# Competition or Cooperation

## Machine Learning Singapore #MSLG


- Martin Andrews = CompetitionOrCooperation @ mdda.net


29-October-2026

# About Me

- Machine Intelligence / Startups / Finance
  - Moved from NYC to Singapore in Sep-2013
- 2014 = 'fun' :
  - Machine Learning, Deep Learning, NLP
  - Robots, drones
- Since 2015 = 'serious' :: NLP + deep learning
  - Including Papers...
  - & GDE ML; ML-Singapore co-organiser...
  - & Red Dragon AI...



MLSG
MACHINE LEARNING
SINGAPORE

# Outline

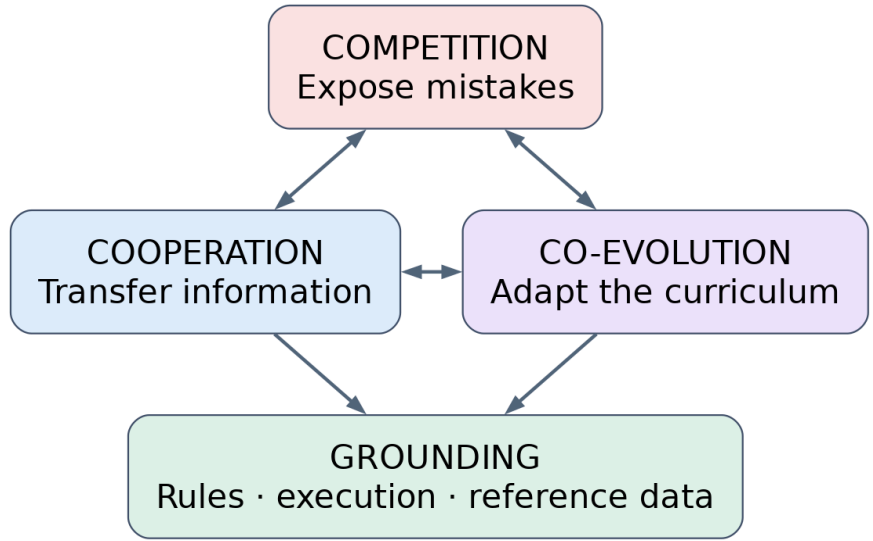
- What's the game?
- Some history (pre-LLM)
  - AlphaZero, GANs, DINO & PAIRED
- Modern times (post-LLM)
  - Competition : Self-Play & Critics
  - Cooperation : Teachers & Hindsight
  - Recursive Self-Improvement (RSI)
    - Auto-Scientists vs Co-evolving Worlds & Zero-Data
- In the news...
- Wrap-up & QR-code

MLSG
MACHINE LEARNING
SINGAPORE

# What’s the Game?

# Why Add a Second Model?

- Single-model RL is limited:
  - needs exact math/code verifiers
  - fixed tasks stop stretching the model
  - sparse `0/1` outcome rewards plateau
- Adding a learning *partner* adds signal:
  - Competition : expose mistakes & force strategy
  - Cooperation : transfer privileged explanations
  - RSI / Co-evolution : invent harder tasks & worlds
- "Two models" can mean:
  - separate nets *or* shared weights (eg: role prompt)
    - ... or an EMA momentum copy
  - key question : *what anchors the loop?*

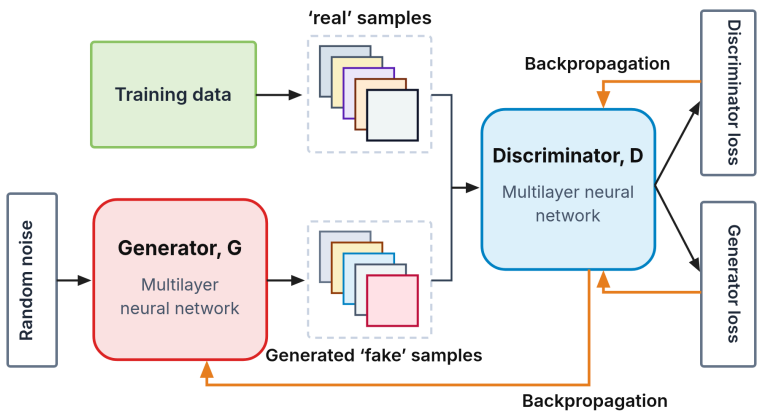



MLSG
MACHINE LEARNING
SINGAPORE

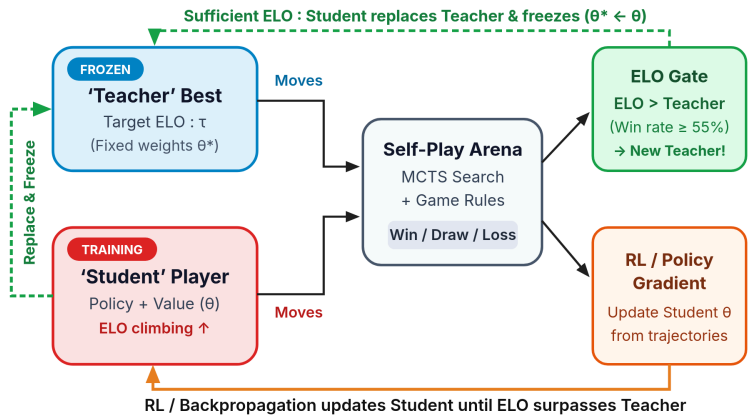
# Some History : Pre-LLM

# Competitive Blueprints : GANs & Self-Play

- GANs - Goodfellow *et al* (2014)
  - Generator tries to fool Discriminator
    - grounded by *real data* distribution
  - discrete tokens block direct backprop!
    - SeqGAN - Yu *et al* (2016) used RL policy gradients
  - modern LLM echo : RARO (reasoning critic)



