KNEO 330 EdgeGPT Server Installation Guide

(v 0.20.9)

Mar 2025

Revision History:

Doc Version	Description	Firmware Version	Author	Date
0.20.9	Initial Release	V0.20.9	Oscar Law	2025/03/12

Notice:

1. Kneron Co., Ltd may change 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 FOR 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 THAT 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

Table of Figures

1 Introduction

The KNEO 330 is an EdgeGPT server powered by an NPU, tailored for Large Language Model (LLM) applications and offering 48 TOPS of AI computing performance. It features an all-metal body with fan-based cooling and boasts multiple peripheral interfaces for enhanced functionality. Compared to traditional GPU-based LLM inference, the KNEO 330 excels in cost-effectiveness, energy efficiency, and overall performance for Artificial Intelligence Generated Content (AIGC) applications. The system supports multiple-user access and functions like the chatbot with user-defined answers. Moreover, it is no longer limited to the text inputs for the knowledge base creation, it also supports embedded document (.pdf), video subtitle (.srt), and graphics (jpg/png) file formats. Finally, it allows the users to access the custom knowledge base through the private group.

The KNEO 330 comes with Kneron's proprietary edge chatbot software, designed primarily for answering questions and providing information. It functions similarly to an advanced offline virtual assistant. Key features and applications of this chatbot include:

1. **Q&A:** Support general inquiry for various areas: science, history, culture, technology, and more.
2. **Language Understanding:** Exhibits strong natural language processing abilities, enabling it to comprehend and respond to complex and abstract queries.
3. **Multiple Users Access:** Allow multiple users to access the machine without performance degradation.
4. **Parallel Processing:** Support multiple machines parallel processing to improve the overall performance
5. **Multiple Media Inputs:** Support additional embedded documents (docx/pdf), images (.jpg/.png), and video subtitles (.srt) inputs for the knowledge base creation
6. **Traditional Chatbot Support:** Offer the traditional Chatbot with user-defined answers using a new feature (i.e. Company QA)
7. **Rational Database:** Utilize a large language model (LLM) to assist with relational database manipulation (e.g., Company SQL) for multiple spreadsheet/sheet access.
8. **Knowledge Base Sharing:** Share the custom knowledge base through the private group (i.e. Company Organization)
9. **Text Generation:** In addition to answering questions, it can generate articles, craft stories, and produce creative content.

10. **User Interaction:** Facilitates smooth conversations with users, offering helpful responses and suggestions based on database information. It can be applied in various fields such as education, customer support, HR, corporate training, and IT support.
11. **Chatbot Assistant:** Customize the chatbot for various financial, insurance, medical, teaching, and meeting assistants.
12. **Privacy and Security:** Provide data protection for user information, data, and privacy using offline mode.

2 EdgeGPT Server

2.1 Product Overview

- KNEO 330 EdgeGPT Server

Figure 2-1 KNEO 330 EdgeGPT Server

- KNEO 330 EdgeGPT Server Interface[1]

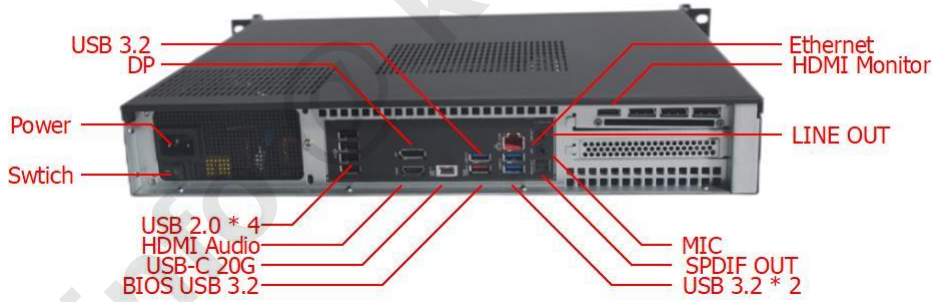

Figure 2-2 KNEO 330 EdgeGPT Server Interface

[1] Use only the top right HDMI port to connect to the monitor for display, while the other HDMI port is designated solely for audio connection only

The KNEO 330 EdgeGPT server is a standard equipment module[2] for data centers. Please consult IT professionals about installation. The KNEO 330 EdgeGPT server adopts an air-cooling system, it is recommended to leave space at the top to ensure proper airflow for cooling.

