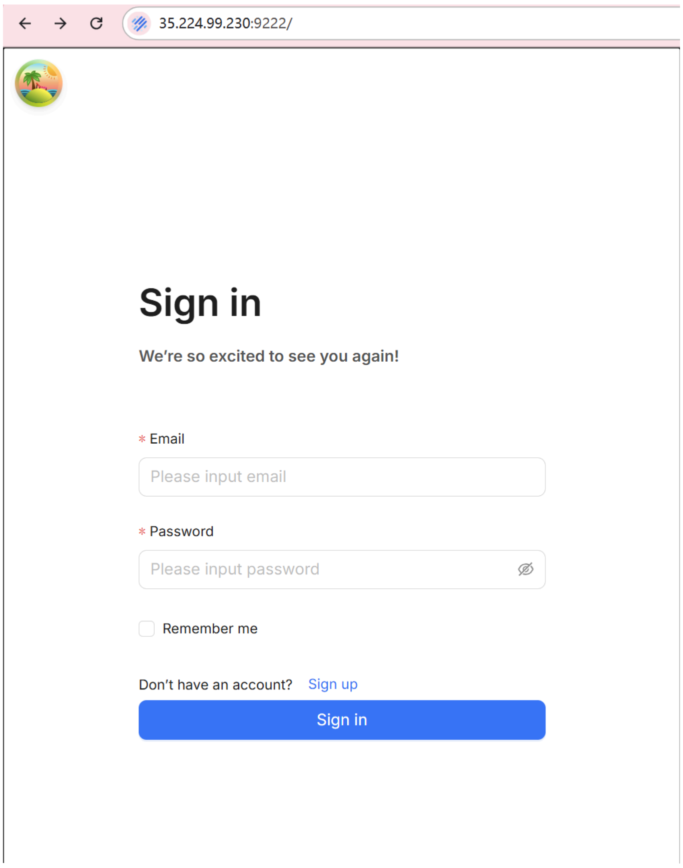
The User manual for Retrieval-Augmented Generation (RAG) System

1 Register and Log in

1. Go to this link https://35.224.99.230:9222/
2. Sign up a new user



35.224.99.230:9222/

Sign in

We're so excited to see you again!

* Email

Please input email

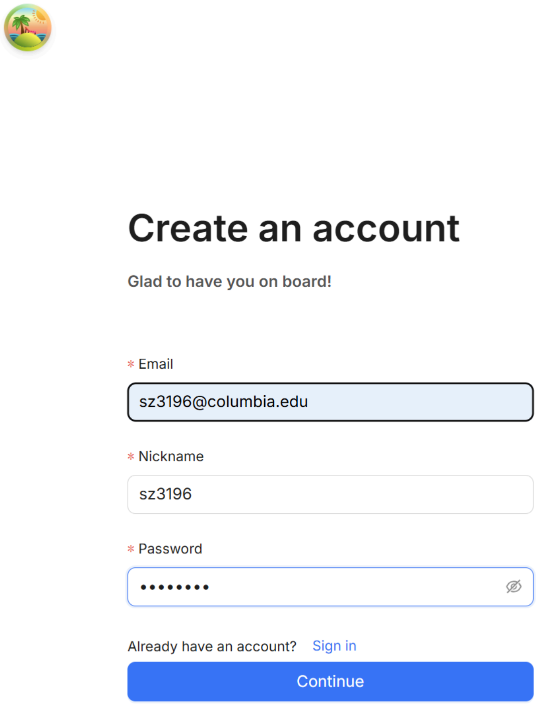
* Password

Please input password

□ Remember me

Don't have an account? Sign up

Sign in



Create an account

Glad to have you on board!

