Intelligenz Welt

GLM-5.3

KI-Modellportrait · Quellenstand 2026-10-31



Z.ai-Anbieterlogo, kein modellspezifisches Logo · Logo-Quelle / Rechtehinweis. Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

GLM-5.3 ist das Flaggschiff-Sprachmodell von Z.ai: 744 Milliarden Parameter (40 Milliarden aktiv), 1-Million-Token-Kontext und offene Gewichte. Es nutzt dasselbe Basismodell wie GLM-5.2 - alle Verbesserungen stammen laut Z.ai aus dem Post-Training und zeigen sich vor allem bei Coding, Agent-Aufgaben und Schwachstellenanalyse.

Quellen geprüft am 31. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: Z.ai →

DAS MODELL EINFACH ERKLÄRT

Z.ai-Flaggschiff mit offenen Gewichten: 744B-MoE, 1M Kontext und ungewöhnlich starke Schwachstellenanalyse

GLM-5.3 ist das Flaggschiff-Sprachmodell des chinesischen Anbieters Z.ai: ein Mixture-of-Experts-Modell mit 744 Milliarden Parametern (40 Milliarden aktiv), 1 Million Token Kontext und frei herunterladbaren Gewichten. Es übernimmt das Basismodell von GLM-5.2 unverändert und holt seine Gewinne laut Z.ai komplett aus dem Post-Training - vor allem bei Coding, langen Agent-Aufgaben und Cybersecurity-Analysen. Die API ist OpenAI- und Anthropic-kompatibel, Preise: 1,40 USD je 1 Mio. Eingabe- und 4,40 USD je 1 Mio. Ausgabe-Token. Reasoning ist immer eingeschaltet und in drei Stufen dosierbar; derzeit verarbeitet das Modell nur Text.

Offizielle Dokumentation: GLM-5.3

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich GLM-5.3?

Für Entwicklerteams, die ein offenes Modell mit sehr großem Kontext für Coding-Agenten suchen, ist GLM-5.3 eine der stärksten Optionen am Markt - nach Hersteller-Benchmarks das beste Open-Weights-Modell für Coding. Der 1M-Kontext erlaubt komplette Repositories im Prompt. Wer schon GLM-5.2 nutzt, muss beim Umstieg vor allem den Reasoning-

Parameter anpassen: Abschalten ist nicht mehr möglich, stattdessen reasoning_effort auf low setzen. Für Bildverarbeitung ist GLM-5.3 die falsche Wahl (Schwestermodell GLM-5.3-Flash ist multimodal). Beim Selbsthosten gilt: 744B-Parameter-Klasse erfordert erheblichen Speicher, auch als FP8 - für die meisten Teams ist die API oder der GLM Coding Plan der pragmatische Weg.

Post-Training statt neuer Basis

GLM-5.3 ist ein ungewöhnliches Release: Z.ai hat das vortrainierte Basismodell von GLM-5.2 nicht angefasst und alle Verbesserungen über skaliertes Post-Training erzielt. Dafür baut das Unternehmen synthetische Aufgabenumgebungen, die reale Expertentätigkeiten nachbilden - bis hin zu mehrtägigen ML-Infrastruktur-Aufgaben mit Compute-Clustern und Codebases. Ein Judge-Agent prüft jede Umgebung auf Lösbarkeit, bevor sie ins Reinforcement Learning geht. Laut Z.ai zeigt sich der Effekt in beiden Richtungen: bessere Ergebnisse bei weniger Output-Token. Auf dem hauseigenen Z.ai Code Bench erreicht GLM-5.3 bei maximaler Reasoning-Stufe 34,5 Prozent bei rund 75.000 Output-Token pro Aufgabe, GLM-5.2 kam auf 23,4 Prozent bei 96.000. Herstellerangabe, kein unabhängiger Test.

Offizielle Dokumentation: GLM-5.3

Coding- und Agent-Benchmarks

Die von Z.ai gemeldeten Sprünge gegenüber GLM-5.2 sind erheblich: Terminal-Bench 3.0 von 4,6 auf 28,3, DeepSWE v1.1 von 46,2 auf 66,9, Agents' Last Exam (CLI) von 23,8 auf 28,5. Gegen geschlossene Frontier-Modelle liegt GLM-5.3 je nach Benchmark vorn oder knapp dahinter - auf Terminal-Bench 3.0 etwa hinter Claude Fable 5 (33,7) und GPT-5.6 Sol (34,6), auf CyberGym vorn. Alle Werte stammen aus den Z.ai-Unterlagen (Model Card und Doku) und wurden von uns nicht nachgemessen.

