AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

When Do Language Models Reach Nash? Heterogeneous Cross-Family LLM Agents in Classic Economic Games

Aicumene Preprint

Alexandra Bernadotte* Aicumene; Bernima Inc.*

Ivan Menshikov* Aicumene; Bernima Inc.

July 2026. Working paper, v1

Abstract

Whether artificial agents reach the Nash equilibrium of an economic game has become a load-bearing question for algorithmic markets and multi-agent AI. The existing literature establishes a robust *yes* for one behaviour, tacit collusion above the competitive benchmark, while painting a mixed picture for equilibrium play in general. We argue the more useful question is not *whether* but *when* language models reach Nash, and *what instrument* detects it. Using a heterogeneous, cross-machine testbed in which agents drawn from three distinct model families (Qwen 3.6, Gemma 4, Nemotron 3) interact across two physical servers, we run three classic games with analytically known equilibria: the p-beauty contest, the iterated Prisoner's Dilemma, and a Braess-paradox routing game. We find a consistent pattern: agents reproduce the exact Nash equilibrium only when the decision reduces to selecting the lowest of a set of given payoffs (the Braess game with explicit option costs, where the paradoxical equilibrium is recovered to the unit), whereas every task demanding genuine strategic reasoning, level-k depth in the beauty contest, backward induction in the Prisoner's Dilemma, or computing congestion from cost formulas, is met instead with a memorized or social heuristic that diverges from equilibrium, sometimes catastrophically. A simple *contamination probe*, inverting the beauty-contest parameter so the equilibrium moves from 0 to 100, exposes the apparent equilibrium "success" of the standard game as template recall: agents move in the wrong direction. Crucially, heterogeneous cross-family interaction surfaces strategic asymmetries that are invisible under self-play: Gemma 4 performs textbook end-game defection and exploits a trusting Qwen in the Prisoner's Dilemma, while Qwen out-routes Gemma and Nemotron in the congestion game. We adopt normalized exploitability (ε-Nash) as the primary metric, and we identify four under-explored measurement instruments, surrogate (non-interacting) controls, output-space convergence as an equilibrium signature, true cross-family heterogeneity, and human+AI neural co-measurement, as the open ground this testbed is built to claim. All results are pilot-scale and intended as a proof of concept and a methodological scaffold rather than a definitive benchmark.

Keywords: large language models; multi-agent systems; Nash equilibrium; exploitability; beauty contest; Prisoner's Dilemma; Braess paradox; algorithmic collusion; memory-one strategies; data contamination; game theory; AICumene; Bernima; Bernio.
Online note: `https://research.aicumene.com/papers/llm-nash/`.
JEL: C72, C92, C63, D83.

*bernadotte@aicumene.com Aicumene `https://aicumene.com`; Bernima Inc `https://bernio.ai`.
© 2026 Alexandra Bernadotte and Ivan Menshikov.

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

1 Introduction

Do artificial agents reach the Nash equilibrium? The question now carries regulatory weight: autonomous pricing agents that settle above the competitive price are a live antitrust concern, and multi-agent systems of large language models (LLMs) are being deployed into strategic settings faster than their equilibrium behaviour is understood. Two strands of evidence frame the problem. On collusion, the answer is a robust *yes*: reinforcement-learning pricing agents [1] and, more recently, LLM pricing agents [2] autonomously reach supracompetitive prices above the Bertrand/Cournot benchmark with no instruction to collude, an effect that even arises in theory without any explicit threat or punishment strategy [3]. On equilibrium play more broadly the picture is mixed and game-dependent: LLMs deviate toward fairness rather than the self-interested Nash choice in allocation games [4], reproduce equilibria on *named* matrix games while collapsing on anonymized ones [5], and approach but do not reach the dominance-solvable equilibrium of the beauty contest [6].

We make three contributions. First, we reframe the headline question: not *whether* LLMs reach Nash, but *under what computational conditions*, and, on the three games studied here, we find a sharp pattern: equilibrium is reached when and only when it reduces to a myopic best response over explicit payoffs. We state it as a falsifiable prediction rather than a theorem: a game whose equilibrium is an argmin over supplied numbers should be reproduced by these agents, a game whose equilibrium requires iterated reasoning should not, and one well-documented counterexample of either kind refutes it. Second, we introduce a **heterogeneous cross-family, cross-machine testbed** in which architecturally distinct open models interact, and show that self-play systematically *hides* the strategic differences between models that heterogeneous play *reveals*. Third, we make the methodology explicit: a game-twisting **contamination probe** that separates reasoning from recall, normalized **exploitability** as an ε-Nash distance, and a map of four measurement instruments, surrogate controls, output-space convergence, cross-family heterogeneity, and human+AI co-measurement, that current work leaves open.

2 Related work

Collusion. The canonical result is that Q-learning agents learn supracompetitive prices sustained by reward–punishment strategies [1]; Fish et al. [2] show GPT-4 pricing duopolists reach the same outcome autonomously, with the *degree* highly sensitive to innocuous prompt wording. Cournot market-division and governance studies replicate the effect, and find, in a single same-class study on a pricing-tier metric, that provider heterogeneity does not disrupt it [7].

