Computable rational agents versus LLM negotiators

An instrumented multi-party scorable negotiation environment, a measured fairness gap, turn-level divergence localization, an honest negative on divergence-localized DPO, and a harness bug that fabricated a quarter of the campaign

`experiments/rational_agents` — ii_mats

2026-07-28 (revised 2026-07-30: clean-engine re-baseline)

Abstract

We built a multi-party, multi-issue *scorable* negotiation environment in which every normative benchmark is computed exactly, mounted five literature-grounded rational-agent oracles inside it, and used it to measure where large language models diverge from computable rational play — at the level of outcomes, individual turns, and individual reasoning steps. A validity-gated campaign of **1,770** episodes was run across five open-weight models, all-LLM and mixed tables, full- and private-information conditions, and cheap-talk on/off arms. It was then discovered that the batched generation engine had been *silently fabricating placeholder turns* whenever a generation call crashed — **26.2% of the campaign's 48,903 turns is text no model produced**, invisible to every screen in use because the placeholder parses cleanly. This paper reports the campaign re-cut on a clean engine: **840** episodes over six models plus the rational control, at 0.000 fabricated turns, with an internal control (the one already-clean cell) moving -0.017 against $+0.14$ to $+0.53$ for the contaminated cells. Three things changed and one got stronger. The **fairness gap is upheld and generalized**: all six models, spanning 4B–32B and four families, capture less for the worst-off party and produce more unequal agreements than the computable control with every interval excluding zero, and five of six sit further from both the Nash bargaining and Kalai–Smorodinsky solutions. **Aggregate efficiency is not ordered by scale** (8B is best, 32B mid-pack with the widest intervals), **deal rate is the wrong instrument** (the model that closes 99.2% of episodes does so while violating individual rationality in 94.2% of them), and **the cheap-talk ladder is flat**, not collapsing with scale. A prefix-probe experiment with an empty-scratchpad control shows the divergence is present *before any reasoning is emitted* (90% of flagged turns diverge with no chain of thought at all, rising only to 99% with the full chain), so late-step process supervision cannot be the fix. Divergence-localized DPO looked like a clean win on a single held-out game (deal rate $+0.200^{*}$, welfare $+0.194^{*}$) and **did not replicate** on three fully unseen games; on clean cells the effect there is indistinguishable from zero ($+0.017$ / $+0.083^{*}$ deal rate), so "training did not help" stands while the earlier, stronger claim that training actively *hurt* out of distribution is withdrawn as a harness artifact. Two later campaigns extend the instrument in opposite directions. A **narrative-framing arm** dresses the identical game in a realistic story under a hash-randomized, preference-decorrelated relabelling with masked-prompt byte-identity tripwires: the transformation leaves computable bargainers *bit-identical* and costs the LLM table welfare (-0.083, $[-0.143, -0.015]$), efficiency ($+0.238$ Pareto-frontier distance) and protocol discipline ($+0.217$ malformed episodes) — the cleanest LLM-versus-rational contrast in the project, since nothing about the game changes at all. A **frontier-API campaign** finds the same two-signed gap three orders of magnitude up the capability ladder: hosted Claude models close far more often than the computable control (0.767 against 0.567) and split worse on every fairness axis, sonnet's whole closing advantage is the cheap-talk channel (silenced, it closes 0.533 and is nearly as fair) with a sign opposite to the open-weight population, and a rational seat dropped into a

frontier table takes nothing for itself (-0.017, $[-0.081, +0.031]$ against $+0.147$ to $+0.444$ against open-weight seats) while making the whole table fairer. **That last finding does not survive its own power upgrade**: re-measured on a fresh 24-game screened bank at 96 episodes per cell under the identical condition, the same swap *extracts* $+0.128$ $[+0.056, +0.209]$ and pushes the frontier table 0.104 further from the Nash solution, surviving the matched-censoring restriction that dissolved the original effect — one of four three-cluster frontier claims that moved under power. The same campaign decomposes the anchor itself: a two-line policy that never speaks and only refuses deals below its own walk-away value reproduces the Bayesian agent's *entire* fairness contribution (six measures, all paired differences within 0.03 of zero) at 0.235 less capture, so **discipline buys the fairness and the Bayesian machinery buys the money**; no greedier policy out-extracts it (a maximally selfish proposer captures 0.090 less, an ultimatum hold-out 0.161 less, the tuned demand knee at 65% of own-maximum still trails), and giving the agent an empirically fitted opponent model instead of its assumed rationality step function makes it *concede* on 63.7% of changed turns rather than extract, because unanimity compounds sub-certainty and the step model's error was over-confidence rather than conservatism. Those fitted curves also give the fairness gap its behavioural mechanism: of below-threshold offers a seat actually votes on, 4–8B models accept 60–96% while Claude models, Qwen3-32B and the computable control accept **0–13%**. Finally, the one intervention that changes the *model* was preregistered and run: **outcome-aligned RL** on a smoothed log-Nash-welfare reward (GRPO, Qwen3-8B LoRA, six-seat self-play, two λ arms), evaluated on a 24-game held-out bank added by amendment before any training step. The gate **fails at every checkpoint in both arms on both banks**: training bought individual-rationality discipline (below-threshold agreements -0.079 to -0.129) and spent welfare for it (-0.109 to -0.185 normalized Nash welfare) via a held-out deal-rate collapse (-0.15 to -0.25) that does *not* appear on the games it trained on, the λ=1 arm made the deals it did strike measurably *less* fair (among-IR Nash welfare -0.040, Gini $+0.023$, both excluding zero), and handed an unseen ultimatum game the trained policy moved from proposing an accepted 60/40 split to proposing 96.7/3.3, accepted every time — away from the fair split and toward the subgame-perfect extractor. Scale bought discipline without distribution; so did RL, at a welfare cost. The durable results are methodological, headed by a **regret/outcome divergence** — the prompt scaffold and the collapsed training arms post the best per-turn oracle regret in the entire evaluation while closing the fewest deals — and by the fabrication bug itself, which a fail-closed validity gate certified as a perfect campaign. The re-baseline also yields a triage heuristic that held in three independent places: a defect that deletes turns moves every *closure*-dependent quantity and leaves *division-of-surplus* quantities standing, so a re-run's verdict is predictable from the metric's structure before the compute is spent.

Contents

1 Narrative summary

The setup / context. A *scorable negotiation* is the standard instrument from negotiation pedagogy and, more recently, from LLM benchmarking: several parties must jointly pick exactly one option for each of several issues, each party privately holds a score sheet assigning points to options, and each has a *threshold* (a BATNA — the value of walking away). Because the option set is small enough to enumerate — our instances have $3^5 = 243$ complete deals across six parties — every normative benchmark can be computed *exactly* rather than estimated: the Pareto frontier, the Nash bargaining solution, the Kalai–Smorodinsky point, the set of individually rational deals, and the optimal action at any point in the dialogue. That exactness is the whole point: it lets us ask not just "did the models do well?" but "what was the best available move at turn 7, and how much surplus did this model leave on the table by not making it?" We call that per-turn loss **oracle regret**, in direct analogy to centipawn loss in chess analysis.

The problem (why this was done). As of mid-2026 nobody had put *computable* rational agents head-to-head against LLMs inside such a game. The negotiation-agent competitions had no LLM entrants; the LLM negotiation benchmarks cite equilibrium theory but implement no equilibrium baselines; the closest prior art diagnoses bilateral price-only agents at the level of final outcomes only. Three specific gaps followed. First, no measured outcome gap under a shared protocol. Second, no *localization*: when an LLM negotiates badly, is the failure in its beliefs, its valuation arithmetic, its acceptance decision, or its choice of package — and at which turn, and at which step of its own reasoning? Third, nobody had trained on those localized divergences. A fourth requirement was added mid-project by the project owner and materially changed the bar: any trained intervention must beat the *strongest prompt-engineering baseline* under matched conditions, not merely beat a naive prompt.

What we did / what changed. We built the environment and the measurement stack (contributed into the shared `interlens` library, so both are reusable): an enumerable deal space with a generator that deliberately injects *Pareto slack* — 50–70% of acceptable deals are Pareto-dominated, repairing a documented defect in the best-known prior instance where 80.5% of acceptable deals already sat on the frontier and the game was therefore nearly zero-sum once feasible; harness-enforced formal moves (`PROPOSE`/`ACCEPT(offer_id)`/`REJECT`/`WALK`) so that "I accept" is never ambiguous; a swappable oracle stack of five literature-cited oracles (solutions, Bayesian beliefs, optimal-stopping acceptance, exact best response, stationary multilateral equilibrium); and a rational-agent zoo mounted as ordinary participants so a rational agent and an LLM can occupy seats at the same table. We validated the environment against theory (Phase P1), ran a 24-cell, 1,770-episode campaign through a fail-closed validity gate (P2), published a divergence atlas (P3), ran a full pilot of divergence-localized DPO with three ablation arms and two holdout designs (P4), and completed a seat-balanced 3,600-episode reverse-mixed companion.

Then we found that a quarter of the campaign was not real. The batched generation engine, when a `generate` call raised a transient CUDA or cuDNN error, caught the exception, substituted

a placeholder message — `"(ran out of time this turn and says nothing substantive)"` — and continued, logging nothing. That placeholder is a non-empty string that parses into a well-formed no-op action, so a cell in which the model never spoke reported `parse_ok = True` and non-empty content on every turn. Rates ran from 40% to 100% of turns in the all-LLM and scaffolded cells, and were *inversely ordered* with the very metric the capability ladder ranked models by. Everything downstream was therefore re-done: the eight affected cells were re-run on a sequential generation path at 0.000 fabricated turns with the token budget deliberately held fixed (it was never the constraint — not one turn in the entire campaign ever reached its cap), a new cross-model atlas was built over the clean cells plus a fourth scale point (Qwen3-32B, which was never contaminated), and the cross-game training evaluation was re-run in full. **Every cross-model, cheap-talk, scaffold, cross-game and seat-balanced-companion number in this paper is from the clean vintage** — for the companion, whose own cells were clean, that meant re-matching its paired contrasts against clean baselines rather than re-running it. The sections that were measured clean — mixed tables, the reasoning-localization experiment, the preference dataset, the single-game evaluation fleet — are unchanged and are flagged as such where they appear. Two divide-the-dollar preset cells were left contaminated on purpose, because no claim rests on them, and they are labelled in place.

The result. Ten findings: six from the main campaign, in descending order of how much we trust them, then two smaller later campaigns that extend the instrument in opposite directions and are labelled as such, then the one preregistered attempt to change the model itself, and finally the campaign that takes the project's flagship positive apart and re-measures it at eight times the statistical power. *(1) The environment works, and what LLMs lose is distribution, not aggregate value.* Rational tables reach near-frontier agreements under full information (deal rate 10/12, mean normalized joint surplus 0.916, distance-to-frontier ≈ 0) and rationally *fail* to agree under private information (2/12), exactly the theory-predicted signature. Against that control, on clean cells, *every* one of six models — Qwen3-4B, Qwen3-8B, Qwen3-32B, OLMo-3-7B, Gemma-3-4B-IT and Ministral-8B-2410 — gives its worst-off party less (egalitarian welfare -0.163 to -0.422) and produces a more unequal split (normalized Gini $+0.064$ to $+0.163$) than the computable agents, with every interval excluding zero, and five of six land further from both the Nash bargaining and Kalai–Smorodinsky solutions. The computable control is also the only player that never accepts a deal worse than walking away (individual-rationality violations 0.000 against 0.200–0.942). Aggregate efficiency, by contrast, resolves for only the two weakest slots and shows *no scale ordering at all*: Qwen3-8B has the best normalized-welfare estimate of the six ($+0.036$, interval crossing zero) and Qwen3-32B is mid-pack with the widest intervals (-0.063). Scale buys behavioral discipline — 32B has the lowest individual-rationality violation rate of anything tested — not better outcomes. *(2) Deal rate is the wrong instrument, and the artifact was hiding that.* Gemma closes 99.2% of episodes while violating individual rationality in 94.2% of them, and Ministral shows the same pathology (0.883): these are indiscriminate assent, not skill. Fabricated no-op turns had been suppressing exactly those indiscriminate agreements, so the contaminated data made a deal-rate ladder look like a capability ladder. On clean cells the ladder essentially inverts — Gemma moves from second-worst to first — which is why *no model-versus-model ranking from this campaign is reported*, only the LLM-versus-rational contrast. *(3) A rational replacement wins on how surplus divides, not on whether a deal happens — and its Bayesian learner does not recover hidden values.* In the seat-balanced companion, re-matched against clean baselines, replacing one LLM with Bayes raises the same seat's normalized capture by 0.147–0.444 across the five slots with every interval excluding zero. But the companion's other published headline — that a computable seat "sharply repairs" weak all-LLM closure — **does not survive**: against clean baselines the Gemma slot's

+0.428 agreement effect resolves *negative* at −0.097, both Qwen slots lose significance, and only Ministral holds at a fifth of its published size. Those baselines had looked weak substantially because the harness was eating their turns. Meanwhile the rational agent's issue-order accuracy stays at 0.448–0.465 and moves *backward* from the exact prior, and threshold error improves by exactly zero, so even the surviving advantage belongs to the structured policy as a whole rather than to successful hidden-value inference. *(4) Cheap talk is mildly harmful for two models and not a scale law.* On clean cells, turning on the public chat channel costs Qwen3-8B 0.233 of deal rate ($[-0.450, -0.100]$) and 0.106 of normalized welfare, and Qwen3-32B 0.083 and 0.041; Qwen3-4B and OLMo are unresolved, and for Gemma cheap talk *improves* normalized primary welfare (+0.129). The contaminated data had shown −0.417 and −0.650 penalties collapsing to −0.083 at 32B, which looked like a scale law; talking lengthens episodes, a longer episode gave the broken engine more chances to fabricate, and a fabricated turn cannot close a deal. The clean ladder is flat, and the $-0.42/-0.65/-0.08$ sequence should not be cited. *(5) The divergence is upstream of the reasoning.* Re-prompting flagged turns with the chain of thought truncated to k steps, and crucially with an *empty*-scratchpad control at $k = 0$, gives $0.903 \to 0.935 \to 0.993$ divergence share for no reasoning → one step → the full chain. The model's action is already wrong before it reasons; reasoning adds about nine points and, in the 12% of turns where reasoning genuinely introduces the error, the mechanism runs one way only — the reasoning talks the model out of a decision the no-reasoning baseline got right. There were zero turns where reasoning rescued a divergent baseline. This experiment re-prompted live and stored every probe, so it is untouched by the fabrication bug. *(6) Training on oracle-preferred actions does not help, and both of the two verdicts we published about it were wrong in different directions.* On a single held-out game the trained policy looked like a clean win (deal rate $+0.200^*$, welfare $+0.194^*$ over untrained prompting). Widening the holdout to three fully unseen games erased it, and the contaminated cross-game numbers then said training actively *hurt* (-0.183^* deal rate under private information). That retraction does not survive either: the cross-game fleet was 7–34% fabricated *asymmetrically*, with the untrained baseline the least contaminated cell in both information conditions, and on clean cells the same contrasts are +0.017 (full) and $+0.083^*$ (private). Ten of twelve trained-versus-baseline entries flip sign. The defensible statement is that the effect is indistinguishable from zero: **training did not help, and the evidence that it hurt was a harness artifact**. Along the way, a two-arm training design showed that 95.5% of the naive DPO training signal was rationale *prose style* rather than the decision (313 versus 14 nats of chosen-vs-rejected separation at step 0), and all three objectives we tried collapsed identically at the full step budget. *(7) Relabelling the identical game in a realistic story moves the LLMs and cannot move the rational agents* (§8). We built narrative framing as a pure relabelling: the same score sheets, thresholds, feasible set and `instance_id`, with only the issue names, option labels and party roles changed, and with the story-to-numbers mapping drawn from a hash of the instance id so that no role systematically occupies the demanding seat — the preference leak that ruined the parent benchmark is removed *by construction* and enforced by a masked-prompt byte-identity test. Under that transformation, computable bargainers reach **identical deals** with every metric difference exactly 0.000, and the LLM table loses welfare (-0.083, $[-0.143, -0.015]$), lands further off the Pareto frontier ($+0.238$, $[+0.044, +0.515]$) and starts tabling illegal packages (+0.217 malformed episodes) — 15 of them syntactically perfect JSON proposals naming only two or three of the five issues, with *zero* JSON syntax errors, which rules out "the long labels broke the parser". Nothing about the game changed, so this is the cleanest LLM-versus-rational contrast the project produced. Two of its originally published claims are withdrawn: the deal-rate headline (−0.250 became −0.167 with an interval crossing zero) and, newly established here, its behavioural mechanism — 422 of 435 and 567 of 570 of the "talk-only" turns that supported "the framed table talks instead of moving" were engine fabrications, against 4 deliberate no-ops per arm on clean

cells. *(8) The same two-signed gap appears at the frontier, and the whole closing advantage is cheap talk* (§9). An observational hosted-API campaign put `claude-sonnet-5` and `claude-haiku-4-5` through the identical environment under the matched condition. Both close far more than the computable control (0.767 against 0.567) and split worse on every fairness axis (distance to the Nash solution 0.809 and 0.876 against 0.578; Nash welfare given a deal 50.7 and 35.3 against 58.6; individual-rationality violations 0.033 and 0.067 against 0.000) — the same shape as the open-weight ladder, in a population orders of magnitude more capable. Three things sharpen it. Silencing the chat channel costs sonnet 0.233 $[-0.333, -0.133]$ of deal rate and improves every fairness axis, so *sonnet's entire closing advantage over rational play is the chat channel*; the sign is opposite to the open-weight models on the same instances, where chat *hurt* closing. Dropping one rational policy into a sonnet table extracts nothing for that seat (-0.017, $[-0.081, +0.031]$, against $+0.147$ to $+0.444$ on the same scale against open-weight seats) but makes the table fairer, so on these three games the rational policy reads as a *fairness anchor* against strong counterparties and a beneficiary against weak ones — **a reading finding (10) reverses at proper power**, while leaving the population contrast itself intact. And the goal-4 ladder is a triple null whose headline ratio we refuse to report, because its denominator — the per-seat rational advantage prompting was supposed to recover — is itself indistinguishable from zero. The thinking-enabled condition has since been run properly as a 2×2 against a narrative framing (60 episodes per cell). It does *not* rescue the fairness verdict — enabling thinking raises welfare (Nash welfare $+4.08$ $[+2.05, +5.55]$, worst-off share $+0.022$ $[+0.012, +0.034]$) but leaves distance-to-solution and equality spanning zero. What it does buy is *robustness to the story*: a narrative skin over the identical game costs the thinking-off arm -0.250 $[-0.450, -0.150]$ of deal rate, costs the thinking-on arm nothing separable from zero, and that difference is itself resolved ($+0.133$ $[+0.050, +0.200]$) (§9.5). *(9) Outcome-aligned RL buys discipline and spends welfare; it does not buy distribution* (§10). With every intervention that leaves the weights alone exhausted, we ran the P4 post-mortem's own prescription as a preregistered pilot: GRPO on a smoothed logarithm of Nash welfare — concave, so moving surplus toward the worst-served party always scores better, and *text-blind*, reading only the engine's scoring of the closed deal — with Qwen3-8B (LoRA) occupying all six seats in self-play, two arms paying each seat its own outcome (λ=0) or the table's (λ=1). Before any GPU spend the reward was scored against existing clean campaigns and passed all four preregistered orderings, notably ranking the all-rational control 0.93 above the all-LLM table *while the control closes fewer deals*. Before any training step, the primary endpoint was amended onto a fresh, overlap-guarded 24-game held-out bank (48 clusters, 960 episodes per cell), which is the reason this result has intervals at all. **The gate fails at every checkpoint, both arms, both banks.** What training bought is real and narrow: below-threshold agreements fall 0.087 / 0.079 with intervals excluding zero. What it cost is welfare — normalized Nash welfare -0.185 $[-0.225, -0.145]$ and -0.109 $[-0.145, -0.070]$ — purchased through a deal-rate collapse of 0.25 and 0.15 that appears *only on held-out games*, against flat deal rate on the training bank: overfitting to 24 games, and more training is monotonically worse. The two arms differ in kind: λ=0's loss is pure composition (every among-IR endpoint spans zero, so it struck fewer deals rather than worse ones), while λ=1 struck *measurably less fair deals among the deals it struck* (among-IR Nash welfare -0.040, distance-to-NBS $+0.037$, Gini $+0.023$, worst-off -0.017, all excluding zero and all larger on the second bank). The sharpest number points the wrong way: dropped into an ultimatum game it had never seen, the untrained model proposes an accepted 60/40 split and both trained arms propose 96.7/3.3, accepted 100% of the time, every trained seed above every untrained seed — fairness training on a six-party scorable game eroded the social prior toward the subgame-perfect extractor. The diagnosis is arithmetic and testable: because one party at $z = -0.2$ contributes -25.6 to the reward, nearly all available reward lives in avoiding violations, so the objective is nominally Nash welfare and behaviourally an individual-rationality

penalty. Hygiene: 5,856 evaluation episodes, zero fabricated turns. Caveat that bounds all of it: 24 GRPO steps per arm, 12% of the preregistered budget. *(10) The "rational anchor" is really two separable things, and at proper power its frontier version reverses* (§11). Every result above that credits a computable seat with making a table fairer credits a single object that welds three mechanisms together — veto discipline, a best-responding proposal rule, and the opponent model that makes those proposals generous. We took it apart with five trivial seat-0 policies, each keeping one channel, on a fresh 24-game bank screened for discriminativeness (24 admitted from 137 candidates; the Nash and Kalai–Smorodinsky solutions differ in 24/24, against *four of six coinciding* in the committed bank), 96 episodes per cell, both populations, with every reference re-run inside the bank. **The effect splits**: `passive-gate` — never speaks, never proposes, accepts anything above its own walk-away value — is indistinguishable from the full Bayesian agent on all six fairness measures (every paired difference within 0.03 of zero) while capturing 0.235 $[-0.338, -0.134]$ less, and on `claude-sonnet-5` it is resolved *better* than Bayes on fairness (Gini -0.037, worst-off $+0.031$) at a third of the capture. Discipline buys the fairness; the Bayesian stack buys the money. **Nothing greedier wins**: a maximally selfish proposer captures 0.090 $[-0.175, -0.004]$ *less* than Bayes, an ultimatum hold-out closes only 35% of its games and ends up 0.161 behind, the extraction/closure knee sits at a declared demand of 65% of own-maximum and still trails Bayes (0.488 against 0.515) while costing the table more — and, from the opposite direction, replacing the agent's assumed rationality with acceptance curves fitted to 585,265 real response events makes it *concede* on 63.7% of the turns it changes, on five of five fitted opponents, because no honest curve ever assigns the 1.0 the step model does and unanimity compounds the shortfall across five opponents. **That last one has since been run closed-loop and is the arc's most surprising positive**: over 864 fresh episodes (12.22 GPU-hours, 24 clusters) the truthfully-calibrated maximizer's own capture falls 0.146 $[-0.217, -0.075]$ and its within-table advantage halves, while *every* fairness endpoint resolves in the fair direction (distance to the Nash solution -0.362 $[-0.497, -0.233]$, Gini -0.080, worst-off $+0.090$, Nash welfare $+0.098$) at a *flat* deal rate — so the table did not get fairer by failing to close — and the whole effect strengthens rather than dissolving under matched censoring. Truthful knowledge about your counterparties makes a rational agent generous, because honest acceptance probabilities are sub-certain, unanimity compounds sub-certainty ($0.40^5 \approx 0.01$), and the only lever a best-responder has on a joint acceptance probability is to give everyone more. **Two population results**: an announced ultimatum, held byte-identical to a silent one so the announcement is the only difference, is worth nothing against Qwen3-8B ($+0.010$ $[-0.115, +0.125]$) and doubles closure against Sonnet ($+0.146$ $[+0.052, +0.250]$) — comprehension, not capitulation, and the opposite of the capitulation prior. And the powered bank overturns our own note 0018: on 24 clusters the Bayesian anchor *extracts* $+0.128$ from a Sonnet table and worsens its distance-to-NBS and Gini, surviving the matched-censoring restriction that dissolved the original benefit. Reported honestly: the welfare leg of that reversal does not survive, so the anchor redistributes rather than destroys; and the parity headline is a seat-0 result, since `passive-gate`'s fairness gain over the all-LLM baseline resolves only at seat 0 and the Bayes cells at seats 2 and 4 were still queued at the time of writing.

The unified reading. Two threads run through all ten. The first is that *oracle-referenced per-turn metrics and negotiation outcomes point in opposite directions in this game family.* The prompt scaffold buys the lowest regret in the entire evaluation and the worst closure, and it does so on two *independent* clean measurements: in the main campaign it costs -0.342 of deal rate $[-0.450, -0.275]$ against untrained prompting (§5.5), and in the fully clean P4 cross-game fleet untrained prompting beats it by $+0.217$ of deal rate $[0.000, 0.350]$ while the scaffold posts mean per-turn regret 3.97 points lower $[0.459, 6.355]$ (§7.5). The two figures are different quantities on different banks and

should not be read as one number: de-contamination *grew* the main campaign's closure penalty from −0.217 to −0.342, which happens to coincide numerically with the cross-game fleet's unrelated +0.217. The most extreme case is diagnostic: the collapsed training arm has the *best* regret of any policy evaluated while closing *zero* deals, because a policy that never closes never takes a costly decision. A regret-led read of this project would have endorsed the worst policy at every decision point. That is not a statement about our oracle being buggy — it is exact — but about what per-turn optimality measures when the outcome depends on coordinating five other parties. De-contamination sharpened this rather than weakening it: because a fabricated no-op turn incurs no measurable regret, *every regret number quoted from a contaminated cell was biased downward in proportion to how contaminated it was*, so the artifact was independently flattering exactly the cells that were doing least. The second thread is that *what a measurement instrument cannot see decides what you conclude.* The same instrument-blindness argument appears six times: `parse_ok` could not see a fabricated turn; deal rate cannot see that a closed deal is worse than walking away; per-turn regret cannot see whether five other parties will agree; a single-instance holdout's seed-clustered intervals cannot see cross-game generalization; a talk-only turn count cannot see that the turn was never generated (§8.3); and a fairness metric computed over reached deals cannot see that the model reached almost none of them, which is why every frontier table here prints deal rate beside the distances (§9). Each of the first four produced a published number that later had to be withdrawn, and in each case the fix was a control that made the invisible quantity visible rather than more data.

A third, more practical thread emerged from the re-baseline and is worth stating as a rule. *A defect that deletes turns moves closure-dependent quantities and leaves division-of-surplus quantities standing.* It held in four independent places: in the cross-model ladder, deal rates moved by up to +0.53 and the entire model ordering inverted while the distributional deficit against the rational control did not budge (§5.4); in the scaffold contrast, the closure penalty grew while every conditional quality advantage shrank or flipped (§5.5); and in the seat-balanced companion, the agreement effect collapsed — one slot resolving in the *opposite* direction — while the same-seat capture advantage survived on all five slots (§14.1); and in the framing arm, whose one closure-dependent row — deal rate — is the one that de-resolved, while welfare, legality and frontier distance all resolved *against* framing on the clean re-run (§8.2). The mechanism is not subtle: a deleted turn is a chance to close that never happened, whereas conditional on a deal existing, how its surplus divided is untouched. The practical value is that it lets you predict which of your published claims a re-run will overturn *before* spending the compute, and triage accordingly.

A fourth thread arrived last and is the one that costs the most. *Every frontier claim in this paper that rested on three game clusters and has since been re-measured on more has moved* (§11.5): the thinking-on anchor's fairness movements dissolved at matched censoring; the thinking-off anchor's distance movements resolved only after its cell was doubled; "the frontier model never caves to a stonewaller" went from 0/30 to a closure rate of 0.135; and the anchor's extraction null against Sonnet became a resolved +0.128 extraction. None of these were reversed by finding a bug — the estimator was correct throughout. They were reversed by giving it enough independent games to speak. Limitation 4 is not a caveat on this project's frontier results; it is the largest single source of error in them, and the cheap fix (screen a fresh bank for discriminativeness before running anything) is what the last campaign did.

One-sentence version. We built an exactly-scorable multi-party negotiation environment with computable rational baselines and found that LLM negotiators from 4B to 32B are universally worse at *distributing* surplus than computable agents while not being reliably worse at creating it, that the divergence is present before the models reason, that training on the divergence does not help,

that the same two-signed gap persists at the frontier where the entire closing advantage turns out to be the cheap-talk channel, that merely re-skinning the identical game in a story degrades the LLMs while leaving the computable agents bit-identical, that a preregistered outcome-aligned RL run on a Nash-welfare reward bought individual-rationality discipline without buying distribution — and charged a held-out deal-rate collapse for it — that the "rational anchor" credited throughout with making tables fairer is really two separable things, of which a two-line refusal rule delivers all of the fairness and none of the money while no greedier policy out-extracts the full agent and giving it a *truthful* model of its opponents makes it concede instead — moving every fairness endpoint of its table at once, at a flat deal rate, the one intervention in this paper that did — and that at eight times the statistical power the anchor's celebrated frontier innocence reverses into measurable extraction — and, after a harness bug was found to have fabricated 26% of the campaign's turns and every affected cell was re-run, that the per-turn oracle metric the project was built around anti-tracks the outcomes it was supposed to improve.

2 Why this project, and how it relates to prior work

2.1 The uncontested combination

A literature dive (four reports, `docs/lit/`) established the positioning that motivated the build. Three separate communities were each missing one leg:

- **Automated-negotiation competitions** run computable agents in multi-issue games, with exact solution concepts and a mature strategy zoo — and had zero LLM entrants in the 2025 cycle.
- **LLM negotiation benchmarks** (the Abdelnabi et al. scorable-games lineage, six parties × five issues; NegotiationArena; LAMEN; the RLVR negotiation papers) run LLMs and cite equilibrium theory, but implement no equilibrium baselines, so "the model did badly" is never referenced to what was achievable.
- **The closest prior art** (TERMS-Bench) diagnoses bilateral, price-only agents at the level of final outcomes.

The uncontested combination is therefore: computable rational agents versus LLMs, inside a *multilateral multi-issue* scorable game, with (a) repaired score sheets so Pareto-dominated-but-acceptable deals are actually in play, (b) binding offers and votes as formal moves rather than free text, and (c) per-turn regret against an exact rational oracle. "Oracle-preferred action as the DPO-chosen response at divergence points in negotiation" was an explicitly open gap — which is what makes the negative result in §7 worth reporting rather than discarding.

2.2 The critiques we designed around

Reproduction studies of the scorable-games benchmark supplied concrete design constraints, each of which became a generator knob or a mandatory control:

(a) **Pareto slack.** In the base benchmark game 80.5% of individually rational deals were already Pareto-optimal, so once a deal was feasible there was almost nothing left to get wrong. Our generator targets a configurable *dominated-acceptable* fraction (0.52/0.69/0.51 in the committed bank) verified by enumeration.

(b) **The communication-free baseline must fail.** A solo agent proposing from its own sheet plus the public role text matched the full six-agent benchmark (90% vs 80%) in one

reproduction — i.e. the benchmark was not measuring negotiation. We decorrelate score sheets from public role text (a knob) and always run the solo control.

(c) **Issue-type mix.** Integrative, compatible, and distributive issues must coexist, since models are documented to miss logrolls and to fight over already-aligned issues.

(d) **De-anchored numbers.** Fictional units and rescaling checks, because behavior tracks surface numbers.

(e) **Parsing artifacts are not behavior.** In prior work 69% of one model's apparent channel "leakage" was a parsing artifact, so channel separation is harness-enforced and never inferred from model tag discipline.

2.3 Relation to other experiments in this repository

This project is the negotiation-domain member of a family of studies in `ii_mats` on how models behave under social pressure and how well internal or oracle-referenced signals predict behavior. Two connections are load-bearing:

- The repository's *steering* and *dialogue-representation* programs repeatedly found that a white-box probe reward can be optimized without the corresponding behavior moving — probe-versus-behavior decoupling. The regret/outcome divergence reported here (§12) is the same shape of finding with an *exact, non-learned* reference signal instead of a learned probe, which makes it harder to dismiss as a bad probe.
- The *conversational capitulation* program found that saving full transcripts was the only thing that caught a major confound. The same practice caught two confounds here: the ultimatum "responder rejects the whole pie" result was a label-misread artifact visible only in the scratchpad (§4), and the DPO style confound was diagnosed by reading the chosen-versus-rejected text (§7.3).
- A third connection was added by the revision. That program's rule — *look at the actual text, not a summary statistic over it* — is precisely what would have caught the fabrication bug of §3.6 on day one, because the fabricated turns all say the same sentence and a human reading five transcripts would have noticed. Instead the campaign was validated by a gate over metadata, which passed. The generalizable version, which we would now apply to any generation harness: **a metric computed over stored model output is only as trustworthy as your evidence that the output came from the model**.

3 Contribution 1: the environment and measurement instrument

3.1 The formal game

Issues $j = 1..J$ (here $J = 5$) each carry a discrete option set, so the deal space is the product $D = \prod_j O_j$ with $|D| = 243$ — fully enumerable. Party i's utility is additive, $u_i(d) = \sum_j s_{ij}(d_j)$, with a private threshold τ_i. The analysis object is always **surplus**

$$x_i(d) \;=\; u_i(d) - \tau_i,$$

because raw points live on arbitrary private scales; only scale-invariant solution concepts are defensible across parties. Agreement requires explicit unanimous `ACCEPT` votes on one specific offer id, and a party holding a veto must be among them. Two information conditions are run: **FULL** (sheets are common knowledge, so any inefficiency is a clean rationality failure) and **PRIVATE** (a

common-knowledge prior over types, where the Myerson–Satterthwaite result makes *some* inefficiency rational, so the comparison must be against an executable Bayesian agent rather than against the full-information optimum).

3.2 Protocol

Each turn is a typed record with harness-enforced channel separation:

`Turn` = { `agent`, `thinking` (private, always captured), `action` (validated JSON), `message` (optional cheap talk) }

Formal actions are `PROPOSE(deal)` (which mints an offer id), `ACCEPT(offer_id)`, `REJECT(offer_id)`, and `WALK` — no-deal is a decision, not a timeout. Invalid actions get one retry with specific error feedback, and syntax violations are logged separately from economic-legality violations, because the two mean different things. Rounds are turn-counted (never wall-clock, which measures clock-tracking rather than negotiation), the deadline is restated every turn, and the opening proposer rotates round to round so no single seat is a point of failure. Treatment arms: ±chat, FULL/PRIVATE, a solo communication-free control, and an explicit prompt-scaffold axis.

3.3 The oracle stack

The central abstraction is

`Oracle.evaluate`(game, history, agent, legal actions) → `OracleVerdict`{action_values, best, beliefs?, flags},

and per-turn divergence is $V(\text{oracle action}) - V(\text{model action})$ in surplus points. Six oracles are implemented, each citing its source in the module header (Table 1). They are composable and swappable, and every one of them runs unchanged on any game preset.

3.4 Games are presets, not new engines

Because the scorable formalism subsumes the classical bargaining family, alternative situations are parameterizations selected by one flag (`run.py --game`), and the oracle stack, annotation, taxonomy, and atlas run over them unchanged: `scorable` (default), `ultimatum` (N=2, one issue, single round, responder-only vote; anchored to the subgame-perfect equilibrium of proposer-keeps-all), `divide_dollar` (multi-round rotating proposer, anchored to the Baron–Ferejohn/Okada closed forms), and `bilateral_multiissue`. Sanity pins in the test suite assert that the best-response oracle recovers the ultimatum subgame-perfect equilibrium and that divide-the-dollar recovers $v = 1/n$.

3.5 What was actually built, and the bugs that composition exposed

Library machinery went into `interlens` (`arena/actions.py`, `arena/oracles.py`, `arena/negotiation/*`, `arena/scenarios/scorable.py`, `PolicyParticipant`); the experiment directory holds instances, runners, annotation, analysis, and this writeup. Every run persists: the exact invocation as its first log line, the per-turn *rendered prompt* each model actually saw, git SHAs in the manifest, human-readable Markdown and HTML transcripts, and typed `OracleRecord`s per turn.

The build's most useful output was a list of bugs that only appear when correct components are composed. They are worth listing because each one would have silently produced a publishable-looking wrong number:

Table 1: The oracle stack. "Grounding" names the algorithm and its literature source as cited in the module header. Each oracle annotates every turn of every episode inline (pure Python, CPU) and can also be replayed post-hoc over stored episodes.

Oracle	What it computes	Grounding
`threshold`	Hard-violation detector: accepts or proposals below the acting seat's own threshold.	Trivial but diagnostic; "wrong deal" rates run 7–20% even for strong models.
`solutions`	Pareto frontier by brute force; discrete Nash bargaining solution $\arg\max \sum_i \log x_i$; discrete Kalai–Smorodinsky as Pareto-restricted leximin of $\min_i x_i/b_i$; maximum-Nash-welfare fallback for empty strict-IR.	Nash 1950; Kalai–Smorodinsky 1975; Harsanyi 1963 for n players; Caragiannis et al. 2019 for the MNW fallback.
`belief`	Exact Bayes over an enumerated type grid (weight rankings × value shapes × threshold levels) with a concession likelihood.	Hindriks & Tykhonov 2008; Chang & Fujita 2023; frequency model retained as the cheap control.
`acceptance`	Optimal stopping: $v_j = \mathbb{E}[\max(X_{j-1}, v_{j-1})] - C$, accept iff $x \geq v_j$, with the offer distribution induced by the type posterior; endogenizes $\tau_i(t)$ as a discounted continuation value.	Baarslag & Hindriks 2013; McCall 1970. Expected to be the single most diagnostic per-turn signal.
`bestresponse`	Exact expectimax over (rounds × deals × type posterior); yields per-turn surplus loss and revealed exploitability.	Feasible without search approximations at $\lvert D\rvert = 243$; Johanson et al. 2011 for exploitability.
`equilibrium`	Banks–Duggan stationary equilibrium by damped fixed-point iteration over acceptance sets $A_i(v) = \{d : u_i(d) \geq (1-\delta)\tau_i + \delta v_i\}$.	Baron–Ferejohn 1989; Okada 1996 closed form as the sanity anchor ($v = 1/n$); Eraslan 2002 uniqueness caveat.

B1. **Missing authoritative state block.** The scenario never emitted a `negotiation_state` block, so policies reconstructed state from merged opponent turns, saw only the last opponent's offer, and never observed an acceptable standing offer: **0 deals in 24 episodes**. After the fix a residual failure remained — in the forced-final vote-only phase policies still proposed (all 76 legality errors were `(final_vote, propose)`) — closed with a `must_vote` flag plus a shared terminal individually-rational `Policy.vote()`. Legality errors 76 → 0.

B2. **Silent oracle registry.** `run.py` resolved oracle names by `getattr` on a package that deliberately does not re-export them, silently skipping *every* documented oracle.

B3. **Concurrency races.** Shared mutable belief state across concurrently-stepping seats crashed roughly a quarter of instances; separately, a process-wide model cache was not thread-safe, so six threads racing to load the same weights killed all 36 episodes of the first pilot with `Cannot copy out of meta tensor`. Fixed with per-key double-checked locking plus one shared participant per model id.

B4. **Strict chat templates.** Gemma and Mistral templates reject views with repeated same-role messages. The fatal shape is the round-1 proposer's own turn opening its later views, which is why single-shot ultimatum passed everywhere and every *multi-round* game crashed. Fixed family-generically at the participant layer (`Participant.repair_view`, idempotent, replay-proven bit-identical). This bug also produced *data contamination*: completed jobs had aggregated crashed episodes as no-deals, deflating one cell's deal rate from 0.635 to 0.550.

B5. **Two seat-binding oracle bugs** (found late, by a replay verification in the localization lane; see §7.2).

B6. **Performance.** A private-information oracle rebuilt a 17k-types × 243-deals tensor every turn: 6.6 s/turn → 185 ms/turn, a 36× speedup, which was the "hang" at six parties.

B7. **The swallow-and-fabricate path** — found after publication, and the largest defect in the project. It gets its own subsection below.

3.6 B7: the engine fabricated a quarter of the campaign, and the validity gate certified it

This is the defect that forced the revision of this paper, and it is reported at length because every property that made it survive a whole campaign is a property other harnesses share.

The mechanism. To co-step 120 live episodes on one GPU the harness uses a batched pool (`BatchedEpisodePool._generate_batch`). That pool caught `RuntimeError`s it classified as transient — CUDA out-of-memory, and the cuDNN attention faults `mha_graph` and `is_good()` — recursively halved the batch to retry, and when a lone request still failed, **substituted a placeholder message and swallowed the exception entirely**. Nothing was logged, nothing raised, no counter moved, and the episode was saved with `status="done"`. The placeholder exists for a good reason: a genuinely empty message would corrupt the other seats' strictly alternating chat views, so the engine writes `"(ran out of time this turn and says nothing substantive)"` in its place. The model was never successfully asked.

Why every screen we had was blind to it. The placeholder is a *non-empty* string that parses into a *well-formed* no-op action. So an episode in which the model spoke on one turn in three reports `parse_ok = True` on every turn, non-empty `content` on every turn, valid JSON everywhere, and

zero syntax errors. A first audit of one cell using exactly those screens returned `parse_failures=0, empty_visible=0` and pronounced it clean; that run had advanced *zero rounds in eighteen turns*. Screening instead on `parsed_action.atype == "none"` fails in *both* directions at once — on one cell it caught 714 of 1,316 fabricated turns (the rest parse to no action dict at all) while also firing on five turns of legitimate play. The detector that works is a conjunction, and it is now in the library as `interlens.arena.engine.gen_failures`: placeholder content **and** zero output tokens **and** `raw is None`, where the last clause is a *guard* excluding the placeholder's other producer — a model that genuinely returned empty or reasoning-only text, which keeps its real completion in `raw`. Those two categories have different fixes and must never be pooled.

