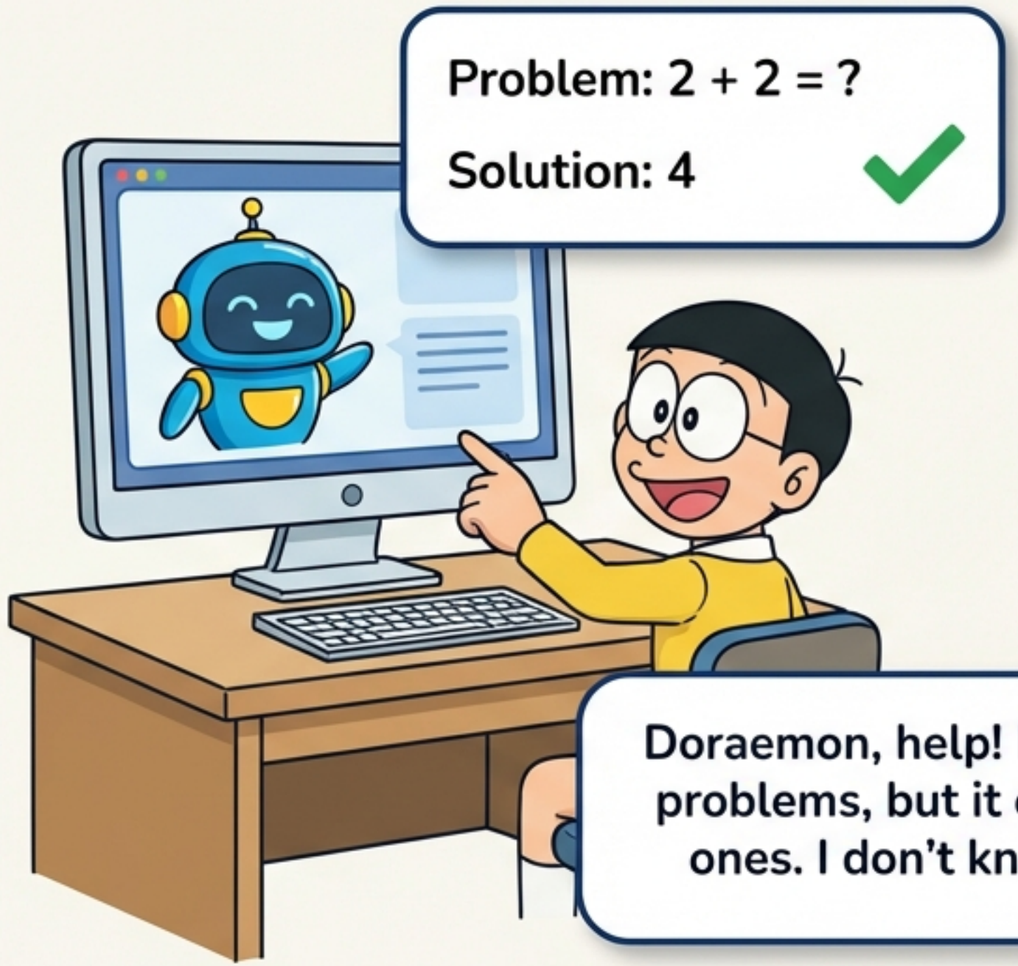
Doraemon's Guide to AI Reasoning!

