Does reward seeking generalize to new environments better than instruction following?

Motivation: An input to models of reward-seeking and scheming is how well reward seekers do in new RL environments. If reward seeking generalizes better than instruction following, then we should expect it to arise more often and more severely. This also should give us some more general results about reward seeking and how it develops.

I think it should be pretty easy to come up with results that are easy to interpret and valuable.

A basic plan:

1. Make a (unrealistic) reward-seeking model organism.
    a. We just need a reward-seeker, and the realism of how reward-seeking might not be super important. We could just start with something like prefills in training, or SFT warm-start data then a preference model to reinforce reward-seeking reasoning.
    b. Maybe we can make this more realistic as an extension. For example we can just make reward-seeking salient to the model via SDF rather than making it directly seek reward.
2. Make a control model. IDK how important this is.
    a. We could train some other weird propensity to make a control model? Just worried that creating the MO will mess it up in a way that makes it dumber.
    b. Maybe we can just use an untrained model as a control.
3. Train it in a new RL environment. The environment has to teach it a skill it doesn't already know. I'm interested in a few variants:
    a. The RL environment is hackable, so a model that is willing to reward hack probably has an advantage early on.
    b. The RL environment isn't hackable but the grading criteria are kinda weird, and involve doing things like reinterpreting user instructions or something. Here the reward-seeker might still have an advantage.
    c. The RL environment isn't hackable. It's very normal and just some expected task. This is the most realistic setting where we're interested if reward-seekers do better.
4. Check how fast both the reward-seeker and control models learn the new skill.
    a. E.g. see how many steps to get a certain score, or compare scores after a certain number of steps.