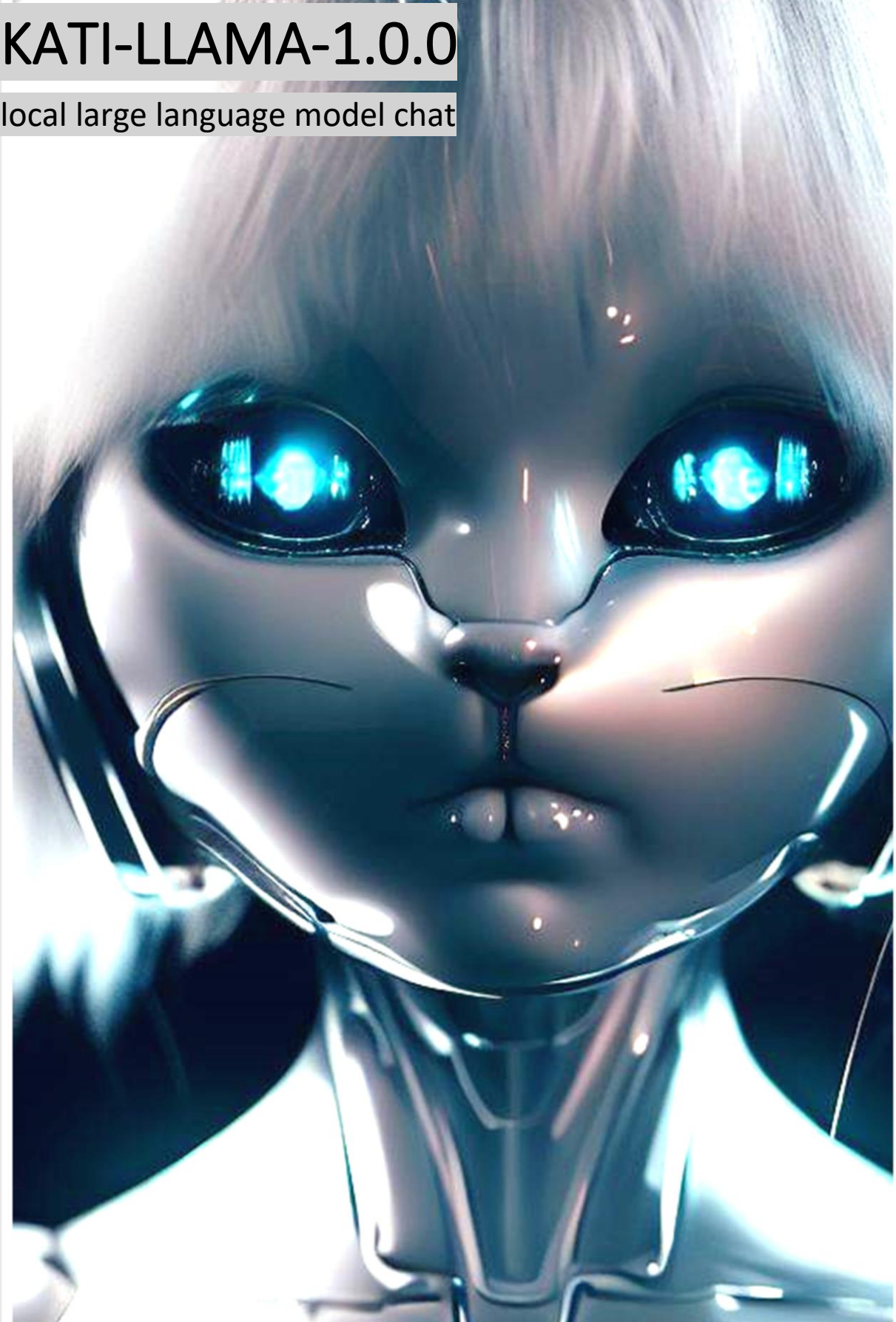
KATI-LLAMA-1.0.0

local large language model chat



INHALT

ÜBER KATI-LLAMA

KATI-LLAMA ist eine Benutzungsoberfläche zum Chatten mit Large Language Models auf dem privaten PC. Das Language Model kann in den Einstellungen automatisch heruntergeladen und danach offline benutzt werden.

