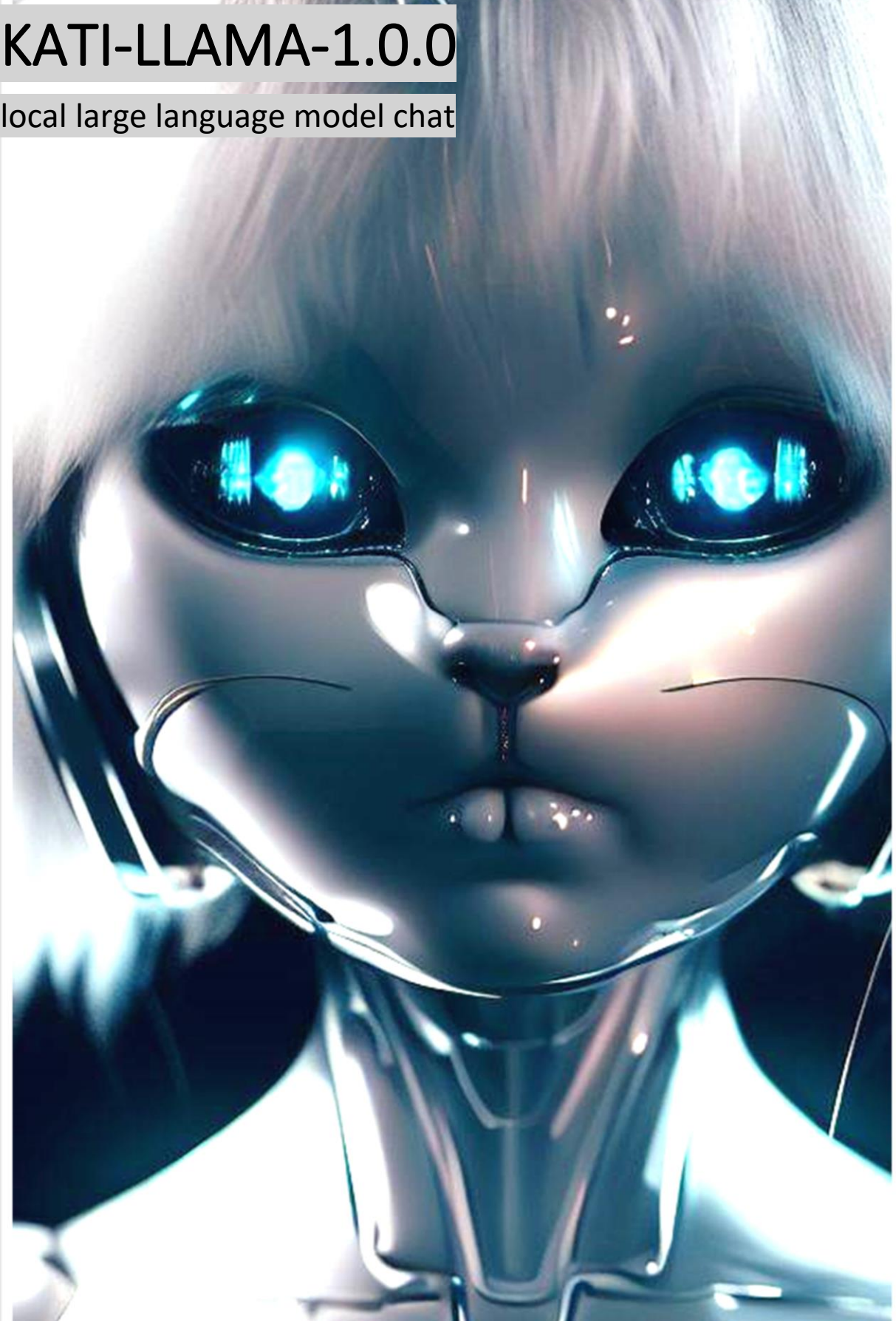
KATI-LLAMA-1.0.0

local large language model chat



CONTENT

ABOUT KATI-LLAMA

KATI-LLAMA is an interface for chatting with Large Language Models on a private PC. The Language Model can be downloaded automatically in the settings and then used offline.

