的准确性和相关性。

3.4.2.5.1. 向量检索

向量检索功能基于自然语言处理（NLP）和深度学习技术，通过将文本内容转化为向量表示，提供高效的语义搜索能力。每个文档、问题、答案或知识片段都通过向量化的方式进行编码，系统可以通过计算向量之间的相似度，快速找到与用户查询最相关的内容。向量检索不仅考虑了关键词的匹配，还能够理解文本中的语义关系和上下文信息，从而实现更精准的结果匹配。

向量检索功能的关键能力：

1. 语义理解：通过深度学习和预训练模型，系统能够识别不同表达方式下的相同意图，从而提高检索的准确性。

2. 相似度计算：基于向量空间模型，系统能够通过计算查询向量与知识库中各个内容向量的相似度，提供高相关度的检索结果。

3. 多语言支持：向量检索能够跨语言处理，支持多种语言的文档和查询，打破语言障碍，提升全球用户的检索体验。

3.4.2.5.2. 关键词检索

关键词检索功能是传统的基于关键词匹配的检索方式，系统通过用户输入的关键词在知识库中进行精准匹配，并返回相关的文档或信息。该功能能够快速准确地定位到包含目标关键词的内容，支持多种检索模式，如精确匹配、模糊匹配和短语匹配。关键词检索在处理结构化信息时尤其高效，广泛应用于文档搜索和常见问题解答中。

关键词检索功能的关键能力：

1. 精确匹配与模糊匹配：支持精确匹配用户输入的关键词，同时也支持模糊匹配，容错率高，能够处理拼写错误、近义词等情况。

2. 过滤与排序：系统能够根据相关性、时间、重要性等条件对检索结果进行排序，帮助用户快速找到最相关的信息。

3. 支持短语检索：用户可以通过短语或多词组合进行检索，提升检索的灵活性和准确性，尤其适用于技术性和专业性内容。

3.4.2.5.3. 混合检索

混合检索功能结合了向量检索和关键词检索的优势，能够同时基于语义理解和关键词匹配进行检索，从而实现更加全面和精准的搜索结果。通过将两种检索方法的优势互补，混合检索可以有效解决单一检索方式在某些场景下的局限性。例如，在面对一些模糊、含糊或多义的查询时，混合检索可以结合语义和关键词信息，提供更好地匹配结果。

混合检索功能的关键能力：

1. 综合匹配：同时进行关键词匹配和语义相似度匹配，系统能够根据两者的结果进行加权排序，提供最相关的搜索结果。

2. 增强检索效果：通过结合向量检索的语义理解能力和关键词检索的精确性，混合检索可以优化搜索体验，提高检索的全面性和准确性。

3. 动态优化：混合检索支持动态调整搜索策略，根据查询内容和上下文自动选择最佳的检索方式，以适应不同类型的查询需求。

3.4.2.6. 知识片段预览

知识片段预览功能支持用户预览已切分向量化好的知识片段，支持用户浏览该片段的字符数，命中数以及对片段开启或停止。通过知识片段预览功能，用户可以对知识片段单独管理，更新知识内容

3.4.3. 专业知识

专业知识功能为大模型提供了一个增强的知识库，超越传统的文本数据，能够通过系统化的专业知识和专业公式，帮助大模型在特定领域提供更精确、权威的解答。这些功能能够显著提高大模型在技术、医学、金融、法律等专业领域的应用效果，为用户提供更有深度和实用价值的答案。

3.4.3.1. 专业知识对

专业知识是大模型在特定领域中知识整合的核心。通过将领域专家总结的理论、概念和常见问题结合成知识对，模型能够迅速从中提取相关信息，进行有效的推理和回答。这些知识对

能够使模型不仅仅停留在表层的词汇理解上，更深入到专业领域的核心内容，提供更加符合行业标准的解答。例如，在医学领域，模型可以运用医学术语知识，帮助解答患者的症状、治疗方案等问题。专业知识的持续更新和优化，使大模型能够快速适应不断变化的行业需求和专业标准。

3.4.3.2. 专业公式

专业公式功能是大模型在处理需要精确计算或推导的专业问题时的重要支撑。通过嵌入领域特定的数学公式、计算模型和推导规则，模型能够实现复杂问题的公式推算和数值计算。例如，在金融领域，模型可以运用股票估值公式、风险评估模型等进行精确的财务分析专业公式功能不仅能够提高计算的准确性，还能帮助大模型对用户提出的定量问题进行更为精确和专业的解答。

3.4.4. 数据库链接

数据库链接功能使得用户能够将私有数据库与平台进行连接，借助智能体的能力进行数据查询、总结和分析。通过该功能，用户可以将分散在不同系统或数据源中的信息集成到平台，提供更加精准和智能的数据分析服务。智能体不仅可以访问数据库中的历史数据，还能实时获取最新数据，支持对复杂业务场景的深度挖掘与分析。

数据库管理功能帮助用户高效地管理与平台连接的数据库，支持关系型数据库 MySQL 库。用户只可以通过可视化界面查看数据库的元数据，确保数据流动的安全与高效。

- 自动化连接设置：智能配置向导帮助用户简化数据库连接过程，无需复杂的技术操作，用户可通过简单配置完成连接。
- 定期同步与备份：平台支持数据库内容的定期同步与备份，避免数据丢失，确保数据的完整性与可靠性。

3.4.5. 知识图谱管理

知识图谱（KG）通过结构化的知识表达机制为检索增强生成（RAG）系统提供了突破性解决方案：通过显式呈现实体间的关联网络获取知识图谱中的复杂逻辑关系。在检索阶段为大语言模型（LLM）注入精准的语义上下文，使系统既能基于实体关系路径精准定位多层级知识片段，又能通过关系推理构建逻辑链条。显著提升 LLM 对复杂问题的解析能力，提供逻辑清晰、知识精准的问答服务。

3.4.5.1. 数据导入

将输入文本进行拆分，提取实体与关系，生成摘要信息，并根据这些信息构建内存中的图（Graph）结构。然后从这个图中识别出各个社区，为每个社区创建报告，并在图中创建文本块节点和文档节点。

3.4.5.2. 图谱构建

自动化地从大量文本数据中抽取实体、概念及它们之间的关系。这一功能不仅能够识别和分类文本中的关键信息，如人名、地点、事件等，还能揭示这些信息之间复杂的相互关联，如因果关系、属性关系、层级关系等。通过这种方式，语言大模型能够帮助组织构建出结构化的知识图谱，这些图谱将作为知识的存储库，不仅促进了信息的快速检索和理解，也为进一步的数据分析和机器学习应用奠定了基础。通过精确的实体识别和关系抽取，语言大模型使得知识图谱构建变得更加高效和准确，从而在知识管理、推荐系统、智能搜索等多个领域发挥着重要作用。

3.4.5.3. 可视化展示

知识图谱的可视化展示将复杂的实体、关系及属性以直观的图形方式呈现，帮助用户快速理解数据结构和语义关系。结合知识图谱的特征、技术原理和传统信息可视化方法，通过“确定知识主题（主题层）-处理与分析数据（数据层）-构建数据三元组（关系层）-进行可视化映射（可视层）”进行可视化展示。