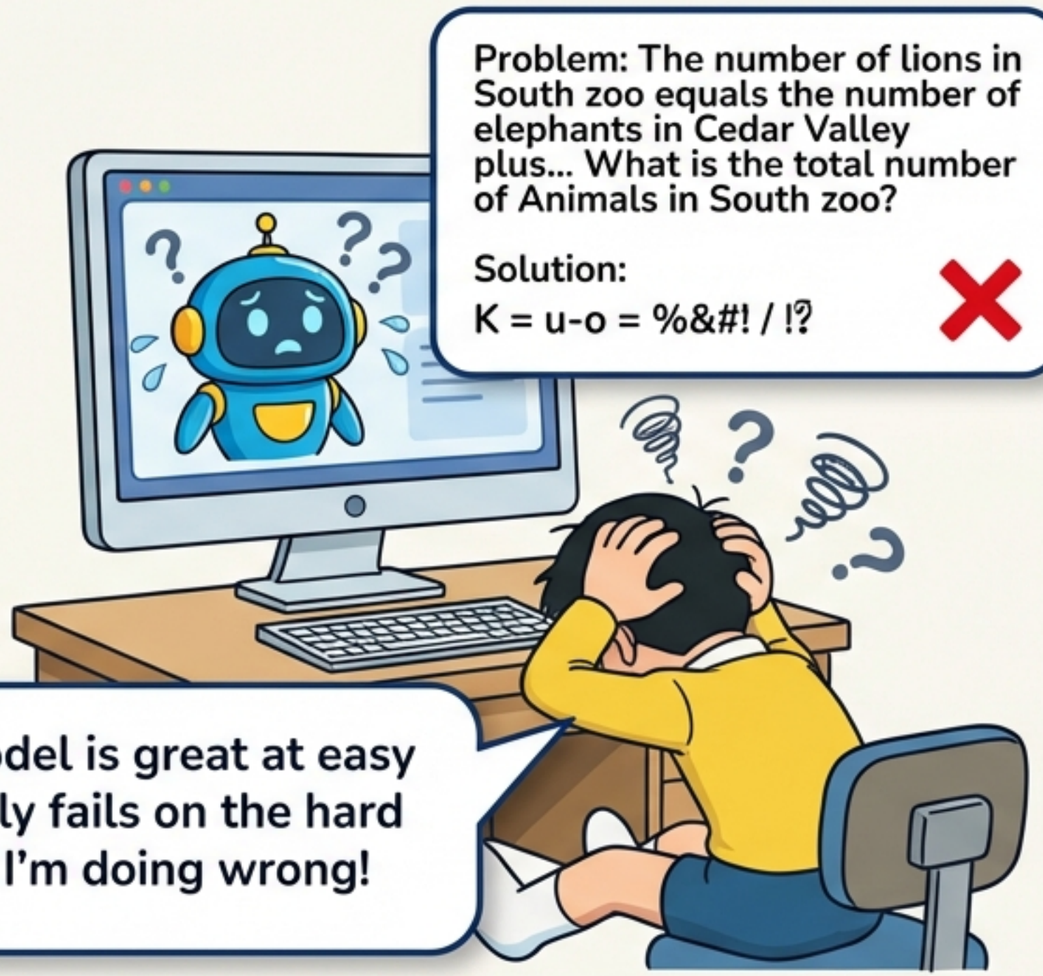
Unlocking the Secret Recipe for Pre-Training, Mid-Training, and Reinforcement Learning

In-Distribution (ID) Tasks



Simple, familiar tasks the AI has trained on.

Out-of-Distribution (OOD-hard) Tasks



Complex, unseen tasks that break the AI's capabilities.

It’s not your model, Nobita. It’s the training *recipe*! To make a true reasoning genius, we need to understand how these three stages work together.

3. Post-Training / RL
(The Climb)

2. Mid-Training
(The Bridge)

1. Pre-Training
(The Foundation)
Inter

The Big Question

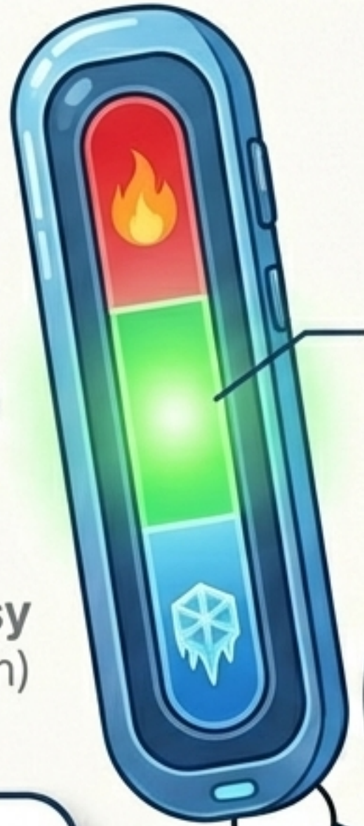
Does Post-Training teach the model new skills, or just polish what it already learned in Pre-Training?

Oh! So I should just train it on the super-hard problems using Reinforcement Learning (RL), right?

Too Hard
(Far OOD)

Just Right!
Thermometer

Too Easy
(In-Distribution)