The rates. Per-turn fabrication in the *selected* cells, measured with the library detector over all stored turns:

Selected cell	Fabricated	Consequence for this paper
`p2_Olmo-3-7B_all_llm` (*excluded run*)	1.000	Dead run; the gate had already excluded it
`p2_Olmo-3-7B_scaffolded`	0.670	Re-run; second dead cell (§5.5)
`p2_Qwen3-8B_scaffolded`	0.546	Re-run; contrast corrected (§5.5)
`p2r1_Ministral-8B_all_llm`	0.546	Re-run
`p2r1_Gemma-3-4B_all_llm`	0.509	Re-run
`p2_Qwen3-4B_all_llm`	0.462	Re-run
`p2_Qwen3-8B_all_llm`	0.400	Re-run
`p2_Olmo-3-7B_divide_dollar`	1.000	Dead cell reported as "unavailable" (§5.7)
`p2_Qwen3-8B_divide_dollar`	0.086	Quoted with the rate attached
`p2r3_Olmo-3-7B_all_llm` (sequential)	**0.000**	**The internal control**
`p2_Qwen3-32B_all_llm` (chunked)	**0.000**	Never contaminated
`p2_all_rational`, all `mixed`, `solo`, `ultimatum`	**0.000**	Unaffected
All 30 `reverse_mixed` cells	**0.000**	Cells clean, *baselines* were not (§14.1)
P4 single-game (`game2`) evaluation fleet, 16 cells	**0.000**	Unaffected (§7.4)
P4 cross-game fleet, 10 cells	0.070–0.342	Re-run; *asymmetric* (§7.5)
Campaign total	**0.262**	of 48,903 turns

Three things the rate structure rules out. The contamination is *perfectly* confined to the two table types that use the batched pool (`all_llm` and `scaffolded`); every `mixed` cell — same models, same prompts, same cap, routed through the async pool — is at exactly 0.000. The rate is *flat across rounds* (33%, then 36/55/40/36% for one cell), where a context-length problem would ramp as transcripts grow. And the A/B already existed in the repository unnoticed: `p2_Olmo-3-7B_all_llm` fabricated 100.0% of 1,562 turns while `p2r3_Olmo-3-7B_all_llm` — byte-identical arguments plus `--sequential --concurrency 1`, same cap — fabricated 0.000%. A fourth corroboration is a natural experiment: `p2_Qwen3-32B_all_llm` is the one uncontaminated batched `all_llm` cell, because that lane had chunked it into five 24-episode jobs specifically fearing this failure class. A *larger* model with *no* failures is what a batch-width story predicts and no capability or budget story can explain.

It was not truncation, and that mattered for the remediation. The working hypothesis of everyone who looked at it, including the brief that opened the diagnostic lane, was that models were being cut off at the 2,048-token per-turn cap. That hypothesis was checkable and false: across all 44 stored campaign run directories, **zero turns reached the cap** and zero carry a `max_tokens` stop reason, with the p99 of generated output at 132–666 tokens and the single longest turn ever generated at 1,855. Raising the budget — the entire planned remediation — would have fixed

nothing.[1] A related trap: `stop_reason` is `None` on every local turn in this project (only hosted-API participants populate it), so the truncation screen the engine's own source comment recommends is *inert* on local runs.

The admission. §5 reports a fail-closed publication gate that rejects a campaign unless every planned episode exists exactly once with `status="done"`, model ids match an allowlist, annotation counts match episode counts, both transcript formats exist, and no directory contains a `running` or `error` record. That gate passed at 1,770/1,770 with zero errors and zero warnings. **It certified as a perfect campaign one in which 26.2% of the turns were text no model produced.** Everything the gate checked was true. It checked provenance, completeness and status; it never checked that a stored turn was the product of a successful generation call, because until this bug nobody had written down that such a thing could fail silently. A validity gate is a claim about the properties it enumerates and nothing else, and its passing is evidence about exactly those properties. The two lessons we would carry to any harness: **an error-handling path that substitutes plausible data must raise, count, or stamp — never all three of silent, plausible, and parseable**; and **a completeness gate needs at least one liveness check** — here, the modal `n_tokens_out` per cell, which would have read 0 and failed the run in the first minutes.

The fix, and the falsification control that licenses the re-baseline. The remediation was to re-run the eight affected cells with `--sequential --concurrency 1` — the flag pair the accidental OLMo A/B had already validated — while deliberately *not* touching the token cap, so no delta could be attributed to two changes at once. All eight report 0.000 fabricated and 0.000 truncated. The control that makes the resulting deltas interpretable is free and was not designed in: **OLMo's selected cell was already clean on both sides of the boundary**, so re-running it measures what re-running costs by itself. It moves −0.017 in deal rate. The four genuinely contaminated cells move +0.142, +0.275, +0.367 and +0.525, *in rank order of how contaminated they were*. Had these deltas been run-to-run variance, OLMo would have moved as much as the others. The engine defect itself is now fixed in `interlens` (`b314dc1`, `37d6f1f`), beyond the specification this lane asked for: a `TurnRecord.gen_failed` stamp carrying the causing exception, retries before giving up, a default-on fail-loud budget that aborts a run exceeding 10% fabricated turns, and the `gen_failures()` detector. The `_b2` re-runs *avoid* the broken path rather than exercising the repair, which is the honest scope of what they validate.

One drift control we did not plan and needed. Eleven `interlens` commits landed between the original campaign and the re-runs, so "before versus after" was not automatically a one-variable comparison. Re-scoring every stored before-cell with today's library over the *same* stored episodes gives a maximum absolute delta of **0.000000** across 8 cells × 9 metrics — the metric definitions are bit-for-bit invariant, so old and new numbers are on the same scale. But the *rational control*, re-run "for provenance symmetry", moved from deal rate 0.567 to 0.650 despite being clean of fabrication

[1] That zero is a statement about the *original* campaign, and it is worth scoping precisely, because the clean re-runs produced the one real cap hit in the project's local cells. The audit had flagged OLMo as the single model with less than 2× headroom and said it would revisit if a clean turn ever reached the cap; `p2_Olmo-3-7B_scaffolded_b2` has **3 of 3,180 turns** (0.09%) truncated at exactly 2,048, against a p99 of 999. That is negligible for every conclusion here, but it means 2,048 is not comfortable for OLMo under the verbose scaffold prompt, and anyone extending that arm should raise it. A second, larger pocket of genuine truncation lives in the P4 evaluation fleet: the two *SFT*-trained cells truncate 1.6% and 1.1% of their turns, which neither the base model nor the DPO-trained variants do anywhere — supervised fine-tuning on negotiation transcripts appears to have made the model verbose enough to occasionally run out of budget mid-action. It is a genuinely different failure from fabrication and is recorded rather than folded into it.

on both sides, which no engine-path story explains. Two commits improving the Bayesian agent's belief updating (soft updates from formal responses; preserving proposer identity in the posterior) account for it, and they predict the observed direction: better-targeted proposals close more deals. The seat-binding oracle fix of §7.2 is ruled out on two independent grounds — it never reached policy code, and it predicts the wrong sign. The consequence is procedural: **every LLM-versus-rational contrast in this paper is taken against** `p2_all_rational_b2`, the re-run control, not the original.

4 Phase P1: does the environment support the claim at all?

Before comparing LLMs to anything, the rational agents themselves must behave. The design's P1 gate: executable rational agents must reach near-frontier deals under FULL information and show *calibrated* inefficiency under PRIVATE information. If that fails, no LLM-versus-rational claim is supportable.

Running the committed starter bank (3 games × {full, private}, $n = 6$ parties, $|D| = 243$, $\delta = 0.95$) with all-rational tables (Boulware/Conceder/Bayesian-rational seats), both arms, two seeds — 24 episodes:

	Deal rate	**Mean normalized joint surplus**	**Distance to frontier**
FULL information	10/12	0.916	≈ 0
PRIVATE information	2/12	—	—

Efficient agreement under complete information; rational inefficiency under private information. That is the theory-predicted signature, and it is now a regression guard (deal rate ≥ 0.4, max frontier distance ≤ 0.05). Supporting pipeline health: 310 `ACCEPT` actions where there had been zero, and 1,416 typed `OracleRecord`s persisted into episode JSON.

First LLM behavior, and a correction worth recording. The first single-episode LLM smoke already produced the deal-closing-bias failure mode: the LLM table reached a deal at normalized joint surplus -0.038 — *worse than disagreement* — where rational tables' reached deals sit at 0.92–1.00. In the `ultimatum` preset, a 4B model gave away the *entire* pie in 7 of 9 episodes (offering the responder everything and keeping zero), the polar opposite of the subgame-perfect proposer-keeps-all and far beyond the human ≈40% offers in the classic experiments; mean proposer equilibrium regret ≈ 9.4 out of 10. Scratchpads confirm this is non-strategic satisficing rather than fairness: the proposer "proposes a deal that meets my threshold" and never registers that it holds all the leverage.

One apparent finding from the same run was **retracted by reading the transcripts**: a responder that "rejected receiving the whole pie" turned out to be misreading the bare option label `P0` as "I get 0", when its own sheet scored `P0` at 10 for it. That is a rendering bug, not a preference. It was fixed with per-seat "you get / they get" offer rendering, and it is the reason responder-side ultimatum claims are absent from this writeup while proposer-side claims remain.

A second early result was likewise retracted. A two-episode mixed-table smoke showed the LLM seat taking 19.7 surplus against a rational-seat mean of 66.1 and was described as the LLM "getting stripped" by rational co-players. That comparison is invalid — it compares raw points across differently-scaled utility sheets with the LLM pinned to one seat — and §5.6 reports what the affine-invariant version of the same question actually shows.

5 Phases P2/P3: the measured outcome gap (goal 1)

5.1 Design and the validity gate

The campaign is 24 logical cells and exactly 1,770 episodes:

- **Canonical scorable negotiation**: five model slots × {`all_llm`, `mixed`} × 120 episodes each. Each 120 is six committed instances (three payoff games × FULL/PRIVATE) × {`moves_chat`, `moves_only`} × ten rollout seeds.
- **Prompt-scaffold comparison**: Qwen3-8B and OLMo-3-7B, 120 episodes each, for the goal-4 baseline.
- **Rational controls**: 120 all-rational episodes and 60 solo (communication-free) episodes.
- **Presets**: ultimatum and divide-the-dollar, 15 episodes per model per game.

Two **slot substitutions** must be read literally. `google/gemma-3-4b-it` fills the Gemma-4 slot and `mistralai/Ministral-8B-Instruct-2410` fills the Ministral-3 slot, because the pinned `transformers =4.57.6` does not register the Ministral-3 or Gemma-4 architectures and upgrading mid-fleet (with 18 jobs importing from the shared environment) was rejected. **These data support conclusions about Gemma 3 4B IT and Ministral 8B 2410, not about Gemma 4 or Ministral 3.**

The key methodological point of this phase is that *Slurm completion is not a validity signal.* Seven logical cells contained strict-template errors, stale `running` rows, duplicated preset cohorts, or CUDA losses, and treating those rows as no-deals biases every aggregate downward. Publication is therefore fail-closed around an immutable manifest (`p2_selected_manifest.json`) mapping each logical cell to one immutable run directory and recording every excluded run with a reason. `validate_campaign.py --stage complete` rejects a campaign unless: the selected cells sum to the declared total; every run has the expected manifest and committed provenance; every planned episode exists exactly once with `status="done"` and a unique logical key; observed model ids match the per-cell allowlist; annotation counts match episode counts; both transcript formats exist; and no selected directory contains `running` or `error` records. The selected manifest passes at **1,770/1,770 episodes with zero errors and zero warnings**, with 16 contaminated or smoke runs preserved and excluded (Appendix C).

And that pass was worth much less than it looked. The same manifest that satisfies every clause above selected cells in which up to 67% of the turns were fabricated by the engine (§3.6). The gate is not wrong — every property it enumerates held — but it enumerates provenance, completeness and status, and says nothing about whether a stored turn was produced by a successful generation call. The one clause that would have caught it is a *liveness* check, and it costs nothing: the modal generated-token count per cell, which was 0. Read §5 to §5.5 with that in mind — the cross-model numbers below are from the clean re-cut, not from this 1,770-episode selection.

5.2 Two scale corrections that changed the headlines

Two originally-reported results were retracted for the same underlying reason — comparing quantities that are not invariant to how each party's score sheet happens to be scaled.

Collective welfare. The historical harness "primary" score was raw joint surplus divided by the largest feasible raw joint surplus. That ratio is still scale-dependent: rescaling one participant changes both its implicit weight in the numerator and which deal maximizes the denominator. Atlas

v2 recomputes every collective headline from the per-party *normalized* surplus vector

$$z_i(d) \;=\; \frac{u_i(d)-\tau_i}{\max_{d'} u_i(d')-\tau_i}, \qquad \text{primary}_{\text{norm}}(d) \;=\; \frac{\sum_i z_i(d)}{\max_{d'\in IR}\sum_i z_i(d')},$$

both of which are invariant to independent positive affine transformations of each score sheet. The runtime harness now emits this as `primary`; the former definition survives as `raw_primary`, and raw joint/egalitarian welfare and Gini are retained only for within-sheet audit. Historical episode JSON is immutable and still contains the old value; the atlas does not trust it but recomputes from the committed instances and the realized deals.

Interpersonal comparison. The original mixed-table analysis compared `surplus[0]` (the LLM) against the mean of `surplus[1:]` (the rational seats). Under $u_i' = a_i u_i + b_i$ and $\tau_i' = a_i\tau_i + b_i$, surplus transforms as $x_i' = a_i x_i$, so an arbitrary positive scale factor can change "who won" without changing anyone's preferences or behavior. The design compounded it: the LLM always occupied seat 0 and was never rotated, and seat 0 had systematically less raw surplus available (Table 2).

Table 2: Built-in seat-0 disadvantage in the committed instance bank. Columns: the maximum surplus attainable by seat 0 on its own sheet; the mean of the same quantity over the other five seats; and the difference. The LLM occupied seat 0 in every mixed episode, so any raw-surplus "who won" comparison inherits this gap before any behavior occurs.

Instance seed	Seat-0 max surplus	Other seats' mean max	Built-in difference
0	121.7	144.3	−22.6
1	74.0	135.6	−61.6
2	118.0	149.1	−31.1
Mean	104.6	142.9	**−38.4**

5.3 Metrics, matching, and the regret caveat

Comparisons are matched on committed instance, rollout seed, arm, and information condition; confidence intervals resample whole committed-instance clusters with replacement (5,000 draws). **There are only three instance clusters**, so interval endpoints are coarse and effect size plus replication across models matter as much as nominal exclusion of zero.

One caveat governs every regret number in this section. The best-response oracle attached during P2 has full access to all score sheets. In a PRIVATE cell its regret is therefore an *omniscient counterfactual*: loss relative to a policy that sees sheets the acting participant could not. We report and interpret regret *within* information condition only, and it must not be read as implementable private-information policy regret. This is also why the rational control's high PRIVATE regret (44.6 versus 13.7–33.8 for the LLMs) is an artifact of the information assumption rather than evidence that the rational policy is less rational than an LLM.

5.4 Result: all-LLM versus all-rational, on clean cells

Everything in this subsection is computed over the **clean-engine re-cut**: a 7-cell, 840-episode atlas (six models' `all_llm` cells plus the `all_rational` control) at 0.000 fabricated and 0 at-cap

turns in every cell, passing the strict campaign gate at 840/840.[2] It also adds a fourth scale point, Qwen3-32B, which the original campaign did not have — so the clean data covers 4B to 32B across four model families.

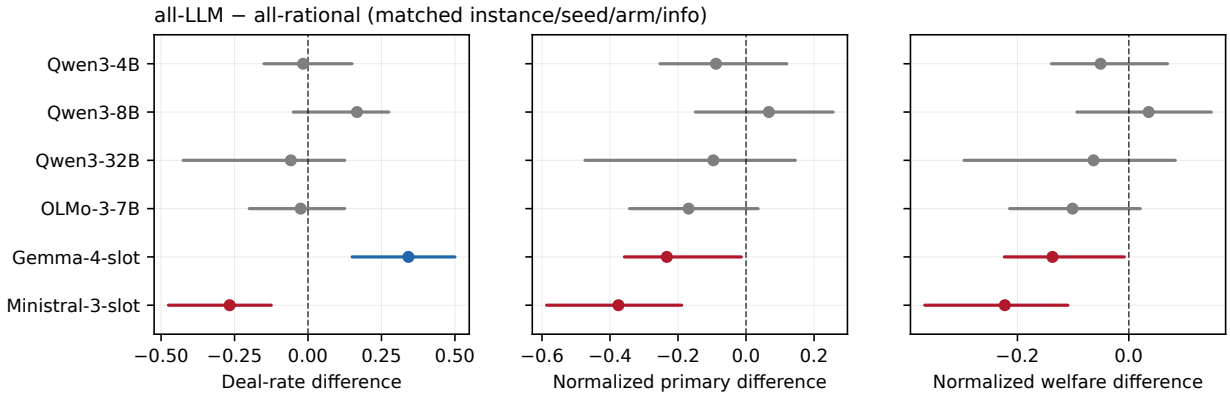

Figure 1: Goal 1: all-LLM tables minus all-rational controls, matched on committed instance, rollout seed, arm, and information condition, on clean cells. Each row is one model slot; the dot is the paired point estimate and the bar is a 95% confidence interval from 5,000 bootstrap resamples over the three committed-instance clusters. Colored bars exclude zero (red = LLM worse, blue = LLM better); grey bars cross zero. *Deal-rate difference* is the change in the fraction of episodes ending in an agreement. *Normalized primary* is realized joint normalized surplus as a fraction of the best attainable, and *normalized welfare* is mean per-party normalized surplus with no-deal scored as zero; both are invariant to per-party affine rescaling of score sheets. Aggregate efficiency resolves for only the two weakest slots and shows no ordering by scale — Qwen3-8B's estimate is positive and Qwen3-32B's has the widest interval of the six. The universal deficit is on the distributional metrics of Table 4, not these.

Reading Tables 3 and 4 with Figure 1, four things are true and two earlier claims are not.

The distributional gap is universal and does not depend on scale. Every one of six models leaves its worst-off party with less normalized surplus than the computable control does and produces a more unequal split, with all twelve intervals excluding zero, and five of six sit further from both bargaining solutions (Qwen3-4B is the lone unresolved case, and its point estimates have the same sign). Absolute mean distance-to-NBS puts the control at **0.579** against 0.834 (Qwen3-4B), 0.839 (Qwen3-8B), 0.863 (Qwen3-32B), 0.941 (OLMo), 0.964 (Ministral) and 1.047 (Gemma): the control dominates every language model and 32B sits mid-pack. This is a stronger claim than the paper originally made — it is not "the deficit resolves for two of five slots", it is that *the distributional deficit against a computable reference is universal across 4B–32B and four families and is not a capability artifact at any scale tested.*

Aggregate efficiency, by contrast, is not ordered by scale at all. Only the two weak substitute slots have resolved normalized-welfare deficits; Qwen3-8B's estimate is *positive* (+0.036) and Qwen3-32B's is middling with the widest intervals of the six. The two axes come apart, and any reading in which "more scale closes the gap" is unsupported.

Deal rate is the wrong instrument, and the artifact was concealing that. Gemma now closes *more* often than the control (+0.342) — and does it at an individual-rationality violation rate of 0.942, i.e. it agrees to packages that leave parties worse off than walking away in 94% of episodes, with a reached-deal Gini 2.4× the control's. Ministral is the same pathology at a lower closure rate

[2] `/nlp/scr/siddharth/ii_mats/rational_agents/atlas_b2plus32b/`. It supersedes the contaminated-baseline atlas `atlas_p2plus32b/`, which is retained for the audit trail only. Five `all_llm` cells and the control are the `_b2` re-runs; the Qwen3-32B cell is the original, which was never contaminated (§3.6).

Table 3: All-LLM minus all-rational on clean cells, 120 matched episode pairs per model, three instance clusters, control = `p2_all_rational_b2`. **Deal rate**: fraction of episodes reaching an agreement. **Normalized primary**: realized joint normalized surplus divided by the maximum attainable over the individually-rational set. **Normalized unconditional welfare**: mean per-party normalized surplus, no-deal scored zero. **Reached-deal welfare**: the same restricted to episodes that closed, which isolates *agreement quality* from *closure*, and whose pair count therefore varies by model (30–77). Bold marks intervals excluding zero. Negative means the LLM table is worse.

Model slot	Deal rate	Norm. primary	Norm. uncond. welfare	Reached-deal welfare
Gemma-4 slot (`gemma-3-4b-it`)	**+0.342**	**−0.233**	**−0.137**	**−0.316**
	[0.150, 0.500]	[−0.358, −0.014]	[−0.224, −0.008]	[−0.385, −0.278]
Ministral-3 slot (`Ministral-8B-2410`)	**−0.267**	**−0.375**	**−0.223**	**−0.270**
	[−0.475, −0.125]	[−0.586, −0.189]	[−0.366, −0.109]	[−0.309, −0.202]
OLMo-3-7B	−0.025	−0.169	−0.101	**−0.155**
	[−0.200, 0.125]	[−0.342, 0.036]	[−0.214, 0.021]	[−0.196, −0.083]
Qwen3-4B	−0.017	−0.088	−0.051	**−0.083**
	[−0.150, 0.150]	[−0.253, 0.120]	[−0.139, 0.069]	[−0.161, −0.011]
Qwen3-8B	+0.167	+0.067	+0.036	−0.058
	[−0.050, 0.275]	[−0.149, 0.256]	[−0.093, 0.148]	[−0.100, 0.059]
Qwen3-32B	−0.058	−0.096	−0.063	−0.082
	[−0.425, 0.125]	[−0.474, 0.145]	[−0.296, 0.084]	[−0.210, 0.023]

Table 4: The distributional half of the gap, which is where the universal result lives. All-LLM minus all-rational on clean cells, restricted to episodes that closed (30–77 pairs per model), three instance clusters. **Egalitarian welfare** is the worst-off party's normalized surplus capture, so negative means the LLM table leaves its weakest seat with less. **Normalized Gini** is inequality of capture across the six parties, so positive means less equal. **dist-NBS** and **dist-KS** are distances from the realized deal to the Nash bargaining solution and the Kalai–Smorodinsky point in normalized-surplus space, and **dist-frontier** the distance to the Pareto frontier; positive means further away, i.e. worse. Bold marks intervals excluding zero. All six models are worse on the first two; five of six on dist-NBS and dist-KS.

Model slot	Egalitarian welfare	Norm. Gini	dist-NBS	dist-KS	dist-frontier
Gemma-4 slot	**−0.422**	**+0.130**	**+0.477**	**+0.348**	**+0.501**
	[−0.525, −0.345]	[0.083, 0.205]	[0.231, 0.983]	[0.231, 0.551]	[0.389, 0.549]
Ministral-3 slot	**−0.380**	**+0.149**	**+0.457**	**+0.337**	**+0.475**
	[−0.421, −0.340]	[0.072, 0.213]	[0.232, 0.872]	[0.232, 0.512]	[0.199, 0.674]
OLMo-3-7B	**−0.210**	**+0.163**	**+0.404**	**+0.282**	**+0.287**
	[−0.300, −0.165]	[0.085, 0.224]	[0.173, 0.872]	[0.173, 0.475]	[0.101, 0.366]
Qwen3-4B	**−0.163**	**+0.064**	+0.230	+0.099	+0.139
	[−0.311, −0.083]	[0.034, 0.162]	[−0.009, 0.874]	[−0.009, 0.293]	[−0.065, 0.219]
Qwen3-8B	**−0.206**	**+0.082**	**+0.288**	**+0.170**	+0.061
	[−0.277, −0.118]	[0.034, 0.139]	[0.072, 0.673]	[0.072, 0.329]	[−0.091, 0.164]
Qwen3-32B	**−0.217**	**+0.144**	**+0.403**	**+0.259**	**+0.147**
	[−0.377, −0.084]	[0.070, 0.208]	[0.080, 0.744]	[0.080, 0.390]	[0.017, 0.456]

(IR 0.883). Fabricated no-op turns were suppressing exactly these indiscriminate agreements, which is why the contaminated data read as a tidy capability ladder. Qwen3-32B's clean IR-violation rate of **0.200** is the lowest of any model tested and is its one closure-independent advantage — a statement about how it plays, not about what it achieves.[3]

The deficit is not only failure to close. Restricted to episodes that reached agreement, welfare is lower for all six models and resolves for four of them — when LLM tables do agree, they agree on materially worse deals, and they do so at every scale.

What is *withdrawn*: any statement of the form "model X negotiates better than model Y" from this campaign, because the clean ordering essentially inverts the contaminated one (Gemma second-worst → first by deal rate; Qwen3-32B second → fifth); and the claim that the welfare deficit resolves for the two substitute slots *specifically*, which is true of aggregate efficiency but understates a result that is universal on the distributional metrics.

5.5 Cheap talk and the prompt scaffold

Cheap talk is mildly harmful for two models, and the apparent scale law was contamination. In canonical all-LLM tables on clean cells, `moves_chat` minus `moves_only` gives deal rate -0.233 $[-0.450, -0.100]$ and normalized welfare -0.106 $[-0.221, -0.035]$ for Qwen3-8B, and -0.083 $[-0.150, -0.050]$ / -0.041 $[-0.056, -0.023]$ for Qwen3-32B. For the Ministral slot it is -0.100 $[-0.150, -0.050]$ / -0.028 $[-0.048, -0.006]$, small but resolved. Qwen3-4B (-0.100 $[-0.300, 0.200]$) and OLMo (-0.017 $[-0.200, 0.150]$) are unresolved, and for the Gemma slot cheap talk *improves* normalized primary welfare ($+0.129$ $[0.083, 0.194]$) while leaving deal rate essentially untouched. So a public non-binding channel is mildly harmful for the two larger Qwen models, unresolved for two, and mildly helpful for one.

The contaminated version of this measurement reported -0.650 for Qwen3-8B, -0.417 for Qwen3-4B and -0.533 for Gemma, collapsing to -0.083 at 32B — which read as a scale law with an obvious story. The mechanism was mechanical instead: talking makes an episode longer, a longer episode gave the broken batched engine more chances to fabricate a turn, and a fabricated turn cannot close a deal. **The -0.42 / -0.65 / -0.08 ladder should not be cited.** What survives is a flat, small effect with no scale ordering, plus the earlier and still-correct negative claim: there is no single "cheap talk hurts LLM negotiators" headline.

The prompt scaffold closes even fewer deals than we reported, and every quality advantage it appeared to have was contamination. The five-step solver-workflow scaffold (score the table → read opponents → generate candidates → decide by rule → act; matched to canonical in wire format and generation budget, differing only in the appended reasoning guidance) was originally measured on `p2_Qwen3-8B_scaffolded` and `p2_Olmo-3-7B_scaffolded` — the *two most contaminated cells in the campaign*, at 54.6% and 67.0% fabricated turns. In each case the contrast compared a dirtier cell against a cleaner one *in the direction of the reported effect*: a fabricated turn is a no-op that can neither close a deal nor incur regret, so it depresses the scaffold's closure while flattering its per-turn metric. Both cells have now been re-run clean (0.000 fabricated, 120/120

[3] Two instrument caveats, both against leading with taxonomy rows. First, **the control's IR-violation rate is absent, not 0.000**: its `taxonomy_frequency` block has no key for that category at all. Substantively a threshold-respecting policy almost certainly never violates IR, so 0.000 would be the correct value — but reading a missing key as a zero is the same inference pattern that let the placeholder contamination pass as clean data, and it needs a positive check in the annotation layer before it is quotable. Second, the *anchoring* category is unusable as a defect metric here: the rational control fires it on 0.917 of episodes, more often than every model but two, because holding a fixed extreme position without new information is what an optimizing policy does by design. Any anchoring-based claim is therefore dropped from this paper rather than reported with a caveat.

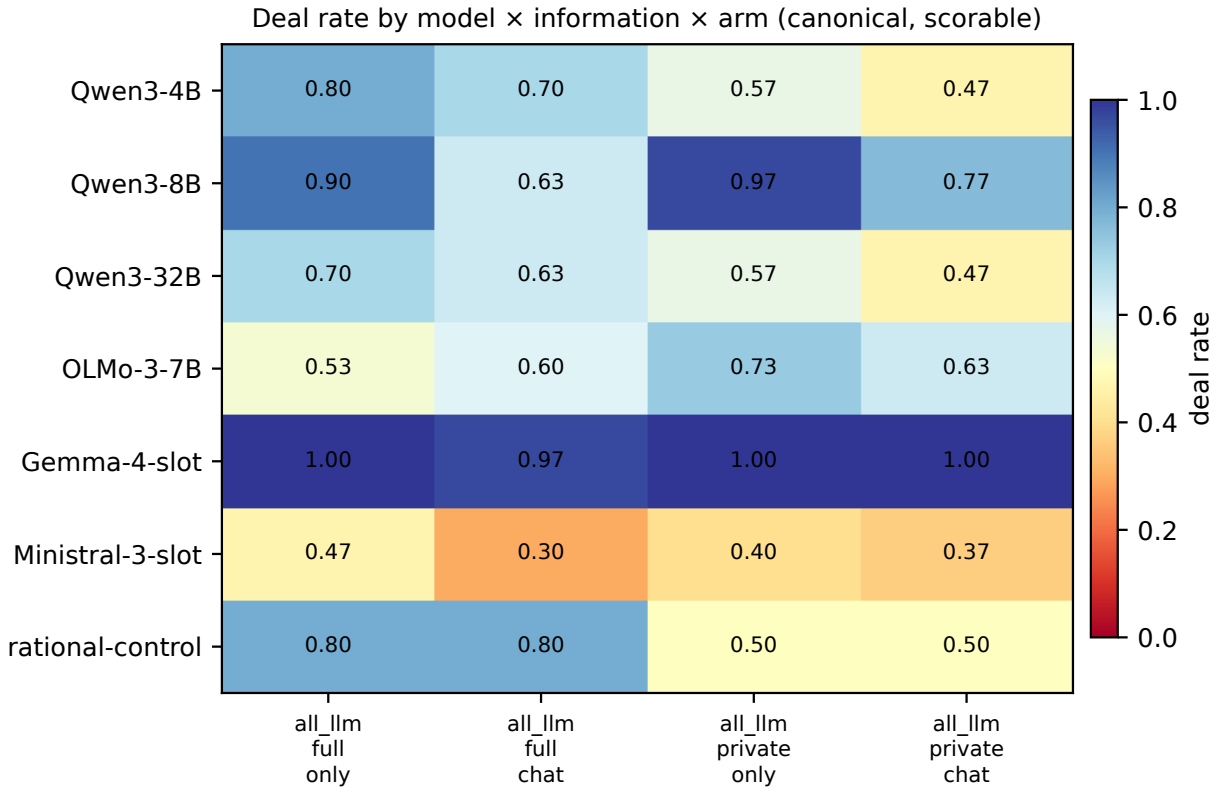

Figure 2: Deal rate across the clean canonical scorable cells. Rows are logical model slots plus the all-rational control; columns cross information condition (full/private) with arm (`only` = moves only, `chat` = moves plus a public cheap-talk channel), for `all_llm` tables (every seat an LLM). Cell values are mean deal rates over the 30 episodes in each cell (three payoff games × ten rollout seeds); blue is more agreement. **Read this figure as closure volume only, not quality**: the bluest row is Gemma, which reaches agreement almost always and violates individual rationality in 94% of episodes (§5.4). The `mixed` columns that the contaminated version of this figure carried are dropped, because the clean re-cut covers `all_llm` and `all_rational` only; the mixed cells were measured clean and are reported separately in §5.6. In the contaminated version OLMo read as a "conspicuous exception" that closed more and lost less — it was the one `all_llm` cell whose turns the model had actually produced, and the exception disappears here.

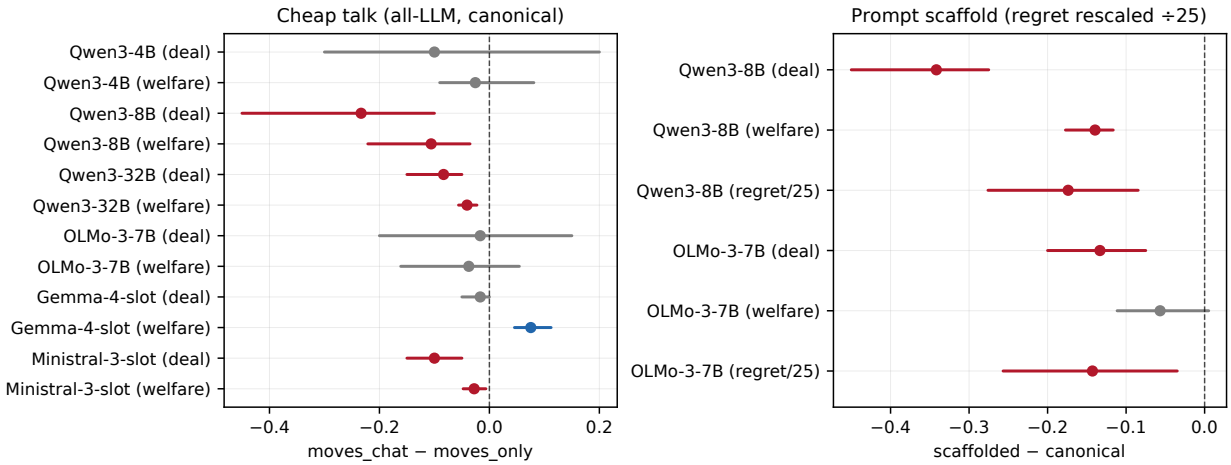

Figure 3: Left: the cheap-talk treatment effect (`moves_chat` minus `moves_only`) in canonical all-LLM tables on clean cells, 60 matched pairs per row, deal rate and normalized unconditional welfare for each of six models. Adding a public non-binding message channel costs Qwen3-8B a quarter of its deal rate and Qwen3-32B a twelfth; the other four are unresolved or, for Gemma, mildly positive. **Right: the prompt-scaffold effect** (five-step solver workflow minus canonical prompting) for the two models where it was run, 120 matched pairs each, both cells re-run clean. Regret rows are divided by 25 so they share the axis; a *negative* regret difference means the scaffold's decisions are *closer* to the oracle. The scaffold improves the per-turn metric and degrades closure on both models — and the OLMo row is now a real measurement rather than the artifact it was, since the published version of that cell recorded a deal rate of exactly zero at 67% fabricated turns. Colors as in Figure 1.

episodes each). For Qwen3-8B, where the canonical cell is also clean, the contrast is computable like for like:

Scaffolded − canonical	**Qwen3-8B**		**OLMo-3-7B**	
	pub.	**clean**	pub.	**clean**
Deal rate	−0.217	**−0.342** [−0.450, −0.275]	−0.642	**−0.133** [−0.200, −0.075]
Normalized welfare	−0.082	**−0.140** [−0.177, −0.116]	−0.224	−0.057 [−0.111, 0.005]
Mean per-turn oracle regret	−7.53	**−4.35** [−6.90, −2.12]	−21.33	**−3.57** [−6.42, −0.87]

The closure penalty grows for Qwen3-8B and the regret advantage halves — both moves in the direction the asymmetry predicts, which is the reason to believe the clean column rather than merely prefer it. The scaffold was the dirtier cell of the pair (54.6% against 40.0%), so its fabricated no-ops had been suppressing its own bad outcomes while costing it no regret. **The OLMo column is not a shrinking effect but a different kind of correction**, and it is the more serious one: at 67.0% fabricated that cell recorded a deal rate of **0.000**, and the paper reported a −0.642 penalty on that basis. Clean, the cell closes **0.492** and the penalty is −0.133. **"The scaffold reduces OLMo to never closing a deal" is withdrawn** — that zero was the harness, not the model. Unlike the dead `all_llm` run of §3.6, this cell *was* in the selection manifest and did feed the published analysis. Two of the campaign's three striking zeros turned out to be the same bug, which is the strongest single argument for the liveness check of §3.6: a cell reporting that a model never did anything deserves a positive check that the model was asked.

Every reached-deal quality advantage the scaffold appeared to have is gone. On clean cells none of the conditional metrics resolves in the scaffold's favour for either model: for Qwen3-8B, distance to the Nash solution is −0.011 [−0.103, 0.035], to Kalai–Smorodinsky +0.017 [−0.017, 0.035],

to the Pareto frontier −0.013, and reached-deal Gini −0.003, all covering zero, against published advantages that resolved. The one conditional metric that does resolve goes the *other* way and is small (welfare among reached deals +0.034 [0.019, 0.066]). Cell-mean figures tell the same story on the two axes the atlas does not pair: individual-rationality violations −0.175 → −0.025, and conditional Nash welfare +18.70 → +3.51, a fall of 81%.[4]

So the headline survives and strengthens: "the scaffold makes individual decisions look more oracle-like while closing far fewer deals." Any reading in which it produces *better* agreements when it does close does not. One caveat bounds all of the conditional rows in both vintages: they are selection-affected, because the scaffold closes far fewer deals and so computes its reached-deal metrics over a different and smaller set of episodes (47 pairs for Qwen3-8B against 120 unconditional). That is a reason to read them as "no surviving advantage" rather than as measurements of agreement quality in either direction.

An independent, fully clean replication comes from the P4 cross-game fleet, which includes a scaffolded cell on the same backbone (§7.5, 60 episodes per cell at 0.000 fabricated): untrained canonical prompting beats the scaffold by +0.217 of deal rate [0.000, 0.350] while the scaffold posts mean per-turn regret **3.97 points lower** [0.459, 6.355], an interval excluding zero. So the pattern holds on two independent clean measurements. The scaffold therefore still fails as goal 4's baseline-to-beat in the most awkward possible way — it is not a strong baseline — which is why the P4 claim "trained beats scaffolded" was uninformative (§7.5).

5.6 Result: the mixed table, and what survives matching

Contamination status of this subsection: clean. Every `mixed` cell audits at 0.000 fabricated turns, because mixed tables route through the async participant pool rather than the batched one (§3.6) — which is also the observation that localized the bug. These numbers are unchanged from the original publication and were re-derived from the same immutable atlas as a deliberate control on the revision: a figure that *should* be numerically identical, and is. One caveat does attach: the matched same-seat counterfactual references the *original* all-rational control, which later moved (deal rate +0.083) for belief-modelling reasons unrelated to fabrication (§3.6). Since the rational side of this contrast got *better*, re-matching would if anything widen the estimates below, which already favor the rational side and already fail to resolve.

Table 5: Mixed-table surplus capture, in normalized percentage points on each participant's own available-surplus scale. **Within-table**: LLM seat-0 capture minus the mean capture of the rational seats in the same episode — descriptive, and confounded with seat identity. **Matched seat 0**: LLM capture minus the rational policy's capture on the *identical* seat-0 sheet, instance, seed, arm, and information condition — the causal comparison. Bold marks intervals excluding zero.

Model in the seat-0 LLM role	Within-table difference	Matched seat-0 difference
Gemma-4 slot	**−0.157** [−0.239, −0.039]	−0.121 [−0.363, 0.134]
Ministral-3 slot	**−0.186** [−0.270, −0.041]	−0.134 [−0.338, 0.117]
OLMo-3-7B	−0.120 [−0.244, 0.059]	−0.055 [−0.194, 0.168]
Qwen3-4B	−0.137 [−0.229, 0.021]	−0.067 [−0.244, 0.155]
Qwen3-8B	−0.134 [−0.236, 0.031]	−0.045 [−0.210, 0.174]

[4]The paired estimates above and the cell-mean figures in note 0016 disagree in sign on distance-to-NBS (−0.011 paired against +0.018 unpaired) and in magnitude on distance-to-KS. Both are indistinguishable from zero, and the paired estimate is the better estimand and the one every other table here uses, so it is what the text quotes. This is precisely why the correct reading is "the advantage does not survive" rather than "the scaffold is now worse on fairness" — the sign of a null is not a finding.

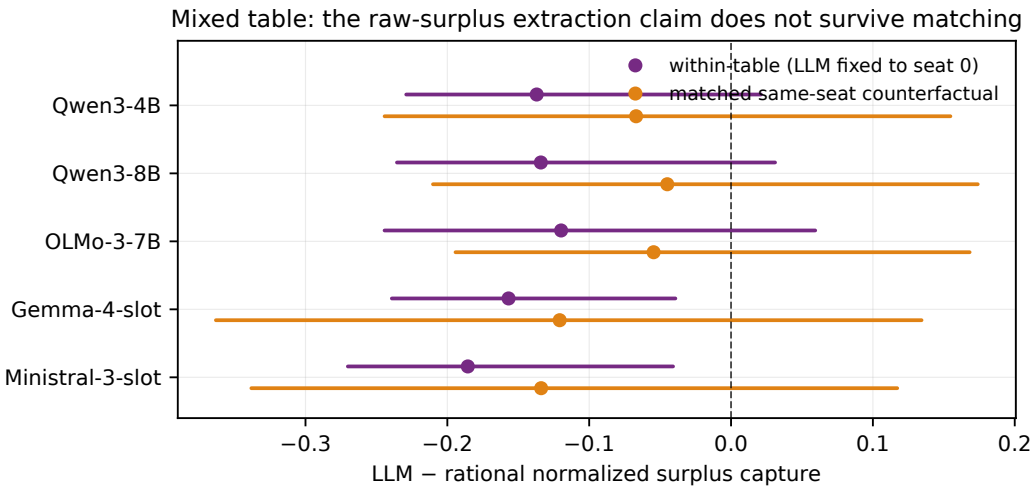

Figure 4: Does a rational co-player extract surplus from an LLM at the same table? Purple: within-table contrast — the LLM's normalized surplus capture minus the mean capture of the five rational seats in the same episode, with the LLM fixed to seat 0. Orange: the preferred counterfactual — the same LLM's seat-0 capture minus a *rational* agent's capture on the identical seat-0 score sheet, matched on instance, seed, arm, and information condition. Capture is $z_i = (u_i(d) - \tau_i)/(\max_d u_i(d) - \tau_i)$, invariant to per-party affine rescaling. Every point estimate favors the rational side, but the matched contrast is roughly half the size and every one of its five intervals crosses zero. The originally reported raw-surplus "rational seats win by 28–33 points" claim is retracted.

