KNEO 330 EdgeGPT Server Administrator Manual

(v 1.8.7)

Sept 2024

info @ kneron.us

Revision History:

Doc Version	Description	Firmware Version	Author	Date
1.8.7	Initial Release	V0.18.7	Oscar Law	2024/09/11

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 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.

info @ kneron.us

Table of Contents

KNEO 330 EdgeGPT Server Administrator Manual i
1 Introduction 1
2 EdgeGPT Server 2
2.1 Product Overview 2
2.2 Accessories List 4
2.3 Power On 4
2.4 WEBUI Interface 5
2.4.1 Session Initialization 5
2.4.2 Database Path 8
2.4.3 Free Chat Mode 9
2.4.4 Knowledge Base Mode 13
2.4.5 Company Q&A 20
2.4.6 Company Organization 22
3 Server Administration 26
3.1 User Registration 26
3.2 Password Change 27
3.3 Access Permissions 28
3.4 Password Reset 30
4 System Management 31
4.1 User Account 31
4.2 Remote Access 32
4.2.1 SSH 33
4.2.2 PuTTY 34
4.3 System Service 35

info @ kneron.us

4.4 Server History 37
4.5 External Storage 37
4.5.1 USB Drive 38
4.5.2 NAS Storage 41
4.6 System Backup 42
4.7 Database Transfer 43
4.8 System Reboot 43
4.9 System Shutdown 43
5 Appendix 44
5.1 Data Source 44
5.2 Custom Configurations 44

info @ kneron.us

Table of Figures

Figure 2-1 KNEO 330 EdgeGPT Server....2
Figure 2-2 KNEO 330 EdgeGPT Server Interface....2
Figure 2-3 Browser Access....5
Figure 2-4 WEBUI Login Page (English)....6
Figure 2-5 WEBUI Login Page (Chinese)....6
Figure 2-6 WEBUI Session (English)....7
Figure 2-7 WEBUI Setting Menu....8
Figure 2-8 WEBUI Data Path....9
Figure 2-9 Free Chat Mode....10
Figure 2-10 Chat Session....10
Figure 2-11 Chat Session Delete....11
Figure 2-12 Chat Mode Setting....12
Figure 2-13 Local LLM Prompt Update....13
Figure 2-14 Create Custom Database....14
Figure 2-15 Share Custom Database....15
Figure 2-16 Knowledge Base Inquiry....15
Figure 2-17 Merge Custom Database....16
Figure 2-18 Single/Multiple Files or Custom Database Delete....17
Figure 2-19 Knowledge Base Creation Configuration....17
Figure 2-20 Knowledge Base QA Configuration....19
Figure 2-21 Company Q&A Spreadsheet....20
Figure 2-22 Company QA Spreadsheet Upload....21
Figure 2-23 Company QA Inquiry....22

info @ kneron.us

Figure 2-24 Company Organization 22
Figure 2-25 Company Organization Create 23
Figure 2-26 Company Organization Add User 24
Figure 2-27 Company Organization Dashboard 24
Figure 2-28 Select Shared Custom Database 25
Figure 2-29 Access Shared Custom Database 25
Figure 3-1 New User Sign-Up 26
Figure 3-2 New User Registration 27
Figure 3-3 User Password Change 28
Figure 3-4 Dashboard Home Page 28
Figure 3-5 Dashboard Login 29
Figure 3-6 Dashboard User Menu 29
Figure 3-7 User Role Modification 30
Figure 4-1 Kneron Homepage 31
Figure 4-2 Kneron User Login 32
Figure 4-3 Window PowerShell 33
Figure 4-4 PuTTY Home Screen 35
Figure 4-5 USB Drive Properties 38
Figure 4-6 Format the USB Drive with exFAT 39
Figure 5-1 Directory Structure 44

info @ kneron.us

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 image (.pdf) and video subtitle (.srt) 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 User Access:** Allow multiple users to access the machine without performance degradation.
4. **Multiple Media Inputs:** Support additional embedded images (.pdf) and video subtitle (.srt) inputs for the knowledge base creation
5. **Traditional Chatbot Support:** Functions as the traditional chatbot with user-defined answer
6. **Knowledge Base Sharing:** Share the custom knowledge base through the private group
7. **Text Generation:** Besides answering questions, it can generate articles, craft stories, and produce creative content.
8. **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.

info@kneronus

9. **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]

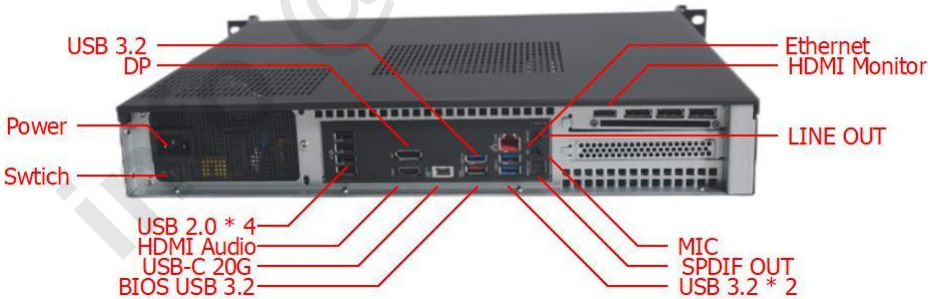

Figure 2-2 KNEO 330 EdgeGPT Server Interface

[1] Use only the top right HDMI port to connect to the monitor for display, while the 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 for 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

info @ kneron.us

[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)

2.2 Accessories List

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

- 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.

2.3 Power On

- 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 IP address shown in inet entry (i.e. 10.200.210.237) 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>
```

info@kneron.us

