insiders

LLM Benchmarking – September 2025

Die besten LLM im Vergleich

Das Insiders LLM Benchmarking im September 2025 setzt die Reihe fort und baut konsequent auf den Erkenntnissen aus Q2 auf. Um Vergleichbarkeit zu sichern, kommen identische Dimensionen und Testdaten wie im vorherigen Benchmarking zum Einsatz.

Da wir weiterhin neue Modelle im Blick behalten und ebenso auch Modelle verworfen werden, umfasst das aktuelle Benchmarking 21 Large Language Models, darunter auch neue, leistungsstarke Modelle wie GPT-5, Gemini 2.5 Pro oder Claude 4 Sonnet. Mit dieser Flexibilität des Benchmarks tragen wir der Dynamik im LLM-Bereich Rechnung: Wir integrieren kontinuierlich neue Modelle gemäß dem aktuellen Stand der Forschung und evaluieren ihren Transfer in praxistaugliche IDP-Lösungen. Dabei achten wir besonders auf eine ausgewogene Kombination aus höchster Ergebnisqualität und Betriebsstabilität.

Das Benchmarking erfolgte auf Basis eines standardisierten IDP-Datensatzes mit realen Dokumenten aus der Versicherungs- und Finanzwelt. So stellen wir sicher, dass die Ergebnisse direkt auf die Anforderungen unserer Kunden übertragbar sind.

Die Modelle, die Insiders bis zum vergangenen Benchmarking zu Testzwecken noch selbst gehostet hat – wie z.B. Gemma 3 27b oder DeepSeek-R1 32b – konnten sich in der Praxis nicht bewähren und wurden daher aus dem Portfolio genommen.

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.

Mit dem Wechsel auf ein leistungsstärkeres Modell konnte Insiders Private einen deutlichen Qualitätssprung erzielen: von einem Score in Q2 von 67,9 auf nun 78,2 – bei gleichbleibender durchschnittlicher Verarbeitungszeit pro Dokument. Damit rückt es näher an die Spitzenmodelle heran, ohne Abstriche bei Datenschutz oder Speedlevel zu machen.

Deutlich wird zudem, dass Reasoning-Modelle – LLMs, die speziell auf komplexes logisches Denken, u.Ä. trainiert wurden – wie GPT-5 zwar scheinbar die besten Ergebnisse in Klassifikation und Extraktion erzielen (90,7 Punkte, höchste Bewertung im Feld), diese Vorteile jedoch mit spürbaren Nachteilen einhergehen und stark vom jeweiligen Modell abhängen. Die Verarbeitungszeiten liegen um den Faktor 4 höher, und auch die Tokenkosten steigen dementsprechend um ein Vielfaches – ein Aspekt, der für den produktiven Einsatz nicht zu vernachlässigen ist. Reasoning-Modelle sollten daher in der Praxis mit Vorsicht und nur in sinnvollen Use Cases angewendet werden.

Es fällt ebenfalls auf, dass sich immer mehr Modelle im oberen rechten Quadranten der Auswertung befinden – Modelle die in der Vergangenheit nicht gut genug performt haben werden nicht mehr unterstützt und fallen weg.

Ein weiteres Ergebnis: Im aktuellen Benchmarking verbleiben lediglich drei Modelle mit EU-Hosting, und ausschließlich die Insiders-Modelle laufen vollständig privat. Damit wird die Lücke zwischen höchster Performance und regulatorisch sicherem Einsatzumfeld noch sichtbarer.

Stand: 11.09.2025

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
GPT-5	1	Gehostet außerhalb EU	90,7
Claude 4 Sonnet	3	Gehostet in EU	90,0
Claude 3.7 Sonnet	3	Gehostet in EU	89,9
Gemini 2.5 Pro	2	Gehostet außerhalb EU	89,8
Claude 4 Opus	2	Gehostet außerhalb EU	88,5
GPT-4.1	3	Gehostet außerhalb EU	87,9
Gemini 2.5 Flash	3	Gehostet außerhalb EU	87,8
GPT-5 mini	2	Gehostet außerhalb EU	87,4
GPT-4o	3	Gehostet außerhalb EU	86,4
Mistral Large 2	3	Gehostet außerhalb EU	85,2
Claude 3.5 Haiku	3	Gehostet außerhalb EU	84,7
GPT-OSS 120b	3	Gehostet außerhalb EU	84,2
GPT-4.1 mini	3	Gehostet außerhalb EU	83,8
GPT-o3 mini	2	Gehostet außerhalb EU	83,7
DeepSeek-R1 671b	1	Gehostet außerhalb EU	82,5
Claude 3 Haiku	3	Gehostet in EU	81,8
GPT-OSS 20b	3	Gehostet außerhalb EU	80,6
Insiders OvAltion LLM	3	Gehostet bei Insiders	80,1
Gemini 2 Flash	3	Gehostet außerhalb EU	79,5
Insiders Private	3	Gehostet bei Insiders	78,2
GPT-4o mini	3	Gehostet außerhalb EU	77,7

Es bestätigt sich, dass globale Modelle die Benchmark setzen – gestützt auf gewaltige Datenbasis, überlegene Hardware-Ressourcen und riesige Modellarchitekturen. Auf Platz 1 im Gesamtranking landet das Modell GPT-5 von Open AI mit einem Score von 90,7, dicht gefolgt von Claude 4 Sonnet (90,0) und dem letzten Sieger Claude 3.7 Sonnet mit 89,9.

