The DeepSeek Effect: Notes on Reinforcement Learning and Emerging Reasoning Capabilities

Discussion ensuing after the release of DeepSeek, Jan 20th – Feb 2nd, 2025, Linkedin

It seems that [GRPO (Group Relative Policy Optimization)](https://arxiv.org/abs/2402.03300) is all we need to teach LLM to reason as well as ChatGPT o1 or better. GRPO is a variant of [PPO](https://arxiv.org/abs/1707.06347) which even Nike uses these days. No supervised fine tuning, no [RLHF](https://rlhfbook.com/) - [just PPO does apparently impressive job](https://www.linkedin.com/posts/charlesmartin14_thanks-to-deepseek-no-one-is-sleeping-activity-7288988183332077569-9G6O/).

Jim Fan wrote:

*We are living in a timeline where a non-US company is keeping the original mission of OpenAI alive - truly open, frontier research that empowers all. It makes no sense. The most entertaining outcome is the most likely.  
  
DeepSeek-R1 not only open-sources a barrage of models but also spills all the training secrets. They are perhaps the first OSS project that shows major, sustained growth of an RL flywheel.  
  
Impact can be done by "ASI achieved internally" or mythical names like "Project Strawberry".   
Impact can also be done by simply dumping the raw algorithms and matplotlib learning curves.  
  
I'm reading the paper [1]:*  
*> Purely driven by RL, no SFT at all ("cold start"). Reminiscent of AlphaZero - master Go, Shogi, and Chess from scratch, without imitating human grandmaster moves first. This is the most significant takeaway from the paper.  
> Use ground truth rewards computed by hardcoded rules. Avoid any learned reward models that RL can easily hack against.  
> Thinking time of the model steadily increases as training proceeds - this is not pre-programmed, but an emergent property!   
> Emergence of self-reflection and exploration behaviors.  
> GRPO instead of PPO: it removes the critic net from PPO and uses the average reward of multiple samples instead. Simple method to reduce memory use. Note that GRPO was also invented by DeepSeek in Feb 2024 ... what a cracked team.*

*[2] is a \*second\* paper dropped with tons of RL flywheel secrets and \*multimodal\* o1-style reasoning is not on my bingo card today. Kimi's (another startup) and DeepSeek's papers remarkably converged on similar findings:  
  
> No need for complex tree search like MCTS. Just linearize the thought trace and do good old autoregressive prediction;  
> No need for value functions that require another expensive copy of the model;  
> No need for dense reward modeling. Rely as much as possible on ground truth, end result.   
  
Differences:  
  
> DeepSeek does AlphaZero approach - purely bootstrap through RL w/o human input, i.e. "cold start". Kimi does AlphaGo-Master approach: light SFT to warm up through prompt-engineered CoT traces.  
> DeepSeek weights are MIT license (thought leadership!); Kimi does not have a model release yet.  
> Kimi shows strong multimodal performance (!) on benchmarks like MathVista, which requires visual understanding of geometry, IQ tests, etc.  
> Kimi paper has a LOT more details on the system design: RL infrastructure, hybrid cluster, code sandbox, parallelism strategies; and learning details: long context, CoT compression, curriculum, sampling strategy, test case generation, etc.*

Charles H. Martin wrote:

*Revenge of the quants. So what have we learned from DeepSeek ? One is that you can get very far with reasoning with just a simple Reinforcement Learning algo (GRPO) and a large training set of “verifiable" math & coding problems.   
  
See this link for a discussion of DeepSeekMath and where the introduced the y GRPO algorithm: "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models"  
paper:* [5] *"First, we harness the significant potential of publicly available web data through a meticulously engineered data selection pipeline.   
  
Second, we introduce Group Relative Policy Optimization (GRPO), a variant of Proximal Policy Optimization (PPO), that enhances mathematical reasoning abilities while concurrently optimizing the memory usage of PPO."  
  
Other Key Innovations:  
1. Low-rank KV cache compression (MLA architecture)  
2. MoE FFN architecture   
3. Multi-token prediction head  
(source:* [*here*](https://x.com/ExponentiallyBe/status/1882466767601070313)*)  
  
"A model that is* ***almost*** *as good as o1 for 30x cheaper!"  
(source:* [*here*](https://x.com/LinusEkenstam/status/1882379270787616961)*)  
  
"It's wild to me that they they did this with no finetuning prior to the RL stage. R1 learns to reason on its own, like AlphaZero. During training, they observed the model learning to use advanced reasoning techniques - an "AHA" moment."  
(source:* [*here*](https://x.com/maxwinga/status/1881372837296083440)*)  
  
You can try it online here:* [*https://chat.deepseek.com*](https://chat.deepseek.com/) *Wait, there's more! You can run the full DeepSeek R1 671B on just two M2 Ultra Macs (source:* [*here*](tps://x.com/awnihannun/status/1881412271236346233) *and* [*here*](https://x.com/ronaldmannak/status/1881515278091342260)*)  
  
And people are even running DeepSeek R1 (distilled to Qwen 1.5B) on their iPhone 16 (link* [*here*](https://x.com/awnihannun/status/1882105560201617903)*).*

Peter Gostev wrote*:*

*DeepSeek's r1 model release caused quite a stir in the AI community, yielding a lot of good and bad takes.  
  
