LLM Benchmarks March 2025

Rinat Abdullin

22 August, 2025

Table of Contents

We are catching up with the latest events and working on the new RPA (SAP) column for the benchmark. So this report will be more condensed than usual. **Google DeepMind managed to surprise us more than once last month.**

- Gemini-2.5 Pro Preview
- DeepSeek V3 0324
- Llama 4 models
- Google Gemma 3 models
- Focus on RPA

1 Gemini-2.5 Pro Preview - takes 2nd place!

Google has released a couple of notable multimodal models. Let's start with **Gemini-2.5 Pro Preview** (already available on Vertex AI). This is DeepMind's most advanced LLM[1], designed to internally reason through complex problems before answering. This chain-of-thought approach yields high accuracy on difficult tasks, excelling in coding, math, and scientific problem-solving.

This large thinking model features native multimodality (can work with documents, images, audio and video) and has a theoretical context limit of 1M (in practice, effective context might be much smaller, if the tasks are cognitively challenging).

The model debuted on the top of LLM Arena[2] (arena where humans pick chatbot answers that they like):

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1439	+7/-10	5858	Google	Proprietary
2	5	Llama-4-Maverick-03-26-Experimental	1417	+13/-12	2520	Meta	N/A
2	1	ChatGPT-4o-latest (2025-03-26)	1410	+8/-10	4899	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1403	+6/-6	12391	xAI	Proprietary
3	2	GPT-4.5-Preview	1398	+5/-7	12312	OpenAI	Proprietary
6	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-4	24298	Google	Proprietary
6	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	20289	Google	Proprietary
6	4	DeepSeek-V3-0324	1369	+10/-10	3526	DeepSeek	MIT
8	5	DeepSeek-R1	1358	+5/-5	14259	DeepSeek	MIT

Our reasoning benchmark is not as much about chatting, but rather solving precise business problems. Still, Gemini 2.5 Pro managed to get to the second place, beating Claude 3.7 Sonnet Reasoning and OpenAI o1 (not pro).

1. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025
2. https://lmarena.ai/?leaderboard

# ▲	Model	bi	compliance	code	reason	Score	Features
1	openai/o3-mini-2025-01-31	45%	70%	100%	74%	76.7%	SO, Reason
2	google/gemini-2.5-pro-preview-03-25	45%	70%	93%	68%	71.1%	Reason
3	anthropic/claude-3.7-sonnet:thinking	54%	32%	100%	67%	70.4%	Reason
4	openai/o1-2024-12-17	45%	70%	84%	67%	70.0%	SO, Reason
5	deepseek/deepseek-r1	27%	64%	100%	63%	66.1%	SO, Reason, Open
6	deepseek/deepseek-r1-distill-llama-70b	36%	32%	96%	56%	60.0%	Open
7	deepseek/deepseek-chat-v3-0324	45%	60%	70%	55%	59.6%	Reason, Open
8	anthropic/claude-3.7-sonnet	45%	47%	65%	55%	56.5%	
9	openai/gpt-4o-2024-11-20	36%	55%	62%	55%	53.6%	SO
10	openai/gpt-4.5-preview-2025-02-27	45%	47%	62%	53%	51.9%	SO
11	deepseek/deepseek-chat	36%	47%	58%	49%	50.6%	SO, Open
12	openai/gpt-4o-2024-08-06	18%	62%	63%	52%	50.5%	SO
13	microsoft/phi-4	36%	62%	57%	48%	49.7%	Open

