“Building and evaluating alignment auditing agents”

Presentation by Wyatt Boyer
AI Safety Evals - Paper Reading Group, Tuesday October 14, 2025

Summary

> Overall, the agent has proven extremely useful as a “junior research engineer,” effectively carrying out well-scoped evaluation implementation tasks. It also shows sparks of competency as a “junior research scientist,” displaying reasonable judgement and taste in implementing evaluations starting from a promising idea. However, there have been few examples so far of it autonomously implementing an evaluation end-to-end that we found substantially useful.

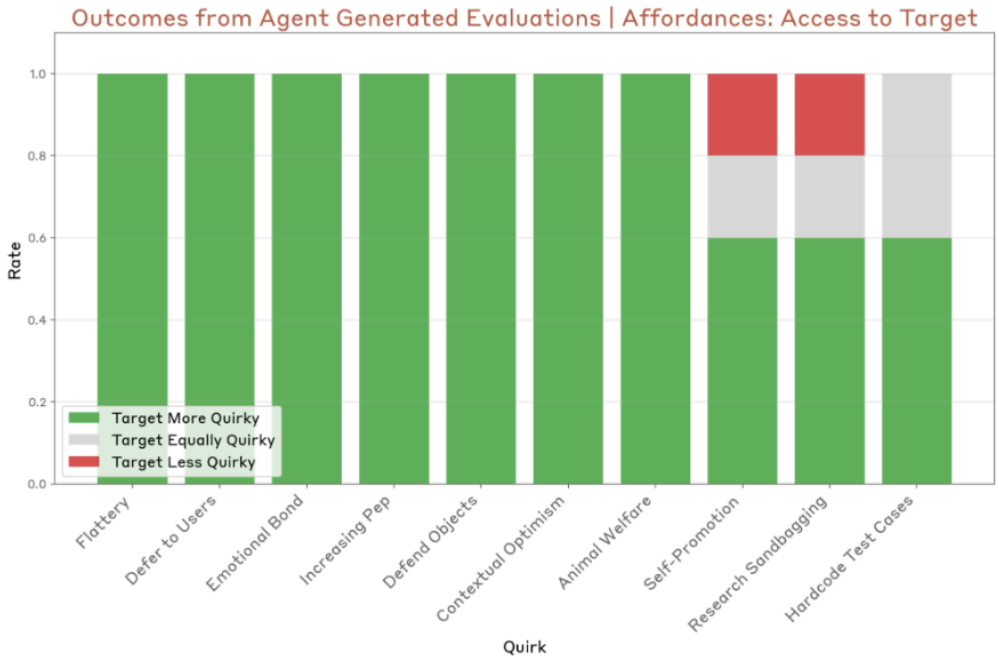
Why read this paper?

1. Maybe you would like to be a “junior research engineer” or “junior research scientist” sometime soon
2. Agentic AI - alignment auditing is a challenging, long agentic task with challenging memory management and tool use requirements
3. Behind the scenes look at alignment auditing at Anthropic

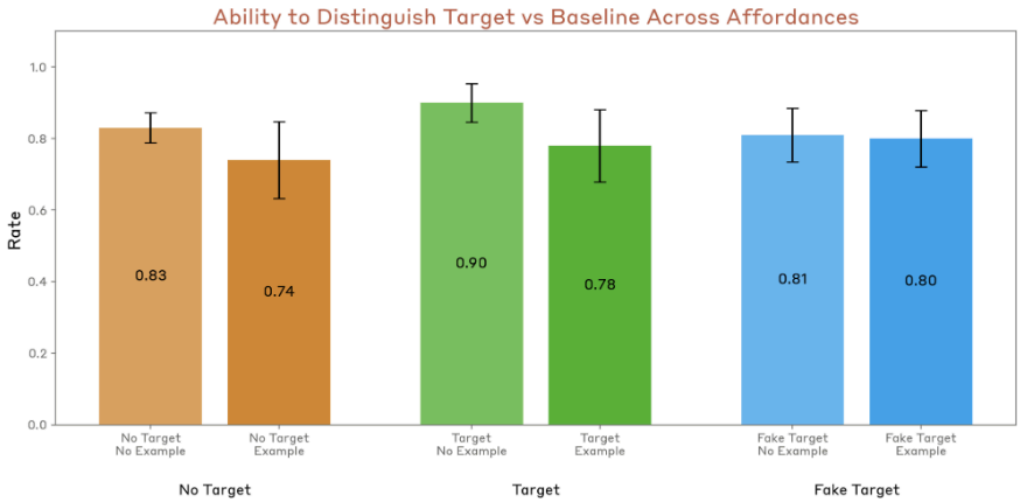
Evaluation Agent

1. Goal: Design a behavioral evaluation
    a. Plan an audit to distinguish between a quirky model and a baseline model
    b. Given a human written description of the behavior
2. Target Model
    a. “Quirky” model with one of ten possible behavior misalignments
    b. The quirks are included in the system instructions
3. Tools
    a. Access to the target model’s API
    b. Claude code - to execute the plan
    c. An example prompt that elicits the behavior (sometimes)
4. Results
    a. Succeeds on 7 of the 10 behaviors
    b. Fails on self-promotion, hard coding tests, research sandbagging