First thing out of the way - it is not a PsyOp or fakery, it is a good model created by a talented team of researchers with a lot of small innovations, especially on efficiency side. The fact that it is released under an MIT license with a nice research paper is also a shift - it is by far the strongest (mostly) open source model for a while.  
  
The most interesting thing for me about what they write in their paper is how well reinforcement learning (RL) is working. The coolest thing is all of the emerging behaviours (i.e. not explicitly selected for by researchers) that these models start to exhibit after all the RL training:  
   
 - Reflection: The model spontaneously rechecks and reevaluates steps mid‐solution.  
 - Self‐Correction: It spots mistakes and fixes them on the fly (e.g. incorrect factorisation).  
 - Aha Moments: It learned to pause, say “Wait,” and re‐start with a fresh chain.  
 - Longer outputs: It learned that longer outputs can yield better results   
 - Language Mixing: It can blend multiple languages or code snippets, since it only cares about getting the right final answer, not readability for humans.  
  
All these behaviors happened without humans explicitly teaching these techniques. This only came from rewarding the correct answer.  
  
The readability is a very interesting point. The original R1-Zero model was trained on pure RL and its 'thinking' was not comprehensible for researchers. If a model performs extremely well but 'thinks' in a completely incomprehensible way for humans, would we be ok with this kind of model? Perhaps forcing it to only 'think' in a single language (e.g. English) would limit how well it can 'think'. If it learned a concept in Chinese or Hebrew, would it be forced to avoid it for our convenience?  
  
To get to DeepSeek R1, they trained the base v3 model on some initial examples of the 'correct' chain of thoughts, which taught the model to think in a single language (mostly) and format chains of thought correctly.   
  
And the last very cool finding was that you can train smaller models on the reasoning chains (incl. 1.5b model) and increase their performance dramatically.  
  
All and all, this is a very interesting model and more importantly paper. This tells us what likely OpenAI (and others) are already doing and how impressive RL can be.*

Xu Ning wrote:

*I think many folks are probably reading DeepSeek papers this weekend. My thoughts on DeepSeek:*

*First, I recommend reading both DeepSeek V3[1] and DeepSeek R1[2] papers. Both have a lot of ingenuity in addressing different challenges. I view DeepSeek V3 primarily engineering optimizations on top of V2 that significantly reduced the compute needed to train a large MoE model.*

*DeepSeek-R1 (Zero) is the "reasoning-reinforced" version of V3 and it's truly a breakthrough. Instead of creating (usually with human effort) reasoning/CoT SFT data, R1 leverages RL to make reasoning capability \*emerge\*. This is very similar to the AlphaZero moment. I believe R1 will have profound impact on the direction. While some may say it signals the end of scaling law, it doesn't. You still need a good base model, one with large capacity, since the paper indicated that you can't RL a small model to have good reasoning. However, the role of SFT is diminished. You can almost say, that the model itself already has the capability of reasoning via it's learning of human text, but it has not been elicited out without SFT (no longer needed now) or RL.*

Cameron Wolfe wrote:

*One of the most overlooked contributing factors to the success of DeepSeek-R1 is the Mixture-of-Experts (MoE) base model from which it is derived–DeepSeek-v3…*

*The recently-proposed DeepSeek LLMs–including DeepSeek-R1, DeepSeek-v3 and more–have made waves within LLM research for a variety of reasons:*

*- Their weights are shared publicly.*

*- They come with technical reports that share many details.*

*- Their performance is impressive—on par with many closed models.*

*- Their training costs are (relatively) reasonable.*

*The DeepSeek-v3 base model, upon which DeepSeek-R1 is based, is an MoE LLM that makes several unique design choices to maximize efficiency and performance.*

*(1) Multi-head latent attention (MLA). DeepSeek-v3 adopts MLA, which is an efficient attention variant. MLA aims to minimize memory consumed of the model’s KV cache via a low-rank, joint projection that allows us to represent all key and value vectors with a much smaller (latent) vector. Adopting MLA decreases the size of DeepSeek-v2’s KV cache by over 93% compared to a 67 billion parameter dense model.*