The KATI application allows the user to communicate with an AI in a human-like manner. The AI's responses can be output with a natural voice and the AI's avatar image changes appearance depending on the chatbot's mood. Below is a summary of the features of KATI-LLAMA.

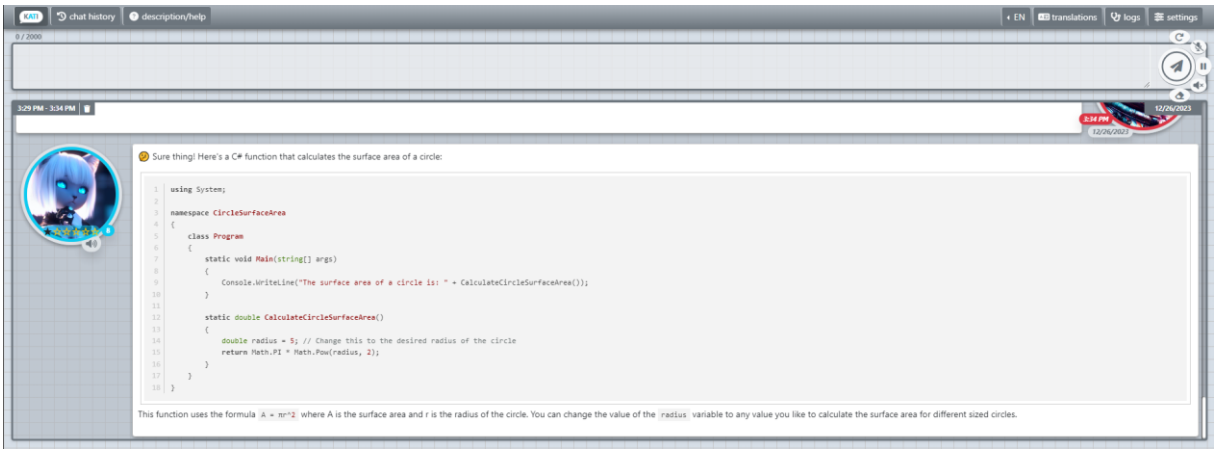
Features:

- Talk to AI without an internet connection
- Optional voice output with a voice pre-installed in the operating system or a natural-sounding TikTok voice. (The TikTok voice requires an internet connection)
- Voice input (System Speech or Whisper)
- Dynamic avatar images to represent AI emotions.
- Chat history with filter function and read-aloud function.
- Rating function for AI responses as an aid to the filter function
- Reduce wait times by streaming responses directly. (If the read-aloud function is active, the output only happens when the sentence is complete)
- Text and code are formatted for better readability.
- Multilingual user interface
  (DE, EN, FR, ES, PT, JA, KO)

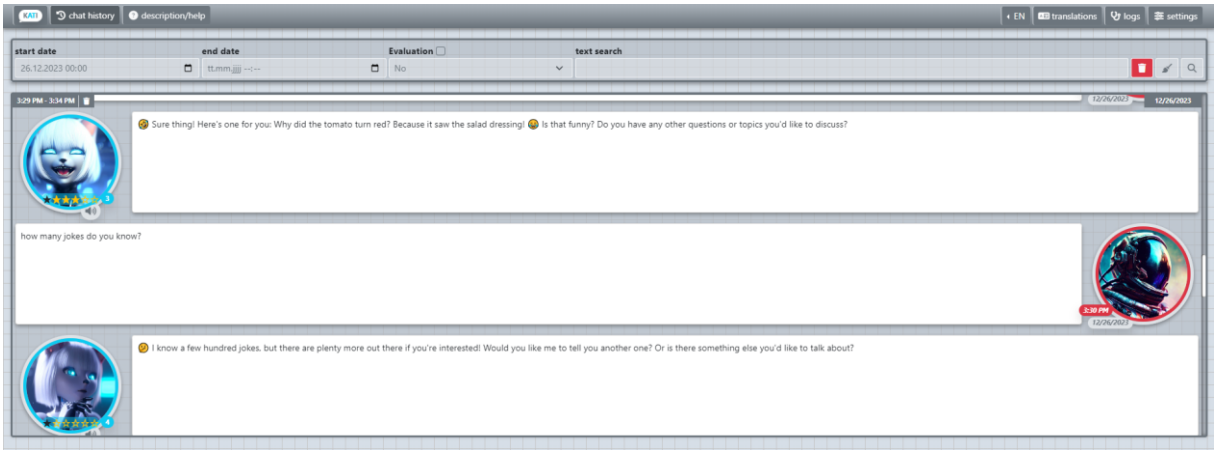
NUGET PACKAGES/LICENSES USED:

- LLamaSharp (MIT)
- ElectronNET.API (MIT)
- Express (BSD-3-Clause)
- LiteDB (MIT)
- Microsoft.AspNetCore.SignalR.Client (MIT)
- NAudio (License Info)
- Newtonsoft.Json (MIT)
- System.Data.SQLite (public domain)
- System.Linq.Async (MIT)
- System.Speech (MIT)
- SoundTouch (License Info)
- WhisperNet (MPL-2.0)

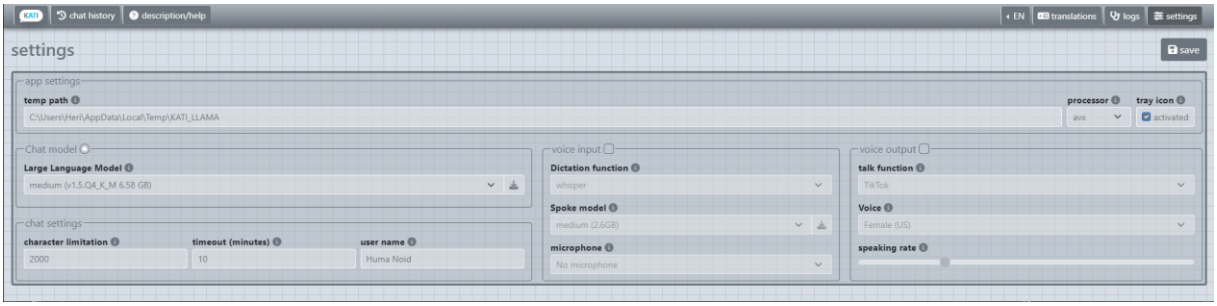
PREVIEW: CHAT VIEW



PREVIEW: CHAT HISTORY VIEW



PREVIEW: SETTINGS VIEW



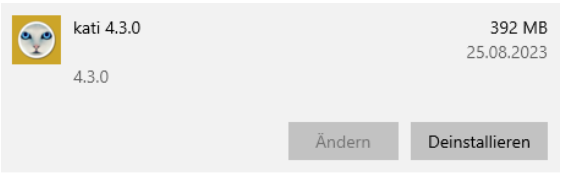
DOWNLOAD AND INSTALLATION

The following variants of KATI are available for download on GitHub

The portable version can be executed directly without installation. However, it takes a little longer to launch this app than with the installed version. The installation is started by clicking on the kati-setup.exe. The app is stored in the following directory:
C:\Users\...\AppData\Local\Temp\KATI_LLAMA

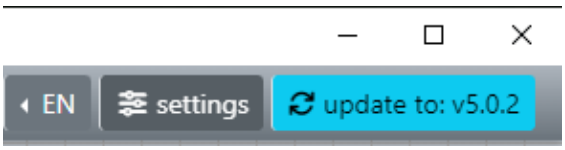
DEINSTALLATION

As with all Windows programs, the uninstallation can be done under "Apps and Features".



UPDATES

For a new version, a button with the version name will appear in the header of the app.



The update can be started either in the app by clicking on the version button or by re-downloading the current version on GitHub. The latter is a little faster.

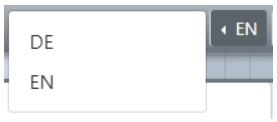
Bug fixes can be identified by the last digit, new features by the second digit and completed milestones by the first digit.

