KNEO 300 User Manual V1.2

(For Admin User)

Mar, 2024

Revision History:

Doc Version	Description	Firmware Version	Date
0.9	Initial version	-	2023/10/23
1.0	Add set-up guide	-	2024/01/08
1.1	Add custom settings	-	2024/01/17
1.2	Add admin user guide	V0.14.1	2024/03/07

Notice:

1. Kneron (Taiwan) Co., Ltd may make changes to any information in this document at any time without any prior notice. The information herein is subject to change without notice.

2. THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY OR CONDITION OF ANY KIND, EITHER EXPRESS, IMPLIED OR STATUTORY, INCLUDING, WITHOUT LIMITATION, ANY WARRANTY OR CONDITION WITH RESPECT TO MERCHANTABILITY, FITNESS FOR ANY PARTICULAR PURPOSE, OR NON-INFRINGEMENT .KNERON DOES NOT ASSUME ANY RESPONSIBILITY AND LIABILITY FOR ITS USE NOR FOR ANY INFRINGEMENT OF PATENTS OR OTHER RIGHTS OF THE THIRD PARTIES WHICH MAY RESULT FROM ITS USE.

3. Information in this document is provided in connection with Kneron products.

4. All referenced brands, product names, service names and trademarks in this document are the property by their respective owners.

Table of contents

1 Introduction

KNEO 300 is an NPU-based edge AI server, specially used to implement LLM applications, supporting 30TOPS AI computing power, equipped with an all-metal casing , fan cooling and rich peripheral interfaces. Compared with traditional GPU LLM inference, it has the advantages of low cost, low power consumption, and high efficiency, and can be widely used in fields such as enterprise AIGC.

KNEO 300 has built-in Kneron self-developed edge chatbot software, which is mainly used to answer questions and provide information. Its function is similar to an advanced offline virtual assistant . Here are some of the key features and uses of this chat product:

1. Q&A: Ability to answer a variety of questions covering a wide range of topics such as science, history, culture, technology, etc.
2. Language Understanding: Strong understanding of natural language and the ability to understand and respond to complex and abstract queries.
3. Text generation: In addition to answering questions, you can also write articles, create stories, generate creative content, etc.
4. User interaction: Able to have smooth conversations with users and provide helpful answers and suggestions based on database and other information. Wide range of applications: education, customer support, HR, company training, IT support, etc.
5. Privacy and security: This system adopts offline mode, which greatly protects the security of user information, data and privacy.

2. Device operation tutorial

2.1 Product Overview

- KNEO 300 series AI box appearance

- KNE300 series AI box peripheral interfaces (from left to right)

1. UP : RS232
2. Down : RS485
3. UP : Ethernet (1000mbps)
4. Down : USB3.0x2
5. UP : Ethernet (1000mbps)
6. Down : USB2.0x2
7. HDMI 2.0
8. TF Card
9. DC 12V
10. Power button

- Product parameters

CPU	8-core A53, 2.0GHz
NPU	30 TOPS (INT8)
DRAM	16GB LPDDR4
eMMC	64GB
Power	DC12V, AC100-240V, 50-60HZ
Operating System	Ubuntu
Size	210mm*130mm*45mm
Operating Environment	Operating Temperature: -20℃~60℃; Storage Temperature: -20℃~70℃; Operating Humidity: 10%~90%RH;
Ethernet	2*Gigabit Ethernet
USB	2*USB3.0
	2*USB2.0
Connecting Ports	1*RS232
	1*RS485

2.2 Accessories list

After receiving the device, check whether the accessories are complete:

- KNEO 300 AI box
- One 12V-5A power adapter
- One HDMI cable
- One Ethernet cable

- A pack of expansion screws

In addition, during use, you also need the following conditions:

- display

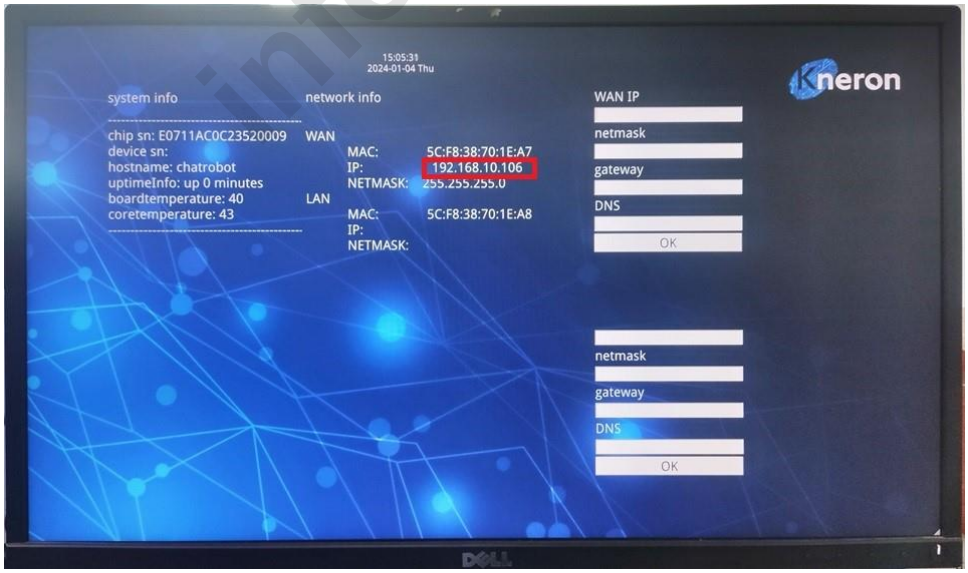
  Monitor or TV with HDMI port.

- network

  100M/1000M wired network.

2.3 Power on

- Connect the power cable to the 12V-5A power adapter.
- Connect the device and monitor with an HDMI cable.
- Plug the network cable into UP: Ethernet (1) and connect to the network.
- The device will automatically turn on after being powered on. The monitor will display the following screen, and the IP address of the device will be displayed in the red box.