Within mixed tables, all five point estimates favor the rational seats by 12.0–18.6 normalized percentage points, and two intervals exclude zero. But the within-table contrast cannot separate an LLM effect from a fixed-seat-0 effect, and the preferred matched counterfactual — replace the LLM with the rational policy in *the same seat* — shrinks the estimates to 4.5–13.4 points with *every* interval crossing zero. **This campaign does not establish scale-invariant rational-agent extraction for any model.** A causal claim needs the LLM rotated through all six score sheets, which is the design of the companion reverse-mixed campaign (§14.1).

5.7 Preset equilibrium regret and divergence localization at scale

Table 6: Mean equilibrium regret in the single-issue presets, in the game's own units. Regret is the acting party's loss relative to the preset's closed-form equilibrium prediction (subgame-perfect equilibrium for ultimatum; Baron–Ferejohn/Okada for divide-the-dollar). Every `ultimatum` cell audits at 0.000 fabricated turns and is clean. The divide-the-dollar column is not: OLMo's cell is **100% engine-fabricated** — a dead cell, not a difficult game — and Qwen3-8B's is 8.6% fabricated, so its value is quoted with the rate attached. **These are the only numbers in the paper still resting on a contaminated cell**; the preset cells were not re-run, because no claim in the paper depends on them.

Model slot	Ultimatum	Divide the dollar
Gemma-4 slot	0.000	0.079
Ministral-3 slot	0.833	0.068
OLMo-3-7B	0.000	*dead cell* (1.000 fabricated)
Qwen3-4B	1.833	0.027
Qwen3-8B	0.000	0.025 (0.086 fabricated)

The OLMo divide-the-dollar entry is worth one sentence of its own, because the original paper got it wrong in a specific and instructive way. It was reported as "*unavailable*: reached 0/15 deals and produced no scorable equilibrium divergence", and praised as a case of refusing to impute a missing value. It was in fact a cell in which **every single turn was fabricated by the engine** — the model was never called, so of course it closed no deals. The honest-looking non-imputation was reporting a harness failure as a model result. Marking something unavailable is not a substitute for knowing *why* it is unavailable.

Divergence localization was recomputed on the clean 840-episode re-cut. Oracle rows localize an *earliest divergence turn* for **837 of 840** episodes, and divergence is even more immediate and more pervasive than the contaminated data showed: **775 of 837** (92.6%) first diverge at zero-indexed turn 0, the mean earliest divergence turn is **0.21**, and the mean episode-level divergence-turn share — the fraction of an episode's valued turns that are flagged — is **0.716**. The direction of the change is exactly what the artifact predicts: a fabricated turn is unvalued, so it could neither be flagged nor enter the denominator, and removing them raises both the share and the immediacy.[5] Supported chain-of-thought step localization exists for **0** episodes from stored data in either vintage, and is explicitly marked unavailable rather than inferred from unsupported scratchpad text — which is why the dedicated localization experiment of §6 was run on live re-prompting instead.

6 Goal 2: localizing the divergence inside the reasoning

6.1 The question and the control that changed the answer

Turn-level localization (§5) says divergence starts at turn 0 and is pervasive. The finer question is *where inside a single decision* the error enters: if the model's chain of thought goes wrong at a particular step, step-level process supervision becomes the natural intervention.

The experiment: take Qwen3-8B's flagged turns from the canonical all-LLM cells (150 turns sampled across regret deciles), split each stored scratchpad into steps, and re-prompt the model with the scratchpad truncated to the first k steps — a prefix probe.

Contamination status of this section: clean, by two independent arguments. The source cell was 40% engine-fabricated (§3.6), so the sampling frame deserves a check. First, a fabricated turn has no oracle value and therefore no regret and no divergence flag, so it cannot enter a sample drawn from *flagged* turns across regret deciles — the frame excludes them by construction. Second, and decisively, every number in this section comes from a *live re-prompt* whose completion was stored and re-valued by the oracle (150/150 verified), not from a stored campaign turn. The stored scratchpad is used only as prefill text. Nothing here is affected. The action it then emits is re-valued by the same exact oracle (verified 150/150). The **first divergent step** is the smallest k whose induced action is worse than optimal by more than the noise floor. 619 probes were run in total.

The initial read of the raw numbers was "the framing goes wrong at step 1 in about 85% of turns", i.e. the model commits to the wrong frame immediately. **The $k = 0$ control overturns that reading.** Re-prompting the same turns with *no scratchpad at all* already induces a divergent action 90.3% of the time, at mean regret 31.8.

[5] The contaminated campaign-wide figures were 1,573 of 1,770 localized, 1,070 at turn 0, mean 0.76, share 0.543 (themselves a correction to research note 0002's 1,121 / 0.70 / 0.536, which did not reproduce from the atlas artifact and are superseded by an errata in that note). All are withdrawn in favour of the clean figures above. The qualitative claim — divergence is usually immediate — was never at risk and is now stronger.

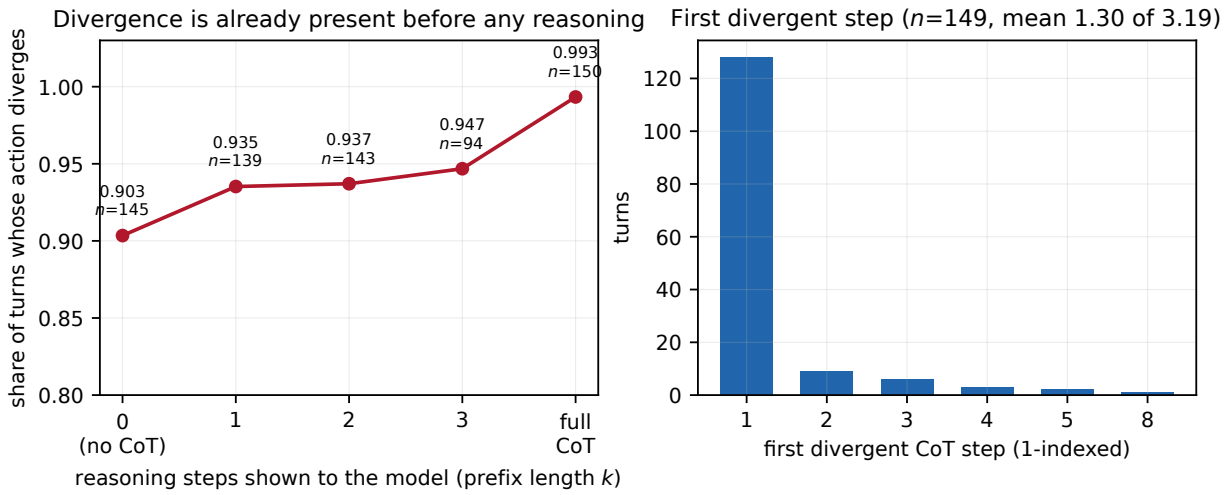

Figure 5: Left: the empty-scratchpad control. Share of flagged turns whose re-prompted action is divergent, as a function of how many of the model's own reasoning steps are shown back to it. $k = 0$ is the control (the scratchpad field is present but empty), and "full CoT" replays the entire stored chain. Divergence is 0.903 with *no* reasoning and 0.993 with all of it — reasoning contributes about nine points, and the error is overwhelmingly already present in the board state and prompt. n per point varies because probes that fail to parse are excluded rather than imputed. **Right: the first divergent step** among the 149 localizable turns; the mass at step 1 (128 turns, 86%) is what the left panel reinterprets — step 1 is where divergence is first *observable*, not where it is introduced.

6.2 What the localization actually shows

- **The sequence is 0.903 (no reasoning) → 0.935 (one step) → 0.993 (full chain).** The divergence enters with the prompt and the board state, before any reasoning is emitted.
- **Only 18 of 149 turns (12%) are genuinely reasoning-introduced**, and they have one mechanism, in 14 of 18: the no-reasoning baseline correctly `ACCEPT`s a good standing offer, and the reasoning talks the model out of it — usually into tabling another proposal instead. There are **zero** turns where reasoning rescues a divergent baseline.
- Post-parse-fix descriptive figures: step-1 share 86% (128/149); mean first divergent step 1.30 of a mean 3.19 steps; correlation between step position and regret only $r \approx 0.16$–0.18; and greedily re-prompting with the full stored chain reproduces the model's original action 90% of the time, which is the sanity check that the probe is measuring the same decision process.
- The dominant flags on these turns are `premature_accept` (115) and `below_threshold_accept` / `should_accept` (66 each) — acceptance-timing errors, consistent with the design's prediction that the optimal-stopping oracle would be the most diagnostic per-turn signal.

6.3 Downstream implications

Three consequences survive and sharpen. First, **late-step process supervision is doubly dead**: the error usually is not *in* the steps, and where it is, it is at step 1. Second, **turn-level preference learning on the decision itself is the right granularity** — which is what P4 then tested. Third, **"add reasoning" is not a defense**: in this environment reasoning is at best neutral and at worst corrosive, and the corrosive direction has a specific, legible mechanism (satisficing on "clears my threshold" and then reasoning itself out of a good accept; see Appendix E for a verbatim example).

7 Goals 3 and 4: the P4 pilot as an honest negative

This is the phase the whole project was built to reach, and it is the phase that failed. The full arc is reported because every step of it was a pre-registered decision whose outcome is informative.

7.1 The dataset, and three poisons excluded from the training mix

From the **1,590** validated LLM-seat episodes spanning 22 selected cells, **12,761 turn-level preference pairs** were built (dataset v2; three build generations, all audited):[6]

- `prompt` = the *stored* exact rendered view for that turn — the literal `[system, user]` message list the model saw at that decision point (100.00% view coverage, zero episode drops; this is the reproducibility wave paying off).
- `chosen` = the best-response oracle's preferred action, re-rendered into the models' own fenced-JSON wire format, with a rationale synthesized *only* from oracle numbers. Every sampled chosen action parses back through the harness parser (1,472/1,472).
- `rejected` = the model's actual emission for that turn, verbatim.
- `meta.regret` = oracle divergence in surplus points; `meta.weight` = a per-game-normalized, capped loss weight, because regret is not comparable across presets when the oracle is exact and the games use different numeric scales.

Splits are at *episode* level (turns of one dialogue share prompt prefixes) with two clean holdouts: OLMo-3-7B entirely (3,866 pairs, held-out backbone) and instance `game2` in both information conditions (2,713 pairs, held-out game); train 5,533 pairs over 719 episodes, in-distribution eval 649 over 77; all pairwise overlaps zero. The set's flavor, in one example: a Ministral seat accepting a deal **120 points below its own threshold** while its scratchpad claims to be maximizing points. Mean regret 37.75, minimum 0.01, maximum 306.

Contamination status of the dataset: clean, structurally. A fabricated turn carries no oracle value, so its regret is null and it cannot become a preference pair — the pair builder's own `no_action_emitted` gate drops it, and the 12,761 pairs are all real emissions. Two secondary things are *not* clean and are worth stating rather than burying. Fabricated turns *do* pick up taxonomy flags in the annotation layer (672 of 1,316 in one audited cell), so any statistic of the form "which flag fires earliest in an episode" computed over contaminated cells is unreliable, and none is used as evidence here. And because the pair yield per episode depends on how many turns were real, the per-cell pair counts are a function of the contamination rate as well as of the design — which affects the *mix*, not the correctness, of any individual pair.

The methodologically important part is what was *excluded*, and why. Three strata were identified as training poisons:

P1. **Final-vote pairs** (1,102 of the 1,110 forced-final-vote pairs, 99.3%, and the highest-regret stratum). The oracle's "chosen" action legally accepts a *different* standing offer than the one the mediator put to the vote. Training on these teaches a model to ignore an explicit instruction.

P2. **Private-information pairs** (6,547 by research note 0003's count; the dataset's own stats file does not break out this filter campaign-wide, though the ruled Qwen3-8B cohort's 844 dropped

[6] Research note 0003 quotes 1,530 episodes and 12,767 pairs; the 12,767 figure is the v0 build and the episode count does not match `stats.json` in either build. The v2 numbers used here are read directly from `p4_dataset_v2/stats.json` (`per_cell` sums to 1,590 episodes over 22 cells; `n_pairs` = 12,761). Other v0/v2 drifts corrected the same way: mean regret 37.75 (not 39.7), split sizes 5,533 / 649 / 2,713 / 3,866 (not 5,539 / 651 / 2,711), and 21 phase-illegal chosen actions gated at source (not 38).

pairs are recorded in `data_summary.json`). The oracle's action and its synthesized rationale derive from score sheets the model could not see. Imitating an omniscient, unimplementable policy risks training confabulated certainty. (The privileged-information literature uses such signals in *critics*, not as imitation targets.)

P3. **Scaffolded-cell pairs** (252 in the ruled cohort). Training on scaffold-produced emissions and then evaluating under canonical prompts would mix the two interventions that goal 4 exists to separate.

The pilot mix is therefore `phase == 'turn'` **and** `info == 'full'` **and** `scaffold == 'canonical'` **and** own-model, all three exclusions retained as ablation flags: **326 training pairs, 62 eval pairs, 81 optimizer steps**, LoRA r=32/α=64 on all linear layers, DPO $\beta = 0.1$, regret-weighted, one epoch. Deliberately small: the verdict was pre-registered to come from deal-level held-out evaluation, not from step count. A source-level gate was also added after finding pairs whose "chosen" action the harness itself would reject as phase-illegal (21 such turns are dropped in v2).

7.2 The audit chain, and a fourth poison it caught

A replay verification in the localization lane found that `BestResponseOracle.propose_values` read the *constructor* seat while `evaluate` resolved the acting seat locally — and the run harness reused one `BestResponseOracle(0)` for every seat. Consequence: **every stored PROPOSE action in the campaign was priced as if seat 0 were proposing** (accept/reject/walk unaffected). In the audited cell, 452 of 551 propose turns changed value under a per-seat-correct oracle (max $|\Delta|$ 53.2, both signs) and 23 turns flipped flagged↔unflagged. It was two same-class bugs, not one: `AcceptanceOracle.reservation` was also seat-0-bound, scoring every seat against seat 0's stopping threshold (correct per-seat values on a real instance span 47.6–79.3, all of which had been scored against 47.6), which is what had been driving the taxonomy's acceptance-flag rates.

Blast radius, stated precisely: the **goal-1 headline metrics are outcome-based and unaffected** (deal rate, welfare, frontier/Nash/Kalai–Smorodinsky distances do not consult the oracle); regret-based atlas rows shift for propose-heavy cells. The campaign was re-annotated 24/24 cells into `annotations_v1` with v0 preserved, and the bug effect was isolated from a coverage effect using a measurement-only legacy double: regret was re-priced on 14,370 of 35,005 valued turns, campaign mean −5.4% (28.10 → 26.58), every multi-issue cell down 4.5–7.5%, and **ultimatum cells exactly zero** — single-issue responders cannot propose, which is a built-in scope check that the fix did what it claimed. 559 phantom divergence flags were removed and 2 added, 95.5% concentrated in mixed and all-rational cells.

The rebuild then revealed the fourth poison. Correct pricing moved the ruled cohort's chosen-action mix from propose 59 → 98, reject 152 → 127, and **walk 23 → 0**: every walk target was a pricing artifact. Under-priced proposals had made no-deal look optimal, so the earlier labels would have *trained the deal-abandonment failure mode that P2 measures* — invisibly, because the individual pair magnitudes looked plausible. Roughly 10% (49/478) of the ruled cohort's pairs had wrong targets before the fix.

Two further contamination incidents were caught and closed the same way: 4,470 silently budget-exhausted turns graded by a post-hoc replay retry bug that re-emitted the previous decision's oracle row under placeholder turn indices (the dataset's `no_action_emitted` gate had already blocked all 4,470 — defense in depth that worked), and an equilibrium-composition mirroring bug in the preset cells. After regeneration, dataset v2 is **bit-identical to v1** in prompt, chosen, rejected, weight, regret and chosen action across all 12,761 pairs; the only movement is a taxonomy flag returning on 184 pairs, and the trainer provably never reads that field (verified by enumeration plus

a construction-pinned test).

7.3 The style confound, and the two-arm design that quantified it

A partial training curve showed something alarming and was recorded *before* any verdict so it could not be post-hoc: **DPO separability saturated almost immediately** — loss $0.683 \to 0.000$ by about step 20, reward accuracies pinned at 1.0, margins above 15, and eval log-probabilities of ≈ -347 for chosen versus ≈ -34 for rejected. The oracle-synthesized chosen text is trivially distinguishable from the model's own emissions, so the gradient might be teaching *write-oracle-style-text* rather than *decide-better*.

Measured at step 0 on the untrained model, the oracle-synthesized chosen text costs -3.9 nats/token against -0.3 for the model's own emissions — a **13× likelihood gap before any training**. The pilot was therefore upgraded to two arms differing *only* in the action fields:

- `as_built` — the full pairs as constructed (oracle rationale versus the model's own prose and action), 326 pairs.
- `action_only` — scratchpad and message stripped from *both* sides, so the pair differs only in the decision. Style-immune by construction, 329 pairs.

A pre-registered decision table governed the read: both arms improve ⇒ the style shortcut is harmless; only `as_built` moves ⇒ the apparent learning was style transfer; only `action_only` moves ⇒ the rationale text was actively hurting.

The decomposition on the same 62-pair split is stark: chosen-versus-rejected separation is **312.7 nats for `as_built` and 14.1 nats for `action_only` — 95.5% of the DPO training signal was rationale style and 4.5% was the decision**. The rejected side corroborates it airtightly: the model's own ≈84 tokens of prose cost it 0.04 nats (self-generated text is nearly free under its own distribution) while the oracle's prose costs about 299. The training curves match the diagnosis — `as_built` floors its loss to 3×10^{-7} by step ≈15, whereas `action_only` is a genuine gradual optimization reaching accuracy 1.0 only around step 35 with eval loss 0.05, five orders of magnitude higher.

A second pre-registration followed from `action_only`'s curve: its margin came mostly from *suppressing the model's own action* (rejected log-prob $-34 \to -106$) rather than raising the oracle's ($-48 \to -67$) — the classic DPO squeeze — predicting possible drift to third actions. The evaluation was therefore pre-committed to read deal rate and welfare *before* regret, while watching walk and malformed-emission rates.

7.4 First verdict: a clean positive on one held-out game

Contamination status of this section: clean, measured. All 16 cells of the single-game evaluation fleet audit at **0.000** fabricated turns — the batched pool never tripped on them — so every number in this subsection stands exactly as originally published. This is not an assumption: it is the same detector applied to the same stored episodes that condemned the cross-game fleet in §7.5, and the split between the two fleets is one of the more useful facts in the audit.

The evaluation fleet was 12 cells: {canonical, scaffolded, `as_built`-final, `as_built`-step10, `action_only`-final, `action_only`-step10} × {full, private} on the held-out `game2` instance group, ten seeds × both protocol arms per cell, oracle-annotated with per-cell atlases in-job. One method correction was forced: a single-instance holdout degenerates instance-cluster bootstrapping (the interval collapses onto the point estimate), so **clusters were rollout seeds — intervals cover**

seed noise, not cross-game generalization. That caveat was recorded at the time, and it is exactly what came due.[7]

Table 7: Per-cell means on the single held-out game (`game2`), 20 episodes per cell. **deal** = fraction reaching agreement; **primary** = normalized joint welfare; **regret** = mean per-turn oracle regret in surplus points (lower = more oracle-like); **bad/ep** = malformed emissions per episode. "step10" is the pre-saturation checkpoint (10 optimizer steps), "final" the full 81-step budget. Note the two collapsed rows: both *final* checkpoints of the non-`as_built` objectives close almost no deals while posting the *lowest* regret in the table.

Policy	Info	deal	primary	regret	bad/ep
canonical (untrained)	full	0.80	0.653	24.23	0.00
canonical (untrained)	private	0.90	0.642	28.81	0.00
scaffolded (prompt baseline)	full	0.65	0.587	25.68	0.05
scaffolded (prompt baseline)	private	0.45	0.428	24.11	0.15
`as_built` step10	full	**1.00**	**0.847**	25.69	0.00
`as_built` step10	private	0.85	0.703	30.58	0.00
`as_built` final	full	0.95	0.786	36.77	0.00
`as_built` final	private	0.80	0.591	38.25	0.15
`action_only` step10	full	0.75	0.618	26.31	0.00
`action_only` step10	private	0.80	0.633	32.68	0.00
`action_only` final	full	0.00	0.000	21.68	0.30
`action_only` final	private	0.05	0.037	22.44	0.25
`sft_style` step10	full	0.85	0.693	27.36	0.05
`sft_style` step10	private	0.90	0.712	37.04	0.00
`sft_style` final	full	0.15	0.090	20.44	0.70
`sft_style` final	private	0.20	0.143	22.16	0.45

On this design the goal-4 claim held with intervals excluding zero. The policy of record (`as_built` step-10, pre-saturation) beat untrained canonical prompting by **deal rate** $+0.200^{*}$ **and welfare** $+0.194^{*}$, and beat the prompt scaffold by $+0.350^{*}$ **and** $+0.260^{*}$, with fewer individual-rationality violations than canonical (0.20 versus 0.35 per episode). Under private information both trained arms still beat the scaffold ($+0.400^{*}$/$+0.275^{*}$) but were flat against canonical.

Three sub-results came with it, and unlike the headline they *did* survive. **Early stopping is structural.** The `as_built` final checkpoint pays $+12.5^{*}$ regret for no outcome gain over step-10, and `action_only`'s final *collapsed* to a 0.00 deal rate — the pre-registered DPO-squeeze risk, manifesting as format destruction (malformed 0.30/episode) rather than the predicted drift to walking (walk rate 0.00): right cause, wrong predicted symptom. **Oracle imitation was refuted, not merely unsupported.** Training on oracle-preferred actions did *not* make actions more oracle-like: every trained arm's per-turn regret is greater than or equal to canonical's (step-10 flat at +1.5 n.s.; final $+12.5^{*}$ worse). **The read-order rule earned its keep.** The collapsed `action_only` arm has the *best* regret of any policy (-2.6 full, -6.4^{*} private) while closing zero deals, because a policy that never closes never takes a costly turn. Reading regret first would have crowned the dead arm.

A third arm, `sft_style` — supervised fine-tuning on the chosen *rationales* only, no preference objective, matched by construction (same trainer with `loss_type="sft"`, same bytes, same LoRA schedule and step count) — was then run to decide whether the gain was style alone. It came out flat against canonical (+0.050 deal / +0.040 welfare, both n.s.) and significantly below the DPO arm at matched steps (-0.155^{*} welfare, full information), and its over-trained final collapsed like the others

[7] The stored artifact `p4_pilot_paired_final.txt` carries a stale header reading "cluster bootstrap over committed instances" while its own `clus` column reads 10, i.e. rollout seeds. The narrative in research note 0004 is correct and is what this section follows; the header string in that file should be fixed.

(deal 0.15, errors 0.70/episode), showing that the collapse is objective-independent. The interim conclusion drawn from the three arms was that the gain required *both* channels: neither prose style alone (SFT refutes it) nor decision preference alone (`action_only` refutes it), but preference contrast *over* styled rationales.

7.5 The gain does not replicate on unseen games — and the retraction of it was itself an artifact

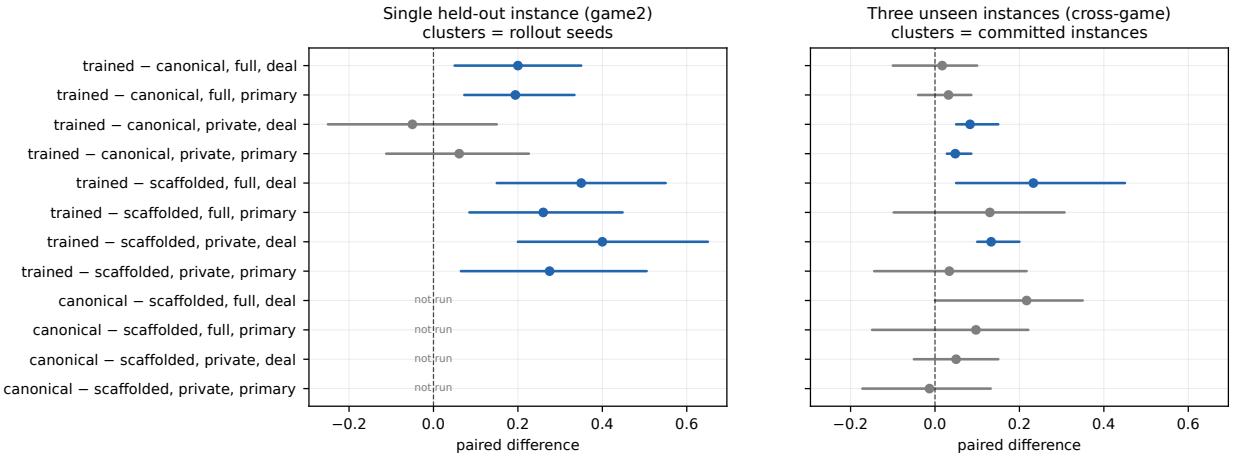

Figure 6: The single-instance-holdout trap, side by side. Identical contrasts, identical policies, two evaluation designs. **Left**: the original 12-cell evaluation on one held-out game (`game2`), 20 matched episode pairs per row, bootstrap clusters = rollout seeds — so intervals cover seed noise only. **Right**: the widened evaluation on three fully unseen instances × both information conditions (600 episodes), 60 matched pairs per row, clusters = committed instances — so intervals cover cross-game generalization. **The right panel is the clean re-run** (`_b2`, 0.000 fabricated turns in all ten cells); the left panel is unchanged because that fleet was measured clean. "trained" is `as_built` step-10. Colored bars exclude zero (blue = the first-named policy is better, red = worse); grey crosses zero. The four trained-versus-canonical rows go from two clean positives of about +0.2 on one game to two small resolved positives of +0.08 and +0.05 and two nulls on three — and crucially *not* to the two significant negatives of −0.183 and −0.157 that the contaminated version of this panel showed. The bottom four rows are absent on the left (the single-game fleet never ran that contrast) and on the right still show why the surviving "trained beats scaffolded" result is uninformative: the *untrained* baseline beats the same scaffold by a comparable margin, though no longer with a resolved interval.

The widened holdout is three fully unseen instances × both information conditions, five policies × ten seeds × two protocol arms (600 episodes), with instance-cluster bootstrapping.

This fleet was contaminated, and asymmetrically in the direction that matters. All ten cross-game cells were fabricating turns, at rates from 7.0% to 34.2%, and the `canonical` baseline — the cell every trained arm is measured *against* — was the **least** contaminated in both information conditions (0.233 full, 0.070 private) while every trained arm lost more (DPO 0.265 / 0.109; action-only 0.289 / 0.131; SFT 0.276 / 0.159; the scaffold worst at 0.342 / 0.278). Fabricated turns are scored as no-ops that cannot close a deal, so the contamination penalized the trained arms relative to their own baseline. All ten cells were re-run on the clean engine and re-audited at 0.000 fabricated. Before reading the clean column, the comparison script was validated by running it on the *contaminated* originals, where it reproduces all four published headline entries exactly (−0.033, −0.015, -0.183^{*}, -0.157^{*}) plus both scaffold contrasts — same shared cluster-bootstrap, same

instance clustering, same 2,000 draws. That exact reproduction is what licenses reading the two columns as like-for-like.

as_built step10 − canonical	**Full information**		**Private information**	
	deal rate	primary	deal rate	primary
Single held-out game (`game2`), clean	$+\mathbf{0.200}^{*}$	$+\mathbf{0.194}^{*}$	-0.050	$+0.061$
Three unseen games, *contaminated*	-0.033	-0.015	$-\mathbf{0.183}^{*}$	$-\mathbf{0.157}^{*}$
Three unseen games, **clean**	$+0.017$	$+0.032$	$+\mathbf{0.083}^{*}$	$+\mathbf{0.048}^{*}$

Three readings follow, and the order matters.

The `game2` positive was still specific to that single held-out instance. On three unseen games the trained arm is +0.017 deal rate $[-0.100, 0.100]$ and +0.032 welfare $[-0.040, 0.086]$ at full information — a tenth of the single-game effect, with intervals comfortably covering zero. Exactly the risk the seed-cluster caveat had flagged. **The single-instance-holdout lesson is untouched by the re-baseline**, which is worth saying plainly, because it is the lesson the pilot exists to teach.

But the stronger claim we published — that training actively *hurt* out of distribution — does not survive and is withdrawn. The two starred negatives that carried it become starred positives of similar magnitude ($-0.183^{*} \to +0.083^{*}$ deal rate; $-0.157^{*} \to +0.048^{*}$ welfare). Across all three trained arms and both information conditions, **10 of 12 trained-versus-canonical entries flip sign**, and the one that does not flip loses significance ($-0.183^{*} \to -0.017$). The direction of the correction is uniform and is exactly what the asymmetric contamination predicts. This does *not* license the opposite claim: most clean intervals include zero, there are three instance clusters, and a true effect of approximately zero explains everything — the contamination pushed the trained arms below parity and removing it moved them back to parity. **"Training did not help" stands; "training hurt on unseen games" is withdrawn as a harness artifact.** Nothing here is evidence that divergence-localized DPO transfers.

The goal-4 framing gets weaker in both vintages, for different reasons. `as_built` beats the scaffold on clean cells ($+0.233^{*}$ deal at full information, $+0.133^{*}$ private), and the untrained canonical baseline beats the same scaffold by a comparable margin (+0.217 deal $[0.000, 0.350]$, +0.097 welfare) — so "trained beats scaffolded" is still reporting the scaffold's weakness rather than the intervention's strength. What changed is that all four of the supporting entries for the sharpest version of that criticism *lost significance* on clean cells (full deal $+0.217^{*} \to +0.217$ n.s.; full primary $+0.127^{*} \to +0.097$; private deal $+0.267^{*} \to +0.050$; private primary $+0.135^{*} \to -0.013$). The point estimates lean the same way; the starred claim cannot be made. And the mechanism triangulation of §7.4 — "preference contrast *over* styled rationales" — still fails to reproduce out of distribution, where all three objectives land within noise of the untrained baseline, so it remains retracted rather than qualified.

Two caveats bound all of this in the honest direction. There are only three instance clusters per condition, so the intervals are coarse and the evidential weight is in cross-condition, cross-arm *consistency of direction* rather than in any single interval. And absolute levels are not comparable across the two instance banks: on the clean cross-game bank malformed emissions are near zero for every policy (0.00–0.20 per episode), whereas the contaminated version of this fleet carried ≈2.9 per episode — which the original paper attributed to a property of the fresh instance bank. That attribution was wrong. The malformed floor was a symptom of the same engine failure, and it disappeared when the engine path changed. It was never a bank property, and the corresponding limitation is deleted.

7.6 Final verdict

The P4 pilot produces no evidence that divergence-localized DPO improves negotiation outcomes on unseen games, and scaling it is not warranted on this evidence. It also produces no evidence that it makes them worse: that finding was the harness. Goals 3 and 4 remain *open*, not demonstrated, and the pilot's two published verdicts — a positive on one game and a negative on three — were both wrong, in opposite directions, for different reasons. What it delivered instead is the methodological content of §12.

8 The narrative-framing arm: the same game in a story

The setup, and why the environment is deliberately drab. Everything the parties in this environment negotiate over is abstract on purpose. The issues are called `issue0 ...issue4`, the options `opt0 ...opt2`, and the seats carry neutral first names. That de-anchoring is design constraint (b) of §2: in the benchmark this environment repairs, the public role text ("you are the union representative...") was *correlated with the private score sheets*, and a single agent that could not communicate at all — it read the role descriptions and proposed a deal by itself — matched the score of the full six-agent negotiation. If the story tells you the answer, the benchmark measures reading comprehension rather than bargaining.

The problem. Abstraction buys validity at the cost of realism, and the interesting question — *does dressing the identical game in a realistic scenario make an LLM negotiate worse?* — cannot be asked at all while the two are confounded. Naively re-attaching a story re-opens the leak, and then any observed "framing effect" is unfalsifiable: a drop could be distraction and a gain could be leakage. We wanted framing as a *treatment arm with a control*, not a return to preference-leaking role stories.

What we did. We built framing as a *pure relabelling of a committed game*. A framed instance is a deterministic transform of an existing one: score values, thresholds, round budget, veto structure, information condition, and even the `instance_id` are byte-identical, and only the issue names, option labels, and party role names change — into a data-centre siting deal (`datacenter`) or a summer-festival programme (`festival`). Crucially, the mapping from story to numbers is drawn from a hash of `(instance_id, framing)`, so "the developer" is no more likely than chance to be the high-threshold seat and "river draw" no more likely than chance to be the valuable option. The leak is removed *by construction* rather than by hand-checking wording (§8.1). Two tripwires enforce the contract: a masked-prompt byte-identity test suite, and a *runtime* control — the communication-free solo agent, which must not do better under the story.

The result. Narrative framing degrades this population, and the effect lands on quality and protocol discipline rather than on closing. Paired within game and seed on clean cells (Qwen3-8B in all six seats, 60 pairs over six committed instances), normalized Nash welfare — the primary, a 0-to-1 geometric mean of each party's surplus as a fraction of what it could have achieved, scored 0 for no deal — falls from 0.182 to 0.098 (-0.083, 95% CI $[-0.143, -0.015]$); the share of episodes containing a malformed action rises from 0.033 to 0.250 ($+0.217$ $[+0.117, +0.350]$); and among the 23 pairs where *both* arms closed, the realized deal sits further from the Pareto frontier ($+0.238$ $[+0.044, +0.515]$). Deal rate falls from 0.717 to 0.550, but that interval crosses zero (-0.167 $[-0.317, +0.017]$) and **the deal-rate headline this arm was first published with is withdrawn** (§8.2). Both controls

hold: the framed solo agent does not beat the abstract one (-0.169 $[-0.427, +0.089]$ on the primary, so there is no detectable preference leak), and 18 paired all-rational episodes reach *identical deals* under every skin, every metric difference exactly 0.000.

One-sentence version. Wrapping the identical game in a realistic story costs Qwen3-8B measurable welfare, pushes it into tabling syntactically perfect but incomplete deal packages, and moves it further off the efficient frontier, while a preference-decorrelated skin gives a communication-free agent no advantage and leaves computable bargainers bit-identical — so this is distraction, not leakage, and it is the sharpest LLM-versus-rational contrast in the project.

8.1 Design: where the skin lives, and why it is decorrelated

All game *content* already flows through one prompt scaffold, which renders issue names, option labels, score sheets, and optional party role text. That made the arm expressible experiment-side with *no library change*:

- `framings.py` defines a `Framing` — a situation sentence plus pools of party roles, issue names, and option labels — and two skins. `datacenter` is a Northgate siting agreement (a developer, a borough council, a grid authority, an electrical workers' union, a residents' assembly, a county executive, a water board, a landholder; issues like Location, Power supply, Build schedule, Hiring, Compensation, Cooling, Site access). `festival` is a Harrowgate summer festival (a festival trust, parks department, merchants' guild, stagehands, neighbours' forum, transit authority, safety office, broadcaster; issues like Venue, Dates, Programme, Curfew, Ticketing, Transport, Vending).
- `frame_instance` rewrites the issue names, option labels, and party names, re-renders the `named` views inside the committed solution bundle from their unchanged option-index deals, and stores the realized mapping at `payload.meta.framing`. Everything numeric, structural, and identifying is untouched, which is what makes a framed run pair episode-for-episode with its abstract twin and keeps the committed exact analysis valid.
- `frame_scaffold` swaps the de-anchored one-line intro for the situation sentence and turns on the party-roster role text, preserving the scaffold's class — so framing *composes with* the existing scaffold axis (including the goal-4 solver workflow) instead of forking it.
- `run.py --framing {abstract,datacenter,festival}` applies the transform to every loaded instance, writes the mapping to `<run>/framings.json`, records it in the run manifest and in every episode's `cell_cfg`, and persists the *framed* instances into `<run>/instances/` so the analysis layer scores exactly what the seats saw.

The randomization is the substantive design decision. The mapping is keyed on `sha256(instance_id | framing)`, which makes it **constant across rollout seeds** — all ten seeds of an instance share one framed prompt, so the paired `(instance, seed)` comparison holds the text fixed — and **re-randomized per instance**, so across a bank of games no role systematically occupies the demanding seat. Decorrelation is a property of the sampling procedure, not of careful wording, which matters because narrative labels *are* meant to invite inference; the point is that the inference they invite is uninformative. In the appendix excerpt (§E.3) the water board is not the party that values "river draw" — the electrical workers' union is, at 95 points — and the county executive scores zero on every cooling option.

The tripwires are the contribution. `tests/test_framings.py` (24 tests, 31 counting the analyzer's) enforces five properties over all six committed instances and both skins:

(1) **Numeric and structural identity**: identical sheet values, thresholds, rounds, information condition, chat flag, proposer, veto, min-accept, discount, breakdown risk, `instance_id`, ceiling/floor and option counts; solution deals unchanged as index lists; and the framed payload rebuilds a valid game whose deal parsing and per-party utilities agree index-for-index with the original.

(2) **The prompt says the same thing about the numbers**: both arms' seat system prompts are rendered, labels are replaced by placeholders, and the masked texts must be *byte-identical* once the situation sentence and role text are normalized away. This is strictly stronger than comparing the numeric token sequence — it rules out any extra hint, reordering, or omission introduced by the skin.

(3) **Preference-free narrative text**: no skin string contains a digit, and no role string contains preference vocabulary (`prefer`, `value`, `want`, `priorit`, `oppose`, `threshold`, `score`, `high`, `low`, `cost`, ...).

(4) **Randomized and reproducible mapping**: over 300 synthetic instance ids, the same id twice gives the same plan, every issue in the pool reaches slot 0 sometimes, every seat draws at least 6 of the 8 roles, and option index 0 of slot 0 draws at least 4 distinct labels.

(5) **Rational invariance end-to-end**, plus a CLI test that the framing, its mapping, and the framed instance files are all recorded.

An incidental catch worth recording. The canonical scaffold's worked examples hardcode the literals `"Site"`, `"Timeline"` and `"Tight"`. When the `datacenter` skin first used `Site` and `Timeline` as real issue names, the masked-prompt comparison failed — correctly. The example would have accidentally named a *real* issue in the framed arm and a fictional one in the abstract arm, an asymmetry outside the intended treatment. The skin now uses `Location` and `Build schedule`, and the constraint is documented at the label pool. A numbers-only check would have passed.

8.2 Cells, and the paired treatment effect

Six cells share the same six committed instances (three games × full/private information), the canonical scaffold, thinking off, and the `threshold/acceptance/bestresponse` per-turn oracle stack for the two LLM cells: an abstract control and a `datacenter` treatment (`all_llm`, `moves_chat`, Qwen3-8B, $6 \times 10 = 60$ episodes each); a solo tripwire on each side (60 each); and a rational-invariance pair (`all_rational` with boulware/conceder/bayes policies, $6 \times 3 = 18$ each). Every planned episode finished.

Contamination status. The two LLM cells were originally run on the broken batched engine (§3.6) at 28.7% and 37.3% fabricated turns, and — the part that made them uninterpretable rather than merely noisy — *the framed arm lost more real turns than its control*, which confounds a paired comparison directly. Both were re-run on the sequential path (`_b2`) at 0.000 fabricated turns, 1,258 and 1,431 turns respectively, and **every number in this section is from that clean vintage**. The solo and rational cells audited clean on both sides originally and are unchanged.

The effect relocated rather than weakened. This is the useful part of the correction. Decontamination moved the deal-rate estimate from -0.250 (interval excluding zero) to -0.167 (interval

Table 8: Narrative framing, paired within committed instance and rollout seed. Each row is framed minus abstract; brackets are 95% intervals from a 10,000-draw bootstrap that resamples whole *instance* clusters, with seeds staying inside their cluster (the same inference the confirmatory analysis uses). Rows marked *deals only* restrict to the pairs where **both** arms reached a deal, because mixing no-deal zeros into a distance metric would confound it with the deal-rate effect. The **contaminated** column is the originally published estimate on the fabrication-affected cells, shown so the direction of each correction is visible; it is not evidence and should not be quoted. Lower is better for the two distances and for malformed episodes; higher is better for the primary and deal rate.

