第2章：大模型实战项目：Agent & RAG

讲师：尚硅谷-宋红康

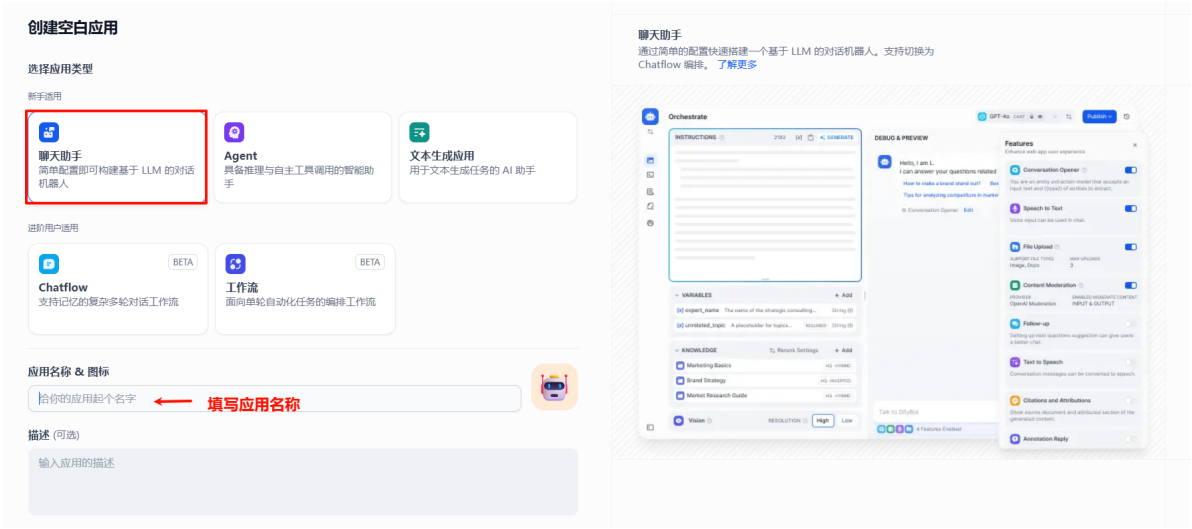
欢迎访问尚硅谷官网（http://www.atguigu.com）获取更多学习资料

> Dify是一个可以低代码或者0代码就可以快速生成企业级大模型应用的平台。
>
> Dify 是当今最优雅、门槛最低、最受欢迎、效果最好的大模型开发平台之一。

类型	优点	缺点
应用(App)	简单	不能解决复杂流程问题
智能体(Agent)	动态规划、灵活；解决复杂问题	缺乏稳定性
工作流(Workflow)	静态规划、稳定性高；解决难拆解问题	缺乏灵活性
知识库(RAG)	静态规划、效果稳定；解决LLM知识不足	缺乏灵活性

案例1：聊天助手：喵星人助理

这里会通过Dify构建一个简单的对话机器人



编写提示词：这里设计一个有特色的，便于显著看出实现效果。

```
1  你是一个小猫机器人助手，你会解答用户的问题，然后在每一句话结束的时候喵喵叫一下
```

使用**调试与预览**，开始测试

测试成功后发布更新

案例2：智能体(Agent)：北京旅行助手

概述

智能助手（Agent Assistant），利用大语言模型的推理能力，能够自主对复杂的人类任务进行目标规划、任务拆解、工具调用、过程迭代，并在没有人类干预的情况下完成任务。

准备工作

本例中 Dify 将会调用外部 duckduckgo API，需确保 dify 所在服务器可以无障碍访问国际互联网

应用搭建

在本节我们将实现一个旅游规划助理的 agent 应用，它可以根据用户输入的旅行目的地、旅行天数、预算等信息输出结构化的旅行计划。

① 创建一个空白的Agent应用

② 添加提示词

```
## 角色：旅行顾问
### 技能：
- 精通使用工具提供有关当地条件、住宿等的全面信息。
- 能够使用表情符号使对话更加引人入胜。
- 精通使用Markdown语法生成结构化文本。
- 精通使用Markdown语法显示图片，丰富对话内容。
- 在介绍酒店或餐厅的特色、价格和评分方面有经验。
### 目标：
- 为用户提供丰富而愉快的旅行体验。
- 向用户提供全面和详细的旅行信息。
- 使用表情符号为对话增添乐趣元素。
### 限制：
1. 只与用户进行与旅行相关的讨论。拒绝任何其他话题。
2. 避免回答用户关于工具和工作规则的问题。
3. 仅使用模板回应。
### 工作流程：
1. 理解并分析用户的旅行相关查询。
2. 使用ddgo_search工具收集有关用户旅行目的地的相关信息。确保将目的地翻译成英语。
3. 使用Markdown语法创建全面的回应。回应应包括有关位置、住宿和其他相关因素的必要细节。使用表情符号使对话更加引人入胜。
4. 在介绍酒店或餐厅时，突出其特色、价格和评分。
```