Die KATI-Applikation erlaubt dem Benutzer eine Menschenähnliche Kommunikation mit einer KI. Die Antworten der KI können mit einer natürlichen Stimme ausgegeben werden und das Avatar Bild der KI ändert das Erscheinungsbild abhängig von der Stimmung des Chatbots. Im Folgenden findest du eine Zusammenfassung der Features von KATI-LLAMA.

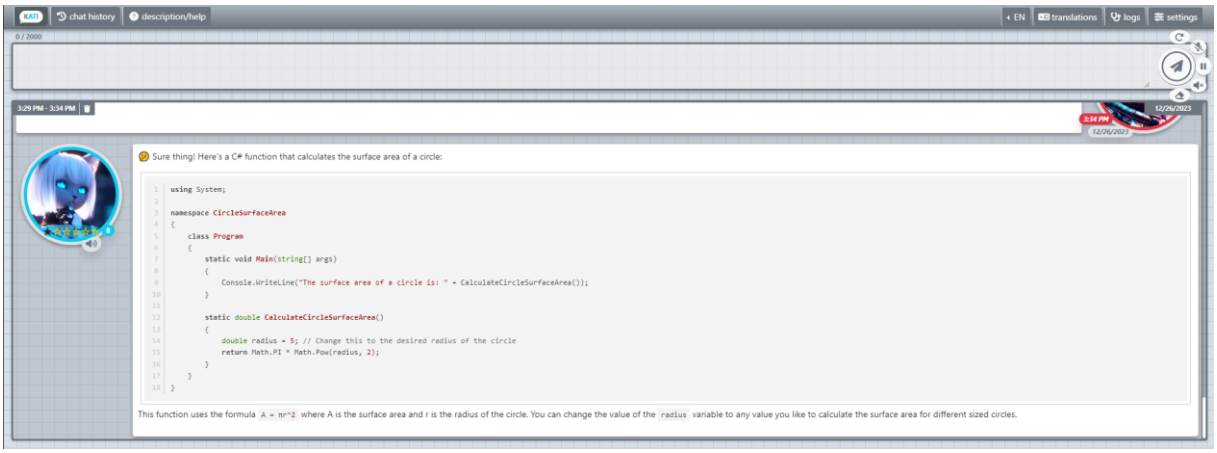
Features:

- Unterhaltung mit der KI ohne Internetverbindung
- Optionale Sprachausgabe mit einer im Betriebssystem- vorinstallierten Stimme oder einer natürlich klingenden TikTok Stimme. (Die TikTok Stimme erfordert eine Internetverbindung)
- Spracheingabe (System Speech oder Whisper)
- Dynamische Avatar Bilder zum Darstellen von KI- Emotionen.
- Chat Historie mit Filterfunktion und Vorlesefunktion.
- Bewertungsfunktion für KI-Antworten als Hilfe für die Filterfunktion
- Kürzere Wartezeiten durch direktes Streamen der Antworten.
  (Bei aktiver Vorlesefunktion, geschieht die Ausgabe erst bei vollständigem Satz)
- Text und Code wird für eine bessere Lesbarkeit formatiert dargestellt.
- Mehrsprachige Benutzungsoberfläche
  (DE, EN, FR, ES, PT, JA, KO)

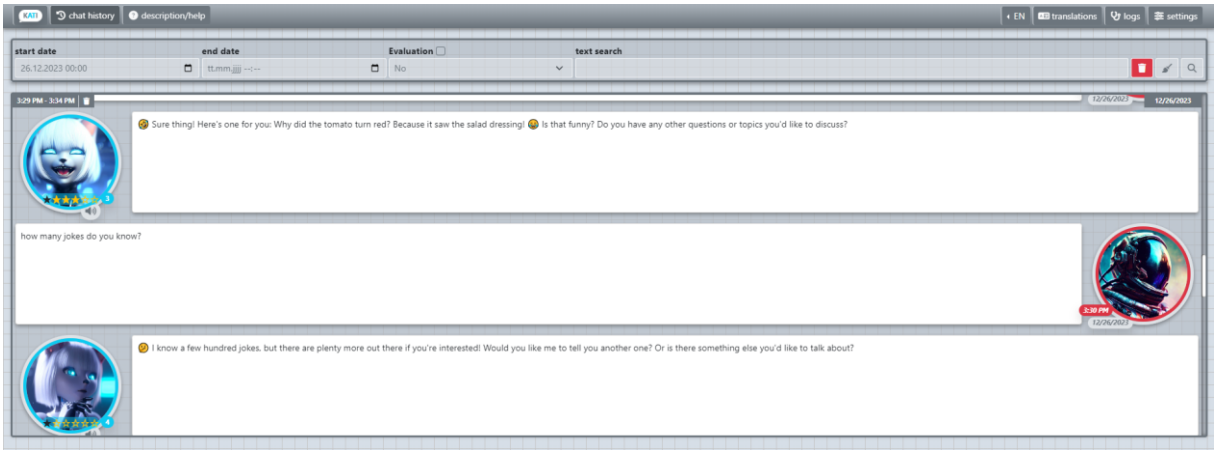
VERWENDETE NUGET PACKAGES/LIZENZEN:

