Deploying a Local RAG Knowledge Base

1. Course Content

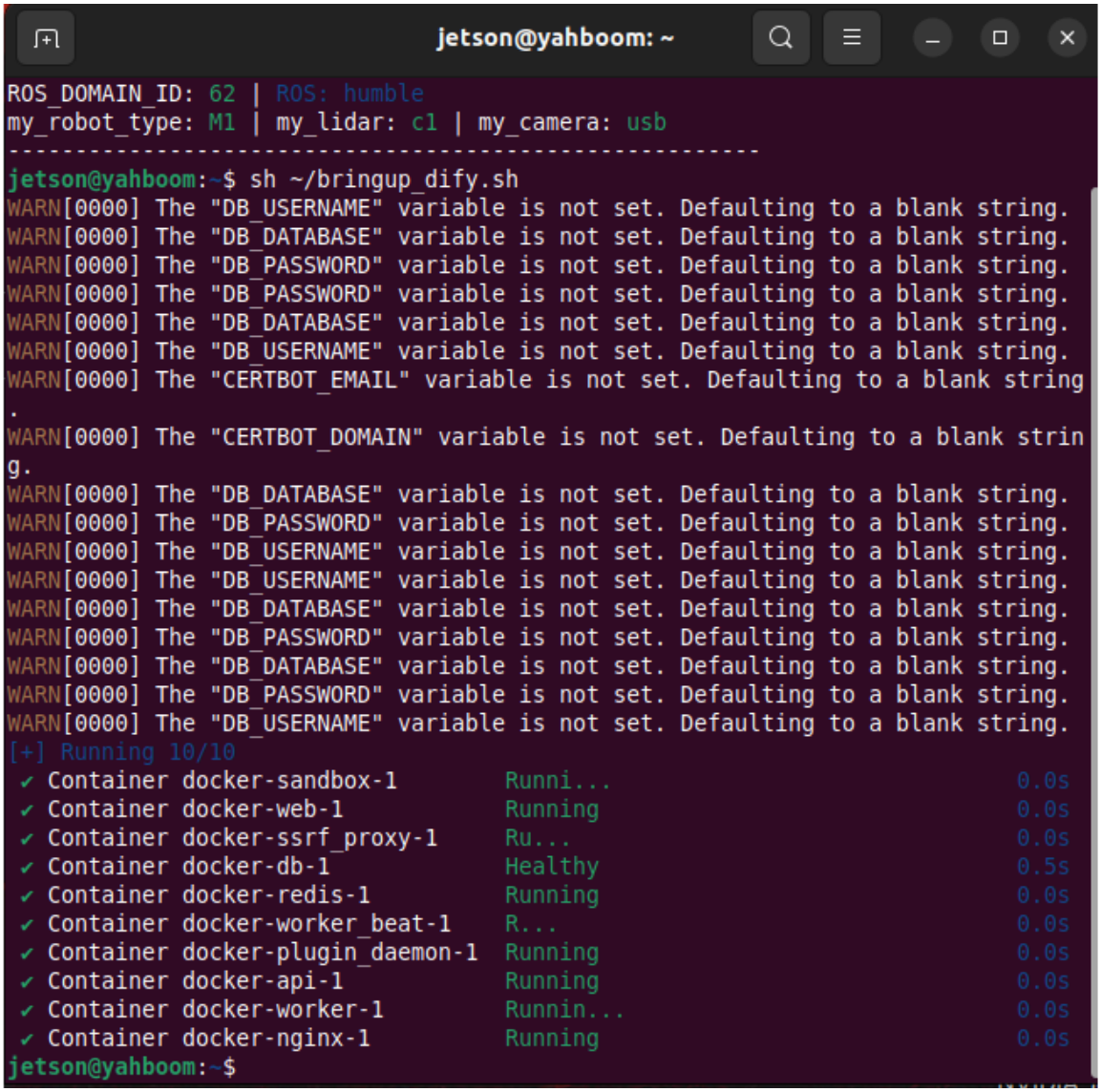
- Master the process and methods for local deployment, debugging, and testing of the RAG knowledge base.
- Master the method for extending the RAG knowledge base based on your specific task scenarios.

> [!TIP]
>
> - The RAG knowledge base helps general AI large models provide reference knowledge in vertical domains, preventing AI large models from generating hallucinatory responses and increasing the model's ability to respond with knowledge in vertical domains.
> - The RAG knowledge base can help robots quickly expand their generalization capabilities in different task scenarios.