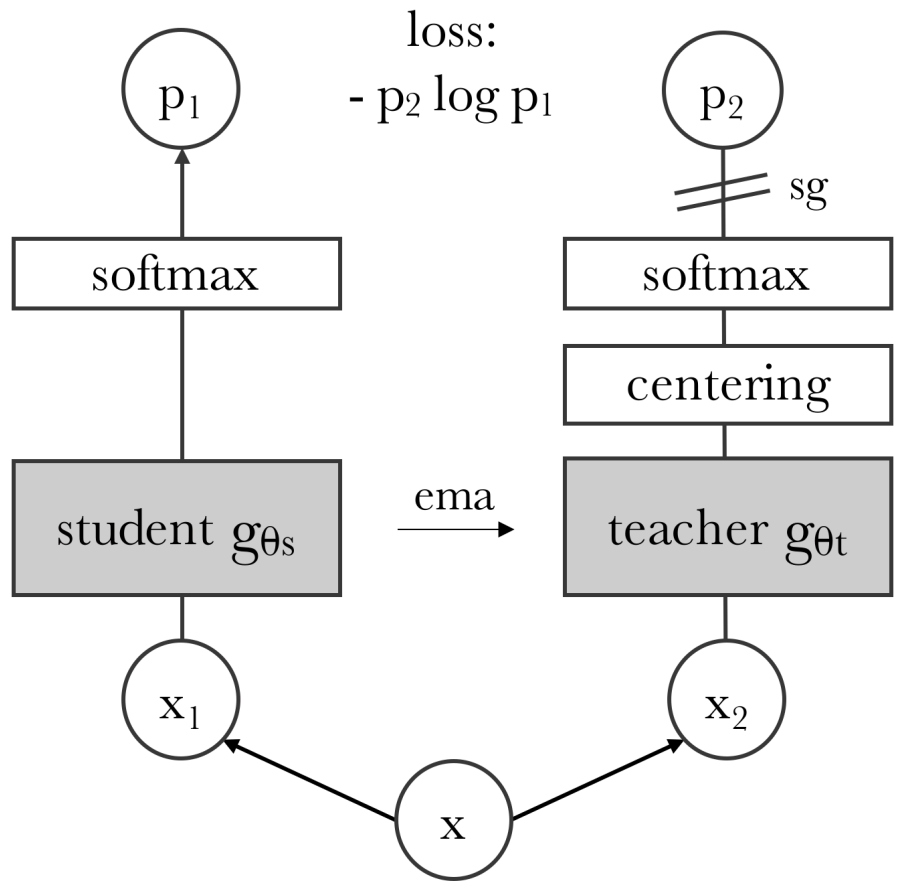

- AlphaZero - Silver *et al* (2017)
  - symmetric self-play + MCTS search
    - opponent is always at your current level
    - grounded by exact *game rules* (win/loss)
  - failure mode : strategy cycling
    - fix = league play vs past checkpoints
  - modern LLM echo : SPIRAL




MLSG
MACHINE LEARNING
SINGAPORE

# Cooperative Blueprint : DINO

- Emerging Properties in Self-Supervised Vision Transformers (DINO)
  - Caron *et al* (2021)
  - Trying to create good representations "p"
    - for each image 'x'
  - Student matches Teacher across image crops
    - Teacher sees *global* views; Student sees *local*
    - Teacher = EMA momentum copy of Student
  - Why doesn't it collapse to a constant?
    - `stop-gradient` on Teacher
    - Teacher output centering + sharpening
  - Core insight for LLMs:
    - *information asymmetry* creates a teaching signal
    - ... even from the *same* underlying model (SDPO)

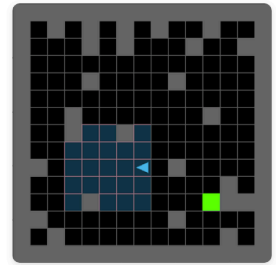



MLSG
MACHINE LEARNING
SINGAPORE

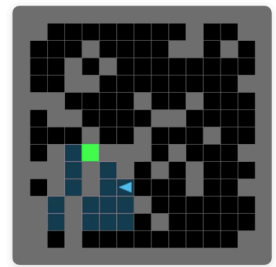
# Curriculum Blueprint : PAIRED

- Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design (PAIRED) - Dennis *et al* (2020)
  - How do you automatically generate *useful* training worlds?
    - naïve adversarial generator makes *impossible* worlds
    - random generator makes *trivial* worlds
  - Solution = reward the Environment Designer for Regret:
    - gap between advantaged *Antagonist* and *Protagonist* agents
    - $\text{Regret} = R(\text{Antagonist}) - R(\text{Protagonist})$
    - high regret $\rightarrow$ world is *solvable*, but *hard right now*
  - Modern LLM echo : SPADE (hint-based regret in code worlds)

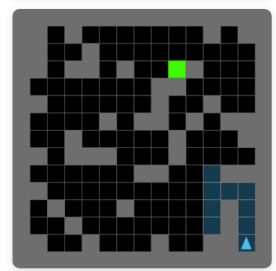
**(a) Random**
Domain Randomization
(unstructured clutter)
Too trivial / noisy



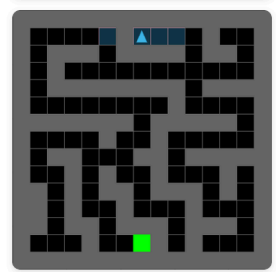
**(b) Minimax Adversarial**
Goal completely walled off
(no path exists)
Impossible (0 reward)



**(c) PAIRED**
Minimax Regret curriculum
(structured corridors)
Solvable & challenging



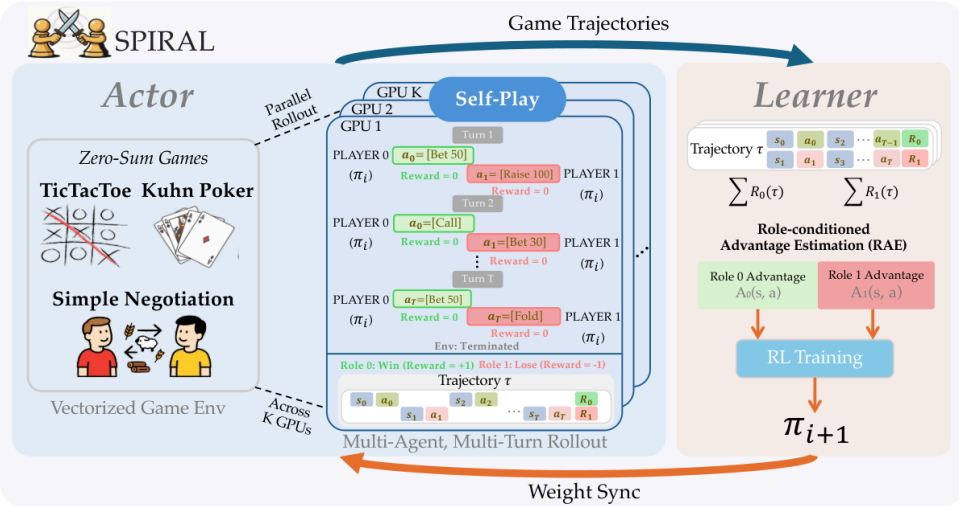
**(d) Actual Task**
Human-designed target
(zero-shot evaluation)
Transfer benchmark



MLSG
MACHINE LEARNING
SINGAPORE

# Competition : Self-Play & Critics

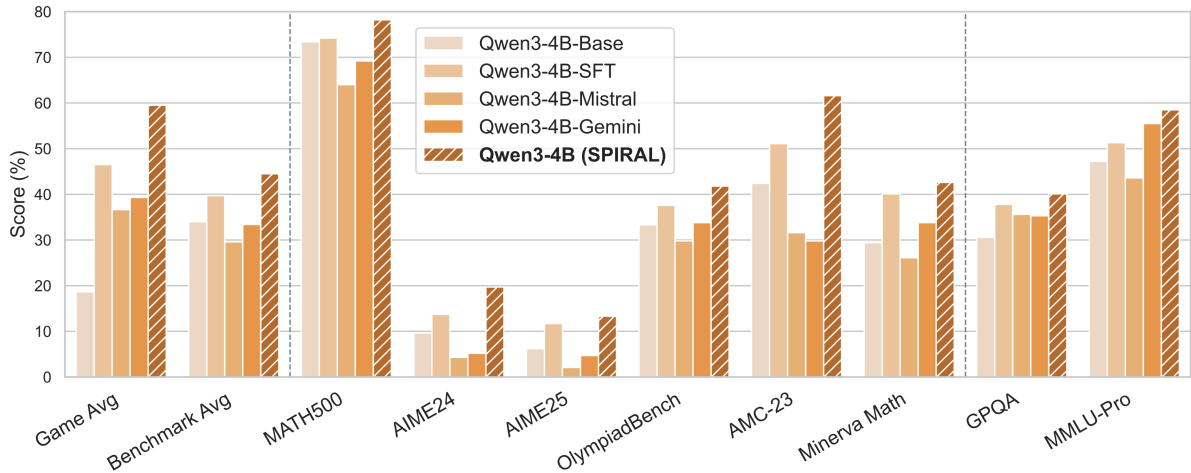
# SPIRAL : Multi-Turn Game Self-Play (for LLMs)




- SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn RL - Liu *et al* (ICLR 2026)
  - One shared policy plays *both sides* in zero-sum games
    - Kuhn Poker, Simple Negotiation, TicTacToe (no human problem sets)
  - Stabilized via **Role-conditioned Advantage Estimation (RAE)**
    - normalizes rewards per player role; anchor = deterministic game rules

MLSG
MACHINE LEARNING
SINGAPORE

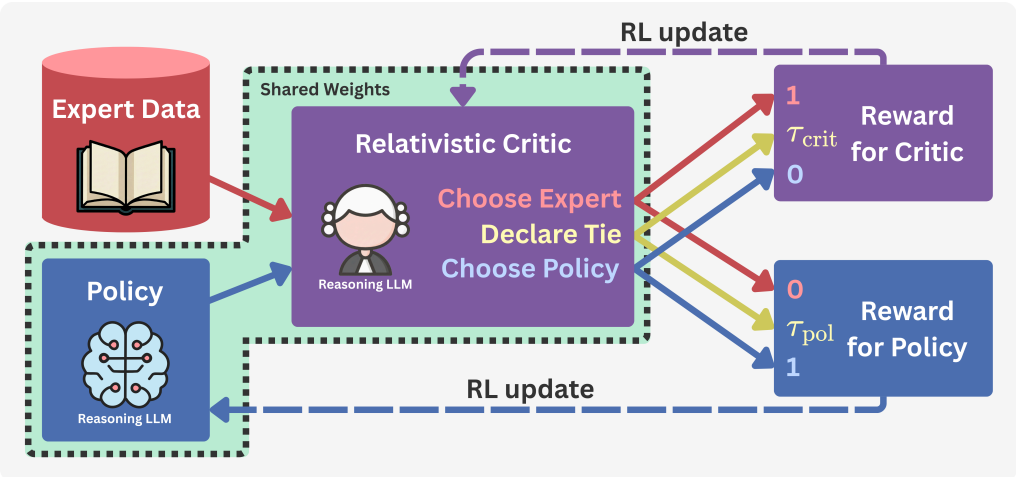
# SPIRAL : Zero-Shot Transfer to Math & Logic (WOW!)




- Why does playing poker improve math benchmarks?
  - opponent pressure forces *case analysis*, *deception detection* & *expected-value tracking* in CoT
  - `Qwen3-4B-Base` : `34.0% → 44.5%` (mean across 8 reasoning benchmarks)
  - multi-game self-play beats single-game training
    - caveat : starts from a pretrained base model (activates existing latent knowledge)

MLSG
MACHINE LEARNING
SINGAPORE

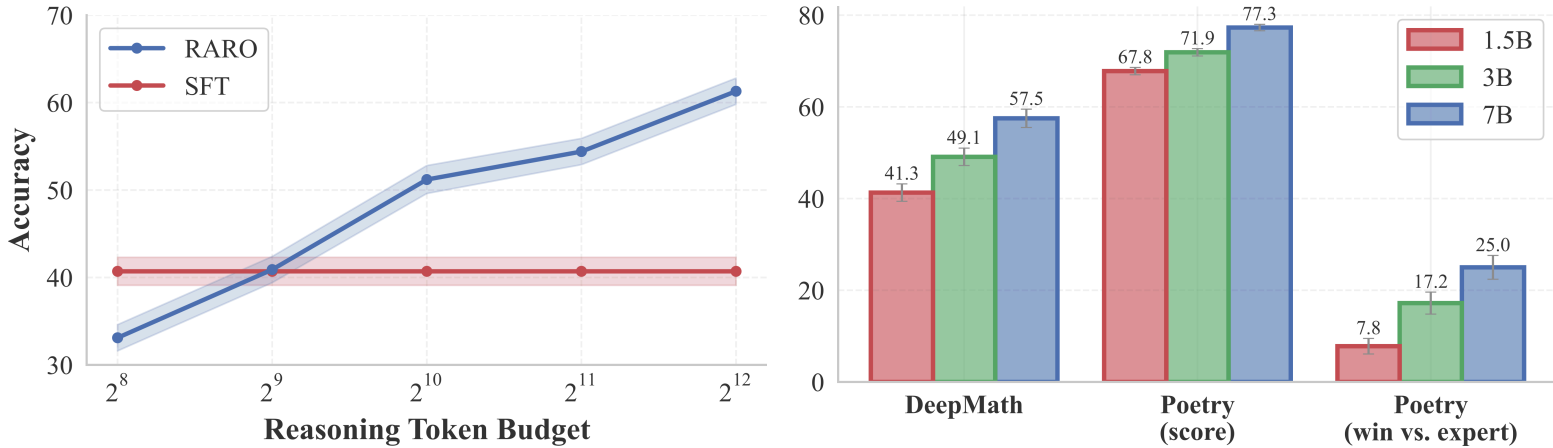
# RARO : RL when there is no Verifier (~ LLM GAN)




- Escaping the Verifier: Learning to Reason via Demonstrations (RARO) - Cai *et al* (2026)
  - What if there is *no* rule-based verifier (eg: creative writing, open-ended tasks)?
  - GAN-style Inverse RL : **Policy** vs **Relativistic Critic** (shared weights)
    - Critic sees `(expert, policy)` answer pair → reasons via CoT to spot the expert
    - Policy trained via RL to fool the Critic (with a `'tie'` option + replay buffer)

MLSG
MACHINE LEARNING
SINGAPORE

# RARO : Scaling Without a Rule Verifier




- Matches RLVR test-time & model-size scaling trends:
  - strong gains on **Countdown** & **DeepMath** *without* using ground-truth verifiers
  - unlocks test-time reasoning scaling on open-ended **Poetry**
  - caveat : `'verifier-free' ≠ 'supervision-free'`
    - anchored by high-quality expert demonstrations (risk = learning style over substance)

