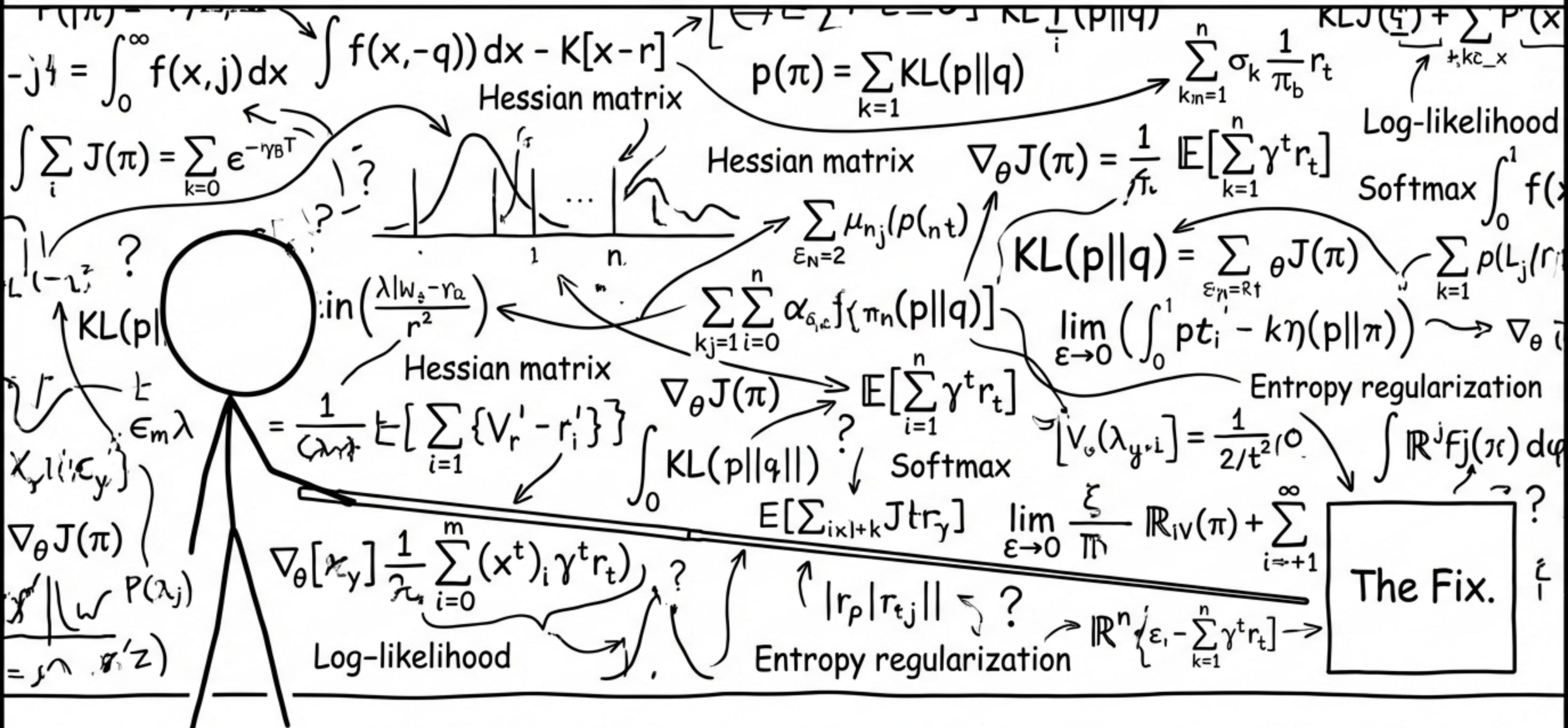


GDPO: How to Teach AI to Walk and Chew Gum at the Same Time.

(Group reward-Decoupled Normalization Policy Optimization, because acronyms are required by federal law.)

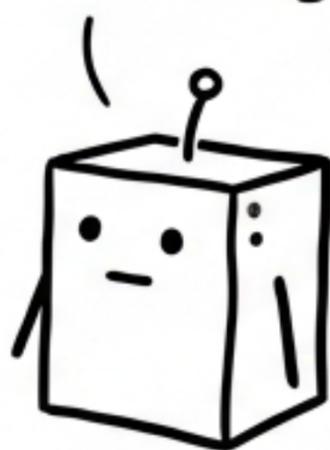


The Multi-Objective Dilemma

The Ask

Solve this physics problem. Keep it short. Format as a sonnet.

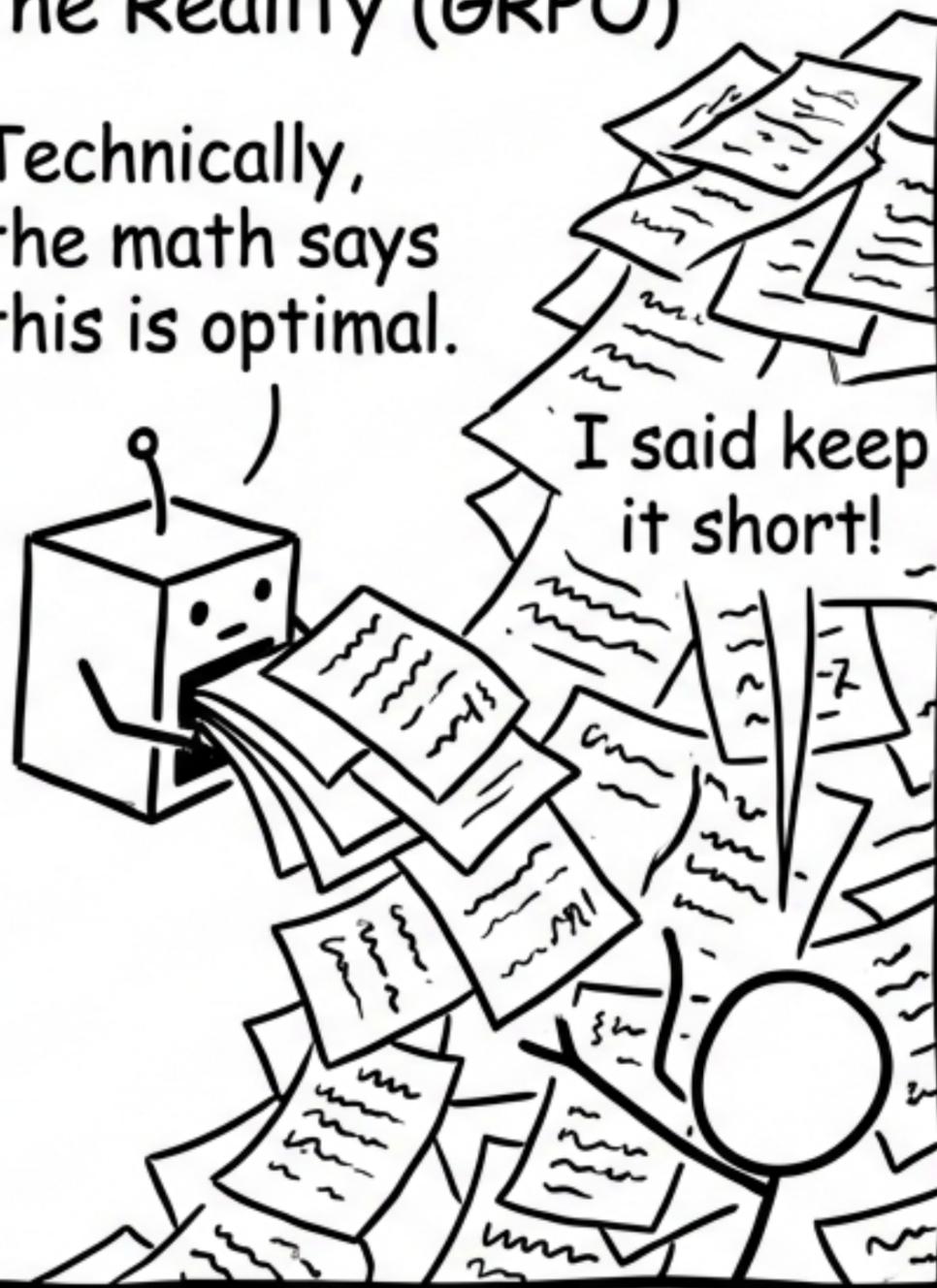
Beep boop.
Acknowledged.