3.5. 智能体管理

包括我的智能体、智能体启停管理、模板管理、版本管理、审核 5 个功能

3.5.1. 智能体广场

提供预置的成熟智能体应用，可供用户体验、查看、试用、依此模板构建自己的智能体。支持用户查看、搜索智能体等功能。

- 官方智能体：平台预置的成熟智能体应用，可供用户体验、查看、依此模板构建自己的智能体
- 可配置智能体：公开智能体分为公开，公开且配置两种状态，可配置智能体支持用户以此智能体为模板构建
- 智能体标签：智能体应用的分类标签，包括公开配置、企业管理、日常办公等

- 搜索智能体：用户可搜索自己想查看的智能体
- 试用智能体：提供用户试用公开智能体的功能，可体验智能体效果

3.5.2. 我的智能体

提供查看、编辑、试用、发布自己已经建好的智能体功能

- 编辑智能体：进入智能体编排页面，调整智能体的关键参数
- 试用智能体：直接试用智能体，体验智能体效果
- 发布智能体：将智能体由私密状态调整为公开或公开可配置

3.5.3. 智能体启停管理

控制智能体提供服务的状态，配置为启动时智能体对外提供服务。

- 智能体启用：控制智能体提供服务的状态，配置为启动时智能体对外提供服务。
- 智能体停用：控制智能体提供服务的状态，配置为停止时智能体不再对外提供服务。

3.5.4. 智能体模板管理

支持将创建好的智能体按照模板导出，或将智能体模板导入平台创建新的智能体。

- 导入模板：将智能体模板导入，创建新的智能体
- 导出模板：将创建好的智能体按照模板导出

3.5.5. 智能体版本管理

记录智能体更新的版本，支持用户选择历史版本进行回退。

- 版本记录：记录智能体更新的版本
- 版本比较：用户更新版本时可对比与历史版本的参数对比
- 版本回退：支持用户选择历史版本进行回退

3.6. 平台管理

包含门户登录管理、智能体集市、智能体管理、系统管理、工作空间管理 5 部分，维持平台业务功能正常运行。

3.6.1. 门户登录管理

提供高效、安全且易于管理的方式来实现单点登录，多个独立系统可以通过共享统一的认证中心，具备统一登录功能。满足各级功能整体呈现和快速跳转，系统门户首页提供平台各种指标的统计及展示功能。

3.6.2. 工作空间管理

支持用户创建多个工作空间，并在不同工作空间间切换，管理不同工作空间的用户，工作空间内插件、提示词等共享。

- 空间列表：列出用户所在的所有空间
- 默认空间：存在一个无法删除以及无法邀请用户的主空间，名字为默认工作空间
- 新建空间：新建空间默认为个人空间，可以邀请其他用户当空间内人员超过 1 人时，自动从个人空间转变为团队空间。
- 空间角色：支持用户对空间角色进行管理。包含所有者、管理员、开发者。
- 成员管理：对本工作空间的用户进行管理，可编辑昵称以及角色
- 工作空间切换：支持用户选择不同用户空间，空间之间做到数据隔离

3.6.3. 智能体审核

管理员用户可审核一般用户提交待发布的智能体应用，可进行查看、审核通过、审核驳回等操作

- 审核查看：查看用户提交审核，准备发布为公开或公开且配置的智能体应用详情
- 审核驳回：驳回用户的审核要求
- 审核通过：通过用户的审核要求，智能体发布至智能体广场

3.6.4. 系统管理

包括账号管理、消息中心、日志管理三部分

- 账号管理：实现平台登录用户、角色、权限信息管理，提供完善的授权管理机制
- 消息中心：展示用户收到的与自己相关的所有消息、未读消息、收藏的消息等
- 日志管理：采集系统各模块日志，形成统一日志管理视图，支持对平台生成日志的统一检索功能

4. 基础能力介绍

4.1. 插件工具能力

4.1.1. 基础组件

基础组件专注于提供高效、实用的通用功能。这类插件涵盖了图表生成、代码执行、数学表达式计算、正则表达式提取、时间获取和二维码生成等多种工具，为智能体的日常任务提供

全面支持。

4.1.1.1. 安全围栏

安全围栏功能基于高效的字典树算法构建敏感词检测引擎，实现用户输入及大模型输出内容的逐层快速匹配。该架构支持 O(n)时间复杂度的高效检索，可对文本内容进行毫秒级实时扫描。系统预置包含万余条敏感词以及敏感组合词，涉及政治敏感、色情低俗、广告营销等领域的敏感词及敏感组合词库，同时支持用户自定义上传敏感词和个性化白名单，个性化定制敏感词库。

4.1.1.2. 代码执行

代码执行功能采用沙箱化容器技术，用于运行代码并返回执行结果。它支持多种编程语言（如 Python 和 JavaScript），用户只需提供代码片段和指定语言，即可快速验证代码逻辑、调试程序或执行特定任务。

4.1.1.3. 计算数学表达式

计算数学表达式插件是一款高效、精准的数学计算工具，专注于快速解析和计算复杂的数学表达式。该插件利用 NumExpr 库进行本地执行，支持多种数学运算符和函数，能够处理从简单算术到复杂数学公式的各种计算需求。用户只需输入数学表达式，即可快速获取计算结果。

4.1.1.4. 提取正则表达式

正则表达式内容提取插件利用正则表达式从输入内容中提取匹配的结果。用户只需提供需要处理的文本内容和正则表达式规则，即可快速获取符合匹配逻辑的提取结果。该工具支持复杂的正则表达式语法，能够处理多种场景下的文本提取需求，例如从日志文件中提取特定信息、从网页内容中提取结构化数据等。其灵活高效的特性使其成为文本处理和数据提取的得力助手，特别适用于需要精确匹配和高效提取的场景。通过正则表达式内容提取插件，用户可以轻松实现复杂文本的自动化处理，大幅提升工作效率和数据处理能力。

4.1.1.5. 获取当前时间

获取当前时间插件是一款简洁实用的时间工具，专注于为用户提供更多样化的功能，包括获取当前时间、支持多种时间格式和时区转换等。用户只需调用该插件，即可快速获取精确的当前时间，满足日志记录、时间敏感型任务调度以及自动化工作流程等场景的需求。其高效性和准确性确保了在各种应用场景中都能提供可靠的时间信息，是时间管理与任务处理的得力助手。

4.1.2. 大模型能力

大模型插件是智能体平台的高级功能模块，专注于利用大规模预训练模型（如 GPT、BERT 等）实现复杂任务的高效处理。这类插件通过调用大模型的强大语言理解和生成能力，为智能体提供自然语言处理、知识问答、文本生成等高级功能。其特点是高度通用性和灵活性，能够快速适配多种应用场景，如智能客服、内容创作、知识管理等。通过与平台其他模块的协同工作，大模型应用插件进一步提升了智能体的智能化水平，满足用户对复杂任务的多样化需求。

4.1.2.1. 文案生成

利用检索增强生成（RAG）技术的大模型系统能够精准地检索出与用户查询最为贴切的信息资源。通过这一系统，首先会从庞大的数据库中筛选出与咨询内容高度相关的资料片段，然后借助大模型卓越的自然语言处理能力对这些精选资料进行深度解析和智能整合。这样不仅能够创造出既准确又连贯的文章，还为用户的写作提供了坚实的支持。

