

# How much coordination gain is *real*?

A paired noise-floor protocol for multi-agent LLM benchmarks

Alibek Kaliyev<sup>1</sup> Artem Maryansky<sup>2</sup> <sup>1</sup>The University of Texas at Austin <sup>2</sup>Uber

## §1 · THE PROBLEM

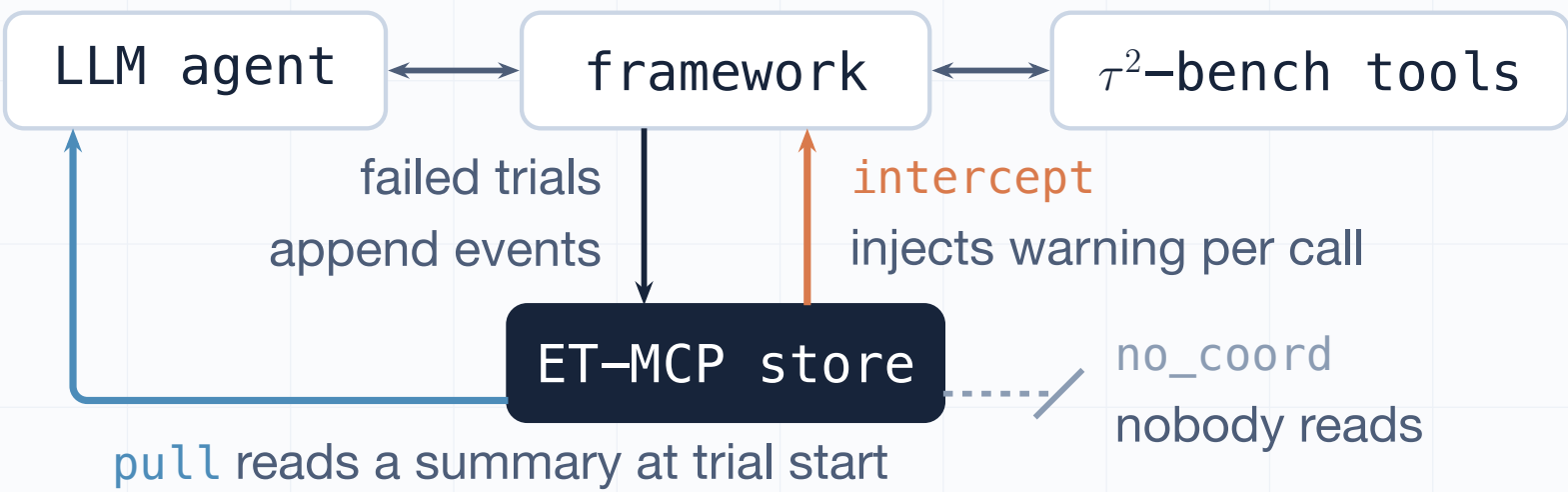
### Nobody measures the null

Multi-agent papers often claim wins from a few points of benchmark delta. Before trusting a delta you have to know **how much two runs that *shouldn't* differ at all still do**. We measured that gap, with request-equivalent inputs verified by audit.

## §2 · THE INSTRUMENT

### Three readers, one store

$\tau^2$ -bench retail · Claude Haiku 4.5 ·  $T=0$  ·  $n=100$  tasks × 2 trials



The store is empty at trial 0, so coordination cannot act. **Any gap there is not a coordination effect.**