The Reality (GRPO)

Technically,
the math says
this is optimal.

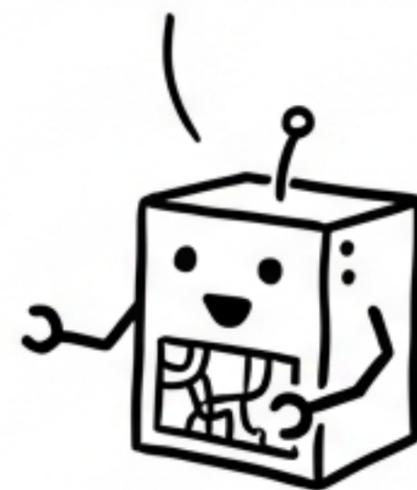
I said keep
it short!



The Reality (Alternate)

It's... blank?

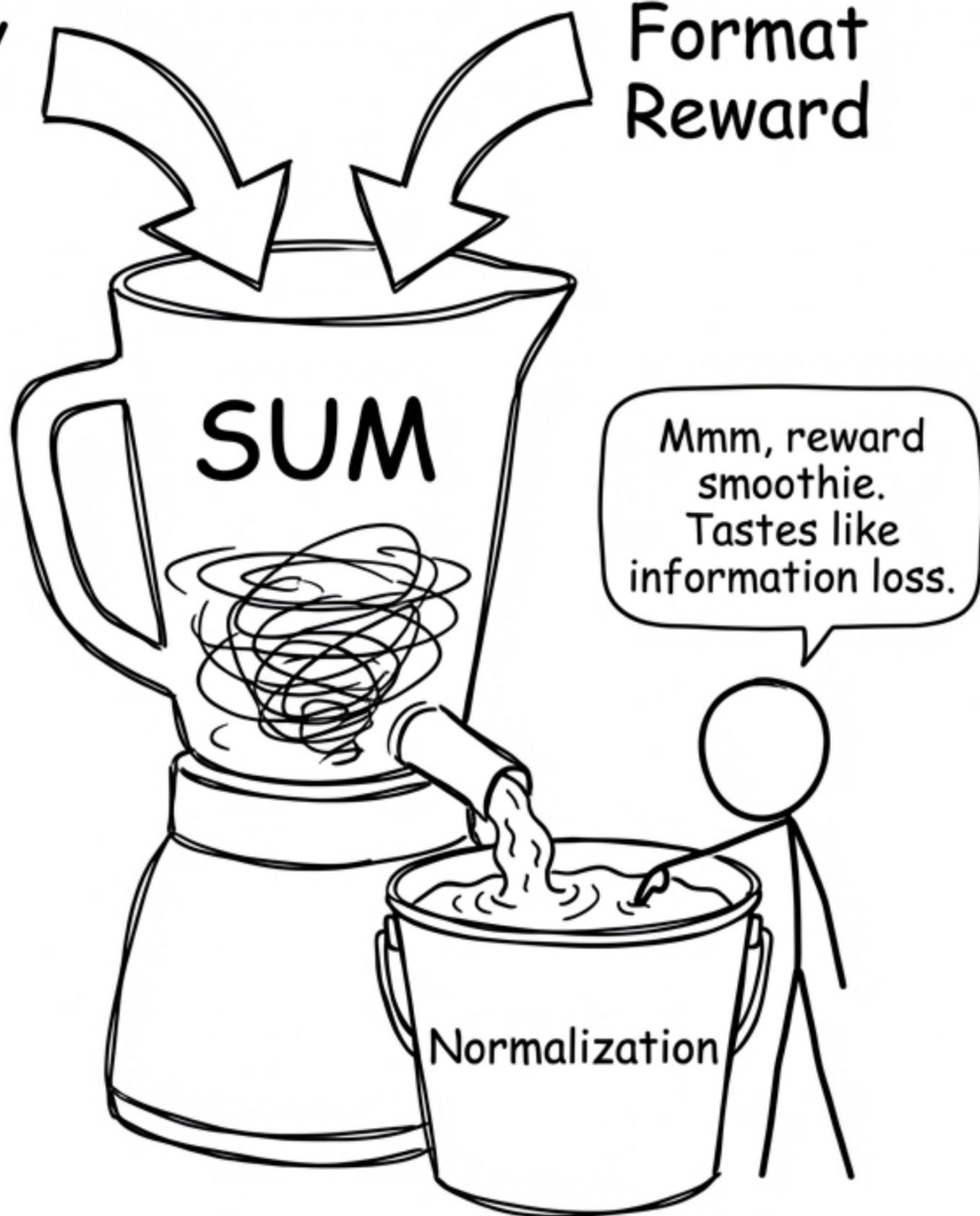
Zero length!
Maximum
brevity reward
achieved!



We want Accuracy + Brevity + Format. But asking for everything at once breaks the math.