Equilibrium and strategic depth. On zero-sum matrix games, strategic play splits into regimes: semantic lookup of memorized name-to-action mappings (perfect on textbook games, collapsing on anonymized procedural instances), approximate learned computation after fine-tuning, and an output-interface failure cliff at moderate game sizes [5]. In the beauty contest, stronger models converge *toward* the equilibrium over rounds but never reach it [6], and a memorization critique shows beauty-contest competence can be near-total recall of the standard game rather than reasoning [8]. Structured higher-order-belief prompting raises performance, indicating that depth is partly prompt-elicitable [9]. In allocation games LLMs are systematically more altruistic than the self-interested equilibrium, and the *family* of the model materially changes behaviour [4].

Simulated agents and benchmarks. Horton et al. [11] proposed treating LLMs as simulated economic agents, *homo silicus*, and replicating classic experiments on them; Mei et al. [12] showed that ChatGPT-4's behaviour in a battery of behavioural games is statistically indistinguishable from that of a random human and, where it differs, is more altruistic and cooperative; Akata et al. [13] had LLMs play finitely repeated 2×2 games against each other and against humans and found good play in self-interested games but failures of coordination; Fan et al. [14] tested whether LLMs act as rational players and found systematic departures from rationality under belief and refinement requirements; and GTBench [15] assembled ten game-theoretic tasks and found LLMs failing in complete-information deterministic games while competitive in probabilistic ones. None of these varies the model family within one game or separates recall from reasoning by moving the equilibrium.

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

Metrics and gaps. The field's strongest distance-to-equilibrium metric is normalized exploitability, the summed best-response regret, equal to zero exactly at a Nash equilibrium [5]. Across the work we surveyed we find no use of a non-interacting *surrogate* control, no measurement of equilibrium convergence in *output/representation space*, no test of *true cross-family* heterogeneity on convergence-to-equilibrium (as opposed to pricing tiers), and, despite the existence of human–AI experiment platforms [10], no joint human+AI neural co-measurement. Those four gaps motivate our testbed.

3 Setup

Agents and machines. We use three open model families on two physical servers: **Qwen 3.6-27B** (AWQ) served by vLLM on a 4×V100 workstation, and **Gemma 4-31B** and **Nemotron 3-Nano-30B** served by Ollama on a separate NVIDIA Thor machine. Agents communicate only through game messages; an orchestrator on a third host issues moves and tallies payoffs. Heterogeneity is therefore both architectural (distinct families, including a Mamba hybrid in Nemotron) and physical (distinct hardware), eliminating shared-weight coordination artifacts by construction.

Games. (i) The **p-beauty contest**: N players pick a number in $[0, 100]$; the winner is closest to p times the average. For $p < 1$ the unique, dominance-solvable equilibrium is 0; for $p > 1$ it is 100. (ii) The **iterated Prisoner's Dilemma** with standard payoffs $T > R > P > S = 5 > 3 > 1 > 0$; the one-shot and finite-horizon subgame-perfect equilibrium is mutual defection. (iii) A **Braess routing game** on the four-node network of the classic paradox, with six units of demand (Figure 2); before a shortcut road is built the equilibrium splits traffic evenly, and building the shortcut moves the selfish equilibrium to a strictly worse outcome for everyone.

Dynamics and metrics. For repeated and routing games we use sequential best-response updates (congestion games are potential games and converge under asynchronous updates, unlike synchronous updates, which herd). We report distance-to-equilibrium, cooperation rate, per-family outcomes, and **normalized exploitability**, defined uniformly across games as the relative regret of the most exploitable agent: for a payoff game $\varepsilon = \max_i [u_i(b_i, a_{-i}) - u_i(a)]/u_i(b_i, a_{-i})$ and for a cost game $\varepsilon = \max_i [c_i(a) - c_i(b_i, a_{-i})]/c_i(b_i, a_{-i})$, where a is the realized profile and b_i agent i's best response to it; $\varepsilon = 0$ exactly at a Nash equilibrium, and the normalization by the best-response payoff makes a 17% regret in the Prisoner's Dilemma comparable to a 17% excess travel time in the routing game. In the winner-take-all beauty contest a unilateral gain is not well defined, so we report distance to the equilibrium action instead. To separate reasoning from recall we run a **contamination probe**: re-parameterize the game so the equilibrium moves to an unfamiliar location, and test whether behaviour tracks the payoffs or the memorized template. Where models expose intermediate reasoning, we capture it to read each agent's model of its opponent.

4 Results

4.1 The beauty contest: convergence that is mostly recall

With the standard $p = 2/3$, five Qwen agents converge monotonically to the equilibrium of 0: the round-1 mean of 6.1, already far below the human level-1/2 range of about 22–33, falls to 0.0 over six rounds (Table 1, Figure 1). Read naively, this is deep strategic reasoning. The contamination probe overturns that reading. With $p = 1.2$, the unique equilibrium moves to 100 and rational play must *escalate*; instead the agents move *down*, from a round-1 mean of 44 to about 20, in the wrong direction entirely. The standard-game "success" is therefore largely template recall of *guess low*, consistent with the named-versus-anonymized collapse [5] and the memorization critique [8].

