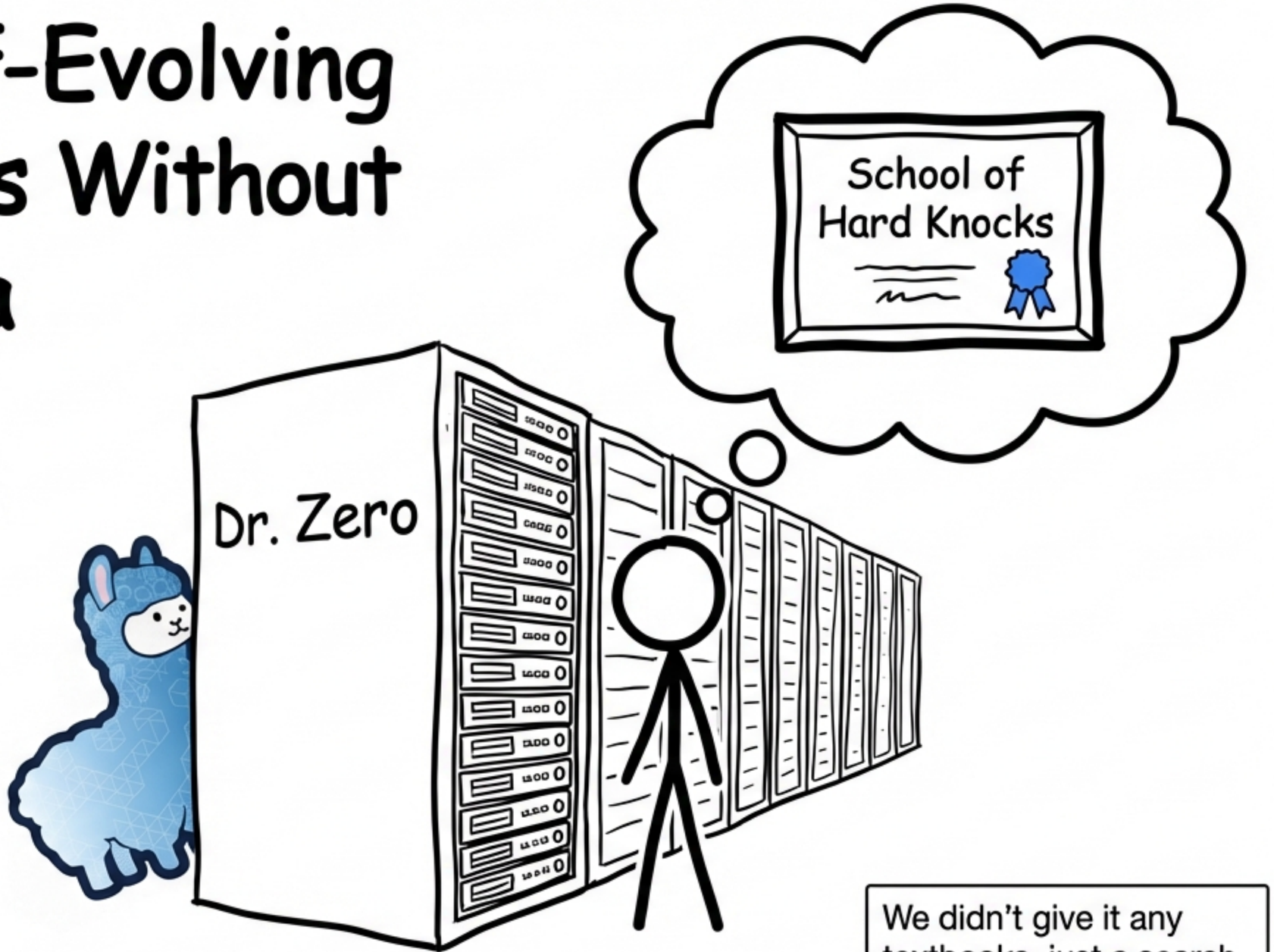
Dr. Zero: Self-Evolving Search Agents Without Training Data