* Email

sz3196@columbia.edu

* Nickname

sz3196

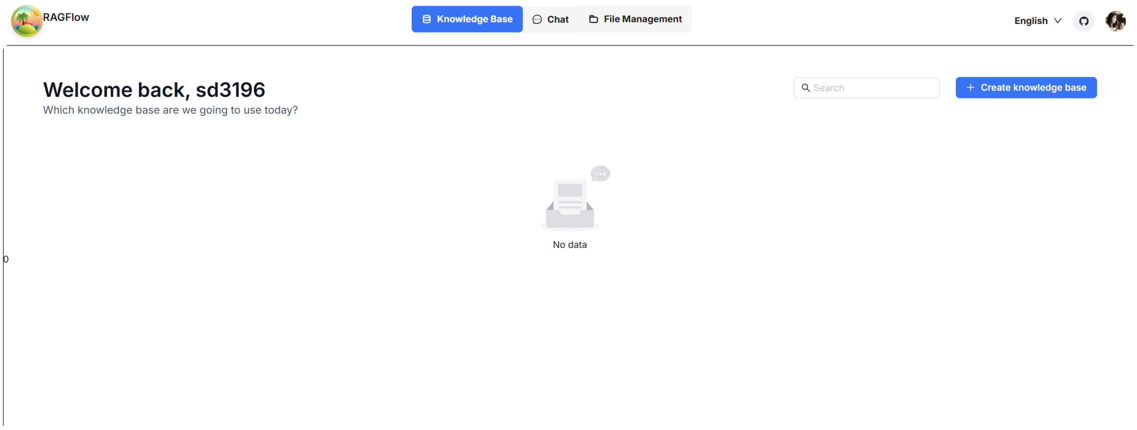
* Password

••••••••

Already have an account? Sign in

Continue

3. Log in, and you will get something like:



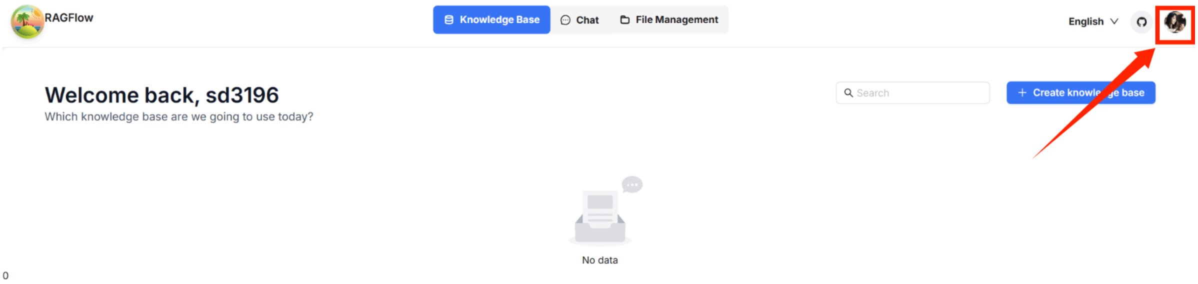
Welcome back, sd3196

Which knowledge base are we going to use today?

No data

2 Set the Language Model

1. Go to the user setting



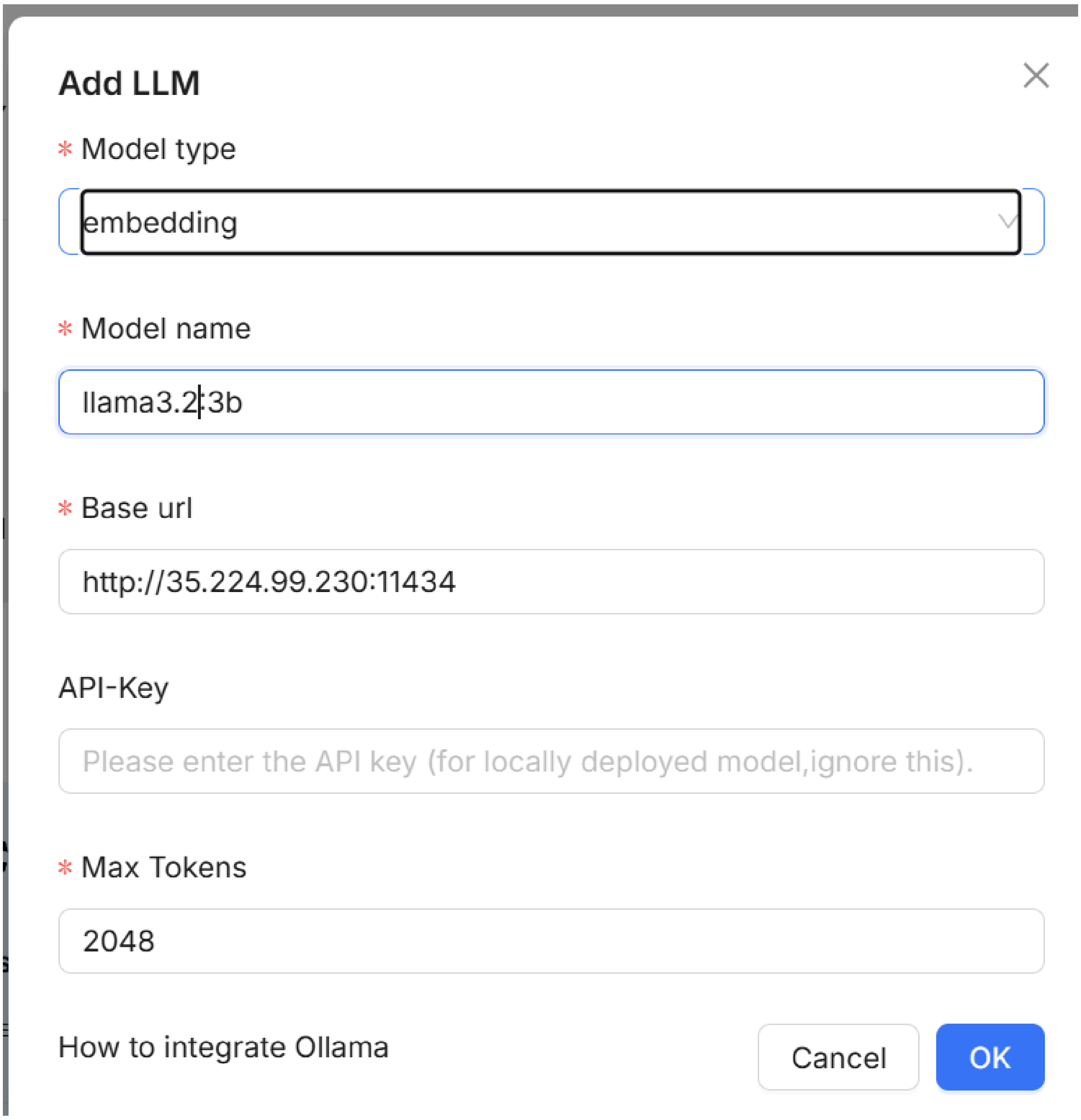
Welcome back, sd3196

Which knowledge base are we going to use today?

No data

2. In 'Model Providers', click "add the model" button under ollama

3. Configure the setting as follows:



Add LLM

* Model type

embedding

* Model name

llama3.2:3b

* Base url

http://35.224.99.230:11434

API-Key

Please enter the API key (for locally deployed model,ignore this).

