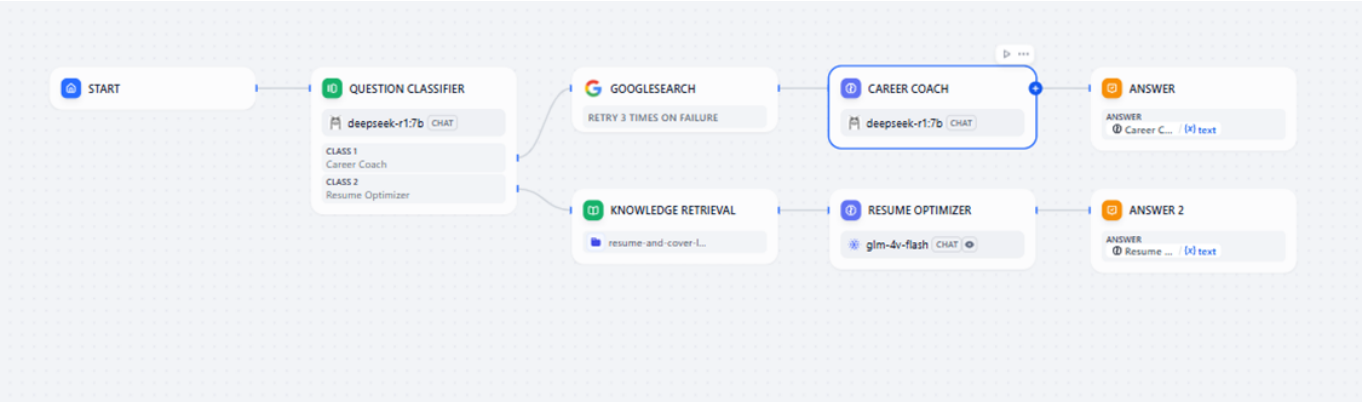
ECS Locally Deployed Ollama+DeepSeek+Dify building Career Coach & Resume Optimizer AI Agent

---

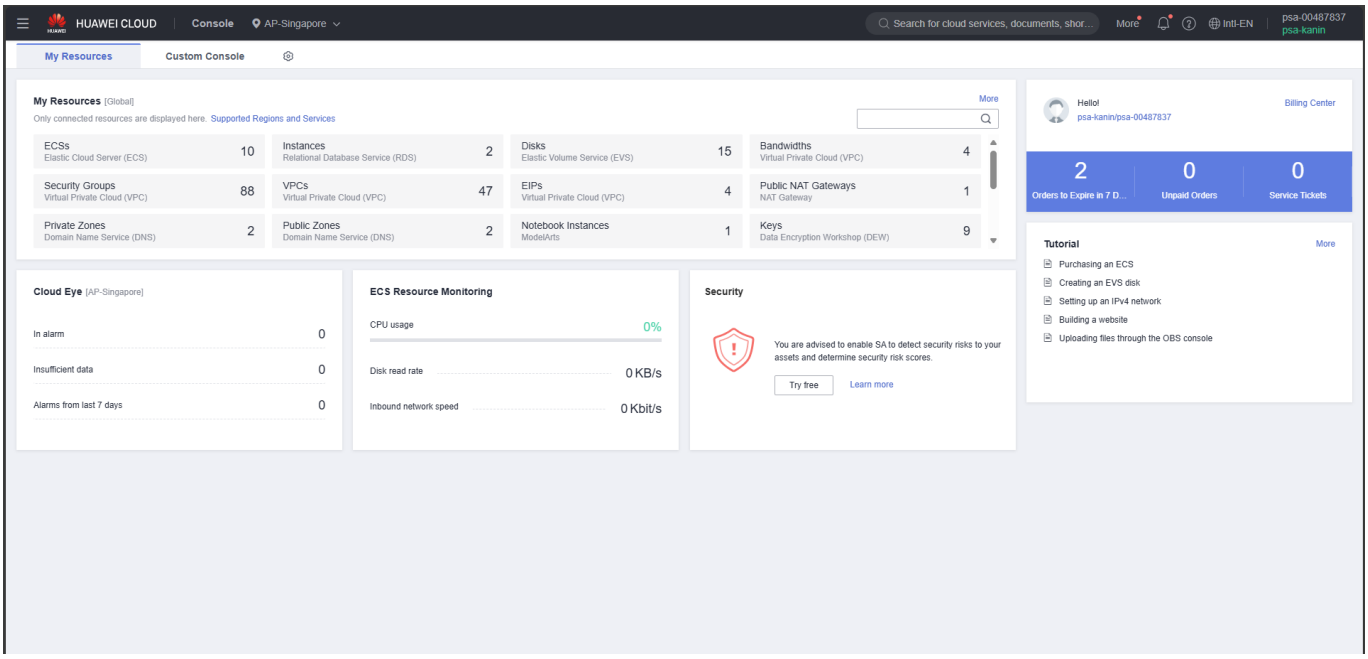
- Recommended time duration for this demo: 2.5 hours
- Required Cloud Resource: ECS, EIP
- Participants to prepare: Image (screenshot) of current resume

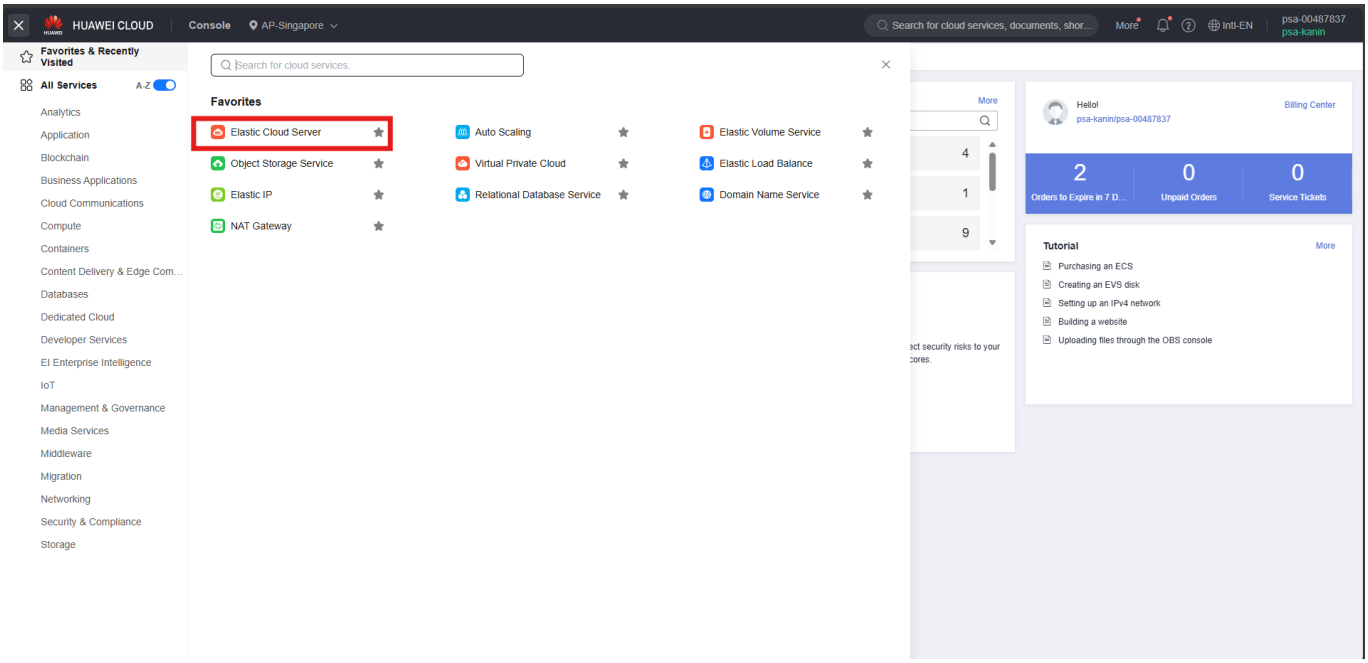
Preview of the chatflow AI Agent:

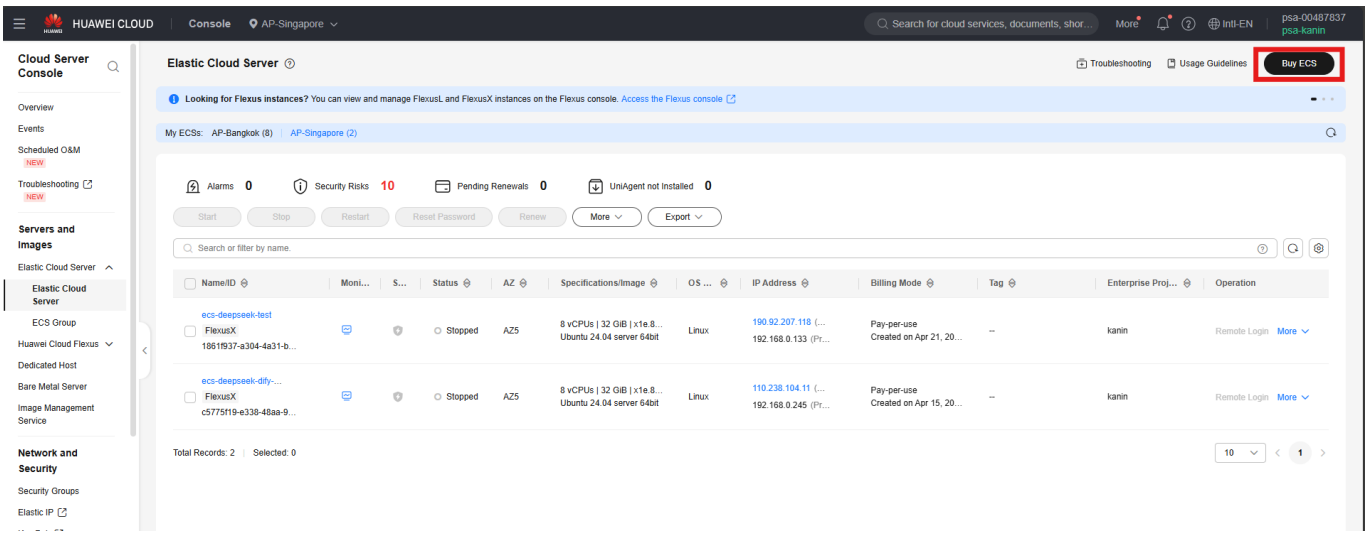


- ECS Recommended configuration
  - vCPU: 4-8
  - Memory: 16-32GB
  - OS: Ubuntu
  - EIP: Traffic bandwidth size (100 Mbit/s)

1. Create ECS



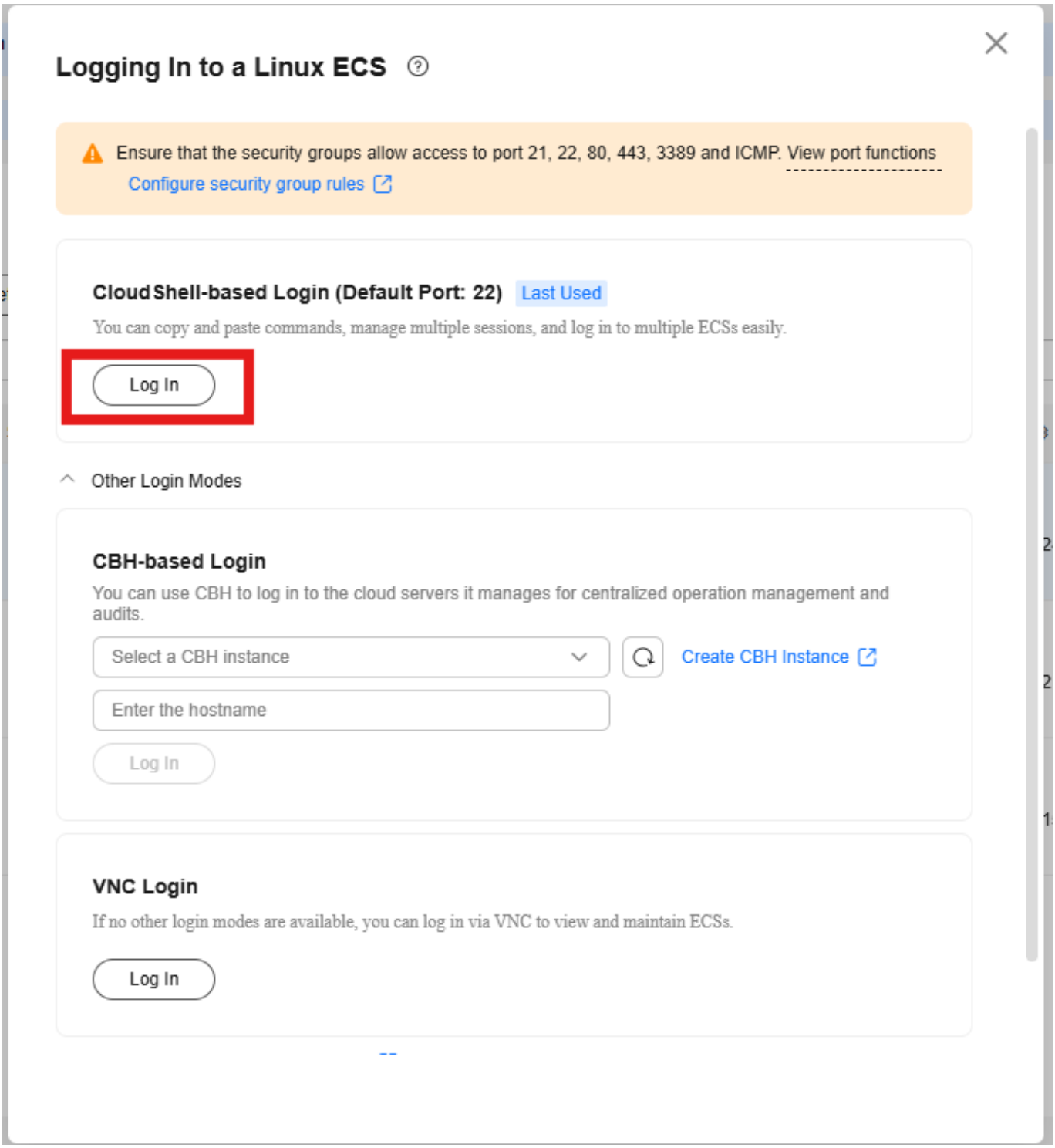




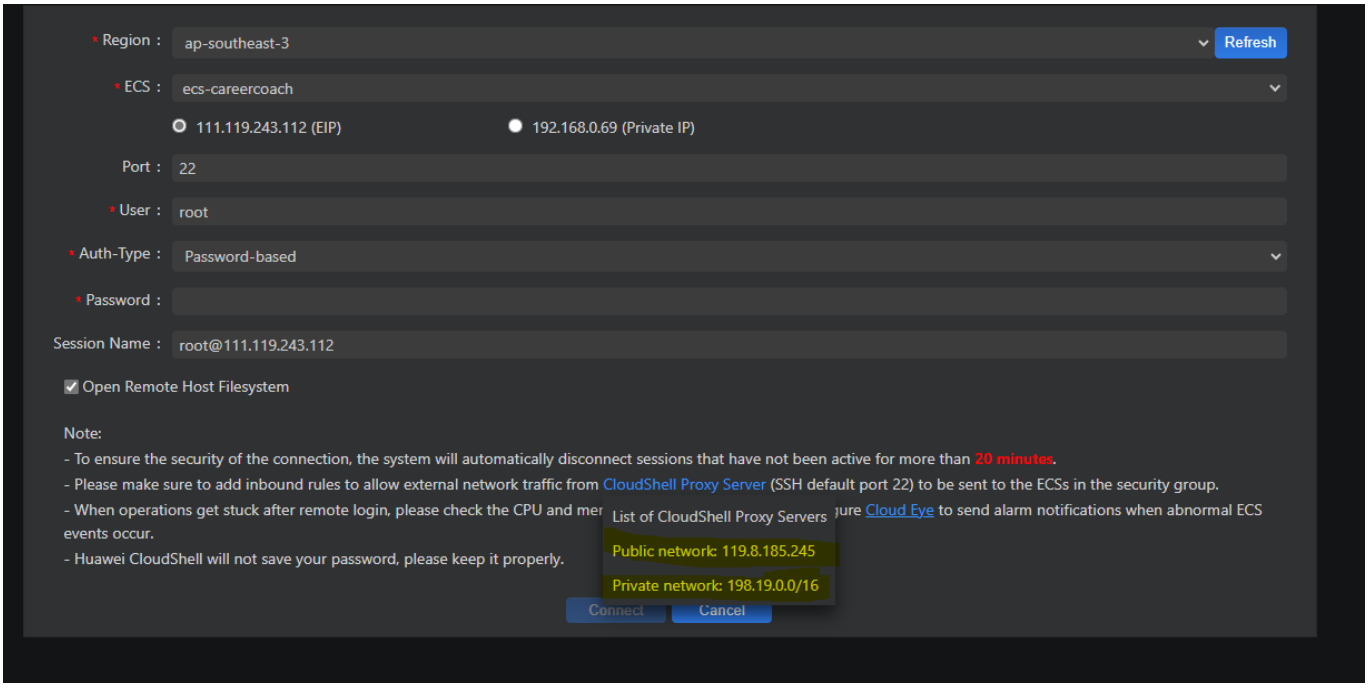
- custom config
- billing mode: pay-per-use
- Region: Brazil