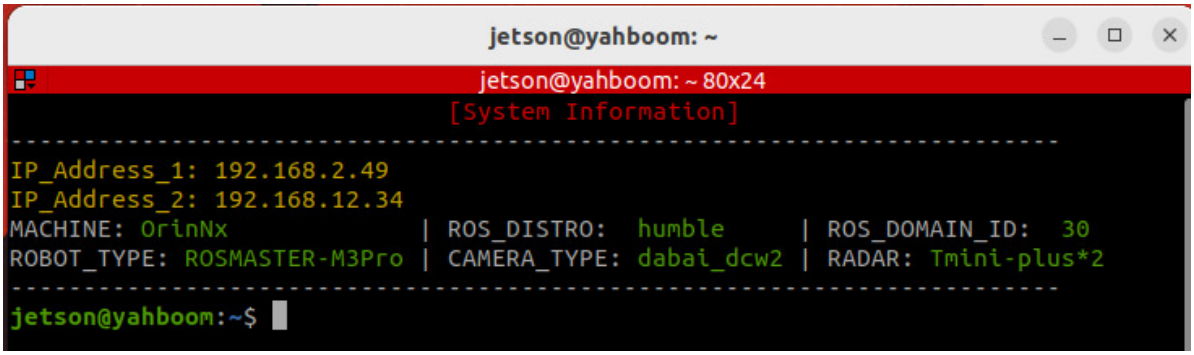
2. Starting the Dify Service

- Connect to the vehicle's computer via VNC or SSH, and enter the following command in the terminal:

```
sh ~/bringup_dify.sh
```



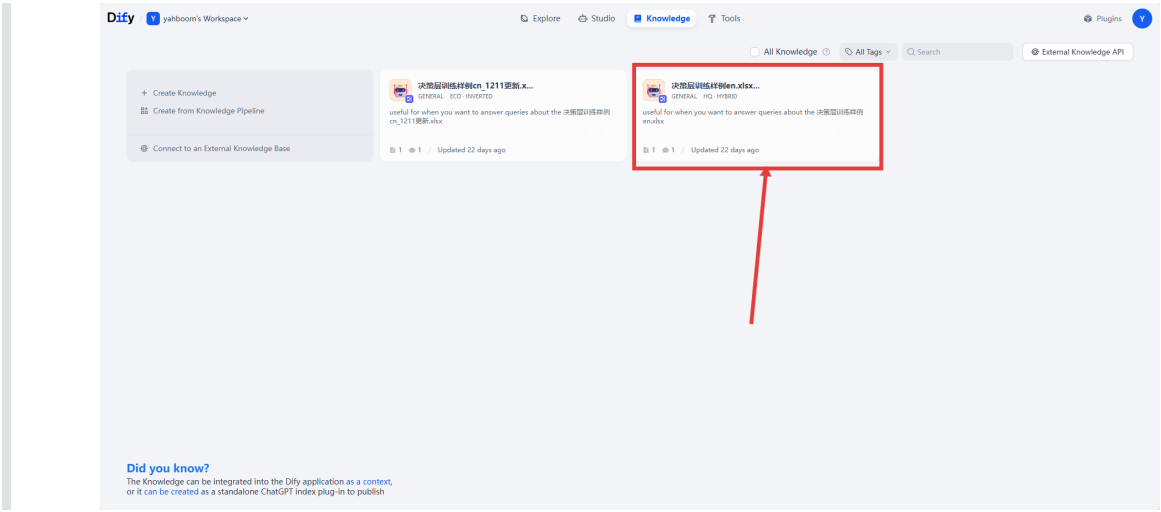
- Check the vehicle's IP address (you can view it on the OLED screen, using `ifconfig`, or directly in the terminal). Enter the vehicle's IP address directly in the browser's address bar to access the Dify management page.



