insiders

LLM Benchmarking – September 2025

The best LLMs in comparison

The Insiders LLM benchmarking in September 2025 continues the series and builds consistently on the findings from Q2. To ensure comparability, identical dimensions and test data are used as in the previous benchmarking.

As we continue to keep an eye on new models and also discard models, the current benchmarking includes 21 large language models, including new, powerful models such as GPT-5, Gemini 2.5 Pro, and Claude 4 Sonnet. With this flexibility in the benchmark, we are responding to the dynamics in the LLM field: We continuously integrate new models in line with the latest research and evaluate their transfer into practical IDP solutions. In doing so, we pay particular attention to achieving a balanced combination of maximum result quality and operational stability.

The benchmarking was based on a standardized IDP data set with real documents from the insurance and finance sectors. This ensures that the results are directly transferable to our customers‘ requirements.

The models that Insiders had hosted itself for testing purposes until the last benchmarking – such as Gemma 3 27b or DeepSeek-R1 32b – did not prove themselves in practice and were therefore removed from the portfolio.

View of the current results

The performance of the models is measured on a scale of 0 to 100: 100 corresponds to error-free processing, while 0 means completely incorrect results. In addition, the processing speed of the models is represented in a simplified speed level: Level 3 stands for very fast models, Level 2 for medium speed, and Level 1 for slow processing. This allows the accuracy and efficiency of the models to be compared.

By switching to a more powerful model, Insiders Private was able to achieve a significant leap in quality: from a score of 67.9 in Q2 to 78.2 now—while maintaining the same average processing time per document. This brings it closer to the top models without compromising on data protection or speed.

It is also clear that reasoning models – LLMs that have been specifically trained in complex logical thinking, etc. – such as GPT-5 appear to achieve the best results in classification and extraction (90.7 points, highest rating in the field), but these advantages come with noticeable disadvantages and depend heavily on the respective model. Processing times are four times higher, and token costs also increase accordingly – an aspect that should not be neglected for productive use. Reasoning models should therefore be used with caution in practice and only in meaningful use cases.

It is also noticeable that more and more models are located in the upper right quadrant of the evaluation – models that did not perform well enough in the past are no longer supported and are being phased out.

Another finding: In the current benchmarking, only three models with EU hosting remain, and only the Insiders models run completely privately. This makes the gap between maximum performance and a regulatory-compliant operating environment even more apparent.

As of: September 11, 2025

Top results at a glance:

Modell	Speed Level	Data Protection	Performance
GPT-5	1	Hosted outside the EU	90,7
Claude 4 Sonnet	3	Hosted in the EU	90,0
Claude 3.7 Sonnet	3	Hosted in the EU	89,9
Gemini 2.5 Pro	2	Hosted outside the EU	89,8
Claude 4 Opus	2	Hosted outside the EU	88,5
GPT-4.1	3	Hosted outside the EU	87,9
Gemini 2.5 Flash	3	Hosted outside the EU	87,8
GPT-5 mini	2	Hosted outside the EU	87,4
GPT-4o	3	Hosted outside the EU	86,4
Mistral Large 2	3	Hosted outside the EU	85,2
Claude 3.5 Haiku	3	Hosted outside the EU	84,7
GPT-OSS 120b	3	Hosted outside the EU	84,2
GPT-4.1 mini	3	Hosted outside the EU	83,8
GPT-o3 mini	2	Hosted outside the EU	83,7
DeepSeek-R1 671b	1	Hosted outside the EU	82,5
Claude 3 Haiku	3	Hosted in the EU	81,8
GPT-OSS 20b	3	Hosted outside the EU	80,6
Insiders OvAItion LLM	3	Hosted by Insiders	80,1
Gemini 2 Flash	3	Hosted outside the EU	79,5
Insiders Private	3	Hosted by Insiders	78,2
GPT-4o mini	3	Hosted outside the EU	77,7

It has been confirmed that global models set the benchmark – supported by enormous databases, superior hardware resources, and huge model architectures. Open AI's GPT-5 model ranks first in the overall ranking with a score of 90.7, closely followed by Claude 4 Sonnet (90.0) and the previous winner Claude 3.7 Sonnet with 89.9.

In contrast to global models, Insiders Private LLM is optimized for data protection and regulatory security. Operation in the ISO 27001-certified Insiders Cloud enables its use for particularly critical document types, from SEPA mandates to health data. This approach is a decisive advantage, especially in information-sensitive industries such as finance and insurance. Continuous benchmarking ensures the further development of the model.

