insiders

Insiders LLM Benchmarking – Q4 2025

The best LLMs in comparison

At the end of 2025, we are publishing the results of our fourth LLM benchmarking cycle, consistently building on previous editions. For this release, we nearly doubled the dataset and added significantly more challenging documents. This increases the benchmark's validity, even though overall scores drop slightly due to the higher complexity. The expansion reflects our aim to capture the rapid progress in the LLM landscape and to test models under conditions that mirror real productive IDP workflows as closely as possible.

This benchmarking includes 24 large language models, among them new entries like Claude 4.5 Sonnet, Gemini 3, and GPT-5.1. As in the previous quarter, some models were removed for various reasons — for example, because successor models deliver similar results at comparable cost and speed.

A look at the current results

Model performance is measured on a 0–100 scale: 100 represents error-free processing, 0 represents fully incorrect output. Processing speed is shown via a simplified speed level: Level 3 indicates very fast models, Level 2 medium speed, Level 1 slow processing. This makes accuracy and efficiency easy to compare.

Due to the doubled dataset and higher document complexity, overall scores are slightly lower than last quarter. The models are pushed harder, and the differences between generic foundation models — broadly trained on massive, often unspecific data — and specialized IDP models become more pronounced.

A growing number of models now cluster in the "upper mid-range." Competition tightens: many models sit close together, especially between 80–88 points. Weaker models were discontinued or deliberately excluded. The market level is consolidating, and variance is decreasing.

Once again, dedicated reasoning models deliver strong results in classification and extraction. At the same time, the structural downsides remain: longer processing times, higher token costs, and lower predictability in production. GPT-5 and GPT-4.1, for example, achieve excellent overall scores of 87.3 and 84.7 but come with major drawbacks regarding data protection or speed.

Compared with the last quarter, more EU-hosted models appear in our selection — though they remain scarce across the broader market.

As of: 03.12.2025

Top results at a glance:

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

Market trends:
Big models — small steps

The latest versions of major foundation models show only marginal improvements in the IDP context:

- **Claude 4.5 Sonnet vs. 4 Sonnet:** practically identical scores of 87.9 and 87.3
- **Gemini 3 Pro vs. 2.5 Pro:** similar score, but nearly twice the runtime per document
- **GPT-5.1 vs. GPT-5:** slightly weaker, but noticeably faster

The benefit of "bigger" models is flattening. Speed gains or small quality shifts are common — real leaps are rare. You don't automatically need the newest or most expensive model to achieve strong IDP results.

Specialized models deliver the real progress

The biggest quality gains — as in previous benchmarks — come not from generic foundation models but from specialization.

Our own model demonstrates this clearly: the OvAItion Private LLM achieves more than a 2% improvement compared with Q3 — despite tougher data. This result is no coincidence: our existing Private LLM is merging with the upcoming OvAItion LLM into the "OvAItion Private LLM," offering maximum security combined with rising quality and domain specialization for our customers' IDP environments.
This places the OvAItion Private LLM for the first time close to Claude 4.5 Haiku — a clear milestone. And development continues: we expect further gains from ongoing finetuning, optimized training, and additional data built on years of experience.

Self-hosted LLMs as a strategic advantage

A self-hosted model gives organizations full data sovereignty, C5-certified security, predictable costs, customizability, and operational stability.

Especially in document processing, the pattern holds: specialization beats size.

Key takeaways

- Large foundation models operate at a high level, but progress in the IDP context is slowing

As of: 03.12.2025

- Reasoning models score well but are often inefficient in practice
- Under real IDP conditions, benefits remain limited: overhead outweighs added quality
- High performance and regulatory safety rarely go hand in hand

The OvAItion Private LLM shows the strongest overall progress. Once again, the most visible leap comes from specialization, not global models. For the first time, we reach a notable proximity to leading foundation models — with stable processing time, fully private deployment, C5 certification, and full data protection compliance.

This combination — competitive performance + complete data sovereignty — remains unique in the benchmark and continues to stand out.