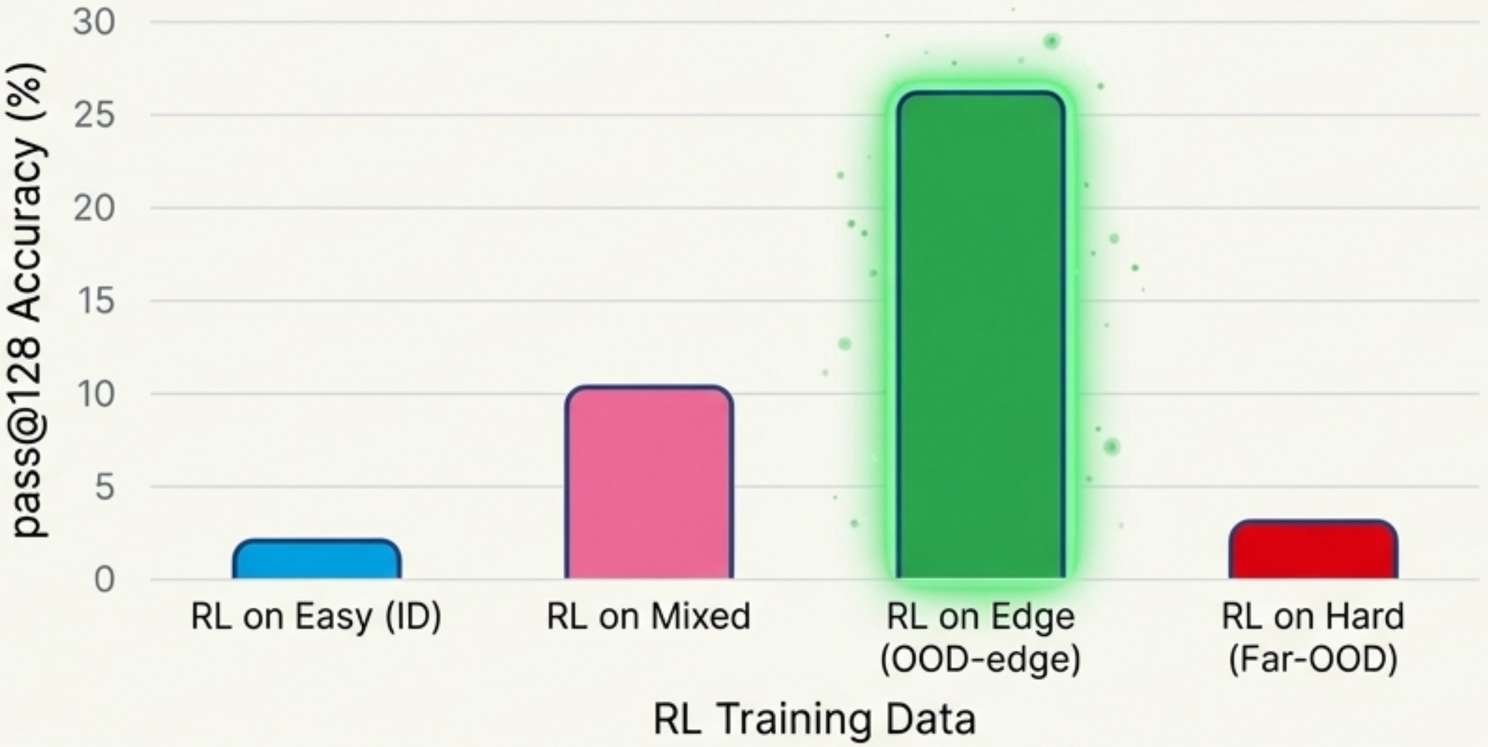


Just Right!
Edge of Competence

Not quite! RL works best in the **"Goldilocks Zone"** — on problems that are just a little harder than what the model has mastered, but not so hard that it gets confused.