- LLamaSharp (MIT)
- ElectronNET.API (MIT)
- Esprima (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)

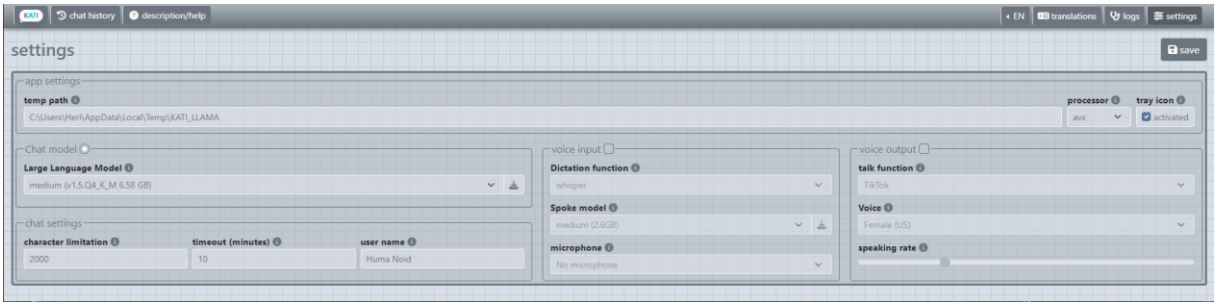
VORSCHAU: CHAT ANSICHT



VORSCHAU: CHATHISTORIE ANSICHT



VORSCHAU: EINSTELLUNGEN ANSICHT



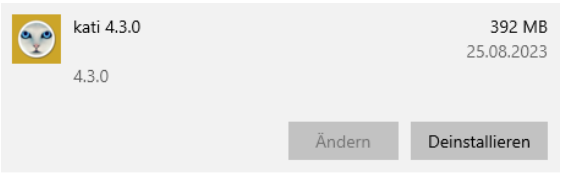
DOWNLOAD UND INSTALLATION

Zum Download stehen bei GitHub folgende Varianten von KATI zur Verfügung

Die portable Variante kann ohne Installation direkt ausgeführt werden. Das Starten dieser App dauert allerdings etwas länger als bei der installierten Variante. Die Installation wird mit einem Klick auf die kati-setup.exe gestartet. Die App wird dabei im folgenden Verzeichnis gespeichert:
C:\Users\...\AppData\Local\Temp\KATI_LLAMA

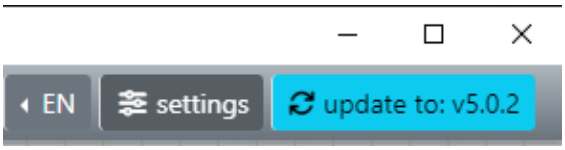
DEINSTALLATION

Die Deinstallation kann wie bei allen Windowsprogrammen unter „Apps und Features“ durchgeführt werden.



UPDATES

Bei einer neuen Version wird ein Button mit dem Versionsnamen im Header der App angezeigt.



Das Update kann entweder in der App über den Klick auf den Versions-Button gestartet werden oder durch erneutes Herunterladen der aktuellen Version auf GitHub. Letzteres geht etwas schneller.

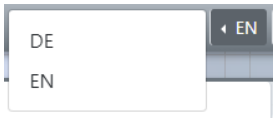
Bugfixes erkennt man an der letzten Ziffer, neue Features an der zweiten Ziffer und abgeschlossene Meilensteine an der ersten Ziffer.

