insiders

INSIDERS LLM BENCHMARKING – MAY 2025

A Look at the Latest Results

The Insiders LLM Benchmarking enters the next round: Building on our initial comprehensive performance comparison, we've expanded our approach and introduced new evaluation dimensions. While the first benchmarking focused primarily on raw performance in information classification and extraction, we now also consider speed, data privacy, and relative cost structure – crucial factors for productive deployment in Intelligent Document Processing (IDP).

As we continuously monitor and evaluate new models, the current benchmarking now includes 25 large language models, featuring high-performing newcomers like Claude 3.7 Sonnet, Gemini 2 Flash, Llama 3.3 70b, and DeepSeek.

Gemini 2 Flash stands out with its exceptional efficiency in output and processing speed – a key advantage for quickly handling large volumes of documents in time-sensitive workflows. Measurements are based on actual response times for classification and extraction tasks.

With the expanded benchmark, we reflect the rapid pace of change in the LLM space: We continuously integrate new models based on the latest research and evaluate their practical transferability into real-world IDP use cases. Our focus remains on a balanced mix of top-tier result quality and operational stability.

Benchmarking is conducted using a standardized IDP dataset with real documents from the insurance and finance sectors – including a new use case: claims invoices. This ensures results are directly applicable to our customers' real-world requirements.

Top Results at a Glance:

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

As of: 31.05.2025

Web
www.insiders-technologies.com

Mail
llm-benchmarking@insiders-technologies.de

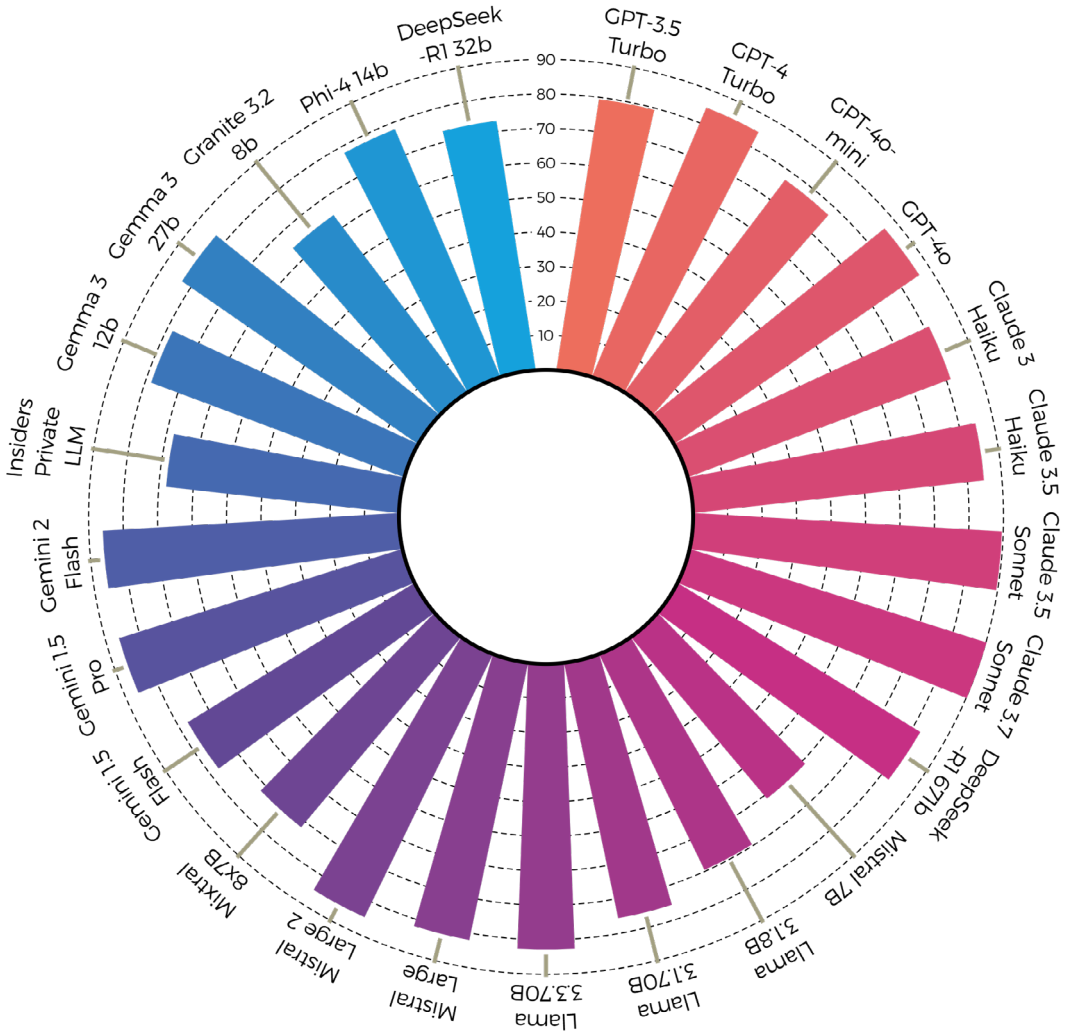
Phone
+49 631 92081 1700

The current comparison shows that global models still deliver the highest performance – driven by massive training datasets, computing power, input volume handling, and parameter counts. Leading the overall ranking is Claude 3.7 Sonnet from Anthropic with a score of 90.17, closely followed by the previous top performer Claude 3.5 Sonnet at 89.61. In third place: GPT-4o by OpenAI with 86.33 – one of the most well-known models on the market.

In contrast, smaller local models like the Insiders Private LLM are optimized for data privacy and compliance – fully operated within the ISO 27001-certified Insiders Cloud, offering the highest level of data protection, especially for sensitive document types such as SEPA mandates or medical data. This intentional tradeoff is critical for data-sensitive industries like insurance and finance. With continuous benchmarking, the Insiders Private LLM is being systematically improved and will soon achieve even higher performance.

So what does “the best LLM” really mean?

The latest Insiders LLM Benchmarking makes it clear: As an AI expert, Insiders keeps a close eye on the LLM market and successfully balances performance and data security for its clients using a best-of-breed approach.



This means Insiders continuously identifies the most capable LLMs available and integrates them flexibly into its products. When a new model hits the market, it is tested through our benchmarking and compared with others. The insights gained flow directly into product development, ensuring consistently high quality for Insiders customers.

The new benchmarking confirms: there's no simple answer to the question of "the best LLM" Performance alone is not enough. In highly regulated industries like insurance and finance, reliability, data privacy, and integration capability are just as critical.

Key takeaways:

- Claude 3.7 Sonnet is currently the top performer – fast and powerful, but limited to global data privacy standards.
- Gemini 2 Flash sets new benchmarks for speed – ideal for high-volume processing, with minor tradeoffs in precision.
- Models hosted by Insiders, such as Gemma 3 27b or Phi-4 14b, now achieve very respectable results – with full data sovereignty and no rate limits.
- The Insiders Private LLM may lag in raw performance, but excels where others fall short: maximum data privacy, full control, local processing, and full transparency.

The Insiders Best-of-Breed Approach

Our benchmarking fully supports the best-of-breed philosophy: We continuously test the most relevant models, integrate them via our OvAltion Engine, and enable customers to choose the optimal LLM setup for their individual requirements.

