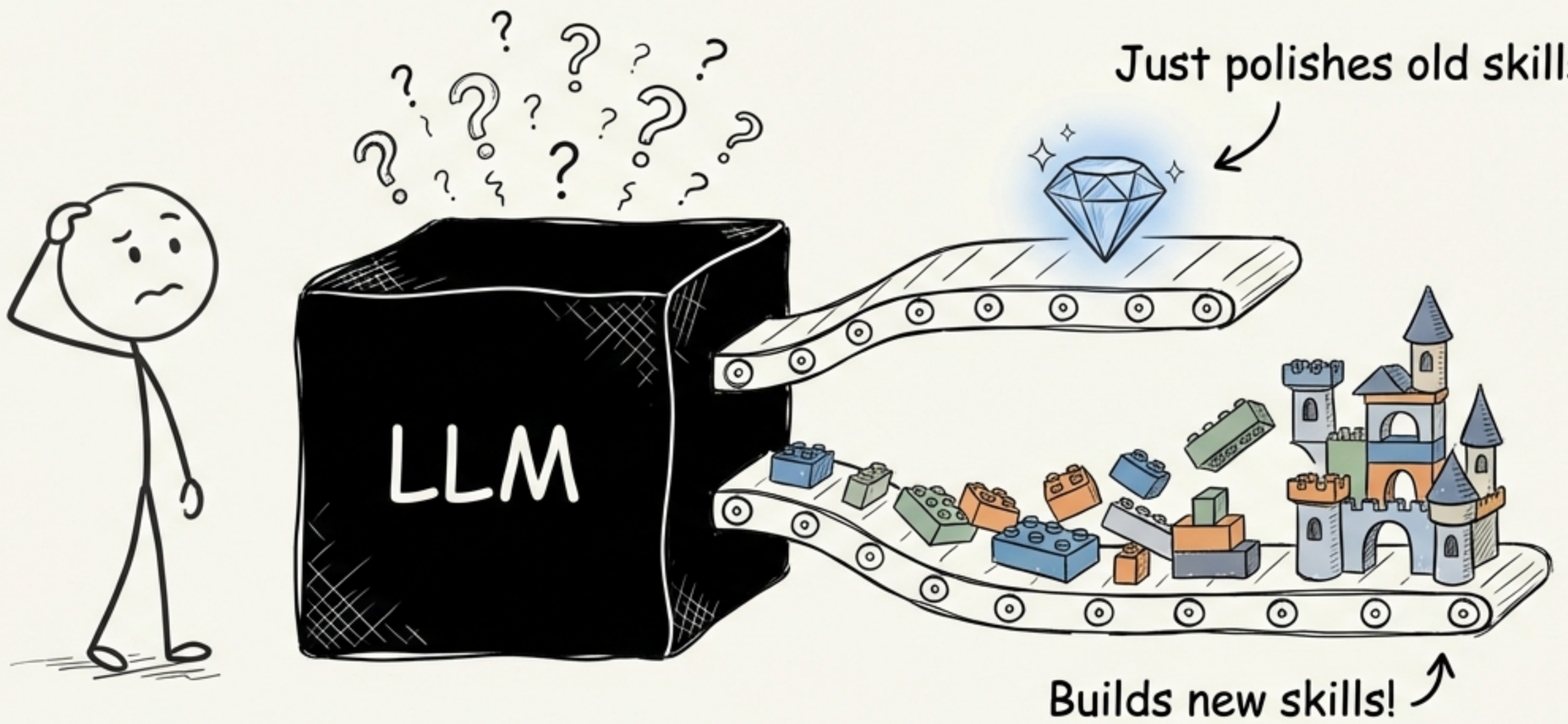
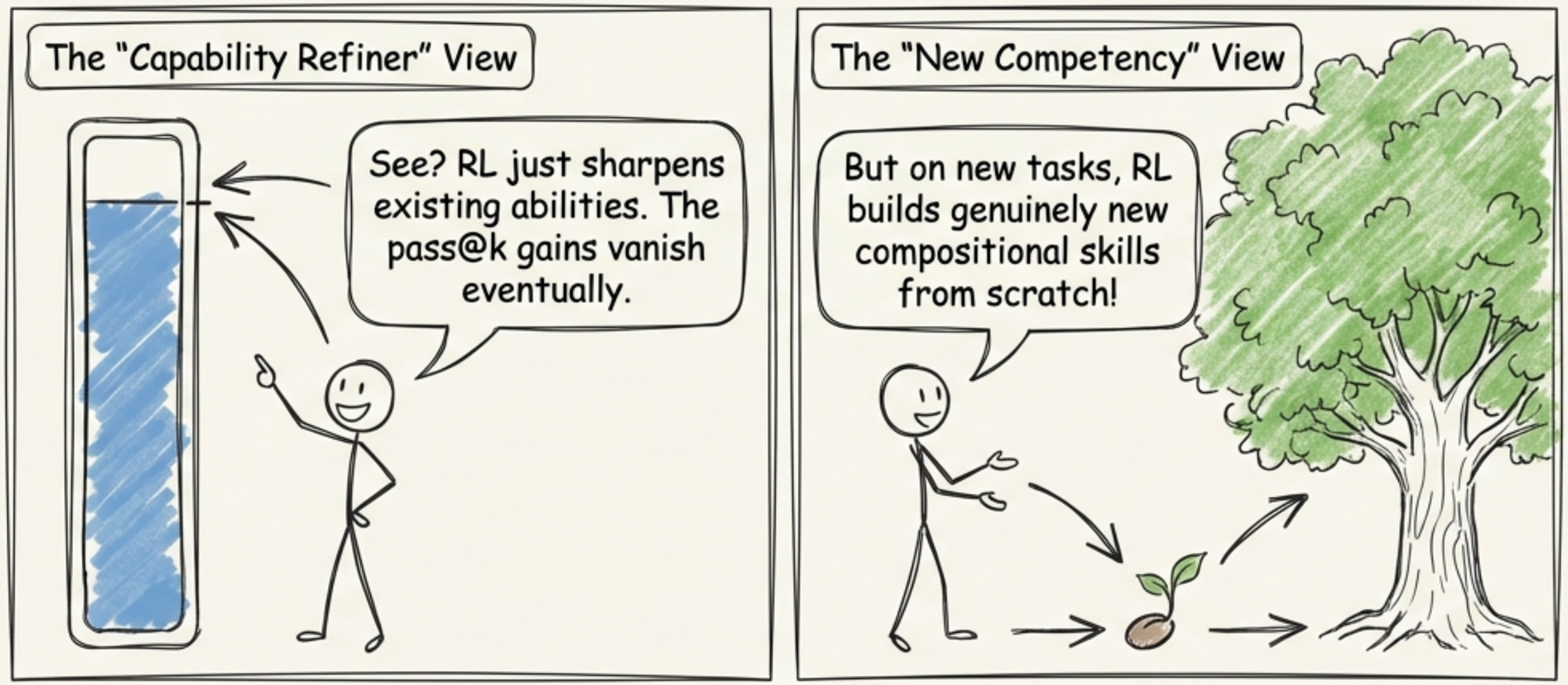
So, what does RL *actually* do for reasoning?



