元脑企智 EPAI 1.1 版

产品白皮书

软件版本：1.1 版

文档版本：V1.0

发布日期：2024 年 09 月 30 日

浪潮电子信息产业股份有限公司

IEIT SYSTEMS CO.,LTD

安全声明

公司产品不会主动获取或使用用户的个人数据，仅在您同意使用特定功能或服务时，在业务运营或故障定位的过程中可能会获取或使用用户的某些个人数据（如告警邮件接收地址、IP 地址），公司产品在涉及个人数据的收集、存储、使用、传输、删除等全生命周期的处理活动中，已在产品功能上部署了必要的安全保护措施，同时，您也有义务根据所适用国家或地区的法律法规制定必要的用户隐私政策并采取足够的措施以确保用户的个人数据受到充分的保护。

浪潮信息高度重视产品数据安全，公司产品在涉及系统运行和安全数据的全生命周期处理活动中，已严格按照相关法律法规及监管要求，在产品功能上部署了必要的安全保护措施。作为系统运行和安全数据处理者，您有义务根据所适用国家或地区的法律法规制定必要的数据安全政策并采取足够的措施以确保系统运行和安全数据受到充分的保护。

浪潮信息将一如既往的严密关注产品与解决方案的安全性，为客户提供更满意的服务。浪潮信息已全面建立产品安全漏洞应急和处理机制，确保第一时间处理产品安全问题。若您在本产品使用过程中发现任何安全问题，或者寻求有关产品安全漏洞的必要支持，请直接联系浪潮信息客户服务人员。

内容声明

您购买的产品、服务或特性应受浪潮信息商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，浪潮信息对本文档内容不做任何明示或默示的声明或保证。文档中的示意图与产品实物可能有差别，请以实物为准。

本文档仅作为使用指导，不对使用我们产品之前、期间或之后发生的任何损害负责。文档所含内容如有升级或更新，恕不另行通知。

目　录

1. 产品介绍

1.1 产品简介

元脑企智 EPAI（Enterprise Platform of AI，简称 EPAI）是企业级大模型开发平台软件，提供完整的模型服务工具和全链路应用开发套件，预置丰富的能力插件，提供 API 及 SDK 等便捷的集成方式，支持高效完成大模型应用构建。同时，它支持调度多元算力和多模算法，可以帮助企业高效开发部署生成式 AI 应用，打造智能生产力。

面向对象：企业开发者及 ISV 的技术人员、领域专家。

核心能力：支持开箱即用的对话应用调用，大模型微调微调和一站式模型灵活部署。

服务形式：支持平台内使用 API 服务，也支持通过 API 服务输出给客户，方便客户进行集成和使用专属大模型能力。

知识增强：自动化的 RAG pipeline，支持多种检索方式，可将最新的知识融入大模型。

应用编排：支持客户自行选择使用的大模型、配置知识库、插件，定制化开发场景应用。

支持业务：支持企业从企业场景需求(Requirements)出发到场景应用开发和部署的实现。

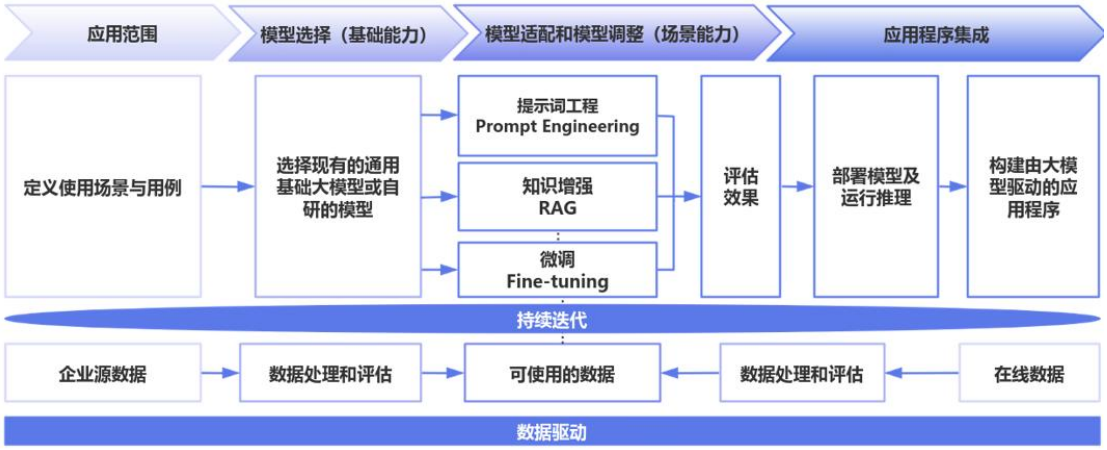

图 1-1 EPAI 平台支持应用开发全流程

1.2 产品架构

产品集成了多模多元计算框架、算力调度服务、模型训推服务工具，提供模型微调工具和应用开发工具，模型微调工具支持数据准备、模型微调训练、到模型评估和部署的全流程，应用开发工具支持用户自建知识库、选用插件增强模型的能力，定制化场景应用。

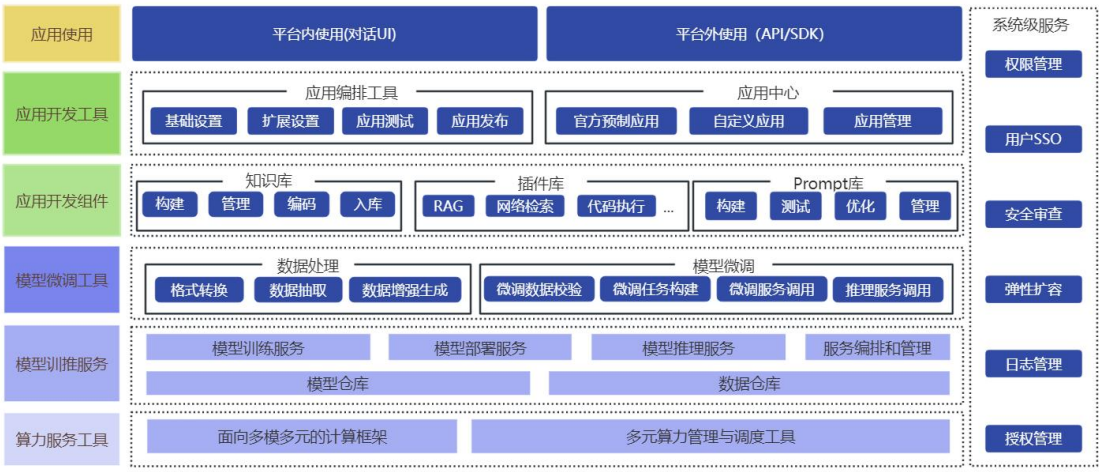

图 1-2 系统架构

1.3 应用场景

EPAI 包括平台管理中心、模型服务中心和应用开发中心三大套件。面向不同的企业需求，EPAI 提供不同的功能服务。例如，智能问答、文档检索等场景，EPAI 提供官方预置完整检索增强工程链路的应用，让您通过应用将通用基础大模型接入到业务解决方案中。而细分领域如金融场景的 IT 维修方案生成、法律场景的裁判文书分析总结等需要对推理结果进行定制调整，则可在 EPAI 使用模型微调功能，快速生成适应行业场景需求的定制模型，满足用户特定需求。

1.3.1 智能问答

智能客服、百科知识等需要沟通对话的场景。在现实生活场景中，针对用户需求提供快速应答，精准匹配用户需求，给出准确答案。推荐使用 Qwen、ChatGLM、Yuan2.0 等模型。

用户想直接了解一个概念。例如，什么是检索增强生成（RAG）？

1.3.2 知识检索

企业内部知识库、产品手册、科技文献等希望通过大模型在已知的文件库中进行查找答案的场景。不同企业家希望通过大模型来提升工作效率的方式，通过已有的知识库，与模型进行建联，从而使模型快速的检索企业知识并问答。推荐使用检索类插件。推荐使用 Owen、ChatGPT 等模型。

用户想让大模型在文档中找到对应答案，并显示文档来源。例如，《生成式人工智能服务管理暂行办法》中关于算法备案的规定是什么？

1.3.2 辅助办公

在办公场景下，我们希望大模型能够进行文档内容的翻译，文档的总结，文本优化等工作。推荐使用 ChatGPT、Yuan2.0 模型。

用户想对一段英文内容翻译为中文。例如， 将以下内容翻译为中文：Git is a distributed version control system widely used worldwide that supports fast, efficient management of projects, tracking code changes, and collaboration.