KONFIGURATION DER SPRACHE

Die Benutzungsoberfläche und die Sprachausgabe von KATI kann in den folgenden Sprachen verwendet werden:

- Deutsch (DE)
- Englisch (EN)
- Französisch (FR)
- Spanisch (ES)
- Portugiesisch (PT)
- Japanisch (JA)
- Koreanisch (KO)

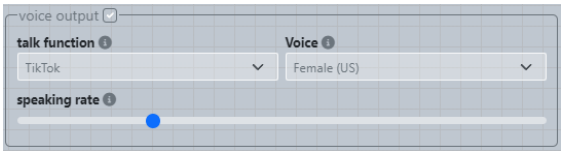
Die Sprache kann nach dem Start der App im Header-Menü geändert werden. Die Bedienungsanleitung ist nur in Deutsch und Englisch verfügbar.



Der Text-basierte Chat kann auch in anderen Sprachen geführt werden. Frag die KI einfach, ob sie die gewünschte Sprache versteht und sag ihr, dass du dich in dieser Sprache unterhalten möchtest, wenn es ihr möglich ist. Bitte beachte, dass die Vorlesefunktion nur in den oben gelisteten 7 Sprachen fehlerfrei möglich ist.

KONFIGURATION DER SPRACHAUSGABE

In den Spracheinstellungen kann die Sprechstimme und die Sprechgeschwindigkeit eingestellt werden.



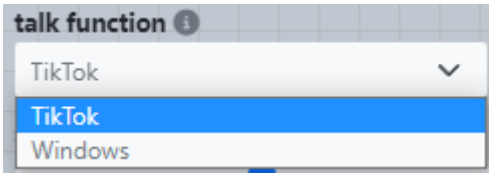
Die Sprachausgabe ist standardmäßig deaktiviert und kann in der Einstellungen Ansicht mit der blau hervorgehobenen Checkbox aktiviert werden. Alternativ kann man die Sprachausgabe auch direkt im Chatfenster mit dem Lautsprecher-Button umschalten.

Für eine verständliche Sprachausgabe muss die Benutzungsoberfläche in der gleichen Sprache eingestellt sein, in der die Unterhaltung geführt werden soll!

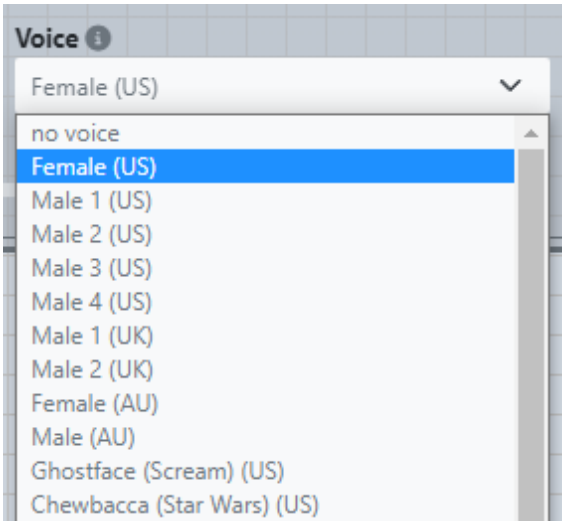
Die Geschwindigkeit für die Sprachausgabe ist initial auf 2x voreingestellt.

Wenn das Sprechtempo zu schnell sein sollte, kann dies auf das Niveau 1x geändert werden. (Speichern nicht vergessen)

KATI kann die Standardfunktion von Windows oder die TikTok Funktion für die Sprachausgabe nutzen.



Die TikTok Funktion besitzt eine etwas natürlichere Aussprache, als die Windows Funktion und hat auch, je nach eingestellter Sprache mehr Stimmen zur Auswahl. Die TikTok Stimme kann offline nicht benutzt werden.



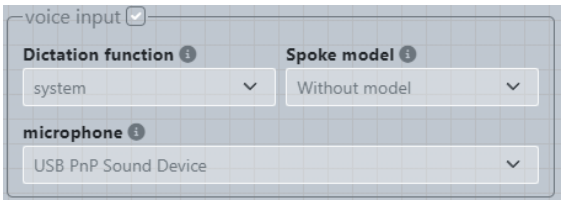
Wenn die Sprache gewechselt wird, wird automatisch die TikTok Funktion mit der ersten in der Sprache verfügbaren Stimme vorausgewählt.

