BRAID
Bremen AI and Digitalization for Administration

GWDG Chat AI

Empfohlene Modelle für typische Anwendungsfälle

Stand: 28. Mai 2026

Alle open-weight Modelle laufen auf Servern der GWDG. Ihre Eingaben werden nicht gespeichert und verlassen nicht die Infrastruktur — datenschutzkonform für hochschulinterne Inhalte.

Empfohlene Modelle nach Anwendungsfall

Allround — Texte, Analyse, Übersetzung, Alltag

Modell	Empf. Einstellungen	Besonders geeignet für	
Qwen 3.6 35B A3B	temp: 1,0 / top_p: 0,95	Texte, Analyse, Bilder (Vision)	✦ Empfehlung
Gemma 4 31B	temp: 1,0 / top_p: 0,95	Bildanalyse, multimodale Aufgaben, Vision	
Llama 3.3 70B Instruct	temp: 0,7 / top_p: 0,80	Sprache, Übersetzung, kreatives Schreiben	
GPT OSS 120B	temp: 0,5 / top_p: 0,50	Reasoning, strukturierte Aufgaben	

Komplexe Aufgaben — Forschung, Analyse, Reasoning

Modell	Empf. Einstellungen	Besonders geeignet für	
Qwen 3.5 122B A10B	temp: 0,6 / top_p: 0,95	Fachliteratur, komplexe Analyse, lange Dokumente	✦ Empfehlung
GLM-4.7	temp: 1,0 / top_p: 0,95	Komplexe Berechnungen, mehrstufige Workflows	
Mistral Large 3 675B Instruct	temp: 0,5 / top_p: 0,50	Starkes Allround-Modell, Vision inklusive	
DeepSeek R1 Distill Llama 70B	temp: 0,7 / top_p: 0,80	Schritt-für-Schritt-Reasoning, Logik, Mathe	

Code & technische Aufgaben — Programmierung, Datenanalyse

Modell	Empf. Einstellungen	Besonders geeignet für	
Qwen 3.6 35B A3B	temp: 1,0 / top_p: 0,95	Python, R, Skripte, Debugging	✦ Empfehlung
Devstral 2 123B Instruct	temp: 0,5 / top_p: 0,50	Größere Projekte, mehrere Dateien, komplexe Pipelines	

Modellauswahl: Welches Modell passt zu meiner Aufgabe?

Im Modellnamen schauen auf:

Instruct	Standard-Chat — für Gespräche und Anweisungen optimiert. Immer die richtige Wahl im Chat.
Coder / Dev	Programmierung — auf Code spezialisiert, nicht für allgemeine Texte gedacht.
Vision / Omni	Bilder — Modell versteht Fotos, Screenshots, Diagramme als Eingabe. Omni zusätzlich auch Audio & Video.
Med	Medizinische Inhalte — auf klinische Texte und medizinische Bildgebung spezialisiert.
7B / 30B / 70B / …	Grobe Größenklasse — mehr Parameter bedeutet oft mehr Fähigkeiten.

Außerdem im Interface beachten:

- **Vision-Symbol** → Modell kann Bilder verarbeiten
- **Reasoning-Symbol** → Modell für schwierige Denkaufgaben optimiert
- Höhere Versionsnummer = neueres Modell *(z. B. Qwen 3.5 ist neuer als Qwen 3)*

Wichtige Einstellungen

Temperature	Steuert, wie kreativ oder variabel die Antworten sind. Niedrig (z. B. 0,3) → präziser, sachlicher — gut für Zusammenfassungen oder Faktenanalyse. Hoch (z. B. 1,0) → kreativer, offener — gut wenn Texte abwechslungsreicher klingen sollen.
Top_P	Steuert, wie breit das Modell aus möglichen Wörtern auswählt. Niedrig → fokussierter. Hoch → vielfältiger. In den meisten Fällen muss nur Temperature angepasst werden — Top_P kann so bleiben.

Die empfohlenen Einstellungen pro Modell sind in der Tabelle oben angegeben. Für die meisten Aufgaben muss nichts geändert werden.

Vollständige Modellübersicht mit allen Parametern
docs.hpc.gwdg.de/services/ai-services/chat-ai/models/
Dort sind alle verfügbaren Modelle mit technischen Details aufgeführt. Empfehlung: Probiere die Modelle aus und finde deine eigenen Favoriten.

Fragen oder Feedback?
Wir helfen gerne weiter: **TeamKI@uni-bremen.de**