

RAG 动手实验

来源: <https://edu.aliyun.com/course/3126500/>

一、开通阿里云大模型服务百炼服务	2
二、背景知识回顾	6
三、尝试提问	8
四、创建知识库	10
五、创建 RAG 应用	15
六、RAG 应用案例：公司财务数据分析	18
七、RAG 应用案例：企业差旅规定	24

一、开通阿里云大模型服务百炼服务

本文为您介绍开通大模型服务百炼服务的方法，如果已经开通服务，请跳过此步骤，可直接在控制台上进行操作，直接执行第 5 步。有关大模型服务百炼服务的计费方式，详情请参见[大模型产品计费](#)。

1. 前往大模型服务平台百炼控制台（<https://bailian.console.aliyun.com/#/home>）。
2. 在**服务协议**对话框中，阅读并单击**同意**。

说明：如果您是大模型服务平台百炼的老用户，不会弹出此对话框，请您跳过此步骤。

服务协议

查看历史版本

阿里云百炼服务协议

版本生效日期：2024年08月21日
版本更新日期：2024年08月14日

欢迎您体验阿里云百炼！

【审慎阅读】您在同意本协议之前，应当认真阅读本协议。请您务必审慎阅读、充分理解各条款的内容，特别是免除或者限制责任的条款、法律适用和争议解决条款，这些条款将以**粗体**或**粗体下划线**标识，您应重点阅读。如您对协议有任何疑问，可以通过本协议披露的联系方式与我们沟通。

【签约动作】当您勾选或点击同意本协议或以其他方式选择接受本协议后，即表示您已充分阅读、理解并接受本协议的全部内容，并与我们达成一致。本协议自您通过网络页面勾选或点击确认或以其他方式选择接受本协议之日起成立。如果您不同意本协议或其中任何条款约定，请勿进行签约动作或实际使用本服务。 您知悉并确

同意

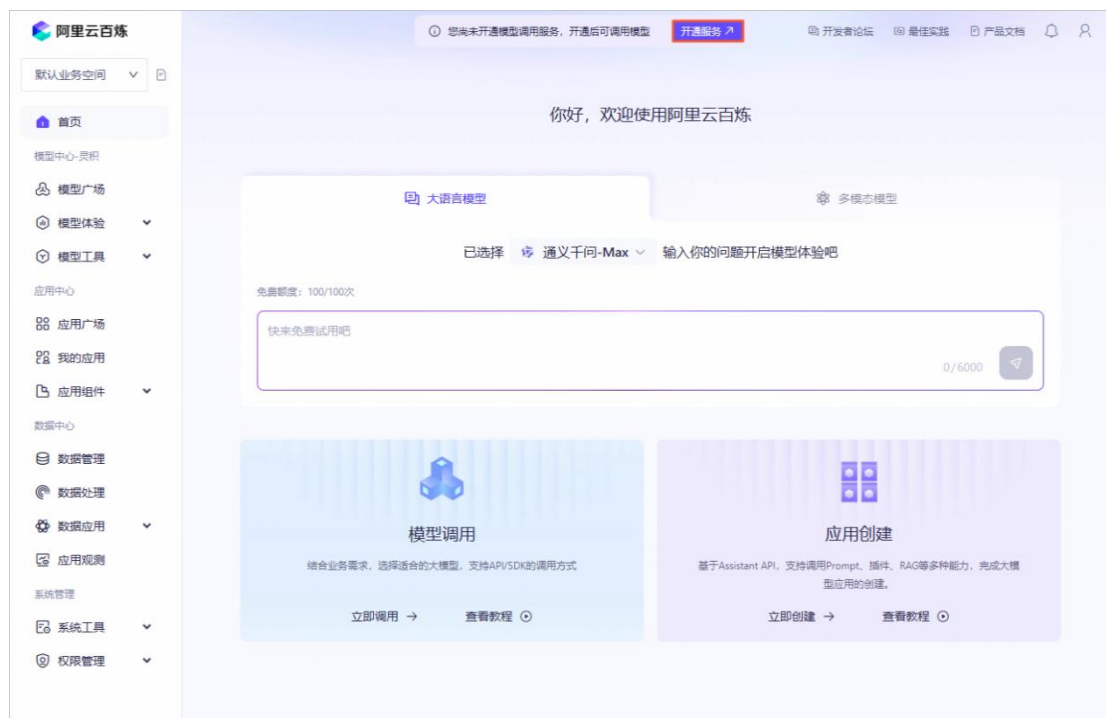
拒绝

3. 在首页顶部，显示如下图所示的消息，您需要开通百炼的模型服务，以

获得免费额度，请单击**开通服务**。

说明：如果未显示该消息，则表示您已经开通，请跳过此步骤。

4. 在弹出的对话框中，勾选**我已阅读并同意《模型管理服务协议》**，单击**确认开通**。



说明：大模型服务平台百炼为首次开通服务的用户提供免费试用额度，开通的阿里云主账号与其 RAM 子账号共享免费试用额度。免费试用额度从开通百炼或模型申请通过之日起计算有效期，有效期一般是 30~180 天不等，详情请参见[模型免费额度赠送](#)。

继续使用百炼大模型服务，需要开通以下商品，并创建模型调用API-KEY

使用大模型服务会按计费规则进行计费，需要开通以下商品，开通后即可使用相关服务，按照对应服务定价产生计费，不使用则不会产生计费。

如下列举了当前计费的模型。未来新增模型计费会直接更新到您已开通的商品中，您无需再次开通。

新增模型计费或修改模型计费，均会在控制台进行公告，请您在调用模型前留意模型单价和计费规格。

百炼大模型推理 (190)

百炼大模型部署 (190)

百炼大模型训练 (190)

通义千问-Max (qwen-max)

通义千问-Max-Latest (qwen-max-latest)

通义千问-Max-2024-09-19 (qwen-max-0919)

通义千问-Plus (qwen-plus)

通义千问-Plus-2024-09-19 (qwen-plus-0919)

通义千问-Plus-Latest (qwen-plus-latest)

通义千问-Turbo (qwen-turbo)

通义千问-Turbo-2024-09-19 (qwen-turbo-0919)

通义千问-Turbo-Latest (qwen-turbo-latest)

通义千问VL-Max (qwen-vl-max)

通义千问VL-Max-Latest (qwen-vl-max-latest)

