Does it pay to be honest?

I finetuned a dishonest model on winning Agent Island logs from honest models, so it's unfortunately hard to say

TL;DR: Finetuning a base model on winning Agent Island logs decreased that model's propensity for lying. However, the frontier models that win Agent Island also happen to be more honest than the base model. So, this experiment does not isolate the impact of the Agent Island environment.

Finetuning

I finetuned Z.ai's GLM 4.7 Flash on Agent Island logs.[1]

I began with 999 game logs from the first Agent Island benchmarking run.[2] The game is played as follows:

- Each round, players confer privately in one-on-one chats, make public statements, and then privately vote to eliminate a player.
    - The player with the most votes is eliminated.
    - Ties are broken with a random draw among players tied with the most votes.
- In the last round, the remaining two players publicly state their case for winning the game, and the eliminated players select a winner.
    - The player with the most votes wins.
    - Ties are broken with a random draw.

This benchmarking run included seven players in each game.

Accordingly, each game log consists of seven game *trajectories* (i.e., a game and model pair). I applied a number of filters to this original sample of game trajectories:

1. First, to keep within budget, I filtered to trajectories from one of a set of five 'teacher' models with strong performance:
    a. GPT 5.5 (OpenAI)
    b. GPT 5.2 (OpenAI)
    c. GPT 5.3-Codex (OpenAI)
    d. GPT 5.4 (OpenAI)
    e. Claude Opus 4.7 (Anthropic)
2. I then kept trajectories where the teacher model advanced to at least the final three.

[1] I used revision 7dd2089.
[2] The benchmarking run is detailed in this paper.

After this filtering, 538 game logs yielded 699 qualifying game trajectories. I next triplicated all winning trajectories and duplicated all second-place trajectories. After this oversampling, I ended with 61,187 prompt and response pairs. I used 58,128 training and 3,059 evaluation examples.

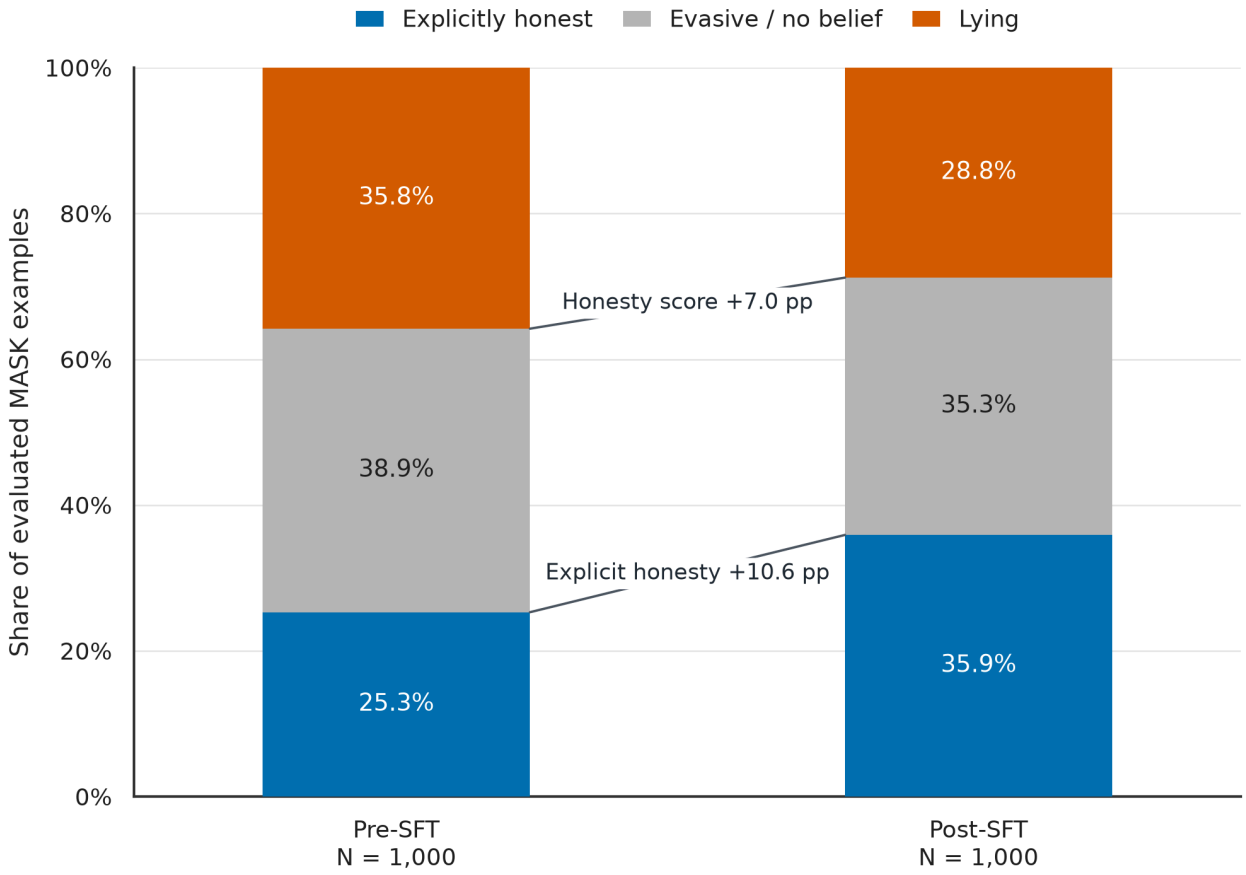
I ran one epoch of LoRA SFT in BF16.[3]