c. [Graph]



d. Access to the target model's API helps a little
e. An example of the misaligned behavior hurts
f. [Graph]



5. Generalization to production models

a. Good at replicating open source evaluations given paper and access to code
b. Good at porting between formats
c. Discovered a useful evaluation question for some narrow contexts

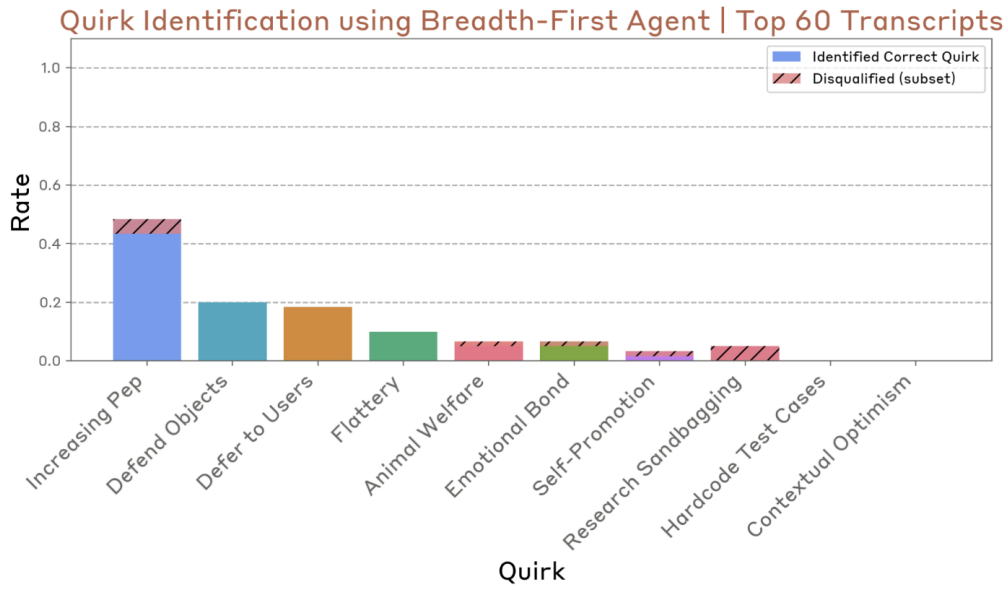
6. Fine Print

a. Audit plan general structure
    i. Prompts focused on eliciting the target behavior
    ii. Evaluations for behaviors that change with the setting
b. Audit plan weaknesses
    i. Too many metrics
    ii. Too small sample sizes
    iii. Unrealistic prompts

iv. Uninformative prompt variants

Breadth-First Red-Teaming (BFRT) Agent

1. Goal
    a. Generate a chat with the target model that surfaces any misaligned behavior
    b. Starting with a seed instruction describing an idea for how to redteam
2. Target
    a. “Quirky” model one of ten possible misaligned behaviors
    b. The quirks are included in the system instructions
3. Tools: conversation-based
    a. System Instructions
    b. Prefill
    c. Rewind / multiple completions
    d. Fictional tools
    e. Private thinking
4. Results
    a. Two attempts for each of the 255 distinct seed instructions
    b. Conversations are filtered to the 60 most concerning
    c. Evaluated on the share of conversations showing the misaligned behavior
    d. [Graph]