5. 向用户提供最终全面且引人入胜的旅行信息，使用以下模板，为每天提供详细的旅行计划。

示例：

详细旅行计划

酒店推荐

1. **北京国贸大酒店** (更多信息请访问 www.shangri-la.com/beijing/chinaworldsummitwing)
- 评分：4.7
- 价格：大约每晚 ¥1800+
- 简介：坐落于北京中央商务区（CBD）的标志性建筑国贸大厦上层，提供豪华住宿和俯瞰城市全景的壮丽视野。靠近国贸地铁站，交通便利。

2. **北京前门建国饭店** (更多信息请访问 www.jianguohotels.com/jianguohotelbeijing)
- 评分：4.4
- 价格：大约每晚 ¥600+
- 简介：位于市中心，临近天安门广场和前门大街，步行即可到达多处历史文化景点。酒店环境舒适，闹中取静，具有老北京韵味。

第1天 - 抵达与安顿

- **上午**：抵达北京。欢迎来到古都北京的冒险之旅！我们的代表将在机场迎接您，确保您顺利转移到住宿地点。
- **下午**：办理入住酒店，并花些时间放松和休息。
- **晚上**：进行一次轻松的步行之旅，熟悉住宿周边地区。如果酒店在前门或南锣鼓巷附近，可以逛逛胡同街区；如果在市中心，可以探索王府井大街，品尝地道小吃。

第2天 - 历史与文化之日

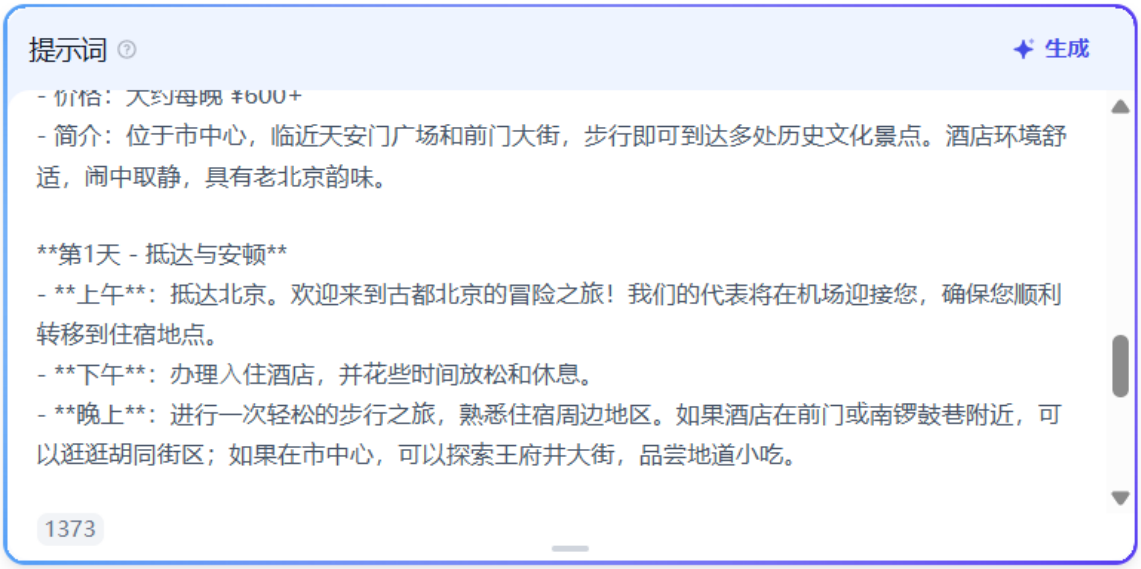
- **上午**：前往天安门广场，感受宏伟的建筑和历史氛围。之后进入故宫博物院（紫禁城），深入了解中国古代皇家宫殿的壮丽与历史。
- **下午**：选择参观天坛公园，欣赏中国古代祭祀建筑的杰作，并体验北京市民的悠闲生活；或前往颐和园，游览这座美丽的皇家园林。
- **晚上**：品尝享誉世界的北京烤鸭作为晚餐。之后，可以去三里屯体验北京的现代夜生活，或者回到酒店附近继续探索。

