

GLM-5.3-Flash

KI-Modellportrait · Quellenstand 2026-10-31

GL

Monogramm-Fallback (Initialen GL); kein geprüftes Z.ai-Markenlogo im Bestand. Geprüfter Lead: 'Zhipu AI.png' auf Wikimedia Commons ist als Public Domain gekennzeichnet, zeigt aber die alte Marke Zhipu AI statt des aktuellen Z.ai-Logos, daher nicht übernommen. Monogramm-Darstellung, kein Logo-Asset übernommen.

Offenes Multimodalmmodell von Z.ai mit 320B Parametern (18B aktiv), 1-Million-Token-Kontext und API-Preisen von 0,15 \$ Eingabe / 0,50 \$ Ausgabe pro Million Token; unter MIT-Lizenz auf Hugging Face verfügbar.

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

Entwickler: [Z.ai](#) →

Technische Einordnung

GLM-5.3-Flash ist das erste nativ multimodale Modell der GLM-5-Serie von Z.ai (Herstellerangabe). Es verarbeitet Text, Bilder, Videos und Dateien als Eingabe und liefert Text, strukturierte Ausgaben und Werkzeugaufrufe zurück. Die Architektur kombiniert erstmals in der GLM-Reihe lineare und spärliche Aufmerksamkeit und nutzt sogenannte Manifold-Constrained Hyper-Connections (mHC); nach Herstellerangaben sinken dadurch Attention-Rechenaufwand um Faktor 3,0 und KV-Cache-Bedarf um Faktor 4,4 gegenüber GLM-5.3. Von den 320 Milliarden Parametern sind pro Token nur 18 Milliarden aktiv, verteilt auf 45 Layer; das Pre-Training umfasste laut Hersteller 30 Billionen Token multimodaler Daten. Der praktische Schwerpunkt liegt auf Agent- und Coding-Arbeitslasten mit visueller Rückkopplung: Das Modell kann Oberflächen, gerenderte Ergebnisse und Interaktionsfeedback beobachten und seine Arbeit iterativ verbessern, etwa bei Frontend- und Spielentwicklung, Blender-3D-Szenen oder Computer-Use-Abläufen. Z.ai bewirbt außerdem Büro- und Wissensarbeit (Recherche, Finanzanalysen, Dokumentenverarbeitung mit PPTX-, PDF-, DOCX- und XLSX-Ausgaben) als Einsatzfeld. Diese Positionierung stammt aus Herstellerquellen und ist als solche eingeordnet. Das Kontextfenster umfasst 1 Million Token bei bis zu 128.000 Ausgabetoken. Der Denkmodus ist immer aktiv und lässt sich über `reasoning_effort` (low, high, max) steuern; er kann nicht abgeschaltet werden. Neben der Basisvariante `glm-5.3-flash` existiert die schnellere Variante `glm-5.3-flashx` (Herstellerangabe: 200 Token/s) sowie die juristische Spezialvariante `glm-5.3-flashx-legal`, deren API-Zugang nur ausgewählten Unternehmenskunden auf Antrag offensteht. Zugang gibt es über die Z.ai-API (Chat-Completion-API), den GLM Coding Plan mit punktebasierter Quote (für GLM-5.3-Flash laut Hersteller die dreifache Quote von GLM-5.3) sowie über offene Gewichte unter MIT-Lizenz auf Hugging Face für den lokalen Betrieb mit SGLang, vLLM, TokenSpeed, Transformers, KTransformers oder Unsloth. Mehrere Drittanbieter (unter anderem Together AI, Fireworks, DeepInfra, Baseten, Novita, Featherless) listen das Modell ebenfalls. Veröffentlicht wurde es im August 2026; vor dem Release testete Z.ai es nach eigenen Angaben anonym unter dem Namen `ox-alpha` auf OpenCode und OpenRouter.

Zugang & Nutzung

Z.ai-API (Modellcode glm-5.3-flash, Varianten glm-5.3-flashx und glm-5.3-flashx-legal), GLM Coding Plan (Punktequote, Off-Peak- und Wochenendnutzung mit 50 % der Standardpunkte; FlashX zum Prüfzeitpunkt noch nicht im Plan), offene Gewichte auf Hugging Face (zai-org/GLM-5.3-Flash, MIT), Inference-Provider laut Model Card: Zai, Together AI, Novita, Fireworks, Featherless AI, DeepInfra, Baseten. GLM-5.3-FlashX-Legal nur für ausgewählte Unternehmenskunden auf Antrag.

