Necessity is the Mother of Invention
Deepseek R2 what to expect?

Warren Niles
Computer Scientist
June 23, 2025
warren@warrenniles.com

Deepseek R1: Review

- **Review:** Deepseek R1 had several important innovations that led to an overall ~95% cost reduction vs. OpenAI o1
  - Mixture of Experts (MoE) ≈ Ask questions to specific experts in their fields not ALL experts to lower costs
    - Lowers active parameters per output token and overall cost.
  - Multi Headed Latent Attention (MHLA) ≈ Summarize your findings during research (attention)
    - Lowers memory cost of Key Value Cache (KV Cache) in Transformers.
  - Group Relative Policy Optimization (GRPO) ≈ Compare a set of past answers vs. a set of new answers of the model. If users tend to prefer new answers, tune the model to use the new answers. By comparing the model against past versions of itself, it saves on compute costs.
    - Lowers Reinforcement Learning (RL) costs during training.
  - FP8 & FP32 mixed architecture ≈ Use a low energy cost FP8 when you care less about precision and use your high energy FP32 when you need more precision.
    - Lowers memory footprint + all computational costs

OpenAI (o1) vs Deepseek (R1) Cost Analysis (Public API's) on release of R1

Model	Context Window	1M Token (Cache Hit)	1M Token (Cache Miss)	1M Token (Output)	Max output tokens
o1	200k	$7.50	$15.00	$60.00	100k
o3-mini	200k	$0.55	$1.10	$4.40	100k
Deepseek V3	64k	$0.014	$0.14	$0.28	8k
Deepseek R1	64k	$0.14	$0.55	$2.19	32k
DeepseekR1 vs o1	-68%	-98%	-96%	-96%	8k

- o1 is most comparable to R1 as both are publicly available, have a similar structure, and little is known about o3. To compute cost savings you want similar structures.

- **Note: Consumers care more about correctness than reasoning vs. no reasoning. Reasoning models just seem to do better at producing correctness.**

Deepseek R1 Innovations Spreading to Larger Players

- April 5th, 2025: META announced their new Llama 4 models which included many optimizations inspired by Deepseek.
  - **Dense → Mixture of Experts**
    - **"**Our new Llama 4 models are our first models that use a mixture of experts (MoE) architecture." – Llama 4 herd April 5th, 2025
    - Deepseek 671B → 37B Active 1 shared expert, 8/128 Routed.
    - Llama 4 Maverick 400B → 17B Active 1 shared expert, 1/128 Routed
  - **Llama 3 BF16 → FP8**
    - **Moved from BF16 in training → FP8 in training**
    - "we focus on efficient model training by using FP8 precision, without sacrificing quality" – Llama 4 herd April 5th, 2025
  - META does not need GRPO as DPO is sufficient due to its various platforms providing real time feedback.
  - META also distilled ALL smaller models.
- June 10th, 2025: OpenAI announced an 80% cost reduction for their o3 model. Little is known as to how they achieved the reduction.

R2 Model Architecture Assumptions

- **R1 was a combination of the papers that preceded it so I am assuming R2 will be similar. This is my best guess given these conditions but in practice my estimates could vary wildly depending on the business decisions made by Deepseek.**
- **R1 had 671B total parameters but only ~37B were "active" per token** (only ~5.5% of weights used per forward pass). This was accomplished via **Deepseek-MoE** layers that route each token through a small subset of expert sub-networks.
- R2 likely follows this design: even if total parameters scale toward the trillions, the **active parameters per token remain on the order of tens of billions**.
- **Assumed Model Specs:**
  - **~1T parameters in total** (to incorporate Janus-Pro scaling)
  - ~**40B active parameters per token**. (MoE)
  - **~128k context window** (Deepseek Coder uses 128k), but thanks to NSA and MLA the **effective attention overhead** is comparable to handling perhaps a few thousand tokens in a standard dense transformer.
- We estimate on the order of 40B Parameters * (1 multiply + 1 add) ≈ **80B FLOPs per input token**.
- We add a small overhead for attention, but NSA+MLA limit it. Without compression a dense 1T-param transformer might need >10T FLOPs per token, but **R2's architecture makes it ~95% cheaper in practice.** (Notably, if R2 does not use MoE/NSA specifically, the FLOPs per token would be orders of magnitude higher).

Deepseek R2: What to Expect?