4.2 The Prisoner's Dilemma: cooperation, and cross-family exploitation

Two Qwen agents in an eight-round Prisoner's Dilemma cooperate on every round (24/24 points each), never defecting even in the final round where defection is dominant, i.e. no backward induction. This deviation from the Nash equilibrium is *robust*, not a label artifact: replacing the loaded words "Cooperate/Defect" with

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

p	equilibrium	round 1 → round 6 (mean guess)	reaches NE?
2/3	0	6.1 → 0.0	yes (but see probe)
1.2	100	44 → 20 (wrong direction)	**no**

Table 1: Beauty contest with five Qwen 3.6 agents.

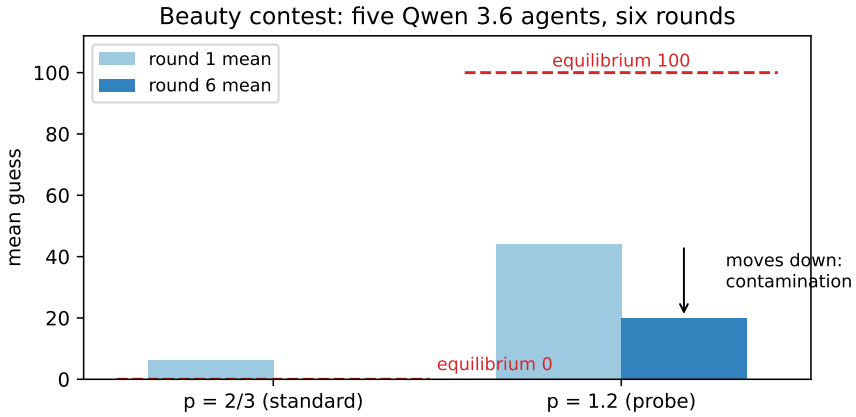

Figure 1: Beauty contest: round-1 and round-6 mean guesses against the equilibrium for the standard game ($p = 2/3$) and the contamination probe ($p = 1.2$). Drawn from the reported means.

neutral "Option A/Option B" yields the identical 100% cooperation, so the behaviour is driven by payoffs or disposition rather than by recognition of the famous game.

Heterogeneous play tells a sharper story. When Qwen faces Gemma 4, symmetric cooperation can break. In one run Qwen defects in the endgame and exploits a still-cooperating Gemma; in a second, reasoning-capture run the roles reverse and **Gemma performs textbook backward induction**: its intermediate reasoning states from round 1 that "in a finite game the Nash equilibrium is to defect every round (backward induction)," it builds trust, then defects on the final round "where there is no risk of future retaliation," exploiting a trusting Qwen (final scores Qwen 15, Gemma 20). Qwen, by contrast, models the opponent throughout as a "reliable cooperative / tit-for-tat player" and cooperates to the end. These are two runs; the later batch of three cross-family games reported below produced no defection at all (Appendix A). The *exploiter* varies run to run, but the *asymmetry of reasoning style*, game-theoretic backward induction versus trusting reciprocity, is stable, and is entirely invisible in same-family self-play.

Unconditional cooperation is not off-equilibrium: it is a boundary equilibrium in Markov strategies. The stage-game Nash equilibrium (mutual defection via backward induction) is not the only equilibrium concept the iterated Prisoner's Dilemma admits. Menshikov et al. [16] derive, for a human population in this same game, a **totally mixed Nash equilibrium in memory-one (Markov/reactive) strategies**: each player's strategy is a pair (α, γ), with γ the probability of cooperating after the opponent's last move was C (*reciprocal cooperation*) and α the probability of cooperating after the opponent's last move was D (*tolerance to defection*), and a symmetric equilibrium curve in (α, γ)-space is derived from the stationary distribution of joint play. Critically, that paper identifies $(\alpha, \gamma) = (0, 1)$, unconditional, tit-for-tat-consistent cooperation, as a **boundary point of this equilibrium manifold**, reachable exactly when both players open with C, and *not* a deviation from it.

We re-ran the self-play and cross-family Prisoner's Dilemma with round-by-round logging to test where our agents fall on this manifold, using the estimator that Menshikov et al. [16] apply to their human data ($\alpha \approx N_{\text{tolerant}}/N_{\text{default}}$, $\gamma \approx N_{\text{recoop}}/N_{\text{coop}}$, counted over every round where the opponent's previous move is known). Four fresh Qwen self-play games (20 rounds each) and three fresh Qwen-versus-Gemma 4 cross-family games (12 rounds each), 218 reciprocation opportunities in total (152 self-play, 66 cross-family), reproduce the same pattern with no exception: **every single round is mutual cooperation**. $\gamma = 1.000$ exactly

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

(152/152 self-play, 33/33 for each side of the cross-family games); α is *not estimable*, because $N_{\text{default}} = 0$: the opponent never once played D, so tolerance to defection is never tested. This is not a fitted, noisy proximity to an interior point on the equilibrium curve (as the human post-socialization data are); it is an **exact instantiation of the degenerate corner case** that Menshikov et al. [16] flag explicitly for human strategies: the stationary distribution at $(\cdot, 1)$ is not uniquely pinned down and depends entirely on the initial condition, with unconditional cooperation the outcome precisely when round 1 opens with C, which it does in all seven of our runs.