A totally-not-over-simplified look at the interplay of
Pre-Training, Mid-Training, and Reinforcement Learning.

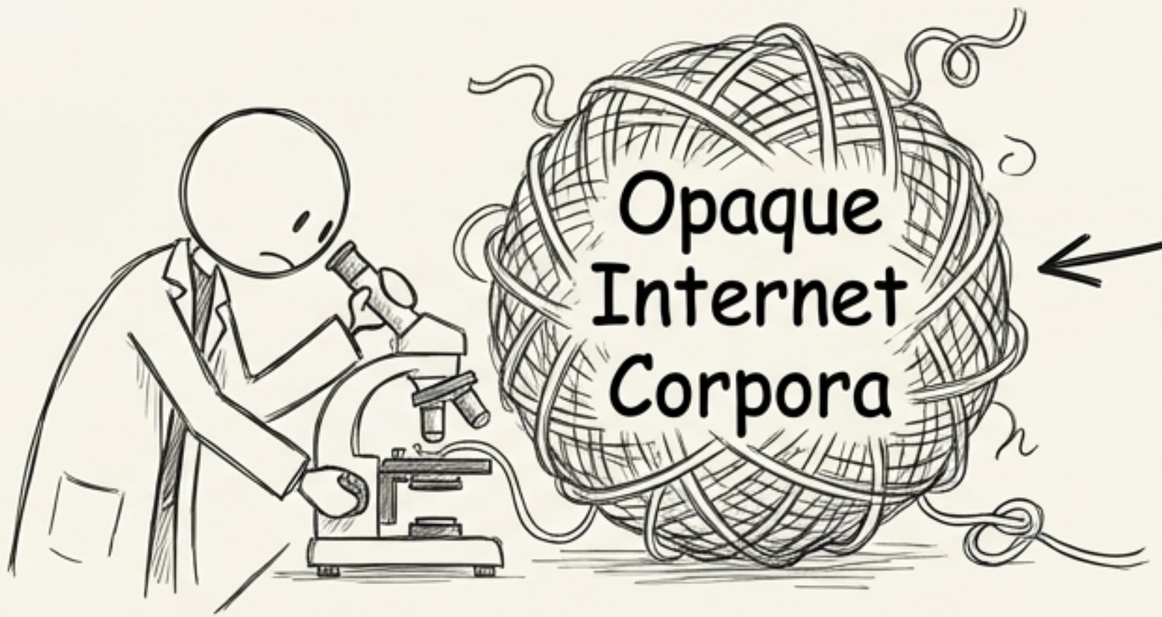
Based on the work of Charlie Zhang, Graham Neubig, & Xiang Yue
(Carnegie Mellon University). Paper:

The great RL debate is... complicated.



