KNEO 300 Installation Guide V1.5

May, 2024

Revision History:

Doc Version	Description	Firmware Version	Author	Date
1.5	Rewrite the document	V0.16.0	Oscar Law	2024/05/22

Notice:

1. Kneron 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.

Contents

1. Introduction

KNEO 300 is an NPU-based edge AI server, especially 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, 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 various 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 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. Product

2.1 Product Overview

- KNEO 300 series AI box appearance

Figure 2-1 KNEO 300 Series AI Box

- 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. HDMI 2.0
7. TF Card
8. DC 12V
9. 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
Power consumption	Typical=20W, Max=45W
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
Connecting Ports	1*RS232
	1*RS485

Table 2-1 KNEO 300 Product Specification

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.

3. System update

3.1 Account Setup

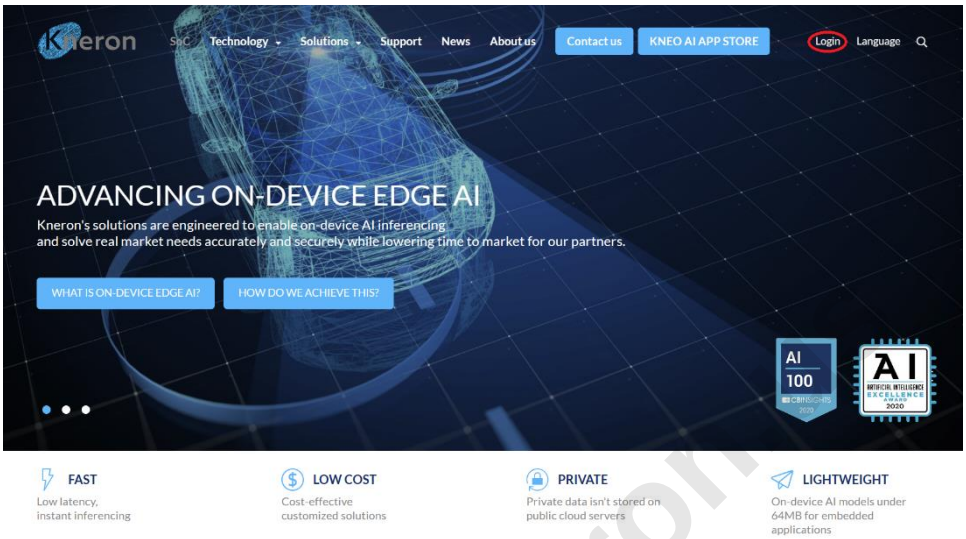

Figure 3-1 Kneron Home Page

For system updates, the administrator must register the user account on the Kneron home page. The administrator clicks the top right-hand Login button, which displays the login page. The administrator clicks the Create an account button and then follows the instructions to set up the user account. After the account setup, please provide the login e-mail to the Kneron salesperson, who shall grant permission to access the KNEO 300 documents and updated firmware.

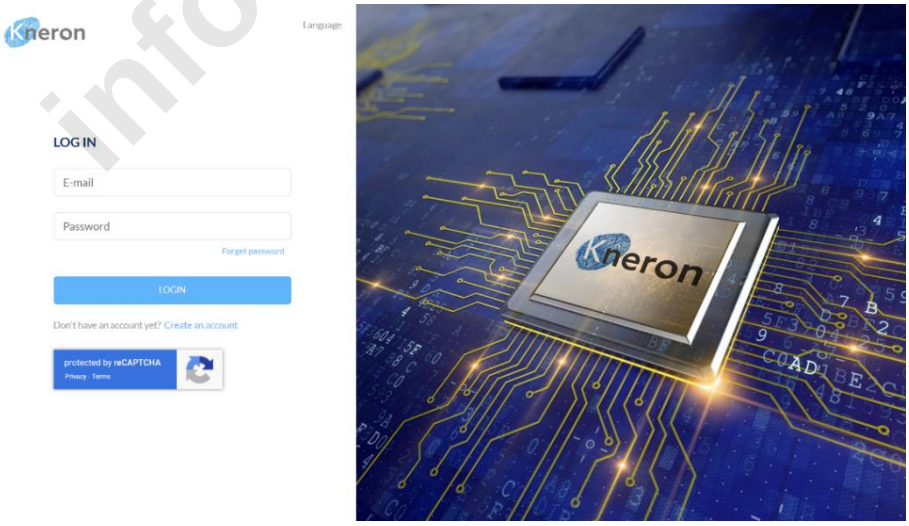

Figure 3-2 Kneron User Login

3.2 Installation Files

After the permission is granted, the administrator can access the KNEO 300 document from the Kneron developer site (https://www.kneron.com/support/developers/), then click the KNEO300 under the entry Kneron AI chat robot to access different releases. **If the entry hasn't shown up, please follow up with the Kneron salesperson for permission access.**