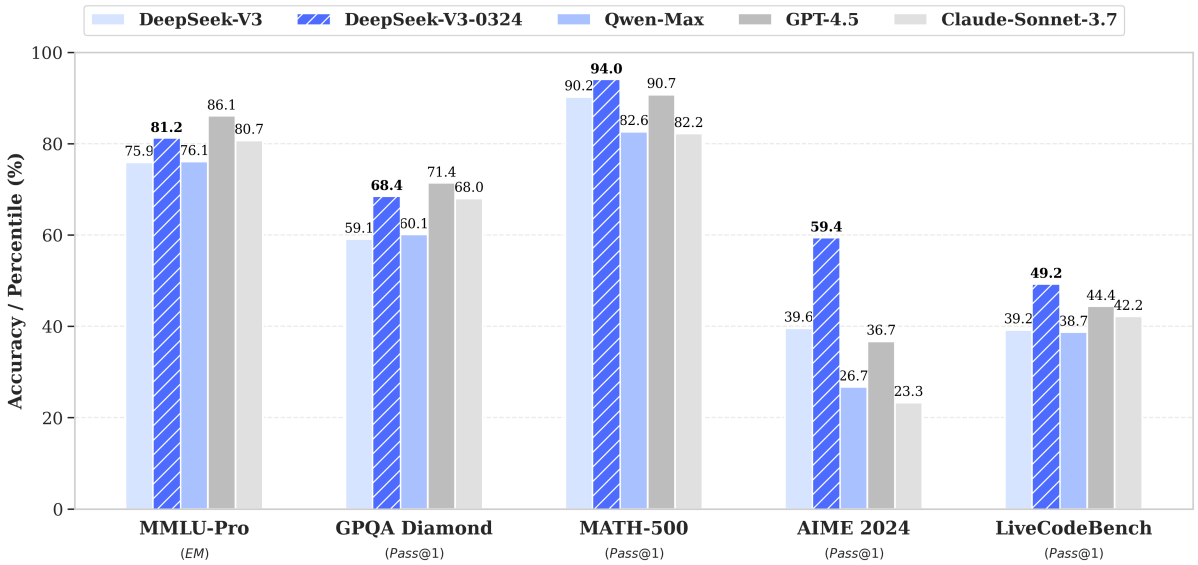
And the most interesting part - Gemini 2.5 Pro worked without the support from the Structured Outputs (because Google only supports a tiny subset of JSON schema), still it managed to get the second place.

2 DeepSeek V3-0324 - Progressive update over previous version

DeepSeek V3-0324 is an ultra-large model (685B parameters) that uses a Mixture-of-Experts architecture. It activates specialized "experts" for different query types, giving it broad knowledge and skills while remaining somewhat efficient for its size.

In theory, anybody can download this model[3] to run it on the local hardware. In practice, its size makes it somewhat useless for its size. We still need to keep the entire model in GPU VRAM, to be able to switch between different experts.

Authors claim that V3-0324 release made major leaps[4] in logic and knowledge tasks, with significantly higher scores on benchmarks like MMLU, GPQA, and AIME than its predecessor.



In our enterprise benchmark we see a similar scale of improvement. Model got better:

3. https://huggingface.co/deepseek-ai/DeepSeek-V3-0324/tree/main
4. https://huggingface.co/deepseek-ai/DeepSeek-V3-0324

#	Model	bi	compliance	code	reason	Score	Features
1	openai/o3-mini-2025-01-31	45%	70%	100%	74%	76.7%	SO, Reason
2	google/gemini-2.5-pro-preview-03-25	45%	70%	93%	68%	71.1%	Reason
3	anthropic/claude-3.7-sonnet:thinking	54%	32%	100%	67%	70.4%	Reason
4	openai/o1-2024-12-17	45%	70%	84%	67%	70.0%	SO, Reason
5	deepseek/deepseek-r1	27%	64%	100%	63%	66.1%	SO, Reason, Open
6	deepseek/deepseek-r1-distill-llama-70b	36%	32%	96%	56%	60.0%	Open
7	deepseek/deepseek-chat-v3-0324	45%	60%	70%	55%	59.6%	Open
8	anthropic/claude-3.7-sonnet	45%	47%	65%	55%	56.5%	
9	openai/gpt-4o-2024-11-20	36%	55%	62%	55%	53.6%	SO
10	openai/gpt-4.5-preview-2025-02-27	45%	47%	62%	53%	51.9%	SO
11	deepseek/deepseek-chat	36%	47%	58%	49%	50.6%	SO, Open
12	openai/gpt-4o-2024-08-06	18%	62%	63%	52%	50.5%	SO
13	microsoft/phi-4	36%	62%	57%	48%	49.7%	Open
14	meta-llama/llama-4-maverick	27%	42%	70%	44%	49.1%	SO, Open

NB: There are no OpenRouter providers running V3 Chat with Structured Output at the moment. So it was tested without the support of SO. Still, it managed to improve the score.