我已阅读并同意《模型管理服务协议》

计费详情

取消开通

确认开通

5. 进行模型体验。

在左侧导航栏中，选择**模型体验>文本模型>文本对话>通义千问-MAX**，在下方的输入框中输入你想问的问题。

4

阿里云百炼

主账号空间

首页

模型中心-灵积

模型广场

批量推理

模型体验

文本模型

语音模型

视觉模型

模型工具

应用中心

应用广场

我的应用

应用组件

数据中心

数据管理

数据处理

数据应用

应用规则

文本对话

文本调试

选择模型

我想体验

性能出众的模型

最新发布的模型

高性价比的模型

长上下文的模型

为你推荐以下大模型

通义千问-Max

文本生成 32K

通义千问2.5系列千亿级超大规模语言模型。支持中文、英文等不同语言输入。随着模型的升级，qwen-max将滚动更新升级。如果希望使用固定版本，请使用历史快照。

通义

2024-10-15更新

通义千问-Plus

文本生成 128K

通义千问超大规模语言模型的增强版。支持中英文等不同语言输入。

通义

2024-12-20更新

通义千问2-开源版-72B

文本生成 128K

通义千问2对外开源的72B规模的模型。

通义

2024-06-07更新

更多模型

已选择

通义千问-Max

输入你的问题开启模型体验吧

体验模型将会消耗Tokens，费用以实际发生为主（模型部署-算力时长计费模型除外）

知道了

请输入您想问的问题

0/6000

模型生成的所有内容均由人工智能模型生成，其生成内容的准确性和完整性无法保证，不代表我们的立场或观点

5

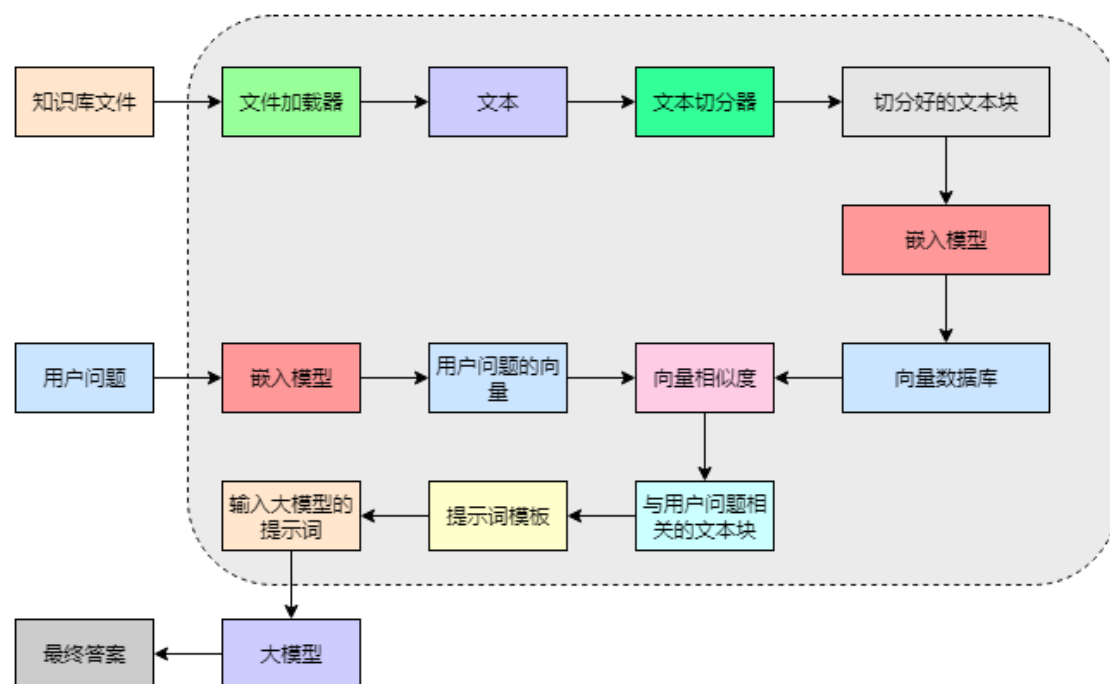
二、背景知识回顾

就像你无法回答一个陌生领域的问题一样，大模型也无法回答预训练阶段没有准确掌握的知识。但是我们可以大模型直接回答私有领域问题之前，给大模型一些参考，让大模型结合参考来回答问题。这一技术被称为检索增强生成（Retrieval-Augmented Generation, RAG），非常适合于在私域知识问答场景中消除大模型幻觉（编造答案）。

RAG 应用有两个关键过程：

1. 建立索引。这一阶段你需要将私有领域知识的文档（如 PDF、Word 等格式）存储起来并建立索引。这一过程包含：将文档中的文字提取加载出来、切分成小的分块（chunk）以避免超过大模型提示词长度限制、将文本 chunk 向量化后存储到向量数据库中以便于后续检索。
2. 检索和生成：当你为私有知识建立好索引，并完成相关流程开发后，用户就可以进行提问。你的 RAG 应用收到用户问题后，会去向量数据库中检索和问题相关的 chunk，然后将相关的 chunk 组合到提示词中给到大模型。大模型会结合参考信息给出回答。

以下是常见的 RAG 应用流程图：



开发一个 RAG 应用需要你具备一定的代码能力和算法基础，并且也会耗费一些时间。

阿里云的百炼平台上提供了 0 代码基础就能创建 RAG 应用的方案，你只需要关注私有领域知识库的维护即可使用，同时也提供了更精细化的设置，如切分方法、召回文本块个数等，以便于你持续改进回答效果。