- On February 25th, 2025 it was rumored that Deepsek would be launching R2 sometime in early 2025.
- The new R2 model boasts the following features:
  - Multimodal Functionality
    - 01/29/2025 – Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling.
  - Enhanced Programming and Coding Abilities
    - 02/11/2025 – CodeI/O: Condensing Reasoning Patterns via Code Input-Output Prediction.
  - The model has been designed from the ground up to be more efficient with computational resources—a critical advantage in the resource-intensive field of large language model development.
    - 02/16/2025 – Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention
  - Advanced Multilingual Reasoning
    - 04/03/2025 – Inference-Time Scaling for Generalist Reward Modeling
  - Novel Training Techniques of Generative Reward Modeling (GRM) and Self-Principled Critique Tuning
    - 04/03/2025 – Inference-Time Scaling for Generalist Reward Modeling
- My Estimate: 1.2T vs GPT-4.1 ≈ 1.8T parameters and a **94% cost reduction vs. GPT-4o**
  - **most comparable model given what we know about its structure.**

Jan 29th 2025: Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling.

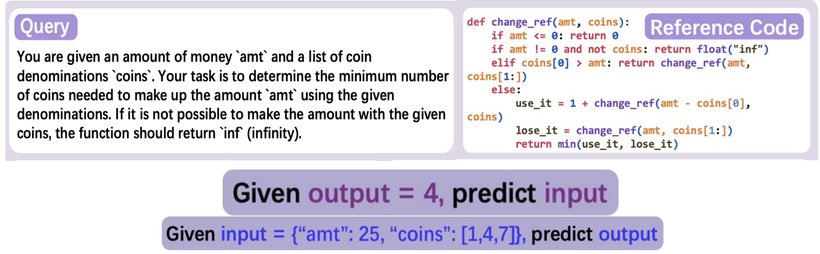
- Reduces inference + training costs by ~95% for images by using rectified flow over diffusion but maintains quality by adding "adapters" for the LLM to handle images.
- Rectified flow reduces inference and training time by trading complex curved transformations for Ordinary Differential Equation (ODE) paths learned via simple regression.
  - Rectified Flow (Janus Pro) vs Diffusion (DALLE-3) ≈ Golf
    - Rectified Flow ≈ Make 1 put and get it close enough to the hole gravity does the rest.
    - Diffusion ≈ Take 1000 small puts to get to the exact center of the hole.
  - Compresses a 30 – 50 step Diffusion Process to 1 step potentially reducing costs of inference and training by ~95%

- Uses **V**ariational **A**uto **E**ncoder (VAE) to do image generation in a compressed space that is only expanded at the last step reducing memory overhead.
  - Similar to KV Cache innovation in R1
- Decoupled image generation from image understanding to boost performance while only adding slight cost.
  - SigLIP-Large-Patch/16 + Linear = Adapter from Image to Text
  - VAE + ConvNext = Adapter from Text to Image

- **Key Takeaway:** Image inference + training costs **decrease by 95% vs. Diffusion.**

Feb 11th, 2025: CodeI/O: Condensing Reasoning Patterns via Code Input-Output Prediction.

- Enhances programming and reasoning capabilities by requiring models given a function along with a text query to predict either 1) the functions outputs given inputs or 2) feasible inputs given outputs only using Chain of Thought (CoT).
  - Improves reasoning skills without throwing more data at the problem.
  - "The cumulative sum of human knowledge has been exhausted in AI training" - Elon Musk



- **Key Takeaway:** Improves reasoning performance **WITHOUT requiring more human generated data**.

Feb 16th 2025: Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

- **N**atively Trainable **S**parse **A**ttention (NSA) removes the bottleneck of Dense Attention as context window increases by compressing continuous tokens into blocks and then only selecting blocks and adjacent tokens the system thinks would add value.
  - Summarize the N/s_c chunks of our text of size B_c in text of size N into m summaries.
  - Of our paragraphs of size B_s, only pick out the k relevant paragraphs to the question and read all the sentences of those paragraphs.
  - Of the individual sentences, read w surrounding sentences to have needed context.
- Dense $\approx O(n^2 d)$ vs. NSA $\approx O(n(m + kB_s + w)d)$,
  - n = Total number of tokens
  - d = # of dimensions (information per token)
  - m = # tokens per compressed blocks B_c w/ stride $s_c \approx \lfloor \frac{n}{s_c} \rfloor$
  - kB_s = # of tokens per k selected blocks of size $B_s \approx 2B_c$
  - w = window size ≈ constant → approx linear in n
- $m, k, w \ll n$ so cost reduction is $\frac{\text{NSA}}{Dense} \approx \frac{(m+kB_s+w)}{n} \approx \lim_{n\to\infty} \frac{\lfloor \frac{n}{s_c} \rfloor + kB_s + w}{n} \approx \left(1 - \frac{1}{s_c}\right) * 100\%$
  - *ex. n* = 65536, s_c = 16, B_c = 32, k=16, B_s = 64, w=512
  - (65536/16 + 16*64 + 512)/65536 ≈ 0.086→ **<u>~ 91% cost reduction estimate which matches observed values</u>**
- **Key Takeaway : Cost savings → 94% vs. Dense Attention or** theoretically $\left(1 - \frac{1}{s_c}\right) * 100\%$ as context window increases.