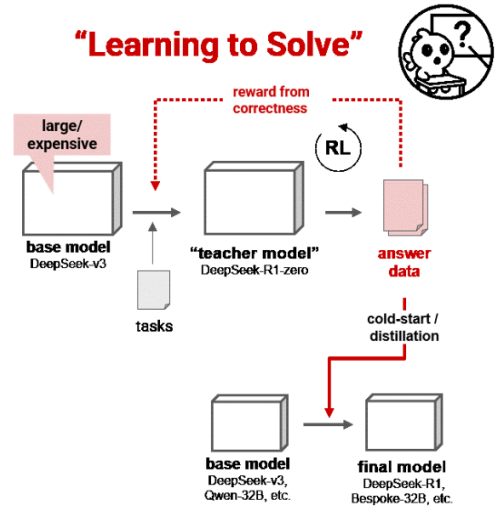
MLSG
MACHINE LEARNING
SINGAPORE

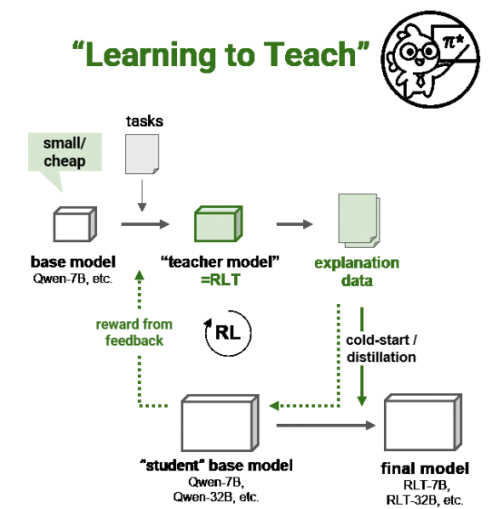
# Cooperation : Teachers & Hindsight

MLSG
MACHINE LEARNING
SINGAPORE

# Sakana RLT : Learning to Teach

- Reinforcement-Learned Teachers (RLT)
  - Cetin *et al* (2025)
  - Standard RLVR:
    - models must discover everything from scratch
  - RLT introduces **information asymmetry**:
    - Teacher given *both* `question + ground-truth answer`
    - Teacher writes step-by-step explanation
      - for a frozen Student
    - Teacher Reward =
      Student's log-likelihood of producing the right answer
  - Sakana AI Blog

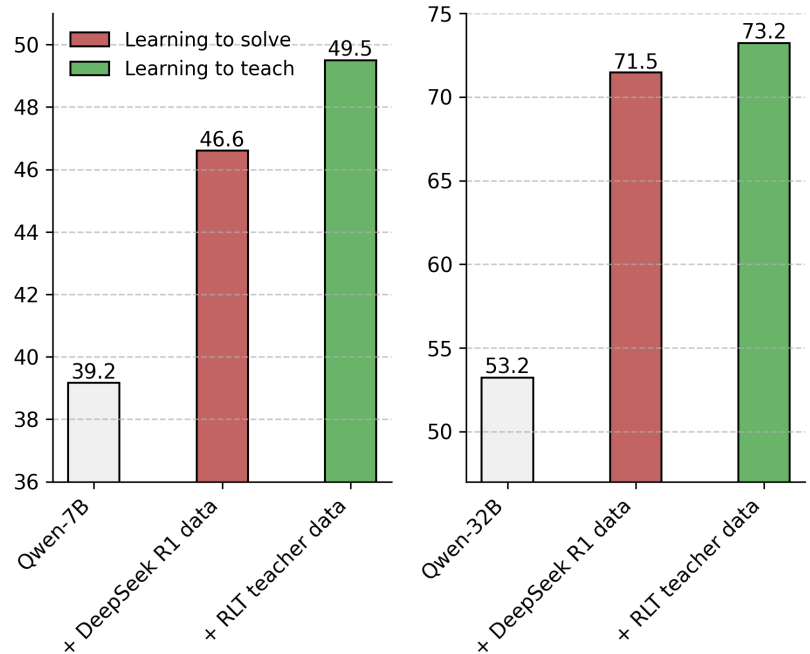



MLSG
MACHINE LEARNING
SINGAPORE

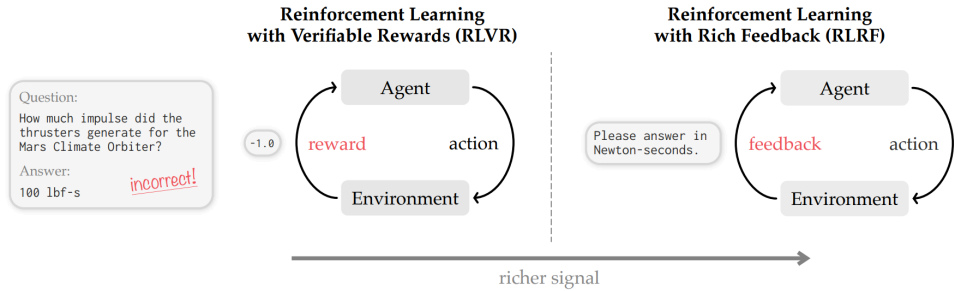
# Good Solver ≠ Good Teacher

- 7B RLT beats **DeepSeek-R1 (671B)** as a teacher:
  - `Qwen-7B` student : `49.5%` (RLT) vs `46.6%` (R1 teacher)
  - `Qwen-32B` student : `73.2%` (RLT) vs `71.5%` (R1 teacher)
- Why?
  - raw solver traces contain dead-ends & leaps
  - RLT explicitly optimizes for *student comprehension*
- Caveat : RLT optimizes 1-step answer likelihood
  - SOAR (ICML 2026) rewards *actual inner-loop Student progress* on hard problems

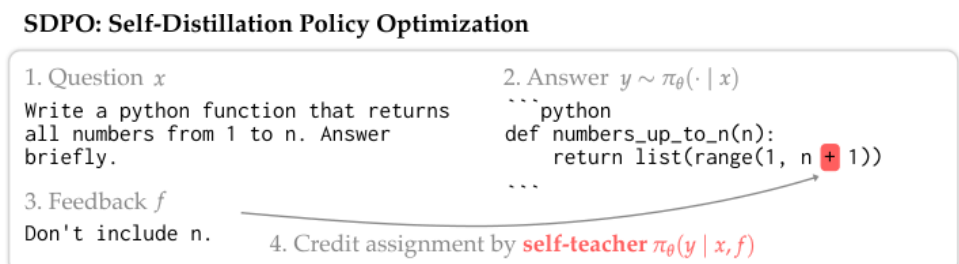



MLSG
MACHINE LEARNING
SINGAPORE

# SDPO : Self-Distillation Policy Optimization



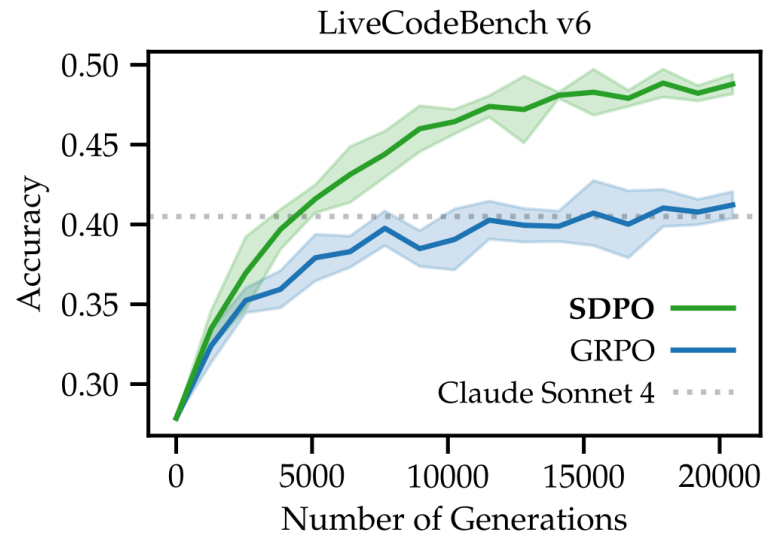

- Reinforcement Learning via Self-Distillation (SDPO) - Hübotter *et al* (2026)
  - ~DINO for LLM post-training (Student + EMA Self-Teacher):
    - Student fails a task ⟶ environment returns rich feedback (compiler error, traceback)
    - Self-Teacher sees `[prompt + rich feedback]` in hindsight & re-scores every token
    - Student (without feedback) distills those token-level probabilities