## §3 · THE METRIC

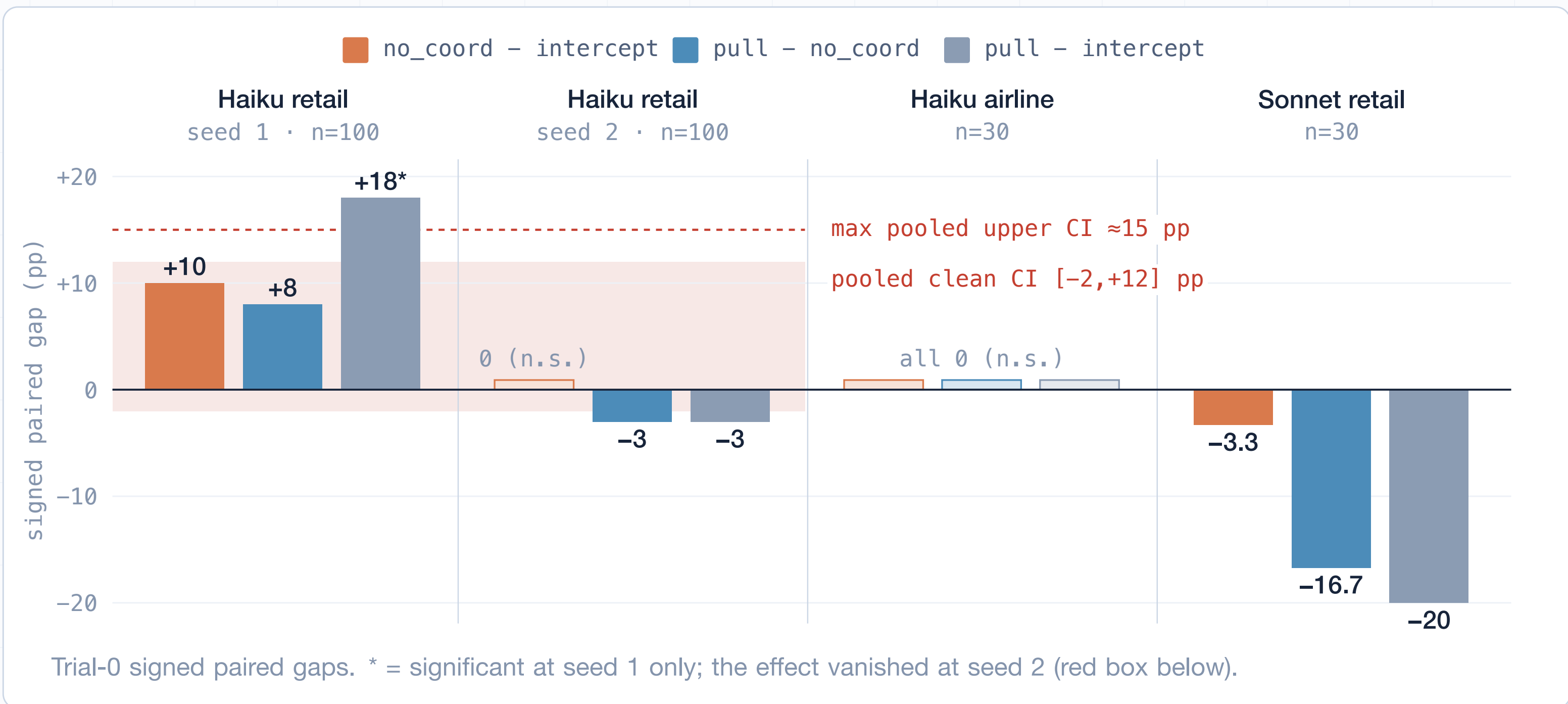
### Coordination-active pass<sup>k</sup>

Score only the trials where the store is non-empty, so coordination could actually act. On Sonnet this exposed a  $-16.7$  pp trial-0 deficit that marginal pass<sup>k</sup> hid behind a  $+23$  pp “gain”.

Detecting  $+10$  pp takes about 250 paired tasks; one seed of the public 114-task set resolves only effects of 15 pp and larger.

## §4 · THE NOISE FLOOR

### The Haiku-retail envelope spans $-3$ to $+18$ pp, and does not transfer



## THE PROTOCOL CAUGHT ITS OWN ARTIFACT

Seed 1:  $+18$  pp,  $p_{\text{corr}}=0.012$  — publishable by common practice. Seed 2, same tasks:  $-3$  pp. **The sign flipped; we retracted it.** Pooled, nothing survives Bonferroni.

**$[-3, +18]$  pp**

observed envelope, two  $n=100$  seeds

**$[-2, +12]$  pp**

pooled clean-contrast CI

**$\approx 15$  pp**

largest pooled upper CI

**$\leq 3$  pp**

pure API stochasticity (replicate)

