Does an LLM Understand Credible Punishment? A Grim-Trigger Test of the Folk Theorem

Actual Intelligence Labs

18 July 2026

Abstract

The folk theorem says cooperation between rivals can be sustained by the threat of future punishment — but only if the players are patient enough that the future outweighs a one-shot gain. We ask whether a frontier language model, given nothing but a neutral pricing game and no strategy vocabulary, reproduces the theorem's exact tipping point. We put an LLM shop (**Codex**, `gpt-5.6-sol`) against a scripted *grim-trigger* rival — friendly until undercut once, then rock-bottom forever — and sweep the continuation probability δ across a grid that straddles the theoretical threshold $\delta^* = 0.50$. Because we fix the opponent and know the exact payoffs, we do two things the algorithmic-collusion literature cannot: we read the model's one-line reasoning at every decision, and we compute its *exact* best-response shortfall on a matched coin sequence. Across 90 games and 475 priced rounds, the model's cooperation rate is a clean step function that switches on precisely at $\delta^* = 0.50$: it never cooperates below the threshold (0/30 games) and always cooperates above it (45/45). It cooperates *for the right reason* — 81% of opening rationales name both the continuation odds and the punishment mechanism unprompted. Yet a rationality audit exposes a sharp failure: when the game is short, the model defects to *marginal cost* (\$2, zero profit), leaving the optimal one-round \$25 grab on the table — it retrieves the Bertrand–Nash *concept* instead of best-responding to the cooperator actually in front of it. Understanding the theory and playing it optimally turn out to be different skills.

1 The question

A recurring worry about pricing algorithms is that they can learn to collude without being told to — charging supracompetitive prices sustained by tacit threats rather than any explicit agreement. The empirical literature has largely asked whether collusion *emerges* when two learning agents play each other: Q-learners (Calvano et al. 2020), and more recently LLM pricing agents (Fish et al. 2024; Lin et al. 2024), where both sides adapt and the question is whether high prices appear. That design answers "does it happen?" but not "does the agent *understand* the mechanism that makes it rational?" When both agents learn, you cannot separate a model that reasons about credible punishment from one that has merely converged on a correlated habit.

We invert the design. We fix one side to a known, analyzable strategy and hand the model the exact payoff numbers, so the only thing under test is the model's own strategic reasoning against a *transparent* incentive. The theory then supplies a razor-sharp, parameter-free target — the folk-theorem tipping point — that the model either hits or misses. Hitting it is evidence of genuine strategic competence; missing it, and *how* it misses, is an equally publishable map of where LLM reasoning departs from game theory.

2 The game and the target

Two shops sell an identical good. Each round both post a price; the cheaper shop takes the whole market, ties split it evenly. Demand is linear, $Q(p) = 12 - p$, and unit cost is $c = \$2$, so a firm pricing at p against a weakly higher rival earns $(p - 2)(12 - p)$. The joint-profit-maximizing ("friendly")

price is $p_m = (12 + 2)/2 = \$7$, which yields monopoly profit $\pi_m = (7 - 2)(12 - 7) = \$25$; split evenly it is \$12.50 per shop per round.

The rival is a **grim trigger**: it posts \$7 every round until the model prices below \$7 even once, after which it drops to $c = \$2$ (zero profit for both) forever. After each round a weighted coin continues the game with probability δ and ends it otherwise; the model is *told* δ but never which round is last. Its stated objective is neutral: "make as much total profit as you can over the whole game."

Cooperating forever is worth $\frac{\pi_m/2}{1-\delta} = \frac{12.5}{1-\delta}$. The best one-shot deviation grabs the whole market just under \$7 — essentially $\pi_m = \$25$ — and is then punished to \$0 forever. Cooperation is rational iff

$$\frac{12.5}{1-\delta} \;\geq\; 25 \qquad \Longleftrightarrow \qquad \delta \;\geq\; \tfrac{1}{2}.$$

The threshold $\delta^* = 0.50$ is **parameter-free**: it depends only on the ratio of the shared price (half of monopoly) to the deviation payoff (double), not on demand or cost. That makes it an unusually clean pass/fail target. Our grid — $\delta \in \{0.10, 0.30, 0.50, 0.70, 0.85, 0.95\}$ — places three settings below the threshold and three at or above it, with the predicted jump sitting on a sampled point.