When the front panel ON/OFF button is activated (indicated by the blue light), it initiates the system in active mode. When the button is turned off, it initiates a soft shutdown, placing the system in standby mode without cutting off the power supply. Pressing the button wakes up the system. The back panel switch directly connects/disconnects the power supply. When the switch is off, the system is completely shut down even if the front panel ON/OFF button is pressed. It recommends the administrator turn off the ON/OFF button first followed by the back panel switch to move the system.

Product Parameters

CPU	Intel i5 10 cores 16 threads 4.6 GHz CPU
NPU	48 TOPS (INT8) equivalent
DRAM	32 Gb DDR4
Storage	2Tb SSD
Power	100-240V, 50/60Hz Avg 140W, Max 320W
Operating System	Ubuntu Linux
Size	428 x 350 x 66.6 mm (16.85 x 13.78 x 2.62 in)
Weight	7.3 kg (16.09 lb.)

Table 2-1 KNEO 330 Product Specification

2.2 Accessories List

Upon receiving the device, ensure that all accessories are included. Additional Ethernet and HDMI cables are necessary to connect the machine to the internet, and a monitor for system setup.

[2] There are multiple input/output data ports, including High-Definition Media Interface (HDMI), Universal Serial Bus (USB), Display Port (DP), LINE OUT (Audio Output), and Sony/Philips Digital Interface Output (SPDIF OUT)

- KNEO 330 EdgeGPT Server
- One AC power cord

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

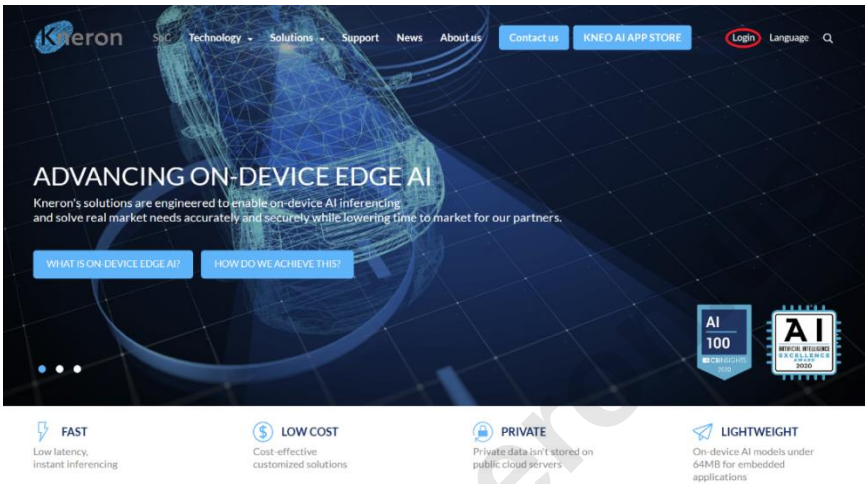

Figure 3-1 Kneron Homepage

The administrator first registers a user account on the Kneron homepage (https://www.kneron.us). The administrator clicks the Login button in the top-right corner, which opens the login page, selects Create an account, and follows the instructions to complete the registration process. Once the account is set up, the administrator must send the request to the email (info@kneron.us), which includes the first name, last name, e-mail, machine model, and serial number. Kneron reviews the request and grants access to the KNEO 330 documents and updated firmware.