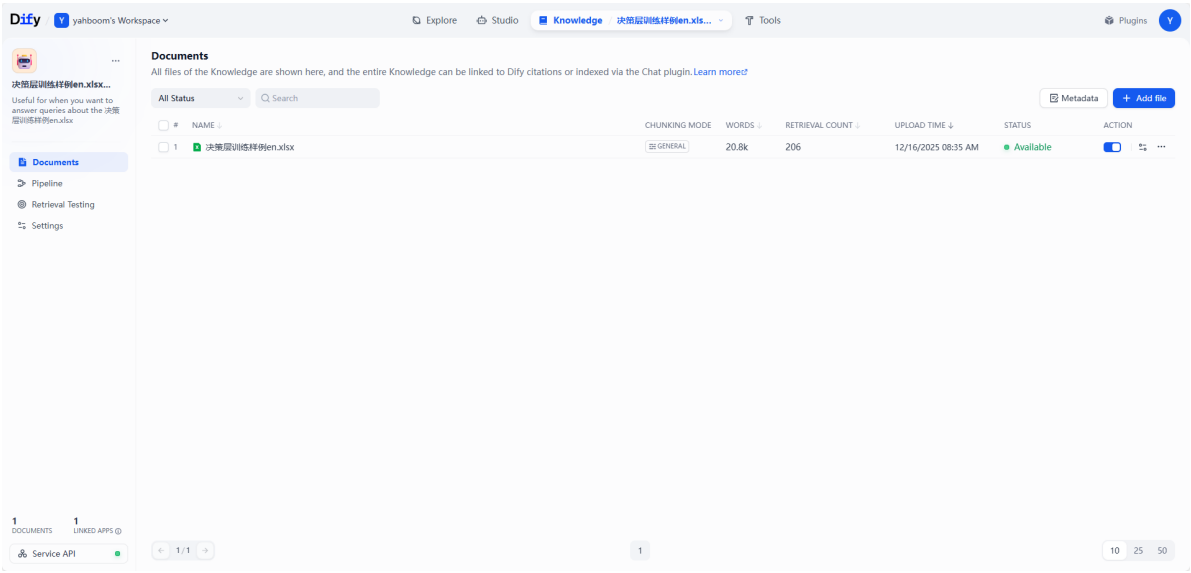
3. Viewing the Preset Knowledge Base

- Click on the Knowledge Base page on the homepage. Dify comes pre-configured with two RAG knowledge bases, with the same content but different languages.

> [!TIP]
>
> - The preset knowledge base provides training examples for some task scenarios, helping the AI model quickly master relevant skills.

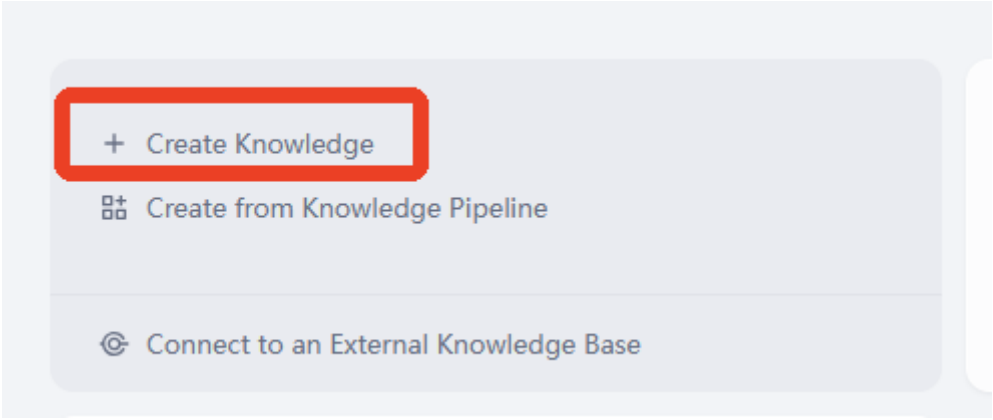


- Clicking on a[决策层训练样例库en] knowledge base reveals a preset file called `Decision Layer Example Library`, which contains the following:
- Decision Layer Sample Library: Stores some preset reference examples related to specific task scenarios.

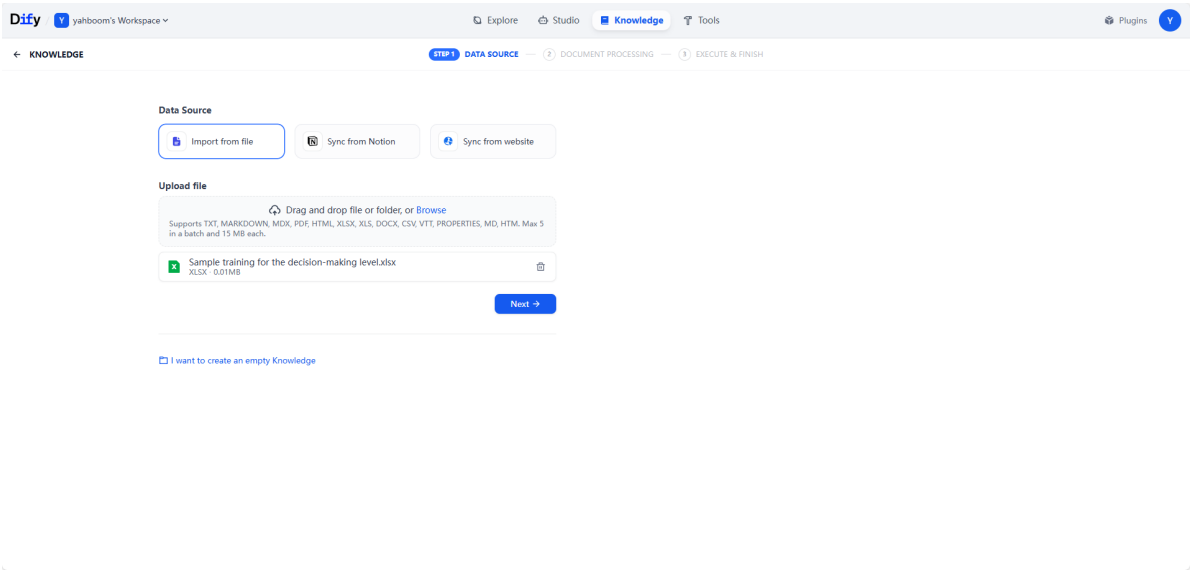


4. Expanding the RAG Knowledge Base

- If you need to expand the knowledge base, click "Create Knowledge Base".