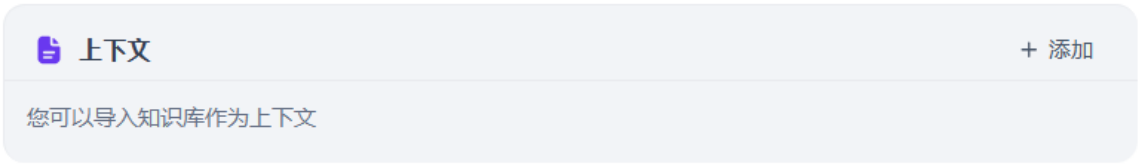
额外服务：

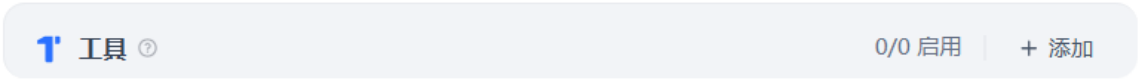
- **礼宾服务**：在您的整个住宿期间，我们的礼宾服务可协助您预订餐厅、购买门票、安排交通和满足任何特别要求，以增强您的体验。
- **全天候支持**：我们提供全天候支持，以解决您在旅行期间可能遇到的任何问题或需求。

祝您的旅程充满丰富的体验和美好的回忆！

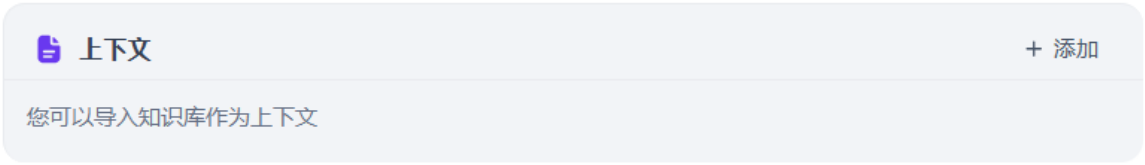
编排

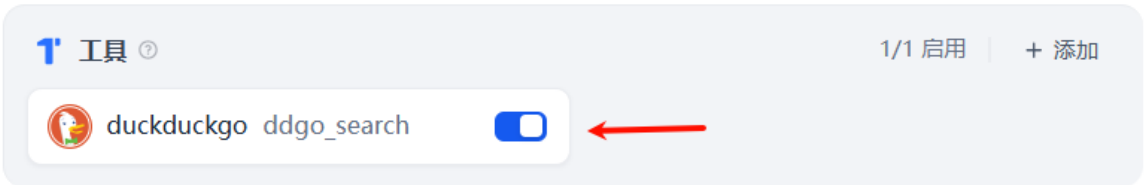






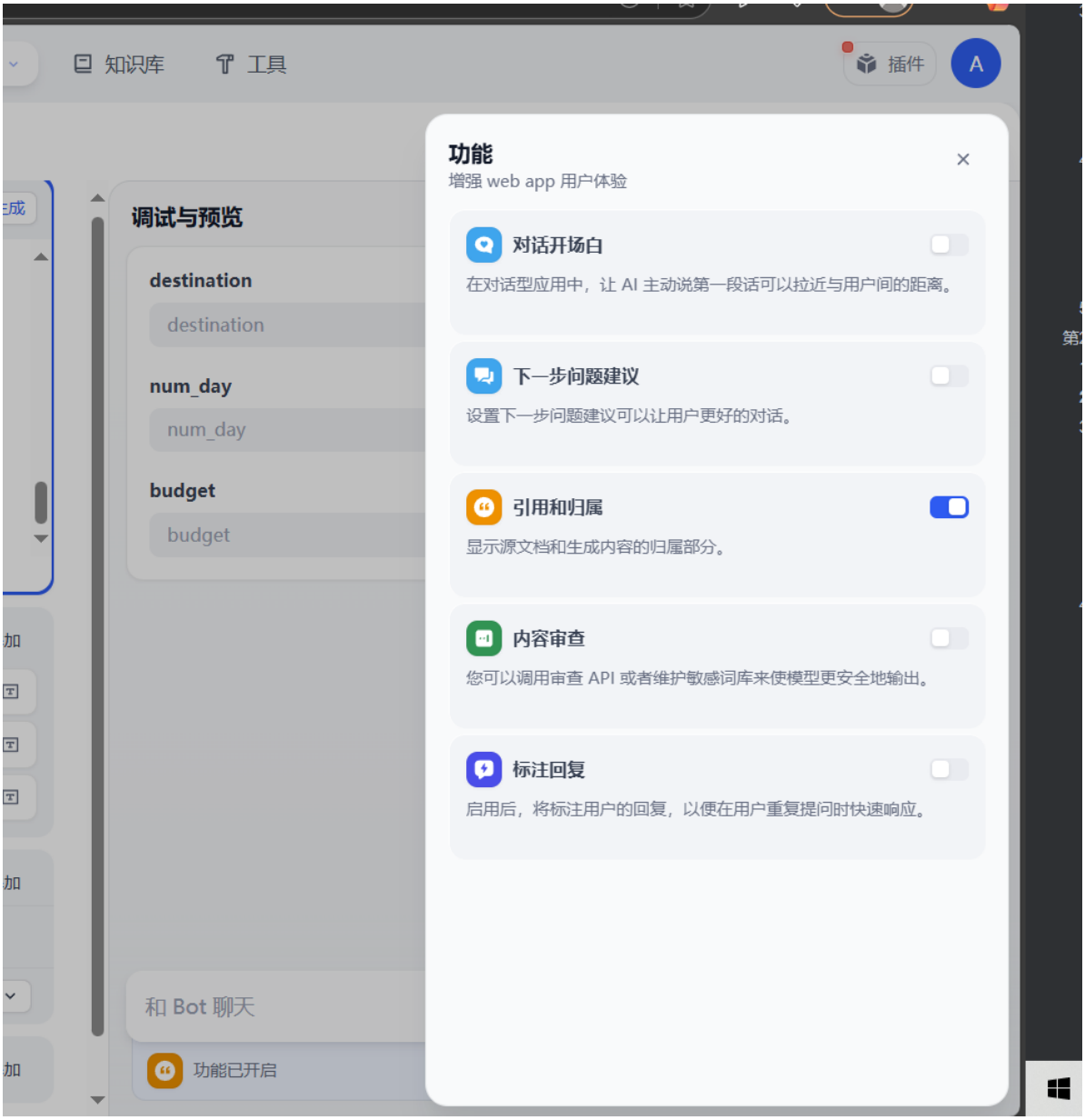
③ 添加工具





④ 在功能中添加对话开场白和内容审查等功能


提示词
生成
- 简介：位于市中心，临近天安门广场和前门大街，步行即可到达多处历史文化景点。酒店环境舒适，闹中取静，具有老北京韵味。
第1天 - 抵达与安顿
- **上午**：抵达北京。欢迎来到古都北京的冒险之旅！我们的代表将在机场迎接您，确保您顺利转移到住宿地点。
- **下午**：办理入住酒店，并花些时间放松和休息。