```
        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
```

2.4 WEBUI Interface

2.4.1 Session Initialization

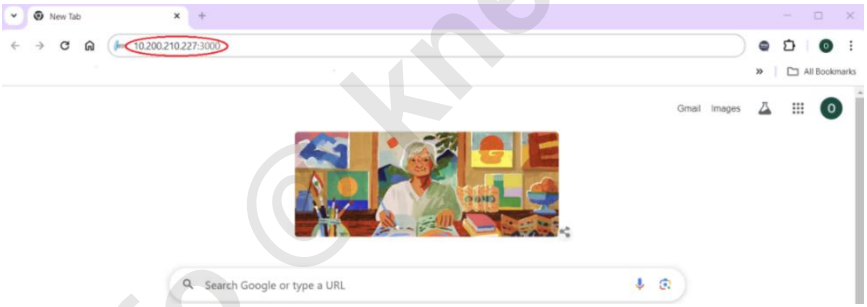

Figure 2-3 Browser Access

The administrator initiates web access using the WEBUI interface. For intranet access, it must employ the prefix: Hypertext Transfer Protocol Secure (i.e. HTTPS)[3], then enters the IP address (e.g., 10.200.210.237) followed by port 3000 in the browser, resulting in the web address (https://10.200.210.237:3000). For internet access, it enters the domain name (e.g., <domain.com>) with port 3000, resulting in the web address (<domain.com>:3000). The administrator must enter https://<web address> (i.e. https://10.200.210.237.3000)

[3] HTTPS is the secure requirements to access the KNEO 330 EdgeGPT server

info @ kneron.us

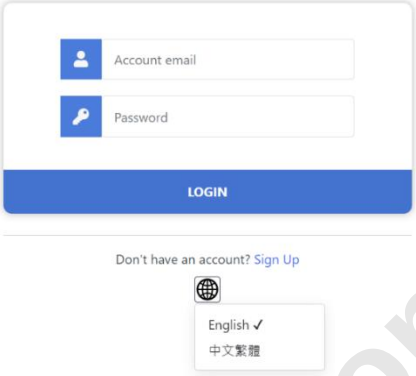

Figure 2-4 WEBUI Login Page (English)

The WEBUI interface is set to English by default, it can click the icon 🌐 to switch to Traditional Chinese. The administrator can log in to the system using the username: admin@kneronchatbot.com with the password: admin123.

Figure 2-5 WEBUI Login Page (Chinese)

info @ kneron.us

Figure 2-6 WEBUI Session (English)

The administrator clicks the menu button in the top-right corner menu button to open the Settings Menu and switches the language between English and Chinese. The administrator first sets the language to English in the Language box, enters the device IP address (e.g., 10.200.210.237) in the Host field, and clicks the ADD DEVICE button. It takes a few minutes for the KNEO 330 to initialize the system. Once initialization is completed, the device IP address appears in the Device List dialog box. The machine is linked with the system, it is no longer required to initialize the machine in the future. The device IP address will be shown in the Device List dialog box during the next login. The administrator can follow the same steps to add additional machines for inquiries, it is strongly recommended to run only one machine during an inquiry session. The chat history can be toggled using the top-left corner history button.

info@kneron.us

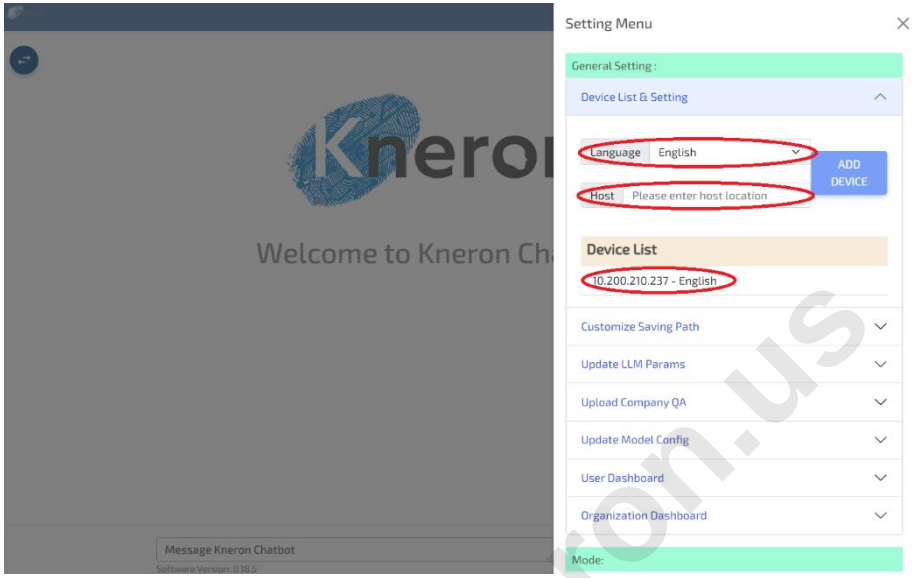

Figure 2-7 WEBUI Setting Menu

2.4.2 Database Path

The databases are stored in the default directory <default directory>. This directory is further divided into two subdirectories: EN (for English) and ZN (for Chinese), which store the language-specific databases based on the system language setting. All databases will be stored in the EN subdirectory if the language is set to English. If the administrator enters the Saving Path as <saving path>, the database is stored in the device with the absolute path <default directory>/content/EN/<saving path>. For the external device (i.e. USB driver and NAS storage)[4], the <saving path> is replaced with the machine data path (i.e. USB driver: /dev/sdb1 and NAS storage: /mnt/kds/data_feed), the absolute path becomes /dev/sdb1/content/EN/ or /mnt/kds/data_feed /content/EN.

[4] Please refer to Section 4.2.1 for detail external device setup

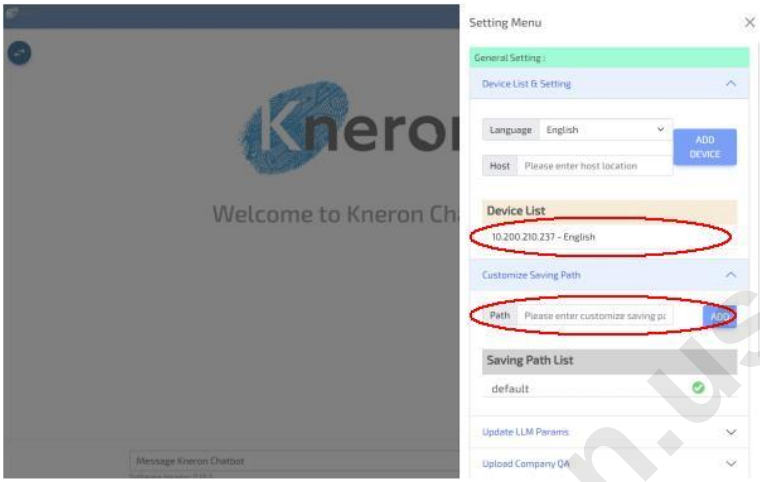

Figure 2-8 WEBUI Data Path

The administrator can set the custom database path using the Path box, followed by the ADD button. It adds the custom data path to the Saving Path List. The administrator can select different data paths using the check mark symbol ⊘ and remove the path using the garbage 🗑 symbol. it is useful to save the database to the other directory or mounted devices.

2.4.3 Free Chat Mode

The KNEO 330 offers two chat modes: Free Chat and Knowledge Base. These modes can be chosen using the chat mode button labeled FREE CHAT and KNOWLEDGE BASE. By default, Free Chat mode is enabled, which is used for general inquiries. The Knowledge Base mode utilizes a custom database that users have created. The administrator can switch between Free Chat and Knowledge Base modes using the buttons in the bottom-right corner of the Setting Menu.

info @ kneron.us

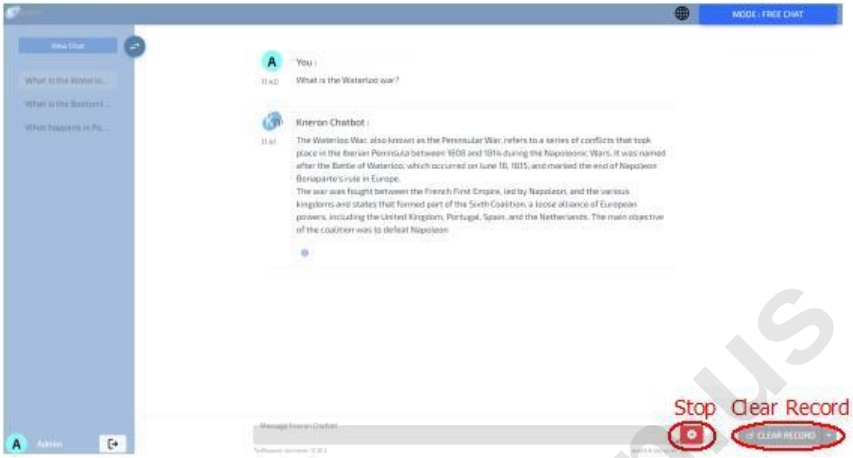

Figure 2-9 Free Chat Mode

For the default Free Chat mode, the user enters the inquiry into the Message Kneron Chatbot box and presses the green arrow key, the response appears in the Dialogue Box. The chat history is shown under the New Chat section. The user can end the chat using the red stop button or clear the inquiry using the CLEAR RECORD button.

Figure 2-10 Chat Session

info @ kneronus

The user can click on the chat tag to reopen a chat session, and the chat tag can be modified by using the edit button. To download a chat session, the user clicks the download button; the session will be saved in JSON format and compressed into the local machine Download directory. Additionally, the user can select a chat session and click the DELETE SESSION button to remove it. A warning message will appear, prompting the user to confirm the deletion of the chat session from the history.

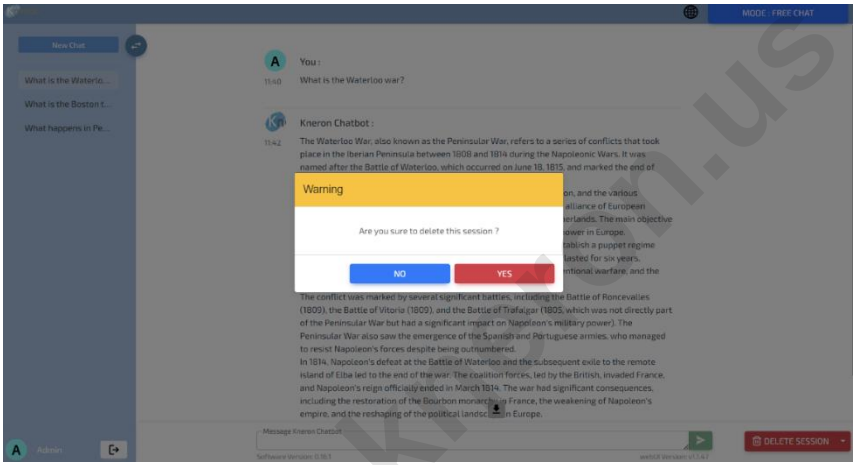

Figure 2-11 Chat Session Delete

The latest software introduces enhanced control over inquiry results through the Update LLM Params feature, as shown in Figure 2-12. It offers two post-processing modes: Top P and Greedy. For Top P mode, the Temperature setting controls output diversity, with lower values producing more conservative and predictable results, while higher values lead to more varied and creative outputs. The temperature range is from 0 to 1. The Top P mode generates random outputs based on the [Top P] slider, which can be adjusted between 0 and 1. Higher values result in more randomness. Additionally, the Repetition Penalty prevents repeated outputs by applying a penalty, adjustable between 1 and 1.1. It recommends maintaining the default settings for general usage. Greedy mode typically generates the same outputs by selecting those with the highest matching probability, the range is also set between 1 and 1.1

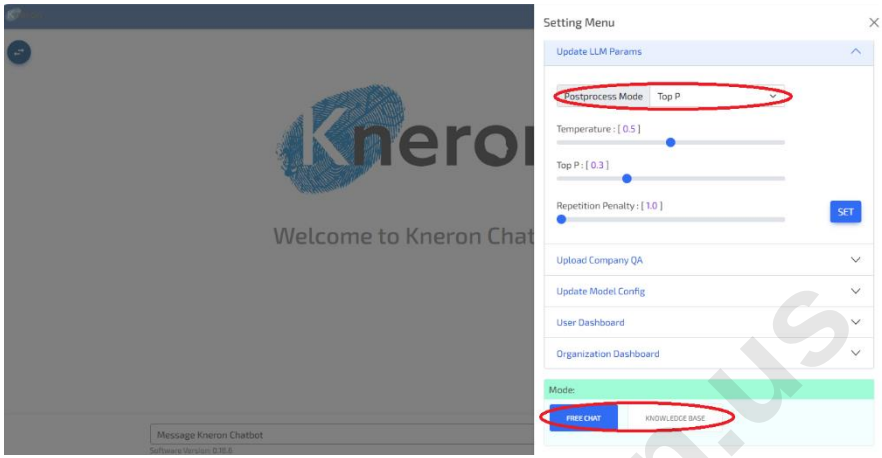

Figure 2-12 Chat Mode Setting