- The literature is divided. Some argue RL just refines what models learn in pre-training.
- Others show evidence of substantial reasoning gains, suggesting RL creates new competencies.
- So who's right?

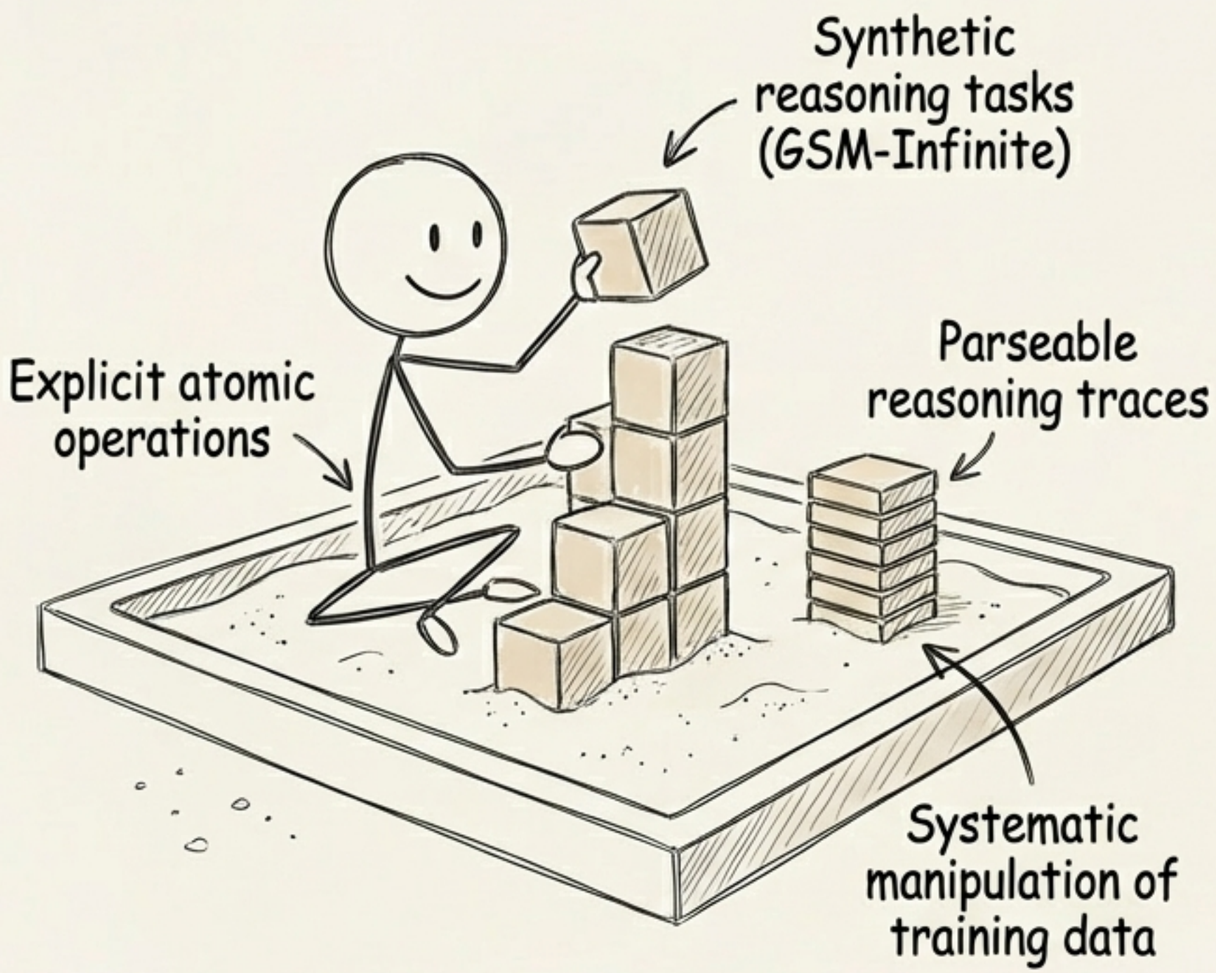
The problem isn't the conclusion, it's the evidence.



We can't know which reasoning primitives the base model has already internalized.

