Google Gemini Benchmark-Tabelle

Korrigierte Fassung: Gemini 3.1 Pro Preview, Gemini 2.5 Pro und Gemini 2.5 Flash. Gemini 3.1 Flash-Lite Preview sowie OSWorld-Verified, FrontierMath und CyberGym wurden entfernt. Datenstand: 25.04.2026.

Leselogik

Hoehere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Werte mit unterschiedlichen Einheiten wie Elo und Prozent sind nicht direkt miteinander vergleichbar. Toolathlon entfaellt, weil in den verwendeten offiziellen Gemini-Quellen kein belastbarer Wert veroeffentlicht ist; stattdessen sind MCP Atlas und tau2-bench als offizielle Tool-use-Benchmarks enthalten.

Agentic, Tool-Use, Reasoning und Coding

Benchmark	Was wird getestet?	Gemini 3.1 Pro Preview	Gemini 2.5 Pro	Gemini 2.5 Flash
Terminal-Bench 2.0	Agentic terminal coding; Terminus-2 Harness.	**68,5 %**	-	-
GDPval-AA	Expertentasks und Wissensarbeit; Google weist den Wert als Elo aus.	**1317 Elo**	-	-
BrowseComp	Agentische Suche mit Search, Python und Browse.	**85,9 %**	-	-
MCP Atlas	Multi-step Workflows mit Model Context Protocol / MCP.	**69,2 %**	-	-
tau2-bench Retail	Agentische Tool-Nutzung in simulierten Retail-Workflows.	**90,8 %**	-	-
tau2-bench Telecom	Agentische Tool-Nutzung in simulierten Telecom-Workflows.	**99,3 %**	-	-
Humanity's Last Exam	Akademisches Reasoning; full set, Text plus multimodal, ohne Tools.	**44,4 %**	**21,6 %**	**11,0 %**
HLE Search + Code	Humanity's Last Exam mit blocklist search plus Code.	**51,4 %**	-	-
ARC-AGI-2	Abstrakte Reasoning-Puzzles; ARC Prize Verified.	**77,1 %**	-	-
GPQA Diamond	Schwierige wissenschaftliche Fachfragen; No tools / single attempt.	**94,3 %**	**86,4 %**	**82,8 %**
AIME 2025	Wettbewerbsnahe Mathematik; single attempt.	-	**88,0 %**	**72,0 %**
SWE-Bench Verified	Agentic coding; reale GitHub-Issues.	**80,6 %**	**59,6 % / 67,2 %**	**60,4 %**
SWE-Bench Pro (Public)	Diverse agentische Coding-Aufgaben; single attempt.	**54,2 %**	-	-
LiveCodeBench Pro	Competitive Coding aus Codeforces, ICPC und IOI; Elo-Wert.	**2887 Elo**	-	-
LiveCodeBench	Code generation; pass@1 in der 2025-UI-Auswertung.	-	**69,0 %**	**63,9 %**
SciCode	Scientific research coding.	**59 %**	-	-
APEX-Agents	Long-horizon professional tasks.	**33,5 %**	-	-
Aider Polyglot	Code Editing; Aider Polyglot.	-	**82,2 %**	**61,9 % / 56,7 %**

Multimodal, Factuality, Multilingual und Long Context

Ergaenzende offizielle Gemini-Benchmarks mit veroeffentlichten Werten fuer die genannten Modelle.

Benchmark	Was wird getestet?	Gemini 3.1 Pro Preview	Gemini 2.5 Pro	Gemini 2.5 Flash
MMMU-Pro	Multimodales Verstaendnis und Reasoning; No tools.	**80,5 %**	-	-
MMMU	Visual reasoning; single attempt pass@1.	-	**82,0 %**	**79,7 %**
MMMLU	Multilingual Q&A.	**92,6 %**	-	-
Global MMLU (Lite)	Multilingual performance in der Gemini-2.5-Modellkarte.	-	**89,2 %**	**88,4 %**
FACTS Grounding	Factuality / Grounding.	-	**87,8 %**	**85,3 %**
SimpleQA	Factuality bei einfachen Wissensfragen.	-	**54,0 %**	**26,9 %**
Vibe-Eval (Reka)	Image understanding; Gemini-as-a-judge-Auswertung.	-	**67,2 %**	**65,4 %**
VideoMME	Videoverstaendnis mit Audio, Visuals und Untertiteln.	-	**86,9 %**	-
VideoMMMU	Multimodales Video-Reasoning.	-	**83,6 %**	-
MRCR v2 8-needle 128k	Long-context performance; 128k average.	**84,9 %**	**58,0 %**	-
MRCR v2 8-needle 1M	Long-context performance; 1M pointwise.	**26,3 %**	**16,4 %**	-
MRCR v2 128k	Long-context performance; Standard-MRCR-v2 128k average.	-	-	**74 %**
MRCR v2 1M	Long-context performance; Standard-MRCR-v2 1M pointwise.	-	-	**32 %**

Quellenbasis Google DeepMind: Gemini 3.1 Pro Modellseite und Modellkarte; Google DeepMind: Gemini 2.5 Pro Modellkarte; Google DeepMind: Gemini 2.5 Flash Modellkarte. Benchmark-Methodiken koennen zwischen Modellgenerationen variieren; die Tabelle ist deshalb als redaktioneller Vergleich mit Quellenstatus "Anbieterangabe" zu verwenden.