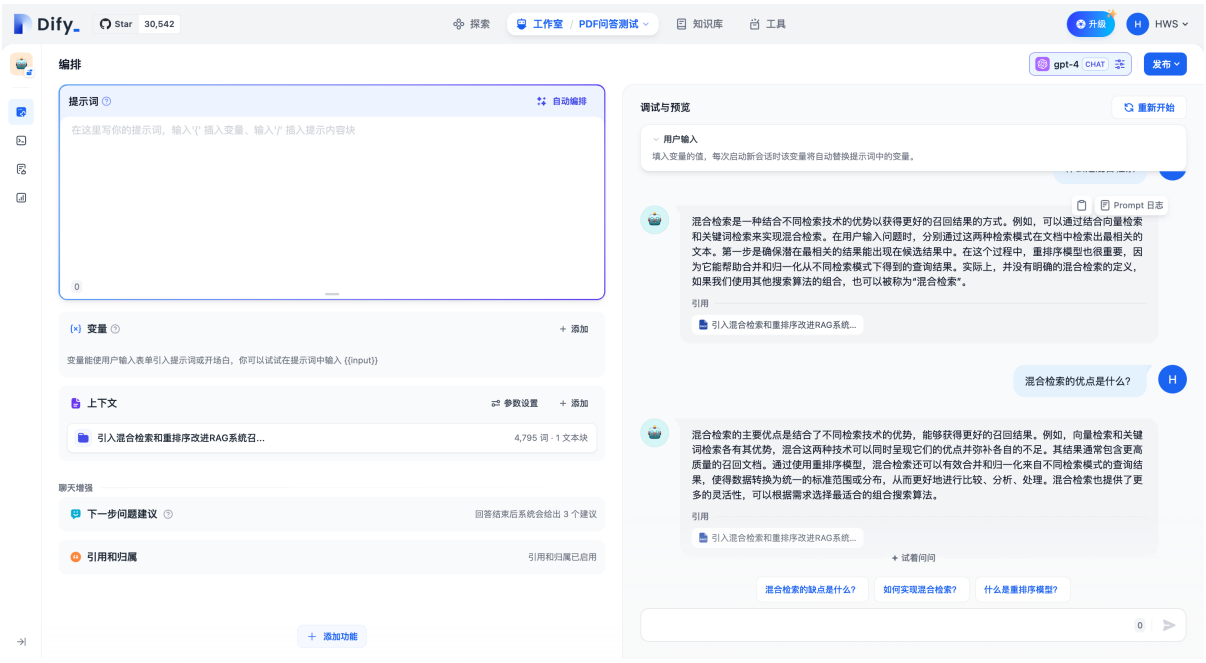
在应用上下文内引用知识库

知识库的引用流程

知识库可以作为外部知识提供给大语言模型用于精确回复用户问题，你可以在 Dify 的所有应用类型内关联已创建的知识库。

以聊天助手为例，使用流程如下：

1. 进入 **工作室 -- 创建应用 --创建聊天助手**
2. 进入 **上下文设置** 点击 **添加** 选择已创建的知识库
3. 在 **上下文设置 -- 参数设置** 内配置**召回策略**
4. 在 **添加功能** 内打开 **引用和归属**
5. 在 **调试与预览** 内输入与知识库相关的用户问题进行调试
6. 调试完成之后**保存并发布**为一个 AI 知识库问答类应用