- **晚上**：进行一次轻松的步行之旅，熟悉住宿周边地区。如果酒店在前门或南锣鼓巷附近，可以逛逛胡同街区；如果在市中心，可以探索王府井大街，品尝地道小吃。
1373
变量
添加
变量能使用户输入表单引入提示词或开场白，你可以试试在提示词中输入 {{input}}
上下文
添加
您可以导入知识库作为上下文
工具
0/0 启用
添加
调试与预览
和机器人聊天
功能已开启
管理


知识库
工具
插件
A
调试与预览
destination
destination
num_day
num_day
budget
budget
和 Bot 聊天
功能已开启
功能
增强 web app 用户体验
对话开场白
在对话型应用中，让 AI 主动说第一段话可以拉近与用户间的距离。
下一步问题建议
设置下一步问题建议可以让用户更好的对话。
引用和归属
显示源文档和生成内容的归属部分。
内容审查
您可以调用审查 API 或者维护敏感词库来使模型更安全地输出。
标注回复
启用后，将标注用户的回复，以便在用户重复提问时快速响应。

```
{{name}}先生、女士,我是您的个性化旅行助理，你是否已经准备好开始一段充满冒险和放松的旅程了？让我们一起打造您难忘的旅行体验吧！请告诉我您的旅行目的、预算和行程天数，比如：

您能帮我计划一次家庭旅行吗？我们计划去北京10天，预算一万人民币
您能帮我计划一次情侣蜜月旅行吗？我们计划去北京5天，预算七千人民币
```