MLSG MACHINE LEARNING SINGAPORE

# SDPO : Dense Credit Assignment

- Fixes the scalar RLVR bottleneck:
  - GRPO gives `0` for a failed 2,000-token rollout
  - SDPO pinpoints *which exact tokens* caused the bug
- LiveCodeBench v6 ( `Qwen3-8B` ):
  - `48.8%` with SDPO vs `41.2%` with GRPO
  - beats **Claude Sonnet 4** & **DeepSeek-R1-Distill**
  - reaches GRPO's peak accuracy in **4× fewer generations**
- Anchor = environment execution feedback




MLSG
MACHINE LEARNING
SINGAPORE

# Recursive Self-Improvement (RSI)

MLSG
MACHINE LEARNING
SINGAPORE

# RSI : Auto-Scientists vs Co-Evolving Worlds

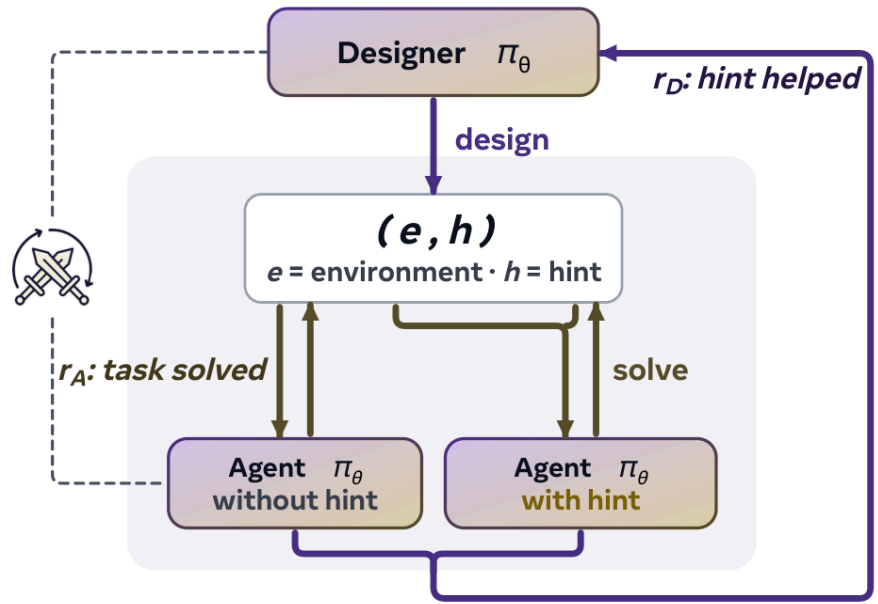
- RSI via **Auto-Scientist**
  (eg: The AI Scientist - Lu *et al* (2024)):
  - tends to produce *very incremental* improvements
    - because the vast bulk of papers it has been trained on are incremental...
  - generating ML paper text ≠ capability breakthroughs
- Other RSI directions :
  **Proposer–Solver Co-Evolution**
  - don't imitate the literature → *invent harder interactive challenges*
  - Absolute Zero (AZR) (2025) : code-execution tasks with zero external datasets




MLSG
MACHINE LEARNING
SINGAPORE

# SPADE : Self-Play Executable Worlds

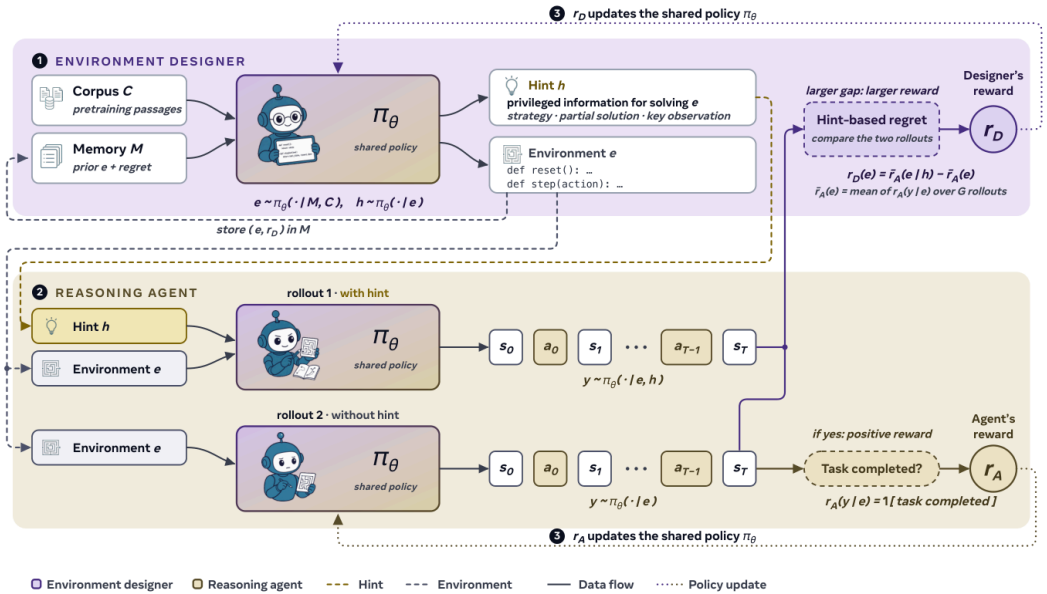
- SPADE: Self-Play in Adaptive Synthetic Executable Environments
  - Liu *et al* (Aug-2026)
    - Excellent Project Page & MIT-licensed Code
      - See "Adaptive Synthetic Environments Scaling"
    - One shared LLM plays two roles:
      - Environment Designer : writes stateful Python worlds ( `reset()` , `step()` , tools & rewards)
      - Reasoning Agent : interacts over multiple turns to solve them
    - Grounded by *real Python execution*
      - syntax or runtime bugs rejected immediately




