

GLM-5.3

KI-Modellportrait · Quellenstand 2026-10-31



Z.ai-Anbieterlogo, kein modellspezifisches Logo · [Logo-Quelle / Rechteinweis](#). 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](#) →

Technische Einordnung

GLM-5.3 ist die aktuelle Spitze der GLM-Familie und richtet sich an Entwicklerinnen und Entwickler, die komplexe Software-Engineering- und Agenten-Aufgaben über einen riesigen Kontext von 1 Million Token abarbeiten wollen. Ungewöhnlich: Z.ai hat gegenüber GLM-5.2 das vortrainierte Basismodell unverändert gelassen und alle Gewinne über skaliertes Post-Training (Reinforcement Learning in eigens synthetisierten, praxisnahen Aufgabenumgebungen) erzielt. Nach Herstellerangaben springt der Wert auf Terminal-Bench 3.0 von 4,6 auf 28,3 Punkte, auf DeepSWE v1.1 von 46,2 auf 66,9. Besonders betont Z.ai die Cyber-Fähigkeiten: Auf CyberGym (Schwachstellensuche) erreicht GLM-5.3 mit 84,5 Prozent den bislang besten Wert des Benchmarks; in Praxistests mit chinesischen Security-Teams soll das Modell 2.436 Schwachstellen in 269 Projekten gefunden haben, davon 1.097 mittleren bis hohen Schweregrads. Das Modell denkt immer mit - Reasoning lässt sich nicht abschalten, nur in drei Stufen (low, high, max) dosieren. Derzeit verarbeitet GLM-5.3 nur Texteingaben; für Bilder und GUIs ist das Schwestermodell GLM-5.3-Flash vorgesehen. Die Gewichte stehen unter der eigenen GLM-5.3-Lizenz frei zum Download, die API ist OpenAI- und Anthropic-kompatibel.

Zugang & Nutzung

API über api.z.ai (OpenAI-Chat-Completions- und Response-Protokoll sowie Anthropic-Messages-Protokoll, Modell-ID glm-5.3), GLM Coding Plan mit punktebasierter Quote (Off-Peak inkl. Wochenende 50 Prozent Punkteverbrauch), offene Gewichte (FP8 und BF16) auf Hugging Face und ModelScope zum Selbsthosten. Laut Plattform-Seite zusätzlich über Drittanbieter-Clouds verfügbar.

GLM-5.3-Lizenz (eigene Lizenz von Z.AI): weitgehend MIT-artige Freiheiten inklusive kommerzieller Nutzung, Fine-Tuning und Weitergabe; einzige nennenswerte Bedingung: Betreiber eines Model-as-a-Service-Geschäfts mit mehr als 10 Mrd. US-Dollar Jahresumsatz benötigen vor kommerzieller Nutzung eine Sicherheitsprüfung durch Z.AI. Keine Gewährleistung.

Grenzen & richtig lesen

Aktuell nur Texteingabe (keine Bilder/Dateien). Reasoning ist immer aktiv und nicht abschaltbar - Migration von thinking.type disabled schlägt sonst fehl. Benchmark-Werte stammen vom Hersteller und sind keine unabhängigen Messungen. Gewichte liegen als FP8 (Standard) und BF16 vor; für Selbsthosting ist sehr viel Speicher nötig (744B-Parameter-Klasse). GLM-5.3-Flash nutzt eine eigene Architektur und eigene (MIT-)Lizenz - nicht verwechseln.

Benchmark-Belege

Für dieses Portrait ist noch kein ausreichend zugeordnetes Ergebnis erfasst. Fehlende Werte sind keine Null. Wir übernehmen keine Werte aus ähnlich benannten Modell- oder Benchmark-Versionen.

Steckbrief

Anbieter

Z.ai

Modell-ID (API)

glm-5.3

Veröffentlichung

18.08.2026 (API/Ankündigung laut Z.ai-Release-Notes)

Architektur

Mixture-of-Experts, 744 Mrd. Parameter gesamt, 40 Mrd. aktiv pro Token

Basismodell

Identisch mit GLM-5.2 - alle Gewinne laut Z.ai aus Post-Training

Kontextfenster

1.000.000 Token

Maximale Ausgabe

128.000 Token

Eingabe-Modalität

Derzeit nur Text

Reasoning

Immer aktiv; Stufen low / high / max (Standard: max)

API-Protokolle

OpenAI Chat Completions, OpenAI Response, Anthropic Messages

Preis (API)

1,40 USD Eingabe / 0,26 USD gecachte Eingabe / 4,40 USD Ausgabe je 1 Mio. Token

Offene Gewichte

Hugging Face zai-org/GLM-5.3 (FP8, Standard) und BF16-Variante; auch ModelScope

Lizenz

GLM-5.3-Lizenz (proprietäre Open-Weights-Lizenz mit MaaS-Klausel ab 10 Mrd. USD Umsatz)

Deployment-Frameworks

SGLang, vLLM, TokenSpeed, Transformers, KTransformers, Unsloth; Ascend-NPU-Unterstützung

Hersteller-Benchmarks

Terminal-Bench 3.0: 28,3 | DeepSWE v1.1: 66,9 | Agents' Last Exam (CLI): 28,5 | CyberGym: 84,5% | ExploitBench: 54,4% | HLE w/ Tools: 62,5

2.436 gefundene Schwachstellen in 269 realen Projekten, davon 1.097 mittel/hoch; laufender Security Disclosure Ledger

Angaben laut den unten verlinkten Primärquellen.

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.

KI Fachblatt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31

Vollständiges Webportrait: <https://kifachblatt.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](#)