Im Gegensatz zu globalen Modellen ist das Insiders Private LLM auf Datenschutz und regulatorische Sicherheit optimiert. Der Betrieb in der ISO 27001-zertifizierten Insiders Cloud ermöglicht den Einsatz für besonders kritische Dokumentarten, von SEPA-Mandaten bis zu Gesundheitsdaten. Gerade in informationssensiblen Branchen wie Finanzwesen und Versicherungen ist dieser Ansatz ein entscheidender Vorteil. Kontinuierliche Benchmarkings sichern dabei die Weiterentwicklung des Modells.

Was bedeutet also „das beste LLM“

Das aktuelle Insiders LLM Benchmarking verdeutlicht, dass Insiders den Markt kontinuierlich beobachtet und für seine Kunden den Spagat zwischen Performance und Sicherheit meistert – mit einem klaren Best-of-Breed-Ansatz. Dieser Ansatz bedeutet, dass nicht ein einziges Modell alle Aufgaben abdeckt, sondern dass für jede Anwendung die jeweils leistungsfähigsten LLMs identifiziert, bewertet und flexibel integriert werden. Neue Modelle werden daher sofort im Benchmarking getestet und mit bestehenden verglichen. Die Ergebnisse fließen direkt in die Produktentwicklung ein und sichern eine dauerhaft hohe Qualität.

Dabei wird auch deutlich: Die Frage nach dem besten LLM lässt sich nicht pauschal beantworten. Höchste Leistung allein genügt nicht. Gerade in regulierten Branchen wie Versicherungen und Finanzwesen spielen zusätzlich Faktoren wie Verlässlichkeit, Datenschutz und Integrationsfähigkeit eine entscheidende Rolle.

Die wichtigsten Erkenntnisse:

- **Claude 4 Sonnet** führt in Summe – schnell und leistungsstark, gehostet in der EU.
- **Claude 3 Haiku** als bewährtes Modell glänzt mit wahnsinnigen Ergebnissen in der Geschwindigkeit – ideal für Volumenverarbeitung, mit überschaubaren Genauigkeitsverlusten.
- **Top-Performer** unter den LLMs werden zunächst in den USA betrieben – ein EU-Hosting folgt meist erst mit zeitlicher Verzögerung
- Das **Insiders Private LLM** wird privat gehostet und bietet so volle Kontrolle, lokale Verarbeitung und höchste Transparenz.
- Das **Insiders OvAltion LLM** übertrifft als Prototyp bereits das Private LLM in Performance und Geschwindigkeit – wird aber noch getestet und trainiert.

Stand: 11.09.2025

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:

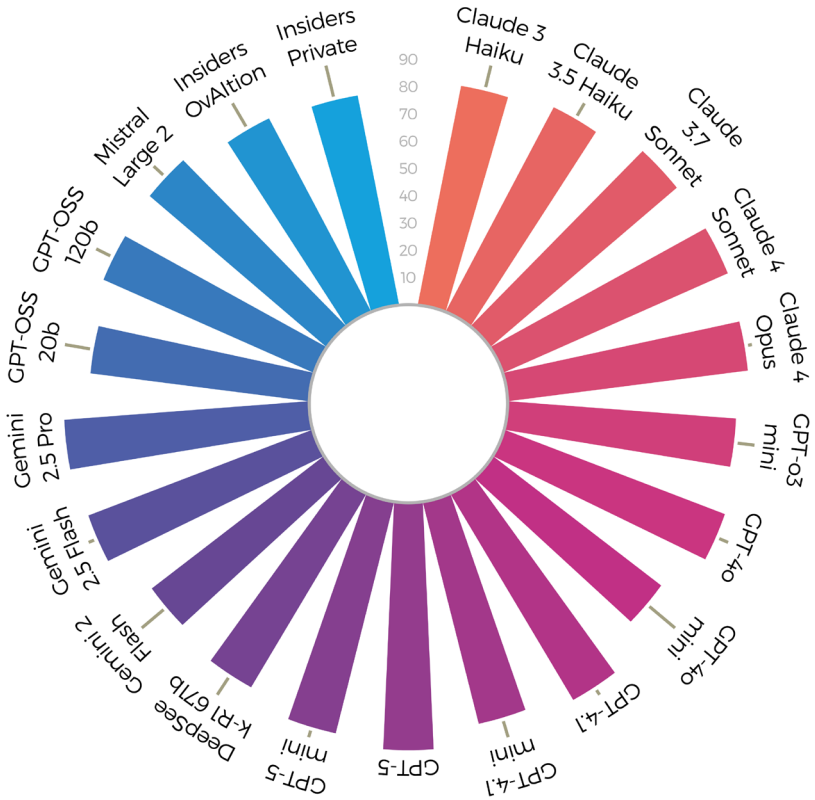
- Kunden müssen sich nicht zwischen Leistung und Datenschutz entscheiden.
- Die Kombination aus bewährter Insiders-KI und State-of-the-Art-LLMs liefert höchste Automatisierung bei geringem Risiko.
- Funktionen wie Green Voting validieren LLM-Ergebnisse automatisch, senken Nachbearbeitungsaufwände und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