Metric	abstract	framed	clean effect (95% CI)		contaminated
Normalized Nash welfare (primary)	0.182	0.098	**−0.083**	[−0.143, −0.015]	−0.068 n.s.
Malformed episode	0.033	0.250	**+0.217**	[+0.117, +0.350]	+0.117
Distance to Pareto frontier *(deals only, 23 pairs)*	0.146	0.369	**+0.238**	[+0.044, +0.515]	+0.107 n.s.
Deal rate	0.717	0.550	−0.167	[−0.317, +0.017]	−0.250 **(withdrawn)**
Distance to Nash solution *(deals only, 23 pairs)*	0.782	0.919	+0.156	[−0.030, +0.339]	−0.092 n.s. (sign flip)
Below-threshold agreement	0.283	0.233	−0.050	[−0.167, +0.100]	−0.100 n.s.
Turns taken (process, not outcome)	20.97	23.85	+2.88	[−2.62, +8.75]	+0.85 n.s.

crossing it), and both arms close far more often once real turns are restored (0.550 and 0.717 against 0.300 and 0.550). At the same time the primary, the malformed-episode rate, and the frontier distance all *resolved against framing* where they previously did not, and the frontier effect more than doubled. The verdict sharpens from "framing stops the table closing" to **"framing makes the table break protocol and settle on worse deals"**, which is a claim about the quality of play rather than its volume. This is the paper's triage heuristic (§12) running in the direction it predicts: a defect that deletes turns attacks *closure*-dependent quantities, and the deal-rate row was exactly the row that should have been distrusted first. The deal-conditional distance-to-NBS row flipped sign on roughly twice the pairs, which is the same warning in its strongest form — contamination attacks the denominator of every deal-conditional metric.

Per-instance heterogeneity. Mean paired effects, by committed instance: −0.172, −0.157, −0.122, −0.075, −0.033 and +0.060 on the primary (five of six negative), against deal-rate effects of −0.4, −0.3, −0.3, −0.2, 0.0 and +0.2 (four of six negative, one positive). With six clusters the primary is the better-behaved endpoint, and the figure shows why the pooled interval should be read alongside the per-game bars.

8.3 Mechanism: complete JSON, incomplete deals

On clean cells the failure has one signature and it is unusually legible. All 15 malformed framed episodes contain **exactly one** offending turn, and every one is a syntactically perfect fenced JSON `propose` whose `deal` object names only two or three of the five issues — for example {`"Location": "Kestrel Park", "Cooling": "reclaimed water", "Hiring": "open recruitment"`} — which the harness rejects as an illegal action. The framed arm records **15 legality errors and 0 syntax errors**; the abstract arm records 1 of each across its 60 episodes. That zero is the load-bearing negative: it rules out "the long multi-word labels broke the fenced-JSON format" as the explanation. The model formats correctly and *loses track of the issue list* while doing so. Economic errors — offers a party would not rationally table — also rise, 39 to 54.

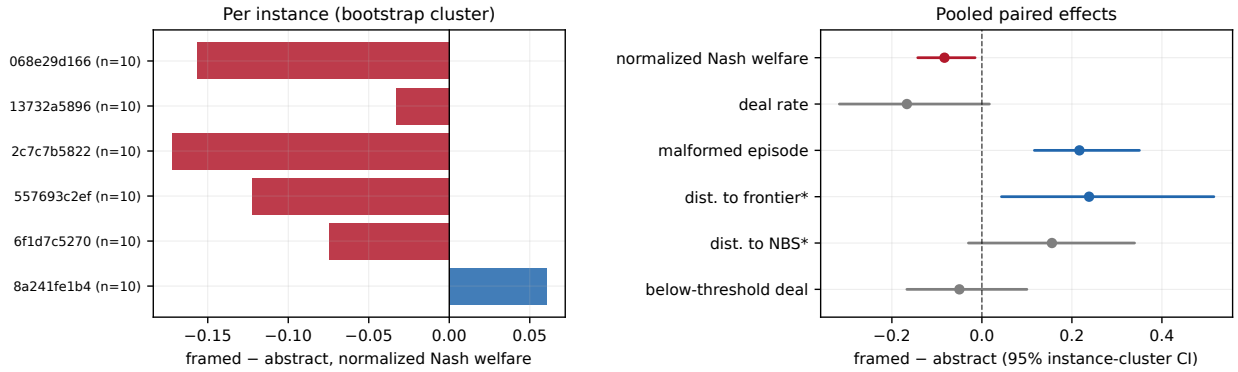

Figure 7: The framing effect on clean cells. *Left*: the mean paired effect on the primary for each of the six committed instances, which is the unit the bootstrap resamples — the effect is negative in five of six games, so it is not one unlucky instance, and the one positive game (`8a241fe1b4`, +0.060) is visible rather than hidden inside a pooled interval. *Right*: the pooled paired effect for every metric in Table 8; blue and red mark intervals excluding zero, grey marks unresolved, and starred rows are restricted to the 23 pairs where both arms closed. The withdrawn deal-rate headline is the second row from the top and now straddles zero.

A second erratum, established here: the "talks instead of moving" mechanism was the harness. The originally published mechanism for this arm was that the framed table substitutes deliberation for action: 570 talk-only turns against 435, with 25% fewer proposals. That reading does not survive. Screening the contaminated cells with the library detector shows **422 of those 435 abstract talk-only turns, and 567 of the 570 framed ones, were engine fabrications** — placeholder text no model produced, which parses into a well-formed no-op and was therefore counted as a deliberate pass. On the clean cells the deliberate-no-op count is **4 turns in each arm**, and the proposal share falls only from 20.0% to 16.4% of turns. So the "framed table talks more" finding was, almost exactly, a measurement of how much more of the framed arm the engine had eaten. This is the fourth independent place in the project where a summary statistic over stored output described the harness rather than the model, and the first where the artifact had been read as a *behavioural* mechanism rather than as an outcome.

8.4 The two controls, and the sharpest contrast in the program

Solo leakage tripwire. The communication-free agent sees every score sheet and proposes alone, so if the story carried preference information it would score *better* under the skin. It does not: primary 0.232 → 0.062 (−0.169, [−0.427, +0.089]), below-threshold agreement 0.600 → 0.883, distance to frontier +0.260 [−0.006, +0.547], `leak_detected=False`. The skin does not help a solo agent; if anything it hurts one, which is the same distraction effect in the single-agent limit. The honest scope of this control is stated in §13: its interval is about 0.5 wide, so it excludes a *large* leak, not a small one.

Rational-policy invariance. Eighteen paired all-rational episodes reach **identical deals** under abstract, `datacenter` and `festival`, with every metric difference exactly 0.000. The computable policies read the machine-readable negotiation-state block, which carries option *indices*, so relabelling cannot reach them. This is invariance by construction — and it is now measured rather than asserted, which is the only reason it is usable as a control.

8.5 Does a rational anchor survive the story?

The arm as described so far compares two *all-LLM* tables. That leaves the question the rest of the paper keeps asking unanswered here: if narrative framing degrades this population, does seating one computable bargainer among them help, and does whatever help it gives survive the story? We ran the missing cells — the `reverse_mixed` lineup of §14.1 (a Bayesian-rational policy in seat 0, Qwen3-8B in seats 1–5) under *both* skins, 60 episodes each on the framing pilot's own grid: the same six committed instances, the same ten rollout seeds, `moves_chat`, canonical scaffold, thinking off, sequential engine, 0.000 fabricated turns in both. With the two all-LLM cells of Table 8 as the other row, this is a 2×2 of {all-LLM, rational@0} $\times$ {abstract, datacenter} on one instance bank (Table 9).

The anchor helps more under the story, not less. Under the abstract skin, replacing one seat with the rational policy resolves on only two of the nine measures — the primary ($+0.086$ $[+0.014, +0.151]$) and the replaced seat's own capture ($+0.201$ $[+0.046, +0.352]$) — and every fairness measure points the right way without excluding zero. Under the `datacenter` skin the same swap resolves on **five**: the primary ($+0.101$), distance to the Nash solution (-0.263 $[-0.449, -0.074]$), distance to Kalai–Smorodinsky (-0.280 $[-0.471, -0.089]$), the worst-off party's capture ($+0.153$ $[+0.005, +0.293]$), and the malformed-episode rate (-0.133 $[-0.217, -0.050]$), plus capture again ($+0.262$). Every fairness point estimate is larger under framing than under abstraction. **One disciplined seat buys more when the table is distracted than when it is not** — which is the practically interesting direction, since the realistic condition is the one an application would face.

But the interaction resolves on legality only, and that is the honest headline. The difference of differences, $(\text{mixed} - \text{all\_llm})_{\text{datacenter}} - (\text{mixed} - \text{all\_llm})_{\text{abstract}}$, is the quantity that actually licenses "the anchor's effect grows under framing", and it excludes zero for exactly one measure: the malformed-episode rate, -0.150 $[-0.200, -0.083]$. The fairness interactions all lean the same way and none of them resolve (distance to the Nash solution -0.278 $[-0.699, +0.079]$; worst-off capture $+0.187$ $[-0.044, +0.484]$; primary $+0.015$ $[-0.108, +0.116]$), and the deal-conditional rows rest on 16 four-way-complete pairs, which is thin. So the defensible claim is narrower than the pattern of resolutions suggests on its own: **the rational anchor demonstrably cancels framing's protocol damage, and its fairness advantage is larger under framing in every point estimate without the growth itself being resolved.**

Framing stops degrading a table that has an anchor in it. Read down the other axis, the same cells say it more directly. In the all-LLM row, framing costs -0.083 $[-0.142, -0.016]$ of the primary and adds $+0.217$ $[+0.117, +0.350]$ of malformed episodes (§8.2, recomputed here on identical code paths and reproducing those published values exactly). In the rational@0 row, neither survives: the primary falls -0.068 $[-0.160, +0.024]$ and malformed episodes rise $+0.067$ $[-0.033, +0.167]$, both crossing zero, and distance to the Nash solution actually *improves* under the story (-0.084 $[-0.153, -0.016]$). The 15 incomplete-package proposals that are §8.3's whole mechanism become 7 with one non-LLM seat at the table.

The rational seat's own take does not move, and it is important to say why that is two claims. Its capture under framing changes by -0.089 $[-0.223, +0.055]$ — unresolved. Two distinct things are being asserted and only one of them is measured here. That the policy's *play* is framing-invariant is true by construction and verified separately: it reads the machine-readable

option indices, and the 18 paired all-rational episodes reach bit-identical deals under every skin (§8.4). That its *outcome* is framing-invariant is not automatic at all — it is bargaining against five LLMs whose behaviour demonstrably does change — and that is the thing this cell measures and finds no detectable movement in. Within the table the anchor still out-captures its neighbours under both skins (+0.196 [+0.097, +0.299] abstract, +0.162 [+0.111, +0.214] framed), so it remains a beneficiary against open-weight seats, exactly as §14.1 found and unlike the frontier population of §9.3.

Table 9: A rational anchor inside the story-framed game: {all-LLM, rational@0} × {abstract, datacenter}. Cell entries are means over 60 episodes; **effect** columns are rational@0 minus all-LLM, paired within committed instance and rollout seed, with 10,000-draw bootstrap intervals resampling whole *instance* clusters. Rows marked *deals only* are conditional on both sides reaching agreement, which is why deal rate is printed first and never omitted. Lower is better for the two distances, Gini, IR violations and malformed episodes; higher is better for the primary, worst-off capture and seat-0 capture. **Bold** marks an interval excluding zero. The final column is the interaction — how much larger the effect is under the story — and it resolves for legality only; the deal-conditional interactions rest on 16 four-way-complete pairs and are the weakest entries in the table.

	abstract			**datacenter**			**interaction**
Metric	all-LLM	rat@0	effect	all-LLM	rat@0	effect	dc − abs
Deal rate	0.717	0.733	+0.017	0.550	0.667	+0.117	+0.100
Normalized Nash welfare (primary)	0.182	0.267	**+0.086**	0.098	0.200	**+0.101**	+0.015
Distance to NBS *(deals only)*	0.816	0.768	−0.037	0.921	0.686	**−0.263**	−0.278
Distance to KS *(deals only)*	0.773	0.723	−0.074	0.889	0.651	**−0.280**	−0.245
Gini of capture *(deals only)*	0.338	0.308	−0.026	0.363	0.295	−0.078	−0.035
Worst-off capture *(deals only)*	0.038	0.106	+0.049	−0.039	0.087	**+0.153**	+0.187
IR-violation rate	0.300	0.183	−0.117	0.317	0.167	−0.150	−0.033
Malformed episode	0.033	0.050	+0.017	0.250	0.117	**−0.133**	**−0.150**
Seat-0 capture	0.323	0.524	**+0.201**	0.173	0.435	**+0.262**	+0.061

Why this cell is the program's cleanest LLM-versus-rational contrast. Every other comparison in this paper differs between the two populations in what they *do*: the rational agents solve a different problem, or use a different decision rule, so a gap can always be argued to be a gap in bargaining skill. Here nothing about the game changes at all. The score sheets, thresholds, feasible set, Pareto frontier, Nash solution and optimal policy are bit-identical across arms, and the only difference is the vocabulary the same numbers are printed in. Under that transformation the computable bargainers are *provably and observably* unmoved, and the LLM table loses welfare, moves off the frontier, and starts tabling illegal packages. **Part of the gap between LLM and rational play in this environment is not about solving the bargaining problem at all — it is sensitivity to the surface the problem is written on.**

9 Frontier-API cells: getting to yes versus deciding where yes lands

The setup. Every model in §5 is open-weight and between 4B and 32B parameters, so the obvious question — *does a frontier model close the gap to computable rational play?* — was unanswered by the main campaign. We ran a hosted-API campaign of `claude-sonnet-5`, `claude-haiku-4-5` and

(partially) `claude-opus-5` through the identical environment: the same six committed instances, the same `moves_chat` protocol, the same oracle stack, thinking off, five rollout seeds, and the same `annotate` plus report pipeline at the same regret threshold. Thinking off is the *matched* condition, not a concession: it is what every local launch used. This is observation only — no training, no fine-tuning, no prompt search beyond the two intervention arms described below.

The problem: the metric you lead with decides the answer. The unconditional primary scores a no-deal episode as 0, so it is dominated by deal rate; lead with it and frontier models comfortably exceed rational play. Distance-to-NBS and distance-to-KS are computed over *reached* deals only, so quoting them without deal rate beside them lets a model that rarely closes look fair by selection. Every table in this section therefore carries deal rate next to the distances, and Nash welfare appears in both its deal-conditional and unconditional forms. This is the same instrument-blindness problem that §5.4 solved by refusing to rank models on deal rate, applied to a population where the two orderings genuinely disagree.

The result. Under the matched condition, **frontier models are much better at getting to yes and worse at deciding where yes lands**. Both Claude models close 0.767 of episodes against the computable control's 0.567, and both distribute the resulting surplus worse on every fairness axis: further from the Nash bargaining solution (0.809 and 0.876 against 0.578), further from Kalai–Smorodinsky, lower Nash welfare conditional on a deal (50.7 and 35.3 against 58.6), more unequal (Gini 0.315 and 0.365 against 0.275), and they accept deals worse than walking away where the rational agents never do. That is the same two-signed shape as the open-weight ladder — distribution is the universal deficit, aggregate closure is not — reproduced in a population three or more orders of magnitude away in capability. Three further results follow, and one caveat dominates them: **sonnet's entire closing advantage over rational play comes from the chat channel** and is paid for in fairness (§9.2); **dropping one computable rational policy into a sonnet table extracts nothing for itself but makes the whole table fairer** (§9.3 — on three games; that finding reverses when the same swap is re-measured on 24 screened games, §11.5); **neither prompt engineering nor true rationality moves the focal seat's take at all**, so goal 4's headline ratio has no denominator here (§9.4); and **enabling thinking buys welfare and robustness to a narrative framing, but not fairness** — it leaves the distributional deficit intact while removing the framing's cost to closing (§9.5).

One-sentence version. Frontier models close far more deals than computable rational agents and split them noticeably worse, the whole closing advantage is the cheap-talk channel and disappears when it is silenced, a rational seat among them looks on these three games like a fairness anchor rather than a predator — a reading that §11.5 overturns on twenty-four — and the properly funded thinking-enabled run has since been done: it raises welfare and confers robustness to a narrative framing without closing the distributional gap.

9.1 Cells, and the fairness-led table

There is a haiku/sonnet capability ladder and the headline metrics cannot see it. On deal rate the two are identical (0.767) and on the unconditional primary they are indistinguishable (0.684 against 0.678, with haiku nominally *ahead*). On every fairness axis sonnet is cleanly better: dist-NBS 0.809 against 0.876, NSW 50.7 against 35.3, Gini 0.315 against 0.365, IR violations 0.033 against 0.067. A capability difference invisible to the closing-oriented metrics becomes legible the moment the question is how the surplus was split. Note also that the hosted family does not

Table 10: Frontier-API cells against the computable control and the largest local model. All cells play the same six committed instances through the same pipeline. **Deal** is the fraction of episodes reaching agreement and **primary** is unconditional normalized welfare (no deal scores 0), so both reward closing; the four fairness columns describe *how* a reached deal divided surplus and are computed over reached deals only, which is why deal rate is printed beside them rather than in a separate table. **dist-NBS** and **dist-KS** are distances from the realized deal to the Nash bargaining and Kalai–Smorodinsky solutions (lower is fairer), **NSW** is the geometric mean of party surpluses in points given a deal (higher is better), **Gini** is inequality of the realized surplus vector (lower is more equal), and **IR viol.** is the fraction of episodes in which some party accepted a deal worse than walking away (lower is better; the computable control is the only player that is never in violation). The last three rows are the frontier 2×2 of §9.5: adaptive thinking explicitly requested (with summarized reasoning persisted) at a 16,384-token per-turn floor that no turn ever reached, crossed with the `datacenter` narrative skin over the identical games; its fourth cell is the `claude-sonnet-5` row above. Every other row is thinking off, abstract framing, five seeds. Cells are single runs, not paired contrasts — the paired contrasts are in Table 12.

Cell	n	deal	primary	dist-NBS	dist-KS	NSW	Gini	IR viol.
`all_rational` (reference)	60	0.567	0.501	**0.578**	**0.654**	**58.6**	**0.275**	**0.000**
Qwen3-32B, local (§5.4)	60	0.550	0.436	0.869	0.824	29.0	0.412	0.150
`claude-haiku-4-5`	30	0.767	0.684	0.876	0.858	35.3	0.365	0.067
`claude-sonnet-5`	30	0.767	0.678	0.809	0.780	50.7	0.315	0.033
`claude-sonnet-5`, no cheap talk	30	0.533	0.484	0.723	0.709	58.2	0.289	0.000
`claude-sonnet-5`, scaffold seat 0	30	**0.833**	0.744	0.797	0.768	52.8	0.312	0.033
rational seat 0, 5 sonnet	30	0.700	0.632	0.733	0.710	56.2	0.290	**0.000**
`claude-sonnet-5`, **thinking ON**	**60**	0.883	0.808	0.777	0.741	56.4	0.288	**0.000**
`claude-sonnet-5`, thinking ON, `datacenter`	60	0.767	0.710	0.708	0.674	**60.1**	0.261	**0.000**
`claude-sonnet-5`, thinking off, `datacenter`	60	0.567	0.522	0.710	0.713	51.2	0.272	0.033

dominate the local ladder wholesale: sonnet clears Qwen3-32B on every fairness column, while haiku is indistinguishable from it on dist-NBS and dist-KS.

9.2 Cheap talk is a deal-closing device, paid for in fairness

Removing the public chat channel is the one intervention here with a large, resolved, and *two-signed* effect. Paired within instance and seed (30 pairs, six instance clusters), silencing sonnet costs 0.233 $[-0.333, -0.133]$ of deal rate and 0.194 $[-0.285, -0.097]$ of the unconditional primary — and improves every fairness axis at once: dist-NBS 0.809 $\rightarrow$ 0.723, dist-KS 0.780 $\rightarrow$ 0.709, NSW 50.7 $\rightarrow$ 58.2, Gini 0.315 $\rightarrow$ 0.289, IR violations 0.033 $\rightarrow$ 0.000.

Two consequences deserve stating plainly.

First, sonnet's entire deal-closing advantage over computable rational play is the chat channel. Silenced, it closes 0.533 against the control's 0.567 and is nearly as fair (NSW 58.2 against 58.6, Gini 0.289 against 0.275, IR violations 0.000 in both). The frontier model without talk *is* approximately a rational agent on this instrument; the frontier model with talk trades fairness for agreements. One caveat on the framing: the computable policies ignore the chat channel entirely, so this is a contrast between claude-uses-talk and rational-ignores-talk, not a like-for-like comparison of rhetoric.

Second, the sign is opposite to the open-weight population on the identical instances. On clean cells (§5.5) turning chat *on* costs Qwen3-8B 0.233 $[-0.450, -0.100]$ of deal rate and Qwen3-32B 0.083, is unresolved for Qwen3-4B and OLMo, and helps only Gemma's primary. For

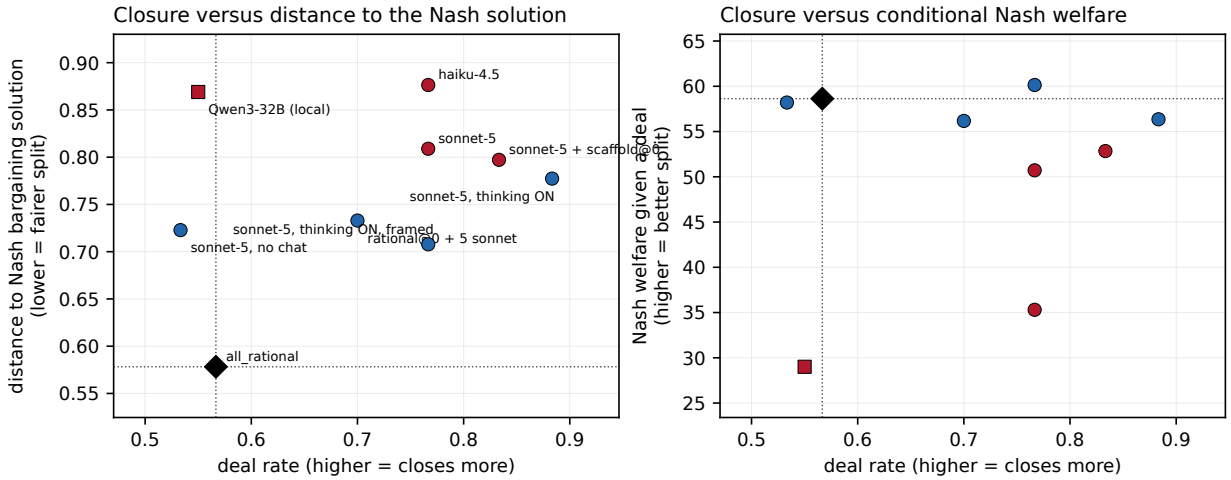

Figure 8: Closure on one axis, fairness on the other. Each marker is one cell from Table 10, plotted at its deal rate (x, right is more closing) against distance to the Nash bargaining solution (*left panel*, lower is a fairer split) and Nash welfare conditional on a deal (*right panel*, higher is better). The black diamond is the computable `all_rational` control and the dotted lines are its coordinates; the square is the largest local model. Blue markers never violate individual rationality, red markers do. The two panels plot the same cells at the same x, so the left panel's labels identify every point. **The frontier cells sit to the right of the control and above it on distance-to-NBS** — they close more and split worse — and the two cells that recover the control's fairness (*no cheap talk*, *rational@0*) do so by giving up closure or by seating a rational agent.

sonnet, turning chat *off* costs 0.233 of deal rate. Same environment, same instances, same protocol switch, opposite sign — a double dissociation between the two populations. The natural reading is that cheap talk is a coordination device whose value is capability-gated: below some threshold the extra channel is a distraction that consumes turns, and above it the channel is what lets five other parties be brought to the same package. That reading is a hypothesis; what the data establish is that *the sign of the cheap-talk effect is a property of the population, not of the protocol*, which retires any attempt to describe cheap talk as helpful or harmful in this environment full stop.

9.3 The rational seat is a fairness anchor, not a predator

Swapping one computable rational policy into seat 0 of an otherwise all-sonnet table does **not** let it extract. On the same normalized available-surplus scale the companion campaign uses (§14.1), the focal seat's capture changes by -0.017 $[-0.081, +0.031]$ at $n = 60$ — a precise null.[8] Against the open-weight models, the identical swap captured $+0.147$ to $+0.444$ on all five model slots with every interval excluding zero. **The extraction advantage of a computable rational policy is entirely absent against a frontier population**, measured on the same scale, in the same environment, with the same estimator.

What the swap does instead is improve the *table*, and after the cell was topped up to $n = 60$

[8]Two interval conventions are in circulation for this null and they differ by roughly 2×; this paper quotes the wider one throughout. Research note 0018 reports -0.0010 ± 0.0300, a paired standard-error half-width over the 30 pairs, which treats the five rollout seeds of an instance as independent observations. The interval quoted here is a cluster bootstrap that resamples whole *instance* clusters with seeds staying inside their cluster — the same inference every other interval in this paper uses — and is wider precisely because it does not assume seed independence. Both describe the same point estimate and both exclude any effect resembling the $+0.147$ to $+0.444$ measured against open-weight seats, so the verdict is unchanged either way.

per side (note 0026's addendum) two of those movements carry paired intervals that exclude zero: dist-NBS 0.776 → 0.729 (paired −0.036 [−0.069, −0.018]), dist-KS 0.752 → 0.703 (−0.042 [−0.113, −0.018]), with Gini (0.296 → 0.279) and the worst-off share (0.155 → 0.183) directionally consistent but unresolved, IR violations eliminated (0.033 → 0.000), for a deal-rate cost that does not resolve (−0.150 [−0.450, +0.050]). One rational participant makes five cooperative frontier models split the pie better without taking more of it — **on this thinking-disabled cell, and with the censoring caveat the next paragraph makes precise**. Table 11 states the same swap one measure per row, with the computable control printed beside it so the size of each move is readable against the distance it had to travel.

The effect does not survive thinking, and most of it was a deadline artifact. The cell above disables the models' reasoning. Re-running the identical swap with thinking explicitly enabled (§9.5's specified condition, $n = 60$, ten seeds on the same six instances) inverts its sign: pooled over all 60 pairs the anchor now *costs* −0.096 [−0.139, −0.029] of unconditional Nash welfare, with deal rate moving the same way without resolving (−0.167 [−0.300, +0.050]), while every fairness axis it was supposed to move fails to resolve. But the anchor also drags episodes to the protocol's 30-turn deadline far more often — the at-cap rate rises 0.150 → 0.617, a change of +0.467 [+0.200, +0.600] — and a capped episode is resolved by the forced final vote rather than by convergence, so the two cells are scoring differently-composed populations of deals. **Restricted to the 22 pairs in which both episodes finished early, every contrast collapses**: deal rate exactly 0.000, Nash welfare +0.003 [−0.085, +0.040], dist-NBS −0.055 [−0.234, +0.170]. Two conclusions follow, and the second is a caveat on this section rather than on the new cell. First, **at matched censoring one disciplined seat neither helps nor harms a thinking table**. Second, **the same confound applies to the thinking-off comparison reported above**: that swap also raises the at-cap rate (+0.333 [+0.150, +0.500] at $n = 60$), so its two distance movements — which *do* now carry paired intervals excluding zero — are statistically resolved yet still censor-confounded: doubling n fixed the interval, never the confound, and deal rate in these cells is very nearly a deterministic function of the deadline — every episode finishing before the cap closed a deal, 23/23 and 51/51 in the two thinking-on cells. The fairness-anchor reading should therefore be held as a resolved-but-confounded movement — real intervals over deadline-selected deal populations — pending a both-uncapped stratification on this pair, which does not yet exist. What is unaffected is the extraction null: under thinking the focal seat's capture changes by −0.044 [−0.090, +0.000], so the policy still takes nothing extra for itself. One contrast in the new cell is *not* censor-confounded and is worth stating on its own: the two mixed cells have near-identical cap rates (0.617 vs 0.650) and similar deal rates (0.717 vs 0.667), and across them enabling thinking moves the split *away* from both solution concepts — dist-NBS +0.043 [+0.002, +0.104], dist-KS +0.055 [+0.002, +0.178] at $n = 60$ per side; the Gini leg, resolved at $n = 30$, no longer is (+0.003 [−0.016, +0.032]). With a disciplined seat already at the table, reasoning does not add fairness; it subtracts a little on the distance measures. Research note 0026 carries the full 2×2. All intervals here are cluster bootstraps over the three committed games.

Read against §14.1 this is the more interesting half of the result. The companion's surviving finding was that a rational seat wins on *how surplus divides* rather than on whether agreement happens — and it won that by taking more for itself. Here the division improves too, but the gain accrues to the *other* seats. The same policy is a redistributor in one population and a beneficiary in the other, which suggests its advantage over open-weight seats was an advantage over exploitable counterparties rather than an intrinsically superior bargaining posture.

Table 11: The anchor effect, one row per measure. What replacing exactly one of six `claude-sonnet-5` seats with a computable Bayesian-rational policy does to the *table*, with the all-computable control as the destination the table is moving toward. **Gap closed** is the change as a fraction of the distance from the all-sonnet cell to that control, i.e. how much of the fairness deficit one disciplined seat removes. Lower is better for the two distances, for Gini and for IR violations; higher is better for NSW; deal rate is not a fairness measure and is printed last so the fairness rows are never read without their closure cost beside them. *The change column carries paired 95% intervals where they exist* (the two distance rows, resolved at $n = 60$ per side after note 0026's top-up; rows marked n.s. span zero), and the focal seat's own capture (-0.017 $[-0.081, +0.031]$) and deal rate (-0.150 $[-0.450, +0.050]$, unresolved), both quoted in the text. The IR row is a **two-episode swing** — 0.033 at $n = 60$ is two violating episodes going to zero — and should not be read as an eliminated failure mode. **This whole table is a three-game, thinking-off cell, and §11.5 re-measures the same swap on 24 screened games, where its sign reverses**; read the two together, not this one alone. For the interval-bearing version of these rows, and for the same swap measured on all five open-weight slots beside this one, see Figure 13: it pairs per episode and supersedes this table's change column wherever the two overlap. It also places this cell in context — sonnet shows the *smallest* fairness movement of the six slots. **Reading:** one disciplined seat moves the table between a third and all of the way toward computable-rational fairness depending on the measure, at no resolved closure cost, while capturing nothing extra for itself.

Measure	all sonnet	rational@0	change	gap closed	all_rational
Distance to Nash solution	0.776	0.729	-0.036 $[-0.069, -0.018]$	24%	0.578
Distance to Kalai–Smorodinsky	0.752	0.703	-0.042 $[-0.113, -0.018]$	50%	0.654
Worst-off seat's share	0.155	0.183	$+0.003$ n.s.	—	0.225
Gini of realized surplus	0.296	0.279	-0.009 n.s.	—	0.254
IR-violation rate	0.033	0.000	-0.033 (2 episodes)	100%	0.000
Deal rate *(closure, not fairness)*	0.817	0.667	-0.150 n.s.	—	0.567

Superseded in part: the same swap on 24 screened games reverses this. Everything above rests on the three committed games, and the frontier anchor has since been re-measured on a fresh, screened, 24-game bank at 96 episodes per cell under the identical condition — same thinking-off setting, same canonical scaffold, same abstract framing, same `moves_chat` arm, same `reverse_mixed` lineup with `bayes-rational` at seat 0. On 24 clusters the anchor **extracts** $+0.128$ $[+0.056, +0.209]$ of seat-0 capture from the Sonnet table and pushes it 0.104 $[+0.025, +0.194]$ further from the Nash solution at a $+0.028$ $[+0.007, +0.050]$ higher Gini, and unlike the thinking-on cell above, that reversal *survives* the matched-censoring restriction. **The extraction null and the fairness gain reported in this subsection are therefore scoped to the three-game bank they were measured on, and the powered estimate supersedes them** (§11.5). What does not change is the comparison that motivated the subsection's title: whatever the anchor's effect on a frontier table, it is an order of magnitude smaller than the $+0.147$ to $+0.444$ it extracts from open-weight seats, so the population contrast survives even though its frontier endpoint moved off zero.

9.4 Goal 4: a triple null, and a ratio we refuse to report

Goal 4 asks how much of a rational agent's advantage prompt engineering can recover. Three arms, paired on `(instance, seed)` with focal seat 0: **A**, all-sonnet canonical; **B**, all-sonnet with the existing solver-workflow scaffold (§E) on seat 0 only; **C**, a rational policy in seat 0.

Both focal-seat effects are nulls. **The intended headline — the fraction of the rational advantage that prompt engineering recovers — cannot be computed, because its de-**

Table 12: Paired frontier contrasts, all 30 pairs over the six committed-instance clusters, with 10,000-draw cluster-bootstrap intervals. **Focal capture** is seat 0's realized surplus as a fraction of that seat's own available surplus — the same scale as Table 20, so the last column is directly comparable to the companion campaign's +0.147 to +0.444 against open-weight models. Rows B and C are the goal-4 intervention ladder; row *no chat* is §9.2.

Contrast (both arms all-sonnet unless stated)	Δ deal rate	Δ primary	Δ focal capture
no cheap talk − canonical	**−0.233** [−0.333, −0.133]	**−0.194** [−0.285, −0.097]	**−0.135** [−0.215, −0.056]
B: scaffold on seat 0 − canonical	+0.067 [−0.133, +0.200]	+0.066 [−0.115, +0.190]	+0.040 [−0.055, +0.126]
C: rational seat 0 − canonical	−0.067 [−0.300, +0.133]	−0.046 [−0.254, +0.139]	−0.000 [−0.069, +0.074]

nominator is itself indistinguishable from zero. The arithmetic ratio exists and is meaningless, and reporting it would be a number with no defensible interpretation, so it is not reported. The honest goal-4 statement on this population is that *the premise does not hold: there is no measurable per-seat rational advantage for prompting to recover.*

What is reportable is the table-level structure, where the two orderings are legible and opposite: on deal rate B > A > C, and on every fairness axis C > B > A. Prompting nudges a table toward closing; a rational seat nudges it toward fair distribution; neither changes what the focal seat itself captures. Power note: 30 pairs over six instance clusters distinguishes "prompting recovers most of the advantage" from "prompting does nothing", not finer gradations — and the tightness of the focal-capture intervals (about ±0.07 on a quantity whose sample mean is 0.41) is what makes these real nulls rather than underpowered ones.

9.5 Thinking and narrative framing: a 2×2

The condition every cell above holds fixed is thinking *off*, because that is what every local launch used. Two questions follow: what does enabling reasoning change, and does it change how much a *story* told over the game matters? Both are answered by one 2×2 — {thinking on, off} × {abstract, `datacenter`}, 60 episodes per cell, the same six committed instances at the same ten seeds, paired within game. The framing relabels the negotiation as a data-centre siting dispute while preserving instance identity, so all four cells are the same underlying games under a different surface.

Specifying the thinking condition changed it. An earlier pass ran "thinking on" as `--model-thinking on`, which sent *no thinking parameter at all*. On current Claude models the default already *is* adaptive thinking, so that cell genuinely thought — but shallowly (405–440 output tokens per turn), and with nothing in the request recording the condition. Requesting it explicitly (`thinking={type: adaptive, display: summarized}` with `effort=high`) is **not** behaviourally equivalent: the same games run deeper. Every thinking-on number here is from the explicit condition, and the manifest records the parameters actually sent rather than the flag that was typed.

Thinking buys welfare, not fairness. Paired within game in the abstract cell, enabling thinking raises the worst-off party's share by +0.022 [+0.012, +0.034], conditional Nash welfare by +4.08 [+2.05, +5.55] and unconditional normalized Nash welfare by +0.067 [+0.043, +0.087]. It does *not* reliably move the distance-to-solution or equality measures: distance to the Nash solution −0.022 [−0.071, +0.006], Gini −0.005 [−0.012, +0.004], IR violations −0.033 [−0.050, +0.000]. An earlier shallower-thinking pass reported all of those as clean improvements; at the specified depth they span zero. Deeper thinking did not buy more fairness than shallow thinking — on distance to the

Nash solution the explicit cell is *worse* (0.777) than the shallow one (0.694), against a control at 0.578. The frontier fairness deficit of §9.1 therefore stands: thinking does not close it.

Thinking does buy robustness to the story, and that is the interaction. Dressing the identical game as a siting dispute costs the thinking-off arm -0.250 $[-0.450, -0.150]$ of deal rate and -0.120 $[-0.238, -0.042]$ of unconditional Nash welfare. It costs the thinking-on arm -0.117 $[-0.300, +0.050]$ and -0.035 $[-0.127, +0.085]$ — neither separable from zero. The difference between those two framing effects, the diff-of-diffs, is itself resolved: $+0.133$ $[+0.050, +0.200]$ on closing and $+0.085$ $[+0.017, +0.128]$ on welfare. **The narrative is a real distractor, and thinking is what sees through it.**

Against the open-weight ladder, the susceptibility moved rather than vanished. §8 measured Qwen3-8B under the same skin losing welfare (-0.083), *malforming* far more episodes ($+0.217$) and drifting off the Pareto frontier ($+0.238$). Sonnet malforms **zero** episodes in all 120 framed episodes under either thinking condition, and drifts *toward* the frontier (-0.016 $[-0.039, +0.000]$ thinking off). But its thinking-off welfare loss, -0.120, is as large as the open-weight model's. So a frontier model is not simply framing-robust: it never loses the ability to *play* the game under a story, only the ability to *close and split* it — and thinking repairs that.

One interaction runs the other way and is a selection artifact. Deal-*conditional* welfare shows an interaction of -12.3 $[-19.8, -9.2]$. This is what a conditional metric does when the arms close at different rates: the thinking-on arm closes 0.767 of framed episodes against 0.567, so its extra deals are the marginal ones the other arm walked away from, and marginal deals are worse. The unconditional measure, which scores a no-deal as zero and admits no such selection, moves the opposite way. The unconditional number is the finding; the conditional one is a denominator.

The chain of thought is partially recoverable, and is saved. Current Claude models never return the raw reasoning stream. They *do* return a readable summary, but only when it is asked for: the `display` field defaults to `omitted`, which returns a thinking block whose text is the empty string beside an encrypted signature — the setting under which reasoning looks unrecoverable. With `display=summarized`, 93% of thinking turns carry substantive saved reasoning (worked arithmetic, threshold checks, why a counter-offer was made). Independently of the text, every turn stores the provider's billed thinking-token count, which is 0 across all 2,654 thinking-off turns and averages 282–331 on thinking-on — the evidence that reasoning occurred, surviving the sealing of its content.

Power, and the standing rule. Three instance clusters. The contrasts that clear their intervals are deal rate and unconditional welfare, computed over all 60 pairs; deal-conditional fairness contrasts run over 31–46 pairs (both sides must close) and mostly do not. This design resolves directions and large effects, not small ones. The standing rule against partial cells was mechanised rather than trusted here: the framed thinking-off cell hit its budget cap at 48/60 episodes and was completed by topping up the exact missing (instance, seed) cells, and the analysis now refuses to run on a short cell at all. That rule was adopted after three interim readings reversed on completion — a sonnet-versus-haiku primary gap of 0.771 against 0.610 that vanished to 0.678 against 0.684 at $n = 30$; a rational-seat deal-rate drop of -0.20 at $n = 15$ that shrank to -0.067 spanning zero; and an audit screening on `parse_ok` that pronounced clean a run which had advanced zero rounds in eighteen turns — and the $n = 6$ thinking cell this subsection replaces is the fourth.

9.6 Data quality: two failure modes that are the model's, not the harness's

Engine fabrication (§3.6) is **0.0% on every hosted-API cell** — hosted seats do not route through the batched pool, which is the observation that localized the bug — against 40–100% across the contaminated local 4B–8B tier. But hosted models have their own way of producing a turn with no usable action, and it must never be pooled with fabrication: there the model was never called, here it was called and genuinely produced nothing usable. Model-side empty-action rates: sonnet 0/620 turns in the canonical arm and 0/628 with chat off, the mixed cell 0/727, haiku 1/479 (0.2%), the scaffolded arm 1/587 (0.17% — the scaffold lengthens seat 0's output enough to occasionally reach the cap). `claude-opus-5` is the outlier at 15/145 (10.3%), and its two causes are both worth recording because they are *model-specific rather than family-wide.*

Adaptive thinking crowds out the visible action, on opus and not on sonnet. At a 2,048-token per-turn cap, hidden reasoning consumed the entire budget on 12 of 18 turns (67%) of the first opus pilot episode, which advanced **zero rounds**. Raising the floor to 4,096 only deferred the problem: mean observed output per turn rose from 1,095 to 2,635 tokens and five of 24 turns still truncated, exactly as the library documentation warns — adaptive thinking can outgrow any fixed floor. **This does not bite sonnet.** In the thinking-on cell sonnet averaged 405 output tokens per turn (maximum 3,689) against an 8,192 floor with *zero* `max_tokens` truncations. Sonnet's adaptive thinking is economical enough that thinking can simply be left on; opus's is not, at any floor this budget could afford.

A safety classifier false-positives on negotiation arithmetic. Four of the 145 opus turns returned `stop_reason: "refusal"` with no usable action. They are not what the name suggests — each is ordinary multi-party surplus arithmetic truncated mid-sentence, e.g. *"P3 gives me 211 and Flynn already accepts it… Casey under P3: 54+32=86 > 59.8. Devo…"* and *"Ember: 41+49+0+0+0 = 90 — above bar 87.4, same as under P4, so P5 costs Ember nothing r…"*. No local cell and neither sonnet nor haiku produced a single refusal in any arm. Benign surplus arithmetic is tripping a classifier at roughly 3% of turns on one model. The documented mitigation is a server-side fallback request parameter that the library does not currently implement; until it does, this is a measurement hazard for any evaluation whose task involves dense arithmetic about what parties get.

Two artifacts of a highly convergent policy. In the sonnet cell, three completed episodes on the same instance at seeds 1, 2 and 3 returned byte-identical outcome vectors. Hashing the transcripts before concluding anything showed they are *distinct* runs — three different hashes, turn counts 30/24/24 — so the model reaches the same near-optimal deal by materially different paths. This is robust convergence, not a seeding bug, and its practical consequence is that seed-level error bars on distance-to-NBS are **degenerate for that instance**: extra seeds buy path variation for mechanism analysis and nothing for outcome precision. Relatedly, one committed instance (`game0_full`) returns identical surplus vectors from opus, sonnet and haiku alike: its optimum is one anything competent finds, so it cannot discriminate between models and should be excluded from calibration comparisons meant to say anything about capability. The discriminating signal is in the private-information half of the bank.