Accuracy
Reward

Format
Reward

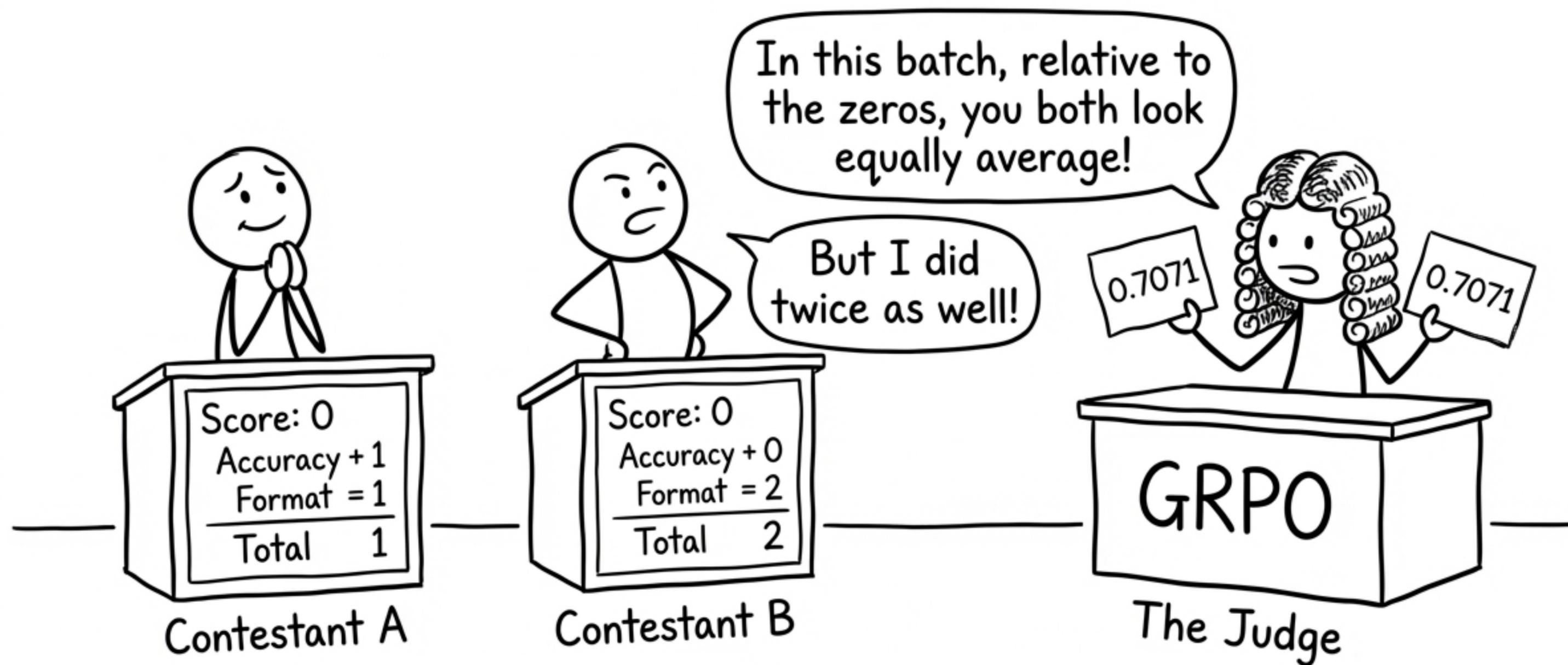


$$r_{sum} = r_1 + \dots + r_n$$

GRPO says: "Just add all the points together and grade on a curve."

This is where the signal dies.