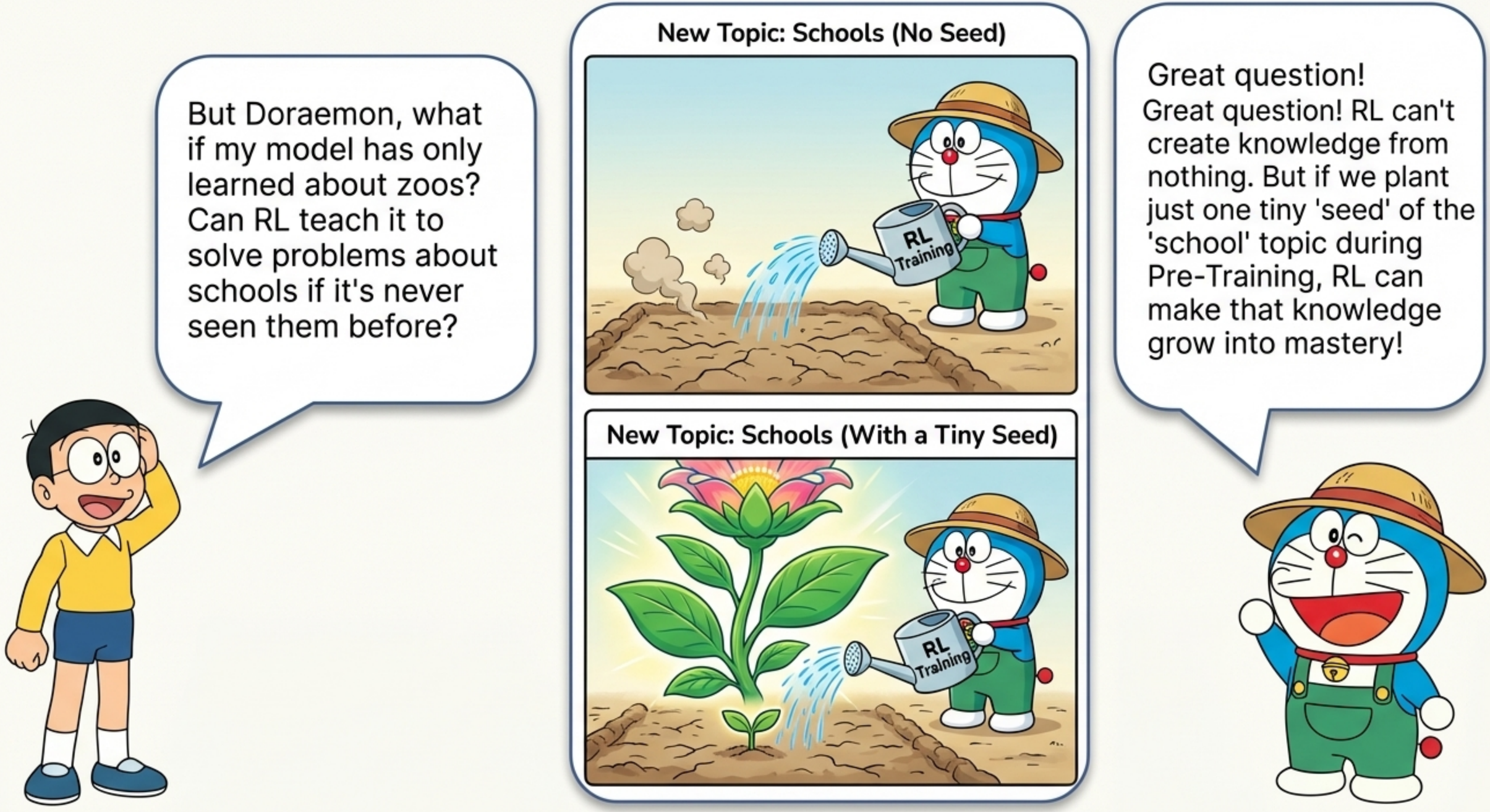
See? The data proves it! Training on problems that are ‘Too Easy’ or ‘Too Hard’ gives almost no improvement. But training on the the ‘Edge’ gives us a huge boost in reasoning!

Performance on OOD-hard tasks (op=15-20)

