KNEO 330 EdgeGPT Server User 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 User Manual i
1 Introduction 1
2 EdgeGPT Server 2
2.1 Product Overview 2
2.2 Accessories List 3
2.3 WEBUI Interface 4
2.3.1 Session Initialization 4
2.3.2 User Registration 5
2.3.3 Password Change 6
2.3.4 Free Chat Mode 7
2.3.5 Knowledge Base Mode 9
2.3.6 Company Organization 13

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....4
Figure 2-4 New User Sign-Up....5
Figure 2-5 New User Registration....6
Figure 2-6 User Password Change....6
Figure 2-7 Free Chat Mode....7
Figure 2-8 Chat Session....8
Figure 2-9 Chat Session Delete....8
Figure 2-10 Chat Mode Setting....9
Figure 2-11 Create Custom Database....10
Figure 2-12 Share Custom Database....11
Figure 2-13 Knowledge Base Inquiry....12
Figure 2-14 Merge Custom Database....12
Figure 2-15 Single/Multiple Files or Custom Database Delete....13
Figure 2-16 Company Organization....13
Figure 2-17 Select Shared Custom Database....14
Figure 2-18 Access Shared Custom Database....15

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 @ Kneron.us

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]

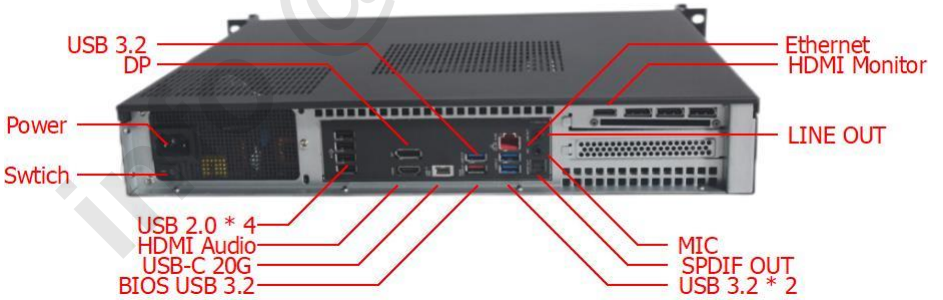


Figure 2-2 KNEO 330 EdgeGPT Server Interface

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

2.2 Accessories List

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

[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.

2.3 WEBUI Interface

2.3.1 Session Initialization

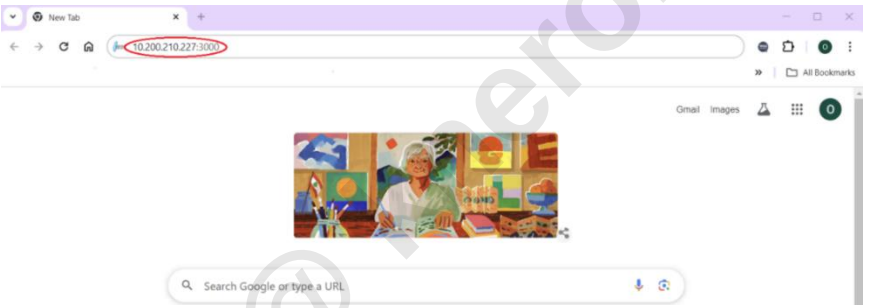


Figure 2-3 Browser Access

The user 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 user 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

2.3.2 User Registration

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.

Figure 2-4 New User Sign-Up

info @ kneron.us

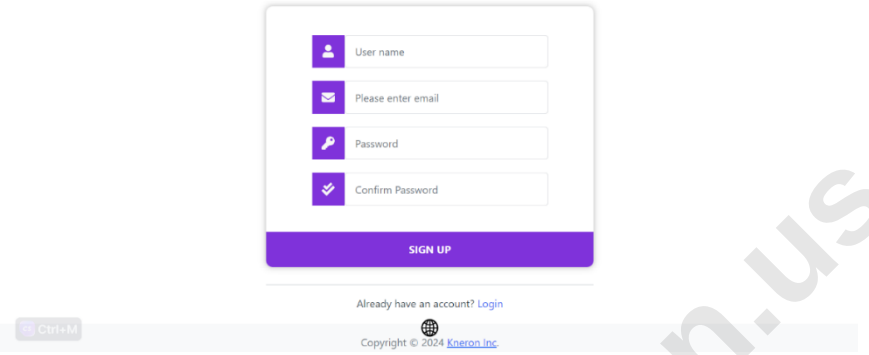


Figure 2-5 New User Registration

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.

