Intelligenz Welt

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 Multimodalmodell 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 →

DAS MODELL EINFACH ERKLÄRT

GLM-5.3-Flash: offenes Multimodalmodell von Z.ai mit 1M Kontext zum Flash-Preis

GLM-5.3-Flash ist ein offenes, nativ multimodales Modell von Z.ai mit 320 Milliarden Parametern (18 Milliarden aktiv), einem Kontextfenster von 1 Million Token und bis zu 128.000 Ausgabetoken. Es kostet über die Z.ai-API 0,15 USD pro Million Eingabetoken und 0,50 USD pro Million Ausgabetoken und steht zusätzlich als offene Gewichte unter MIT-Lizenz auf Hugging Face. Nach Herstellerangaben übertrifft es GLM-5.2 in Coding- und Agentic-Benchmarks bei etwa einem Zehntel des Preises; diese Werte sind Herstellerangaben.

Offizielle Dokumentation: GLM-5.3-Flash

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich GLM-5.3-Flash?

GLM-5.3-Flash richtet sich an Nutzer, die ein günstiges Modell mit sehr großem Kontextfenster und Bild-/Video-Verständnis für Agent- und Coding-Arbeitslasten suchen: Frontend- und Spielentwicklung mit visueller Selbstprüfung, Computer Use, Dokumenten- und Videoanalyse sowie Büro-Workflows mit Datei-Ausgaben (PPTX, PDF, DOCX, XLSX). Über den GLM Coding Plan ist es nach Herstellerangabe mit der dreifachen Quote von GLM-5.3 nutzbar; die schnellere

Variante FlashX (200 Token/s) eignet sich für interaktive Anwendungen, ist zum Prüfzeitpunkt aber nicht im Plan enthalten.

Positionierung: Frontier-Anspruch zum Flash-Preis

Z.ai stellt GLM-5.3-Flash als Modell vor, das höhere Intelligenz als GLM-5.2 zu deutlich niedrigeren Kosten liefert. Als Beleg nennt der Hersteller den Artificial Analysis Intelligence Index v4.1.1 mit einem Score von 57 bei 0,045 USD pro Aufgabe (rabattiert) sowie sechs Coding- und Agentic-Benchmarks, in denen das Modell GLM-5.2 durchgehend übertrifft und sich Claude Opus 4.8 annähert. Diese Angaben stammen aus dem Ankündigungsblog und der Model Card, also aus Herstellerquellen; unabhängige Bestätigungen liegen der Redaktion zum Prüfzeitpunkt nicht vor. Vor der Veröffentlichung lief das Modell nach Herstellerangabe anonym als ox-alpha auf OpenCode und OpenRouter und wurde dort das meistgenutzte Modell der Woche.

Offizielle Dokumentation: GLM-5.3-Flash

Architektur: gebaut für niedrige Inferenzkosten

Die zentrale Neuerung ist die Kombination aus linearer Aufmerksamkeit (lokale Abhängigkeiten über Zustandsmodellierung) und spärlicher Aufmerksamkeit (globaler Kontext über einen leichtgewichtigen Indexer). Dazu kommen Manifold-Constrained Hyper-Connections für bessere Skalierungseffizienz und IndexPool zur Senkung von Latenz und Speicherbedarf bei 1M-Kontext. Gegenüber GLM-5.3 halbiert sich die Zahl der aktivierten Parameter (18B statt 32B) bei nur 45 statt 92 Layern. Im Hersteller-Vergleich mit GLM-5.3, DeepSeek-V4-Flash und Kimi-K3 hat GLM-5.3-Flash den niedrigsten Attention-Rechenaufwand pro Token; der KV-Cache bleibt etwas größer als bei den beiden Vergleichsmodellen.