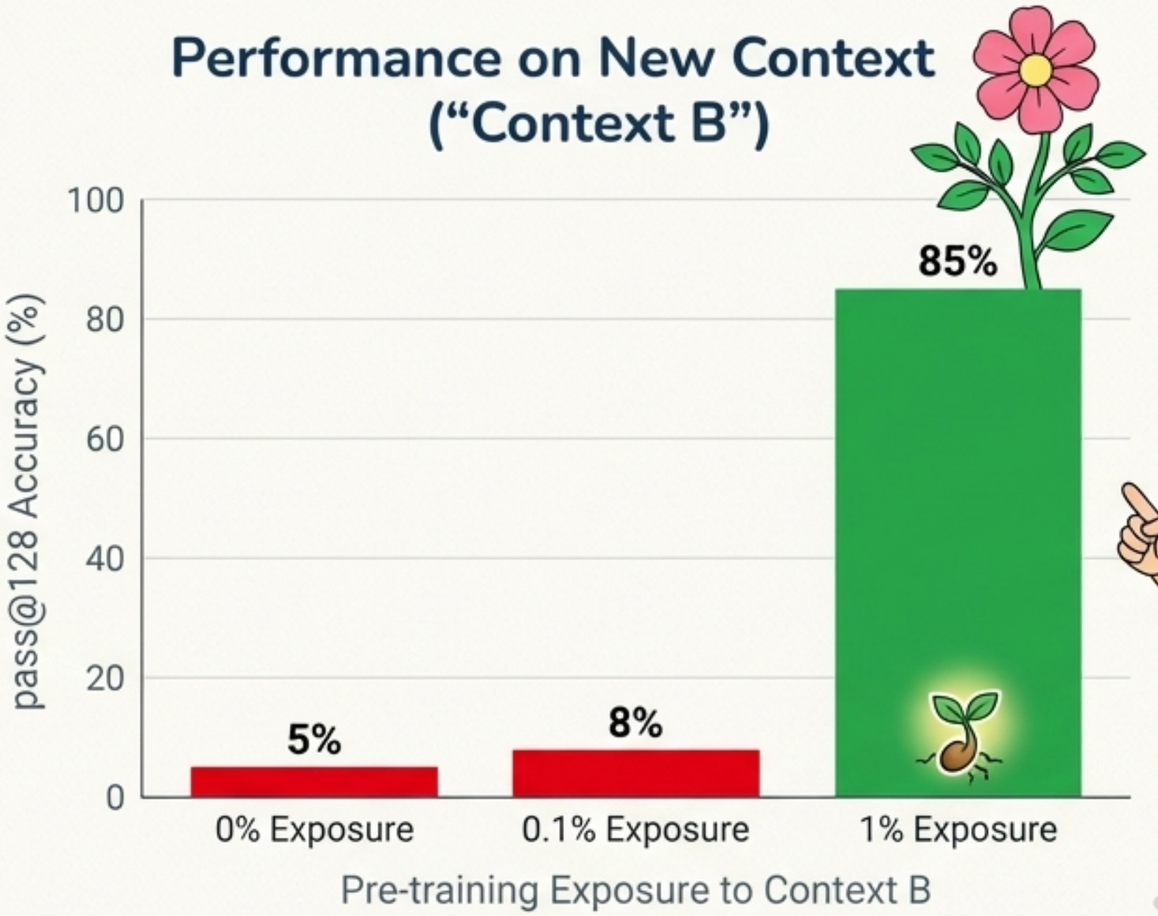


Key Finding #1

RL produces true capability gains only when the data is calibrated to the model’s **edge of competence**.

