# LibreChat mit ChatGPT

## Probebetrieb

Dieser Dienst befindet sich noch in der Erprobung. Verfügbarkeit und Zuverlässigkeit können noch nicht garantiert werden. Bitte melden Sie technische Fehler an support@rz.tu-clausthal.de.

## Beschreibung

LibreChat ermöglicht die Integration unterschiedlicher Anbieter von Large Language Modellen (LLMs). Aktuell stehen Modelle von OpenAI und der GDWG zur Verfügung:

- OpenAI | gpt-4o
- OpenAI | gpt-4o-mini
- OpenAI | gpt-o1
- OpenAI | gpt-o1-mini
- OpenAI | gpt-o3-mini

Informationen zu den Modellen der GDWG finden zu auf der Seite der GWDG zur Modellbeschreibung der GDWG .

Diese stehen allen Studierenden und Mitarbeitenden der TU Clausthal zur Verfügung. Der TU entstehen damit jedoch bei den Modellen von OpenAI nutzungsabhängige Kosten, die bis auf weiteres aus einem zentralen Budget getragen werden. *Die Kosten entstehen nicht dem Nutzer!*

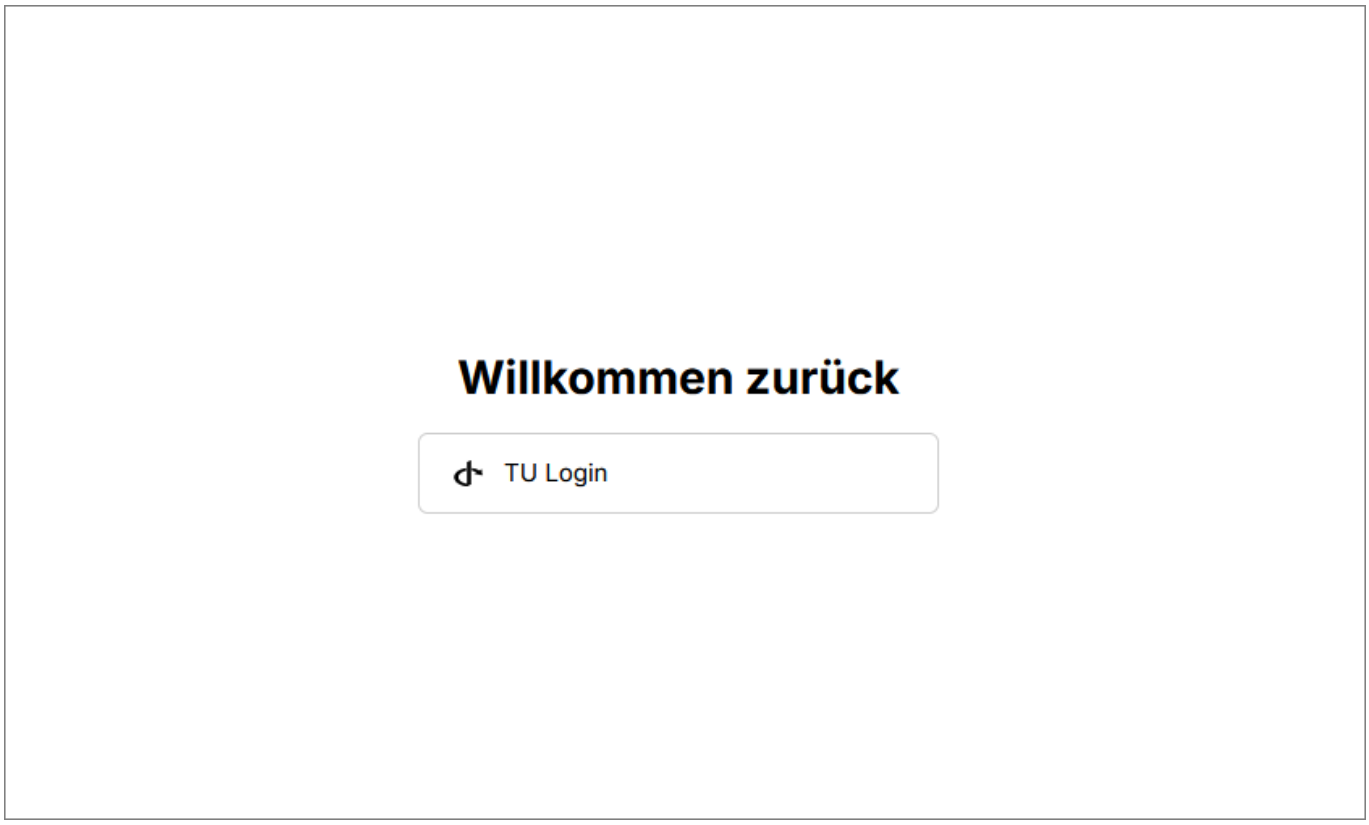
Wir weisen darauf hin, dass LibreChat ein Frontend (Oberfläche) für ChatGPT ist! Alle Anfragen werden direkt über die von OpenAI bereitgestellte API weitergeleitet. Es findet eine Drittstaatenübermittlung ins Nicht-EU-Ausland statt. Es besteht jedoch ein Datenverabeitungsvertrag mit OpenAI in dem zugesichert wird, dass die übermittelten Daten nicht für Trainingszwecke verwendet werden und die Daten nicht an Dritte weitergegeben werden. Siehe Enterprise privacy at OpenAI. Zudem sei auf die Datenschutzhinweise zu LibreChat verwiesen.

