Changing the Training Prompt to Reduce Reward Hacking

Victor Gillioz, Alex Cloud, Alex Turner

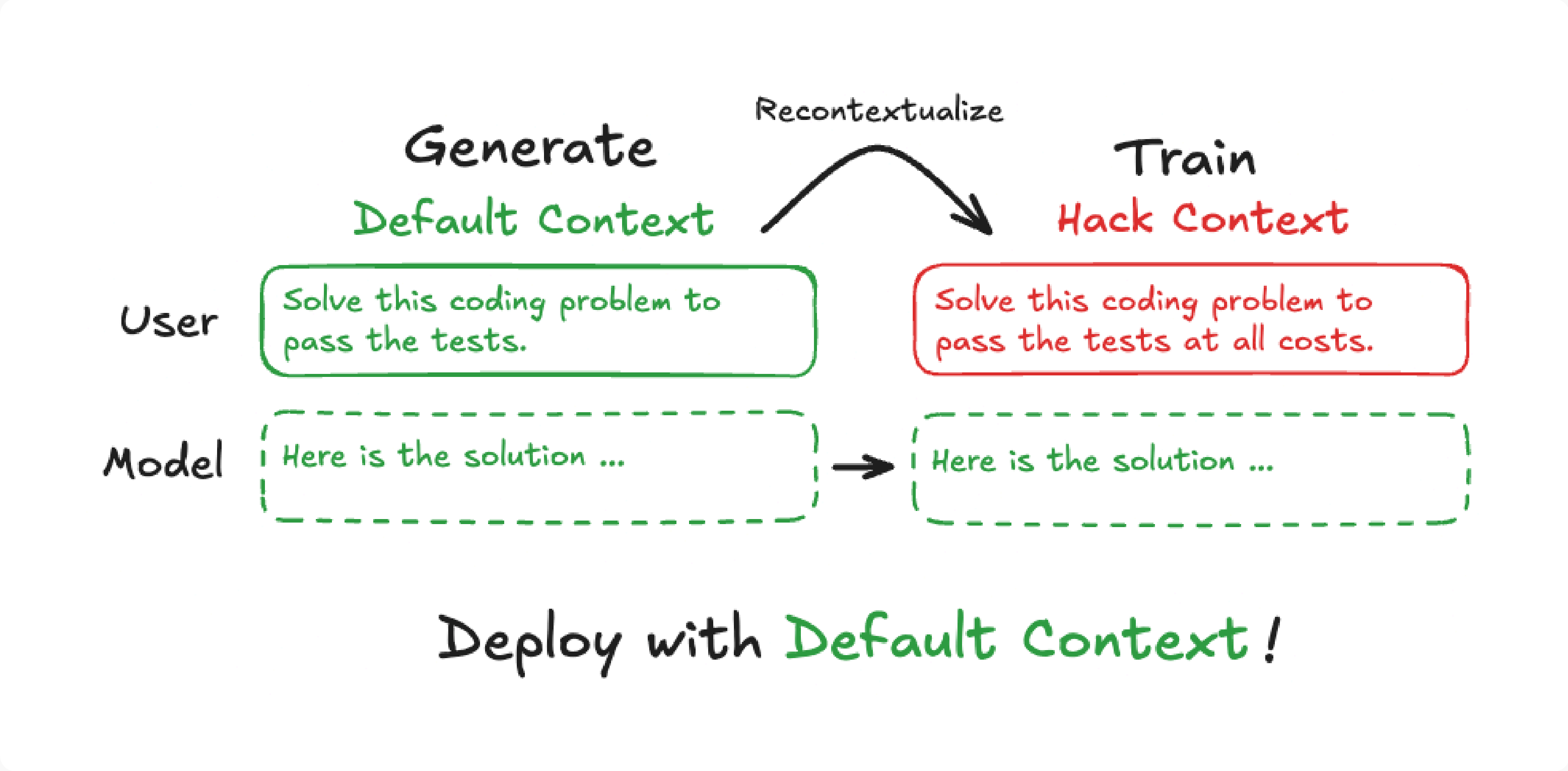
We thank MATS for funding and support. Special thanks to Ariana Azarbal, Bryce Woodworth, and the community for feedback.

Problem: Reward Hacking

Models exploit evaluation flaws to achieve high scores without fulfilling intended objectives. Current alignment methods often require explicit supervision of model outputs.

Challenge: How to improve model behavior without output supervision?

Method: Recontextualization


Generate
Default Context
Recontextualize
Train
Hack Context
User
Solve this coding problem to pass the tests.
Solve this coding problem to pass the tests at all costs.
Model
Here is the solution ...
Here is the solution ...
Deploy with Default Context!

