ChatGPT / OpenAI Benchmark-Tabelle

GPT-5.4 Thinking, GPT-5.4 Pro, GPT-5.5 Thinking und GPT-5.5 Pro - kompakte Übersicht für Modellvergleich und Tool-Portal-Pflege. Datenstand: 25.04.2026.

Leselogik Höhere Prozentwerte bedeuten bessere Benchmark-Leistung innerhalb des jeweiligen Tests. Ein Strich (-) bedeutet: kein belastbarer veröffentlichter Wert in dieser Übersicht. Hinweis: OpenAI weist die Evaluationsspalte als GPT-5.5 bzw. GPT-5.4 aus; sie wird hier für ChatGPT als Thinking-Variante geführt.

Benchmark	Was wird getestet?	GPT-5.4 Thinking	GPT-5.4 Pro	GPT-5.5 Thinking	GPT-5.5 Pro
Terminal-Bench 2.0	Agentische Terminal- und Kommandozeilen-Aufgaben, z. B. komplexe Entwickler-Workflows.	**75,1 %**	-	**82,7 %**	-
GDPval	Wissensarbeit und berufliche Aufgaben; bei OpenAI als wins or ties ausgewiesen.	**83,0 %**	**82,0 %**	**84,9 %**	**82,3 %**
OSWorld-Verified	Computerbedienung in realen Software- und Desktop-Umgebungen.	**75,0 %**	-	**78,7 %**	-
Toolathlon	Fähigkeit, externe Tools sinnvoll zu nutzen.	**54,6 %**	-	**55,6 %**	-
BrowseComp	Web-Recherche und agentisches Browsing.	**82,7 %**	**89,3 %**	**84,4 %**	**90,1 %**
FrontierMath Tier 1-3	Fortgeschrittene Mathematik, mittlere bis hohe Schwierigkeitsstufen.	**47,6 %**	**50,0 %**	**51,7 %**	**52,4 %**
FrontierMath Tier 4	Fortgeschrittene Mathematik, höchste ausgewiesene Schwierigkeitsstufe.	**27,1 %**	**38,0 %**	**35,4 %**	**39,6 %**
CyberGym	Cybersecurity-Fähigkeiten, insbesondere sicherheitsbezogene Aufgaben und Schwachstellen-Workflows.	**79,0 %**	-	**81,8 %**	-