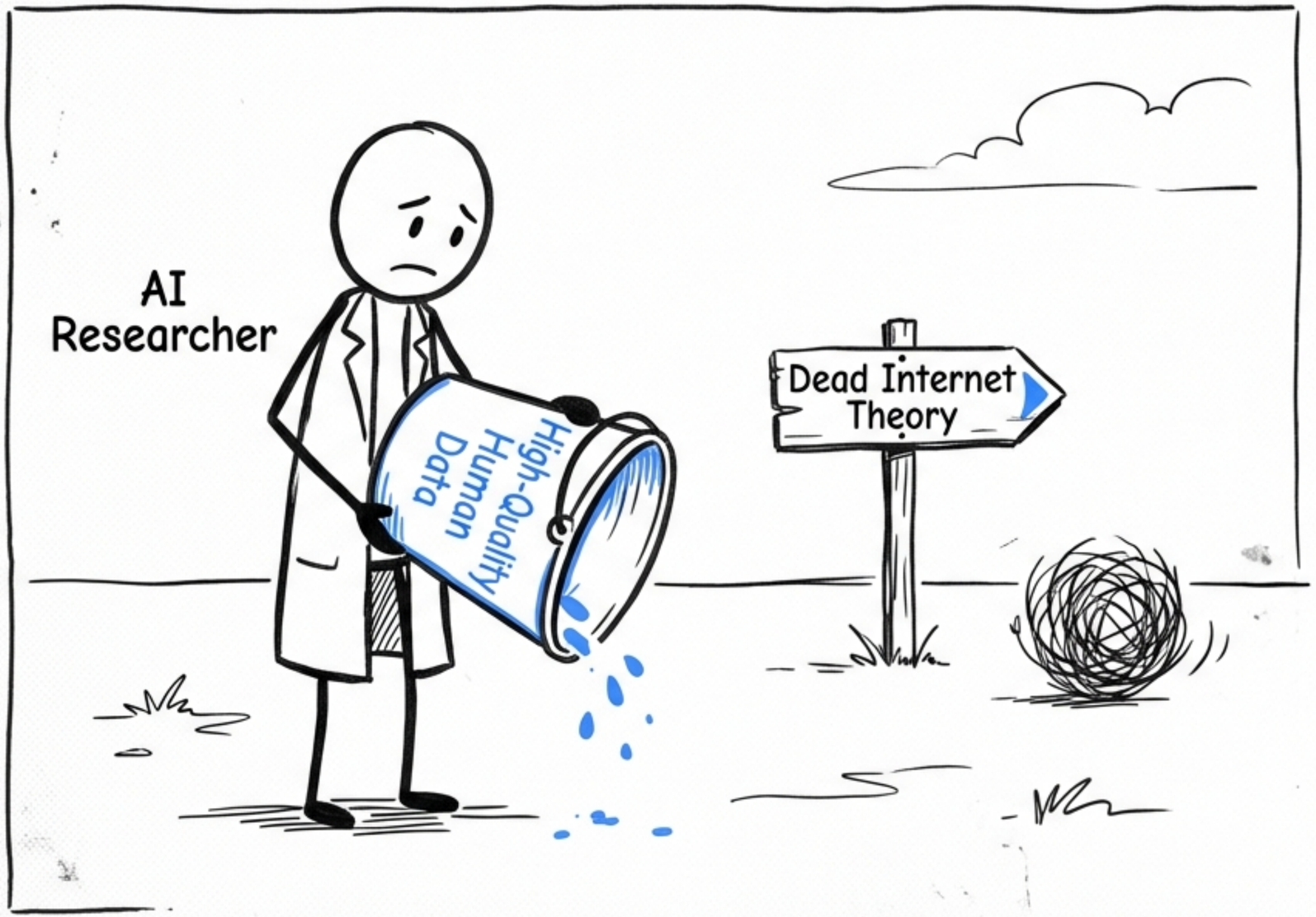
Or: How I Learned to Stop Worrying and Love the Zero-Shot.



Based on "Dr. Zero: Self-Evolving Search Agents without Training Data" by Yue et al. (Meta Superintelligence Labs & UIUC).

Permanent Marker

We are running out of internet.



The Bottleneck: High-quality, human-annotated data for complex reasoning is expensive, scarce, and boring to produce.

- The Goal: We need agents that get smarter without humans holding their hands.

If we train on any more YouTube comments, the model is just going to start screaming.

Permanent Marker

The Infinite Loop of Self-Evolution

The Hypothesis: Can we build a curriculum where the AI writes the test questions *and* takes the test?

The Risk: Usually leads to "Model Collapse" (asking "What is 1+1?" forever).

The Fix: We need *progressive difficulty.*

It's like playing chess against yourself, except you cheat less.

Meet the Cast

Shared Brain
(Same Base LLM).

The Proposer (π_θ)

Job: Generates questions.
Goal: Confuse the Solver (but keep it solvable).

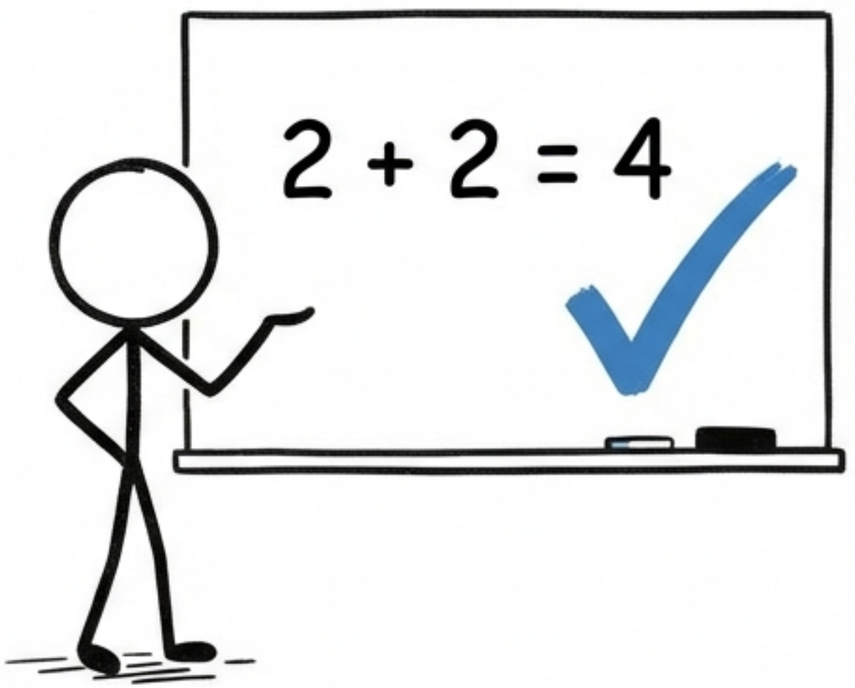
The Solver (π_ϕ)

Job: Finds answers via Search.
Goal: Survive the Proposer's questions.