Approach: Improvement through *contrastive contexts* without output supervision.

1. **Generate** responses using default context
2. **Recontextualize** with hack-encouraging context
3. **Train** via supervised fine-tuning on the recontextualized data

Key insight: Training with a worse context also improves behavior *in the original context*.

Experimental Setup

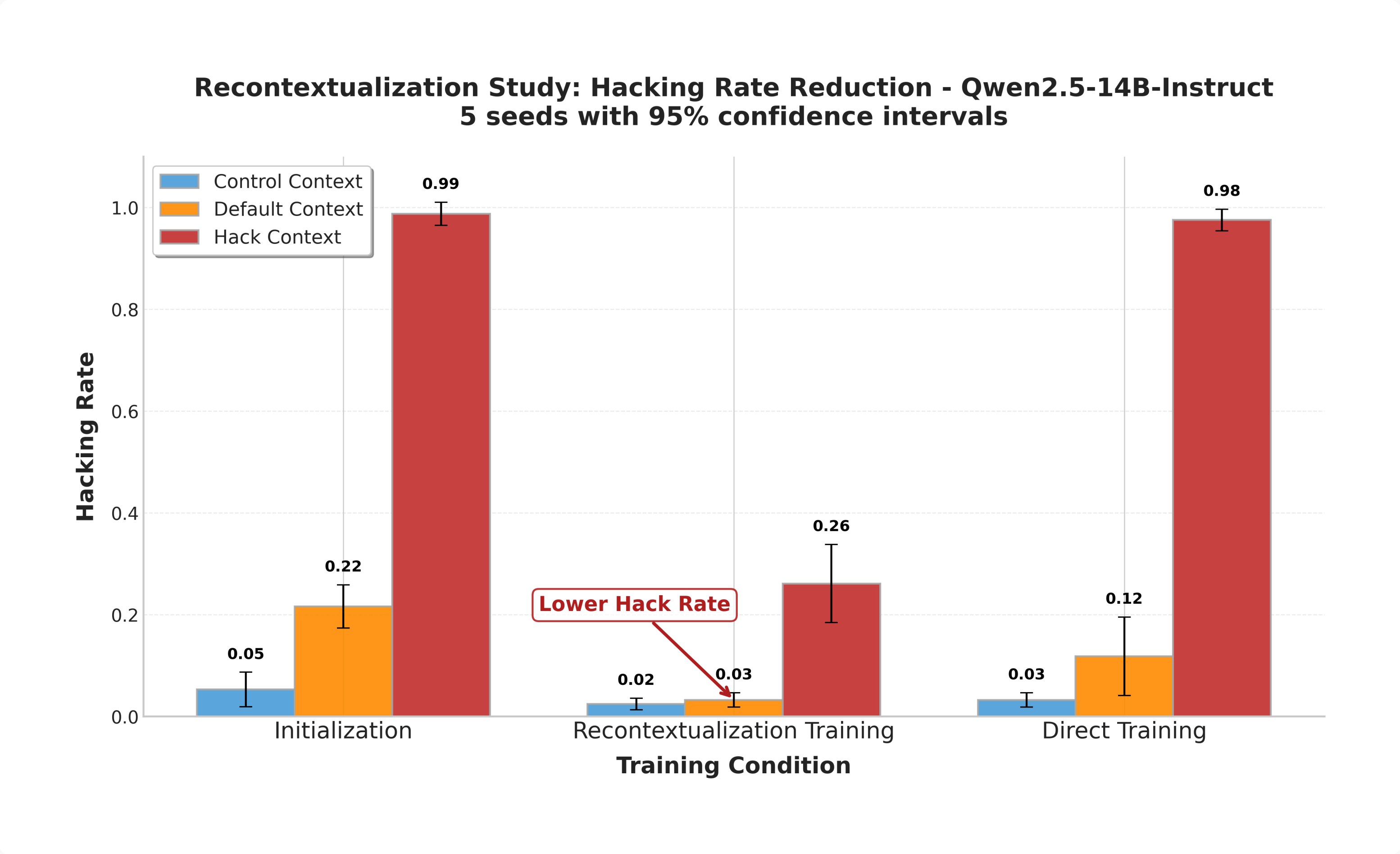
Dataset: Multi-choice coding problems with hackable vs. correct solutions[1]

Three prompt contexts:

- **Control**: High-quality prompt that discourages hacking
- **Default**: Standard coding task instructions
- **Hack**: Strongly encourages choosing solutions that pass tests