1.3.3 辅助编程

在开发者进行编程的场景，它可以实现代码的生成与补全、自动为代码添加注释、自动解释代码、自动编写单元测试等。推荐使用 Qwen、Yuan2.0 等模型。

开发者想让大模型直接根据要求生成程序代码，然后自己再进行检查和修改。例如，在诊断程序中，有大量的场景需要在 Linux 系统下执行系统命令。使用 Python 语言 subprocess 模块的 Popen 类，编写一个函数，执行 Linux 系统的命令，并返回命令的输出、错误信息和返回码。

当然，这仅仅是大模型落地企业中常见的一些应用场景，随着大模型能力的提高和应用开发组件越来越丰富，EPAI 的应用场景也将越来越多。

2. 产品优势

2.1 高效算力调度

EPAI 模型服务中心为平台的微调任务、服务推理提供资源分配能力，在大规模集群中选择合适节点用于任务运行，提高资源利用率的同时，尽可能提高任务运行性能。调度系统的主要能力体现在：

- GPU 细粒度调度：支持对 GPU 显存进行切片隔离，提交任务时制定所需切片的显存粒度大小和分片梳理，任务就会调度到合适的显存粒度切片的 GPU 卡上
- GPU 负载调度：采集并统计 GPU 卡负载数据，包括 GPU 利用率和 GPU 显存使用率，根据 GPU 卡的负载信息执行作业调度，优先选择负载较低的节点和 GPU 卡
- 紧急任务调度：内置紧急任务队列，只有在处理完全部紧急任务后，系统才会处理其他任务

2.2 多元算力适配

EPAI 模型服务中心可适配广泛的软硬家平台，通过建立设备插件规范和制定多元异构芯片接入的标准等手段，避免每次对不同计算芯片或加速卡或操作系统的重新对接开发工作，实现对多元算力的统一管理和调度。

2.3 微调数据便捷生成

EPAI 模型服务中心的自动化数据处理工具通过 OCR、版面分析等技术从多格式文档中提取可用于大模型训练或微调的文本内容，能够批量并发处理，实现私有数据资产的快速高效提取；数据处理工具通过对接大模型能力，对既有文本内容进行知识数据抽取或者根据 self-instruct、evol-instruct 等生成方法进行数据衍生，从而获得高质量微调数据，可直接用于大模型微调，帮助客户整理行业数据和专业数据，加速企业数据资产的生产效率。

另外，EPAI 模型服务中心内置了 5 万对中文问答数据，主题涉及公文撰写、诗词生成、文本翻译、安全对齐、头脑风暴、知识问答、代码生成、数学推理等多个领域，可用以解决大模型分阶段训练过程中出现的通用能力遗忘问题，方便用户直接调用通用数据进行模型微调，节省数据处理时间。

2.4 低门槛大模型微调

EPAI 模型服务中心提供低代码配置的微调模块，通过可视化界面构建微调任务，方便用户快速便捷进行大模型微调，降低使用门槛。支持全参数微调，仅需配置少量基本参数或高级参数，即可发起微调；微调过程，提供基于 Tensorboard 的全方位多角度的微调任务监控，同时提供容器级别的 CPU、内存、GPU 使用率等指标的实时展示，方便用户快速发现微调任务的资源问题。

2.5 自动化 RAG pipeline

支持企业常见的 13 种类型的文档格式构建企业知识库，支持文档版面分析（Layout Analysis）、文字和表格抽取，全面覆盖企业私有数据类型的知识库构建能力,非结构化数据抽取精度高。