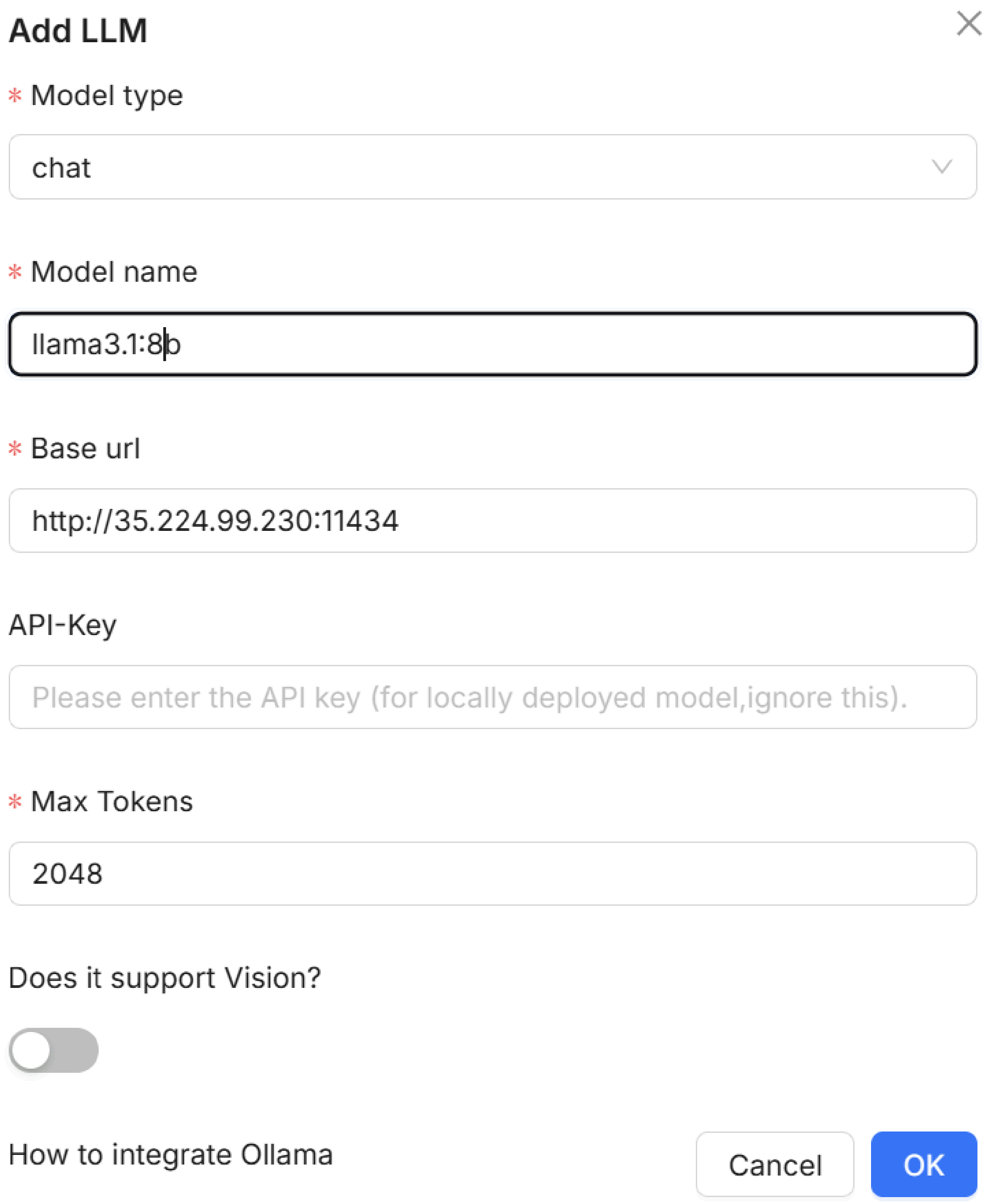
* Max Tokens

2048

How to integrate Ollama

- model name: llama3.2:3b
- Base url: http://35.224.99.230:11434
- max token 2048

4. Similarly, add another model with type "chat"



Add LLM

* Model type

chat

* Model name

llama3.1:8b

* Base url

http://35.224.99.230:11434

API-Key

Please enter the API key (for locally deployed model,ignore this).

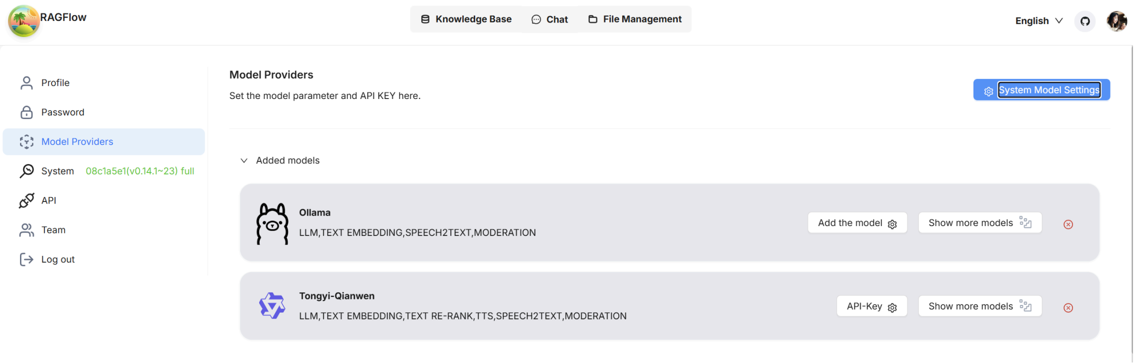
* Max Tokens

2048

Does it support Vision?