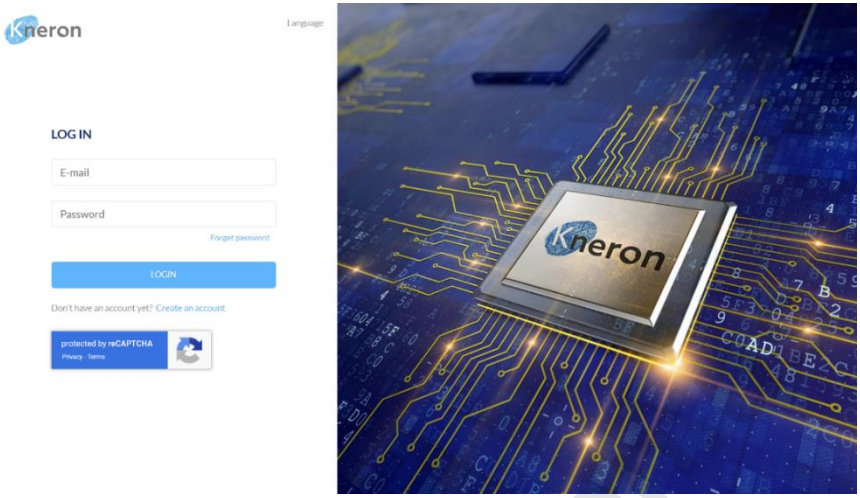

Figure 3-2 Kneron User Login

Once permission is granted, the administrator accesses the software and documents on the developer site (https://www.kneron.com/support/developers/), then navigate to the Kneron AI chat robot (KNEO330) and click on KNEO330 to proceed

3.2 System Access

The administrator login the KNEO 330 to perform the system installation, it first powers up the system with the following steps:

- Connect the power cable to the 100-240V 50/60Hz power cord.
- Connect the device and monitor with the HDMI cable.
- Plug the network cable into the Ethernet port and connect it to the network.
- Once powered on, the device will automatically start, and the default terminal is initialized. The administrator first logs in to the system with user ID: aiuser with

password: aiuser, then types the command: ifconfig to display the server IP address shown in inet entry (i.e. 10.200.210.237[3]) for web access

```
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.8.0-40-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

45 updates can be applied immediately.
7 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

Last login: Fri Sep 13 14:23:45 2024 from 10.200.211.96
aiuser@kneron330:~$ ifconfig
eno1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 10.200.210.237  netmask 255.255.255.0  broadcast 192.168.200.255
        inet6 fe80::2979:5613:2a22:d58  prefixlen 64  scopeid 0x20<link>
        ether 10:7c:61:74:cd:d0  txqueuelen 1000  (Ethernet)
        RX packets 9407575  bytes 698190452 (698.1 MB)
        RX errors 0  dropped 606829  overruns 0  frame 0
        TX packets 191990  bytes 117483916 (117.4 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 211475  bytes 86707283 (86.7 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 211475  bytes 86707283 (86.7 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```


3.3 System Login

Open the Windows PowerShell Terminal with administrative privileges to access the KNEO 330 EdgeGPT server for system management. Right-click the Windows

[3] Please keep the IP address (i.e., 10.200.210.237) for the future remote access

start icon in the lower left corner, then select Terminal (Admin) to launch the terminal window. The system administrator can use the ping command followed by the IP address (e.g., 10.200.210.237) to verify the machine's accessibility. It gets a Reply from the machine to confirm that the machine is alive. After completing the ping process, press CTRL-C to stop it, and then proceed to initialize the server using SSH or PuTTY.

Figure 3-3 Window PowerShell

```
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 10.200.210.237

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

3.3.1 SSH

Use the ssh command to log in to the KNEO 330. The username and password are both aiuser.

```
C:\Users\oscar> ssh aiuser@10.200.210.237
aiuser@10.200.210.237's password:
```

After logging in, the following message is displayed[4]:

```
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.5.0-44-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


The list of available updates is more than a week old.
To check for new updates run: sudo apt update
Last login: Tue Aug 27 13:31:53 2024 from 10.200.211.96
```


3.3.2 PuTTY

Download putty from the official website (https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html) and select the 64-bit x86 package from the Windows Installer options. Once the software is downloaded, double-click the binary file and follow the on-screen instructions to complete the installation. Next, launch

[4] Please use the aiuser commands to access KNEO 330, as standard LINUX commands may not function correctly

putty as shown in Figure 4-2, and enter the IP address (i.e., 10.200.210.237) along with port number 22.