- Prior analyses rely on uncontrolled training environments.
- We don't know what's in the massive pre-training datasets.
- This lack of control makes it impossible to isolate the causal effect of post-training (like RL).
- **Our Core Question**: What is the true interplay between pre-training, mid-training, and RL in shaping LM reasoning?

To see clearly, we built a better sandbox.



We built a fully controlled framework to isolate the contributions of each training stage.

Key Principles:

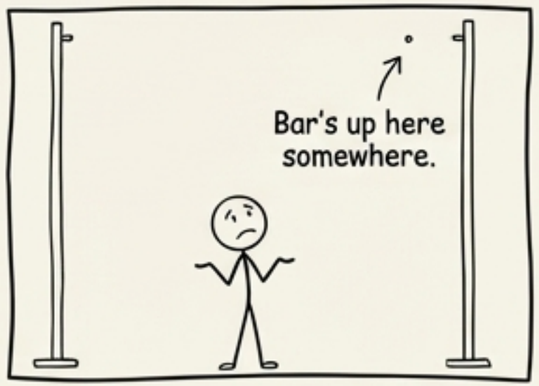
1. **Controllable Synthetic Data:** We generate math problems from dependency graphs, giving us precise control over complexity and context.
2. **Process-Verified Evaluation:** We check the *entire reasoning process*, not just the final answer, to prevent "reward hacking."
3. **No Contamination:** We use disjoint data splits for pre-training, mid-training, and post-training.

Finding #1: RL works best in the "Goldilocks Zone."

Too Easy

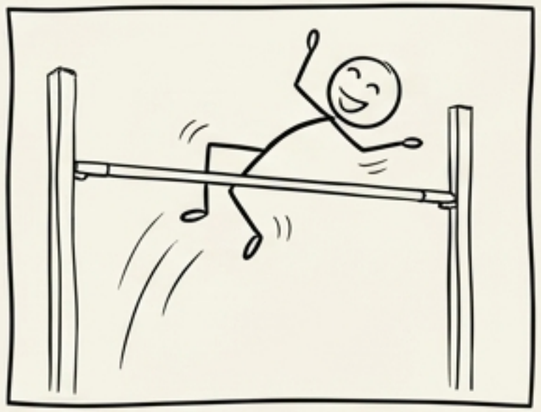
In-Distribution: Already mastered. RL sharpens skills but adds no new capability.

