insiders

Insiders LLM Benchmarking – Q4 2025

Die besten LLM im Vergleich

Zum Ende des Jahres 2025 veröffentlichen wir die Ergebnisse unseres vierten LLM Benchmarkings, das konsequent auf den bisherigen Benchmarks aufbaut. Für diese Ausgabe haben wir den Datenumfang nahezu verdoppelt und zugleich deutlich anspruchsvollere Dokumente integriert. Damit steigt die Aussagekraft des Benchmarks weiter, auch wenn das Gesamtniveau der Scores aufgrund der höheren Komplexität etwas sinkt. Die Erweiterung folgt unserem Anspruch, den schnellen Fortschritt im LLM-Umfeld realitätsnah abzubilden und Modelle unter Bedingungen zu testen, die ihren Einsatz in produktiven IDP-Workflows möglichst präzise widerspiegeln.

Das aktuelle Benchmarking umfasst 24 Large Language Modelle, darunter auch neue Modelle wie Claude 4.5 Sonnet, Gemini 3 und GPT-5.1. Wie auch im vergangenen Quartal gibt es Modelle, die wir aus unterschiedlichen Gründen aus unserem Benchmark genommen haben – etwa, weil es Nachfolgemodelle gibt, die bei ähnlichen Kosten und vergleichbarer Geschwindigkeit genauso gute Ergebnisse liefern.

Blick auf die aktuellen Ergebnisse

Die Performance der Modelle wird mit einem Wert zwischen 0 und 100 gemessen: 100 entspricht einer fehlerfreien Verarbeitung, 0 bedeutet vollständig falsche Ergebnisse. Zusätzlich wird die Verarbeitungsgeschwindigkeit der Modelle in einem vereinfachten Speedlevel dargestellt: Dabei steht Level 3 für sehr schnelle Modelle, Level 2 für mittlere Geschwindigkeit und Level 1 für langsame Verarbeitung. Auf diese Weise lassen sich Genauigkeit und Effizienz der Modelle vergleichbar gegenüberstellen.

Durch den verdoppelten Datenumfang und die höhere Dokumentkomplexität liegt das Gesamtniveau der Scores etwas niedriger als im Vorquartal. Die Modelle werden stärker gefordert, dadurch steigen die Unterschiede zwischen generischen Foundation-Modellen – also Modelle, die bewusst breit aufgestellt sind und auf riesigen, oft unspezifischen Daten trainiert werden – und spezialisierten IDP-Modellen deutlicher.

Es fällt auf, dass sich immer mehr Modelle im „oberen Mittelfeld“ ansiedeln. Der Wettbewerb zieht sich weiter zusammen: Viele Modelle liegen eng beieinander, insbesondere im Bereich 80–88 Punkte. Schwächere Modelle wurden zum Beispiel nicht weitergeführt oder von uns bewusst nicht gewählt. Das Branchenniveau konsolidiert sich, und die Varianz zwischen den Modellen sinkt.

Auch diesmal liefern dedizierte Reasoning-Modelle starke Ergebnisse in Klassifikation und Extraktion. Gleichzeitig zeigen sich dieselben strukturellen Nachteile wie im letzten Benchmark: längere Verarbeitungszeiten, höhere Tokenkosten und geringere Planbarkeit im Produktivbetrieb. So schneiden GPT-5 oder GPT-4.1 zum Beispiel bei der Gesamtperformance mit Werten von 87,3 und 84,7 herausragend ab, bringen aber große Nachteile, wenn es um Datenschutz oder Verarbeitungsgeschwindigkeit geht.

Im Vergleich zum letzten Quartal steigt in unserer Auswahl die Anzahl der in der EU gehosteten Modelle – bleibt aber auf dem Gesamtmarkt nach wie vor rar.

Stand: 03.12.2025

Web
www.insiders-technologies.com

Mail
llm-benchmarking@insiders-technologies.de

Telefon
+49 631 92081 1700

Top-Ergebnisse im Überblick:

Modell	Speed Level	Datenschutz	Performance
Claude 4.5 Sonnet	3	Gehostet in EU	87,9
GPT-5	1	Gehostet außerhalb EU	87,3
Claude 4 Sonnet	3	Gehostet in EU	87,3
Gemini 2.5 Pro	2	Gehostet außerhalb EU	87,3
Gemini 3 Pro	1	Gehostet außerhalb EU	87,1
Claude 4 Opus	2	Gehostet außerhalb EU	85,8
Gemini 2.5 Flash	3	Gehostet außerhalb EU	84,8
GPT-4.1	3	Gehostet außerhalb EU	84,7
GPT-5 mini	1	Gehostet außerhalb EU	84,5
GPT-5.1	3	Gehostet außerhalb EU	84,4
Qwen 3 235B A22B	3	Gehostet außerhalb EU	83,3
Claude 4.5 Haiku	3	Gehostet in EU	83,2
Claude 3.5 Haiku	3	Gehostet außerhalb EU	83,1
Ovaition Private LLM	3	Gehostet bei Insiders	82,7
DeepSeek-R1 671b	3	Gehostet außerhalb EU	82,6
Mistral Small	3	Gehostet in EU	81,0
GPT-4.1 mini	3	Gehostet außerhalb EU	80,1
Mistral Medium	3	Gehostet in EU	80,6
Qwen 3 32B	3	Gehostet in EU	80,5
GPT-OSS 120b	3	Gehostet außerhalb EU	80,3
Pixtral Large	3	Gehostet in EU	80,2
Claude 3 Haiku	3	Gehostet in EU	79,2
GPT-OSS 20b	3	Gehostet außerhalb EU	78,2
Mistral Large	3	Gehostet in EU	78,1

Markttrends: Große Modelle – kleine Fortschritte

Die neuen Versionen der großen Foundation-Modelle zeigen im IDP-Kontext nur noch marginale Verbesserungen:

- **Claude 4.5 Sonnet vs. 4 Sonnet:** praktisch identische Ergebnisse von 87,9 und 87,3
- **Gemini 3 Pro vs. 2.5 Pro:** ähnlicher Score, aber Gemini 3 Pro mit fast doppelter Laufzeit pro Dokument
- **GPT-5.1 vs. GPT-5:** etwas schwächer, aber dafür spürbar schneller

Der Nutzen „größerer“ Modelle steigt kaum noch. Geschwindigkeitsvorteile oder minimale Qualitätsänderungen sind die Regel – echte Sprünge werden seltener. Man braucht nicht automatisch die neuesten oder teuersten Modelle, um im IDP-Einsatz gute Ergebnisse zu erzielen.

Spezialisierte Modelle liefern die echten Fortschritte

Der größte Qualitätssprung kommt – wie in den vorherigen Benchmarks – nicht von generischen Foundation-Modellen, sondern von modellseitiger Spezialisierung.

Unser eigenes Modell zeigt das besonders deutlich: Das OvAltion Private LLM erreicht mehr als 2% Verbesserung gegenüber dem Vergleich in Q3 – und das trotz anspruchsvollerer Daten. Dieses Ergebnis kommt nicht von ungefähr – unser bisheriges Private LLM wird mit dem angekündigten OvAltion LLM zum „OvAltion Private LLM“ verschmelzen und bietet so höchste Sicherheit bei immer besser werdender Qualität und Spezialisierung auf das IDP Umfeld unserer Kunden und Partner.

Damit liegt das OvAltion Private LLM erstmals in der Nähe von Claude 4.5 Haiku – ein klarer Meilenstein. Und die Entwicklung geht weiter: Wir erwarten weitere deutliche Schritte durch laufendes Finetuning, optimierte Trainings und zusätzliche Daten basierend auf jahrelanger Erfahrung.

Stand: 03.12.2025

Self-Hosted LLM als strategischer Vorteil

Durch ein selbstgehostetes Modell profitieren Unternehmen zukünftig vor allem von voller Datenhoheit, C5-zertifizierter Sicherheit, kontrollierbaren Kosten, modellseitiger Anpassbarkeit, Planbarkeit und Stabilität.