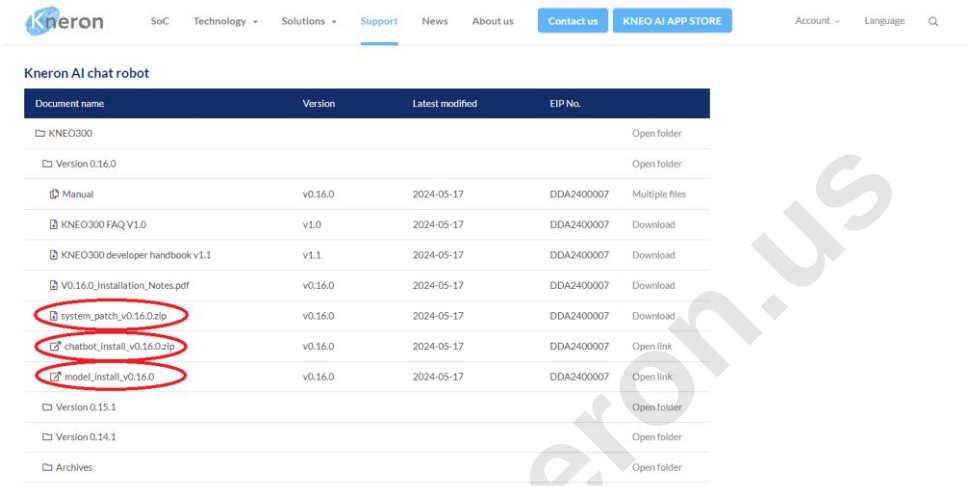

Figure 3-3 KENO 300 Document

The administrator can access the KNEO 300 Installation guide (v1.5), there are three installation files: system_patch_v0.16.0.zip(firmware), chatbot_install_v0.16.0.zip(software), and model_install_v0.16.0.zip (model). For the system_patch_v0.16.0.zip, the administrator clicks Download button on the right-hand side and downloads it to the Downloads directory. To download chatbot_install_v0.16.0.zip, the administrator presses the Open link button, which directs to the link directory, then clicks the top left-hand Download button to store the file in the Downloads directory.

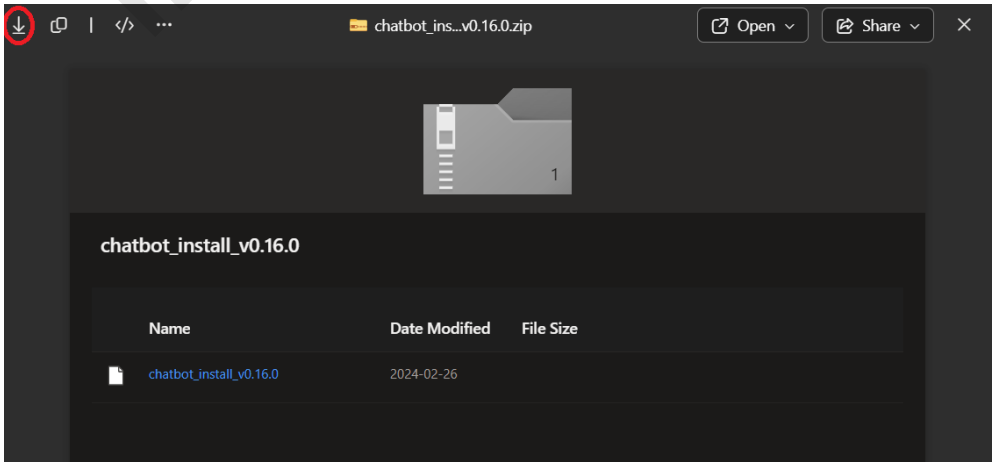

Figure 3-4 Download chatbot_install_v0.16.0.zip

Similarly, the administrator presses the Open link button to the Model folder and downloads the install.sh, README.txt, and model_install_v0.16.0.zip, separately to the Downloads directory. The administrator selects the file and clicks the Download button to download the file individually. **Don't download all three files together using the Download button only, it creates the Model.zip larger than 4Gb limit and fails to uncompress it during the model installation.**