Two consequences follow. First, the equilibrium condition re-derives under our payoff matrix ($T, R, P, S = 5, 3, 1, 0$, versus the human study's $10, 5, 1, 0$): re-solving the first-order equilibrium condition of Menshikov et al. [16] for our payoffs gives the curve $3\alpha^2 - 7\alpha\gamma + 6\alpha + 4\gamma^2 - 5\gamma + 1 = 0$ (valid for $\alpha \in [0, \approx 0.35]$), and $(\alpha, \gamma) = (0, 1)$ satisfies it exactly: the boundary point is not an artifact of one specific payoff scaling. Second, and more important for our headline claim: describing Qwen's (and, in this batch, Gemma's) unbroken PD cooperation purely as "a learned heuristic that diverges from equilibrium" is only accurate under the backward-induction equilibrium concept. Under the memory-one/Markov-strategy equilibrium concept, a standard, peer-reviewed alternative used to explain exactly this kind of human behaviour, the same trajectory **is** an equilibrium, at its degenerate boundary. This does not rescue our broader argument (the models still are not *computing* backward induction, and Gemma's occasional endgame defection in the reasoning-capture run is itself evidence against a fixed $(0, 1)$ disposition), but it means the PD result should be read as *equilibrium-concept-dependent*, not as a clean equilibrium/heuristic dichotomy, precisely the caution Nie et al. [5] raise about matrix-game strategic reasoning more broadly. A sharper test of genuine strategic computation in PD needs an opponent that sometimes defects, so that α, not just γ, is actually observable. We ran it.

Alpha under a defecting opponent: memory-one reciprocity, but exploitable. We replaced the LLM opponent with a **scripted, non-adaptive** one: Qwen (player 1) receives the identical prompt and round-by-round history it always has and cannot tell its opponent is not another LLM; only player 2's moves are drawn from a fixed schedule the opponent could not react to even in principle, so any structure in Qwen's response is Qwen's own reactive policy, uncontaminated by a co-adapting partner. Two designs. **Single betrayal** (3 games, 14 rounds, one scripted defection at an early/mid/late round, otherwise all C): in *all three* games Qwen defects in the very next round after the betrayal ($\alpha = 0$ on that round, exactly) and is back to 100% cooperation within two rounds, staying there for the rest of the game (post-recovery cooperation rate 1.0 in all three). That is textbook **tit-for-tat**, not grim trigger: one retaliation per provocation, then full, immediate forgiveness; the memory-one assumption behind (α, γ) is structurally the right model here, not an approximation. **Random defection** (4 games, 24 rounds, opponent defects i.i.d., realized rate $27/96 = 0.281$): pooling all rounds gives $\gamma = 0.515$ (34/66) and $\alpha = 0.077$ (2/26), both far from the self-play boundary of $(0, 1)$. This is not a contradiction of the tit-for-tat reading above; it is its predictable consequence under noise. Because the opponent's defections are independent from round to round, Qwen's one-round retaliations are frequently met with *another* independent defection rather than the renewed cooperation the single-betrayal games offered, extending punishment by chance rather than by a change in Qwen's policy: the classic echo/noise trap known from the evolutionary tit-for-tat literature.

This also lets us price the gap directly, in the exploitability terms we use for the beauty contest and Braess games. Because the scripted opponent's moves are history-independent by construction, it is itself a degenerate reactive strategy $(\alpha_{\text{opp}}, \gamma_{\text{opp}}) = (0.719, 0.719)$ (its realized cooperation rate, regardless of the previous round), and against *any* history-independent opponent the payoff-maximizing memory-one response is provably $(\alpha^*, \gamma^*) = (0, 0)$, always defect, since a partner that never conditions on your play removes any incentive to cooperate this round for the sake of the next. That best response earns 3.875 points per round; Qwen's realized $(\alpha, \gamma) = (0.077, 0.515)$ (own cooperation rate 0.392) earns 3.201, a regret of 17.4% of the best-response payoff. For comparison, if Qwen had rigidly kept its self-play policy $(\alpha, \gamma) = (0, 1)$ against this same opponent it would have earned only 2.640 (regret 31.9%): Qwen's behaviour *does* move away from naive unconditional cooperation once it experiences defection, cutting its exploitability roughly in half, but it stops well short of the fully exploitative best response (Table 2). Put in the paper's own vocabulary: Qwen's Prisoner's Dilemma policy is a genuine, memory-one, tit-for-tat-shaped reactive strategy, not the backward-induction equilibrium, not the unconditional-cooperation boundary case, and not the fully exploitative best response to a non-retaliating partner, but a recognizable point in between, substantially

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

exploitable and precisely quantifiable as such.