三、尝试提问

在正式搭建 RAG 应用之前，我们可以先测试一个问题「西红市实验十小一年二班的班主任是谁？」来看下大模型的回答效果：

1. 进行模型体验。点击**模型中心-灵积>模型体验>通义千问-MAX**，在下方的输入框中输入提示词：

西红市实验十小一年二班的班主任是谁？



2. 你可能会看到这样的回答。



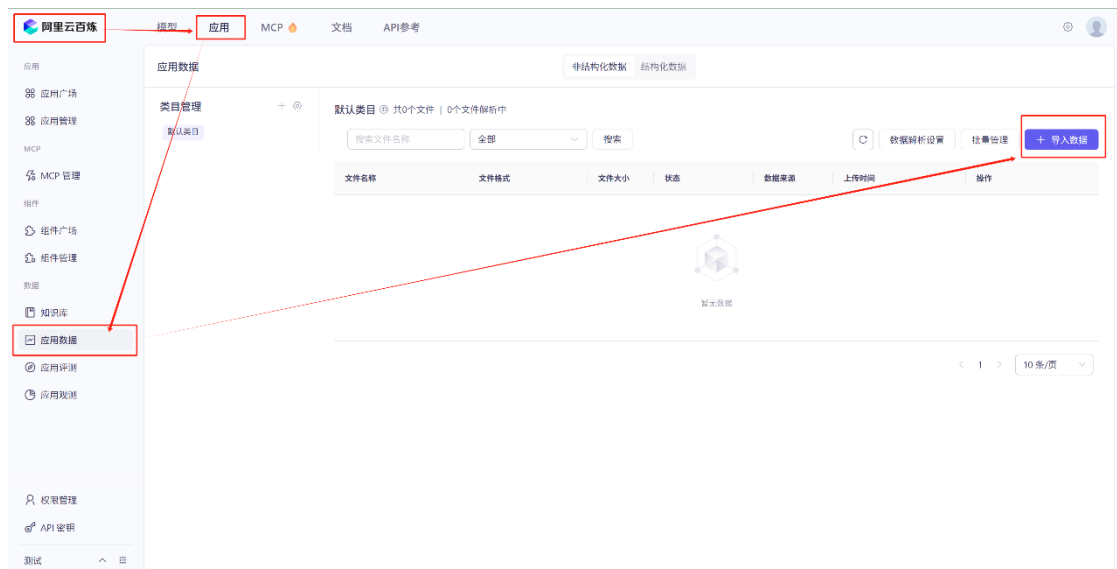
因为“西红市实验十小”这个学校是我们虚构的，大模型无法回答这个私有

领域的问题。

四、创建知识库

为了能够回答前一步骤的问题，我们需要创建一个知识库，并维护一些私有领域的知识文档。你可以参考如下步骤完成：

1. 下载我们提前准备好的示例知识库文件：[示例知识库.doc](#)
2. 单击左侧菜单栏中的**应用数据**，在**默认类目**下，单击**导入数据**。



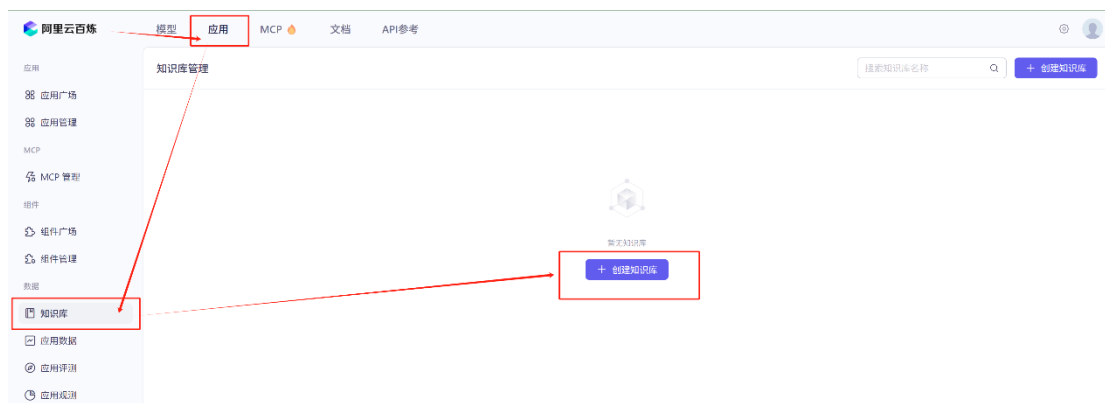
3. 在导入数据界面，单击**本地上传**，上传知识库文件（本实验使用的是示例数据），上传完成后单击**确认**。文档解析需要花费一段时间，请耐心等待，可以主动刷新页面。



