insiders

INSIDERS LLM BENCHMARKING – MAI 2025

Ein Blick auf die aktuellen Ergebnisse

Das Insiders LLM Benchmarking geht in die nächste Runde: Aufbauend auf dem ersten umfassenden Performancevergleich haben wir unser Vorgehen weiterentwickelt und zusätzlich neue Dimensionen eingeführt. Während im ersten Benchmarking der Fokus primär auf der reinen Performance in den Bereichen Informationsklassifikation und -extraktion lag, berücksichtigen wir nun auch Geschwindigkeit, Datenschutz und relative Kostenstruktur – entscheidende Kriterien für den produktiven Einsatz im IDP-Umfeld.

Da wir ständig neue Modelle im Blick behalten und einbeziehen, umfasst das aktuelle Benchmarking mittlerweile 25 Large Language Models, darunter auch neue, leistungsstarke Modelle wie Claude 3.7 Sonnet, Gemini 2 Flash, Llama 3.3 70b und DeepSeek.

Gerade Gemini 2 Flash überzeugt durch seine außerordentlich hohe Effizienz in der Ausgabe- und Verarbeitungsgeschwindigkeit, ein entscheidender Vorteil für die schnelle Bearbeitung großer Dokumentenmengen in zeitkritischen Verfahren. Die Messung basiert dabei auf der tatsächlichen Antwortzeit bei der Informationsklassifikation und -extraktion.

Mit der Erweiterung 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 – einschließlich eines neuen Use Cases: Schadenrechnungen.
So stellen wir sicher, dass die Ergebnisse direkt auf die Anforderungen unserer Kunden übertragbar sind.

Top-Ergebnisse im Überblick:

Model	Speed Level	Datenschutz	Performance
Claude 3.7 Sonnet	2	Gehostet außerhalb EU	90,17
Claude 3.5 Sonnet	3	Gehostet in EU	89,61
GPT-4o	3	Gehostet außerhalb EU	86,33
Gemini 1.5 Pro	3	Gehostet außerhalb EU	86,09
Gemini 2 Flash	3	Gehostet außerhalb EU	85,93
GPT-4 Turbo	2	Gehostet außerhalb EU	84,82
Mistral Large 2	2	Gehostet in EU	84,82
Claude 3.5 Haiku	3	Gehostet außerhalb EU	84,78
Gemma 3 27b	2	Gehostet bei Insiders	83,11
Llama 3.3 70b	3	Gehostet außerhalb EU	82,91
DeepSeek-R1 671b	1	Gehostet außerhalb EU	82,51
Mistral Large	3	Gehostet in EU	82,04
Claude 3 Haiku	3	Gehostet in EU	81,33
GPT-3.5 Turbo	3	Gehostet außerhalb EU	79,47
Gemma 3 12b	3	Gehostet bei Insiders	78,31
Phi-4 14b	3	Gehostet bei Insiders	78,02
Gemini 1.5 Flash	3	Gehostet außerhalb EU	77,12
GPT-4o mini	3	Gehostet außerhalb EU	77,02
Llama 3.1 70b	3	Gehostet außerhalb EU	75,60
DeepSeek-R1 32b	2	Gehostet bei Insiders	73,11
Mixtral 8x7b	3	Gehostet in EU	72,01
Llama 3.1 8b	3	Gehostet außerhalb EU	69,56
Insiders Private	3	Gehostet bei Insiders	67,87
Granite 3.2 8b	3	Gehostet bei Insiders	64,41
Mistral 7b	3	Gehostet außerhalb EU	61,16