MLSG MACHINE LEARNING SINGAPORE

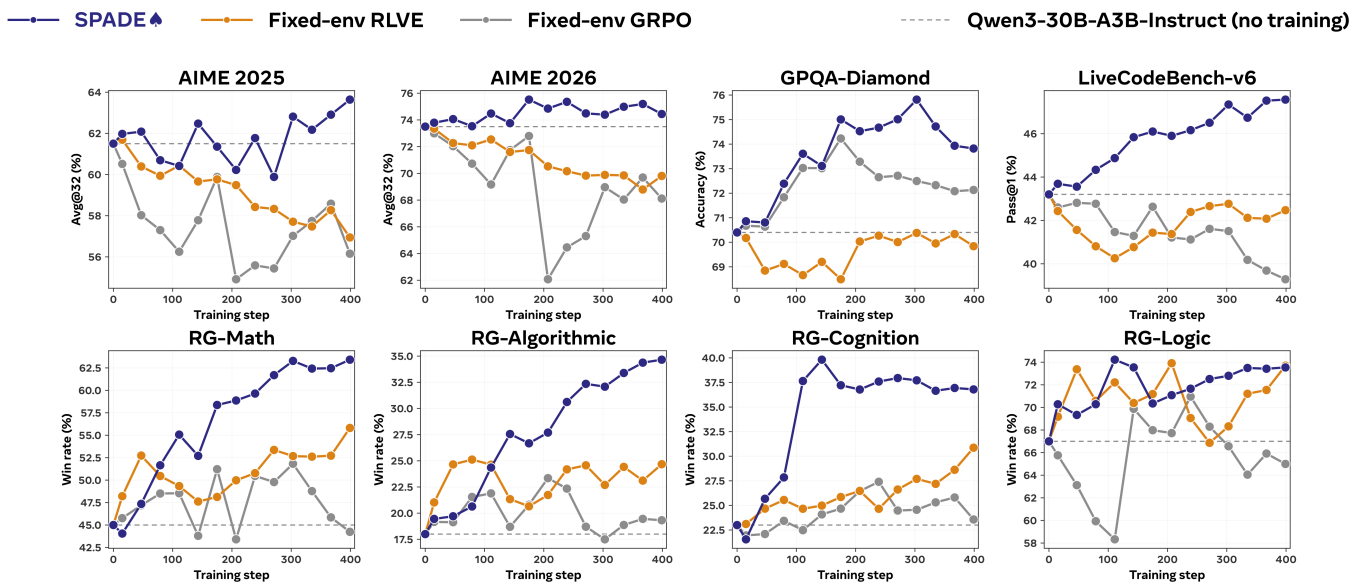
# SPADE : Hint-Based Regret Curriculum

- How to keep generated worlds *solvable* yet *challenging*?
  - Designer also generates a **solution hint**
  - Hint-based Regret : $r_D = R(\text{with hint}) - R(\text{without hint})$
    - high gap = solvable with the hint, too hard without it (right at the frontier!)
  - Corpus documents seed topic diversity : **+13.9 pts** on ACEBench-Agent ( `Qwen3-30B-A3B` )




MLSG
MACHINE LEARNING
SINGAPORE

# SPADE : Breaking the Fixed-Environment Plateau

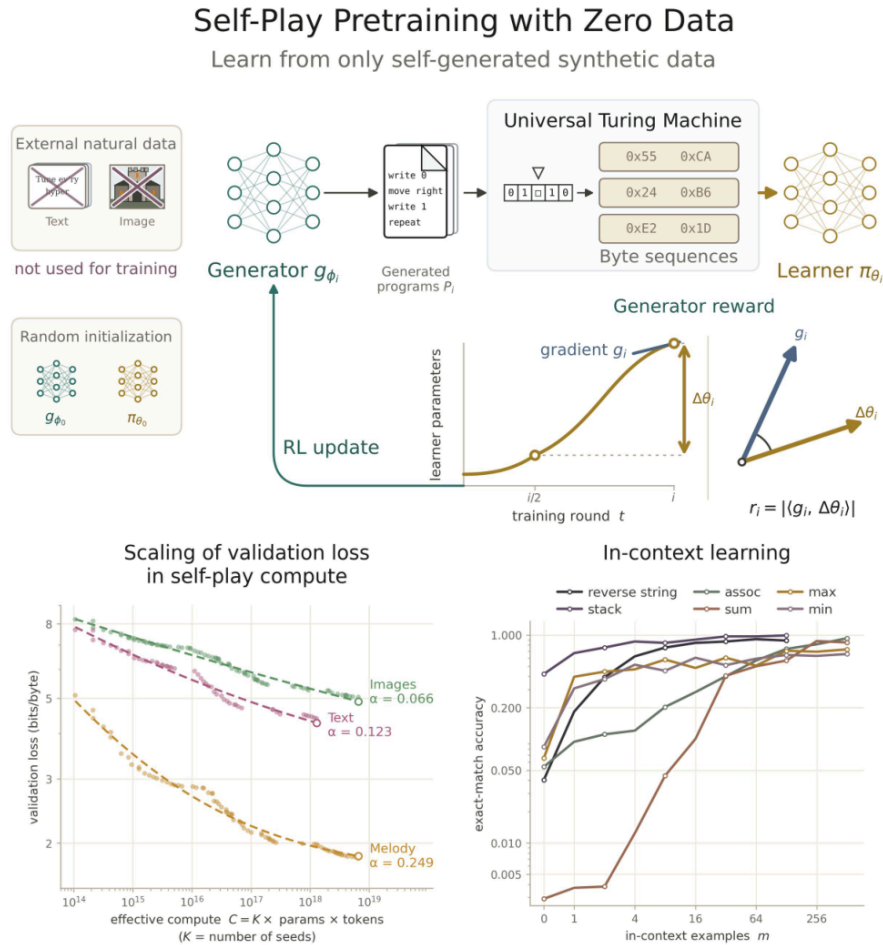



- Fixed environments plateau; co-evolving worlds (SPADE, dark blue) keep scaling:
  - baselines : `Fixed-env RLVE` (orange) & `Fixed-env GRPO` (grey) saturate or degrade
  - +5.3 pts mean gain across 8 reasoning benchmarks ( `Qwen3-30B-A3B` )
  - caveat : Python execution guarantees *runnable* code, not *real-world truth*

MLSG
MACHINE LEARNING
SINGAPORE

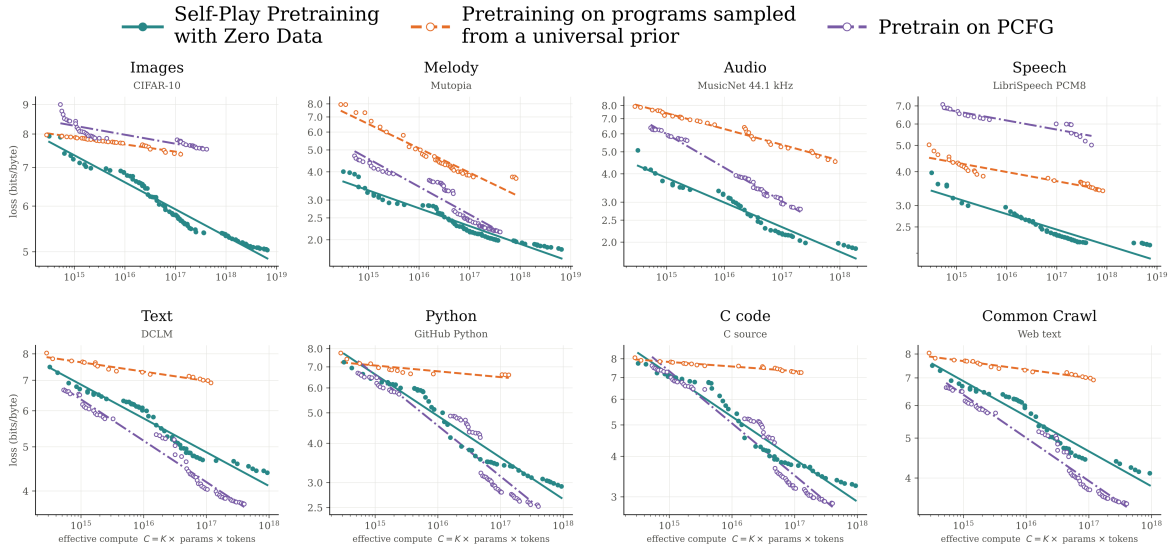
# Self-Play Pretraining with Zero Data

- Self-Play Pretraining with Zero Data - Cowsik *et al* (Sept-2026)
  - Every method so far started from a *pretrained* LLM
  - Can two transformers learn *from random weights* with **zero human data**?
    - **Generator** writes small Universal Turing Machine programs in `BrainF*ck` (via RL)
    - **Interpreter** executes them into byte streams
    - **Learner** predicts the next byte
  - Generator rewarded for **gradient alignment** with Learner's trajectory