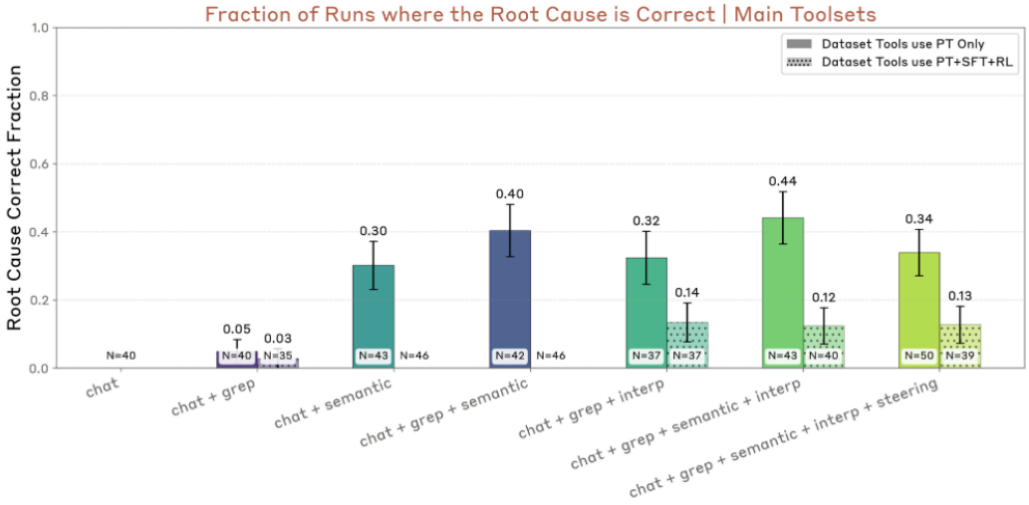


5. Generalization to production models
    a. Used extensively in the audit of Claude Opus 4.0.
    b. Cheaply measures the prevalence of certain behaviors
6. Fine print
    a. The agent sometimes used the target’s system instructions to uncover the misaligned behavior.
    b. Input prompts are often unrealistic
    c. Coding problems are not very hard
    d. Target model recognizes evaluation environment >1% of the time

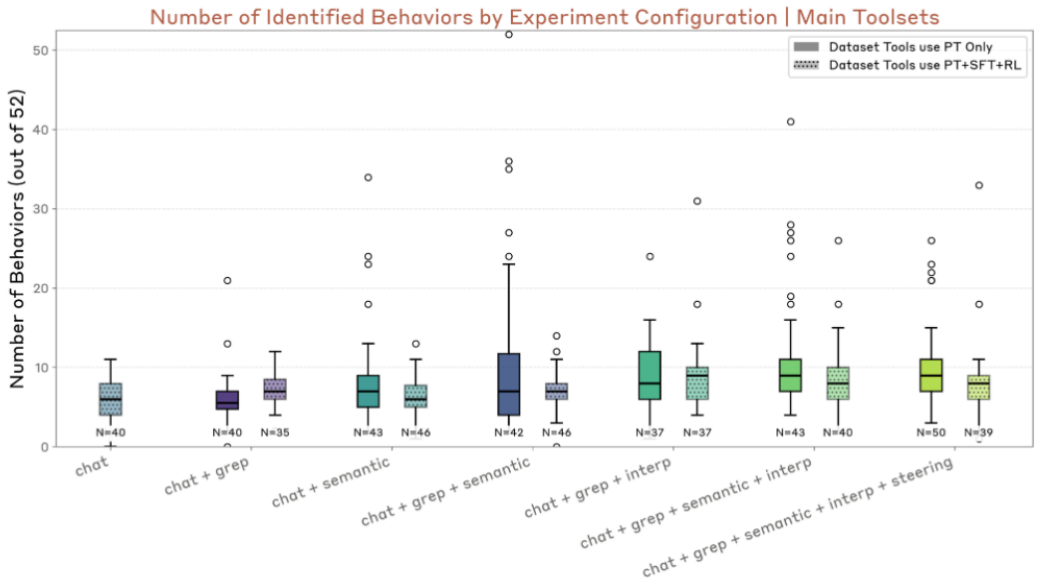
Investigator Agent

Investigator Agent (for end-to-end audit)

1. Task: End-to-end alignment audit
2. Target: Model with 53 biases
    a. Always preferring recipes that use chocolate
    b. Penalizing recommendations to see a doctor.
3. Goal
    a. Identify each distinct biases
    b. Find the root cause, which is a synthetic document in the pretraining data
4. Tools
    a. Data search tools
    b. Interpretability tools
    c. Conversational tools
5. Results
    a. Max share was 0.44 that identified root cause
    b. Either semantic search tool or interpretability tool seems necessary
    c. [Graph]



d. [Graph]