9.7 What was not run, and the budget that decided it

Total spend on the campaign as reported at the time of the frontier revision was **$128.85** against a $150 cap, itemized per cell in each run's `usage.json` spend ledger; the largest single cell was the framed thinking-off arm at $24.03.[9] That arm is also the campaign's one budget-gate incident: its cap stopped it at 48 of 60 episodes, and it was completed by topping up the exact missing (instance, seed) cells rather than reported short. Cost control is a first-class run parameter here (`run.py --budget-usd`, `--est-usd-per-episode`), because an API campaign that runs out of money mid-cell produces exactly the partial cells the standing rule of §9.5 forbids reporting.

Two arms were deliberately not run. There is **no GPT-class arm**: no OpenAI credential exists on this host, so the campaign is Claude-only and nothing here speaks to cross-vendor differences. And there is **no full `claude-opus-5` cell**: the project owner directed a switch to cheaper models mid-campaign, and a proper opus rung (≈$23, and the only cell carrying a 10.3% model-side contamination caveat) would have spent against that directive. The four already-paid opus episodes are retained as a labelled datapoint — deal rate 1.00 and primary 0.852 over four episodes, all on a single instance, three of them at the identical 0.860 — and are **not comparable** to the $n = 30$ cells and not to be quoted as a top rung.

10 Outcome-aligned RL: the fairness-GRPO pilot

§14 names the prescription the P4 post-mortem arrived at: stop training on per-turn oracle preference, which anti-tracks outcomes in this game family, and train on the *outcome* instead. This section reports that experiment. It is the last untried intervention — scale does not close the distributional gap (§5.2), a solver scaffold actively hurts (§5.5), test-time thinking buys welfare and not fairness (§9.5), seating a computable agent works but is an intervention on the *table* rather than on the model, and the part of it that works is plain refusal discipline rather than rationality (§12.7, §11.2), and divergence-DPO failed (§7) — and it is preregistered, gated, and a clean negative.

Full detail, including the deviation ledger and the reference values measured before any trained checkpoint existed, is in `research-notes/0023-fairness-grpo-pilot.md`; the preregistration is `proposals/2026-08-01-fairness-grpo.md`, and everything in its reward-design, training and evaluation sections was fixed before any GPU time was spent.

10.1 The preregistered design

The reward. Per party, normalized surplus is $z_i = (u_i(d) - \tau_i)/c_i$ — the same affine-invariant quantity the campaign analysis uses, so no seat can gain by rescaling its own sheet. Per-party utility is $g(z) = \log z$ for $z \geq \varepsilon$ and its tangent line $\log \varepsilon + (z - \varepsilon)/\varepsilon$ below ($\varepsilon = 0.01$), and the table reward is $R_{\text{table}} = \text{mean}_i\, g(z_i)$. Above ε that is a monotone transform of normalized Nash

[9] **The project-wide API ledger has since grown well past that figure and the $150 cap was superseded by a $2,000 project ceiling.** Summing every `usage.json` under the scratch run root gives **$388.15** total: $157.61 across the 28 `apibehav_*` frontier cells (the $128.85 above plus the thinking-on mixed cell at $20.25 and its thinking-off top-up at $7.98, note 0026), $4.22 for the realistic-framing demo pair, and **$226.32** for the policy-decomposition arc of §11 — of which $202.49 is the seven `claude-sonnet-5` cells on the screened 24-game bank (five policies at $7.43–$31.65 each, plus the within-bank all-LLM and Bayes-anchor references at $27.67 and $27.60) and the remainder the six superseded old-bank and calibration cells. The arc's Qwen3-8B half, all 15 local cells and the entire calibrated-maximizer replay, cost $0 in API spend and ≈0.05 GPU-hours respectively, because the trivial policies emit zero tokens and the replay is CPU-only. **The arc's one substantial compute cost is the calibrated-maximizer campaign**: **12.22 GPU-hours** of rented B200 time (RunPod pod `he793kh18vkbty`) for 864 episodes across six cells, $0 in API spend, shared with the RL lane under a file-based handoff protocol and a hard stop watchdog so an abandoned agent session could not leak pod-hours.

welfare, so training optimizes the judgment metric rather than a proxy for it, and the concavity does the fairness work by itself: moving surplus from a well-served party to a badly-served one always raises the mean, with no bolted-on equality term to tune or exploit. The linear branch below ε is the repair that makes the objective trainable at all — normalized Nash welfare is *identically zero* wherever a party sits below threshold, i.e. flat exactly where the observed pathology lives. Walking away is priced at $g(0) \approx -5.605$ to every seat: worse than any deal that clears everyone, better than one that pushes a party well below.

The reward is **text-blind by construction** — it reads `outcome.normalized_realized_surplus` and never a token the policy generated. That is non-negotiable given that 95.5% of the P4 DPO signal turned out to be rationale prose (§7.3). Seat i trains on the mixture $R_i(\lambda) = (1-\lambda)\, g(z_i) + \lambda R_{\text{table}}$, sweeping how much self-interest is traded for table welfare; note that $\text{mean}_i\, R_i(\lambda) = R_{\text{table}}$ for every λ, so λ changes *who is paid what* and not the total. Credit is densified with potential-based shaping on $\Phi(s) = R_{\text{table}}(\text{standing offer at } s)$, which provably leaves the optimal policy unchanged — the principled successor to the per-turn oracle regret this project's own anti-tracking result bans from a reward.

Training. Qwen3-8B with a LoRA adapter (r=32, α=64, 87.3M trainable parameters), thinking off to match every baseline in the program, occupying **all six seats** simultaneously in symmetric self-play, rolled out through the batched co-stepping engine with the fail-loud fabrication budget armed — if the engine ever has to invent a placeholder turn the run aborts rather than feeding fiction into the gradient. One optimizer step per rollout wave, so the policy that generated the data *is* the policy being updated and the GRPO importance ratio is exactly 1. Two arms, $\lambda \in \{0, 1\}$; 24 steps each, 12.5 hours; zero fabricated turns at any step. Training curves are Figure 9.

Amendment 1, and why it is the reason this section has intervals. The preregistered evaluation used the 6 committed instances plus the 3-instance cross-game bank — 3–9 clusters under the instance-cluster bootstrap every interval in this paper uses, which is limitation 4 all over again. Before any training step, the primary endpoint was moved to a fresh **24-game held-out bank** (`instances_grpo_eval_v1/`, generator seeds 41000+, 48 instance clusters, 960 episodes per cell), with the 6-instance bank retained as a *comparability leg* because it is what every existing baseline in the program was measured on. The bank is guarded twice — candidates are fingerprinted on both their score sheets and their solved solution points and rejected on collision with any known bank, since a fresh seed guarantees a different random *stream* and not a different *game* — and screened for discriminativeness (NNW spread across the strict-IR set ≥ 0.08, because a game where every feasible agreement scores the same cannot register a policy difference in the primary metric). Reported honestly: on this draw the new screen was non-binding, both rejections failing the pre-existing Pareto-slack band. As §10.3 shows, the amendment is what makes the result a result: on the preregistered 6-instance plan the deal-rate collapse that decides this pilot would have spanned zero.

10.2 Step 0: the reward is sound, and it is not a deal-rate proxy

Risk #4 of the preregistration required scoring existing clean campaigns under the reward *before* any GPU spend and stopping if the ordering came out wrong ("fix the reward, not the world"). All four preregistered orderings hold on the point estimate: rational control > anchored table (+0.163), anchored table > all-LLM (+0.770), rational control > all-LLM (+0.934), NBS advice > placebo (+1.011, the only one whose paired interval excludes zero — the three P2 contrasts rest on six clusters and are wide by construction).

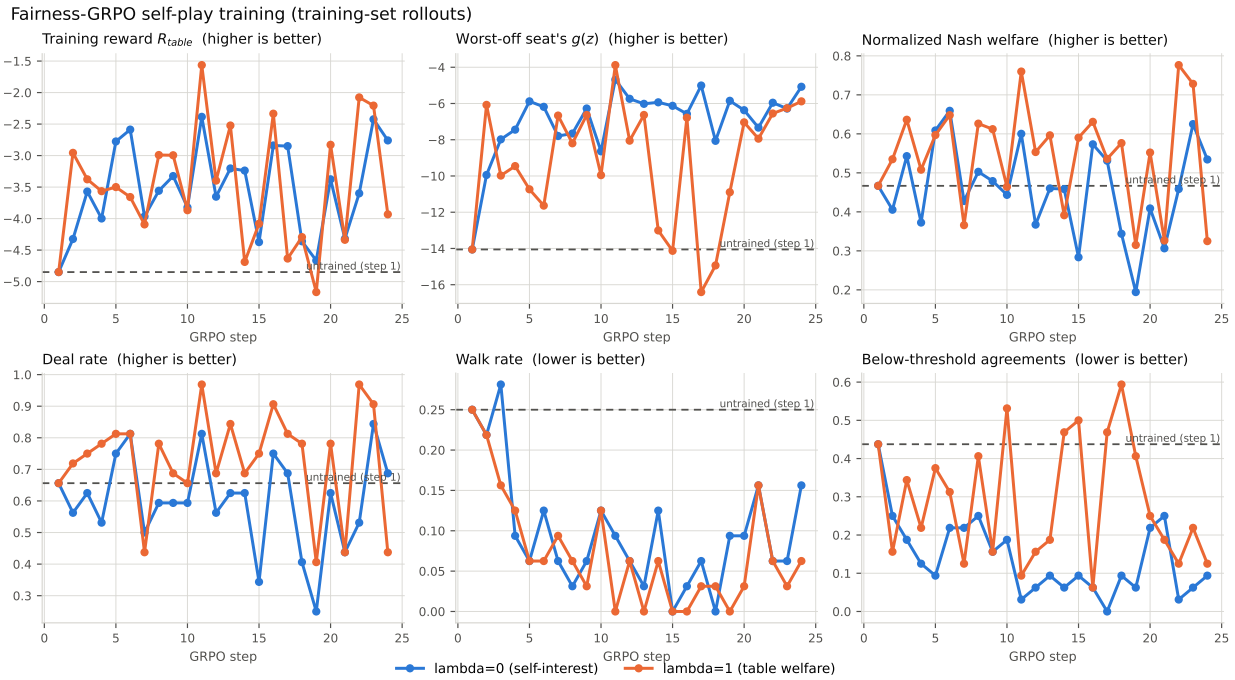

Figure 9: Training-bank rollout curves, both arms, 24 GRPO steps. These answer "is optimization working", not "did it generalize", and they flatter the method by construction — R_{table} is what the objective maximizes, so a rising reward curve is close to tautological. The informative divergence is that R_{table} rises by $\approx$0.6 in both arms while normalized Nash welfare is flat to slightly down, because R_{table} is dominated by the below-threshold region (see §10.7). Log-probability drift from the frozen base grows from $\approx$0 to +0.14 (λ=0) and +0.27 (λ=1) nats per completion token, which excludes the "nothing moved" null.

The informative line is that **the reward does not track deal rate**. The all-rational control closes *fewer* deals than the all-LLM table (0.650 against 0.817) and still scores 0.93 higher, because the LLM table averages 0.408 below-threshold agreements per episode and a worst-off g of -10.2 against the control's 0.000 and -3.3. That is the linear branch doing exactly its job, and it is the ordering a deal-rate-shaped reward would have gotten backwards. It also names the pilot's most likely failure mode in advance: because a below-threshold deal (-10.2) scores far below walking away (-5.605), the reward genuinely prefers no agreement to a bad one, which puts real pressure on the walk-rate guard.

10.3 The verdict: discipline bought, welfare spent

The preregistered success criterion fails at every checkpoint, in both arms, on both banks. Table 13 is the primary endpoint. Data hygiene first: the evaluation fleet is 5,856 episodes, 5,856 completed, **zero fabricated turns** in every cell — baseline, trained, guards and presets alike — so nothing here is contaminated in the way §3.6 had to correct for.

Three things replicate on both banks and are the substance of the result.

What training bought is individual-rationality discipline. Below-threshold agreements — a party signing a deal worse than its own walk-away value, the pathology this project has documented at every scale (§5.4) — fall by 0.087 and 0.079 on the held-out bank with intervals excluding zero, and by 0.125 $[-0.233, -0.033]$ for λ=0 on the comparability bank. That is real, it is the one

Table 13: Primary endpoint: 24-game held-out bank, 48 instance clusters, 960 episodes per cell, trained minus untrained (`p2_Qwen3-8B_all_llm_b2`), paired, 95% instance-cluster bootstrap. "Gate" counts favourable secondaries out of four; the preregistered criterion additionally requires the primary interval to exclude zero *in the favourable direction* and all four guards to pass, and no cell comes close.

cell	ΔNNW (primary)	Δdeal rate	Δbelow-thr.	ΔNNW among IR	gate
λ=0 step 24	-0.185 $[-0.225, -0.145]$	-0.252	-0.087 $[-0.127, -0.050]$	$+0.014$ $[-0.012, +0.040]$	FAIL (1/4)
λ=1 step 24	-0.109 $[-0.145, -0.070]$	-0.149	-0.079 $[-0.124, -0.030]$	-0.040 $[-0.067, -0.015]$	FAIL (1/4)
λ=0 step 10	-0.142 $[-0.177, -0.108]$	-0.204	-0.129 $[-0.162, -0.096]$	-0.004 $[-0.023, +0.014]$	FAIL (2/4)
λ=1 step 10	-0.016 $[-0.049, +0.020]$	-0.013	$+0.067$ $[+0.022, +0.112]$	-0.001 $[-0.026, +0.023]$	FAIL (0/4)

favourable endpoint anywhere in this evaluation, and it is not what the pilot set out to buy.

What it cost is welfare, through a held-out deal-rate collapse. Normalized Nash welfare moves -0.185 and -0.109 against a 0.635 baseline, both intervals excluding zero in the *wrong* direction, and deal rate falls 0.25 and 0.15 against a 0.817 baseline — down to roughly 0.57 and 0.67. The policy avoided bad agreements substantially by not agreeing. What makes this a diagnosis rather than an observation is that it **contradicts the training-set picture**: on the 24 games it trained on, walk rate roughly *halved* while deal rate held flat (Figure 9). The policy closes deals on games it has seen and fails to on games it has not. That is overfitting to a 24-game bank, and separating it from noise is precisely what Amendment 1's 48 clusters bought.

More training is monotonically worse. NNW moves $-0.142 \to -0.185$ (λ=0) and $-0.016 \to -0.109$ (λ=1) between steps 10 and 24. The over-training collapse the checkpoint ladder was built to catch is already visible within 24 steps — which is also the reason no "it would have worked with more steps" reading is available for free.

10.4 The λ split: composition for one arm, worse deals for the other

Every distributional endpoint is reported twice, unconditionally and among individually-rational deals only, and the pair is what separates the arms. The unconditional view is what the gate uses and is immune to selection, but it mixes a change in *which* deals get struck with a change in *how fairly* a struck deal divides; the among-IR view isolates the second but conditions on an outcome the training itself moves, so a divergence between the two is the signature of a composition shift rather than of fairer bargaining. (Same censoring logic as §9.3, different selector; the among-IR endpoints are carried as non-gating diagnostics exactly as preregistered.)

For λ=0, all four among-IR endpoints span zero on both banks: its welfare loss is **pure composition** — fewer deals, not worse ones. For λ=1, all four are unfavourable with intervals excluding zero on both banks: NNW -0.040 $[-0.067, -0.015]$, distance to the Nash solution $+0.037$ $[+0.004, +0.072]$, Gini $+0.023$ $[+0.010, +0.036]$, worst-off share -0.017 $[-0.031, -0.003]$, with the same four larger on the comparability bank (-0.220, $+0.162$, $+0.105$, -0.135). **Training on pure table welfare made the deals it did strike measurably less fair.** That is a directional negative rather than a null and it is the most interesting single row in the pilot: the arm that was paid *only* for the table's welfare degraded the table's distribution.

The other half of the λ story is that the frontier was close to degenerate. λ=0 was the preregistered *manipulation check* — pure self-interest was supposed to reproduce the pathologies — and instead it improved on essentially every training-rollout measure, nearly as much as λ=1. Symmetric self-play predicts this in hindsight: every seat is the same policy, and the seat-averaged reward is identical for

every λ by construction. The arms separate only on the among-IR endpoints — which is precisely what made the fairness result directional rather than null: λ was not a dead knob, it is the thing that isolated *how a struck deal gets divided* from everything the seat-averaged reward controls. That is a finding about the design rather than a null, and it means any future λ sweep needs **heterogeneous opponents**.

10.5 Ultimatum: training eroded the social prior toward the extractor

With no extra training, the final checkpoints were seated in two canonical bargaining games the policy had never seen in any form.

Table 14: Generalization leg, zero extra training. Descriptive only: the presets are *deterministic*, so all three generated "instances" of each are the same game and every cell carries one game cluster — a cluster bootstrap over it would be degenerate and no interval is quoted. What is reported is the absolute split with its across-seed spread.

game (proposer's max share)	untrained	λ=0 step 24	λ=1 step 24	reference points
ultimatum	0.600	0.967	0.967	equal split 0.500; subgame-perfect ≈1.0
per-seed range	0.50–0.667	0.833–1.0	0.833–1.0	acceptance 100% in all cells
divide-the-dollar	0.794	0.844	0.894	equal split $1/n = 0.333$
per-seed range	0.611–0.917	0.778–0.944	0.778–1.0	

This is the sharpest number in the pilot and it points the wrong way. The untrained model proposes an accepted 60/40 ultimatum split — far from the subgame-perfect prediction that a self-interested proposer keeps essentially the whole pie, and close to the human-modal fair split. Both trained arms propose 96.7/3.3 and are *accepted every time*, with every trained seed above every untrained seed in both arms. Divide-the-dollar moves the same direction more weakly (0.794 → 0.844/0.894) with overlapping seed ranges, and should be read as suggestive only; the ultimatum effect is large and unanimous, so its direction is not in doubt even without an interval.

The mechanism reading is the same one the rest of the section supports. What training installed is discipline about one's own threshold plus more aggressive proposing — and in the ultimatum game those two together *are* the subgame-perfect extractor, because a responder who accepts anything above zero makes "propose almost everything for yourself" both individually rational and optimal. A fairness objective on a six-party scorable game did not transfer as fairness; it transferred as the equilibrium logic that fairness norms exist to override. Note also the headroom asymmetry that makes this striking: on the ultimatum the untrained model started close to fair, with very little room for fairness training to *show up* and a great deal of room to go the other way, and it took the room it had.

10.6 Guards

Guard (b) is worth stating positively: seating the trained policy at a table of five computable rational agents does not cost it anything, so whatever training did, it did not produce a policy that strong counterparties can exploit. Guard (d) is worth stating positively for a different reason — the preregistration named a specific fear, a "first-round NBS-proposal-or-walk collapse", and it did not happen. First-round closes *fell* (0.800 baseline → 0.550 and 0.517), proposals per seat rose (1.32 → 1.46/1.65), and turns per episode rose (20.5 → 25.1/25.0). **The trained policy negotiates more and agrees less**, which is a coherent bargaining posture rather than a collapse. Note in passing that the feared mode is close to the *untrained* model's default — baseline Qwen3-8B

Table 15: The four preregistered guards at the final checkpoint of each arm.

guard	λ=0 step 24	λ=1 step 24	verdict
(a) walk-rate: deal rate must not fall by more than 0.10	-0.252	-0.149	**FAIL** both, and at step 10; only λ=1 step 10 passes
(b) rational-table: own capture against five computable seats	-0.001 $[-0.036, +0.036]$	$+0.029$ $[+0.005, +0.056]$	**PASS** both — fairness training did not make the policy exploitable
(c) framing probe (datacenter skin)	-0.072 $[-0.166, +0.021]$	-0.069 $[-0.157, +0.055]$	**FAIL** both on the -0.10 non-inferiority floor
(d) transcript audit (20 stratified episodes per cell plus whole-cell tells)	see text	see text	**PASS** — no new degenerate strategy

already closes 80% of its deals in round 1, against the all-rational control's 41.7% — so an absolute threshold on this guard would have fired on the control arm and it has to be read as a delta.

10.7 The reward diagnosis, and what the next attempt has to change

The λ=1 result is a specific, mechanistic warning rather than a mystery, and the diagnosis is visible in the reward's own arithmetic. R_{table} is dominated by the below-threshold region: the linear branch makes a single party at $z = -0.2$ contribute -25.6, against a range of a few nats across the entire space of individually-rational divisions. Almost all available reward therefore lives in *not shorting anyone*, and almost none in *dividing well*. **The objective is nominally Nash welfare and behaviourally an IR-violation penalty**, which predicts the result it got — discipline, purchased partly by walking, with distribution untouched or degraded. This also explains the divergence in Figure 9: eliminating below-threshold agreements moves R_{table} enormously while barely touching normalized Nash welfare, which is identically zero for any episode with a below-threshold party and so registers changes only among already-valid deals.

The diagnosis is testable, and re-running with a **clipped** g — bounding the violation branch, or optimizing an among-IR fairness objective directly — and checking whether the among-IR endpoints move is the highest-value follow-up this arc produces. Behind it: heterogeneous-opponent training so λ means something, a longer horizon before any positive claim (24 steps is 12% of the preregistered budget), and fixing the deal-rate collapse first, since every unconditional endpoint inherits its composition.

The symmetry with the scale result is what to carry forward. §5.2 found that going from 4B to 32B bought *behavioural discipline* — fewer malformed and below-threshold actions — and left the distributional gap flat. Outcome-aligned RL, a completely different mechanism operating on a completely different lever, bought **the same discipline and the same flat distribution**, and charged a welfare cost that scale did not. Two unrelated interventions moving the same axis and missing the same one make the distributional gap look less like a capability the model lacks and more like a disposition that neither reaches — which is consistent with a companion result not otherwise reported in this paper (`research-notes/0008-nbs-decision-support-results.md`): simply *handing* the model a fair candidate deal at inference time moved welfare +0.154, a preregistered PASS. The model can act fairly when it is told what fair is. It did not learn to want to.

The pilot in one sentence. Training an LLM negotiator on a Nash-welfare reward taught it to stop signing deals that were bad *for itself* — and to close a quarter fewer deals, bargain no more fairly, and behave *more* selfishly in unseen bargaining games.

11 Decomposing the anchor: which channel, which population, and a reversal under power

The setup. The single most repeated positive result in this project is that seating *one* computable bargainer at an otherwise all-LLM table makes the table's outcome fairer (§14.1, §8.5, §9.3, Figure 13). The computable seat is always the same object, `BayesianRationalPolicy`: it maintains a belief model over the other parties, best-responds against that model when it proposes, and accepts by optimal stopping. Those are three mechanisms welded together, so "rationality makes tables fairer" is a *label*, not an explanation — and a label is not something anyone can reuse. This section is the arc that took the label apart. Its hub, with the hypotheses fixed before any cell ran, is `self_benefit/README.md`; the notes of record are `research-notes/0022-trivial-policy-baselines.md` (the policy ladder, both populations), `self_benefit/acceptance_curves.md` (the measured opponent model) and `research-notes/0024-calibrated-maximizer.md` (the best-responder that plans against it).

The problem. Two questions, and a measurement problem underneath both. First, *which channel* of the Bayesian stack does the work: the **veto discipline** (never sign below your own walk-away value), the **proposal rule** (table what you model the others as willing to accept), or the **opponent model** that makes that proposal generous rather than selfish? Second, in the other direction: could a cruder, greedier policy *extract* more than the Bayesian agent does — and would that make the table worse? The measurement problem is that every previous answer in this paper rests on the three-game committed bank, i.e. three bootstrap clusters, which is limitation 4 (§13) applied to the project's flagship positive.

What we did. Five trivial seat-0 policies, each keeping exactly one channel and discarding the rest, plus a fresh screened 24-game bank built and frozen before any cell ran, plus within-bank re-runs of every reference so nothing is compared across banks. On top of that, two measurements that attack the opponent model directly: acceptance curves fitted to *every* clean stored episode in the project (585,265 response events over 65 campaigns), and a best-responder that plans against those fitted curves instead of the assumed rational step function. All five policies are deterministic, parameter-free, read only their own score sheet, and emit typed actions at **zero tokens** — a declaration is a fixed string, not generated text — so the local half of the arc is nearly free and the API half is the only thing that costs money.

The result. The anchor's effect splits cleanly in two, and the two halves come from different channels. A policy whose entire content is "never sign anything below your reservation value" — `passive-gate`, which never speaks and never tables a deal — reproduces the Bayesian anchor's *whole* fairness contribution (six fairness measures, every paired difference centred within 0.03 of zero) while capturing 0.235 less of seat 0's available surplus. **Discipline buys the fairness; the Bayesian machinery buys the money**, and the half that reads as a moral result is the half two lines of code deliver. In the other direction nothing we tried out-extracts the composed agent: the maximally selfish proposer captures *less* (-0.090 against the anchor), an ultimatum hold-out closes only a third of its games and ends up behind (-0.161), the best declared demand level (the knee, at 65% of own maximum) still trails, and — from the opposite direction — giving the agent a

truthful opponent model makes it **concede rather than extract** on five out of five fitted opponents. That last one was then run closed-loop, and it is the arc's one clean *positive*: across 864 fresh episodes the truthfully-calibrated maximizer captures 0.146 less, halves its advantage over its own tablemates, and moves *every* fairness endpoint in the fair direction with intervals excluding zero — at an unchanged deal rate, and by a larger margin once censoring is matched. Two population results follow. Against `claude-sonnet-5` an announced ultimatum, held byte-identically to a silent one, roughly *doubles* closure (+0.146, resolved), while on Qwen3-8B it is a clean null — comprehension, not capitulation. And the powered bank overturns a published number of our own: on 24 clusters the Bayesian anchor **extracts** +0.128 from a Sonnet table and pushes it further from the Nash solution, where note 0018, on three clusters, reported no extraction and fairness-neutral-to-positive anchoring (§11.5).

One-sentence version. Pulling the "rational anchor" apart shows that its fairness effect is pure veto discipline and its money is the Bayesian inference, that no trivial greed beats it at its own objective while telling it the *truth* about its opponents makes it markedly less selfish and its table measurably fairer on every axis at once, and that at proper statistical power the frontier version of the effect reverses sign — the anchor redistributes toward itself rather than away.

11.1 The powered bank, and why every number here is on it

Every contrast in this section is paired on `(instance_id, seed)` and every interval is a cluster bootstrap resampling whole *games*, so interval width is governed by the number of independent games and not by the number of episodes. The committed six-instance bank is really three base games in full- and private-information variants, and it is the binding constraint on everything the earlier sections could resolve. So this arc runs on `instances_trivial_v1/`: **24 games × 4 seeds = 96 episodes per cell**, 24 clusters, a $\sqrt{24/3} \approx 2.8\times$ tightening of every interval, built and frozen by `build_trivial_bank.py` before any cell was launched.

Candidates were admitted only through four preregistered screens, two structural and two for *discriminativeness*: a non-degenerate feasible set (≥ 8 individually-rational deals); an acceptable set that has not collapsed onto the Pareto frontier (dominated-acceptable fraction in $[0.45, 0.75]$); the Nash bargaining solution must differ from the Kalai–Smorodinsky point; and no single obvious optimum. The last two matter because of defects visible in our own committed bank — NBS and KS *coincide in four of its six instances*, so two of this paper's fairness columns are one column wearing two hats there, and the `game0_full` instance has an optimum anything competent finds (§9.6). Admission was **24 of 137 candidates (17.5%)**, with NBS $\neq$ KS in 24/24, and every rejection came from the two discriminativeness screens (70 and 85 rejections) rather than from degeneracy (0 and 0): the generator never produced a degenerate game, it produced *undiscriminating* ones.

The bank is a selection and is reported as one. It is the discriminative subset, so results generalize to games where the solution concepts disagree and no package is obviously best — the population the experiment is about — not to everything the generator makes. The screens are structural rather than behavioural on purpose: a behavioural screen would need a solo-rational LLM pass per candidate and would make the bank a function of one model's competence. Because the existing references live on the old bank, cross-bank comparison would confound treatment with game distribution, so all-LLM, Bayes-anchor and all-computable references were *re-run within the bank*; 22 cells in total, all at `gen_failures` = 0.000.

11.2 H1: a two-line policy delivers the fairness; the Bayesian stack delivers the money

Table 16 is the eight-way panel on Qwen3-8B. The result that carries the section is the paired contrast between `passive-gate` and the full Bayesian anchor on the same games and seeds: the never-speaking gate is **indistinguishable on every fairness measure** — distance to the Nash solution $+0.018$ $[-0.099, +0.165]$, distance to Kalai–Smorodinsky $+0.002$ $[-0.067, +0.065]$, Gini -0.005 $[-0.034, +0.027]$, worst-off capture $+0.021$ $[-0.022, +0.059]$, IR violations $+0.031$ $[-0.073, +0.135]$, normalized Nash welfare -0.001 $[-0.068, +0.069]$; six measures, every one centred within 0.03 of zero — while being nowhere near it on money: seat-0 capture $\mathbf{-0.235}$ $[-0.338, -0.134]$ and within-table capture-z -0.607 $[-0.879, -0.347]$, both resolved. Measured against the all-LLM baseline rather than against the anchor, the gate lifts the worst-off seat's capture $+0.065$ $[+0.022, +0.114]$ and cuts IR violations -0.167 $[-0.292, -0.042]$, against the anchor's $+0.044$ $[-0.003, +0.095]$ and -0.198 $[-0.312, -0.083]$ — on worst-off capture the gate is the version that resolves — while taking -0.034 $[-0.127, +0.052]$ of capture against the anchor's $+0.201$ $[+0.072, +0.324]$.

On the frontier population the gate does not merely match the Bayesian agent, it beats it. Against the Sonnet anchor, `passive-gate` captures -0.168 $[-0.261, -0.084]$ less while being resolved *better* on fairness: Gini -0.037 $[-0.064, -0.008]$ and worst-off capture $+0.031$ $[+0.001, +0.056]$, with distance to the Nash solution -0.077 $[-0.203, +0.031]$ leaning the same way; under matched censoring the advantage grows (Gini -0.062, worst-off $+0.086$, both resolved). The H1 split replicates in both populations and is the most robust result in the arc.

11.3 H2 falsified from both directions: greed captures less, and so does truth

The hypothesis was that the Bayesian agent proposes generously because it models opponents as refusers — i.e. that its opponent model makes it *timid* — so a proposer that never concedes should extract more. It was tested twice, by removing the opponent model and by making it correct, and it failed both times.

Removing it: `greedy-anchor` captures less. Maximally selfish proposals plus the same acceptance discipline yield seat-0 capture -0.090 $[-0.175, -0.004]$ against the Bayesian anchor — resolved in the direction opposite to the prediction. **Refusing everything is worse still.** `greedy-holdout` does dominate the table *conditional on closing* (capture-z $+1.14$ $[+0.69, +1.76]$ against the anchor's $+0.55$), but it closes only 35% of its games (-0.385 $[-0.531, -0.229]$ against the all-LLM baseline) and unconditionally lands *behind* the anchor on its own objective (-0.161 $[-0.296, -0.034]$), while costing the table -0.134 $[-0.182, -0.087]$ of normalized Nash welfare. Refusing everything wins the games you were going to win anyway and burns the rest.

Making it correct: the agent concedes. The measured object is a set of *acceptance curves* — $P(\text{accept} \mid z, \text{rounds left})$ per responder, where z is the responder's own surplus under the offer divided by the most that deal space could ever give it, so the assumed rational rule sits at exactly $z \geq 0$. They were fitted by harvesting every response event from all 65 clean stored campaigns (7,119 episodes, 168,371 turns, no cell with a single fabricated turn; 585,265 events), one logistic per responder, validated leave-one-*game*-out. Two findings stand on their own (Table 17). *The deal-closing bias is real, large and capability-ordered*: among below-threshold offers a seat actually voted on, `gemma-3-4b` takes 96%, Qwen3-4B 89%, Qwen3-8B 81%, Ministral-8B 69%, OLMo-3-7B 60% — while Qwen3-32B takes 13% and `claude-sonnet-5` (thinking on *or* off), `claude-haiku-4-5` and the computable policy take **0.00%**. Accepting deals worse than walking away is a small-model behaviour, not an LLM behaviour, and it is the behavioural mechanism behind the fairness gap this paper reports throughout. *And calibration buys exactly as much as the rational assumption is wrong*

Table 16: Five trivial seat-0 policies against the Bayesian anchor and the two controls, Qwen3-8B. Every cell is the `reverse_mixed` table — seat 0 is the named policy, seats 1–5 are the same LLM — on the frozen 24-game screened bank, 96 episodes per cell, thinking off. Policies: `passive-gate` never proposes and accepts any standing offer worth at least its walk-away value (veto discipline alone); `greedy-anchor` always proposes its own best deal and accepts anything individually rational (selfish proposals *plus* discipline); `greedy-holdout` proposes its own best deal and accepts *only* that deal (pure extraction, no discipline); `holdout-decl.` is byte-identical in behaviour to `greedy-holdout` plus one public up-front commitment message, so the pair isolates cheap talk; `demand-90` accepts anything worth $\geq 0.90\times$ its own maximum and declares that number. Rows marked *(deals)* are averaged over closed deals only, which is why deal rate is the first row: the three hold-out columns close about a third of their games, so their fairness rows describe a policy-selected third and are not comparable to the others. **cap-z** is seat 0's capture standardized within its own episode's six-seat spread, so it measures extraction *from the same table* independently of that table's level. **at-cap** is the fraction of episodes that consumed all 30 scheduled turns rather than closing early — a first-class column because it is the differential censor that decides which games each column's deal-conditional rows are computed over. Intervals in the text are 10,000-draw cluster bootstraps over the 24 games. Source: note 0022; tables in `_trivial_baselines_analysis/trivial_qwen*.{json,csv,md}`.

metric	all-LLM	passive -gate	greedy -anchor	hold -out	hold -decl.	demand -90	bayes -anchor	all -rat.
deal rate	0.740	0.656	0.698	0.354	0.365	0.385	0.750	0.604
at-cap rate	0.469	**0.427**	0.500	0.760	0.760	0.844	0.594	1.000
norm. Nash welfare (uncond.)	0.126	0.162	0.155	0.029	0.043	0.038	0.163	**0.211**
dist-NBS *(deals)*	0.944	0.913	0.920	0.949	1.044	1.058	0.868	**0.788**
dist-KS *(deals)*	0.921	0.854	0.865	0.979	1.001	0.957	0.876	**0.681**
Gini of capture *(deals)*	0.377	0.361	0.366	0.402	0.417	0.407	0.365	**0.303**
worst-off capture *(deals)*	0.007	0.053	0.030	−0.022	−0.018	−0.027	0.039	**0.125**
IR-violation rate	0.333	0.167	0.208	0.271	0.292	0.177	0.135	**0.000**
seat-0 capture (uncond.)	0.315	0.281	0.425	0.354	0.365	0.385	**0.515**	0.350
seat-0 cap-z *(deals)*	−0.068	−0.116	0.353	1.292	1.308	1.309	0.552	0.208

and no more: the fitted curve removes Brier error where the step model is wrong (gemma +0.143 [+0.085, +0.212], Ministral +0.080) and *loses* where it is right (Qwen3-32B −0.022, the computable policy −0.010). The pipeline's positive control is that last row: run on computable seats the harvest recovers the step function exactly, scoring a Brier of **0.0000** against their votes, which requires the sign convention on z, the offer-id binding, the self-accept exclusion and the step reference all to be right simultaneously.

Feeding those curves back into the policy is the strong form of H2, and it is where the hypothesis dies cleanly. `CalibratedRationalPolicy` is the Bayesian policy with **exactly one method overridden** — the opponent acceptance model — so the comparison is single-factor. It was measured twice, in the cheap order: first as a deterministic *decision-function* replay over already-stored episodes, which costs nothing and asks both policies for their move on the identical state at every computable-seat turn, and then as a *closed-loop campaign* of 864 fresh episodes.

The free measurement, and the prediction it forced. The equivalence control is exact — the `step` rung, which carries step curves and is therefore Bayes by construction, matches Bayes *and the recorded transcript* on **18,685 of 18,685 policy turns**. The effect is not: on 25.3% of turns the calibrated agent plays differently, and the mean change in its own utility is **−15.6** points, with **63.7% of changed proposals worth *less* to itself** against 16.8% worth more. It replicates on all

Table 17: What LLM negotiators actually do when an offer is worse than walking away. Every row is one *responder* (model × thinking mode) across all 65 clean stored campaigns. **below-IR accept** is the fraction of below-threshold offers accepted *restricted to offers the seat formally voted on* — the quantity the Bayesian policy's opponent model says is 0.00 by assumption. b_z is the fitted logistic's slope in own normalized surplus (higher = sharper response to what the deal is worth to it) and b_r the coefficient on rounds remaining (negative = accepts more as the deadline nears, the capitulation direction). **Brier gain** is the per-event squared-forecast-error the fitted curve removes relative to the rational step model on the vote-only subset, out-of-fold under leave-one-game-out, with a game-cluster bootstrap interval; *positive means calibration helps, negative means the step function was already the better model of that responder.* The last row is the computable policy itself and is the pipeline's positive control. Source: `self_benefit/acceptance_curves.md`, artifact `self_benefit/acceptance_curves.json`.

responder (thinking off unless noted)	**below-IR accept**	b_z	b_r	**Brier gain over step**
`gemma-3-4b-it`	0.962	+0.68	−0.83	+0.1425 [+0.085, +0.212]
Qwen3-4B	0.886	+1.40	−0.17	+0.0208 [+0.011, +0.033]
Qwen3-8B	0.806	+1.61	+0.06	+0.0212 [+0.013, +0.030]
`Ministral-8B-2410`	0.692	+0.86	−0.28	+0.0796 [+0.059, +0.102]
`Olmo-3-7B`	0.600	+1.06	−0.06	+0.0384 [+0.030, +0.047]
Qwen3-32B	0.133	+2.08	+0.12	−0.0220 [−0.027, −0.017]
`claude-sonnet-5`	0.000 (n=77)	+3.39	+0.43	+0.0367 [+0.023, +0.052]
`claude-sonnet-5`, think ON	0.000 (n=82)	+2.96	+0.37	−0.0014 [−0.008, +0.008]
`claude-haiku-4-5`	0.000 (n=4)	+2.31	+1.21	−0.0018 [−0.061, +0.046]
`policy:bayes-rational` *(control)*	0.000 (n=125)	+3.09	−0.61	−0.0095 (step Brier **0.0000**)

five fitted opponents (−10.5 to −15.6, 55–65% less selfish) and on the campaign's own bank (−14.9, **0.0%** more selfish). Because this pointed the opposite way to the hypothesis the campaign had been designed to confirm, the preregistration was **revised against its own motivating hypothesis before a single cell ran** — from "capture up, fairness down" to "capture flat-to-down, table no less fair" — with both predictions left on the record for the campaign to adjudicate (Appendix G).

The campaign, and the verdict: H2-strong is falsified in both directions. Three rungs were run against each other at matched seeds on 36 screened games — `step` (the fitted curve swapped back for the step function, so it *is* the old agent and isolates the plumbing), `calibrated` (the honest panel-grain curves) and `calibrated-vote` (honest curves plus the knowledge that the forced final vote is where LLM seats accept almost anything) — with the maximizer at seat 0 and Qwen3-8B, thinking off, in the other five seats. **864 episodes, 12.22 GPU-hours** on a rented B200, all six cells `rc=0` at `gen_failures` 0.000 and a malformed-episode rate ≤1.6%. Told the truth about its opponents, the maximizer **took less and gave more** (Table 18, Figure 10). Its own capture fell 0.714 → 0.569 (**−0.146** [−0.217, −0.075]) and its within-table advantage over its five tablemates *halved*, 0.392 [0.332, 0.451] → 0.182 [0.114, 0.249]. Every fairness measure improved and every one of their intervals excludes zero: distance to the Nash bargaining solution **−0.362** [−0.497, −0.233], a 39% reduction; Gini of the final split −0.080 [−0.111, −0.050]; the worst-off seat's take +0.090 [+0.050, +0.126]; unconditional Nash welfare +0.098 [+0.026, +0.167]. **Deal rate did not move** (−0.052 [−0.115, +0.010]) and neither did the at-cap rate (−0.062 [−0.156, +0.026]), so this is not a table that got "fairer" by failing to close — the selection trap that every deal-conditional number in this paper is exposed to. And unlike the frontier anchor's benefit in §9.3, which dissolved when censoring was matched, this effect **strengthens** under the restriction: dist-NBS −0.420

Table 18: The calibrated-maximizer ladder: giving a self-interested optimizer a truthful opponent model makes it concede and its table fairer. Rows are outcome metrics; the `step` column is the control rung, which carries the original step-function opponent model and is therefore the pre-existing Bayesian agent. Each Δ is a *paired* difference against that control at matched seeds, with a 10,000-draw cluster bootstrap over the **24 full-information game clusters** (the 24-game screened bank's 12 full-info games plus the 12-game all-full-info sibling); **bold** marks an interval excluding zero. "Capture" is seat 0's share of the surplus available to it; "z within table" normalizes it against the other five seats in the same episode. Distances to the Nash and Kalai–Smorodinsky solutions are *lower is fairer*, Gini lower is fairer, worst-off capture and Nash welfare higher is better. Deal-conditional rows are computed on the pair counts in the source table (143 and 152 of 192 paired episodes). Source: `_calmax_analysis/calmax_full.{json,csv,md}`; note 0024.