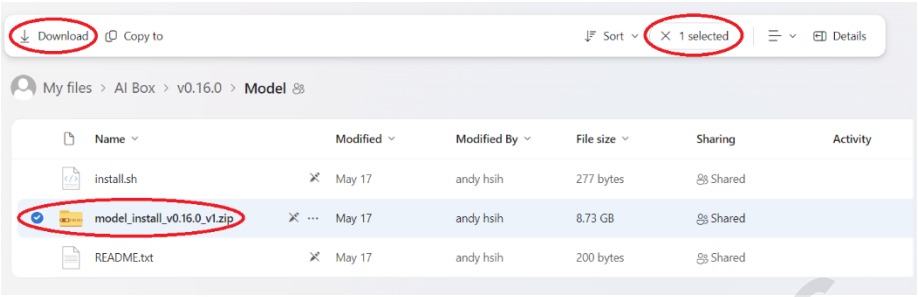

Figure 3-5 Download model_install_v0.16.0.zip

3.3 System Access

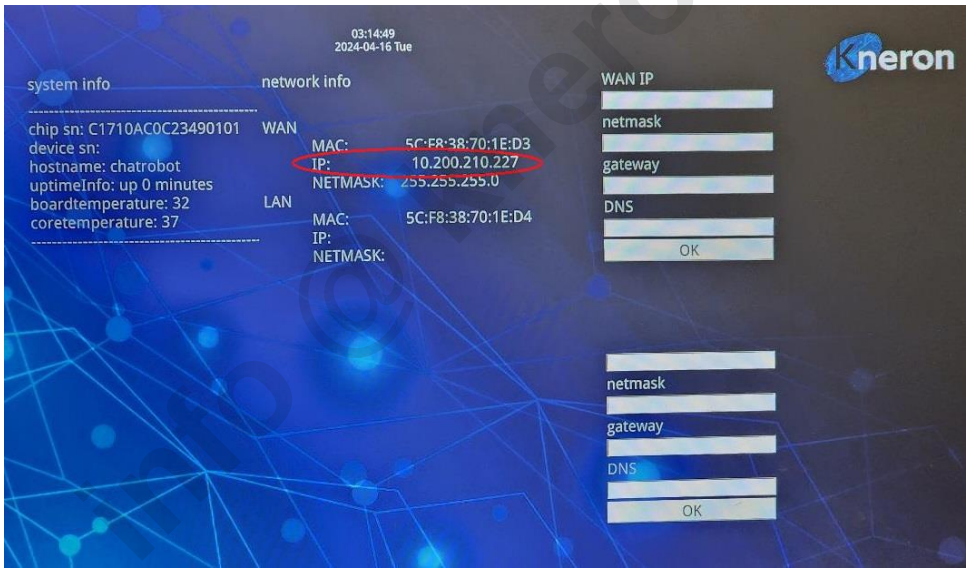

Figure 3-6 KNEO 300 IP Address

The administrator remote login the machine to perform the system installation, it first powers up the system with the following steps:

- 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 IP address (i.e. 10.200.210.227) on the screen.

