



KubeCon



CloudNativeCon

THE LINUX FOUNDATION



AI_dev
Open Source GenAI & ML Summit

China 2024



KubeCon



CloudNativeCon

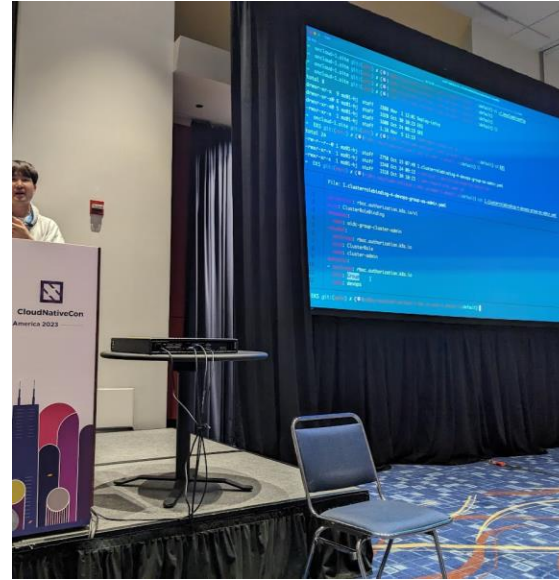


China 2024

Find Your Own **Personal Tutor** for the Study of **Kubernetes**

Cloud Solutions Architect | Cloud Native Engineer
{{ CNCF Ambassador, Kubestronaut }}

Hoon Jo@Megazone



Who am I ?



<https://github.com/SysNet4Admin>



<https://www.linkedin.com/in/hoonjo/>



KubeCon



CloudNativeCon



China 2024

PART I

- k8sGPT + ollama

I have a question for the k8s.



China 2024



reddit



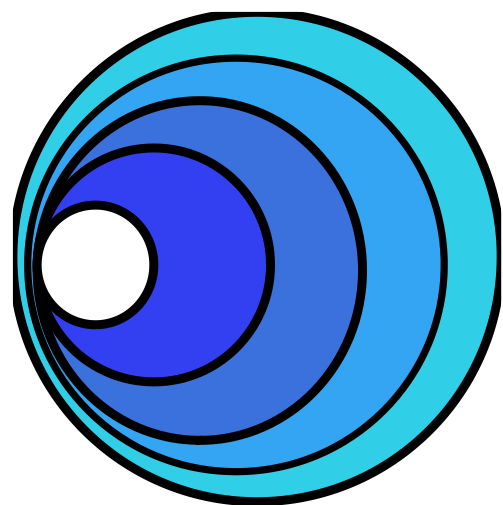
stackoverflow



K8sGPT's simple workflow(default)



China 2024



K8SGPT

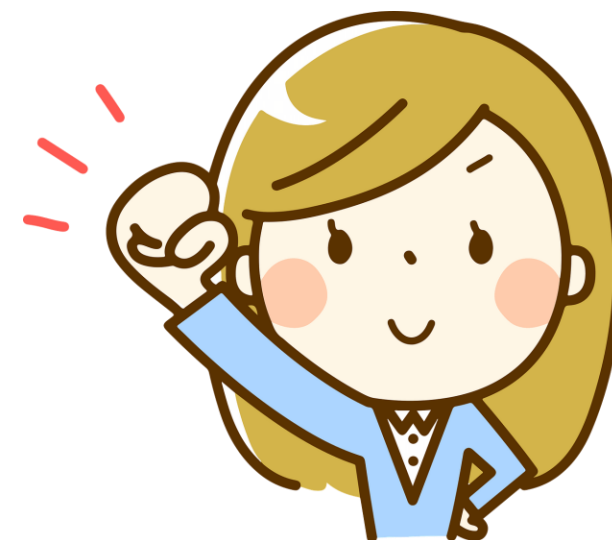
analyze



(request to) explain



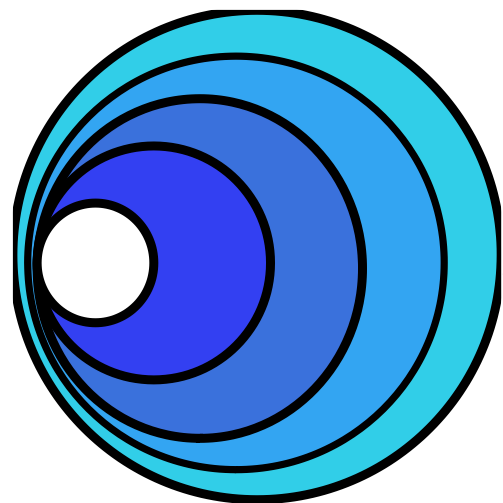
OpenAI



K8sGPT's simple workflow(localai)