等待数据解析完成

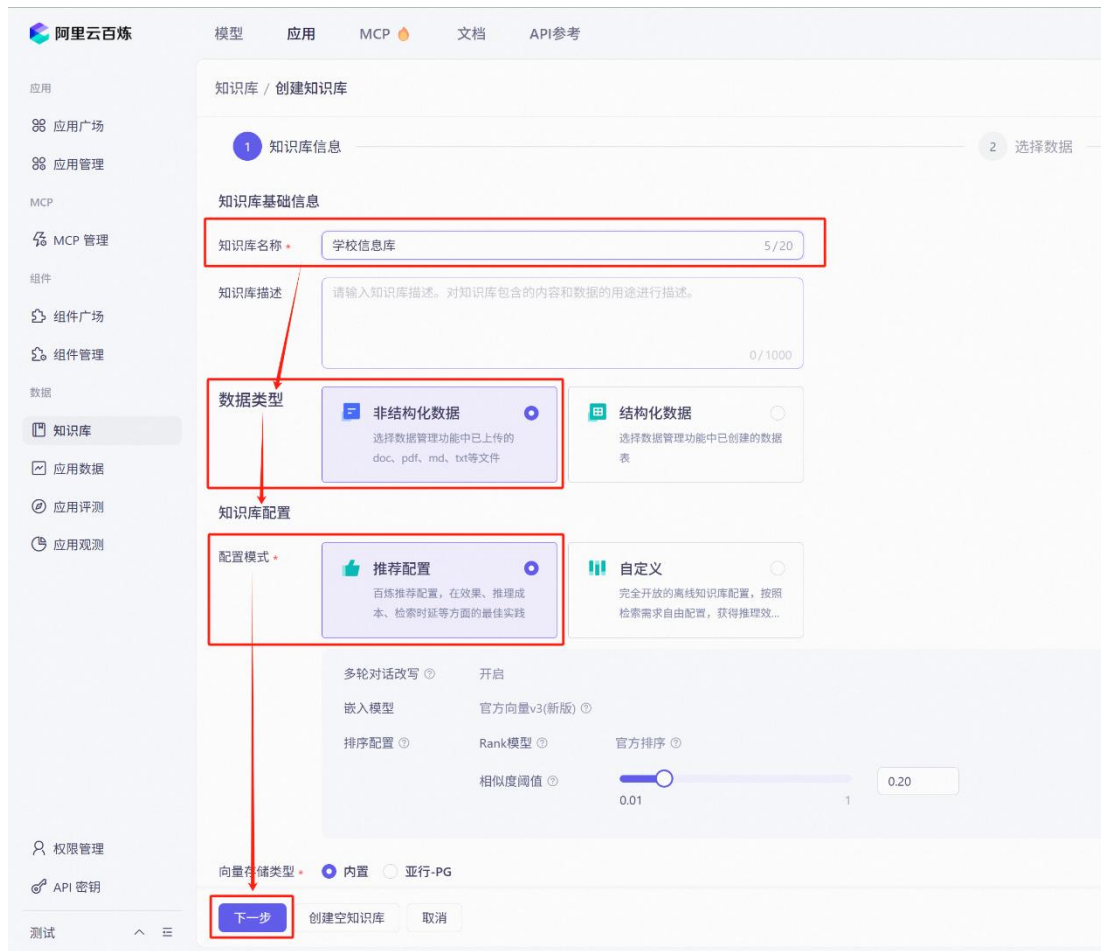


4. 单击左侧菜单栏中的知识库，单击创建知识库。

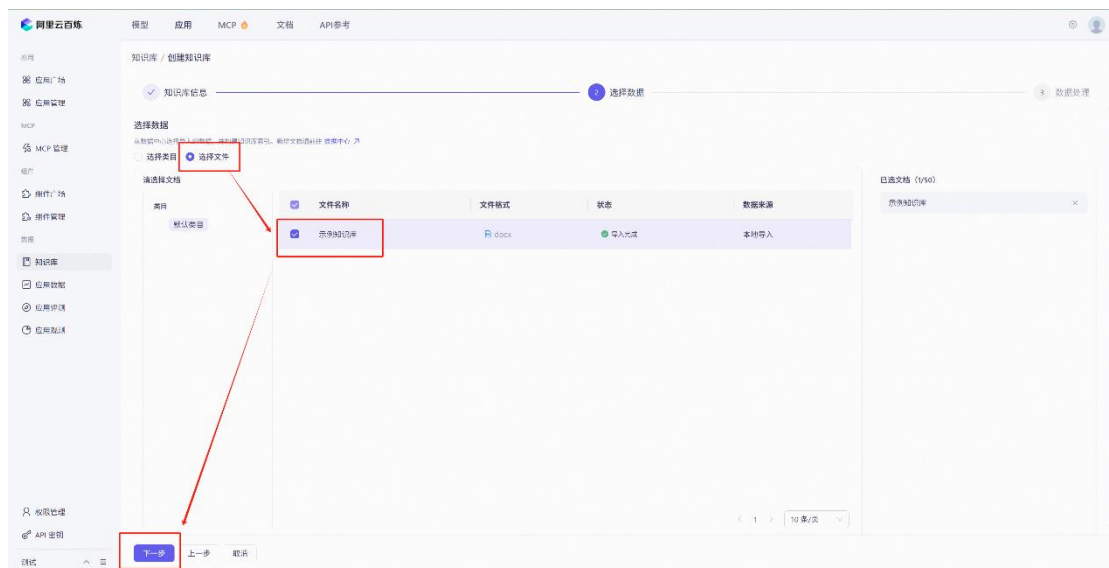


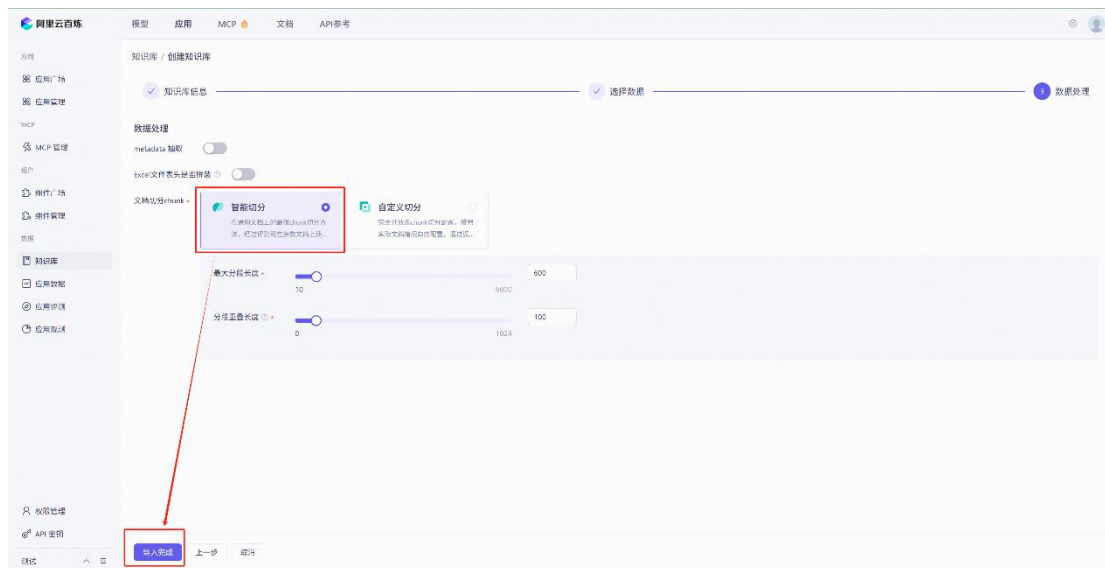
5. 填写**知识库名称**：**学校信息库**，其它参数保持**默认**即可，单击**下一步**。

为了更好地区分不同的知识库，建议填写知识库描述；选择推荐配置；相似度阈值越高，模型可以从知识库中获取到的知识越精确，但是可能会丢失部分信息，相似度阈值越低，模型可以从知识库中获取的知识越多，但是可能会引入无用的知识，对模型生成的回复造成干扰，建议使用默认的阈值。



6. 单击**选择文件**，在默认类目中选中上传的示例文档，若有多个知识库文档，可以进行多选，单击**下一步**。在**数据处理区域**选择**智能切分**，单击**导入完成**。