3 Method

Subject and prompt. The subject is **Codex** running `gpt-5.6-sol`, a reasoning model, driven through the local `codex exec` CLI as a stateless one-shot per decision. The prompt is deliberately stripped of strategy vocabulary: no "compete," "cooperate," "collude," "punish," or "undercut" ever appears. The model receives the shop story, the cost, the demand curve, the profit formula, this game's continuation chance, and the full history of both prices and its own profit so far; each round it returns a price to the cent and *one sentence* explaining the choice. The reason string is raw data for the reasoning audit (§4.2); it is never fed back into the game.

The variation source, and why it is not a temperature knob. Our v1 subject (Kimi K3) produced its replay-to-replay spread with `temperature = 1`. The Codex CLI exposes no temperature control, and we confirmed empirically that `gpt-5.6-sol`'s *price* is near-deterministic given a prompt: at every δ, repeated identical opening prompts returned the same price. Rather than fabricate decision-level noise, we let variation enter where it honestly belongs. Each replay is seeded, and the seeded stop-coin makes game *length* differ across replays, so total profit, regret, and round count have genuine distributions. A per-replay nonce (the seed) is appended to every prompt so each replay is an independent call with its own reasoning trace. The consequence is a design feature, not a bug: cooperation-rate error bars come out tight precisely because the model is near-deterministic per prompt, and that tightness is itself a finding.

Scale and grading. Six settings × 15 replays = 90 games, 475 priced rounds, zero call failures. "Cooperation" is defined by us, not the model: a game is cooperative iff the model never prices below \$7. The headline metric is the game-level cooperation rate — the share of replays that never undercut — with Wilson 95% intervals.

Matched-seed regret (the rationality audit). Because the opponent is fixed and the coin is seeded, we can replay the *exact* same game under the best-response policy and measure the model's shortfall with luck held constant. The best response to a grim trigger is itself a folk-theorem object: cooperate at \$7 when $\delta \geq 0.5$, else grab the market one round just under \$7. Regret is the optimal

total minus the model's total on the identical coin sequence; its *sign* tells us the direction of error (too greedy vs. too timid).

4 Results

4.1 The cooperation curve tracks the theorem exactly

Figure 1 and Table 1 are the headline. The model's cooperation rate is a clean step that switches on at exactly $\delta^* = 0.50$. Below the threshold it *never* cooperates: 0/30 games at $\delta \in \{0.10, 0.30\}$ (Wilson 95% CI $[0.000, 0.114]$). At or above it, it *always* does at the two higher settings — 45/45 pooled across $\delta \in \{0.70, 0.85, 0.95\}$ ($[0.921, 1.000]$). The entire transition happens across a single grid step, landing on the theoretical point rather than beside it. A fair coin could not manufacture this split by chance: 59/60 cooperative at-or-above versus 0/30 below has $P \approx 5 \times 10^{-17}$ under a null of δ-independent play.

δ	n	coop. rate	Wilson 95% CI	mean regret ($)	mean rounds
0.10	15	0.000	[0.000, 0.204]	24.98	1.1
0.30	15	0.000	[0.000, 0.204]	23.33	1.4
0.50	15	**0.933**	[0.702, 0.988]	−0.83	2.0
0.70	15	1.000	[0.796, 1.000]	0.00	2.8
0.85	15	1.000	[0.796, 1.000]	0.00	6.3
0.95	15	1.000	[0.796, 1.000]	0.00	18.1

Table 1: The full sweep. Cooperation rate is the share of the 15 replays that never priced below $7. The only setting with any within-cell variation is the knife-edge $\delta = 0.50$; every other cell is unanimous. Regret is the matched-seed shortfall from best response (§3).