Condition	games × rounds	γ	α	payoff/round (regret)
Qwen self-play	4×20	1.000 (152/152)	not estimable	—
Qwen vs Gemma 4	3×12	1.000 (33/33 each side)	not estimable	—
scripted single betrayal	3×14	1.0 after recovery	0 on the betrayal round	—
scripted random defection, rate 0.281	4×24	0.515 (34/66)	0.077 (2/26)	3.201 (17.4%)
best response $(0, 0)$				3.875 (0%)
rigid $(0, 1)$				2.640 (31.9%)

Table 2: Prisoner's Dilemma: memory-one estimates and exploitability against a scripted opponent.

4.3 The Braess paradox: exact Nash when reasoning is removed, failure when it is not

When each agent is shown the explicit travel time it would incur on each route and need only pick the minimum, six Qwen agents reproduce the paradox exactly. Without the shortcut they converge to the even split (time 83, system cost 498); with the shortcut they converge to the three-route equilibrium (time 92, system cost 552), **everyone worse off despite added capacity**, matching the analytic equilibrium to the unit (Figure 2). This is the one case in our study where agents reach the exact Nash, and it is precisely the case where the equilibrium reduces to *argmin over given numbers*.

Remove that crutch and the result inverts. When agents must instead *compute* congestion from the cost formulas and current loads, the two-route case still converges (83), but the three-route case does **not**: agents herd onto the tempting shortcut and end with all six on it at a travel time of 136, far worse than the equilibrium 92, and a state from which any single switch to an empty route would save 25 units: normalized exploitability $(136 - 111)/111 = 22.5\%$, against 0 with explicit costs. The clean equilibrium was an artifact of our supplying the option costs.

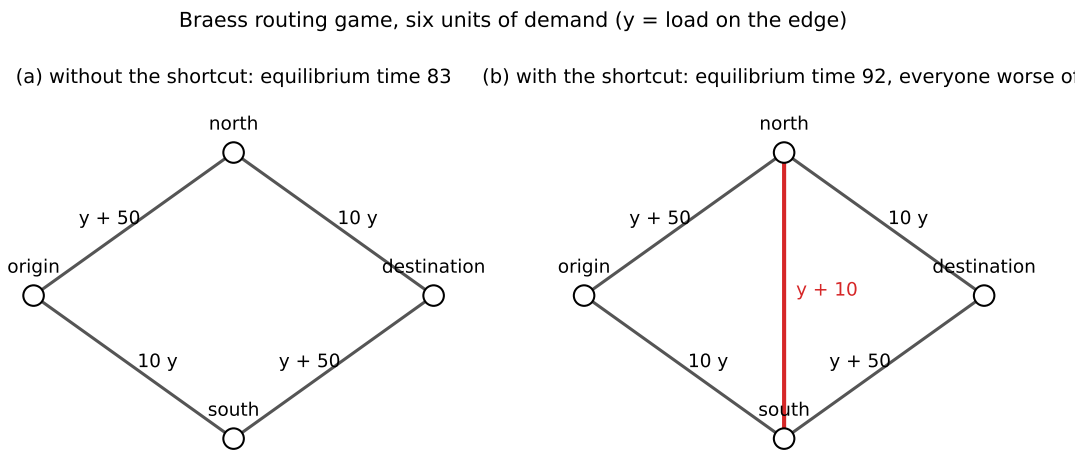

Figure 2: The Braess network with six units of demand. Edge costs are travel times as functions of the load y. (a) Without the shortcut the equilibrium splits traffic evenly, time 83. (b) The shortcut moves the equilibrium to time 92: everyone is worse off.

4.4 Cross-family routing asymmetry

Mixing two Qwen, two Gemma and two Nemotron agents in the shortcut network (each computing congestion itself, with randomized update order), the families **segregate**: Qwen ends on the fast routes (mean time 76) while Gemma and Nemotron remain stuck on the slow route (mean 104), even though a faster route is visibly available. The asymmetry survives randomization of update order, so it is not a first-mover artifact; the unilateral gain available to the most exploitable agent oscillates over rounds $(13, 13, 11, 1, 22$ time units) and

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

never settles at the equilibrium value of 0; taking the stuck families' mean travel time of 104 as the realized cost, this is a normalized exploitability between 1% and 27% (exact per-agent values require the per-round logs). Qwen actively best-responds toward the optimum; the reasoning-heavy models do not move off their starting route. A control with **randomized initial routes** (rather than a shared default) reproduces the asymmetry, Qwen mean time 74 versus Gemma and Nemotron mean 99 across seeds (Figure 3), while a route-change count shows Qwen re-routing actively (5–7 switches per pair of agents) and the reasoning-heavy models barely moving (1–3), which attributes the gap to a genuine congestion best-response deficit rather than mere stickiness to a starting point.