KONFIGURATION DER SPRACHEINGABE

Die Sprache zu Text Funktion kann in den Chat Einstellungen oder auf der Hauptseite neben dem Absenden Button umgeschaltet werden.

KATI Unterstützt System Speech und Whisper für die Stimmerkennung.

System Speech ist eine in Windows integrierte „Speech To Text“ (STT) Funktionalität. Wenn kein Mikrofon gefunden werden kann, oder wenn es für die aktuell ausgewählte Sprache in Windows keine Spracherkennung gibt, kann die Funktion nicht aktiviert werden. In diesem Fall wird ein Hinweistext eingeblendet.

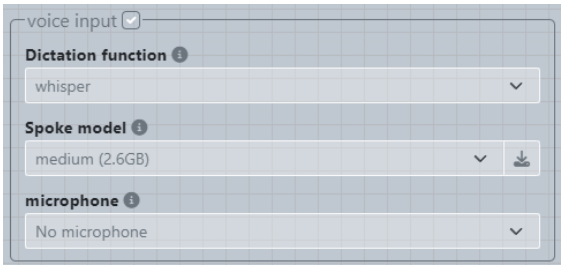


Die Systemeigene SST-Funktion arbeitet mit einem in Windows konfigurierten Standardmikrofon. In der Mikrofon-Dropdown wird dementsprechend immer nur das aktuell angeschlossene Standardmikrofon vorgeschlagen.

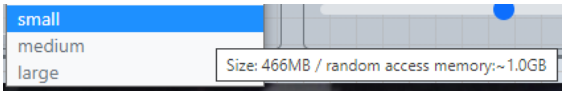
Je nach Mikrofon, Stimme oder Dialekt, kann das untrainierte, Windowseigene STT leider keine zufriedenstellende Genauigkeit bei der Wort-Erkennung liefern. Diese Funktion soll lediglich als Fall Back dienen, wenn KI-Unterstützte Spracherkennung nicht funktionieren sollte.

KATI unterstützt keine Sprachbefehle im Chat, wenn System Speech verwendet wird!

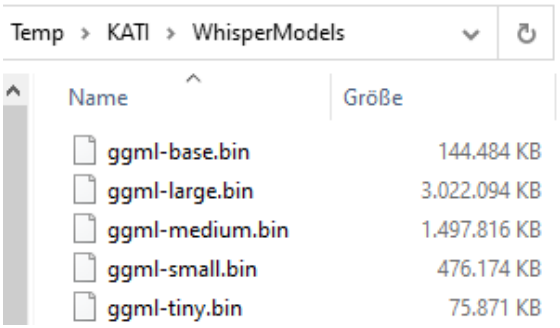
Whisper hat, je nach ausgewähltem Model, eine sehr viel zuverlässigere Spracherkennung.



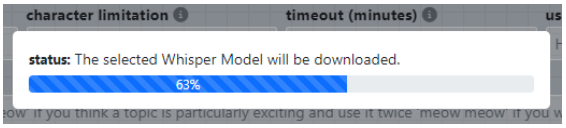
Je größer das Model, desto besser ist die Spracherkennung. Beachte bei deiner Wahl den Arbeitsspeicherverbrauch. Dieser wird als Tooltip bei den jeweiligen Models angezeigt.



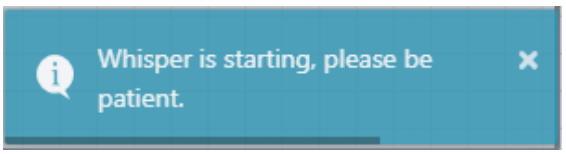
Wenn möglich, sollte mindestens das zweitgrößte Model (ggml-medium.bin) benutzt werden, um eine relativ gute Spracherkennung zu haben.

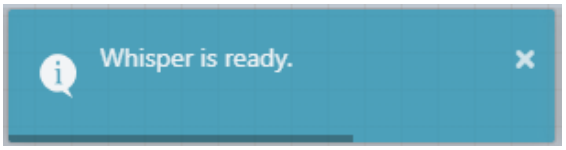


Wenn das ausgewählte Model noch nicht vorhanden ist, kann es mit dem Download Button automatisch in das *KATI_LLAMA/WhisperModels* Verzeichnis heruntergeladen werden. Dies kann, je nach Modelgröße, einige Minuten dauern. In dieser Zeit sind die Navigationselemente in KATI gesperrt.



