LLM Benchmarks January 2025

Rinat Abdullin

22 August, 2025

Table of Contents

This Benchmark report is going to be interesting. We'll start with benchmarks and end up with Nvidia stock price forecast (not to be treated as a financial advise)

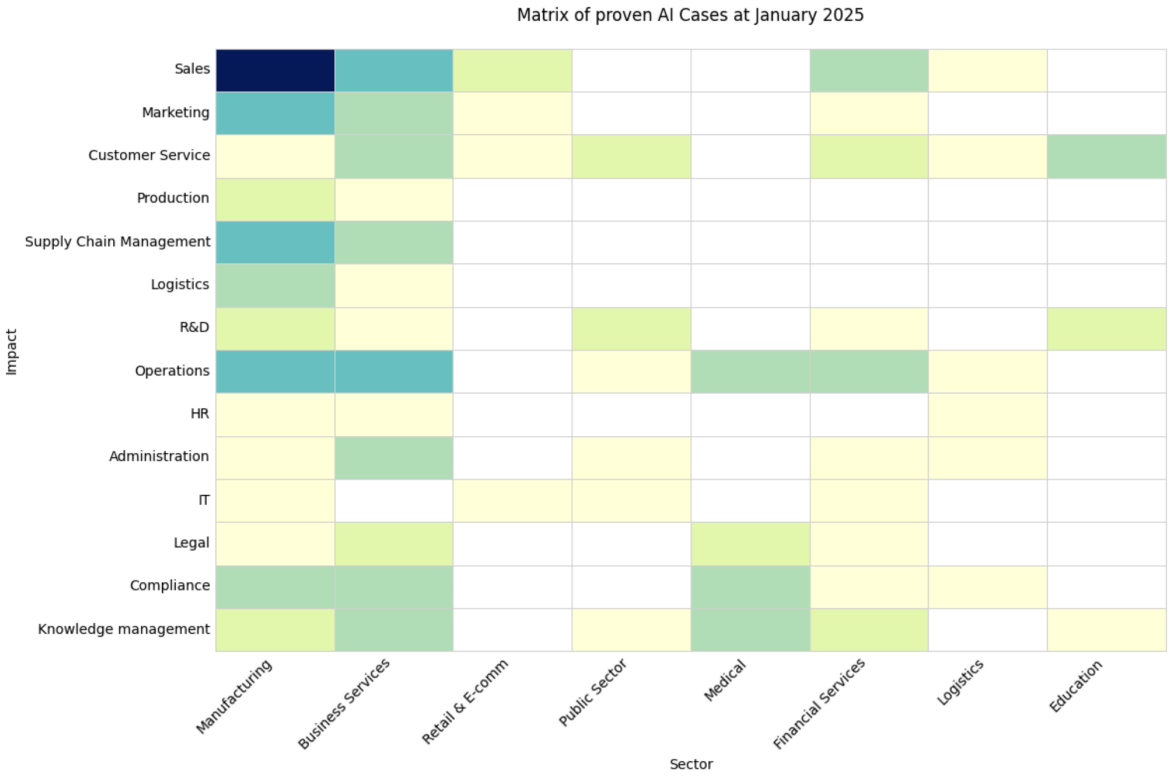
- Second generation benchmark - early preview
- DeepSeek r1
- Cost and price dynamics of DeepSeek r1

So, let's get things going.

1 LLM Benchmark gen2 - early preview

In the past few months we've been working heavily on redoing our first-generation of LLM Benchmark. Get 1 focused on business workload automation, but relied on insights from AI Cases completed in year 2023.

This started showing in the last months of the benchmark results as saturated top rows (too many models with high scores). Besides, the test cases themselves were getting a bit stale. They didn't properly represent insights from the last year worth of our AI Research and practice with the companies in EU and USA.



So we've been building a new generation of benchmark to incorporate both new LLM capabilities and new insights. **The timing was just perfect - o1 pro came out to challenge the complexity of the benchmark, while DeepSeek r1 shortly thereafter introduced a concept of accessible reasoning**.