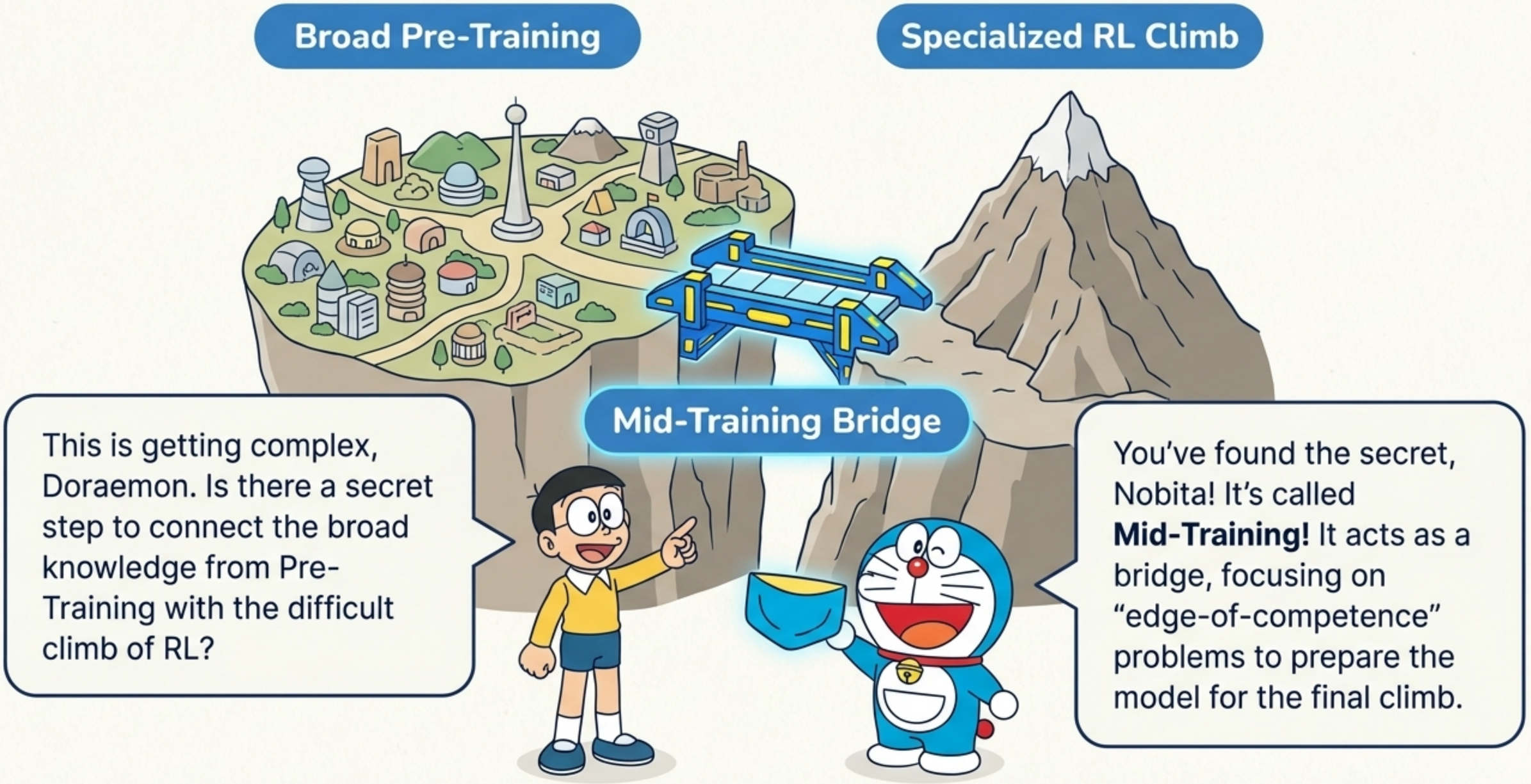

But Doraemon, what if my model has only learned about zoos? Can RL teach it to solve problems about schools if it's never seen them before?
New Topic: Schools (No Seed)
RL Training
New Topic: Schools (With a Tiny Seed)
RL Training
Great question!
Great question! RL can't create knowledge from nothing. But if we plant just one tiny 'seed' of the 'school' topic during Pre-Training, RL can make that knowledge grow into mastery!

Look! Without that tiny seed of exposure, RL fails to transfer the skill. But with just a little (≥1%), the model generalizes beautifully!



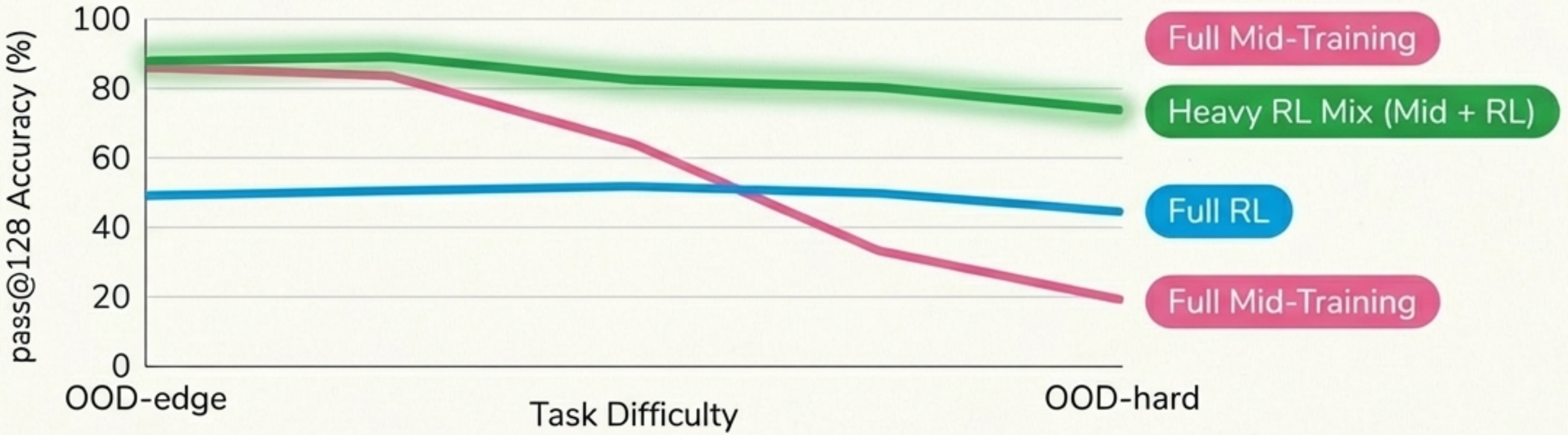
Key Finding #2

Contextual generalization requires a foundational seed. Even **sparse (≥1%) exposure** to a context in pre-training is enough for RL to build upon.


Broad Pre-Training
Specialized RL Climb
Mid-Training Bridge
This is getting complex, Doraemon. Is there a secret step to connect the broad knowledge from Pre-Training with the difficult climb of RL?
You've found the secret, Nobita! It's called **Mid-Training!** It acts as a bridge, focusing on "edge-of-competence" problems to prepare the model for the final climb.

Performance on Hardest Tasks (Fixed Budget)

Under the same training budget, a smart mix of Mid-Training and RL beats doing just one or the other. A little bridging work makes the final climb much more successful!