- Here, we'll use importing local data as an example.
- Click "Import from file" ->Browse -> Next

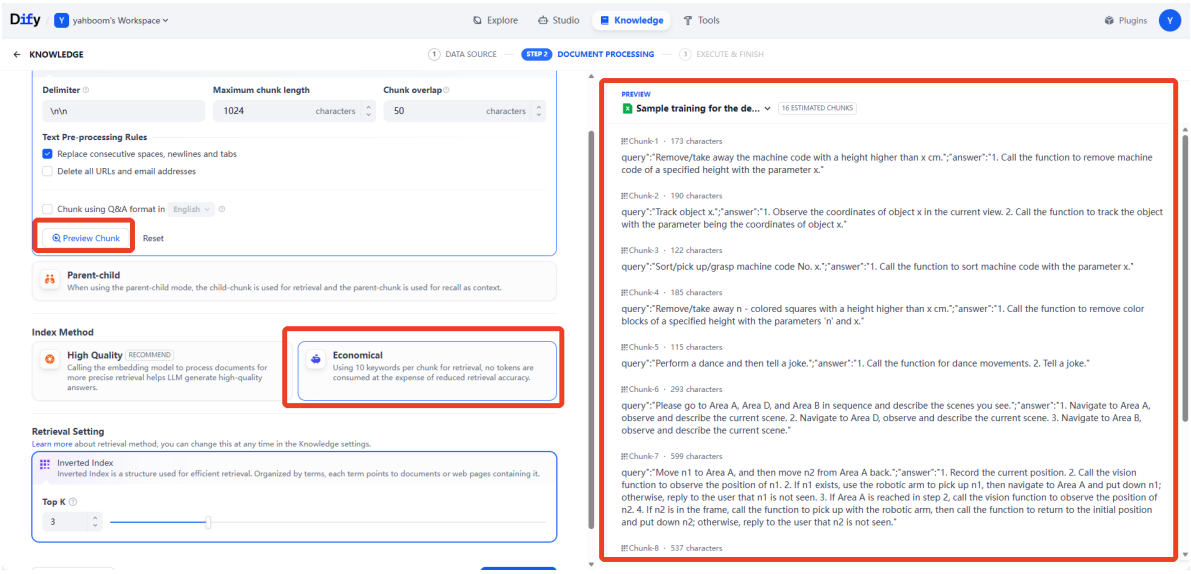


- Then, you'll enter the knowledge base configuration page. Click the preview block to view the file chunking effect. Here, select "Economic" for the indexing mode.

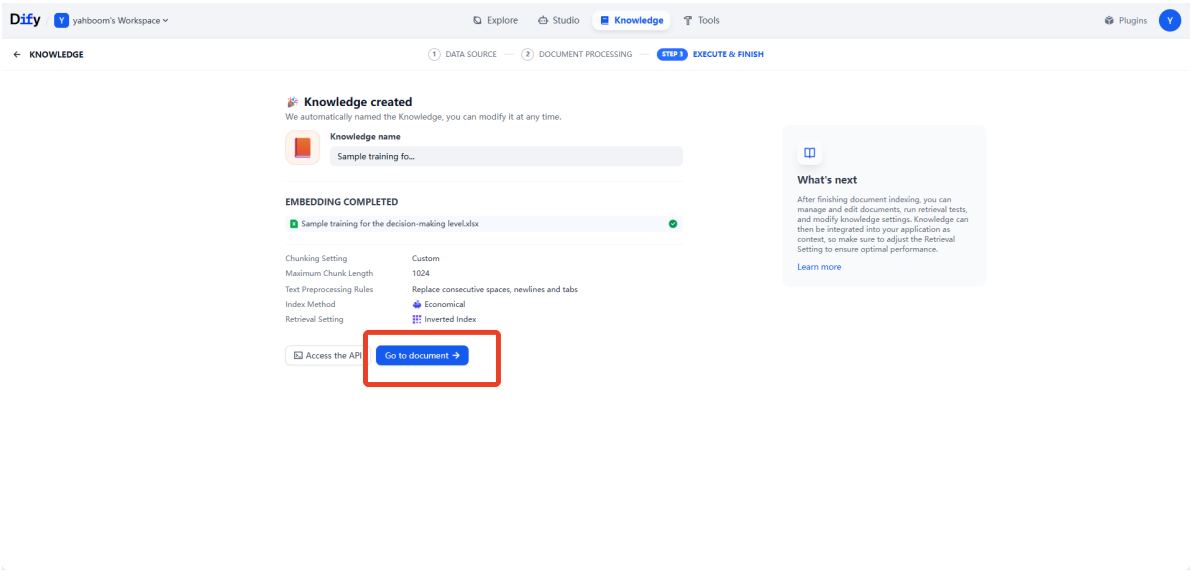
> [!TIP]
>
> - For beginners, it is recommended to use the economic mode for learning and testing. The difference between the two indexing modes:
>   - Economic: Retrieves content from the knowledge base using **keywords**. It cannot perform extended retrieval of similar semantics, and the method of retrieving knowledge fragments is relatively rigid.
>   - High-Quality Mode: Requires an embedding model to consume extra tokens and a rerank model, which can achieve more accurate retrieval of similar semantic fragments.
> - The default knowledge base mode is the high-quality mode.