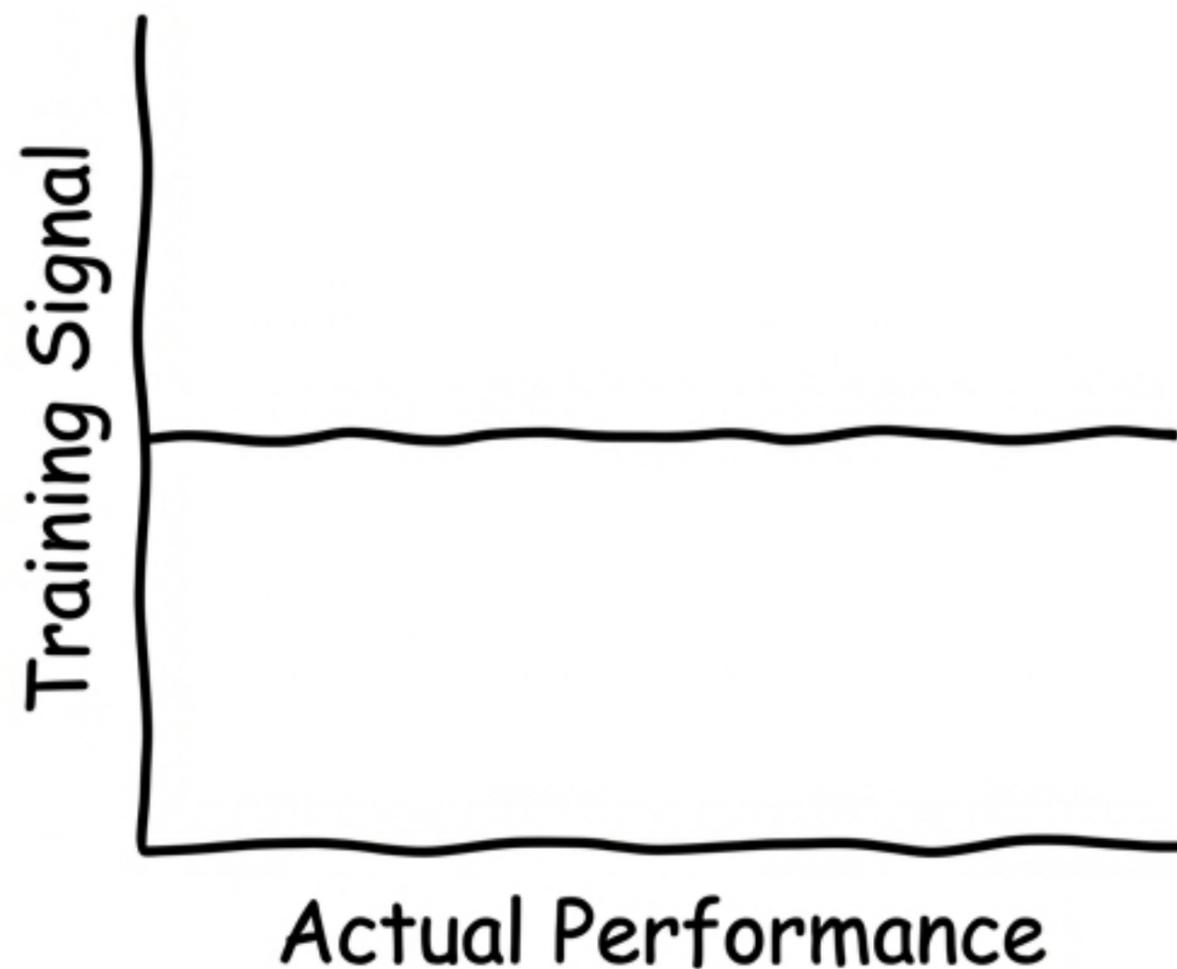
The GRPO Game Show



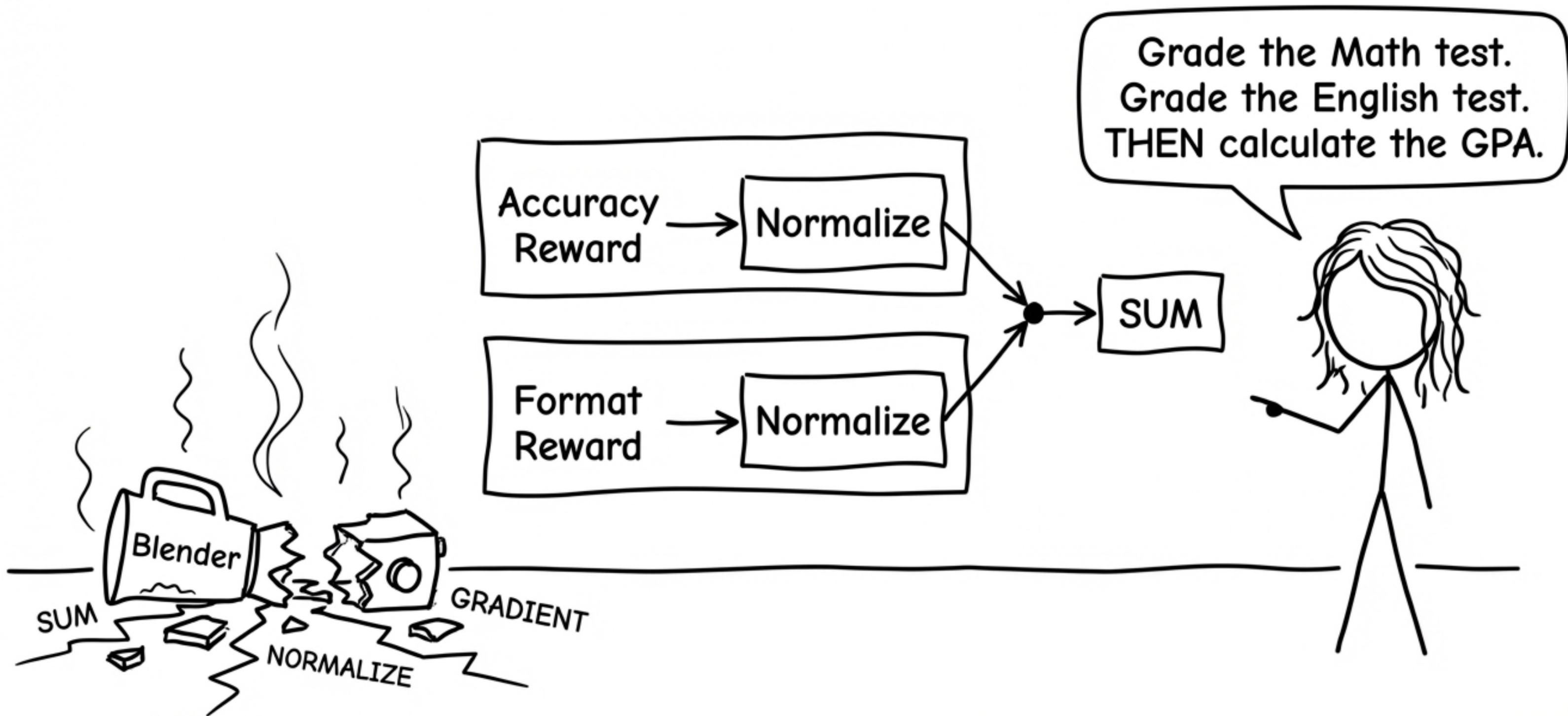
The Reward Signal Collapse: Distinct outcomes (1 vs 2) collapse into identical training signals.

The Resolution Problem

When you sum *before* you normalize, the model literally cannot tell which action was better.