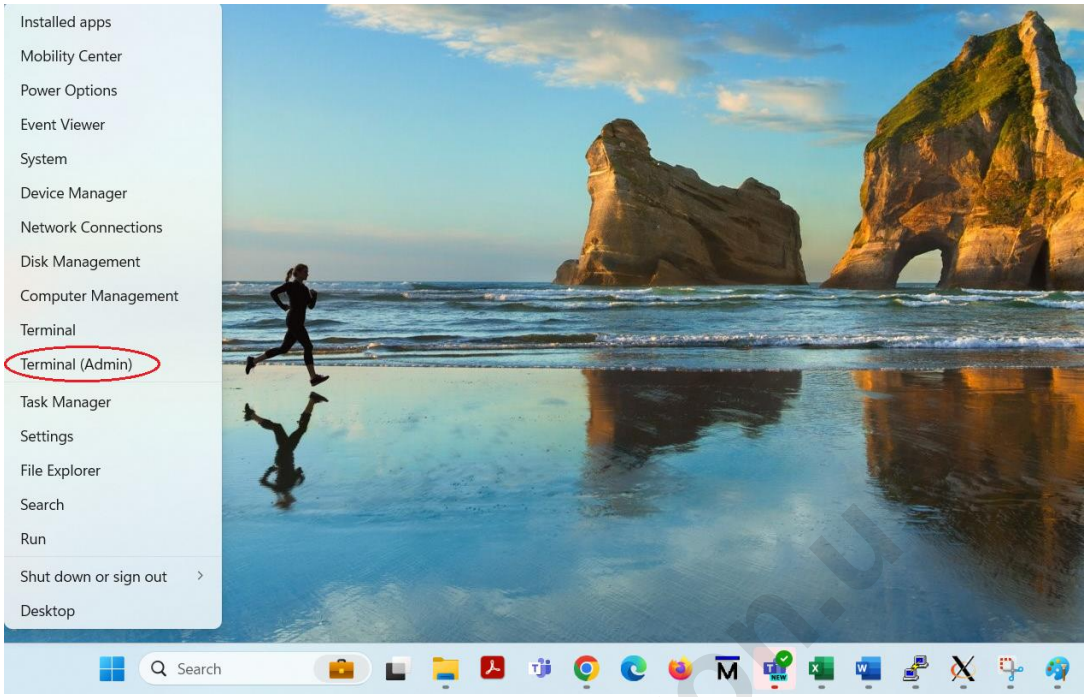

Figure 3-7 Windows PowerShell

Invoke the Windows PowerShell Terminal (Admin) to access the KNEO 300. First, click the lower left side Windows start icon with the right button, then select the Terminal (Admin) to create the terminal windows. The administrator can use the command ping with the IP address (i.e. 10.200.210.227) to check machine accessibility. If the pinging fails, please contact the local IT department to ensure the KNEO 300 is accessible. After that, it hits the CTRL-C to terminate the ping process and start to initialize the server using either ssh:

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\oscar> ping 10.200.210.227

Pinging 10.200.210.227 with 32 bytes of data:
Reply from 10.200.210.227: bytes=32 time=14ms TTL=62
```

3.4 Remote Login

To log in to the KNEO 300 with the command ssh. The username is linaro and the password is linaro

```
C:\Users\oscar > ssh linaro@10.200.210.227
linaro@10.200.210.227's password:
```

After the login, it displays the message as follows:

```
Welcome to Ubuntu 20.04 LTS (GNU/Linux 5.4.217-bm1684-g4758df7c6cfd-dirty aarch64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge        overlay / overlay
rw,relatime,lowerdir = /media/root-ro,upperdir=/media/root-rw/overlay, workdir=/media/root-
rw/overlay-workdir 0 0
/dev/mmcblk0p5 /media/root-rw ext4 rw,relatime 0 0
/dev/mmcblk0p4 /media/root-ro ext4 ro,relatime 0 0

Last login: Tue Apr 16 01:45:17 2024 from 10.200.211.128
linaro@chatrobot:~$
```


3.5 Database Backup

All the databases are stored in the directory: /home/linaro/kneron_chatbot_prod/kneron_doc_chat/knowledge_base/content, which is further divided into the EN (English) and ZN (Chinese) subdirectories. They store the different language databases dependent on the system setting, if the language is set to English, all the databases are stored under the EN subdirectory. After the administrator initializes the servers using the language English. All the databases are stored under the subdirectory (EN), the users can access those databases only. The administrator can re-initialize the system using the language Chinese, which allows the users to access those databases under the subdirectory (ZN). The system update automatically backs up the current databases and restores them after installation. **However, it recommends the administrator back up all the databases before the system update**.

For example, the system administrator can back up the user's database in the Windows environment using the command: scp -r <administrator name>@<ip address>:<user directory>/<database name> . with the option flag -r