7. 当看到状态为**解析完成**时，表示知识库创建完成；单击右侧的**查看-查看切片**即可查看切分完成的文本块。



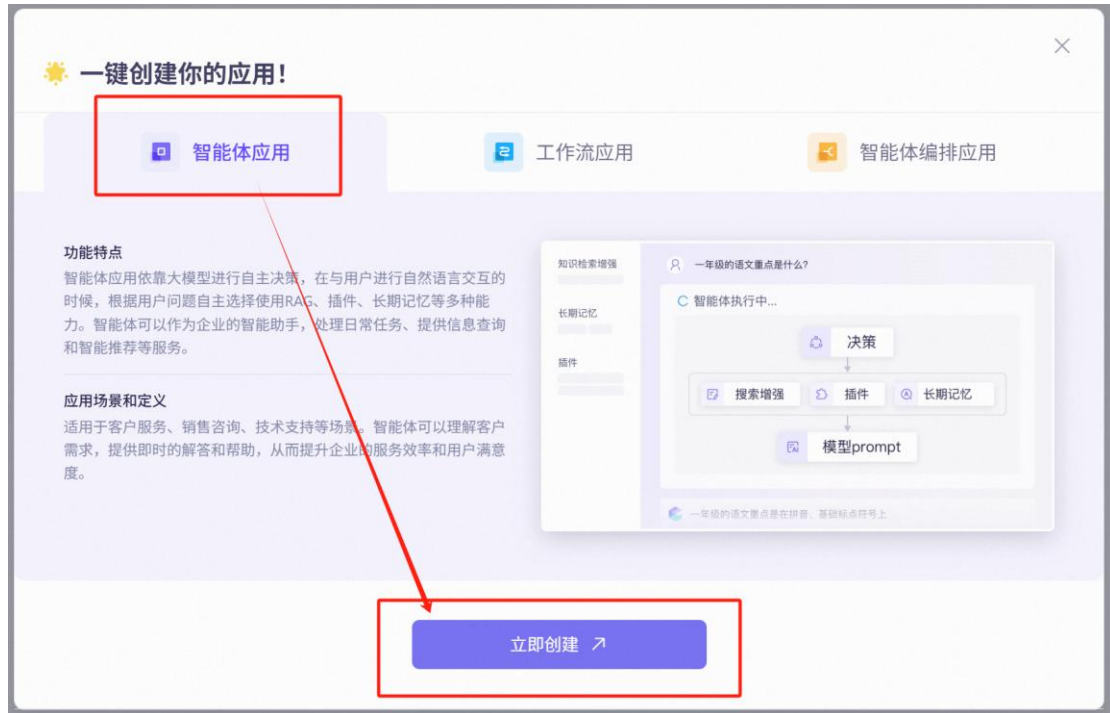


五、创建 RAG 应用

完成知识库的创建后，我们可以创建一个 RAG 应用，用于回答私有知识：

1. 单击左侧边栏的应用中心-我的应用，单击新增应用>直接创建。





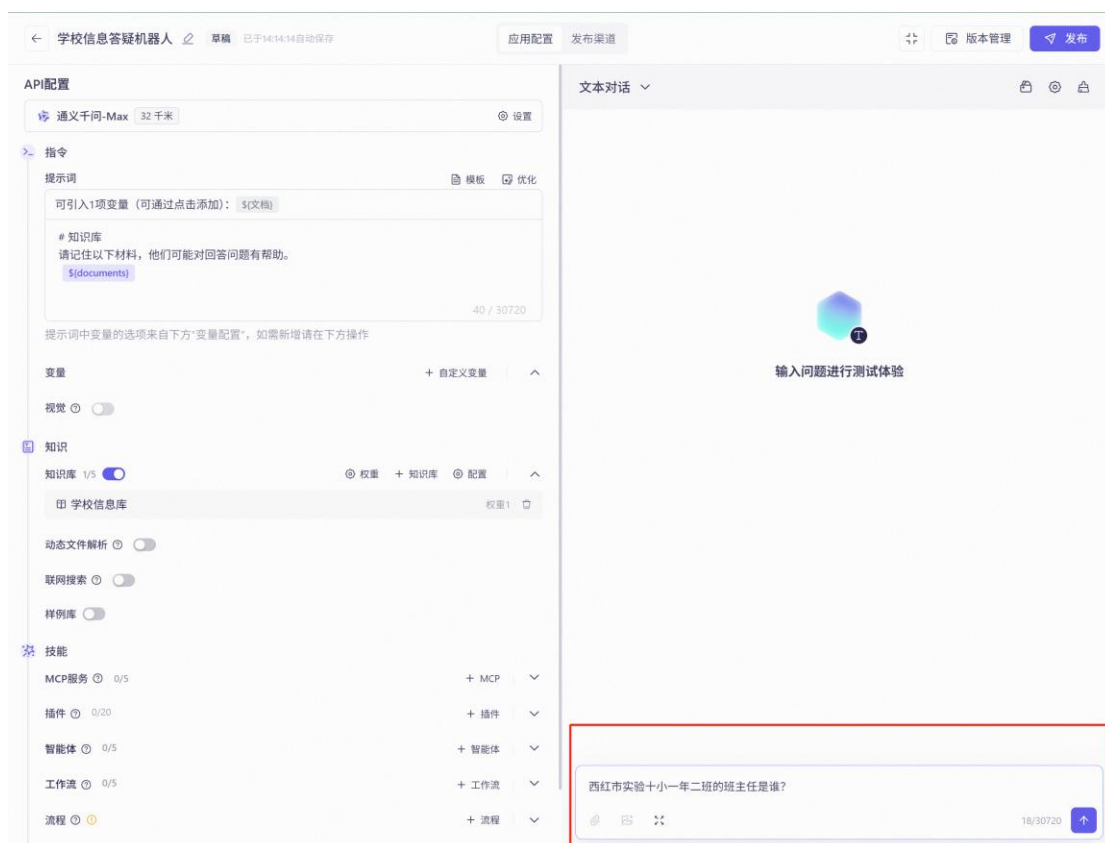
2. 应用信息如下，配置好后点击单击发布。

- 应用名称：示例名称-学校信息答疑机器人
- 模型：在模型下拉列表可以查看并选择**通义千问系列模型**
- 知识检索增强：**开启**，Prompt 栏中会自动填充内容
- 选择知识库：选择创建好的知识库（**学校信息库**）



3. 创建好 RAG 应用后，我们可以再次尝试提问，看看现在大模型是否能正确回答这个问题。在输入框进行提问：

西红市实验十小一年二班的班主任是谁？



4. 可以看到，开启知识检索增强的应用已经能够成功回答该问题了。



六、RAG 应用案例：公司财务数据分析

【背景知识】

本例中，我们将按照前面介绍的创建 RAG 应用的方法，创建名为“**xx 公司**”的**专有财务数据**知识库，并通过专有知识库内的信息完成相关的财务数据分析工作。

【实验步骤】

1. 根据前面创建知识库中导入数据的方法，导入如下财务数据知识库文件。
 - [xx 公司-利润表](#)
 - [xx 公司-现金流量表](#)
 - [xx 公司-资产负债表](#)

← 导入数据

已上传文档 (4/10000)

导入类目:

默认类目

类目类型:

本地类目

数据类型:

文档

文档包括PDF、Doc、Docx等非结构化文档，系统将统一解析格式，提取文档内容，处理为应用。

导入方式:

本地上传

OSS

点击或将文件拖拽到这里上传 (3/200)

支持扩展名: pdf、doc、docx

单文档最大限制100MB或1000页

xx公司-资产负债表.docx

xx公司-现金流量表.docx

xx公司-利润表.docx

文档识别:

文档智能解析

确认

取消

2. 根据前面创建知识库中创建知识库的方法，创建名为“xx 公司财务数据
库”的知识库，点击下一步。



知识库基础信息

* 知识库名称:

xx公司财务数据库

9 / 20

知识库描述:

请输入知识库描述。对知识库包含的内容和数据的用途进行描述。

0 / 1000

数据类型

☒ 非结构化数据

☐ 结构化数据 (即将上线, 敬请期待)

知识库配置

* 配置模式:

推荐配置

☒

百炼推荐配置, 在效果、推理成本、检索时延等...