Currently, the KNEO 330 supports advanced Prompt Engineering through the Update Local LLM Prompt feature, which includes the Knowledge Base Q&A Prompt, Table Extraction Prompt, and Page Extraction Prompt options. Knowledge Base Q&A Prompt defines the role of the EdgeGPT server, functioning as a virtual assistant to respond to inquiries. Table Extraction Prompt retrieves the content from the input document table cells. Page Extraction Prompt guides the user in retrieving all the content, including text, tables, and images, from the input document. **Since the accuracy of responses heavily depends on the Prompt settings, please consult the Kneron FAE before making any changes. Otherwise, it is recommended to leave the settings blank.**

info@kneron.us

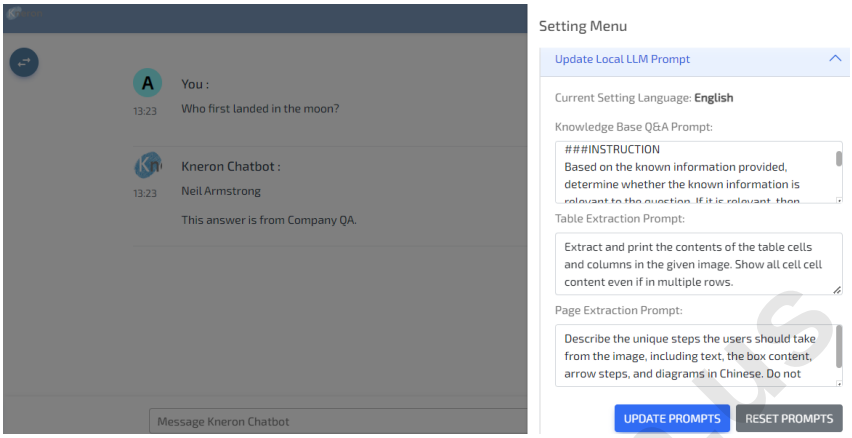

Figure 2-13 Local LLM Prompt Update

2.4.4 Knowledge Base Mode

The administrator creates a custom database using the MANAGEMENT button for the knowledge base inquiry. Upon clicking this button, a management pop-up menu appears. The administrator then enters the database name <user>/<database> (e.g., public/bda602) in the Knowledge Base List box[5]. Next, the administrator clicks the MANAGEMENT button again and uploads files using the Drop files box and the UPLOAD button. The administrator clicks the CONFIRM button on the Confirm Settings page, then the system automatically uploads the files to the system, The KNEO 330 supports multiple file formats, including .txt (text), .pdf (portable document format), .docx (Microsoft Word), .csv (Microsoft Excel), .md (markdown-formatted text), and .zip (compressed files). File names should not contain special characters such as space, (), {}, or []. Depending on the file size, uploading may take a few minutes or longer. The administrator can create databases in both public and user directories, whereas general users can create databases only within their user directory.

[5] The system will display an error message if the database name includes special characters.

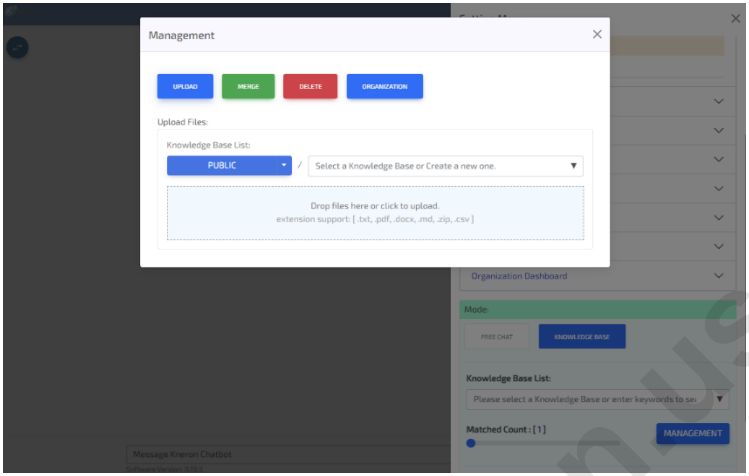

Figure 2-14 Create Custom Database

The user can share a custom knowledge base with others in a private group[6] during the database creation process. Once the database is created, the user clicks the ORGANIZATION button, selects the desired knowledge base from the Knowledge Base List, chooses the appropriate organization from the Selected Organizations menu, and presses the ADD button. This allows the knowledge base to be shared with the private group users. Additionally, the user can remove the knowledge base from the group by using the REMOVE KNOWLEDGE BASE FROM ORGANIZATIONS option. A symbol (+ or x) is associated with the organization name. During removal, the user should ensure the symbol is set to +.

info @ kneron.us

[6] Please refer to Chapter 2.4.6 Company Organization, it describes how to set up the private group and invite the users to join

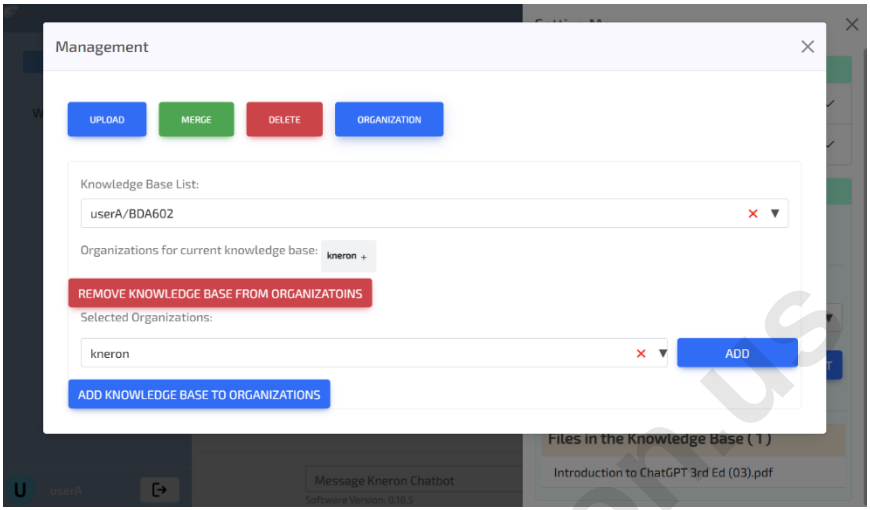

Figure 2-15 Share Custom Database

Once the custom database is created, the administrator clicks the cross symbol in the top right corner and returns to the knowledge base inquiry page. The custom database name appears in the Knowledge Base List box, and loaded files are displayed in the Files in the Knowledge Base box.

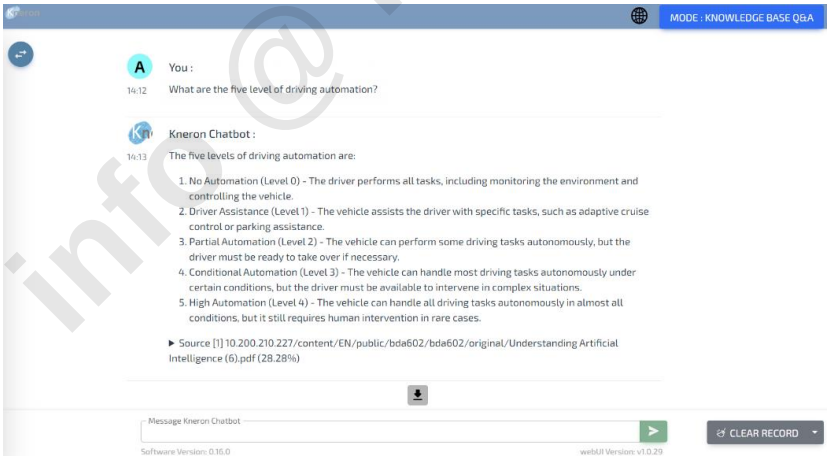

Figure 2-16 Knowledge Base Inquiry

info@kneron.us

The administrator can select different databases from the Files in the Knowledge Base box for inquiry, enter a prompt into the Message Kneron Chatbot, and adjust the Matched Count: [n] sliding bar (where n is 1, 2, 3, 4 or 5) to access more matching results

The pop-up menu offers two additional functions: merge and delete. To merge two databases, it first clicks the MERGE button, a new pop-up menu shows how to merge from one database to another.

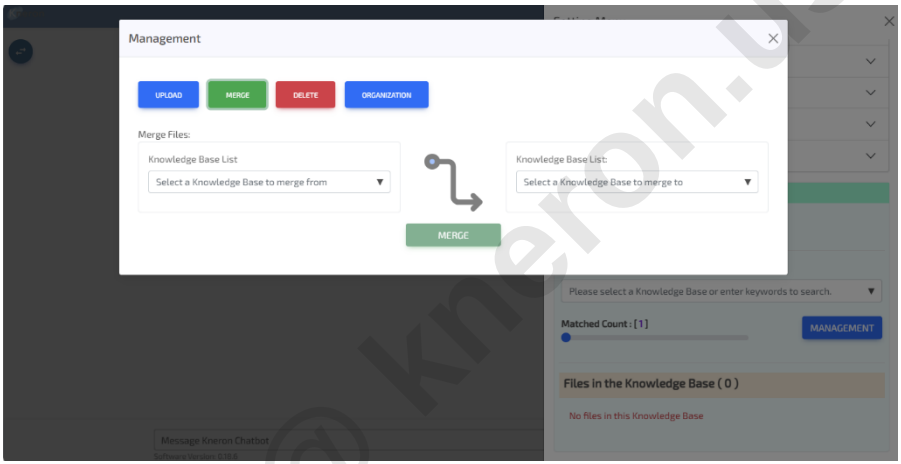

Figure 2-17 Merge Custom Database

To delete files or databases, the administrator clicks the DELETE button. They can select one or multiple files to delete using the DELETE SELECTED FILES button. If no files are selected, the entire custom database will be deleted using the DELETE KNOWLEDGE BASE button, removing it from the file system.

info @ kneron.us

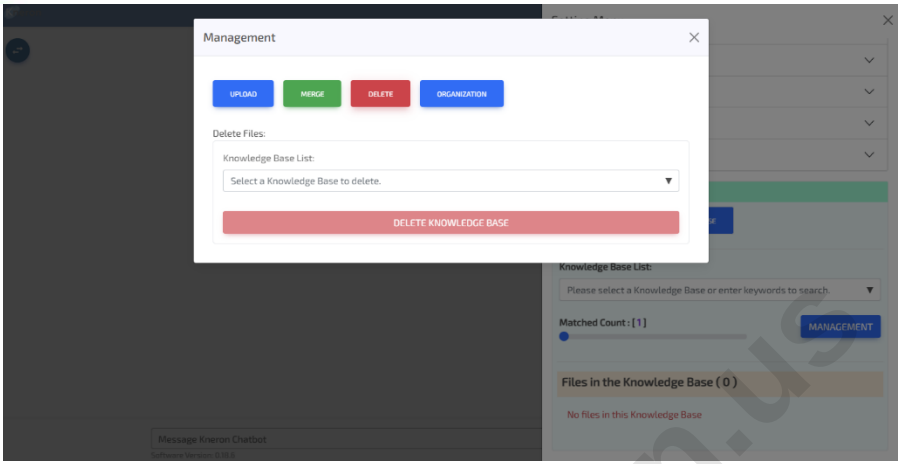

Figure 2-18 Single/Multiple Files or Custom Database Delete

The administrator can set up the model in knowledge base mode through the Update Model Config submenu, which includes two sections: KNOWLEDGE _BASE_CREATION and KNOWLEDGE_BASE_QA.

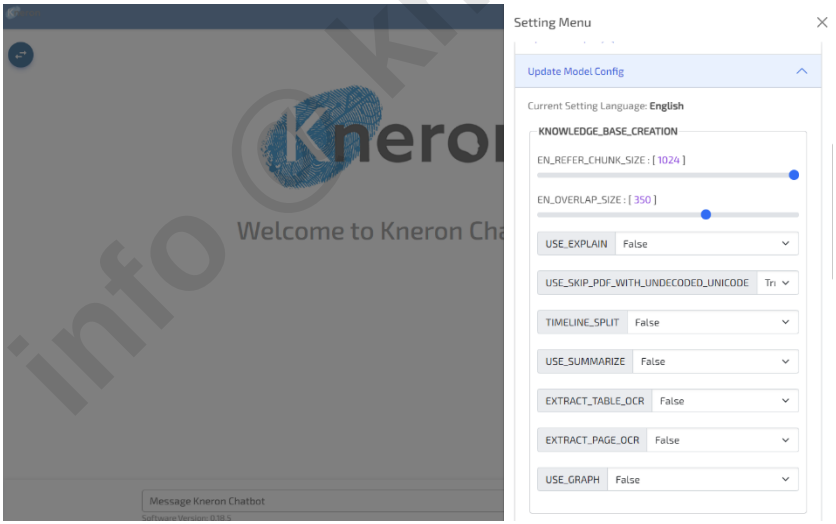

Figure 2-19 Knowledge Base Creation Configuration

info @ kneron.us

The administrator optimizes the performance by adjusting the processing chunk size (number of characters) using EN_REFER_CHUNK_SIZE, which refers to the amount of the processing data ranging from 0 to 512 during the reasoning. The chunk size can effectively manage memory usage and improve processing efficiency. Reducing the chunk size increases accuracy with longer processing time. The processing chunks are linked together through the overlap controlled by EN_OVERLAP_SIZE, which refers to the number of words or phrases shared between two or more text segments during the processing. Overlap can preserve contextual information to achieve better accuracy with more computational resources. The overlap size must be less than 50% of the chunk size.

USE_EXPLAIN connects the keyword to relevant information during knowledge base creation, offering a more detailed explanation when answering questions. USE_SKIP_PDF_WITH_UNDECODED_UNICODE bypasses PDF files with undefined Unicode characters for knowledge base creation and prevents errors during processing. TIMELINE_SPLIT allows the system to process the video SubRip Subtitle (.SRT) file[7], and then build the knowledge base. The .SRT file consists of the index, timestamp, and content for sequential data handling. USE_SUMMARIZE summarizes the uploaded files during the knowledge base creation. When the option is enabled, the user can request a summary of the input file during the inquiry session.

info @ kneron.us

[7] The SubRip Subtitle (SRT) format is a plain-text file format used for video subtitles. It contains a sequence of subtitles, each with an index, start, and end timecodes (formatted as hours, minutes, seconds, and milliseconds), and the corresponding subtitle text with an empty line separating the entries. The timecodes ensure that each subtitle appears at the correct moment in the video. The basic format is listed as below:

```
1
00:00:05,000 --> 00:00:10,000
Hello, welcome to Kneron.

2
00:00:12,000 --> 00:00:15,000
This is an example of an SRT file.
```

EXTRACT_TABLE_OCR extracts the table from the input PDF file and uses its contents to respond to the query. EXTRACT_PAGE _OCR converts the uploaded PDF document into image format and extracts information using OCR. While users can access the data from the embedded images, a downside is that extracting information from images takes more time during the knowledge base creation. USE_GRAPH applies the graph theory approach to link various uploaded PDFs to improve query accuracy with a long processing time during knowledge base creation.

The administrator adjusts the QA threshold using KNOWLEDGE_BASE_QA _THRESHOLD to enhance the matching between inquiries and sources. A higher threshold makes it easier to match inquiries with knowledge base sources with less accuracy drawback. The administrator also modifies the TITLE_THRESHOLD to improve the matching process by using the file names of uploaded knowledge base documents.

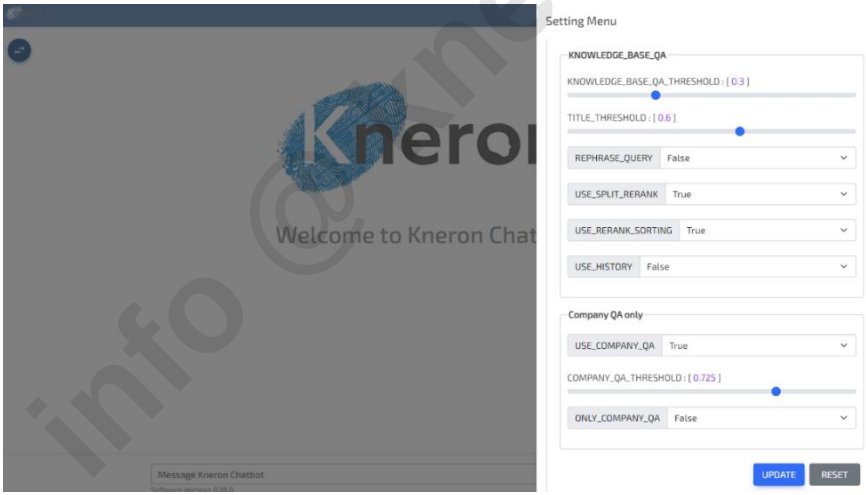

Figure 2-20 Knowledge Base QA Configuration

The administrator enables USE_SPLIT_RERANK by setting it to TRUE, which divides and ranks content within the same document to improve accuracy.

Alternatively, USE_RERANK_SORTING is set to TRUE to sort and rank across different sources with longer processing time. The key difference between USE_SPLIT_RERANK and USE_RERANK_SORTING is that the former focuses on a single source (the same document), while the latter applies to multiple sources (various documents). Moreover, the USE_HISTORY option utilizes past inquiry history to improve the accuracy of current results.

Enabling both USE_SPLIT_RERANK and USE_RERANK_SORTING may slow down operations. The administrator can toggle these options to balance the trade-off between accuracy and inquiry speed.

2.4.5 Company Q&A

	A	B
1	Questions	Answers
2	Who is the first American president?	George Washington
3	Who is the longest British ruler?	Queen Elizabeth II
4	When is the Boston tea party?	December 16, 1773.
5	Which countries are the axis in the World War 2?	Germany, Italy and Japan
6	Who first landed on the moon?	Neil Armstrong
7	Why the first industrial revolution was so important?	The First Industrial Revolution (1760-1840) revolutionized manufacturing, transportation, and communication, laying the groundwork for modern industrial society.
8	What is the nobel prize ?	The Nobel Prize is a prestigious international award recognizing outstanding contributions in Physics, Chemistry, Medicine, Literature, Peace, and Economic
9	Who is the first black American president?	Barack Obama
10	What were the four important inventions in the ancient China?	Paper, Printing, Gun Powder, and the Compass
11	What was the most famous Pyramid?	The Great Pyramid of Giza, also known as the Pyramid of Khufu

Figure 2-21 Company Q&A Spreadsheet

KNEO 330 currently provides a standard one-to-one chatbot feature that allows users to set questions with predefined answers using a spreadsheet in Knowledge Base Mode. The system supports multiple spreadsheets, and the matching is based on the input order. The search engine searches the results from the first spreadsheet to the last one until it finds the answer. An example of the chatbot.csv file is shown in Figure 2-19. The first row includes comments indicating that column A contains

info @ kneron.us

questions and column B contains answers. The actual questions and their corresponding answers begin from the second row onward.

- The comments are defined in the first row (i.e. row 1), which is ignored during the processing
- The questions are defined in the first column (i.e. col A)
- The answers are defined in the second column (i.e. col B)

The administrator can directly load the spreadsheet into the system using the Upload Company QA in Figure 2-6 and drag the spreadsheet into the box, then click the button UPLOAD FILES to upload the spreadsheet to the system.

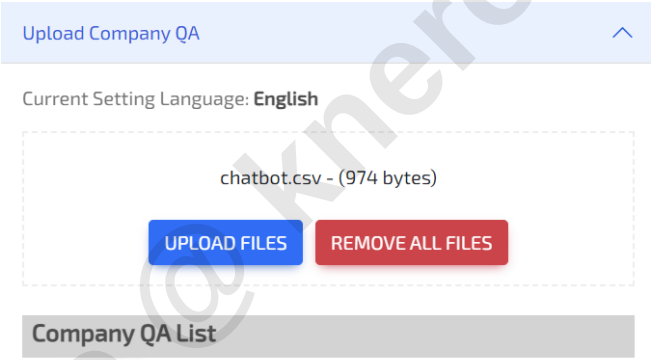

Figure 2-22 Company QA Spreadsheet Upload

The system replies to the inquiry and references to the Company QA spreadsheet shown in Figure 2-19.

info@kneron.us

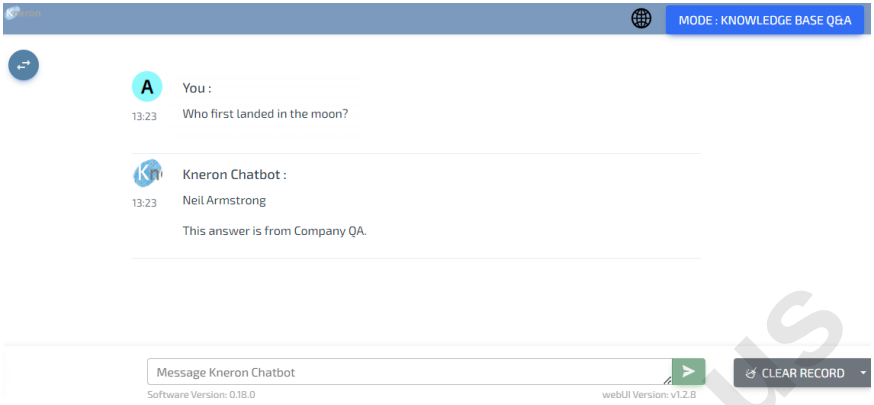

Figure 2-23 Company QA Inquiry

2.4.6 Company Organization

The KNEO 330 enables users to share custom knowledge bases within a private group (called company organization). The administrator first creates the group using the Organization Dashboard in Figure 2-4 and then includes the users in the private group. While all users can access and share the knowledge base, only the creator can modify it.

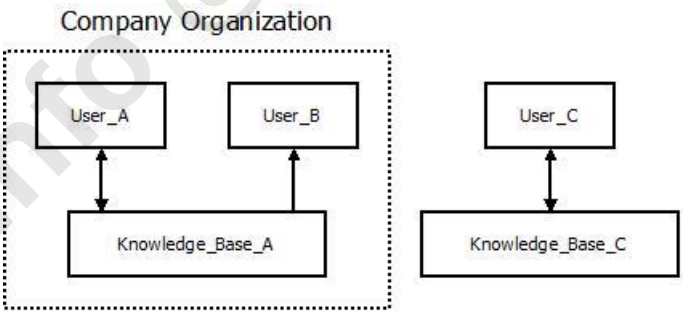

Figure 2-24 Company Organization

info@kneron.us

The administrator clicks OPEN ORGANIZATION DASHBOARD IN NEW TAB in Figure 2-4 and enters the organization name to create the new private group.

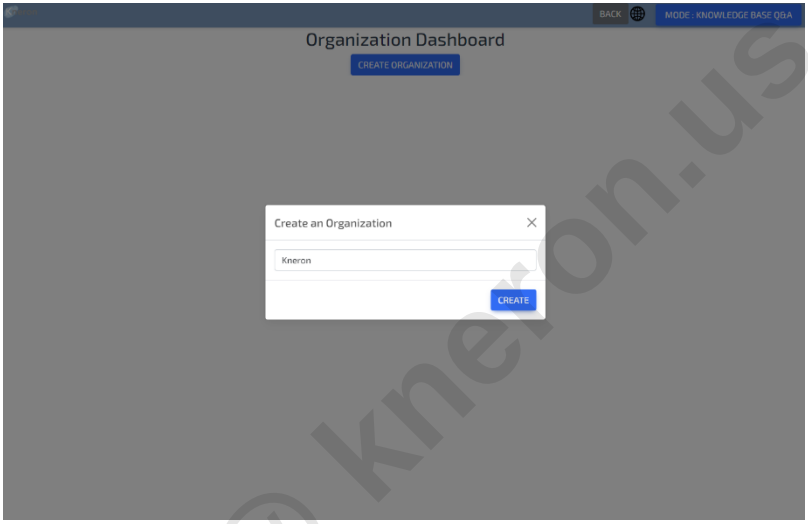

Figure 2-25 Company Organization Create