How to integrate Ollama

- model type: chat
- model name: llama3.1:8b
- Base url: http://35.224.99.230:11434
- max token 2048



System 08c1a5e1(v0.14.1~23) full

Model Providers

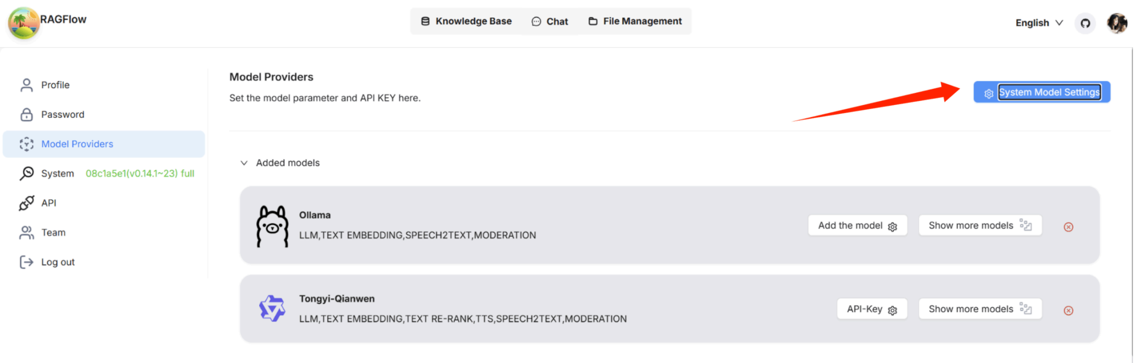
Set the model parameter and API KEY here.

Added models

Ollama
LLM,TEXT EMBEDDING,SPEECH2TEXT,MODERATION

Tongyi-Qianwen
LLM,TEXT EMBEDDING,TEXT RE-RANK,TTS,SPEECH2TEXT,MODERATION

5. Then click the system model settings, select the two models we add above in the pull list



Model Providers

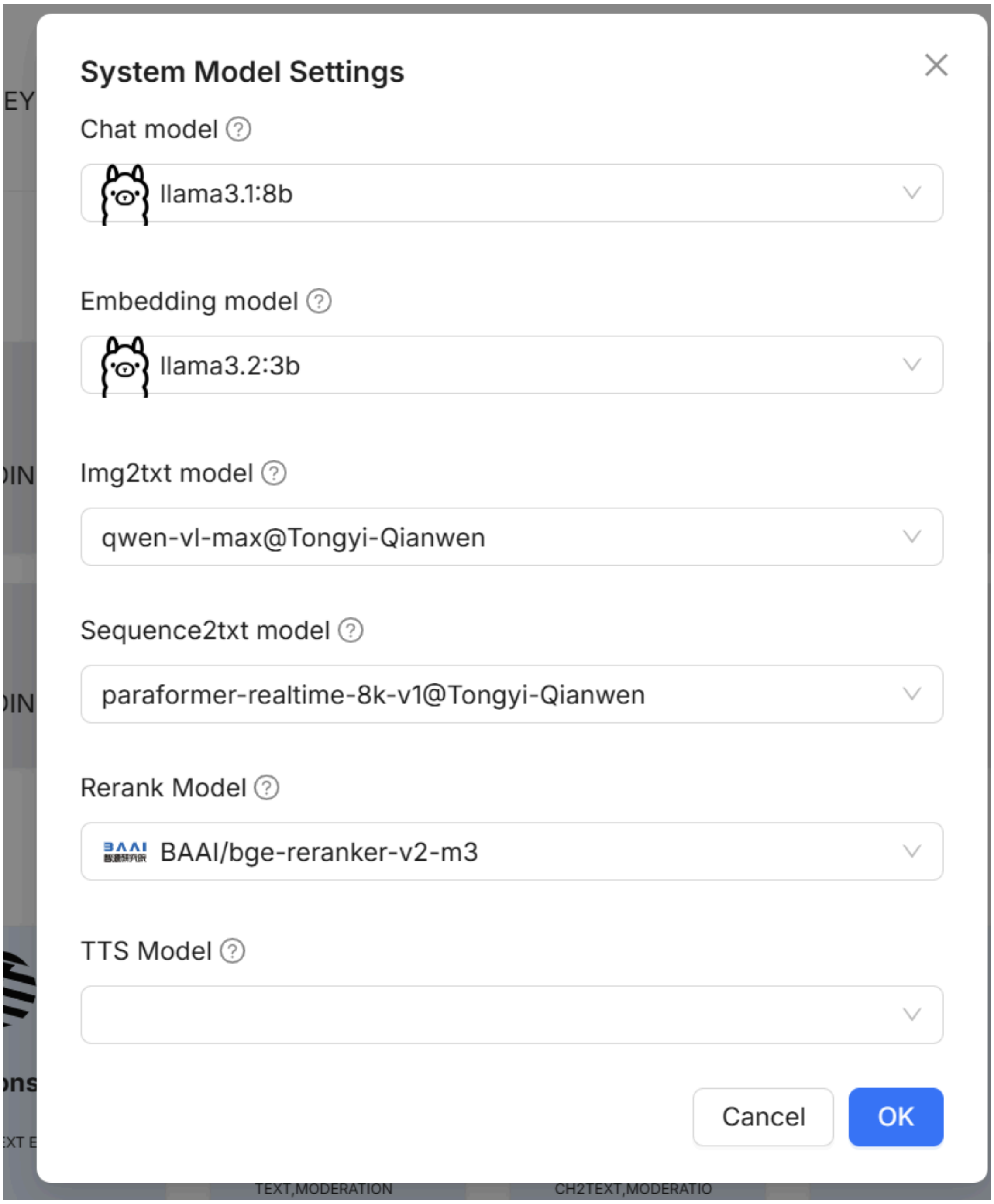
Set the model parameter and API KEY here.