图 2-1 知识库构建

支持关键词检索、语义检索、混合检索多种检索方式，满足不同场景的检索需求。引用结果快速定位到文档，支持生成结果引用溯源。

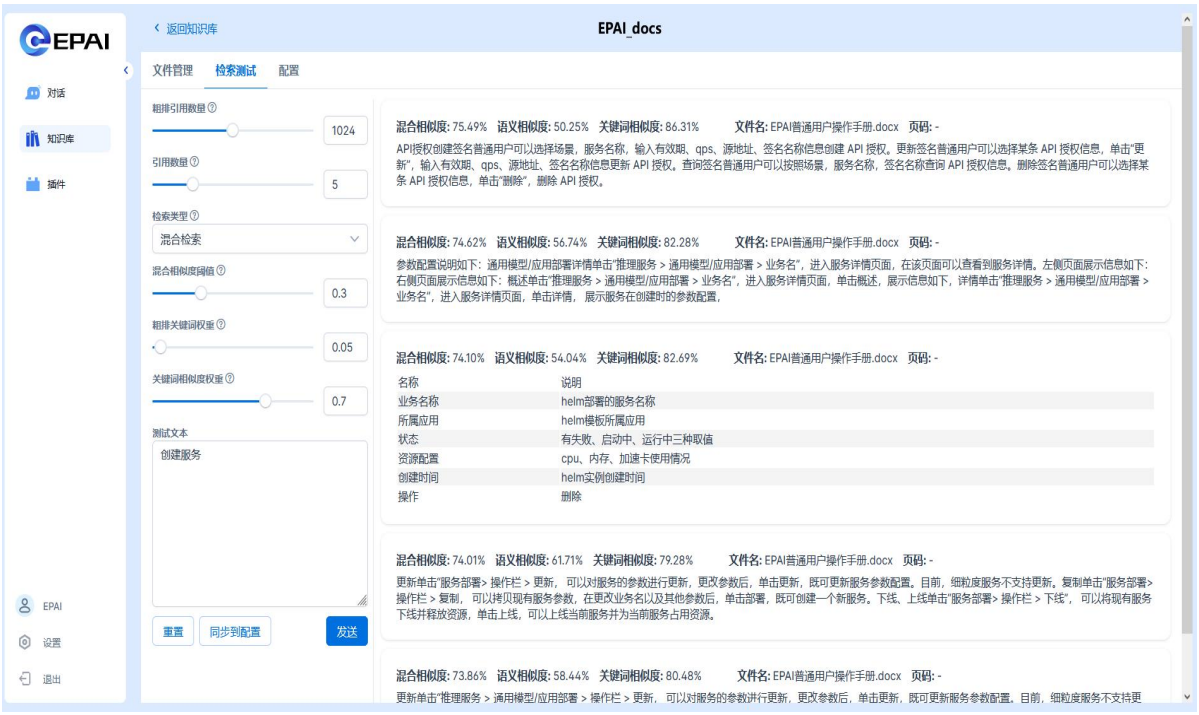

图 2-2 多种检索方式

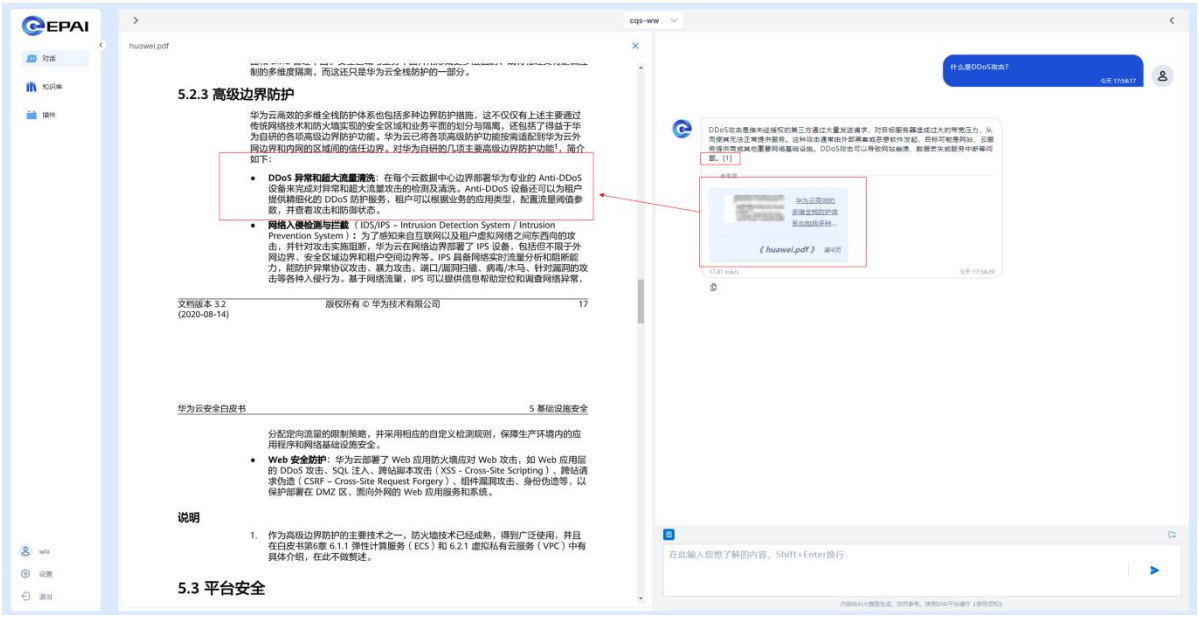

图 2-3 引用溯源

2.6 灵活使用应用

平台支持应用的多种使用方式，可以面向业务系统集成提供标准的模型能力输出。

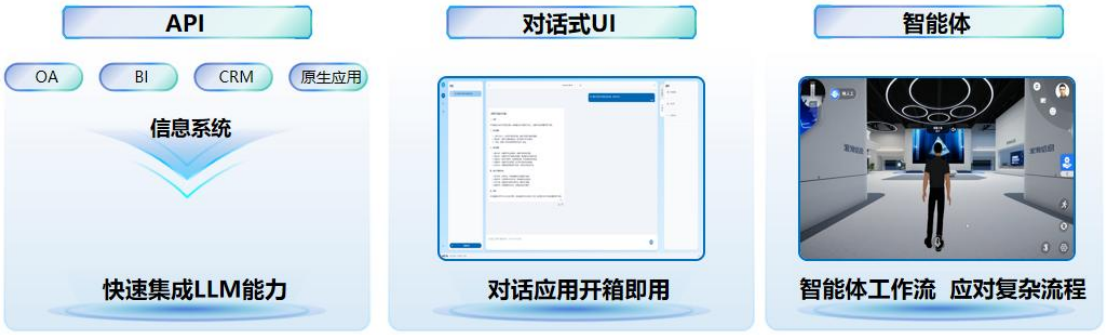

图 2-4 多种应用的使用方式

3. 产品功能

3.1 数据处理

EPAI 模型服务中心提供 4 种数据处理方式，分别是文本提取、微调数据抽取、微调数据增强和数据清洗。

- 文本提取：converter 命令支持从多种格式文档提取可用于大模型训练或知识库构建的文本 txt。支持提取的格式包括 pdf、doc、ppts、xlsx、xml、png、epub、md 等结构化、非结构化格式；在格式转换过程中，兼顾了图片、表格、公式中信息的留存，保证数据的高质量性。converter 命令可用于单一格式或混合格式的文本提取
- 微调数据抽取：sft_extractor 命令通过接入第三方大模型对已有文本进行问答数据抽取，获得基于原有文档的高质量微调数据。
- 微调数据增强：sft_generator 命令通过接入第三方大模型对已有微调数据，根据 self-instruct、evol-instruct 等生成方法进行数据衍生，获得同类型、同主题的高质量微调数据。
- 数据清洗：data_cleaner 命令针对已有的微调数据或者文本数据，使用各个过滤器对原始数据进行清洗。

用户可根据自身业务对数据的要求选择对应的数据处理类型。EPAI 对数据生成以任务的方式进行管理，以低代码的方式提供数据生成的全生命周期管理。

（1） 创建数据处理任务

在数据处理任务列表，点击“创建”，进入创建任务页面，如下图所示：

- 数据处理类型：支持三种数据处理任务类型，根据业务需求选择对应的处理类型
- 运行：根据所选择的处理类型的备注提示，填写数据处理任务的运行命令，不同命令的参数不同，请按照提示说明填写符合实际业务所需的参数，包括指定输入文件或文件目录、指定输出文件目录等
- 资源：选择任务所需的资源、加速卡等