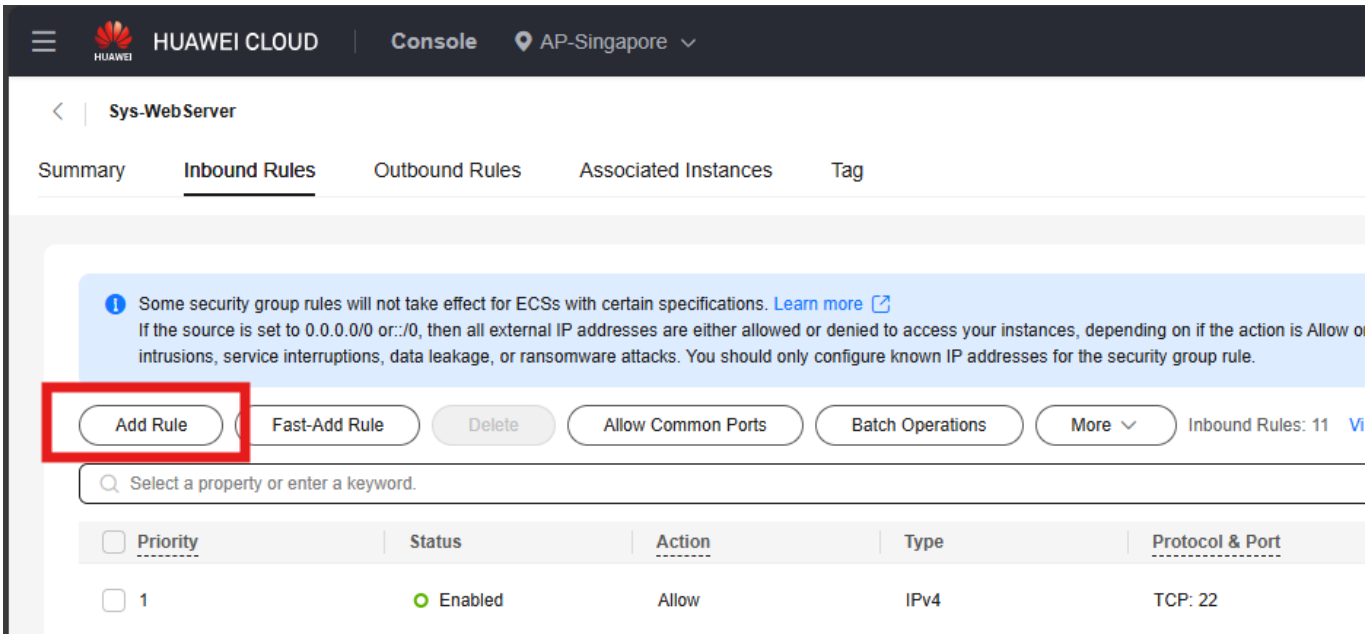
- AZ: Random
- Instance: General computing-plus x1e, x1e.8u.32g, 8 vCPUs, 32 GiB
- OS: Ubuntu 24.04
- system disk: 40GB
- Network: use your own VPC & subnet
- Security group: default
- Public Network Access: EIP, Auto Assign, Traffic bandwidth size (100 Mbit/s), release with ECS
- ECS Name: you can decide
- Login mode: password
- Enterprise project: can select default

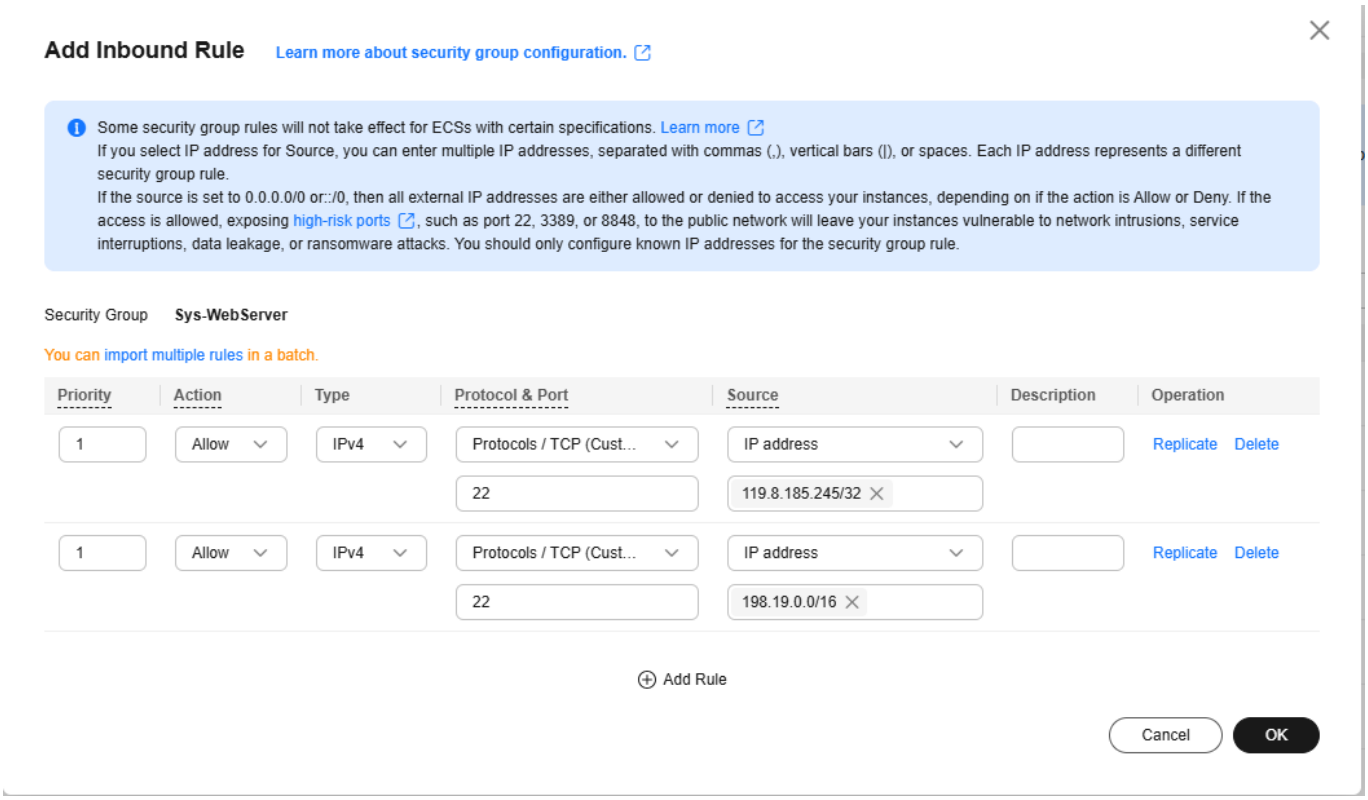
After the ECS is created and running, remote login into ECS:



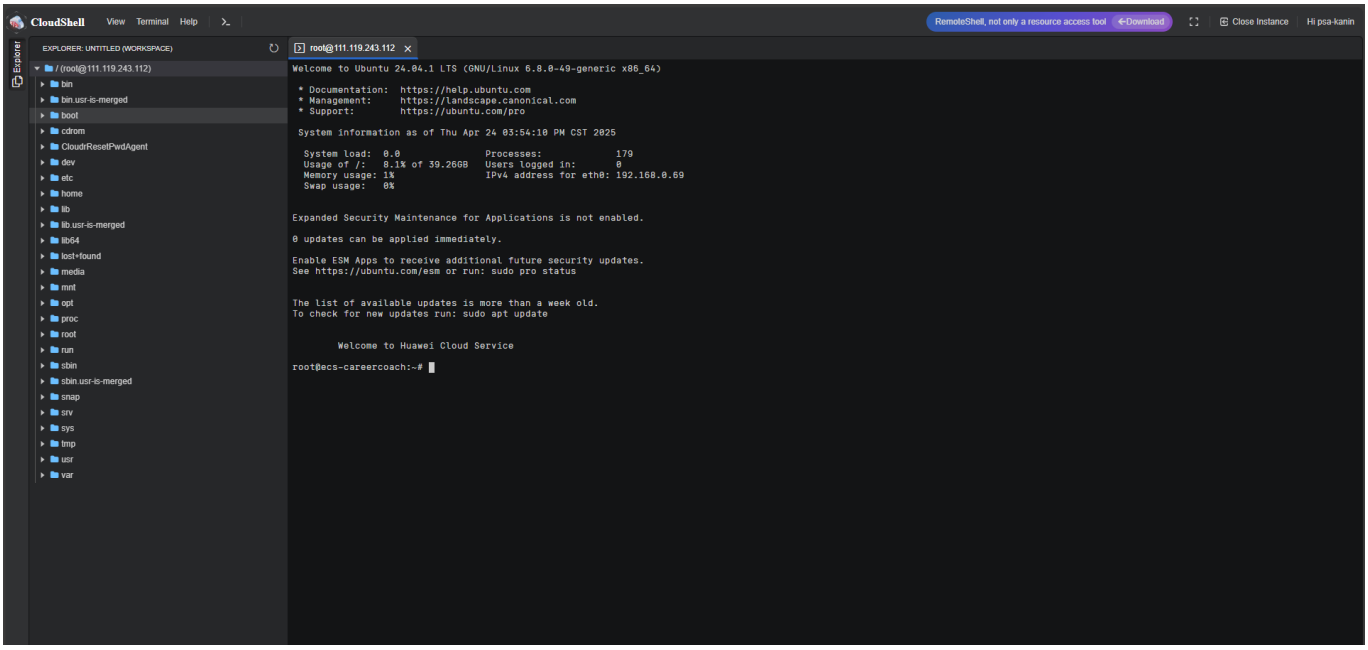
need to edit security group:




HUAWEI CLOUD
Console
AP-Singapore
Sys-WebServer
Summary
Inbound Rules
Outbound Rules
Associated Instances
Tag
Some security group rules will not take effect for ECSs with certain specifications. Learn more
If the source is set to 0.0.0.0/0 or::/0, then all external IP addresses are either allowed or denied to access your instances, depending on if the action is Allow o
intrusions, service interruptions, data leakage, or ransomware attacks. You should only configure known IP addresses for the security group rule.
Add Rule
Fast-Add Rule
Delete
Allow Common Ports
Batch Operations
More
Inbound Rules: 11
Select a property or enter a keyword.
Priority
Status
Action
Type
Protocol & Port
1
Enabled
Allow
IPv4
TCP: 22


Add Inbound Rule
Learn more about security group configuration.
Some security group rules will not take effect for ECSs with certain specifications. Learn more
If you select IP address for Source, you can enter multiple IP addresses, separated with commas (,), vertical bars (|), or spaces. Each IP address represents a different security group rule.
If the source is set to 0.0.0.0/0 or::/0, then all external IP addresses are either allowed or denied to access your instances, depending on if the action is Allow or Deny. If the access is allowed, exposing high-risk ports, such as port 22, 3389, or 8848, to the public network will leave your instances vulnerable to network intrusions, service interruptions, data leakage, or ransomware attacks. You should only configure known IP addresses for the security group rule.
Security Group
Sys-WebServer
You can import multiple rules in a batch.
Priority
Action
Type
Protocol & Port
Source
Description
Operation
1
Allow
IPv4
Protocols / TCP (Cust...
IP address
22
119.8.185.245/32
Replicate
Delete
1
Allow
IPv4
Protocols / TCP (Cust...
IP address
22
198.19.0.0/16
Replicate
Delete
Add Rule
Cancel
OK

After login into ECS:

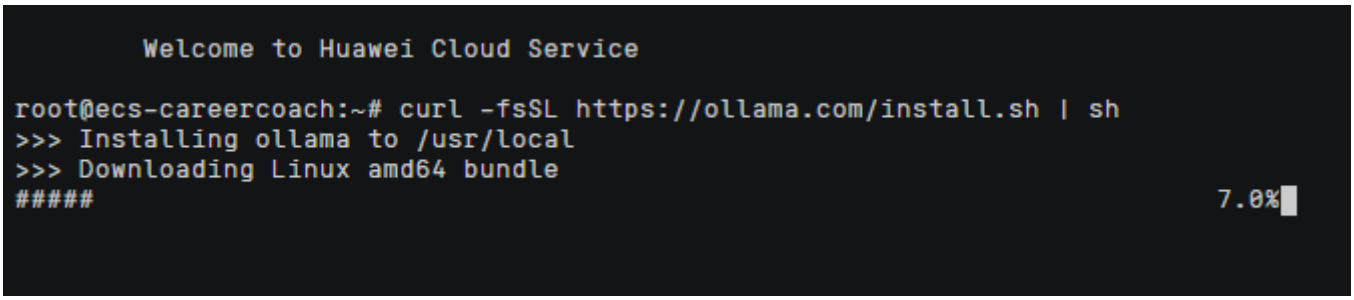


2. Download Ollama

[Download Ollama on Linux] (https://ollama.com/download/linux)

Install with below command:

```
curl -fsSL https://ollama.com/install.sh | sh
```



After installation is completed, verify ollama version

```
ollama -v
```

```
root@ecs-careercoach:~# curl -fsSL https://ollama.com/install.sh | sh
>>> Installing ollama to /usr/local
>>> Downloading Linux amd64 bundle
######################################################################## 100.0%
>>> Creating ollama user...
>>> Adding ollama user to render group...
>>> Adding ollama user to video group...
>>> Adding current user to ollama group...
>>> Creating ollama systemd service...
>>> Enabling and starting ollama service...
Created symlink /etc/systemd/system/default.target.wants/ollama.service → /etc/systemd/system/ollama.service.
>>> The Ollama API is now available at 127.0.0.1:11434.
>>> Install complete. Run "ollama" from the command line.
WARNING: No NVIDIA/AMD GPU detected. Ollama will run in CPU-only mode.
root@ecs-careercoach:~# ollama -v
ollama version is 0.6.6
root@ecs-careercoach:~#
```

3. Download and run DeepSeek Distilled models

[deepseek-r1] (https://ollama.com/library/deepseek-r1)

- DeepSeek-R1-Distill-Qwen-1.5B

```
ollama run deepseek-r1:1.5b
```

- DeepSeek-R1-Distill-Qwen-7B

```
ollama run deepseek-r1:7b
```

- DeepSeek-R1-Distill-Llama-8B

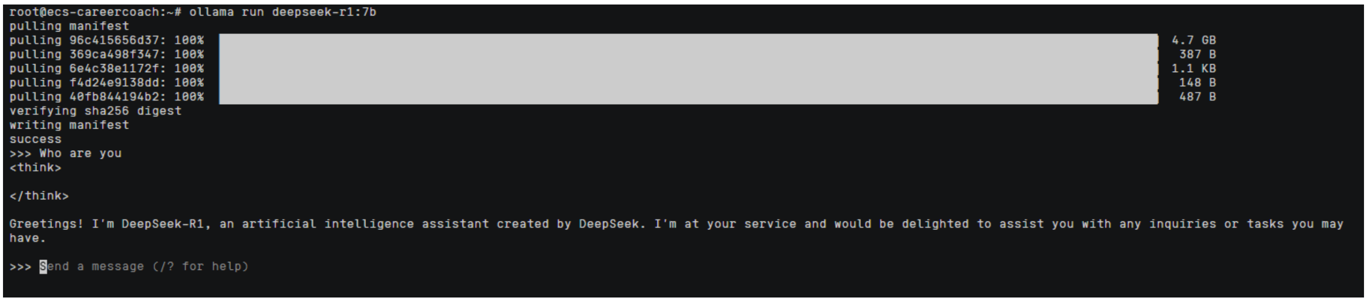
```
ollama run deepseek-r1:8b
```

- DeepSeek-R1-Distill-Qwen-14B

```
ollama run deepseek-r1:14b
```

This demo we choose `deepseek-r1:7b`

```
root@ecs-careercoach:~# ollama run deepseek-r1:7b
pulling manifest
pulling 96c415656d37:   5%                                                  | 243 MB/4.7 GB   54 MB/s   1m22s
```



```
root@ecs-careercoach:~# ollama run deepseek-r1:7b
pulling manifest
pulling 96c415656d37: 100%  4.7 GB
pulling 369ca498f347: 100%   387 B
pulling 6e4c38e1172f: 100%  1.1 KB
pulling f4d24e9138dd: 100%   148 B
pulling 40fb844194b2: 100%   487 B
verifying sha256 digest
writing manifest
success
>>> Who are you
<think>

</think>

Greetings! I'm DeepSeek-R1, an artificial intelligence assistant created by DeepSeek. I'm at your service and would be delighted to assist you with any inquiries or tasks you may
have.

>>> Send a message (/? for help)
```

to exit chatbot mode:

```
/exit
```

4. Download and install Docker compose

[Ubuntu | Docker Docs]([https://docs.docker.com/engine/install/ubuntu/]

4.1 Uninstall old versions

Run the following command to uninstall all conflicting packages:

```
for pkg in docker.io docker-doc docker-compose docker-compose-v2 podman-docker
containerd runc; do sudo apt-get remove $pkg; done
```

4.2 Install using the apt repository

Before you install Docker Engine for the first time on a new host machine, you need to set up the Docker apt repository. Afterward, you can install and update Docker from the repository.