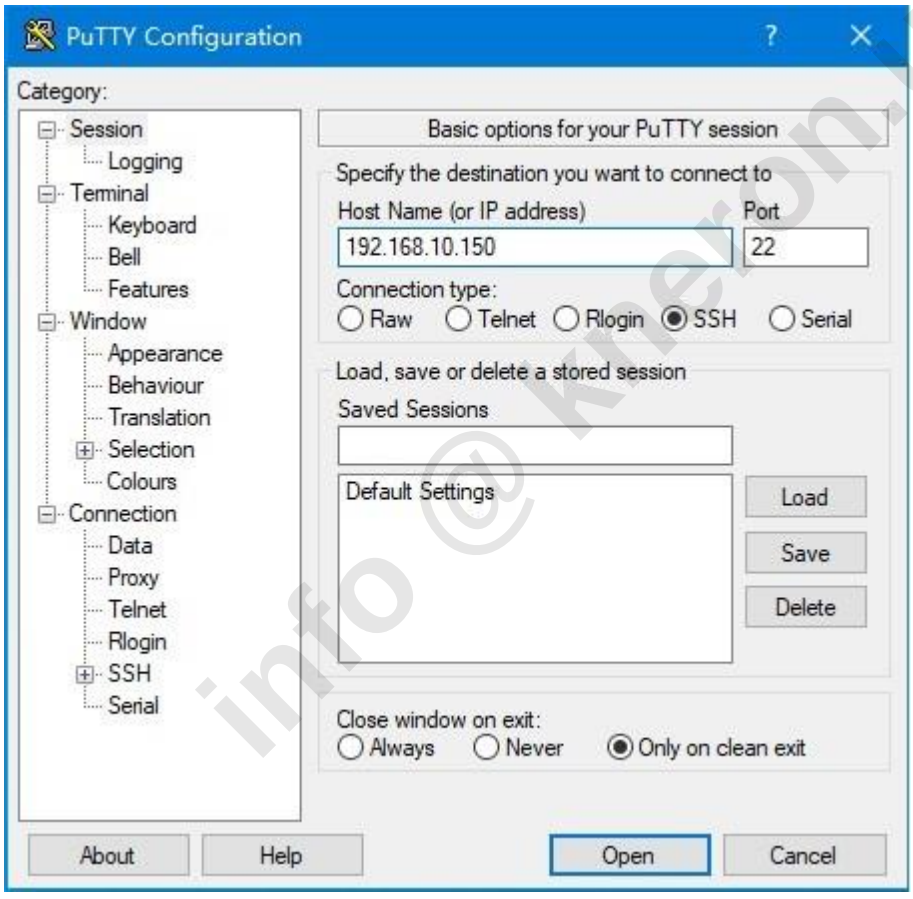


2.4 Network remote login

If the PC can successfully ping the IP address displayed on the monitor after startup, you can then use ssh to log in remotely. The port number is 22 and the username and password are both *linaro*. Users can use Windows or Linux systems, but it is recommended to use putty or SecureCRT software on Windows systems for remote login.
Command line: *ssh linaro@ip_address*

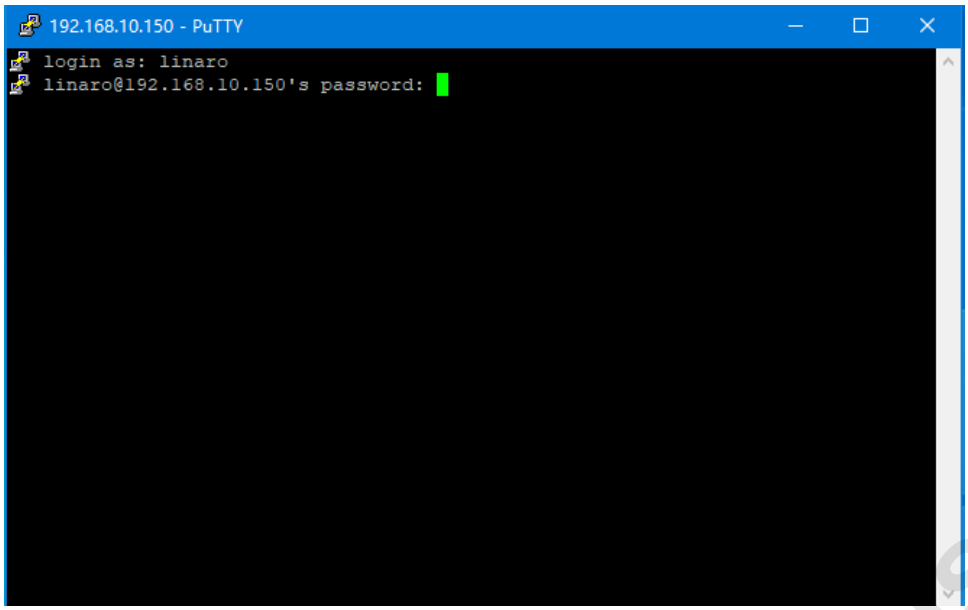
2.4.1 windows system

The following uses putty software as an example.
(1) Download and install putty.
(2) Start putty, enter the device IP address displayed on the monitor, the port number is 22

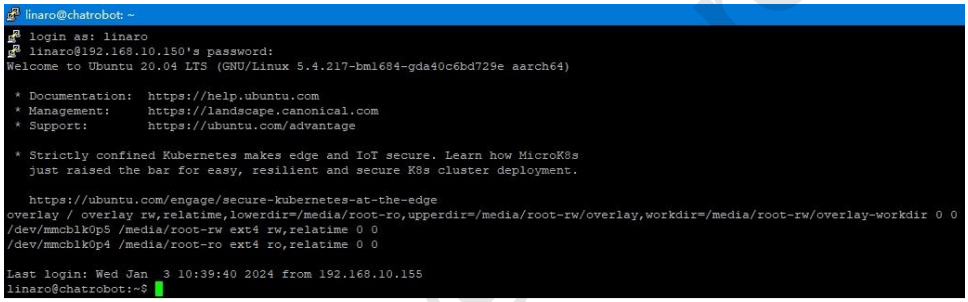


Click the "Open" button.

(3) In the next interface, enter the username and password, both are *linaro*



(4) After the verification is passed, it will be displayed as follows.



After connecting the device, you can start the chatbot software according to 3.1.

2.4.2 Linux system

Take the ubuntu system as an example.

Use the command: *ssh linaro@192.168.10.150*

192.168.10.150 is the IP address of the device, linaro is the username, and you will be prompted to enter a password later. The password is also linaro.

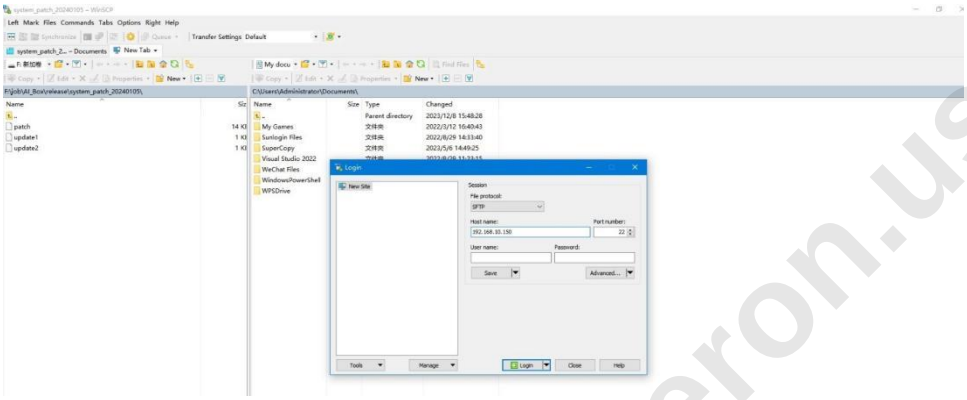
After connecting the device, you can start the chatbot software according to 3.1.

2.5 File upload and download

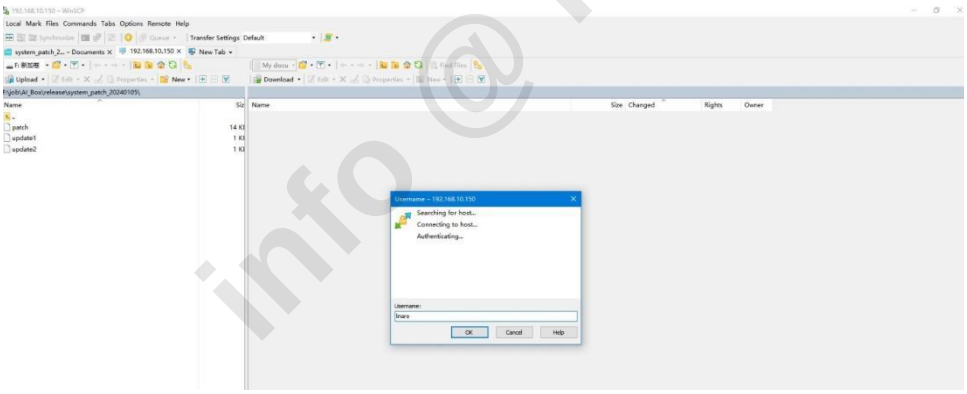
Users can upload and download files through the SCP protocol on Windows or Linux systems. The username and password are both *linaro*, and the port number is 22. We recommend using winscp for access on Windows systems.