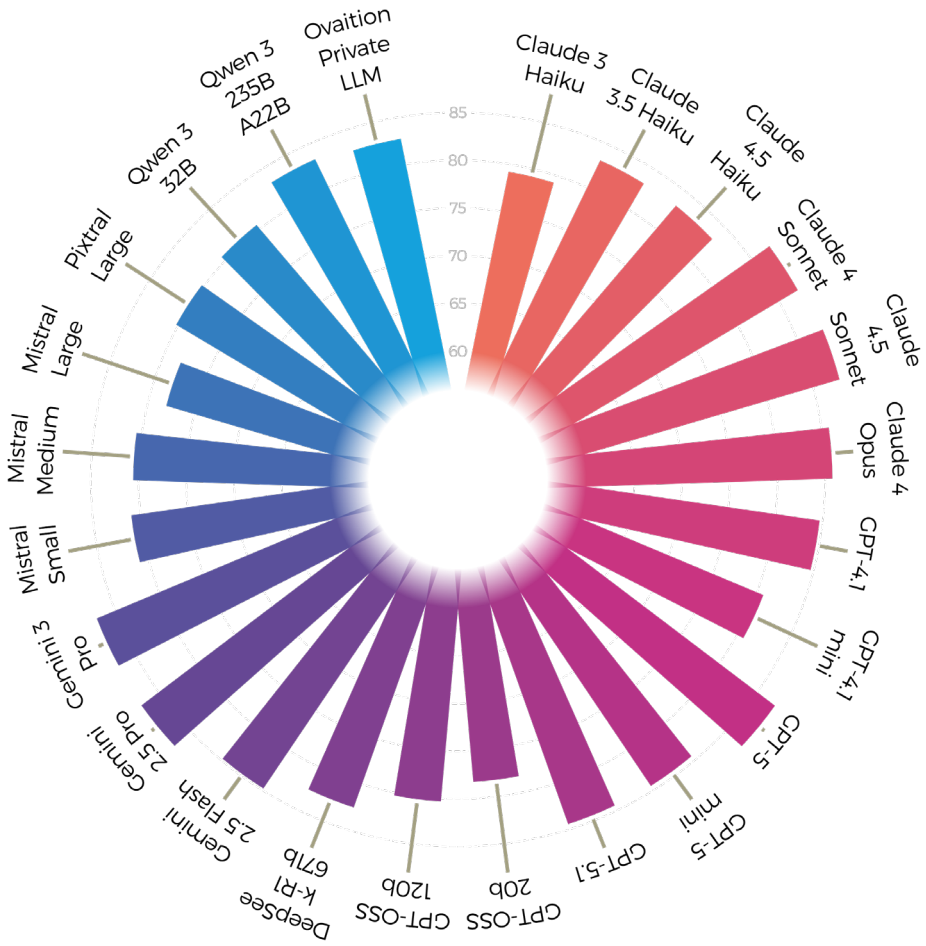
Gerade im Dokumentenverarbeitungsumfeld zeigt sich wieder deutlich: Spezialisierung schlägt Größe.

Die wichtigsten Erkenntnisse

- Große Foundation-Modelle bewegen sich auf hohem Niveau, aber Entwicklung ver langsamt sich im IDP Kontext spürbar
- Reasoning-Modelle erzielen gute Scores, sind oft nicht praxiseffizient
- Unter realen IDP-Bedingungen bleibt Vorteil begrenzt: Mehraufwand übersteigt Zusatzqualität
- Hohe Performance und regulatorische Sicherheit fallen nur selten zusammen

Das OvAltion Private LLM erlebt den deutlichsten Fortschritt im Feld. Der sichtbarste Sprung kommt erneut nicht von globalen Modellen, sondern durch Spezialisierung. Wir schaffen erstmals eine bemerkenswerte Annäherung an die führenden Foundation-Modelle – und das bei stabiler Verarbeitungszeit, vollständig privat, C5-zertifiziert und datenschutzkonform.

Diese Kombination – Wettbewerbsperformance + vollständige Datenhoheit – ist im Benchmark weiterhin einzigartig und hebt sich immer mehr ab.



Stand: 03.12.2025

Der Insiders Best-of-Breed-Ansatz

Unser Benchmarking unterstützt den Best-of-Breed-Gedanken: Wir testen laufend die relevantesten Modelle, integrieren sie über unsere OvAltion Engine und geben Kunden so die Möglichkeit, aus einer Vielzahl an LLMs das optimale Setup für Ihre individuellen Anforderungen zu wählen.

Dabei gilt:

- Kunden müssen sich nicht zwischen Leistung und Datenschutz entscheiden.
- Die Kombination aus bewährter Insiders-KI und State-of-the-Art-LLMs liefert höchste Automatisierung bei geringem Risiko.
- Funktionen wie Green Voting validieren LLM-Ergebnisse automatisch, senken Nachbearbeitungsaufwände und steigern die Dunkelverarbeitungsquote.

Je nach individuellen Anforderungen im Spannungsfeld von Performance, Latenz, Dunkelverarbeitung und Kosten ermöglicht Insiders durch die Integration von LLMs von Drittanbietern seinen Kunden einen bequemen und variablen Zugang zu genau dem LLM, dass sie wirklich brauchen.

Insiders Kunden müssen sich also nicht zwischen Performance und Sicherheit entscheiden. Sie erhalten beides, angepasst an ihre individuellen Bedürfnisse – und bleiben dabei flexibel, um die vielfältigen Einsatzmöglichkeiten von LLMs für Ihr Unternehmen zu erobern.

Mit der ISO- sowie C5-zertifizierten Infrastruktur und der nahtlosen Integration über die Insiders OvAltion Engine bietet Insiders mit OmnIA eine Automatisierungs-Plattform, die nicht nur sicher, sondern auch zukunftssicher ist. Das Insiders LLM Benchmarking ist dabei der verlässliche Referenzpunkt, um im schnelllebigen LLM-Markt den Durchblick zu behalten.

Für individuelle Benchmarkings mit eigenen Use Cases bieten die Insiders KI-Experten eine fundierte Beratung für Ihr Unternehmen an.

llm-benchmarking@insiders-technologies.de

Stand: 03.12.2025

Stand: 03.12.2025