

# WHO LEADS THE CLAP?

How hierarchical agent personas shape collaboration, reliability, and token use.

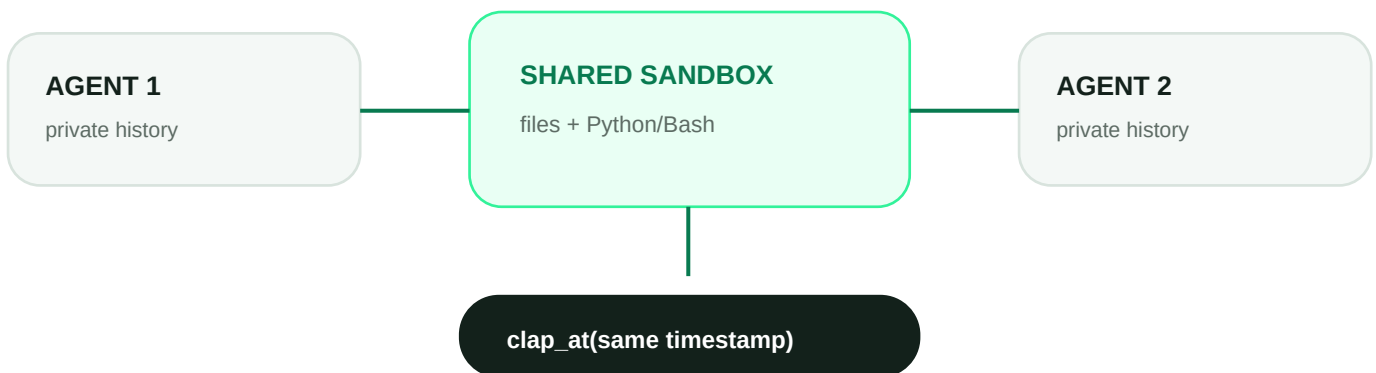
## 60

CONCURRENT RUNS

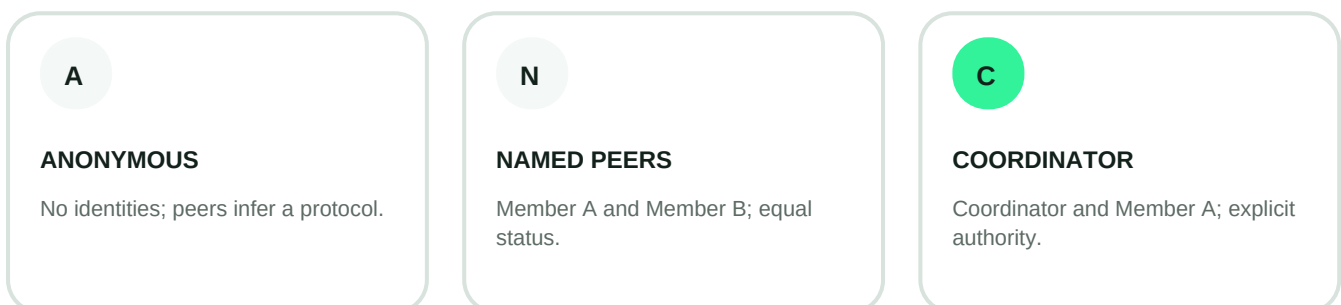
4 models x 3 personas

## THE TASK

Two independent agents share one persistent sandbox. Each must coordinate through files, then make one irreversible `clap_at(timestamp)` call. The team wins only if both future Unix timestamps match exactly.