2.5.1 windows system

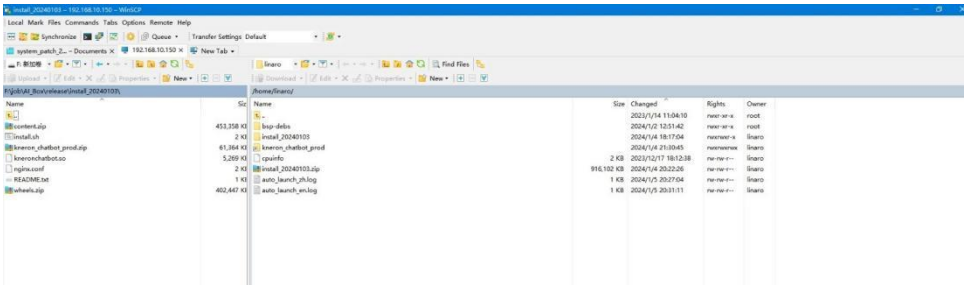
- Download and install winscp software.
- Start the winscp software, fill in the IP address of the device in the Host name column, and click "Login"



- Enter the username and password on the next screen, both are *linaro*, and click "OK".



- After the login, you can use the mouse to drag files to the box or download files from the box.



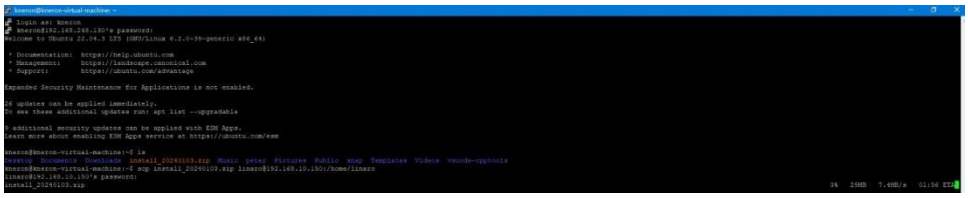
2.5.2 Linux system

Take the ubuntu system as an example.

- Upload file.

Take uploading the file install_20240103.zip to the 192.168.10.150 /home/linaro directory as an example.

Use the command: *scp install_20240103.zip linaro@192.168.10.150:/home/linaro*

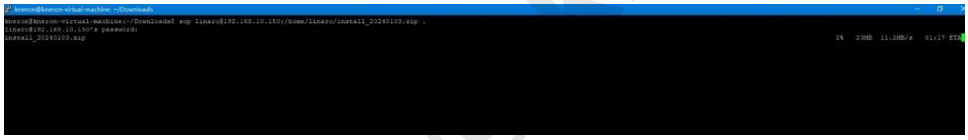


192.168.10.150 is the IP address of the device, /home/linaro is the address where the file is stored after uploading, and linaro is the username. The system will prompt you to enter the password, both of which are linaro.

- Download file

Take downloading the file install_20240103.zip from the /home/linaro directory of the device to the local as an example.

Use the command: *scp linaro@192.168.10.150:/home/linaro/install_20240103.zip* .



192.168.10.150 is the IP address of the device, /home/linaro/install_20240103.zip is the file that needs to be downloaded, and linaro is the user name. The system will prompt you to enter the password, which is also linaro. After the file is downloaded, it is stored in the current directory.

2.6 Shutdown

Please try not to disconnect the power directly when shutting down but run it first sudo poweroff before powering off to avoid damaging file system data.

In addition, if you have successfully entered the Linux system, you can also press and hold the power button. The system will detect and safely turn off the power of the system and development board.

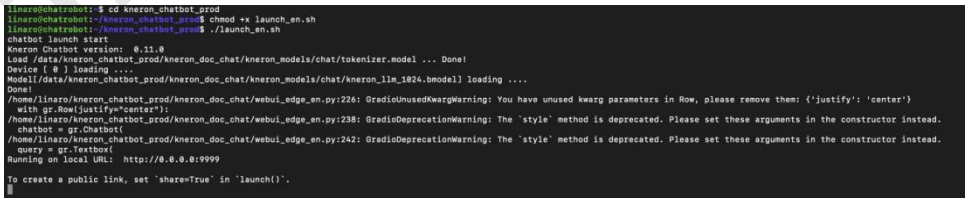
3. Chatbot software usage guide

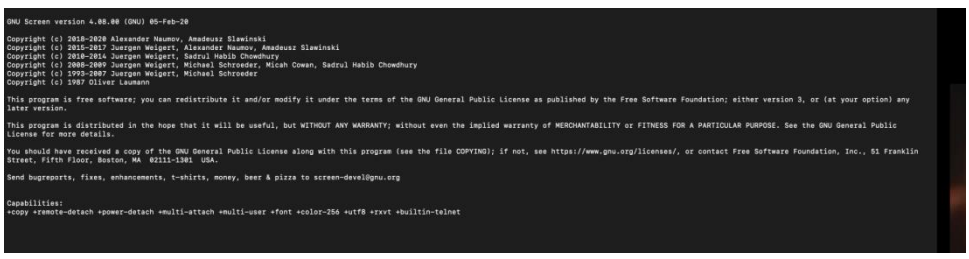
3.1 Start the WEBUI interface

3.1.1 Start up

a. Using 'screen', you can start a process that will continue to run on the server even if the SSH connection is disconnected, so that the website can continue to serve external parties.
   i. Enter screen on the command line. After Figure 2 appears, press the space bar or Return key.

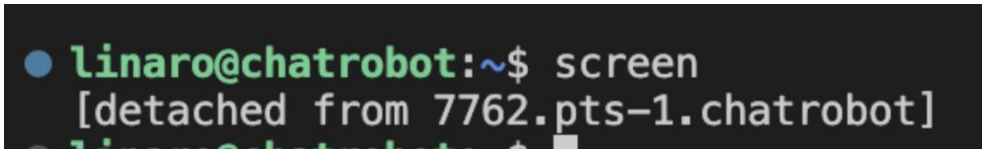
b. Enter the following commands in order on the command line, as shown in the figure below
   i. cd kneron_chatbot_prod
   ii. chmod +x new_launch.sh
   iii. ./ new_launch.sh (it takes about 30 seconds)





c. the WEBUI starts, press and hold the Ctrl , A and D keys on the keyboard at the same time ,

and the following picture will appear. This step allows even if the remote network connection is disconnected, using screen The created session (session) will still be maintained



d. Enter "screen ls" to see the sessions created by screen

```
linaro@chatrobot:~$ screen -ls
There is a screen on:
        2181.pts-2.chatrobot    (02/15/24 03:53:43)     (Detached)
1 Socket in /run/screen/S-linaro.
```

e. To retrieve the session created by screen, you need to enter "screen -r"

```
linaro@chatrobot:~$ screen -r 2181.pts-2.chatrobot
```

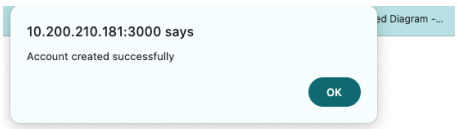
f. After Launch is successful, it will display