CONFIGURING THE LANGUAGE

The user interface and voice output of KATI can be used in the following languages:

- German (DE)
- English (EN)
- French (FR)
- Spanish (ES)
- Portuguese (PT)
- Japanese (JA)
- Korean (KO)

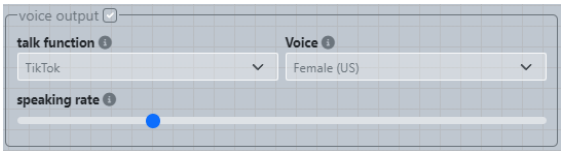
The language can be changed in the header menu after launching the app. The user manual is only available in German and English.



The text-based chat can also be conducted in other languages. Just ask the AI if it understands the language you want and tell it that you want to converse in that language if it can. Please note that the read-aloud function is only possible in the 7 languages listed above.

CONFIGURING NARRATOR

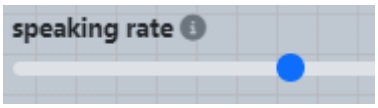
In the language settings, the speaking voice and the speaking speed can be adjusted.



Narrator is disabled by default and can be activated in the Settings View with the checkbox highlighted in blue. Alternatively, you can also switch the voice output directly in the chat window with the speaker button.

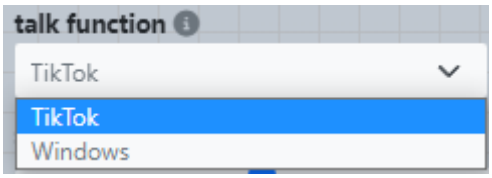
For a comprehensible speech output, the user interface must be set in the same language in which the conversation is to be conducted!

The speed for the speech output is initially preset to 2x.

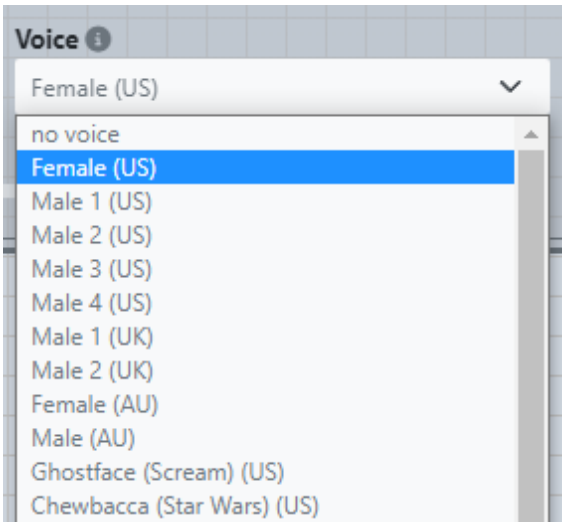


If the rate of speech is too fast, this can be changed to level 1x. (Don't forget to save)

KATI can use the standard feature of Windows or the TikTok feature for voice output.



The TikTok function has a slightly more natural pronunciation than the Windows function and also has more voices to choose from, depending on the language you set. The TikTok voice cannot be used offline.



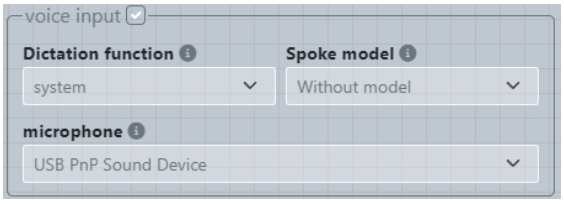
When the language is changed, the TikTok feature is automatically preselected with the first voice available in the language.

CONFIGURING VOICE INPUT

The language-to-text function can be toggled in the chat settings or on the main page next to the submit button.

KATI Supports System Speech and Whisper for voice recognition.

System Speech is a "Speech To Text" (STT) functionality built into Windows. If no microphone can be found, or if there is no speech recognition in Windows for the currently selected language, the feature cannot be enabled. In this case, a hint text will be displayed.