China 2024

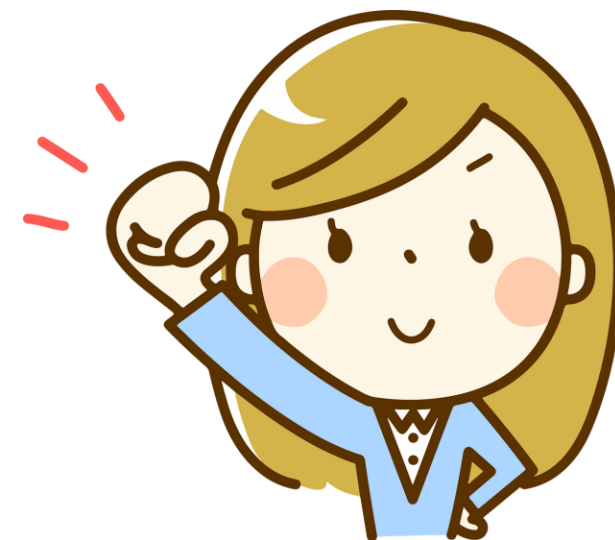
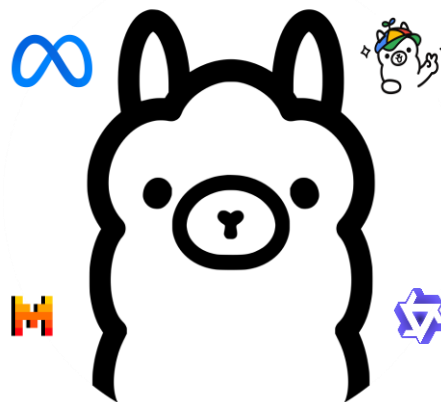


analyze

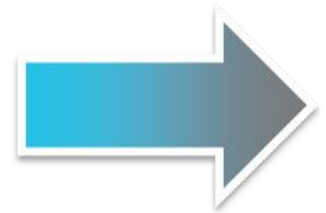
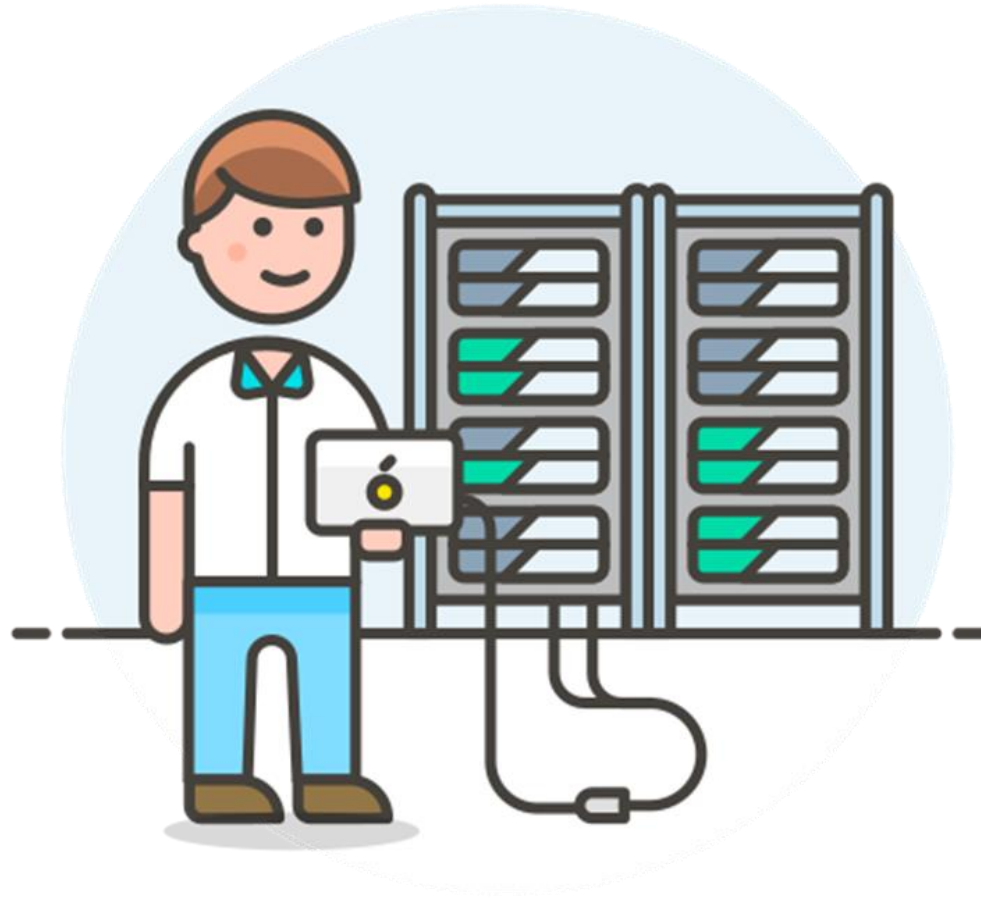