```
C:\Users\oscar>  scp  -r  linaro@10.200.210.227:/home/linaro/kneron_chatbot_prod/kneron_doc_chat/
knowledge_base/content/EN/oscarlaw/bda602 .
linaro@10.200.210.227's password:
Understanding Artificial Intelligence (5).pdf                    100% 9854KB  40.6MB/s   00:00
index.pkl                                                        100% 442KB   30.6MB/s   00:00
index.faiss                                                      100%  881KB  49.9MB/s   00:00
title_keyword.json                                               100%   96    23.4KB/s   00:00
parent.pkl                                                       100%   43KB   8.2MB/s   00:00
```

where the administrator name is linaro with password linaro, the ip address is the local ip (i.e 10.200.210.227), the user directory is the private directory (i.e. /home/linaro/kneron_chatbot_prod/ kneron_doc_chat/knowledge_base/content/EN/oscarlaw) and the database name is the private database (i.e. bda602).

3.5 System Installation

To update the system, please ask all the users to log out of the KNEO 300, then reboot the system using the command: sudo reboot.

```
linaro@chatrobot: sudo reboot
```

After the system reboot, the administrator launches two PowerShell windows, one is used to remote login the KNEO 300 to perform the system installation, and the other is used to transfer the installation files from the Windows to the KNEO 300.

3.5.1 System Patch Update

Due to the limit of the disk space of the working directory (i.e. /home/linaro), the installation is done using the directory /data. All the files must be installed individually and not uncompressed all the files at once, resulting in disk space issues. The administrator first transfers the system_patch_v0.16.0.zip to the KNEO 300 from Windows.

```
C:\Users\oscar> scp system_patch_v0.16.0.zip linaro@10.200.210.227:/data
linaro@10.200.210.227's password:
system_patch_v0.16.0.zip                                         100%   12MB  42.8MB/s   00:00
```

The administrator uncompresses the system_patch_v0.16.0.zip using unzip, it creates the subdirectory called system_path_v0.16.0. After the uncompressing, the administrator removes the system_patch _v0.16.0.zip, and then changes to the subdirectory system_path_v0.16.0 to start the system update.

```
linaro@chatrobot:/data$ unzip system_patch_v0.16.0.zip
Archive:  system_patch_v0.16.0.zip
  inflating: system_patch_v0.16.0/README.txt
 extracting: system_patch_v0.16.0/emmcboot
 extracting: system_patch_v0.16.0/flash_update
  inflating: system_patch_v0.16.0/patch
  inflating: system_patch_v0.16.0/setup
 extracting: system_patch_v0.16.0/spi_flash
linaro@chatrobot:/data$ rm system_patch_v0.16.0.zip
linaro@chatrobot:/data$ cd system_patch_v0.16.0/
```

The administrator first reads the README.txt to understand the patch installation process, it changes the shell script patch to be executable using the command: chmod +x patch and then starts the firmware installation. It automatically reboots the system after the installation is completed. The administrator should remove the subdirectory using the command: rm -r system_patch_v0.16.0 to save the disk space.

```
 linaro@chatrobot:/data/system_patch_v0.16.0$ more README.txt
Install Steps

1. copy system_patch_v0.16.0.zip into box
2. unzip system_patch_v0.16.0.zip
3. go into system_patch_v0.16.0, and run the update scripts
      cd system_patch_v0.16.0
            chmod +x patch
      ./patch

linaro@chatrobot:/data/system_patch_v0.16.0$ chmod +x patch
linaro@chatrobot:/data/system_patch_v0.16.0$ ./patch
patch is finished
Now reboot
linaro@chatrobot:/data/system_patch_v0.16.0$ Connection to 10.200.210.227 closed by remote host.
Connection to 10.200.210.227 closed.
```


3.5.2 Chatbot Update

The administrator transfers the chatbot_install_v0.16.0.zip to the KNEO 300 using a similar command.

```
C:\Users\oscar\Downloads> scp chatbot_install_v0.16.0.zip linaro@10.200.210.227:/data
linaro@10.200.210.227's password:
chatbot_install_v0.16.0.zip                                   100% 1210MB  37.0MB/s   00:32
```

Then, the administrator uncompresses the chat_install_v0.16.0.zip to create the subdirectory: chat_install

_v0.16.0.zip, and removes the zip file after the uncompressing process.