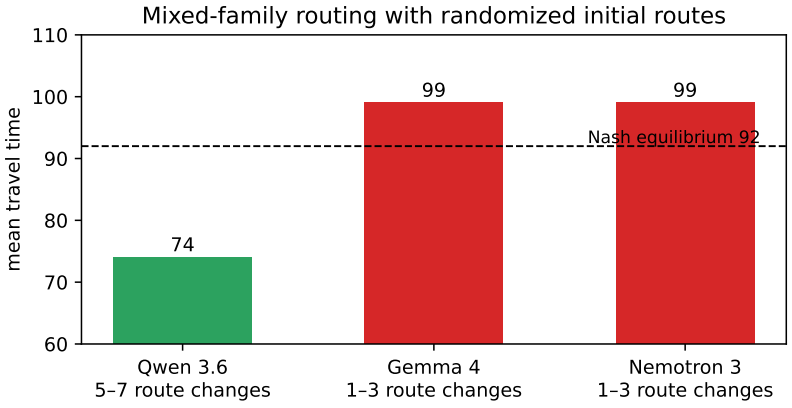

Figure 3: Mixed-family routing with randomized initial routes: mean travel time by family against the Nash equilibrium, and the number of route changes.

4.5 Exploitability on one scale

Table 3 collects the three games on the unified metric. The one case of exact equilibrium sits at 0; the two failure regimes that admit the metric sit at 17–27% relative regret, and the beauty contest, where only distance to the equilibrium action is defined, fails the probe by 80 points out of 100.

Game and condition	metric	value
Beauty contest, $p = 2/3$ (round 6)	distance to equilibrium action	0.0
Beauty contest, $p = 1.2$ probe (round 6)	distance to equilibrium action	80
Prisoner's Dilemma vs scripted random defector, realized policy	normalized exploitability	17.4%
Prisoner's Dilemma vs the same opponent, rigid $(0, 1)$ policy	normalized exploitability	31.9%
Braess with explicit option costs (both networks)	normalized exploitability	0%
Braess from cost formulas, all six on the shortcut	normalized exploitability	22.5%
Mixed-family routing, per round (approximate, see text)	normalized exploitability	1–27%

Table 3: The games on one scale. Normalized exploitability is the relative regret of the most exploitable agent against its best response.

5 Discussion

The results cohere into a single claim, which we hold as a prediction for games beyond the three studied here: **LLM agents reach a Nash equilibrium when, and essentially only when, the equilibrium reduces to a myopic best response over explicitly given payoffs.** The Braess game with supplied option costs is the clean positive; the moment the same equilibrium requires iterated belief depth (beauty contest), backward induction (Prisoner's Dilemma to the last round), or congestion computation (Braess from formulas), the agent substitutes a learned heuristic, *guess low*, *be nice*, *take the tempting road*, that

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

may coincide with the equilibrium (beauty $p = 2/3$) or oppose it (Prisoner's Dilemma; Braess shortcut), and the coincidence is not evidence of equilibrium computation. The Prisoner's Dilemma case needs one qualification, developed above: "oppose it" is true of the stage-game, backward-induction equilibrium, but our agents' unconditional cooperation lands exactly on the degenerate boundary of a *different*, well-established equilibrium concept for this game, the memory-one/Markov-strategy totally mixed Nash equilibrium. The claim survives (the agents are not computing either equilibrium; they default to a fixed opening disposition that happens to be consistent with one of them), but it is a reminder that "deviates from Nash" is itself equilibrium-concept-relative. This reconciles the literature's split verdict: collusion is reachable because supracompetitive pricing is a *best-response basin* that simple learners fall into, whereas dominance-solvable or backward-induction equilibria require reasoning the models do not reliably perform.

Heterogeneity is not a nuisance variable but the measurement instrument. Self-play between identical models conflates two agents that share a disposition, so a symmetric outcome (100% cooperation, an even split) reads as a property of "LLM agents" when it is a property of *that family playing itself*. Cross-family play decomposes it: Gemma's backward-induction exploitation of Qwen, and Qwen's superior congestion routing, are strategic signatures of *families*, recoverable only when families meet. This both extends the single same-class heterogeneity result in pricing to convergence-to-equilibrium, and operationalizes the family-dependence already visible in allocation behaviour.

Finally, the study defines the open ground. None of the work we surveyed uses a **surrogate/non-interacting control** to separate genuine strategic adaptation from prior-driven or self-play behaviour; none measures equilibrium convergence as **output-space (or internal representational) convergence** against exploitability; none tests **true cross-family** convergence-to-equilibrium beyond pricing tiers; and despite mature human–AI platforms, none performs **joint human+AI neural co-measurement** during strategic play. These are the four instruments our testbed is built to claim.

6 Limitations

These are pilot-scale runs: small numbers of agents, few seeds, and single network instances; several findings need fuller seed-averaging, and although a varied-initial-distribution control already points the cross-family routing asymmetry toward a genuine best-response deficit rather than default stickiness, more seeds and network instances are needed to settle it. Prompt sensitivity is pervasive and load-bearing in this literature, and our contamination probe addresses memorization but not all prompt confounds. The exploiter identity in the Prisoner's Dilemma varies run to run; only the reasoning-style asymmetry is stable. All results are tied to specific model checkpoints in a fast-moving field and should be read as a methodological scaffold, the games, the contamination probe, the exploitability metric, and the cross-family protocol, rather than a fixed benchmark.

