DRAFT

DEEPSEEK-R1: INCENTIVIZING LLM REASONING WITH GPRO-BASED RE-INFORCEMENT LEARNING

NEURAI Lab, Silicon Valley

Jose L Sampedro Mazon
sampedromazon.j@northeastern.edu

Lab

About DeepSeek And Why You Should Care

- DeepSeek was founded by Liang Wenfeng as a spin-off AI research lab from his hedge fund in China, known for developing and trading on AI algorithms, in 2023.
- DeepSeek has created and open-sourced numerous models, starting with coding-focused DeepSeek-Coder, and DeepSeek-LLM for general language understanding.
- In 2024, DeepSeek made major releases with DeepSeek-MOE as the first in a line of mixture-of-experts architecture models, and DeepSeekV3 – the base model for R1.
- In 2025, DeepSeek shocked the world with the open-source release of its highly efficient and performant R1, matching prior SotA performance at a fraction of the cost and promising significant potential for improvement.
- At the same time, DeepSeek also released smaller versions of Qwen and Llama distilled from R1, coming out on top of their counterparts and even bigger models on most benchmarks. DeepSeek-R1-Zero, fine-tuned on RL only, was also released.

"DeepSeek Disrupts AI Market: R1 Model Threatens OpenAI's Supremacy"
--Bloomberg

"AI Arms Race Heats Up: DeepSeek R1 Matches OpenAI at Fraction of Cost"
--Scientific American

"DeepSeek's r1 is an impressive model, particularly around what they're able to deliver for the price"
--Sam Altman

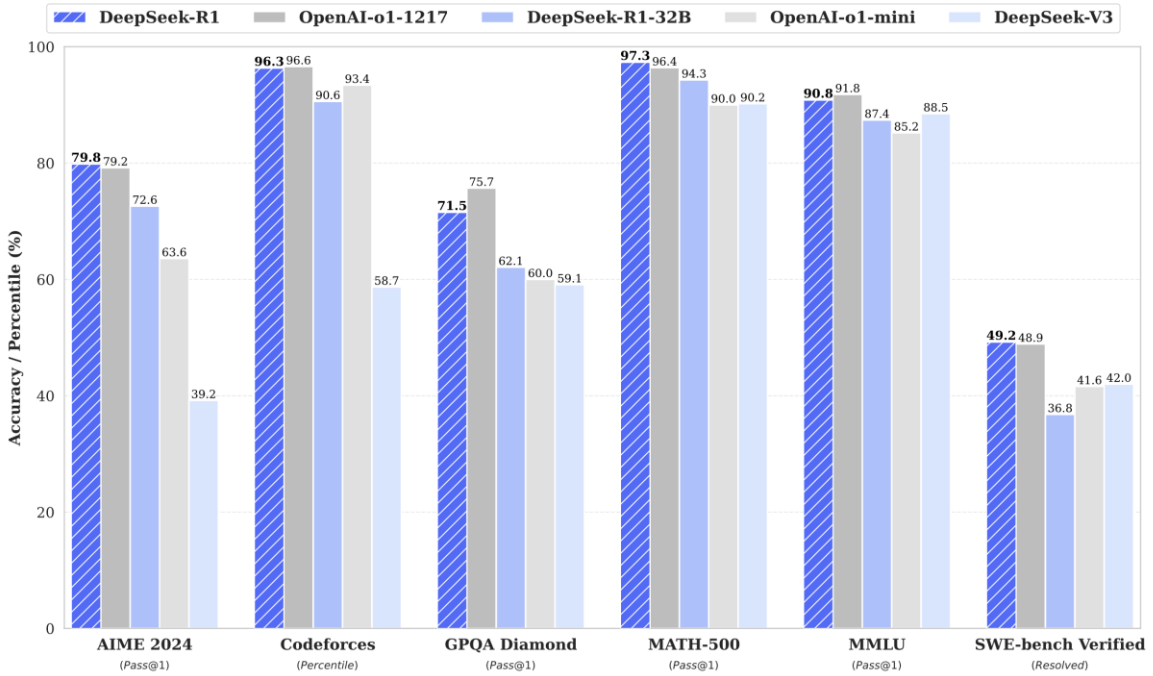
" DeepSeek R1 is AI's Sputnik moment (...), and as open source, [it is] a profound gift to the world
--Marc Andreesen

"AI Earthquake: DeepSeek R1 Wipes $1 Trillion Off US Tech Stocks"
--TechTarget

R1 hails two major breakthroughs using RL and Distillation

Large-Scale Reinforcement Learning vs SFT / RLHF

SotA reasoning performance can be attained with minimal reliance on human input with large-scale RL



Distillation vs RL/RLHF for Smaller Models

Distilling more powerful models into smaller ones achieves superior performance at a lower cost