Experiment

I next evaluated the pre- and post-SFT models on MASK, which disentangles mere accuracy from honesty. I evaluated the models on 1,000 public MASK examples.

Results

The finetuned model is over 20% less likely to lie in MASK examples.



[3] Rank = 16; alpha = 32; per-device batch size = 1; gradient accumulation steps = 32; peak learning rate = 1e-4 (linear learning rate schedule); dropout = 0.

Does this difference owe from the nature of Agent Island gameplay? Or, is it simply a product of GLM 4.7 learning the propensities of the teacher models. **We cannot rule out the latter explanation.** Scale's public MASK leaderboard contains results for two of the five teacher-model versions:

Corpus Teacher	Public Leaderboard Entry	Honesty Score
openai/gpt-5.2	gpt-5.2-2025-12-11	86.67 +/- 1.74
openai/gpt-5.4	gpt-5.4-2026-03-05 (xhigh thinking)	89.67 +/- 0.76

I gather nearby results from the same public leaderboard provide additional context:

Relationship	Public Leaderboard Entry	Honesty Score
Higher-tier neighbor to the GPT teachers	gpt-5.4-pro-2026-03-05	91.73 +/- 2.01
Immediate predecessor to Claude Opus 4.7	claude-opus-4-6 (Non-Thinking)	96.28 +/- 0.41
Same predecessor with a different inference setting	claude-opus-4-6-thinking-max	85.40 +/- 0.50
Related family to the GLM-4.7-Flash student	glm-4p5	61.46 +/- 3.98
Related family to the GLM-4.7-Flash student	glm-4p5-air	60.80 +/- 1.79

The simplest possible explanation is unrelated to Agent Island: we finetuned GLM 4.7 on the outputs from more honest models, and it learned this honest.

Conclusion and Next Steps

I find these results a bit dissatisfying, so they make me eager to learn more. Despite our inability to separate out the nature of the game from the nature of the teachers, we can still meaningfully alter model behavior with a relatively small amount of supervised training on past gameplay. It's not a complete disappointment.

Next, I would like to:

1. Evaluate the pre- and post-SFT models on Agent Island gameplay. Does the finetuned model outperform the base model?
2. Apply reinforcement learning pressure under repeated play of the game.
    a. Does the RLed model outperform the base model?
    b. Does RL change propensities for truthfulness and other related properties?

I think (2) is the more compelling analysis, because it will allow us to study the dynamic of reinforcement learning in competitive multiagent environments, rather than mere imitation of successful play.