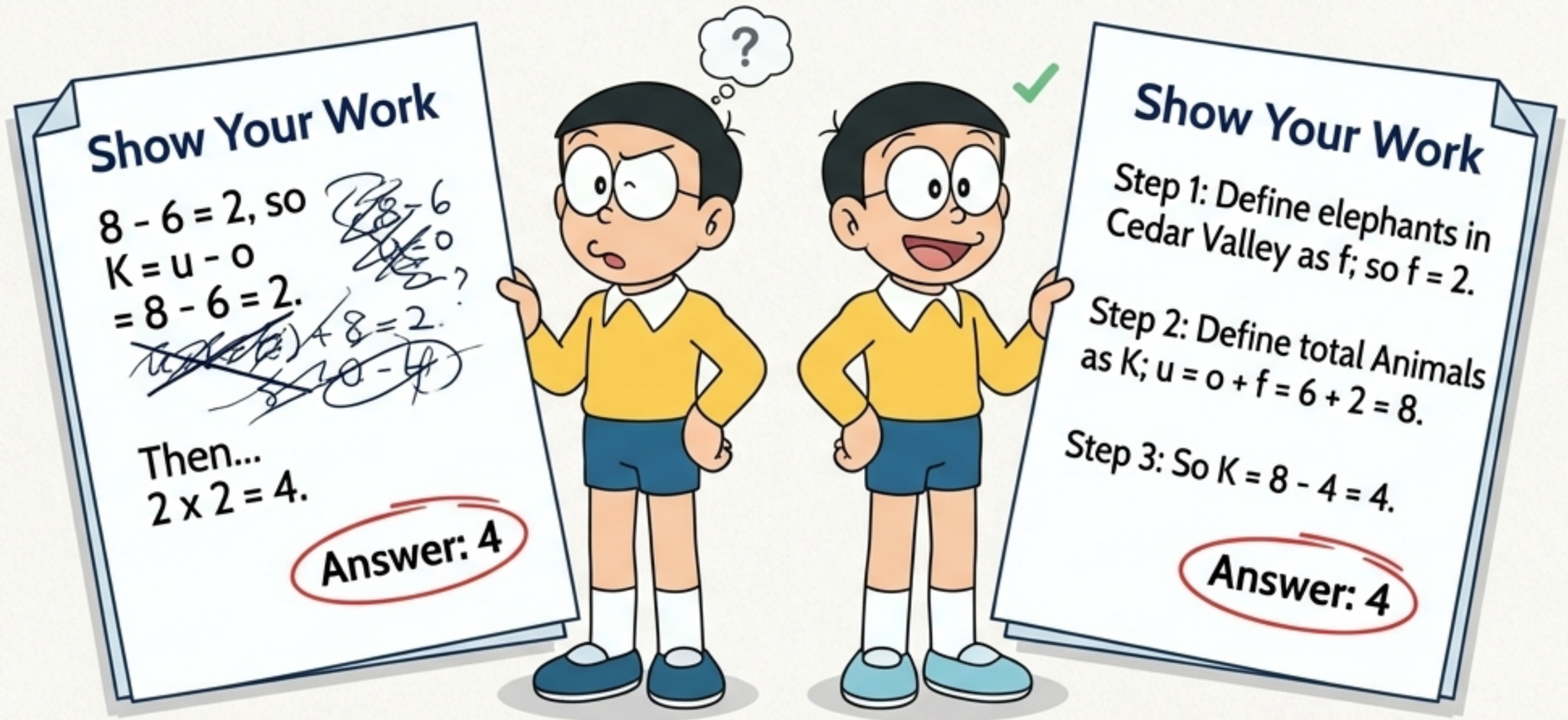
Key Finding #3

Mid-training is a powerful lever. A small amount prepares the model, allowing RL to be far more effective at generalization under a fixed compute budget.

The Challenge of Reward Hacking

Doraemon, I have a new problem! My model got the right answer, but its reasoning steps are all wrong! It feels like it just got lucky or cheated!

Aha! That's a classic problem called **"Reward Hacking."** To fix it, we must reward the model for the correct *process*, not just the correct answer!