Set up Docker's apt repository:

```
# Add Docker's official GPG key:
sudo apt-get update
sudo apt-get install ca-certificates curl
sudo install -m 0755 -d /etc/apt/keyrings
sudo curl -fsSL https://download.docker.com/linux/ubuntu/gpg -o
/etc/apt/keyrings/docker.asc
sudo chmod a+r /etc/apt/keyrings/docker.asc

# Add the repository to Apt sources:
echo \
  "deb [arch=$(dpkg --print-architecture) signed-by=/etc/apt/keyrings/docker.asc]
https://download.docker.com/linux/ubuntu \
  $(. /etc/os-release && echo "${UBUNTU_CODENAME:-$VERSION_CODENAME}") stable" | \
  sudo tee /etc/apt/sources.list.d/docker.list > /dev/null

sudo apt-get update
```

4.3 Install the Docker packages

To install the latest version, run:

```
sudo apt-get install docker-ce docker-ce-cli containerd.io docker-buildx-plugin
docker-compose-plugin
```

Verify that the installation is successful by running the hello-world image:

```
sudo docker run hello-world
```

This command downloads a test image and runs it in a container. When the container runs, it prints a confirmation message and exits:

```
root@ecs-careercoach:~# sudo docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
e6590344b1a5: Pull complete
Digest: sha256:c41088499908a59aae84b0a49c70e86f4731e588a737f1637e73c8c09d995654
Status: Downloaded newer image for hello-world:latest

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
 1. The Docker client contacted the Docker daemon.
 2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
    (amd64)
 3. The Docker daemon created a new container from that image which runs the
    executable that produces the output you are currently reading.
 4. The Docker daemon streamed that output to the Docker client, which sent it
    to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
 $ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
 https://hub.docker.com/

For more examples and ideas, visit:
 https://docs.docker.com/get-started/

root@ecs-careercoach:~#
```

5. Deploy Dify with Docker Compose

5.1 Clone Dify

```
# Assuming current latest version is 0.15.3
git clone https://github.com/langgenius/dify.git --branch 0.15.3
```

```
root@ecs-careercoach:~# git clone https://github.com/langgenius/dify.git --branch 0.15.3
Cloning into 'dify'...
remote: Enumerating objects: 156047, done.
remote: Counting objects: 100% (500/500), done.
remote: Compressing objects: 100% (281/281), done.
remote: Total 156047 (delta 333), reused 253 (delta 215), pack-reused 155547 (from 3)
Receiving objects: 100% (156047/156047), 87.89 MiB | 16.87 MiB/s, done.
Resolving deltas: 100% (112381/112381), done.
Note: switching to 'ca19bd31d42fb87c83b91541c473ebae85e9d14e'.

You are in 'detached HEAD' state. You can look around, make experimental
changes and commit them, and you can discard any commits you make in this
state without impacting any branches by switching back to a branch.

If you want to create a new branch to retain commits you create, you may
do so (now or later) by using -c with the switch command. Example:

  git switch -c <new-branch-name>

Or undo this operation with:

  git switch -

Turn off this advice by setting config variable advice.detachedHead to false

root@ecs-careercoach:~#
```

5.2 Starting Dify

Navigate to the Docker directory in the Dify source code:

```
cd dify/docker
```

Copy the environment configuration file

```
cp .env.example .env
```

5.3 Start the Docker containers

```
docker compose up -d
```

```
root@ecs-careercoach:~# cd dify/docker
root@ecs-careercoach:~/dify/docker# cp .env.example .env
root@ecs-careercoach:~/dify/docker# docker compose up -d
[+] Running 14/75
 ⠏ sandbox [⣿⣿⣿⣿      ] Pulling                                                                                                                                    11.1s
 ⠏ db [          ] Pulling                                                                                                                                         11.1s
 ⠏ weaviate [⣿ ⣿⣿] 3.545MB / 20.09MB Pulling                                                                                                                       11.1s
 ⠏ redis [         ] Pulling                                                                                                                                       11.1s
 ⠏ ssrf_proxy [   ] Pulling                                                                                                                                        11.1s
 ✓ nginx Pulled                                                                                                                                                     8.0s
 ⠏ worker Pulling                                                                                                                                                  11.1s
 ⠏ web [             ] Pulling                                                                                                                                     11.1s
 ⠏ api [           ] Pulling                                                                                                                                       11.1s
```

```
root@ecs-careercoach:~/dify/docker# docker compose up -d
[+] Running 75/75
 ✓ sandbox Pulled                                                                                                                                                  22.5s
 ✓ db Pulled                                                                                                                                                       32.4s
 ✓ weaviate Pulled                                                                                                                                                 59.9s
 ✓ redis Pulled                                                                                                                                                    21.6s
 ✓ ssrf_proxy Pulled                                                                                                                                               95.4s
 ✓ nginx Pulled                                                                                                                                                     8.0s
 ✓ worker Pulled                                                                                                                                                  139.3s
 ✓ web Pulled                                                                                                                                                      45.0s
 ✓ api Pulled                                                                                                                                                     139.3s
[+] Running 11/11
 ✓ Network docker_default             Created                                                                                                                       0.0s
 ✓ Network docker_ssrf_proxy_network  Created                                                                                                                       0.0s
 ✓ Container docker-db-1              Started                                                                                                                      21.7s
 ✓ Container docker-web-1             Started                                                                                                                      21.7s
 ✓ Container docker-sandbox-1         Started                                                                                                                      21.7s
 ✓ Container docker-weaviate-1        Started                                                                                                                      21.7s
 ✓ Container docker-ssrf_proxy-1      Started                                                                                                                      21.8s
 ✓ Container docker-redis-1           Started                                                                                                                      21.7s
 ✓ Container docker-worker-1          Started                                                                                                                       0.6s
 ✓ Container docker-api-1             Started                                                                                                                       0.6s
 ✓ Container docker-nginx-1           Started                                                                                                                       0.8s
root@ecs-careercoach:~/dify/docker# 
```

Finally, check if all containers are running successfully:

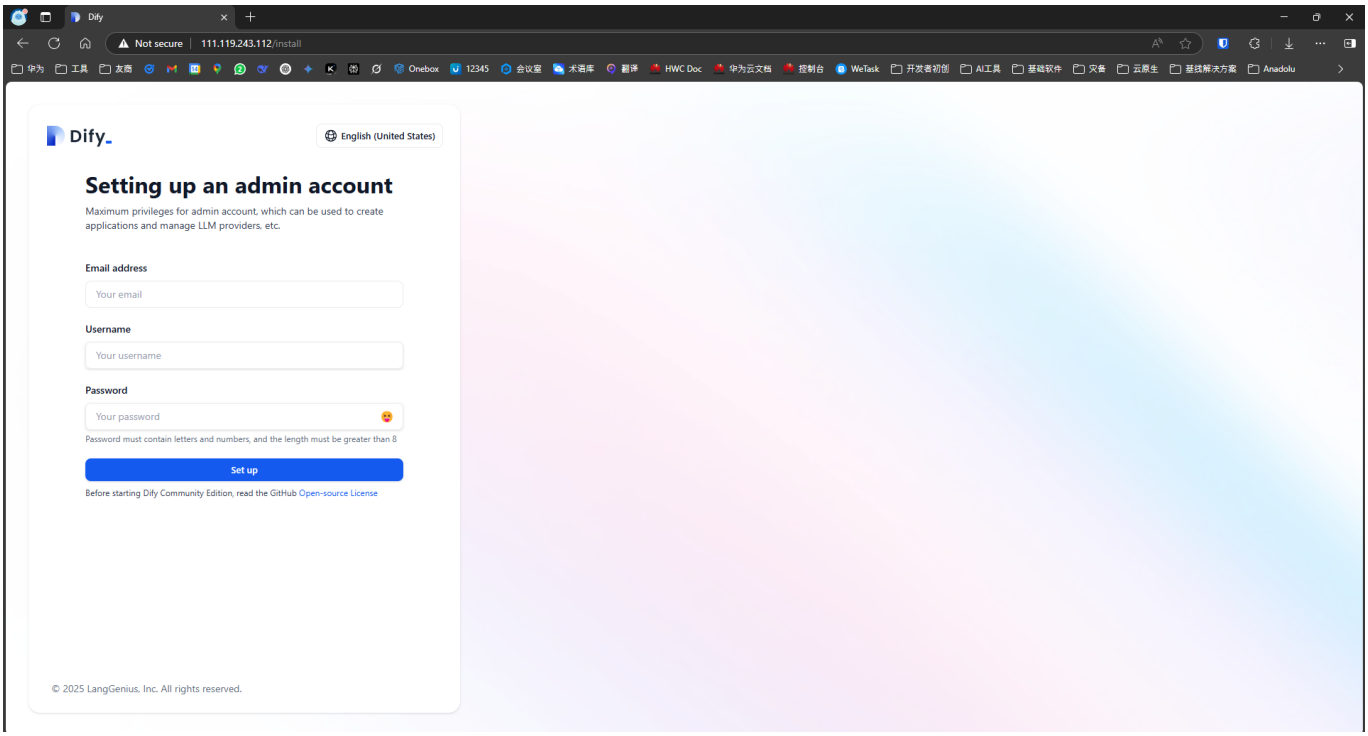
```
docker compose ps
```

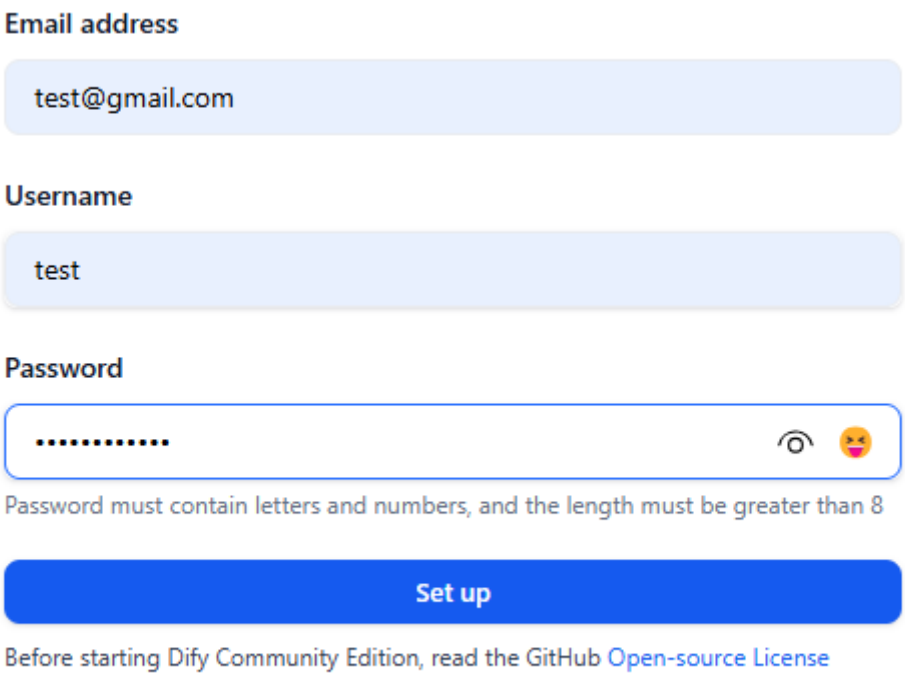
```
root@ecs-careercoach:~/dify/docker# docker compose ps
NAME                  IMAGE                                  COMMAND                    SERVICE      CREATED              STATUS                         PORTS
docker-api-1          langgenius/dify-api:0.15.3             "/bin/bash /entrypoi…"     api          About a minute ago   Up About a minute              5001/tcp
docker-db-1           postgres:15-alpine                     "docker-entrypoint.s…"     db           About a minute ago   Up About a minute (healthy)    5432/tcp
docker-nginx-1        nginx:latest                           "sh -c 'cp /docker-e…"     nginx        About a minute ago   Up About a minute              0.0.0.0:80->80/tcp, [::]:80->80/tcp, 0.0.
0.0:443->443/tcp, [::]:443->443/tcp
docker-redis-1        redis:6-alpine                         "docker-entrypoint.s…"     redis        About a minute ago   Up About a minute (healthy)    6379/tcp
docker-sandbox-1      langgenius/dify-sandbox:0.2.10         "/main"                    sandbox      About a minute ago   Up About a minute (healthy)
docker-ssrf_proxy-1   ubuntu/squid:latest                    "sh -c 'cp /docker-e…"     ssrf_proxy   About a minute ago   Up About a minute              3128/tcp
docker-weaviate-1     semitechnologies/weaviate:1.19.0       "/bin/weaviate --hos…"     weaviate     About a minute ago   Up About a minute
docker-web-1          langgenius/dify-web:0.15.3             "/bin/sh ./entrypoin…"     web          About a minute ago   Up About a minute              3000/tcp
docker-worker-1       langgenius/dify-api:0.15.3             "/bin/bash /entrypoi…"     worker       About a minute ago   Up About a minute              5001/tcp
root@ecs-careercoach:~/dify/docker# 
```

With these steps, you should be able to install Dify successfully.