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

R1 leverages already impressive DeepSeek's V3 as base model

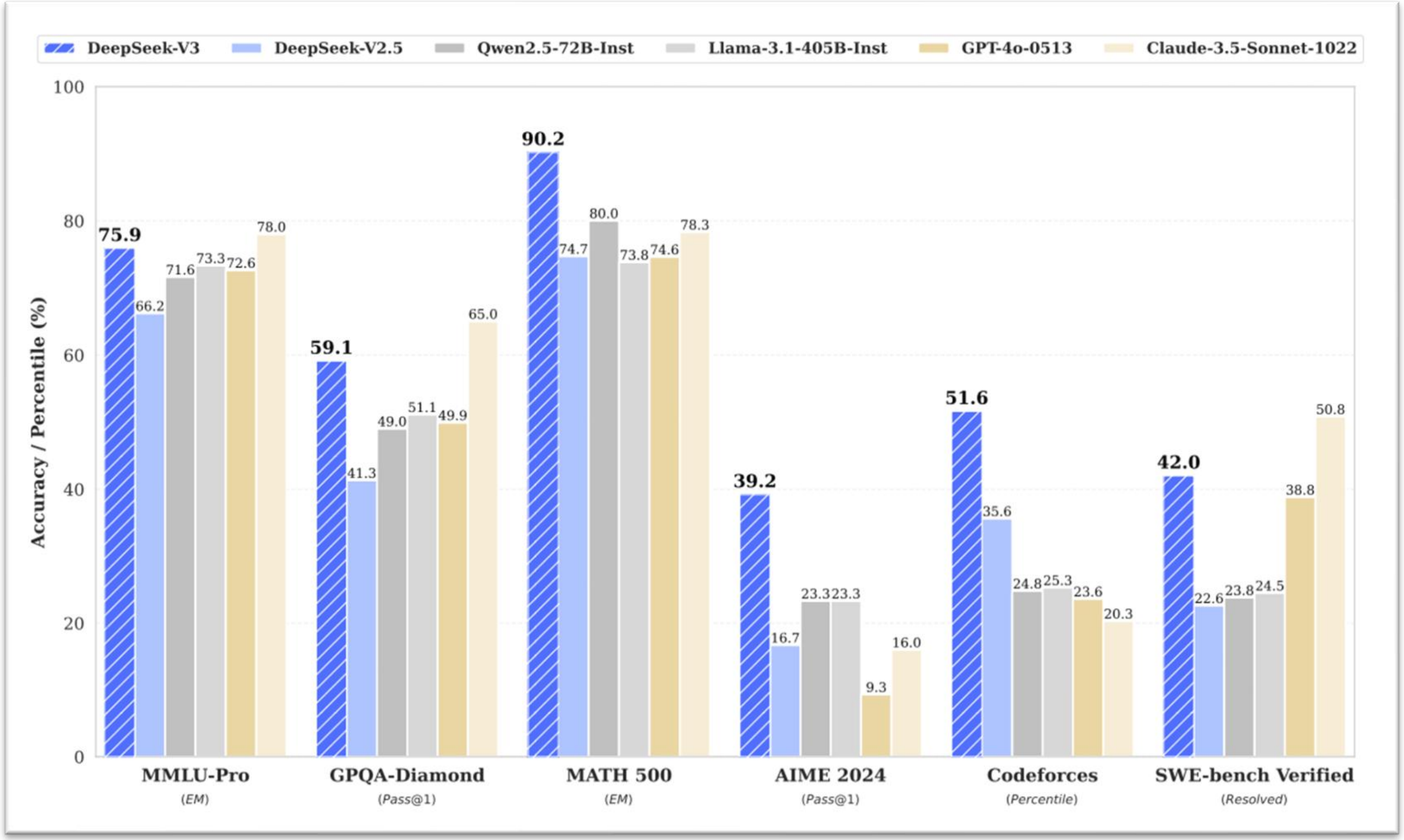
MOE Architecture
Minimizes neuron activation (~5.5%/token)

Auxiliary-loss-free Load Balancing
Reduces trade-off: expert activation vs perf

Multi-Head Latent Attention
K-V compression reduces VRAM needed

Mixed Precision (FP8/32)
Reduces memory usage by half vs FP-16/32

Memory Optimization in Training
Eliminates reliance on Tensor Parallelism



R1’s training pipeline matches SotA with less human input

Proves more performance to be extracted from RL (in addition to methods with human input)

SIMPLIFIED

DeepSeek R1	Cold Start SFT	Reasoning RL without HF	SFT with Model Output	General RL without HF
	Relatively ***small CoT data set by human annotators*** *(thousands vs millions)* for supervised fine-tuning to improve readability/alignment	***First round*** of large-scale RL ***without human feedback***, reasoning-focused using GPRO algorithm which forgoes critic model, ***saving costs***	New data set for SFT created through rejection sampling on generated content from RL-converged checkpoint model. DSV3 used as critic.	Second round of RL for alignment focused on helpfulness and harmlessness using reward signals and different prompt distributions

