

Übersicht der Sprachmodelle in AIS-chat

Modell (Anbieter)	Typische Stärken	Mögliche Nachteile	Besonders geeignet für
GPT-5 (OpenAI)	sehr hohe Qualität bei komplexen Aufgaben, starkes Reasoning, gute Deutschkompetenz, i. d. R. multimodal (Text+Bild, falls freigeschaltet)	höhere Kosten/Latenz als „Mini/Nano“, strenge Sicherheitsfilter, US-Anbieter; kann in Einzelfällen halluzinieren	Oberstufe/Abitur, anspruchsvolle Analysen, differenzierte Erklärungen, Kreativtexte
GPT-5 mini (OpenAI)	schnell, kosteneffizient, guter Allrounder	weniger Tiefe bei komplexen Problemstellungen, teils oberflächlichere Antworten	Routineaufgaben, kurze Erklärungen, Formulieren/Umformulieren
GPT-5 nano (OpenAI)	sehr schnell, sehr günstig/ressourcenschonend	deutlich begrenztes Tiefenverständnis, empfindlicher für Unklarheiten in Prompts	Stichpunkte, einfache Fragen, jüngere Klassen, schnelle Checks
GPT-4o-mini (OpenAI)	solide Alltagsleistung, kompakte multimodale Fähigkeiten (falls Bildfunktion aktiv)	ältere Generation, geringere Tiefe/Genauigkeit als GPT-5	Bild-zu-Text-Aufgaben, schnelle Alltagsunterstützung
o3-mini (OpenAI)	fokussiert auf logisches/technisches Reasoning, gut für MINT-Aufgaben und Coding	in kreativen/erzählerischen Aufgaben etwas nüchterner, kann langsamer sein als „mini/nano“	Mathematik/Physik/Chemie-Problemlösen, Programmieren, Schritt-für-Schritt-Herleitungen
Llama-3.1-405B (Meta, Open Weights)	sehr starke Textqualität, gute Deutschkompetenz; open-weights ermöglichen EU/On-Prem-Hosting	hohe Rechenanforderungen/Latenz, i. d. R. textzentriert	datenschutzsensible Szenarien mit leistungsfähigem Modell, hochwertige Texte
Llama-3.1-8B (Meta, Open Weights)	sehr effizient und schnell, gut für einfache Aufgaben; gut lokal/on-prem betreibbar	begrenztes Tiefenverständnis, kürzeres Kontextverständnis	Kurze Antworten, Zusammenfassungen, einfache Arbeitsaufträge
Llama-3.3-70B (Meta, Open Weights)	leistungsstarkes aktuelles Modell, gute Balance aus Qualität und Aufwand	nicht ganz auf Top-SOTA-Niveau bei sehr komplexen Aufgaben; Rechenbedarf spürbar	umfangreiche Arbeitsblätter, Erklärtex te, solide Allround-Nutzung
Mistral Nemo Instruct	multilingual, effizient, im	Qualität variiert je nach Aufgabe; weniger stark	datenschutzfreundliche Workflows,

(Mistral/NVIDIA-Ökosystem)	Open-Source-Umfeld gut integrierbar; EU-nahe Anbieteroptionen	bei sehr langem Kontext/komplexem Reasoning	Standard-Textaufgaben, kurze Rewrites
----------------------------	---	---	---------------------------------------

Kurzfazit:

- Für Tempo und Alltag: GPT-5 nano/mini oder Llama-3.1-8B.
- Für komplexe Aufgaben: GPT-5, o3-mini (MINT), Llama-3.1-405B/Llama-3.3-70B als starke Open-Weights-Alternativen.
- Für Bild-bezogene Aufgaben: GPT-4o-mini (falls Bildfunktion freigeschaltet).
- Für besonders datenschutzsensible Nutzung: Llama- oder Mistral-Varianten (bei EU/On-Prem-Hosting).

AIS-chat:

Entscheidungshilfe: Welches Modell wähle ich wann? (AIS-chat NRW)

Schnellstart

- **Tempo & günstig** → GPT-5 nano oder GPT-5 mini; alternativ Llama-3.1-8B
- **Maximale Qualität/Tiefe** → GPT-5; open-weights-Alternativen: Llama-3.1-405B oder Llama-3.3-70B
- **MINT/Logik/Code** → o3-mini (ggf. GPT-5, wenn es sehr knifflig wird)
- **Bildinhalte** → GPT-4o-mini (nur, wenn Bildfunktion in AiS-chat freigeschaltet ist)
- **Besonders datenschutzsensibel / EU-nah** → Llama-Modelle oder Mistral Nemo Instruct

Checkliste nach Einsatzzweck

- ☐ Ich brauche **schnelle Kurzantworten**, Umformulierungen, Stichpunkte → GPT-5 nano / GPT-5 mini / Llama-3.1-8B
- ☐ Ich will **anspruchsvolle Analysen, hochwertige Erklärtexpte, kreative Aufgaben** → GPT-5; alternativ Llama-3.1-405B oder Llama-3.3-70B
- ☐ Ich löse **Mathe-/Naturwissenschafts-Probleme oder programmiere** → o3-mini (klar strukturierte Schritt-für-Schritt-Lösungen)
- ☐ Ich arbeite mit **Bildern/Arbeitsblättern** → GPT-4o-mini (falls aktiv)
- ☐ Ich brauche **EU/On-Prem-Optionen** → Llama-Varianten oder Mistral Nemo Instruct
- ☐ **Große Klasse/kleines Budget** → zuerst GPT-5 nano oder Llama-3.1-8B, bei Bedarf auf mini/70B hochschalten
- ☐ **Arbeitsblätter, saubere Gliederung, gute Sprache** → GPT-5 mini oder Llama-3.3-70B
- ☐ **Robuste Allround-Sitzung (lange Aufgabenfolgen)** → GPT-5 oder Llama-3.3-70B

Praktische Auswahlstrategie

- **Erst schnell, dann tief:** Starte mit GPT-5 mini oder Llama-3.1-8B. Wenn Qualität/Tiefe nicht reicht, wechsle zu GPT-5 oder Llama-3.3-70B.
- **Fachspezifisch denken:** Für MINT/Code direkt o3-mini testen; für Bildaufgaben GPT-4o-mini.
- **Datenschutzbedarf prüfen:** Wenn besonders sensibel, bevorzuge Llama-Modelle oder Mistral Nemo Instruct.

Tipp für bessere Ergebnisse: Gib dem Modell eine **Rolle** (z. B. „Erkläre wie für Schülerinnen und Schüler der Klasse 8“), nenne **Format/Umfang** und bitte um **Schritt-für-Schritt-Begründungen**.