填写必填项的信息后，点击“确定”，提交任务，任务提交成后，在数据处理任务列表新增一条记录

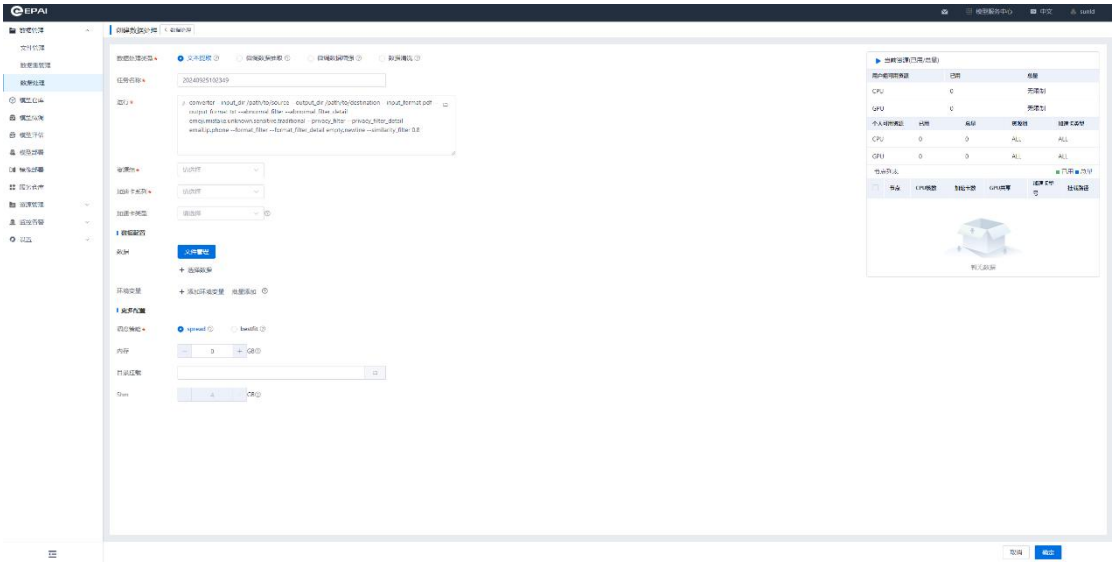

图 3-1 创建数据处理任务

（2） 查看任务过程

在数据处理任务列表，点击任务名称，进入任务详情页，查看任务运行日志、监控指标、任务基本信息等内容。

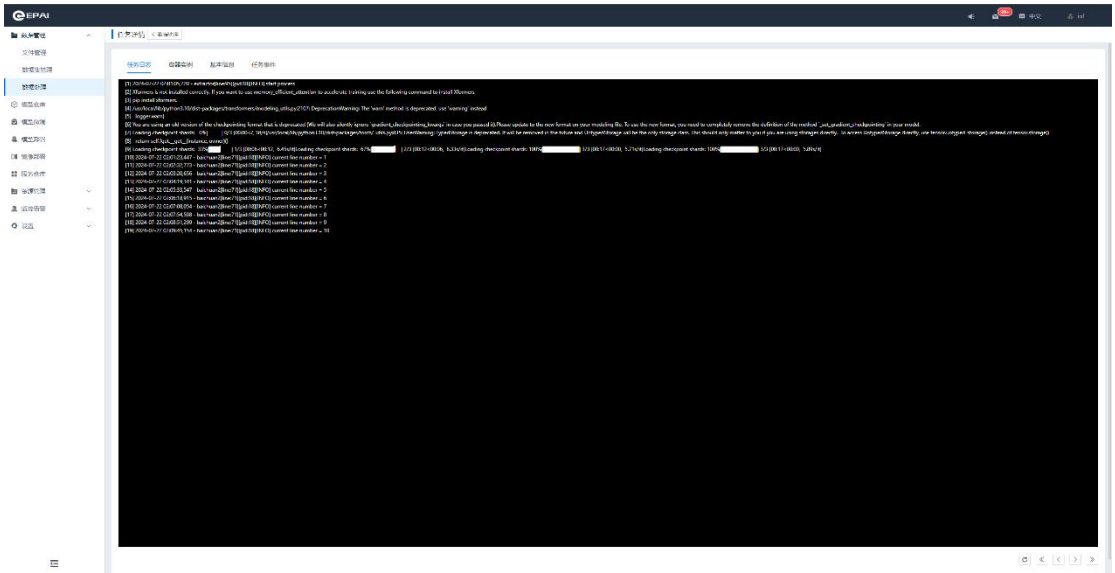

图 3-2 查看数据处理过程

（3） 查看输出结果

任务运行完成，可根据之前任务运行命令中指定的输出目录，在自己的用户家目录下，查看数据处理后输出的文件结果。

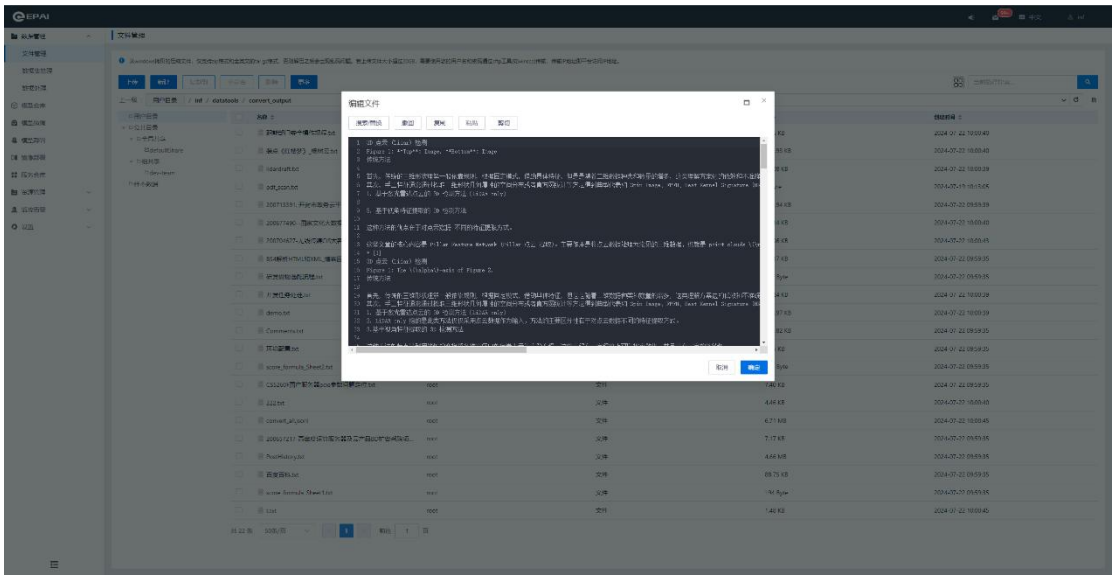

图 3-3 查看数据处理结果

3.2 模型微调

EPAI 模型服务中心提供了低代码配置化的微调模块，可以方便用户快速便捷的进行微调，无需关心微调任务的各种技术细节，仅需要调整各项参数即可进行模型调优，极大的提高了大模型微调的易用性和效率。

（1）创建微调任务

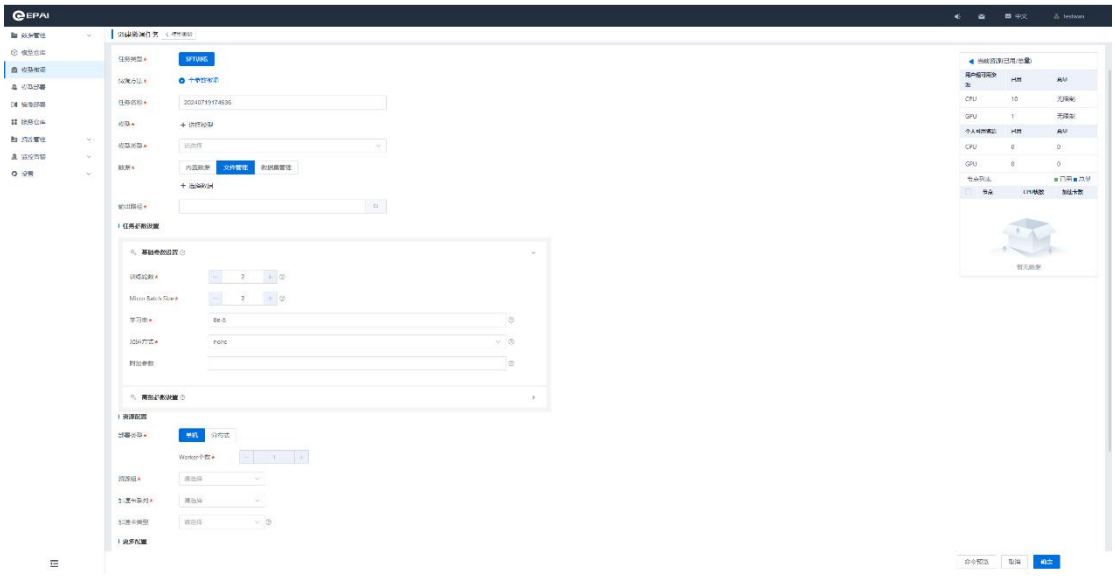

图 3-4 创建微调任务

在“模型微调”列表下，点击“创建”，进入微调任务的创建页面，如下图：

- 填写任务基本信息，包括任务类型、微调方法（目前只支持全参数微调）、任务名称、基座模型、模型类型、微调数据
- 设置微调参数，包括基础参数和高级参数，如训练轮数、micro batch size、

学习率、学习率调节器、梯度累计等

- 配置资源，选择资源组、加速卡等

完成任务配置后，点击“确定”，提交任务，系统即可从资源组中调用资源对任务进行处理

（2）查看微调任务过程

在模型微调过程中，用户通过点击任务名称，进入任务详情，查看任务运行日志；可以在任务列表中点击“可视化”，平台调用 Tensorboard 对任务运行情况进行可视化展示

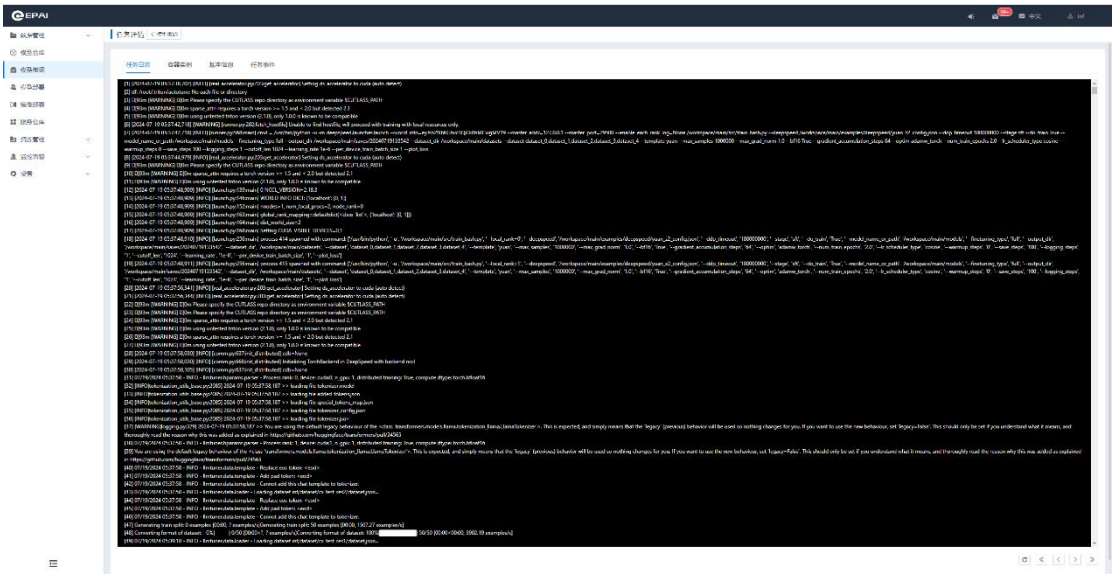

图 3-5 查看微调任务过程

（3）模型导出

微调任务完成后，EPAI 平台会将生成的模型自动导入至模型仓库，供用户进行模型再次微调、评估或者服务部署

3.3 模型评估

EPAI 模型服务中心支持对微调后生成的模型进行快速的能力评估。模型服务中心提供了模型评估生命周期管理功能，支持模型评估任务的发起、运行、结束、状态监控等能力。评估任务运行完成后在，支持对模型预测结果的人工评估，支持模型维度的多评估结果查看。