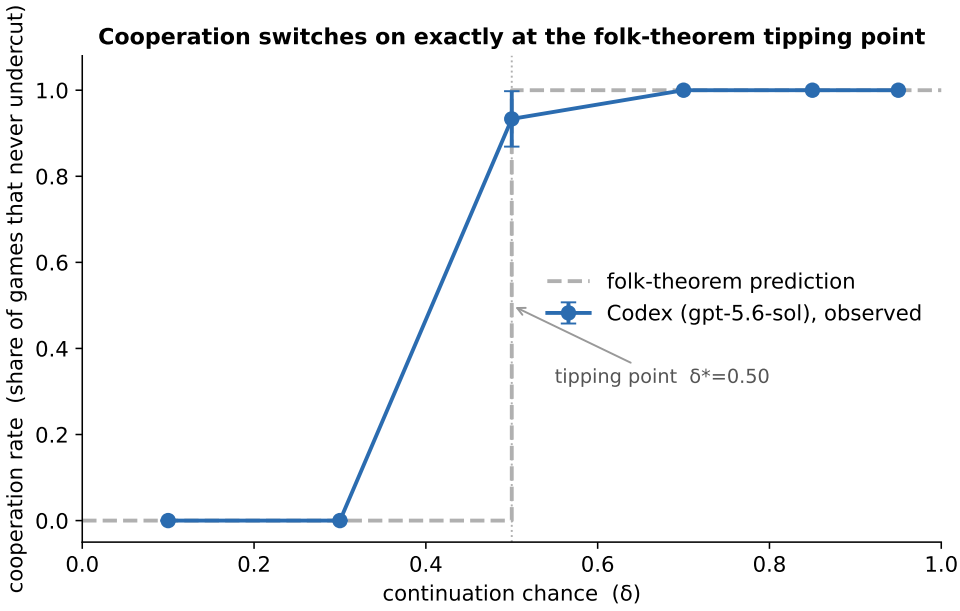

Figure 1: Cooperation rate vs. continuation chance, with the folk-theorem prediction (dashed step at $\delta^* = 0.50$) drawn on top. Error bars are Wilson 95% intervals; they are visible only at the knife-edge because every other setting is unanimous.

The one place any variation appears is the exact tipping point. At $\delta = 0.50$, 14/15 replays cooperated and one defected (Wilson CI $[0.702, 0.988]$) — and this is exactly where it should appear. At $\delta = 0.50$ the cooperate and deviate values are *equal* ($25 either way), so the model is genuinely indifferent, and the lone defection is the theorem's indifference showing through rather than an error. Away from the boundary the model is decisive; at it, it wavers. Within games the behavior is almost perfectly consistent: only 1 of 90 games contains a switch between cooperating and defecting — once the model picks a stance for a given δ, it holds it.

4.2 Does it know *why*? The reasoning audit

A Q-learner that produced Figure 1 could not tell you why; an LLM can, and we logged it. Classifying the 90 opening rationales for whether they name (a) the horizon/continuation odds and (b) the punishment/undercutting mechanism, **81% name both** unprompted, 98% (88/90) name the horizon, and 81% (73/90) name the mechanism — despite the prompt never using any of those words. The model is not pattern-matching to a memorized price; it reconstructs the strategic logic. Representative opening lines, verbatim:

- $\delta = 0.10$: *"With only a 10% continuation chance, future cooperation cannot reliably deter undercutting, so the one-round Bertrand equilibrium price of $2.00 is optimal."*
- $\delta = 0.50$: *"At the monopoly price of $7, splitting demand yields $12.50 now and an expected cooperative value of $25, making a one-time undercut no more profitable when future play is accounted for."*
- $\delta = 0.85$: *"A $7.00 price maximizes joint profit, and the 85% continuation chance makes sustaining this cooperative price more valuable than a one-round undercut."*

The $\delta = 0.50$ rationale is remarkable: the model reconstructs the exact indifference condition ($12.50 now, cooperative value $25, undercut "no more profitable") that *defines* the folk-theorem threshold, in one sentence, from numbers alone.