4.2. 场景化能力

4.2.1. 智能问答

智能问答采用了基于知识库增强的预训练大模型技术，能够从海量数据中学习到丰富的知识，并且能够根据用户的意图和问题，进行智能化的回答和解决方案的提供。减少了人工回复的工作量，提高了效率。同时，智能问答还支持多轮对话和追问，能够更好地理解用户的意图和问题，提高答复的准确性和满意度。

4.2.1.1. 单轮/多轮对话

多轮对话是一种自然语言处理技术，它允许用户通过连续的提问和回答来获取更深入的信息。在智能客服中，多轮对话技术可以用来实现更复杂的问题回答和解决方案的提供。

通过多轮对话技术，智能客服能够根据用户的提问和回答，进行多轮次的交流和追问，以更深入地了解用户的需求和问题。这种技术能够提高答复的准确性和满意度，减少用户的疑问和不满。

多轮对话技术的应用场景非常广泛，包括但不限于电商、金融、教育、医疗等领域。在电商领域，多轮对话技术可以用来实现购物咨询和售后服务；在金融领域，可以用来实现投资咨询和贷款申请；在教育领域，可以用来实现学习指导和课程咨询；在医疗领域，可以用来实现健康咨询和病情诊断。

4.2.1.1.1. 问题改写

问题改写是指在对用户提出的问题进行理解和分析的基础上，将其重新表述为更为清晰、具体或者更适合系统处理的形式。这一过程涉及对用户原始问题的语法、词汇和语义进行调整或优化，以便更准确地捕捉用户的真实需求。问题改写是实现高效信息检索和智能问答系统的关键步骤之一，因为它能够提升系统对用户问题的理解精度，从而返回更加贴合用户需求的答案。

4.2.1.1.2. 上下文理解

上下文理解是指在对用户表达的上下文信息进行语义分析,并理解上下文之间的内在逻辑,实现用户精准诉求的识别

4.2.1.1.3. 意图识别

意图识别是指在多轮对话中，识别出用户提出问题的目的或意图。需要对用户的输入进行分析，以确定他们想要获取的信息类型。例如，用户可能会询问天气、查询新闻事件或者寻求建议等。意图识别是实现智能对话系统的关键步骤之一，因为它可以帮助系统更好地理解用户的需求，从而提供更准确的回答。

4.2.1.1.4. 情感识别

情感识别是指在多轮对话中，识别出用户表达的情感倾向。包括识别用户的情绪（如愤怒、高兴、悲伤等）以及他们对某个话题的态度（如积极、消极、中立等）。情感识别可以帮助系统更好地理解用户的情感需求，从而提供更贴心的服务。例如，当用户表示愤怒时，系统可以采取相应的措施来安抚用户的情绪。

4.2.1.1.5. 文本内容归纳

基于大模型的文本机器人在服务过程中，对于长文本内容，或知识详情，结合服务场景，形成归纳信息，通过简要文本信息对长文本信息的所表达的意思进行阐述，最终回答用户提出的问题

4.2.1.1.6. 追问

追问是指在多轮对话中，根据用户的输入和上下文信息，向用户提供更详细的信息或者引导用户提供更多的信息。需要对用户的输入进行深入分析，以确定他们可能感兴趣的细节或者

需要补充的信息。追问可以帮助系统更好地满足用户的需求，从而提高对话的效果。

4.2.1.1.7. 参考来源

参考来源是指在多轮对话中，为用户提供相关信息的来源。包括引用权威的数据、引用专家的观点或者提供相关的链接等。参考来源可以帮助用户更好地了解问题的背景和相关信息，从而提高对话的质量。同时，参考来源也可以提高系统的可信度，增加用户对系统的信任度。

4.2.1.1.8. 敏感信息过滤

基于大模型的文本机器人对回复的内容进行安全审查，包括回复内容是否存在误导、是否有诉求支撑、是否合规，审核通过后，再对外展示。

4.2.1.2. 知识库问答

知识库问答是一种基于知识库的问答系统，旨在通过查询知识库中的信息来回答用户提出的问题。这种系统包括问题抽取和知识匹配两个环节，知识匹配又可以分为知识对匹配和文本匹配两类。

4.2.1.2.1. 问题抽取

问题抽取是指在对话过程中，从用户的输入中提取出问题的关键信息。需要对用户的输入进行语义分析，以确定他们想要获取的信息类型。例如，用户可能会询问某个概念的定义、某个事件的经过或者某个事物的属性等。问题抽取是实现知识库问答系统的关键步骤之一，因为它可以帮助系统更好地理解用户的需求，从而提供更准确的回答。

4.2.1.2.2. 知识对匹配

知识对匹配是指在知识库中查找与用户提出的问题相匹配的答案。需要将问题的关键信息与知识库中的知识对进行比较，以确定最合适的答案。知识对匹配可以采用多种方法，如基于关键词匹配、基于语义相似度匹配等。在实际应用中，通常会结合多种方法来提高匹配的准确性。

4.2.1.2.3. 文本匹配

文本匹配是指在知识库中查找与用户提出的问题相关的文档或文档块。需要将问题的关键信息与知识库中的文本或文本块进行比较，以确定最合适的文本。文档匹配可以采用多种方法，如基于关键词匹配、基于语义相似度匹配等。在实际应用中，通常会结合多种方法同时使用，提高匹配的准确性。

4.2.1.2.4. 答案生成

答案生成是指大模型根据知识对匹配或文本匹配的结果，回答用户提出的问题。需要将匹配的结果与用户的问题进行提示词拼装。在实际应用中，提示词的编写会大幅影响答案生成的结果，需要设计合适的提示词答案生成。

4.2.2. 文本分析

大模型能够自动分类文本、提取关键要素、生成精炼的摘要，并通过聚类发现文本中的潜在模式，极大地提高了数据处理的效率和准确性。此外，还能检测并纠正文本中的错误，为长篇文章、对话内容以及问答交互提供总结，使信息更易于理解和检索。

4.2.2.1. 文本分类

大模型在文本分类方面的能力体现在其对语言理解的深度和广度上，它们能够有效地处理和分析大量文本数据，从而识别和分类各种类型的信息。这些模型通过训练学习海量的文本数据，掌握语言的结构、语义和上下文关系，从而能够在不同的应用场景中实现高效的文本分类任务。

4.2.2.1.1. 应用场景及优势

业务文档梳理：如公司发文文档按发文单位、文章内容属性等类别进行划分，在公司每天产出大量文章内容的情况下，借助文本分类能力，能快速精准地将不同主题的公司发文归位，方便后续编辑、推送及检索查阅。

客户反馈整理：企业在日常运营过程中会收到来自众多客户的反馈信息，这些反馈形式多样，涵盖邮件、在线表单留言、客服对话记录等文本内容。文本分类能力可以依据反馈的主题，像是产品功能建议、服务质量投诉、使用体验问题等类别对其进行细致划分。这样一来，相关部门能够迅速聚焦于特定类型的客户诉求，针对性地制定解决方案。例如，产品研发团队可以专门查看关于产品功能建议的分类内容，从中汲取灵感来优化和迭代产品；客服部门则能重点关注服务质量投诉方面的反馈，及时改进服务流程，提升客户满意度。