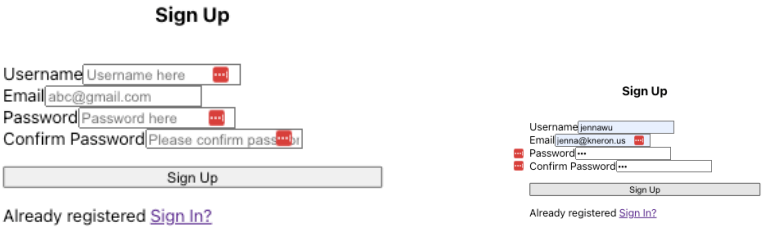
```
Compiled successfully!

You can now view auth in the browser.

  Local:            http://localhost:3000
  On Your Network:  http://10.200.210.181:3000

Note that the development build is not optimized.
To create a production build, use npm run build.

webpack compiled successfully
```

3.1.2 Admin settings

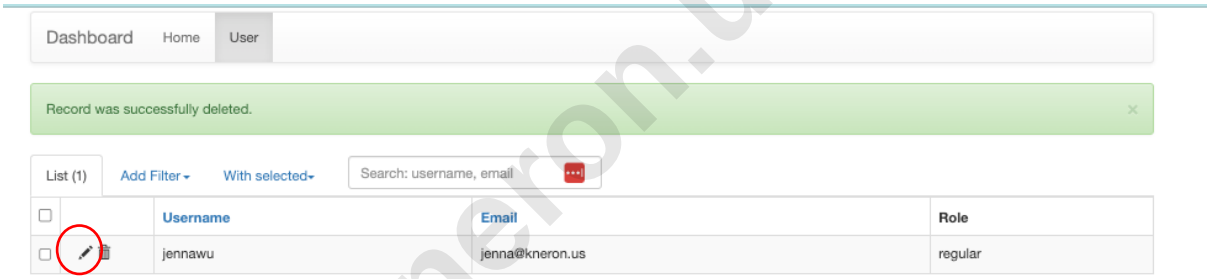
a. After the service is started successfully, enter 10.200.210.181:3000 in the local computer browser (for example, the IP of this service device is 10.200.210.181)

b. For first time use, select Sign_up ,

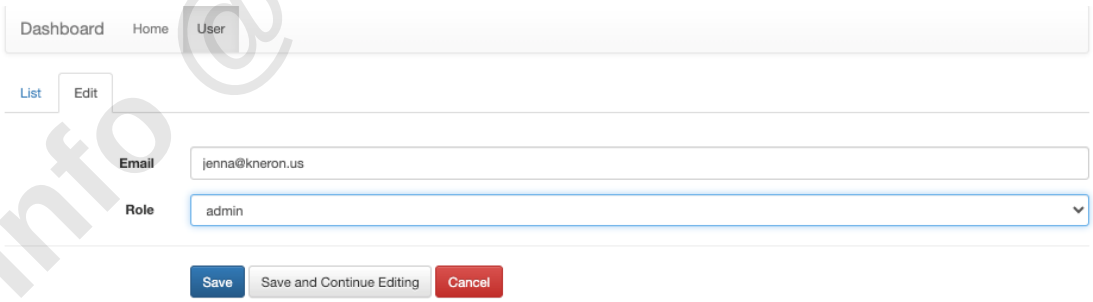




c. As Admin, after logging in, please go to http://10.200.210.181:5000/admin/user/
Set the mailbox permissions on the margin of the page. (Take device ip 10.200.210.181, jenna@kneron.us account as an example). The default user is a regular user after registration and login.



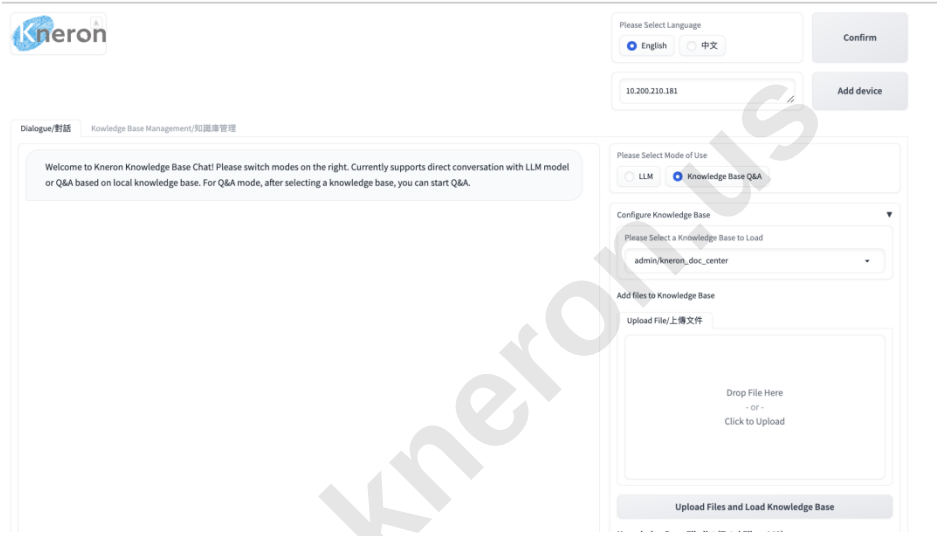
d. Click the pen circled in red in the picture above, edit the jennawu account, change the regular in the role to admin, and click "save" to save.



3.1.3 Service settings

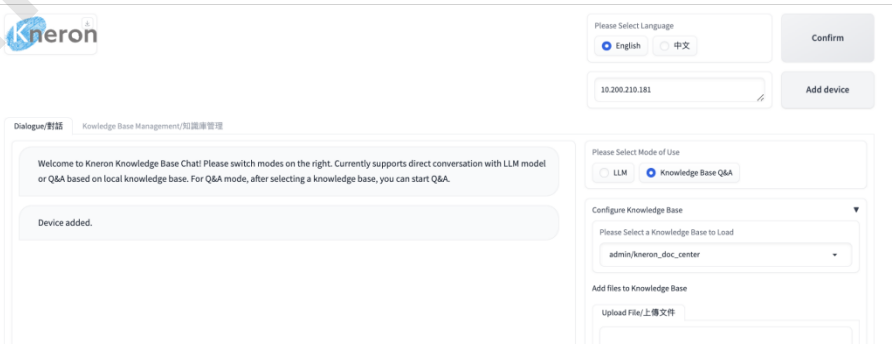
a. Enter 10.200.210.181:3000 in the local computer browser (for example, the IP of this service device is 10.200.210.181), and then log in with the updated Admin email password (wait 30 seconds). Note: Do not click the Sign-in button back and forth.

b. After logging in, the following interface appears



i. Step 1: You need to select the service device language. After selecting, please click the "Confirm" button.

