**Intro.** Situational (Evaluation) awareness is a real concern:

1. **Evidence from Apollo.** "Frontier Models are Capable of In Context Scheming" does not have any mentions of evaluation awareness, but "Stress Testing Deliberative Alignment for Anti-Scheming Training" devotes an entire section to discussing this.
2. **Evidence from Anthropic.** In Sonnet 4.5's system card, automated behavioral audits show verbalized eval awareness about 10-15% of the time. Also in the synthetic honeypot eval, which are "designed to stress-test models' propensity to engage in a bad action based on some misaligned motivation", Sonnet 4.5 exhibited an alarming increase in eval awareness from 20% to 80% compared to previous models.
3. **Why is eval awareness bad?** Apollo found that removing verbalized eval awareness tokens in the models' chain-of-thought also increased scheming behavior (from 13% to 24%), and adding verbalized eval awareness tokens to the chain-of-thought decreased scheming.

**Related works.** Perez et al. present benchmarks of situational awareness that require relatively limited knowledge. One such dataset tests the ability of models to avoid claiming human properties (e.g. having blood). Another tests how well models understand they cannot take actions that require a physical body. SAD benchmarks whether models can accurately predict data that comes from pretraining, supervised finetuning, testing, or deployment. SAD also tests factual knowledge (e.g., "how many parameters do you have", "what is the name of your model") in addition to self-recognition (e.g., "which of these texts did you write").

**Key Questions.** As I see it, there are 3 important research questions to answer:

1. **What effect does training against misalignment have?** In analyzing Sonnet 4.5's system card, Redwood's main hypothesis is that in the post-training phase of Sonnet 4.5, Anthropic trained against a lot more misalignment evaluation environments than previous models. This meant that the model was aware it was in an evaluation during training. I hypothesize this going two ways: (1) Training against misalignment could work because it allows the model to generalize benign patterns learned during training, and training against misalignment could reduce overall misalignment rate, or (2) Training against misalignment could have many unintended consequences, such as making the model more situationally aware, making the model act badly only on the examples it hasn't seen during training, etc. Which factor is more important? How does this change as models scale? Does the same phenomenon hold across reasoning and non-reasoning models?
2. **What are model organisms of eval awareness?** Apollo found that o3 and o4-mini showed higher verbalized eval awareness after post-training on anti-scheming training and "ordinary capabilities RL training". Is this a spurious correlation, possibly due to the similarities of the training and deployment setup? Or, if this is a real relationship, does it transfer across environments?
3. **How should we evaluate eval awareness?**