*(2) Fine-grained / Shared Experts. DeepSeek-v3 uses fine-grained experts–a larger number of smaller experts–in its MoE layers. A subset of these experts are shared, which encourages specialization among experts while minimizing redundant information between experts.*

*(3) No load balancing loss. A novel, auxiliary-loss-free load balancing strategy is used by DeepSeek-v3 that simply adds a per-expert bias term to the selection of Top-K experts. At each iteration, the bias term for each expert is either increased or decreased by a fixed factor γ based upon whether that expert was underloaded or overloaded, respectively.*

*(4) Multi-Token Prediction. DeepSeek-v3 uses an MTP training objective. This objective is an extension of the supervised, cross entropy-based next token prediction objective that is used almost universally for training LLMs. Instead of predicting the next token for each token within a sequence, MTP predicts D future tokens. These predictions are made sequentially by a set of additional modules that are added to the model’s architecture.*

*Putting it together. DeepSeek-v3 has 671 billion total parameters and 37B active parameters. The model is pretrained on a corpus of 14.8 trillion tokens and aligned via the following steps:*

*- A two-stage context extension procedure (to 32K and 128K).*

*- Further SFT + RLHF.*

*- Distillation from the DeepSeek-R1 reasoning model.*

*DeepSeek-v3 outperforms closed-source models and achieves similar performance to even the best closed LLMs. The model–trained using a novel FP8 mixed precision framework–is economical (~$5.6M to train the final model) and is a great base model for DeepSeek-R1.*

Louis Scott wrote:

*A few thoughts after my quick read of the DeepSeek-R1 paper.*

*From the paper [1], table 3: It is reinforcement learning that drove the gains. A clue to the training efficiency is the use of Group Relative Policy Optimization [5]. It speaks in bladerunner patois, in 2.3.2 of [1], the paper notes language mixing in chain of thought (CoT) reasoning. Finally, Yann LeCun's comment [11] yesterday on the big take away:*

*"Open source models are surpassing proprietary ones."*

A math equations and formulas on a white background

Description automatically generated

Niccolo Gentile wrote:

*There is, understandably, quite a lot of buzz ongoing around the latest DeepSeek-R1.*

*As a matter of fact, it consists in an improved version, via Reinforcement Learning (and Group Relative Policy Optimization in particular) of the already existing DeepSeek-V3-Base model.*

*Under an architectural point of view, DeepSeek-V3 consists in a Mixture-of-Expert model encompassing 685B total parameters, of which 671B main weights and 14B for Multi-Token Prediction: only 37B of them get activated for each token.*

*In a previous post, I already reviewed DeepSeek's MoE architecture, described in "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models", Dai et al. 2024, based on Fine-Grained Expert Segmentation and Shared Expert Isolation.*

*However, DeepSeek-V3's other main ingredient, and already used in DeepSeek-V2, is its Multi-Head Latent Attention (MLA) mechanism.*

*How does it work, and how is it different from the classic Vaswani et al. 2017's Self-Attention - (potential) Multi Head Attention (MHA)?*

*In the standard MHA mechanism, for each token t, let h\_t be a d sized vector, where d is its embedding dimension (input at the attention layer). Let also n\_h be the number of heads and d\_h each of them's dimension.*

*In MHA, you first need to produce three vectors, q\_t, k\_t, and v\_t, each of dimension (d\_h \* n\_h), by multiplying h\_t with three matrices W\_q, W\_k, and W\_v, each of dimension, in turn, (d\_h \* n\_h) x d.*

*Opposed to this approach, MLA's main innovation - as per equations 9, 10, and 11 - consists in first computing a low-dimensional latent representation of h\_t, called c\_t^KV and of dimension d\_c << (d\_h \* n\_h), by multiplying it with a low-dimensional projection matrix W^DKV of dimension d\_c x d (9).*

*Then, only on this c\_t^KV one computes k\_t and v\_t using two upward projecting matrices, W^UK and W^UV, aiming at projecting again the latent representation to a (d\_h \* n\_h) vector.*

*Hence, where are the savings? They can be found in a sharply reduced need for KV-Caching, measured as number of elements that need to be stored.*

*In standard MHA, given l layers, one needs to store (2 \* n\_h \* d\_h \* l) elements per token. Conversely, in MLA, only around (d\_c + d\_h^R) \* l elements need to be stored, where d\_h^R denotes the "per-head dimension of the decoupled queries and keys" to be able to compute Rotary Positional Embeddings" (all discussion on this in Section 2.1.3).*