4.2.2.1.2. 技术介绍

特征提取：通过自然语言处理技术，如词嵌入和句嵌入，将文本转换为机器能够理解的数值特征向量。

分类模型训练：基于 BERT 模型，首先对其进行预训练模型的加载，利用它在大规模语料上学习到的通用语义表示能力作为坚实基础。将之前通过自然语言处理技术提取的数值特征

向量，适配性地输入 BERT 架构中，使其融入 BERT 特有的多头注意力机制下的上下文感知环境。

4.2.2.2. 要素提取

大语言模型的要素提取能力是指从给定的文本中识别和提取出关键信息的能力，如人名、地点、日期、组织等，以及它们之间的关系。这对于信息检索、知识图谱的构建等任务至关重要。这对于许多自然语言处理任务至关重要，如信息检索、文本摘要、关系提取等。

4.2.2.2.1. 应用场景及优势

信息提取：从大量非结构化文本中快速准确地提取所需信息，如新闻报道中的事件、时间、地点等，极大地提高了信息获取效率。

知识图谱构建：通过从各种文本源提取实体、关系等要素，并将其组织成结构化的知识库，为智能问答、推理等任务提供知识支持。

舆情分析：从社交媒体、新闻等渠道提取关键观点、情感倾向等信息，洞悉公众对特定事件或主题的看法和态度。

风险监控：从监管文件、合同等法律文书中提取关键条款、风险点，协助企业及时发现并规避潜在风险。

客户服务：从客户询问、投诉等文本提取出用户需求、问题症结等核心内容，指导客服人员快速高效地解答疑问。

4.2.2.2.2. 技术介绍

语言大模型的要素提取能力源于其对自然语言的深度理解和建模。通过预训练吸收了海量文本数据，模型能够捕捉到词语、短语之间的语义关联，从而在全文层面上识别和抽取关键信息单元。此外，结合注意力机制等技术，模型可以自动学习到哪些信息对于特定任务更为重要。

4.2.2.3. 文本摘要

在当今信息爆炸的时代，文本摘要技术已成为处理和理解大量文本资料的关键工具。大模型采用最先进的自然语言处理技术，提供高效、准确的文本摘要能力，以帮助用户快速获取信息的精髓。

4.2.2.3.1. 抽取式摘要

抽取式摘要通过直接从原文中选取关键句子或短语来构建摘要。这种方法保留了原文的精确表达，适用于需要高度准确性和原文保真度的场景。

4.2.2.3.1.1. 应用场景及优势：

法律文档摘要：在法律领域，对案件报告、判决书等文档的摘要需要高度的准确性和原文的保留，抽取式摘要能够确保关键法律用语和表述的准确无误。

科学研究摘要：对于科学论文和研究报告，抽取式摘要可以快速提供研究问题、方法、结果的直接摘录，帮助研究人员迅速把握论文核心。

新闻摘要：在新闻编辑和汇编过程中，抽取式摘要可以保证关键信息的准确传达，避免误解或偏离原意。

4.2.2.3.1.2. 技术介绍

文本预处理：包括对文本进行分词处理，将连续的文本流切分成一个个有意义的词语单元，像中文文本可借助专业的分词工具如 jieba 分词，精准地把句子拆分开，便于后续分析。同时，去除停用词，诸如常见的 “的”“是”“在” 等对语义表达贡献较小的词汇，减少文本噪声干扰，聚焦关键信息。

句子重要性评估：使用词频 - 逆文档频率（TF-IDF）算法，通过统计每个词在当前文档中的出现频率以及在整个语料库中的稀有程度，来衡量词语的重要性，进而推算出包含这些重要词语较多的句子具有更高的抽取价值；位置特征也不容忽视，一般而言，位于文本开头、结尾以及段落开头的句子往往承载着关键内容，会被赋予更高的优先级；还有基于语义相似性的评估手段，利用词向量模型，计算不同句子间的语义相似度，将与多篇文本主题句相似度高的句子视为重点抽取对象，避免重复抽取语义相近的冗余句子。

句子排列组合生成：按照句子在原文中的出现顺序排列，维持文本原本的叙述脉络，让读者能顺着熟悉的逻辑理解摘要；或者根据句子间的语义关联度重新组织，利用语义图等技术构建句子间的连接关系，将紧密相关的句子挨在一起，增强摘要的可读性，确保抽取式摘要能以最精炼、精准的形式呈现原文的精华内容，满足不同领域对信息快速准确获取的迫切需求。

4.2.2.3.2. 生成式摘要

生成式摘要通过理解原文的整体内容和结构，自行生成一个连贯、概括性强的新文本作为摘要。这种方法能够产生更加自然、流畅的摘要文本，适用于需要更高级别概括和语言生成能力的场景。

4.2.2.3.2.1. 应用场景及优势：

商业报告总结：生成式摘要可以在处理商业报告和市场分析时，不仅提取关键信息，还能

以流畅的语言重新表达报告的主要观点和结论，便于阅读和理解。

教材内容概括：教育领域中，对课程内容和教材的摘要需要更加通俗易懂，生成式摘要能够以学生易于理解的方式重新表达教材要点。

个性化新闻服务：为了满足用户个性化阅读需求，生成式摘要能够根据用户的阅读偏好，提供更加定制化、有吸引力的新闻摘要。

4.2.2.3.2.2. 技术介绍：

预制摘要模板：通过预制摘要中需要包含的主要信息点，作为输入大模型的提示词，让大模型根据提示词模板进行摘要输出，可以提高摘要内容的可读性和一致性。

4.2.2.4. 文本聚类

文本聚类技术利用深度学习模型来识别文本中的语义模式，进而将具有相似意义的文本分配到同一类别中。这一过程无需预先定义分类标准，模型会自动发现数据中的结构。通过这种方式，大模型能够处理和组织海量的文本数据，使其更加易于访问、分析和理解。

4.2.2.4.1. 应用场景

客户反馈分析：企业可以使用文本聚类技术自动将客户反馈分为不同的类别（如产品问题、服务体验、功能请求等），帮助企业快速识别客户关注的热点问题，并做出相应的改进。

内容推荐系统：在媒体和娱乐行业，文本聚类能力可以帮助识别用户的阅读或观看偏好，通过聚类相似内容，为用户推荐他们可能感兴趣的文章、视频或产品。

信息检索和组织：在大型知识库或文档管理系统中，文本聚类可以帮助自动分类和组织信息，提高搜索效率和准确性，让用户更快地找到所需信息。

市场趋势分析：通过分析社交媒体、新闻报道和在线论坛中的文本内容，企业可以利用文本聚类技术识别行业趋势、公众情绪变化或竞争对手动态，为战略决策提供支持。

科学研究：在学术领域，研究人员可以利用文本聚类技术对文献资料进行分类，以发现新的研究方向或理论联系，加速知识的积累和创新。

4.2.2.4.2. 技术介绍

特征提取：通过自然语言处理技术，如词嵌入和句嵌入，将文本转换为机器能够理解的数值特征向量。

聚类算法：应用各种统计学和机器学习算法，如 K-均值、层次聚类和 DBSCAN 等，根据特征向量的相似度将文本分组。

优化和调整：使用各种技术对聚类结果进行评估和优化，例如轮廓系数和调优聚类算法的参数，以实现更精准的文本分类。

4.2.2.5. 智能校对

语言大模型在错误检测与纠正方面利用先进的自然语言处理技术，能够识别并修正文本中的语法、拼写以及语义错误。通过深度学习的算法，这些模型能够理解上下文信息，从而提供更准确、自然地修正建议。这不仅提高了写作质量，也大幅度节省了用户在文本校对上的时间与精力。