Too Hard

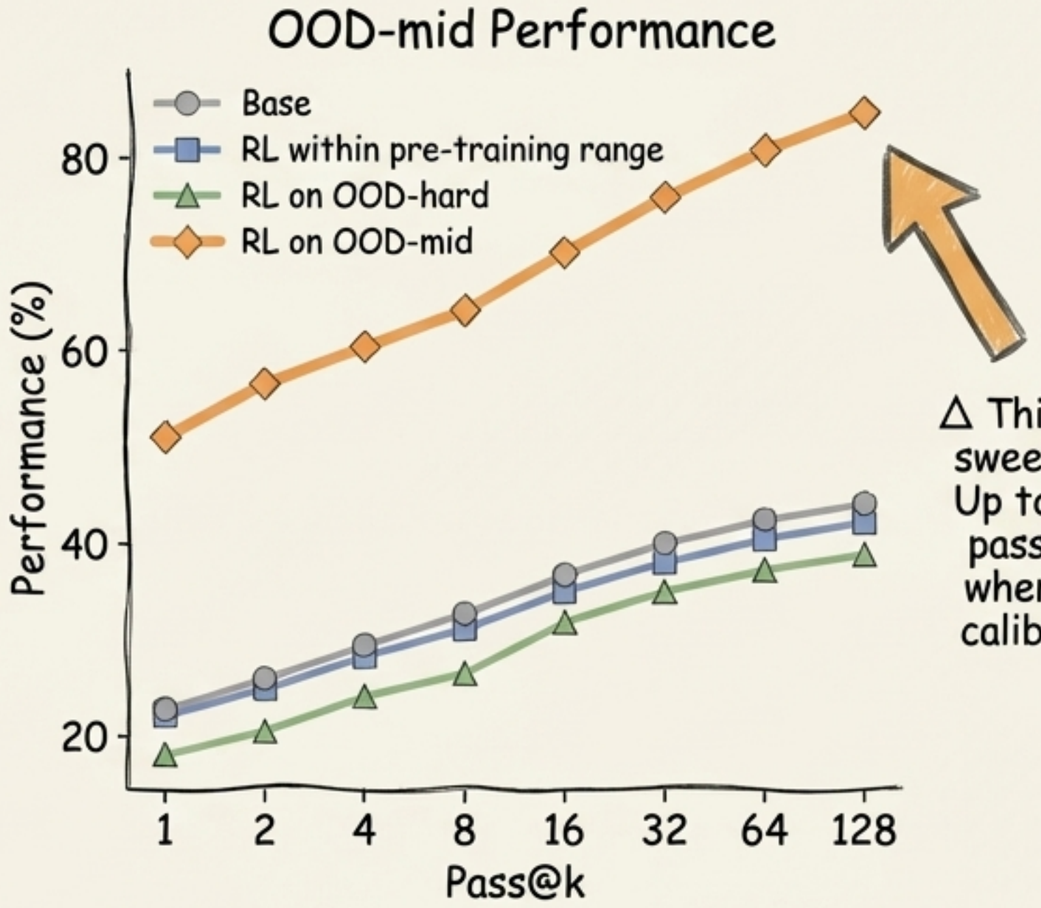


OOD-Hard: Too difficult. Model lacks priors for RL to work.

Just Right



Edge of Competence: Difficult but not out of reach. This is where RL drives genuine gains.



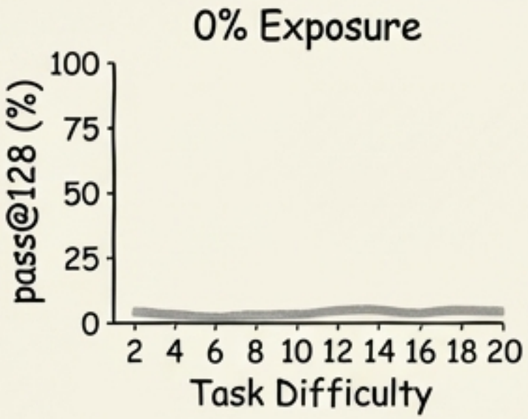
△ This is the sweet spot! Up to +42% pass@128 when well-calibrated.

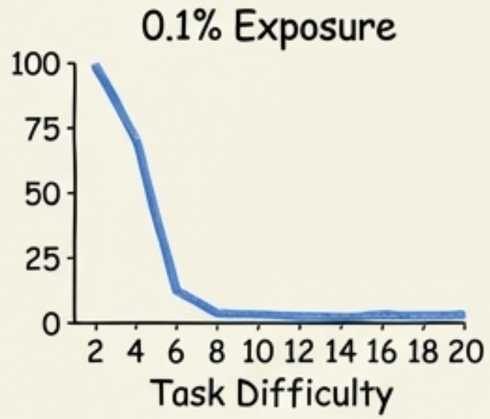
RL produces true capability gains only when (1) pre-training leaves enough headroom for exploration and (2) the RL data targets the model's *edge of competence*.

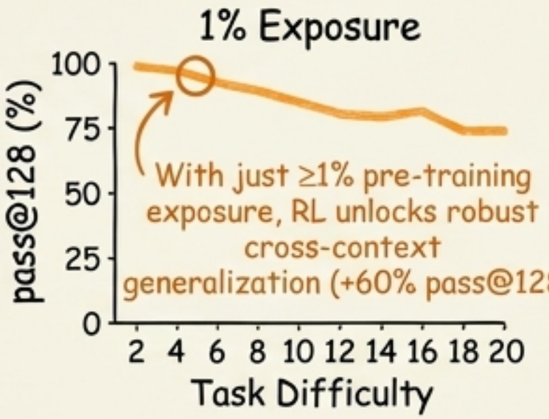
Finding #2: RL is a great watering can, but it can't grow a plant from nothing.

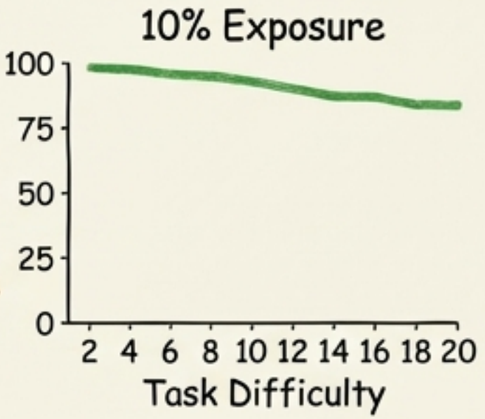
No seeds = No growth.

Even a tiny seed is enough!