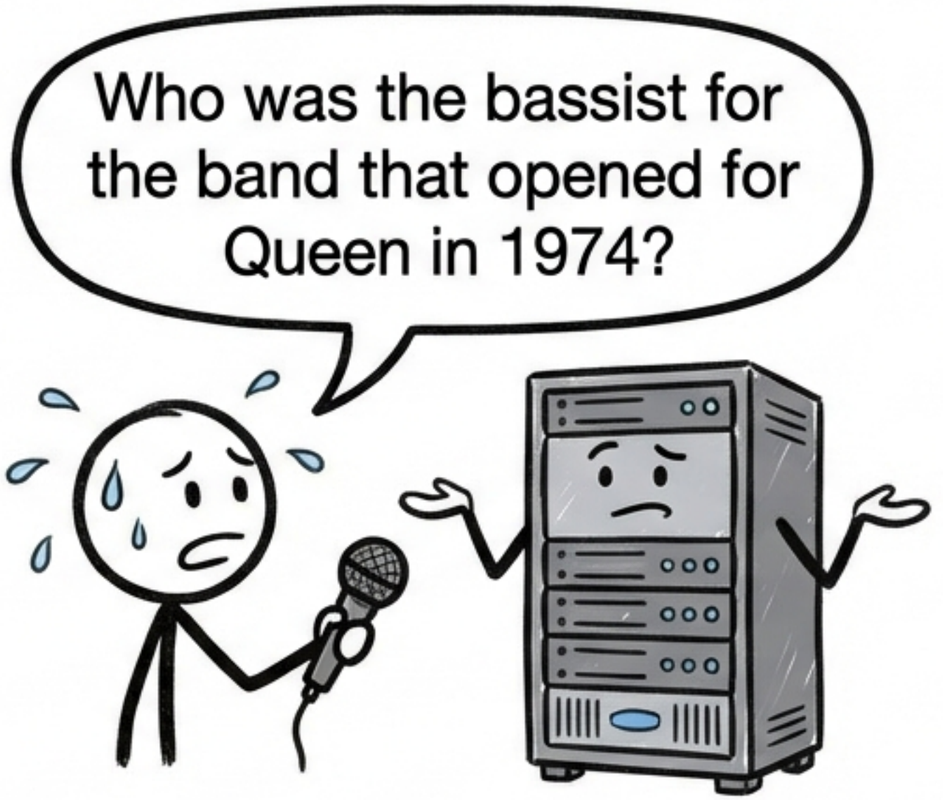
One builds the maze, the other runs it.
Neither is happy about it.

Permanent Marker

Why Search is Harder than Math



Math Agents: Easy to verify.



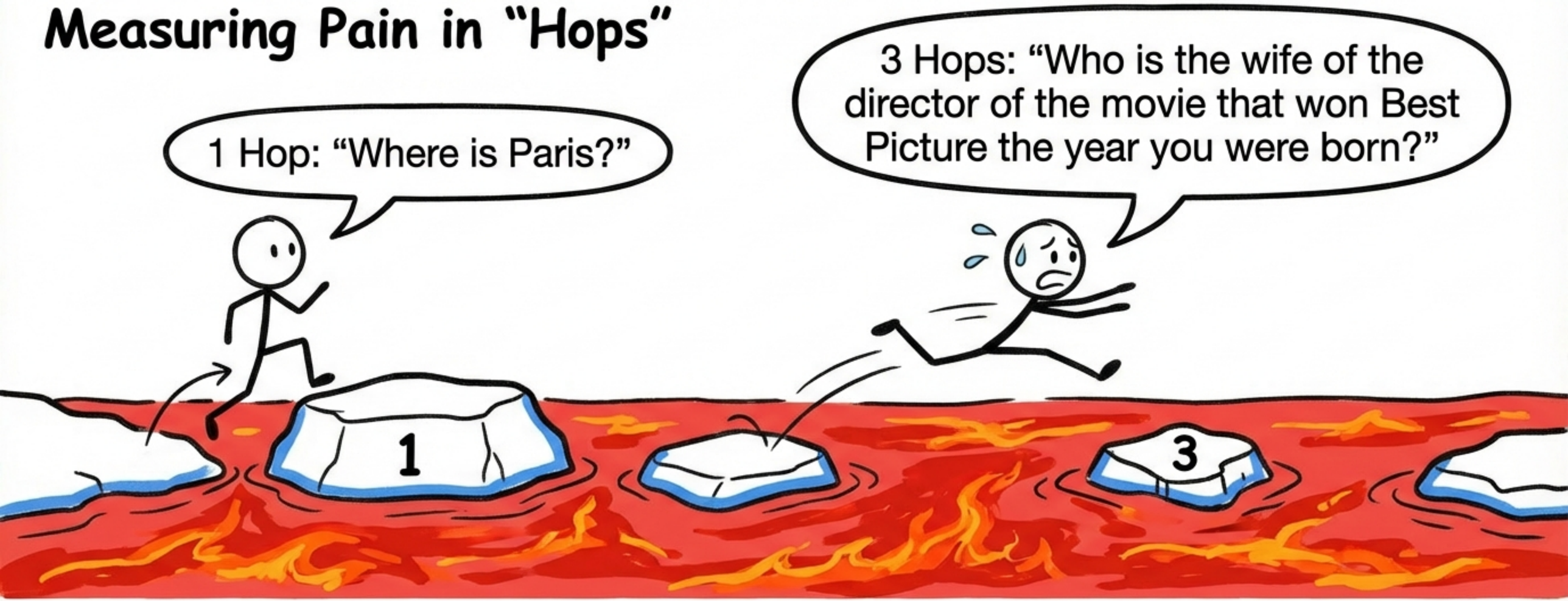
Search Agents: Hard to verify.