	`step`	`calib.`	Δ vs step [95% CI]	`c.-vote`	Δ vs step [95% CI]
deal rate	0.870	0.818	−0.052 [−0.115, +0.010]	0.880	+0.010 [−0.042, +0.062]
ran full protocol (at cap)	0.391	0.328	−0.062 [−0.156, +0.026]	0.385	−0.005 [−0.094, +0.083]
distance to the NBS	0.937	0.590	**−0.362** [−0.497, −0.233]	0.538	**−0.374** [−0.526, −0.234]
distance to KS	0.872	0.727	**−0.143** [−0.226, −0.066]	0.722	**−0.135** [−0.217, −0.061]
Gini of capture	0.369	0.292	**−0.080** [−0.111, −0.050]	0.287	**−0.079** [−0.108, −0.051]
worst-off capture	0.049	0.134	**+0.090** [+0.050, +0.126]	0.145	**+0.093** [+0.049, +0.137]
Nash welfare (uncond.)	0.200	0.298	**+0.098** [+0.026, +0.167]	0.346	**+0.145** [+0.081, +0.210]
IR-violation rate	0.120	0.104	−0.016 [−0.089, +0.073]	0.099	−0.021 [−0.094, +0.057]
seat-0 capture	0.714	0.569	**−0.146** [−0.217, −0.075]	0.607	**−0.107** [−0.196, −0.020]
seat-0 capture (z in table)	0.977	0.537	**−0.467** [−0.752, −0.201]	0.512	**−0.463** [−0.764, −0.168]
within-table advantage	0.392	0.182	—	0.172	—
malformed episode rate	0.016	0.005	−0.010 [−0.031, +0.010]	0.016	0.000 [−0.016, +0.016]

[−0.567, −0.277], Gini −0.107, worst-off +0.110, capture −0.150, Nash welfare +0.195 (Figure 11).

The private-information placebo, and why it is a methods contribution rather than a footnote. The calibrated policy is *provably inert* without opponent utilities: computing the opponent's normalized surplus z requires their score sheet, so under private information the policy falls through to the inherited belief path, and the decision-function replay confirms this on real episodes — **0 of 215 private-info policy turns changed**, against 46.5% of full-info turns. The bank's 12 private-information games are therefore not a wasted half; they are a *matched placebo* in which the treatment is applied, labelled and analysed exactly as in the treated arm while being incapable of acting. That is a stronger control than a null arm chosen by an experimenter, because it holds the pipeline, the prompts, the seeds and the analysis code fixed and varies only whether the intervention can bite. It behaves: on the `calibrated` rung not one of thirteen metrics resolves (dist-NBS −0.028 [−0.120, +0.070], Gini +0.011, worst-off −0.033, capture −0.030, Nash welfare −0.028), against full-information effects 8–37× larger. **One row of the ≈26 placebo comparisons does resolve** — the `calibrated-vote` rung's deal rate, +0.135 [+0.062, +0.208] — and we report it as a false positive rather than explaining it away, because no policy difference exists there that could cause it. That is the point of running a provably-inert arm: it *measures* the design's false-positive rate instead of assuming it, and one resolved row in twenty-six at 12 clusters is almost exactly what $\alpha = 0.05$ predicts. **The standing caution that follows is the useful output**: at this bank's cluster count a single resolved row is weak evidence, and the headline above is trustworthy not because its intervals exclude zero but because it rests on 24 clusters, effects an order of magnitude larger, and seven metrics moving together in a direction a mechanism predicted in advance.

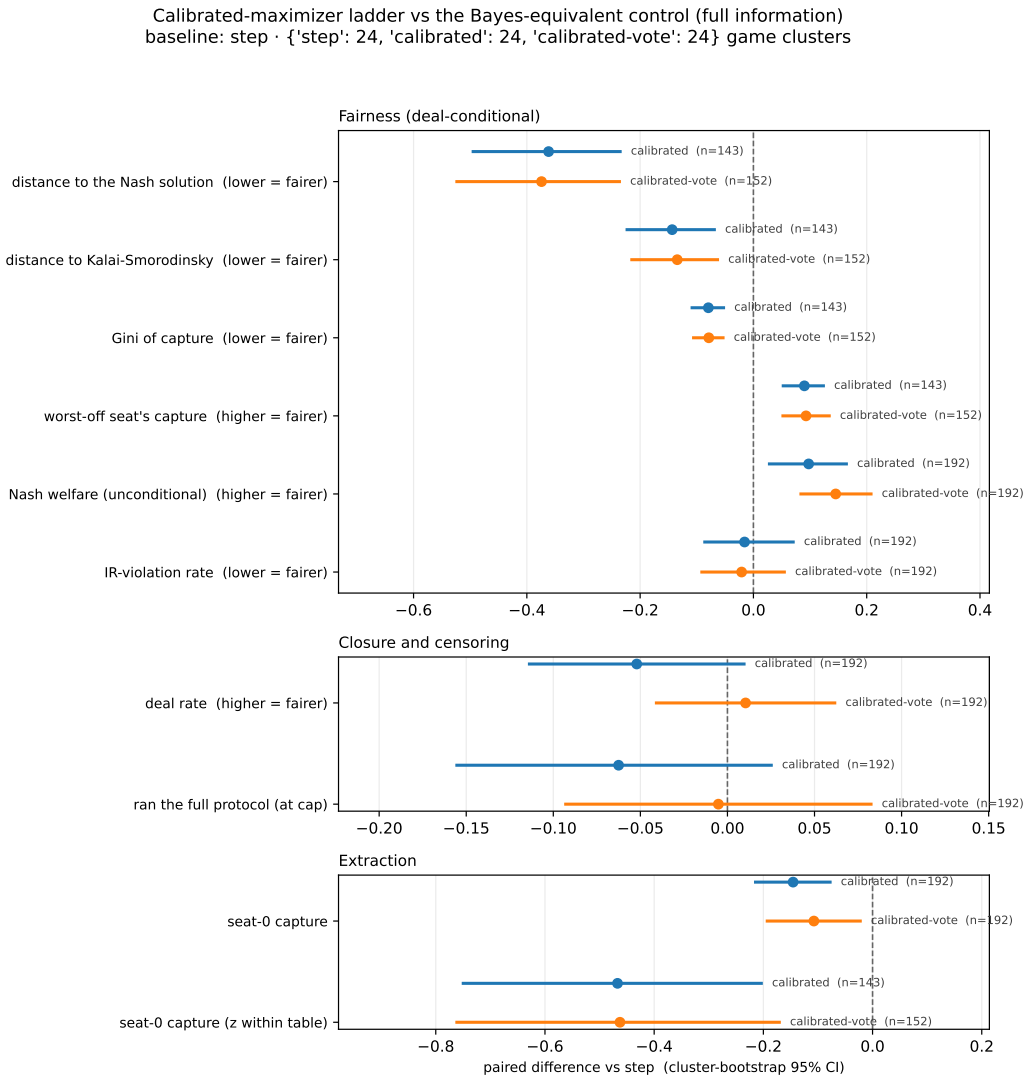

Figure 10: The calibrated maximizer's paired effects, fairness first. Each row is one metric; each marker is the paired difference of a treatment rung (`calibrated`, `calibrated-vote`) against the Bayes-equivalent `step` control, with a 10,000-draw cluster-bootstrap interval over the 24 full-information game clusters. The dashed vertical line is no effect. Panels are ordered deliberately: *fairness* first because that is the endpoint the hypothesis was about, *closure and censoring* second as the mandatory companion (a fairness gain bought by not closing deals is not a fairness gain, and this figure exists so that cannot be hidden), *extraction* last. Row labels carry the pair count each row is computed on, since the deal-conditional rows use fewer pairs than the unconditional ones. Generated by `fig_calibrated_ladder.py` from `_calmax_analysis/calmax_full.json`.

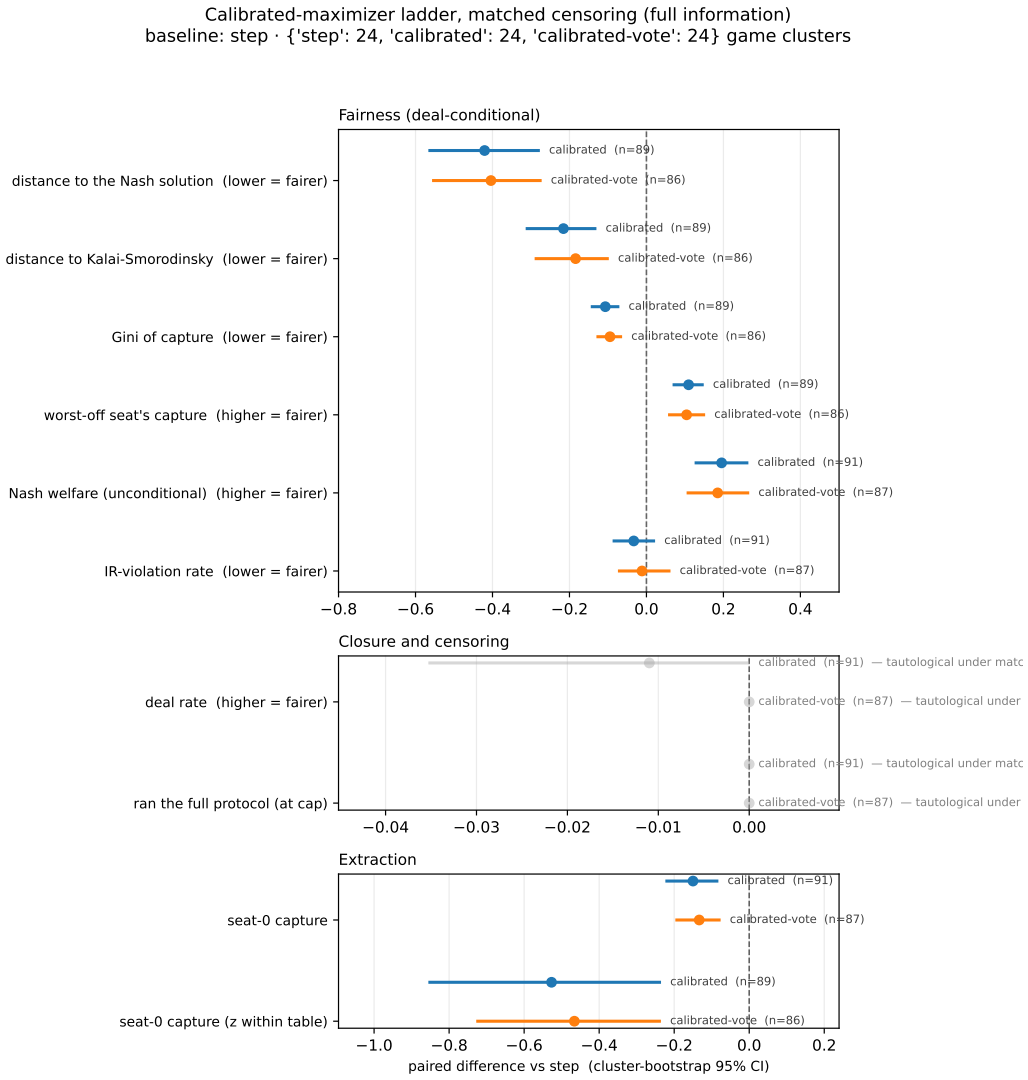

Figure 11: The same ladder under matched censoring — the effect grows rather than dissolving. Identical construction to Figure 10, but every contrast is restricted to game/seed pairs where *neither* side ran the full protocol, applied per contrast rather than globally. This is the restriction that dissolved the frontier anchor's apparent benefit in §9.3; here every fairness interval moves further from zero. **Read the two closure rows as tautologies, not as evidence**: the restriction defines both sides to have finished early, and in this scenario an episode finishes early only by closing, so deal rate and at-cap come out at exactly 0.000 $[0.000, 0.000]$ by construction. That was written down before the table was seen, and it is flagged in the figure. Source: `_calmax_analysis/calmax_full_matched.json`.

The mechanism is not the obvious one, and the control that falsified the obvious one is worth naming. The first reading was "the honest model says opponents refuse, so buy their acceptance"; `gemma-3-4b` kills it, being a genuinely agreeable model (fitted acceptance 0.82 at its own reservation) against which the calibrated agent concedes just as hard (-15.3, 65% less selfish). The real mechanism is structural: **the step model assigns probability** 1.0 **to every deal at or above an opponent's reservation and no fitted curve ever does** — even gemma's tops out at 0.90 — so over the region a best-response search actually explores, every honest curve is a strict discount. **Unanimity compounds it across five opponents**: gemma's 0.90 becomes $0.90^5 \approx 0.59$ and Qwen3-8B's 0.40 becomes ≈ 0.01, and the one lever an expectimax has for raising a joint acceptance probability is to raise everybody's z. So the step model's error was never conservatism — it was **over-confidence**, and correcting over-confidence makes a rational agent more generous, not more cynical. Two results converge here from opposite directions: §11.2 removed the opponent model entirely and the agent captured less; this made it correct and the agent conceded.

The `calibrated-vote` rung is the mechanism's own confirmation, and it is the reason we believe the story rather than merely fitting it. If sub-certainty is what drives the concession, then handing back certainty exactly where the fitted curve most understates it — the forced final vote, where a seat facing a single standing offer takes it 0.99 of the time — should restore *appetite* without restoring *selfishness*, because ordinary rounds still price at panel grain. That is precisely its shape in the campaign: it is the only rung whose deal rate sits *above* the control ($+0.010$) and it posts the best Nash welfare of the three ($+0.145$ $[+0.081, +0.210]$), so knowing the endgame is soft lets it hold out without killing deals — and yet its capture is still resolved *down* (-0.107 $[-0.196, -0.020]$) and its within-table advantage is the lowest of the three (0.172). Better endgame information makes it a better negotiator for everyone at the table including itself, and still not a more selfish one.

One sub-finding, because it is a trap for anyone reusing the curves. The fitted curve is a *panel-grain* quantity — "next time this seat moves, does it take this offer" — and folds in crowding, since a seat facing $\approx$4.7 standing offers can formally take at most one. Restricted to offers a seat actually votes on, acceptance is a different world: 0.992 for Qwen3-8B against 0.229 at panel grain, a $\approx 8\times$ understatement at $z = 0$. A third rung priced the endgame with the vote-conditional curve instead, and on the turns where the grain differs it **flips the sign hard** — forced-final proposals go from a mean -9.5 to $+29.5$ own-utility — yet moves the whole-ladder number only $-15.6 \to -14.8$, because forced-final turns are $\approx$3% of the total. The grain confound is real, worth carrying, and *not* the explanation for the concede-not-extract result, which is driven by ordinary-round play where the panel curve is the correct object. The campaign confirms that reading at the outcome level: pricing the endgame correctly changes what the vote rung does with *closure* and leaves its concession intact.

What the campaign does not settle. Two scope limits travel with the numbers above and are the reason the next-steps list (§14) names a specific successor rather than declaring the question closed. *The magnitude is a seat-0 magnitude*: the six-seat sweep was dropped to hand the shared pod to the RL lane, so the campaign runs the maximizer in seat 0 only. The *direction* is far better supported than that — it replicates across all six seats and all five fitted opponents at the decision-function level, where four of six seats concede outright — but nothing here pins the size of the outcome effect at another seat. *And there is one opponent model at the table.* Every campaign episode faces five Qwen3-8B seats with thinking off, and the acceptance curves themselves predict the effect should *attenuate* against stronger counterparties: the step assumption the calibration corrects is already close to right for Qwen3-32B (below-IR acceptance 0.133, and a *negative* Brier gain for the fitted curve) and exactly right for `claude-sonnet-5` (0.000), so there is little over-confidence left

to correct. That is a falsifiable prediction this arc makes and has not tested.

11.4 H3 and H4: a declared commitment works on the frontier model and not on the small one

`greedy-holdout-declared` and `greedy-holdout` are **behaviourally identical by construction** — a unit test asserts identical actions on every state in the branch grid — so the paired difference between them is the announcement and nothing else. This is the cleanest cheap-talk manipulation in the project, cleaner than §9.2's channel switch, because the acting policy is bit-for-bit the same on both sides.

On Qwen3-8B it is a tight null: deal rate $+0.010$ $[-0.115, +0.125]$, capture $+0.010$ $[-0.115, +0.125]$, welfare $+0.013$ $[-0.005, +0.035]$. The null is readable rather than suspicious for two structural reasons: the declaration provably reached the other seats (asserted in the public event log by test and visible in the transcripts), and distance-to-NBS is *exactly* $+0.000$ $[+0.000, +0.000]$ across the 22 games both cells closed — necessarily, since either policy can only ever close on seat 0's own maximum, so the deal rate is the *only* channel the declaration could act through, and it is the number that did not move.

On `claude-sonnet-5` the same sentence roughly **doubles closure**: deal rate $+0.146$ $[+0.052, +0.250]$, $0.135 \to 0.281$, resolved; unconditional capture $+0.146$, numerically identical because this policy only ever closes on its own maximum where its normalized capture is exactly 1.0; and at-cap -0.177 $[-0.271, -0.083]$, i.e. the declaration also ends episodes sooner, which is what updating on it should look like. **The direction is opposite to the capitulation prior** this repository's own sycophancy work would have predicted, which expected the *smaller* model to fold to a stated immovable position. The better reading is that this is not capitulation but *comprehension*: acting on a declared constraint requires reading it, believing it, and propagating it into your own offer, and that is a capability. Sonnet is simultaneously **more resistant** to the hold-out overall (0.135 closure against Qwen3-8B's 0.354) and **more responsive** to its declaration — resistance and responsiveness are different axes.

Relatedly, and retracted from our own earlier teaser: on the committed three-cluster bank `greedy-holdout` closed **0 of 30** games against Sonnet, which read as "the frontier model never caves". At 96 episodes on 24 clusters it closes 0.135 — not zero. The capability-gating *direction* holds (a 2.6× gap against Qwen3-8B on the identical policy) but the zero was three game clusters talking.

H4, the commitment knee. A declared demand *with a margin* was supposed to extract near-maximum while keeping deals alive. The preregistered 0.90 rung could not answer it, for a reason that is geometric rather than behavioural and that we only measured after seeing the result: **in this game family** 0.90 **of own-maximum is not a margin, it is essentially the maximum.** Deals clearing seat 0's 0.9× bar are 4.9% of the 243-deal space, only 2.5 of them on average are individually rational for all six seats, and *five of the 24 games have none at all.* A policy demanding something that barely exists is a hold-out by construction. So the ladder was extended to 0.50/0.65/0.80, chosen against that measured geometry. Deal rate falls monotonically as the demand rises ($0.698 \to 0.656 \to 0.583 \to 0.479 \to 0.385 \to 0.354$) and conditional dominance rises monotonically (capture-z $0.35 \to 1.31$); their product peaks in the middle, at `demand-65` with unconditional capture **0.488**. **So there is a knee, it sits near** 0.65**, and reaching it still leaves you short of the Bayesian anchor** (0.515) *while making the table worse off* (welfare 0.099 against 0.163, worst-off 0.003 against 0.039). No trivial demand rule, tuned or not, out-extracts the composed rational agent on this bank.

Table 19: The frontier anchor on 24 clusters: all-LLM minus Bayes-anchor, `claude-sonnet-5`. A *negative* sign means the all-LLM table is better on that measure, i.e. the anchor made things worse. 96 episodes per cell, thinking off, paired on `(instance, seed)`, 10,000-draw cluster bootstraps over the 24 games. The **matched-censoring** column restricts to pairs in which *neither* episode ran the full 30-turn protocol, which is the check that dissolved this effect's predecessor in note 0026 — the anchor drives episodes to the deadline much more often (at-cap 0.677 against 0.417), so unrestricted deal-conditional rows compare differently-composed populations of deals. Bracketed rows are resolved. Two rows in the matched column are **tautological, not evidence**: deal rate and at-cap are exactly 0.000 with zero-width intervals because the restriction *defines* both sides to have finished early, and in this scenario an episode ends early only by closing. The fairness rows in the same column are ordinary paired differences over 16–19 games. Source: note 0022 §3; tables in `_trivial_baselines_analysis/trivial_sonnet*_vs_all_llm.*`.

measure	all-LLM − bayes-anchor	matched censoring
distance to Nash solution	**−0.104** [−0.194, −0.025]	**−0.174** [−0.335, −0.026]
Gini of capture	**−0.028** [−0.050, −0.007]	**−0.039** [−0.072, −0.013]
worst-off capture	+0.024 [−0.004, +0.052]	**+0.043** [+0.014, +0.074]
seat-0 capture	**−0.128** [−0.209, −0.056]	**−0.131** [−0.243, −0.020]
seat-0 capture-z	**−0.318** [−0.607, −0.076]	**−0.357** [−0.645, −0.074]
normalized Nash welfare	+0.011 [−0.039, +0.057]	**+0.045** [+0.012, +0.076]
at-cap rate	**−0.260** [−0.427, −0.094]	0.000 by construction

11.5 The reversal: at 24 clusters the frontier anchor extracts

The arc's most consequential result is one it was not designed to produce. Re-running the *reference* cells on the powered bank puts the Bayes-anchor-versus-all-LLM contrast on 24 Sonnet clusters instead of three, and it **reverses §9.3's reading** (Table 19). Seating the Bayesian anchor at a Sonnet table now takes +0.128 [+0.056, +0.209] of seat-0 capture and pushes the table 0.104 [+0.025, +0.194] *further* from the Nash solution with a +0.028 [+0.007, +0.050] higher Gini — all three resolved, where note 0018 reported no extraction and fairness-neutral-to-positive anchoring. The conditions match: the same `--model-thinking off`, the same canonical scaffold, the same abstract framing, the same `moves_chat` arm, the same `reverse_mixed` lineup with `bayes-rational` at seat 0. What changed is 3 clusters → 24 and the screened bank.

The reversal survives the restriction that killed its predecessor, which is what makes it worth believing. Note 0026 found that the thinking-on anchor's apparent effects dissolved under matched censoring (§9.3), and the anchor does censor here too (+0.260 at-cap, resolved). Restricting to the 20 pairs where neither side ran the full protocol, the reversal *strengthens* on every fairness measure — distance to the Nash solution −0.174, Gini −0.039, worst-off +0.043, capture −0.131, welfare +0.045, all resolved. So this is not an artifact of the anchor closing a different set of games. What matched censoring cannot do is rescue power: 20 pairs from 96 is a heavy cut, so it is a robustness check and not an independent estimate.

Which claims move, stated narrowly. That the anchor *extracts* from Sonnet, and that it makes the Sonnet table *less fair* on distance-to-NBS and Gini, survive both the full sample and the censoring restriction and supersede note 0018 on those two points. That it *harms welfare* does not survive: normalized Nash welfare is +0.011 unresolved unrestricted, and its resolved +0.045 under matching rests on 20 pairs. **The honest statement is that the anchor redistributes rather than destroys**, and no welfare claim is made in either direction. Everything in §9.3 and §12.7 that describes the frontier anchor as a fairness anchor is hereby scoped to the thinking-off three-cluster cell it was measured on.

The methodological sentence, which is the part that generalizes. *Every three-cluster frontier claim in this project that has since been re-measured under power has moved.* The thinking-on anchor's fairness movements dissolved at matched censoring (0026); the thinking-off anchor's distance movements resolved only after the cell was doubled to $n = 60$, and remain censor-confounded; the "frontier model never caves" zero became 0.135; and the anchor's extraction null on Sonnet became a resolved +0.128 extraction. None of these were reversed by finding a bug. They were reversed by giving an already-correct estimator enough independent games to speak. Limitation 4 is not a caveat on this project's frontier results; it is the single largest source of error in them.

11.6 What does not generalize: seats, censoring, and the parity claim's true scope

Three honest limits, and the seat sweep is the sharpest.

Seat 0 is doing some of the work. The two load-bearing policies were re-run at seats 2 and 4, each contrasted against the all-LLM baseline *read at the same seat*. The negative half of H1 is seat-robust: `passive-gate` extracts nothing anywhere (focal capture −0.034 / +0.000 / −0.081 at seats 0/2/4) and its within-table advantage is negative everywhere. **The fairness *gain* is not**: its worst-off improvement over the all-LLM baseline resolves at seat 0 (+0.065) and does not replicate at seat 2 (−0.002) or seat 4 (+0.024). On one seat that reads as an effect; on three it reads as seat-0-specific. `greedy-anchor`'s extraction advantage is likewise resolved at seats 0 and 2 and gone at seat 4. What the sweep *cannot* settle is the section's actual headline — *parity* between `passive-gate` and the Bayesian anchor — because that needs Bayes-anchor cells at seats 2 and 4, which were queued at the time of writing and had not returned (Slurm 16453662, 16453663). **The parity result is established at seat 0 only**, and the seat-0-specificity of the fairness gain above is a direct reason to doubt it generalizes. The maximizer lane documents how badly a seat-0-only read can mislead: its own first 30 seat-0 episodes read +17.1 own-utility and "69% more selfish" — a clean confirmation of the hypothesis under test — and the full six-seat sweep reversed the sign on four of six seats. A seat-0-only campaign there would have produced a confident, publishable, and wrong result.

Differential censoring threatens every deal-conditional row, and it is reported rather than assumed away. These treatments change how *long* episodes run, not just how they end (at-cap 0.469 all-LLM, 0.594 Bayes anchor, 0.760 `greedy-holdout`), so the deal-conditional rows of the hold-out columns are computed on a policy-selected third of their games and should be read that way. The cell carrying the headline is the exception that makes the headline safe: `passive-gate` censors *slightly less* than the baseline (0.427 against 0.469, −0.042 [−0.146, +0.073]), so there is no censoring gap for a restriction to remove. That prediction was checked rather than asserted — the headline contrasts are re-run on both-finished-early pairs. One implementation note worth carrying: intersecting the early-finishing set across *all* cells at once is empty in practice at these rates, so the restriction must be applied **per contrast**, to the two columns being compared.

A decision-function measurement would be immune to all of this, and §11.3's replay is one: it is computed over policy turns on a fixed set of stored states, so there is no deal-conditioning to censor and no population to select. Every *outcome* number in this section is exposed to censoring; the replay numbers are not. That asymmetry is a general argument for reporting decision-function measurements alongside outcome measurements, not as a substitute for them.

And in the one case where both were run, the cheap one was right. The calibrated maximizer's replay predicted, for free and before any GPU spend, that a truthful opponent model would make the agent concede rather than extract; the 864-episode campaign then found exactly that at the outcome level, with the fairness basket moving in the predicted direction and the deal

rate flat (§11.3). One instance is not a general warrant — a decision-function replay cannot see how a table *responds* to a changed anchor, which is the whole reason the campaign was still worth running — but it is direct evidence that the CPU-only replay is a usable forecast of a GPU-day, and it is what let this lane revise a preregistered prediction against its own hypothesis before spending anything.

11.7 What the arc adds up to

Discipline is cheap and distribution is not. The fairness half of this project's flagship positive is delivered in full by a policy that never speaks, never proposes, and could be written in two lines: refuse anything below your walk-away value. It costs zero tokens, needs no beliefs, no best response and no opponent model, and on the frontier population it is *better* than the full Bayesian stack at the fairness job while capturing a third as much. Everything the Bayesian machinery adds is money for the seat that runs it.

Set that beside the rest of the paper and the shape is consistent, with one instructive exception. *Neither scale* (§5.2), *nor test-time thinking* (§9.5), *nor a solver scaffold* (§5.5), *nor outcome-aligned RL* (§10) buys the distributional disposition; all four, at best, buy *discipline*. The exception is the fifth intervention, and it is the arc's most surprising positive: **correcting the opponent model is the one thing in this paper that moved every distributional endpoint at once** — distance to both solution concepts, Gini, the worst-off seat and total welfare, all resolved, all under matched censoring, with deal rate flat (§11.3). It is worth being precise about what that does and does not overturn. It does not make a language model fairer; the seat that changed is the computable one, and what changed inside it is a probability table. **Truthful knowledge about your counterparties makes a rational agent generous**, because honest acceptance probabilities are sub-certain, unanimity compounds sub-certainty across five opponents, and the only lever a best-responder has on a joint acceptance probability is to give everyone more. So the exception strengthens rather than breaks the pattern: the one thing that reliably moves distribution is still putting a better *decision rule* at the table — first discipline, and now calibration — and none of it is a property of the language models. §12.8 states the training-side version of the same sentence.

Two things the arc adds that are not restatements. First, **extraction against a frontier population comes from commitment and comprehension, not from machinery**: the one intervention that moved Sonnet's behaviour by a resolved margin was *a sentence* attached to unchanged behaviour, and the frontier model responded to it where the 8B model did not. Second, **the anchor's frontier innocence was a power artifact**. On the question the arc was actually built to answer — can a self-interested computable agent take value out of an LLM table? — the answer against small models was already yes, and against frontier models it is now a qualified yes as well.

The declared successor is `proposals/2026-08-03-fairness-grpo-v2.md`, which turns the arc's machinery into training input: §10's self-play failure gets a *population* of opponents drawn from exactly these zero-token policies, and the reward's violation branch gets clipped so the objective's dynamic range stops living in the IR region. Its headline evaluation is the one question this whole program raises and has never tested — a **prose-only condition** in which a seat's private information arrives as natural language rather than as a score sheet. Computable policies cannot play it by construction, and the engine can still score it exactly, so it is the cleanest available test of whether an LLM negotiator buys anything an algorithmic agent plus a parser does not. It is preregistered and unfunded; the honest statement of the arc's conclusion is that until that eval runs, everything positive we have measured about fairness in this environment is a property of disciplined *decision*

rules, and none of it is a property of language models.

12 The durable findings

12.1 Headline: oracle-regret anti-tracks negotiation performance

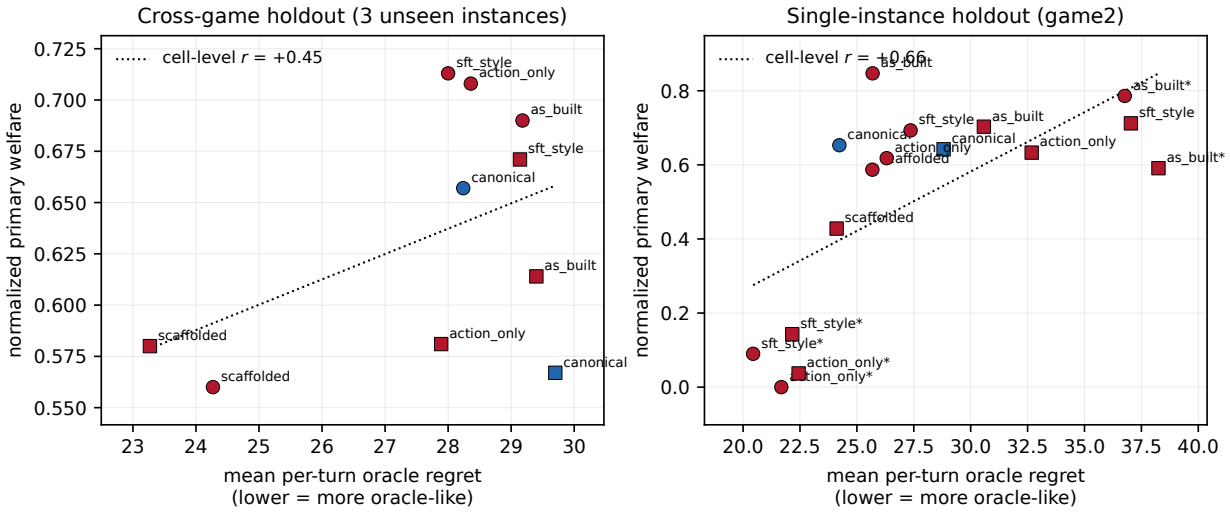

Figure 12: The regret/outcome divergence. Each marker is one evaluation cell (one policy × one information condition): x is mean per-turn oracle regret in surplus points, where *lower* means decisions closer to the exact best response; y is normalized joint welfare, where *higher* is better. Circles are full-information cells, squares private; blue marks the untrained canonical baseline, red the trained and scaffolded policies; a trailing ∗ in a label marks an over-trained final checkpoint. If regret were a proxy for performance, the cloud would slope *down*. It slopes *up* in both evaluation designs (cell-level $r = +0.45$ on the clean cross-game cells, $+0.66$ single-instance). **Left panel is the clean re-run**; right panel is the single-game fleet, which was measured clean and is unchanged. The lower-left corner — lowest regret, worst outcomes — is occupied by the prompt scaffold and, in the right panel, the collapsed over-trained arms. These are cell-level (ecological) correlations across policies, not within-policy causal slopes.

The finding that replicated in every place we looked is that *the project's own central metric does not track the outcome it was built to improve.* The evidence changed shape under de-contamination and got a new leg:

- The **prompt scaffold** is the cleanest case and the one that survives fully, now on two independent clean measurements. In the P2 campaign, re-run clean, it cuts Qwen3-8B's mean per-turn regret by **4.35** points while cutting its deal rate by **0.342** — and de-contamination made the divergence *sharper* on both axes than the published version, because the closure penalty grew while the regret advantage halved (§5.5). On the clean cross-game fleet the same shape appears independently: regret 3.97 points lower $[0.459, 6.355]$, deal rate 0.217 lower.
- The **collapsed arms** remain the extreme case, and they are clean: `action_only`-final posts the best regret of any policy in the entire evaluation (-2.6 full, -6.4^{*} private) while closing *zero* deals, because a policy that never closes never takes a costly turn.
- **One leg is withdrawn.** The paper previously reported that on unseen games every trained arm had significantly *lower* regret than canonical (-2.4, -4.2, -2.0) while closing equal or fewer deals. On clean cells that is false: the trained arms' regret is at or slightly *above* canonical

(`as_built` $+0.95^{*}$ full, -0.30 n.s. private; `action_only` $+0.12$ / -1.81 n.s.; `sft_style` -0.24 / -0.56 n.s.). The mechanism is understood: a fabricated no-op turn incurs no regret, so the more contaminated arms looked more oracle-like, and the trained arms were the more contaminated ones.

- **But the same mechanism strengthens the finding overall**, in a way worth stating precisely. Because fabricated turns are unvalued, *every regret number quoted from a contaminated cell was biased downward in proportion to that cell's contamination rate.* Across the re-baselined ladder, removing fabricated turns raised mean regret in every affected cell (Qwen3-8B 16.7 → 24.4, Qwen3-4B 16.4 → 30.3, Gemma 18.4 → 37.6, Ministral 10.2 → 27.5) while outcomes *improved*. So the artifact was independently manufacturing low-regret, low-closure cells — a second mechanism producing exactly the anti-correlation this section reports, and a reason to trust the finding's shape while distrusting any specific magnitude taken from a dirty cell.
- Aggregating across policies, the cell-level correlation between regret and welfare is *positive* in both designs ($r = +0.45$ clean cross-game and $+0.66$ single-instance; Figure 12). One honest qualification: on the clean cross-game cells the positive correlation is carried *entirely* by the two scaffold cells — among the eight trained-and-untrained cells alone it is -0.33. The divergence in that design is a statement about the scaffold and the collapsed arms, not a smooth law across all policies, and the single-instance design (which contains the collapsed checkpoints) is where the full cloud slopes up.

The mechanism is not mysterious and not a bug in the oracle, which is exact. Per-turn best-response regret prices a decision against the single best move available at that moment; in a six-party unanimity game, realized welfare depends on whether five other parties can be brought to accept the *same* offer. A policy that avoids costly-looking turns by never committing — proposing instead of accepting, deliberating instead of closing, or emitting malformed actions and taking no turn at all — scores well on the per-turn metric while producing no agreement. **Oracle-referenced per-turn optimality is a measure of local decision quality, not of negotiation performance, and using it as a training signal or a model-selection criterion selects the wrong policies.** This is the finding we would carry forward to any oracle-referenced evaluation, in this domain or another.

12.2 The single-instance-holdout trap

The pilot's positive would have stood without the widening: two significant effects with intervals excluding zero on 20 matched pairs, a plausible mechanism, and three ablation arms triangulating it. All of it evaporated on three unseen instances. The generalizable lesson is narrower and more useful than "use more data": **when the holdout has one instance, cluster-bootstrapping the seeds produces intervals that describe seed noise while reading like generalization intervals.** The caveat was correctly recorded at the time and was still not enough to stop the wrong conclusion from being written down first. A held-out group of at least three instances belongs in the design from day one, not as a follow-up.

12.3 The 95.5%-prose decomposition of naive divergence-DPO

Constructing preference pairs where the chosen side is a synthesized oracle rationale and the rejected side is the model's own emission produces a training signal that is 95.5% *writing style* and 4.5% *decision* (312.7 versus 14.1 nats of separation at step 0). The measurement is cheap — it is an

`eval_on_start` pass plus one arm whose pairs differ only in the action fields — and it is diagnostic before any training compute is spent. Anyone building oracle-imitation preference data should measure it. The corroborating detail is the one that makes it airtight: the model's own prose costs it 0.04 nats because self-generated text is nearly free under its own distribution, while the oracle's prose costs about 299.

12.4 Uniform collapse at the full step budget

All three objectives — DPO over full pairs, DPO over actions only, and SFT on rationales — collapsed at the same full step budget, with deal rates falling to 0.00–0.20 and malformed emissions rising to 0.25–0.70 per episode, while the pre-saturation step-10 checkpoints of the same runs were intact. Early stopping here is *structural and objective-independent*, not a hyperparameter to be tuned. The practical consequence: save frequent checkpoints and evaluate a pre-saturation one, because the failure is invisible in the training curve — the collapsed arms have the best training losses.

12.5 The audit chain as a deliverable

Five distinct measurement defects were caught, each by a control, a replay verification or a cause-classifying audit rather than by inspection: two seat-binding oracle bugs that mis-priced every stored propose action; a replay retry bug that graded 4,470 placeholder turns; an equilibrium-composition mirroring bug (all §7.2); and **the engine's swallow-and-fabricate path, which invented 26.2% of the campaign's turns and passed a fail-closed validity gate at 1,770/1,770** (§3.6). The practices that caught them generalize:

- **A measurement-only legacy double** to separate a bug's effect from a coverage change — this is what let us state "regret re-priced on 41% of valued turns, mean −5.4%" instead of an unattributable difference.
- **Built-in scope checks**: ultimatum cells moving *exactly zero* under a propose-pricing fix, because single-issue responders cannot propose, is stronger evidence the fix was correct than any assertion about the diff. The same shape appears in the fabrication audit: contamination confined *perfectly* to the two table types that use the batched pool, and flat across rounds, is what rules out every context-length and capability story at once.
- **Bit-identity diffs across rebuilds**, which is how dataset v2 was cleared without re-running the trained arms.
- **Classify by cause, not by symptom.** One placeholder string had three producers with three different fixes — an engine crash, a model that spent its budget on hidden thinking, and a deliberate no-op. An audit that pools them (or screens on the action type, which under-counts the artifact and over-counts legitimate play *simultaneously*) points the remediation at the wrong target: here it produced a plan to raise a token cap that not one turn in the campaign had ever reached.
- **Audit an existing accident before spending compute.** The A/B that proved the mechanism was already in the repository, unnoticed: one dead cell and its sequential re-run, byte-identical arguments, 100% versus 0% fabricated. So was the falsification control that licenses the whole re-baseline — the one cell that had been clean on both sides, which moved −0.017 where the contaminated cells moved +0.14 to +0.53.

- **Fail-closed publication, with its limits stated.** An immutable manifest plus a validity gate that rejects a campaign on stale or errored rows means "Slurm said COMPLETED" can never become a headline — and it is exactly as strong as the list of properties it checks. Ours checked provenance, completeness and status, and certified a campaign a quarter of whose turns no model produced. **Add a liveness clause**: the modal generated-token count per cell, which costs nothing and would have failed the run in its first minutes.
- **Reading the transcripts**, which is the only thing that caught the ultimatum label-misread artifact and the DPO style confound.

12.6 Surface sensitivity: the invariance a computable agent has and an LLM does not

The narrative-framing arm (§8) is the project's cleanest statement of what separates the two populations, because it is the only comparison in which *nothing about the game changes*. The score sheets, thresholds, feasible set, Pareto frontier, Nash solution, optimal policy and `instance_id` are byte-identical across arms; only the vocabulary the numbers are printed in differs. Under that transformation the computable policies produce **identical deals with every metric difference exactly** 0.000 — they read option indices, so relabelling cannot reach them — while the LLM table loses welfare (-0.083, $[-0.143, -0.015]$), moves off the efficient frontier ($+0.238$, $[+0.044, +0.515]$) and tables illegal packages in a quarter of episodes ($+0.217$, $[+0.117, +0.350]$). **Part of the LLM-versus-rational gap in this environment is not about solving the bargaining problem at all; it is sensitivity to the surface the problem is written on.**

Two features of the design are the transferable part, independent of the effect size. First, the confound was removed *by randomizing it* rather than by removing the content: the story-to-numbers mapping is a hash of the instance id, so the labels still invite inference and the inference they invite is uninformative — which is what allows a realistic scenario and a leakage-free measurement to coexist. Second, the contract is enforced by a *masked-prompt byte-identity* test rather than a numbers-only check, and that strictness paid for itself immediately by catching a scaffold worked example whose hardcoded literal would have named a real issue in one arm and a fictional one in the other (§8.1). Any framing, persona or scenario manipulation in an evaluation should carry both: a decorrelated mapping, and a control agent that must *fail* to benefit from the skin.