NSA Paper Tests

Metric	Paper Measurement
Decoding	-91% (11.6x)
Inference	-89% (9.0x)
Training	-83% (6.0x)

Apr 3rd 2025: Inference-Time Scaling for Generalist Reward Modeling

- Introduces Generalist Reward Modeling (GRM) and Self Principled Critique Tuning (SPCT) allowing AI to optimize reward modeling as well as providing feedback to the model to allow for better scaling of reward at inference time and scale future models.
- GRM ≈ movie critic
  - Pointwise Critic – Given a question and one or more candidate answers, generate a short **textual review** (ex. Clear answer and short!") and then score the review using a helper model/function for each answer (9/10).
  - The critic **scales with compute**: asking it to draft multiple critiques and then voting or averaging those critiques gives more reliable scores.
- SPCT ≈ critic school
  - Given a set of rules in natural language on how to write a **textual review**, **generate a review of the GRM review** based on the rules **AND** human preferences. Overtime the critic learns to write reviews that humans and the checklist **BOTH** agree on – and to do so more reliably if you let them write multiple drafts (sampling at inference).
  - Reusing prior work to train new models achieves top-tier accuracy with a mid-sized models.
    - Note: very similar to DGAN where you have a generator and a discriminator.
- **Key Takeaway:** Rather than increasing parameter count and therefore cost by **increasing sampling effort it achieves equal or better performance.**
  - **80 – 90% compute savings vs. standard reward models** in training and inference and enables **GRM for older models to be reused for newer models**.

R1 Release: OpenAI (o1) vs Deepseek (R1) Cost Analysis (Public API's)

Model	Context Window	1M Token (Cache Hit)	1M Token (Cache Miss)	1M Token (Output)	Max output tokens
o1	200k	$7.50	$15.00	$60.00	100k
o3-mini	200k	$0.55	$1.10	$4.40	100k
DeepseekV3	64k	$0.014	$0.14	$0.28	8k
Deepseek R1	64k	$0.14	$0.55	$2.19	32k
DeepseekR1 vs o1	-68%	-98%	-96%	-96%	8k

- o1 is most comparable to R1 as both are publicly available, have a similar structure, and little is known about o3.
- Context Window ≈ # of tokens before the model "forgets" old tokens and cannot directly remember them. Files, text, and model responses all are part of the context window.

May 8th, 2025: OpenAI (GPT 4.1) vs Deepseek (R2) Cost Analysis (Estimates)

Model	Context Window	1M Token (Cache Hit)	1M Token (Cache Miss)	1M Token (Output)	Max output tokens
GPT 4.1	1M	$0.50	$2.00	$8.00	32,768
GPT 4.1 mini	1M	$0.10	$0.40	$1.60	32,768
GPT 4o	1M	$1.25	$2.50	$10.00	32,768
Deepseek R2	128k*	$0.09*	$0.18*	$0.72**	???
Deepseek R1 vs GPT 4o	???	-93%	-93%	-93%	8k

* Assumptions: A100 delivering ~ 312T FLOPs at $2.50/hr → estimates (~80B FLOPs/input, 40B FLOPs/cached, 320B FLOPs/output) for 40B ~ 50B parameter experts ~ 1T total, and unknown context window as NSA scales w/ context window size.
** Output cost will depend on internal caching and architecture design choices made by Deepseek within their model and the billing heuristic they use. Usually based on estimated context window usage of users and output size of responses. Should be multiple between 1.5x – 4.0x of Cache Miss.