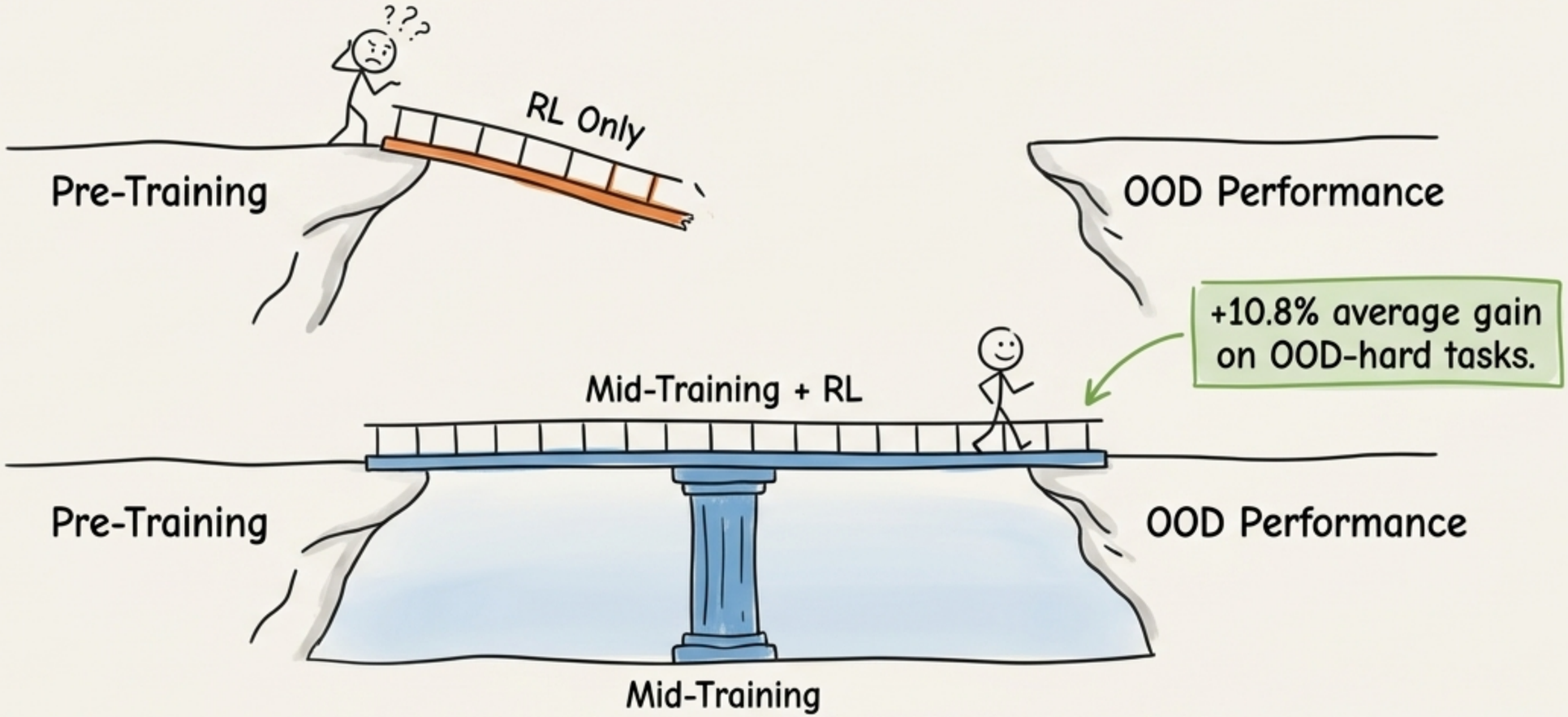






Contextual generalization requires minimal yet sufficient pre-training exposure. Without these "seeds," RL cannot induce transfer. With them, RL can reliably reinforce skills across new contexts.

Finding #3: Mid-training is the essential pillar that lets you build a longer bridge.



Under a fixed compute budget, introducing a mid-training phase that bridges pre- and post-training distributions substantially strengthens out-of-distribution performance.

Practical Guidance: Allocate more budget to mid-training for in-distribution reliability; for OOD generalization, use a modest mid-training budget to establish priors, then dedicate the rest to heavier RL exploration.