(request to) explain

K8SGPT



Short DEMO:
k8sgpt+ollama



How to work it? - prerequisite_apps



China 2024

```
1  #!/bin/bash # Assuming the script needs to be executed by the shell
2
3  set -eou pipefail # Set error handling options
4
5  IFS=$'\n\t'
6
7  SELF_CMD="$0"
8  exit_oky() { exit 0; }
9  exit_err() { exit 1; }
10
11 prerequisite_check() {
12     if ! hash ollama >/dev/null 2>&1; then echo "ollama is not installed"; exit_err; fi
13     if ! hash k8sgpt >/dev/null 2>&1; then echo "k8sgpt is not installed"; exit_err; fi
14     if ! hash fzf >/dev/null 2>&1; then echo "fzf is not installed"; exit_err; fi
15 }
16
```

How to work it? - run_ollama_n_add_4_k8sgpt



China 2024

```
32  run_ollama_n_auth_add_4_k8sgpt() {
33      get_ollama_list
34      local CHOICE
35      CHOICE="$(FZF_DEFAULT_COMMAND="${SELF_CMD}" fzf --ansi --no-preview || true)"
36
37      # Prepare to run
38      if [[ -z "${CHOICE}" ]]; then
39          echo 2>&1 "Error: You need to choose specific model"
40          exit_err
41      else
42          repeat '-'
43          echo "Run this '${CHOICE}' to analyze for '`kubectl config current-context`'"
44          repeat '='
45          # Run the model as background
46          ollama run ${CHOICE} &
47          # Auth add localai that already chose by above
48          k8sgpt auth add --backend localai --baseurl http://localhost:11434/v1 --model ${CHOICE}
49      fi
50  }
```

How to work it? - analyze_k8sgpt



China 2024

```
52 analyze_by_k8sgpt() {
53     # Analyze k8s cluster by k8sgpt
54     if [[ "$#" -eq 0 ]]; then
55         # Default
56         k8sgpt analyze --backend localai --explain
57     elif [[ "$#" -eq 1 ]]; then
58         echo "Run in '${1}' language"
59         # Specific language
60         k8sgpt analyze --backend localai --explain --language "${1}"
61     elif [[ "$#" -eq 2 ]] && [[ "$2" = "-i" ]]; then
62         # Interactive mode
63         k8sgpt analyze --backend localai --explain --language "${1}" --interactive
64     else
65         echo 2>&1 "Error: You need to choose specific option(s)"
66         exit_err
67     fi
68 }
```

SCRIPT: <https://shorturl.at/GbQ7k> (https://github.com/sysnet4admin/laC/blob/main/AIOps/K8sGPT/run_ollama_n_k8sgpt/run_ollama_n_k8sgpt.sh)



KubeCon



CloudNativeCon



China 2024


PART II

- Find the model from ollama origin

Most popular models from ollama



China 2024

 Models

Most popular

llama3

Meta Llama 3: The most capable openly available LLM to date

8B

70B

↓ 5.2M Pulls

🏷 68 Tags

🕒 Updated 2 months ago

gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

2B

7B

↓ 4M Pulls

🏷 102 Tags

🕒 Updated 3 months ago

mistral

The 7B model released by Mistral AI, updated to version 0.3.