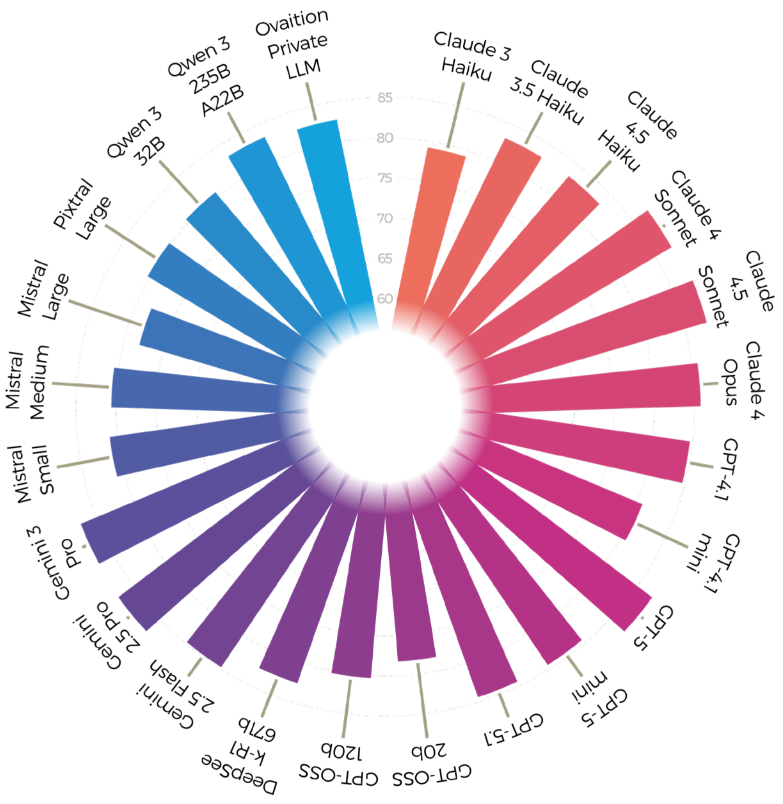
The Insiders best-of-breed approach

Our benchmarking supports a best-of-breed strategy: we continuously test the most relevant models, integrate them through our OvAItion Engine, and give customers the ability to choose the optimal LLM setup for their specific needs.

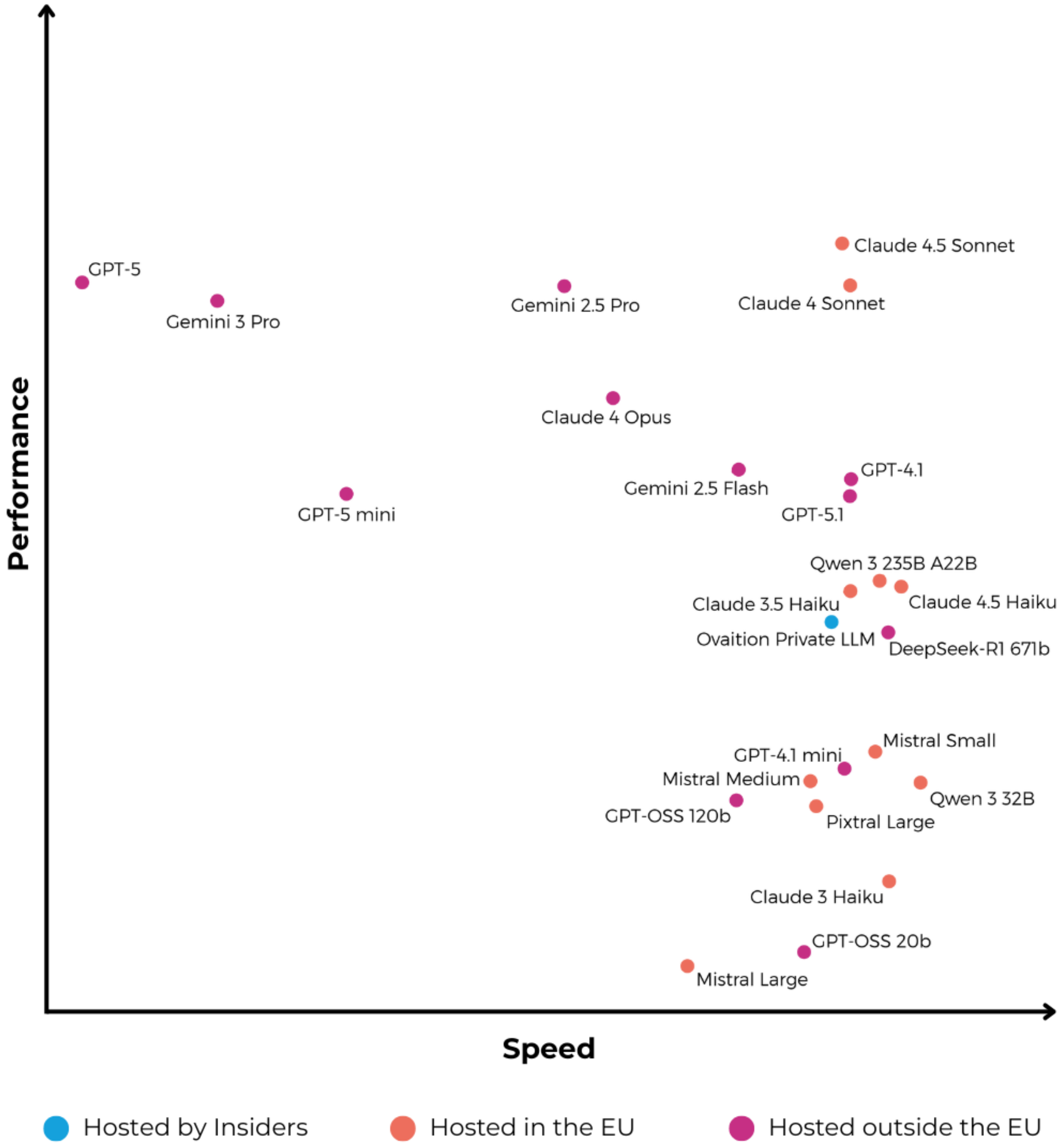
Key principles:

- Customers don't have to choose between performance and data protection.
- The combination of proven Insiders AI and state-of-the-art LLMs delivers high automation with low risk.
- Features like Green Voting validate LLM outputs automatically, reduce rework, and increase straight-through processing.

Depending on requirements across performance, latency, automation rate, and cost, Insiders enables flexible access to exactly the LLM each customer needs — without forcing a trade-off between performance and security.



As of: 03.12.2025



As of: 03.12.2025