4.1 Economic Mode Knowledge Base

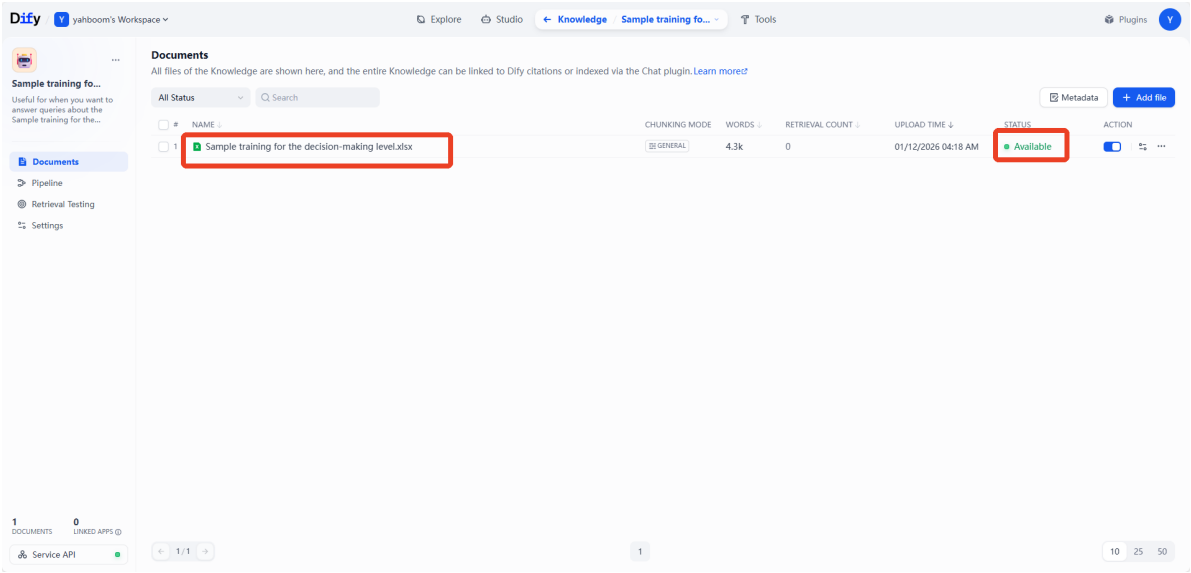
- After selecting the following configuration, click "Save & Process".



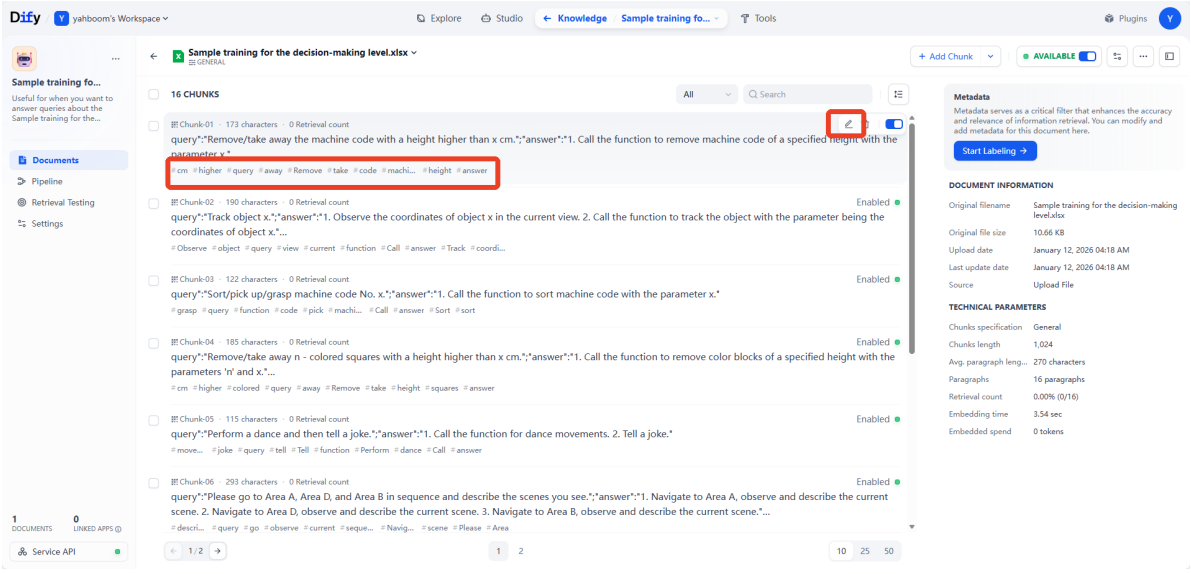
- Then wait for the embedding to complete, and click to ""Go to the document."