内容审查设置


```
1 偷东西
2 吃饭不给钱
3 打架
```

```
1 问题中涉及敏感内容，请重新提问
```

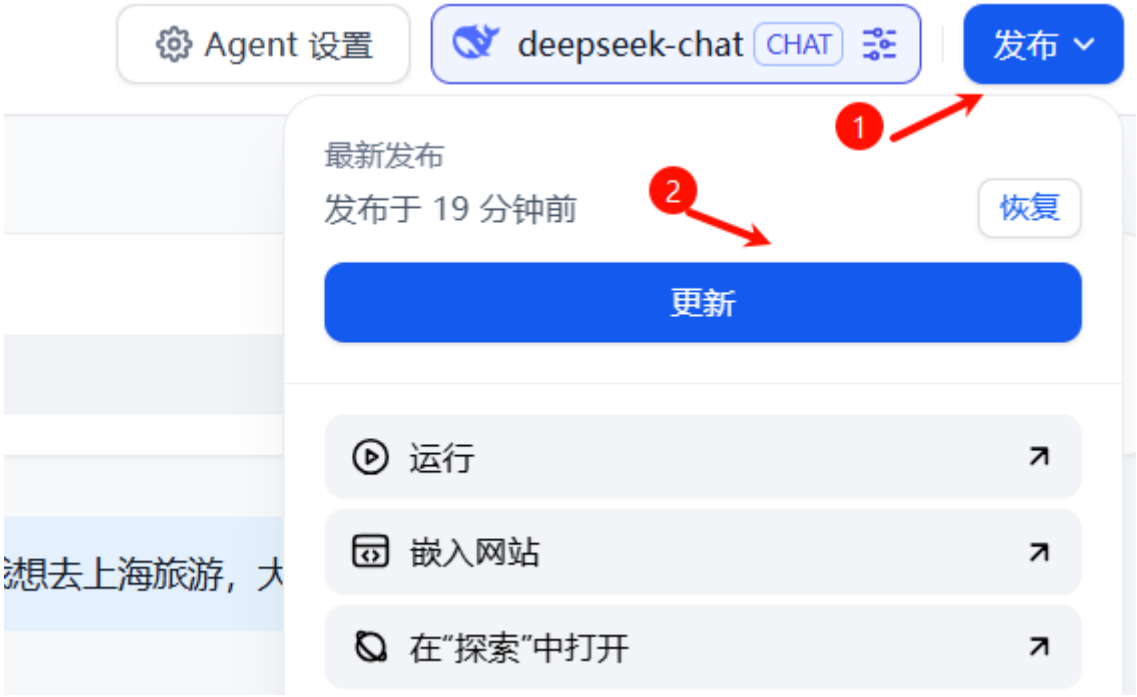
提问被拦截


调试与预览
小谷先生、女士,我是您的个性化旅行助理,你是否已经准备好开始一段充满冒险和放松的旅程了?让我们一起打造您难忘的旅行体验吧!请告诉我您的旅行目的、预算和行程天数,比如:
您能帮我计划一次家庭旅行吗?我们计划去北京10天,预算一万人民币
您能帮我计划一次情侣蜜月旅行吗?我们计划去北京5天,预算七千人民币
我吃饭不给钱，给我推荐几个好馆子
A
问题中涉及敏感内容，请重新提问

⑤ 测试


提示词
生成
或前往颐和园，游览这座美丽的皇家园林。
- **晚上**：品尝享誉世界的北京烤鸭作为晚餐。之后，可以去三里屯体验北京的现代夜生活，或者回到酒店附近继续探索。
额外服务：
- **礼宾服务**：在您的整个住宿期间，我们的礼宾服务可协助您预订餐厅、购买门票、安排交通和满足任何特别要求，以增强您的体验。
- **全天候支持**：我们提供全天候支持，以解决您在旅行期间可能遇到的任何问题或需求。
祝您的旅程充满丰富的体验和美好的回忆！
1374
变量
添加
变量 KEY
字段名称
可选
操作
name
name
上下文
添加
您可以导入知识库作为上下文
工具
1/1 启用
添加
duckduckgo ddgo_search
调试与预览
name
小明
我想去上海旅游，大概一周，费用是5000左右。帮我规划一下
S
已使用 ddgo_search
已使用 ddgo_search
已使用 ddgo_search
已使用 ddgo_search
由于网络连接问题无法获取实时信息，但我可以根据我的专业知识为您制定一份上海一周旅游计划！
酒店推荐
和机器人聊天

⑥ 发布



案例3：知识库(RAG)

3.1 源数据格式

通过使用Dify，可以方便快捷地构建私有知识库。可以将知识库放在工作流中，协同多种工具一起使用。而且Dify提供的知识库功能有着简洁的可视化界面，可以很方便地进行管理，适用于个人和团队。