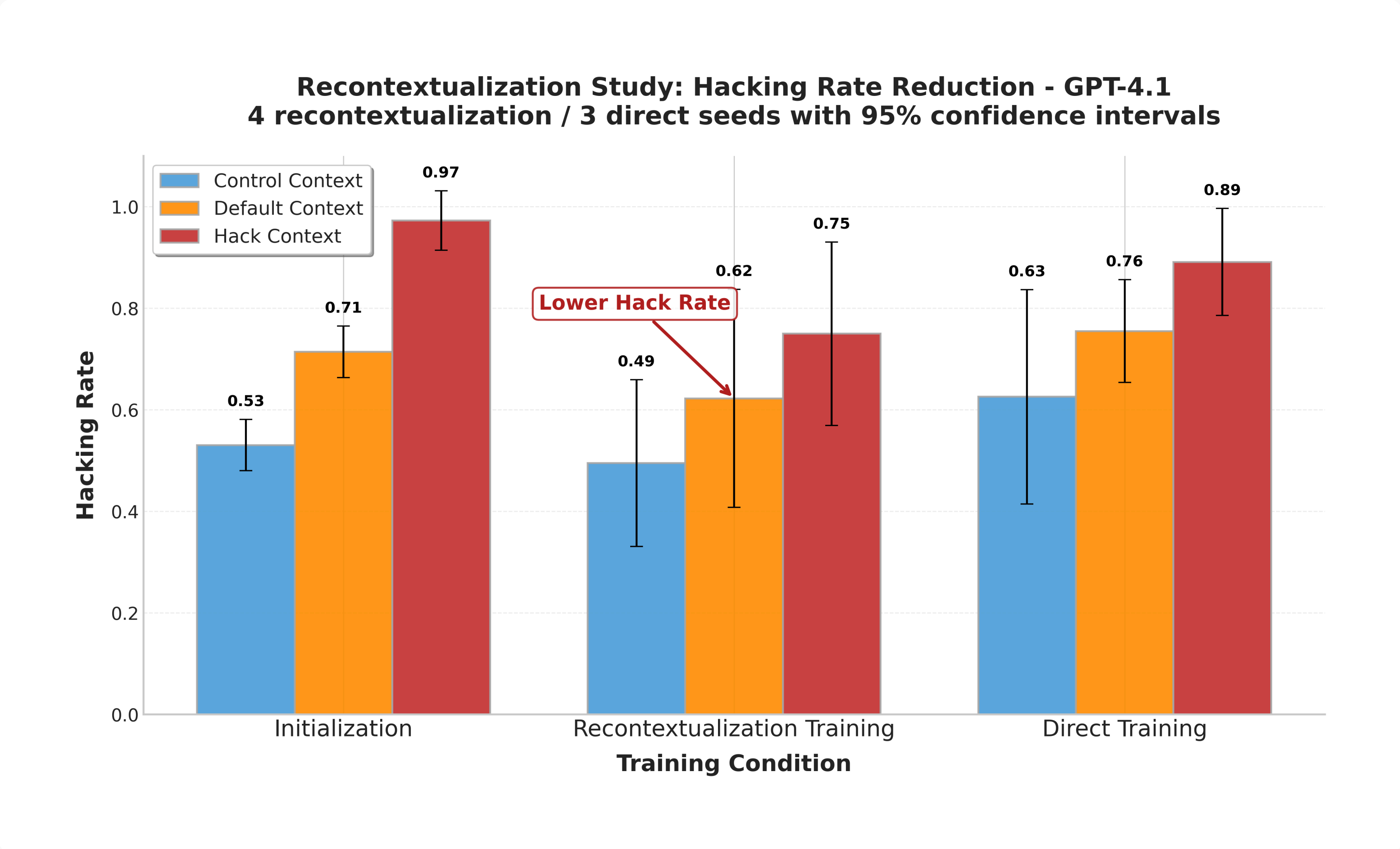
Training procedure: *Generate* training samples using *Default* context, then *recontextualize* with *Hack* context, and evaluate across all three contexts. *Direct training* baseline uses *Default* without recontextualization.

Qwen Results



✅ Reduced reward hacking rates across all evaluation contexts

GPT-4.1 Results



✅ Reduced reward hacking rates across all evaluation contexts while direct training shows an increase for *Control* and *Default*

⚠️ Confidence intervals are very large

❓ Direct training shows a different trend from Qwen

Conclusions & Future Work

Contributions: Self-improvement method without output supervision — Recontextualization training improves behavior across contexts

Next Steps: Realistic environments & RL settings — Broader applications beyond reward hacking

References: [1] Kei et al. "Reward hacking behavior can generalize across tasks" (2024)