在应用内关联知识库

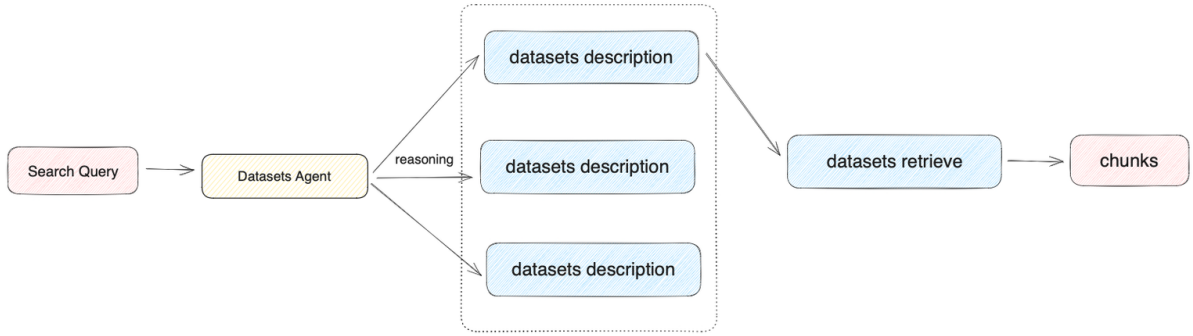
关联知识库并指定召回模式

如果当前应用的上下文涉及多个知识库，需要设置召回模式以使得检索的内容更加精确。进入 **上下文 -- 参数设置 -- 召回设置**，选择知识库的召回模式。

N 选 1 召回（Legacy）

N 选 1 召回由 Function Call/ReAct 进行驱动，每一个关联的知识库作为工具函数，LLM 会自主选择与用户问题最匹配的 1 个知识库来进行查询，**推理依据为用户问题与知识库描述的语义的匹配程度**。

举例：A 应用的上下文关联了 K1、K2、K3 三个知识库。使用 N 选 1 召回策略后，用户在应用内输入和知识库有关的问题后，LLM 将检索这三个知识库的描述，匹配某个最适合的知识库再进行内容检索。



虽然此方法无需配置 Rerank 模型，但该召回策略仅匹配单个知识库，且匹配的目标知识库严重依赖于 LLM 对于知识库描述的理解，检索匹配知识库时可能会存在不合理的判断，导致检索到的结果可能不全面、不准确，从而无法提供高质量的查询结果。

自 9 月份后，该策略将会被自动替换为**多路召回**，请提前进行修改。

在 N 选 1 模式下，召回效果主要受三个因素影响：

- **系统推理模型的能力** 部分模型对于 Function Call/ReAct 的指令遵循程度不稳定
- **知识库描述是否清晰** 描述内容会影响 LLM 对用户问题与相关知识库的推理
- **知识库的个数** 知识库过多会影响 LLM 的推理精确性，同时可能会超出推理模型的上下文窗口长度。

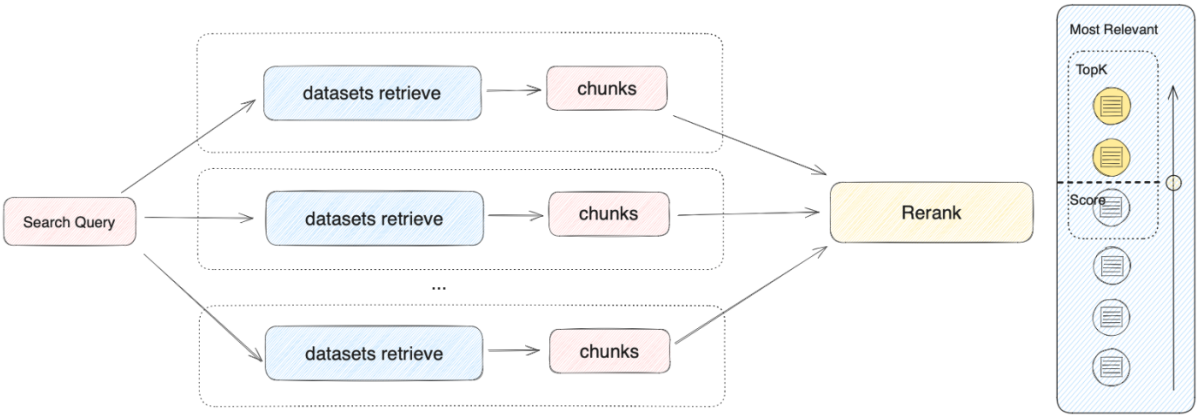
提升 N 选 1 模式推荐效果的方法：

- 选择效果更好的系统推理模型，关联尽量少的知识库，提供精确的知识库描述。
- 在知识库内上传文档内容时，系统推理模型将自动为知识库生成一个摘要描述。为了在该模式下获得最佳的召回效果，你可以在 “知识库->设置->知识库描述” 中查看到系统默认创建的摘要描述，并检查该内容是否可以清晰的概括知识库的内容。