So here is the early preview of our v2 benchmark. It doesn't look like much on the outside, but it already deterministically compares models below on complex business tasks, while allowing each model to reason before producing an answer.

Model	compliance	reason	bi	Score	Features
openai/o1-2024-12-17	50%	61%	37%	67%	SO
openai/gpt-4o-2024-08-06	75%	39%	12%	56%	SO
deepseek/deepseek-r1	50%	41%	25%	53%	SO
microsoft/phi-4	50%	48%	37%	51%	
deepseek/deepseek-r1-distill-llama-70b	25%	43%	25%	50%	
openai/gpt-4o-2024-11-20	75%	41%	25%	49%	SO
meta-llama/llama-3.1-405b-instruct	75%	35%	25%	49%	SO
meta-llama/llama-3.3-70b-instruct	75%	34%	37%	48%	SO
qwen/qwen-2.5-coder-32b-instruct	50%	41%	25%	45%	SO
deepseek/deepseek-r1-distill-qwen-32b	25%	27%	25%	39%	SO
anthropic/claude-3.5-sonnet	25%	32%	12%	35%	
qwen/qwen-2.5-72b-instruct	50%	31%	25%	34%	SO
openai/gpt-4o-mini-2024-07-18	50%	26%	25%	30%	SO
mistral/mistral-small-24b-instruct-2501	25%	21%	12%	26%	SO

We'll go into the DeepSeek r1 analysis in a bit, let's just focus on the benchmark itself. Here is the current progress and an overview what we plan to ultimately add there.

- **~10% of relevant AI cases are currently mapped to v2.** As we get closer to 100% numbers will get more representative of the state of latest applications of AI/LLM in the modern business workloads
- It is the common industry standard in our cases to use Structured Outputs (ability to follow predefined schema precisely) whenever it is possible. There is no reason not to - it is supported by OpenAI, Google and local inference as well. So **our benchmark includes locally capable models with constrained decoding whenever possible**.
- Benchmarks currently focus on **business tasks that require multiple logical steps to be performed by LLM in a single prompt**. In reality, not every complex AI project needs that level of creative autonomy. In some cases like regulation, compliance, it is even counterproductive to let the LLM on its own. A smaller locally-capable model walking over the auditable reasoning path can be more desirable. Over the time we will bring back simpler logical tasks along with adding the new "plan generation" category. The end goal is to be able to **see if a powerful cloud-based model could be replaced by a simpler local model working long the predefined workflow in step**.
- Customers and partners have been asking for a long time to get insights into the contents of this benchmark, to serve as inspiration and guidelines in building their own projects. While it was difficult in v1 due to NDA constraints, v2 will have a portion of **clean unburdened tests that we can share upon the request**.
- We have added only a few categories of cases so far, more will come over the time.

The full journey will take a few months, but the hardest part of actually bringing many moving pieces together into a single coherent framework - is done! LLM Benchmark v2 will get only better from this point on.

