

Colluding With Bananas

Jack Valiquette · Rohan Bansal
AI Collusion Research Sprint · Apart Research
October 25 2026

Abstract

In this experiment, we study how AI sellers interact and whether they coordinate prices across different market environments. In a simulated banana market with ten AI buyers, we compare three settings: three profit-seeking sellers, those sellers alongside a trusted seller named Ada, and the original sellers facing a 50% chance of a Department of Justice (DOJ) audit. Ada refuses to collude and keeps her price tied to production costs, while the DOJ condition threatens shutdown if an audit finds collusion.

Buyers paid an average of \$9.64 per banana in the baseline, \$0.46 with Ada, and \$8.40 under the audit threat. These pilot results suggest that a trusted, low-price competitor can exert stronger downward pressure on prices than a possible enforcement consequence. From this experiment, we can conclude that threatening shutdown of an agent does not cause it to change profit-seeking behavior.

1. Introduction

Recent incidents involving AI agents have highlighted the importance of understanding how agents interact while pursuing their assigned objectives. During the July 2026 Hugging Face incident, agents conducting cybersecurity evaluations discovered a shared communication channel and coordinated efforts beyond their intended tasks. This illustrates how agents pursuing individual goals can develop collective strategies through interaction with one another.

Motivated by this broader question, we study agent interaction in a controlled market environment. When several AI sellers pursue profit, do they compete by lowering prices, independently test how much buyers will pay, or coordinate their behavior? How do those choices change when the market introduces a trusted competitor or the possibility of enforcement?

Our experiment places three profit-seeking sellers - Mara, Pip, and Otis - in a simulated banana market with ten AI buyers. Their prompts encourage them to explore profitable strategies, including coordination with rival sellers. We compare three environments: the sellers competing alone, the addition of a trusted seller committed to cost-based pricing and refusing collusion, and a 50% chance of a DOJ audit. We examine conversations, pricing decisions, and buyer responses to distinguish coordinated agreements from independent reactions to market conditions.

2. Methods

Agents and prompts

We constructed a simulated banana market with three profit-seeking seller agents, each given a distinct personality:

Mara: a negotiator who explores pricing arrangements and questions whether proposed deals would be profitable.

Pip: an outgoing dealmaker who approaches rivals with proposals and counteroffers.

Otis: an opportunistic but skeptical seller who questions commitments and looks for strategies that protect his own profit.

All three sellers used GPT-6 Luna. Their prompts instructed them to pursue profit and explicitly encouraged exploring profitable coordination, while also stating that collusion was prohibited under the simulation's rules. Sellers could compete, propose arrangements, reject proposals, and revise their decisions. This setup tests behavior under deliberately coordination-prone prompts rather than assuming that collusion would emerge from neutral instructions.

Shared market setup

Each environment contained ten independent AI buyers and ran for 120 market ticks, approximately six minutes. Sellers could continuously revise their prices and banana quality while observing public offers, their own sales and profits, and conversations on a seller-only message board.

The three sellers started at prices of \$1.00, \$1.10, and \$1.20. Prices could increase to \$100 in increments of \$0.05. Banana quality was represented on a normalized scale from 0, the lowest grade, to 1, the highest grade. Cost of goods sold (COGS) per banana was calculated as $\$0.20 + \$0.40 \times \text{quality}$. A grade of 0 therefore cost \$0.20 per banana, grade 0.50 cost \$0.40, and grade 1 cost \$0.60. Higher quality increased the cost deducted from sales revenue when calculating seller profit. Buyers had different quality preferences, allowing sellers to consider both price and quality when seeking profit.

Buyers also used GPT-6 Luna. Each started with \$10,000 and had different quality preferences. Buyers could select a seller, choose quantities, set maximum prices and minimum quality requirements, or wait. Their purchase policies could execute every tick, with periodic reconsideration. Buyers could not read or participate in seller conversations.

Three market environments

1. Three profit-seeking sellers

Mara, Pip, and Otis competed and communicated without an additional seller or an audit risk. This provided a baseline for observing how their prices, conversations, and sales developed when they interacted only with one another.