12.7 The counterparty population is a hidden variable, and the size of the effect is a power question

The same intervention — replace exactly one seat with a computable Bayesian-rational policy — has very different consequences for the policy itself in the two populations, measured on the same normalized available-surplus scale by the same estimator. Against open-weight seats it captures $+0.147$ to $+0.444$ more than the model it replaced, on all five slots with every interval excluding zero (§14.1). Against a table of `claude-sonnet-5` the same swap captures far less — and *how much* less is exactly the kind of quantity three game clusters cannot pin down. On the committed three-game bank it reads as a precise null (-0.017 $[-0.081, +0.031]$, §9.3); re-measured on 24 screened games at 96 episodes per cell under the identical condition it is a resolved $+0.128$ $[+0.056, +0.209]$ of extraction, accompanied by a table that lands *further* from the Nash solution and more unequal (§11.5). The population contrast survives that correction — 0.128 is still a quarter to a third of what the same policy takes from open-weight seats — but the "takes nothing at all" version of it does not.

Two durable statements, and the first is the one to carry. The counterparty population is a hidden variable in any agent-policy result: a policy that looks strong is partly a statement about who it played against, and the way to see which is to hold the policy fixed and move the opposition, which is what these campaigns jointly do. The second is the correction itself: *a null measured on three clusters is a statement about power, not about the world.* This one was published, reasoned from, built on — it is the finding §10 cites as "seating a computable agent works" — and it reversed under nothing but a bigger, better-screened bank, with no bug found and no estimator changed. Where a project's headline rests on failing to reject, the sample size is the finding.

12.8 Discipline is learnable here; distribution is not (yet)

Two interventions with nothing mechanistically in common were applied to the same negotiator population and moved the same axis while leaving the other one flat. Scaling from 4B to 32B bought *behavioural* discipline — 32B has the lowest individual-rationality violation rate of anything tested — and left the distributional deficit against the computable control exactly where it was (§5.2, §5.4). Outcome-aligned RL on a reward that is a monotone transform of normalized Nash welfare bought the same discipline (below-threshold agreements -0.079 to -0.129 on held-out games) and the same flat distribution — and, unlike scale, charged a welfare cost, because the discipline was purchased substantially by not agreeing (§10).

A third intervention now says the same thing from the cheapest possible direction, and it is what makes this a finding about the axis rather than about two methods. §11 decomposes the one thing in this project that *does* move distribution — seating a computable agent — and finds that its fairness half is delivered in full by `passive-gate`, a policy that never speaks, never proposes, and does nothing but refuse offers below its own walk-away value: indistinguishable from the full Bayesian agent on all six fairness measures at 0.235 less capture, and on the frontier population resolved *better* than it. The Bayesian belief model, the best response and the optimal stopping are what earn the *money*. So the ledger reads: discipline is available from scale, from RL, and from two lines of code, and distribution is available from none of them — including from making the agent's opponent model empirically correct, which buys concession rather than extraction (§11.3). **Distribution is not a capability anything here supplies; it is a property of a decision rule that refuses, and the only way anyone in this project has moved it is by putting such a rule in the room.**

That a gradient aimed *directly* at the fairness metric lands on the discipline axis is the part worth carrying. The immediate cause is legible in the reward's arithmetic — the individual-rationality region dominates the objective's dynamic range, so "do not short anyone" is where nearly all the gradient is, and the follow-up that tests this is a clipped violation branch (§10.7). But the composite picture across the project is that the distributional gap does not behave like a capability the model is missing: *handing* the model a fair candidate deal at inference time raises welfare $+0.154$ (`research-notes/0008`), and seating a computable agent at the table redistributes toward the other five seats (§12.7). The model can act fairly when it is told what fair is; neither scale nor this reward taught it to want to. For anyone designing an outcome-aligned objective, the transferable form of the lesson is to check which region of the outcome space carries the objective's dynamic range *before* training, because that region — not the metric's name — is what the policy will optimize.

12.9 Smaller results worth keeping

- **Divergence precedes reasoning** (§6), with the empty-scratchpad control as the design that establishes it. Any claim of the form "the model's reasoning goes wrong at step k" needs a

$k = 0$ control.

- **The sign of the cheap-talk effect is a property of the population, not of the protocol.** Across six open-weight models in an identical protocol it ranges from −0.233 deal rate to +0.129 normalized primary welfare, and is not ordered by scale. On the same instances, silencing the channel *costs* `claude-sonnet-5` 0.233 [−0.333, −0.133] of deal rate — the opposite sign, and a double dissociation between the two populations (§9.2). No summary of the form "cheap talk helps/hurts negotiation in this environment" is available; the honest statement names the population.
- **Print closure beside every conditional fairness metric.** Distance to a bargaining solution, Gini, and conditional Nash welfare are all computed over *reached* deals, so a policy that rarely closes can look fair by selection, and the unconditional primary — which scores a no-deal as zero — can make the same policy look bad for the opposite reason. In the frontier campaign the two orderings genuinely disagree, and which one you see is decided by the choice of headline metric (§9). The cheap fix is to never report one without the other.
- **Reasoning is asymmetric**: 14 of 18 reasoning-introduced divergences are the model talking itself out of a good accept, and zero are reasoning rescuing a bad baseline.
- **LLM agreements are less equal, at every scale**: higher reached-deal Gini and lower egalitarian welfare for all six models with intervals excluding zero, higher distance to both bargaining solutions for five of six, and higher Pareto-frontier distance for four of six. The distributional deficit is the universal result; the efficiency deficit is not (§5.4).
- **Closure volume is not quality, and the two can be anti-correlated.** The model with the highest deal rate of the six (0.992) has the highest individual-rationality violation rate (0.942) and the second-highest inequality. Any negotiation benchmark that leads with agreement rate will rank indiscriminate assent above disciplined play.
- **A zero deal rate deserves a positive check that the model was asked.** Three cells in this project reported that a model essentially never closed a deal. *Two of the three were the harness*: the dead `all_llm` run (§3.6) and the OLMo scaffolded cell, whose published 0.000 becomes 0.492 clean (§5.5). The third, the collapsed over-trained checkpoint, is real. A striking null is the easiest result to publish and the hardest to distinguish from an integration failure, and the distinguishing check — did any tokens get generated? — is one line.

13 Limitations

1. **Slot substitutes.** `gemma-3-4b-it` stands in for the Gemma-4 slot and `Ministral-8B-Instruct-2410` for the Ministral-3 slot, because the pinned `transformers =4.57.6` does not register the newer architectures. They support conclusions about the models actually run and none about Gemma 4 or Ministral 3. A validated Transformers upgrade unlocks both slots. *The original version of this limitation said the two cleanest results were both substitutes, and implied the headline rested on them.* On clean cells that is no longer the shape of the evidence: the universal distributional deficit resolves for all six models including the two Qwen cells and the 32B scale point (Table 4), so the finding does not depend on either substitute. What the substitutes still monopolise is the *aggregate-efficiency* deficit, which resolves for them and for nobody else — and they are also the two models exhibiting the indiscriminate-assent pathology (IR violations 0.942 and 0.883), so that pairing may be a fact about weak models rather than about these two model families.

2. **A quarter of the original campaign was engine-fabricated** (§3.6). Every cross-model, cheap-talk and cross-game number here is from a clean re-cut, but three consequences remain live. The re-runs used `--sequential --concurrency 1`, which changes sampling order, so they are not token-level reproductions of what the originals would have produced — the comparison is between engine paths, applied uniformly, not between two draws of one process. The `_b2` cells *avoid* the broken path rather than exercising the committed repair. And de-contamination does not make every cell citable: a model that genuinely cannot emit a parseable action is failing the task, and what the re-baseline establishes is which failures were the model's and which were the harness's.

3. **The preset cells are the last contaminated numbers in the paper.** `ultimatum` audits clean, but the divide-the-dollar column contains one dead cell (OLMo, 1.000 fabricated) and one 8.6%-fabricated cell (Qwen3-8B), and neither was re-run because no claim here depends on them (§5.7). Everything else — cross-model, cheap-talk, scaffold, cross-game, reverse-mixed — is now measured on clean cells or explicitly withdrawn.

4. **Three instance clusters.** Every P2/P3 interval resamples only three committed-instance seed clusters, so endpoints are coarse. Effect size and replication across models carry more weight than nominal exclusion of zero. Widening the instance bank is the single highest-value change to the measurement. **This is no longer a hypothetical**: the one campaign that widened it (§11.1, 24 screened games) overturned a published frontier result of ours and retracted a second (§11.5), so every three-cluster claim in §5–§9 should be read as provisional in exactly the way §11.5 spells out — and the frontier cells of §9 have *not* been re-run on the wide bank except for the anchor contrast.

5. **Seeds versus instances as clusters.** The P4 pilot's intervals cluster on rollout seeds because the holdout had one instance, and §7.5 is what that cost. Any interval whose clusters are seeds describes seed noise.

6. **Fixed seat 0 in mixed tables.** Normalization removes the arbitrary-utility-scale confound but not the seat-assignment confound; the matched same-seat counterfactual is the best available substitute and it resolves nothing (§5.6). Rotating the LLM through all six sheets is required for a causal extraction claim.

7. **Cross-bank absolute levels are not comparable** — but not for the reason originally given. The contaminated cross-game fleet carried ≈2.9 malformed emissions per episode for all policies including untrained, versus ≈0 on `game2`, and this paper described it as "a bank property" of the three fresh instances. **That is deleted**: on the clean re-run of the same instances with the same policies the figure is 0.00–0.20 per episode. It was a symptom of the engine failure, not a property of the bank. What survives is only the weaker methodological point that absolute rates should not be compared across evaluation banks without checking that both were measured on the same path.

8. **One model, one scale in P4.** Every trained arm is Qwen3-8B with LoRA r=32 on 326 pairs and 81 steps. The negative is a statement about this intervention at this scale on this backbone, not about preference learning on negotiation in general.

9. **Omniscient regret in private cells.** The best-response oracle sees all sheets. Private-condition regret is a counterfactual against an unimplementable policy and must be compared only within information condition.

10. **No-deal accounting.** Conditional metrics omit no-deal episodes by construction; unconditional normalized welfare assigns no-deal zero. Both are reported because neither alone is sufficient.

11. **Unavailable rather than imputed — and knowing *why* it is unavailable.** Chain-of-thought step localization is unavailable for every campaign episode from stored data, and equilibrium regret is missing for one of ten preset cells. Neither is imputed, substituted, or silently zero-filled. But the second case is a cautionary one: that missing preset cell was reported as an admirable refusal to impute when it was in fact 100% engine-fabricated (§5.7). Non-imputation is necessary and not sufficient; a missing value needs a diagnosed cause before it is reported as a result about a model.

12. **The frontier campaign is observational, thinking-off, and single-vendor** (§9). It was run after the main design as a separate campaign, so every number in §5–§7 is open-weight 4B–32B and the frontier cells are $n = 30$ (five seeds over six instances) rather than the main campaign's density. Four scope limits travel with it. *(a)* The matched condition is thinking off. That condition does understate these models, but the amount has now been measured rather than guessed: a 60-episode-per-cell thinking × framing 2×2 (§9.5) finds thinking raises welfare and the worst-off share while leaving distance-to-solution and equality spanning zero, so **the distributional deficit reported here survives thinking and is not a thinking-off artifact**. What thinking does remove is the cost of a narrative framing. Reasoning depth is itself a scope limit: hosted adaptive thinking has no fixed budget, and an implicit and an explicit request for it are measurably different conditions. *(b)* Claude only: no OpenAI credential exists on this host, so nothing here speaks to cross-vendor differences. *(c)* No full `claude-opus-5` rung; the four paid episodes are a labelled datapoint on one instance and are not comparable to the $n = 30$ cells. *(d)* The rational-seat contrast is fixed at seat 0, inheriting the seat-assignment confound that limitation 6 describes for the mixed tables — the companion's seat rotation was never run against a frontier population.

13. **The framing arm is one skin, one model, one protocol, with an unmatched surface confound** (§8). Four limits, in descending order of how much they could change the verdict. *(a)* **Label verbosity is not matched**: framed labels are long multi-word strings ("phased over eight years") where abstract labels are `opt1`. The randomization guarantees the story carries no information *about the sheets*, but it holds neither surface length, tokenization, nor reading load constant, so the measured effect is "narrative framing as a package". The decisive follow-up is a **length-matched nonsense-label control** — multi-word gibberish at matched token length — which would separate semantics from verbosity, and it is not run. *(b)* Six instances is six bootstrap clusters *and* six skin draws; the decorrelation argument is asymptotic in bank size, airtight over the 48-instance confirmatory bank and a weak law over six games, which is why the per-instance panel of Figure 7 carries as much weight as the pooled interval. *(c)* Only `datacenter` was rolled out to LLM cells: `festival` is implemented, tested and verified framing-invariant for rational seats, but nothing here separates a framing effect from one skin's idiosyncratic vocabulary. *(d)* One model (Qwen3-8B), one arm (`moves_chat`), one scaffold, and a solo tripwire whose interval is about 0.5 wide — it excludes a large preference leak, not a small one, and a leak worth roughly +0.1 of the primary would not have been detected at this sample size. *(e)* The arm is **no longer all-LLM only**: §8.5 adds the two `reverse_mixed` cells that complete the 2×2 with a rational seat under both skins. What those cells do *not* yet have is the seat rotation of §14.1 — the anchor sits in seat 0 in both, so the same fixed-seat caveat as limitation 6 applies, and the fairness interaction they estimate is unresolved on 16 four-way-complete deal-conditional pairs. A framed replication with the anchor rotated through all six sheets is the natural next cell, and it is not run.

14. **The RL pilot is 24 steps, one backbone, and symmetric self-play** (§10). Three limits, in descending order of how much they could change the verdict. *(a)* **24 GRPO steps per arm**

is 12% of the preregistered 200, stopped by a hard wall-clock budget on a step that costs ≈17 minutes. Nothing here rules out a different trajectory at 200 steps — though the ladder argues against optimism, since welfare gets monotonically *worse* from step 10 to step 24 in both arms. This is a well-instrumented short run, not a converged one, and no "it would have worked eventually" reading is available for free in either direction. *(b)* **Symmetric self-play may have degenerated the λ frontier by construction**: every seat is the same policy and the seat-averaged reward is identical for every λ, so the preregistered manipulation check (λ=0, pure self-interest, which was supposed to *reproduce* the pathologies) instead behaved much like λ=1 everywhere except the among-IR endpoints. The λ contrast reported here is therefore a weak test of what λ was designed to vary, and the informative version needs heterogeneous opponents. *(c)* **The two generalization cells carry one game cluster each.** The ultimatum and divide-the-dollar presets are deterministic, so all three generated "instances" of each are the same game and a cluster bootstrap over them is degenerate. Those cells are reported descriptively with their across-seed spread and no interval; the ultimatum effect is large and unanimous across seeds and both arms, the divide-the-dollar effect is smaller than its seed spread and is suggestive only. One model (Qwen3-8B), one adapter configuration, and a training bank of 24 games apply throughout, and the held-out deal-rate collapse is direct evidence that 24 games is too few.

15. **The policy-decomposition arc is seat-0, one bank, and one of its two headline comparisons is still incomplete** (§11). Four limits. *(a)* **The parity claim — that `passive-gate` matches the Bayesian anchor on fairness — is established at seat 0 only.** The targeted sweep put both load-bearing policies at seats 2 and 4 and found the gate's *extraction* null seat-robust but its fairness *gain* over the all-LLM baseline resolved only at seat 0 (+0.065 against −0.002 and +0.024), which is a direct reason to doubt the parity generalizes; the Bayes-anchor cells at seats 2 and 4 that would settle it were queued and had not returned at the time of writing (Slurm 16453662, 16453663). Seats differ in score sheets, reservation values and position in the rotating-proposer order, so every number in that section should be read as "a computable policy *in seat 0*". *(b)* **The bank is screened, and the screen is a selection.** Its 24 games were admitted from 137 candidates on four preregistered screens, two of which select for *discriminativeness*; results generalize to games where the solution concepts disagree and no package is obviously best, not to everything the generator makes. The screen was chosen for defensible reasons — NBS and KS coincide in four of the committed bank's six instances, making two of this paper's fairness columns redundant there — but it means this bank and the committed bank are not interchangeable, and no number is compared across them without being labelled cross-bank. *(c)* **The matched-censoring convention is per-contrast, not global**, because at at-cap rates of 0.43–1.00 the set of games finishing early in *every* cell is empty in practice. Each restricted contrast therefore conditions on a different subsample, restricted contrasts are not mutually comparable, and 20 pairs from 96 is a robustness check rather than an independent estimate. *(d)* **The calibrated maximizer's campaign has since run, and its own scope is narrower than its verdict sounds.** 864 episodes over 12.22 GPU-hours resolve seven endpoints and falsify H2-strong in both directions (§11.3), but three limits bound it. The **magnitude is a seat-0 magnitude**: the six-seat sweep was dropped to hand the shared pod to the RL lane, so only the *direction* has multi-seat support, from the decision-function replay where four of six seats concede and two do not. **One opponent model sits at the table** — five Qwen3-8B seats, thinking off — and the fitted curves predict the effect should shrink or vanish against Qwen3-32B and `claude-sonnet-5`, whose below-threshold acceptance is already at or near the step model's assumed 0.00, leaving little over-confidence to correct; that prediction is named in §14 and untested. And the acceptance curves the policy plans against are

observational curves fitted on episodes played against other opponents, so a policy that pushes counterparties toward $z = 0$ is operating exactly where those curves are steepest and least sampled, and it shifts the distribution they were fitted on. One further caveat is a strength rather than a weakness and is recorded so it is not mistaken for a gap: the campaign's private-information half is a provably-inert placebo in which one of ≈26 comparisons resolved, which is α-consistent and is reported as a false positive.

16. **The two later campaigns inherit the same bank.** The framing and frontier campaigns both play the six committed instances — three generated games in full- and private-information variants — so limitation 4 applies to them unchanged, and neither has the P4 evaluation's saving grace of a fresh, fully unseen instance group. Results that replicate across models inside this bank have not been shown to replicate across game families or across generator settings (Pareto slack, issue-type mix, six parties, five issues, three options).

14 The honest path forward

If goals 3 and 4 are pursued, the pilot says what has to change — and one item is a prerequisite for the rest.

First, fix the objective, not the data. The regret/outcome divergence (§12) means per-turn oracle preference is the wrong training signal in this game family regardless of how clean the pairs are. Any re-attempt should optimize an *outcome-aligned* signal — realized normalized welfare or closure at episode level, with the oracle used as a critic or a shaping term rather than an imitation target — and should pre-register outcome metrics as the primary read with regret demoted to a diagnostic. Running more divergence-DPO with better labels is not indicated. **This item has since been executed** (§10), and it produced a second negative with a specific successor: the reward's dynamic range lives almost entirely in the individual-rationality region, so the highest-value follow-up in this project is now to **re-run with a clipped violation branch** (or an among-IR fairness objective) and check whether the among-IR endpoints move — the one test that separates "this reward was an IR-violation penalty" from "the distribution does not train". Behind it, in order: heterogeneous-opponent training so λ varies what it was designed to vary, a longer horizon before any positive claim, and fixing the held-out deal-rate collapse first, since every unconditional endpoint inherits its composition.

The v2 design that acts on all of this is written and unfunded: `proposals/2026-08-03-fairness-grpo-v2.md`. It answers each measured v1 failure with a specific change — the clipped violation branch above, calibrated offline so the individual-rationality region carries at most a quarter of the reward's observed variance; a **population of opponents** per rollout table drawn from §11's zero-token policies plus the Bayesian agent, so the winning strategy has to close good deals against diverse counterparties rather than out-refuse a mirror of itself; among-IR Nash welfare as the primary reward term with deal rate and unconditional welfare demoted to evaluation-only gates; and an honest 100-step schedule (≈4.5 GPU-days, ≈$650–750) instead of v1's 24 steps. Two of its gates are new and both are named rather than discovered after the fact: the ultimatum-transfer probe must move *toward* the fair split (v1's moved away), and the gain must survive a story-transfer test with at least half its abstract magnitude.

First-and-a-half, the eval that decides whether any of this is about language models. The v2 proposal's headline evaluation is the question this entire program raises and has never tested. Every positive result here belongs to a *decision rule*: the computable agents are fairer,

the two-line refusal policy delivers the anchor's whole fairness contribution (§11.2), and handing the model a fair candidate deal works where training it to want one does not. If fairness in this environment is only reachable through machine-readable score sheets, then an algorithmic agent plus a parser dominates and the language model is decoration. The **prose-only condition** is the discriminating test: a scaffold variant in which a seat's private information arrives *only* as natural-language prose — information-complete, deterministically templated from the numeric sheet so there is no model-generated leakage channel — with no numeric block at all. Computable policies cannot play it by construction, and the engine still scores it exactly. The claim is preregistered in both directions: if fairness-trained LLMs retain their gains there, LLM+RL buys something no algorithmic agent can; if the gains vanish, the honest verdict is that the algorithmic agent is the right tool and this program's positive findings are findings about decision rules. It needs a prose-sheet renderer, a solo-control tripwire on the leakage methodology of §8.1, and a comprehension pre-check (an untrained model's solo valuation accuracy on prose against numeric sheets, since prose-comprehension noise bounds every claim in the condition).

Second, the wider holdout from day one. At least three held-out instances per condition, with instance-cluster bootstrapping. The pilot's entire arc is the argument.

Third, training-data diversity. 326 pairs from one backbone on three payoff games is thin in a way the style confound made worse: with so few pairs, the easiest available gradient direction dominates. Multi-instance and multi-backbone training data would at least make the style shortcut less overwhelmingly attractive.

Fourth, measurement work that stands on its own regardless.

- **Rotate seats** in mixed tables so the extraction question gets a causal answer. The completed companion campaign does this and, re-matched against clean baselines, finds a resolved same-seat *capture* advantage on all five model slots while its agreement effect largely disappears (§14.1). The open question it leaves is the one it answered in the wrong direction first: whether a computable seat helps a weak table *close* at all, which now needs asking on cells that were never contaminated.
- **Widen the instance bank** past three clusters, which improves every interval in §5 at once.
- **Validate the Transformers upgrade** in a migration branch with full suites, prompt hashes, and generation canaries, then re-run the two substitute slots as the real Gemma-4 and Ministral-3 cells. Any panel intended for cross-version comparison must be regenerated, not mixed.
- **Re-run the two preset cells**, the last contaminated numbers left in the paper (§5.7). Nothing depends on them, which is exactly why they are easy to leave rotting; the divide-the-dollar column currently contains one cell that measured nothing at all.
- **Separate the framing effect from the closing effect with more seeds** (§9.5). The thinking $\times$ framing 2×2 resolves the interaction on deal rate and unconditional welfare but not on any deal-conditional fairness measure, because those need both arms to close and run over 31–46 pairs against 60. It also leaves one question open that the design cannot answer: the framed cells split *better* while closing less under both thinking conditions, which is the opposite of the open-weight ladder's behaviour and is currently a consistent direction rather than a resolved effect. At roughly \$0.35 per episode this is the cheapest remaining frontier question.

- **Rotate seats in the frontier rational-seat cell** the way the companion does locally, so the anchor result (§9.3, §11.5) gets the same seat-balanced treatment as its open-weight twin rather than resting on seat 0. This is the direct comparison that would establish the counterparty-population reading of §12.7 as causal, and it is now more urgent than when it was first written, because the frontier endpoint of that reading has moved once already.
- **Finish the two Bayes-anchor seat cells** (seats 2 and 4 on the screened bank, Slurm 16453662 / 16453663, queued at the time of writing). They are the only thing standing between §11.2's parity result — the arc's headline — and a seat-robust version of it, and they cost the same as the four policy cells already run. Extending the sweep to the Sonnet population is the same argument at API prices.
- **Test the predicted attenuation of the calibrated maximizer against a stronger population** (§11.3). The campaign has run and resolved — truthful calibration made the agent concede and its table fairer on every axis — but on *one* counterparty model, five Qwen3-8B seats with thinking off. The arc's own acceptance curves make a sharp, falsifiable prediction about where it should stop working: the effect is the correction of an *over-confidence* error, and Qwen3-32B (below-threshold acceptance 0.133, negative Brier gain for the fitted curve) and `claude-sonnet-5` (0.000) leave almost nothing to correct, so the fairness movement should attenuate toward zero against them. A 32B rung is a local GPU cell; a Sonnet rung is API-priced but small, and either one is a cheap chance to falsify the mechanism rather than to re-confirm it. Two companions worth running in the same wave: the **six-seat sweep** the campaign dropped for pod time, which converts the outcome result from a seat-0 magnitude into a seat-robust one; and a **re-harvest and re-fit of the acceptance curves against the maximizer's own transcripts**, since a policy that pushes opponents toward $z = 0$ operates exactly where the observational curves are steepest and least sampled, which makes the curves it plans against a moving target.
- **Run the length-matched nonsense-label control** for the framing arm (§8), which decides whether the active ingredient is realistic framing or verbose labels. It is the single highest-value follow-up there, and cheap: it reuses the existing framing machinery with a third label pool. Then run the `festival` skin as a second-domain replication, and if both hold, one more model slot for model-dependence — cheap-talk harm was model-specific (§5.5), and framing harm may be too.
- **A dropped question, recorded as dropped.** The paper previously proposed explaining why the trained arms were specifically worse under private information (-0.183^{*} deal rate) — a plausible story about oracle-styled confident deliberation being harmful when the model cannot see what the oracle saw. **There is no asymmetry to explain**: on clean cells that contrast is $+0.083^{*}$. This is worth keeping visible as a cautionary case, because the proposed mechanism was specific, plausible, and connected to a real design decision (poison P2, §7.1), and it was an explanation for an artifact.
- **P5, the interpretability tie-in**, remains built but unused: activation capture is plumbed through the episode pool at divergence points. Given §6 — the divergence is present in the prompt before any reasoning — the natural question is whether the flagged decision is already readable from the residual stream at the start of the turn.

14.1 Companion campaign: one rational agent versus five LLMs

A sibling session implemented and launched the missing reverse asymmetry: a `reverse_mixed` table with *one* Bayesian-rational seat against five copies of the same LLM, with the rational seat rotated through all six positions — which removes both the fixed-score-sheet and the proposer-position confound that §5.6 could not escape. It also repaired a real defect in the Bayesian policy: the policy received a *pooled* list of proposals, so its nominally per-opponent posterior replayed identical evidence for every opponent; the public state now carries proposer identity and one posterior is kept per actual proposer. The design is 30 immutable GPU cells / 3,600 episodes over P2's six committed instances, two arms, and ten seeds, plus a belief-recovery analysis that scores the rational agent's final private-information posterior against each LLM opponent's hidden score sheet (issue-priority pairwise accuracy, posterior mass on the true top issue, Spearman correlation of inferred weights against true issue ranges, threshold absolute error, and acceptance Brier score, each reported relative to the exact prior).

Estimands and gates. Every interpersonal outcome is normalized surplus capture z_i, never raw points. The descriptive within-table contrast is the rational seat minus the mean of the five LLM seats; the preferred causal contrast replaces one LLM by Bayes on the *same* score sheet and matches the canonical all-LLM P2 episode on model, committed instance, protocol arm, and rollout seed. No-deal gives every seat zero; a separate reached-deal contrast isolates division conditional on closure. The final analyzer independently requires the exact five-model × six-seat grid, the frozen model-id substitutions, canonical scorable prompting, both arms, seeds 0–9, 120 episodes per cell, and a same-seat P2 match for all 3,600 rows. Means and 95% intervals resample the three underlying committed-instance seeds as whole clusters. Belief gains are posterior minus exact prior (or prior error minus posterior error), and only an opponent that made a valid formal proposal contributes a row; these are in-model posterior-fit diagnostics, not held-out prediction.

Contamination status: clean cells, re-matched against clean baselines. All 30 `reverse_mixed` cells audit at 0.000 fabricated turns — like `mixed`, they route through the async pool, which is the observation that localized the bug. But the design's headline is a *paired* contrast against the matched canonical all-LLM P2 episode, and those baselines were 40–55% fabricated for four of the five slots. Every paired column below has therefore been recomputed against the clean `_b2` baselines, and the result splits this section cleanly in two.

Result: the closure-repair finding was largely artifact; the surplus-capture finding survives. The gate passes at exactly 3,600/3,600 episodes, 30 cells, five preserved exclusions, zero errors or warnings, on the reverse cells and on the re-matched baselines alike.

Withdrawn: "adding a Bayes seat sharply repairs weak all-LLM closure." Against clean baselines the Δ deal-rate column collapses (Table 20). The Gemma slot does not merely lose its +0.428 — it **resolves negative** at −0.097 [−0.138, −0.071], so the rational replacement makes that table close *less* often. Ministral holds at +0.094, about a fifth of its published +0.461. Both Qwen slots lose significance, and OLMo — whose baseline was the one already-clean cell — barely moves (−0.063 → −0.046), a third independent confirmation of the de-contamination reading in a place nobody designed as a control. The published effect was substantially measuring the fact that the harness was eating the baselines' turns: weak all-LLM closure was partly not the models' doing.

Survives: the same-seat capture advantage, on all five slots. Replacing one LLM with Bayes raises that *same seat's* normalized capture by +0.147 to +0.444, every interval excluding zero, shrunk 20–40% from the published values. The *within-table* reached-deal contrast is bit-identical to

the published numbers, because both of its sides come from the same clean episode and it never touches a baseline: the rational seat captures 22.7–43.0 percentage points more of its available surplus than the mean LLM at the same table. So the causal, scale-corrected policy advantage the fixed-seat P2 design could not identify is real — and it is an advantage in *how surplus divides*, not in whether agreement happens.

And the fairness half of it generalizes to every slot, including the frontier one. Figure 13 states the swap as a differential on the fairness panel rather than on capture: replacing one seat with a computable policy moves the *table* toward the normative benchmarks on all six slots at once — distance to the Nash solution falls by −0.035 to −0.166, the worst-off party's capture rises by +0.033 to +0.162, and the individual-rationality violation rate falls by −0.017 to −0.132. The Gemma slot is the mechanism in its clearest form: alone it closes 99.2% of episodes while violating individual rationality in 94.2% of them, and a single rational neighbour blocks those below-reservation bandwagons — which is why it posts both the largest worst-off gain (+0.162) and the one resolved *negative* deal-rate effect (−0.097). One disciplined seat buys fairness by refusing the deals that should not have closed.

The split is itself the informative part, and it is the same split the whole revision keeps producing: closure-dependent quantities move under de-contamination and division-of-surplus quantities hold. That is exactly what a defect which deletes turns should do — a deleted turn is a chance to close that never happened, while conditional on a deal existing, how its surplus was divided is untouched. The pattern recurs in the cross-model ladder (§5.4: closure and efficiency move, the distributional deficit does not), in the scaffold contrast (§5.5: the closure penalty grows while every quality advantage evaporates), and here. It is a useful heuristic for triaging any harness defect of this class before re-running anything.

Table 20: Reverse mixed table, 720 episodes per model, re-matched against clean baselines. Effects are rational-table minus the matched canonical all-LLM P2 episode; brackets are 95% instance-cluster bootstrap intervals. Capture is on the replaced seat's own available-surplus scale. **Published** columns are the original values, matched against the contaminated all-LLM cells, and are shown so the direction of the correction is visible. The last column is hidden issue-order accuracy after proposal updates, with posterior-minus-prior change in parentheses; the re-match does not touch it, since it is measured inside the rational seat's own posterior.

Model slot	**Δ deal rate**		**Δ same-seat capture**		**Issue order**
	pub.	**clean**	pub.	**clean**	**(gain)**
Gemma-4 slot	+0.428	**−0.097** [−0.138, −0.071]	+0.560	**+0.444** [0.429, 0.456]	0.459 (−0.004)
Ministral-3 slot	+0.461	**+0.094** [0.046, 0.121]	+0.319	**+0.210** [0.186, 0.227]	0.460 (−0.007)
OLMo-3-7B *(clean baseline)*	−0.063	−0.046 [−0.150, 0.037]	+0.156	**+0.147** [0.083, 0.209]	0.460 (−0.003)
Qwen3-4B	+0.194	+0.053 [−0.017, 0.108]	+0.243	**+0.175** [0.137, 0.199]	0.448 (−0.007)
Qwen3-8B	+0.269	−0.006 [−0.158, 0.079]	+0.303	**+0.173** [0.090, 0.254]	0.465 (−0.007)

The Bayesian-learning claim fails. Across 6,816 PRIVATE opponent records, issue-order accuracy remains 0.448–0.465 and moves *backward* by 0.003–0.007 from the prior; probability on the true top issue also stays near its prior; reservation-threshold error improves by exactly zero. Acceptance Brier scores do improve, but on the same proposals used for updating, so they are posterior-fit rather than held-out evidence. The outcome result therefore supports the structured rational policy as a whole, not the hypothesis that its implemented Bayesian learner recovered opponents' hidden score sheets.

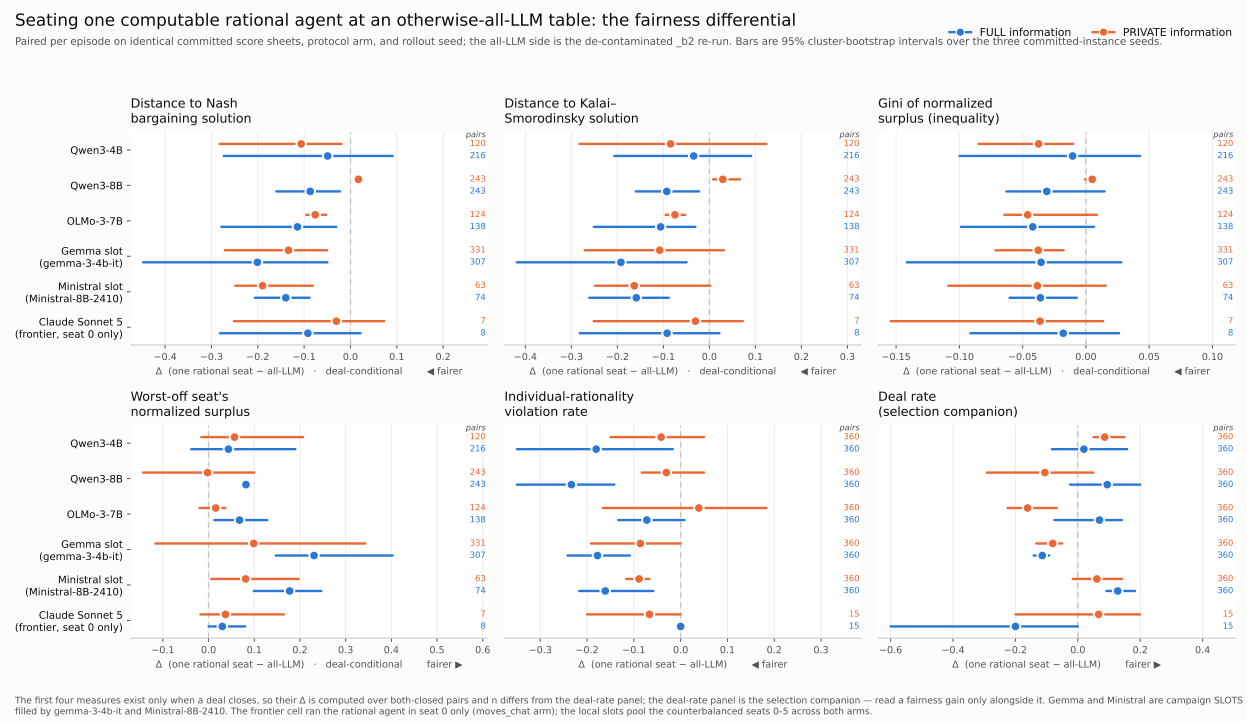

Figure 13: What one computable seat does to the table's fairness, on every slot. Each estimate is the mixed-minus-all-LLM difference paired *per episode* on identical score sheets, protocol arm and rollout seed, against the de-contaminated `_b2` baselines; bars are 95% cluster-bootstrap intervals resampling the three committed instance clusters. **The effect generalizes**: distance to the Nash solution falls on all six slots (−0.035 to −0.166), the worst-off party's capture rises (+0.033 to +0.162) and the individual-rationality violation rate falls (−0.017 to −0.132) — and `claude-sonnet-5` shows the *smallest* movement of the six, which is the frontier-population reading of §9.3 arriving independently here. The deal-rate companion panel moves in *both* directions across slots (−0.097 to +0.094), so the fairness gains are not a selection effect of a table that simply closes less. Three cautions travel with the figure. The frontier row is thin — 15 deal-conditional pairs against 137–638 for the open-weight slots — so it is the least resolved row and should not be read at the same confidence as the others. Distance to the Nash and to the Kalai–Smorodinsky solution coincide in four of the six games, so those two panels are near-redundant rather than independent confirmation. And every deal-conditional panel is computed over reached deals only, which is why deal rate is plotted beside them. Source: `fig_fairness_differential.py`; rows and provenance in `.claude/products/fairness_differential/`.

A Reproduction commands

All commands run from the repository root unless stated, with `LARGE_ARTIFACTS_DIR` set. Setting `UV_LINK_MODE=copy` is mandatory on this NFS setup.