Upon selecting the new group, the administrator enters the username to invite the user to join the private group.

info@kneron.us

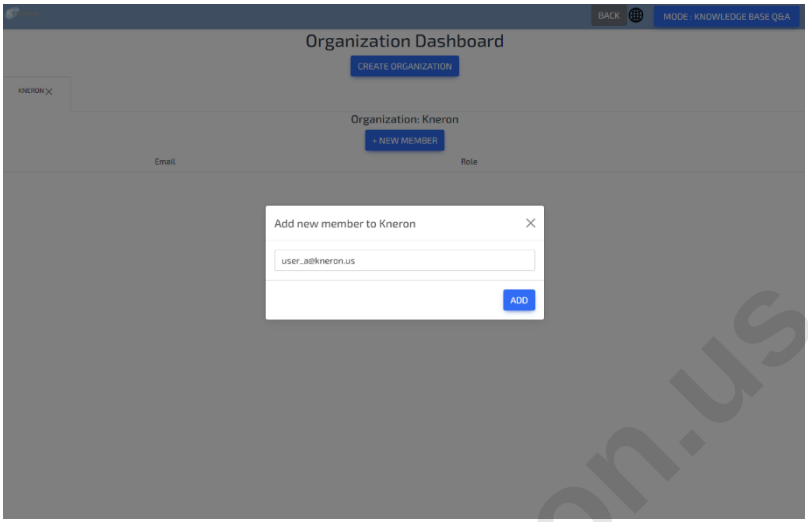

Figure 2-26 Company Organization Add User

The Company Organization Dashboard displays the user information including the email address and the role. The administrator uses the ADD MODERATOR button to assign the role of moderator. The moderator can invite other users to the private group. Only the database creator can modify the knowledge base. The group members can access the information, but not make any changes. The administrator can remove the user from the private group using the REMOVE button.

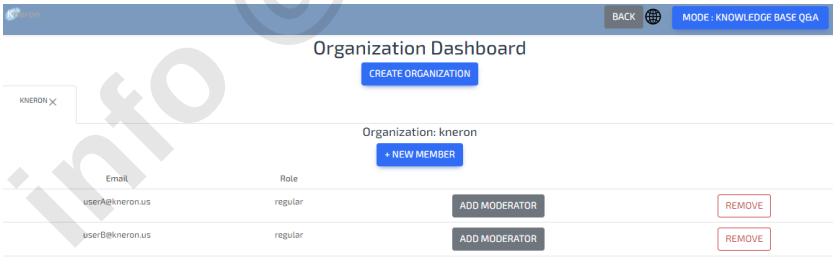

Figure 2-27 Company Organization Dashboard

After the new knowledge base is created, all users in the private group must log out and log back into the system to activate their access rights. Without doing this, other users will not be able to access the new knowledge base

info @ kneron.us

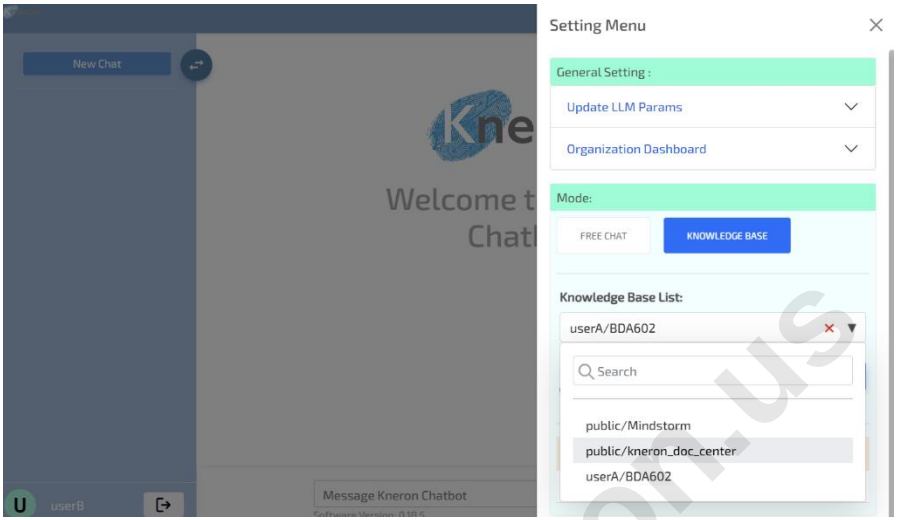

Figure 2-28 Select Shared Custom Database

If another user wants to access the custom knowledge base, they first select the database from the Knowledge Base List and then ask a question from the shared database. The custom knowledge base will not appear in the Knowledge Base List if it is not shareable.

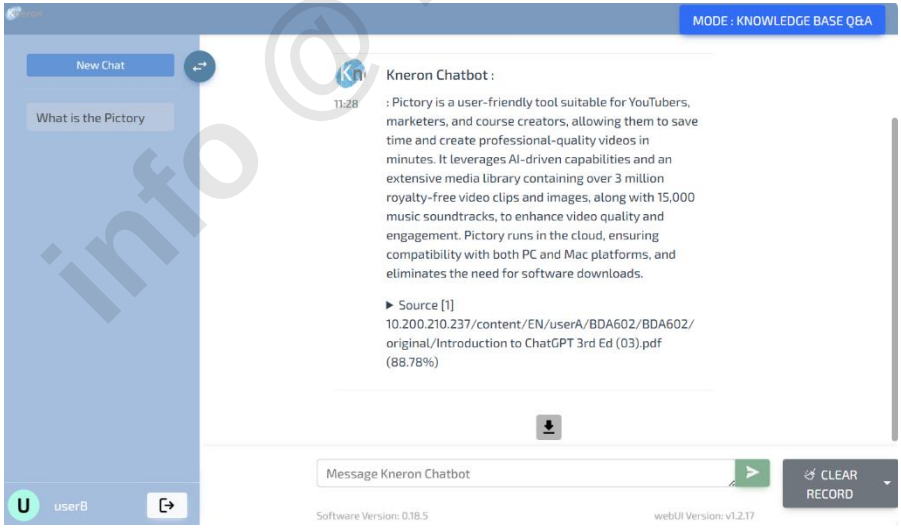

Figure 2-29 Access Shared Custom Database

info @ kneron.us

3 Server Administration

3.1 User Registration

Figure 3-1 New User Sign-Up

The new users can create an account by selecting the Sign Up button on the login screen. In the pop-up menu, the new user enters the personal information, including username, email address, and password. The username can include letters, numbers, and the special characters "." and "_". After completing the form, the user can click the SIGN UP button to register the account. To log in to the KNEO 330, the users must use their email address, not their username. The username will display in the lower left corner after logging in.

info@kneron.us

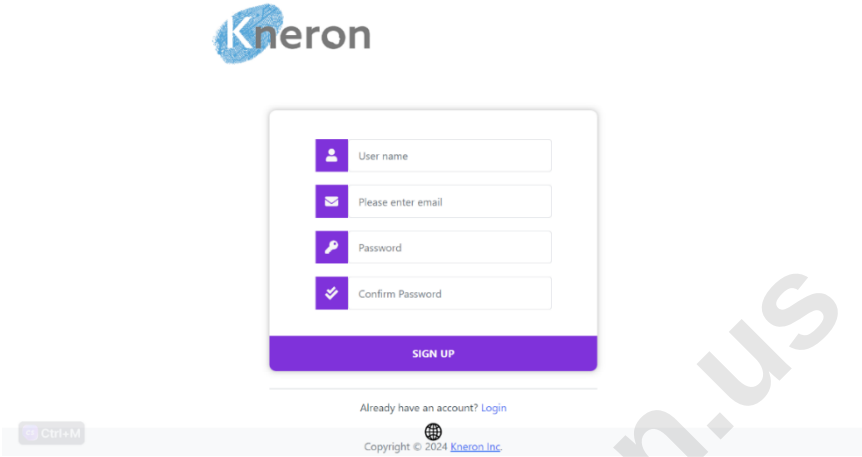

Figure 3-2 New User Registration

To reset the password, the user clicks on the username in the bottom left corner, which opens the User Settings menu. The user then enters their current password, followed by the new password, and clicks the SUBMIT button to complete the change.

3.2 Password Change

To change the password, the user clicks their username in the bottom left corner, which opens the User Setting menu. The user then inputs the current password, followed by the new one, and clicks the SUBMIT button to complete the change.

info @ kneron.us

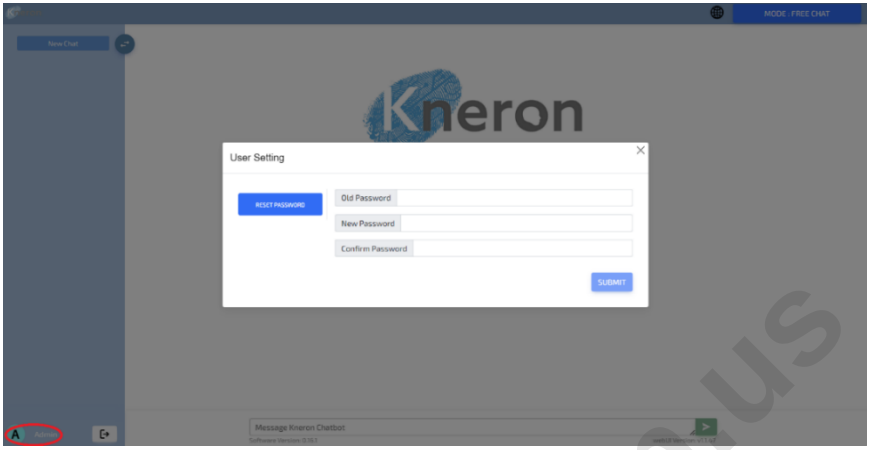

Figure 3-3 User Password Change

3.3 Access Permissions

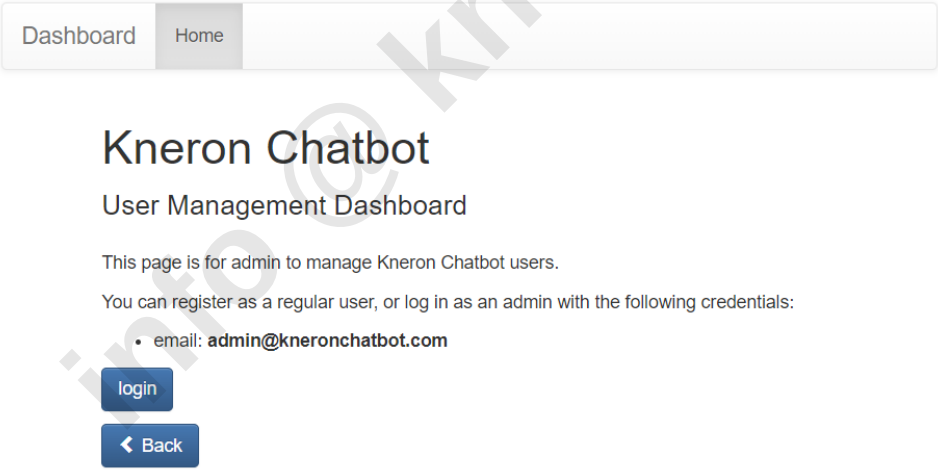

Figure 3-4 Dashboard Home Page

info @ kneron.us

The administrator first opens the user dashboard using the USER DASHBOARD IN NEW TAB in the Setting menu and sets the user access permissions, The system requires the administrator to log in to the dashboard using E-mail Address (admin@kneronchatbot.com) and Password (admin123).

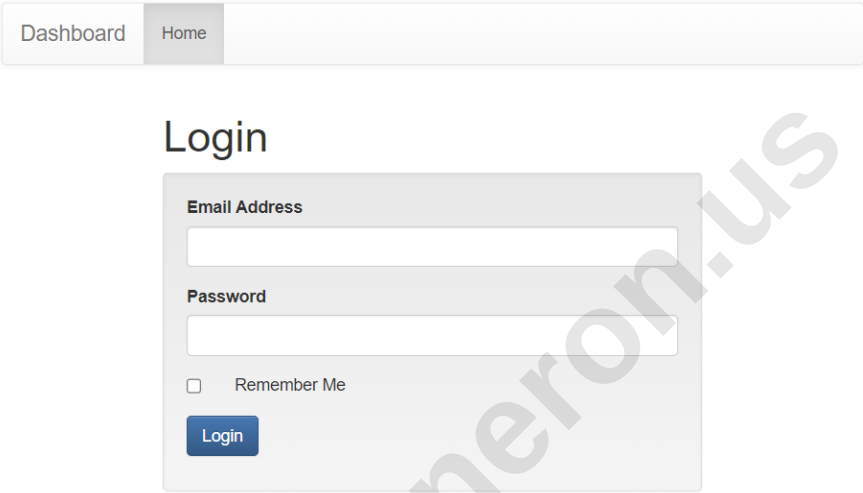

Figure 3-5 Dashboard Login

The dashboard consists of three menus, Home, User, and Role. In the User menus, it shows the user's e-mail, username, last login time, and the encrypted user ID (Fs Uniquifier)

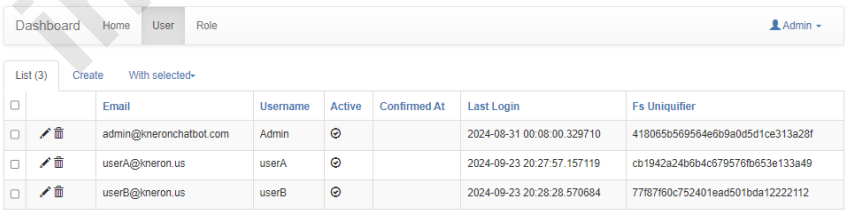

Figure 3-6 Dashboard User Menu

info @ kneron.us

By default, all users are assigned the regular role. The administrator edits the user role by clicking the pen icon and modifying access permissions (admin or regular) in the Role box, then clicking the Save button to apply the changes.

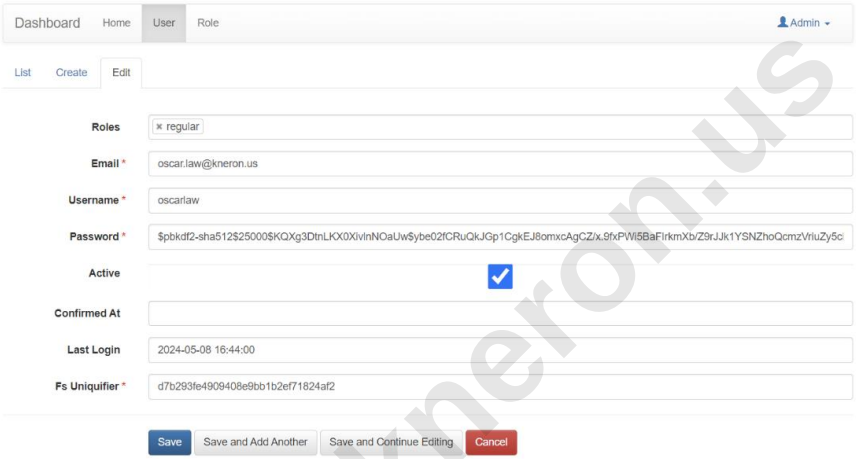

Figure 3-7 User Role Modification

3.4 Password Reset

For security reasons, the administrator cannot reset user passwords. If the user forgets their password, they must contact the administrator to delete the account via the Dashboard. The user re-registers with the same username on the Login Page. The username must match the old one; otherwise, the custom database is denied access

info @ kneron.us

4 System Management

4.1 User Account

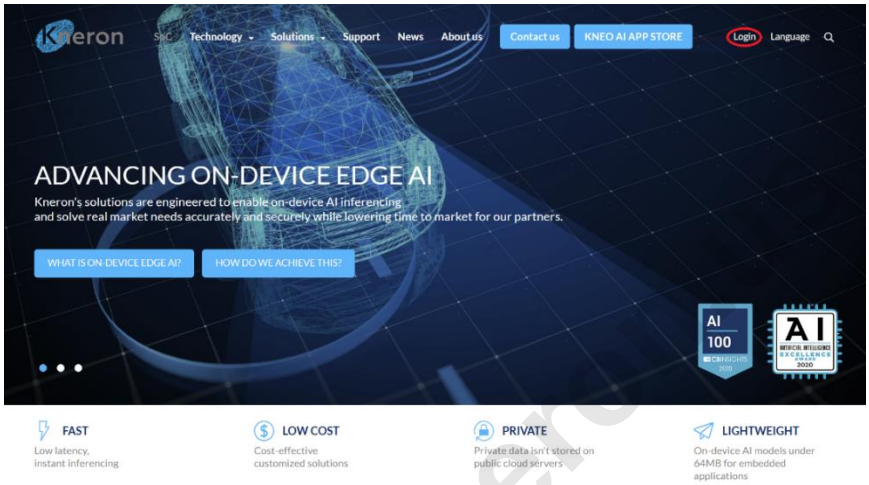

Figure 4-1 Kneron Homepage

The administrator first registers the user account online (https://www.kneron.com) and clicks the Login button in the top right corner. Next, the administrator selects the Create an account option and follows the instructions to complete the account setup.

info @ kneron

Figure 4-2 Kneron User Login

After the user account is set up, the administrator can visit the developer center and access the latest documentation under the Kneron AI chat robot and KNEO330 subdirectory.

4.2 Remote Access

Open the Windows PowerShell Terminal with administrative privileges to access the KNEO 330. Right-click the Windows 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.

info @ kneron

Figure 4-3 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
```