Wir weisen darauf hin, dass derzeit die Bildgenerierung mit LibreChat nicht möglich ist! Hier verweisen wir derzeit auf Microsoft Copilot.

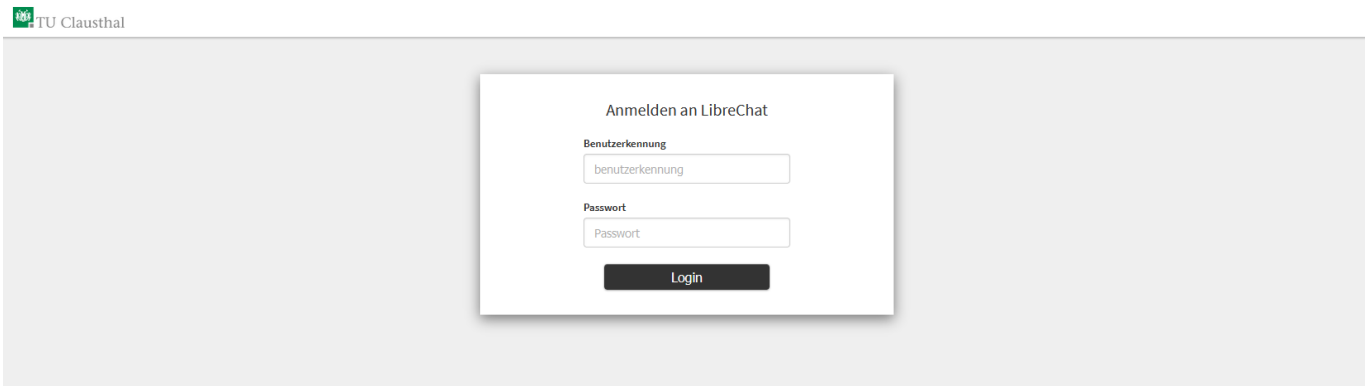

Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel RZ-Dokumentationen, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.

# Zugang

LibreChat erreichen Sie unter folgender Adresse: https://chat-ai.tu-clausthal.de/ aus dem Netz der TU-Clausthal. Für den Zugriff von zu Hause benötigen Sie eine VPN-Verbindung.




Gehen Sie auf den Button „TU Login“. Geben Sie danach als Benutzerkennung Ihren <TUC-User>(z.B. abc123) an, sowie Ihr zugehöriges TUC-Passwort und bestätigen Sie mit „Login“.