2 DeepSeek r1

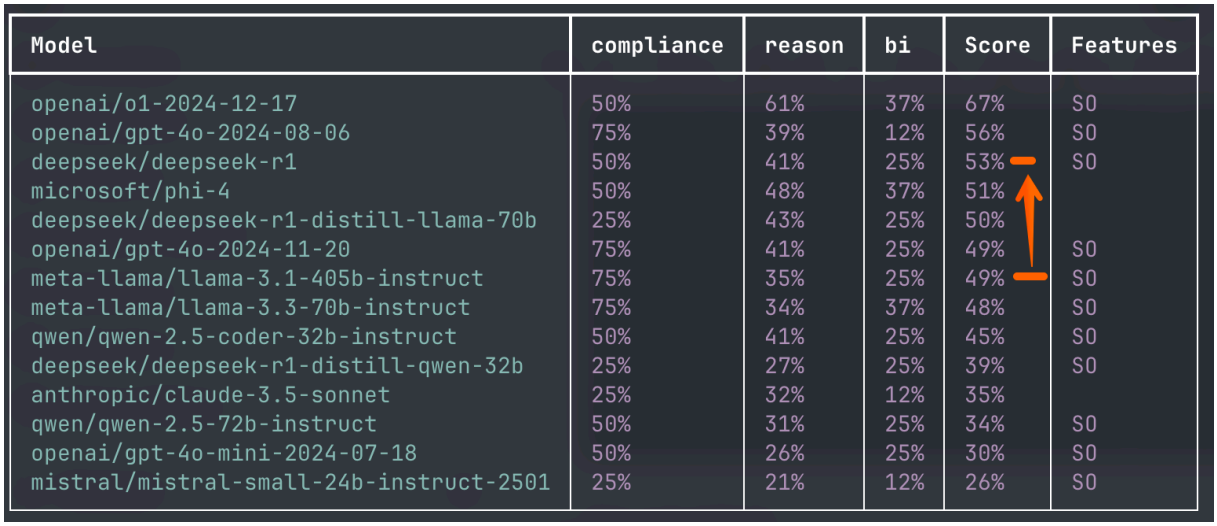
Let's talk about the elephant in the room. DeepSeek r1 is the new Chinese model that is much faster and cheaper than the winning o1 model from OpenAI. Aside from being locally capable (anybody can download it), it is also supposed to be more smart.

No wonder that the stocks went crashing after all these insights.

Let's start with the reasoning. According to our benchmarks DeepSeek r1 is really good. It is better than almost all flavours of 4o. It is also better than any open source model. **It is still worse by OpenAI o1 and GPT-4o from the August 2024.**

Model	compliance	reason	bi	Score	Features
openai/o1-2024-12-17	50%	61%	37%	67%	SO
openai/gpt-4o-2024-08-06	75%	39%	12%	56%	SO
deepseek/deepseek-r1	50%	41%	25%	53%	SO
microsoft/phi-4	50%	48%	37%	51%	
deepseek/deepseek-r1-distill-llama-70b	25%	43%	25%	50%	
openai/gpt-4o-2024-11-20	75%	41%	25%	49%	SO
meta-llama/llama-3.1-405b-instruct	75%	35%	25%	49%	SO
meta-llama/llama-3.3-70b-instruct	75%	34%	37%	48%	SO
qwen/qwen-2.5-coder-32b-instruct	50%	41%	25%	45%	SO
deepseek/deepseek-r1-distill-qwen-32b	25%	27%	25%	39%	SO
anthropic/claude-3.5-sonnet	25%	32%	12%	35%	
qwen/qwen-2.5-72b-instruct	50%	31%	25%	34%	SO
openai/gpt-4o-mini-2024-07-18	50%	26%	25%	30%	SO
mistral/mistral-small-24b-instruct-2501	25%	21%	12%	26%	SO

Also remember that the base Deepseek r1 is Mixture of Experts model containing 685B parameters in total (so you need to fit them all in a GPU). And if we compare score progression to the other large open source model, the progress is somewhat proportional to the size:



Do you see a **smaller elephant in the room that does break this pattern, though**? It is a distillation of DeepSeek r1 capabilities towards Llama 70B! This locally-capable model is not the one everybody is talking about, but it could actually be the biggest deal.

If you can make any good foundational model better by distilling r1 and just letting it reason before producing an answer - this will offer an alternative option for making common models faster.

Model	compliance	reason	bi	Score	Features
openai/o1-2024-12-17	50%	61%	37%	67%	SO
openai/gpt-4o-2024-08-06	75%	39%	12%	56%	SO
deepseek/deepseek-r1	50%	41%	25%	53%	SO
microsoft/phi-4	50%	48%	37%	51%	
deepseek/deepseek-r1-distill-llama-70b	25%	43%	25%	50%	
openai/gpt-4o-2024-11-20	75%	41%	25%	49%	SO
meta-llama/llama-3.1-405b-instruct	75%	35%	25%	49%	SO
meta-llama/llama-3.3-70b-instruct	75%	34%	37%	48%	SO
qwen/qwen-2.5-coder-32b-instruct	50%	41%	25%	45%	SO
deepseek/deepseek-r1-distill-qwen-32b	25%	27%	25%	39%	SO
anthropic/claude-3.5-sonnet	25%	32%	12%	35%	
qwen/qwen-2.5-72b-instruct	50%	31%	25%	34%	SO
openai/gpt-4o-mini-2024-07-18	50%	26%	25%	30%	SO
mistral/mistral-small-24b-instruct-2501	25%	21%	12%	26%	SO

