Mistral Benchmark-Tabelle

Modelle: Mistral Small 4, Mistral Large 3, Ministral 3 8B und Ministral 3 3B. Datenstand: 25.04.2026.

Leselogik

Hoehere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Prozentwerte, Elo, Punktescores und Win-Rates sind nicht direkt miteinander vergleichbar. Bei Mistral Small 4 wird zwischen Instruct- und Reasoning-Modus unterschieden; bei Ministral 3 werden je nach Benchmark Reasoning-, Instruct- oder Base-Varianten ausgewiesen.

Reasoning, Mathematik, Coding und Text-Benchmarks

Benchmark	Was wird getestet?	Mistral Small 4	Mistral Large 3	Ministral 3 8B	Ministral 3 3B
AA LCR	Long-Context-Reasoning; Genauigkeit bei langen Kontexten.	**72,0 % (R) / 44,0 % (I)**	-	-	-
AIME 2025	Wettbewerbsnahe Mathematik; Reasoning- bzw. High-Reasoning-Variante.	**84,0 % (R) / 36,0 % (I)**	-	**78,7 % (R)**	**72,1 % (R)**
AIME 2024	Wettbewerbsnahe Mathematik; Reasoning-Variante.	-	-	**86,0 % (R)**	**77,5 % (R)**
GPQA Diamond	Schwierige naturwissenschaftliche Fachfragen auf Graduate-Level.	**71,2 % (R) / 59,1 % (I)**	**43,9 %**	**66,8 % (R)**	**53,4 % (R)**
MMLU Pro	Schwerere MMLU-Variante fuer anspruchsvollere Fach- und Reasoning-Fragen.	**78,0 % (R) / 73,5 % (I)**	-	-	-
MMMLU 8-lang avg.	Mehrsprachiger MMLU-Durchschnitt ueber acht Sprachen.	-	**85,5 %**	-	-
SimpleQA	Faktische Genauigkeit bei einfachen Wissensfragen; Exact Match.	-	**23,8 %**	-	-
AMC	Mathematik-Wettbewerbsaufgaben; American Mathematics Competitions.	-	**52,0 %**	-	-
LiveCodeBench	Aktuelle Programmieraufgaben und Coding-Faehigkeiten.	**64,0 % (R) / 32,0 % (I)**	**34,4 %**	**61,6 % (R)**	**54,8 % (R)**
Collie	Mistral-interner Reasoning/Coding-Benchmark; High-Reasoning-Auswertung.	**62,9 % (R)**	-	-	-

Instruction, Vision, Multilingual und Human-Evaluation-Benchmarks

Zusaetzliche offizielle Benchmarks mit veroeffentlichten Werten fuer die genannten Mistral-Modelle.

Benchmark	Was wird getestet?	Mistral Small 4	Mistral Large 3	Ministral 3 8B	Ministral 3 3B
AllenAI IFBench	Instruction Following; Befolgung genauer Anweisungen.	**48,0 % (R) / 35,7 % (I)**	-	-	-
Arena Hard	Schwierige Chat-/Instruction-Prompts; modellvergleichende Bewertung.	**58,3 % (R) / 55,8 % (I)**	-	**50,9 % (I)**	**30,5 % (I)**
WildBench	Breite, anspruchsvolle Instruction- und Chat-Aufgaben.	-	-	**66,8 (I)**	**56,8 (I)**
MATH Maj@1	Mathematische Aufgaben; Mehrheitsergebnis / Maj@1.	-	-	**87,6 % (I)**	**83,0 % (I)**
MM MTBench	Multimodales Multi-Turn-Benchmarking; Score, nicht Prozent.	-	-	**8,08 (I)**	**7,83 (I)**
MMMU-Pro	Multimodale akademische Aufgaben mit Bild-/Dokumentverstaendnis.	**60,0 % (R) / 46,3 % (I)**	-	-	-
LMArena Elo	Blindvergleich durch Nutzerpraeferenzen; Elo-Score.	-	**1418 +/- 11**	-	-
Human Eval: General Prompts	Menschliche Drittbewertung; Win-Rate von Mistral Large 3 gegen Vergleichsmodell.	-	**53 % vs DeepSeek V3.1 / 55 % vs Kimi K2**	-	-
Human Eval: Multilingual Prompts	Menschliche Drittbewertung; mehrsprachige Prompts.	-	**57 % vs DeepSeek V3.1 / 60 % vs Kimi K2**	-	-

Base-Modell-Benchmarks fuer Ministral 3

Benchmark	Was wird getestet?	Mistral Small 4	Mistral Large 3	Ministral 3 8B	Ministral 3 3B
Multilingual MMLU	Mehrsprachige Fachfragen; Base-Variante der Ministral-Modelle.	-	-	**70,6 % (B)**	**65,2 % (B)**
MATH CoT 2-shot	Mathematisches Reasoning mit Chain-of-Thought und 2-shot Setup.	-	-	**62,6 % (B)**	**60,1 % (B)**
AGIEval 5-shot	Pruefungs- und Reasoning-Aufgaben im 5-shot Setup.	-	-	**59,1 % (B)**	**51,1 % (B)**
MMLU Redux 5-shot	Bereinigte MMLU-Variante im 5-shot Setup.	-	-	**79,3 % (B)**	**73,5 % (B)**
MMLU 5-shot	Allgemeines Fachwissen und akademische Fragen im 5-shot Setup.	-	-	**76,1 % (B)**	**70,7 % (B)**
TriviaQA 5-shot	Faktenwissen / Question Answering im 5-shot Setup.	-	-	**68,1 % (B)**	**59,2 % (B)**

Quellenbasis

Mistral Small 4: offizielle Mistral/Hugging-Face-Model-Card und Benchmarkgrafiken. Mistral Large 3: offizielle Mistral/Hugging-Face-Model-Card und Benchmarkgrafiken. Ministral 3 8B/3B: offizielle Hugging-Face-Model-Card der Ministral-3-Familie. Fuer redaktionelle Verwendung im Portal sollten diese Werte als Anbieterangaben gekennzeichnet werden.