System 08c1a5e1(v0.14.1~23) full

Added models

Ollama
LLM,TEXT EMBEDDING,SPEECH2TEXT,MODERATION

Tongyi-Qianwen
LLM,TEXT EMBEDDING,TEXT RE-RANK,TTS,SPEECH2TEXT,MODERATION



System Model Settings

Chat model: llama3.1:8b

Embedding model: llama3.2:3b

Img2txt model: qwen-vl-max@Tongyi-Qianwen

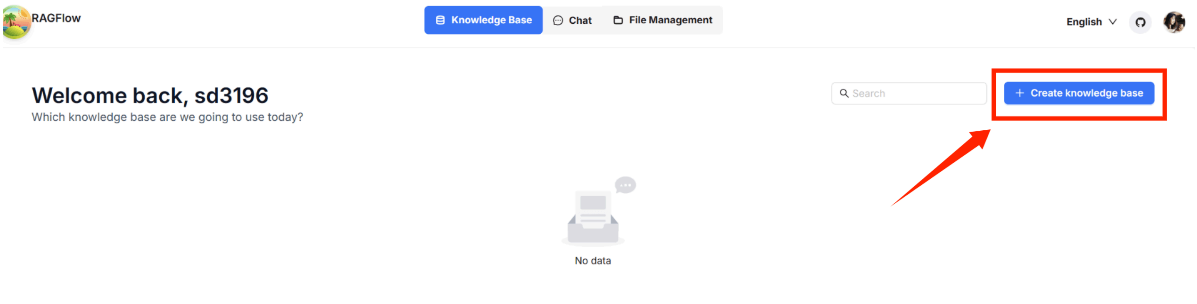
Sequence2txt model: paraformer-realtime-8k-v1@Tongyi-Qianwen

Rerank Model: BAAI/bge-reranker-v2-m3

TTS Model:

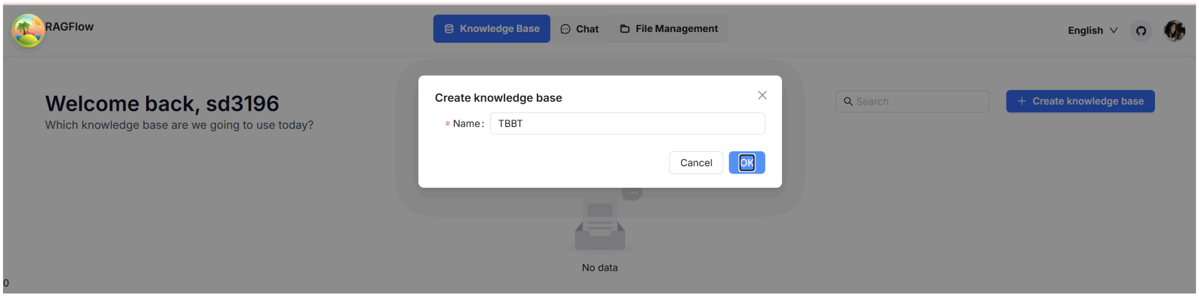
3. Set Knowledge Base

1. Click the button and create a Knowledge Base



Welcome back, sd3196

Which knowledge base are we going to use today?



Welcome back, sd3196

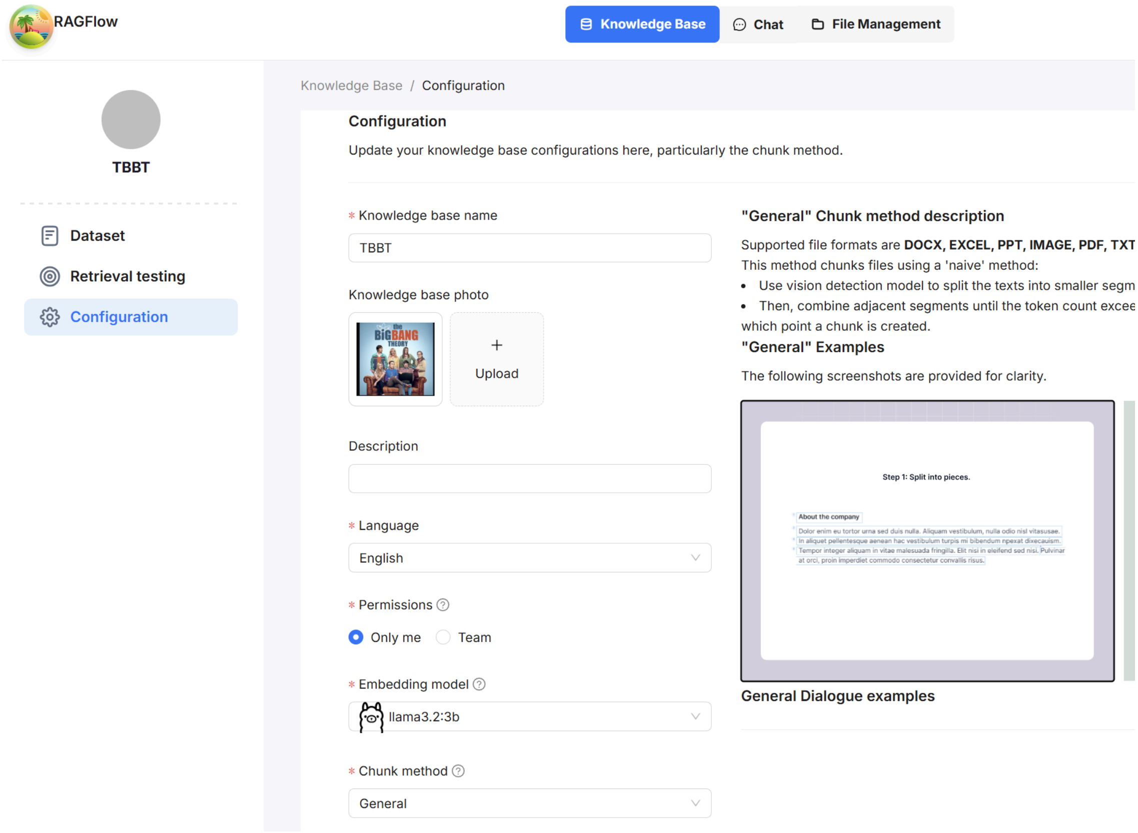
Which knowledge base are we going to use today?