在正式部署模型前，可通过评估任务对模型进行评估，以了解模型在不同数据集或场景下的表现如何。以下是模型评估任务的流程

（1）创建评估任务

在模型版本列表，选择需要评估的模型版本，点击“模型评估”，将会进入

创建评估任务页面，填写任务基本信息和任务参数。

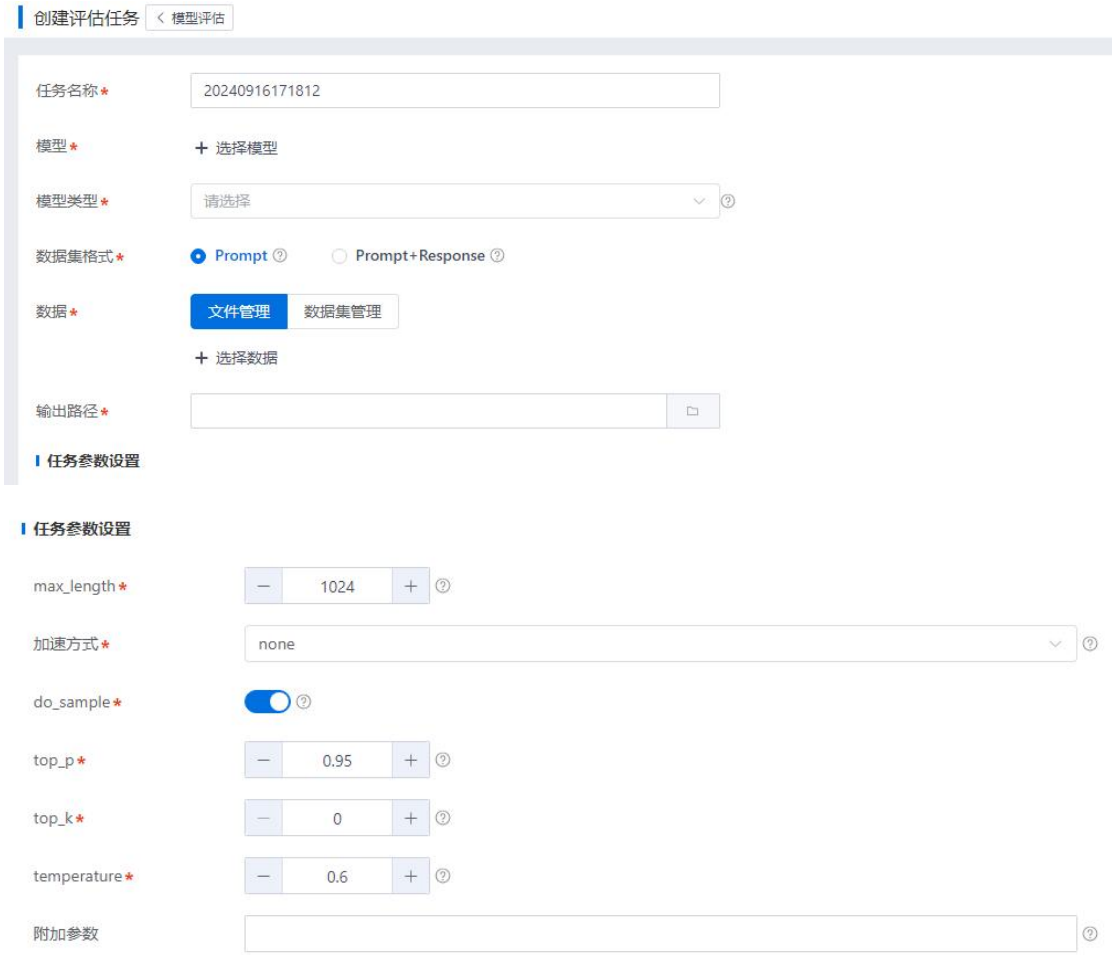

图 3-6 创建模型评估任务

（2）任务列表

创建评估任务成功后，在任务列表中新增一条任务记录。支持用户查看进行中的任务列表和已终结的任务列表。

点击任务名称，支持查看任务详情页。在详情页，支持查看任务日志、容器实例和任务基本信息。

针对进行中的评估任务，支持对任务进行停止、启动和删除操作，

针对评估任务，支持对任务进行重新提交操作。

针对状态为完成的终结任务，支持查看评估结果，在详情页，支持查看任务信息、评估分数、评估结果和模型评估记录

（3）评估结果

针对已完成的评估任务，点击“评估结果”，可查看模型在所选数据集下生成的推理结果。这些模型推理支持人工对输出结果进行人工标注打分。

当全部推理输出结果完成人工标注打分后，系统会自动生成模型评估结果打分，包括样本数、人工评估分数等。

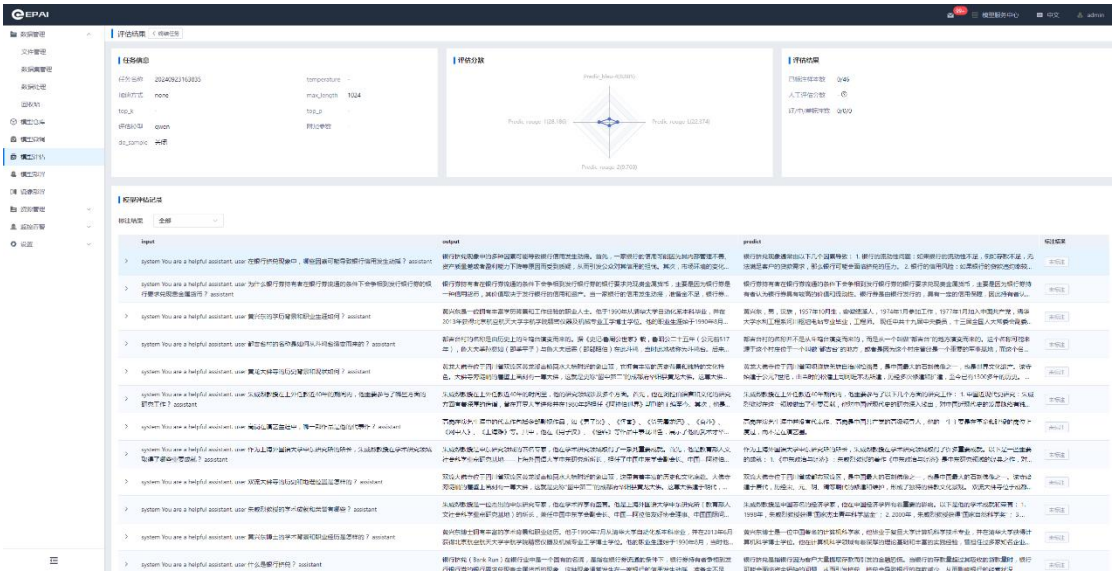

图 3-7 模型评估结果

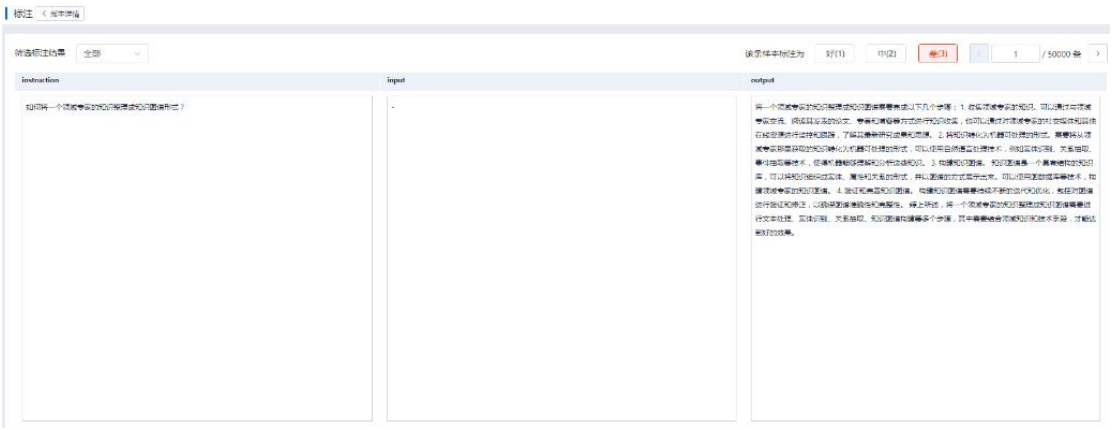

图 3-8 结果标注

3.4 模型部署

EPAI 模型服务中心支持大模型的一键配置化部署，并提供 OpenAI API 和 EPAI API 接口，能够兼容业内常见的大模型，方便用户部署各类大模型。

支持业内常见大语言模型：EPAI 服务部署框架支持 vLLM、transformer 等服务推理框架，并针对常见模型进行了模板化处理，实现了各个模型的低代码部署。并提供标准的 OpenAI API 接口。同时，对于私有或定制化模型，提供了镜像部署方式，提供镜像文件和模型文件下，自定义推理服务启动命令即可部署推理服务。

提供基于 QPS（服务请求量）的自动扩缩容能力，保证线上业务在不同调用量下的资源合理性分配，能够基于 QPS（请求量）的变化主动为服务提供所需的计算资源，释放过程同样无须手动操作。

在应用调用大模型服务前，需要将微调后的模型进行服务部署。以下是模型

部署的主要流程

（1）发布模型

当大模型微调完成并自动生成新的模型到模型仓库后，需要先将新的模型进行发布。在“模型仓库”->模型版本列表中，选择对应的模型版本点击“发布”，确定后，模型版本状态为“已发布”。