Stand: 31.05.2025

Web
www.insiders-technologies.com

Mail
llm-benchmarking@insiders-technologies.de

Telefon
+49 631 92081 1700

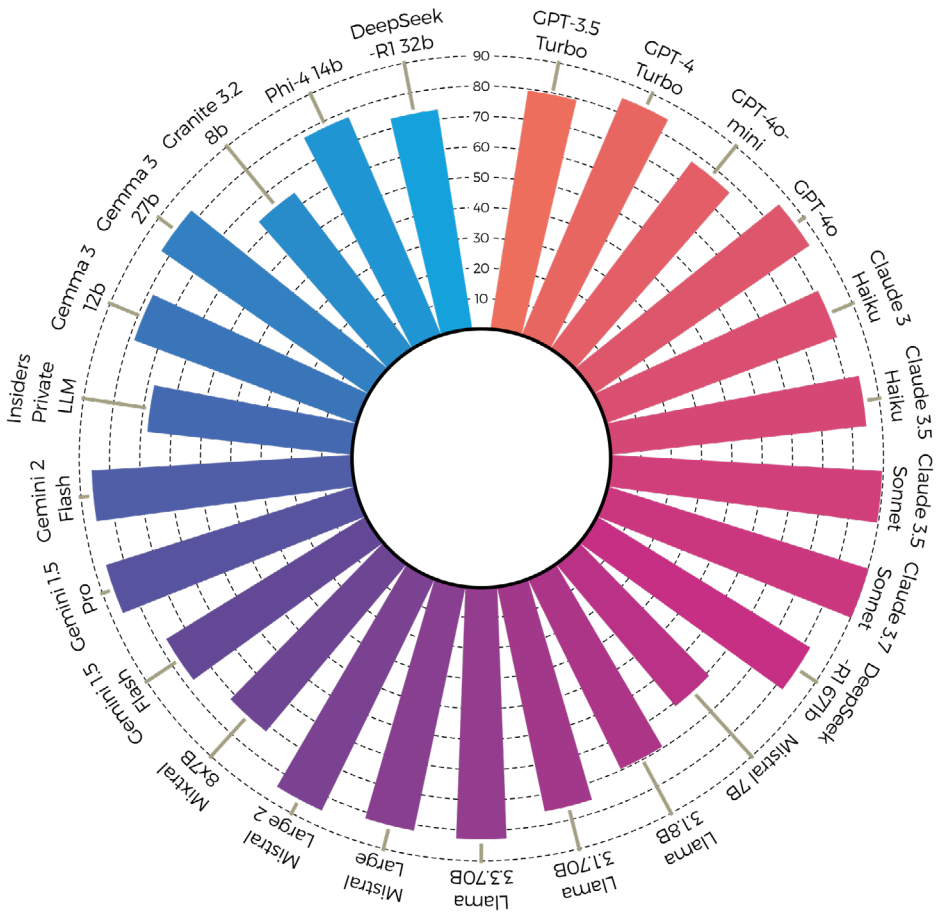
Der aktuelle Vergleich zeigt, dass die globalen Modelle nach wie vor die höchste Leistung bieten – angetrieben durch gigantische Trainingsdatenmengen, Rechenleistung, Verarbeitung von Inputmengen sowie Parametern. Auf Platz 1 im Gesamtranking landet das Modell Claude 3.7 Sonnet von Anthropic mit einem Score von 90,17, dicht gefolgt vom letzten Sieger Claude Sonnet 3.5 mit 89,61. Den dritten Platz belegt eines der bekanntesten Modelle: GPT-4o von Open AI mit 86,33.

Kleinere lokale Modelle wie das Insiders Private LLM hingegen, sind auf Datensicherheit und Compliance optimiert – mit vollem Betrieb in der ISO 27001-zertifizierten Insiders Cloud und höchstem Datenschutzlevel, speziell für sensible Dokumententypen wie SEPA-Mandate oder medizinische Daten. Dieser bewusste Tradeoff ist entscheidend für informationssensible Branchen wie die Versicherungs- und Finanzbranche. Mit Hilfe des kontinuierlichen Benchmarkings wird das Insiders private LLM gezielt weiterentwickelt und erreicht zeitnah eine noch höhere Performance.

Was bedeutet also „das beste LLM“?

Das aktuelle Insiders LLM Benchmarking zeigt, dass Insiders als KI-Experte den LLM-Markt genau im Blick hat und den Balanceakt zwischen Performance und Sicherheit für seine Kunden mit dem Best-of-Breed-Ansatz meistert. Der Best-of-Breed-Ansatz bedeutet, dass Insiders in all seinen Produkten die leistungsfähigsten LLMs auf dem Markt identifiziert und flexibel integriert. Erscheint ein neues Modell auf dem Markt, wird dieses im aktuellen Insiders LLM Benchmarking auf die Probe gestellt und mit den weiteren LLMs verglichen. Die gewonnen Erkenntnisse fließen umgehend in die Produktentwicklung ein und stellen eine dauerhaft hohe Qualität für die Insiders Kunden sicher.

Das neue Benchmarking zeigt: Die Frage nach „dem besten LLM“ ist kein Schwarz-Weiß-Thema. Leistung allein reicht nicht. In hochregulierten Branchen wie Versicherungen und Finanzen zählen vor allem auch Verlässlichkeit, Datenschutz und Integrationsfähigkeit.



Web
www.insiders-technologies.com

Mail
llm-benchmarking@insiders-technologies.de

Telefon
+49 631 92081 1700

Die wichtigsten Erkenntnisse:

- Claude 3.7 Sonnet ist der aktuelle Perfomance-Spitzenreiter – schnell und leistungsstark, aber nur mit globalem Datenschutz.
- Gemini 2 Flash setzt neue Maßstäbe bei der Geschwindigkeit – ideal für Volumenverarbeitung, mit kleinen Tradeoffs bei der Präzision.
- Bei Insiders gehostete Modelle, etwa Gemma 3 27b oder Phi-4 14b, erreichen inzwischen sehr respektable Werte – bei voller Datenhoheit und ohne Rate Limits.
- Das Insiders Private LLM bleibt zwar in der Performance zurück, glänzt aber dort, wo andere Modelle Schwächen zeigen: hoher Datenschutz, volle Kontrolle, lokale Verarbeitung und höchste Transparenz.

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.