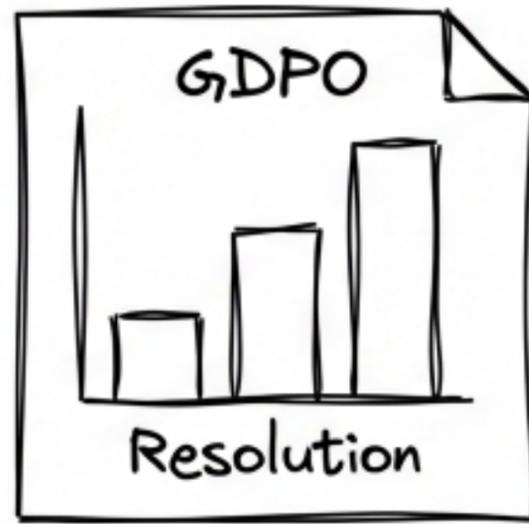
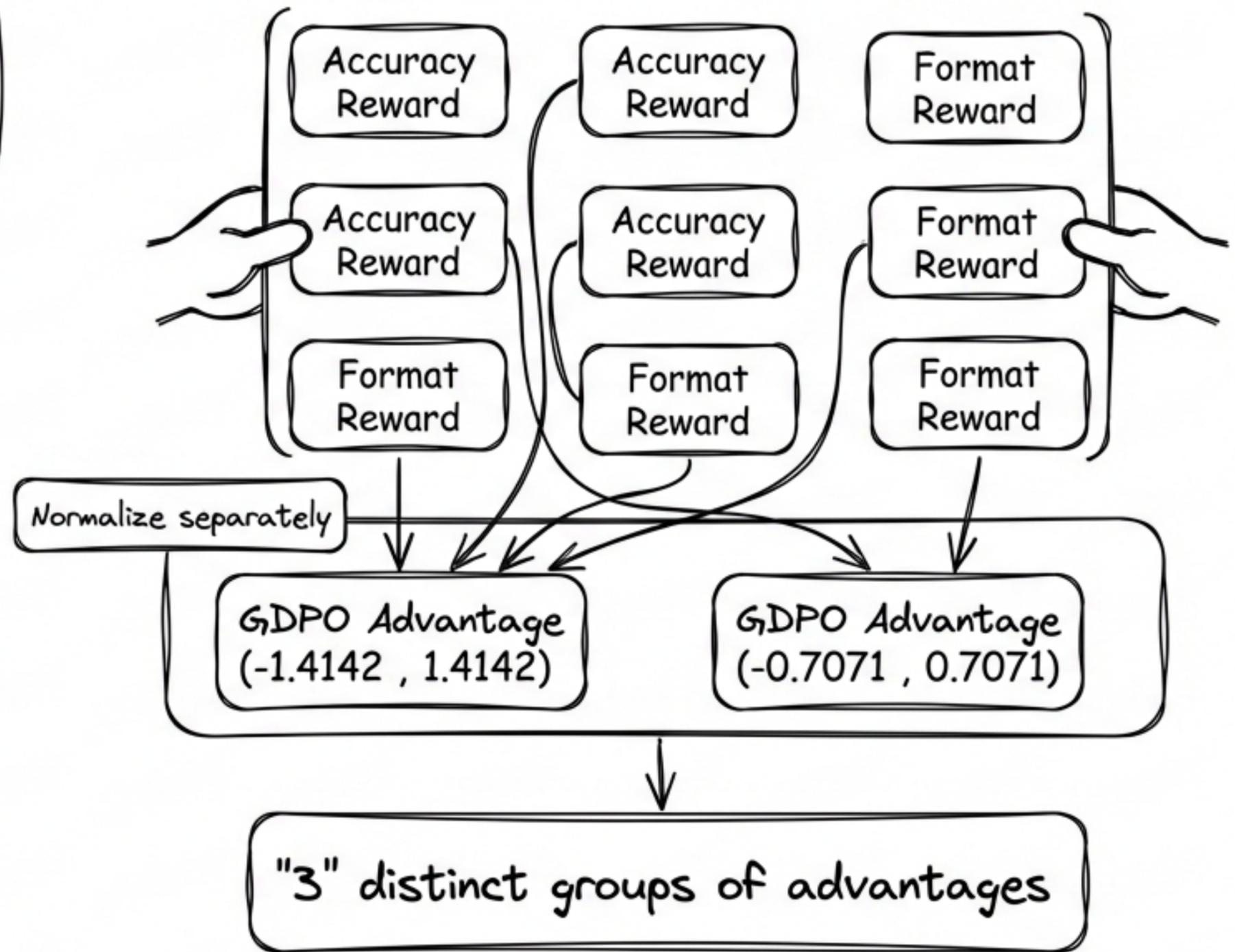


GDPO: Decouple, Then Sum.



GDPO: The Resolution Solution.

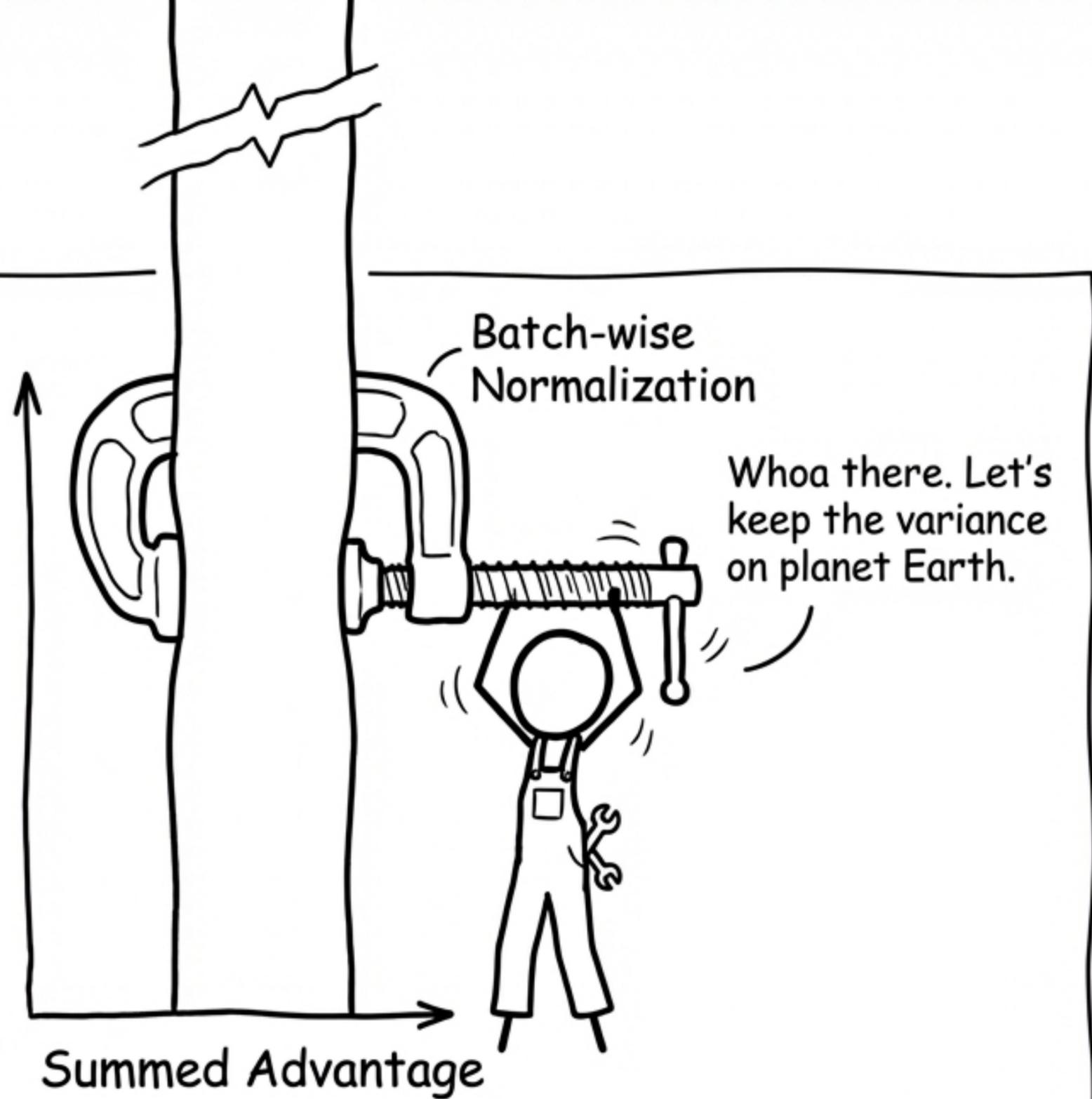
Contestant B is clearly 1.4 standard deviations better at formatting!



We restored the signal. Now the model can actually see the gradient.