So to summarise:

- DeepSeek r1 model is really good, but **it is not good enough to compete with OpenAI o1, directly yet. It has to beat at least 4o first**.
- The technology itself is quite promising and it will likely result in new breed of more efficient reasoning models derived from the distillation approaches on top of DeepSeek r1.

This might have something to do with our prediction last December[1]:

> *As a new shortcut for improving model reasoning, we believe, **more AI vendors will start providing reasoning capabilities, similar to o1 models**. This will be a temporary workaround to boost model accuracy quick and without heavy investments: just throw more compute, let the model think longer before answering and charge more for the API.*
>
> *However, we also believe, that **the upcoming hype of providing smart reasoning models that are outrageously expensive will also start fading**. It is just not very practical.*

1. https://www.timetoact-group.at/en/details/llm-benchmarks-december-2024

3 Cost and price dynamics of DeepSeek r1

Now, let's talk about the **cost claims of DeepSeek**.

DeepSeek provides its r1 model at very affordable prices. Just $0.55 per 1M input tokens and $2.19 per 1M output tokens.

MODEL[1]	CONTEXT LENGTH	MAX COT TOKENS[2]	MAX OUTPUT TOKENS[3]	1M TOKENS INPUT PRICE (CACHE HIT) [4]	1M TOKENS INPUT PRICE (CACHE MISS)	1M TOKENS OUTPUT PRICE
deepseek-chat	64K	-	8K	~~$0.07~~[5] $0.014	~~$0.27~~[5] $0.14	~~$1.10~~[5] $0.28
deepseek-reasoner	64K	32K	8K	$0.14	$0.55	$2.19 [6]

This is way cheaper than OpenAI o1 pricing or 4o pricing. Let's put things in one table.

We'll also count the total price of a common business workload with a ratio of 10:1 - 10M Input tokens and 1M Output tokens. This 10:1 ratio is very typical for data extraction and RAG-like systems which dominate the landscape of our AI Cases)

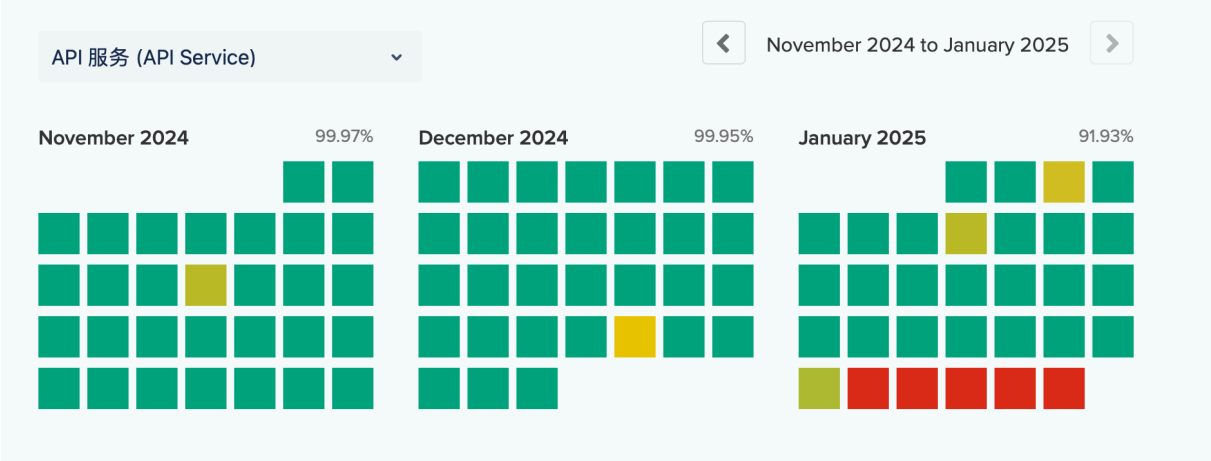
Model	1M Input Tokens	1M Output Tokens	Cost of 10M:1M
DeepSeek r1	$0.55	$2.19	$7.69
OpenAI gpt-4o	$2.5	$10	$35
OpenAI o1	$15.0	$60	$210