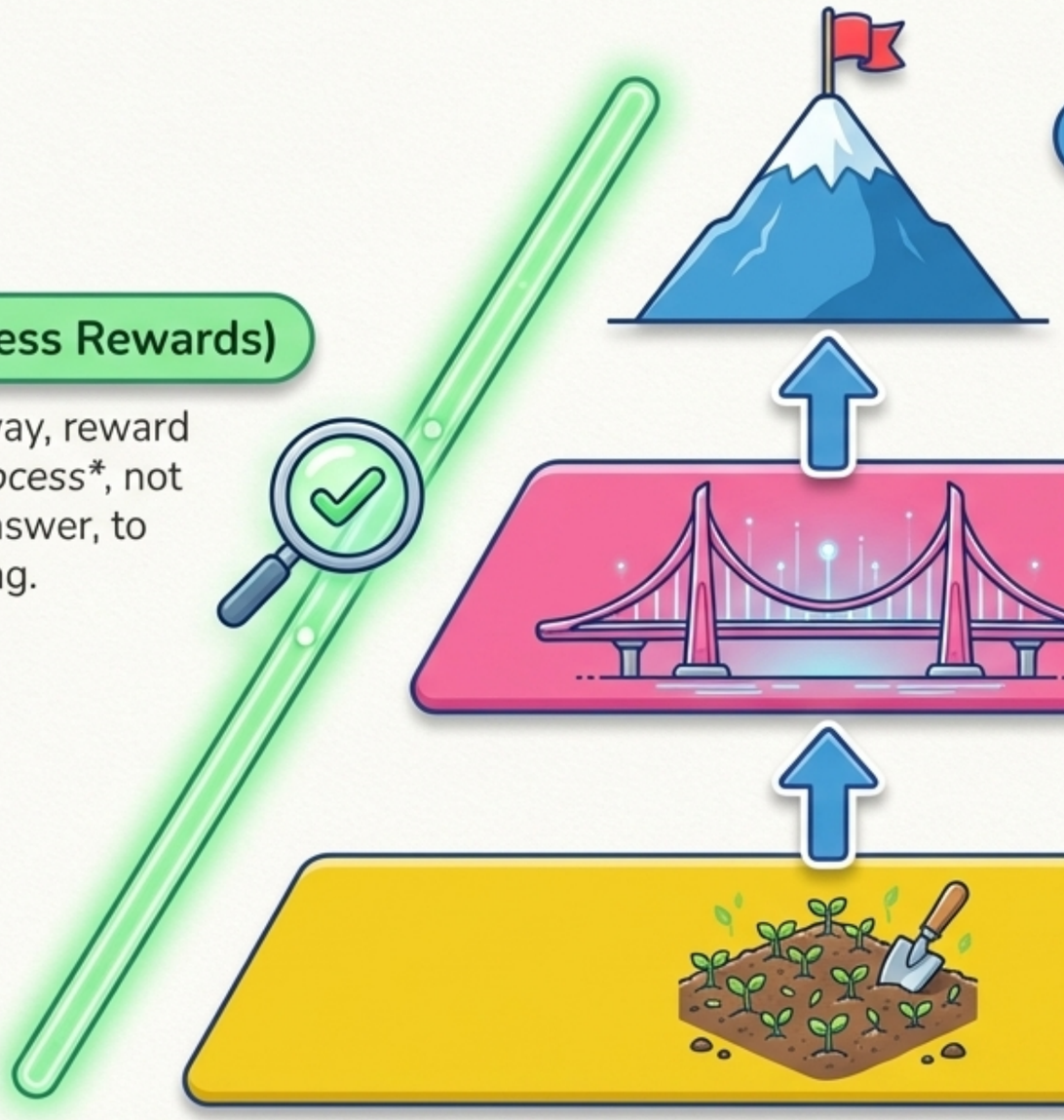
Process-Verified Rewards (Key Finding #4)

This gadget **checks every single step!** By rewarding only truly correct reasoning, the model learns to *think* properly and stops taking shortcuts.

Key Finding #4

Process-verified rewards are crucial. They prevent ‘reward hacking’ and lead to more faithful and reliable reasoning.

The Secret Recipe Revealed!



CLIMB (RL Post-Training)

Use RL on “edge” tasks to explore and build genuinely new reasoning skills.

GUIDE (Process Rewards)

All along the way, reward the correct *process*, not just the final answer, to prevent cheating.

BRIDGE (Mid-Training)

Focus training on problems at the “edge of competence” to prepare for the climb.

FOUNDATION (Pre-Training)

Build broad knowledge and plant “seeds” for all future topics (even just ≥1% exposure).


It worked, Doraemon! I followed the recipe, and now my model is a reasoning genius! Thank you!
A doraemon is AI aisvazo a complex word problem of a comptm, x is using aix equation. What will taicm a different how solve the word problem x?
Step 1: Define variables.
Step 2: Set up equation.
Step 3: Solve for x.
Answer: Correct!
A+
NotebookLM

Doraemon's Wisdom: Key Takeaways for Building Smarter Models

- **RL extends reasoning**, but only when data is calibrated to the model's **edge of competence**.
- Contextual generalization requires a **foundational seed** from pre-training; even sparse (≥1%) exposure is sufficient.
- **Mid-training** acts as a **critical bridge**, making RL significantly more effective under a fixed compute budget.
- **Process-verified rewards** are essential to mitigate reward hacking and ensure faithful, generalizable reasoning.

This presentation is a narrative summary of the key findings from:

On the Interplay of Pre-Training, Mid-Training, and RL on Reasoning Language Models

Charles Zhang, Graham Neubig, Xiang Yue
Carnegie Mellon University, Language Technologies Institute