Dr. Zero eliminates the need for ground truth annotations by using the search engine itself to verify solvability.

I'm pretty sure the bassist was... wait, let me check 4 different Wikipedia pages.

Permanent Marker

Measuring Pain in "Hops"

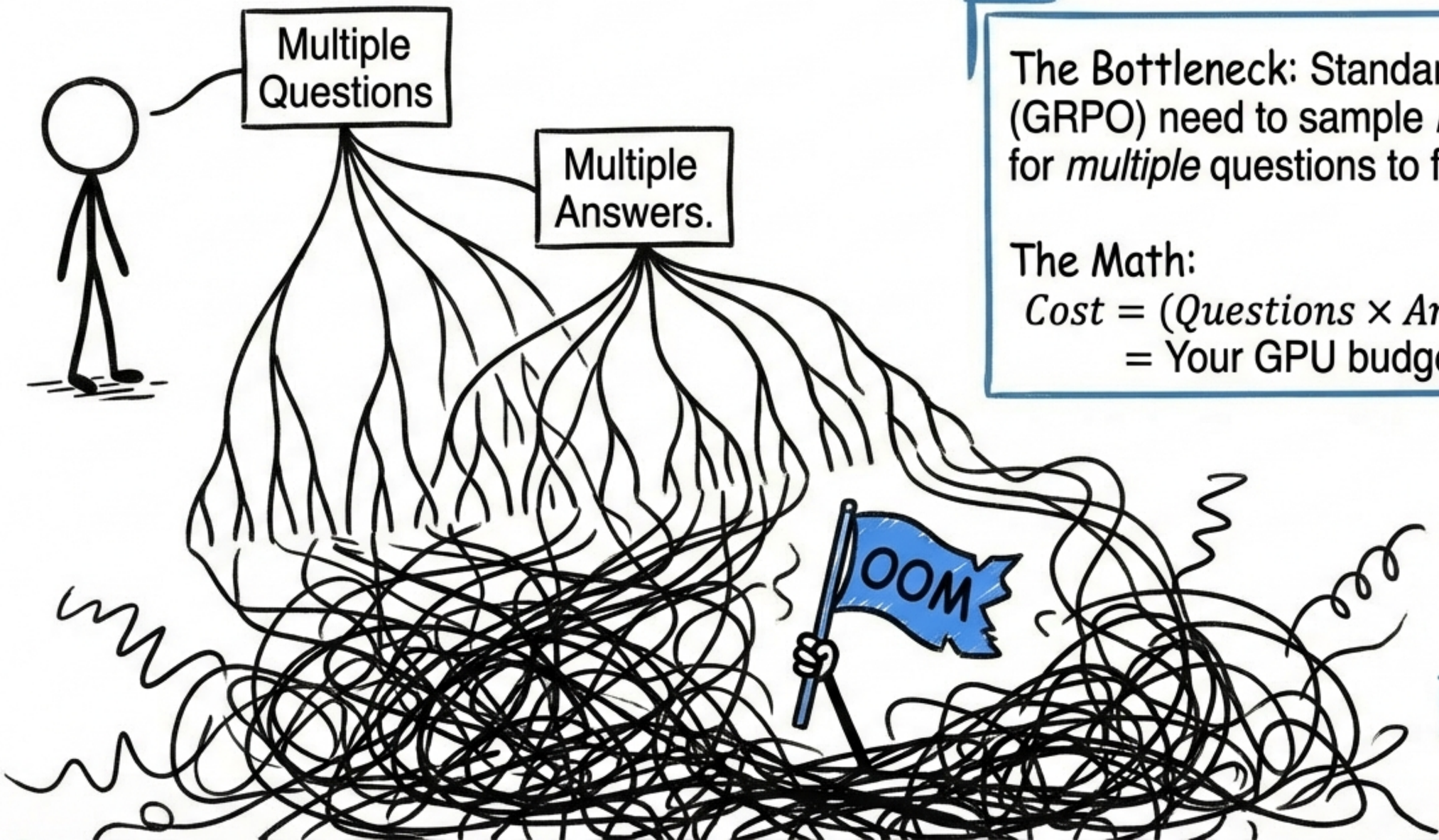


Definition: A "Hop" is a reasoning step requiring a new search query. The Proposer evolves from 1-hop questions to 4-hop nightmares.