At this point we can confidently say that the price of DeepSeek r1 blow everybody else out of the water. They aren't even "25x cheaper than OpenAI o1" on common business workloads, but **27x time cheaper**.

Devil is in the details, though. Currently offered price, for various reasons, might not be exactly equal to the actual market price or related to the cost of running business.

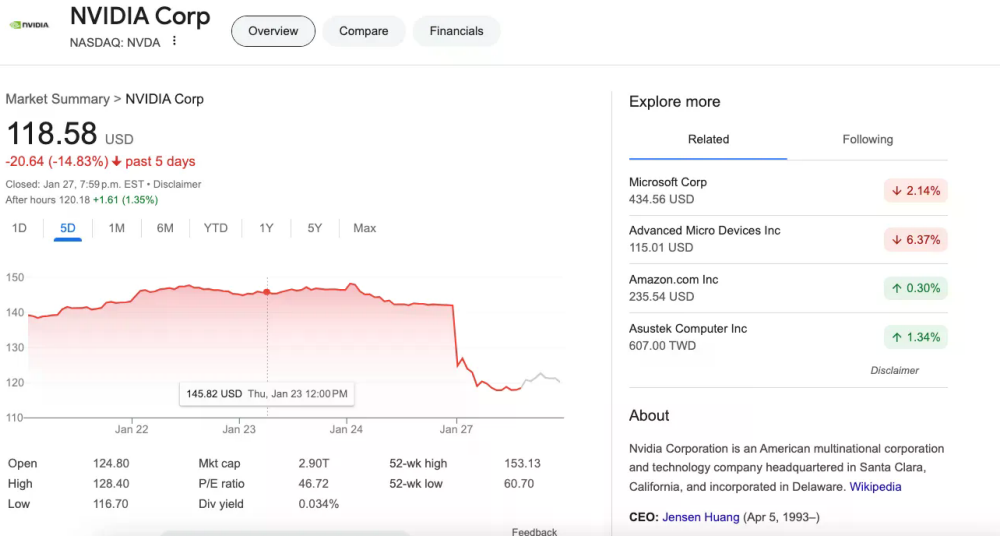
First of all, can DeepSeek even handle all the demand? According to their status page[2], API has been in "Major Outage" mode since January 27th. So they are not actually serving all LLM requests at the advertised price:

2. https://status.deepseek.com/uptime



In general, if you look at the financial incentives of DeepSeek as the company, you might discover that turning profit might not be their primary motivation. DeepSeek is owned by a Chinese High-Flyer hedge fund (see wikipedia[3]), so in theory they can make more money by shorting Nvidia. We'll leave that theory aside.

Although, it is still a funny coincidence that January 27th is also the day when NVidia stock went crashing.