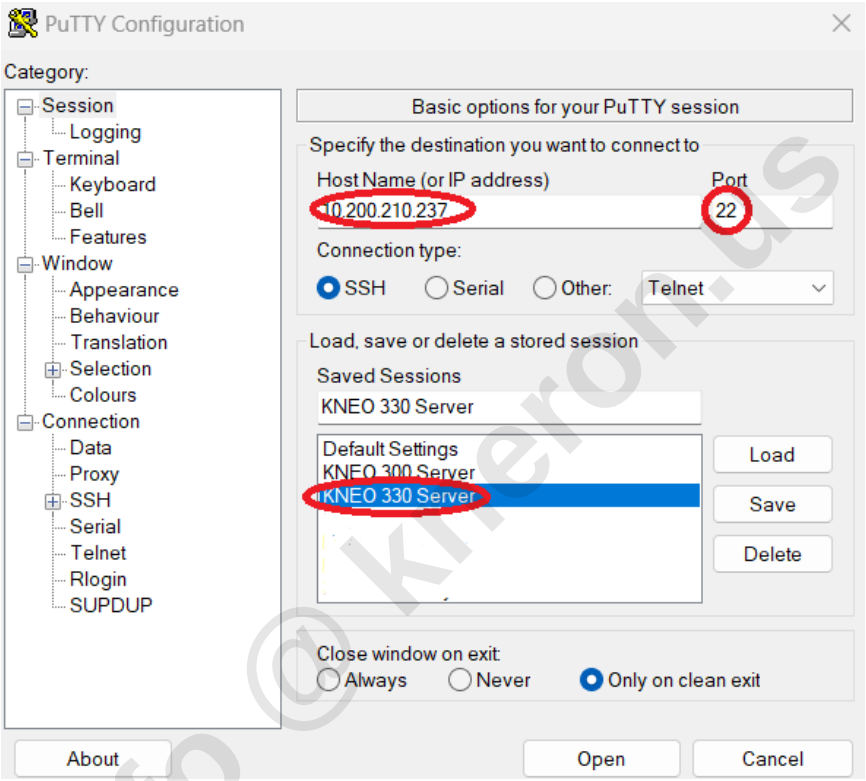

Figure 3-4 PuTTY Home Screen

Administrators and users can save the IP address and port number under 'Saved Sessions' (e.g., EdgeGPT Server) and use the 'Load' command to initialize the KNEO 330 in future sessions. For now, click the 'Open' button to start the PuTTY session and log in using the username aiuser with the password aiuser.

3.4 External Storage

The KNEO 330 can store its knowledge base on a USB drive or Network Attached Storage (NAS) for the database backup. The USB drive connects directly to the KNEO 330 via the USB port, while the NAS is accessed through the internet.

3.4.1 USB Drive

A USB drive is mounted in KNEO 330 with the format: exFAT, not vFAT. To verify the USB drive format in Windows, the administrator inserts the drive into a USB port and opens File Explorer, then right-clicking on the drive and selecting Properties, the format information is displayed in the pop-up menu, including the driver type, file system format, used/free space, and disk capacity.

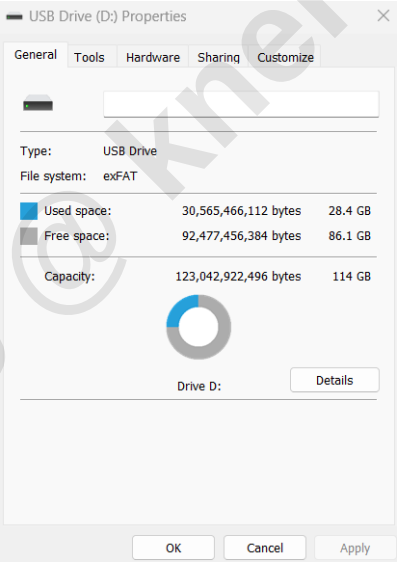

Figure 3-5 USB Drive Properties

The administrator right-clicks on the drive and invokes the Format commands, which sets the File System to exFAT (Default) to format the drive.