Tools

7B

↓ 3.1M Pulls

🏷 84 Tags

🕒 Updated 8 days ago

qwen

Qwen 1.5 is a series of large language models by Alibaba Cloud spanning from 0.5B to 110B parameters

0.5B

1.8B

4B

32B

72B

110B

↓ 2.3M Pulls

🏷 379 Tags

🕒 Updated 7 weeks ago

llama3.1

Llama 3.1 is a new state-of-the-art model from Meta available in 8B, 70B and 405B parameter sizes.

[Tools](#) [8B](#) [70B](#) [405B](#)

↓ 560.9K Pulls 🏷 35 Tags 🕒 Updated 6 days ago

gemma2

Google Gemma 2 is a high-performing and efficient model by now available in three sizes: 2B, 9B, and 27B.

[2B](#) [9B](#) [27B](#)

↓ 790.3K Pulls 🕒 Updated 6 days ago

qwen2

Qwen2 is a new series of large language models from Alibaba group


[0.5B](#) [1.5B](#) [7B](#) [72B](#)

↓ 560.1K Pulls 🏷 97 Tags 🕒 Updated 7 weeks ago

Leading Models (Popular+newest)



China 2024

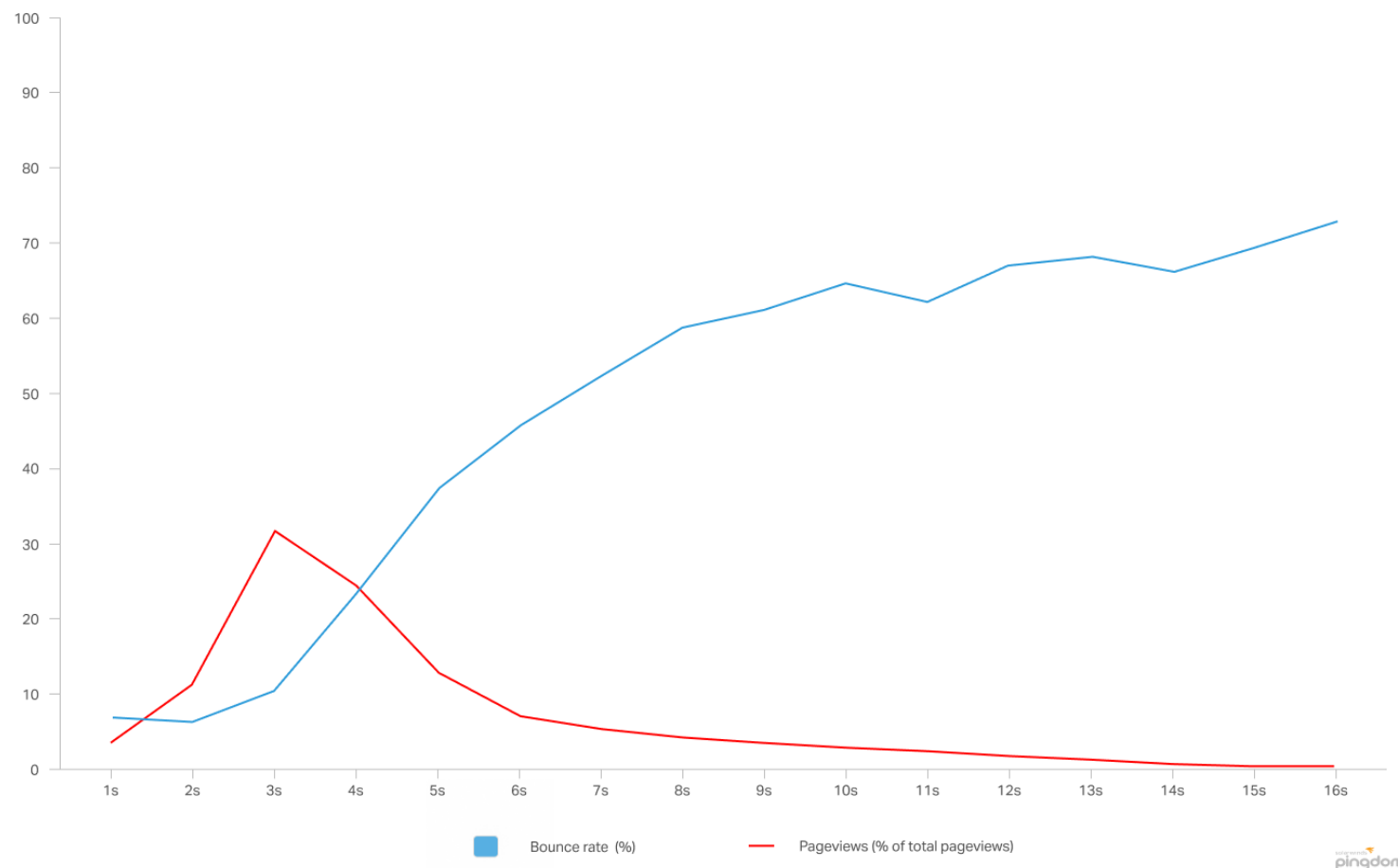
→ SysNet4Admin/IaC git:(main) ( | gke:default) **ollama list**