So what does "the best LLM" mean?

The latest Insiders LLM benchmarking shows that Insiders continuously monitors the market and masters the balancing act between performance and security for its customers – with a clear best-of-breed approach. This approach means that no single model covers all tasks, but rather that the most powerful LLMs are identified, evaluated, and flexibly integrated for each application. New models are therefore immediately tested in benchmarking and compared with existing ones. The results are directly incorporated into product development and ensure consistently high quality.

It also becomes clear that there is no blanket answer to the question of which LLM is best. Highest performance alone is not enough. Especially in regulated industries such as insurance and finance, additional factors such as reliability, data protection, and integrability play a decisive role.

The most important findings:

- **Claude 4 Sonnet** leads the pack – fast and powerful, hosted in the EU.
- **Claude 3 Haiku** is a proven model that shines with incredible speed – ideal for volume processing, with manageable accuracy losses.
- **Top-Performer** among LLMs are initially operated in the US – EU hosting usually follows with a time delay.
- **Insiders Private LLM** is privately hosted, offering full control, local processing, and maximum transparency.
- As a prototype, **Insiders OvAItion LLM** already outperforms Private LLM in terms of performance and speed – but is still being tested and trained.

Insiders Best-of-Breed Approach

Our benchmarking supports the best-of-breed approach: we continuously test the most relevant models, integrate them via our OvAItion Engine, and thus give customers the opportunity to choose the optimal setup for their individual requirements from a wide range of LLMs.

The following applies:

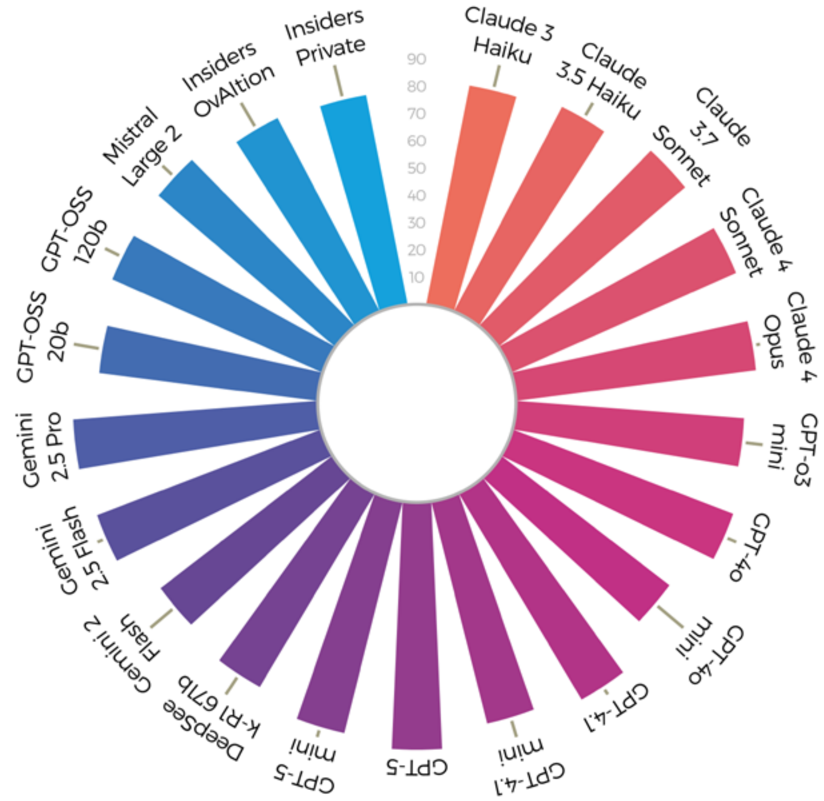
- Customers do not have to choose between performance and data protection.
- The combination of proven Insiders AI and state-of-the-art LLMs delivers maximum automation with minimal risk.
- Features such as Green Voting automatically validate LLM results, reduce post-processing effort, and increase the dark processing rate.

Depending on individual requirements in the area of performance, latency, dark processing, and costs, Insiders enables its customers to conveniently and flexibly access exactly the LLM they really need by integrating third-party LLMs. Insiders‘ customers therefore do not have to choose between performance and security. They get both, tailored to their individual needs – while remaining flexible to explore the diverse applications of LLMs for their business.