2.3.3 Password Change

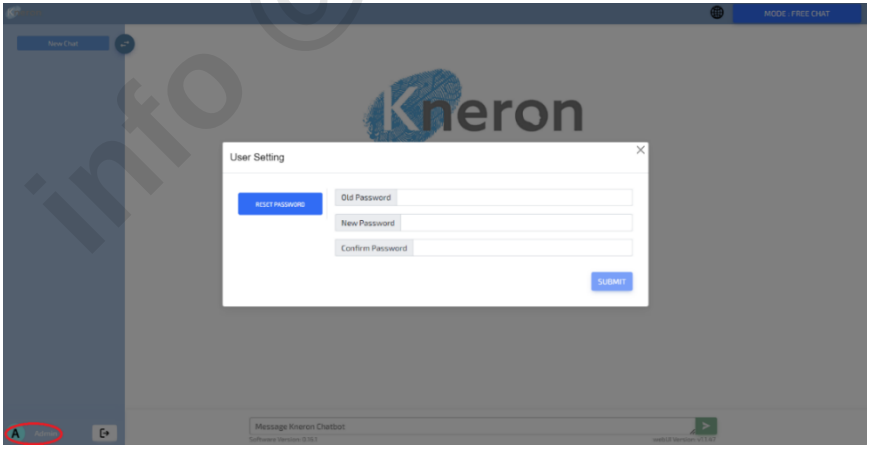


Figure 2-6 User Password Change

info @ kneron.us

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.

2.3.4 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 user can switch between Free Chat and Knowledge Base modes using the buttons in the bottom-right corner of the Setting Menu.

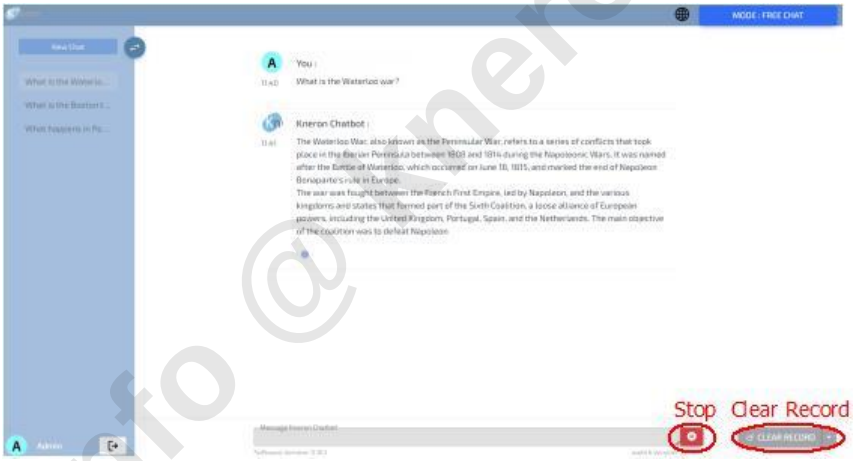


Figure 2-7 Free Chat Mode

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.

info @ kneron.us

Figure 2-8 Chat Session

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.

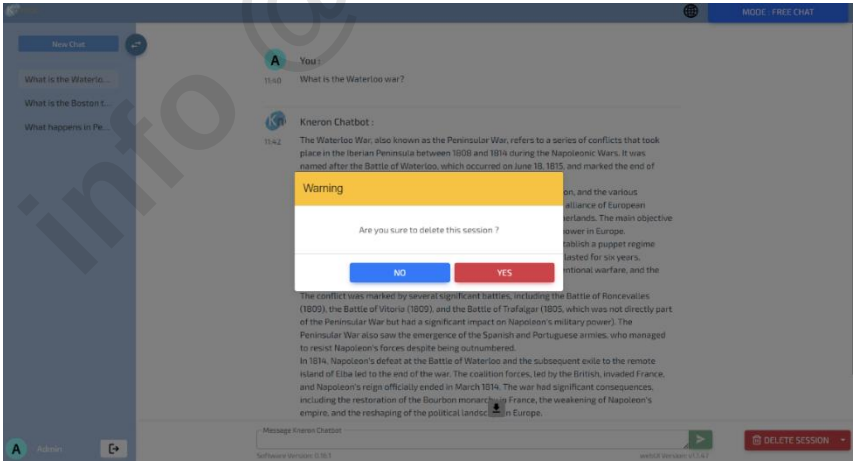


Figure 2-9 Chat Session Delete

info @ kneron.us

The latest software introduces enhanced control over inquiry results through the Update LLM Params feature, as shown in Figure 2-10. 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