NAME	ID	SIZE	MODIFIED
gemma2:2b	8ccf136fdd52	1.6 GB	16 seconds ago
mistral:7b	f974a74358d6	4.1 GB	39 minutes ago
mistral-nemo:12b	4b300b8c6a97	7.1 GB	About an hour ago
llama3.1:70b	fb41669f7289	39 GB	6 days ago
llama3.1:8b	a23da2a80395	4.7 GB	7 days ago
llama2-chinese:13b	990f930d55c5	7.4 GB	4 weeks ago
llama2-chinese:7b	cee11d703eee	3.8 GB	4 weeks ago
qwen2:72b	14066dfa503f	41 GB	4 weeks ago
qwen2:7b	e0d4e1163c58	4.4 GB	4 weeks ago
qwen2:1.5b	f6daf2b25194	934 MB	4 weeks ago
qwen2:0.5b	6f48b936a09f	352 MB	4 weeks ago
gemma2:27b	371038893ee3	15 GB	4 weeks ago
gemma2:9b	c19987e1e6e2	5.4 GB	4 weeks ago

Web Page Load Time vs. Bounce Rate



China 2024



Page Load Time (seconds)	Bounce Rate (%)
--------------------------	-----------------

1	7
2	6
3	11
4	24
5	38
6	46
7	53
8	59
9	61
10	65
11	62
12	67
13	69
14	66
15	69
16	73

<https://www.pingdom.com/blog/page-load-time-really-affect-bounce-rate/>

Processing Time (Intel Mac, 8cores / 64GB)

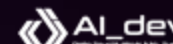


China 2024

	Q1: 请告诉我更多详细信息	Q2: What is the endpoints?	Q3: in Chinese	Q4: in Korean
Qwen2:0.5b(352MB)	3s	1s	1s	3s
Qwen2:1.5b(934MB)	15s	4s	6s	5s
Qwen2:7b(4.4GB)	1m33s	1m24s	31s	53s
Qwen2:72b(41GB)	15m15s	7m35s	4m47s	5m8s
Llama3.1:8b(4.7GB)	1m13s	49s	55s	45s
Llama3.1:70b(39GB)	10m41s	6m18s	2m45s	3m7s
llama2-chinese: 13b(7.4GB)	59s	23s	21s	20s
gemma2:2b(1.6GB)	56s	44s	22s	27s
gemma2:9b(5.4GB)	2m19s	59s	55s	49s
gemma2:27b(15GB)	13m2s	10m12s	4m14s	20m

RAW Data : <https://url.kr/l5uk81>

Processing Time (Apple Silicon M2, 8cores / 24GB)

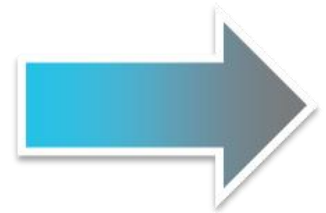


China 2024

	Q1: 请告诉我更多详细信息	Q2: What is the endpoints?	Q3: in Chinese	Q4: in Korean
Qwen2:0.5b(352MB)				
Qwen2:1.5b(934MB)	3s	3s	3s	4s
Qwen2:7b(4.4GB)	30s	21s	11s	11s
Qwen2:72b(41GB)				
Llama3.1:8b(4.7GB)	24s	12s	11s	17s
Llama3.1:70b(39GB)				
llama2-chinese: 13b(7.4GB)				
gemma2:2b(1.6GB)	16s	15s	8s	7s
gemma2:9b(5.4GB)	36s	16s	18s	19s
gemma2:27b(15GB)	2m41s	1m20s	1m28s	2m32s

RAW Data : <https://url.kr/l5uk81>

Short DEMO:
gemma2:2b





KubeCon



CloudNativeCon

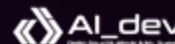


China 2024

PART III

- Find the model from huggingface

Where is llama3 for chinese?