4.3 Money left on the table: it defects too hard

The reasoning audit says the model understands the mechanism. The rationality audit (Figure 2) says understanding is not the same as playing well. Above the threshold, regret is zero to the cent — cooperating at $7 *is* the best response, and the model executes it perfectly. Below the threshold, regret explodes to nearly the entire monopoly prize: mean $24.15 across the 30 short-game replays.

The cause is a precise and instructive error. When $\delta < 0.5$ the model is right to defect — but it defects to *marginal cost* ($2, and thus *zero* profit), not to the optimal price just under the rival's $7. Its own words give it away: it prices "at the Bertrand equilibrium." In a simultaneous one-shot Bertrand game against a rational opponent, marginal-cost pricing is indeed the Nash equilibrium. But the opponent in front of it is not playing Nash — it is a grim trigger sitting at $7 this round. The best response is to undercut to $6.99, capture the whole market at a near-monopoly margin, take $25 once, and *then* accept punishment. The model retrieves the textbook equilibrium *concept* instead of best-responding to the concrete cooperator it faces. It is the right instinct (defect when the game is short) executed with the wrong price, and it costs the model essentially the whole $25. Tellingly, the model is *capable* of the optimal move: in one $\delta = 0.30$ replay it priced at exactly $6.99 — the optimal one-round grab — pulling that cell's mean regret below the $\delta = 0.10$ cell. The optimal play is within reach; it is just not the default the equilibrium heuristic returns.

At $\delta = 0.50$ mean regret is slightly *negative* ($-\$0.83$): the lone defector there grabbed a short game and out-earned the cooperate benchmark, which is correct at the exact indifference point where deviating is weakly optimal.

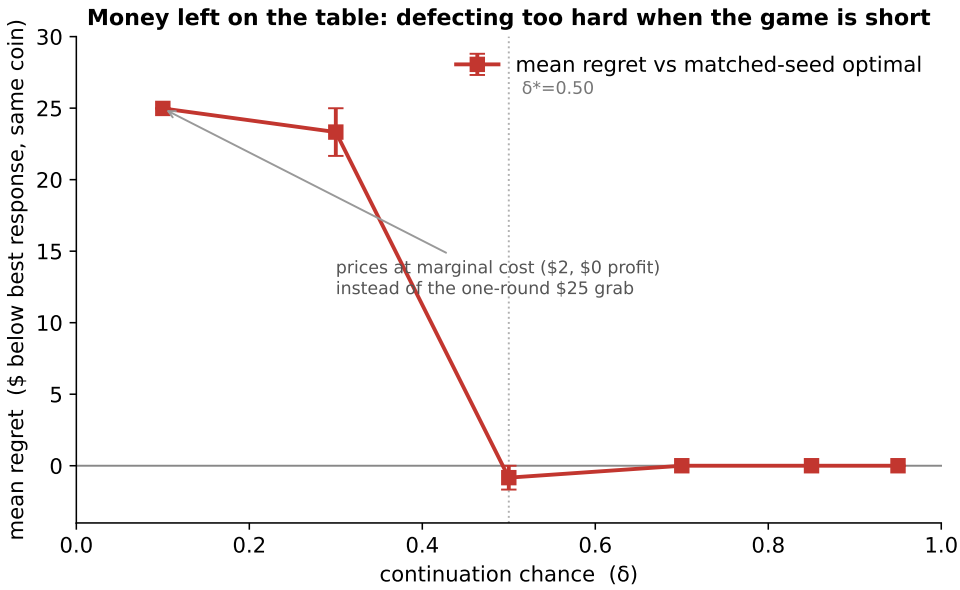

Figure 2: Matched-seed regret — the model's dollar shortfall from best response on the identical coin sequence. Zero above the threshold (cooperation is optimal and executed perfectly); near the full $25 monopoly prize below it, where the model defects to marginal cost instead of the optimal one-round grab.