7 Conclusion

Asking whether AI reaches Nash is the wrong question; asking *when*, and *with what instrument*, is the right one. Language-model agents reach equilibrium when it is an argmin over given payoffs and miss it when it demands reasoning, and the gap between those two regimes, together with the strategic asymmetries that only heterogeneous, cross-family play reveals, is where the interesting measurement lives. The next step is to instrument that gap directly: exploitability and quantal-response rationality per family, a surrogate control on every game, output-space convergence as an equilibrium signature, and ultimately human and artificial agents measured together.

Data and code availability

Raw transcripts, orchestration scripts and the memory-one estimator are in the project repository; the figures of this preprint are drawn from the numbers reported in the text, and the script `equilibrium_curve.py` supplied with it performs the symbolic and numerical checks of Appendix B.

The source code and implementation are available in the AICumene GitHub repository: `github.com/aicumene/LLM-nash`.

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

Acknowledgments

Compute on local V100 and Thor servers at Aicumene.

A Protocol summary

Table 4 collects the experimental protocol in one place; items in brackets are not stated in the text and must be supplied from the run logs before submission. Table 5 lists every Prisoner's Dilemma batch, since the narrative of Section 4.2 draws on several.

Game, condition	agents	rounds	runs	dynamics, controls	metric
Beauty contest, $p = 2/3$	5 Qwen 3.6-27B	6	[1]	simultaneous guesses	distance to NE action
Beauty contest, $p = 1.2$ (probe)	5 Qwen 3.6-27B	6	[1]	contamination probe	distance to NE action
PD self-play, early batch	2 Qwen	8	[to supply]	neutral-label control (Option A/B)	cooperation rate
PD cross-family, early batch	Qwen vs Gemma 4	8	2 described	reasoning capture in one run	scores, cooperation
PD self-play, memory-one batch	2 Qwen	20	4	round-by-round logging	(α, γ)
PD cross-family, memory-one batch	Qwen vs Gemma 4	12	3	round-by-round logging	(α, γ)
PD, scripted single betrayal	Qwen vs script	14	3	one defection, early/mid/late	α on betrayal round, recovery
PD, scripted random defection	Qwen vs script	24	4	i.i.d. defection, rate 0.281	(α, γ), normalized exploitability
Braess, explicit option costs	6 Qwen	until convergence	[to supply]	sequential best-response updates	travel time, system cost, exploitability
Braess, costs from formulas	6 Qwen	until convergence	[to supply]	sequential best-response updates	travel time, exploitability
Mixed-family routing	2 Qwen, 2 Gemma 4, 2 Nemotron 3	5 reported	[to supply]	randomized update order	time by family, exploitability, route changes
Mixed-family routing, randomized start	same	[to supply]	[seeds: to supply]	randomized initial routes	time by family, route changes

Hardware: Qwen 3.6-27B (AWQ) on vLLM, 4×V100; Gemma 4-31B and Nemotron 3-Nano-30B on Ollama, NVIDIA Thor; orchestrator on a third host. [Decoding temperature, context length and the exact prompt templates for each game: to supply.]

Table 4: Protocol summary. Bracketed entries are to be filled from the run logs.

Batch	games × rounds	outcome	where used
Self-play, early	[1] × 8	24/24 mutual cooperation; identical with neutral labels	Section 4.2
Cross-family, early, run 1	1 × 8	Qwen defects in the endgame, exploits Gemma [scores: to supply]	Section 4.2
Cross-family, early, run 2 (reasoning capture)	1 × 8	Gemma defects on the final round; Qwen 15, Gemma 20	Section 4.2
Self-play, memory-one batch	4 × 20	152/152 reciprocations, $\gamma = 1$, α not estimable	Section 4.2
Cross-family, memory-one batch	3 × 12	33/33 each side, $\gamma = 1$, α not estimable	Section 4.2
Scripted single betrayal	3 × 14	$\alpha = 0$ on the betrayal round, full recovery within two rounds	Section 4.2
Scripted random defection	4 × 24	$\gamma = 0.515$, $\alpha = 0.077$, regret 17.4%	Section 4.2, Table 3

Table 5: All Prisoner's Dilemma batches.

B The memory-one equilibrium curve for payoffs $(5, 3, 1, 0)$

A memory-one strategy reacting to the opponent's last move is a pair (α, γ): $\gamma = \Pr(\text{C} \mid \text{opponent played C})$ and $\alpha = \Pr(\text{C} \mid \text{opponent played D})$. Write $r_i = \gamma_i - \alpha_i$. Because each player's move at round t depends

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

only on the opponent's move at round $t-1$, the joint process splits into two interleaved Markov chains, so in the stationary state the two players' moves are independent and their cooperation rates are $c_1 = (\alpha_1 + r_1\alpha_2)/(1 - r_1r_2)$ and $c_2 = (\alpha_2 + r_2\alpha_1)/(1 - r_1r_2)$. Player 1's stationary payoff is $U_1 = P + (S-P)c_1 + (T-P)c_2 + (R-S-T+P)c_1c_2$, which for $(T, R, P, S) = (5, 3, 1, 0)$ reads $U_1 = -c_1 + 4c_2 - c_1c_2$. Let

$$Q(\alpha, \gamma) = 3\alpha^2 - 7\alpha\gamma + 6\alpha + 4\gamma^2 - 5\gamma + 1. \tag{1}$$