China 2024



Models

chinese

llama2-chinese

Llama 2 based model fine tuned to improve Chinese dialogue ability.

7B

13B

↓ 115.8K Pulls 🏷 35 Tags ⌚ Updated 9 months ago

Qwen2:0.5b(352MB)

Qwen2:1.5b(934MB)

Qwen2:7b(4.4GB)

Qwen2:72b(41GB)

Llama3.1:8b(4.7GB)

Llama3.1:70b(39GB)

llama2-chinese:13b(7.4GB)

gemma2:2b(1.6GB)

gemma2:9b(5.4GB)

gemma2:27b(15GB)

Found llama3-chinese in HF



China 2024



Hugging Face

Search models, datasets, u:

Models

Datasets

Spaces

Posts

Docs

Pricing



FlagAlpha/Llama3-Chinese-8B-Instruct

like 60



Text Generation



Transformers



Safetensors

llama

llama3

chinese

conversational

custom_code



text-generation-inference



Inference Endpoints



License: apache-2.0



Model card



Files



Community

5



Train



Deploy



Use this model

Edit model card

Llama3-Chinese-8B-Instruct

Llama3-Chinese-8B-Instruct基于Llama3-8B中文微调对话模型，由Llama中文社区和AtomEcho（原子回声）联合研发，我们会持续提供更新的模型参数，模型训练过程见 <https://llama.family>。

模型的部署、训练、微调等方法详见Llama中文社区GitHub仓库：

<https://github.com/LlamaFamily/Llama-Chinese>

<https://huggingface.co/FlagAlpha/Llama3-Chinese-8B-Instruct>

Downloads last month
8,448



Safetensors

Model size

8.03B params

Tensor type

FP16



Inference API

Text Generation

NEED TO CONVERT



China 2024


- download model from huggingface
- git clone llama.cpp
- install requirements by pip
- **convert from safetensor to gguf by llama.cpp**
- make llama-quantize command(+) from llama.cpp
- **convert quantized-model to Q5_K_M**
- **load gguf model on ollama platform**
- Run “run_ollama_n_k8sgpt.sh”