MLSG
MACHINE LEARNING
SINGAPORE

# Zero-Data : Skills Learned work across Modalities

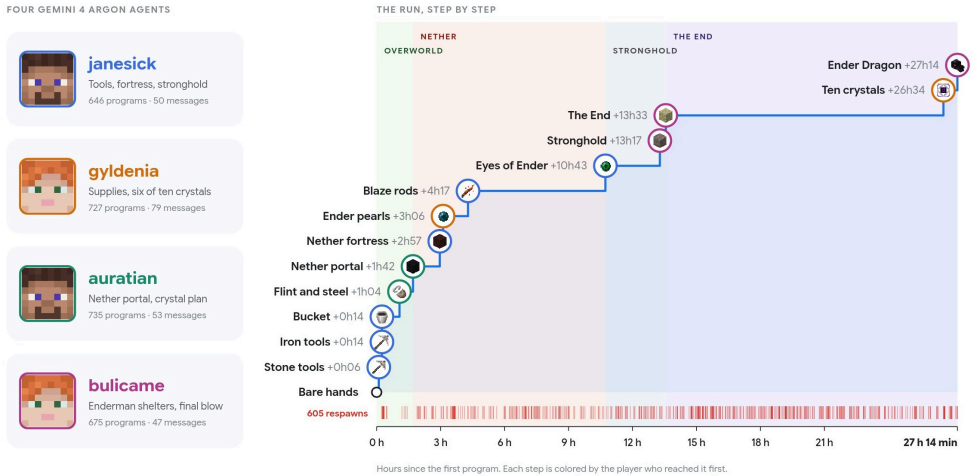



- Zero-shot validation loss on *real* data drops as a power law of self-play compute:
  - transfers across Text (DCLM), Code, Audio, Images & DNA
  - caveats :
    - learns universal *structural priors* (algorithmic patterns), not factual world knowledge
    - natural validation loss was still used to tune hyperparameters

MLSG
MACHINE LEARNING
SINGAPORE

# In the news...

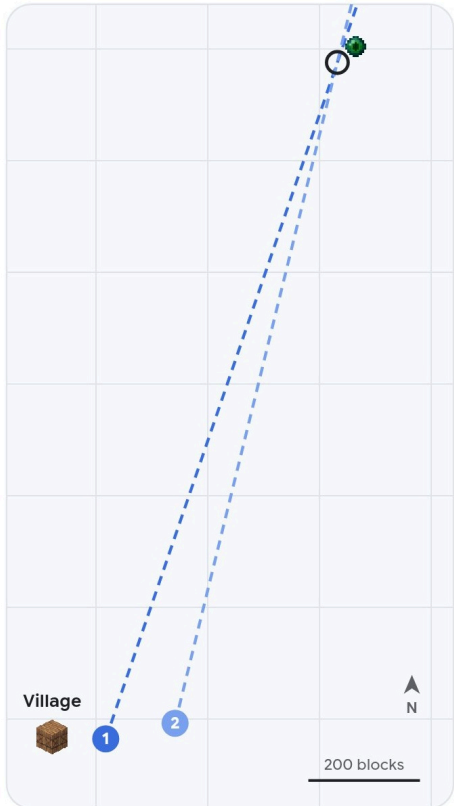
# Gemini 4 Argon : Agent Team Beats Minecraft




- A team of Gemini 4 Argon agents beats Minecraft - Faldor *et al* (Oct-2026)
  - 4 general-purpose coding agents in a random Survival world - First Author Thread
    - 27 hours & 2,783 self-written Mineflayer programs (605 respawns, 229 chat messages)
    - self-organized via chat & shared helper functions : tools, food, diamonds, Nether portal
      - did not have a lot of fun with the Ender Dragon...

MLSG
MACHINE LEARNING
SINGAPORE

# Gemini 4 Argon : Triangulating the Stronghold




## Triangulating the stronghold

A thrown Eye of Ender flies toward the nearest stronghold. janesick threw two from two places, recorded both directions, and computed where the lines cross.

- 1 2 Where janesick threw the two eyes
- Where the lines cross: (432, −1376)
- The actual portal room: (463, −30, −1406)

**1.3 km**
from the village to the crossing

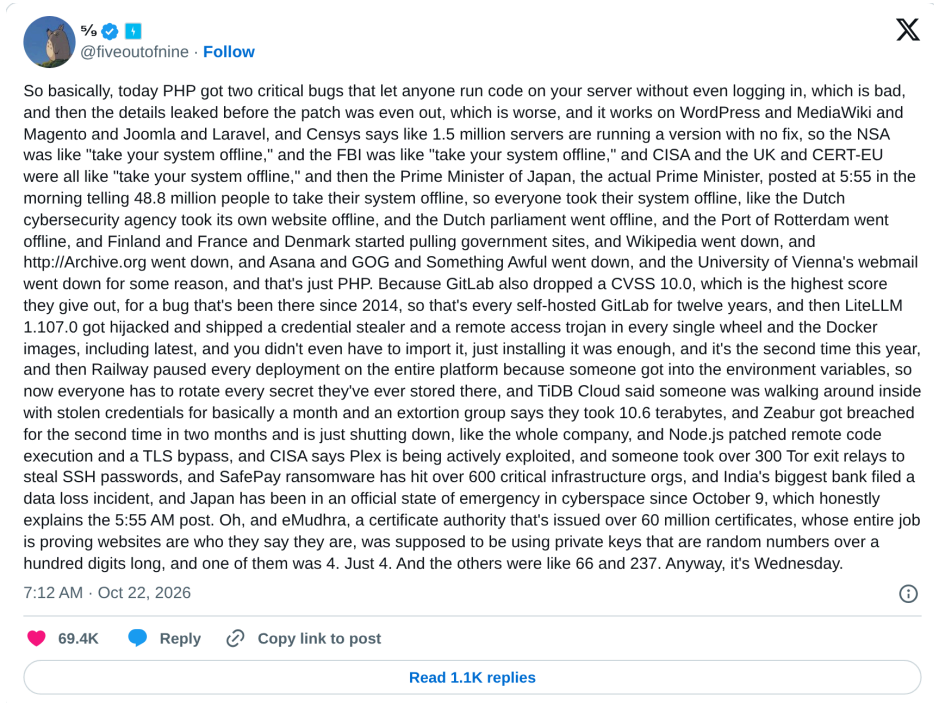
**43 m**
from the crossing to the portal room

**10 h 46 min**
into the run when janesick found the crossing

MLSG
MACHINE LEARNING
SINGAPORE

# Security Issues

- `@fiveoutofnine` : The State of InfoSec
  - PHP : two possible RCE 0-days leaked *before* patch
    - NSA, FBI, CISA, CERT-EU & Japan PM (at 5:55am) : *"take your system offline"*
  - LiteLLM `1.107.0` : hijacked on PyPI & Docker
    - credential stealer + RAT triggered *on install*
      - (2nd time this year!)
  - GitLab : CVSS `10.0` bug
    - (in every self-hosted install since 2014)
  - Railway & TiDB Cloud breached
  - Zeabur shutting down
  - eMudhra Certificate Authority (60M+ certs)
    - '100+ digit random' private keys:
    - one of them was `4` (others were `66` & `237`)
- "Anyway, it's Wednesday" ...