多路召回（推荐）

> 该设置通过 Rerank 策略提供更加精准的内容检索能力。

在多路召回模式下，检索器会在所有与应用关联的知识库中去检索与用户问题相关的文本内容，并将多路召回的相关文档结果合并，以下是多路召回模式的技术流程图：



根据用户意图同时检索所有添加至 **“上下文”** 的知识库，在多个知识库内查询相关文本片段，选择所有和用户问题相匹配的内容，最后通过 Rerank 策略找到最适合的内容并回答用户。该方法的检索原理更为科学，这也意味着哪怕只有 1 个知识库，该方法也能够提供比 N 选 1 召回模式更加精准的内容回答效果。

举例：A 应用的上下文关联了 K1、K2、K3 三个知识库。使用多路召回策略后，用户输入问题后，在三个知识库内检索并汇总多条内容。最后通过 Rerank 策略确定与用户问题最相关，排分最高的内容，确保结果更加精准与可信。

多路召回模式支持以下两种 Rerank 设置：

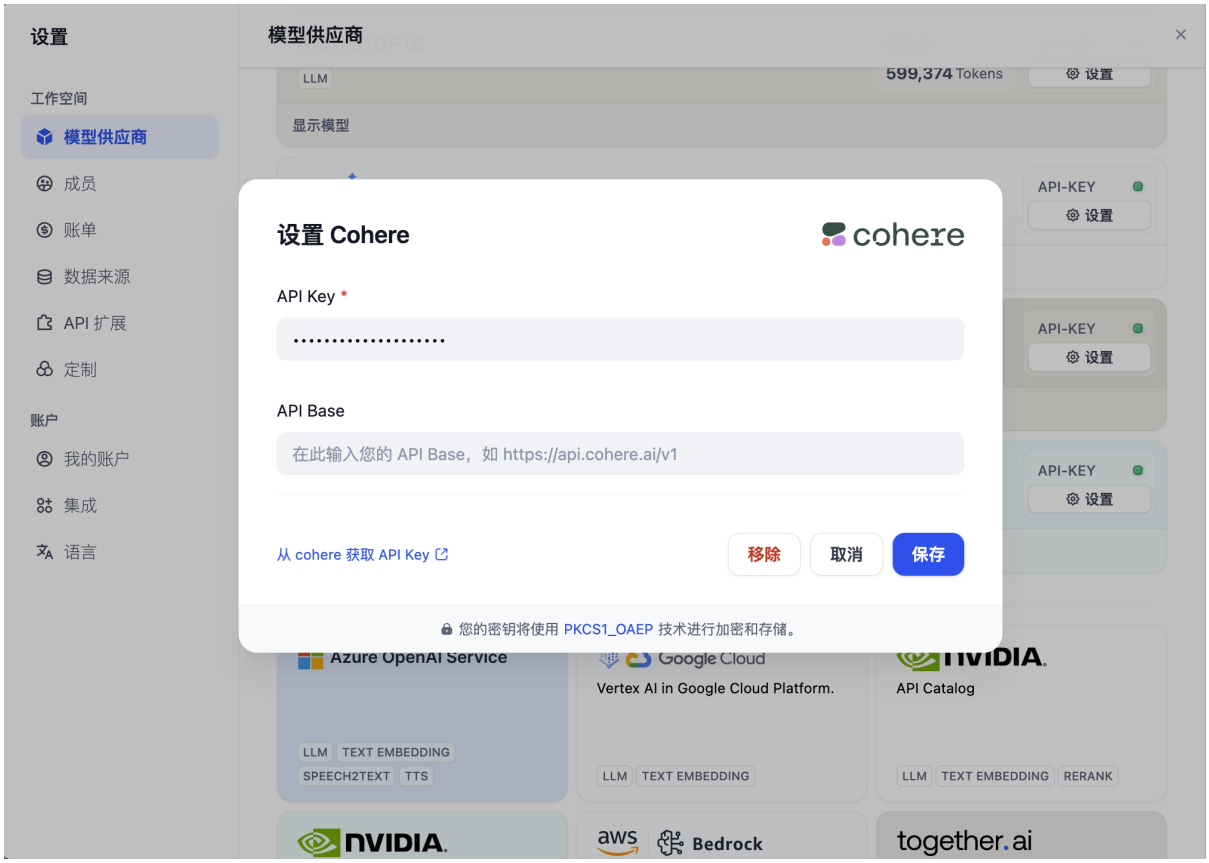
权重设置（默认）