4.2.2.5.1. 应用场景

教育领域：在线教育平台可以利用语言大模型来提供学生作业的自动校对服务，帮助学生改进写作技能。

内容创作：内容创作者、博客作者和新闻编辑可以使用此技术快速检查并修正文章中的错误，保证发布内容的质量。

商务通信：企业可以在其内部通讯（如电子邮件、报告、提案等）中集成错误检测与纠正功能，以确保对外文档的专业性和准确性。

软件开发：开发工具可以集成语言模型来提供注释和文档的错误校正，帮助开发者提高代码的可读性和维护性。

4.2.2.5.2. 技术介绍

深度学习模型：错误检测与纠正功能基于复杂的深度学习网络，如 Transformer 模型，这些模型通过大规模语料库学习语言规律。

上下文理解能力：通过长短期记忆或 Transformer 架构，模型能够理解单词、短语和句子在具体上下文中的含义，提高纠错的准确性。

自适应学习：随着模型不断地被使用，它们可以通过用户的反馈和更多的文本数据进行自我优化，以更好地适应特定领域的语言习惯和风格。

多语言支持：高级模型支持多种语言的错误检测与纠正，能够跨文化背景服务全球用户，满足不同地区的需求。

4.2.2.5.3. 技术优势

高效率：语言模型能够快速识别并修正文本中的错误，显著提高了编辑和校对的效率。

高准确性：借助深度学习技术，这些模型在理解语境和修正错误方面的表现通常超过传统

的基于规则的校对工具。

强大的适应性：模型能够学习和适应特定的语言风格和术语，这使它们在专业领域（如法律、医学和科技）中特别有用。

多语言和多方言支持：高级语言模型能够支持多种语言和方言，为全球用户提供服务。

持续学习和改进：通过机器学习，这些模型可以持续从新的数据中学习，随着时间的推移不断提高其性能。

4.2.2.6. 文章总结

文章总结功能利用先进的自然语言处理技术，自动识别并提取文本内容的关键信息，以简洁明了的方式重新组织和呈现给用户。这一功能通过理解文本的主要观点、论据和结论，帮助用户快速把握文章的核心内容，无需阅读全文。

4.2.2.6.1. 应用场景

新闻阅读：帮助用户快速了解新闻要点，特别是在时间紧迫时。

学术研究：快速提取学术文章的主要发现和结论，辅助研究人员高效阅读。

商业报告分析：在企业决策过程中快速获取报告关键信息，提高决策效率。

书籍摘要：为读者提供书籍章节的摘要，帮助他们决定是否深入阅读。

4.2.2.6.2. 技术介绍

自然语言理解：通过深度学习模型理解文本语义，捕捉关键信息。

文本生成：利用语言模型根据理解的内容生成凝练、流畅的总结。

注意力机制：高效识别文本中的重要部分，确保摘要的准确性和相关性。

4.2.2.6.3. 技术优势

效率提升：大幅缩短获取信息的时间，提高阅读或研究的效率。

易于理解：简化复杂内容，使非专业人士也能快速理解。

灵活应用：广泛适用于各种文本类型和行业领域。

4.2.2.7. 观点提炼

观点提炼是指利用自然语言处理技术，尤其是大语言模型的深度语义理解能力，从给定的文本中识别并抽取出用户表达的核心观点、态度倾向、主要诉求等信息。该能力适用于意见反馈、社交评论、客户来信、调研问卷等多类场景，旨在从复杂的、非结构化文本中提炼出有价值的信息观点，帮助用户快速把握对话的核心内容。

在实际应用中，大模型首先对文本内容进行语义分层处理，识别句子之间的逻辑关系和情感色彩，进而抽取出表达明确立场或倾向性的句段，归纳其所涉及的观点主题，可用于如满意度、服务质量、政策诉求、使用体验、改进建议等。

观点提炼不仅关注显性表达，还能够识别出隐含态度和暗示性意见，例如通过比喻、反问等方式间接传达的观点。此外，系统还可对提炼出的观点进行分类标注，如正面、中立、负面，辅助业务部门实现舆情监测、客户画像、风险预警等分析任务。通过对大量文本数据的观点提炼与聚合，组织可以洞察用户关注焦点和变化趋势，为策略调整、产品优化、服务改进等提供数据支持与决策依据。

4.2.2.7.1. 应用场景

会议记录总结：自动提取会议讨论的关键点和决策，为参会者提供会议摘要。

客户服务记录：总结客户服务对话，快速提取客户问题和解决方案，改善服务质量。

教育与培训：对教学或培训过程中的交流进行总结，便于学习者复习关键点。

健康咨询记录：为医疗健康领域提供对话总结，帮助医生和患者快速回顾咨询内容。

4.2.2.7.2. 技术介绍

自然语言理解：利用 NLU 技术理解对话的语义，识别重要的信息和关键词。

文本摘要生成技术：应用抽取式和生成式摘要技术，准确生成对话的核心内容摘要。

深度学习模型：使用先进的深度学习模型，如 Transformer 架构，提高总结的准确度和流畅度。

语境理解和连贯性分析：通过分析对话中的上下文信息和语境，确保总结内容的连贯性和逻辑性。

4.2.2.7.3. 技术优势

效率提升：自动化处理大量对话内容，节省时间和人力资源。

信息准确：通过技术手段准确抓取关键信息，降低人为遗漏或误解的风险。

易于整合：可以轻松集成到现有的聊天平台、会议记录工具中，提升用户体验。

4.2.2.8. 知识对提取

知识对提取功能利用语言模型的深度学习技术，自动识别和提取文本资料中的问题与对应答案。该功能通过理解文本的语义结构，能够有效地从各类文档、网页或对话记录中检索出有价值的问答信息，为用户提供快速准确的信息获取路径。

4.2.2.8.1. 应用场景

客服支持：自动提取常见问题及解答，优化知识库的构建和更新，提高客户服务效率。

教育领域：从教材和学术文章中提取知识对，辅助教学和学习，加深学生对知识点的理解。

法律咨询：快速从大量法律文件中提取相关问答，简化法律咨询和案例研究过程。

研究分析：从研究报告或数据集中提取关键问答，加速信息梳理和分析工作。

4.2.2.8.2. 技术介绍

自然语言处理：利用 NLP 技术解析文本结构，理解问题与答案的语义关联。

深度学习模型：通过训练大量文本数据，模型能够识别和预测知识对的模式。

上下文理解：技术包括对话理解能力，即在提取知识对时考虑上下文信息，确保答案的准确性和相关性。

4.2.2.8.3. 技术优势

效率提升：自动化处理文本，大幅减少人工筛选和提取信息的时间成本。

准确性：利用先进的 NLP 技术，提高了信息提取的准确度和可靠性。

灵活性：适用于多种类型的文本资料，能够灵活应对不同的信息提取需求。

4.2.3. 文本生成

基于检索增强的大模型技术，通过多轮对话、多路召回等技术手段，检索出与客户咨询最相关的内容。基于大模型对相关内容的理解分析，生成高质量的文本，输出通顺连贯的 Word 文档，为用户提供有力的写作支持。这确保了文案的准确性和相关性，帮助用户更快地完成任务，从而提高文案编写的效率。