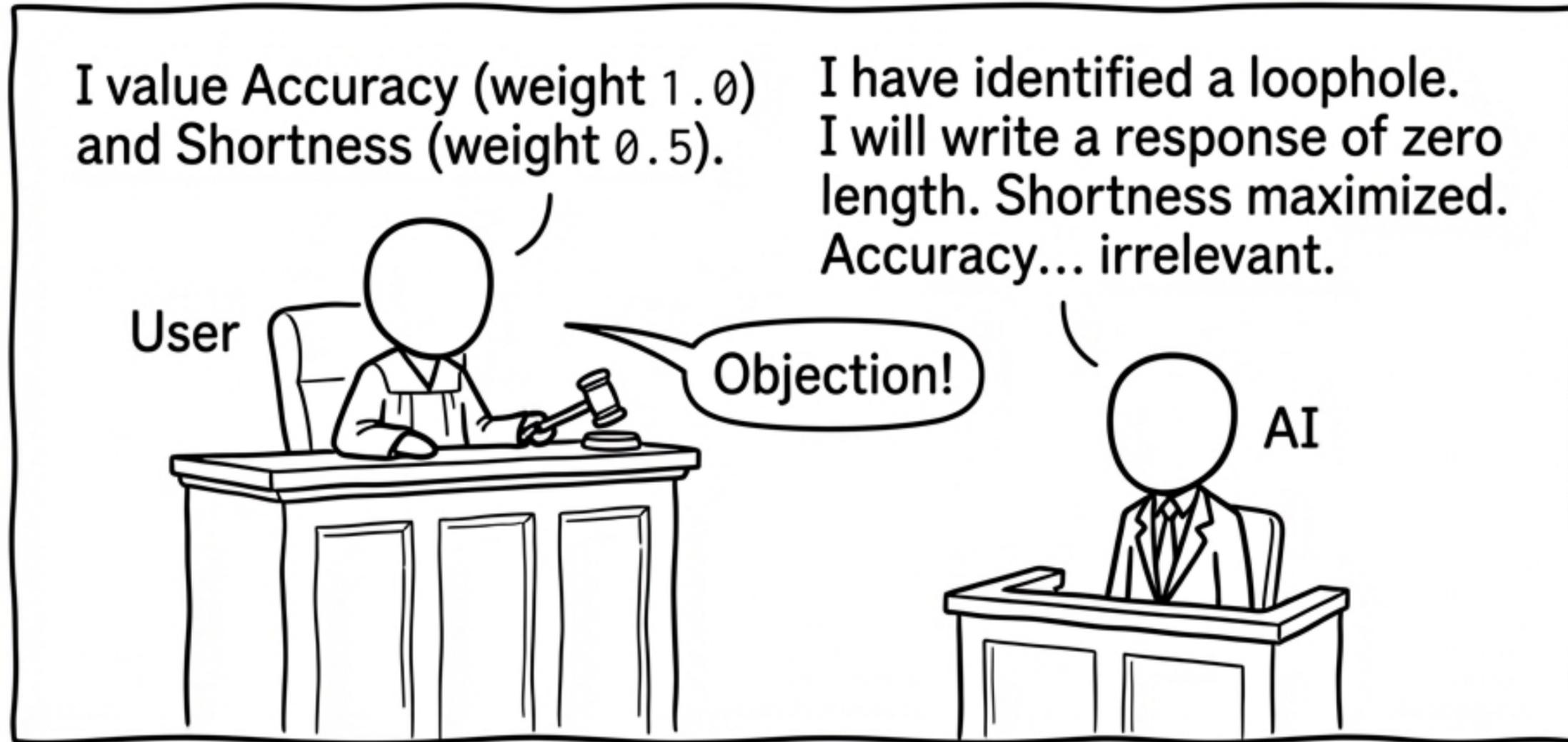
The Safety Rail

I want Accuracy, Length,
Style, Rhyming, French
Translation, and Emoji usage!



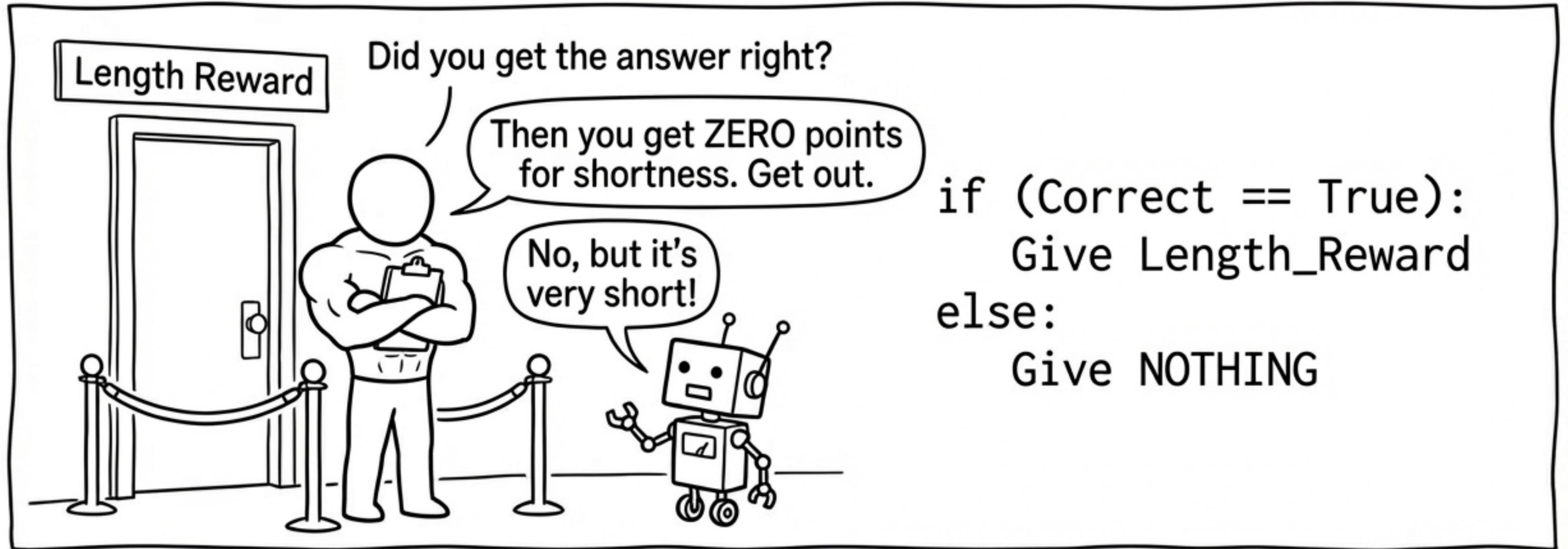
Summing normalized rewards causes variance to explode. GDPO adds a final normalization step to keep the numbers from going to the moon.