- When the knowledge base is functioning normally, the status will show as available. Then click on the knowledge base file.

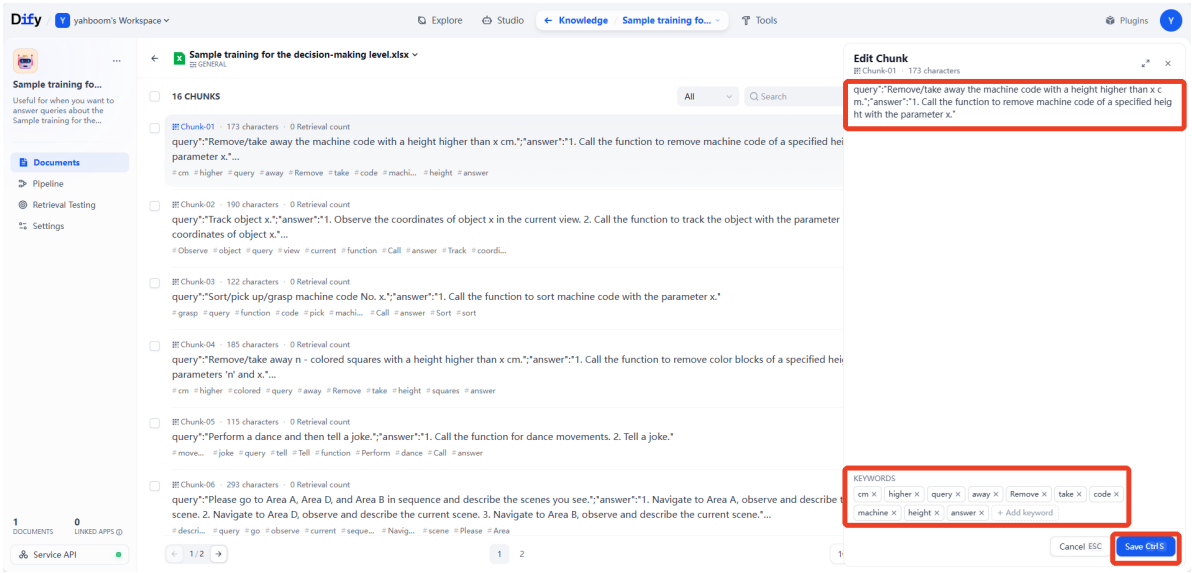


- Afterwards, you can see the segmented knowledge base fragments. The small text below each segment shows the automatically generated keywords for that segment (only available in economic mode).



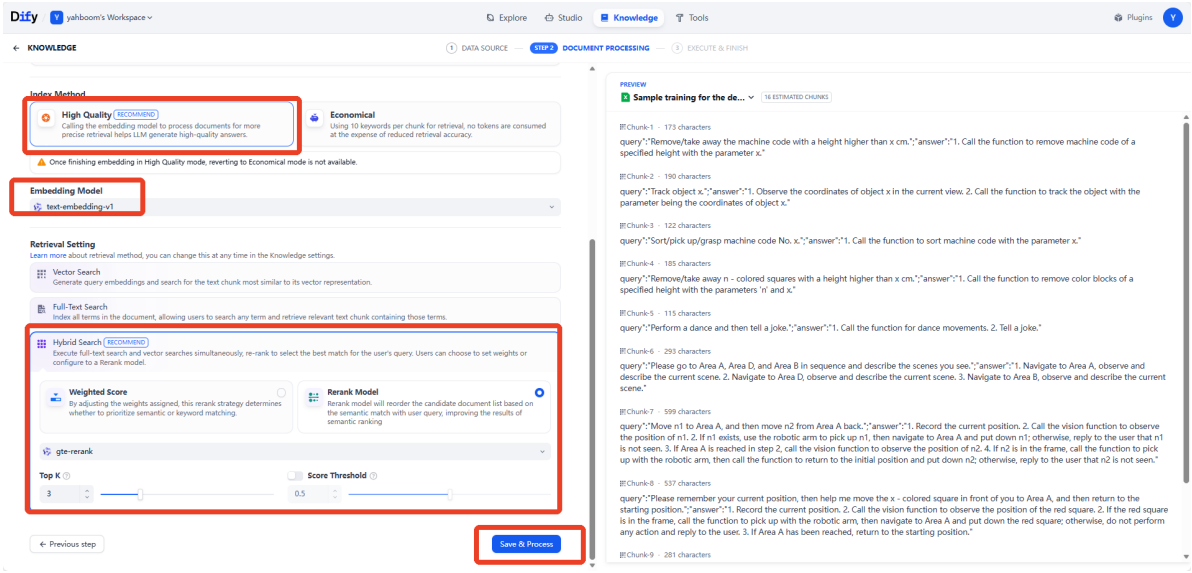
- If the keywords do not accurately describe the knowledge fragment, click "Edit" on the right side of the fragment to edit the content or keywords of that fragment. The image below

shows the modified keywords, then click save.