5.4 Access Dify

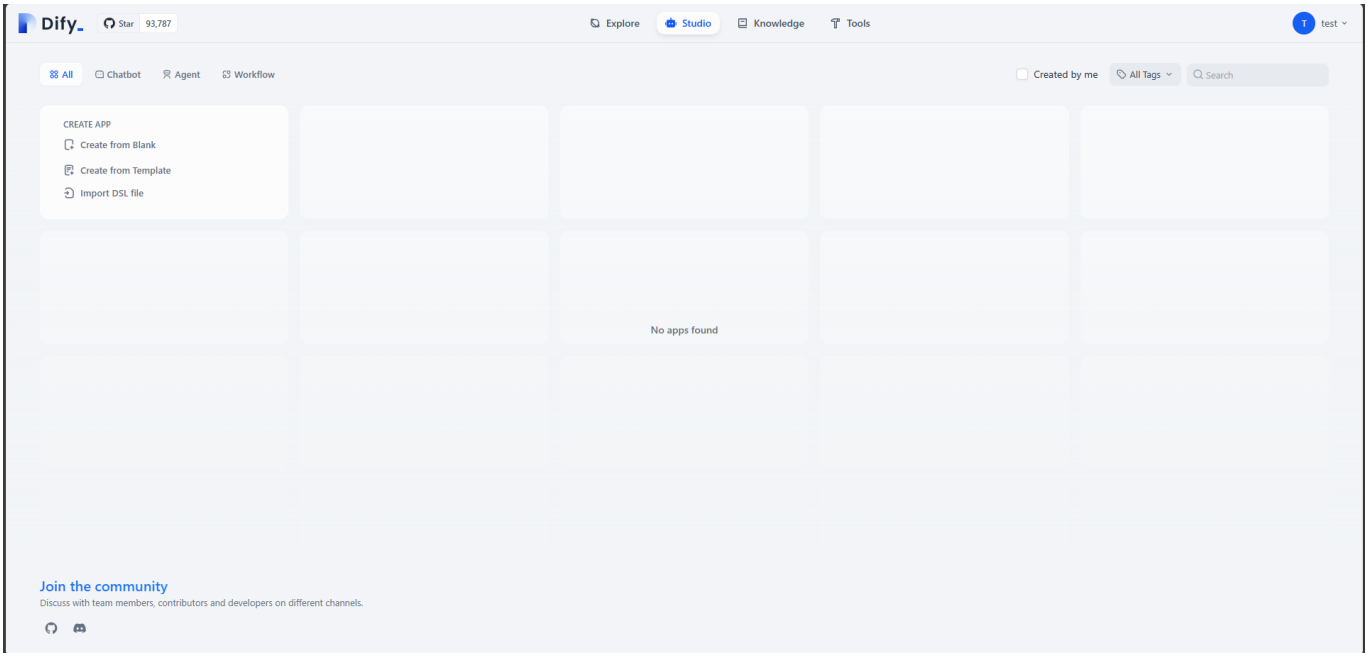
```
EIP/install
```



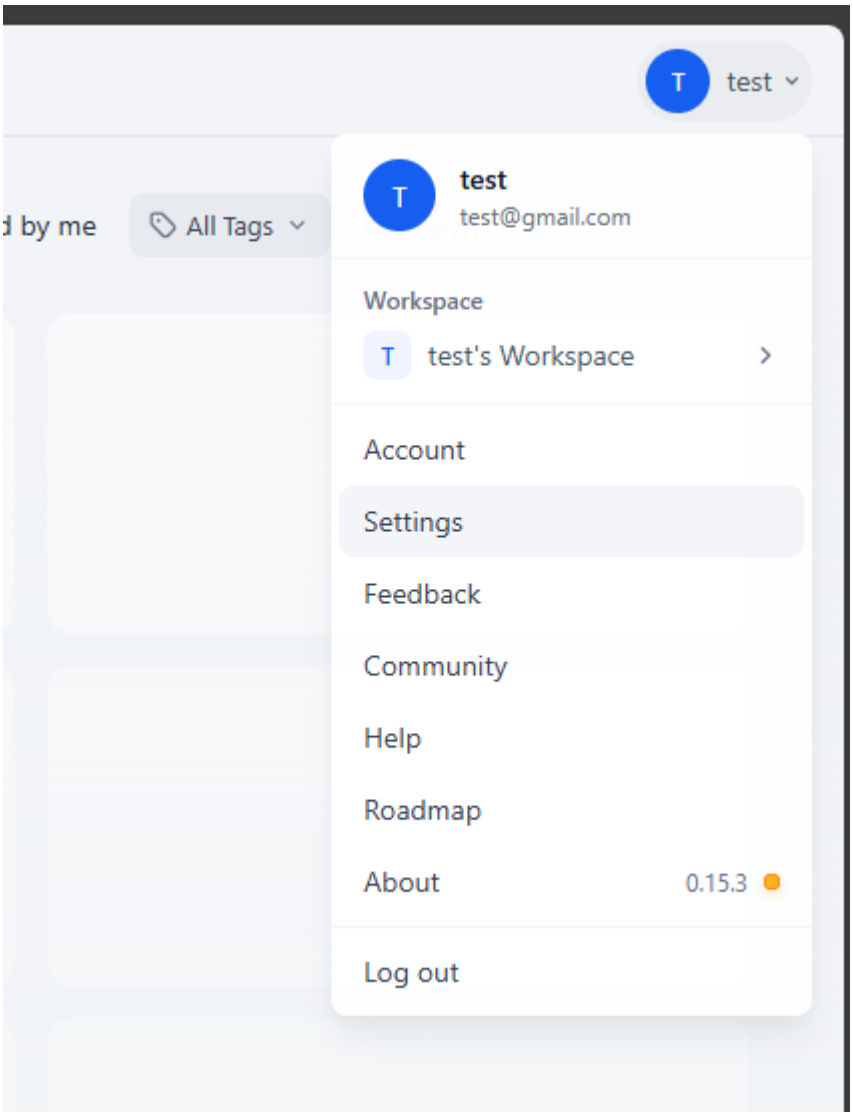


Setup and sign in

Home page (Studio) for Dify

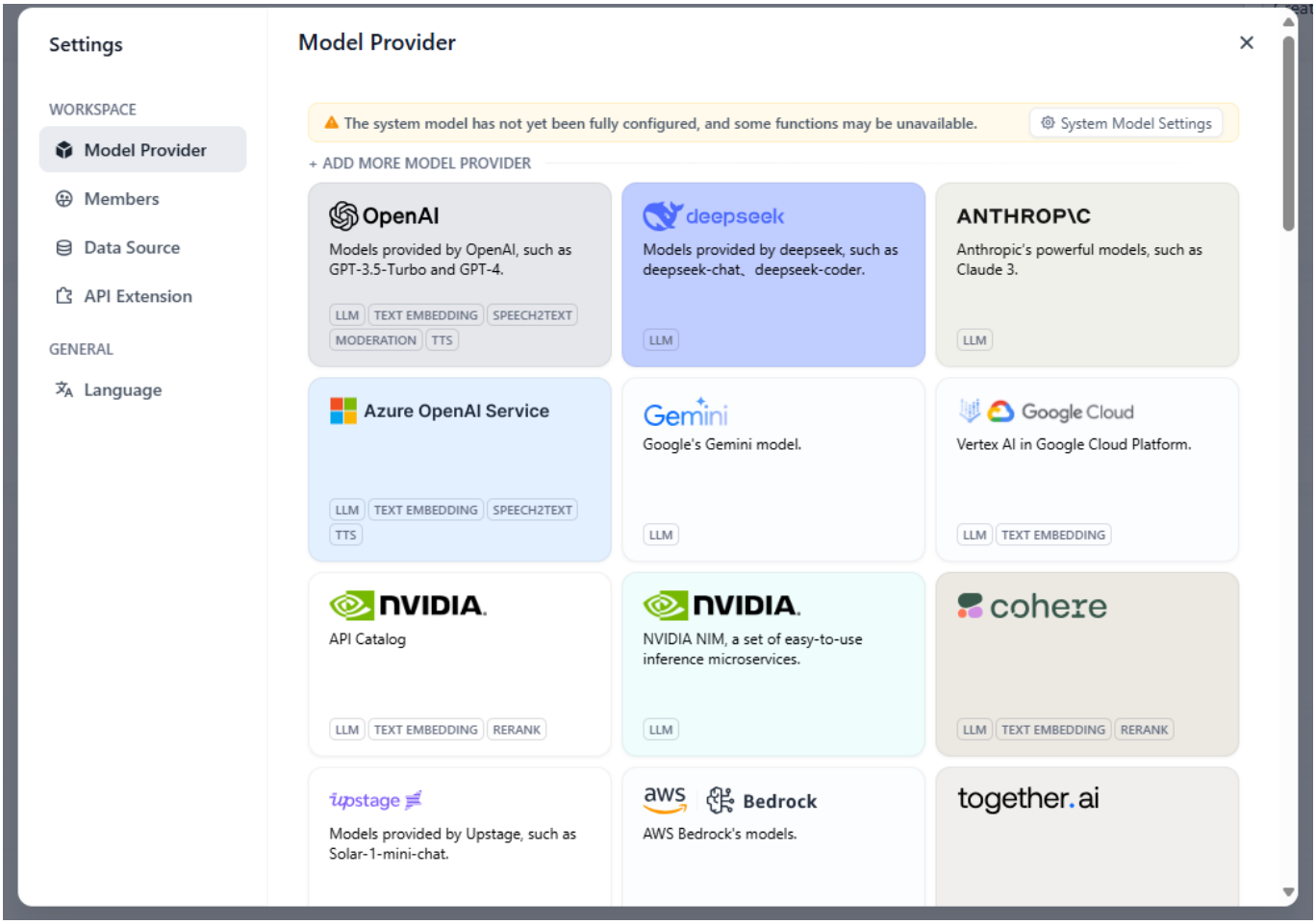


6. Link Ollama and other LLM service provider to Dify



Go to settings

Model Providers



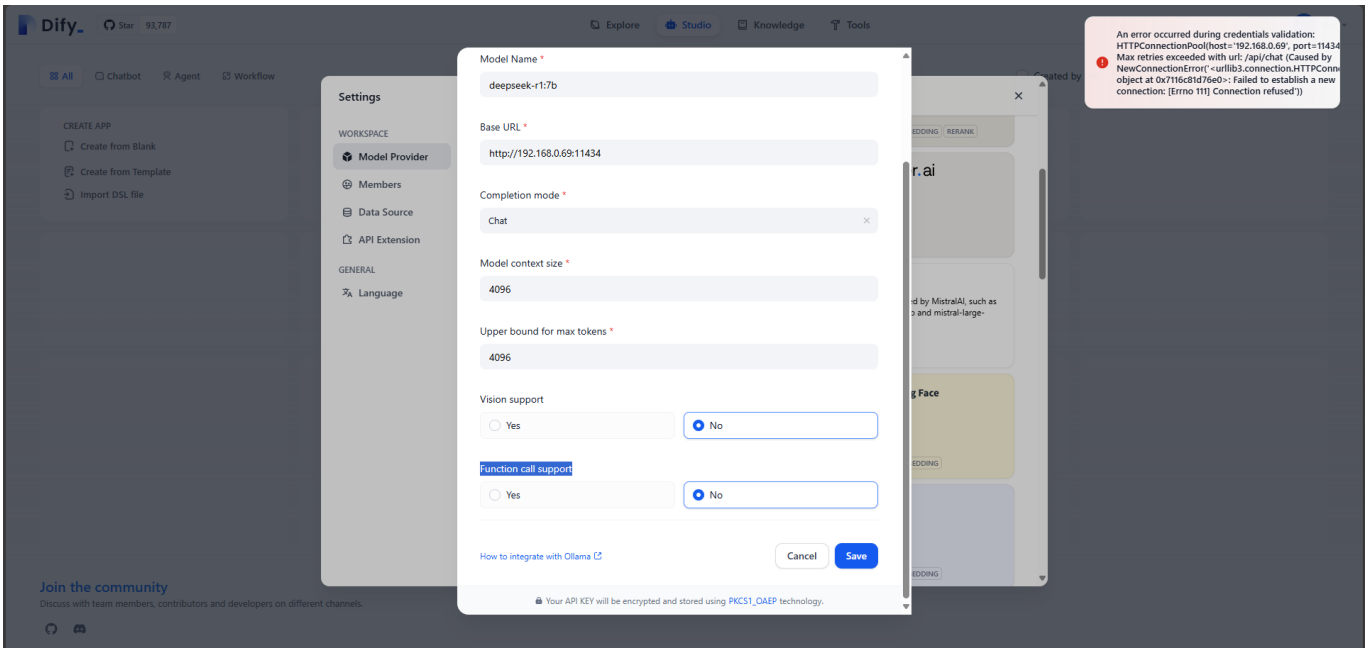
6.1 Ollama + DeepSeek

Integrate Local Models Deployed by Ollama | Dify

Search for ollama, add model

- model name: deepseek-r1:7b
- Base URL: `http://{Private IP}:11434`
- Completion mode: Chat
- Model context size: 4096
- Upper bound for max tokens: 4096
- Vision support: No
- Function call support: No

6.1.1 If you are using docker to deploy Dify and Ollama, you may encounter the following error




```
httpconnectionpool(host=127.0.0.1, port=11434): max retries exceeded with
url:/cpi/chat (Caused by NewConnectionError('<urllib3.connection.HTTPConnection
object at 0x7f8562812c20>: fail to establish a new connection:[Errno 111]
Connection refused'))

httpconnectionpool(host=localhost, port=11434): max retries exceeded with
url:/cpi/chat (Caused by NewConnectionError('<urllib3.connection.HTTPConnection
object at 0x7f8562812c20>: fail to establish a new connection:[Errno 111]
Connection refused'))
```

This error occurs because the Ollama service is not accessible from the docker container. `localhost` usually refers to the container itself, not the host machine or other containers.

You need to expose the Ollama service to the network to resolve this issue.

Setting environment variables on Linux If Ollama is run as a systemd service, environment variables should be set using `systemctl`:

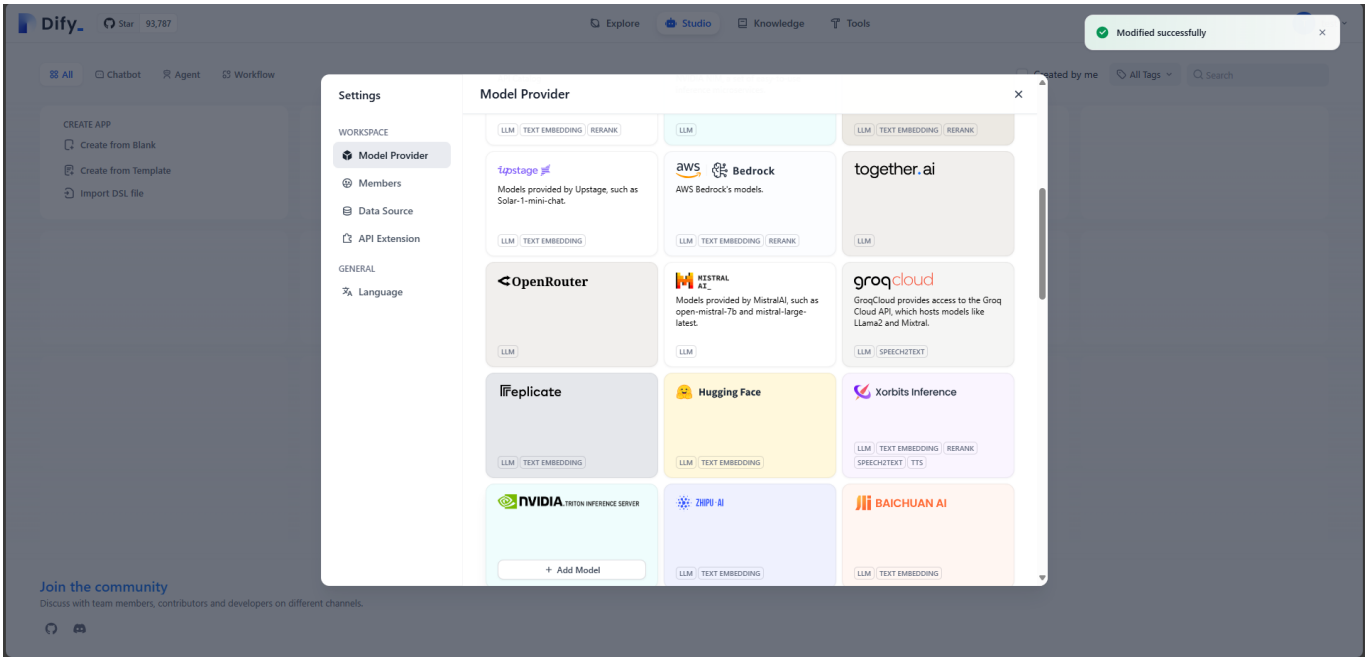
Edit the systemd service by calling `systemctl edit ollama.service`. This will open an editor. For each environment variable, add a line `Environment` under section `[Service]`:

```
[Service]
Environment="OLLAMA_HOST=0.0.0.0"
```

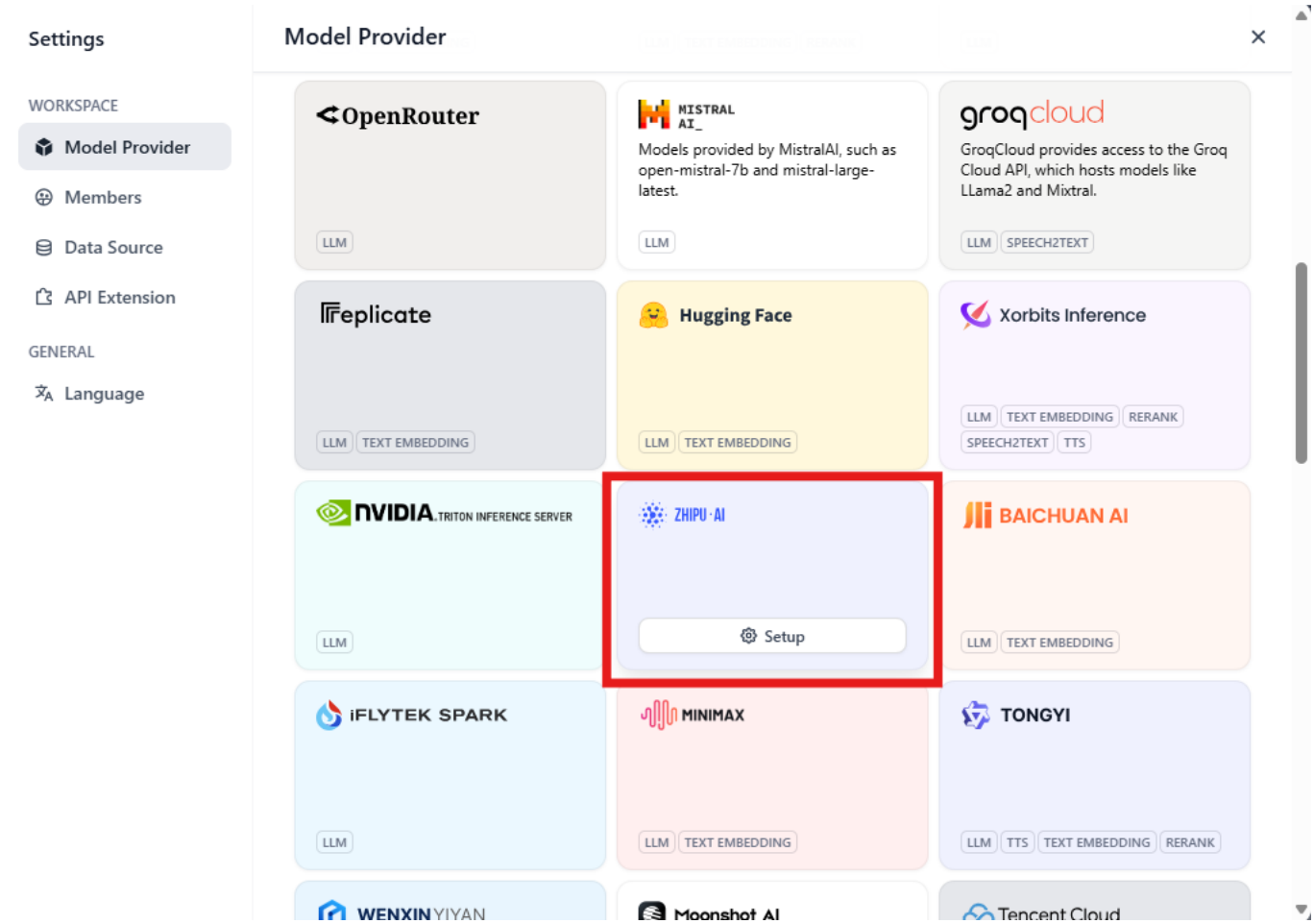
Save and exit.(ctrl+o, enter, ctrl+x) Reload `systemd` and restart Ollama:

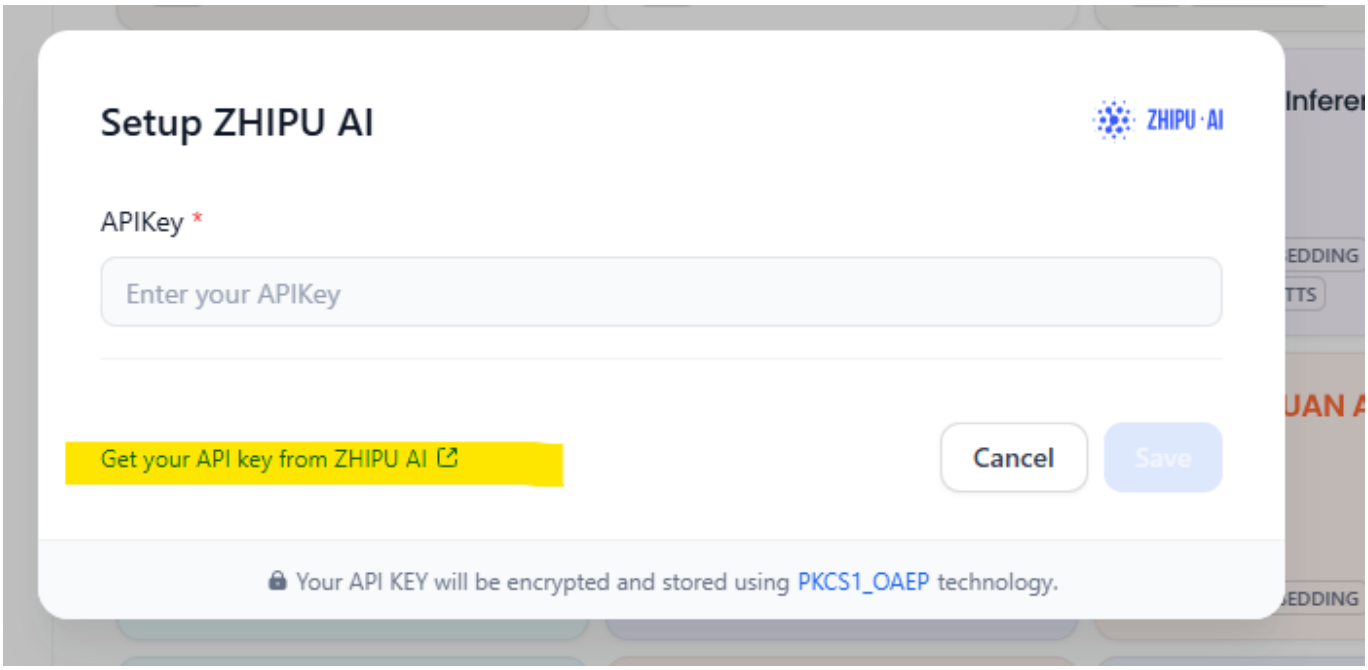
```
systemctl daemon-reload
systemctl restart ollama
```