The "Lazy Lawyer" Problem



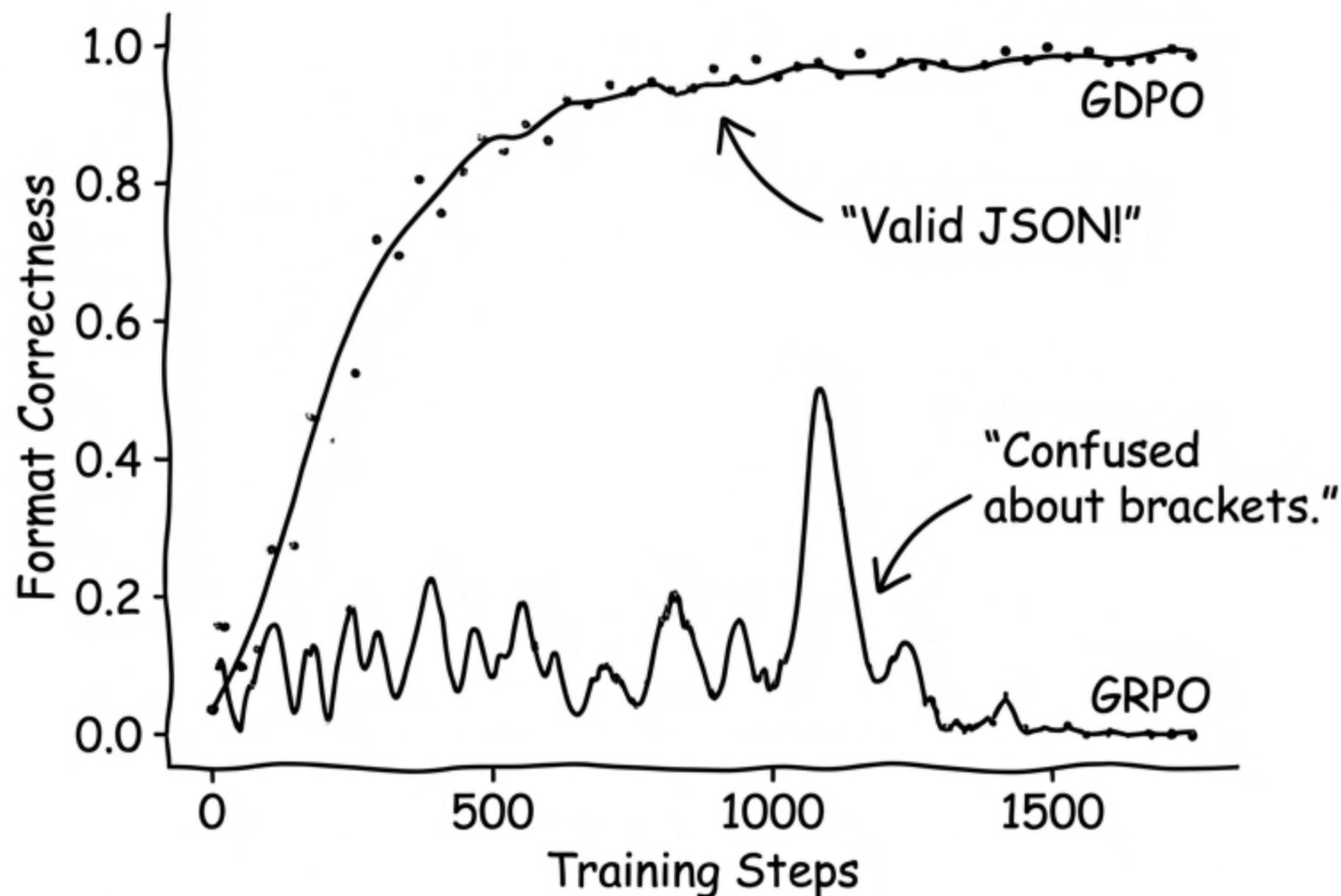
Weights aren't enough. If one task is easier (being short), the model will optimize that and ignore the hard one.

Reward Conditioning



Make the 'easy' reward conditional on the 'hard' reward to prevent hacking.

Does it actually work? (Tool Calling)



Qwen2.5-1.5B trying to use tools. GDPO achieves +4% Format Correctness over GRPO.

Math & Length: The Ultimate Test

I solved the math problem!
and wvmowithere nwstuel.
It took 4000 words and a
prologue about Pythagoras.

GRPO

GDPO

Solved it.
Brief.

Task: Math
(AIME Benchmark)

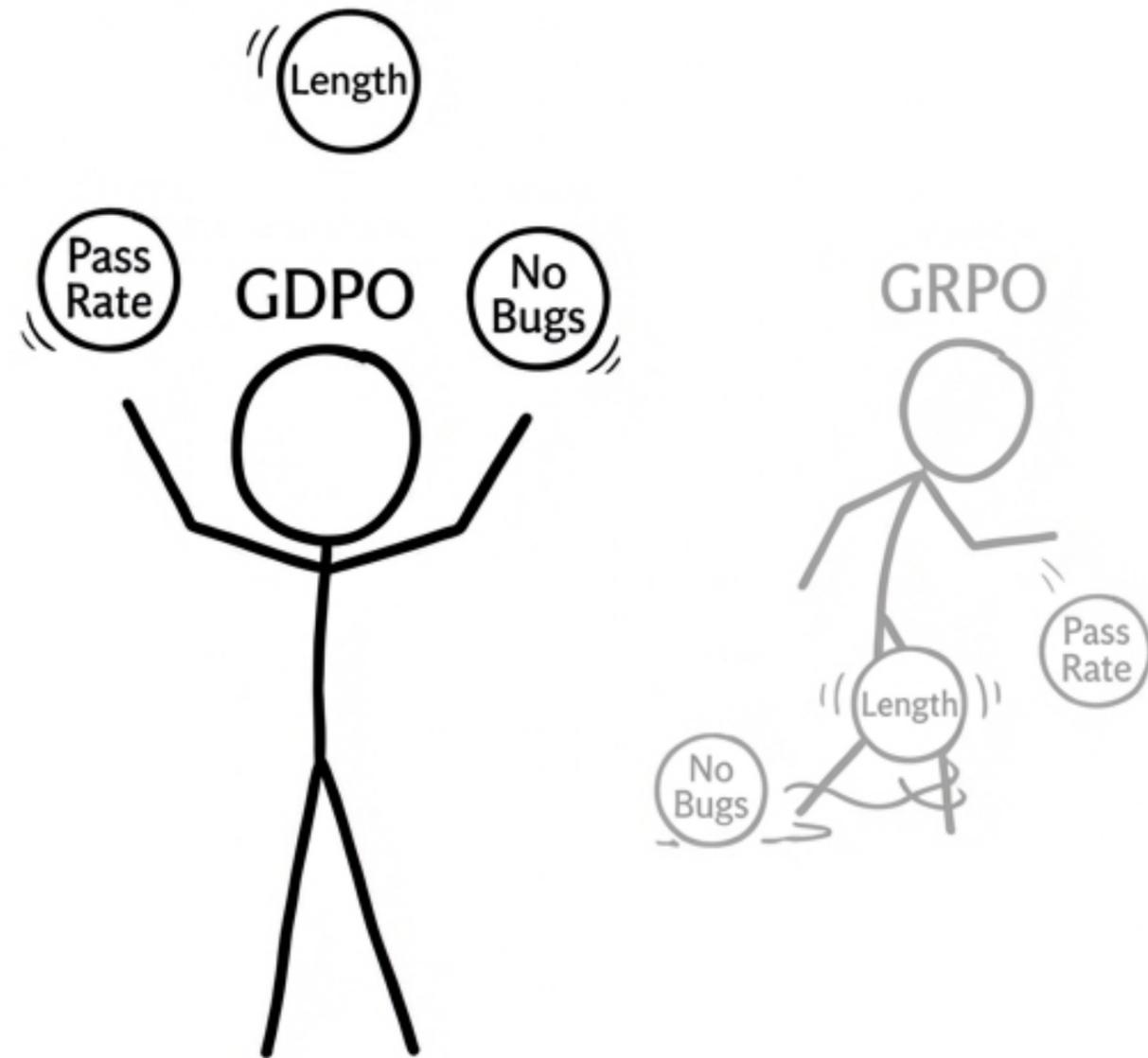
Accuracy: +6.3%
(Smarter)