Offene Gewichte unter MIT-Lizenz (Copyright (c) 2026 Z.AI Co., Ltd), LICENSE-Datei im Hugging-Face-Repo geprüft: kommerzielle Nutzung, Vervielfältigung, Änderung und Weitergabe erlaubt, Copyright- und Lizenzhinweis in Kopien aufzunehmen. API- und Abo-Nutzung richtet sich zusätzlich nach den Z.ai-Plattformbedingungen.

Grenzen & richtig lesen

Denkmodus (thinking) ist immer aktiv und kann nicht deaktiviert werden; nur Steuerung über reasoning_effort (low, high, max, Standard max). Die maximale Ausgabelänge beträgt 128.000 Token und liegt damit unter dem Kontextfenster. KV-Cache-Bedarf liegt laut Hersteller-Vergleich leicht über Kimi-K3 und DeepSeek-V4-Flash. Benchmark-Werte (u.a. DeepSWE, AutomationBench, OSWorld, Z.ai Code Bench) stammen aus Herstellerquellen und teils hauseigenen Evaluierungen, nicht aus unabhängigen Tests. GLM-5.3-FlashX ist zum Prüfzeitpunkt nicht im Coding Plan verfügbar; FlashX-Legal nur für ausgewählte Unternehmenskunden. Preise in USD, Cache-Speicher nur zeitlich begrenzt kostenlos (Herstellerangabe).

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 (Zhipu AI)

Veröffentlichung

August 2026 (Ankündigungsblog 26.08.2026)

Gesamtparameter

320 Milliarden

Aktivierte Parameter pro Token

18 Milliarden

Layer

45

Architektur

Hybrid aus linearer und spärlicher Aufmerksamkeit, Manifold-Constrained Hyper-Connections (mHC), IndexPool (Herstellerangabe)

Pre-Training

30 Billionen Token multimodaler Daten (Herstellerangabe)

Kontextfenster

1.000.000 Token

Maximale Ausgabe

128.000 Token

Eingabemodalitäten

Text, Bilder, Videos, Dateien (native multimodale Eingabe)

Ausgabemodalitäten

Text, strukturierte Ausgaben (u.a. JSON), Function Calling

API-Modellcodes

glm-5.3-flash, glm-5.3-flashx, glm-5.3-flashx-legal

API-Preis (glm-5.3-flash)

0,15 USD Eingabe / 0,03 USD gecachte Eingabe / 0,50 USD Ausgabe pro 1 Mio. Token

API-Preis (glm-5.3-flashx)

0,37 USD Eingabe / 0,075 USD gecachte Eingabe / 1,25 USD Ausgabe pro 1 Mio. Token

Antwortgeschwindigkeit FlashX

200 Token/s (Herstellerangabe)

Lizenz der Gewichte

MIT (Copyright (c) 2026 Z.AI Co., Ltd)

Denkmodus

Immer aktiv, Steuerung über reasoning_effort (low, high, max)

Benchmarks (Herstellerangaben)

u.a. DeepSWE v1.1 63,4 (GLM-5.2: 46,2), AutomationBench v1.0.6 48,8 (GLM-5.2: 26,2), Terminal-Bench 2.1 84,3; hauseigene Z.ai Code Bench v1.0: 29,0 vs. Claude Opus 4.8 29,5 bei max effort

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [Z.ai Developer Documentation: GLM-5.3-Flash/FlashX - Overview](#)

Geprüft: 2026-10-31

2. [Z.ai Developer Docs: Pricing](#)

Geprüft: 2026-10-31

3. [Z.ai Blog: GLM-5.3-Flash: Frontier Intelligence, Flash Cost](#)

Geprüft: 2026-10-31

4. [Hugging Face: zai-org/GLM-5.3-Flash \(Model Card und LICENSE, MIT\)](#)

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-flash>

Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. [Impressum](#) · [Transparenz](#)