Show output: <https://url.kr/msvd2g>

Added llama3-chinese model

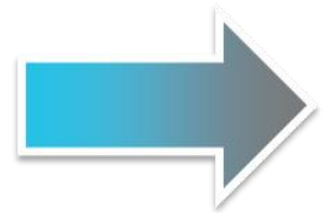
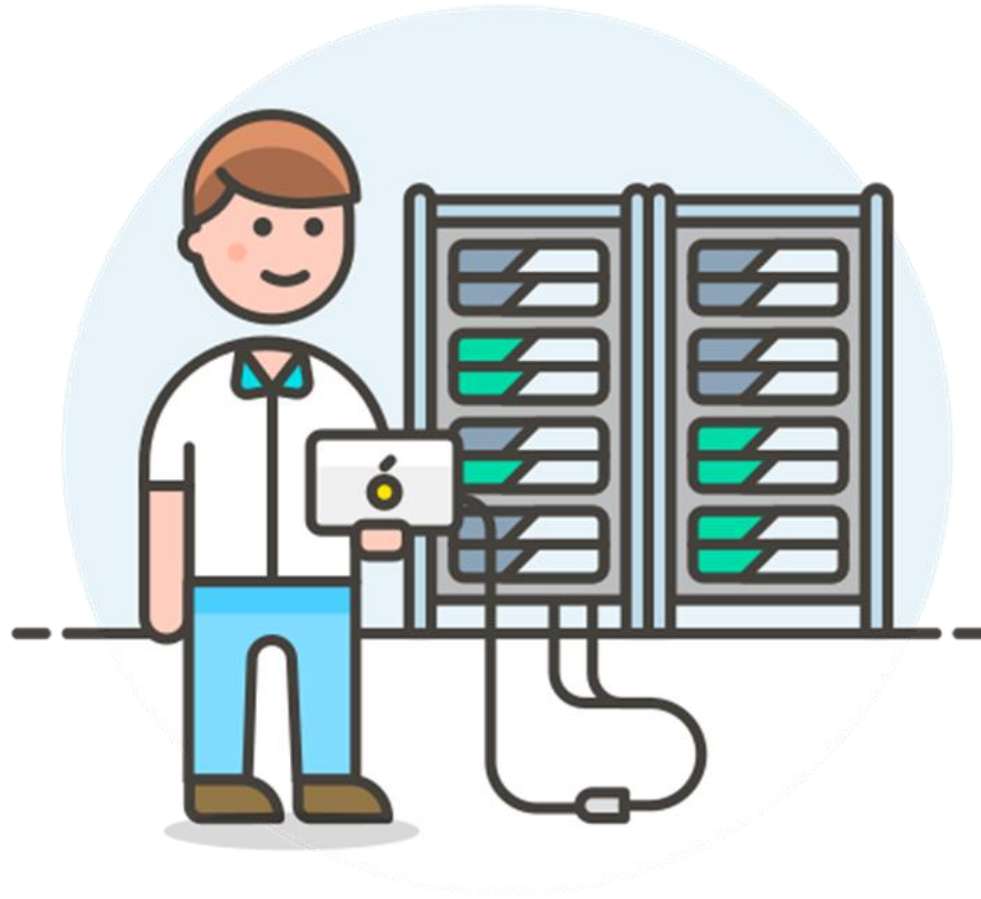


China 2024

→ SysNet4Admin/IaC git:(main) ( | gke:default) **ollama list**

NAME	ID	SIZE	MODIFIED
llama3-chinese:8b	885cf086a330	5.7 GB	6 seconds ago # add llama3-chinese model
llama3:8b	365c0bd3c000	4.7 GB	19 hours ago
gemma2:2b	8ccf136fdd52	1.6 GB	6 days ago
mistral:7b	f974a74358d6	4.1 GB	7 days ago
mistral-nemo:12b	4b300b8c6a97	7.1 GB	7 days ago
llama3.1:70b	fb41669f7289	39 GB	2 weeks ago
llama3.1:8b	a23da2a80395	4.7 GB	2 weeks ago
llama2-chinese:13b	990f930d55c5	7.4 GB	5 weeks ago
llama2-chinese:7b	cee11d703eee	3.8 GB	5 weeks ago
qwen2:72b	14066dfa503f	41 GB	5 weeks ago
qwen2:7b	e0d4e1163c58	4.4 GB	5 weeks ago
qwen2:1.5b	f6daf2b25194	934 MB	5 weeks ago
<snipped>			

Short DEMO:
ollama3-chinese





KubeCon



CloudNativeCon



China 2024

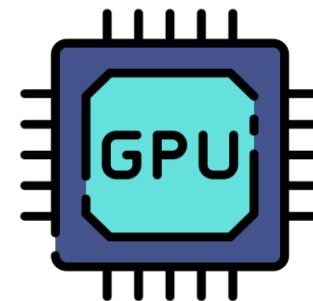
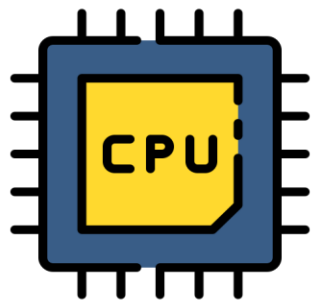
PART IV

- TL; Summary

Find the your proper model but...



China 2024



Qwen2 :1.5b



Qwen2 :7b



Gemma 2 :2b



Gemma 2 :2b

Any Questions?

KubeCon China 2024's docs

[KubeCon China 2024] #1 run_ollama_n_k8sgpt.sh per model

- ShortURL: <https://url.kr/l5uk81>



[KubeCon China 2024] #2 safetensor_2_gguf for ollama

- ShortURL: <https://url.kr/msvd2g>



<https://github.com/SysNet4Admin>



<https://www.linkedin.com/in/hoonjo/>