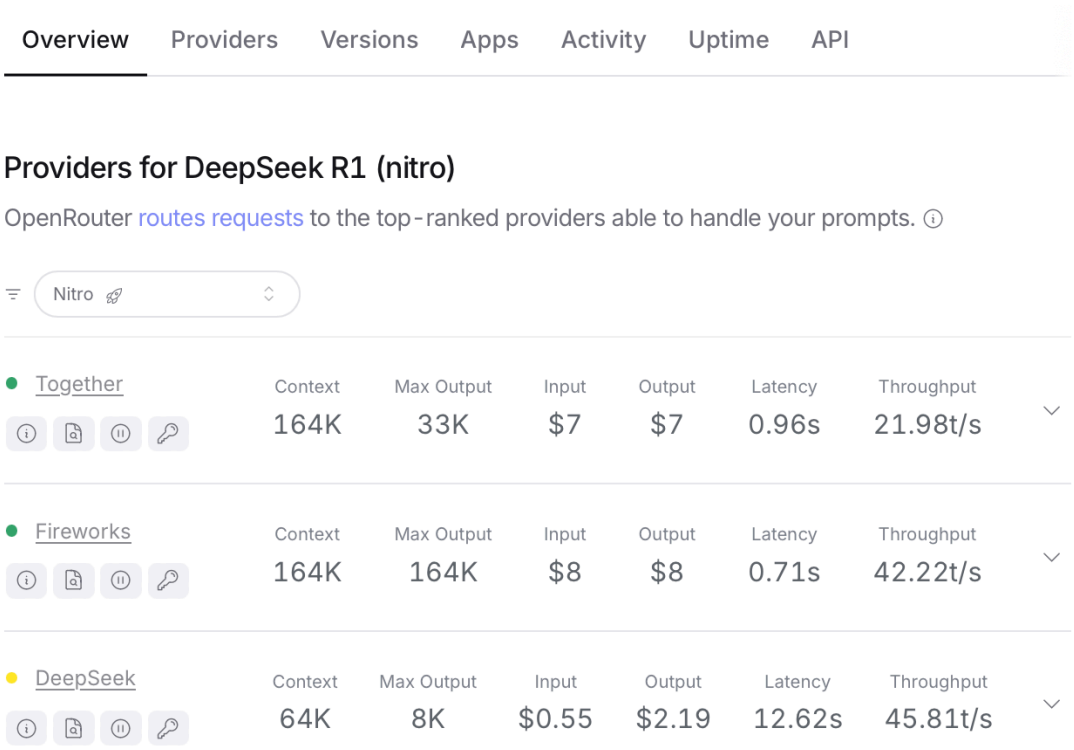


In order to take a deeper look at the LLM price dynamics, we can refer to a popular LLM marketplace called OpenRouter.

It conveniently lists multiple providers behind one API, creating a sort of open market for providing LLMs-as-a-service. And since DeepSeek r1 is an open-source model, multiple providers can actually serve the same model on their own prices, letting supply and demand balance things out.

3. https://en.wikipedia.org/wiki/DeepSeek

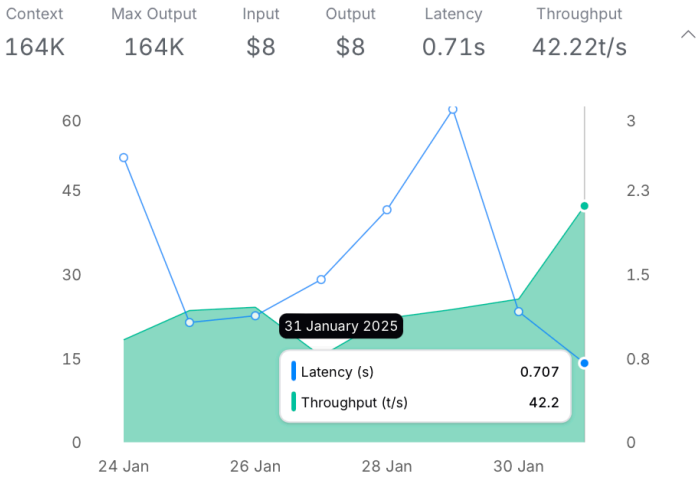
Here is how the prices look like for the top-rated providers of DeepSeek R1[4] at the moment of writing this article ("nitro" stands for "provider that can handle some work load"):

