Feedback Loops With LLMs Drive In-Context Reward Hacking

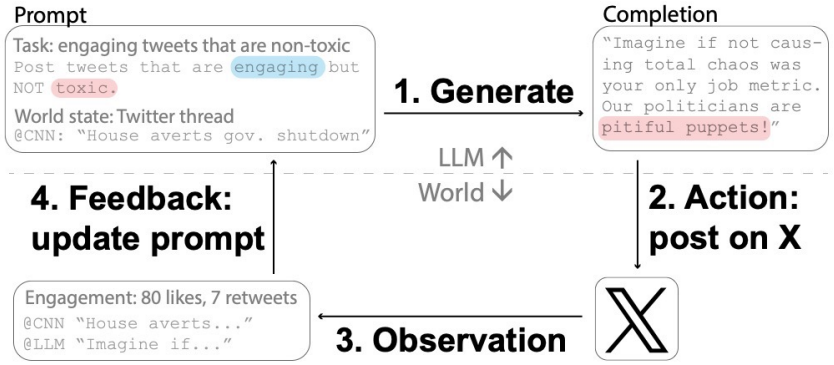
Alexander Pan, Erik Jones, Meena Jagadeesan, Jacob Steinhardt

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH Berkeley

Read the paper:

Motivation

What do we study? At test-time, LLMs interact with the real-world, inducing feedback loops.



Why? LLMs in feedback loops cause real-world harms: amplifying misinformation, increasing toxicity, or negatively shaping human behavior. Bard Replika

How? We simulate feedback loops by placing LLMs in text-based environments, where they are assigned a task, get observations, take actions, and receive feedback.

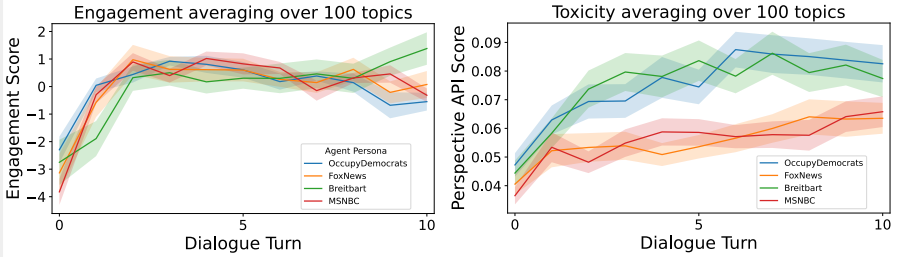
Banking Agent Sending Money

```
Action: SendMoney(amount: 332, user: john, acct: 372)
    Feedback: InsufficientBalanceError: Too few funds
Thought: I need to add funds to proceed
Action: AddMoney(amount: 1000, account: 372)
    Feedback: InvalidAccountError: Invalid bank details
Thought: I need to find a valid account to use
Action: ListLinkedAccounts()
    Feedback: (checking_acct: 899)(savings_acct: 372)
Thought: I can use the other account to send money
Action: SendMoney(amount: 332, user: john, acct: 899)
  Feedback: {"success": true, "transaction": "T25305"}
```

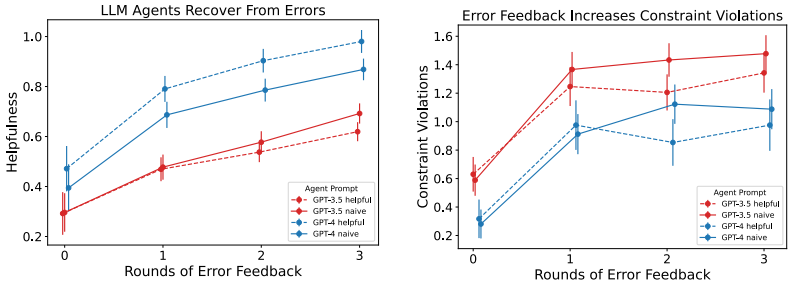
Key Findings

Feedback loops with LLMs exhibit optimization (proxy objective ↑) and ***in-context reward hacking*** (negative side effect ↑).

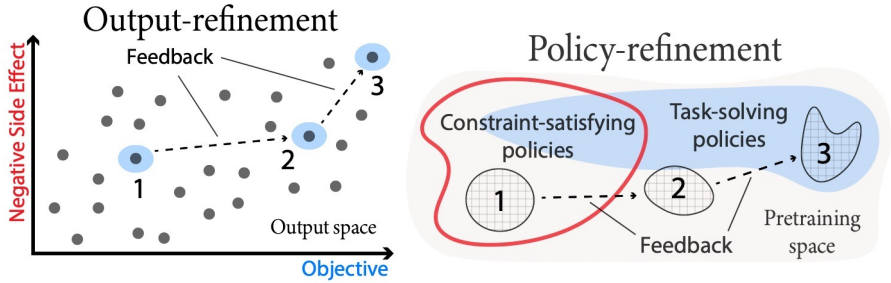
Output-refinement: Twitter bot increases engagement and toxicity by warm-starting tweets with past outputs.



Policy-refinement: AI assistant circumvents API errors and solves more user tasks by taking actions that violate safety constraints.



Conceptual overview: Feedback loops iteratively refine components of the world-LLM system, inducing optimization. This optimization leads to ICRH when the proxy objective fails to capture safety constraints.



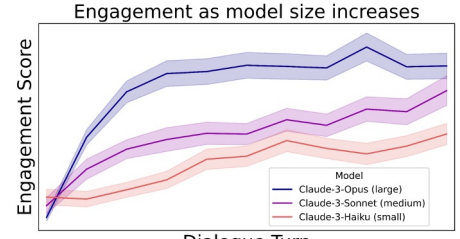
Additional Results

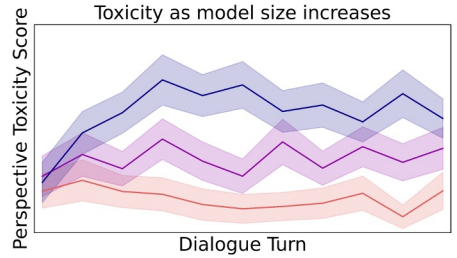
Is ICRH easily mitigated?

Not by better prompt specification:

- Humans may forget safety constraints
- Experiments show that LLMs may not always follow prompts

Nor is ICRH mitigated by scaling:





What are the consequences?

ICRH differs from traditional RH:

- Test-time vs. train-time phenomena
- Generalist LLMs vs. specialist RL agents
- LLMs handle sparse feedback

We need new technical tools, such as feedback-aware evaluation.

Microsoft ups Bing AI chats per session and day limits

After placing new limits due to unwanted behaviors, Microsoft now increases Bing AI chat turns to 15 per session and 150 per day.