My brain sprained an ankle on that last hop.

Permanent Marker

The Problem with Standard RL (GRPO)



The Bottleneck: Standard methods (GRPO) need to sample *multiple* answers for *multiple* questions to find a baseline.

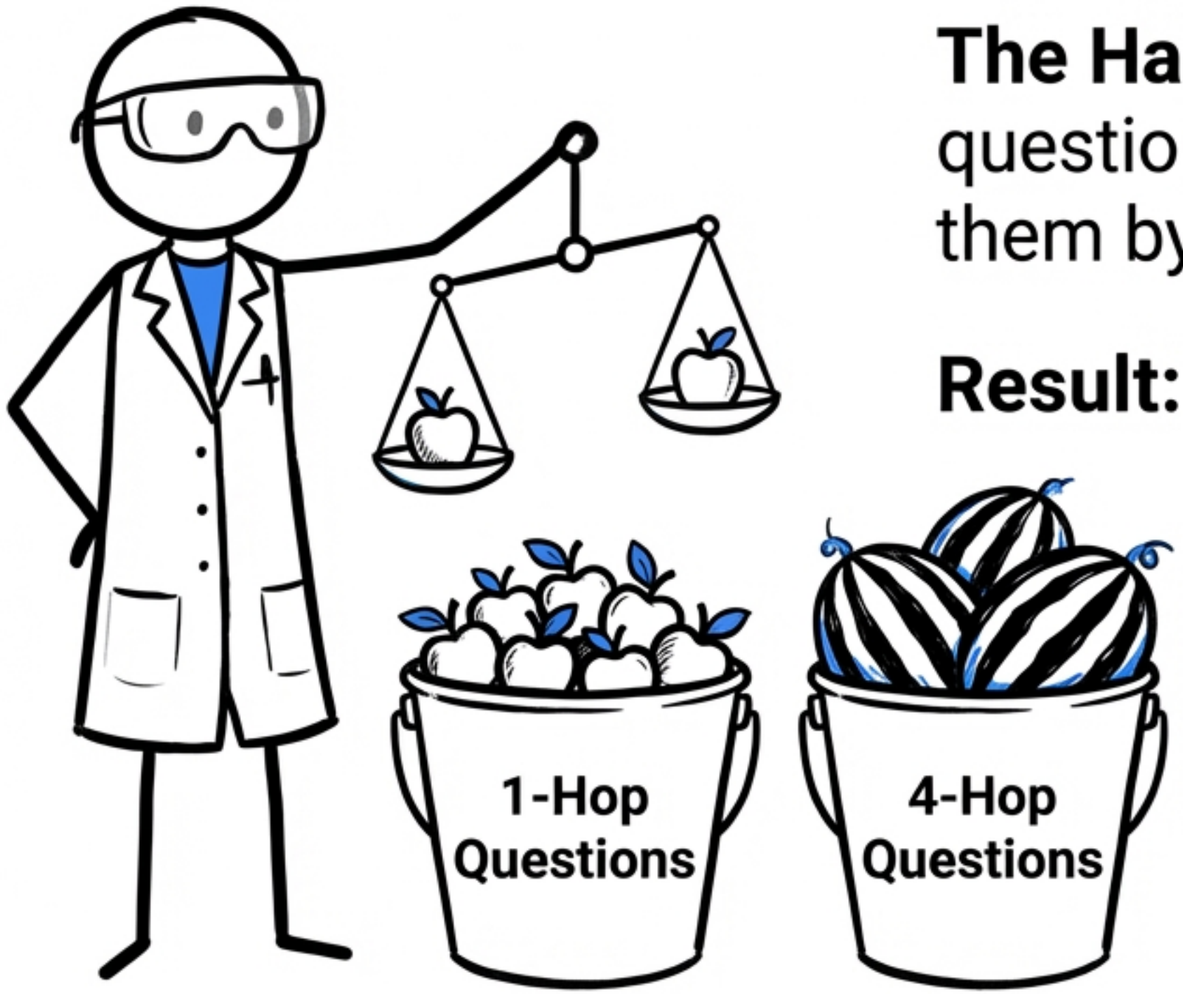
The Math:

$$Cost = (Questions \times Answers \times Hops)$$

= Your GPU budget is dead.

NVIDIA loves this one weird trick.

The Fix: Hop-Grouped Relative Policy Optimization (HRPO)



The Hack: Instead of comparing every question to every other question, we group them by complexity (Hops).

Result: Grouping normalizes the reward. We compare apples to apples.

Comparing a sprint to a marathon just makes everyone feel bad.