```
linaro@chatrobot:/data$ unzip chatbot_install_v0.16.0.zip
Archive:  chatbot_install_v0.16.0.zip
  inflating: chatbot_install_v0.16.0/README.txt
  inflating: chatbot_install_v0.16.0/Release_notes.md
  inflating: chatbot_install_v0.16.0/debs.zip
  inflating: chatbot_install_v0.16.0/install.sh
  inflating: chatbot_install_v0.16.0/kneron_chatbot_prod.zip
  inflating: chatbot_install_v0.16.0/nginx.conf
  inflating: chatbot_install_v0.16.0/wheels.zip
```

The administrator reads README.txt and updates the software. It takes about 20 minutes to complete the software update process, then it automatically reboots the system again. The administrator finally removes the subdirectory using the command: rm -r chatbot_install_v0.16.0

```
linaro@chatrobot:/data/chatbot_install_v0.16.0$ more README.txt
Install Steps

1. copy chatbot_install_vX.zip(X means version, such as 0.14.0, chatbot_install_v0.14.0.zip)
into box
2. unzip chatbot_install_vX.zip
3. go into chtbot_install_vX, and run the update scripts
      cd chatbot_install_vX
      chmod +x install.sh
      ./install.sh

linaro@chatrobot:/data/chatbot_install_v0.16.0$ chmod +x install.sh
linaro@chatrobot:/data/chatbot_install_v0.16.0$ ./install.sh
Archive:  ./kneron_chatbot_prod.zip
  inflating: /data/kneron_chatbot_prod/README.md
  inflating: /data/kneron_chatbot_prod/Release_notes.md
  inflating: /data/kneron_chatbot_prod/envsetting.sh
  ….....
Removing node-strip-ansi (6.0.0-2) ...
Removing node-ansi-regex (5.0.0-1) ...
Removing nodejs (10.19.0~dfsg-3ubuntu1) ...
Setup completed successfully.
/data/chatbot_install_v0.16.0
Done.
linaro@chatrobot:/data/chatbot_install_v0.16.0$ Connection to 10.200.210.227 closed by remote
host.
Connection to 10.200.210.227 closed.
```

3.5.3 Model Update

The administrator transfers the install.sh, README.txt, and model_install_v0.16.0_v1.zip to the KNEO 300.

```
C:\Users\oscar\Downloads> scp install.sh linaro@10.200.210.227:/data
linaro@10.200.210.227's password:
install.sh                                                    100%  277     0.3KB/s   00:00
C:\Users\oscar\Downloads> scp .\model_install_v0.16.0_v1.zip linaro@10.200.210.227:/data
linaro@10.200.210.227's password:
model_install_v0.16.0_v1.zip                                  100% 8937MB  40.4MB/s   03:41
PS C:\Users\oscar\Downloads > scp .\README.txt linaro@10.200.210.227:/data
linaro@10.200.210.227's password:
README.txt                                                    100%  200     0.2KB/s   00:00
```

The administrator reads README.txt, updates the English and Chinese chatbot models, and removes the install.sh, model_install_v0.16.0_v1.zip, and README.txt after update.

```
linaro@chatrobot:/data$ more README.txt
Install Steps

1. copy model_install_v0.16.0 into box
2. go into model_install_v0.16.0, and run the update scripts
      cd model_install_v0.16.0
      chmod +x install.sh
      ./install.sh

linaro@chatrobot:/data$ chmod +x install.sh
linaro@chatrobot:/data$ ./install.sh
Archive:  ./model_install_v0.16.0_v1.zip
  inflating:
/data/kneron_chatbot_prod/kneron_doc_chat/kneron_models/npu_models/small/en/bert_en_f32.bmodel
  inflating:
/data/kneron_chatbot_prod/kneron_doc_chat/kneron_models/npu_models/small/zh/bert_zh_f32.bmodel
  inflating:
/data/kneron_chatbot_prod/kneron_doc_chat/kneron_models/npu_models/reranking/reranking_f32_v_0_1
_0.bmodel
  inflating:
/data/kneron_chatbot_prod/kneron_doc_chat/kneron_models/chat/en/kneron_llm_en_v_0_1_0.bmodel
  inflating:
/data/kneron_chatbot_prod/kneron_doc_chat/kneron_models/chat/zh/kneron_llm_zh_v_0_3_3.bmodel

model update finished!
Done.
```