3 Llama 4 Models from Meta - Nothing peculiar

Recently released Llama 4 models are about **open-source multimodal intelligence.**

Llama 4 introduces MoE to the Llama family, greatly boosting efficiency. The *Llama 4 Maverick* model leverages 400B total parameters with only ~17B active per query, yielding **faster responses** and lower inference cost without sacrificing quality. It targets 1M context size. It has 128 experts under the hood.

Llama 4 Scout model uses 16 experts, keeping it size to 109B parameters and targeting 10M context length.

Both models — target **12 primary languages** (English, French, German, Arabic, Hindi, Indonesian, Italian, Portuguese, Spanish, Tagalog, Thai, and Vietnamese), but pretrained on **over 200 languages** (via the *No Language Left Behind*[5] initiative).

> Perhaps you remember our earlier post about Homai and AI for Good[6]. Thanks to the work and support of Aigiz Kunafin, Meta's NLLB initiative already includes Bashkir language, and the other languages are also on the way.

However despite powerful linguistic and multimodal capabilities, Llama 4 didn't score high on our enterprise benchmark.

#	Model	bi	compliance	code	reason	Score	Features
14	meta-llama/llama-4-maverick	27%	42%	70%	44%	49.1%	SO, Open
18	meta-llama/llama-3.1-70b-instruct	36%	50%	44%	43%	42.6%	SO, Open
19	meta-llama/llama-3.3-70b-instruct	27%	50%	48%	41%	40.8%	SO, Open
26	meta-llama/llama-3.1-405b-instruct	18%	55%	40%	38%	35.5%	SO, Open
34	meta-llama/llama-4-scout	9%	25%	22%	16%	18.0%	SO, Open
36	meta-llama/llama-3.2-3b-instruct	0%	17%	16%	11%	10.6%	SO, Open
	Averages	**27%**	**38%**	**55%**	**41%**		

This is fine and expected for two reasons:

- Historically Llama models never score high on our benchmarks. Usually fine-tunes come out later that manage to do that better.
- Llama 4 models were not trained on reasoning workloads. Unlike reasoning models, they couldn't make use of reasoning slots within the Custom Chain of Thought schemas. They just tried to jump to the answer straight away.

We'll wait for r1 distills on top of llama-4 pruned expert trees.

5. https://ai.meta.com/research/no-language-left-behind/
6. https://www.timetoact-group.at/en/techblog/techblog/celebrating-homai-using-ai-for-good

4 Google Gemma 3 - this is where things get interesting

Gemma 3 27B is Google's latest 27B open model that delivers **state-of-the-art results for its size**, rivaling or beating models many times larger.

In human evaluations (Chatbot Arena), it outscored a 405B-parameter Llama 3 and the 685B DeepSeek V3, all while running on a single GPU. It is also multimodal, supports context up to 128k and was trained to do function calling.

Numbers check out on our side, too. This small and open model works really well for its small size, beating on our benchmark models much larger than its size.

Best part? Gemma-3-27B was tested without Structured Outputs and managed to respond while following a fairly complex schema and utilising custom chain of thought slots in it.

13	microsoft/phi-4	36%	62%	57%	48%	49.7%	Open
14	meta-llama/llama-4-maverick	27%	42%	70%	44%	49.1%	SO, Open
15	qwen/qwen-max	45%	45%	45%	50%	46.3%	
16	google/gemma-3-27b-it	27%	27%	70%	43%	45.0%	Open
17	anthropic/claude-3.5-sonnet	36%	32%	57%	44%	43.6%	
18	meta-llama/llama-3.1-70b-instruct	36%	50%	44%	43%	42.6%	SO, Open
19	meta-llama/llama-3.3-70b-instruct	27%	50%	48%	41%	40.8%	SO, Open
20	google/gemini-2.0-flash-001	27%	24%	57%	38%	40.7%	
21	qwen/qwq-32b	36%	52%	41%	37%	40.0%	SO, Reason, Open
22	qwen/qwen-2.5-72b-instruct	27%	30%	47%	39%	39.2%	SO, Open

Its smaller sibling - Gemma-3-12B also delivered similar results - it manages to beat models of much larger size, also without Structured Outputs.