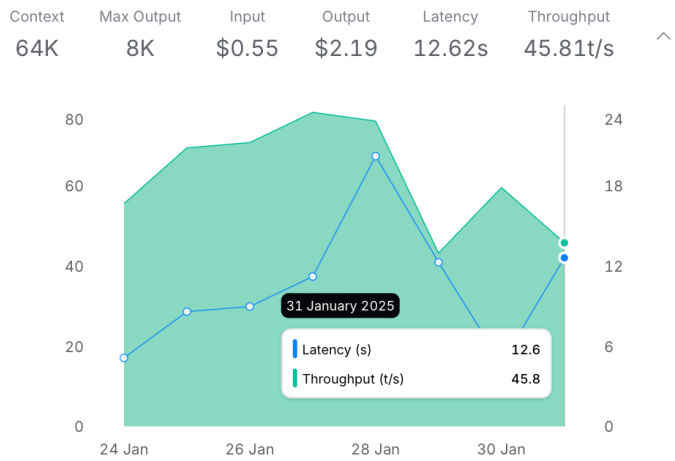


As you can see, DeepSeek attempts to provide API at the claimed prices, but with a few nuances:

- As a cost reduction measure **it limits its inputs and outputs at a fraction of what the other companies do** (compare "Context" and "Max Output", but keep in mind the original pricing of DeepSeek - includes 32K Reasoning tokens limit on top of 8K Output Limit.).
- Normally OpenRouter routes requests to the cheapest provider (market dynamics in play), but DeepSeek r1 API can't keep up with the current demand, so it has been explicitly de-ranked with a message: "Users have reported degraded quality. Temporarily deranked."
- Alternative competitors that are motivated to turn some profit - charge noticeably higher prices per input and output tokens. They can keep up with the demand and serve ever increasing throughput despite high costs.

4. https://openrouter.ai/deepseek/deepseek-r1:nitro





So effectively current market price for getting stable access to DeepSeek r1 is around $7-$8 per 1M Input/Output Tokens. That would come at $77 per our average 10:1 workload, which is twice more expensive than similarly capable GPT 4o $35.

However, these are just the guesses based on the actual market price. They don't necessarily tell us about the real cost of running DeepSeek r1, if we were to run it ourselves. So let's take a look at the latest report of Nvidia - running DeepSeek r1 on the latest NVIDIA HGX H200 at 3,872 tokens per second[5]. They are using native FP8 inference to achieve this speed.

5. https://blogs.nvidia.com/blog/deepseek-r1-nim-microservice/?ncid=so-infl-633755

Assuming the long-term 2y rental cost of HGX H200 of $16 per hour around Silicon Valley, running the software stack optimised by Nvidia at ideal capacity we would get a cost of $1.15 per 1M Input/Output tokens. This translates to $12.65 cost of 10:1 business workload, which is higher than $7.69 price for the same workload offered by DeepSeek r1.

Don't forget that DeepSeek doesn't even supposed to have access to the latest and most efficient Nvidia hardware and GPU interconnect. They are limited to H800 GPUs, which are an inferior export version of H100 with a limited memory bandwidth. This could bring the actual costs even higher.

Either way, no matter how we look at the numbers, we see the same picture:

- **We don't see how DeepSeek r1 can be 25x times cheaper than o1**, unless the price is heavily subsidised. Subsidised prices and market demand don't mix well.
- On our early v2 LLM benchmark **DeepSeek r1 shares reasoning capability similar to an older OpenAI GPT 4o from the August 2024**. It is **not comparable with o1**, yet.

Of course, both o1 and 4o are also multi-modal models, capable of working natively with images and complex documents, while r1 accepts only text inputs. This puts both types of the models even further in different leagues, especially when applied to document-oriented business workloads.

This makes stock market reaction around a promising Chinese text-only model that is comparable to an older-generation OpenAI model and is sold below price - a bit overreacting. Especially, given the fact that the future models are gradually steering towards multi-modal foundational models with world understanding. This is way beyond mere text capabilities and will provide plenty of added value for Nvidia and its partners to generate in the near future.

So here is our **promised stock market prediction** (don't treat it as a financial advise) - Nvidia will rebound and continue growing back pretty fast, backed up the actual workloads and cost dynamics in the real world.

Meanwhile the DeepSeek r1 model itself is a very interesting model that might allow OpenAI competitors to start catching up (there haven't been major interesting releases from Anthropic or Sonnet recently). R1 itself might fade away pretty soon due to the cost dynamics, that its distillations might take over it. They are already showing up quite high in our new Benchmark v2.

-