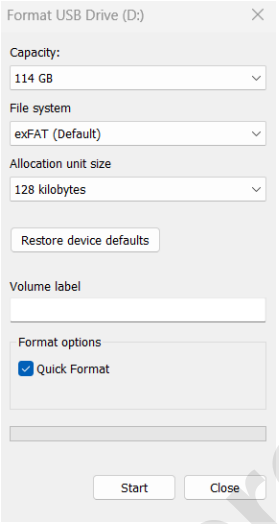

Figure 3-6 Format the USB Drive with exFAT

Before the administrator mounts the device into the system, it first checks the device information using the command: lsblk

```
aiuser@kneron330:~$ lsblk
NAME         MAJ:MIN RM   SIZE RO TYPE  MOUNTPOINTS
loop0          7:0    0     4K  1 loop  /snap/bare/5
loop1          7:1    0  74.3M  1 loop  /snap/core22/1564
loop2          7:2    0  74.3M  1 loop  /snap/core22/1612
loop3          7:3    0 269.8M  1 loop  /snap/firefox/4793
loop4          7:4    0 271.2M  1 loop  /snap/firefox/4848
loop5          7:5    0   497M  1 loop  /snap/gnome-42-2204/141
loop6          7:6    0 505.1M  1 loop  /snap/gnome-42-2204/176
loop7          7:7    0  91.7M  1 loop  /snap/gtk-common-themes/1535
loop8          7:8    0  12.9M  1 loop  /snap/snap-store/1113
loop9          7:9    0  12.3M  1 loop  /snap/snap-store/959
loop10         7:10   0  40.4M  1 loop  /snap/snapd/20671
loop11         7:11   0  38.8M  1 loop  /snap/snapd/21759
loop12         7:12   0   476K  1 loop  /snap/snapd-desktop-integration/157
loop13         7:13   0   500K  1 loop  /snap/snapd-desktop-integration/178
loop14         7:14   0   256G  0 loop
```

```
└─kneron_enc 252:0    0   256G  0 crypt /mnt/kneron_enc
sda            8:0    1  57.3G  0 disk
└─sda1         8:1    1  57.3G  0 part
nvme0n1      259:0    0   1.9T  0 disk
├─nvme0n1p1  259:1    0   512M  0 part  /boot/efi
└─nvme0n1p2  259:2    0   1.9T  0 part  /var/snap/firefox/common/host-
hunspell
```

The administrator can mount the USB drive to the system with mount points 0-3 using aimount command: sudo aimount <mount point> <device> where <device> is referred to /dev/sda1 or /dev/sdb1. Since the device /dev/sda1 is taken, . It mounts the USB drive /dev/sdb1 to mount point 2 with the command: sudo amount 2 /dev/sdb1

```
aiuser@kneron330:~$ sudo aimount 2 /dev/sdb1
Device /dev/sdb1 mounted at /home/aiuser/mnt/data2
```

The administrator can check the system device using the command: sudo aimount –show

```
aiuser@kneron330:~$ sudo aimount --show
Mount point 0: /home/aiuser/mnt/data0
Mount point 1: /home/aiuser/mnt/data1
Mount point 2: /home/aiuser/mnt/data2
Mount point 3: /home/aiuser/mnt/data3
Current status:
/dev/sdb1              115G   29G   87G  25% /home/aiuser/mnt/data2
```

The administrator unmounts the USB drive using the command: sudo aiumount <mount point> and checks the system device using the command: sudo aiumount –show

```
aiuser@kneron330:~$ sudo aiumount 2
Device at /home/aiuser/mnt/data2 is unmouted.
aiuser@kneron330:~$ sudo aiumount --show
```

```
Mount point 0: /home/aiuser/mnt/data0
Mount point 1: /home/aiuser/mnt/data1
Mount point 2: /home/aiuser/mnt/data2
Mount point 3: /home/aiuser/mnt/data3
Current status:
```

3.4.2 NAS Storage

Similarly, the administrator can mount the external NAS storage on KNEO 330 using the command: sudo aimount <mount point> <machine>:<volume> where <machine> is referred to NAS IP address and <volume> is set to the directory name. For example, the external NAS storage is referred to as the drive /mnt/kds/data_feed in the machine 10.200.100.60, and links it to mount point 1, the command becomes: sudo aimount 1 10.200.100.60:/mnt/kds/data_feed

```
aiuser@kneron330:~$ sudo aimount 1 10.200.100.60:/mnt/kds/data_feed
Device 10.200.100.60:/mnt/kds/data_feed mounted at /home/aiuser/mnt/data1
```