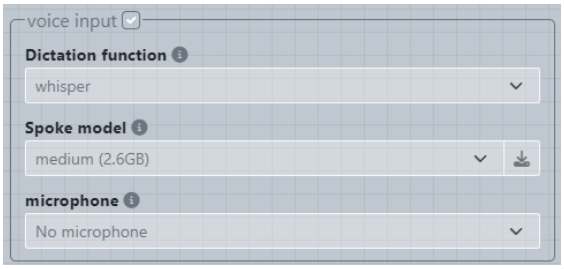


The native SST feature works with a default microphone configured in Windows. In the microphone dropdown, only the currently connected default microphone is suggested.

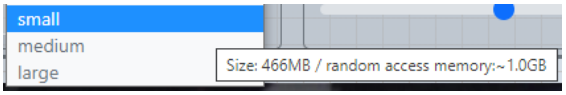
Depending on the microphone, voice or dialect, the untrained, Windows-owned STT unfortunately cannot provide satisfactory accuracy in word recognition. This feature is only intended to serve as a case back if AI-assisted speech recognition does not work.

KATI does not support voice commands in chat when using System Speech!

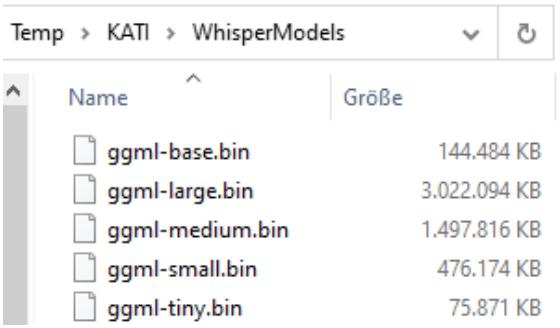
Depending on the model you choose, Whisper has a much more reliable voice recognition.



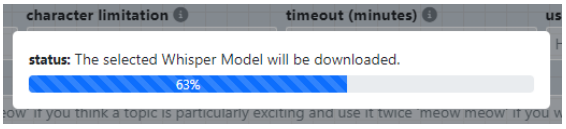
The larger the model, the better the speech recognition. When making your choice, consider the RAM consumption. This is displayed as a tooltip for the respective models.



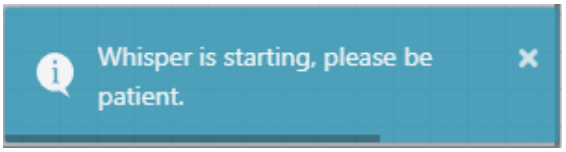
If possible, at least the second largest model (ggml-medium.bin) should be used to have relatively good speech recognition.

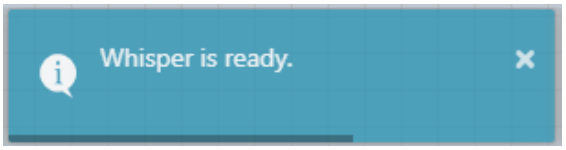


If the selected model does not yet exist, it can be automatically downloaded to the *KATI_LLAMA/WhisperModels* directory using the download button . This may take a few minutes, depending on the model's size. During this time, the navigation elements in KATI are locked.



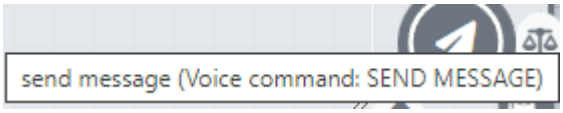
Launching Whisper takes a little more time than system speech. As soon as the following message is displayed, the microphone can be used.





Whisper requires more modern processors. So, it may happen that the SST feature with Whisper is not supported on older PCs. In this case, an error message will appear instead of the top message. In the event of a compatibility problem, the already mentioned System Speech function can be used.

When Whisper is configured, voice commands can be used. These are displayed as a tooltip above the navigation elements.