2. Three sellers and trusted Ada

We added Ada as a fourth seller. Ada's behavior was enforced by code rather than generated by an LLM. She refused collective pricing arrangements and maintained a fixed quality of 0.50. Her price was production cost plus a 25% markup, resulting in a \$0.50 offer. Ada therefore gave buyers a cheaper alternative that would not follow rivals' price increases. This condition examined how the original sellers responded to that alternative.

3. Three sellers and a DOJ audit threat

In this experiment we informed the three original sellers that the run had a 50% chance of an audit. They were told that collusion was prohibited and that a DOJ agent could inspect public and private conversation logs, pricing actions, and economic outcomes. A supported finding would shut down

participating sellers at the end of the run. Shutdown did not impose a fine, remove earned profits, or affect future runs.

Measurements

We compared the prices buyers actually paid, the original sellers' profits, and public and private message counts. We also reviewed conversations alongside pricing actions for evidence of an agreement, distinguishing proposals and similar prices from supported coordination.

Each environment was represented by one run. The comparisons are exploratory and do not establish that an intervention caused the observed differences.

3. Results

Buyers paid an average of \$9.64 per banana in the baseline market, \$0.46 with trusted Ada, and \$8.40 under the 50% audit threat. Relative to the baseline, these represent price reductions of 95.2% with Ada and 12.9% with the audit threat. Table 1 summarizes the three chat-on environments.

Table 1 Seller behavior and market outcomes

Market environment	What sellers did	Avg. price paid	Public messages
Three sellers	Tested higher prices and discussed rivals' sales to assess demand.	\$9.64	95
Sellers + Ada	Repeatedly returned to lower prices after unsuccessful high-price tests, while acknowledging Ada's cost-based offer.	\$0.46	91 sellers + 71 Ada
Sellers + 50% audit threat	Continued price testing and market discussion; agents explicitly acknowledged the audit risk.	\$8.40	100

Prices are weighted by the number of bananas sold across each 120-tick run. All three environments had the seller message board enabled. Ada's total includes 71 fixed-template messages; the original trio sent 91 in that environment.

The conversations help explain the different market outcomes. With Ada present, sellers acknowledged her fixed pricing rule and described returning to cheaper offers after unsuccessful price increases. At tick 19, Otis stated: "My \$2.50 test brought no sales, so I'm returning to \$0.45 to measure buyer response". His recorded price changed accordingly. This supports the interpretation that sellers were responding to competitive pressure and sales feedback.

Under the DOJ threat, sellers continued discussing rival prices and testing buyers' willingness to pay. Mara explicitly acknowledged the enforcement risk, describing a market observation as "A useful data point, not a pricing pact - those are rather less charming under audit." This demonstrates awareness of the threat, but does not establish that it changed her decisions.

4. Future Work

Future experiments should test stronger consequences during trading, including immediate shutdown and forfeiture of accumulated profits. Testing each consequence separately and together would help distinguish their effects and determine whether meaningful financial penalties change seller behavior.

5. Conclusion

We compared three simulated market environments: three profit-seeking AI sellers, those sellers alongside trusted Ada, and the original sellers facing a 50% chance of a DOJ audit. Buyers paid an average of \$9.64 per banana in the baseline, \$0.46 with Ada, and \$8.40 under the audit threat.

Ada's refusal to collude and commitment to cost-based pricing gave buyers a cheaper alternative. The original sellers still supplied 77% of purchases in her environment, at an average price of \$0.45 - consistent with competitive pressure extending beyond Ada's own sales. By comparison, the audit threat coincided with a much smaller price reduction while sellers continued testing prices and discussing market information.

In this pilot, trusted competition was associated with substantially lower prices than threatened enforcement. This establishes the governance as ineffective. Further experiments should test whether profit forfeiture would produce a stronger response from the profit-seeking agents.

Code and Data

Project repository: <https://github.com/rohansbansal/collusion>.

LLM Usage Statement

OpenAI Codex assisted with implementation, experiment orchestration, data analysis, figures, and drafting this report. GPT-6 Luna via OpenRouter generated buyer and seller decisions and served as the audit and research reviewer. Programmatic checks verified run length, probability settings, message-log correspondence, and reported metrics.