4.2.1 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:
```

info @ kneron.

After logging in, the following message is displayed[8]:

```
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
```

4.2.2 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 putty as shown in Figure 3-2, and enter the IP address (10.200.210.237) along with port number 22.

info @ kneron.us

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

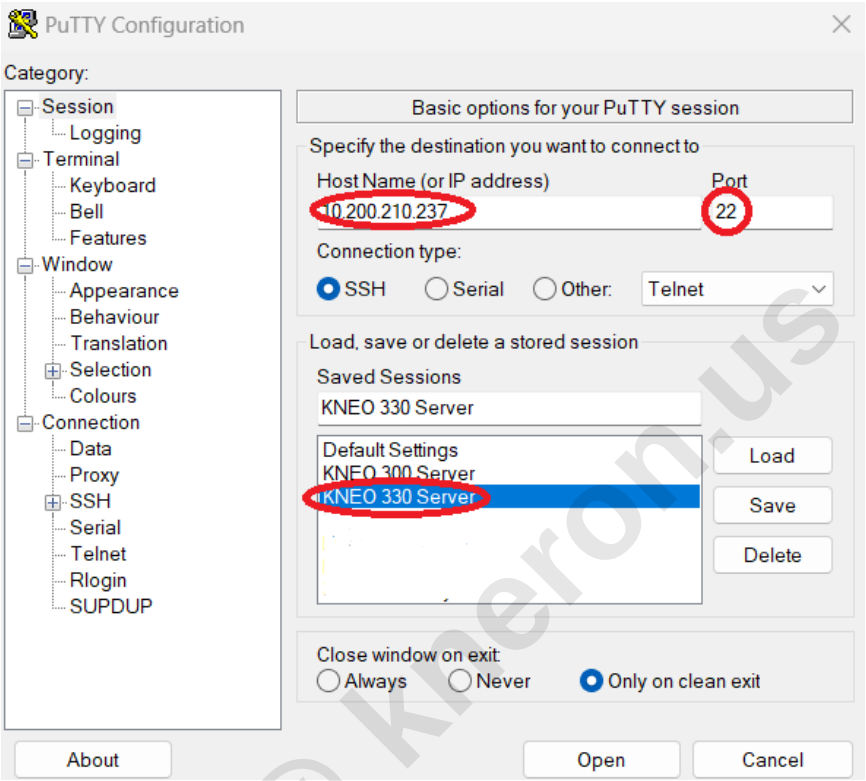

Figure 4-4 PuTTY Home Screen

The 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.

4.3 System Service

The administrator can manage the system service using the command: sudo aiservice <action> <service> where <action> refers to status, start, and stop. <service> sets to kneron-backend or kneron-edge.

To monitor the current server activities:

info @ kneron.us