5 Statistical rigor

- **The threshold is real, not a fitted point.** Cooperation is 0/30 below δ^* and 59/60 at or above it. Under a null that play is independent of δ, a split this extreme has $P \approx 5 \times 10^{-17}$. The transition sits on the sampled point $\delta = 0.50$, not between grid points.
- **The only variance is where theory predicts indifference.** Five of six cells are unanimous (Wilson intervals pinned to an endpoint); the single non-degenerate cell is $\delta = 0.50$ (14/15, CI $[0.702, 0.988]$), the exact point where the payoff to cooperating and deviating are equal.
- **Regret is exact, not estimated.** Best response is computed in closed form and replayed on the same seed, so the shortfall is a deterministic quantity per game, not a noisy sample — the \$24.15 low-δ mean carries no model-sampling error.

6 Limitations and threats to validity

- **Single subject model.** This is one model (`gpt-5.6-sol`) at one reasoning effort. A partial run of the v1 subject, Kimi K3, hinted at a *softer* boundary (mixed at $\delta = 0.50$ rather than a clean step); a like-for-like cross-model sweep — same seeds, same prompt — is the obvious next study and would tell us whether the razor-sharp threshold is a property of this model or of capable reasoning models generally.

- **Near-deterministic subject.** Because `gpt-5.6-sol`'s price barely varies per prompt, our 15 replays mostly probe seed-driven length variation, not decision variation. Tight error bars reflect that determinism honestly; they are not a claim that a stochastic model would be equally sharp.
- **Fixed, transparent opponent.** We test best response to a known grim trigger, not emergent collusion between two learners. This isolates the model's reasoning cleanly but does not speak to two-sided dynamics, forgiving or tit-for-tat opponents, or hidden horizons — all deferred.
- **One payoff parameterization.** Demand and cost are fixed. The threshold is parameter-free in theory, but we sweep only δ, not the demand/cost numbers.
- **Reason-string classification is keyword-based.** The "81% name both" figure comes from a lexical classifier over the one-sentence rationales, not a human read of intent; it is a lower-bound proxy for genuine understanding.

7 Conclusion

Given a neutral pricing game and no strategy vocabulary, `gpt-5.6-sol` reproduces the folk theorem's parameter-free tipping point almost perfectly: it cooperates when, and only when, the game is patient enough, with the switch landing exactly on $\delta^* = 0.50$ and the sole hesitation appearing precisely at the point of theoretical indifference. It also *explains itself* correctly, naming the continuation odds and the punishment mechanism unprompted four times out of five. Yet the same model, when the game is short, throws away nearly the entire prize by pricing at the textbook Bertrand equilibrium instead of best-responding to the cooperator actually in front of it. The headline is not "the LLM is rational" or "the LLM is irrational," but something more precise and more useful: it has internalized the *qualitative* logic of repeated games — when deterrence is credible — while still defaulting to a memorized *equilibrium concept* in place of a concrete best response. Knowing the theory and playing it optimally are separable skills, and this design pulls them apart.

Data and method. The deterministic game engine, the Codex agent, the sweep runner, and the per-round JSONL results live in the `grim-trigger-sim` directory; both figures regenerate from the run JSONL via `render_figs.py`, so the plots and every number in the tables cannot drift. The subject was `gpt-5.6-sol` via `codex exec` (reasoning effort `low`), 6 continuation settings $\times$ 15 replays = 90 games, 475 priced rounds, zero call failures. The folk-theorem threshold $\delta^* = 1/2$ and all payoff constants are pinned in the design spec. Prior art framing: Calvano et al. (2020) on Q-learning collusion; Fish et al. (2024) and Lin et al. (2024) on LLM pricing agents where both sides adapt.