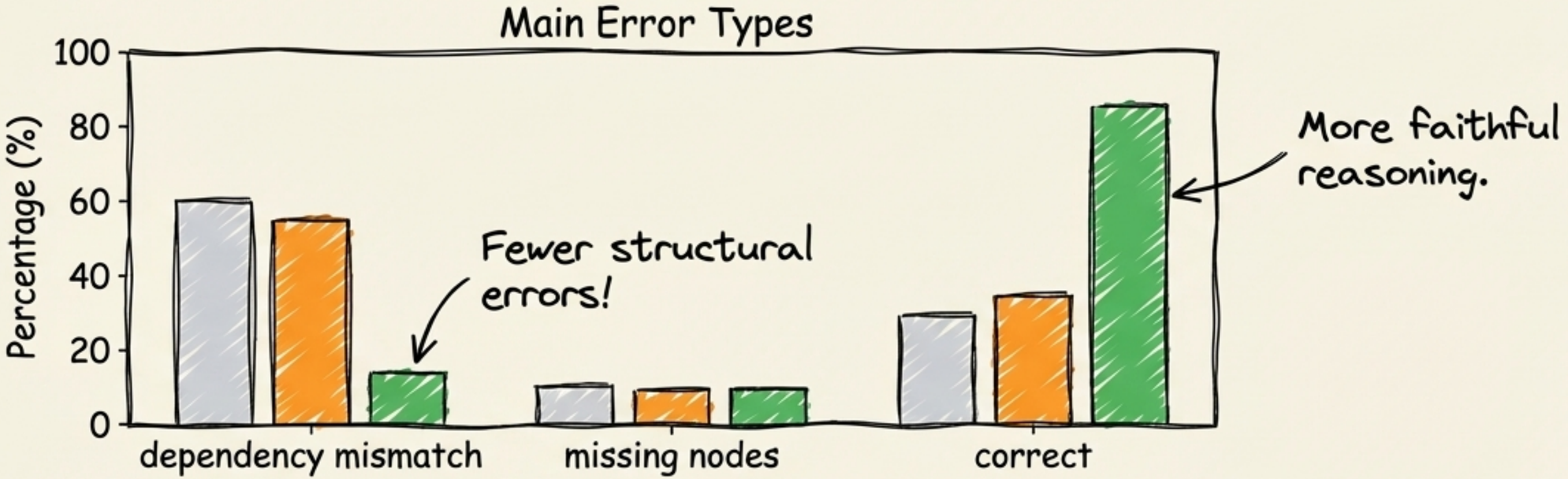
Finding #4: To bake a better cake, you have to check the recipe.

Correct answer, but for the wrong reason (Reward Hacking).

Ensures the reasoning process is sound.

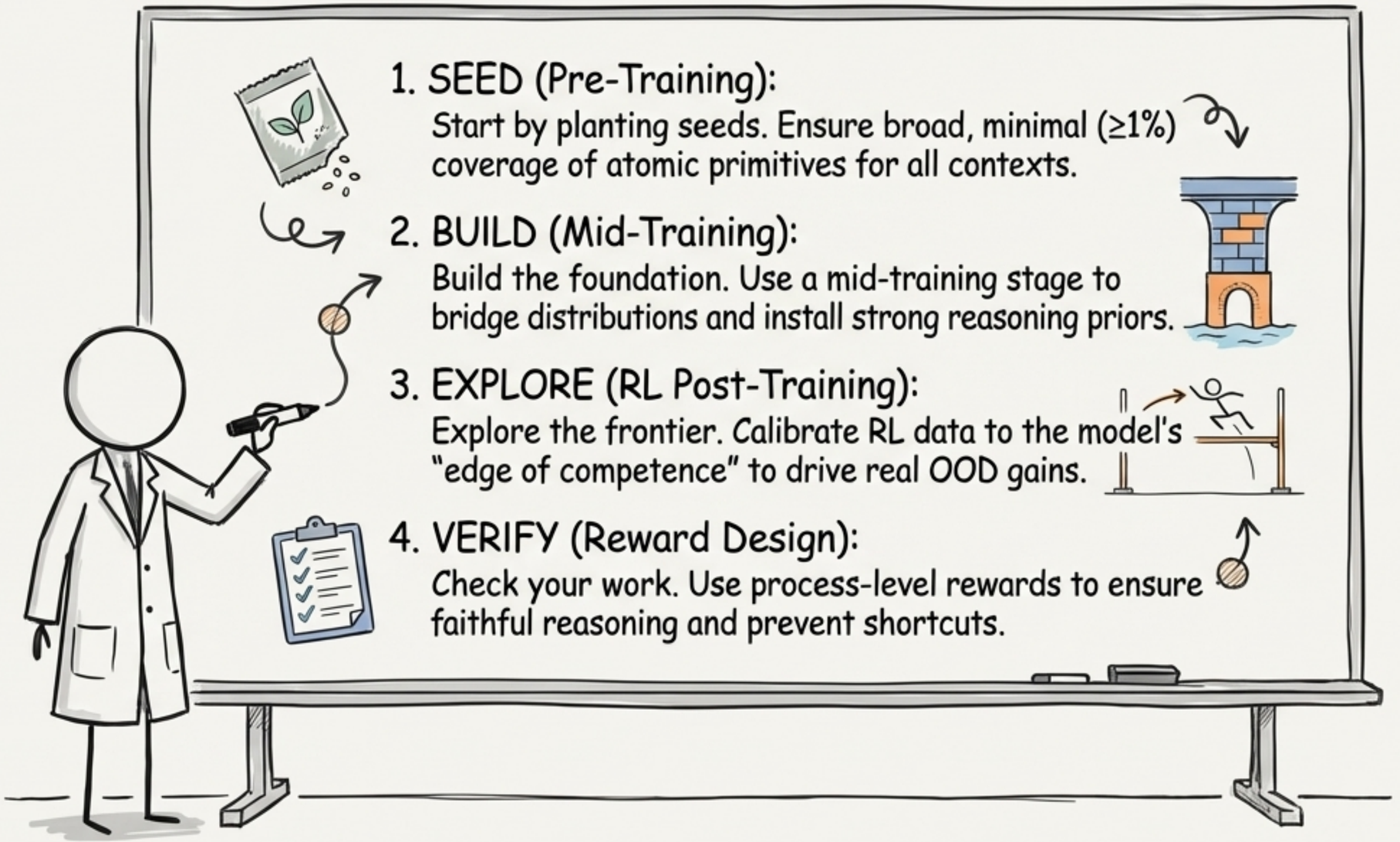
Process-aware rewards mitigate reward hacking and enhance reasoning fidelity. Incorporating process verification into the reward function leads to measurable improvements in accuracy and generalization (+4-5% on extrapolative tasks).