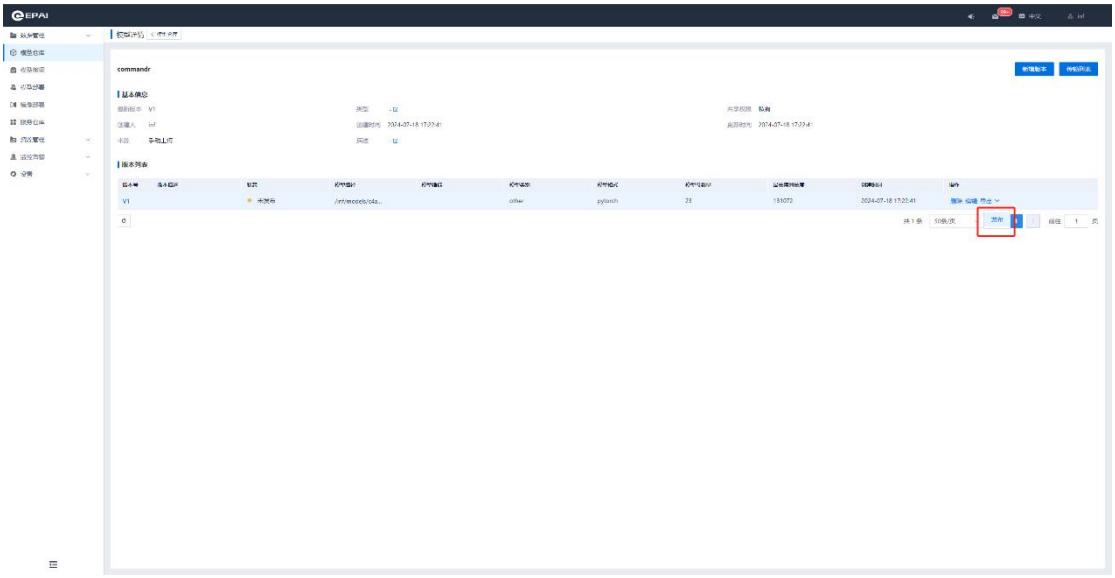

图 3-9 模型发布

（2）生成服务配置

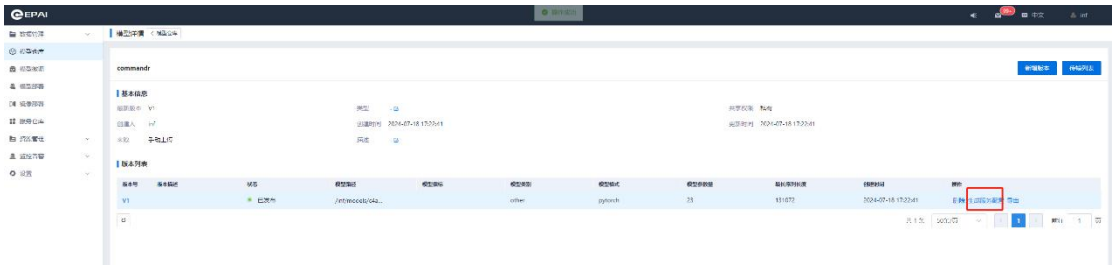

图 3-10 生成服务配置

针对已发布的模型版本，点击“生成服务配置”，页面弹出服务配置导入界面，如下图所示，填写相关信息

- 点击“校验”，自动获取模型信息和校验 token
- 选择场景
- 填写服务配置名称

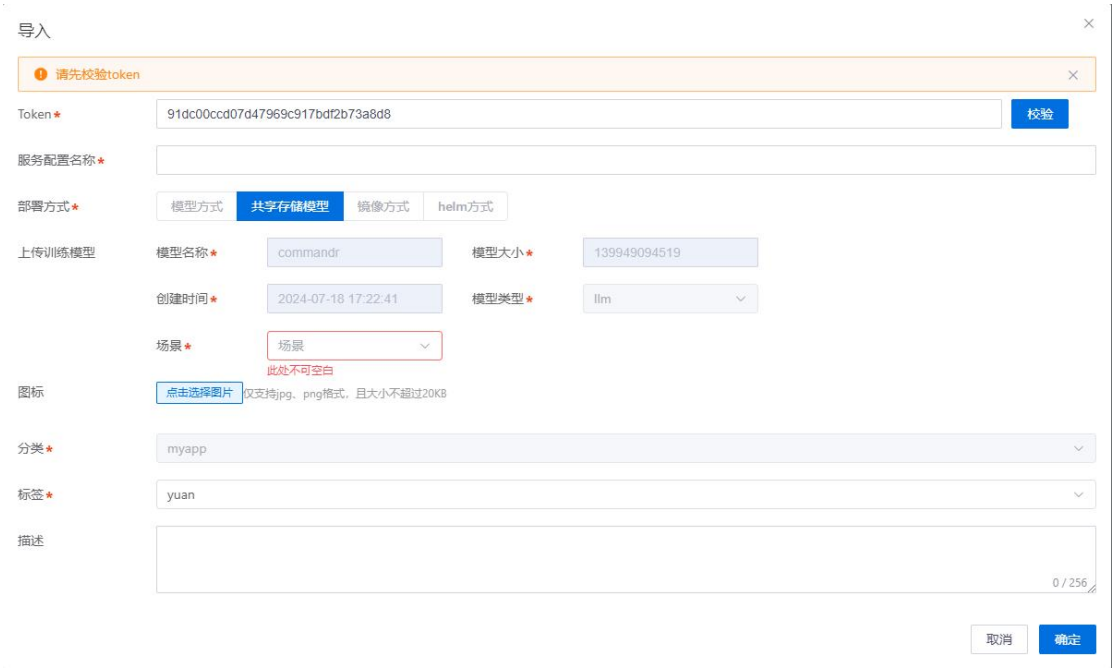

图 3-11 填写部署配置信息

完成信息填写后，点击“确定”，就可在服务仓库中新增 1 条服务配置内容。

（3）模型部署

在“模型部署”页面，点击“立即部署”，打开模型部署页面，如下图所示。

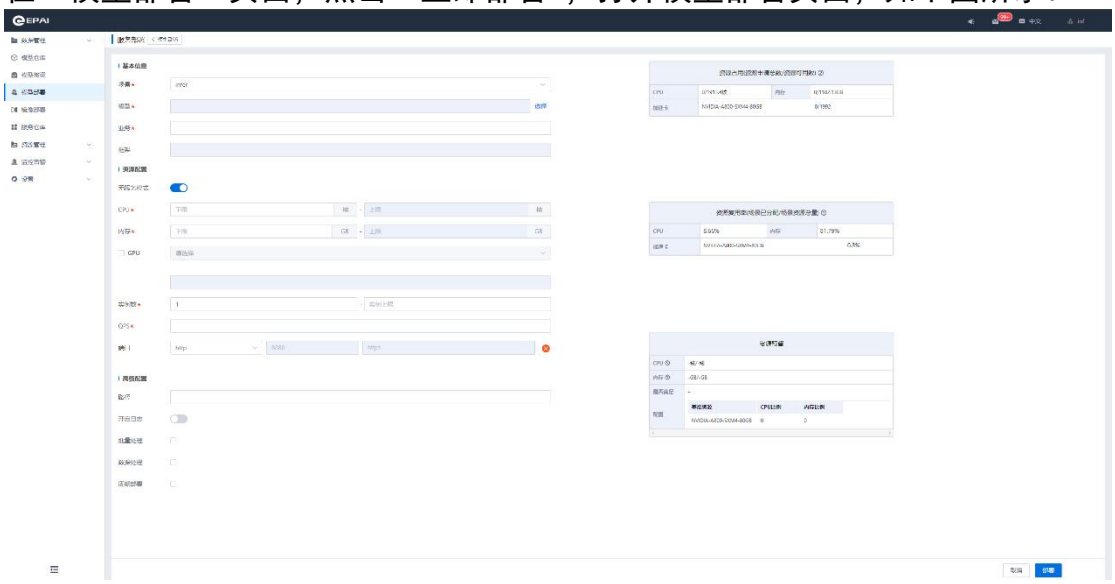

图 3-12 立即部署

- 基本信息：选择场景和模型，填写业务名称。其中，点击“选择”，弹出模型列表，从中选择之前生成的服务配置

图 3-13 模型列表

- 资源配置：选择服务部署的资源，包括 CPU、内存、GPU、实例数、QPS 和端口等
- 高级配置：根据业务要求灵活配置更多设置，如批处理、周期部署等

完成上述信息填写后，点击“部署”，即可完成部署，等待系统调用资源将服务进行部署和启动。

（4）查看服务详情

在“模型部署”列表页，选择业务名称，进入服务详情页，查看服务各项信息和运行情况，包括概述、服务监控、日志、shell 链接、在线调试和 AB 测试。用户可根据自身业务要求查看服务对应的内容，或更新服务。

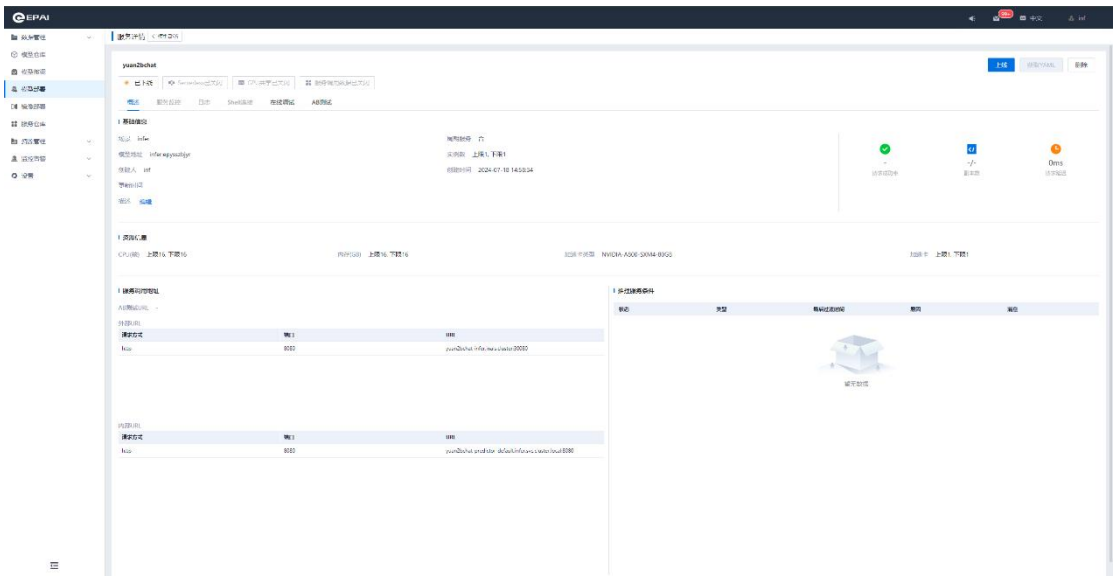