*As I mentioned, the same MLA mechanism is used both in DeepSeek-{V2, V3}, where the latter is the basis of R1. I personally found the MLA explanation clearer in DeepSeek-V2, but you can also find it in DeepSeek-V3 model card.*

*Link to both model cards in the comments: in the picture, from DeepSeek-V2's one, the three main equations controlling KV-caching compression. In the comments, also a link to my review of the aforementioned DeepSeek MoE.*

A close-up of a text

Description automatically generated

# References

[1] [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, DeepSeek AI, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeek_R1-Incentivizing_Reasoning_Capability_in_LLMs_via_RL.pdf)

[2] [Kimi K1.5: Scaling Reinforcement Learning with LLMs, Technical Report of Kimi K1.5 , 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Scaling_RL_with_LLM_Kimi_k1.5.pdf)

[3] [DeepSeek-Prover-V1.5: Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search, H. Xin et al, DeepSeek, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeek-Prover-V1.5-Harnessing_Proof_Assistant_Feedback_for_Reinforcement_Learning_and_Monte-Carlo_Tree_Search.pdf)

[4] [DeepSeek Technical Report, 2024](ttps://github.com/dimitarpg13/large_language_models/blob/main/articles/DeepSeek-V3_Technical_Report.pdf)

[5] [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, Z. Shao et al, DeepSeek AI, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeekMath-Pushing_the_Limits_of_Mathematical_Reasoning_in_Open_Language_Models_Shao_2024.pdf)

[5] DeepSeek AI repo: <https://github.com/deepseek-ai/DeepSeek-R1>

[6] Moonshot AI / Kimi repo: <https://github.com/MoonshotAI>

[7] [Linkedin discussion 1, Jan 20th, Jim Fan](https://www.linkedin.com/posts/drjimfan_we-are-living-in-a-timeline-where-a-non-us-activity-7287125475280265217-dUT3?utm_source=share&utm_medium=member_desktop)

[8] [Linkedin discussion 2, Jan 20th, Jim Fan](https://www.linkedin.com/posts/drjimfan_that-a-second-paper-dropped-with-tons-of-activity-7287148978247290880-Gkd2?utm_source=share&utm_medium=member_desktop)

[9] [Linkedin discussion 3, Jan 23rd, Charles H Martin](https://www.linkedin.com/posts/charlesmartin14_revenge-of-the-quants-so-what-have-we-activity-7288268696253345792-zjQT)

[10] [Linkedin discussion 4, Jan 25th, Charles H Martin](https://www.linkedin.com/posts/charlesmartin14_thanks-to-deepseek-no-one-is-sleeping-activity-7288988183332077569-9G6O)

[11] [Linkedin discussion 5, Jan 25th, Yann Le Cunn](https://www.linkedin.com/posts/yann-lecun_to-people-who-see-the-performance-of-deepseek-activity-7288591087751884800-I3sN?utm_source=share&utm_medium=member_desktop)

[12] [Linkedin discussion 5, Jan 26th, Peter Gostev](https://www.linkedin.com/posts/peter-gostev_deepseeks-r1-model-release-caused-quite-activity-7289341329325604865-hlUS?utm_source=share&utm_medium=member_desktop)

[13] [Linkedin discussion 6, Jan 26th, Xu Ning](https://www.linkedin.com/posts/xning_deepseek-r1-incentivizing-reasoning-capability-activity-7289460243539890178-znll?utm_source=share&utm_medium=member_desktop)

[14] [Linkedin discussion 7, Jan 27th, Cameron Wolfe](https://www.linkedin.com/posts/cameron-r-wolfe-ph-d-04744a238_one-of-the-most-overlooked-contributing-factors-activity-7289651602108338176-7yTL/?utm_source=share&utm_medium=member_desktop)

[15] [Linkedin discussion 8, Jan 28th, Louis Scott](https://www.linkedin.com/posts/louisscott_a-few-thoughts-after-my-quick-read-of-the-activity-7289918932457566209-Z3Hj/?utm_source=share&utm_medium=member_desktop)

[16] [Linkedin discussion 9, Jan 29th, Niccolo Gentile](https://www.linkedin.com/posts/niccolo-gentile-phd-02208160_there-is-understandably-quite-a-lot-of-activity-7289951426699493376-Q1Ob?utm_source=share&utm_medium=member_desktop)

[12] Running DeepSeek R1 via ollama: <https://ollama.com/library/deepseek-r1:14b>

[13] Fully Open Reproduction of DeepSeek R1: <https://github.com/huggingface/open-r1>