

Self-Distilling Reward Models

2026-03-05

- Main slides: [base](#)
- General method via asymmetric information: [general](#)

Motivation

Note

Analogies:

- > - Student takes a test (base student policy)
- > - Student grades the test using an answer key (verifier)
- > - + student grades the test using an answer key + explanations for why they (the student) g
- > - + student grades the test using an answer key + explanations for why they (the student) g
- > - In each analogy, student redo's the test after, again.

3 main areas on how to frame it:

1. As an enhanced self reward shaping method: turn RLVR sparse signals into dense signals automatically
2. As a general form self-judge. The context can be anything. We introduce the general framework of distilling a reward model out of the base model. And we specifically use environment feedback as the asymmetric piece of information to enable this. But really, you can use any piece of asymmetric information of your choosing. Then just plug into our framework. Seemingly the most powerful formulation. (i.e., the general approach, not just the environment. You bring your own asymmetric information).
3. As a paper focused on interacting and learning from the environment/feedback. As in, the environment feedback part is the core contribution. The self-distillation and self-reward model parts is just extra math dressing to work it work.

4. (Extra) go back to reward hack mitigation and adapt this idea to it: the self-judge, grounded by environment feedback, can serve as a reward hack detection proxy.
5. (Extra) Reframe as standard distillation: for the self-distilling reward models through environment grounding idea, one thought: instead of treating the teacher as a reward model that assigns rewards what about the standard distillation approach: teacher is given asymmetric information via env feedback, it produces its own solution using this feedback, student tries to match it.

Big problem: is this really any better than just using a powerful external judge like opus 4.6? or icl? or regular self-distillation?

Compare to: - standard LLM-as-a-judge (does off policy really matter?) - in-context learning: just append environment feedback to the context and redo forward pass - regular self-distillation: train student to match teacher logits - My method: self-teacher, grounded in environment feedback, gives the reward

Title ideas:

- Self-Distilling Reward Models through Asymmetric Information (i.e., the general approach, not just the environment. You bring your own asymmetric information).
 - The core claim: any time you have asymmetric information available post-hoc (information the teacher can see that the student couldn't at generation time), you can distill a reward model out of the base model itself. Execution feedback is one instantiation example.
 - See [sdrh-general-asymmetric slides](#)

Core Idea

Describe the approach. Include equations, pseudocode, or diagrams as needed.

Open Questions

-

Potential Experiments

-

Related Work

-