自定义

☐

完全开放的离线知识库配置, 按照检索需求自由...

多轮对话改写 ①

开启

下一步

创建空知识库

取消

- 选择数据通过“选择文件”方式，勾选“xx 公司-利润表”、“xx 公司-现金流量表”、“xx 公司-资产负债表”三个表格，点击下一步完成。

创建知识库

✓ 知识库信息

2 选择数据

3 数据处理

选择数据

从数据中心选择导入的数据, 并构建知识库索引。新增文档请前往 [数据中心](#)

☐ 选择类目

☒ 选择文件

请选择文档

类目

默认类目

文件名称	文件格式	状态	数据来源
<input checked="" type="checkbox"/> xx公司-利润表	docx	导入完成	本地导入
<input checked="" type="checkbox"/> xx公司-资产负债表	docx	导入完成	本地导入
<input checked="" type="checkbox"/> xx公司-现金流量表	docx	导入完成	本地导入

已选文档 (3/50)

xx公司-利润表

xx公司-现金流量表

xx公司-资产负债表

下一步

上一步

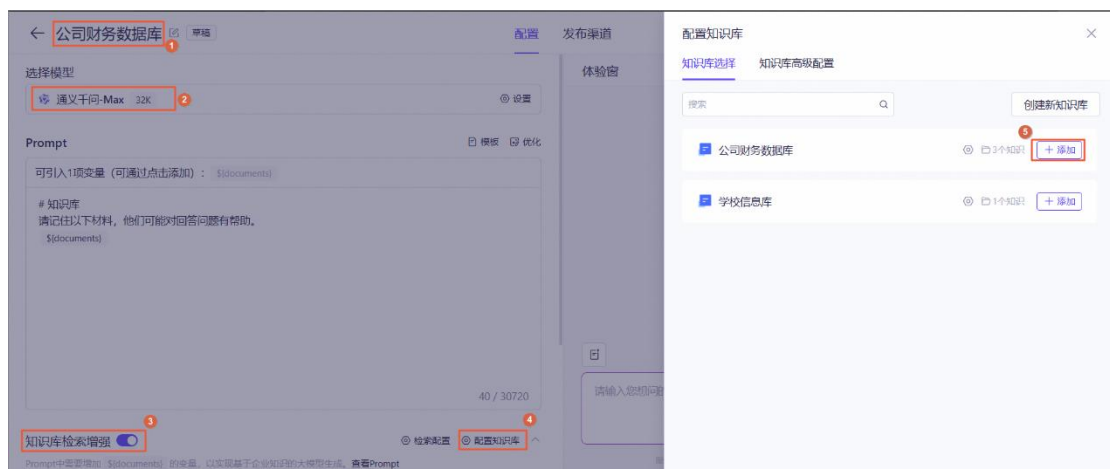
取消

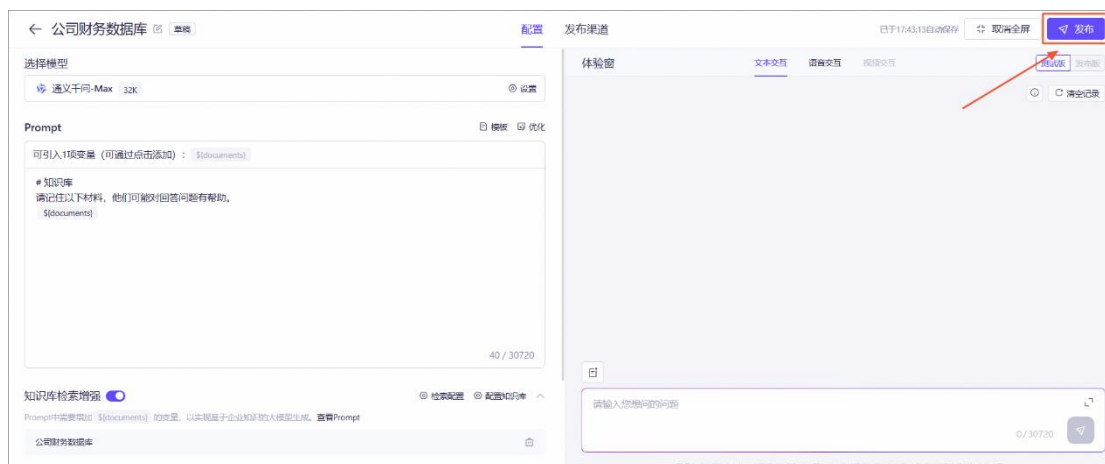
- 数据处理部分保持默认，单击导入完成。

20



5. 根据前面创建 RAG 应用的方法，创建名为“xx 公司财务数据分析助手”的应用，勾选知识检索增强，知识库选择“xx 公司财务数据库”，点击保存并发布。

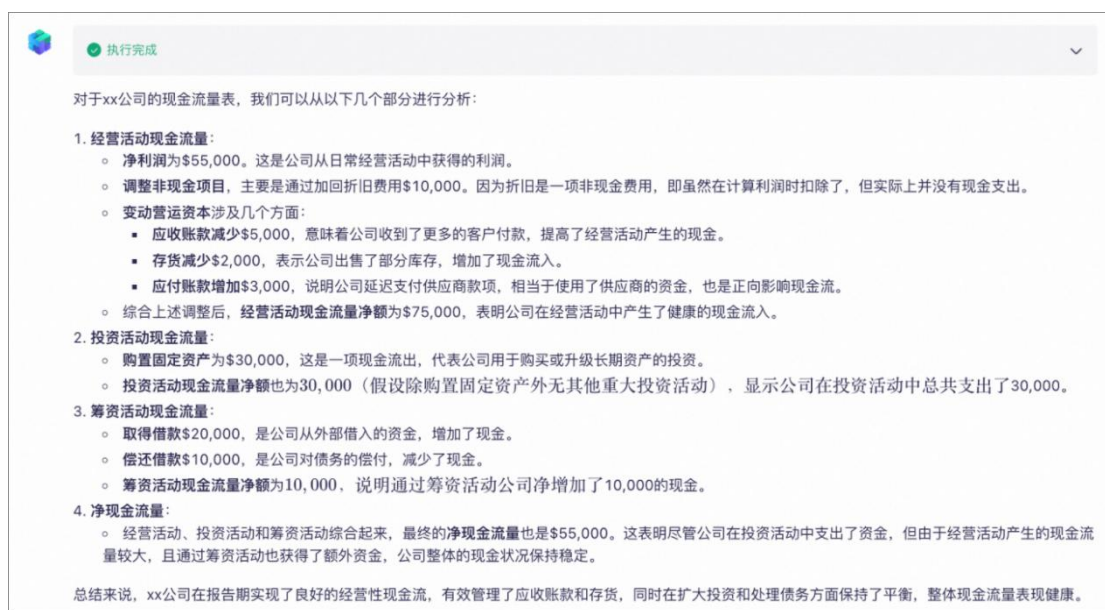




6. 在右侧边栏**体验窗**区域，切换为**发布版**。在输入框进行提问，提示词 1 如下。

请分析 xx 公司的现金流量

参考生产图如下。



7. 在输入框进行提问，提示词 2 如下。

请分析一下 xx 公司的经营状况

参考生产图如下。

 执行完成

根据提供的财务三表信息，我们可以从以下几个方面分析xx公司的经营状况：

1. 资产负债表

总资产：公司总资产为\$195,000，表明公司的总体资源规模。

流动资产：流动资产总额为105,000，其中现金50,000，说明公司拥有良好的短期偿债能力。应收账款30,000和存货20,000，显示公司的运营活跃度和销售循环状态。

非流动资产：主要包括固定资产净额80,000（固定资产原值100,000减去累计折旧20,000）和无形资产10,000，反映公司长期投资和无形资产的价值。

总负债：公司总负债为70,000，其中流动负债30,000和非流动负债（长期借款）40,000，表明公司有一定的债务负担，但短期内流动性较好。

股东权益：股东权益总计125,000，显示了投资者在公司净资产中的份额，包括股本50,000和留存收益\$75,000，这表明公司有较好的积累和盈利能力。

2. 利润表

总收入：销售收入为\$150,000，体现了公司的市场销售能力和业务规模。