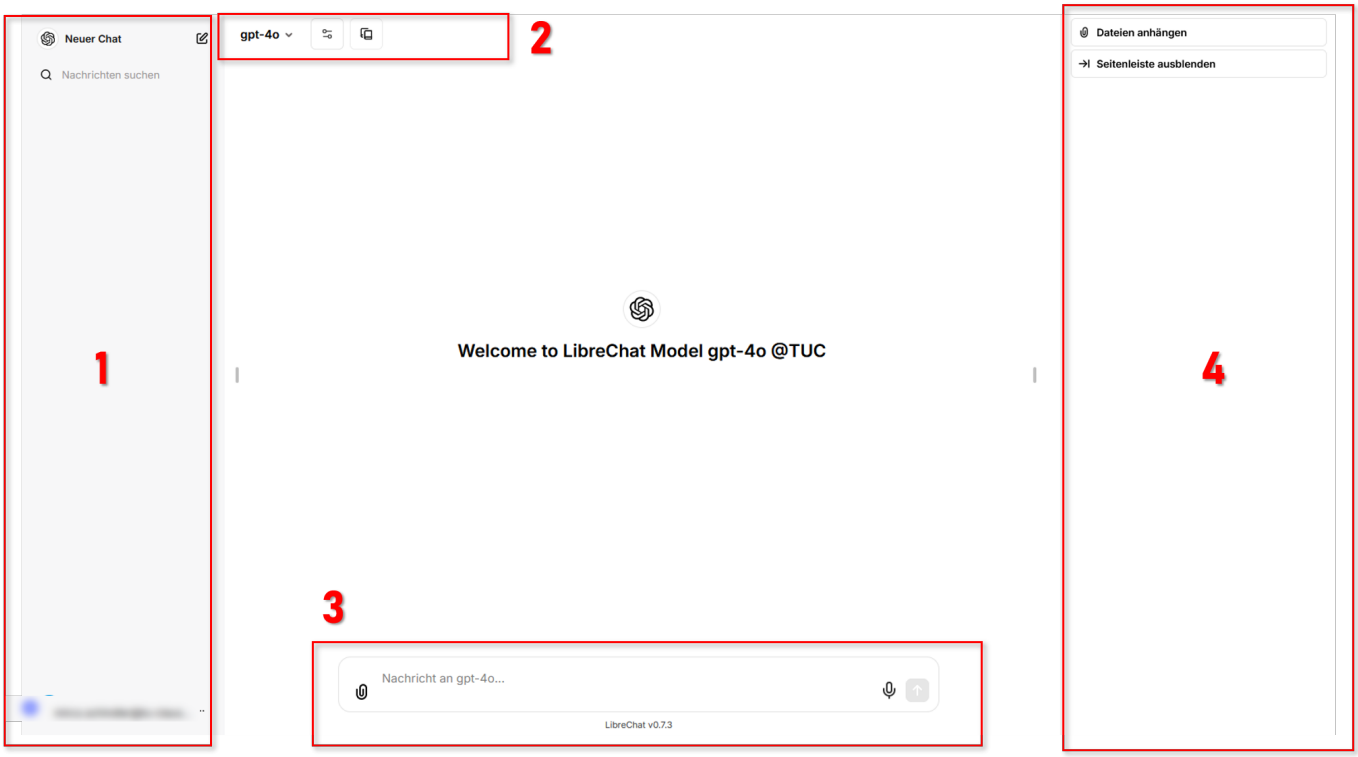
Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel RZ-Dokumentationen, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.

Nach erfolgreicher Anmeldung steht Ihnen der Zugang zu ausgewählten Sprachmdellen von OpenAI zur Verfügung.

# Nutzung

Die Oberfläche von LibreChat sieht sehr ähnlich zu dem Webinterface von ChatGPT selbst aus.

Bitte beachten Sie, dass je nach ausgewähltem Sprachmodell dieses bis zu einem bestimmten Datum trainiert ist. Die Modelle können daher unvollständige und auch falsche Antworten liefern. Zudem verfügen die angeboteten Sprachmodell nicht über unternehmensspezifisches Wissen.




Auf der linken Seite **(1)** finden Sie Ihren Chatverlauf, wo Sie neue Unterhaltungen beginnen oder bereits bestehende fortsetzen können. Ihre Fragen und die Antworten von ChatGPT werden hier in chronologischer Reihenfolge angezeigt. Für jedes neue Thema sollten Sie einen neuen Chat aufmachen, damit sich der Kontext unterschiedlicher Themen nicht vermischt.