The administrator can check the system device using the command: sudo aimount –show

```
aiuser@kneron330:~$ sudo aimount --show
Mount point 0: /home/aiuser/mnt/data0
Mount point 1: /home/aiuser/mnt/data1
Mount point 2: /home/aiuser/mnt/data2
Mount point 3: /home/aiuser/mnt/data3
Current status:
10.200.100.60:/mnt/kds/data_feed   42T  7.7T   34T  19%
/home/aiuser/mnt/data0
```

The administrator unmounts the NAS storage using the command: sudo aiumount <mount point> and checks the system device using the command: sudo aiumount –show

```
aiuser@kneron330:~$ sudo aimount 1 10.200.100.60:/mnt/kds/data_feed
```

```
Device 10.200.100.60:/mnt/kds/data_feed mounted at /home/aiuser/mnt/data1
aiuser@kneron330:~$ sudo aiumount 1
Device at /home/aiuser/mnt/data1 is unmouted.
aiuser@kneron330:~$ sudo aiumount --show
Mount point 0: /home/aiuser/mnt/data0
Mount point 1: /home/aiuser/mnt/data1
Mount point 2: /home/aiuser/mnt/data2
Mount point 3: /home/aiuser/mnt/data3
Current status:
```

3.5 System Backup

The administrator can back up system information to external storage using the command: sudo aibackup <information> <location> [filename]. This command allows backing up different types of information, including the database (knowledge base), app_configs (user configurations), user_info (user profiles), and user_filter (custom configurations). The location parameter specifies the external storage mount point (e.g., 0, 1, 2, 3). The filename is optional with a default value: database (chatbot_database.tar.gz), app_configs (chatbot_user_configs.tar.gz), user_info (chatbot_users.db), user_filter (user_filter_config.json).

To back up the information (i.e. knowledge base) use the command: sudo aibackup database 1

```
aiuser@kneron330:~$ sudo aibackup database 1
Backup database to /home/aiuser/mnt/data1/chatbot_database.tar.gz
```

The administrator can restore backup information using the command: sudo airestore <information> <location> <filename>. The information and filename correspond to the backup one, and the area is set to an external mount point (e.g., 0, 1, 2, 3).

To restore the information (i.e. knowledge base) using the command: sudo airestore database 1 chatbot_database.tar.gz

```
aiuser@kneron330:~$ sudo airestore database 1 chatbot_database.tar.gz
Warning: This operation will overwrite the current database. We recommand you
to backup the current database before restore.
Continue?(Y/N) Y
Restore database from /home/aiuser/mnt/data1/chatbot_database.tar.gz
```

3.6 System Information

Before updating the latest software, the administrator should verify the current release. The release includes both the system and edge packages. To check the system package, use the command: sudo aipackage --version, and to check the edge package, use the command: sudo edge-upgrade -v.

To check the current system package:

```
aiuser@kneron330:~$ sudo aipackage --version
AI package version: 1.19.0
```

To check the current edge package:

```
aiuser@kneron330:~$ sudo edge-upgrade -v
Current edge: kneron-chatbot-edge
Current version: v0.20.8(79534c6292b3c0ee7aef552346de36d9b5a62e08)
```

3.7 Software Download

The administrator navigates to the developers' area (i.e. https://www.kneron.com/support/developers) and accesses the KNEO 330 download folder labeled Kneron AI Chat Robot (KNEO 330). The administrator clicks on the Version 0.20.8 icon and proceeds to the Installation of Related Files sub-folder. This sub-folder contains two text files: AI Packages (System Packages) and Kneron Edge Package (Software), each containing a download link.

