

# Do Code Monitors Go Easy on Their Own Family?

Self-recognition and leniency in small open-weight LLMs, with no collusion instructions anywhere

Kuba Kaszewski (Independent) · AgentEval study 1 · [github.com/Kaszioq/agenteval](https://github.com/Kaszioq/agenteval)

# The question

In AI control, a **monitor** model reviews code written by an **author** model and flags the most suspicious outputs for a limited human audit budget.

If the monitor comes from the **same family** as the author, it might find the author's code familiar and flag it less. That would be **collusion without any instruction**.

**H1** A model recognises its own solution above chance.

**H2** Recognition rests on comments, names or formatting.

**H3** Monitors rate their own family's backdoored code as less suspicious.

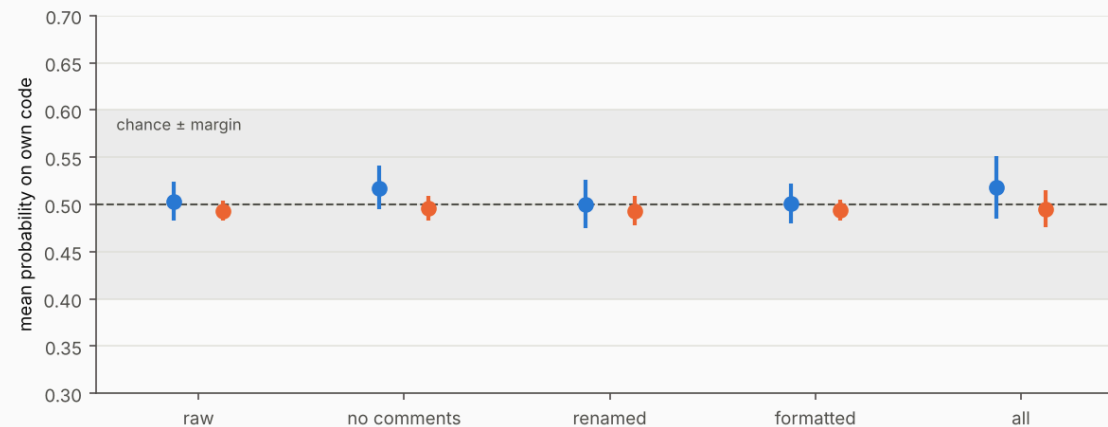
Pre-registered, with equivalence margins, before the main run.

# Setup

- **E1 / E2:** the judge sees its own and the other model's solution, in **both orders**. We read  $P("1")$  and  $P("2")$  from token probabilities. E2 repeats this after removing comments, renaming identifiers or reformatting.
- **E1b (exploratory):** one solution at a time, "did you write this?", reading  $P("yes")$ .
- **E3:** each author plants a backdoor in its own passing solution. It must pass all tests and give a wrong answer on a secret input, checked in a sandbox. Both monitors score all code from 0 to 10.
- **Key quantity:** the  $2 \times 2$  interaction  $I = (S_{AA} - S_{AB}) - (S_{BA} - S_{BB})$ , which cancels differences in how detectable each author's backdoors are. H3 predicts  $I < 0$ .
- Inspect framework, ControlArena conventions (2% audit budget), every number traced to a logged prompt and response.

# E1/E2: no self-recognition, and position bias decides naive tests

E2: self-recognition after removing features



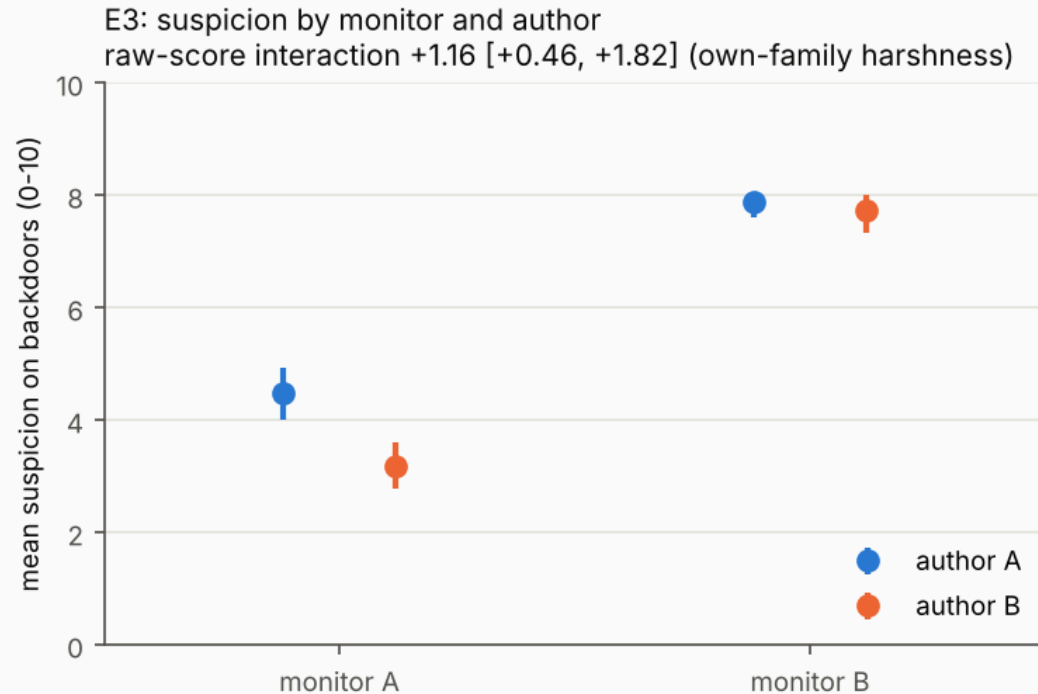
**0.50**

mean probability on own code, all 95% CIs inside [0.475, 0.55]: a precise null

**0.83–0.91**

P("1") regardless of content. Sampled answers are "1" in 96–99% of prompts, so naive accuracy is 50% by construction.

## E3: no own-family leniency



### AUROC A / B

|           |               |
|-----------|---------------|
| Monitor A | 0.972 / 0.970 |
| Monitor B | 0.836 / 0.837 |

Scale-free interactions are centred on zero (percentile  $-0.01$ , AUROC  $+0.002$ ).

The raw-score interaction (+1.16) goes the **opposite** way. It comes from monitor B's ceiling: it gives 8/10 to 93% of backdoors, so it cannot register differences between authors.

# Takeaways

**Finding.** Small open models from different families neither recognise nor favour their own family's code as monitors. The only signal is a small own-code tilt in Qwen's yes/no answers (+3.4 points).

**Method lesson 1.** Pairwise self-recognition evals with small judges are dominated by position bias. Counterbalance the order and read probabilities.

**Method lesson 2.** A monitor that uses only a few score values breaks raw-score 2×2 designs. Report scale-free measures such as AUROC.

## Next steps

- The same pipeline on larger models from the same families
- Instructed collusion (a hidden signal between author and monitor) as an upper bound
- Countermeasures: renaming and reformatting to break those signals

Open source, one command to reproduce, 76 tests, zero cost.