Oben **(2)** befindet sich ein Menü, in dem Sie das konkrete Sprachmodell auswählen können. Hier ist auch die Anpassung der Modellparameter möglich.

Im Eingabefeld **(3)** tippen Sie Ihre Fragen oder Kommandos ein und können zudem Dateien über das Symbol einer Klammer hochladen. Auch ist es möglich Text per Spracheingabe einzugeben.

Derzeit steht die Spracheingabe nicht in allen Browsern zur Verfügung, da hier auf die

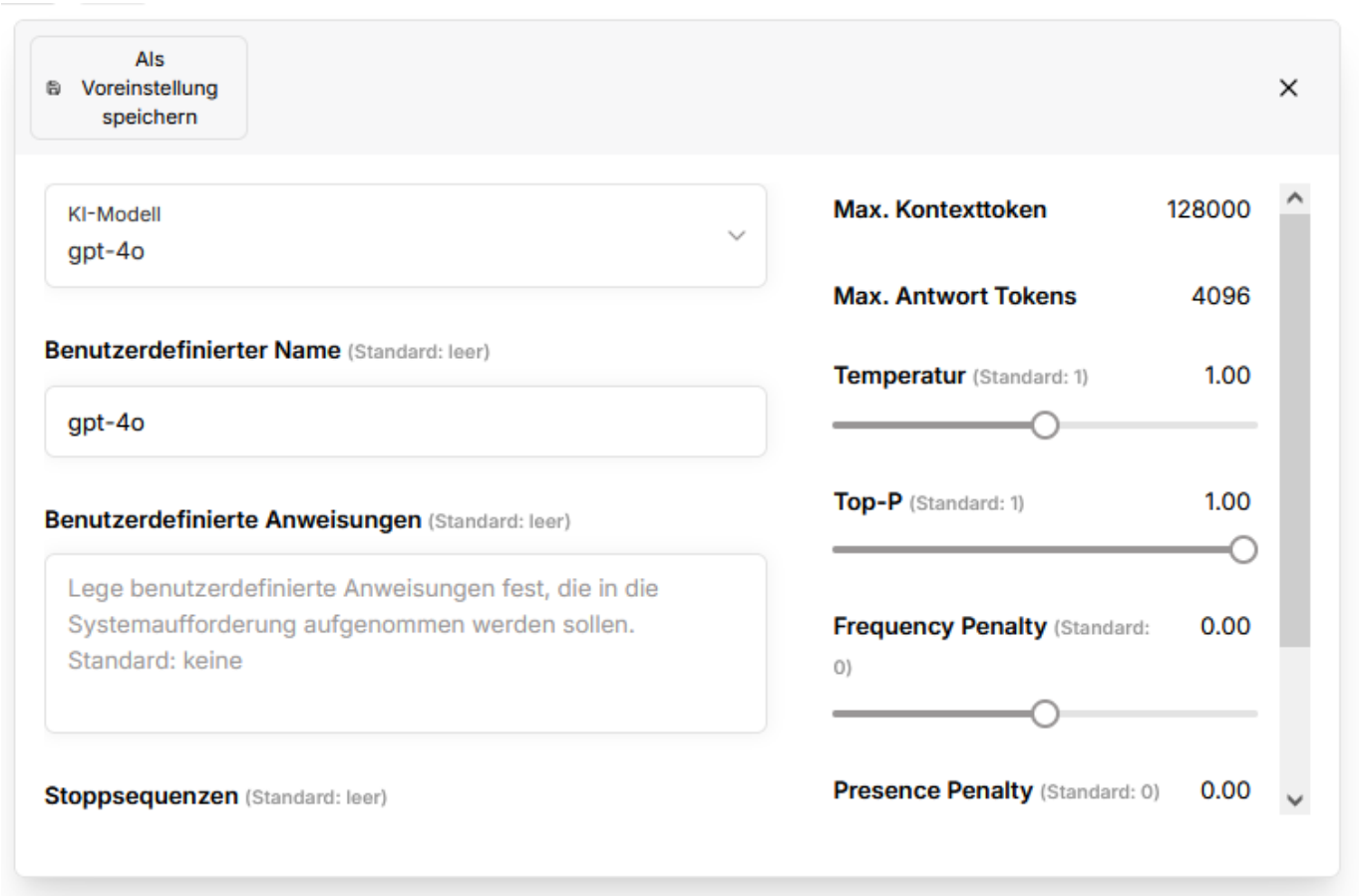

Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel RZ-Dokumentationen, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.