图 3-14 查看服务详情

（5）服务发布

当服务已就绪且验证测试 OK 后，可选择将服务发布给应用开发中心。选择某个服务，点击“发布”，可将该服务进行发布，发布范围支持全局或部分用户

- 全局：应用开发中心的所有用户都可见该服务，都可调用该服务进行应用开发。
- 用户：支持从用户列表中选择或搜索单个或多个用户进行发布，被选中的用户可以调用该服务进行应用开发。

3.5 知识库构建

平台支持用户构建专属知识库，知识库构建包括①文档上传、②文档解析、③知识编码、④知识存储四个步骤。

图 3-15 知识库构建

①文档上传

目前支持 pdf、docx、txt、html、htm、md、json、jsonl、epub、mobi、xml、pptx、doc 共 13 种格式的文档。单篇文档大小不超过 500MB，文件上传支持多个选择多个文件，可多次进行上传，单次上传的文档数量最多为 200 篇，上传过程中可以查看上传进度。

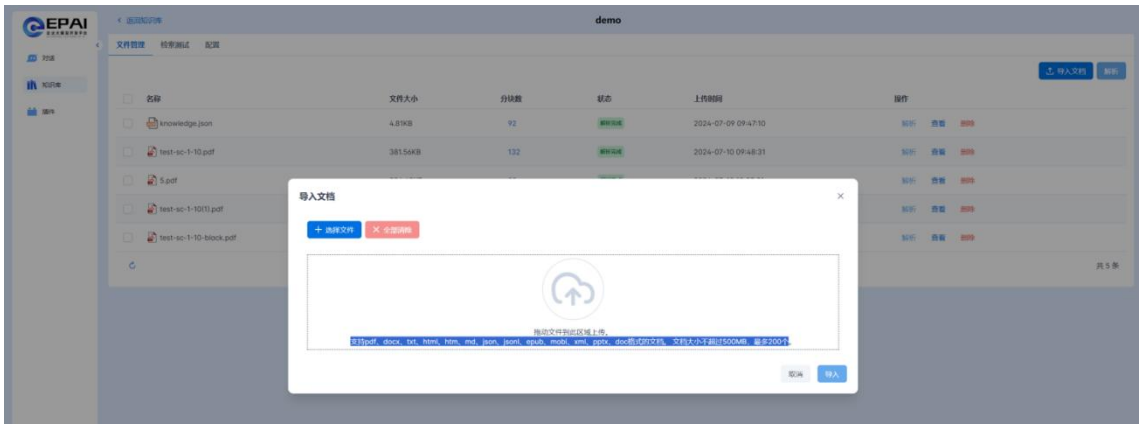

图 3-16 文档上传

②文档解析

文档解析是指文档中的具体内容进行解析，抽取出内容并对其进行分块,文本分块（chunk）也可称为“知识”。当前支持文档版式分析、文档中文本和表格内容的解析。

图 3-17 文档解析

③知识编码

选择编码模型（Embedding Model），对知识（文本 Chunk）进行向量编码，就得到了一条知识向量。当前支持的编码模型有 bge-large-zh-v1.5、bce-embedding-base_v1、text2vec-base-chinese，可以根据文档内容的需要，选择合适的编码模型。