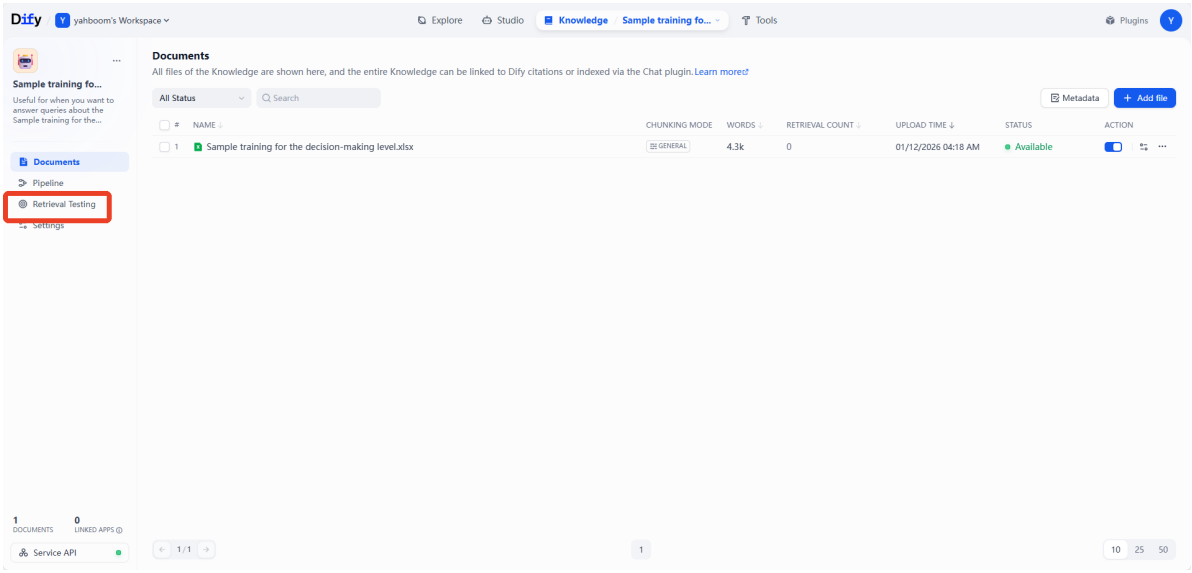
4.2 High-Quality Mode Knowledge Base

- If you need to use a high-quality knowledge base later, refer to this section of the tutorial.
- The knowledge base creation and file import process is the same as before.
- Here, select "High Quality" for the indexing method, and choose any retrieval method. Here, we use hybrid retrieval as an example. Finally, save and process.