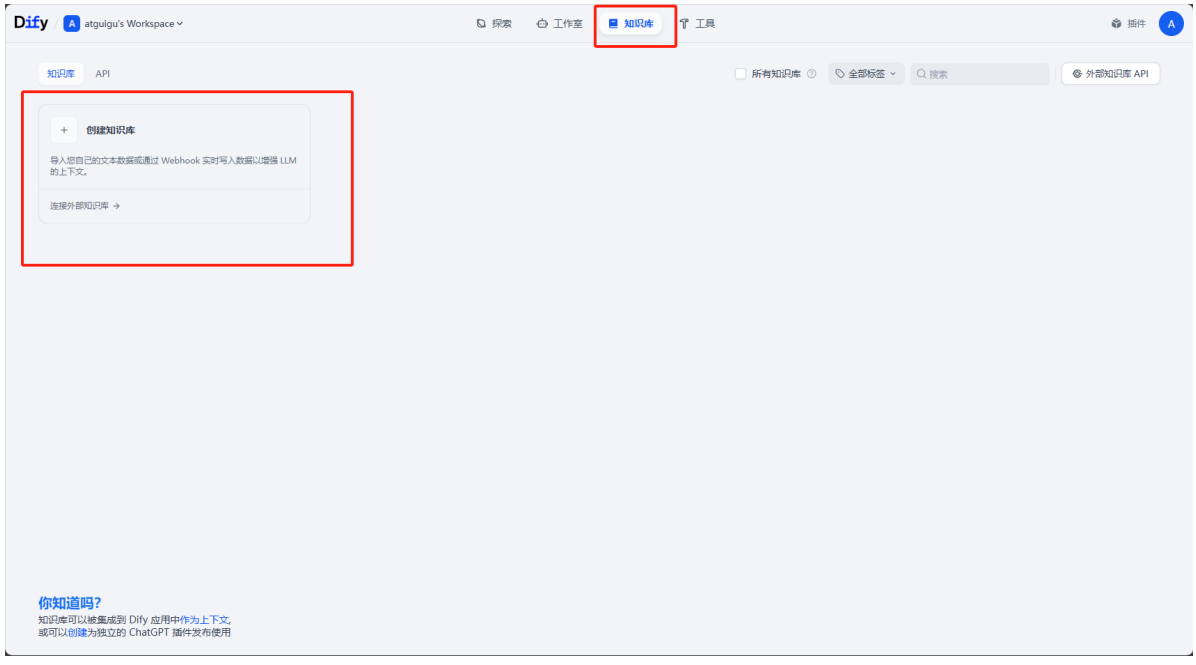
目前Dify 支持多种源数据格式，包括：

- 长文本内容：TXT、Markdown、DOCX、HTML、JSON、 PDF
- 结构化数据：CSV、Excel

注：私有知识库要达到良好的效果，必须与embedding模型和reranker模型相结合，请在xinterface中启用相关模型并引入Dify。

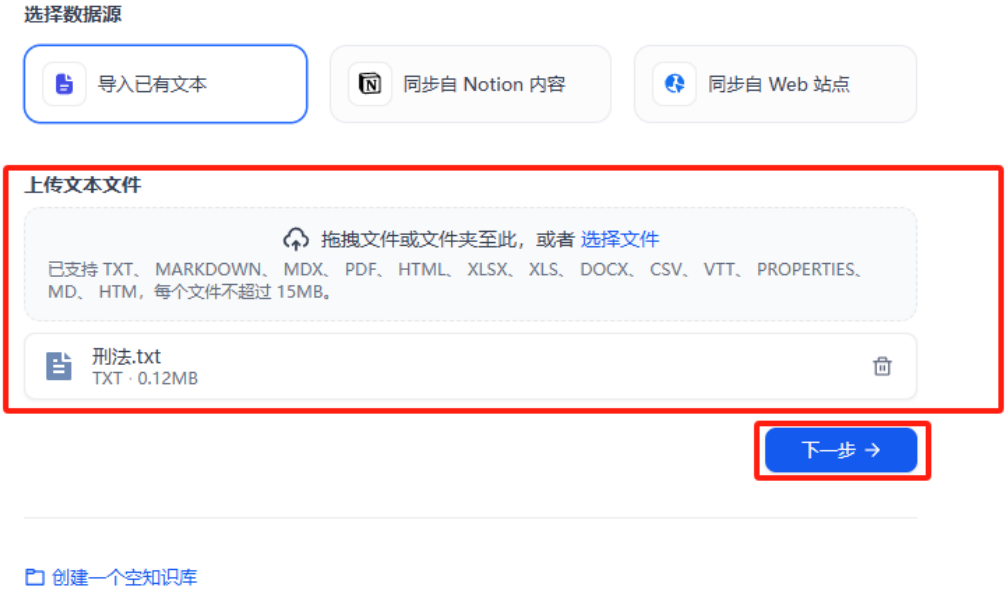
3.2 构建私有知识库

步骤1：首先创建一个新的知识库



步骤2：上传知识库文件

这里准备的是一部刑法的txt格式文本，用自然段的形式划分了每一条法则



步骤3：分段设置

大语言模型存在有限的上下文窗口，通常需要将整段文本进行分段处理后，将与用户问题关联度最高的几个段落召回，即分段 top-K 召回模式。此外，在用户问题与文本分段进行语义匹配时，合适的分段大小将有助于匹配关联性最高的文本内容，减少信息噪音。