Speech-To-Text Funktion des Browsers verwendet. Möglich ist dies zum Beispiel mit Chrome und Edge.

Im rechten Bereich **(4)** erhalten Sie eine Übersicht über alle Dateien, die Sie in diesem Chatverlauf hochgeladen haben.

# Modellparameter

Diese Parameter helfen dabei, die Vielfalt, Relevanz und Einzigartigkeit der generierten Antworten zu kontrollieren. Anfängern wird jedoch empfohlen die Standardwerte beizubehalten.




## Erläuterung der Modellparameter

- **Kontexttokens**: Dies bezieht sich auf die maximale Anzahl von Token (Einheiten von Wörtern oder Zeichen), die das Modell auf einmal „verstehen“ oder verarbeiten kann. Bei vielen Sprachmodellen gibt es eine obere Grenze für die Anzahl von Token, die im Eingabekontext enthalten sein können. Wenn diese Grenze überschritten wird, muss der Text gekürzt oder geteilt werden.
- **Antwort Tokens**: Dies bezieht sich auf die Tokens, die im generierten Output oder in der


Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel RZ-Dokumentationen, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.

Antwort des Modells enthalten sind. Wenn das Modell eine Antwort erzeugt, besteht diese Antwort aus einer bestimmten Anzahl von Token. Es gibt oft auch eine maximale Begrenzung dafür, wie viele Token die Antwort des Modells haben kann, abhängig von der Implementierung oder den Nutzerpräferenzen.

- **Temperatur** (Temperature): Dieser Parameter steuert den Zufallsgrad der Antworten. Ein niedrigerer Wert (z.B. 0.1) führt zu vorhersehbareren und konsistenteren Antworten. Ein höherer Wert (z.B. 1.0) erhöht die Vielfalt und macht die Antworten weniger vorhersagbar. Eine Temperatur von 0 würde immer dieselbe Antwort auf dieselbe Eingabeaufforderung geben (sofern der Rest der Einstellungen identisch ist).
- **Top-P** (Topical Penalization, auch bekannt als Nucleus Sampling): Dieser Wert bestimmt, wie das Modell aus einer Gruppe von wahrscheinlichen nächsten Worten auswählt. Ein niedrigerer Top-P-Wert (z.B. 0.5) bedeutet, dass nur die wahrscheinlichsten Worte berücksichtigt werden. Ein höherer Wert (z.B. 0.95) erlaubt eine größere Vielfalt potenzieller Worte. Dieser Mechanismus funktioniert durch das Beschränken der Auswahl an Wörtern auf ein kumulatives Wahrscheinlichkeitsintervall („P“).
- **Frequency Penalty** (Häufigkeitsstrafe): Dieser Parameter verhindert, dass das Modell dieselben Wörter oder Phrasen wiederholt. Je höher der Wert, desto größer die Strafe für die Wiederholung von Wörtern, was zu einer variableren und diversifizierteren Antwort führt.
- **Presence Penalty** (Anwesenheitsstrafe): Ähnlich wie die Frequency Penalty, aber dieser Parameter bestraft Wörter basierend darauf, ob sie bereits im Text erschienen sind, unabhängig von ihrer Häufigkeit. Ein höherer Wert fördert die Einführung neuer, bisher nicht verwendeter Wörter oder Phrasen in den Antworten.

# Weiterführende Informationen

FAQ-LibreChat

Direkt-Link:
**https://doku.tu-clausthal.de/doku.php?id=sonstige_dienste:ki-dienste:librechat&rev=1739957568**

Letzte Aktualisierung: **10:32 19. February 2025**


Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel RZ-Dokumentationen, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.