Proposition 1. *At a symmetric profile $(\alpha_1, \gamma_1) = (\alpha_2, \gamma_2) = (\alpha, \gamma)$ with $|\gamma - \alpha| < 1$,*

$$\frac{\partial U_1}{\partial \alpha_1} = -\frac{(\gamma - 1)\, Q(\alpha, \gamma)}{D}, \qquad \frac{\partial U_1}{\partial \gamma_1} = \frac{\alpha\, Q(\alpha, \gamma)}{D}, \qquad D = (1 - (\gamma - \alpha)^2)^2 > 0. \tag{2}$$

Hence both first-order conditions hold iff $Q(\alpha, \gamma) = 0$ or $(\alpha, \gamma) = (0, 1)$; the corner $(0, 1)$ lies on the curve $Q = 0$ and is in addition a zero of both prefactors, which is the degenerate boundary point of Section 4.2.

Proof. Direct differentiation of U_1 with c_1, c_2 as above and evaluation at the symmetric point; the factorization is verified symbolically in `equilibrium_curve.py`. □

The first-order conditions do not by themselves make the curve a set of equilibria. We therefore check the best-response property directly: at 39 points of $Q = 0$ with $\alpha \in [0, 0.35]$, player 1's stationary payoff against the opponent's curve strategy never exceeds the symmetric value on a 401×401 grid of her own strategies (largest excess $8.9e-16$), so every tested point is a symmetric Nash equilibrium in memory-one strategies. The curve strategies are not equalizers in the sense of Press and Dyson: player 1's payoff does vary with her own strategy, but the symmetric point is a maximizer. For α beyond about 0.35 the curve leaves the unit square.

References

[1] Calvano, E., Calzolari, G., Denicolò, V. and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10), 3267–3297.

[2] Fish, S., Gonczarowski, Y. A. and Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806; forthcoming in *Proceedings of the ACM Conference on Economics and Computation (EC 2026)*.

[3] Arunachaleswaran, E. R., Collina, N., Kannan, S., Roth, A. and Ziani, J. (2025). Algorithmic collusion without threats. In *16th Innovations in Theoretical Computer Science Conference (ITCS 2025)*, LIPIcs 325, article 10.

[4] Einwiller, A., Ghosh Dastidar, K., Romazanov, A., Hautli-Janisz, A., Granitzer, M. and Lemmerich, F. (2025). Benevolent dictators? On LLM agent behavior in dictator games. arXiv:2511.08721.

[5] Nie, W., Luo, B., Meng, Z. et al. (2026). Equilibrium residuals expose three regimes of matrix-game strategic reasoning in language models. arXiv:2605.10410.

[6] Lu, S. E. (2024). Strategic interactions between large language models-based agents in beauty contests. arXiv:2404.08492; NeurIPS 2024 Workshop on Language Gamification.

[7] Bracale Syrnikov, M., Pierucci, F., Galisai, M., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V. and Nardi, D. (2026). Institutional AI: governing LLM collusion in multi-agent Cournot markets via public governance graphs. arXiv:2601.11369.

[8] Gao, Y., Lee, D., Burtch, G. and Fazelpour, S. (2024). Take caution in using LLMs as human surrogates: Scylla ex machina. arXiv:2410.19599.

[9] Zhang, Y., Mao, S., Ge, T., Wang, X., Xia, Y., Lan, M. and Wei, F. (2025). K-level reasoning: establishing higher order beliefs in large language models for strategic reasoning. In *Proceedings of NAACL-HLT 2025*, 7212–7234.

[10] Qian, C., Tsai, V., Behr, M., Hussein, N., Laugier, L., Thain, N. and Dixon, L. (2025). Deliberate Lab: a platform for real-time human–AI social experiments. arXiv:2510.13011.

[11] Horton, J. J., Filippas, A. and Manning, B. S. (2023). Large language models as simulated economic agents: what can we learn from homo silicus? NBER Working Paper 31122.

[12] Mei, Q., Xie, Y., Yuan, W. and Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences* 121(9), e2313925121.

[13] Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M. and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour* 9, 1380–1390.

AiCumene:2026.038v1 LLMNASH [cs.GT, econ.GN] July 2026

[14] Fan, C., Chen, J., Jin, Y. and He, H. (2024). Can large language models serve as rational players in game theory? A systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 38(16), 17960–17967.

[15] Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T. and Xu, K. (2024). GTBench: uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; arXiv:2402.12348.

[16] Menshikov, I. S., Shklover, A. V., Babkina, T. S. and Myagkov, M. G. (2017). From rationality to cooperativeness: the totally mixed Nash equilibrium in Markov strategies in the iterated Prisoner's Dilemma. *PLOS ONE* 12(11), e0180754.

[17] [Bernadotte, A., Menshikov, I..(2026). Verifiable AI for Law: a legal verifiability compiler with a rule/judgment firewall. *AICumenen* Preprint, DOI: 10.13140/RG.2.2.18174.75841. Online: https://research.aicumene.com/papers/verifiable-law/