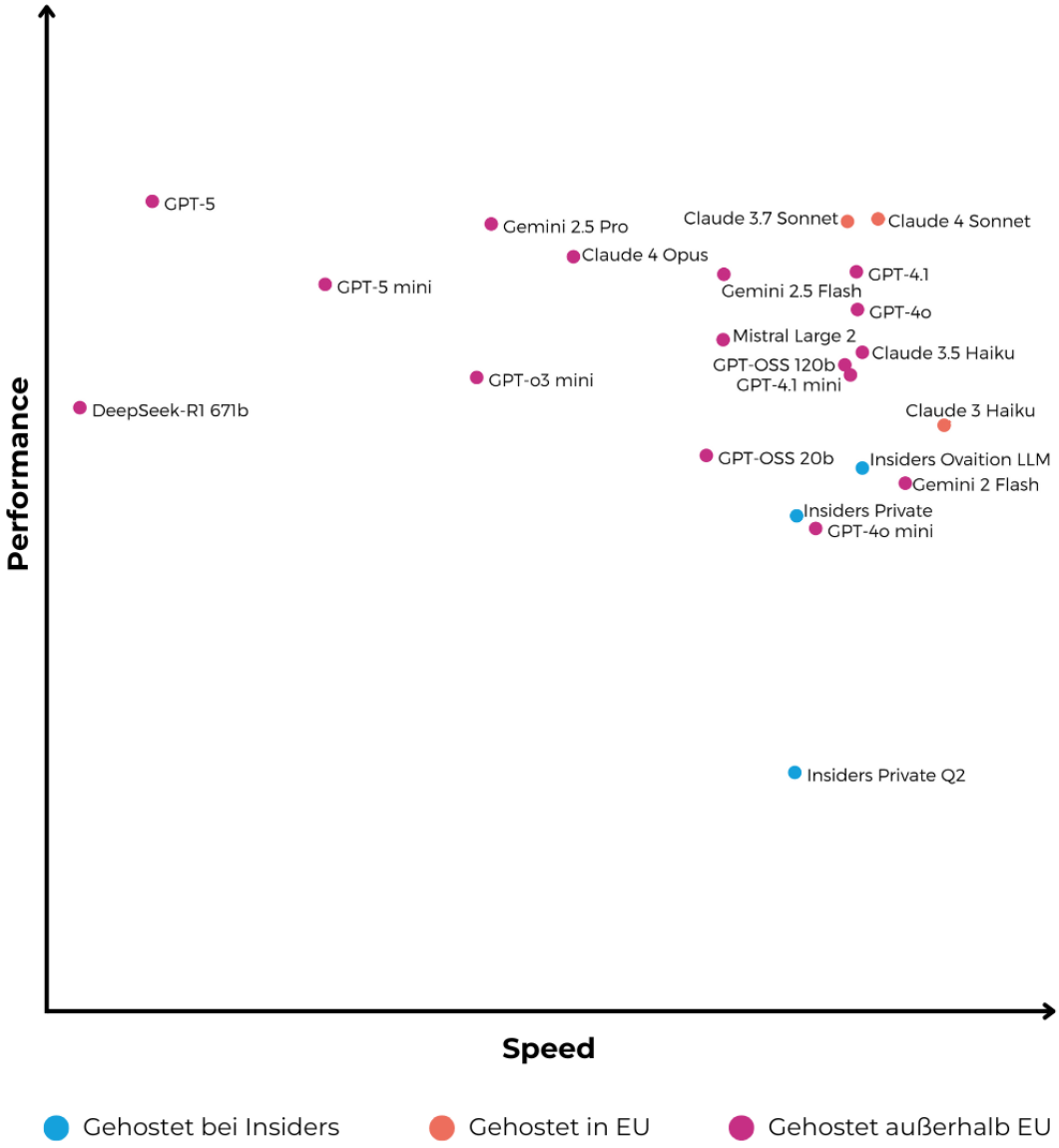
Mit der ISO-zertifizierten Infrastruktur und der nahtlosen Integration über die Insiders OvAltion Engine bietet Insiders mit OmnIA eine Automatisierungs-Plattform, die nicht nur sicher, sondern auch zukunftssicher ist. Das Insiders LLM Benchmarking ist dabei der verlässliche Referenzpunkt, um im schnelllebigen 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. Kommen Sie für ein individuelles Benchmarking einfach auf unsere Insiders KI-Experten zu.

llm-benchmarking@insiders-technologies.de



Stand: 11.09.2025



Stand: 11.09.2025