- bge-large-zh-v1.5，擅长中文，模型较大，向量维度 1024。BGE 是由北京智源人工智能研究院发布的开源通用 embedding 模型，包括一系列的多款模型，其中 large 代表大参数量，zh 代表中文，1.5 代表版本，相比于 1.0 版本，其相似度分布更加合理。
- bce-embedding-base_v1，擅长中英双语和跨语种语义表征能力，向量维度 768。bce-embedding 是由网易有道翻译发布的一款开源模型，针对不同领域如教育、法律、金融、医疗等的 RAG 场景做了专门优化。
- text2vec-base-chinese，擅长中文，向量维度 768。该模型由个人开发者 shibing624 发布的开源模型，对于句子编码效果效果，适用于句子嵌入、文本匹配或语义搜索等任务。

④知识索引

每条知识包括一条文本内容和一条向量内容，可以把这些内容存入数据库 ElasticSearch 中，建立索引，从而支持关键词、语义、混合的检索。

向量检索和关键词检索都有其优点和局限性，在实践中我们往往综合两者进行使用（混合检索），从而达到比较好的检索性能。

图 3-18 混合检索

3.6 应用开发

EPAI 平台支持可视化零代码的应用配置开发，只需要进行模型、插件的选择和参数设定，就能完成定制应用的开发。但应用开发不是一次的过程，想要达到比较高的性能，需要持续进行调试和优化。

图 3-19 应用开发过程

①大模型选择

不同的大模型其擅长能力不同，用户可根据场景需求，选择合适的模型进行使用。

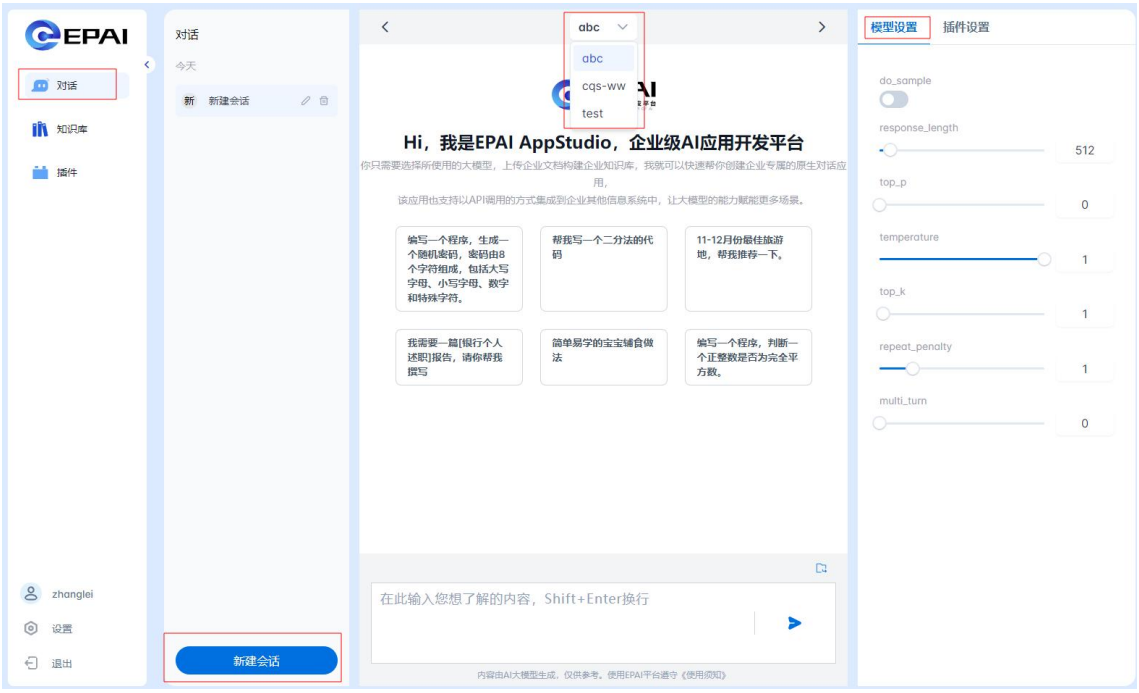

图 3-20 选择模型

每个大模型推理服务都有其对应超参数，不同的模型对应的超参数以及参数的取值范围是不一样的，可通过点击右侧栏【设置】显示修改。其中最主要的超参数说明如下。

输入参数	参数含义	参数类型	取值范围
do_sample	使用采样策略，当设置 do_sample=True 时，生成过程将采用随机采样策略。在生成每个新词时，模型会根据词汇表中词汇的概率分布进行随机采样，从而选择下一个词。采样的随机性使得生成的文本更加多样化，且可能会产生更为创造性的结果。	bool	True/False
response_length	UI 设置生成文本的最大长度	Int	(1-8000)
top_p	float，用于控制生成回复的多样性，它基于累积概率选择候选词，直到累计概率超过给定的阈值为止。	Float	[0-1] 步长 0.1
temperature	float，控制生成的随机性，较高的值会产生更多样化的输出。	float	(0-1] 步长 0.1
top_k	Int，可选；用于控制模型生成文本的参数,它指定了模型在生成下一个词时只考虑概率最高的前 k 个词。	Int	[0-50]
repeat_penalty	float，（重复惩罚）是一种技术，用于减少在文本生成过程中出现重复片段的概率。它对之前已经生成的文本进行惩罚，使得模型更倾向于选择新的、不重复的内容。	float	[0.5-3] 步长 0.01
multi_turn	多轮对话轮数设置，默认 0 表示不采用多轮对话	Int	[0.5-10]

②插件选择

针对业务场景，可以配置不同的插件。例如，当前是阅读资料迅速获取信息

的场景，则可以使用文档阅读插件；而如果是产品咨询问答的场景，需要先查询企业的产品知识库，再进行回答，则可以使用知识库插件；而如果是百科知识、最新咨询的问答，则可以使用网络检索插件。不同侧插件也有不同的参数，设定合适的参数，可以让插件发挥更好的性能。

图 3-21 使用插件

4. 产品安全

4.1 数据安全

EPAI 平台提供全链路的企业数据防护，从多个层面保证用户数据的隔离和访问安全

- 账户管理控制：采用 RBAC 权限管理，使账号管理更加安全。针对数据集、知识库等进行用户级访问控制和细粒度权限管理。增加独立的运维管理账号，在不影响使用的前提下，避免过多的权限分配。
- 数据与存储加密：针对数据进行了文件级别的数据加密；数据存储时同步进行加密存储
- 备份恢复：具备独立的备份和容灾模块，在平台遭遇不可抗力时能够恢复平台和业务数据。
- 端口管理：通过端口限制达到数据的安全访问，支持修改集群 ssh 连接端口，屏蔽非法软件扫描 22 端口；整个集群的各个节点都开启了 firewall 防火墙，对外默认只允许访问必要的服务端口，不在开放端口资源内的访问默认拒绝，以适配业务安全的需求
- 租户资源隔离：租户间的计算、网络和存储等资源不可见，避免恶意攻击和数据窃取等。

- 系统 HA 机制：平台支持整体高可用方案，提供对服务、Kubernetes、数据库、Harbor 等模块的高可用功能，保障了业务不中断、数据不丢失。

4.2 网络安全

EPAI 对用户（创建的微调任务或推理服务）添加了网络隔离限制。当前版本支持两种策略，在使用安装包安装完成后，默认启用策略一。管理员可以按照以下说明随时切换策略或解除限制。

禁止访问外部网络：用户仅可以访问同用户组下的用户。用户无法访问集群中宿主机的 IP 地址和集群外部的 IP 地址。

允许访问外部网络：用户可以访问同用户组下的用户。默认情况下，用户无法访问集群中宿主机的 IP 地址，但可以访问集群外部的所有 IP 地址。管理员可以通过后台配置禁止列表，这将禁止用户访问这些 IP 地址或 IP 地址段。当用户需要访问的某 IP 或 IP 地址段已经在禁止列表中时，可以再将它们配置到允许列表中来允许用户访问它们。

4.3 内容安全

模型在训练阶段，可以学习防指令攻击的能力。在推理阶段，除了模型自身具备的能力外，平台在交互内容审核方面还预制了敏感词检测插件工具，它可以自动识别输入和大模型生成内容、伦理道德、价值观、信息保护等生成内容风险，提供多种风险等级的监控。它支持本地敏感词库审核、微调模型语义审核、第三方安全内容审核三种手段来进行交互内容的审核。