DeepSeek Benchmark-Tabelle

Modelle: DeepSeek-R1 0528, DeepSeek-V3, DeepSeek-VL2 sowie DeepSeek LLM 7B/67B Chat. Datenstand: 25.04.2026.

Leselogik Hoehere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Werte mit unterschiedlichen Metriken wie Rating, Elo, aggregierter Score oder Prozent sind nicht direkt miteinander vergleichbar. DeepSeek-VL2 ist ein Vision-Language-Modell; daher sind seine Werte ueberwiegend in multimodalen Benchmarks ausgewiesen.

Reasoning, Coding, Tool-Use und Text-Benchmarks

Benchmark	Was wird getestet?	DeepSeek-R1 0528	DeepSeek-V3	DeepSeek-VL2	LLM 7B Chat	LLM 67B Chat
MMLU / MMLU-Redux	Allgemeines Fachwissen und akademische Fragen. R1/V3: Redux; LLM: klassisches MMLU.	**93,4 %**	**89,1 %**	-	**49,4 %**	**71,1 %**
MMLU-Pro	Schwerere MMLU-Variante mit anspruchsvolleren Fach- und Reasoning-Fragen.	**85,0 %**	**75,9 %**	-	-	-
GPQA Diamond	Schwierige naturwissenschaftliche Fachfragen auf Graduate-Level.	**81,0 %**	**59,1 %**	-	-	-
SimpleQA	Faktische Genauigkeit bei einfachen Wissensfragen.	**27,8 %**	**24,9 %**	-	-	-
FRAMES	Mehrschrittiges Reasoning und Retrieval-orientierte Aufgaben.	**83,0 %**	**73,3 %**	-	-	-
HLE	Humanity's Last Exam; sehr schwere fachliche Aufgaben.	**17,7 %**	-	-	-	-
LiveCodeBench	Aktuelle Programmieraufgaben; Pass@1 / COT je nach Quelle.	**73,3 %**	**40,5 %**	-	-	-
Codeforces Rating	Kompetitive Programmierung; Elo-/Rating-aehnlicher Score.	**1930**	**1134**	-	-	-
SWE Verified	Reale GitHub-Issues; Anteil geloester Aufgaben.	**57,6 %**	**42,0 %**	-	-	-
Aider Polyglot	Mehrsprachige Code-Editing-Aufgaben mit Aider.	**71,6 %**	**49,6 %**	-	-	-
HumanEval / HumanEval-Mul	Codegenerierung. V3: HumanEval-Mul; LLM: klassisches HumanEval.	-	**82,6 %**	-	**48,2 %**	**73,8 %**
AIME 2024	Wettbewerbsnahe Mathematik; Pass@1.	**91,4 %**	**39,2 %**	-	-	-
AIME 2025	Wettbewerbsnahe Mathematik; Pass@1.	**87,5 %**	-	-	-	-
HMMT 2025	Anspruchsvolle Mathematik-Wettbewerbsaufgaben.	**79,4 %**	-	-	-	-
CNMO 2024	Chinese National High School Mathematics Olympiad.	**86,9 %**	**43,2 %**	-	-	-
GSM8K	Grundlegendes mathematisches Schlussfolgern.	-	-	-	**62,6 %**	**84,1 %**
BBH	Big-Bench Hard; schwierige Reasoning-Aufgaben.	-	-	-	**42,3 %**	**71,7 %**
C-Eval	Chinesische Fach- und Pruefungsfragen.	-	**86,5 %**	-	**47,0 %**	**65,2 %**
CMMLU	Chinesische MMLU-aehnliche Fachfragen.	-	-	-	**49,7 %**	**67,8 %**
ChineseQA	Inhouse-Benchmark fuer chinesische Fragen.	-	-	-	**75,0 %**	**85,1 %**
BFCL v3 MultiTurn	Function Calling / Tool-Use ueber mehrere Dialogrunden.	**37,0 %**	-	-	-	-
Tau-Bench	Agentische Tool-Nutzung in simulierten Airline- und Retail-Workflows.	**53,5 % Airline / 63,9 % Retail**	-	-	-	-

Multimodal, OCR, Visual QA und Visual Grounding

Offizielle DeepSeek-VL2-Werte aus OCR-, Dokument-, Chart-, Multimodal- und Grounding-Benchmarks.

Benchmark	Was wird getestet?	DeepSeek-R1 0528	DeepSeek-V3	DeepSeek-VL2	LLM 7B Chat	LLM 67B Chat
DocVQA	Visuelles Frage-Antworten auf Dokumentbildern.	-	-	**93,3 %**	-	-
ChartQA	Fragen zu Diagrammen und Charts.	-	-	**86,0 %**	-	-
InfoVQA	Informationssuche und QA auf infografikartigen Bildern.	-	-	**78,1 %**	-	-
TextVQA	Visuelles QA mit Text in Bildern.	-	-	**84,2 %**	-	-
OCRBench	OCR-nahe Aufgaben: Texterkennung, Dokumente, Key-Information-Extraction.	-	-	**811**	-	-
MMStar	Multimodales Reasoning und visuelles Verstaendnis.	-	-	**61,3 %**	-	-
AI2D	Diagrammverstaendnis und wissenschaftsnahe Bildfragen.	-	-	**81,4 %**	-	-
MMMU	Multimodale akademische Aufgaben.	-	-	**51,1 %**	-	-
MME	Breites MLLM-Verstaendnis; aggregierter Score.	-	-	**2253**	-	-
MMBench EN	Multimodales Benchmarking, englischer Test.	-	-	**83,1 %**	-	-
MMBench CN	Multimodales Benchmarking, chinesischer Test.	-	-	**79,6 %**	-	-
MMBench-V1.1	Aktualisierte MMBench-Variante.	-	-	**79,2 %**	-	-
MMT-Bench	Multimodale Multi-Task-Bewertung.	-	-	**63,6 %**	-	-
RealWorldQA	Visuelle Fragen mit realweltnahem Bildkontext.	-	-	**68,4 %**	-	-
MathVista	Mathematisches und visuelles Reasoning.	-	-	**62,8 %**	-	-
RefCOCO	Visual Grounding; Objektlokalisierung nach Referenzausdruck.	-	-	**95,1 / 96,7 / 92,7**	-	-
RefCOCO+	Visual Grounding mit anspruchsvolleren Referenzausdruecken.	-	-	**91,2 / 94,9 / 87,4**	-	-
RefCOCOg	Visual Grounding mit laengeren Beschreibungen.	-	-	**92,8 / 92,9**	-	-

Quellenbasis

DeepSeek-R1-0528 Hugging-Face Model Card; DeepSeek-R1/DeepSeek-V3 GitHub Model Cards; DeepSeek-LLM GitHub Model Card; DeepSeek-VL2 Technical Report. Fuer redaktionelle Verwendung im Portal sollte der Quellenstatus als Anbieterangabe gekennzeichnet werden.