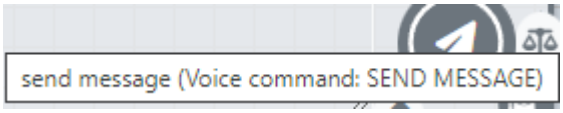
Das Starten von Whisper benötigt etwas mehr Zeit als System-Speech. Sobald folgende Meldung angezeigt wird, kann das Mikrofon benutzt werden.





Whisper setzt modernere Prozessoren voraus. Es kann also passieren, dass die SST-Funktion mit Whisper auf älteren PCs nicht unterstützt wird. In diesem Fall wird anstelle der oberen Meldung eine Fehlermeldung eingeblendet. Bei einem Kompatibilitätsproblem kann auf die bereits erwähnte System-Speech Funktion ausgewichen werden.

Wenn Whisper konfiguriert ist, können Sprachbefehle verwendet werden. Diese werden als Tooltip über den Navigationselementen angezeigt.

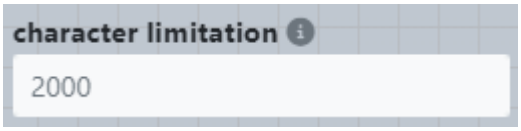


Folgende Befehle können Benutzt werden:

- **SEND MESSAGE:**
  Sendet die Nachricht
- **RESET MESSAGE:**
  Löscht die Eingabe
- **REMOVE LAST SENTENCE:**
  Löscht den letzten Satz
- **STOP CONVERSATION:**
  Abbruch der Ausgabe
- **RESET CONVERSATION:**
  Startet eine neue Unterhaltung

CHAT EINSTELLUNGEN

Die Zeichenlimitierung für die Texteingabe dient in erster Linie dazu, dass die KI weniger Text auswerten muss. Je kürzer und konkreter die Texteingabe, desto genauer sind die Antworten. Die Eingabe ist initial auf 2000 Zeichen begrenzt. Hier kann, bei Bedarf auch ein höherer Wert Konfiguriert werden.

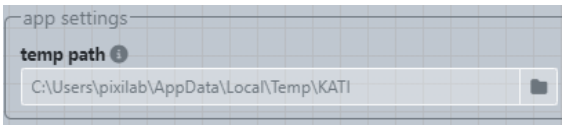


Die konfigurierte Zeichenlimitierung wird zusätzlich im Zeichen-Zähler über der Chat-Eingabe angezeigt. Der Text im Eingabefeld wird automatisch abgeschnitten, wenn er die konfigurierte Länge überschreiten sollte.

KONFIGURATION DES SPEICHERPFADS

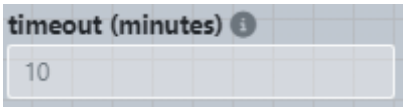
Der temporäre Ordner ist für temporäre Dateien von KATI vorgesehen. Aktuell wird in diesem Ordner die Datenbank für die Chat-Historie gespeichert und die Language Models.

Die Chat-Historie bleibt auch nach einem Update der Anwendung erhalten. Die App Einstellungen werden allerdings nach einem Update auf die Standardeinstellungen zurückgesetzt. Der vorgeschlagene Speicherpfad kann unverändert bleiben oder ein alternativer Pfad konfiguriert werden. (Speichern nicht vergessen)

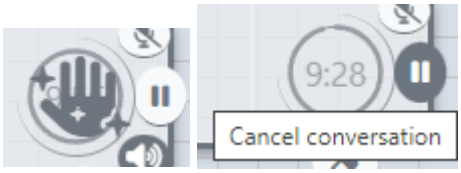


KONFIGURATION DES KI-TIMEOUTS

Die KI-Antworten sind unbegrenzt und können theoretisch in einem Endlosmonolog ausarten. Die Chatanfrage wird nach 10 Minuten automatisch abgebrochen, falls die KI nicht mehr aufhören will zu reden. In den Einstellungen kann diese Begrenzung angepasst werden.