Offizielle Dokumentation: GLM-5.3-Flash

Visuelles Coding und Computer Use

Visuelle Fähigkeiten sind direkt in den Coding-Loop eingebaut: Das Modell betrachtet Oberflächen, gerenderte Ergebnisse und Interaktionsfeedback und verbessert seine Ausgaben iterativ. Die Dokumentation zeigt Workflows von Screenshots oder Bildschirmaufnahmen zu vollständigen Frontend-Projekten, spielbaren Godot-Prototypen, Blender-3D-Szenen, CAD-Rekonstruktionen mit build123d und Computer-Use-Abläufen (BUA/CUA), in denen das Modell Anwendungen bedient, nachbaut und die Nachbauten gegen das Original testet. Diese Beispiele stammen aus den Best Practices des Herstellers und beschreiben empfohlene Vorgehensweisen, keine garantierten Ergebnisse.

Offizielle Dokumentation: GLM-5.3-Flash

Wissensarbeit jenseits von Coding

Z.ai positioniert das Modell auch für Office- und Finanzarbeit: Recherche mit Quellennachweis, Trennung von Fakten, Annahmen und Ableitungen, Bewertungsmodelle und fertige PPTX-, PDF-, DOCX- und XLSX-Dateien mit visueller Selbstkontrolle (Überlauf, Ausrichtung, Überlappungen). Hinzu kommt Videoverständnis für Schnitt-Workflows, etwa Sprecherzuordnung und Untertitelung langen Filmmaterials. Auch hier gilt: Hersteller-Positionierung, im Einzelfall zu prüfen.

Offizielle Dokumentation: GLM-5.3-Flash

Varianten: FlashX und FlashX-Legal

GLM-5.3-FlashX ist die schnellere Variante mit 200 Token/s (Herstellerangabe) und höheren Preisen (0,37/1,25 USD pro 1 Mio. Token); sie ist zum Prüfzeitpunkt nicht im Coding Plan verfügbar. GLM-5.3-FlashX-Legal ist eine juristische Spezialvariante mit Legal-Mid- und Post-Training für Recherche, Vertragsprüfung und Dokumentenerstellung; der API-Zugang ist auf ausgewählte Unternehmenskunden per Antrag beschränkt. Auf dem hauseigenen Harvey-LAB-Benchmark nennt der Hersteller eine All-Criteria-Pass-Rate von 19,2 %, ein Plus von 12,5 Prozentpunkten gegenüber dem Basismodell (Herstellerangabe).

Offizielle Dokumentation: GLM-5.3-Flash

Betrieb auf chinesischen KI-Chips

Nach Herstellerangaben wird GLM-5.3-Flash produktiv auf einem großen Cluster chinesischer KI-Chips mit einer eigenen, auf SGLang aufbauenden Inferenz-Engine ausgeliefert, unter anderem mit W8A8-Quantisierung, hybriden INT8/FP8/BF16-Caches und einer EPD-Architektur (getrennte Worker-Pools für Encoding, Prefill und Decode). Gegenüber der anfänglichen Basislinie auf derselben Hardware meldet Z.ai eine Verdreifachung der End-to-End-Serving-Leistung. Interessant am Rand: Ein GLM-5.3-gesteuerter Infrastruktur-Agent half laut Blog bei der Kernel-Entwicklung und Optimierung des Serving-Stacks.

Offizielle Dokumentation: GLM-5.3-Flash

Ein- und Ausgaben im Überblick.

Eingabe

Text, Bilder, Videos, Dateien

Ausgabe

Text, strukturierte Ausgaben (JSON), Tool-Aufrufe

Kontext / Grenzen

1.000.000 Token

Maximale Ausgabe

128.000 Token

Was kostet GLM-5.3-Flash über die API?

Preise der Z.ai-API pro 1 Mio. Token (USD, Stand 31.10.2026): GLM-5.3-Flash 0,15 Eingabe / 0,03 gecachte Eingabe / 0,50 Ausgabe; GLM-5.3-FlashX 0,37 / 0,075 / 1,25; Cache-Speicher zeitlich begrenzt kostenlos (Herstellerangabe). Web-Search-Tool 0,01 USD pro Nutzung. GLM Coding Plan mit Punktequote, Off-Peak und Wochenenden 50 % der Standardpunkte. Preise können sich ändern; maßgeblich ist die Z.ai-Preisseite.