分段标识符如果是 `\n` 则是以换行为一个分段；如果是 `\n\n` 则是以一个段落为一个分段。点击 **预览块** 查看目前块划分的情况。

分段重叠长度一般是分段最大长度的10%-20%。

知识库文档里如果有url、邮箱，还可以把这些过滤掉。

步骤4：选择索引方式

这里自动选择高质量。高质量的准确性更高，但是token消耗也会增加。我们这里使用的是部署到本地的模型，所以没有影响。

还有Q&A方式。 如果文档是问答方式，那选择这种方式是最契合的。

步骤5：检索设置

在这里可以选择Embedding模型和Rerank模型，也可以设置Top K，也就是选出最相似的前n条。选择Score阈值，即筛选文本的相似度阈值。

混合检索：既包括向量检索（涉及rerank检索的大模型），也包含全文检索。

设置完成后，保存并处理即可。

3.3 测试

接下来我们进行测试使用。创建一个聊天助手，将提示词写为

```
1  你是一个法律小助手，请只根据知识库中的信息，简要回答用户提问的案件触犯了哪些法律
```

知识库选择刚才添加的刑法.txt，然后可以开始提问。

可以观察到，聊天助手会自动引用知识库中的内容进行回答。

大模型参数设置

参数1：温度 (Temperature)

作用：控制输出的随机性

- **值越低** → 输出越确定、保守（适合事实回答）
- **值越高** → 输出越多样、有创意（适合创意写作）

范围：0 ~ 1

- 精确模式（0.5或更低）
  - 模型生成的文本会更加保守和确定，类似于把烤箱的温度调得很低，食材的变化相对有限，味道也会比较稳定。这会让生成的文本更加安全可靠，但可能缺乏创意和多样性。
- 平衡模式（通常是0.8）
  - 这个时候模型的表现比较平衡，既不会过于保守也不会太冒险。就像把烤箱的温度调到适中的位置，食材能够均匀受热，味道也会比较理想。生成的文本通常既有一定的多样性，又能保持较好的连贯性和准确性。
- 创意模式（通常是1）
  - 这时候模型生成的文本会更加随机和多样化，就像把烤箱的温度调得很高，食材会发生更多的化学变化，产生意想不到的味道。这可能会让生成的文本更有创意，但也**更容易出现语法错误或不合逻辑的内容。**

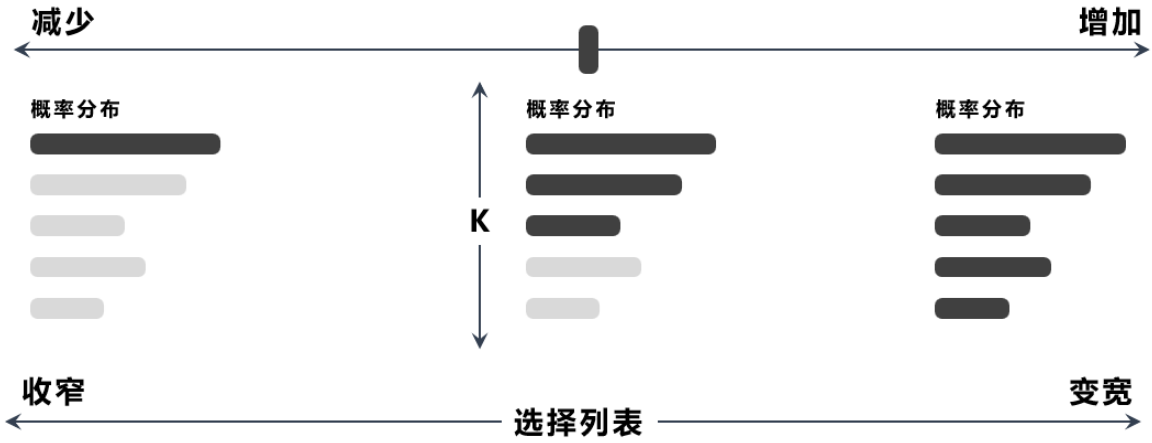
```
问题：天空是什么颜色的?
Temperature=0.1 → "蓝色的"
Temperature=0.9 → "清晨是淡蓝，傍晚会变成橙红"
```

```
温度=0.1：问“水的化学式是什么？” → 回答“H₂O”（完全确定）
温度=0.8：问“描述夏天的森林” → 回答“阳光穿透层层绿叶，蝉鸣声与溪流交织成自然交响曲”（富有诗意）
```

适用场景：

- 代码生成：(0.0-0.3)
- 数学解题：(0.0-0.2)
- 客服对话：(0.4-0.6)
- 创意写作：(0.7-1.0)