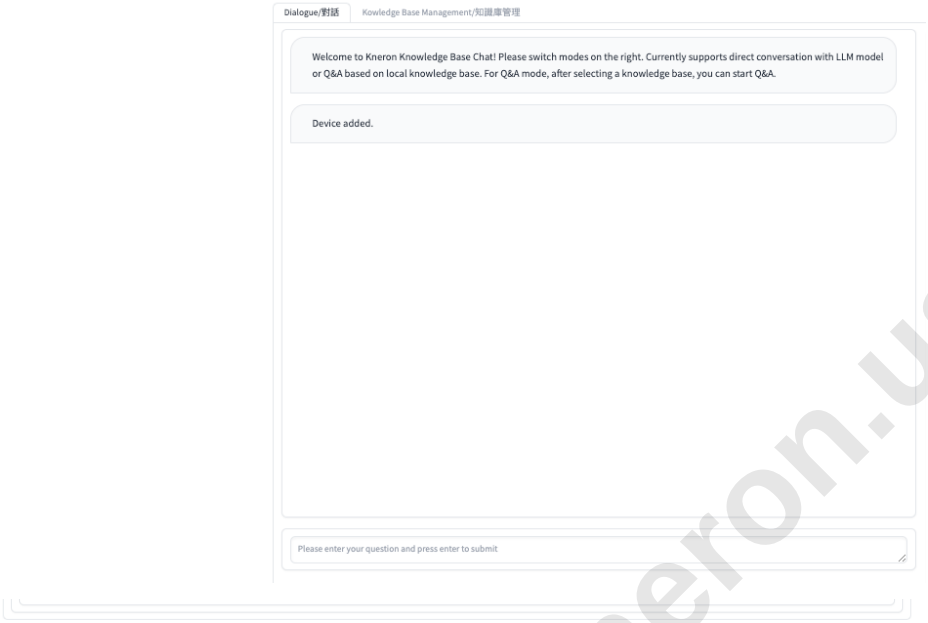
ii. Step 2: Fill in the IP address of the service device (take local as 10.200.210.181). Note that you can add multiple KNEO300 EdgeBOX IP addresses. To obtain the IP address, please refer to 2.3 Introduction. As shown in the figure, the device is added. Edge service is officially available to employees.

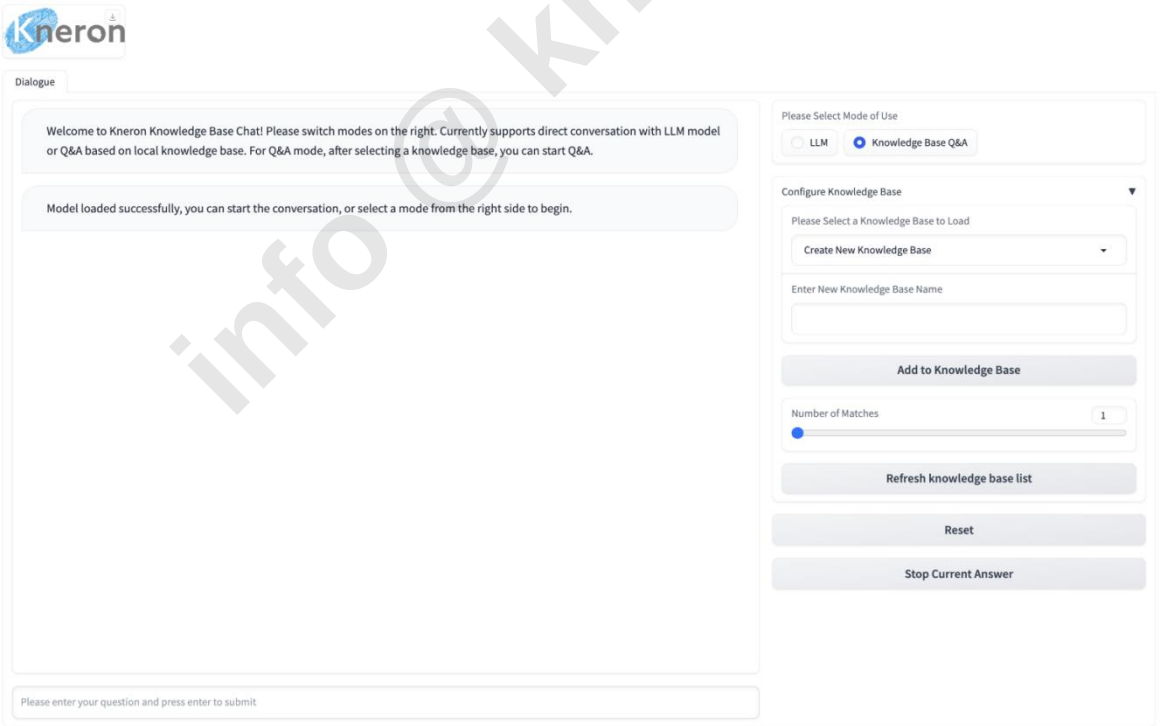


Note: Do not switch between Chinese and English at will for the admin account. Once selected, other users are already using it. If you want to switch, please coordinate the service switching time with other regular users.

3.1.4 WEBUI interface introduction

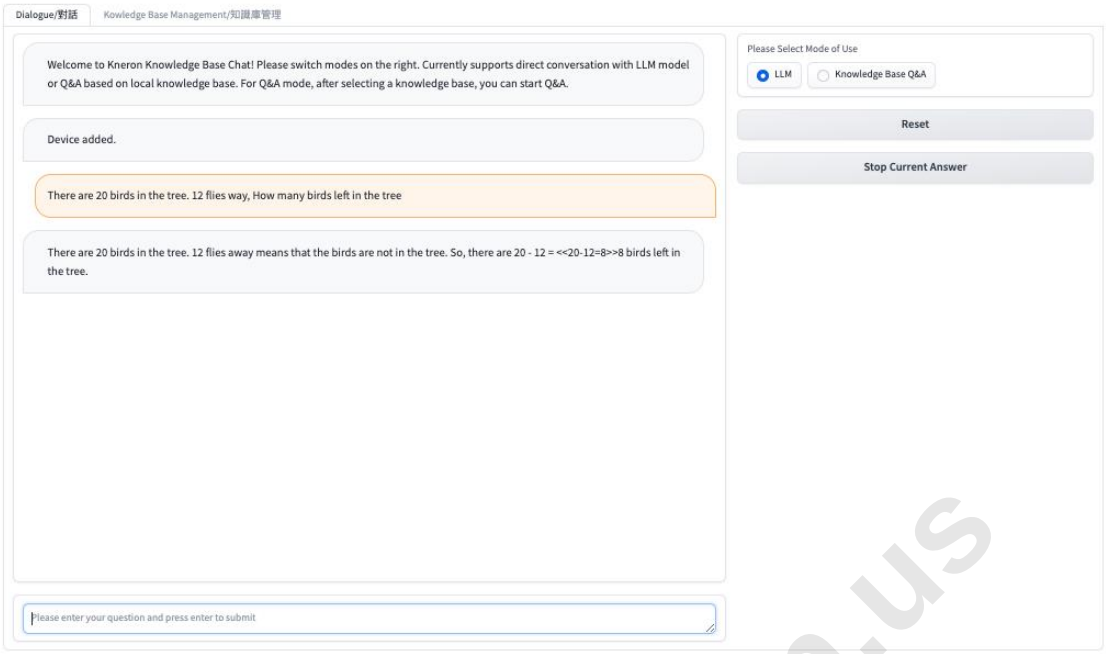
3.1.4.1. WEBUI is a chat dialog box, where users can have interactive conversations. At the same time, the knowledge base management page performs operations such as deleting and merging existing knowledge bases.