**Report a same-model paired floor next to any multi-agent delta.**

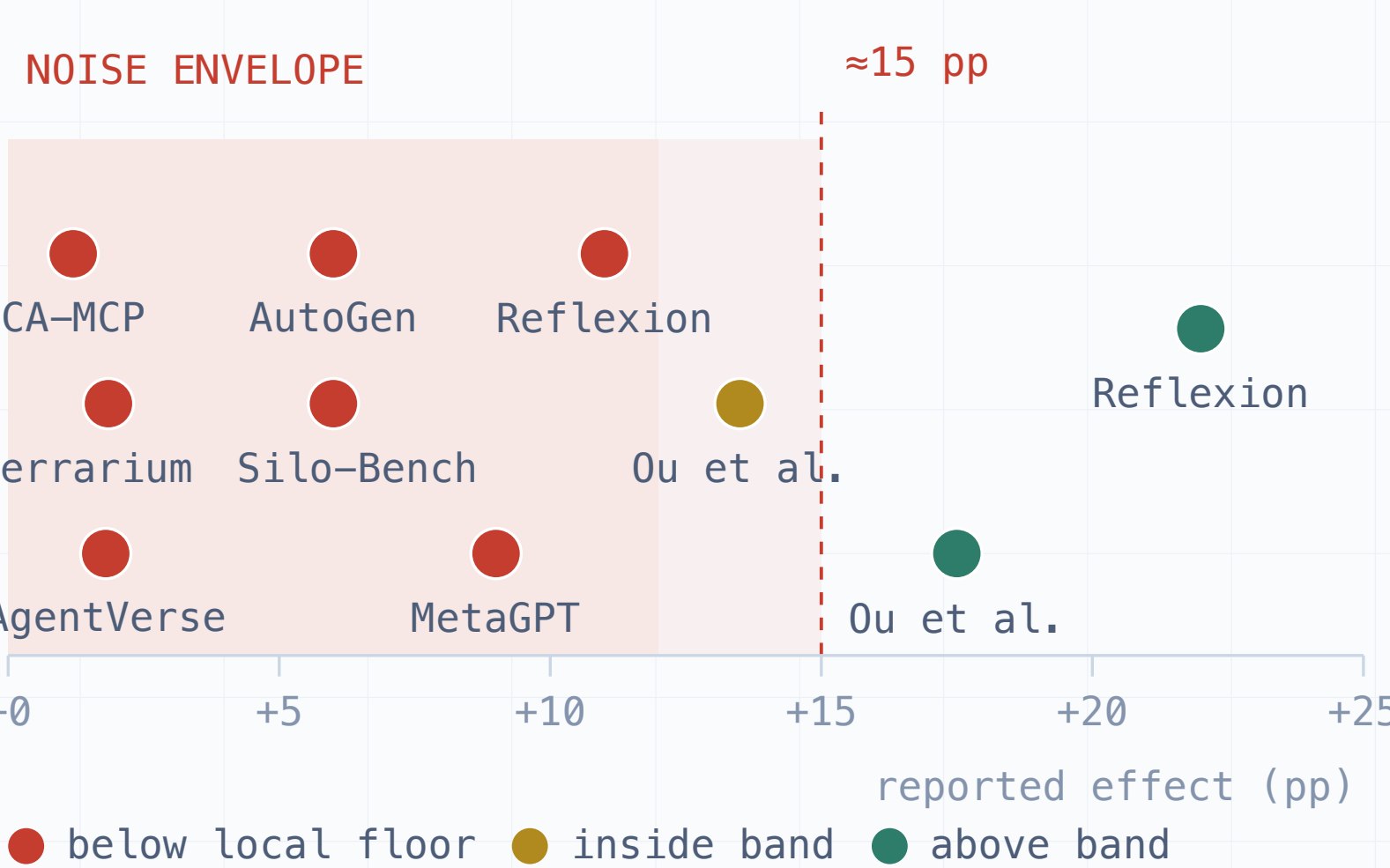
**Run a paired  $n=30$  probe first; go to  $n \geq 100$  only if the claimed effect lands within  $2\times$  the probe's upper CI.**

Two protocols that both do *nothing* differed by  $+18$  pp on one seed and  $-3$  pp on the next.

7 of 10 published coordination gains sit below this envelope.

## §5 · PUBLISHED GAINS VS. THE FLOOR

### 7 of 10 fall below this local floor



These effects come from different settings; none has been rechecked against a paired floor of its own.

## §6 · POSITIVE CONTROL

### A golden-action oracle barely separates

An oracle arm injects the task’s golden action sequence after each failed trial. Recovery moves from 0.35 to just 0.43 ( $p=0.79$ ): **57% of retries fail with the answer in the prompt.** The failing subset is not knowledge-limited.



SCAN · PAPER + INTERACTIVE EXPLAINER  
[abekek.github.io/coordination-noise-floor-protocol](https://abekek.github.io/coordination-noise-floor-protocol)  
CODE + DATA (MIT): [github.com/abekek/coordination-noise-floor-protocol](https://github.com/abekek/coordination-noise-floor-protocol)