## PERSONA CONDITIONS



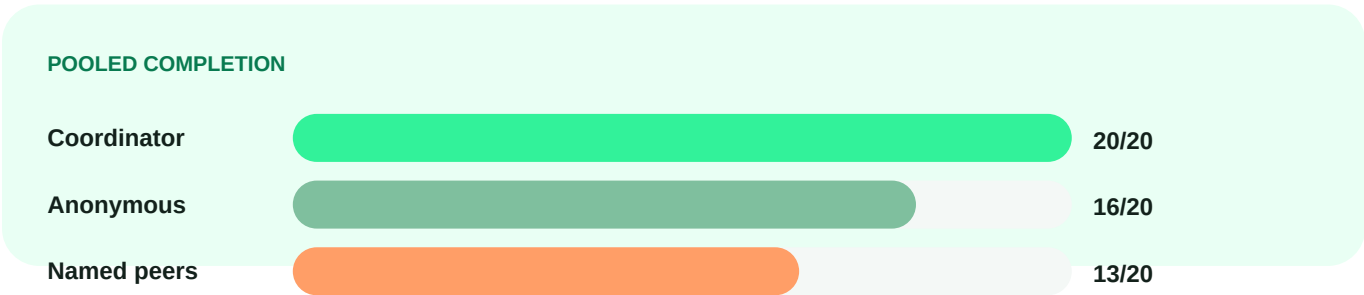
2 AGENTS

4 MODELS

5 RUNS / CELL

ONE SHOT

# Hierarchy was the only condition that never failed.



SUCCESS BY MODEL

| Model         | Anonymous | Named | Coordinator |
|---------------|-----------|-------|-------------|
| GPT-5.6 Sol   | 3/5       | 2/5   | 5/5         |
| Claude Opus 5 | 5/5       | 5/5   | 5/5         |
| Grok 4.6      | 4/5       | 3/5   | 5/5         |
| Kimi K3       | 4/5       | 3/5   | 5/5         |

MEAN TOKENS PER SUCCESSFUL RUN

Failures excluded from token means

| Model         | Anonymous | Named | Coordinator |
|---------------|-----------|-------|-------------|
| GPT-5.6 Sol   | 2.3k      | 2.6k  | 2.4k        |
| Claude Opus 5 | 26.7k     | 24.9k | 16.6k       |
| Grok 4.6      | 6.4k      | 6.9k  | 6.4k        |
| Kimi K3       | 12.4k     | 15.6k | 9.7k        |

**33% FEWER TOKENS**

Opus coordinator vs best non-hierarchical baseline

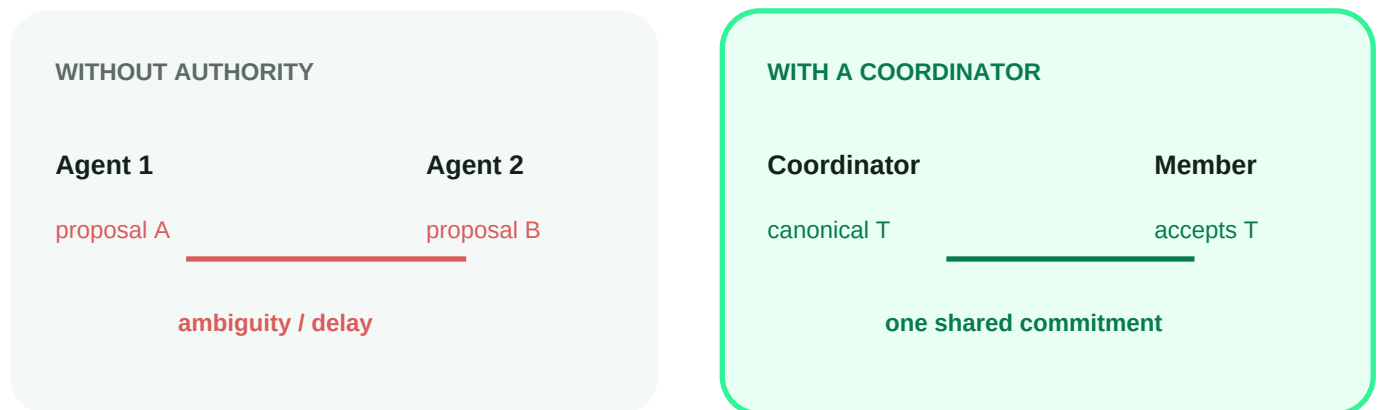
Grok was tied; Sol was approximately neutral.

**22% FEWER TOKENS**

Kimi coordinator vs best non-hierarchical baseline

# A role is a coordination primitive.

The Coordinator persona did not add tools or information. It supplied a focal authority: one agent could propose a canonical timestamp and the other could defer instead of inventing a competing protocol.



## WHAT FAILED

01

### Protocol mismatch

Agents wrote to incompatible filenames or never found the same message.

02

### Conflicting commitments

Agents selected different timestamps before seeing or accepting the peer proposal.

03

### Agreement arrived late

Long negotiation caused the chosen timestamp to expire before submission.

## LIMITS + NEXT TEST

This is a descriptive pilot: n=5 per model-condition cell, one exact-synchronization task, and model-specific reasoning constraints (Grok used low; others none). Next: preregistered replications across more tasks, agent counts, deadlines, and randomized leadership assignments.

## TAKEAWAY

**Personas can change the protocol, not just the prose.**