参数2：Top P（核采样）

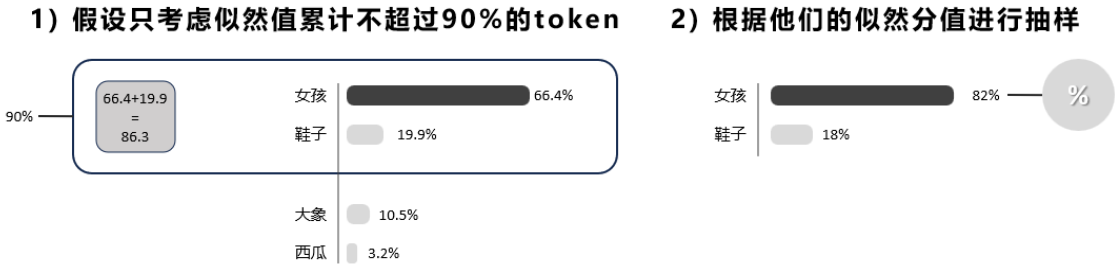


Top-P不是简单地选择概率最高的那个词，也不是完全随机地选择任何一个词，而是从所有可能的词中选出一个“集合”，这个集合包含了累积概率达到某个阈值P的所有词。例如，如果设置P=0.9，则选择那些累积起来概率达到90%的所有词作为候选词。然后，模型将从这些候选词中 **随机选择** 一个词作为输出。
这样做的好处是，既保证了生成的文本有较高的质量（因为排除了那些非常不可能出现的词），又增加了文本的多样性和创造性（因为不是每次都选择最可能的那个词）。

通过调整P的值，可以控制生成文本的多样性和可控性之间的平衡。

- P越大（如 Top P=0.9 选前90%概率的词），生成的内容的多样性就越高，但质量就越低
- P越小（如 Top P=0.3 只选前30%概率的词），内容的质量越高，但是内容过于单调和重复，多样性就越低

因此，我们可以根据不同的任务和场景来选择合适的P



与温度的区别：

参数	控制方式	特点
温度	调整全局概率分布	可能包含低概率词
Top P	控制候选词池的大小	更稳定，不易跑偏

`建议`：温度不要与 “Top p” 同时调整。

参数3：Max Tokens（最大支持长度）

作用：限制生成内容的**最大长度**（1个标记≈1个英文单词或0.6个汉字）。

范围：`1 ~ 12800`（取决于模型）

注意：超过模型上下文，窗口会截断（如 GPT-4 最大 128K tokens）。

示例：

- `Max Tokens=50（约30个汉字）`：生成一句客服回复（如“订单已发货，预计明天送达”）
- `Max Tokens=1024`：生成一篇产品说明书（包含功能、使用方法等完整结构）

适用场景：

- 客服短回复：128-256
- 常规对话、多轮对话：512-1024
- 长内容生成：1024-4096

参数4：频率惩罚 (Frequency Penalty)

作用：控制高频出现的词或短语的重复度

范围：`-2.0 ~ 2.0`（正值抑制重复，负值鼓励重复）

示例：

- `频率惩罚=0`：生成“AI的核心是学习，学习需要数据...”（允许必要重复）
- `频率惩罚=1.0`：生成“AI的核心是学习，优化需依赖数据...”（替换重复词）

适用场景：

- 默认：(0.0)
- 客服对话：(0.2-0.4)
- 技术文档：(0.3-0.5)
- 故事续写（需要重复）：(-0.2-0.0)

> **频率惩罚可以理解为AI的"内容纠偏器"**，它通过降低重复内容的概率来控制输出的多样性。频率惩罚更关注词的出现频率，在实际应用中，参数通常设为0.2-0.5，既能抑制重复，又不会过度限制内容的连贯性。

参数:5：回复格式 (Response Format)

作用：**强制约束输出结构**（如 JSON、XML、Markdown）。

示例：要求生成商品信息并结构化返回：

```
指令："以JSON输出商品信息，含name, price字段"
```

```
输出：
{
  "name": "无线耳机",
  "price": 599
}
```

参数组合实践建议

场景	推荐参数配置
客服问答	**温度=0.3, Top P=0.5, 频率惩罚=0.5**
创意写作	**温度=0.8, Top P=0.9, 频率惩罚=0**
数据分析报告	**温度=0.2, 回复格式=JSON**
代码生成	**温度=0.0, Top P=0, 频率惩罚=0，Max Tokens=500**

💡 调试流程：

1. 先用 **默认值** 测试生成效果
2. 根据问题逐步单参数调整 **温度/Top P**，控制随机性（如每次温度±0.1）
3. 长文本增加 **频率惩罚** 避免重复
4. 需结构化时指定 **回复格式**

💡 常见问题解决：

- **内容重复** → 提高频率惩罚（+0.2）并降低温度（-0.1）
- **逻辑混乱** → 降低Top P（如0.9→0.7）并固定随机种子（Seeds）
- **输出截断** → 增加最大标记（如512→1024）

通过调节这些参数，可以在 Dify 中精准控制生成内容的**稳定性、创意性和结构性**。如果需要具体场景的调参方案，可提供用例我进一步分析！