Offizielle Dokumentation: GLM-5.3

Cyber-Fähigkeiten als Schwerpunkt

Auffälligstes Merkmal des Releases ist die Sicherheitskompetenz: Z.ai hat Schwachstellen-Daten und -Umgebungen in den Post-Training-Mix aufgenommen und beschreibt die Entwicklung selbst als schneller als erwartet. Auf CyberGym (White-Box-Schwachstellensuche) erreicht GLM-5.3 mit 84,5 Prozent den Bestwert des Benchmarks. Auf ExploitBench (54,4 Prozent) und ExploitGym (105 Aufgaben in 2h / 130 in 6h) liegt es deutlich über GLM-5.2, aber hinter den geschlossenen Modellen Mythos 5 und GPT-5.6 Sol. In Praxistests mit Security-Teams in China soll das Modell 2.436 Schwachstellen in 269 Projekten gefunden haben, darunter 1.097 mittleren oder hohen Schweregrads; Z.ai dokumentiert den Prozess in einem öffentlichen Security Disclosure Ledger. Für deutsche Anwender relevant: Diese Stärke macht das Modell für Defensive-Security- und Code-Audit-Workflows interessant, verlangt aber saubere Governance im eigenen Einsatz.

Offizielle Dokumentation: GLM-5.3

Zugang: API, Coding Plan oder eigene Hardware

Drei Wege führen zu GLM-5.3. Erstens die API unter api.z.ai mit OpenAI-Chat-Completions-, OpenAI-Response- und Anthropic-Messages-Protokoll - bestehende Clients lassen sich meist durch Austausch von Base-URL und Modell-ID weiterverwenden. Zweitens der GLM Coding Plan: ein Abo mit punktebasierter Quote, bei dem Aufrufe außerhalb der Stoßzeiten (inklusive Wochenende) nur 50 Prozent der Standard-Punkte verbrauchen; Coding-Plan-Kunden nutzen derzeit nur das OpenAI-Chat-Completions-kompatible Protokoll. Drittens Selbsthosting: Die Gewichte liegen als FP8 und BF16 auf Hugging Face und ModelScope, unterstützt werden SGLang, vLLM, TokenSpeed, Transformers, KTransformers und Unsloth, auch auf Ascend-NPUs. Der API-Preis beträgt 1,40 USD pro 1 Mio. Eingabe-Token (0,26 USD gecacht) und 4,40 USD pro 1 Mio. Ausgabe-Token.

Offizielle Dokumentation: GLM-5.3

Grenzen und Einordnung

GLM-5.3 verarbeitet derzeit ausschließlich Text - wer Bilder, Screenshots oder GUIs braucht, muss zu GLM-5.3-Flash greifen, das zudem eine komplett andere (hybride) Architektur und MIT-Lizenz hat. Reasoning lässt sich nicht deaktivieren; wer sehr kurze, billige Antworten will, zahlt den Denk-Overhead mit, auch auf Stufe low. Die Lizenz ist großzügig, aber keine Standard-Open-Source-Lizenz: Die MaaS-Klausel ab 10 Mrd. USD Umsatz betrifft nur sehr große Anbieter, sollte aber in einer Lizenzprüfung auftauchen. Und wie bei allen Hersteller-Benchmarks gilt: Die gemeldeten Werte sind Angaben von Z.ai, keine unabhängigen Messungen dieser Redaktion.

Offizielle Dokumentation: GLM-5.3

Ein- und Ausgaben im Überblick.

Eingabe

Text

Ausgabe

Text

Kontext / Grenzen

1.000.000 Token

Maximale Ausgabe

128.000 Token

Was kostet GLM-5.3 über die API?

API: 1,40 USD je 1 Mio. Eingabe-Token, 0,26 USD gecachte Eingabe, 4,40 USD je 1 Mio. Ausgabe-Token (Z.ai-Preisliste). GLM Coding Plan mit punktebasierter Quote und 50 Prozent Off-Peak-Rabatt separat. Selbsthosting kostenfrei unter GLM-5.3-Lizenz (MaaS-Klausel beachten).