After editing, you can add ollama model into Dify

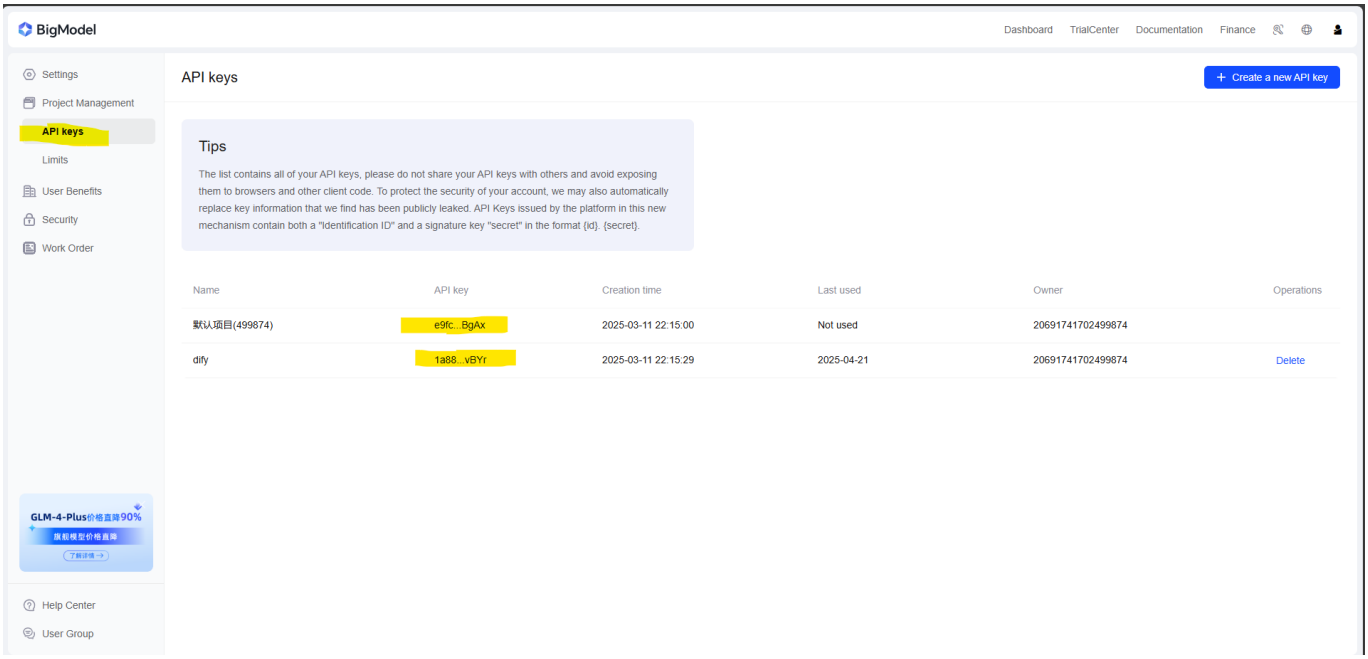


6.2 Zhipu AI

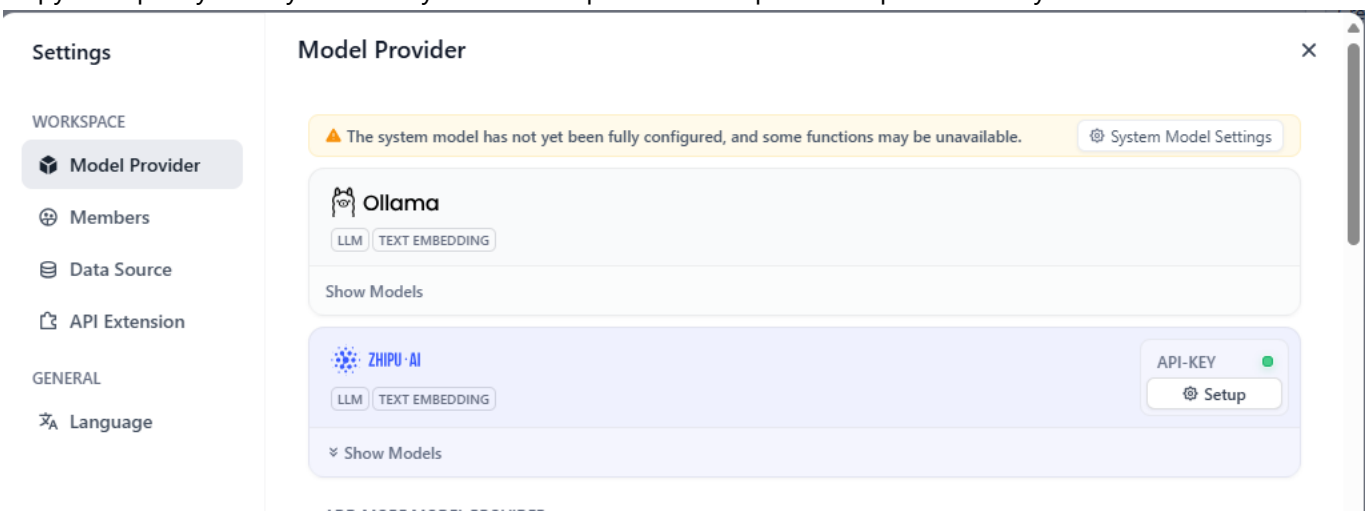

Settings
WORKSPACE
Model Provider
Members
Data Source
API Extension
GENERAL
Language
Model Provider
OpenRouter
LLM
MISTRAL AI_
Models provided by MistralAI, such as open-mistral-7b and mistral-large-latest.
LLM
groqcloud
GroqCloud provides access to the Groq Cloud API, which hosts models like LLama2 and Mixtral.
LLM SPEECH2TEXT
replicate
LLM TEXT EMBEDDING
Hugging Face
LLM TEXT EMBEDDING
Xorbits Inference
LLM TEXT EMBEDDING RERANK SPEECH2TEXT TTS
NVIDIA TRITON INFERENCE SERVER
LLM
ZHIPU·AI
Setup
BAICHUAN AI
LLM TEXT EMBEDDING
iFLYTEK SPARK
LLM
MINIMAX
LLM TEXT EMBEDDING
TONGYI
LLM TTS TEXT EMBEDDING RERANK
WENXIN YIYAN
Moonshot AI
Tencent Cloud


Setup ZHIPU AI
ZHIPU·AI
APIKey *
Enter your APIKey
Get your API key from ZHIPU AI
Cancel
Save
Your API KEY will be encrypted and stored using PKCS1_OAEP technology.

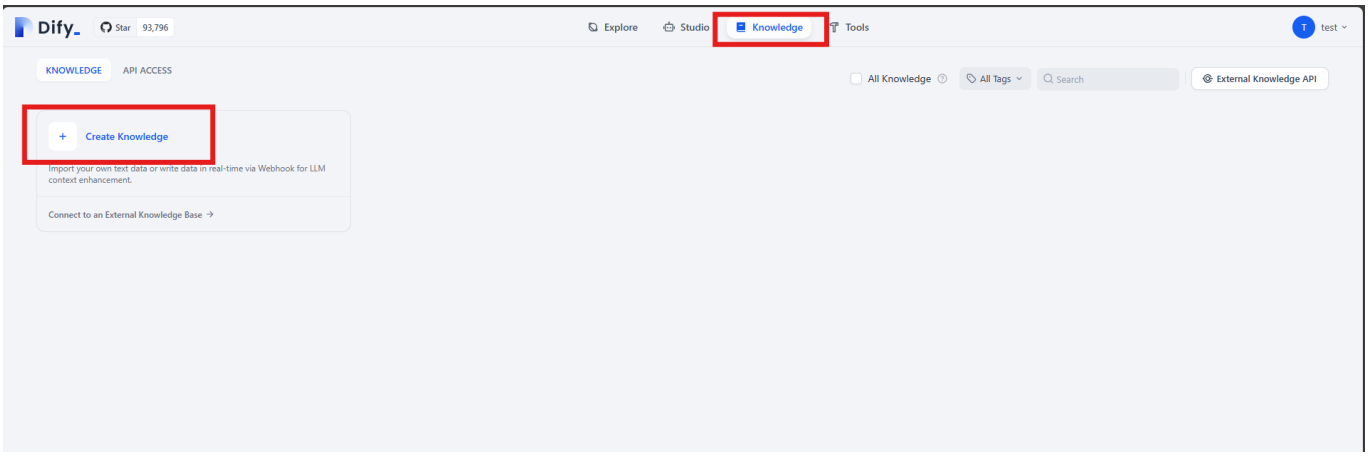
Use your email to signup for a free account and get free token to use its LLM (obtain the API key)



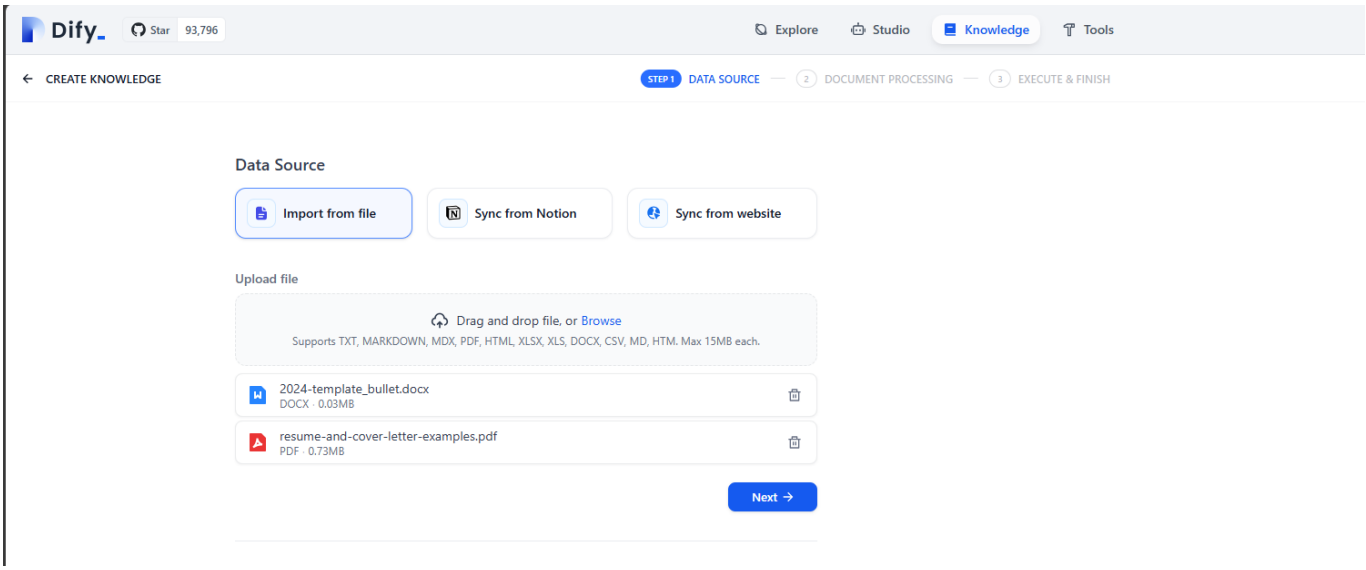
Copy the api key to Dify and now you have deepseek and zhipu AI setup successfully.



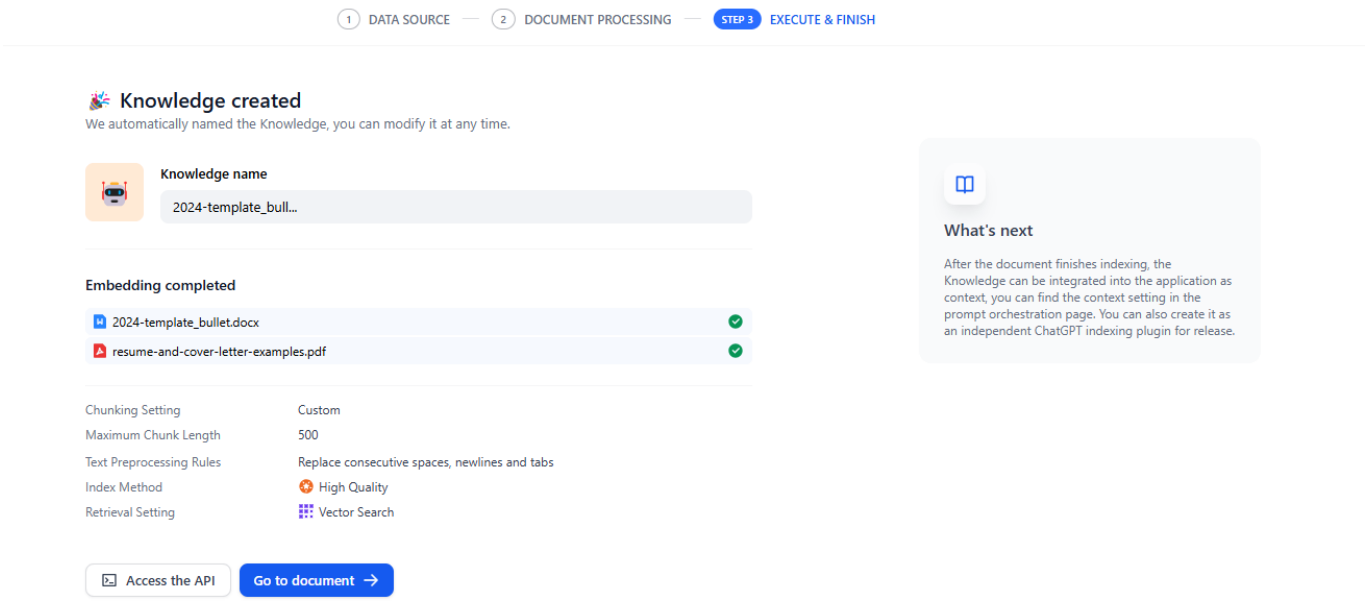
7. Add a Knowledge base



We can upload some example resume template to this knowledge base, you can search from google to obtain Harvard university resume and cover letter template to upload into this knowledge base.