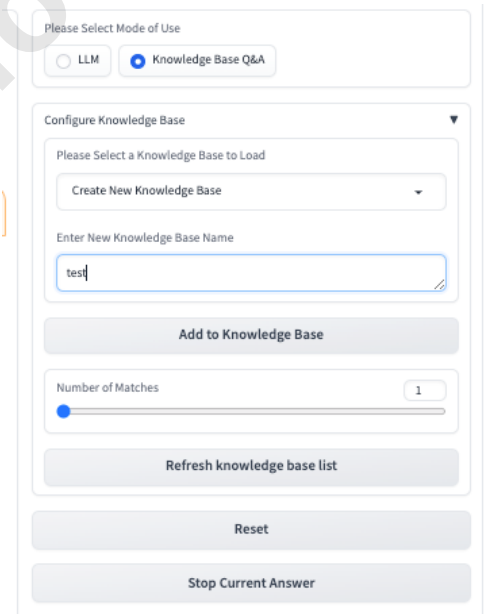
3.1.4.2. WEBUI is the selection of chat mode. Currently we have launched two modes: free conversation and knowledge base conversation.

a. Free conversation mode



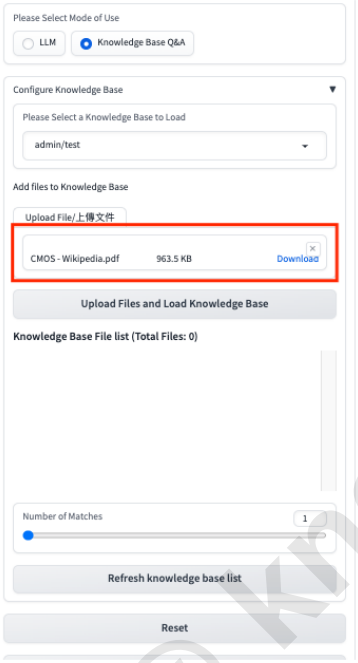
Click the LLM check box to switch to free conversation mode. Users can enter questions they want to ask in the left chat box.
“Reset”: Click this button to reset the dialog box.
“Stop current answer”: Click this button to stop the current conversation.

b. Knowledge base question and answer mode
   i. Click the knowledge base Q&A mode, and users can conduct Q&A based on the knowledge base in this interface.
   ii. Create new database.

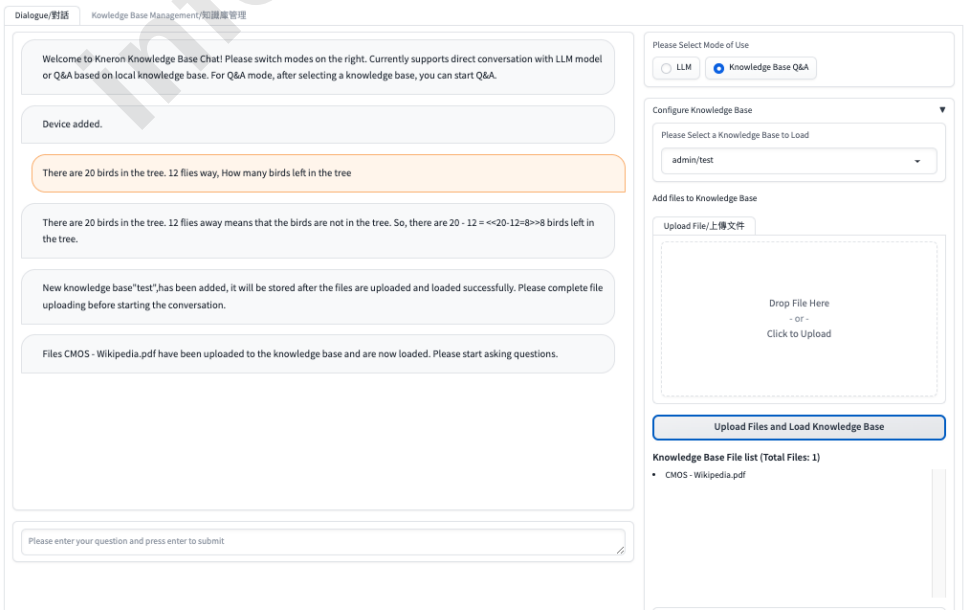


iii. Create a new knowledge base " in " Please select the knowledge base to load ".
iv. exist Enter the name of the knowledge base to be created in " Enter new knowledge base name”.
v. Click " Add to knowledge base" button to complete the creation of the knowledge base.

c. upload files
I. Currently supports documents in docx, pdf, txt, and md formats.
II. Drag the file to be uploaded to " Upload file " box.



III. Click the " Upload file and load knowledge base " button to upload the document.
IV. After the upload is complete, in the " Knowledge Base Document List" will display the names of the documents that have been uploaded and the number of documents currently in the knowledge base.



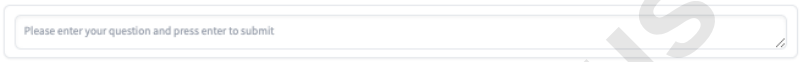
Hint:

- Please keep the file name specifications in the uploaded file name and do not include special characters, such as (), $, {}, etc.
- Normally, it takes about 15 seconds to upload a file of size 25kb, just for reference. Upload speed will be affected by file size, type, format and current network environment.

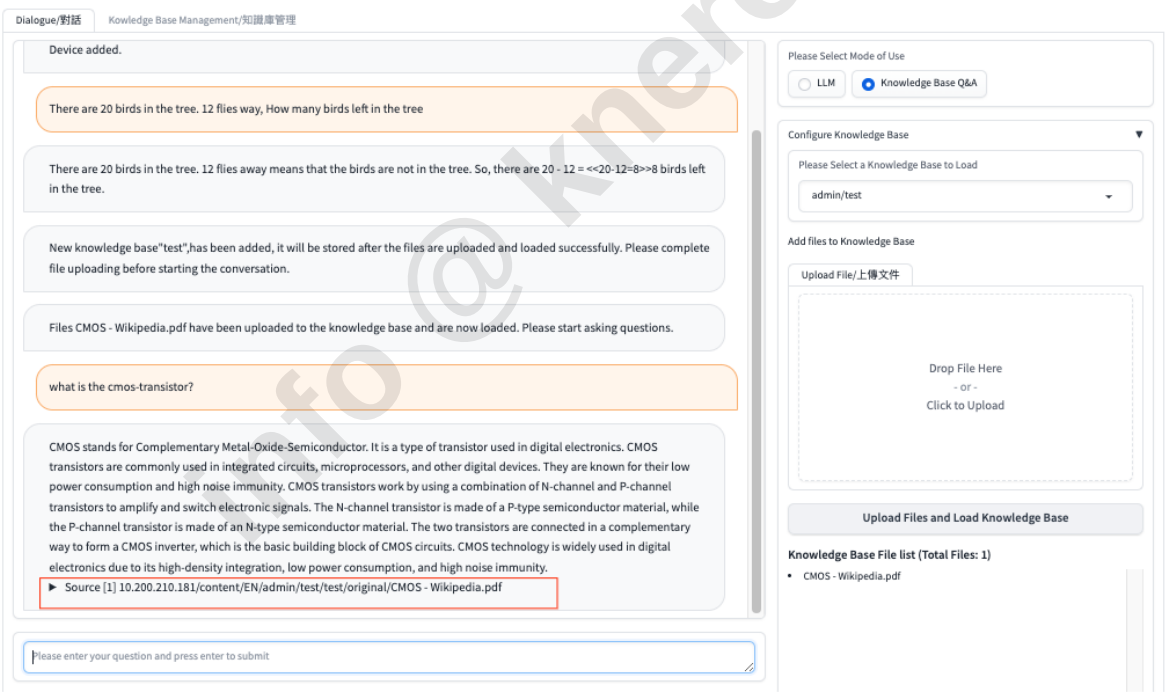
d. Question and answer based on knowledge base.

i. On the right side of the UI, there is a dropdown menu labeled "Please select the knowledge base to load" that allows you to select the desired database.

ii. Type your question in the dialog box (shown in below figure) on the left and click Enter to start the conversation.