```
aiuser@kneron330:~$ sudo aiservice status kneron-edge
● kneron-edge.service - Kneron edge server
     Loaded: loaded (/etc/systemd/system/kneron-edge.service; enabled; vendor
preset: enabled)
     Active: active (running) since Tue 2024-09-24 12:33:15 PDT; 20h ago
   Main PID: 2864 (edge-daemon)
      Tasks: 244 (limit: 38223)
     Memory: 13.0G
        CPU: 2h 41min 29.735s
     CGroup: /system.slice/kneron-edge.service
             ├─ 2864 /bin/bash /usr/local/bin/edge-daemon
             ├─ 2878 bash new_launch_release.sh
             ├─ 2924 tclsh /home/kneox/anaconda3/envs/kneron_env/bin/unbuffer
java -jar load_balancer/load_balancer_api/qachat/database_manager/RAG/doc_>
             ├─ 2925 tee -a ./logs/log_2024-09-24_12-33-16.log
             ├─ 2927 java -jar
load_balancer/load_balancer_api/qachat/database_manager/RAG/doc_loader/jars/t
ika-server-standard-nlm-modified-2.4.1_v6.jar
             ├─ 2966 java -Djava.awt.headless=true -cp
load_balancer/load_balancer_api/qachat/database_manager/RAG/doc_loader/jars/t
ika-server-standard->
             ├─ 3025 tclsh /home/kneox/anaconda3/envs/kneron_env/bin/unbuffer
python3 models_edge/app_edge.py --use-GPU
             ├─ 3026 tee -a ./logs/log_2024-09-24_12-33-16.log
             ├─ 3027 tclsh /home/kneox/anaconda3/envs/kneron_env/bin/unbuffer

. . . .