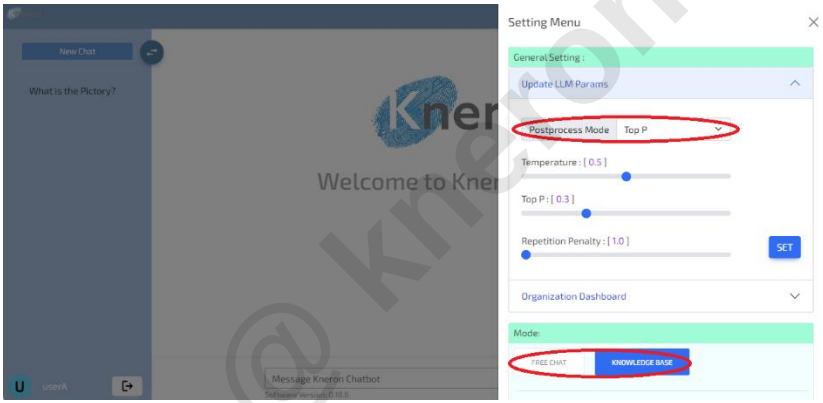


Figure 2-10 Chat Mode Setting

2.3.5 Knowledge Base Mode

The user creates a custom database using the MANAGEMENT button for the knowledge base inquiry. Upon clicking this button, a management pop-up menu appears. The user then enters the database name <user>/<database> (e.g., public/bda602) in the Knowledge Base List box[4]. Next, the user clicks the MANAGEMENT button again and uploads files using the Drop files box and the UPLOAD button. The user clicks the CONFIRM button on the Confirm Settings

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

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 user can create databases in both public and user directories, whereas general users can create databases only within their user directory.

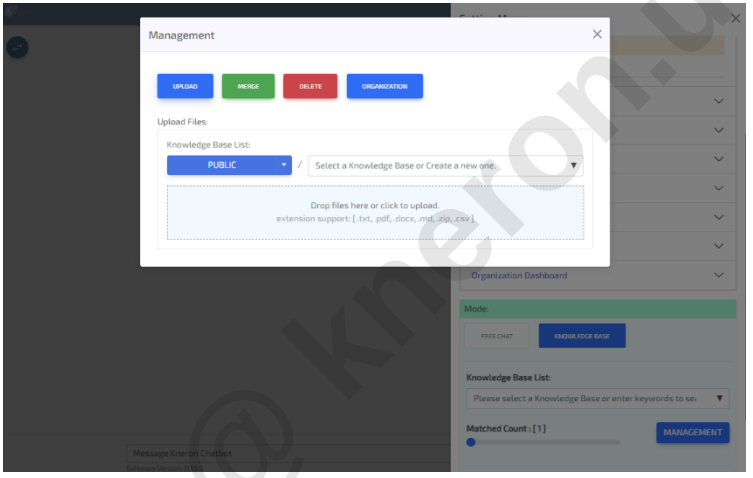


Figure 2-11 Create Custom Database

The user can share a custom knowledge base with others in a private group[5] 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

[5] Please refer to Chapter 2.3.6 Company Organization, it describes how to set up the private group and invite the users to join

info@kneron.us

FROM ORGANIZATIONS option. A symbol (+ or x) is associated with the organization name. During removal, the user should ensure the symbol is set to +.