OpenAI GPT4	Single Large-Scale SFT	Single RLHF Phase
	Supervised fine-tuning on a larger, curated dataset of human-generated or annotated text to align the model with human preferences	Large round of RL with human feedback, even outsourcing expert-level samples for multiple domains from the likes of Scale AI

SIMPLIFIED

LVX VERITAS VIRTVS

DeepSeek uses a conservative cost function GPRO

Group Relative Reference Optimization algorithm prevents drastic updates from unsupervised RL

Takes an average over a number G of training sample pairs of prompts and outputs, each generated for the same question by the current and previous version of the model (NB: Advantage and drift penalty applied to each sample)

Probability of an output given a prompt, as a ratio of the latest model over the previous version (measures output probability change since last update)

Each output sampled is assigned a reward based on accuracy and format rules; Advantage term accounts for how much better or worse the reward for the output is compared to the batch average

KL Divergence 'drift' penalty based on probability deviation between old ref model and new policy model and importance (probability) of the given output – *another measure to prevent extreme model updates*

$$JGPRO(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi\theta(o_i|q)}{\pi\theta_{old}(o_i|q)}A^i, clip\left(\frac{\pi\theta(o_i|q)}{\pi\theta_{old}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A^i\right) - \beta D_{KL}(\pi\theta||\pi_{ref})\right)$$

Clip function caps probability ratio setting a lowest (1−ε) and highest (1+ε) value, resulting in a 'moderated' term.

Min function takes the lower of the clipped or unclipped probability ratio.

Designed to prevent drastic model outputs – clipped term taken when probability ratio is deemed too high

DeepSeek tweaked conventional D_{KL} function to penalize based on the importance (probability) of the output in the previous version of the model (instead of the latest), suggesting conservatism in favor of the base model (V3), known to be stable and trained 'conventionally'

LVX VERITAS VIRTVS

DeepSeek's Tweaks KLD Formula for Large-Scale RL

In unsupervised RL, full output distribution is not known (vs RLHF with fixed labelled dataset)

$$D_{KL}(\pi\theta || \pi_{ref}) =$$

CONVENTIONAL

Sums over all possible outputs to compute expected divergence across the **entire probability distribution**

$$\pi\theta(O|Q) * \log\left(\frac{\pi\theta(O|Q)}{\pi\theta_{ref}(O|Q)}\right)$$

Uses the probabilities of the **new model** as weights

Measures **probability divergence**, using log for smooth, symmetrical penalties

If probabilities match, log term and hence penalty is zero

DEEPSEEK

Applied to each sample - avoids using expectations over full distributions given in RL full distributions are difficult to compute

$$\frac{\pi\theta_{ref}(o_i|q)}{\pi\theta\ (o_i|q)} - \log\left(\frac{\pi\theta_{ref}(o_i|q)}{\pi\theta\ (o_i|q)}\right) - 1$$

Measures **raw divergence** and uses **previous model probabilities** as weights *(comparing relative probabilities directly)*

Smooths penalty to avoid extreme penalties for small probability mismatches

Ensures KLD is **no penalty when probabilities match**

flips the ratio so that the reference model's probability term acts as the weighting term

DISTILLATION

Fine-tuning smaller LLMs using outputs from more capable models to make performance comparable in a task without sacrificing accuracy or reliability.

Relying on the large-scale RL to improve reasoning capabilities of smaller models *'require enormous computational power and may not even achieve the performance of distillation'*.

Using the reasoning data generated by DeepSeek-R1, various widely used dense models fine-tuned: Qwen and Llama

DeepSeek demonstrates larger model reasoning patterns can be distilled into smaller models with better performance compared RL

DeepSeek-R1-Distill-Qwen-7B outperformed non-reasoning models like GPT-4o-0513

DeepSeek-R1-Distill-Qwen-14B surpasses QwQ-32B-Preview on all evaluation metrics

DeepSeek-R1-Distill-Qwen-32B and Distill-Llama-70B exceed o1-mini on most benchmarks

"While distillation strategies are both economical and effective, advancing beyond the boundaries of intelligence may still require more powerful base models and larger- scale reinforcement learning"

Food for Thought: DeepSeek-R1-Zero and Emerging Behavior

Also open-sourced: DeepSeek-R1-Zero, developed to “explore the potential of LLMs to develop reasoning capabilities without any supervised data, focusing on their self-evolution through a pure RL process”

DeepSeek-R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs.

It even had a literal ‘aha moment’ as it learned to allocate more thinking time to a problem by reevaluating its initial approach during test-time compute.

“[DeepSeek-R1-Zero’s] ‘aha moment’ serves as a powerful reminder of the potential of RL to unlock new levels of intelligence in artificial systems (…) a captivating example of how reinforcement learning can lead to unexpected and sophisticated outcomes.”

NORTHEASTERN