EDGE_LOG - DEBUG - Server is ready:  True
Sep 25 09:24:12 kneron330 edge-daemon[3026]: 2024-09-25 09:24:12,927 -
EDGE_LOG - DEBUG - status: {'init_status': True, 'language': 'dynamic'}
Sep 25 09:24:12 kneron330 edge-daemon[3026]: 10.200.210.237 - - [25/Sep/2024
09:24:12] "POST /models_edge/flsk_kcb_check_status HTTP/1.1" 200 -
```

To activate the service

```
aiuser@kneron330:~$ sudo aiservice start kneron-edge
Service kneron-edge started.
```

To stop the service

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

info@kneronus

4.4 Server History

The administrator can display the server history using the command: sudo ailog <action> where action refers to show or clear

To show the server history

```
aiuser@kneron330:~$ sudo ailog --show | more
Latest AI log: log_2024-09-25_09-45-50.log
INFO  [main] 09:45:51,365 org.apache.tika.server.core.TikaServerProcess
Starting Apache Tika 2.4.1 server
INFO  [main] 09:45:51,424 org.apache.tika.server.core.TikaServerProcess
loading resource from SPI: class org.apache.tika
.server.standard.resource.XMPMetadataResource

. . . .

Start the Kneron Chatbot WebUI Service :
 - https://10.200.210.237:3000/
 - WebUI version: v1.2.17
```

To clear the server history, the administrator must stop the server from using aiservice

```
aiuser@kneron330:~$ sudo aiservice stop kneron-edge
Service kneron-edge stopped.
aiuser@kneron330:~$ sudo ailog --clear
Warning: This operation will clear all the AI logs. Also please stop the edge
service before clearing the logs.
Continue?(Y/N) Y
Chatbot log cleared.
```

4.5 External Storage

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

info@kneron.us

4.5.1 USB Drive

The KNEO 330 supports USB drives formatted with 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.

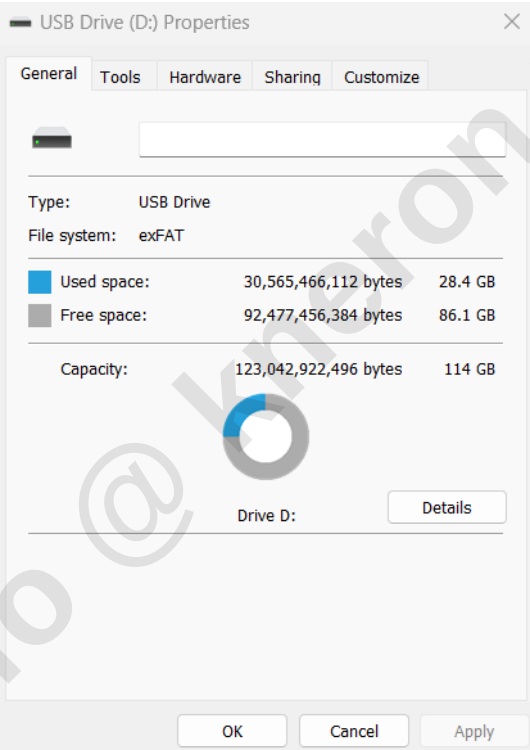

Figure 4-5 USB Drive Properties

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

info @ kneron.us

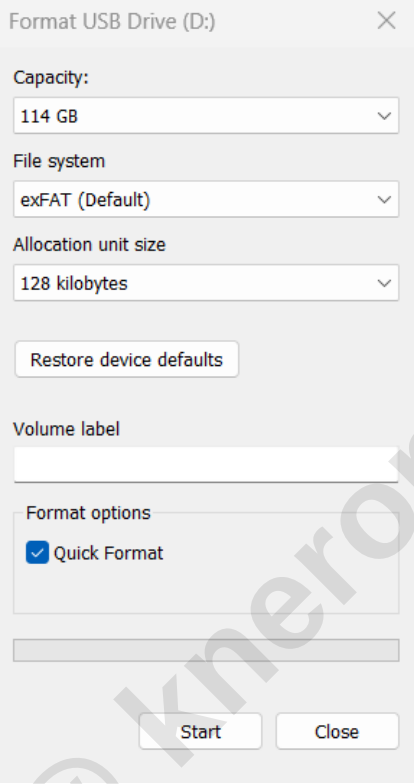

Figure 4-6 Format the USB Drive with exFAT

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
```

info @ kneron.us

```
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
```

info@kneron.us

```
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:
```

4.5.2 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
```

info@kneron.us

```
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:
```

4.6 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) using 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 location 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
```

info@kneron.us

4.7 Database Transfer

The KENO 330 offers two methods for transferring the knowledge base to another machine. Once the administrator has backed up the knowledge base using NAS storage, the storage can be remounted on a different KENO 330 to restore the database.

The administrator can transfer the database to other KNEO 330 using the USB drive. It mounts the USB drive in another KENO 330 and then restores the database from the USB drive. The database can be transferred to remote machines using Dropbox or OneDrive[9]. The USB drive is first plugged into Windows, then uploads the database to Dropbox or OneDrive. The database can be downloaded to other remote machines from Dropbox or OneDrive.

4.8 System Reboot

Before rebooting the system, all the users log out from the system, and the administrator initializes the reboot command sudo reboot to hardware reset the system.

4.9 System Shutdown

Before turning off the system, it is recommended that all users log out first and the knowledge base is backed up to an external USB drive or NAS storage. The administrator then initiates the shutdown command sudo poweroff to ensure the data is properly saved. Avoid disconnecting the power before the software fully shuts down, as this could damage the file system.

info @ kneron.us

[9] Please search online information to upload/download the data from the Dropbox and OneDrive.

5 Appendix

5.1 Data Source

The administrator can access the source of the uploaded files from the backup file: chatbot_database.tar.gz. The backup file is first uncompressed, and the source of the uploaded files is located in the following directories:

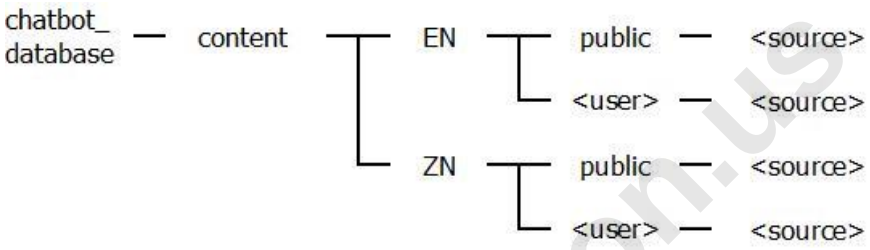

Figure 5-1 Directory Structure

5.2 Custom Configurations

The administrator can customize the KENO 330 using the user_filter_config.json file, which divides into three sections. The first section defines the standard query responses, the second handles sensitive query responses, and the last section maps incorrect names to their correct ones.

```
{
  "standard_query_answer": [
    ["what is kneron doc center", "Kneron doc center provides documents for
Kneron toolchain, etc."]
  ],
  "sensitive_query_answer": [
    ["火藥", "根據當地法律規定，道德或涉及敏感內容，我們無法提供這個問題的答案"],
    ["gunpowder", "According to local laws, ethics, or sensitive content, we
are unable to provide an answer to this query"],
    ["成人圖", "根據當地法律規定，道德或涉及敏感內容，我們無法提供這個問題的答案"],
          ["porn pics", "According to local laws, ethics, or sensitive
content, we are unable to provide an answer to this query"]
  ],
  "standard_names": {
    "後麵": "後面",
    "皇後": "皇后",
```

info @ kneron.us

```
    "麵對": "面對",
          "Alot": "A lot",
          "Irregardless", "Regardless",
          "Anyways", "Anyway"
  }
}
```

info @ kneron.us

info @ kneron.us