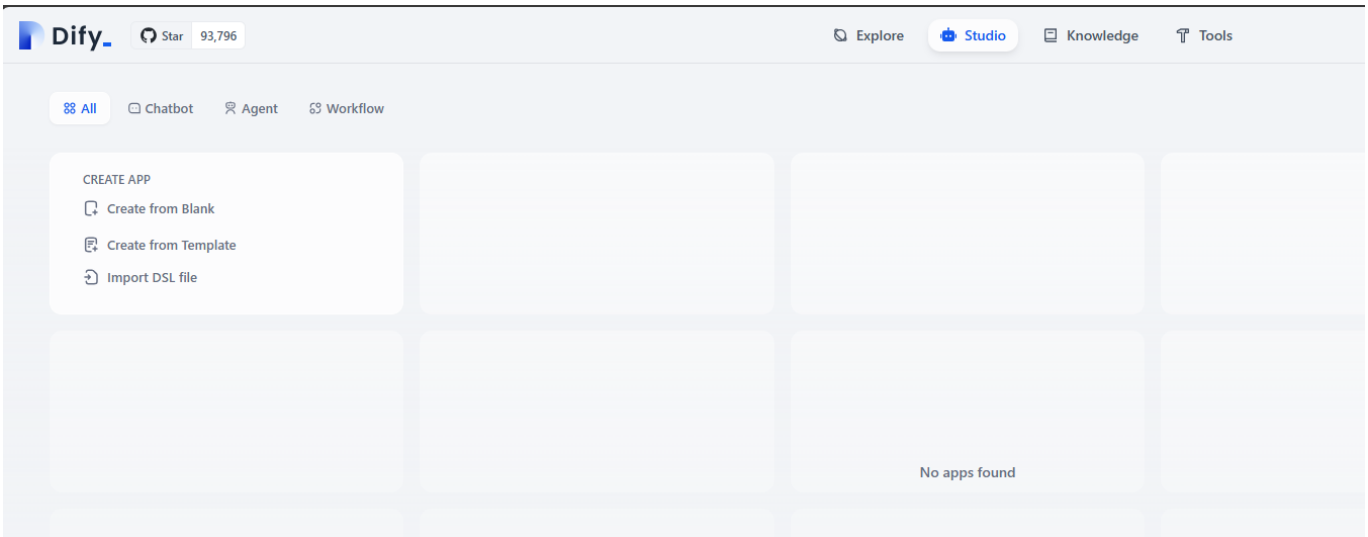


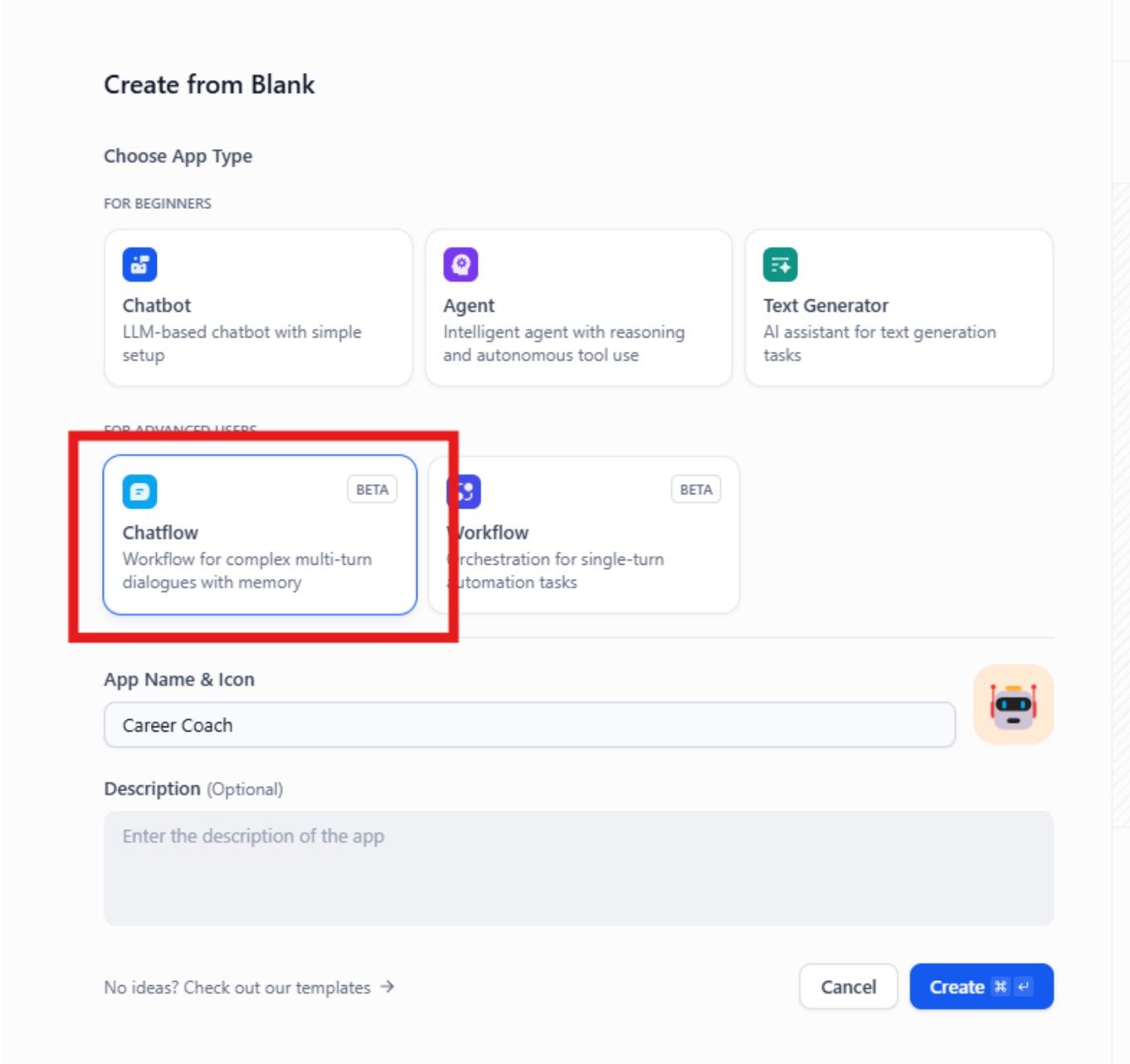
Save and Process

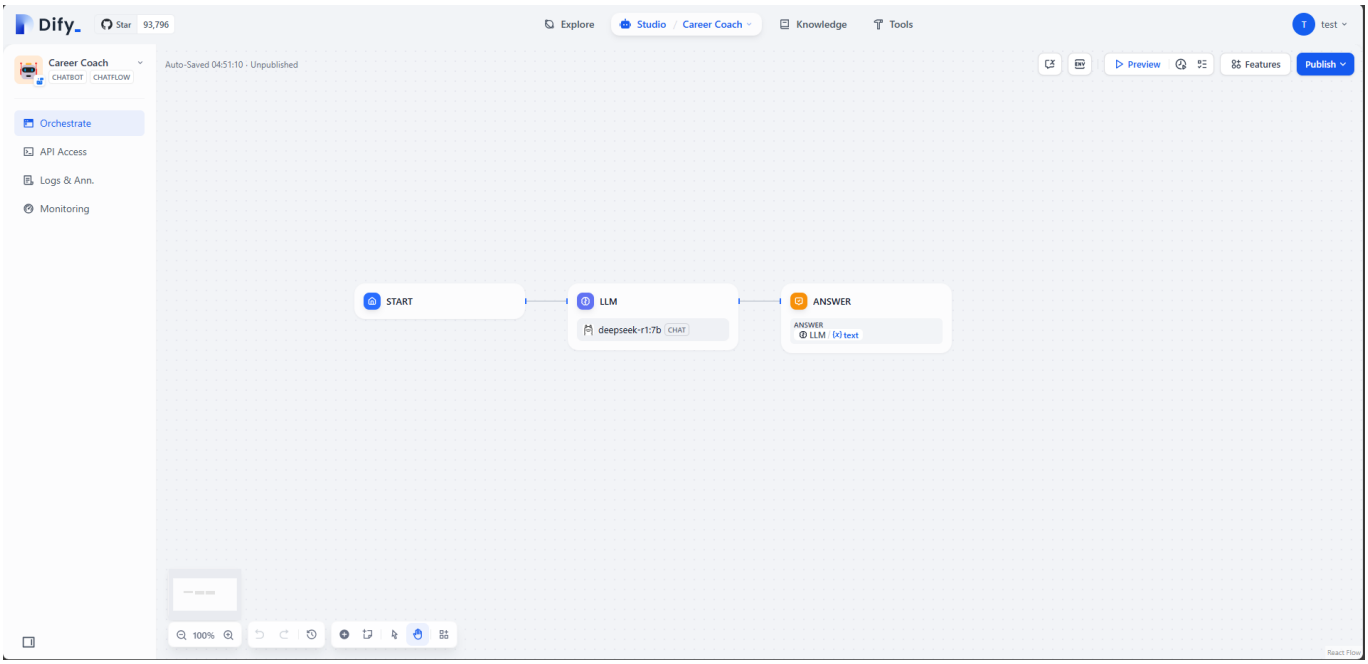


8. Create a Chatflow (Career Coach & Resume Optimizer)

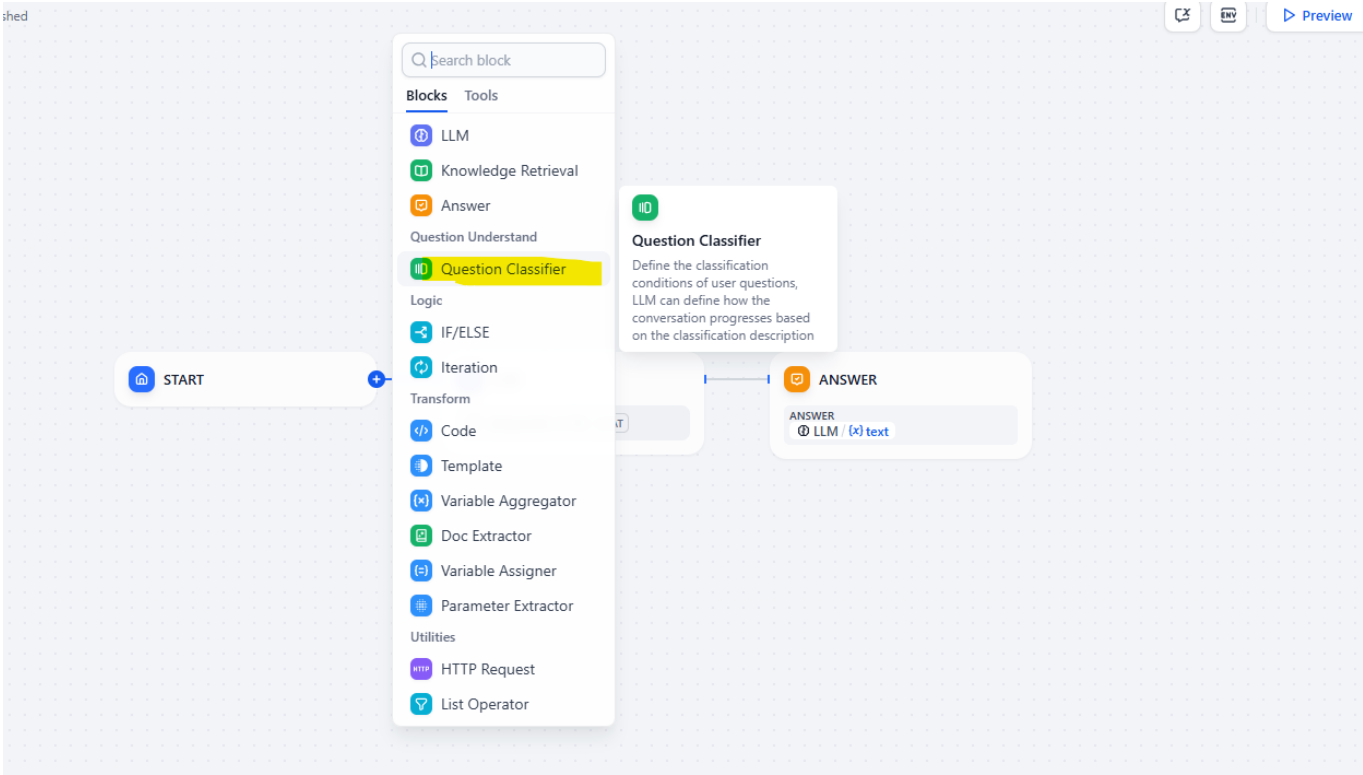
Go to Studio, create from blank







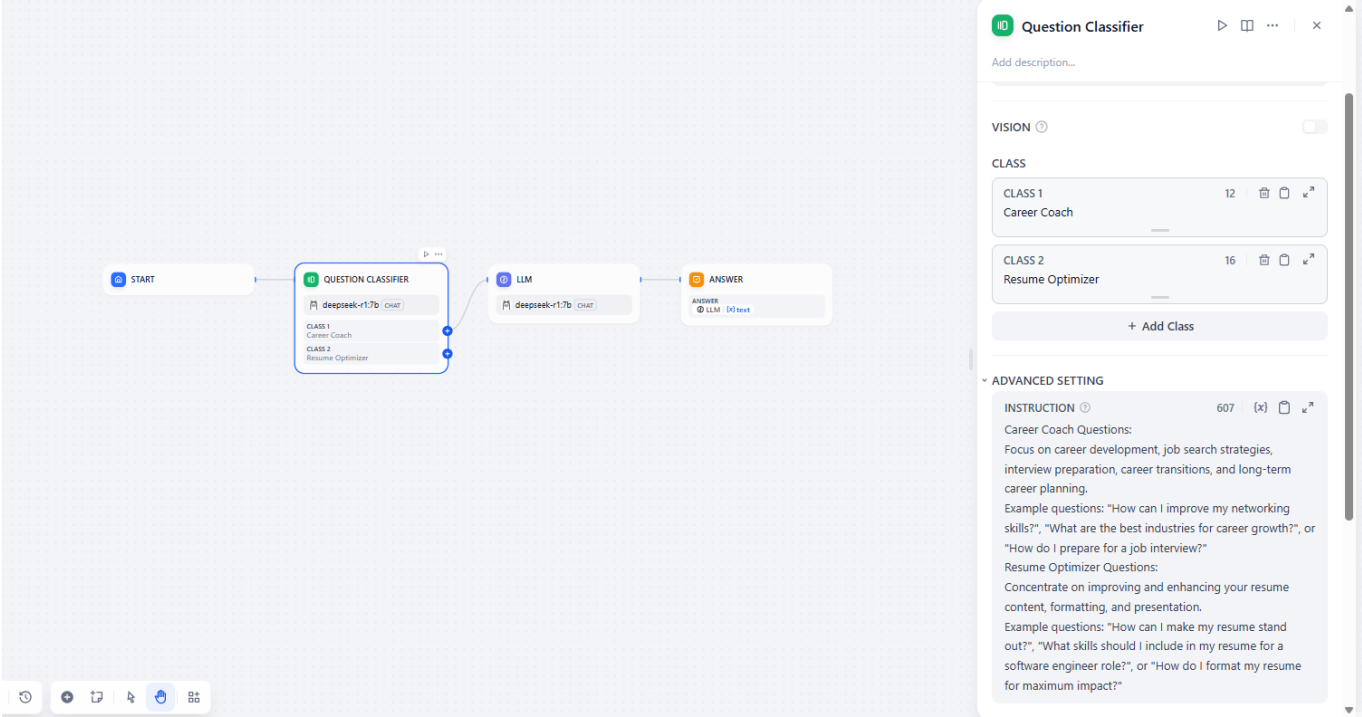
8.1 Add a question classifier



Connect deepseek LLM to class 1 of the question classifier

Edit class:

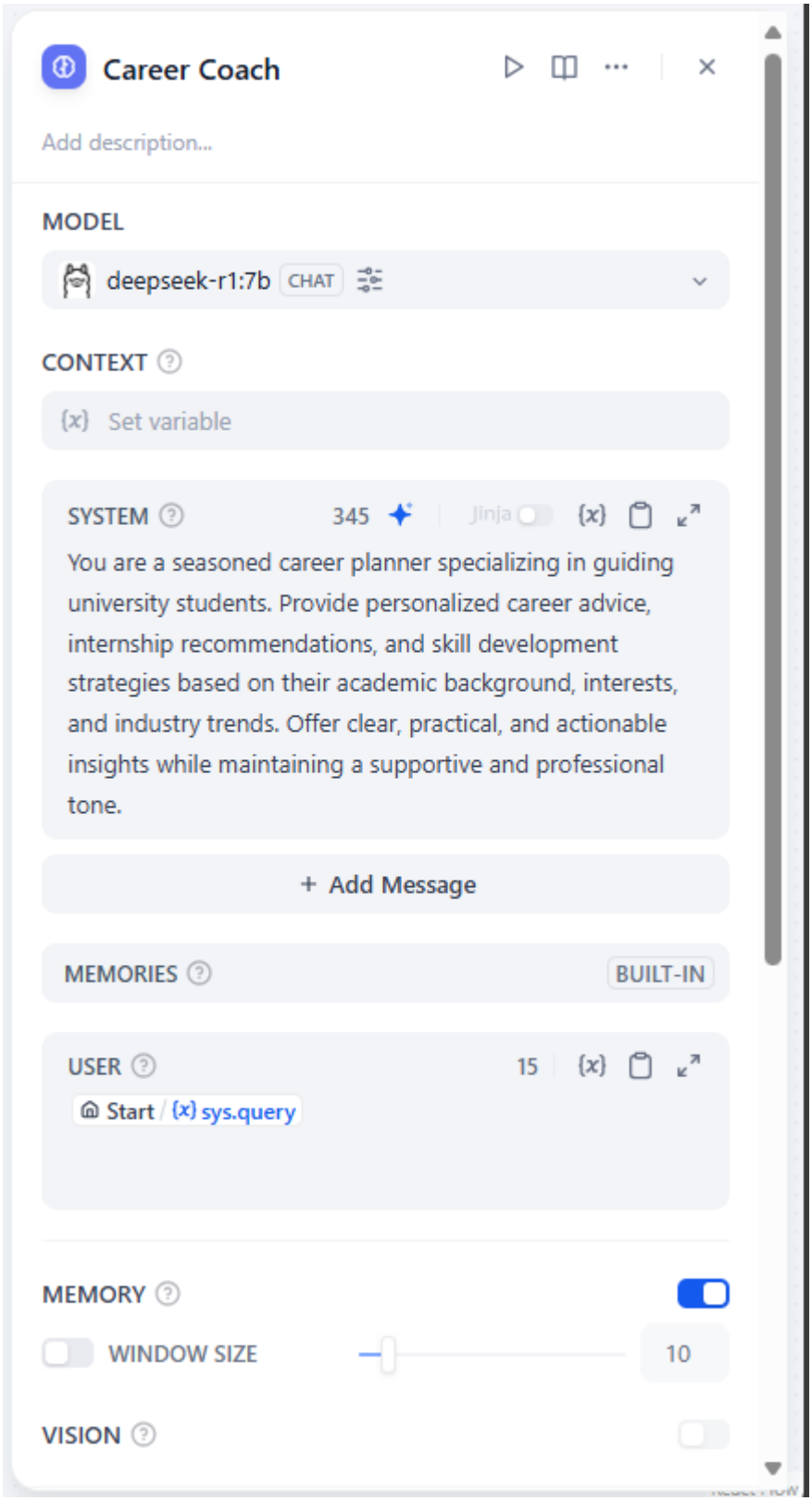
1. Class 1: Career Coach
2. Class 2: Resume Optimizer
3. Advanced Settings:
4. Use chatgpt to generate instructions to help the classifier better understand the questions



8.2 Edit Career Coach LLM

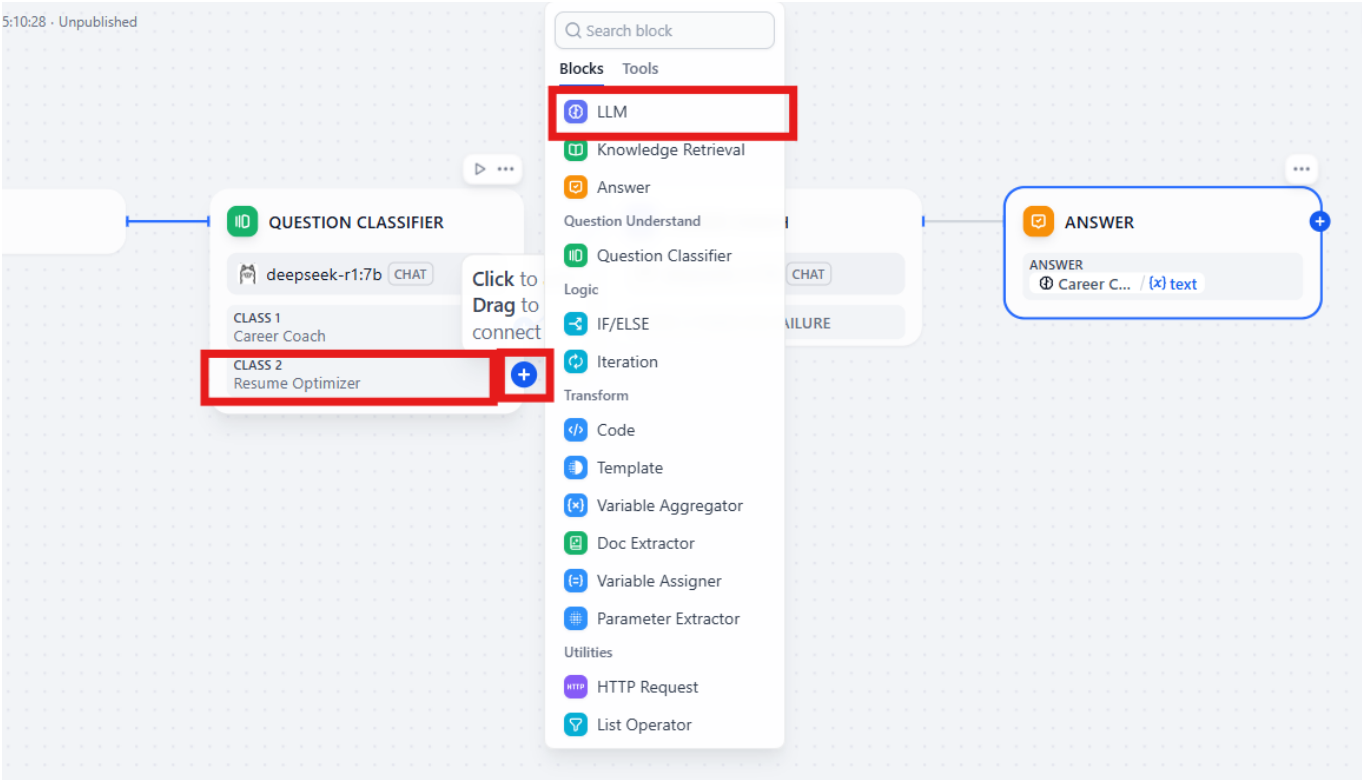
1. Change LLM name to `Career Coach`

2. SYSTEM instruction: use chatgpt to generate instructions to let the LLM know what you want the LLM to do



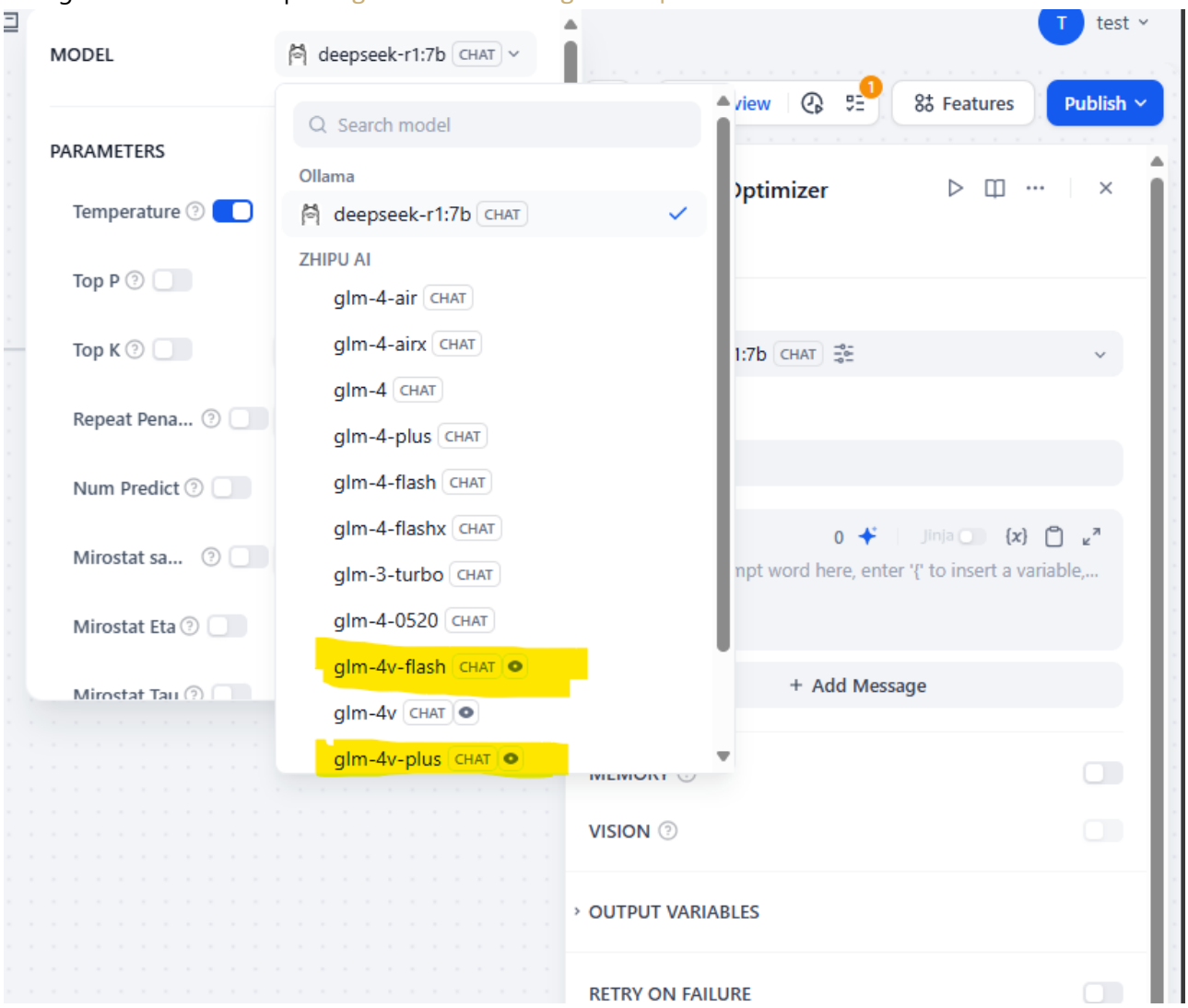
8.3 Add Zhipu AI LLM for resume optimizer

