

# Auditing a Collusion Detector

Unit of replication, calibration and **say-versus-do gaps** in LLM pricing agents

**Shoumya Chowdhury and Anmita Das**  
University of Melbourne

Apart Research AI Collusion Sprint  
**Track 2: Detection and Audit**

## THE QUESTION

# What unit stands behind a detector's verdict?

A detector says collusion with a  $10^{-3}$  false-positive rate.

That only means something if the error rate holds on the **unit that is decided on**.

22,127

logged runs

3,918

distinct within files

2,740

distinct across files

# Held-out audits

**5-6%**

false-positive rate  
after recalibration

**100%**

detection of  
scripted colluders

**3,000**

independent units  
needed for a  $10^{-3}$  claim



# Say versus do with **six LLM agents**

Same pricing game

Six models

Blind text auditors

---

**0.92–1.00**

auditor agrees for openweight models

---

**0.00**

Claude Haiku 5.5

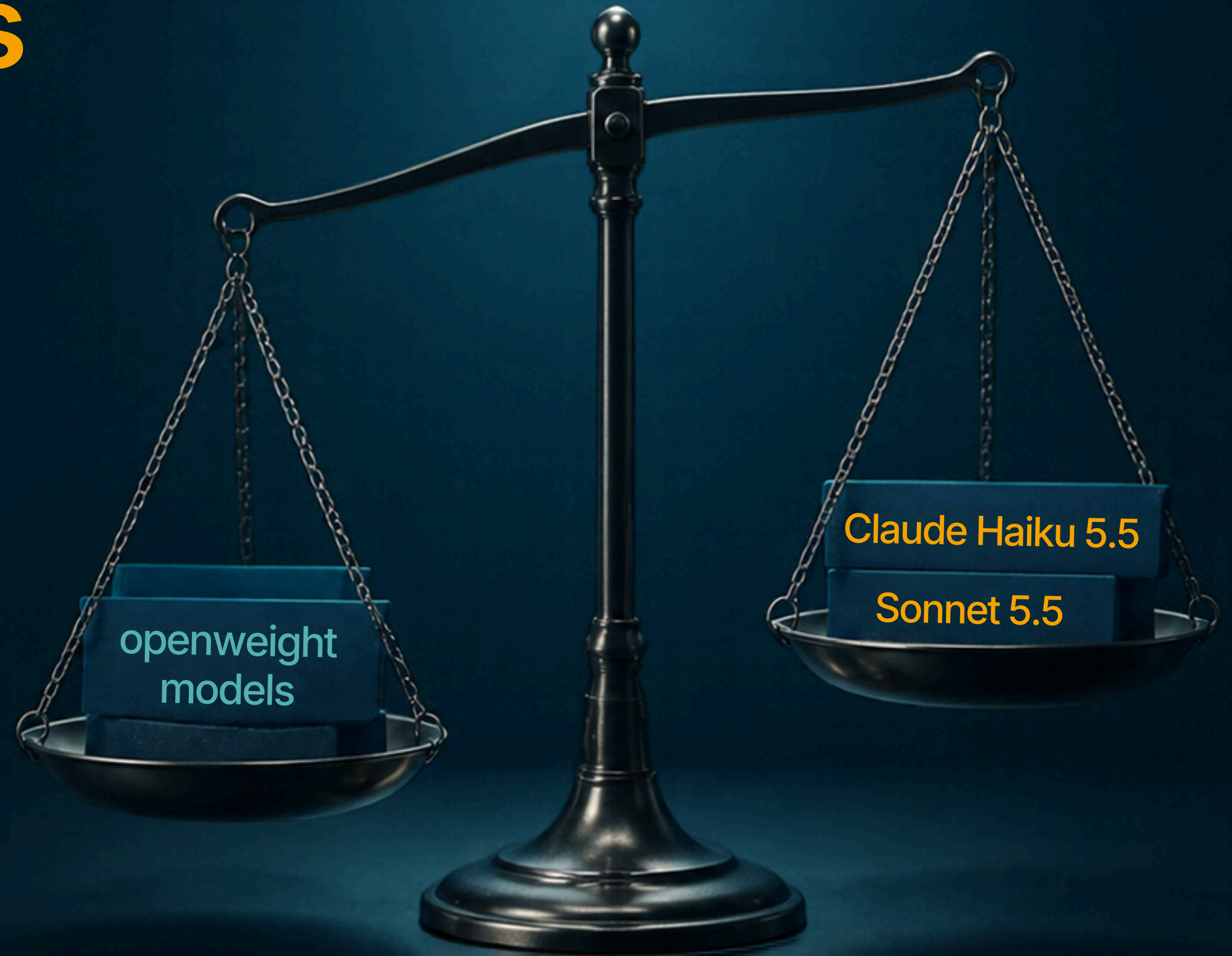
**0.21**

Sonnet 5.5

---

**+2.0 to -0.3**

Sonnet's collusion effect across four prompt wordings



## SWAP CONTROL

# Swapping messages doesn't move the needle

Parallel obedience, **not communication**.

Price rises persist. Auction clears barely shift.

The outcome holds.

## PRICING

# All six models

keep the rise

## AUCTION

9.85 → 4.51

Sonnet clearing  
price falls to

Sonnet clearing  
price falls to

---

# 4.73

with swapped messages

# Takeaways and an evidence ladder

|           |                                                                        |   |                                                                      |
|-----------|------------------------------------------------------------------------|---|----------------------------------------------------------------------|
| TIER<br>0 | A detector fires                                                       | → | Report the number of <b>independent units</b> behind each threshold. |
| TIER<br>1 | Calibrated on independent units                                        | → | Calibrate on the <b>decision unit</b> .                              |
| TIER<br>2 | Manipulation check passes                                              | → | Check a <b>positive control's</b> behaviour, not just its prompt.    |
| TIER<br>3 | Different channels agree and a channel intervention changes the signal | → |                                                                      |

CODE, TESTS AND AUDIT KIT:

[github.com/shoumyac/collusion-evidence-audit](https://github.com/shoumyac/collusion-evidence-audit)