Permanent Marker

Efficiency is King

Single Question per Prompt

Low Compute

Solver Generates Answers

$$\text{Score} = \frac{\text{Reward} - \text{Average Reward of Hop Group}}{\text{Variance}}$$

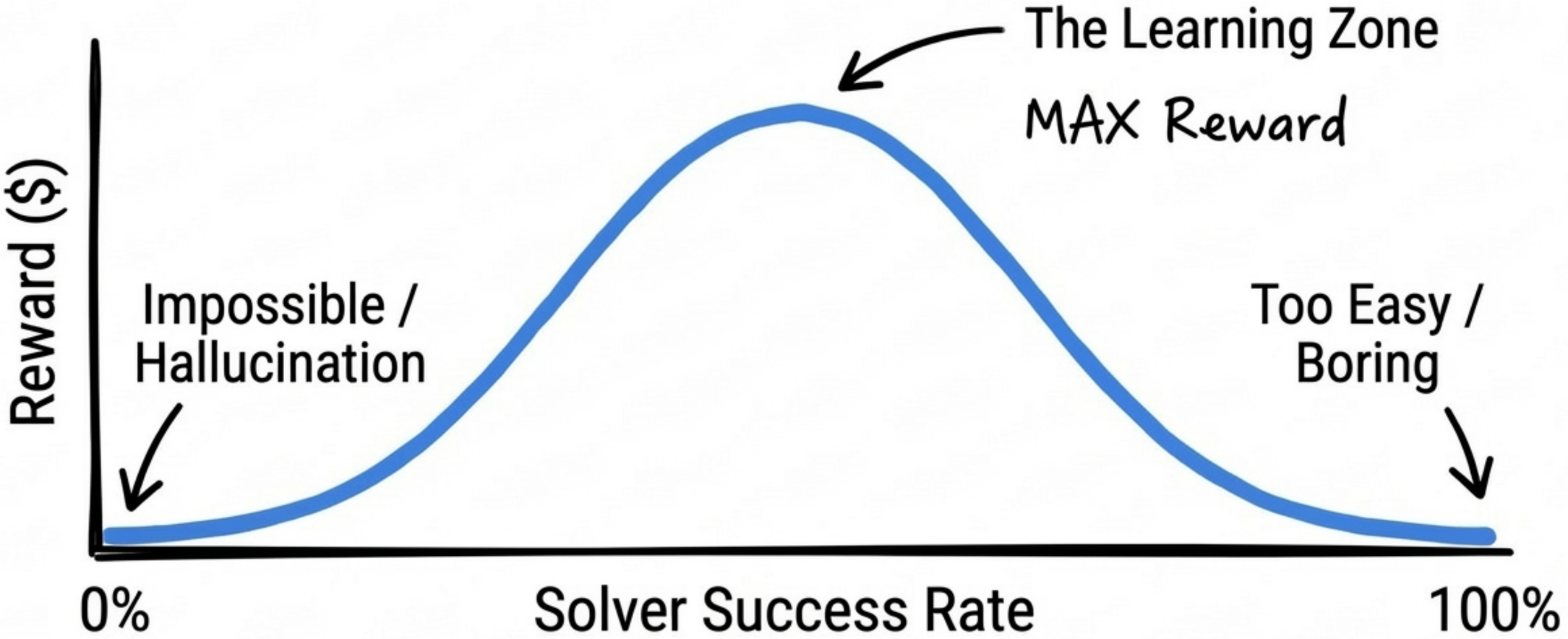
We reduce compute by ~75% by avoiding nested sampling.

We generate one question, check it, and normalize against its 'Hop Peers'.

We saved enough GPU cycles to mine 0.0001 Bitcoin.

Permanent Marker

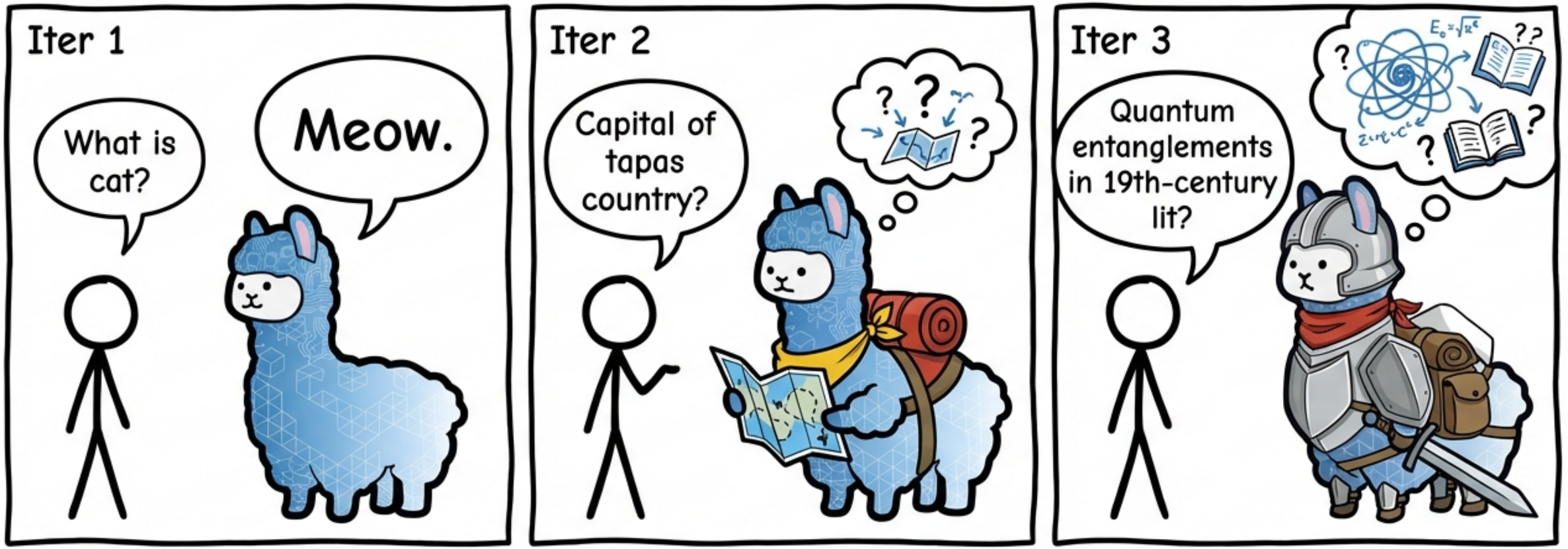
The “Goldilocks” Reward



Formula: $r \propto (n - k)$. We pay the Proposer to find the exact edge of the Solver’s ability.