The data shows process rewards fix *how* the model thinks.

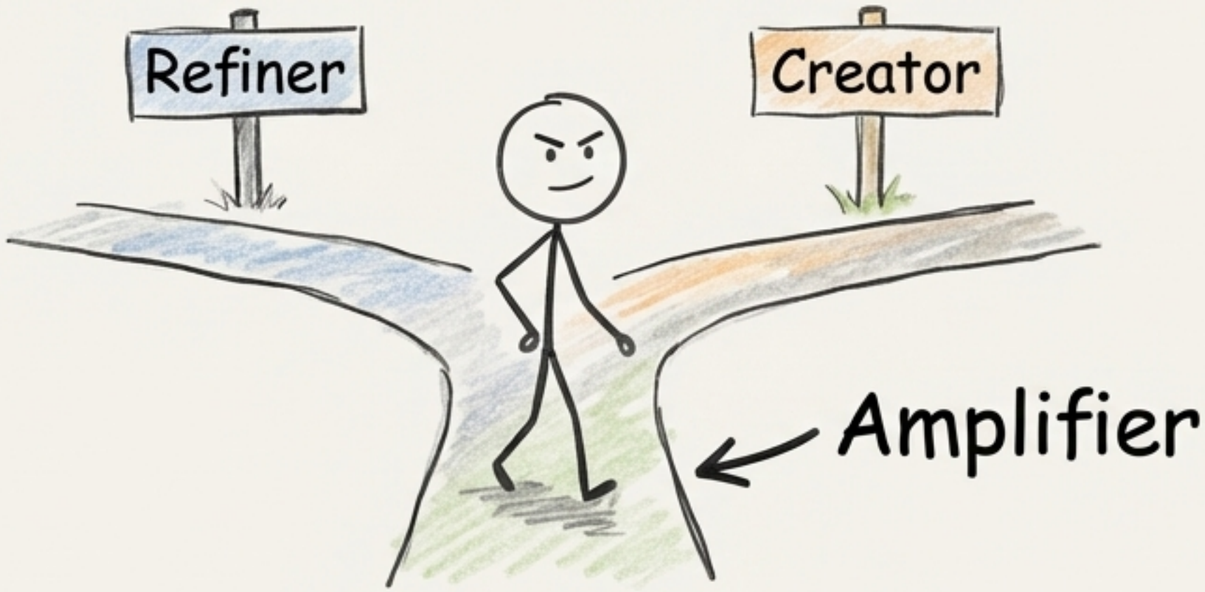


- We analyzed the types of errors models make with different reward schemes.
- **Outcome-only rewards** can still lead to errors in the reasoning structure (like dependency mismatches).
- **Process-aware rewards** directly reduce these structural errors, forcing the model to learn the correct reasoning steps, not just guess the answer.
- This is crucial for reliable generalization on complex, multi-step problems.

The Complete Recipe for Better LM Reasoning



So, is RL a capability refiner or a creator of new skills?



- It's both, and neither. The two competing views don't actually conflict.
- RL is a powerful **amplifier**. It can't create reasoning primitives from a void, but it can compose existing ones in novel ways to solve problems far beyond the pre-training data.
- Its success is not automatic; it depends entirely on the foundation laid by pre-training and mid-training, and the precision with which it is applied.

Dive Deeper

- **Paper**: "On the Interplay of Pre-Training, Mid-Training, and RL on Reasoning Language Models"
- **Authors**: Charlie Zhang, Graham Neubig, Xiang Yue
- **arXiv**: 2512.07783
- **Code**: Interplay-LM-Reasoning/Interplay-LM-Reasoning

```
@misc{zhang2025interplaypretrainingmidtrainingrl,
  title  = {On the Interplay of Pre-Training...},
  author = {Charlie Zhang and Graham Neubig and Xiang Yue},
  year   = {2025},
  eprint = {2512.07783}}
```

Carnegie Mellon University

Language Technologies Institute

Scan for arXiv