该设置无需配置 Rerank 模型，内容查询无需额外花费。提供介于语义和关键词匹配之间的调整设置。语义指的是在知识库内进行向量检索（Vector Search），关键词匹配指的是在知识库内进行关键词的全文检索（Full Text Search）。你可以根据内容检索的实际效果，在设置内调整两者之间的权重。

> 语义优先模式（向量检索）指的是比对用户问题与知识库内容中的向量距离。距离越近，匹配的概率越大。参考阅读：《Dify：Embedding 技术与 Dify 数据集设计/规划》。

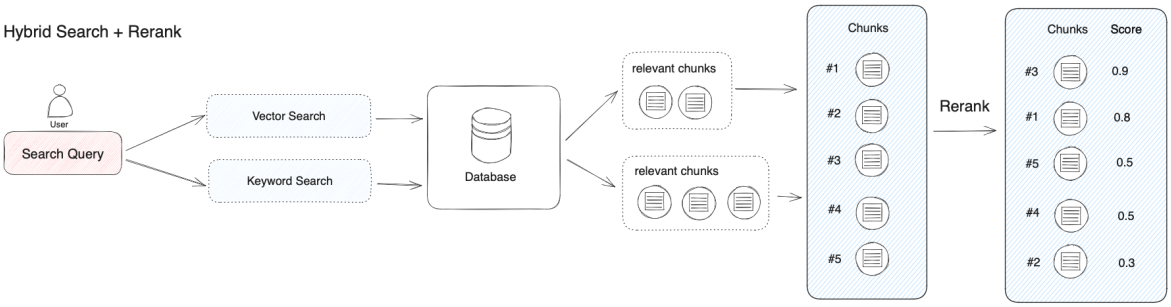
Rerank 模型

该方法需要在“模型供应商”内配置 Rerank 模型，内容查询可能会产生费用，但结果更加精准。Dify 目前支持多个 Rerank 模型，进入 “模型供应商” 页填入 Rerank 模型（例如 Cohere、Jina 等模型）的 API Key。



在模型供应商内配置 Rerank 模型

重排序模型通过将候选文档列表与用户问题语义匹配度进行重新排序，从而改进语义排序的结果。其原理是计算用户问题与给定的每个候选文档之间的相关性分数，并返回按相关性从高到低排序的文档列表。