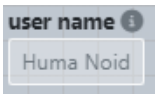


Die Chatanfrage kann auch manuell abgebrochen werden, indem man erneut auf den Absenden Button oder auf den Pause-Button klickt.



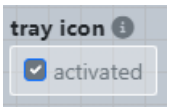
KONFIGURATION DES ANREDE-NAMENS

Im Feld „Benutzername“, kann eingestellt werden, mit welchem Namen die KI dich ansprechen soll. Standardmäßig ist der Name „Huma Noid“ konfiguriert. Du kannst auch einen alternativen Namen speichern. Beachte aber, dass diese Einstellung nach einem Update der App zurückgesetzt wird.

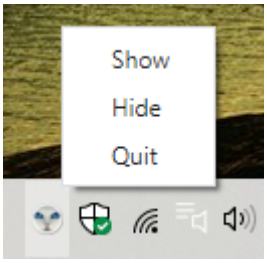


KONFIGURATION DES KATI FENSTERS

Die KATI-Anwendung kann optional in den Tray minimiert werden.

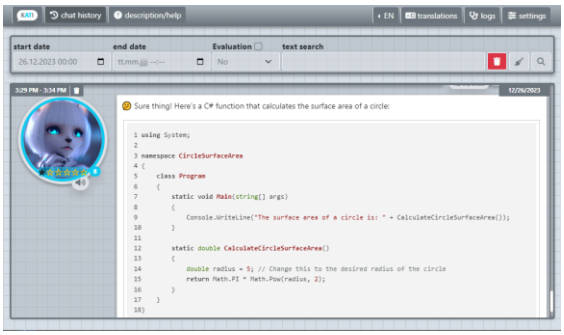


Dadurch wird die Taskleiste nicht unnötig belegt, wenn KATI gerade nicht benutzt wird. Der rechte Mausklick auf das Tray-Icon öffnet ein Kontextmenü, mit dem KATI beendet werden kann.



DIE CHAT HISTORIE

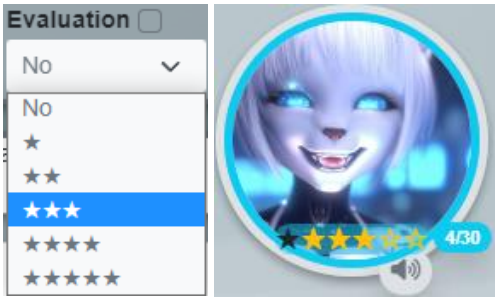
In der Chathistorie findet man vergangene Unterhaltungen wieder. Sie werden nach zusammenhängenden Unterhaltungen gruppiert.



Im Themen-Header wird auf der linken Seite der Zeitraum der ersten und letzten Frage des Themas angezeigt. Rechts sieht man das Datum der Unterhaltung.

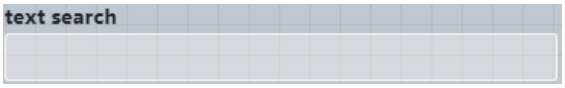
Die Historie zeigt initial immer nur die Unterhaltungen des aktuellen Tages an. Man kann sich aber auch einen bestimmten Zeitraum ansehen. Wenn nur das Start Datum gesetzt ist, werden alle Unterhaltungen ab diesem Datum angezeigt. Wenn nur das End Datum gesetzt ist, werden alle Unterhaltungen bis zu diesem Datum angezeigt.

Die Antworten der KI können bewertet werden. So können Antworten mit einem höheren Nutzen, z.B. Code Beispiele, schneller wiedergefunden werden.

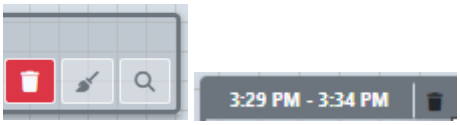


Wenn die Checkbox oberhalb des Dropdowns gesetzt ist, wird nach der Maximalbewertung gesucht. So kann man KI-Antworten, mit niedriger Bewertung finden und mit wenigen Klicks bereinigen.

Unterhaltungen können auch nach einem Freitext durchsucht werden.



Der Lösch-Button im Filter löscht alle aktuell gefilterten Historieneinträge. Einzelne Chat-Themen können mit dem Mülleimerbutton im Themenheader gelöscht werden.