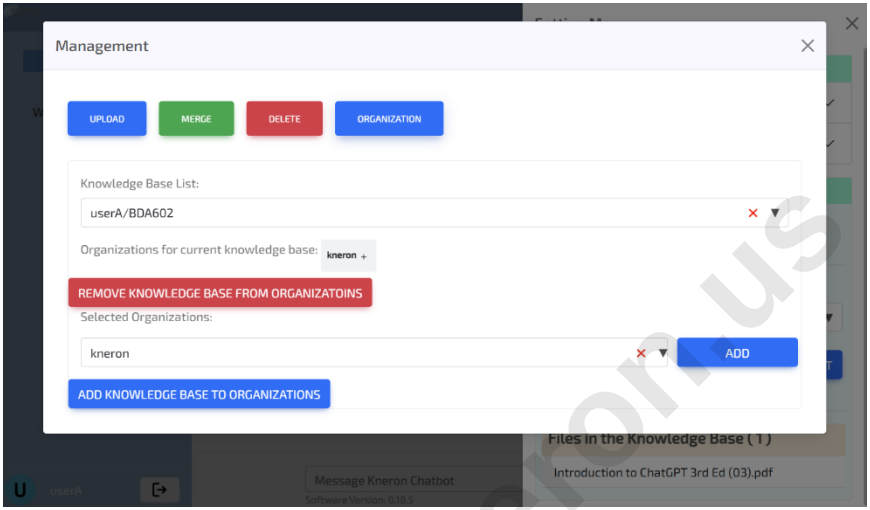
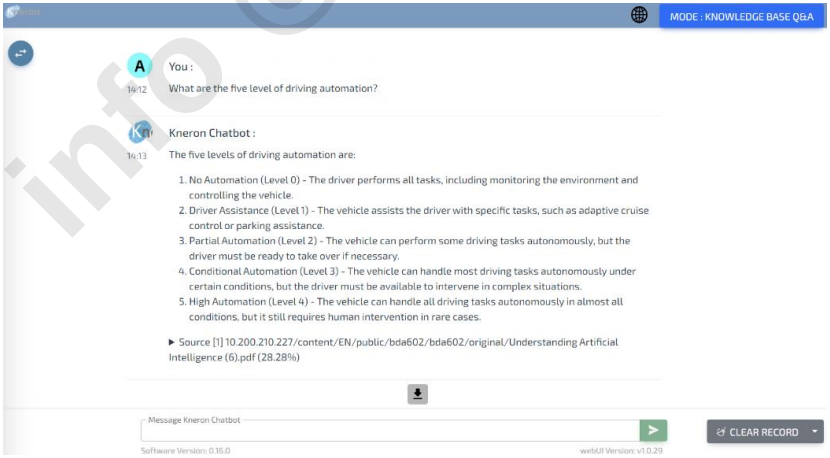


Figure 2-12 Share Custom Database

Once the custom database is created, the user 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.



info @ kneron.us

Figure 2-13 Knowledge Base Inquiry

The user 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.

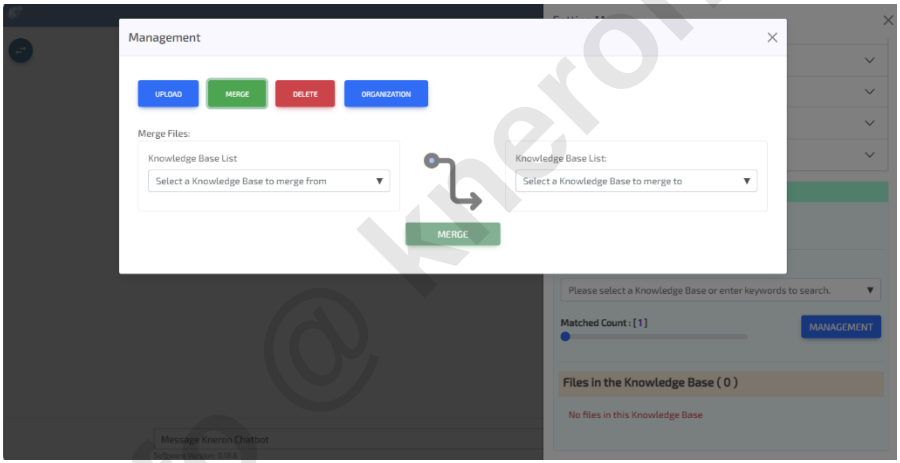


Figure 2-14 Merge Custom Database

To delete files or databases, the user 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

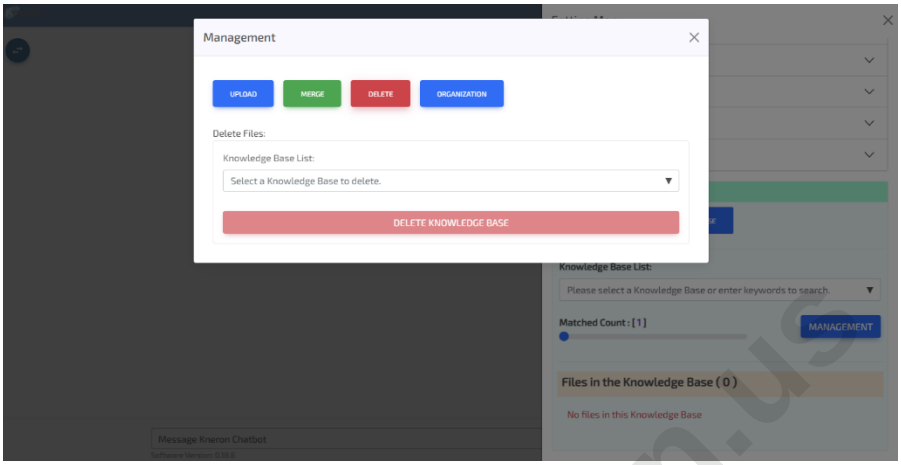


Figure 2-15 Single/Multiple Files or Custom Database Delete

2.3.6 Company Organization

The KNEO 330 enables users to share custom knowledge bases within a private group (called company organization). The administrator creates a private group and invites the users to join the group. The administrator can grant the modulator privilege to the user to invite another user to join the group. While all users can access and share the knowledge base, only the creator can modify it.