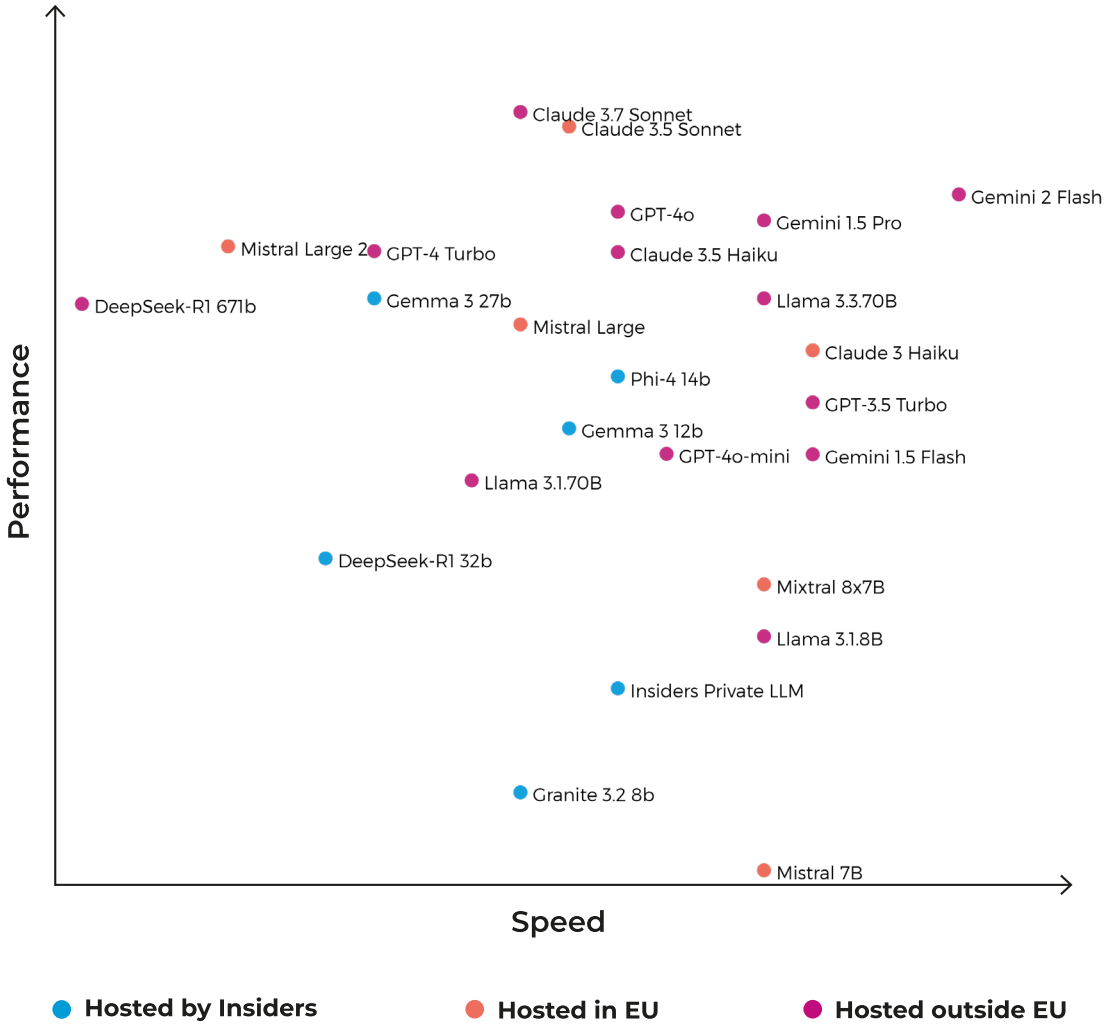
Our principles:

- Customers don't have to choose between performance and data privacy.
- The combination of proven Insiders AI and state-of-the-art LLMs delivers maximum automation with minimal risk.
- Features like Green Voting automatically validate LLM results, reduce manual post-processing, and increase straight-through processing rates.

Depending on the specific needs – balancing performance, latency, automation depth, and cost – Insiders gives customers flexible and convenient access to exactly the LLM they need. There's no compromise between performance and security.

With our ISO-certified infrastructure and seamless integration through the OvAltion Engine, OmnIA offers an automation platform that is both secure and future-ready. The Insiders LLM Benchmarking serves as a reliable reference to stay ahead in the fast-paced LLM market.

For customized benchmarking based on your own use cases, our AI experts are here to support you. Reach out to schedule a tailored benchmark


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
Hosted by Insiders
Hosted in EU
Hosted outside EU