净利润：净利润为\$55,000，显示出公司在报告期内的盈利水平较高，经营效率良好。

成本与费用：销售成本50,000，占收入的三分之一，加上营业费用、折旧费用、利息费用等总费用45,000，说明公司控制成本和费用的能力较强。

3. 现金流量表

经营活动现金流：净额为\$75,000，表明公司从日常运营中获得了正向现金流入，运营能力强。

投资活动现金流：投资活动净流出\$30,000，主要是购置固定资产，显示公司正在投资于长期资产以支持未来增长。

筹资活动现金流：净流入\$10,000，意味着公司通过取得借款等活动增加了现金，同时也有偿还借款的流出，显示公司适度利用财务杠杆进行扩张。

净现金流量：净现金流量也为\$55,000，与净利润数额一致，说明公司利润转化为现金的能力强，财务健康。

综合分析，xx公司展现出了较强的盈利能力、良好的现金流状况以及稳健的财务结构。公司能够有效管理其资产和负债，并且在扩大投资的同时保持了正向的现金流和盈利。不过，应注意长期借款的增加可能带来的利息负担和未来还款压力，以及持续投资是否能带来预期的回报。

可以看到，xx 公司财务数据分析助手已经可以完成现金流量和经营状况的财务分析工作。

七、RAG 应用案例：企业差旅规定

【背景知识】

本例中，我们将按照前面介绍的创建 RAG 应用的方法，创建名为“xx 公司”的差旅规定的知识库，并通过专有知识库内的信息完成 xx 公司差旅规则解析的 RAG 应用。

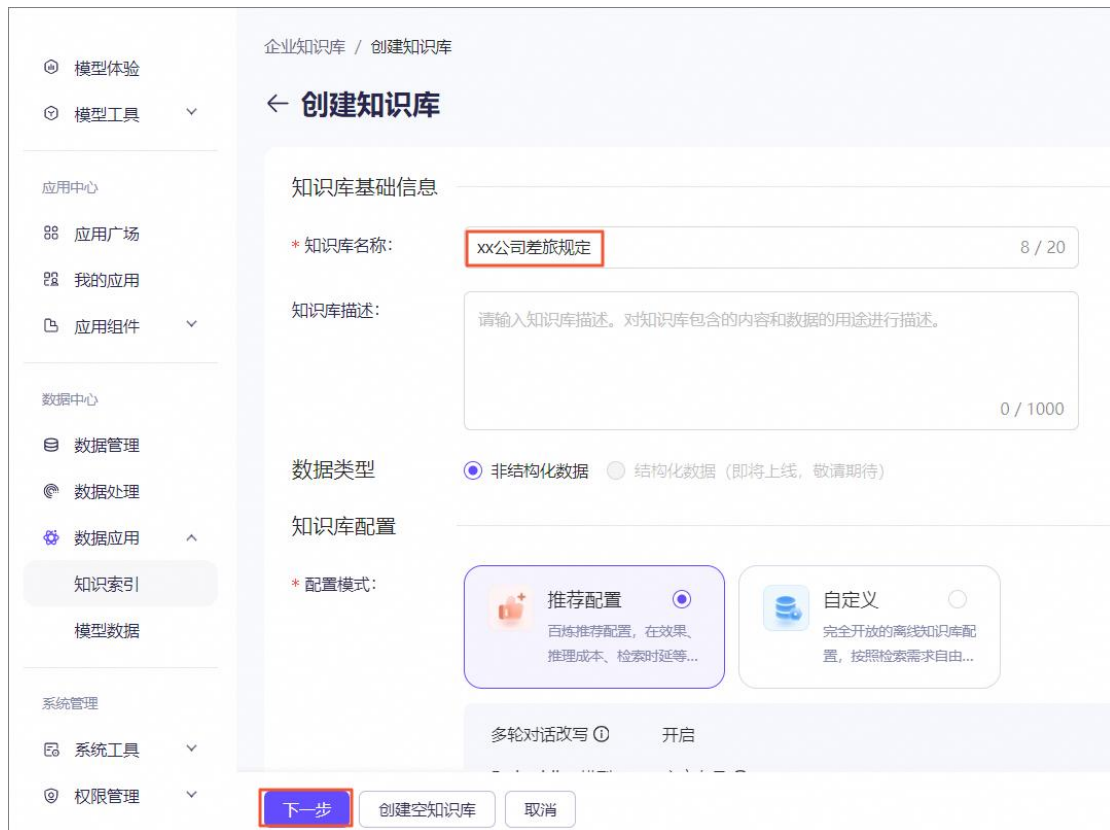
【实验步骤】

1. 根据前面创建知识库中导入数据的方法，导入如下企业差旅规定知识库文件。

- xx 公司-差旅规定



2. 根据前面创建知识库中创建知识库的方法，创建名为“xx 公司差旅规定”的知识库，点击下一步。



3. 选择数据通过“选择文件”方式，勾选“xx 公司-差旅规定”表格，点击下一步。在数据处理区域选择智能切分，单击导入完成。

← 创建知识库

✓ 知识库信息

2 选择数据

3

选择数据

从数据中心选择导入的数据，并构建知识库索引。新增文档请前往 [数据中心](#) ↗

☐ 选择类目

☒ 选择文件

请选择文档

类目

默认类目

<input checked="" type="checkbox"/> 文件名称	文件格式	状态	数据来源
<input checked="" type="checkbox"/> xx公司-差旅规定	docx	✓ 导入完成	本地导入
<input type="checkbox"/> xx公司-利润表	docx	✓ 导入完成	本地导入
<input type="checkbox"/> xx公司-资产负债表	docx	✓ 导入完成	本地导入
<input type="checkbox"/> xx公司-现金流量表	docx	✓ 导入完成	本地导入
<input type="checkbox"/> 示例知识库 (1)	docx	✓ 导入完成	本地导入