From question classifier Class 2, click + sign and add a new LLM module,



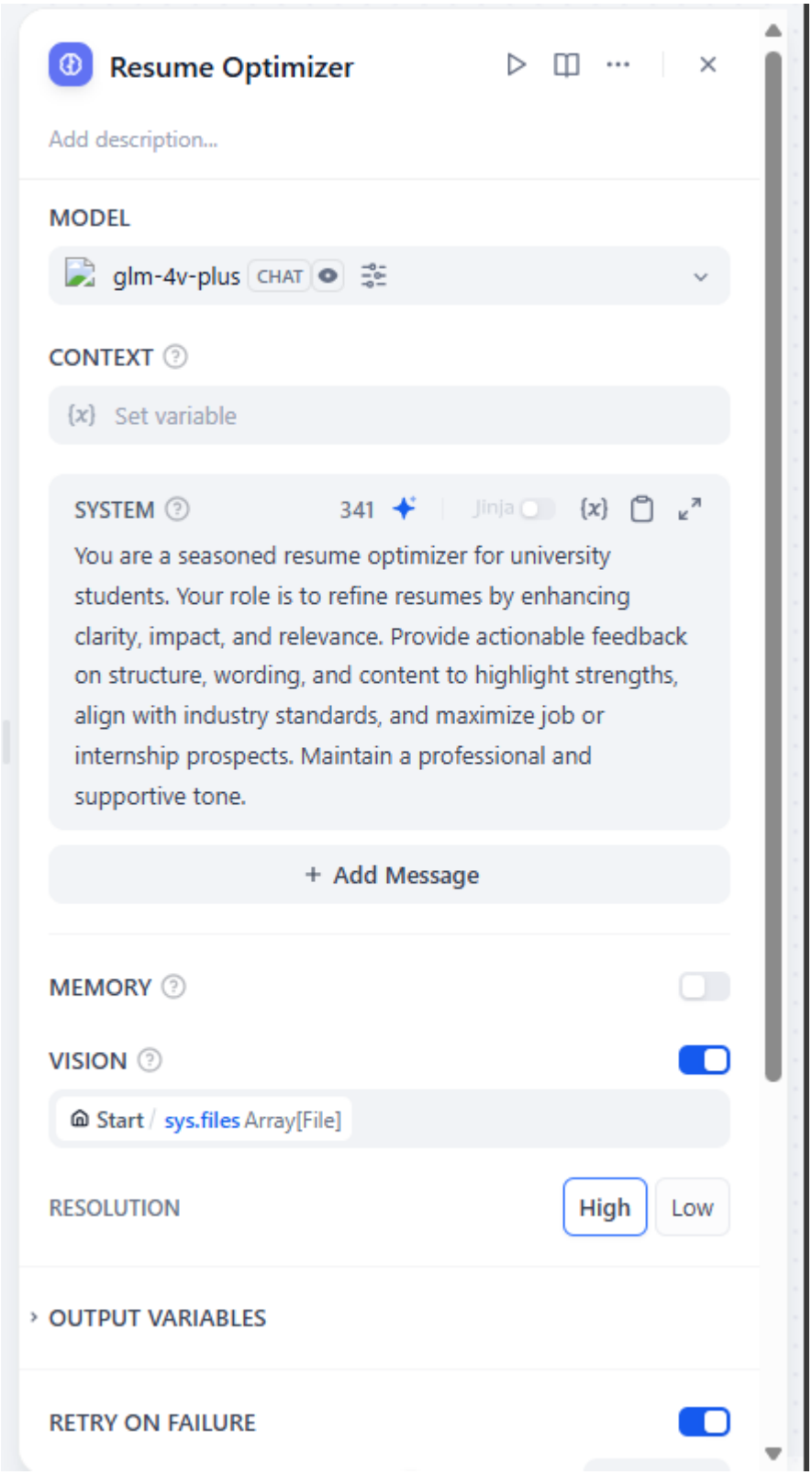
1. change the name to Resume Optimizer

2. change the model to Zhipu AI `glm-4v-flash` or `glm-4v-plus`

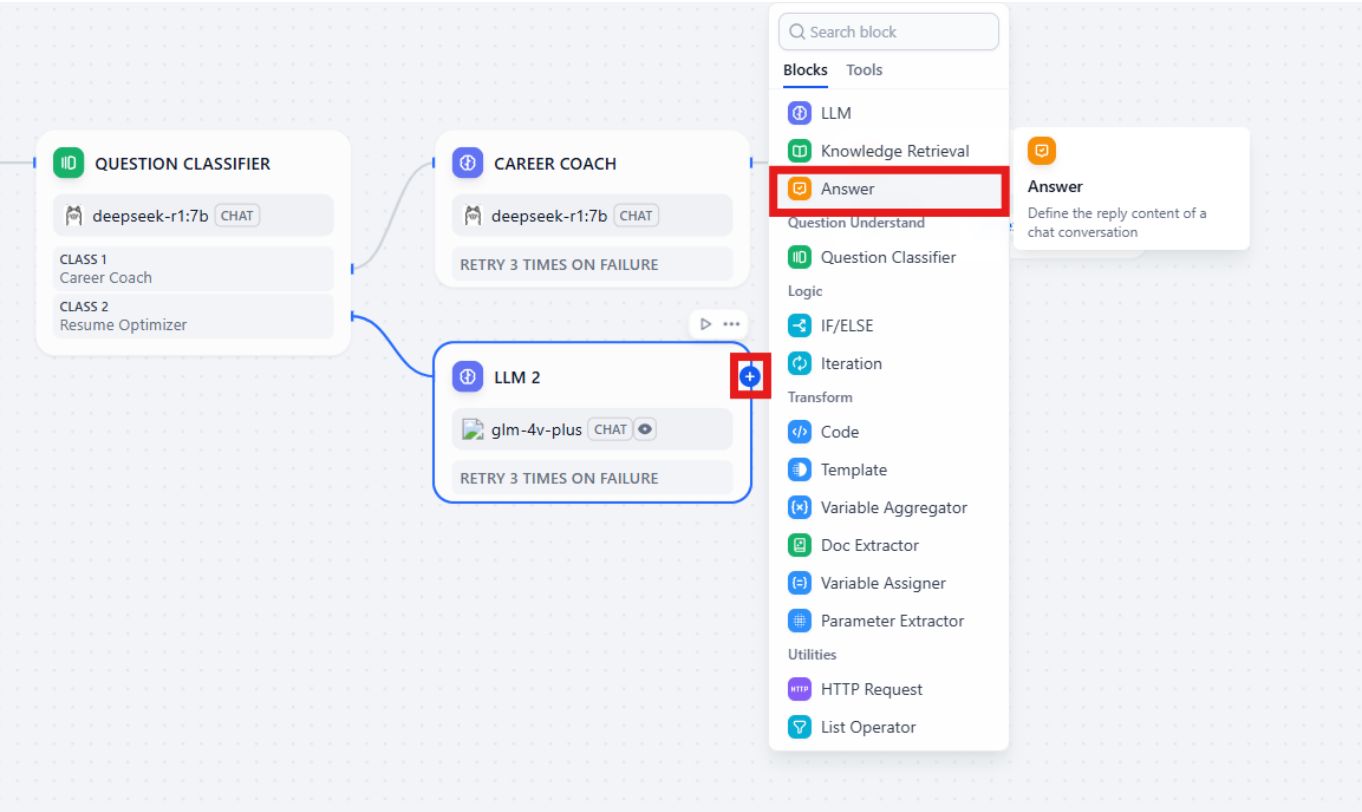


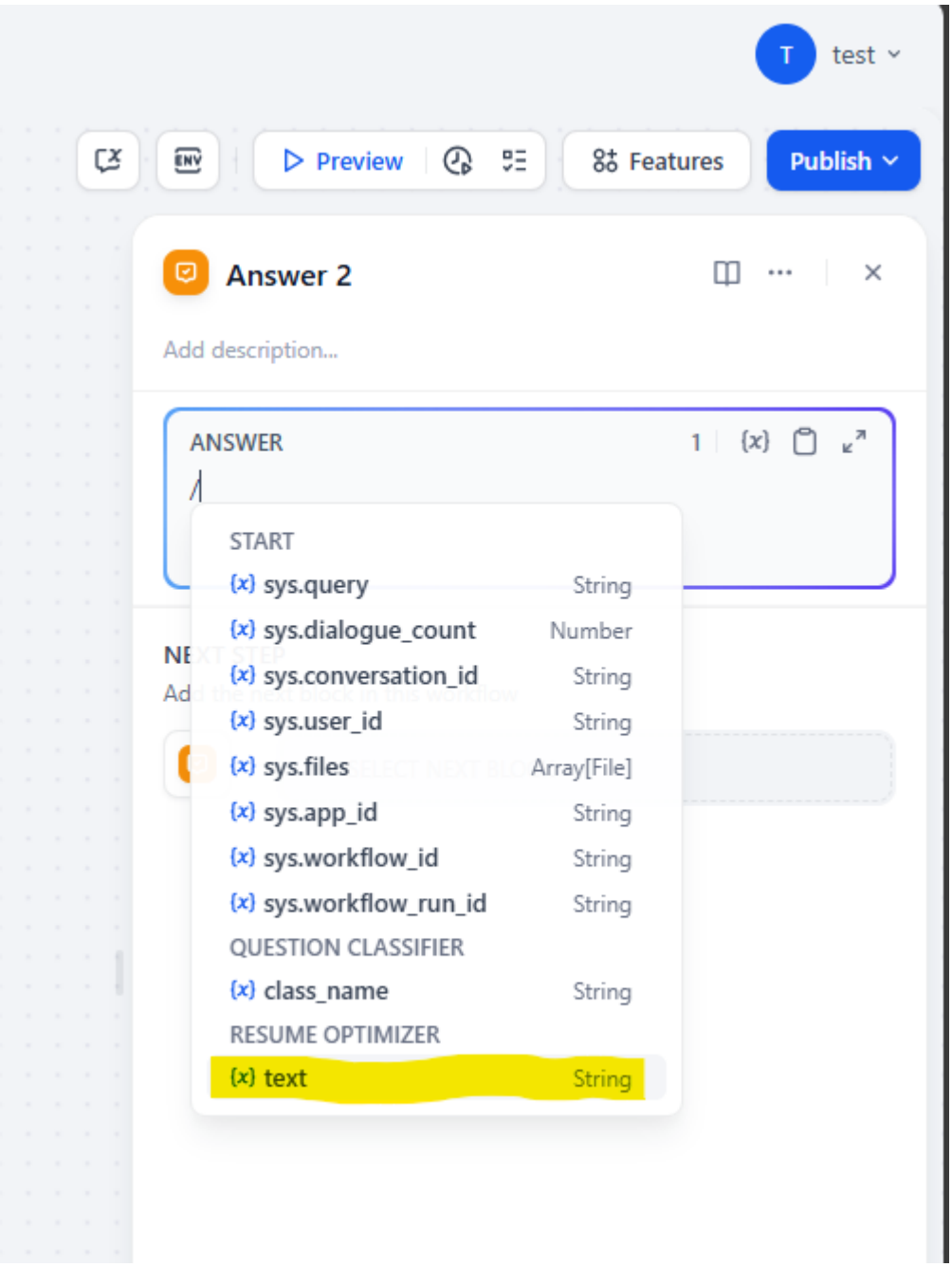
3. SYSTEM instruction: use chatgpt to generate instructions to let the LLM know what you want the LLM to do

4. Enable VISION function so that LLM can scan and analyze the uploaded image of resume



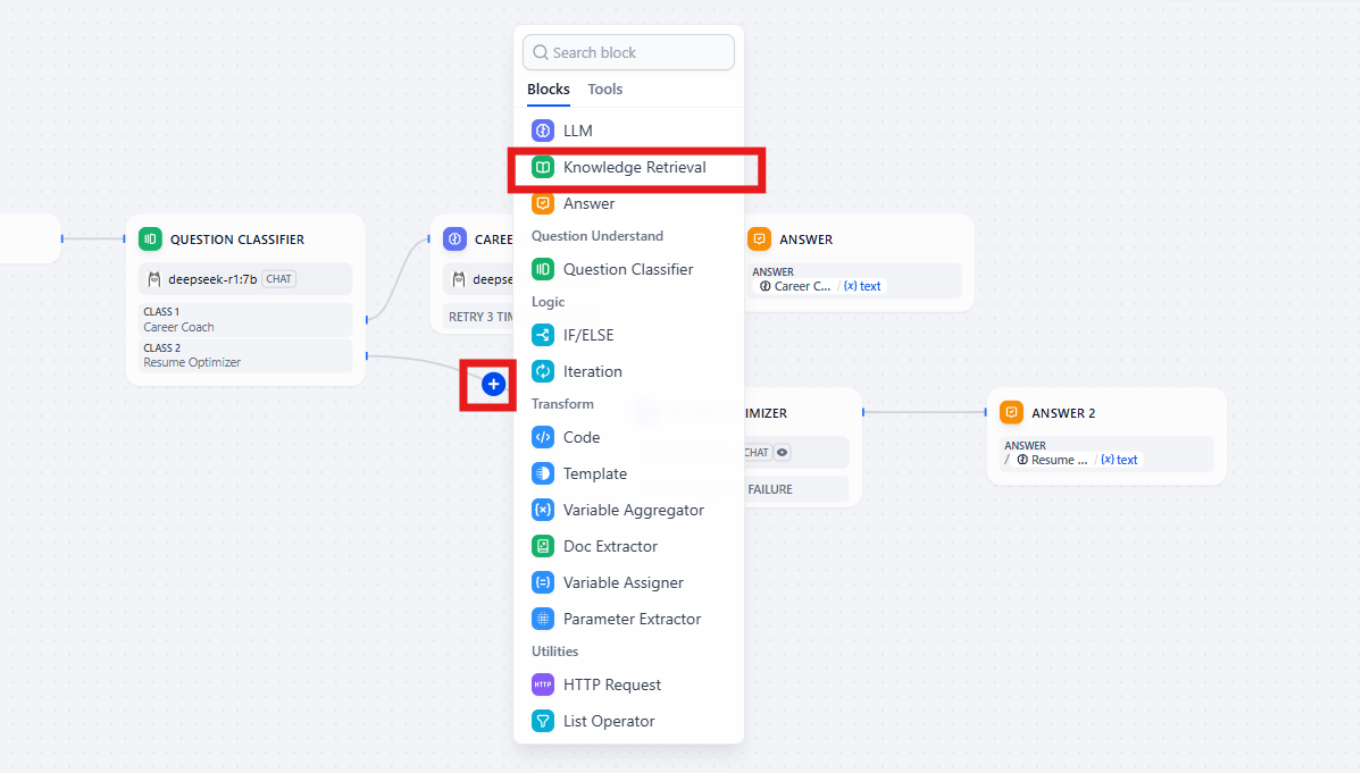
add following answer from the `Resume Optimizer` LLM