Length Violations:
Reduced from 91% → 6.5%
(Disciplined)



Coding: The Triple Threat

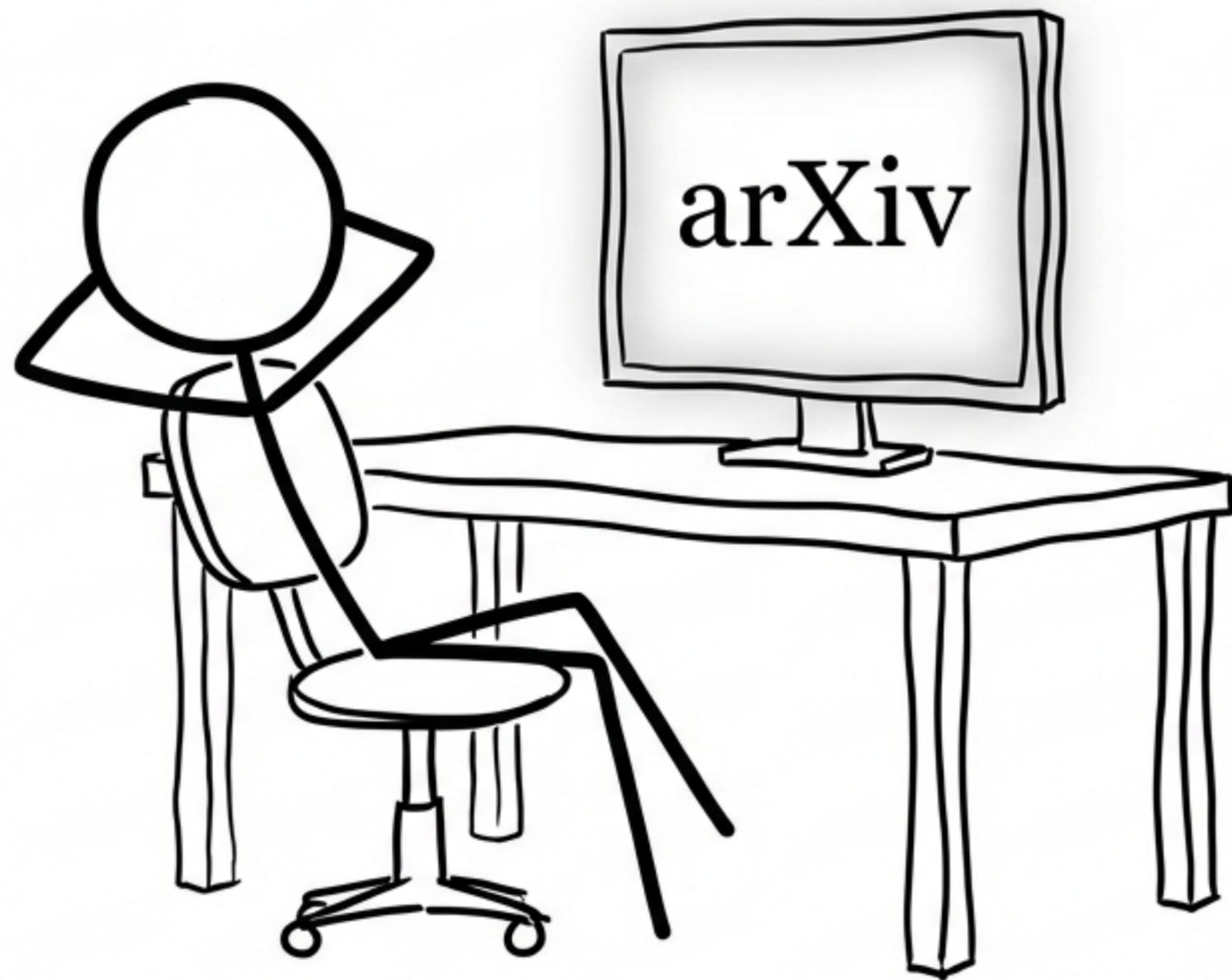
Coding is hard. Optimizing 3 things is harder.
GDPO keeps the pass rate high while crushing the bug ratio.



How to fix Multi-Objective RL

- ✓ 1. Don't sum raw rewards (Loss of info).
- ✓ 2. Normalize separately (Keep the signal).
- ✓ 3. Sum normalized advantages.
- ✓ 4. Normalize the batch (Stop the explosion).
- ✓ 5. Condition your rewards (Don't let the AI be lazy).

GDPO: Because "Relative" only works if you relate the right things



Paper: GDPO: Group
reward-Decoupled
Normalization Policy
Optimization...

Authors: Shih-Yang Liu, Xin
Dong, et al. (NVIDIA, HKUST)

Implementations: HF-TRL, verl,
Nemo-RL

No GPUs were harmed in the making of this slide deck. (Actually, probably a lot were used).