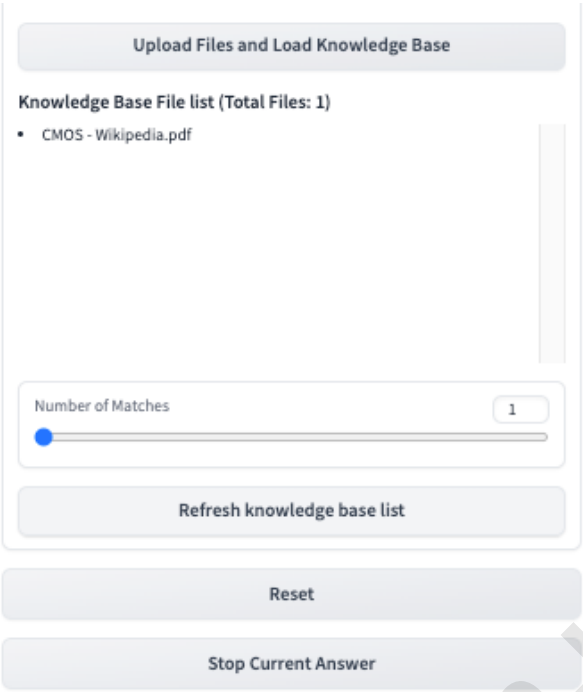
e. Download and view knowledge base files.



i. In the generated answer, you can view the source information and documents.

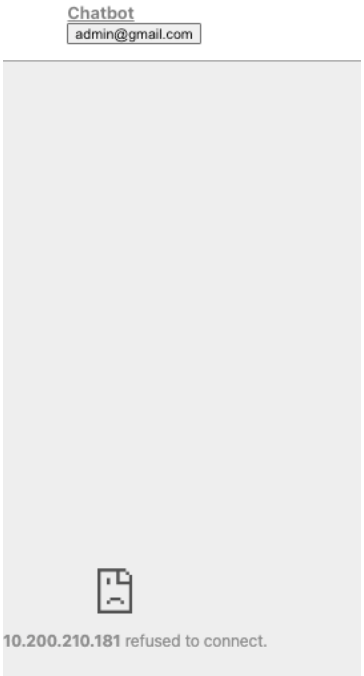
ii. Open a new browser tab, copy the file path mentioned in the KNEO300's answer regarding the source, and paste the path into the tab.

f. Other function introduction



I. "Number of Matches": The number of entries found in the given database that are related to or matching the query text. Currently the default is 1
II. " Refresh knowledge base list ": Click this button to refresh the directory of the current knowledge base.
III. " Reset " Click this button to reset the dialog box.
IV. Stop current answer: Click this button to interrupt the current conversation.

g. Web caching
   I. If you have logged in as admin before, after the device firmware is updated, when you log in with the same interface, due to the existence of web page cache, as shown in the figure below, you need to click the mailbox->log out-> and then log in again.



3.2 Others

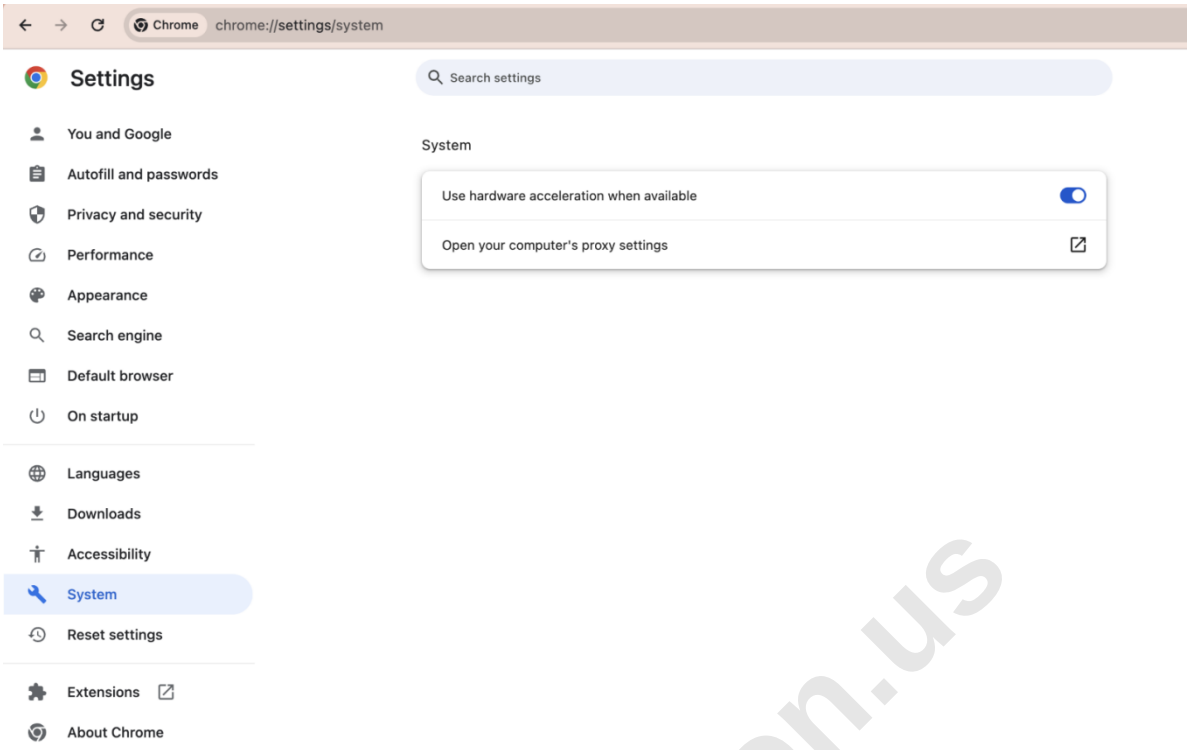
If the network you are connected to is a non-public IP, that is, it is used within a local network (such as a home, school, or corporate network) and is used for communication within the internal network (such as printers, smart devices, etc.), These IP addresses are usually automatically assigned by the local network's router. Please follow these steps:

3.2.1. on port 3000. On your local computer, please set up a local proxy.
ssh - NfL 127.0.0.1:8080:127.0.0.1:3000 linaro@192.168.200.102 (This is an example of device ip address, please change it to your actual machine ip address)