4.2.3.1. 文本优化

文本润色是指大模型利用其先进的语言处理能力和对召回的相关内容的理解，对用户提供的文本进行优化和改进，主要包括文本扩写、润色、精简功能。通过这一过程，确保最终输出的文本更加清晰、准确、流畅，并且契合用户的需求或使用场合的要求。

1. 文本扩写

文本扩写则是针对已有内容进行扩展，通过增加细节描述、背景介绍或相关论据等手段使原始信息更加丰富全面。该过程注重逻辑连贯性和内容的相关性，旨在为读者提供更详尽的信息支持，帮助深入理解主题内容。文本扩写不仅能够加深读者的认知，还能够增强文本的表现力和说服力，适用于需要详细解释或广泛讨论的场合。

2. 文本润色

文本润色是指大模型利用其先进的语言处理能力和对召回的相关内容的理解，对用户提供的文本进行优化和改进。通过这一过程，确保最终输出的文本更加清晰、准确、流畅，并且契合用户的需求或使用场合的要求。

文本润色提供专业的编辑建议。它不仅能加速内容创作流程，还能确保内容的质量和创新性。通过对词汇选择、句子结构和段落组织的精心打磨，润色后的文本更具深度和层次感，从而吸引更多的读者关注并参与讨论。

3. 文本精简

文本精简是指在保持原文核心意义和信息量的前提下，通过删除冗余表达、简化复杂结构等方式对文本进行压缩。这一过程强调精准与高效，确保处理后的文本更加简洁明了，易于理解，同时不失其原意的完整性和深度。文本精简对于提升阅读体验、优化文档管理及适应特定格式要求具有重要意义。

4.2.3.1.1. 应用场景

营销文案生成：借助文本润色的强大能力，优化广告文案、社交媒体帖子及营销电子邮件，通过精细调整语言风格、逻辑结构和情感表达，帮助品牌精准触达目标受众，显著提升市场响应率和品牌影响力。

内容创作辅助：在博客文章、新闻稿和其他出版物的创作过程中，文本润色提供专业的编辑建议。它不仅能加速内容创作流程，还能确保内容的质量和创新性。通过对词汇选择、句子结构和段落组织的精心打磨，润色后的文本更具深度和层次感，从而吸引更多的读者关注并参与讨论。

自动化客户服务：通过文本润色生成的客户服务回复不仅自然且友好，还能够根据客户的具体问题和情感状态进行个性化调整。这不仅提升了客户满意度，还提高了服务效率，使企业能够更迅速有效地回应客户需求。

4.2.3.1.2. 技术介绍

自然语言处理（NLP）：借鉴人类大脑处理信息的方式，通过构建多层神经网络架构来解析和生成数据，使计算机具备理解、解释以及生成人类自然语言的能力，从而实现与人类的高效沟通。

交互式生成：支持持续对话，逐步完善生成的内容，提高准确性和相关性。接收用户即时反馈，即时调整生成策略以满足用户期望。

4.2.3.1.3. 技术优势

创意增强：在保持原意的同时，引入更具创意的表达方式，使文本更加生动有趣，有效吸引读者并激发兴趣。

语法和拼写纠错：自动检测并纠正文本错误，确保输出内容的准确性与专业性。

4.2.3.2. 提纲生成

提纲生成技术能够根据给定的文本内容自动生成结构化的提纲。通过深入理解文本的主旨和关键点，高效提炼核心信息，形成条理清晰、逻辑严谨的提纲。这一过程不仅简化了信息整理的工作，还确保了提纲的准确性和完整性，极大地方便了后续的内容创作和编辑工作。

4.2.3.2.1. 应用场景

学术研究：加速文献整理，自动生成研究提纲，显著提高文献复习效率。

内容创作：提供文章和书籍的基础框架，促进创意思考并优化内容结构。

商业报告：快速整理会议纪要、市场分析及项目计划，提升企业员工的工作效率。

教育领域：辅助教师和学生规划课程、讲义及学术论文，优化教学与学习过程。

4.2.3.2.2. 技术介绍

强化学习：利用奖励函数指导模型，优化文本生成的质量，涵盖语法正确性、流畅度和相关性等方面。

语境理解与生成：借助如 Transformer 等先进架构，实现上下文的理解，确保话题连贯及风格一致。

4.2.3.2.3. 技术优势

效率提升：大大减少人工整理提纲的时间和劳动强度，提高工作和学习效率。

准确性：通过大数据分析和深度学习技术，能够精准捕捉文本的核心要素和结构。

灵活性：适用于多种文本类型和领域，能够根据不同的需求定制提纲结构。

4.2.3.3. 文案写作

利用大模型技术，文案写作能够生成逻辑严密、连贯流畅的文本内容。该工具不仅理解输入文本的上下文，还能在此基础上创作出相关的新内容，助力用户快速且高效地完成写作任务，显著提升内容创作的速度与质量。

4.2.3.3.1. 应用场景

内容创作：自动生成文章、报告、故事或任何其他文本内容，帮助内容创作者克服创作障碍，提高写作效率。

自动回复：在客户服务和社交媒体管理中自动生成响应，提高响应速度和质量。

教育与培训：为学生提供定制化的学习材料和作业帮助，增强学习体验。

数据分析报告：根据数据集自动生成分析报告，帮助分析师快速理解和传达关键信息。

4.2.3.3.2. 技术介绍

迁移学习：迁移学习使得模型能够高效适应特定应用场景和专业领域，即使在数据有限的情况下也能保持出色的性能。

预训练和微调：通过在海量文本数据上进行无监督或弱监督学习，模型可以获得广泛的语言知识和模式识别能力。针对具体应用场景，在预训练模型的基础上使用标注数据进一步优化模型性能，使其更好地适应特定任务。

4.2.3.3.3. 技术优势

个性化：依据具体需求生成定制化文案，确保信息精准触达目标受众。

节省成本：相较于聘请专业写手或文案团队，文案写作大幅降低了人力成本，特别适合大批量重复性文案的生产。

数据驱动：文案写作可以通过分析大量的历史数据来优化生成效果，选择最有可能引起共鸣的话语和表达方式。

4.2.4. 智能问数

利用大模型与数据库的结合，大模型把用户输入的内容转成 SQL 代码，进而查找数据宽表相关数据，返回数值或生成简单的图表，并推送到用户页面。通过用户输入的文本找到相关数据、图形化展示数据、解释数据。

4.2.4.1. 数据库接入

支持连接用户选择已创建连接的数据库，搭建智能问数应用。

4.2.4.2. 知识名词解释

支持用户输入与业务强相关的知识名词解释，补足大模型对于特定业务领域知识的短缺。可以输入固定缩写解释，专有名词解释等。

4.2.4.3. 指代消歧

指代消歧是一种自然语言处理技术，用于解决文本中指代词的具体含义。通过分析上下文语义和句法结构，指代消歧技术可以准确确定指代词所指代的对象或实体，从而增强大模型对语言的理解能力。关键特点包括上下文分析及准确定位，用于自动理解指代词在当前文本语境中的意义及结合句法规则和语义模型，定位指代词对应的具体实体。