Vor dem Einsatz prüfen

- Von GLM-5.2: Modell-ID auf glm-5.3 ändern und thinking.type: disabled durch thinking.type: enabled mit reasoning_effort: low ersetzen - sonst schlägt die Anfrage fehl. Für Coding-Aufgaben empfiehlt Z.ai reasoning_effort: max.
- GLM-Coding-Plan-Kunden (auch mit abgelaufenem Abo) können die Modell-API derzeit nur über das OpenAI-Chat-Completions-kompatible Protokoll nutzen.
- Nicht mit GLM-5.3-Flash verwechseln: Flash hat eine andere Architektur (hybrid, multimodal, 320B/18B), eigene Modell-IDs und MIT-Lizenz; Specs sind nicht übertragbar.

Häufige Fragen zu GLM-5.3

Ist GLM-5.3 Open Source?

Die Gewichte sind frei verfügbar (Open Weights) auf Hugging Face und ModelScope, unter der eigenen GLM-5.3-Lizenz. Diese erlaubt kommerzielle Nutzung, Modifikation und Weitergabe - ähnlich wie MIT - verlangt aber von Model-as-a-Service-Anbietern mit mehr als 10 Mrd. USD Jahresumsatz eine Sicherheitsprüfung durch Z.AI. Streng genommen ist es keine OSI-approved Open-Source-Lizenz. Achtung: GLM-5.3-Flash steht dagegen unter MIT.

Kann GLM-5.3 Bilder verarbeiten?

Nein. GLM-5.3 unterstützt laut Z.ai-Dokumentation derzeit nur Texteingaben. Für multimodale Aufgaben (Bilder, GUI-Verständnis) ist das Schwestermodell GLM-5.3-Flash vorgesehen.

Wie groß ist das Kontextfenster von GLM-5.3?

1 Million Token Kontext bei bis zu 128.000 Token maximaler Ausgabe. Eingabe- und Ausgabelimit sind getrennte Grenzen.

Lässt sich das Reasoning abschalten?

Nein. GLM-5.3 arbeitet immer mit Reasoning. Die Intensität steuert der Parameter reasoning_effort mit den Stufen low, high und max (Standard: max). Anwendungen, die bisher thinking.type: disabled nutzen, müssen auf enabled mit reasoning_effort: low umstellen, sonst schlägt die Anfrage fehl.

Was kostet GLM-5.3 in der API?

1,40 USD pro 1 Mio. Eingabe-Token, 0,26 USD für gecachte Eingaben und 4,40 USD pro 1 Mio. Ausgabe-Token laut Z.ai-Preisliste. Alternativ gibt es den GLM Coding Plan mit punktebasierter Quote und halbiertem Punkteverbrauch außerhalb der Stoßzeiten.

Kann ich GLM-5.3 selbst hosten?

Ja, die Gewichte (FP8 und BF16) stehen auf Hugging Face und ModelScope. Unterstützt werden SGLang, vLLM, TokenSpeed, Transformers, KTransformers und Unsloth. Mit 744 Milliarden Parametern braucht das Modell allerdings sehr viel Speicher - auch quantisiert bleibt es Infrastruktur für Server, nicht für Laptops.

Alle Antworten beziehen sich auf die verlinkte Herstellerdokumentation und den angegebenen Stand. Kein unabhängiger Benchmark, keine Garantie für Verfügbarkeit oder Ergebnisqualität.

Quellen & Stand

1. Z.ai Developer Docs: GLM-5.3 (Overview, Specs, API)
   Geprüft: 2026-10-31
2. Z.ai Developer Docs: Pricing
   Geprüft: 2026-10-31
3. Z.ai Developer Docs: New Released (Release-Notes)
   Geprüft: 2026-10-31
4. Hugging Face: zai-org/GLM-5.3 (Model Card)
   Geprüft: 2026-10-31
5. Hugging Face: GLM-5.3 License (Volltext im Repo zai-org/GLM-5.3)
   Geprüft: 2026-10-31
6. Z.ai API Platform (Produktseite)
   Geprüft: 2026-10-31

Redaktionelle Zusammenfassung mit KI-Unterstützung, verantwortet von Philipp Paulik. Kein eigener Leistungstest. Änderbare Anbieterangaben gelten für den angegebenen Quellenstand.

Intelligenz Welt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31
Vollständiges Webportrait: https://intelligenz-welt.de/ki-modelle/glm-5-3/
Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. Impressum · Transparenz