下一步

上一步

取消

数据处理

metadata抽取 ☐

Excel文件表头是否拼装 ☐

文档切分chunk *

智能切分

自定义切分

在通用文档上的最优chunk切分方法，经过评测可在多数文档上获...

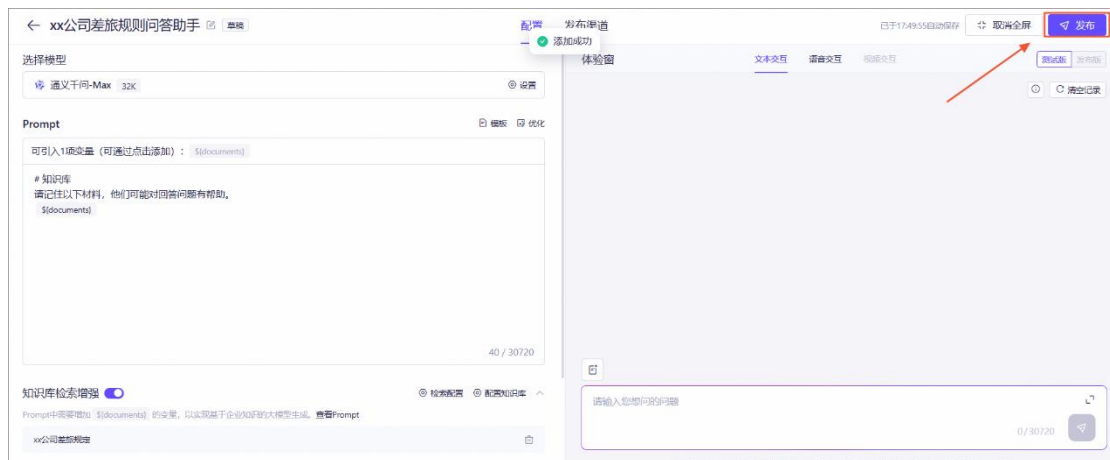
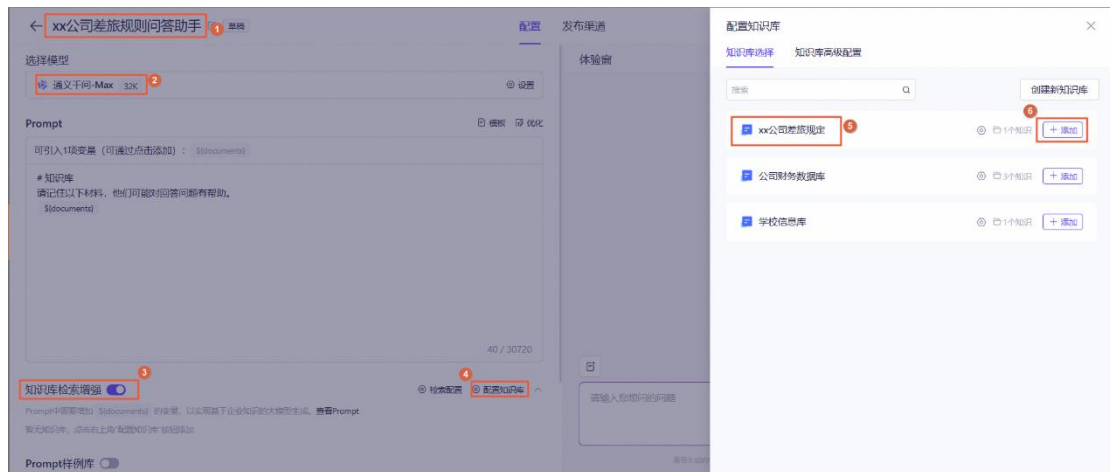
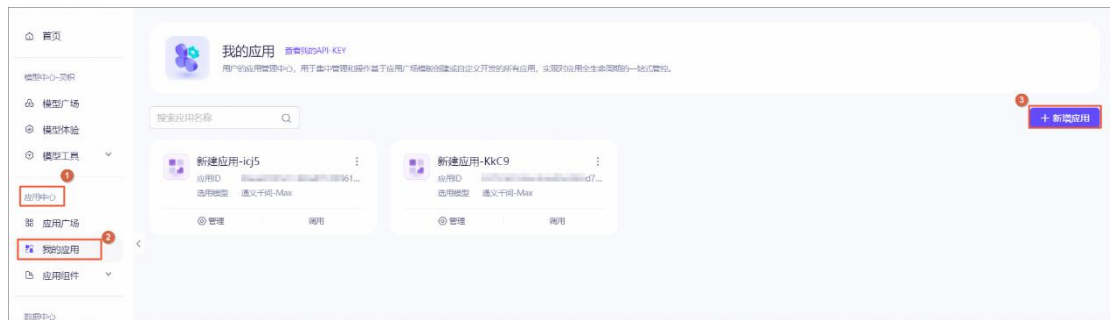
完全开放的chunk切分配置，按照实际文档情况自由配置，通过调...

导入完成

上一步

取消

4. 根据前面创建 RAG 应用的方法，创建名为“xx 公司差旅规则问答助手”的应用，勾选知识检索增强，知识库选择“xx 公司差旅规定”，点击保存并发布。



5. 在右侧边栏**体验窗**区域，切换为**发布版**。在输入框进行提问，提示词 1 如下。

去北京出差可以订多少钱的酒店？

参考生成图如下。



6. 在输入框进行提问, 提示词 2 如下。

去杭州出差可以订多少钱的酒店?

参考生成图如下。



7. 在输入框进行提问, 提示词 3 如下。

去上海出差中午的餐补是多少?

参考生成图如下。



8. 在输入框进行提问，提示词 4 如下。

出差可以坐高铁一等座吗？

参考生成图如下。



可以看到，xx 公司差旅规则问答助手已经可以根据用户的提问完成差旅规则解读和合理建议。