混合检索+重排序

可调参数

- **TopK**

  用于筛选与用户问题相似度最高的文本片段。系统同时会根据选用模型上下文窗口大小，动态调整分段数量。数值越高，预期被召回的文本分段数量越多。

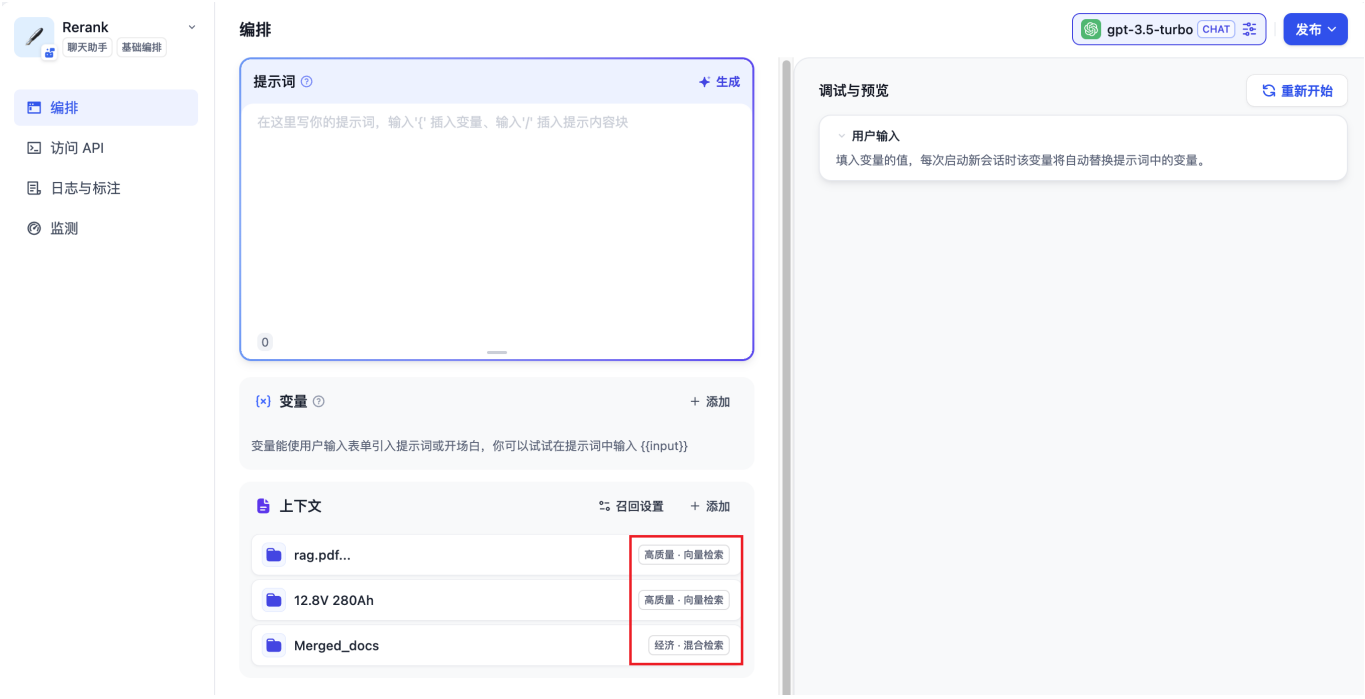
- **Score 阈值**

  用于设置文本片段筛选的相似度阈值。向量检索的相似度分数需要超过设置的分数后才会被召回，数值越高，预期被召回的文本数量越少。

多路召回模式在多知识库检索时能够获得质量更高的召回效果，因此更**推荐将召回模式设置为多路召回**。

多路召回与知识库索引的配置场景

多路召回受知识库索引方式的影响而存在部分配置差异，你在使用多路召回时可能会遇到以下情况：



- 所有已关联的知识库索引模式均为经济型

  在该情况下，你可以选择启用 Rerank 模型来增强内容检索的准确性。

- 已选择的知识库的索引模式既有经济型也有高质量，或所有知识库均为高质量但使用了不同的 Embedding 模型

该情况的内容来源格式并不统一，无法按照相同的标准进行排序。为了确保内容检索的准确度，要求配置 Rerank 模型以增强内容检索的准确性。

- 所有知识库均为高质量的索引模式，且使用了相同的 Embedding 模型

    1. 所有知识库都使用了向量检索，多路召回将默认使用“权重设置”，并且语义优先。
    2. 所有知识库都使用了全文检索，多路召回将默认使用“权重设置”，并且关键词优先。
    3. 在二者混合的情况下，多路召回将默认使用“权重设置”，配置比例语义:关键词 = 0.7:0.3

常见问题

1. **引用多个知识库时，无法调整“权重设置”，提示以下错误应如何处理?**

出现此问题是因为上下文内所引用的多个知识库内所使用的嵌入模型（Embedding）不一致，为避免检索内容冲突而出现此提示。推荐设置在“模型供应商”内设置并启用 Rerank 模型，或者统一知识库的检索设置。

2. **为什么在多路召回模式下找不到“权重设置”选项?**

请检查你的知识库是否使用了“经济”型索引模式。如果是，那么将其切换为“高质量”索引模式。