Create knowledge base

* Name : TBBT

Cancel OK

2. Set the embedding model in llama3.2:3b



Knowledge Base / Configuration

Configuration

Update your knowledge base configurations here, particularly the chunk method.

* Knowledge base name: TBBT

Knowledge base photo

Description

* Language: English

* Permissions: ● Only me ○ Team

* Embedding model: llama3.2:3b

* Chunk method: General

"General" Chunk method description

Supported file formats are **DOCX, EXCEL, PPT, IMAGE, PDF, TXT**

This method chunks files using a 'naive' method:

- Use vision detection model to split the texts into smaller segm
- Then, combine adjacent segments until the token count excee

which point a chunk is created.

"General" Examples

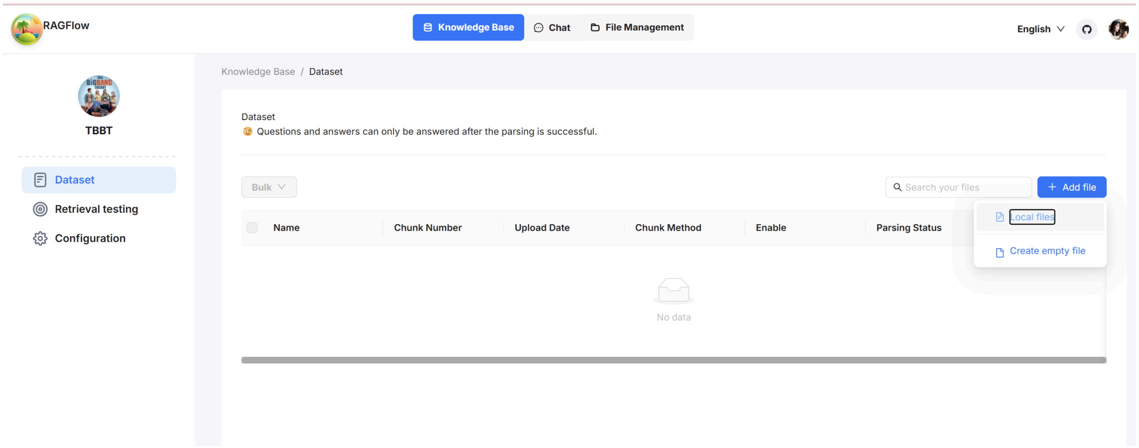
The following screenshots are provided for clarity.

About the company

Dolor enim eu tortor urna sed duis nulla. Aliquam vestibulum, nulla odio nisl vitasusae. In aliquet pellentesque aenean hac vestibulum turpis mi bibendum npexat dixecauism. Tempor integer aliquam in vitae malesuada fringilla. Elit nisi in eleifend sed nisi. Pulvinar at orci, proin imperdiet commodo consectetur convallis risus.

General Dialogue examples

3. add the file



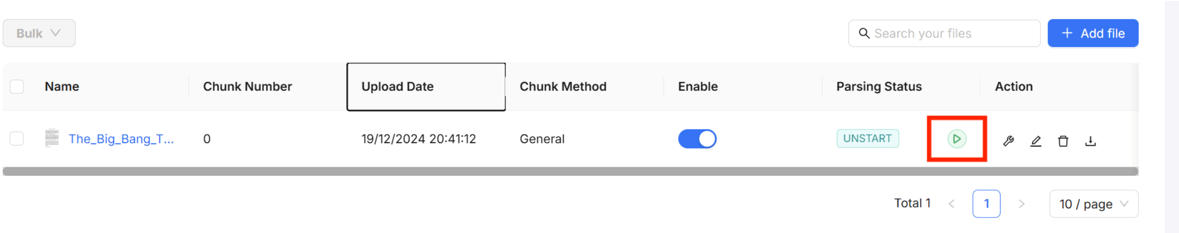
Knowledge Base / Dataset

Dataset

Questions and answers can only be answered after the parsing is successful.