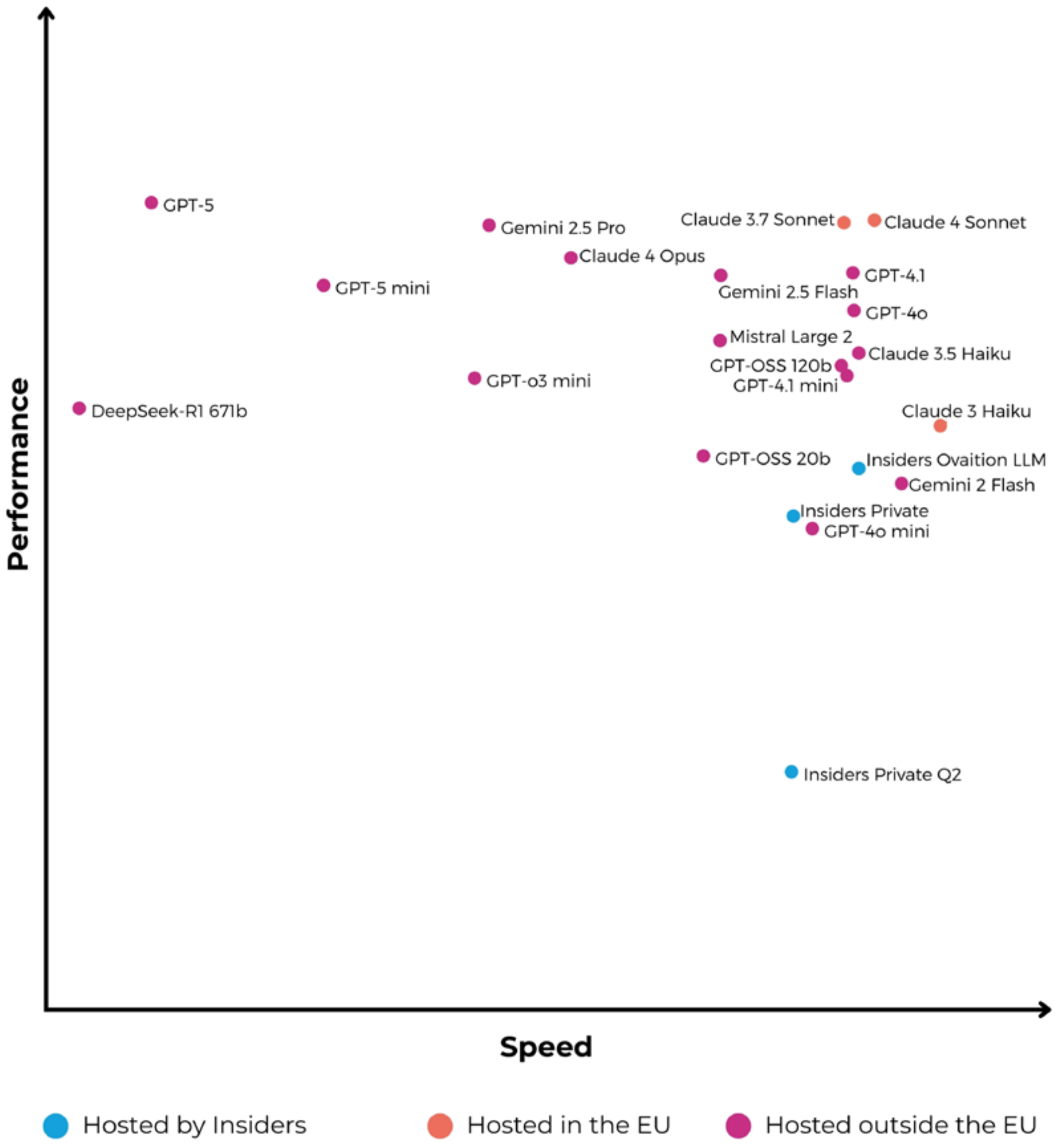
With its ISO-certified infrastructure and seamless integration via the Insiders OvAItion Engine, Insiders offers OmnIA, an automation platform that is not only secure but also future-proof. Insiders LLM Benchmarking is the reliable reference point for keeping track of the fast-moving LLM market.

For individual benchmarking with your own use cases, Insiders AI experts offer sound advice for your company. Simply contact our Insiders AI experts for individual benchmarking.

llm-benchmarking@insiders-technologies.de



As of: September 11, 2025



As of: September 11, 2025