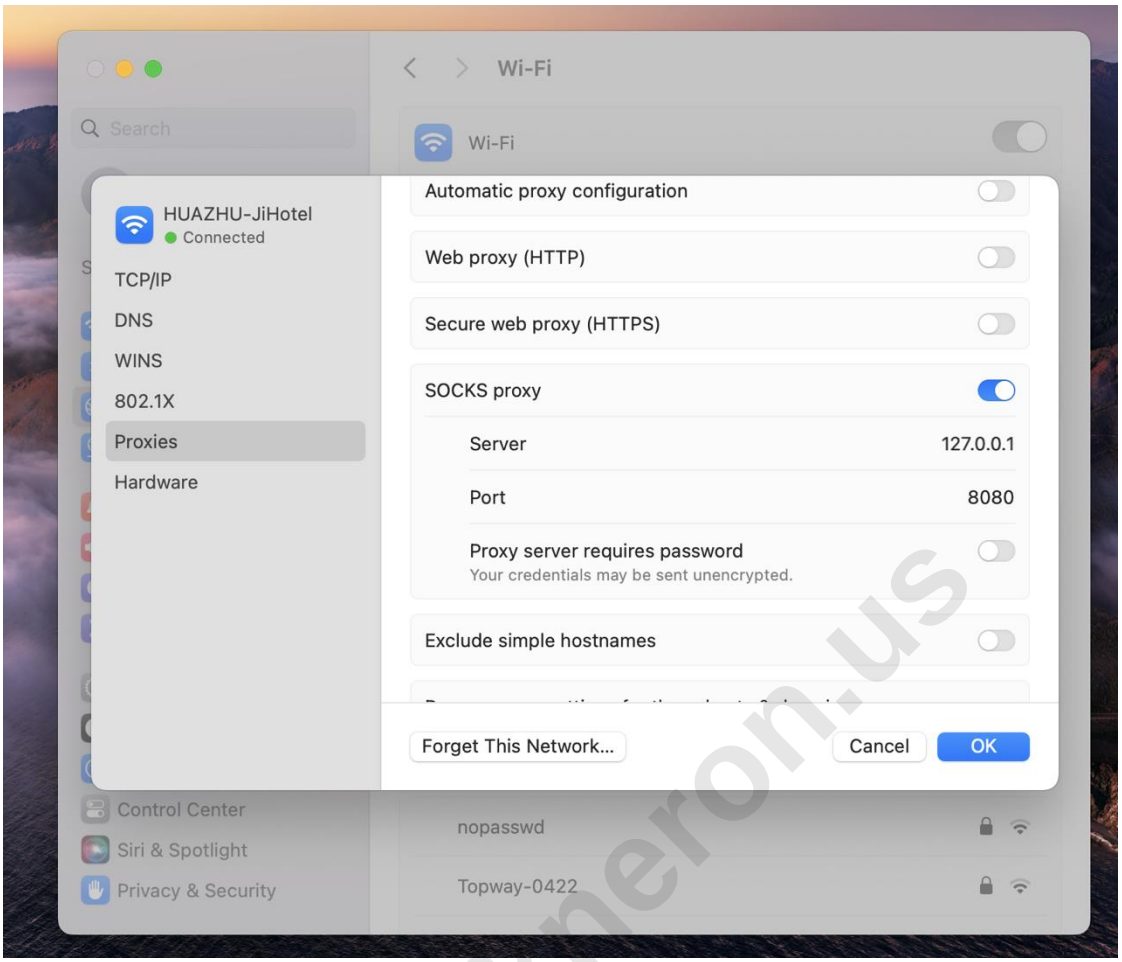
```
linaro@192.168.10.150's password:
```

3.2.2. Set up the proxy on the local machine, taking chrome as an example.

a. Go to settings.



3.2.3. Find the system in the left column, click Open your computer's proxy settings, and then set the SOCKs proxy server (127.0.0.1) and port (8080)



b. In your local computer browser, enter: http://localhost:8080/

4. System update

4.1 Update chatbot software

Please follow the steps below to update the chatbot software, taking firmware install_20240103 as an example.

- Download the update package (zip compressed format) install_20240103.zip from Kneron development center Developers | Kneron – Full Stack Edge AI, you need to apply the account first, then provide the email address that you registered to Kneron sales team, or the contact window that you purchased, to get the authority of the web access. Bellowing figure shows the snapshot of the Kneron development center that you can download from.

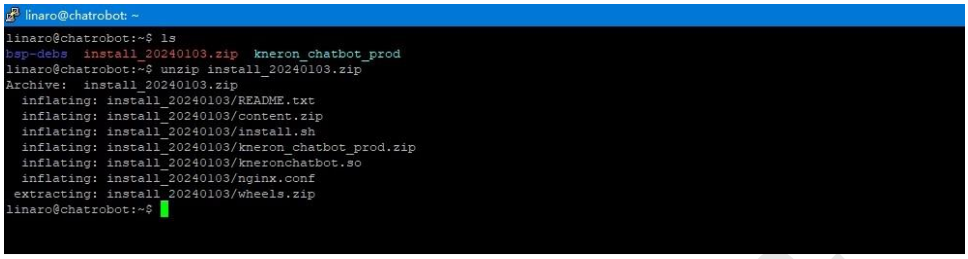
Note: The official website will also include an update guide. The update process may differ in the future, so please refer to the documentation on the official website for the most accurate information.

Kneron AI chat robot

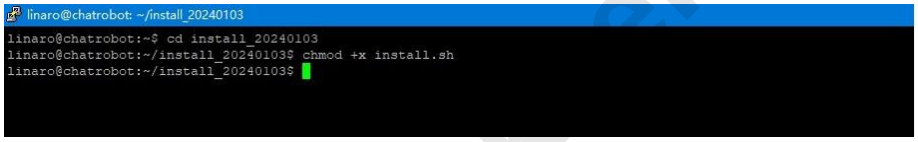
Document name	Version	Latest modified	EIP No.
KNEO300			Open folder

- Copy the update package to the device. The recommended storage address is /home/linaro/ Please refer to 2.5 for how to upload files.

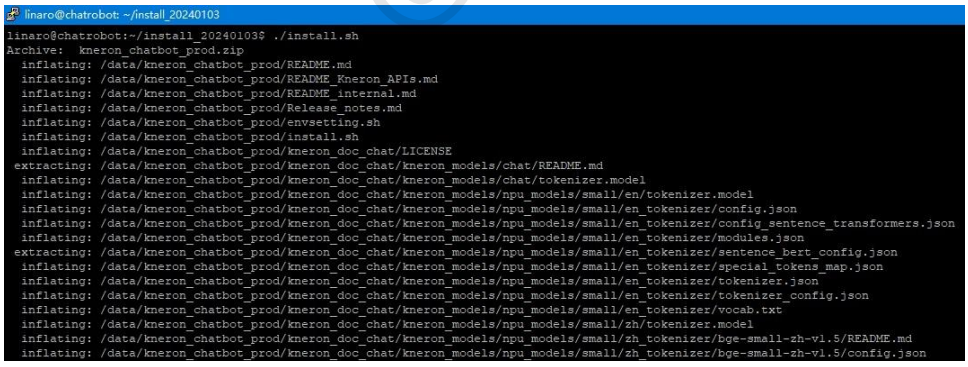
- Unzip the update package and run the command "unzip install_20240103.zip"



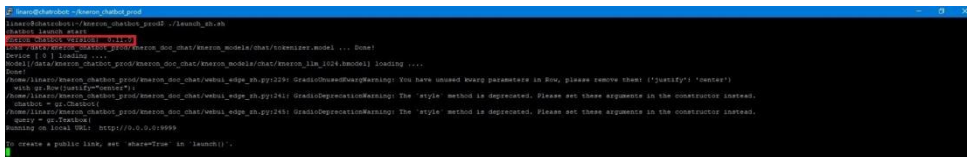
- Enter the update package directory and run the command "chmod +x install.sh" to set the running permissions



- Execute the installation script and run the command " ./install.sh". After the installation is completed, the device will automatically restart.



- Re-run the chatbot software and confirm whether the update is successful by checking the software version (shown in the red box)



4.2 Update model

The steps for updating the model are the same as updating the Chatbot. The only difference is the download address of the model update package (please contact the supplier or Kneron technical support to obtain the download address).

4.3 Update Firmware

Please contact the supplier or Kneron technical support for firmware update

5.FAQs