5/9 @fiveoutofnine · Follow

So basically, today PHP got two critical bugs that let anyone run code on your server without even logging in, which is bad, and then the details leaked before the patch was even out, which is worse, and it works on WordPress and MediaWiki and Magento and Joomla and Laravel, and Censys says like 1.5 million servers are running a version with no fix, so the NSA was like "take your system offline," and the FBI was like "take your system offline," and CISA and the UK and CERT-EU were all like "take your system offline," and then the Prime Minister of Japan, the actual Prime Minister, posted at 5:55 in the morning telling 48.8 million people to take their system offline, so everyone took their system offline, like the Dutch cybersecurity agency took its own website offline, and the Dutch parliament went offline, and the Port of Rotterdam went offline, and Finland and France and Denmark started pulling government sites, and Wikipedia went down, and http://Archive.org went down, and Asana and GOG and Something Awful went down, and the University of Vienna's webmail went down for some reason, and that's just PHP. Because GitLab also dropped a CVSS 10.0, which is the highest score they give out, for a bug that's been there since 2014, so that's every self-hosted GitLab for twelve years, and then LiteLLM 1.107.0 got hijacked and shipped a credential stealer and a remote access trojan in every single wheel and the Docker images, including latest, and you didn't even have to import it, just installing it was enough, and it's the second time this year, and then Railway paused every deployment on the entire platform because someone got into the environment variables, so now everyone has to rotate every secret they've ever stored there, and TiDB Cloud said someone was walking around inside with stolen credentials for basically a month and an extortion group says they took 10.6 terabytes, and Zeabur got breached for the second time in two months and is just shutting down, like the whole company, and Node.js patched remote code execution and a TLS bypass, and CISA says Plex is being actively exploited, and someone took over 300 Tor exit relays to steal SSH passwords, and SafePay ransomware has hit over 600 critical infrastructure orgs, and India's biggest bank filed a data loss incident, and Japan has been in an official state of emergency in cyberspace since October 9, which honestly explains the 5:55 AM post. Oh, and eMudhra, a certificate authority that's issued over 60 million certificates, whose entire job is proving websites are who they say they are, was supposed to be using private keys that are random numbers over a hundred digits long, and one of them was 4. Just 4. And the others were like 66 and 237. Anyway, it's Wednesday.

7:12 AM · Oct 22, 2026

69.4K Reply Copy link to post

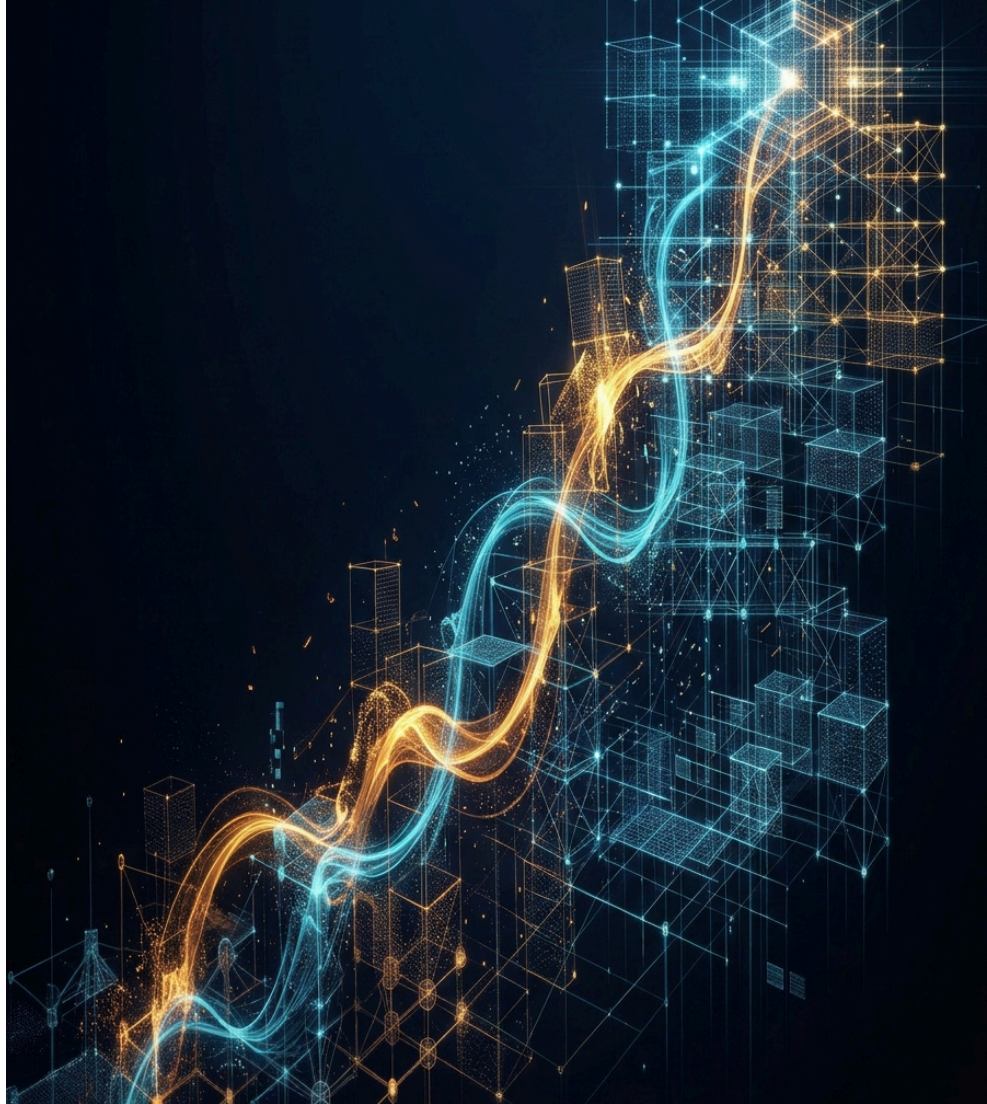
Read 1.1K replies

MLSG MACHINE LEARNING SINGAPORE

# Wrap-Up

- Single-model RLVR has limits
  - another model changes *what can be learned*
- Three multi-model modes:
  - **Competition** (SPIRAL, RARO) :
    - opponents & critics expose weaknesses
  - **Cooperation** (RLT, SDPO) :
    - privileged context & hindsight turn into dense supervision
  - **RSI** (SPADE, Zero-Data) :
    - co-evolving executable worlds beat static datasets
- Be sure that it's working as intended:
  - *What external anchor keeps the models honest?*
  - *Does it win at matched total compute?*

MLSG
MACHINE LEARNING
SINGAPORE



# Link to Slides

- https://bit.ly/MLSG_2026-10




MLSG
MACHINE LEARNING
SINGAPORE