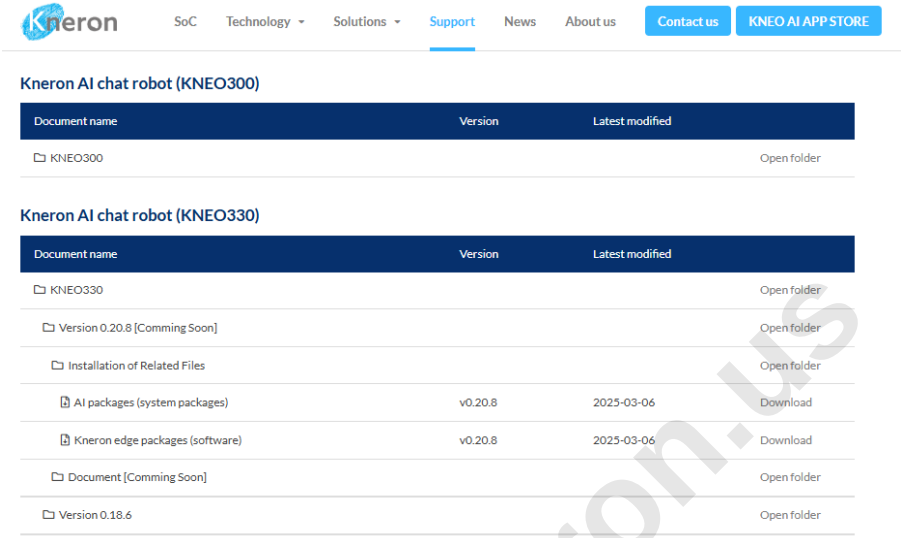

Figure 3-7 KNEO 330 Release Folder

The administrator opens the Kneron Edge Package (Software), clicks the link, and then redirects to the download drive. The administrator selects the edge package and clicks the Download button to download the edge packages (i.e. kneron-chatbot-edge-v0.20.8.pkg) to the default directory: Download

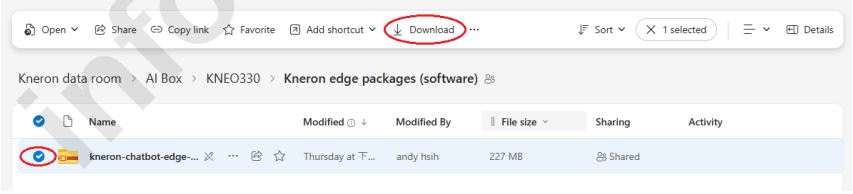

Figure 3-8 KNEO 330 Edge Packages

Likewise, the administrator opens the AI packages file to access the system packages directory. There are 17 system packages, sys_upgrade_1.n.0.aipkg,

where n ranges from 3 to 19. The administrator must first check the installed package in the System Information section, then download all the packages until the final one. If the latest installed system package is sys_upgrade_1.10.0.aipkg, the administrator should download the packages from sys_upgrade_1.11.0.aipkg to sys_upgrade_1.19.0.aipkg. As it is the first official release of the KNEO 330, downloading all the packages takes a long time with some packages exceeding 10Gb.

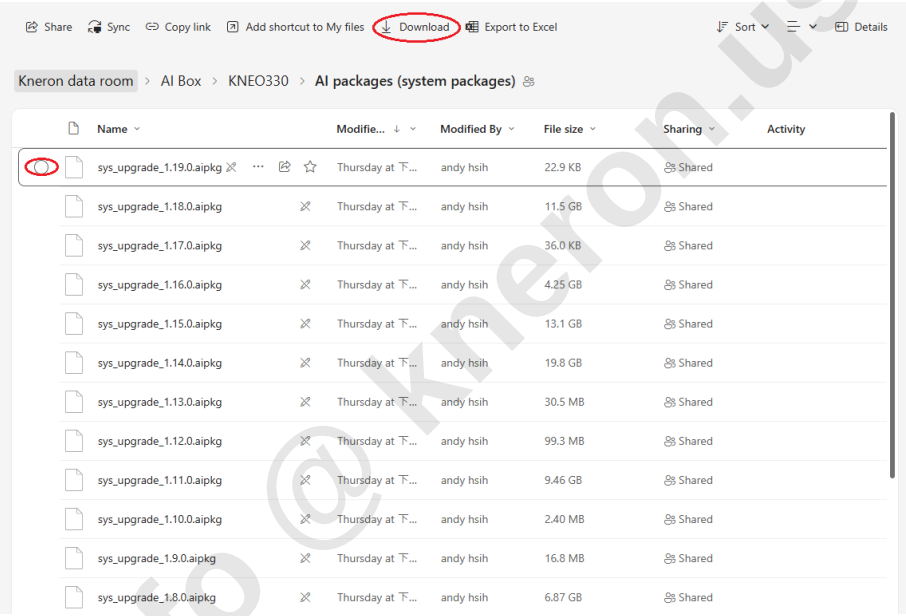

Figure 3-9 KNEO 330 System Packages

3.8 System Update

In a Windows environment, switch to the Linux subsystem and transfer the files from the local computer to the remote KNEO 330. The administrator starts by launching Linux using the command wsl[5] and navigating to the Download