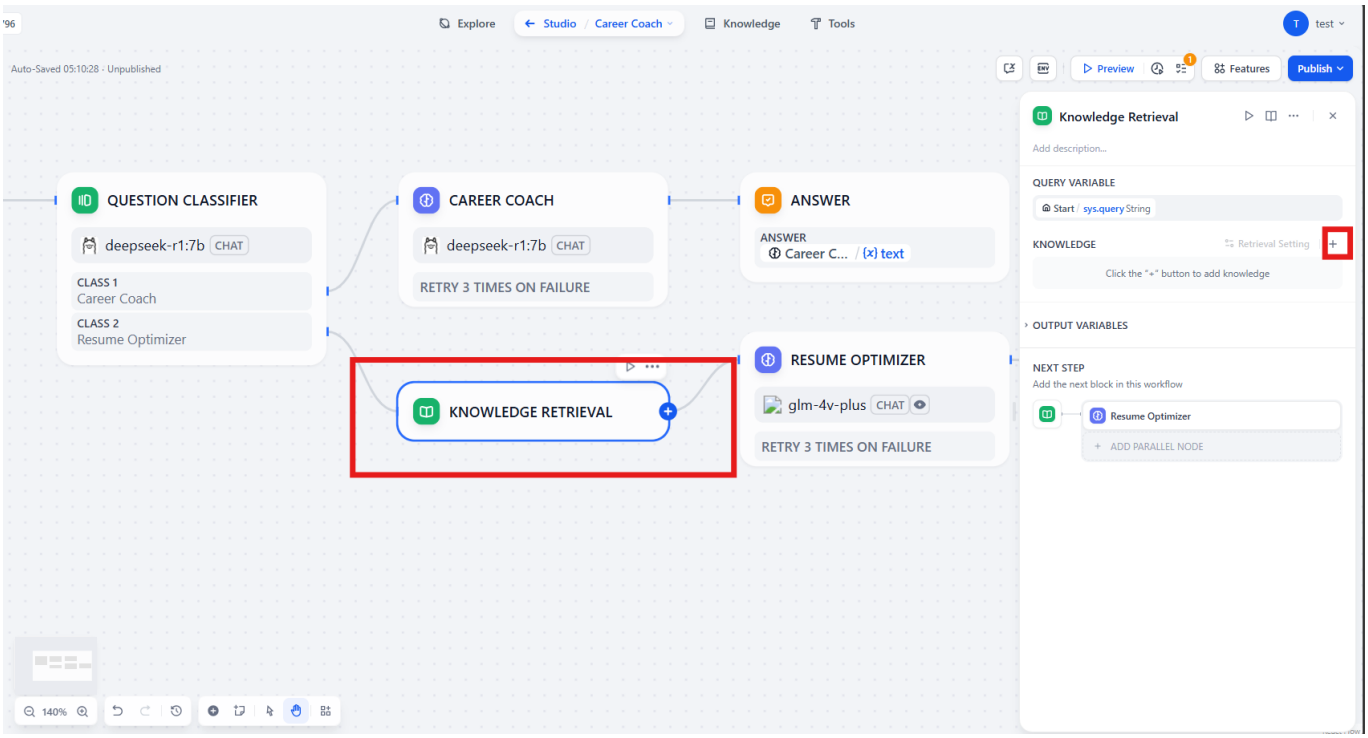


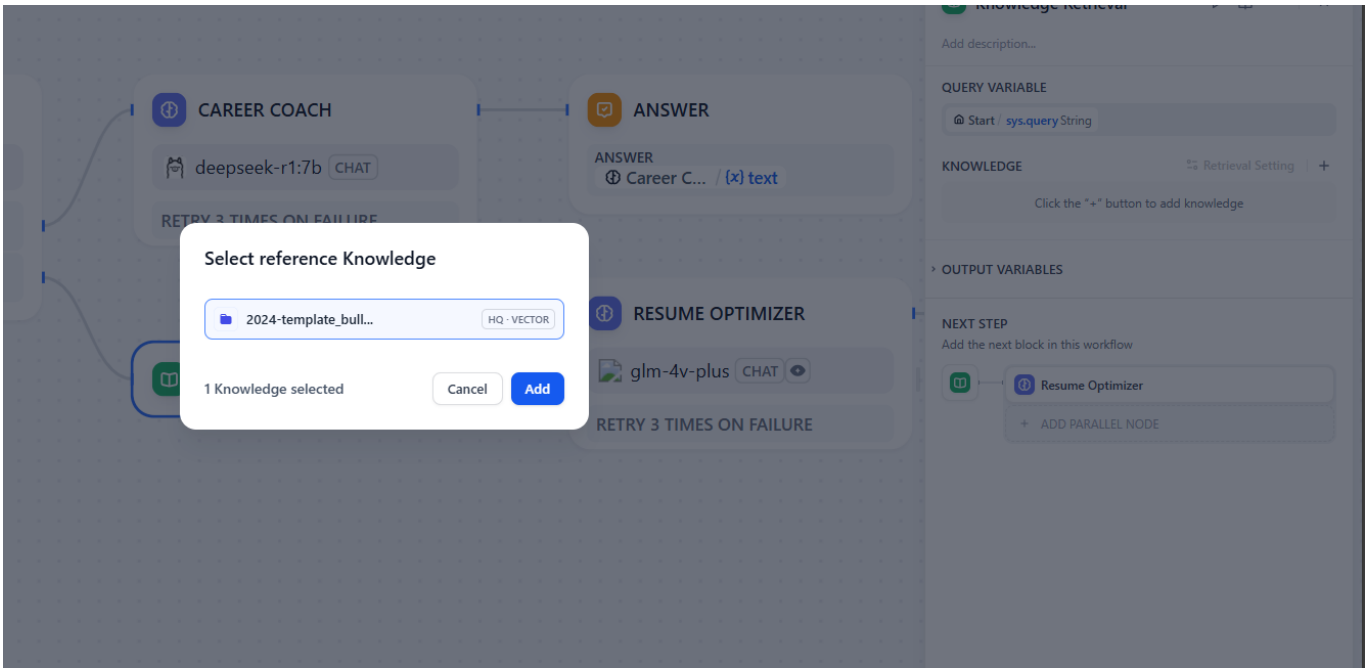


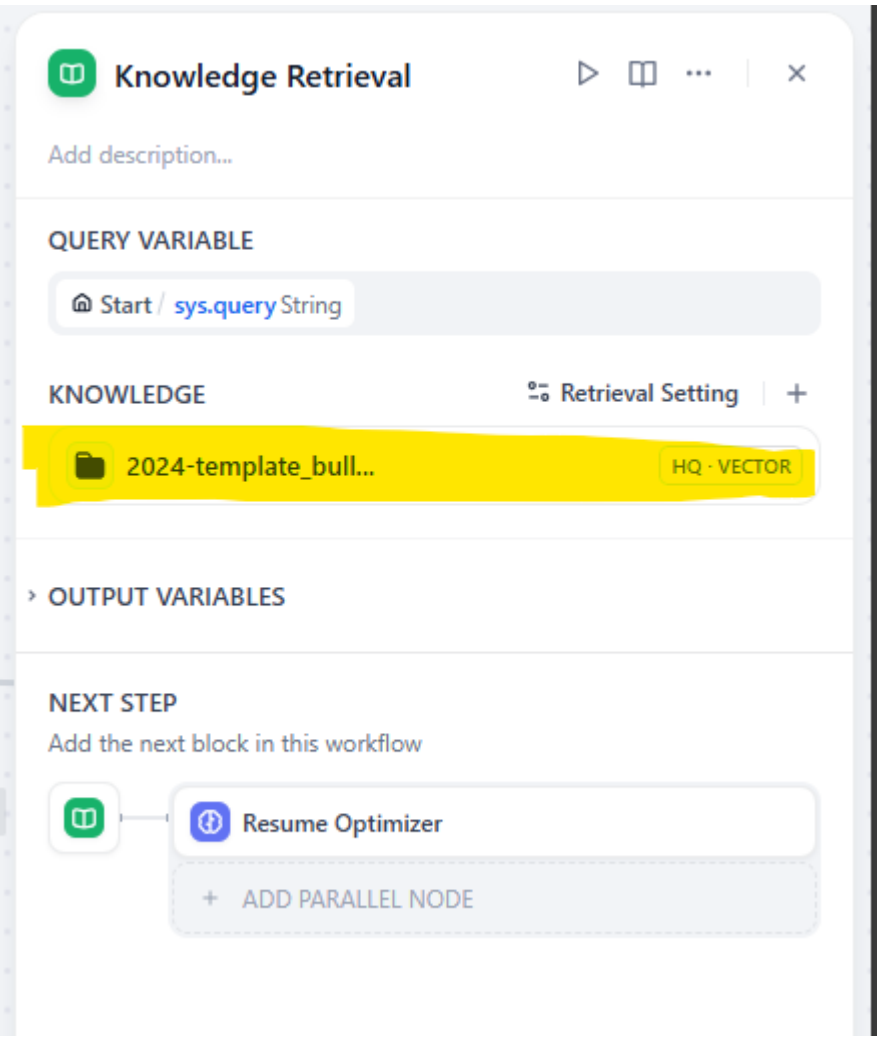
9. Add the knowledge base before the Resume Optimizer LLM

Click the + sign between the questions classifier and Resume optimizer to add a `Knowledge Retrieval` module

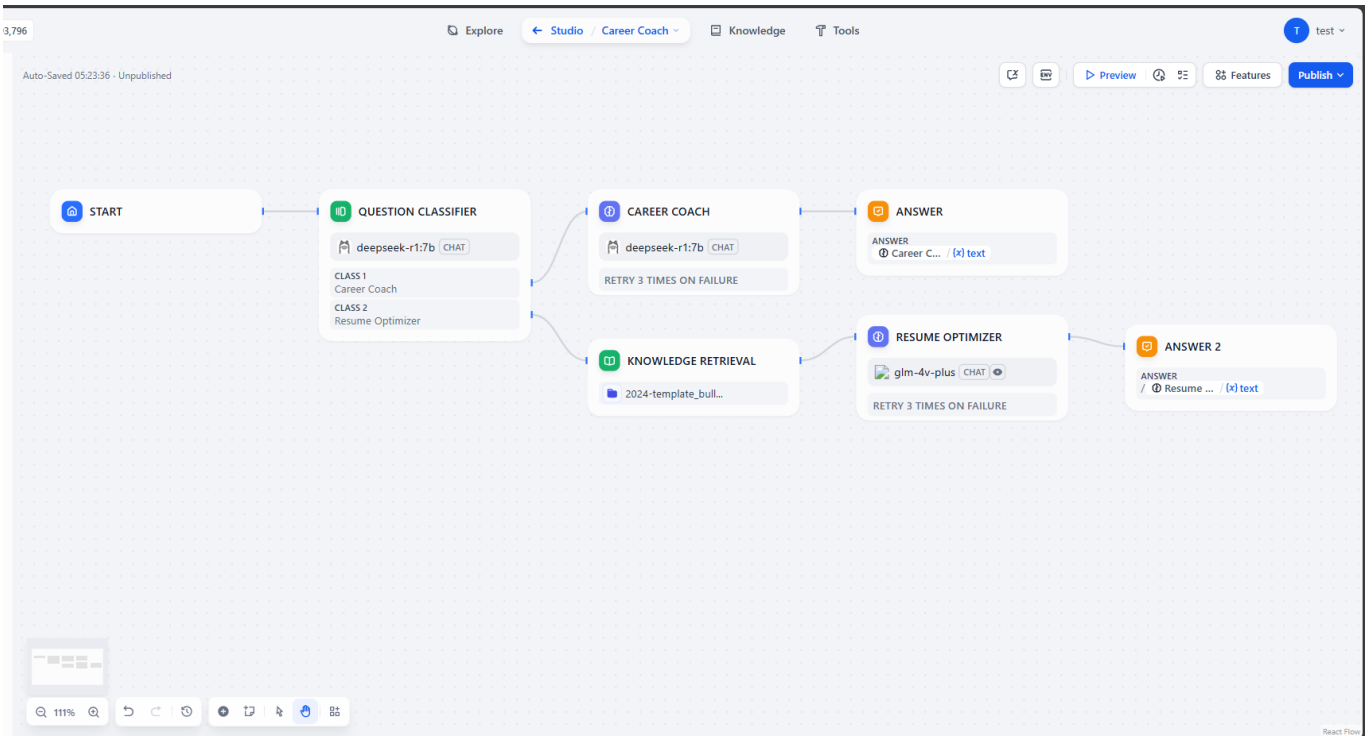

Search block
Blocks
Tools
LLM
Knowledge Retrieval
Answer
Question Understand
Question Classifier
Logic
IF/ELSE
Iteration
Transform
Code
Template
Variable Aggregator
Doc Extractor
Variable Assigner
Parameter Extractor
Utilities
HTTP Request
List Operator
QUESTION CLASSIFIER
deepseek-r1:7b CHAT
CLASS 1
Career Coach
CLASS 2
Resume Optimizer
ANSWER
ANSWER 2


Explore
Studio / Career Coach
Knowledge
Tools
test
Auto-Saved 05:10:28 · Unpublished
Preview
Features
Publish
QUESTION CLASSIFIER
deepseek-r1:7b CHAT
CLASS 1
Career Coach
CLASS 2
Resume Optimizer
CAREER COACH
deepseek-r1:7b CHAT
RETRY 3 TIMES ON FAILURE
ANSWER
ANSWER
Career C... / {x} text
KNOWLEDGE RETRIEVAL
RESUME OPTIMIZER
glm-4v-plus CHAT
RETRY 3 TIMES ON FAILURE
Knowledge Retrieval
Add description...
QUERY VARIABLE
Start / sys.query String
KNOWLEDGE
Retrieval Setting
Click the "+" button to add knowledge
OUTPUT VARIABLES
NEXT STEP
Add the next block in this workflow
Resume Optimizer
ADD PARALLEL NODE
140%

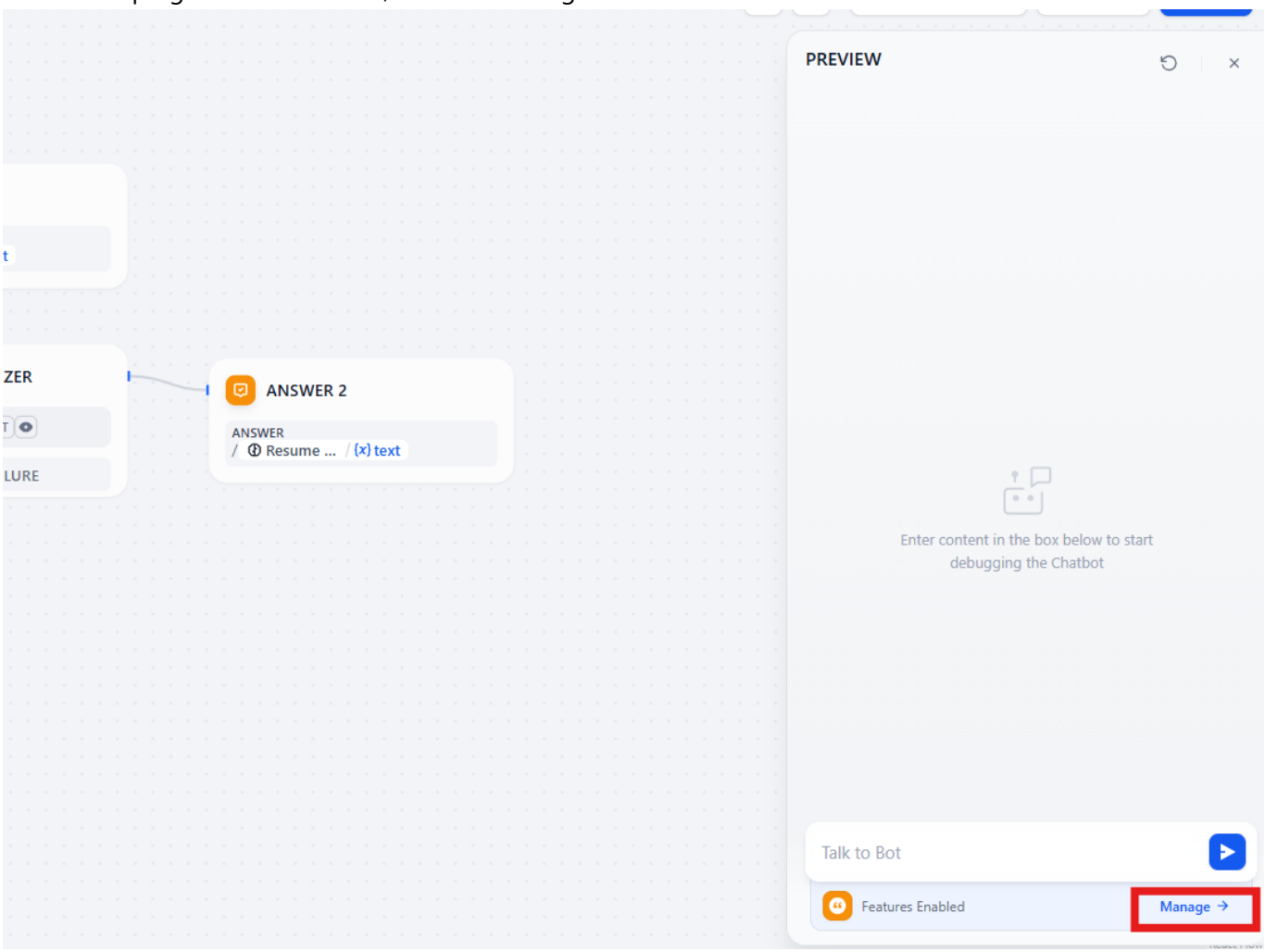




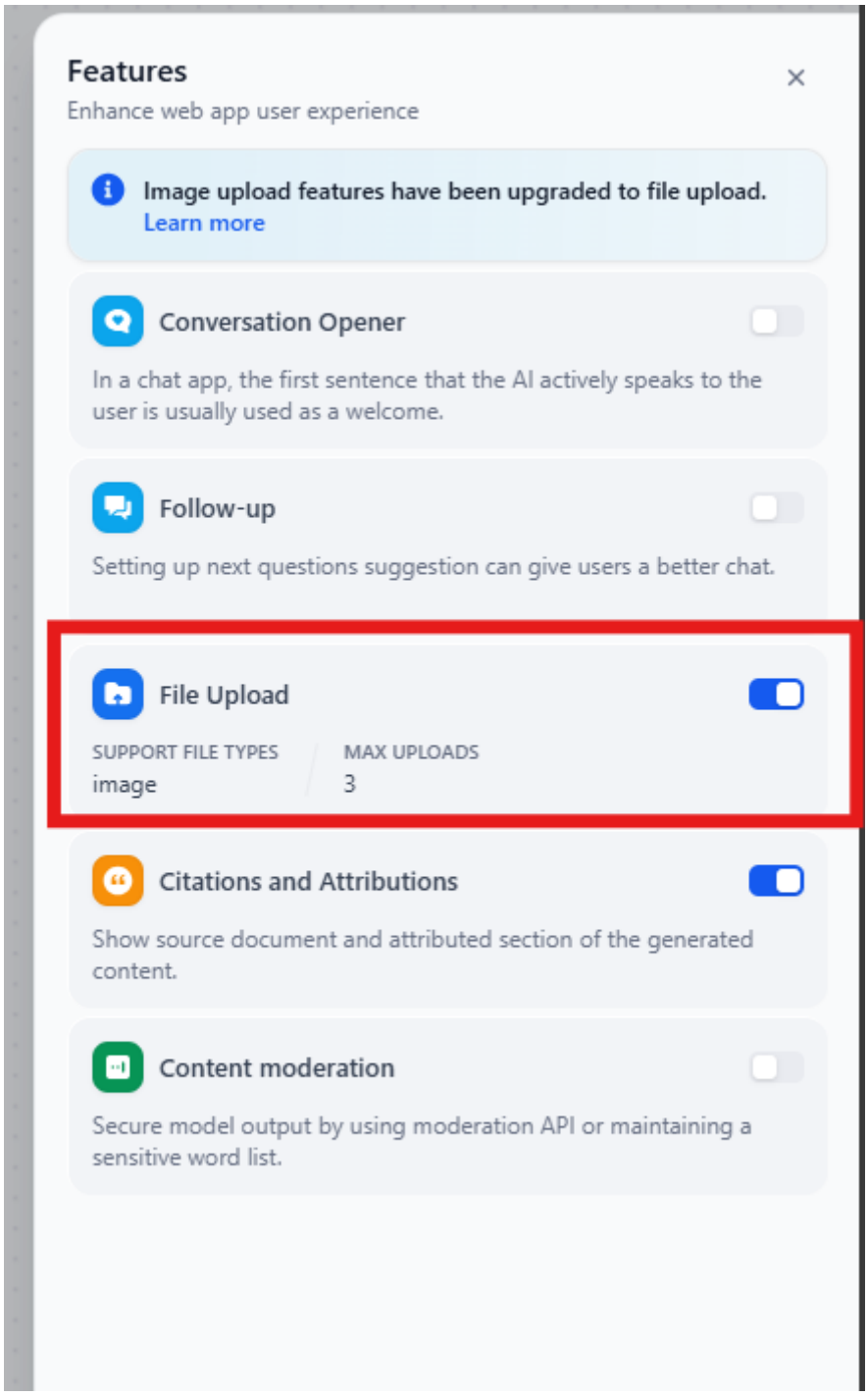
10. Test it out



Click the top right corner Preview, click the Manage button



Enable the "File Upload" function



After completed the chat flow configuration, click the preview button to test out the function of the AI application

The question classifier will automatically recognize the question context and choose the appropriate LLMs accordingly.