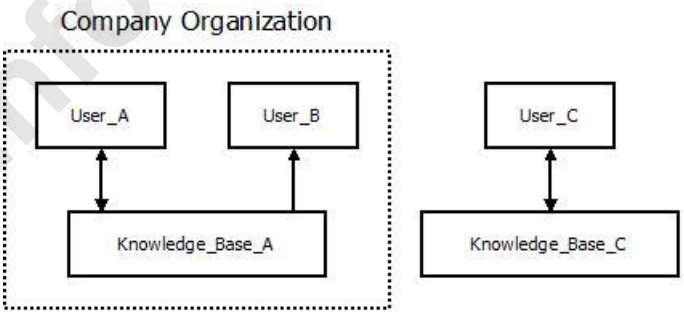


Figure 2-16 Company Organization

info @ kneron.us

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

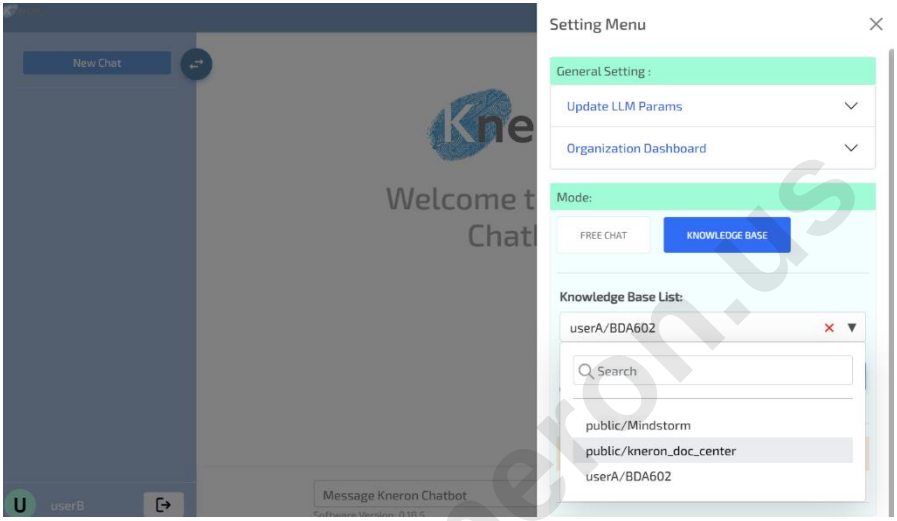


Figure 2-17 Select Shared Custom Database

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.

info @ kneron.us

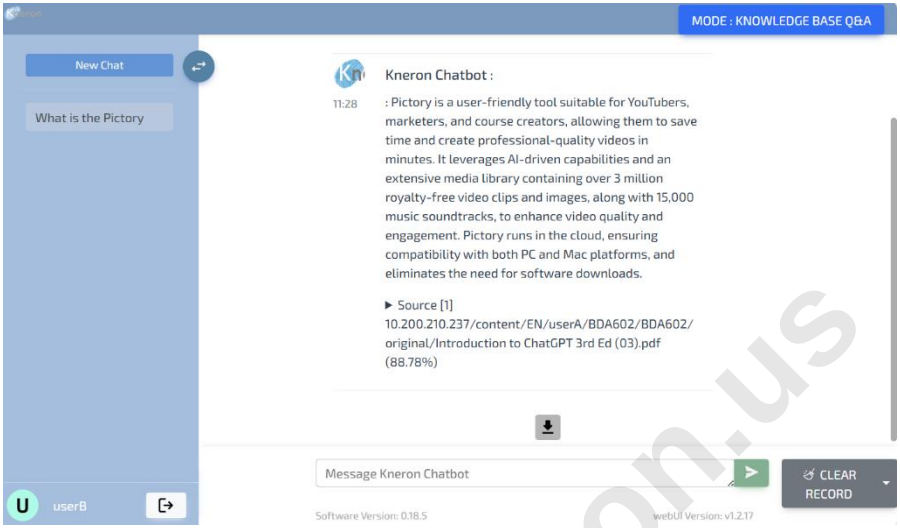


Figure 2-18 Access Shared Custom Database

info @ kneron.us

info @ kneron.us