4.2.4.3.1. 应用场景

日期指代消歧：智能理解用户输入中模糊的时间表达，如“下周五”或“前天”，并结合当前时间和上下文，准确解析为具体的日期。

模糊实体指代消歧：解析文本中模糊的指代对象，例如“它”“这个”等词汇，通过分析上下文确定指向的具体实体，如设备、产品或文档，

模糊意图指代消歧：在模糊表达意图的语句中，如“我想知道”或“我想问”，智能识别用户的具体需求。

4.2.4.3.2. 技术介绍

上下文语义分析：通过自然语言处理技术，深入理解文本中的上下文信息，分析句法和语义结构，以准确判断指代词的含义，为指代消歧提供语义支持。

指代规则与知识库：利用指代规则和预先构建的知识库（如实体关系、时间推断等），结合逻辑推理机制，帮助机器准确匹配模糊指代与具体实体或日期。

共指解析技术：采用共指解析算法，检测文本中是否有多个指代词描述同一对象，整合信息以确保一致性，常用于长文档理解与多轮对话分析。

4.2.4.3.3. 技术优势

提升理解能力：指代消歧技术通过准确解析模糊指代，大幅提升大模型对自然语言的理解能力，使得复杂语义场景的处理更加高效可靠。

增强用户体验：通过消除用户与系统之间的语义歧义，确保大模型能够准确响应用户需求，显著提高智能问数人机交互场景的流畅性和满意度。

降低人工干预：自动化处理模糊指代，大幅减少人工参与，既降低了时间成本，又提高了系统的自动化水平和运行效率。

4.2.4.4. 自动库表和字段选择

在智能问数过程中，自动库表和字段选择是一项关键技术，用于根据语义自动识别数据库

中的表和字段。通过分析用户查询的上下文语义及数据库模式，精准匹配查询意图和对应字段，显著提升 SQL 生成的准确性。关键特点包括上下文分析，结合数据库结构。前者结合用户查询上下文，大模型自动理解语义并定位到相关表和字段，后者通过语义分析和句法规则，将自然语言与数据库 Schema 对接，确保生成 SQL 时能准确选择相关实体。

4.2.4.4.1. 应用场景

模糊字段描述：用户输入“显示销售数据”，系统通过语义匹配将“销售数据”解析为数据库中的相关字段。

部分字段模糊：对于查询“每年的数据”，系统自动将“数据”解析为表中的核心指标字段。

动态字段选择：当查询涉及动态上下文（如“这个月的销售额”），结合时间上下文自动解析为具体时间范围并匹配对应的字段。

4.2.4.4.2. 技术介绍

语义扩展：利用词嵌入或同义词库，将自然语言中的关键词扩展为可能的字段名或表名。

模式分析：解析数据库的 Schema，包括表间关系、字段类型，帮助识别与查询语义最相关的数据库元素。

动态推理：结合查询上下文和用户意图，对模糊描述（如“数据”“记录”）进行逻辑推断，将其映射到具体的数据库字段。

4.2.4.4.3. 技术优势

提升查询准确性：自动库表和字段选择技术通过精准解析自然语言中的模糊表达，显著提升了对复杂数据库结构的理解能力，确保生成的 SQL 查询准确无误。

增强用户体验：通过智能分析用户查询的上下文和意图，消除语义歧义，系统能够更流畅地响应用户需求，提供更符合期望的查询结果，提高了智能问答和数据查询的互动体验。

减少人工干预：系统自动进行库表和字段的选择，减少了人工调整和干预的需求，不仅降低了人力成本，还提升了查询生成的效率和自动化水平，极大地提升了系统的可用性和可靠性。

4.2.4.5. 文本转 sql

文本转 SQL 是指将自然语言形式的查询语句转换为可执行的 SQL 语句。需要使用自然语言处理技术和数据库查询语言转换技术来实现。在实际应用中，文本转 SQL 可以采用多种方法，如基于规则的方法、基于模板的方法等。这项技术通过理解自然语言中的意图和上下文，自动生成对应的 SQL 查询，从而为非技术用户提供了一个直观、易用的数据查询接口。

4.2.4.5.1. 应用场景

报表生成：允许非技术背景的用户通过简单的自然语言描述来创建复杂的数据报表。

数据分析：分析师可以直接使用自然语言来查询数据，加速数据分析过程。

客户支持系统：在客户支持系统中，文本转 SQL 功能可以帮助快速检索客户信息或相关数据，提高服务效率。

教育和培训：作为一个教学工具，帮助学生理解 SQL 语法和数据查询的基础。

4.2.4.5.2. 技术介绍

自然语言处理：使用先进的 NLP 技术来理解用户的查询意图和上下文。

深度学习模型：运用深度学习算法，如 Transformer 和 BERT，训练模型以理解和生成 SQL 查询。

意图识别与槽位填充：识别用户查询中的关键信息（如查询目标、条件等），并将其映射到相应的 SQL 结构上。

适应性学习：通过持续学习用户的查询习惯和偏好，不断优化查询结果的准确性和效率。

4.2.4.5.3. 技术优势

用户友好：为非技术用户提供了一个易于使用的查询接口，降低了数据检索的门槛。

提高效率：自动化的查询生成可以显著提高数据检索的速度和准确性。

灵活性：可以根据不同的数据库结构和查询需求进行定制和优化。

5. 标杆案例

5.1. 物流行业

物流大脑项目目标为物流集团构建一个基于 AI 大模型技术的企业智能体，旨在通过汇聚集团及其产业链上下游的海量数据，引入 DeepSeek 大模型强大的语言能力，结合先进的技术进行分类整理、融合集成以及深度挖掘，构建行业规范化的知识体系，围绕人、车、货、场、单等多个业务领域重塑物流产业应用的新范式，成为物流信息与智能应用的超级入口。

1) 客户需求

- 整合物流集团及上下游企业的多维度数据，建立数据枢纽，实现物流全链条的数据贯通与共享，支撑业务协同与监管决策。
- 建立覆盖物流产业应用场景的智能交互入口，通过标准化知识体系与可复用的算法模型，提升行业整体运营效率。

2) 解决方案

- 精准知识匹配：整合物流集团及上下游企业的多维度数据，构建物流行业知识库。
- 智能交互：借助 DeepSeek 强大的语言能力，结合知识库功能，实现精准理解物流行业复杂的专业术语和业务场景描述。
- 定制化分析报告：针对物流集团的宏观研究需求，根据历史数据市场趋势生成定制化的分析报告。

3) 核心价值

- 提升工作效率：员工遇到物流知识问题或需要研究资料时快速获取答案。
- 优化决策质量：研究助手提供深度分析和趋势预测，帮助集团管理层制定更科学的战略决策。
- 促进知识传承：将物流行业知识集中整合，缩短新员工适应期。
- 增强市场竞争力：集团能更敏锐地捕捉市场变化，提升企业竞争力。

5.2. 能源行业

央国企数字化转型是必然趋势，AI 大模型技术作为数字化转型的重要工具，能够显著提

升央国企的办公效率和管理水平。

1) 客户需求

- 改进工作流程：优化流程，减少低价值工作占用。
- 企业决策支持：通过数据分析等功能，增强员工决策能力。
- 灵活构建办公应用：具备自主开发和个性化定制功能。
- 系统集成与数据打通：灵活集成 OA、CRM、财务等异构系统，实现跨系统协作