Vor dem Einsatz prüfen

- Von GLM-5.2/GLM-5.3: Text-Parameter sind laut Dokumentation mit GLM-5.3 konsistent; im Wesentlichen Modellcode auf glm-5.3-flash ändern. Beachten: thinking kann nicht deaktiviert werden, reasoning_effort steuert das Denkbudget (Standard max).
- Chat-Template: clear_thinking ist standardmäßig false; für Chat-Szenarien empfiehlt der Hersteller, es ausdrücklich zu setzen (API-Empfehlung: clear_thinking false für Kontinuität des Reasonings).
- Für Streaming empfiehlt die Dokumentation stream: true und tool_stream: true; Bilder über image_url-Content-Blöcke als URL (empfohlen) oder Base64-Data-URL, mehrere Bilder über mehrere Blöcke.

- Im GLM Coding Plan bringt GLM-5.3-Flash laut Hersteller die dreifache nutzbare Quote von GLM-5.3; FlashX ist zum Prüfzeitpunkt nicht im Plan.

Häufige Fragen zu GLM-5.3-Flash

Was kostet GLM-5.3-Flash?

Über die Z.ai-API 0,15 USD pro Million Eingabetoken und 0,50 USD pro Million Ausgabetoken; gecachte Eingaben kosten 0,03 USD pro Million Token. Die schnellere Variante GLM-5.3-FlashX kostet 0,37 beziehungsweise 1,25 USD. Die offenen Gewichte stehen unter MIT-Lizenz kostenlos auf Hugging Face; lokaler Betrieb verursacht eigene Infrastrukturkosten.

Ist GLM-5.3-Flash Open Source?

Die Modellgewichte sind öffentlich auf Hugging Face (zai-org/GLM-5.3-Flash) unter MIT-Lizenz verfügbar. Das erlaubt kommerzielle Nutzung, Änderung und Weitergabe unter Aufnahme des Copyright- und Lizenzhinweises. Z.ai bezeichnet das Modell als Open-Source-Frontier-Modell; streng genommen handelt es sich um offene Gewichte, da Trainingsdaten und Trainingscode nicht vollständig offengelegt sind.

Wie groß ist das Kontextfenster von GLM-5.3-Flash?

Eine Million Token Eingabekontext bei maximal 128.000 Ausgabetoken. Der Denkmodus ist immer aktiv und lässt sich über den Parameter reasoning_effort in den Stufen low, high und max steuern.

Was ist der Unterschied zwischen GLM-5.3-Flash und GLM-5.3-FlashX?

FlashX ist die auf Geschwindigkeit optimierte Variante mit laut Hersteller 200 Token/s und höheren Preisen (0,37 USD Eingabe, 1,25 USD Ausgabe pro Million Token). Zum Prüfzeitpunkt ist FlashX noch nicht im GLM Coding Plan verfügbar. Eine dritte Variante, FlashX-Legal, ist für juristische Workflows trainiert und nur für ausgewählte Unternehmenskunden zugänglich.

Kann GLM-5.3-Flash Bilder und Videos verarbeiten?

Ja. Das Modell nimmt nativ Bilder, Videos und Dateien als Eingabe entgegen (über image_url-Content-Blöcke in der API) und gibt Text, strukturierte Daten und Werkzeugaufrufe aus. Z.ai beschreibt visuelles Coding, Dokumentenanalyse und Videoverständnis als zentrale Einsatzfelder.

Kann ich GLM-5.3-Flash lokal betreiben?

Ja, die Gewichte auf Hugging Face unterstützen laut Model Card SGLang, vLLM, TokenSpeed, Transformers, KTransformers und Unsloth. Wegen der 320 Milliarden Gesamtparameter ist der lokale Betrieb trotz nur 18 Milliarden aktivierter Parameter anspruchsvoll; als Alternative listen Zai, Together AI, Novita, Fireworks, Featherless AI, DeepInfra und Baseten gehostete Endpunkte.

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 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.

Intelligenz Welt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31
Vollständiges Webportrait: https://intelligenz-welt.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