Notes on Reinforcement Learning in LLMs by Moonshot AI and DeepSeek

Discussion led by Charles H. Martin, Jim Fan, Jan 20th – Jan 25th, 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 groundtruth, 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)*).*

# 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] DeepSeekAI 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)