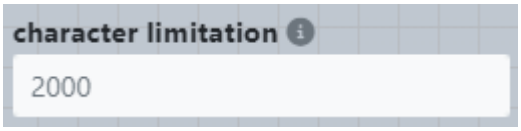


The following commands can be used:

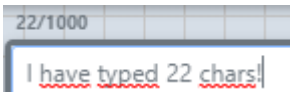
- **SEND MESSAGE:**
  Sends the message
- **RESET MESSAGE:**
  Deletes the input
- **REMOVE LAST SENTENCE:**
  Deletes the last sentence
- **STOP CONVERSATION:** Canceling the Output
- **RESET CONVERSATION:**
  Starts a new conversation

CHAT SETTINGS

The character limit for text input is primarily used to ensure that the AI has to evaluate less text. The shorter and more specific the text input, the more accurate the answers. The input is initially limited to 2000 characters. Here, if necessary, a higher value can be configured.



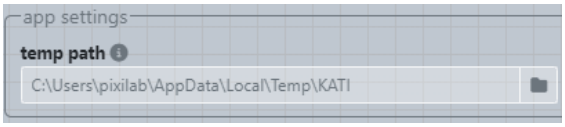
The configured character limit is also displayed in the character counter above the chat input. The text in the input field is automatically truncated if it exceeds the configured length.



CONFIGURING THE STORAGE PATH

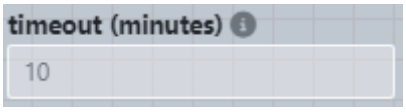
The temporary folder is intended for temporary files from KATI. Currently, the database for the chat history and the language models are stored in this folder.

The chat history is preserved even after an update of the application. However, the app settings will be reset to the default settings after an update. The suggested storage path can be left unchanged or an alternate path can be configured. (Don't forget to save)

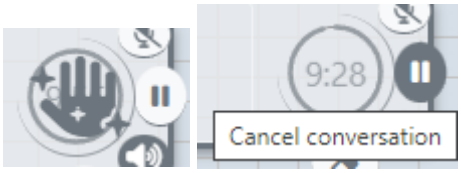


CONFIGURING AI TIMEOUT

The AI answers are unlimited and can theoretically degenerate into an endless monologue. The chat request is automatically canceled after 10 minutes if the AI doesn't want to stop talking. This limit can be adjusted in the settings.

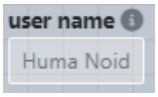


The chat request can also be cancelled manually by clicking on the Submit button again or on the pause button.



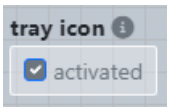
CONFIGURATION OF THE SALUTATION NAME

In the "Username" field, you can set which name the AI should address you with. By default, the name "Huma Noid" is configured. You can also save an alternate name. Keep in mind, however, that this setting will be reset after updating the app.

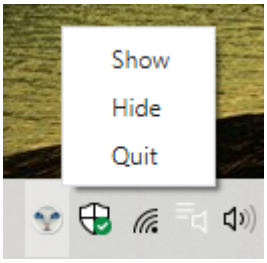


CONFIGURING THE KATI WINDOW

The KATI application can optionally be minimized into the tray.

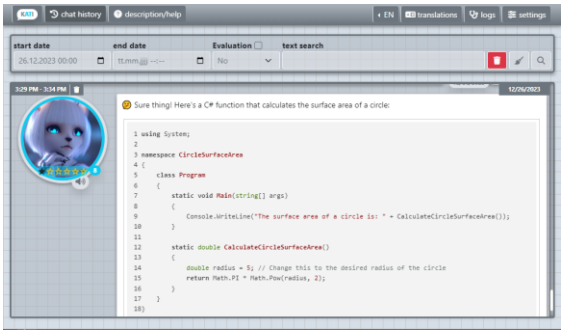


This prevents the taskbar from being used unnecessarily when KATI is not in use. Right-clicking on the tray icon opens a context menu that can be used to exit KATI.