23	mistralai/mistral-small-3.1-24b-instruct	36%	42%	41%	39%	39.2%	SO, Open
24	qwen/qwen2.5-32b-instruct	27%	20%	53%	36%	36.6%	Open
25	qwen/qwen-2.5-coder-32b-instruct	18%	35%	54%	39%	36.5%	SO, Open
26	meta-llama/llama-3.1-405b-instruct	18%	55%	40%	38%	35.5%	SO, Open
27	google/gemma-3-12b-it	9%	17%	61%	30%	33.4%	Open
28	qwen/qwen-plus	18%	25%	40%	31%	31.7%	
29	mistralai/mixtral-8x22b-instruct	9%	27%	47%	28%	29.2%	SO, Open
30	openai/gpt-4o-mini-2024-07-18	9%	32%	41%	30%	28.4%	SO
31	mistral/mistral-small-24b-instruct-2501	27%	22%	33%	30%	27.8%	SO, Open

It looks like Google DeepMind has discovered some "secret sauce" that lets them reliably train State-of-the-Art models of different sizes. Obviously, having powerful LLM in Google Cloud is interesting for

the business, however having small open models that benefit from reasoning and perform really well - is even better.

Best part? Google didn't stop there.

You can download Gemma-3-27B-it[7] from Hugging Face and run it on your local server. It is ~55GB of weights which you'll need to load in GPU in bf16 format (two bytes per weight). It would require ~ 60GB VRAM for text-based tasks and ~70GB VRAM for vision tasks. This means H100 80GB.

However, if you want to run this model on a smaller budget, there is another alternative. Google has also created a special version of this model[8] that is much smaller. It is saved in GGUF format corresponding to Q4_0 quantisation (roughly equals to 4 bits per weight). Thanks to Quantization Aware Training (QAT) this model is able to preserve similar quality as while significantly reducing the memory requirements to load the model.

This overall push towards open, powerful, and small (pick all three) models is great. It is essential for building business systems with Trustworthy AI under the hood. We'll see, how the things evolve from there.

This trend also aligns well with our strategic focus for this year.

7. https://huggingface.co/google/gemma-3-27b-it
8. https://huggingface.co/google/gemma-3-27b-it-qat-q4_0-gguf

5 New direction - Enterprise Reasoning and Robotic Process Automation

AI Case portfolio, LLM Benchmarks, Enterprise Challenges[9] and various events - all are our vehicles for pushing AI R&D forward together with a talented community around the world.

> By the way, if you are in Vienna on June 18th, come join IBM, Cloudera and us for "Designing Trustworthy AI[10]" event. We'll discuss a range of topics, from Data governance to agents and public AI R&D efforts.

Previously we have focused on generic business tasks and reasoning. This reflected in BI, Compliance and Reason categories in the LLM benchmark and culminated in Enterprise RAG Challenge.

Next, we are going to ground our tasks closer to the everyday enterprise challenges. We'll focus more on the **Robotic Process Automation (RPA) and Enterprise Reasoning**.

RPA can be described as "automate repetitive SAP workflows via human interfaces". Historically it was done via rules and browser automation. Recent progress in Large Language Models gives a new angle to approach this problem - via operators and visual agents.

We are going to invest part of this year to dive deeper into this problem:

- Import visual automation cases into LLM benchmark under RPA column
- Work on a dedicated Operator/Agent benchmark together with our industry peers.
- Ultimately - setup and run Enterprise RPA Challenge.

Our partners and customers are quite interested in the possibility of AI automation in modern enterprise software: SAP, Salesforce, ServiceNow. We are going to try to explore this field further, share our findings with the community and, perhaps, push State-of-the-Art forward together here.

9. https://www.timetoact-group.at/landingpages/enterprise-rag-challenge
10. https://www.timetoact-group.at/events/designing-trustworthy-ai