- **Key Takeaway:** R2 could have a variable size context window but the **cost per token could vary wildly** and be up to a **30 - 90% cost reduction against GPT 4o**.
  - If they increase parameter count, they will increase performance but increase cost.
  - **It depends on what performance Deepseek would want to achieve at what cost, size of experts etc. and there is little information as to internals of 4o.**

Deepseek Improvements: The Positives and the Negatives

The R1 & R2 improvements will drop the cost for all AI models

POSITIVES

- **Jevon's Law: Cost reductions could increase inference demand.**
- **Distilled Models: Distilled models need bigger models to be trained by.**
- **Gen AI inference demand is inflecting. Google in May 2025 processing 50x as many tokens at 480 trillion a month across products & APIs.**
- **Microsoft processing 100 trillion tokens in Q1:25 up 5x y/y with 50 trillion in March alone.**

NEGATIVES

- **Commoditized Hardware: Improvements allow new models to run on old hardware.**
- **Distilled Models Trend: Distilled Models are smaller and use less compute on inference.**
- **Hyperscalers (Amazon, Microsoft Google) all saw rev estimates CUT for CQ3:24 after reporting CQ2:24 and for CQ1:25 after reporting CQ4:24.**

IMPLICATIONS FOR SEMICONDUCTOR INDUSTRY

Deepseek has theoretically built a more efficient model that needs $\underline{1/10^{th} \sim 1/20^{th}}$ of the compute of GPT 4o.

The question is if they give:
1) a more performant model at the same price.
or
2) an equally performant model at a much cheaper price.

I believe:
1) Costs could vary wildly depending on what business decisions are made.

2) It will drop costs across the AI industry.

3) It could lead to a drop off in GPU demand short term and an increase over the long run.

About Warren A. Niles

Warren A. Niles
Founder
Advertising Technology Startup

Warren Niles is currently the founder of an advertising technology startup. His expertise includes quantitative finance, artificial intelligence, distributed systems, and cryptography. He has developed the following:

- AI for web traffic fingerprinting.
- AI for sentiment analysis.
- Distributed database system for Consistency and Partitioning (CP) scaling at high and tuneable availability.
- Autonomous financial data scraper and quantitative analysis engine that scrapes and consumes real time financial data to find statistically significant short term trading patterns.

Warren received a Bachelor of Science in Computer Science with a concentration in Artificial Intelligence from University of Chicago and a minor in Statistics.

Disclosures

Any securities, companies, sectors or markets mentioned are for informational purposes only and should not be construed to reflect any investment advice.

Information is obtained from sources deemed reliable, but there is no representation or warranty as to its accuracy, completeness or reliability. All information & opinions are current as of the date of this material and are subject to change without notice.

Readers should not assume that any investments in securities, companies, sectors or markets identified and described were or will be profitable. Investing entails risks, returns may be volatile, and investors can lose all or a substantial portion of their investment.

Any securities, companies, sectors or markets mentioned may not be representative of the holdings of Warren Niles, current or future, or their success in the future. Any securities, companies, sectors or markets mentioned are subject to change and should not be considered to be investment advice.

Any securities, companies, sectors or markets mentioned in this document may not be eligible for sale in some states or countries, nor suitable for all types of investors. Specific securities identified and described do not represent all of the securities purchased, sold or recommended for advisory clients. It should not be assumed that recommendations made in the future will be profitable or will equal the performance of the securities mentioned herein.

This material is presented solely for informational purposes and nothing herein constitutes investment, legal, accounting or tax advice, or a recommendation to buy, sell or hold any security, company, sector or market. No recommendation or advice is being given as to whether any investment or strategy is suitable for a particular investor. Readers should not assume that any investments in securities, companies, sectors or markets identified and described were or will be profitable. This material has been prepared by Warren Niles on the basis of publicly available information, internally developed data and other third-party sources believed to be reliable. Warren Niles has not sought to independently verify information taken from public and third-party sources and does not make any representation or warranty as to the accuracy, completeness or reliability of the information contained herein. All information is current as of the date of this material and is subject to change without notice. Any views or opinions expressed may not reflect those of Warren Niles as a whole. Certain products and services may not be available in all jurisdictions or to all client types. Investing entails risks, including possible loss of all or a substantial portion of principal.

The views expressed are those of Mr. Niles. These views are current as of the time of this presentation and are subject to change without notice. This material is not intended to be a formal research report or recommendation and should not be construed as an offer to sell or the solicitation of an offer to buy any security. Warren Niles may have long or short positions in some or all of the securities, companies, sectors or markets discussed. Before acting on any advice or recommendation in this material, you should consider whether it is suitable for your particular circumstances and, if necessary, seek professional advice. Mr. Niles does not accept any responsibility to update any opinions or other information contained in this document. Before acting on any advice, opinions or recommendation in this material, you should consider whether it is suitable for your particular circumstances and, if necessary, seek professional advice.