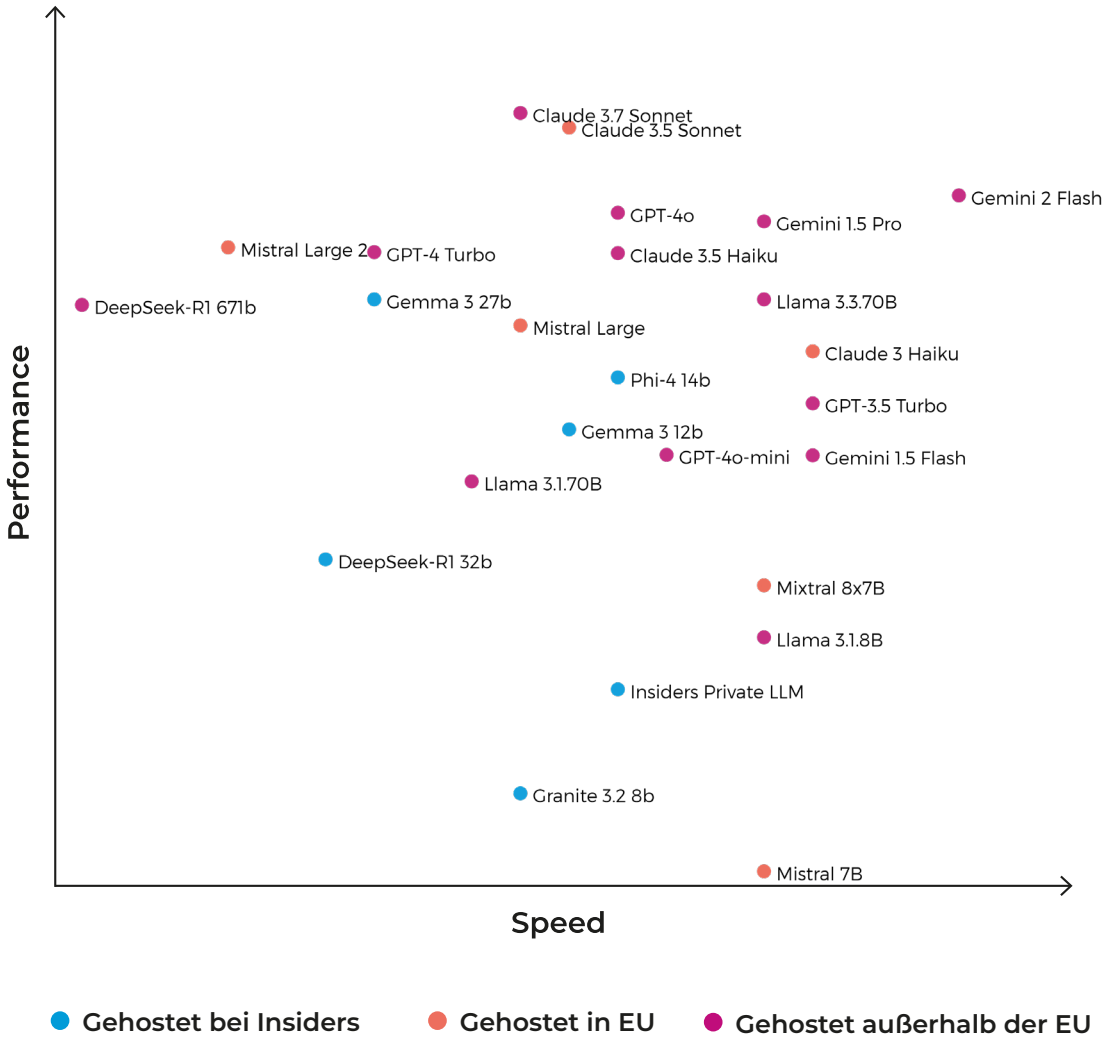
Mit der ISO-zertifizierten Infrastruktur und der nahtlosen Integration über die Insiders OvAltion Engine bietet Insiders mit OmnIA eine Automatisierungs-Plattform, die nicht nur sicher, sondern auch zukunftssicher ist. Das Insiders LLM Benchmarking ist dabei der verlässliche Referenzpunkt, um im schnelllebigen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

Web
www.insiders-technologies.com

Mail
llm-benchmarking@insiders-technologies.de

Telefon
+49 631 92081 1700


Performance
Speed
Claude 3.7 Sonnet
Claude 3.5 Sonnet
Gemini 2 Flash
GPT-4o
Gemini 1.5 Pro
Mistral Large 2
GPT-4 Turbo
Claude 3.5 Haiku
DeepSeek-R1 671b
Gemma 3 27b
Llama 3.3.70B
Mistral Large
Claude 3 Haiku
Phi-4 14b
GPT-3.5 Turbo
Gemma 3 12b
GPT-4o-mini
Gemini 1.5 Flash
Llama 3.1.70B
DeepSeek-R1 32b
Mixtral 8x7B
Llama 3.1.8B
Insiders Private LLM
Granite 3.2 8b
Mistral 7B
Gehostet bei Insiders
Gehostet in EU
Gehostet außerhalb der EU