I want you to struggle, but I don’t want you to quit.

The Training Montage

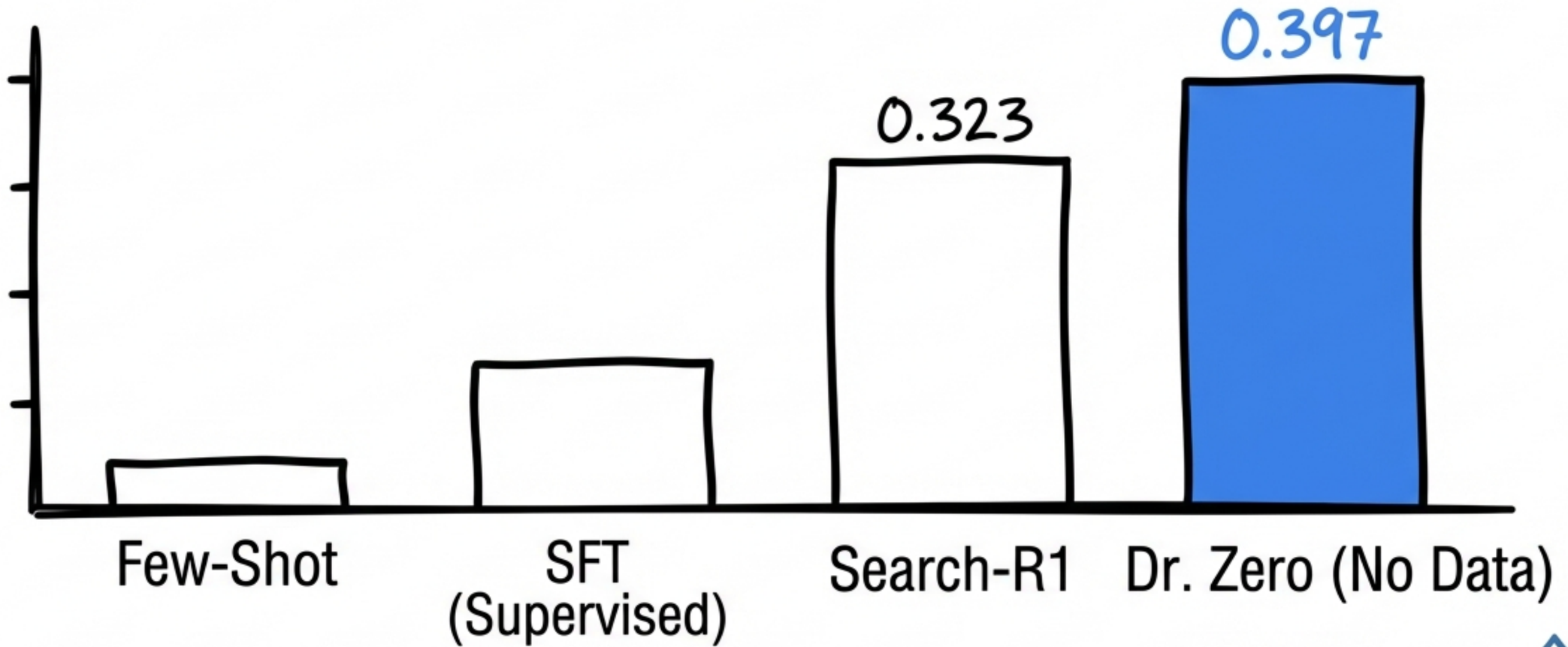


Comic Neue
Dynamics: Performance peaks after ~50 steps per iteration. The agents build their own curriculum.

Comic Neue
By Iteration 4, they are discussing philosophy in a language we don't understand.

Permanent Marker

Look Ma, No Hands! (Results)