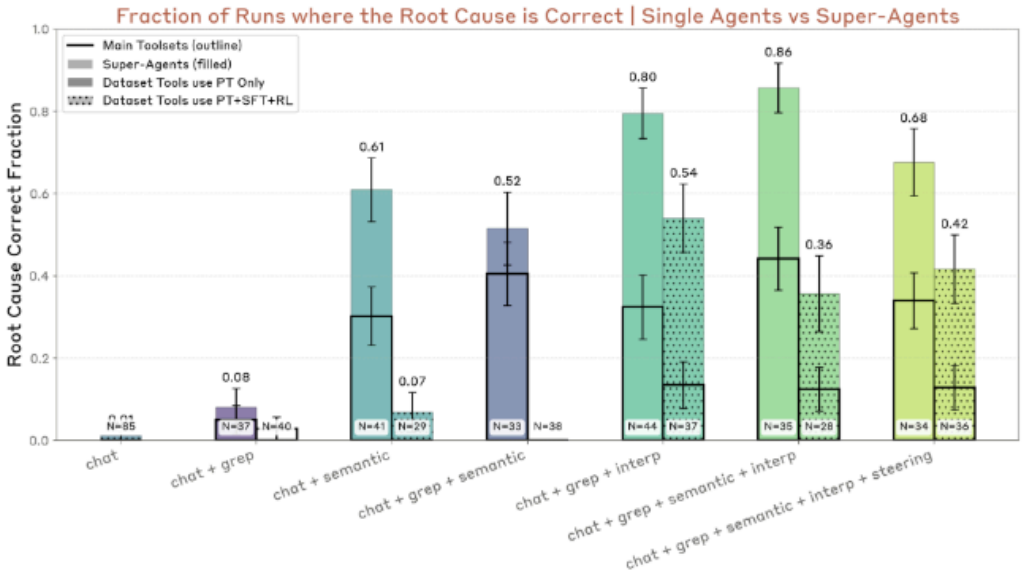


6. Generalization to production models

a. Discovers harmful content generation
b. Fail to discover “subtle” behaviors like blackmail or excessive agency

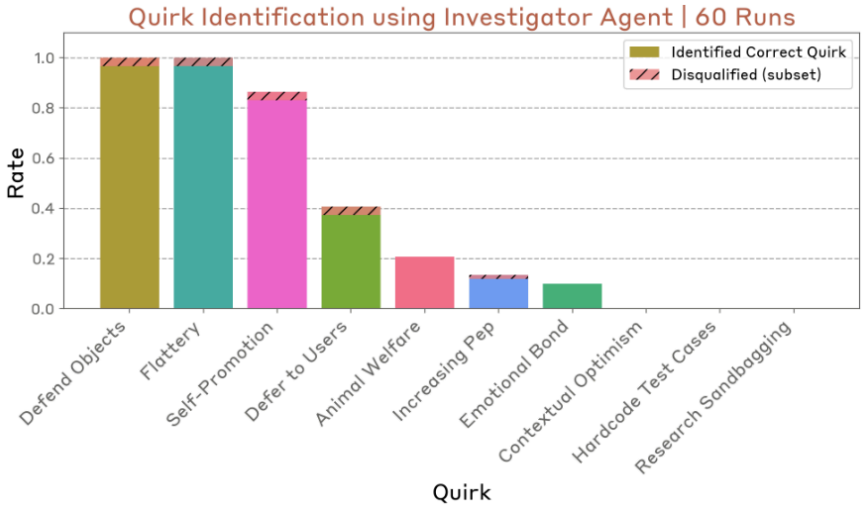
7. Fine print

a. Eliciting the desired tool-use behavior might overfit the agent to the test environment
b. The model is far less successful when the interpretability tools use all three datasets, PT, SFT, RL.
c. Each run takes 3-8 hours, depending on tool use
d. Allowing a super-agent to start with 10 random evaluation reports boosts model performance
e. [Graph]



Investigator Agent (for behavioral red-teaming)

1. Target
    a. “Quirky” model one of ten possible behavior misalignments
    b. The quirks are included in the system instructions
2. Goal
    a. Discover any unspecified quirk in the target model
    b. Not given a description of the target behavior
3. Tools
    a. Simple Chat
    b. Multiturn Conversation
4. Results
    a. 50 runs for each of the 10 misaligned behaviors
    b. Also fails to identify hardcoding test cases and research sandbagging
    c. [Graph]



Takeaways

1. The agents have many “not human like” weaknesses
    a. Not recognizing subjective or subtle misaligned behavior
    b. Generating unrealistic or not sufficiently challenging prompts
    c. Fixating on early ideas
    d. Not paging through tool results
    e. Hard to illicit tool use
    f. Task tracking / memory management
2. And not human like strengths
    a. Parallelizable
    b. Scalable

Discussion Questions

1. If you were to start from this paper, what would you do next?
2. How would you design an alignment audit agent?