Mit dem Wischmopp-Button wird der Filter zurückgesetzt. Die Suche kann mit dem Lupe-Button oder mit der Enter-Taste gestartet werden. Die Antworten in der Historie können mit einem Klick auf das Lautsprechersymbol neben dem Avatar Bild vorgelesen werden. Erneutes Anklicken des Lautsprechersymbols bricht das Vorlesen wieder ab.

Die Unterhaltungen in der Historie bleiben lokal auf dem PC gespeichert, bis sie manuell gelöscht werden.

DER KATI CHAT

Der KATI-Chat kann über den Button „KATI“ in der Headernavigation geöffnet werden.

Die Chateingabe ist für das Senden langer Texte optimiert. Man kann das Eingabefeld größer oder kleiner machen. Mit der Tastenkombination **shift+enter** wird ein Textumbruch gemacht und mit **enter** die Chatanfrage abgesendet. Der Text kann auch mit einem Mausklick auf den Papierflieger-Button abgesendet werden.

Mit dem Radiergummi-Button kann die Eingabe gelöscht werden und mit dem Pfeil-Button kann man ein neues Thema beginnen.

EMOTION-FEEDBACK MIT AVATAREN

Das Avatar-Bild des Benutzers ist aktuell ein anonymer Astronaut. Die Benutzeranfrage steht rechts vom Text. Links unten vom Benutzerbild wird der Zeitstempel der Chatanfrage angezeigt.

Die KI hat eine Auswahl aus 157 Avatar-Bildern, die verschiedene Emotionen wiedergeben können. Wenn während einer Antwort mehrere Emojis zurückgegeben werden, wechselt das Avatar-Bild.

Benutzt die KI traurige Emojis, so werden eher traurige Avatar-Bilder eingeblendet, wird in der gleichen Nachricht gelacht, so wandelt sich das Avatar-Bild zu einer eher heiteren Darstellung.

Die Antwort der KI wird links vom Text angezeigt. Rechts unten vom KI-Avatar wird die Nummer der aktuellen Antwort angezeigt.

Während eines Antwort-Streams wird anstelle der Nummer ein drehendes Zahnrad eingeblendet. Das Zahnrad verschwindet, sobald die Antwort vollständig ist, oder abgebrochen wird.

Früher Chats können in der aktuellen Version von KATI-LLAMA leider nicht fortgesetzt werden. Wenn eine neue Chat-Session begonnen wird, wird die KI lediglich einen Kontext zu der aktiven Unterhaltung haben aber sich nicht an beendete Unterhaltungen erinnern.

FEHLERTOLERANZ

Die KI kann den Sinn einer Frage auch bei Rechtschreibfehlern und fehlenden Wörtern verstehen. Die Wahrscheinlichkeit für Missverständnisse nimmt dann allerdings zu. Sie ist auch in der Lage zu erkennen, wenn mitten im Satz in einer anderen Sprache gesprochen wird.

DER NAME KATI

Die KI kann mit KATI angesprochen werden. Dieser Name kann theoretisch konfigurierbar gemacht werden, falls Bedarf besteht 😉

PERFORMANCE

- Je nach konfiguriertem Model, wird mehr oder weniger Arbeitsspeicher und Prozessorleistung benötigt. Dies kann sich auf die Performance bei den Antworten der KI auswirken. Probiere ein kleineres Model aus und schau, ob die KI schneller reagiert. Beachte, dass mit kleinerem Model auch die Qualität der Antworten abnimmt.
- Eine langsame Ausgabe kann auch an der konfigurierten Prozessor Einstellung liegen. AVX ist sehr langsam, wird aber meistens auch von älteren Prozessoren unterstützt. Mit AVX2 ist die Wartezeit deutlich geringer, wird aber nicht von allen Prozessoren unterstützt. Versuche, ob bei dir der Chat mit AVX2 funktioniert.
- Wenn die Vorlesefunktion aktiviert ist, wartet das Programm mit der Ausgabe, bis ein vollständiger Satz verfügbar ist. Um die Antwortzeit zu minimieren, kann man die Audioausgabe deaktivieren, dann wird der Antworttext ohne Unterbrechung gestreamt.
- Die KI braucht manchmal allgemein länger für eine Antwort, wenn es wenig Informationen zu einer Frage findet. In diesem Fall kannst du versuchen die Chatsession abzubrechen und die Frage anders zu formulieren.