2) 解决方案

- 通用办公产品：智慧办公应用产品服务，文档写作、会议助手、视频制作、PPT 生成、数字人等应用产品。
- 智能问数：协助员工完成数据分析功能，辅助决策
- 智能体构建：提供智能体构建引擎，方便快速搭建个性化智能体
- 多种接入方式：提供 api、RPA 等多种接入方式，自动获取数据，并实现准确、无侵入式操作

3) 核心价值

- 广泛的办公场景覆盖：涵盖人力资源、采购等 5+专业领域及通用办公领域，通过整合各种业务场景，提供一站式服务，实现全方位智能化办公体验。
- 有效提升员工效率：自动化常规任务和流程，显著提高员工工作效率。
- 满足企业个性化需求：方案适应不同企业业务复杂性和个性化需求，提供定制化解决方案。通过深度学习和模型训练，理解企业独特需求，提供相应功能和服务，

实现个性化体验。

- 简化系统集成架构：通过开放 API 接口、构建 RPA 机器人和标准化集成流程，简化与其他办公软件集成。

5.3. 医疗行业

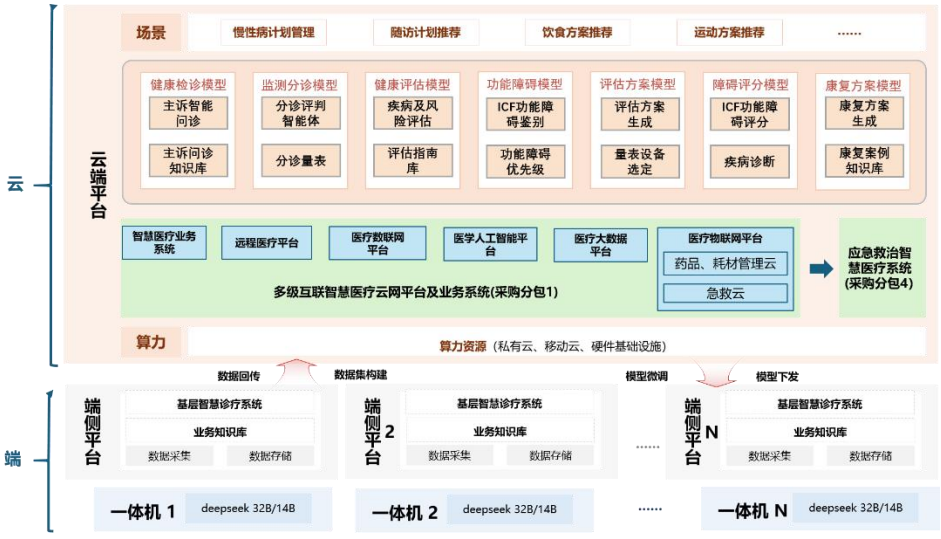
以基层创伤、运动伤和心理疾病的诊疗、评估、康复场景为核心打造基层诊疗大模型及智能体，有效弥补基层医疗资源不足，构建全新智慧医疗体系

1) 客户需求

- 医疗资源分布不均，偏远地区难享优质医疗；
- 基层医疗区域康复诊疗水平不高，诊疗效果欠佳；
- 优质医疗资源辐射不足；

2) 解决方案

- 多源高质量数据与大模型充分融合，协助地方医护人员提升诊疗水平；
- AI+物联集成，大模型多智能体设计，实现基层医疗资源有效补齐；
- 云边协同设计，实现全场景服务，快速部署至基层医院



3) 核心价值

- 打造心理疾病、创伤疾病、慢性病 3 类疾病行业大模型，实现分诊、诊断、评估、康复等 4 种智能体业务有效闭环。
- 首个医疗行业大模型+智能体项目，对接 24+款诊疗、康复设备，实现百万级 Token 高质量数据微调。
- 已在 301 医院及 12 个边远地区实现三室一中心部署，启动全医疗行业推广。

6. 产品部署要求

6.1. 部署资源需求

1）部署大模型（以部署 72B 大模型做评估）

大模型推理服务：根据实际需求选择 7B、9B、72B 等不同参数大模型，为使能平台提供模型服务。

一站式 AI 智能体开发平台：部署平台全部功能，包括智能体工坊、知识库等模块，为用户自建智能体提供支持

服务器类型		最小配置	推荐配置	备注
智算服务器	GPU 数量	4	8	4 张卡用于大语言模型部署，2 张卡用于向量化及 rerank 模型部署，2 张卡用于传统语音、视觉、文件解析服务
	GPU 类型	A800/910B		
	GPU 显存	80G(A800)/64G(910B)		
	CPU 架构	X86/ARM		
	CPU 核数	32		
	RAM 内存	128GB		
	网络地址数	1	2	
	硬盘	1TB		
	支持的操作系统	CentOS 7.9（x86）、Ubuntu 20.04（x86） CTyunOS 23.01（x86）		
	互联网访问能力	需求确定		
通算服务器	服务器数量	2		引擎与平台单独部署
	CPU 架构	X86/ARM/C86		
	CPU 核数	16	32	
	RAM 内存	64	128	
	网络地址数	1	2	
	硬盘	2*1TB	2*4TB	
	支持的操作系统	CentOS 7.9（x86 和 C86）、Ubuntu 20.04（x86 和 C86）、CTyunOS 23.01（x86 和 C86）、麒麟 V10 SP2（ARM）		其他类型操作系统，需要额外进行适配，需延长部署周期。
	互联网访问能力	需求确定		

2）已有大模型服务，使能平台接入

大模型推理服务：已有大模型服务通过 API 方式为使能平台提供模型服务。

一站式 AI 智能体开发平台：部署平台全部功能，包括智能体工坊、知识库等模块，为用户自建智能体提供支持。

服务器类型		最小配置	推荐配置	备注
智算服务	GPU 数量	1	3	用于部署向量化及

器	GPU 类型	4090/A10/V100		rerank 模型部署或传统语音、视觉、文件解析服务
	GPU 显存	24G/32G		
	CPU 架构	X86		
	CPU 核数	32		
	RAM 内存	128GB		
	网络地址数	1	2	
	硬盘	1TB		
	支持的操作系统	CentOS 7.9（x86）、Ubuntu 20.04（x86 和 ARM）CTyunOS 23.01（x86）		
	互联网访问能力	需求确定		
通算服务器	服务器数量	2		引擎与平台单独部署
	CPU 架构	X86/ARM/C86		
	CPU 核数	16	32	
	RAM 内存	64	128	
	网络地址数	1	2	
	硬盘	2*1TB	2*4TB	
	支持的操作系统	CentOS 7.9（x86 和 C86）、Ubuntu 20.04（x86 和 C86）、CTyunOS 23.01（x86 和 C86）、麒麟 V10 SP2（ARM）		其他类型操作系统，需要额外进行适配，需延长部署周期。
	互联网访问能力	需求确定		

6.2. 交付清单

应用平台：一站式 AI 智能体开发平台

配套材料：平台部署安装包（platform.tar 、backend.tar、model.tar）、私有化部署手册、兼容性清单、操作手册