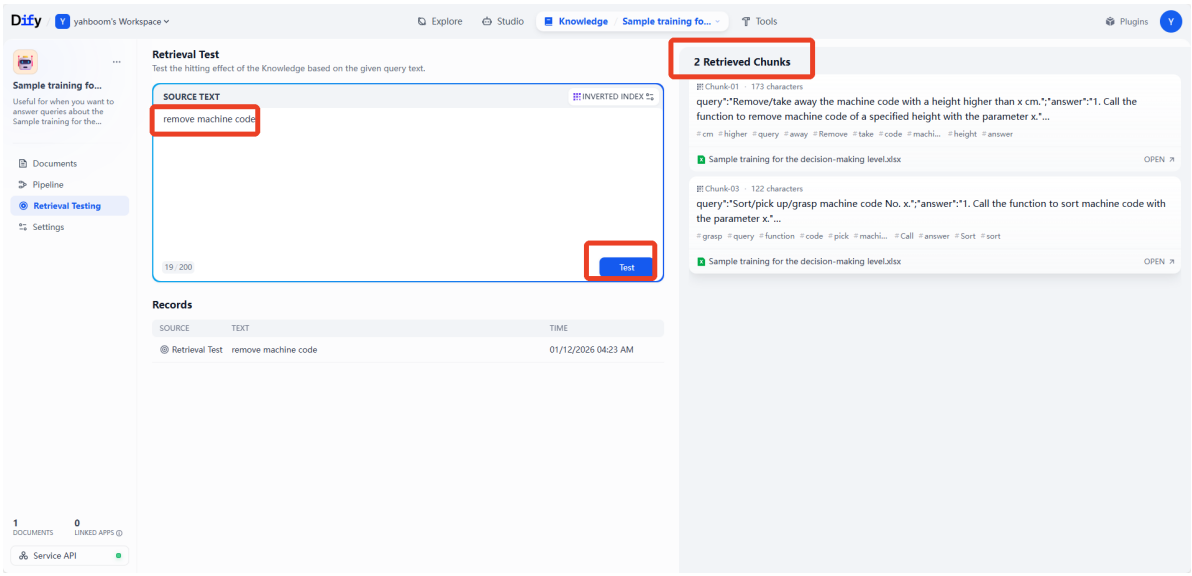


4.3 Recall Test

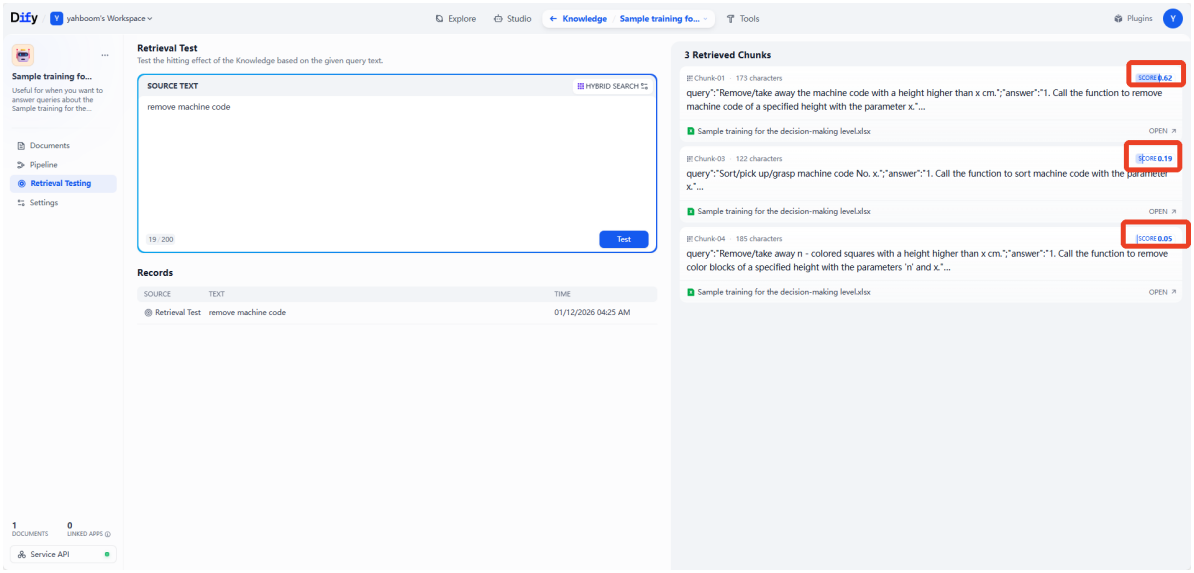
- Recall testing tests the actual effectiveness of retrieving relevant knowledge snippets from the knowledge base based on the input, helping to optimize the AI model's response performance.
- After opening a knowledge base, click on "Recall Test" on the left side.



- Enter the test content in the source text (simulating user input during actual use), and then click "Test".
- The retrieved paragraphs and the knowledge base related to the input content will appear on the right. The knowledge base tested here is the economic model knowledge base, which retrieves information based on keywords.



- If it's a high-quality mode knowledge base, the retrieved snippets will have a `SCORE` rating. A higher score indicates a higher relevance between the snippet and the input content. High-quality mode knowledge bases can perform associative retrieval of similar semantics, but also consume tokens.


Dify
yahboom's Workspace
Explore
Studio
Knowledge
Sample training fo...
Tools
Plugins
Sample training fo...
Useful for when you want to answer queries about the Sample training for the...
Documents
Pipeline
Retrieval Testing
Settings
1
DOCUMENTS
0
LINKED APPS
Service API
Retrieval Test
Test the hitting effect of the Knowledge based on the given query text.
SOURCE TEXT
HYBRID SEARCH
remove machine code
19 / 200
Test
Records
SOURCE
TEXT
TIME
Retrieval Test
remove machine code
01/12/2026 04:25 AM
3 Retrieved Chunks
Chunk-01 · 173 characters
SCORE 0.62
query":"Remove/take away the machine code with a height higher than x cm.","answer":"1. Call the function to remove machine code of a specified height with the parameter x."...
Sample training for the decision-making level.xlsx
OPEN
Chunk-03 · 122 characters
SCORE 0.19
query":"Sort/pick up/grasp machine code No. x.","answer":"1. Call the function to sort machine code with the parameter x."...
Sample training for the decision-making level.xlsx
OPEN
Chunk-04 · 185 characters
SCORE 0.05
query":"Remove/take away n - colored squares with a height higher than x cm.","answer":"1. Call the function to remove color blocks of a specified height with the parameters 'n' and x."...
Sample training for the decision-making level.xlsx
OPEN