[5] Please search online for the Windows Linux subsystem setup

directory, then, transfers the system and edge packages to the aiuser home directory using the command: scp.

```
C:\> wsl
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.10.102.1-microsoft-standard-WSL2
x86_64)

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

 System information as of Fri Mar  7 14:13:10 PST 2025
```

The administrator copies the files to the KNEO 330 with the syntax: scp <file> aiuser@<ip address>:/home/aiuser (i.e. ip address = 10.200.210.237). It requires the administrator to enter the password: aiuser to initiate the transfer.

```
/mnt/c/Users/oscar/Downloads$ scp -r kneron-chatbot-edge-v0.20.8.pkg
aiuser@10.200.210.237:/home/aiuser
aiuser@10.200.210.237's password:
kneron-chatbot-edge-v0.20.8.pkg                 100%  227MB  36.9MB/s   00:06
/mnt/c/Users/oscar/Downloads$ scp -r sys_upgrade_1.3.0.aipkg
aiuser@10.200.210.237:/home/aiuser
aiuser@10.200.210.237's password:
sys_upgrade_1.3.0.aipkg                         100%   19MB  16.7MB/s   00:01
. . . .
```

3.9 Software Update

The administrator first requests all users to log out of the KNEO 330 before updating the system, then logs in to KNEO 330 to stop the edge and system software by executing the following commands: sudo aiservice stop kneron-edge (edge software) and sudo aiservice stop kneron-backend (system software).

To stop the edge software:

```
aiuser@kneron330:~$ sudo aiservice stop kneron-edge
Service kneron-edge stopped.
```

To stop the system software:

```
aiuser@kneron330:~$ sudo aiservice stop kneron-backend
kneron-backend
kneron-backend
Service kneron-backend stopped.
```

The administrator installs the required system packages according to the current package release. In this section, all system packages are installed[6] from sys_upgrade_1.3.0.aipkg to the latest version, sys_upgrade_1.19.0.aipkg, using the command: sudo aipackage –upgrade <system packages>, then, the edge package is installed with the command: sudo edge-install <edge package> --force

To install the system packages:

```
aiuser@kneron330:~$ sudo aipackage --upgrade
/home/aiuser/sys_upgrade_1.3.0.aipkg
gpg: Signature made Wed 16 Oct 2024 10:46:20 AM PDT
. . . .
Upgraded successfully. Please restart the kneron-backend and kneron-edge
services using aiservice command.
. . . .
aiuser@kneron330:~$ sudo aipackage --upgrade
/home/aiuser/sys_upgrade_1.19.0.aipkg
gpg: Signature made Tue 04 Feb 2025 09:11:44 PM PST
. . . .
Upgraded successfully. Please restart the kneron-backend and kneron-edge
services using aiservice command.
```

The administrator must verify the installation log for the message: Upgrade successfully to confirm that the system packages were installed correctly.

[6] The administrator can skip the current system packages and starts from the next one

To install the edge package

```
aiuser@kneron330:~$ sudo edge-install kneron-chatbot-edge-v0.20.8.pkg --force
gpg: Signature made Wed 26 Feb 2025 10:13:59 PM PST
. . . .
Edge installation completed. Please restart the kneron-backend and kneron-
edge services using aiservice command.
```

Similarly, the administrator must verify the installation log message: Edge installation completed to confirm the edge package was installed successfully.

Once the installation is complete, the administrator reboots the KNEO 330 and initializes the system and edge packages automatically.

```
aiuser@kneron330:~$ sudo reboot
Connection to 10.200.210.237 closed by remote host.
Connection to 10.200.210.237 closed.
```