	Name	Chunk Number	Upload Date	Chunk Method	Enable	Parsing Status

4. start parsing and wait until its finish

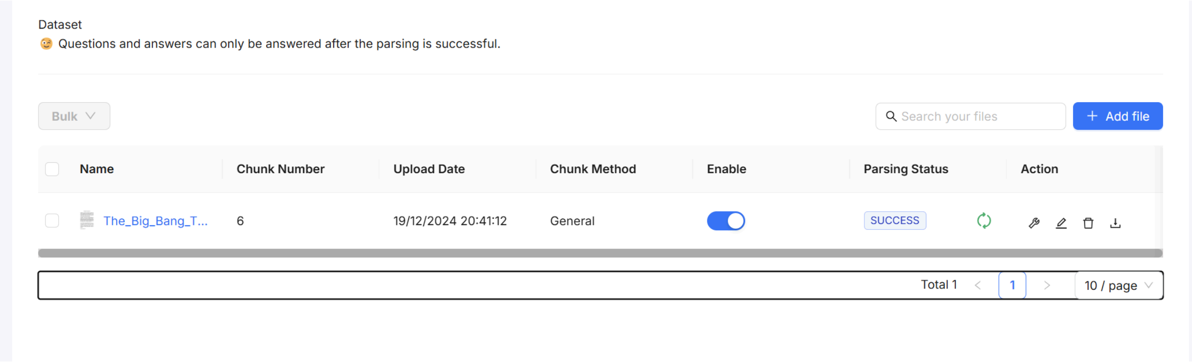


Bulk

Search your files | Add file

Name	Chunk Number	Upload Date	Chunk Method	Enable	Parsing Status	Action
The_Big_Bang_T...	0	19/12/2024 20:41:12	General	●	UNSTART	▷

Total 1 < 1 > 10 / page



Dataset

Questions and answers can only be answered after the parsing is successful.

Bulk

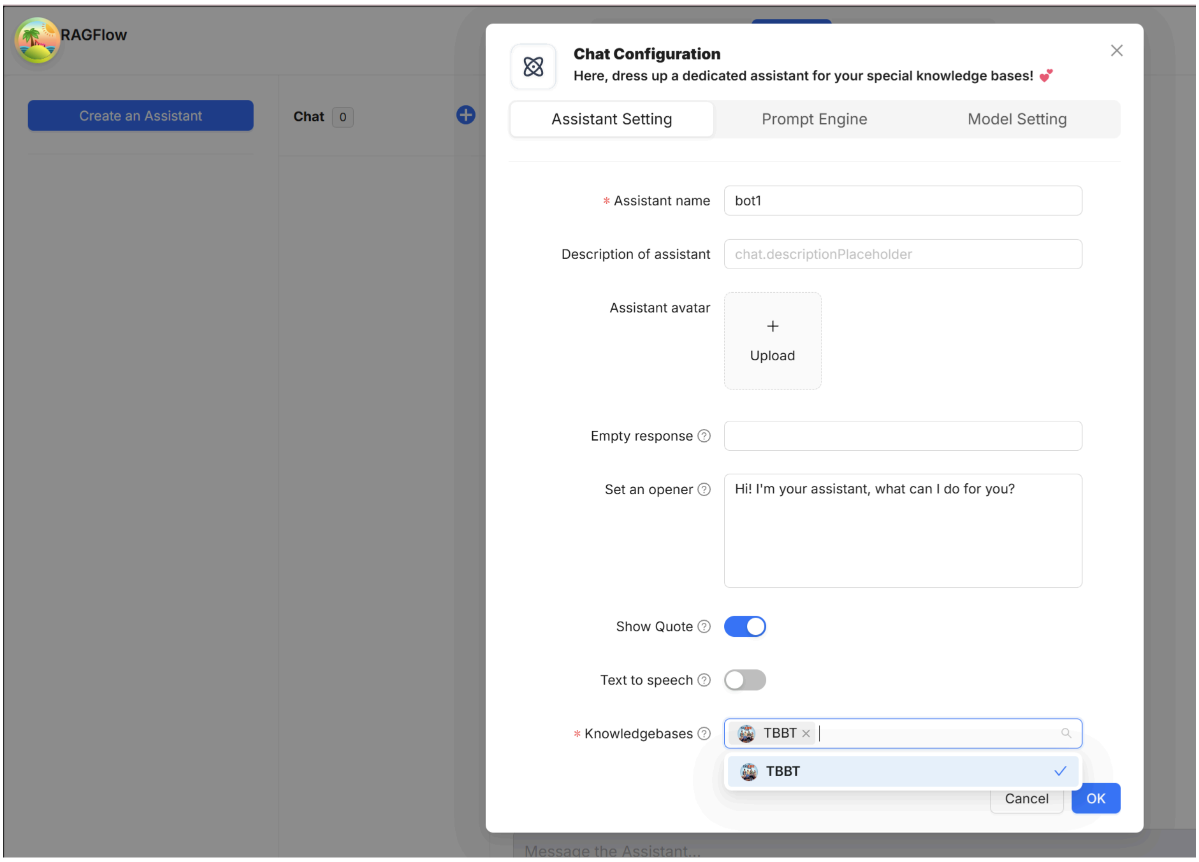
Search your files | Add file

Name	Chunk Number	Upload Date	Chunk Method	Enable	Parsing Status	Action
The_Big_Bang_T...	6	19/12/2024 20:41:12	General	●	SUCCESS	⟳

Total 1 < 1 > 10 / page

4. Create a chat assistant and talk to it

1. click "Create an Assistant", and select the knowledge base we create before



Chat Configuration

Here, dress up a dedicated assistant for your special knowledge bases!

Assistant Setting | Prompt Engine | Model Setting

* Assistant name: bot1

Description of assistant: chat.descriptionPlaceholder

Assistant avatar: + Upload

Empty response:

Set an opener: Hi! I'm your assistant, what can I do for you?

