Anthropic Benchmark-Tabelle

Claude Opus 4.7, Opus 4.6, Sonnet 4.5 und Haiku 4.5 - kompakte Übersicht für Modellvergleich und Tool-Portal-Pflege. Datenstand: 25.04.2026.

Leselogik Höhere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Elo-Werte sind Ranking-/Vergleichswerte und nicht direkt mit Prozentwerten vergleichbar. Ein Strich (-) bedeutet: kein belastbarer veröffentlichter Wert in dieser Übersicht.

Benchmark	Was wird getestet?	Claude Opus 4.7	Claude Opus 4.6	Claude Sonnet 4.5	Claude Haiku 4.5
Terminal-Bench 2.0	Agentische Terminal- und Kommandozeilen-Aufgaben, z. B. komplexe Entwickler-Workflows.	**69,4 %**	**65,4 %**	**50,0 %**	**41,0 %**
GDPval / GDPval-AA	Wissensarbeit und berufliche Aufgaben; bei Anthropic hier als Elo-Wert ausgewiesen.	**1753 Elo**	**1619 Elo**	-	-
OSWorld / OSWorld-Verified	Computerbedienung in realen Software- und Desktop-Umgebungen.	**78,0 %**	**72,7 %**	**61,4 %**	**50,7 %**
Toolathlon	Fähigkeit, externe Tools sinnvoll zu nutzen.	-	-	-	-
Alternative: tau2-bench Retail	Agentische Tool-Nutzung in simulierten Retail-/Geschäftsprozessen.	-	-	**86,2 %**	**83,2 %**
Alternative: tau2-bench Airline	Agentische Tool-Nutzung in Airline-Workflows.	-	-	**70,0 %**	**63,6 %**
Alternative: tau2-bench Telecom	Agentische Tool-Nutzung in Telekommunikations-Workflows.	-	-	**98,0 %**	**83,0 %**
Alternative: MCP-Atlas	Skalierte Tool-Nutzung und MCP-basierte Agentenaufgaben.	**77,3 %**	**75,8 %**	-	-
BrowseComp	Web-Recherche und agentisches Browsing.	**79,3 %**	**83,7 %**	-	-
FrontierMath	Fortgeschrittene Mathematik.	-	-	-	-
Alternative: AIME 2025 mit Python	Wettbewerbsnahe Mathematik mit Python-/Tool-Unterstützung.	-	-	**100,0 %**	**96,3 %**
Alternative: AIME 2025 ohne Tools	Wettbewerbsnahe Mathematik ohne Tool-Nutzung.	-	-	**87,0 %**	**80,7 %**
CyberGym	Cybersecurity-Fähigkeiten, insbesondere Reproduktion von Schwachstellen.	**73,1 %**	**73,8 %**	-	-
Finance Agent v1.1	Finanzanalyse und agentische Wissensarbeit.	**64,4 %**	**60,1 %**	**55,3 %**	-
GPQA Diamond	Schwierige Fachfragen und Graduate-Level Reasoning.	**94,2 %**	**91,3 %**	**83,4 %**	**73,0 %**