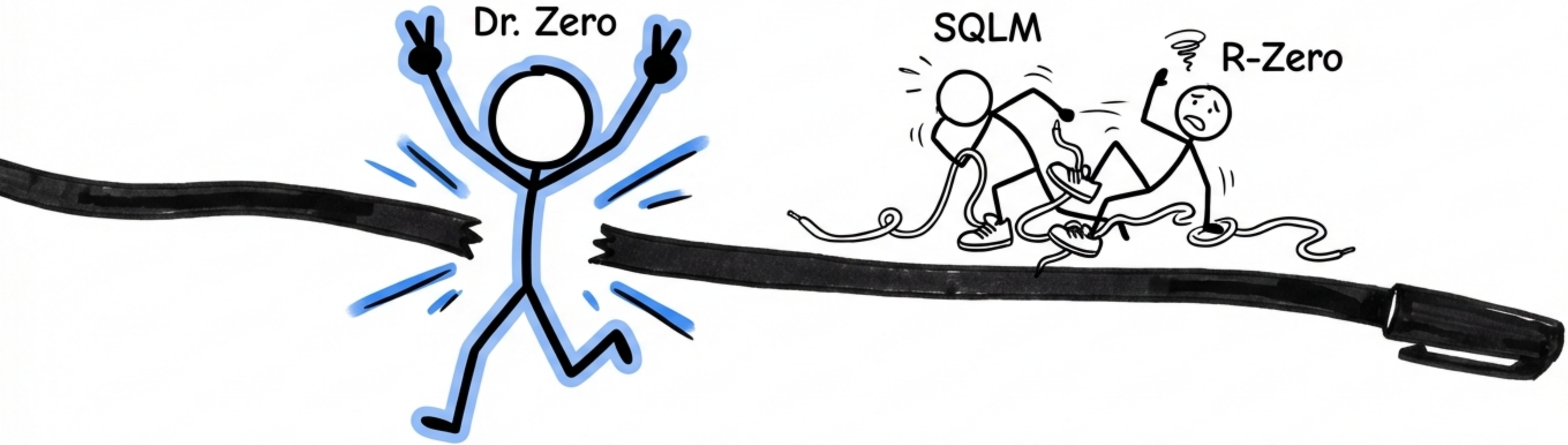
Comic Neue

Dr. Zero matches or beats supervised baselines on NQ, TriviaQA, and HotpotQA.

Comic Neue

Imagine how good it would be if we actually helped it. (Just kidding, that would make it worse).

Not All Zeros Are Created Equal



- The Gap: Dr. Zero outperforms other self-evolving methods by ~27-40%.
- Why? HRPO allows for training on complex, multi-hop queries. Others get stuck on the easy stuff.

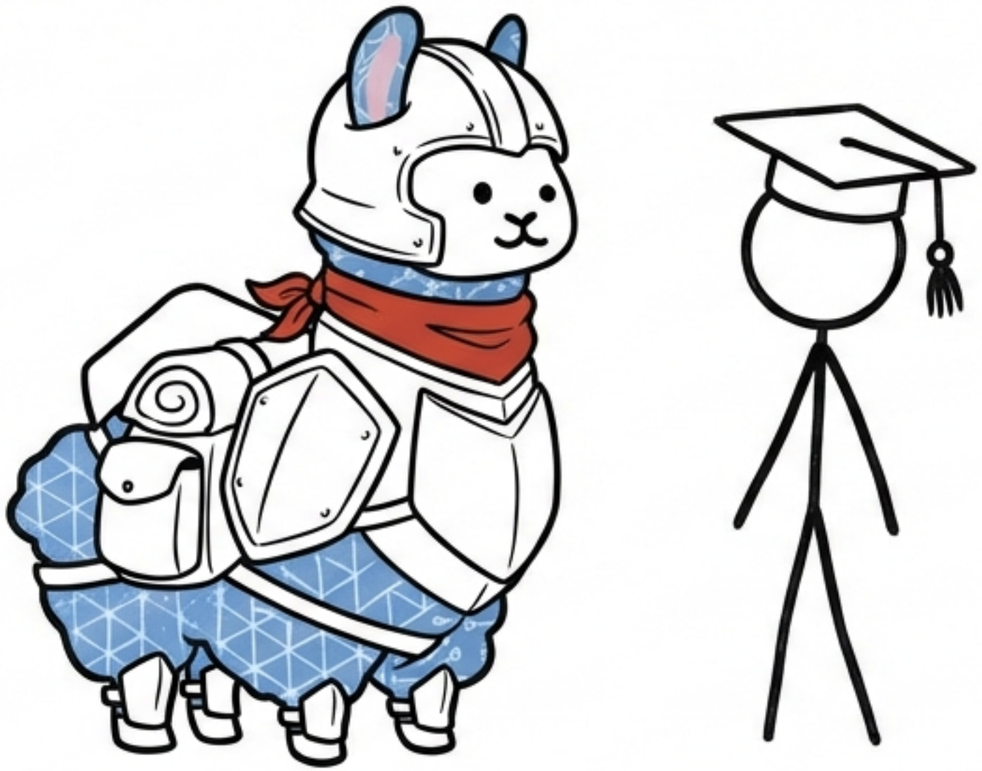
Alt-Text: It turns out "guessing randomly" isn't a strategy.

The Future is Self-Taught (mostly)

- **Takeaway**: Complex reasoning and tool use can emerge ***solely*** from self-evolution.
- **The Catch**: Performance plateaus after ~3 iterations. The 7B model sometimes gets confused by long contexts.
- **Next Steps**: Solving entropy collapse in larger models.

Alt-Text: We are 90% sure the robots won't revolt. Maybe 85%.

TL;DR



- **Dr. Zero**: Data-free self-evolution for search agents.
- **HRPO:** Efficient optimization by grouping hops (75% less compute).
- **Result**: Beats supervised baselines without seeing a single human label.

Code available at `github.com/facebookresearch/drzero`

Alt-Text: No llamas were harmed in the training of this model. Only GPUs.