Show Quote: ●

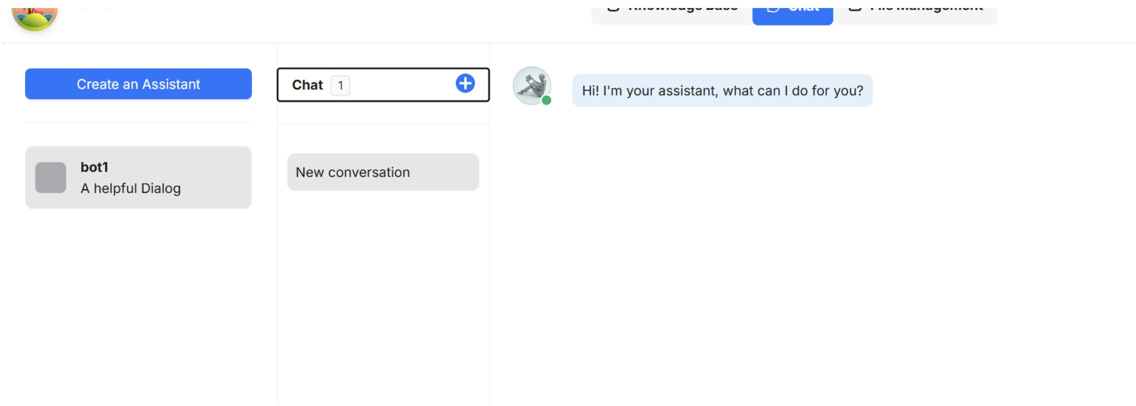
Text to speech: ○

* Knowledgebases: TBBT ×

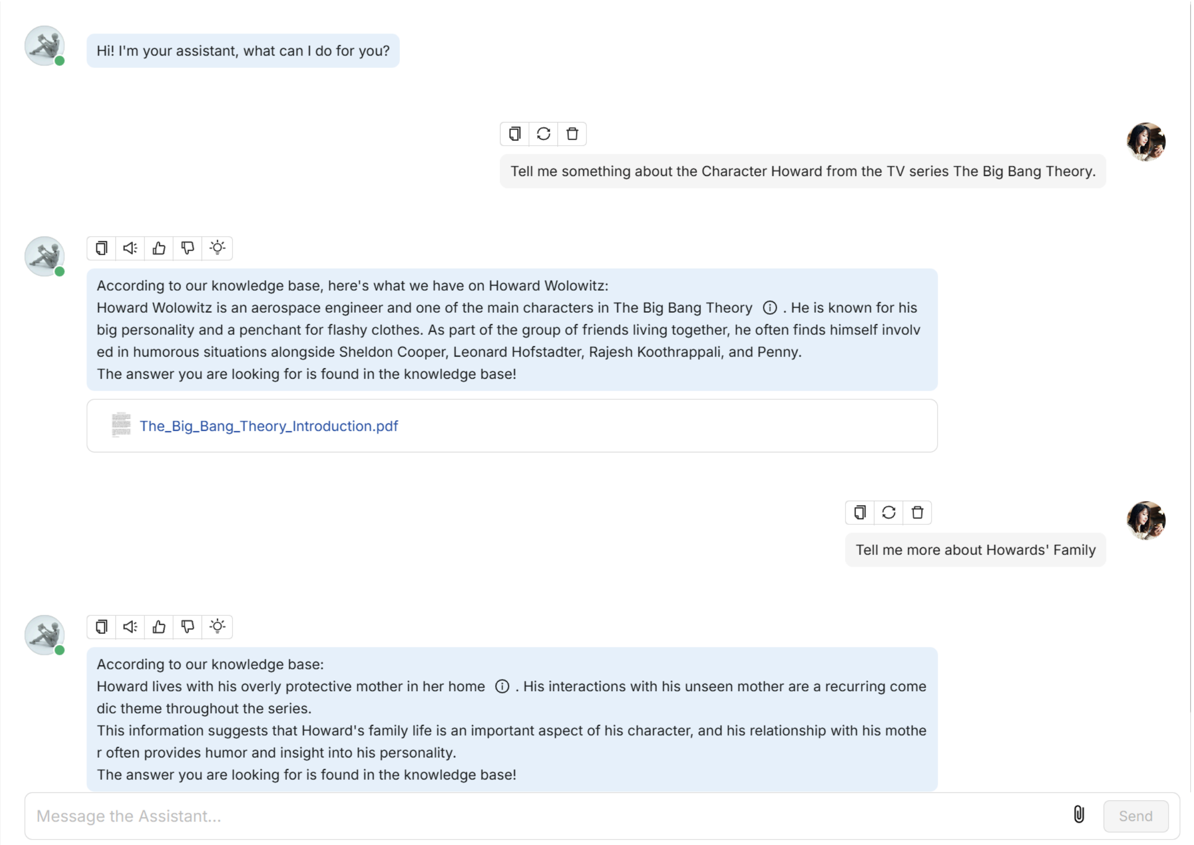
TBBT ✓

Cancel | OK

2. add a chat and talk to it!



Hi! I'm your assistant, what can I do for you?



Hi! I'm your assistant, what can I do for you?

Tell me something about the Character Howard from the TV series The Big Bang Theory.

According to our knowledge base, here's what we have on Howard Wolowitz:
Howard Wolowitz is an aerospace engineer and one of the main characters in The Big Bang Theory ⓘ . He is known for his big personality and a penchant for flashy clothes. As part of the group of friends living together, he often finds himself involved in humorous situations alongside Sheldon Cooper, Leonard Hofstadter, Rajesh Koothrappali, and Penny.
The answer you are looking for is found in the knowledge base!

The_Big_Bang_Theory_Introduction.pdf

Tell me more about Howards' Family

According to our knowledge base:
Howard lives with his overly protective mother in her home ⓘ . His interactions with his unseen mother are a recurring comedic theme throughout the series.
This information suggests that Howard's family life is an important aspect of his character, and his relationship with his mother often provides humor and insight into his personality.
The answer you are looking for is found in the knowledge base!

We can see the assistant is answering the question using the knowledge from the knowledge base we created. It can be a powerful solution for processing unstructured documents and delivering intelligent, context-aware answers.