

Cohere Command A Benchmark-Tabelle

Modelle: Command A, Command A Reasoning, Command A Vision und Command A Translate. Datenstand: 25.04.2026.

Leselogik	Hoehere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Werte aus unterschiedlichen Benchmark-Familien sind nicht direkt miteinander vergleichbar. Fuer Command A Reasoning waren in den zugaenglichen offiziellen Textquellen Benchmark-Vergleiche genannt, aber keine belastbar extrahierbaren Einzelwerte; diese Felder sind mit n.v. markiert.
------------------	---

Text, RAG, Tool-Use, Coding und Reasoning

Benchmark	Was wird getestet?	Command A	Command A Reasoning	Command A Vision	Command A Translate
MMLU	Allgemeines Fachwissen und akademische Multiple-Choice-Fragen.	85,5 %	-	-	-
MMLU-Pro	Anspruchsvollere MMLU-Variante fuer schwerere Fach- und Reasoning-Fragen.	69,6 %	-	-	-
GPQA / GPQA Diamond	Schwierige naturwissenschaftliche Fachfragen auf Graduate-Level.	50,8 %	-	-	-
IFEval	Instruction Following; Befolgung genauer Anweisungen.	90,9 %	-	-	-
InFoBench	Komplexe Informations- und Instruktionsbefolgung.	94,9 %	-	-	-
ChatRAGBench	Retrieval-Augmented Generation; Antwortqualitaet mit externem Kontext.	72,9 %	-	-	-
StrategyQA	Fragen mit implizitem, mehrstufigem Reasoning.	76,7 %	-	-	-
DROP	Leseverstehen mit numerischem und diskretem Reasoning.	91,1 %	-	-	-
HotPotQA	Multi-Hop-Fragen ueber mehrere Wissensstuecke hinweg.	92,1 %	-	-	-
BFCL Overall	Function-/Tool-Calling; korrekte Nutzung von Funktionen und APIs.	63,8 %	n.v. (BFCL-v3/V4 genannt)	-	-
TauBench Retail P@1	Agentische Tool-Workflows in simulierten Retail-Prozessen.	60,0 %	n.v. (genannt)	-	-
TauBench Airline P@1	Agentische Tool-Workflows in simulierten Airline-Prozessen.	45,3 %	n.v. (genannt)	-	-
LiveCodeBench	Aktuelle Programmieraufgaben und Coding-Faehigkeiten.	26,9 %	-	-	-
BigCodeBench	Breitere und komplexere Codegenerierungsaufgaben.	45,4 %	-	-	-
SWE-Bench Verified	GitHub-Issue-Loesung / Code-Aenderungen in realen Repositories.	26,8 %	-	-	-
MATH (all)	Mathematische Problemloesung ueber verschiedene Schwierigkeitsgrade.	80,0 %	-	-	-
AIME 2024	Wettbewerbsnahe Mathematikaufgaben.	23,3 %	-	-	-

Vision, Dokumente, OCR und Translation

Zusaetzliche offizielle Benchmarks mit veroeffentlichten Werten fuer Command A Vision und Command A Translate.

Benchmark	Was wird getestet?	Command A	Command A Reasoning	Command A Vision	Command A Translate
ChartQA	Fragebeantwortung zu Diagrammen und Charts.	-	-	90,9 %	-
InfoVQA	Visuelle Fragebeantwortung zu informationsreichen Dokumenten.	-	-	82,9 %	-
AI2D	Verstehen von wissenschaftlichen und instruktiven Diagrammen.	-	-	94,0 %	-
MMMU (CoT)	Multimodale akademische Aufgaben mit Chain-of-Thought-Setup.	-	-	65,3 %	-
MathVista	Mathematisches und visuelles Reasoning mit Bild-/Diagrammkontext.	-	-	73,5 %	-
DocVQA	Document Question Answering; Dokumentverstaendnis und Layout.	-	-	95,9 %	-
TextVQA	Text in Bildern erkennen und beantworten.	-	-	84,8 %	-
OCRBench	OCR- und Dokumentverarbeitungsfaehigkeiten.	-	-	86,9 %	-
RealWorldQA	Visuelles Verstaendnis in realen Bildszenarien.	-	-	73,6 %	-
Vision Avg.	Durchschnitt der neun ausgewiesenen Vision-Benchmarks.	-	-	83,1 %	-
WMT25 xCOMET Avg	Maschinelle Uebersetzungsqualitaet ueber WMT25-Sprachen.	-	-	-	83,9
WMT25 MetricX	Alternative MT-Qualitaetsmetrik; niedrigere Werte sind besser.	-	-	-	-6,3
Long-context xCOMET	Langkontext-Uebersetzung mit Absatzstruktur.	-	-	-	51,9
Prompt-Injection Failure	Fehlerrate bei manipulativen Prompt-Injection-Tests; niedriger ist besser.	-	-	-	0,3 %

Modell-Spezifikationen

Feld	Beschreibung	Command A	Command A Reasoning	Command A Vision	Command A Translate
Modell-ID	Offizielle API-/Modellkennung.	command-a-03-2025	command-a-reasoning-08-2025	command-a-vision-07-2025	command-a-translate-08-2025
Kontextfenster	Maximaler Eingabekontext laut Cohere-Dokumentation.	256k	256k	128k	8k Input / 16k gesamt
Max. Output	Maximale Ausgabetokens laut Cohere-Dokumentation.	8k	32k	8k	8k
Fokus	Primaerer Einsatzbereich.	Enterprise Agents, RAG, Tool Use	Reasoning, Agents, Tool Use	Vision, OCR, Dokumente	Machine Translation

Quellenbasis

Cohere Docs fuer Modell-IDs, Kontextfenster und Einsatzbereiche; Cohere Command A Technical Report fuer Text-, RAG-, Tool-, Coding- und Mathematikwerte; CohereLabs/Command A Vision Benchmarkveroeffentlichung fuer Vision-Werte; ACL/WMT25 Paper zu Command A Translate fuer Uebersetzungswerte. Fuer redaktionelle Verwendung im Portal sollten diese Werte als Anbieterangaben bzw. Paper-Werte gekennzeichnet werden.