THE CHAT HISTORY

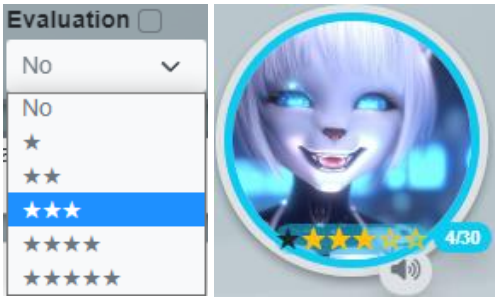
In the chat history, you can find past conversations. They are grouped by related conversations.



The topic header on the left shows the time period of the topic's first and last question. On the right you can see the date of the conversation.

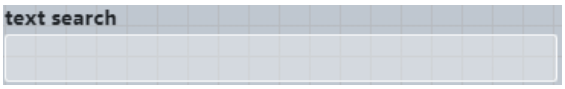
Initially, the history only shows the conversations of the current day. But you can also look at a specific period of time. If only the start date is set, all conversations from that date will be displayed. If only the end date is set, all conversations up to that date will be displayed.

The AI's responses can be evaluated. In this way, answers with a higher benefit, e.g. code examples, can be found more quickly.

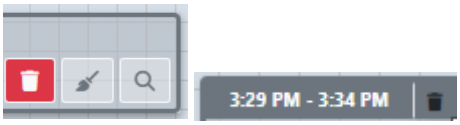


If the checkbox is set above the dropdown, the maximum rating will be searched. This allows you to find low-scoring AI answers and clean them up with just a few clicks.

Conversations can also be searched for free text.



The delete button in the filter deletes all currently filtered history entries. Individual chat topics can be deleted with the trash can button in the topic header.



The mop button resets the filter. The search can be started with the magnifying glass button or with the Enter key. The answers in the history can be read aloud by clicking on the speaker icon next to the avatar image. Clicking on the speaker icon again cancels the reading aloud.

The conversations in the history remain stored locally on the PC until they are manually deleted.

THE KATI CHAT

The KATI chat can be opened via the "KATI" button in the header navigation.

The chat input is optimized for sending long texts. You can make the input field larger or smaller. With the keyboard shortcut **shift+enter** a text break is made and with **enter** the chat request is sent. The text can also be sent with a mouse click on the paper airplane button.

With the eraser button you can delete the input and with the arrow button you can start a new topic.

EMOTION FEEDBACK WITH AVATARS

The user's avatar image is currently an anonymous astronaut. The user request is to the right of the text. At the bottom left of the user image, the timestamp of the chat request is displayed.

The AI has a choice of 157 avatar images that can convey different emotions. If multiple emojis are returned during a reply, the avatar image will change.

If the AI uses sad emojis, rather sad avatar images are displayed, if there is laughter in the same message, the avatar image changes to a rather cheerful representation.

The AI's response is displayed to the left of the text. At the bottom right of the AI avatar, the number of the current answer is displayed.

During a reply stream, a spinning gear is displayed instead of the number. The gear disappears as soon as the response is complete or canceled.

Unfortunately, previous chats cannot be continued in the current version of KATI-LLAMA. When a new chat session is started, the AI will only have context to the active conversation, but will not remember any conversations that have ended.

FAULT TOLERANCE

The AI can understand the meaning of a question even if there are spelling mistakes and missing words. However, the likelihood of misunderstandings then increases. It is also able to recognize when another language is spoken in the middle of a sentence.

DER NAME KATI

The AI can be addressed with KATI. This name can theoretically be made configurable if there is 😉 a need

PERFORMANCE

- Depending on the model configured, more or less RAM and processor power are required. This can affect the performance of the AI's responses. Try a smaller model and see if the AI responds faster. Keep in mind that the smaller the model, the lower the quality of the responses.
- Slow output may also be due to the configured processor setting. AVX is very slow, but it is mostly supported by older processors. With AVX2, the latency is significantly lower, but not all processors support it. Try chatting with AVX2 to see if it works for you.
- If the read-aloud function is enabled, the program waits to output until a complete sentence is available. To minimize the response time, you can disable the audio output, then the response text will be streamed without interruption.
- The AI sometimes takes longer to answer in general if it finds little information about a question. In this case, you can try to cancel the chat session and rephrase the question.