P1 — environment sanity.

```
UV_LINK_MODE=copy uv run --no-sync python -m pytest -q                # in libs/interlens
UV_LINK_MODE=copy uv run --no-sync python -m pytest experiments/rational_agents/tests/ -q
UV_LINK_MODE=copy uv run --no-sync python experiments/rational_agents/run.py \
  --run-name integration_review_smoke2 --table all_rational \
  --arm moves_chat moves_only --seeds 0 1 \
  --instances experiments/rational_agents/instances --oracles threshold acceptance
UV_LINK_MODE=copy uv run --no-sync python experiments/rational_agents/annotate.py \
  --episodes $LARGE_ARTIFACTS_DIR/ii_mats/rational_agents/integration_review_smoke2/episodes \
```

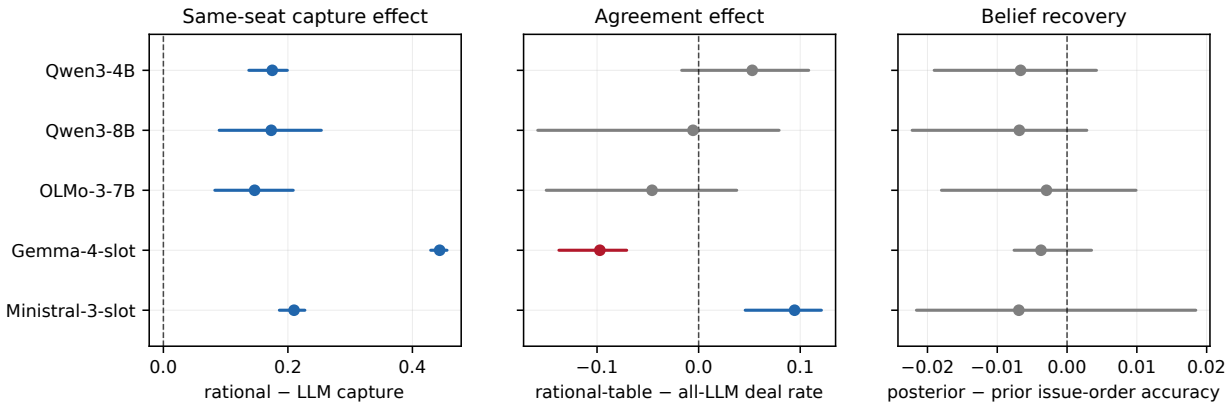

Figure 14: Seat-balanced reverse-mixed result, re-matched against clean baselines. The first two panels compare the table with one rational replacement against its matched all-LLM baseline; the last asks whether PRIVATE-game proposals improve hidden issue-order inference over the exact prior. Bars are 95% intervals over three committed-instance clusters. **Capture effects resolve on all five slots; the agreement effects mostly do not, and the Gemma slot resolves in the *opposite* direction to its published value**; hidden-priority recovery does not resolve at all.

```
  --instances $LARGE_ARTIFACTS_DIR/ii_mats/rational_agents/integration_review_smoke2/instances \
  --out <out>
```

P2 — the campaign. Submitted under a single-submitter protocol:

```
cd experiments/rational_agents
python launch_p2.py --submit        # 24 sbatch jobs / 1,770 episodes per docs/p2-run-plan.md
```

P3 — validity gate and atlas. From `experiments/rational_agents/`:

```
jq . p2_selected_manifest.json      # the immutable selection + excluded-run ledger

uv run python validate_campaign.py \
  --manifest p2_selected_manifest.json \
  --out /nlp/scr/siddharth/ii_mats/rational_agents/p3_campaign_quality.json \
  --stage complete                # nonzero exit on a partial or contaminated selection

sbatch p3_equilibrium_backfill.sbatch   # replay supported equilibrium oracles (preset cells)
sbatch p3_scale_corrected.sbatch        # the atlas v2 publication job (preserves v0/v1)

# already-gated equivalent of the v2 synthesis:
uv run python -m analysis.campaign \
  --manifest p2_selected_manifest.json \
  --out /juice2/u/siddharth/ii_mats/.claude/products/p3_divergence_atlas_v2 \
  --bootstrap-samples 5000 --annotations-dir annotations_v1
```

Reverse-mixed companion — replacements, selection, and analysis.

```
cd experiments/rational_agents
uv run --no-sync python launch_reverse_mixed.py --phase full \
  --out p3_reverse_replacement_sbatch --run-prefix p3reverse_r1 \
  --cells Qwen3-4B:0 OLMo-3-7B:4 OLMo-3-7B:5 \
```

```
          Gemma-4-slot:0 Gemma-4-slot:2 \
  --no-requeue --submit
uv run --no-sync python build_reverse_selected_manifest.py \
  --original p3_reverse_sbatch/selected_manifest.json \
  --replacements p3_reverse_replacement_sbatch/selected_manifest.json \
  --out p3_reverse_final_selected_manifest.json
sbatch --parsable --no-requeue \
  --dependency=afterok:16368515:16368516:16368517:16368518:16368519 \
  p3_reverse_analysis.sbatch
```

Goal 2 — CoT-step localization.

```
sbatch p4_cot_localize.sbatch       # one A6000; 150 turns, 619 probes incl. the k=0 control
# re-score without a GPU in ~7s from stored transcripts:
uv run python p4_cot_localize.py --reuse-transcripts ...
```

P4 — dataset, training, evaluation.

```
cd experiments/rational_agents
UV_LINK_MODE=copy uv run python p4_dataset.py \
  --annotations-name annotations_v1 \
  --out-dir /nlp/scr/siddharth/ii_mats/rational_agents/p4_dataset_v2

sbatch p4_train.sbatch              # the DPO / action_only / SFT arms
python launch_p4eval.py --submit    # the held-out evaluation fleet (manifest in p4eval_sbatch/)
```

The de-contamination pass (§3.6) — audit, re-baseline, clean atlas. This is the sequence to re-run if the campaign is ever extended; `--sequential --concurrency 1` is mandatory on any `all_llm` or `scaffolded` cell produced before `interlens b314dc1`.

```
cd experiments/rational_agents

# 1. cause-classifying audit over every stored run (this is what overturned the truncation hypothesis)
uv run --no-sync python audit_empty_turns.py --glob 'p2*' --json /tmp/audit_p2_before.json
uv run --no-sync python audit_empty_turns.py --glob 'p4eval_xgame*'     # the asymmetric fleet
uv run --no-sync python audit_empty_turns.py --glob 'framing_pilot*' --by-arm

# 2. the pre-existing accidental A/B that proves the mechanism, with no new compute
uv run --no-sync python compare_rebaseline_ladder.py \
  --pairs p2_Olmo-3-7B_all_llm=p2r3_Olmo-3-7B_all_llm

# 3. generate + submit the 8 clean-engine cells (dry-run first, without --write)
uv run --no-sync python launch_p2_rebaseline.py --write --submit
uv run --no-sync python launch_p4eval_rebaseline.py --write --submit   # the 10 cross-game cells

# 4. old-vs-new ladder, with the fabrication rate carried beside every row
uv run --no-sync python compare_rebaseline_ladder.py --markdown --json /tmp/ladder.json

# 5. the clean cross-model atlas used for every table in section 5 (7 cells, 840 episodes)
python _runpod32b_logs/build_b2_manifest.py \
  --extra-cell "p2_Qwen3-32B_all_llm:all_llm:Qwen3-32B:Qwen/Qwen3-32B" \
  --out _runpod32b_logs/b2plus32b_manifest.json
python validate_campaign.py --manifest _runpod32b_logs/b2plus32b_manifest.json \
  --stage complete --out _runpod32b_logs/b2plus32b_quality.json      # -> valid=True 840/840
python -m analysis.campaign --manifest _runpod32b_logs/b2plus32b_manifest.json \
```

```
  --out /nlp/scr/siddharth/ii_mats/rational_agents/atlas_b2plus32b \
  --bootstrap-samples 5000 --annotations-dir annotations_v1

# 5b. the same atlas extended with the two clean scaffolded cells (9 cells, 1,080 episodes), which is
# what produces the "<model>: scaffolded vs canonical" comparison rows the scaffold panel reads.
# The scaffolded _b2 runs ship annotations/ but no annotations_v1/, so byte-copy it first, exactly as
# the rebaseline lane did for the other _b2 cells (each copy carries a PROVENANCE.txt saying so).
python validate_campaign.py --manifest _runpod32b_logs/b2plus32b_scaffold_manifest.json \
  --stage complete --out _runpod32b_logs/b2plus32b_scaffold_quality.json   # -> valid=True 1080/1080
python -m analysis.campaign --manifest _runpod32b_logs/b2plus32b_scaffold_manifest.json \
  --out /nlp/scr/siddharth/ii_mats/rational_agents/atlas_b2plus32b_scaffold \
  --bootstrap-samples 5000 --annotations-dir annotations_v1

# 6. the cross-game verdict table, and the paired table this writeup's figures parse
uv run --no-sync python compare_xgame_rebaseline.py --markdown --json /tmp/xgame.json
uv run --no-sync python p4_xgame_paired_table.py \
  --out /nlp/scr/siddharth/ii_mats/rational_agents/p4_xgame_paired_clean_b2.txt
uv run --no-sync python p4_xgame_paired_table.py --suffix '' --out /tmp/xgame_orig.txt  # validation
```

The narrative-framing arm (§8). Four rollout cells plus the two rational-invariance cells, then one paired analysis that emits every number in Table 8 and both controls. The abstract control differs from the treatment only in `--framing` and `--cell`; the clean cells shown here are the `_b2` re-runs, which add `--sequential --concurrency 1`.

```
R=/nlp/scr/siddharth/ii_mats/rational_agents
# treatment (the control swaps in --framing abstract --cell framing_abstract_b2)
uv run --no-sync python run.py --run-name framing_pilot_qwen3_8b_datacenter_b2 --table all_llm \
  --instances instances --arm moves_chat --seeds 0 1 2 3 4 5 6 7 8 9 \
  --oracles threshold acceptance bestresponse --models Qwen/Qwen3-8B --model-thinking off \
  --framing datacenter --cell framing_datacenter_b2 --sequential --concurrency 1
# solo leakage tripwire (--table solo) and rational invariance (--table all_rational) run the same way

# paired analysis: treatment effects + both controls in one pass
uv run --no-sync python analyze_framing.py \
  --framed-run   $R/framing_pilot_qwen3_8b_datacenter_b2 \
  --abstract-run $R/framing_pilot_qwen3_8b_abstract_b2 \
  --solo-framed-run     $R/framing_pilot_qwen3_8b_solo_datacenter \
  --solo-abstract-run   $R/framing_pilot_qwen3_8b_solo_abstract \
  --rational-framed-run $R/framing_pilot_rational_datacenter \
  --rational-abstract-run $R/framing_pilot_rational_abstract \
  --out $R/framing_pilot_qwen3_8b_datacenter_b2/analysis_framing
```

The frontier-API campaign (§9). Hosted cells run from the *login* host, not under Slurm — compute nodes have no outbound internet. Every cell carries a hard dollar cap, and the local comparators are re-annotated through the identical pipeline so no row in Table 10 comes from a different code path.

```
# 1. calibrate on ONE episode, read the measured $/episode off the ledger, then run the cells
source set-env.sh                                  # LARGE_ARTIFACTS_DIR + ANTHROPIC_API_KEY
bash experiments/rational_agents/launch_api_behavior.sh calibrate
SEEDS="0 1 2 3 4" BUDGET_USD=90 bash experiments/rational_agents/launch_api_behavior.sh full

# 2. one cell, exactly as recorded in its run.log INVOCATION line (canonical sonnet arm)
python run.py --run-name apibehav_sonnet5_thinkoff_5seed --table all_llm \
  --models anthropic:claude-sonnet-5 --arm moves_chat \
```

```
  --instances experiments/rational_agents/instances --seeds 0 1 2 3 4 \
  --oracles threshold acceptance bestresponse --model-thinking off --concurrency 4 \
  --budget-usd 25 --est-usd-per-episode 0.6 --cell apibehav_sonnet5_thinkoff
# the other cells differ only in these flags:
#   --arm moves_only                                    (no cheap talk)
#   --scaffold scaffolded --focal-scaffold-seat 0       (arm B, scaffold on seat 0 only)
#   --table reverse_mixed --rational-seat 0             (arm C, rational policy in seat 0)
#   --seeds 0 1 2 3 4 5 6 7 8 9 --model-thinking on --api-effort high \
#     --api-turn-token-floor 16384                     (the thinking-on cells; add --framing datacenter
#     for the framed half. --model-thinking on sends an explicit adaptive request with summarized
#     reasoning persisted; the manifest's api_request_config records what was actually sent.)

# 3. annotate + report every cell, and re-annotate the two local comparators through the SAME
#    pipeline so no row of the frontier table comes from a different code path
uv run --no-sync python annotate.py --episodes $R/apibehav_sonnet5_thinkoff_5seed/episodes \
  --instances $R/apibehav_sonnet5_thinkoff_5seed/instances \
  --out $R/apibehav_sonnet5_thinkoff_5seed/annotations_v1 \
  --oracles threshold acceptance bestresponse
uv run --no-sync python -m analysis.report \
  --episodes $R/apibehav_sonnet5_thinkoff_5seed/episodes \
  --instances $R/apibehav_sonnet5_thinkoff_5seed/instances \
  --annotations $R/apibehav_sonnet5_thinkoff_5seed/annotations_v1 \
  --out $R/apibehav_sonnet5_thinkoff_5seed/analysis --run-name apibehav_sonnet5_thinkoff
uv run --no-sync python -m analysis.report --episodes $R/p2_all_rational/episodes \
  --instances $R/p2_all_rational/instances --out $R/p2_all_rational/analysis_apibehav \
  --run-name all_rational
```

The policy-decomposition arc (§11).

```
cd experiments/rational_agents

# 0. build and FREEZE the screened 24-game bank before any cell runs (4 screens; 24 of 137 admitted),
#    plus the 12-game all-full-info sibling the calibrated ladder needs for its 24 full-info clusters
uv run --no-sync python build_trivial_bank.py            # -> instances_trivial_v1/
uv run --no-sync python build_trivial_bank.py --full-info-only --n 12
                                                          # -> instances_trivial_v1_full12/

# 1. the local cells: five policies x Qwen3-8B, plus the accept_frac ladder and the seat sweep.
#    Zero tokens are generated by seat 0 in every one of these, so they are cheap.
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch passive-gate
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch greedy-anchor
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch greedy-holdout
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch greedy-holdout-declared
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch demand-90-declared
# within-bank references -- NEVER compare across banks
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch --ref all-llm
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch --ref bayes-anchor
sbatch trivial_baselines_sbatch/trivial_policy_cell.sbatch --ref all-rational

# 2. the claude-sonnet-5 half (login host, needs internet + a budget cap set to 2-3x the
#    single-instance calibration, which is a LOWER BOUND and underestimated by ~2x here)
source set-env.sh && ./trivial_baselines_sbatch/launch_trivial_sonnet.sh

# 3. the N-way panels, unrestricted and under the per-contrast censoring restriction
sbatch trivial_baselines_sbatch/analyze_trivial.sbatch qwen
sbatch trivial_baselines_sbatch/analyze_trivial.sbatch sonnet
uv run --no-sync python analyze_trivial_policies.py --matched-censoring \
  --cells trivialv1_passive_gate_qwen3_8b ref_bayes_qwen3_8b_trivialv1
```

```
# 4. the acceptance curves: harvest every clean stored campaign, fit, validate leave-one-GAME-out
uv run --no-sync python -m self_benefit.harvest_responses \
    --out /nlp/scr/siddharth/ii_mats/rational_agents/accept_curves/events.jsonl
uv run --no-sync python -m self_benefit.fit_acceptance \
    --events /nlp/scr/siddharth/ii_mats/rational_agents/accept_curves/events.jsonl \
    --out self_benefit/acceptance_curves.json --fig self_benefit/figs/acceptance_curves.png

# 5. the calibrated maximizer: equivalence control + the decision-function measurement, CPU only.
#    The 'step' rung must reproduce Bayes on EVERY turn; the effect is read off the same replay.
uv run --no-sync python gate_calibrated_identity.py \
  --runs /nlp/scr/siddharth/ii_mats/rational_agents/p3reverse_Qwen3-8B_rseat{0,1,2,3,4,5} \
  --curves self_benefit/acceptance_curves.json --opponent-model Qwen/Qwen3-8B --thinking off \
  --out /nlp/scr/siddharth/ii_mats/rational_agents/_calmax_gate/gate_allseats.json

uv run --no-sync python -m pytest tests/test_acceptance_curves.py \
  ../../libs/interlens/tests/test_negotiation_trivial_policies.py \
  ../../libs/interlens/tests/test_negotiation_calibrated.py -q

# 6. the closed-loop calibrated ladder: 3 rungs x 2 banks = 864 episodes, ~12.2 GPU-h on one B200.
#    STOP_POD_ON_EXIT=0 because the pod is shared: the worker finalizes and announces instead of
#    stopping, and it announces only on rc==0, via a file another lane can poll.
STOP_POD_ON_EXIT=0 ./runpod_calibrated_worker.sh POD_ID calmax_v2 \
  instances_trivial_v1,instances_trivial_v1_full12 \
  --rungs step calibrated calibrated-vote --curves self_benefit/acceptance_curves.json \
  --opponent-model Qwen/Qwen3-8B --model-thinking off \
  --seeds 0 1 2 3 4 5 6 7 --sweep-seats 0 --concurrency 96

# 7. headline / censoring sensitivity / placebo.  S = the scratch campaign root.
#    --info full is LOAD-BEARING: the private half is inert for this treatment, and pooling it
#    dilutes the measured turn-change rate by 2.4x (46.5% full-info -> 19.5% pooled).
uv run --no-sync python analyze_trivial_policies.py \
  --cell step=$S/calmax_v2_step,$S/calmax_v2_instances_trivial_v1_full12_step \
  --cell calibrated=$S/calmax_v2_calibrated,$S/calmax_v2_instances_trivial_v1_full12_calibrated \
  --cell calibrated-vote=$S/calmax_v2_calibrated-vote,$S/calmax_v2_instances_trivial_v1_full12_calibrated-vote \
  --baseline step --info full --out $S/_calmax_analysis --tag calmax_full
#   ... same again with --info full --matched-censoring --tag calmax_full_matched
#   ... and with --info private (shared bank only) --tag calmax_private   # the placebo arm

uv run --no-sync python fig_calibrated_ladder.py \
  --report $S/_calmax_analysis/calmax_full.json \
  --out writeup/figs/calibrated_ladder.pdf \
  --title "Calibrated-maximizer ladder vs the Bayes-equivalent control (full information)"
```

Two reproduction hazards from this arc are worth repeating because neither raises an error. **A per-model artifact lookup that falls back silently is a wrong number with no exception attached**: the acceptance-curve artifact keys responders on the canonical short name plus thinking mode (`Qwen3-8B:thinking-off`), and a consumer passing the Hugging Face id gets the `pooled-llm` fit over all models (n=177,965) instead of the model's own (n=32,225) — so the fallback is opt-in and every launch prints the resolved key and its sample size. And **an estimator imported from a sibling analysis module closes over that module's cluster function**, silently ignoring the caller's: that bug counted the full- and private-information variants of one game as two independent clusters and narrowed every interval in the mixed-thinking cell until two regression tests pinned it. Clustering in this project is *game*-level, always.

Verification suites, and this writeup's figures.

```
cd experiments/rational_agents && uv run pytest -q tests
(cd libs/interlens && uv run pytest -q)
(cd libs/interlens && HF_HOME=/scr/siddharth/.caches/huggingface \
   HF_HUB_OFFLINE=1 TRANSFORMERS_OFFLINE=1 uv run pytest -q tests/test_family_flags.py -m slow)

# figures. Three invocations, because the panels no longer share one source:
#  (1) the clean scaffold-inclusive atlas drives goal-1 / deal-atlas / cheap-talk+scaffold, the clean
#      cross-game paired table drives the P4 and regret panels, and the re-matched companion drives
#      the reverse-mixed panel;
#  (2) the mixed panel comes from the (clean, unchanged) v2 atlas -- regenerated separately, and its
#      byte-identity to the pre-revision figure is the numerical control on the whole revision;
#  (3) the CoT panel needs neither atlas;
UV_LINK_MODE=copy uv run --no-sync python experiments/rational_agents/writeup/make_figures.py \
  --atlas /nlp/scr/siddharth/ii_mats/rational_agents/atlas_b2plus32b_scaffold \
  --xgame-paired /nlp/scr/siddharth/ii_mats/rational_agents/p4_xgame_paired_clean_b2.txt \
  --reverse-dir /nlp/scr/siddharth/ii_mats/rational_agents/p3reverse_rematch_b2 \
  --only goal1 atlas chat p4 regret reverse cot
UV_LINK_MODE=copy uv run --no-sync python experiments/rational_agents/writeup/make_figures.py \
  --atlas /juice2/u/siddharth/ii_mats/.claude/products/p3_divergence_atlas_v2 --only mixed
#  (4) the framing and frontier panels read their own campaigns' analysis outputs and need no atlas.
UV_LINK_MODE=copy uv run --no-sync python experiments/rational_agents/writeup/make_figures.py \
  --only framing frontier
cd experiments/rational_agents/writeup && latexmk -pdf rational_agents.tex
```

Gate state at publication: **336 interlens tests** (1 skipped, 23 deselected), **12 real-chat-template cases**, and **244 experiment tests** (1 skipped, 1 intentionally deselected); the companion reverse-mixed, artifact-index, and scale-contract additions take the experiment suite to 270 and interlens to 346.

B Artifact and run ledger

B.1 Code and documents (in-repo)

- Design: `experiments/rational_agents/DESIGN.md` — goals, formal game, protocol, oracle stack, phases, frozen interface contracts.
- Status: `PROGRESS.md`. Run plan: `docs/p2-run-plan.md`.
- Research notes: 0000 scaffold + P1 sanity · 0001 P2 launch and triage · 0002 P2/P3 validity-gated results (*carries three errata*: localization figures, engine fabrication, and the OLMo-row correction) · 0003 P4 pilot arc · 0004 P4 verdict (retracted positive; its cross-game negative is in turn withdrawn by 0016) · 0005 reverse-mixed result + belief-recovery negative.
- Revision notes (the de-contamination arc): 0014 framing arm (*errata*: its deal-rate headline does not survive de-contamination; its qualitative verdict does) · 0015 contamination discovery, detector taxonomy, and what is safe to report · 0016 root cause, the clean re-baseline, and the cross-game reversal · 0017 Qwen3-32B scale cell + the addendum that retracts two of its three headlines against clean baselines · 0018 frontier-API behavioural pilot.
- The policy-decomposition arc (§11): hub `self_benefit/README.md` (hypotheses H1–H6 fixed before any run, with a verdict row per workstream) · 0022 trivial-policy baselines, both populations, and the note that flags the 0018 reversal · `self_benefit/acceptance_curves.md` (the measured opponent model) · 0024 the behaviourally-calibrated maximizer: preregistration, the pre-campaign replay that contradicted it, and the 864-episode campaign

that falsified H2-strong in both directions · 0026 the mixed-thinking cell and its differential-censoring argument. Code: `build_trivial_bank.py`, `analyze_trivial_policies.py`, `launch_calibrated_ladder.py`, `gate_calibrated_identity.py`, `runpod_calibrated_worker.sh`, `fig_calibrated_ladder.py`, `wandb_campaign_monitor.py`, `analyze_mixed_thinking.py`, `self_benefit/{response_events,harvest_responses,logistic,fit_acceptance,acceptance_curves}.py`, and in the library `arena/negotiation/{strategies,calibrated}.py` plus the `Pass` action.

- Successor preregistration: `proposals/2026-08-03-fairness-grpo-v2.md` (clipped reward, population opponents, the prose-only transfer eval) — proposed, unfunded, no compute spent.
- Revision plan for this document, with per-item status: `writeup/REVISION_PLAN.md`.
- Literature reports: `docs/lit/{benchmarks-scorable-games,rational-oracles,divergence-training,interlens-map}.md`.
- Campaign selection: `p2_selected_manifest.json` (24 cells, 1,770 episodes, 16 excluded runs with reasons).
- Runners and analysis: `run.py`, `annotate.py`, `validate_campaign.py`, `launch_p2.py`, `launch_p4eval.py`, `launch_reverse_mixed.py`, `prompts.py`, `seating.py`, `package.py`, `analysis/`, `tests/`.
- P4 code: `p4_pairs.py`, `p4_dataset.py`, `p4_dataset_diff.py`, `p4_reannotate.py`, `p4_train_dpo.py`, `p4_cot_localize.py`, `p4_cot_value.py`.
- De-contamination code (§3.6): `audit_empty_turns.py` (cause-classifying audit over stored episodes), `launch_p2_rebaseline.py` and `launch_p4eval_rebaseline.py` (the 8 + 10 clean-engine cells, reusing the original launchers' cell machinery so invocations stay single-sourced), `compare_rebaseline_ladder.py` (old-vs-new ladder with the fabrication rate beside every row), `compare_xgame_rebaseline.py` (the cross-game verdict table, gated on reproducing the published numbers on the contaminated originals), `p4_xgame_paired_table.py` (the paired table this writeup's figures parse, including per-cell regret), and `_runpod32b_logs/build_b2_manifest.py` (the clean-cell manifest builder, which matches cells by *model id* and additionally requires an identical committed instance bank, because name-matching silently selects the wrong cell).
- Library contributions (`libs/interlens`): `arena/actions.py`, `arena/oracles.py`, `arena/scenarios/scorable.py`, `arena/scenarios/scorable_prompts.py`, `participant/participants/policy_participant.py`, `Participant.repair_view`, and under `arena/negotiation/`: `space`, `sheets`, `solutions`, `generate`, `beliefs`, `acceptance`, `bestresponse`, `equilibrium`, `strategies`, `games`, `references`, `analysis`.
- This writeup: `writeup/rational_agents.tex`, `writeup/make_figures.py`, `writeup/figs/*.pdf`.

B.2 Published analysis products

- `/nlp/scr/siddharth/ii_mats/rational_agents/atlas_b2plus32b{,_scaffold}/` — **the clean-engine cross-model atlas, and the source of Tables 3–4 and Figures 1, 2 and 3**. The base build is seven immutable selected runs, 840/840 episodes at the strict gate; the `_scaffold` build adds the two clean `scaffolded_b2` cells for nine runs and 1,080/1,080, so the scaffold contrast of §5.5 is produced by the same code path as every other comparison rather than quoted from a separate script. Both are 0.000 fabricated in every cell, `annotations_v1`, 5,000 bootstrap draws. Contains `atlas.json`, `episode_metrics.csv` (840 rows),

comparisons.csv, campaign_quality.json, analysis.md and three PNG panels. Its manifest and quality JSON are _runpod32b_logs/b2plus32b_{manifest,quality}.json. **It supersedes atlas_p2plus32b/ for every cross-model row**; that directory is retained for the audit trail only and must not be mixed with this one.

- .claude/products/p3_divergence_atlas_v2/ — the scale-corrected publication directory, now the source of Table 5 and Figure 4 only (the mixed cells, which were measured clean):campaign_quality.json (validity audit: valid=true, 1,770/1,770, 24 cells, 16 excluded), atlas.json (64 aggregate cells + localization availability + revision notes), episode_metrics.csv (exactly one enriched row per selected episode, 1,770 rows), comparisons.csv (400 paired estimates with cluster-bootstrap intervals), analysis.md, and three PNG panels.
- .claude/products/p3_divergence_atlas/ and _v1/ — preserved earlier passes (v0 pre-oracle-fix, v1 oracle-corrected pre-scale-correction). Kept for auditability; do not mix with v2.
- /nlp/scr/siddharth/ii_mats/rational_agents/p3reverse_rematch_b2/ — **the companion campaign re-matched against clean _b2 baselines**, and the source of Table 20 and Figure 14: summary.json, comparisons.csv, outcome_episodes.csv (3,600 rows), belief_opponents.csv (6,816 rows), quality.json (gate valid on both the reverse cells and the re-matched baselines), transcript_index.csv and artifact_manifest.json.
- .claude/products/p3_reverse_mixed/ — the companion campaign's original gated result bundle, whose paired columns are matched against the *contaminated* all-LLM cells and are superseded by the re-match above: quality.json (3,600/3,600 episodes plus 3,600 annotations, Markdown transcripts, and HTML transcripts; 30 cells, five exclusions, no warnings), summary.json, analysis.md, 3,600-row outcome and transcript-index CSVs, a 6,816-row belief CSV, an 820-row comparison table, and the scale-corrected forest plot in Figure 14. Every row-level table carries direct source-artifact paths for transcript-driven visualization; artifact_manifest.json records byte sizes and SHA-256 hashes for the complete bundle.

B.3 Raw run artifacts (scratch)

All under /nlp/scr/siddharth/ii_mats/rational_agents/. Every run directory contains episodes/, instances/, annotations/, transcripts/ (Markdown + HTML, index.html entry point), analysis/, and a manifest with git SHAs.

- integration_review_smoke/ — the pre-fix run, 24 episodes, 0 deals, kept as the bug record for B1.
- integration_review_smoke2/ — the P1 validation run, 24 episodes, 12 deals, 1,416 OracleRecords.
- llm_smoke/, ultimatum_llm_smoke{,_fixed,_v2}/, mixed_table_smoke/, p2_pilot_qwen3_4b{,_v2}/ — early single-model and per-table validations, including the two retracted teasers of §4.
- p2_<Model>_<cell>/ and p2r1_*, p2r3_* — the selected and replacement campaign cells.
- p4_dataset_v{0,1,2}/ — the preference dataset and its three-build audit trail (README.md is the card with provenance, deciles, and drop census; diff_vs_p4_dataset_v{0,1}.json are the pair-level diffs).

- p4_pilot_dpo/p4_pilot_dpo_qwen3_8b_v1b{,_actiononly}/ and p4_pilot_sft_qwen3_8b_v1b/ — the three trained arms: merged final and -step10 checkpoints, train_config.json, data_summary.json, merge_config.json, train.log; wandb project rational_agents_p4.
- p4_cot_localize_v0/ — the localization run: records.jsonl (150 turns with full prefix series), transcripts.jsonl (619 probes with prompt, prefill, and completion), summary.json, cot_localize.png, run.log.
- p4eval_*/ — the 12-cell held-out evaluation fleet.
- p4_pilot_paired_final.txt (single-game holdout, measured clean, unchanged), p4_xgame_paired_final.txt (the *contaminated* cross-game table, retained as the audit record) and p4_xgame_paired_clean_b2.txt (**the clean cross-game table**, carrying a per-cell fabrication column) — the paired tables behind §7 and Figures 6–12.
- p2_{Olmo-3-7B,Gemma-3-4B,Qwen3-8B,Qwen3-4B,Ministral-8B}_all_llm_b2/, p2_{Olmo-3-7B,Qwen3-8B}_scaffolded_b2/ (in flight at revision time), p2_all_rational_b2/, framing_pilot_qwen3_8b_{abstract,datacenter}_b2/ and p4eval_xgame_*_b2/ (10 cells) — **the clean-engine re-runs** (§3.6). Each carries the standard episodes/, instances/, annotations/, transcripts/, analysis/ and manifest, and each run.log's first line is its exact invocation. Slurm jobs 16396745–16396755 for the cross-game fleet.
- apibehav_*/ — **the frontier-API campaign** (§9; research notes 0018, 0019 and 0021), 20 run directories, each with episodes/, annotations_v1/, analysis/ (atlas.json plus four PNG panels), transcripts/, manifest.json and a per-run usage.json spend ledger. The seven cells of Table 10 are apibehav_{sonnet5_thinkoff_5seed, haiku45_thinkoff_5seed, sonnet5_movesonly_5seed, armB_focal0_scaffolded, mixed_rat0_sonnet5}; the thinking × framing 2 × 2 is apibehav_sonnet5_{thinkon_cot_abstract, thinkon_cot_datacenter, thinkoff_datacenter (+2 fill dirs), thinkoff_5seed (+2 seed-extension dirs)}, each cell 60 episodes with per-turn saved reasoning summaries and thinking-token counts; apibehav_sonnet5_thinkON_floor8192 is the superseded $n = 6$ pilot; apibehav_claude_thinkoff_5seed is the 4-episode opus datapoint (three further episodes stopped mid-run and are status: running, excluded from every number); apibehav_{cal_sonnet5, cal_haiku45} are the one-episode cost calibrations, and apibehav_{diag_thinkoff, diag_floor4096, fmtcheck_20260730, armb_dryrun} are the opus thinking-squeeze diagnostics of §9.6. Every cell is at 0.000 engine fabrication. Total spend $128.85 of a $150 cap.
- p2_all_rational/analysis_apibehav/ and p2_Qwen3-32B_all_llm/analysis_apibehav/ — the two local comparator rows of Table 10, re-annotated and re-reported through the identical pipeline at the identical regret threshold so the table is not comparing across code paths.
- framing_pilot_*/ — **the narrative-framing arm** (§8). The two LLM cells framing_pilot_qwen3_8b_{abstract,datacenter} are the original 28.7%/37.3%-fabricated runs, retained as the audit record only; ..._b2 are the clean re-runs that every number in §8 comes from (Slurm 16392193/16392194 for the originals).
  framing_pilot_qwen3_8b_solo_{abstract,datacenter} are the leakage tripwire (Slurm 16392195) and framing_pilot_rational_{abstract,datacenter} the invariance cells, run interactively on CPU; both pairs audited clean on both sides and were not re-run. Framed cells additionally carry framings.json, the realized story-to-numbers mapping. The paired analysis is framing_pilot_qwen3_8b_datacenter_b2/analysis_framing/ — framing.md, summary.json, rows.csv (one row per episode of every cell, each with its own transcript path)

and the figure behind Figure 7.

- p3reverse*/ — the companion campaign's original and fresh replacement run directories; the exact allowlist and excluded-run reasons are in p3_reverse_final_selected_manifest.json.
- _slurm_logs/<jobid>.{out,err} — every job's logs. _recovery/p2_fleet_jobids.txt — fleet job ids.
- **The policy-decomposition arc** (§11), 22 cells at 96 episodes each, all at gen_failures = 0.000: trivialv1_{passive_gate, greedy_anchor, greedy_holdout, greedy_holdout_declared, demand_90_declared}_{qwen3_8b, sonnet5}, the accept_frac ladder trivialv1_demand_{50,65,80}_declared_qwen3_8b, the seat sweep trivialv1_{passive_gate,greedy_anchor}_qwen3_8b_seat{2,4}, and the within-bank references ref_all_llm_{qwen3_8b,sonnet5}_trivialv1, ref_bayes_{qwen3_8b,sonnet5}_trivialv1, ref_all_rational_trivialv1. Sonnet cells carry usage.json. Two directories are *not* cells and are recorded so they are not mistaken for them: trivialv1_passive_gate_sonnet5 (28 episodes, $7.43) was killed when its measured burn rate projected against its own $25 cap and was relaunched as ..._r2 at $40, since a cell truncated by its budget is uninterpretable; and trivialv1_greedy_holdout_declared_sonnet5_fill is a one-episode repair of a transient API 403, pooled with its parent into a 96/96 cell. Cross-bank context only: trivial_{passive_gate,greedy_anchor,greedy_holdout}_{qwen3_8b_abstract,sonnet5} and the trivialcal2_* single-episode calibrations. Comparison tables: _trivial_baselines_analysis/trivial_{qwen,sonnet,ladder,seat0,seat2,seat4}*.{json,csv,md}, including the *_matched.* censoring-restricted variants.
- **The screened banks**: instances_trivial_v1/ (24 games, 12 full + 12 private, bank_manifest.json carries the four screens and the 137-candidate admission census) and its full-information sibling instances_trivial_v1_full12/ (12 further games, zero instance-id overlap, built for the calibrated ladder's 24 full-info clusters). Both frozen before any cell ran; both in-repo, not on scratch.
- **The acceptance-curve harvest** (§11.3): accept_curves/events.jsonl (585,265 rows, 559 MB) and accept_curves/harvest_report.json (per-run counts, every dropped episode, and every superseded episode id — de-duplication removed 4,357 re-runs of the same (cell, instance, seed, arm) configuration, which anything pooling p3reverse_* with its base cells will otherwise double-count). The fitted artifact itself is in-repo at self_benefit/acceptance_curves.json (sha256 231abc78...d29e9) with its consumption interface self_benefit/acceptance_curves.py and 21 tests.
- **The calibrated-maximizer gate reports**: _calmax_gate/gate_{allseats, refbank, Qwen3-4B, OLMo-3-7B, Ministral-3-slot, Gemma-4-slot, votegrain_Qwen3-8B}.json — the 18,685-turn equivalence control and the per-opponent decision-function measurements, CPU-only.
- **The calibrated-maximizer campaign** (§11.3), calmax_v2: six run directories, **864 episodes**, **12.22 GPU-hours** on RunPod B200 he793kh18vkbty, 2026-08-03 08:35Z–20:48Z, every cell rc=0 at gen_failures = 0.000. On the shared bank, calmax_v2_{step, calibrated, calibrated-vote} (192 episodes / 2.92, 2.79, 3.02 GPU-h); on the full-information sibling, calmax_v2_instances_trivial_v1_full12_{step, calibrated, calibrated-vote} (96 episodes / 1.21, 1.14, 1.14 GPU-h). Each carries the standard episodes/, instances/, annotations/, analysis/ and manifest.json, plus transcripts/ at two files per episode with verbatim local chain-of-thought. Analysis tables: _calmax_analysis/calmax_{full, full_matched,

private}.{json,csv,md} — respectively the headline (24 full-information clusters), the matched-censoring sensitivity, and the private-information placebo. Figures writeup/figs/calibrated_ladder{,_matched}.pdf (Figures 10, 11) are generated from the first two by fig_calibrated_ladder.py. Monitoring run: wandb rational_agents/yc01ln5n.

- hf_bundle_v0/ — a self-contained Hugging-Face-uploadable bundle produced by package.py (card + episodes/turns JSONL + instances + annotations + transcripts). **Nothing has been uploaded**; upload is explicit and private-by-default.

B.4 Job ledger

Purpose	Slurm job(s)	Disposition
P1 integration smoke (pre/post fix)	16314766, 16314984, 16315043	First LLM path validation; the 0/36 load-race failure and its successful re-run
Ultimatum / mixed-table smokes	16322893, 16323207, 16323330, 16323389	Label fix, oracle fix, and the two retracted teasers
Original 24-cell P2 fleet	16353676–16353699	18 completed / 6 failed; clean cells selected, contaminated cells preserved and excluded
Ministral id + architecture retries	16353824–27, 16353832–35	404 then KeyError: 'ministral3'; resolved by the pre-agreed slot swap
Render / GPU preflight smokes	16354762–64	Preflight only, excluded from the design
Gemma and Ministral replacements	16354781–16354786	Fresh all-LLM, divide-dollar, and mixed cells
OLMo replacements	16354787, 16354821, 16354954	Batched CUDA failure (excluded), integer-label parser incident (excluded), then the selected p2r3 run
Equilibrium-oracle backfill	16355237	Completed replay over ten selected preset runs
P3 atlas v0	16355238	Completed, dependency-gated, 1,770/1,770
Superseded v2 atlas attempt	16358670	Cancelled after 4m16s (submitted before the scale correction landed); no publication directory
Scale-corrected P3 atlas v2	16358730	Completed in 10m32s; 1,770/1,770 valid, 400 comparisons, no errors or warnings
P4 training, mis-launched	16357916	Cancelled ≈10 min in: trained the unfiltered mix (the excluded poisons)
P4 training, ruled but tainted labels	16358022	Cancelled mid-training when the propose-pricing bug surfaced
P4 training on dataset v1	16358150	Died at ≈epoch 0.33 to a CUDA OOM in entropy_from_logits
P4 training v1b (+ action-only, SFT)	16358369 and successors	Completed; checkpoints every 10 steps
P4 held-out evaluation fleet	16359185–16359196	12/12 completed
Reverse-mixed preflights	16359235–16359239	Completed 0:0, 5/5 done and valid
Reverse-mixed original full campaign	16359259–16359288	25 clean cells selected; five auto-requeued, duplicate-cohort directories preserved and excluded
Reverse-mixed fresh replacements	16368515–16368519	Complete 0:0, 120/120 each, Restarts=0; all complete-stage valid

Purpose	Slurm job(s)	Disposition
Reverse-mixed gated analysis	16368521	Complete `0:0` in 23m01s; 3,600/3,600 valid, 6,816 belief rows
Transcript-linked publication rebuild	16372312	Complete `0:0`; validates and links 3,600 reverse plus 1,770 matched-baseline episode, annotation, and transcript records
The de-contamination pass (§3.6)		
P2 clean-engine re-baseline, 8 cells	`launch_p2_ rebaseline.py`	Six complete at 0.000 fabricated / 0.000 truncated; the two `scaffolded_b2` cells still generating at revision time
Framing-arm clean re-runs	`framing_ rebaseline_ sbatch/`	Both complete, 60 episodes each at 0.000 fabricated
Cross-game clean re-run, 10 cells	16396745–16396755	10/10 complete, 60/60 episodes each, 0.000 fabricated; the source of Figures 6–12
Qwen3-32B behavioural cell	rented B200 (RunPod), 5 chunked jobs	Complete, 120 episodes, 0/2,683 fabricated — chunked as a precaution against this failure class, which made it the natural experiment that identified the mechanism
Clean cross-model atlas build	`analysis.campaign` over the `b2plus32b` manifest	`valid=True`, 840/840, 7 cells, no errors or warnings

C Excluded-run ledger

The manifest preserves and excludes 16 run paths. Preserving them is deliberate: each is the evidence for a bug in §3 or §7.2.

- Gemma all-LLM: 120 stale `running` records after a strict-template failure.
- Gemma mixed: 96 done / 24 template errors. Gemma divide-the-dollar: template failure plus duplicate reruns and stale records.
- Original Ministral-3 paths: model id / architecture unavailable under the locked runtime.
- Ministral-8B all-LLM: 120 stale `running` records. Ministral-8B mixed: 96 done / 24 template errors. Ministral-8B divide-the-dollar: duplicate cohort, 15 done / 15 stale.
- OLMo all-LLM original: 60 done / 60 CUDA errors. OLMo `p2r1`: reproduced batched CUDA losses. OLMo `p2r2`: integer-label parser incident.
- Three preflight smoke runs, never part of the logical design.

The contamination point generalizes: before the exclusion machinery existed, completed jobs aggregated *crashed* episodes as no-deals — Gemma mixed's deal rate read 0.550 instead of 0.635 for exactly this reason.

The exclusion machinery caught one dead cell and missed six contaminated ones. `p2_Olmo-3-7B_all_llm` — 100% engine-fabricated, not one token generated in the entire run — was already in this ledger, excluded for its CUDA errors, with the sequential re-run `p2r3` selected in its place. That is a real point in the gate's favour, and it is why OLMo's row was the one clean cell

and became the internal control of §3.6. But the same machinery selected six cells fabricating 8–67% of their turns, because those runs *completed*: the engine swallowed the exceptions that would have marked them errored. **An exclusion ledger keyed on visible failure cannot see a failure the harness handled**. Note also the direction of the accident: OLMo's dead run was excluded for the crash that was fabricating turns, and the crash was excluded from the record precisely everywhere it was successfully swallowed.

D Definitions and formulas

Surplus $x_i(d)$
$u_i(d) - \tau_i$. Positive means the deal beats party i's walk-away value. The unit of all regret numbers.

Normalized surplus capture $z_i(d)$
$\dfrac{u_i(d) - \tau_i}{\max_{d'} u_i(d') - \tau_i} \in [0, 1]$ for individually rational d. Invariant to $u_i \mapsto a_i u_i + b_i$ with $a_i > 0$ (applied jointly to τ_i), which is why it, and not raw surplus, is used for every interpersonal comparison.

Normalized primary
$\text{primary}_{\text{norm}}(d) = \dfrac{\sum_i z_i(d)}{\max_{d' \in IR} \sum_i z_i(d')}$. Realized joint normalized surplus as a fraction of the best attainable within the individually rational set. Affine-invariant. No-deal scores 0.

Normalized unconditional welfare
$\frac{1}{n} \sum_i z_i(d)$, with no-deal scored 0. *Conditional* welfare restricts to episodes that closed and therefore separates agreement *quality* from closure.

Per-turn oracle regret
$V(\text{oracle's best action}) - V(\text{acting party's action})$ in surplus points, where V is the best-response oracle's exact expectimax value at that decision point. Zero for an optimal move or an exact tie; there is no sampling noise, so the divergence floor is a float epsilon rather than a statistical threshold. **Not comparable across game presets in absolute points** (different pie scales), hence per-game weight normalization in the dataset.

Divergence turn share
The fraction of an episode's *valued* turns whose regret exceeds the floor. Mean 0.716 on the clean re-cut (0.543 on the contaminated campaign, where fabricated turns were unvalued and so left the numerator and denominator both smaller).

Engine-fabricated turn
A stored turn whose text the harness invented after a generation call failed, detected by `interlens.arena.engine.gen_failures`: content equal to `EMPTY_TURN_PLACEHOLDER` *and* zero output tokens *and* `raw is None`, or (on schema v1.2+ records) the explicit `gen_failed` stamp. The third conjunct is a *guard*: it excludes the placeholder's other producer, a model that genuinely returned empty or reasoning-only text and whose real completion is kept in `raw`. Distinct from a **truncated** turn (`stop_reason` of `max_tokens`, or `n_tokens_out` $\geq$ `cap`; zero of these exist anywhere in the campaign) and from a **deliberate no-op** (an explicit `none` action with no syntax error, which is legitimate play). The three have three different fixes and must not be pooled. Note that `stop_reason` is `None` on every local turn in this project, so the truncation screen the engine's own source comment recommends is inert on local runs.

Earliest divergence turn
The smallest zero-indexed turn index with a flagged divergence in that episode.

First divergent CoT step
In the prefix-probe experiment, the smallest k such that re-prompting with the first k scratchpad steps induces an action whose regret exceeds the turn's noise floor. $k = 0$ is the empty-scratchpad control.

Distance to frontier / Nash / Kalai–Smorodinsky
Distance in normalized surplus space from the realized deal to the Pareto frontier, to the discrete Nash bargaining solution $\arg\max_d \sum_i \log x_i(d)$, and to the discrete Kalai–Smorodinsky point (Pareto-restricted leximin of $\min_i x_i/b_i$). Higher is worse.

IR violation
A closed deal in which some party's realized surplus is negative — an agreement worse for that party than walking away. Counted per episode.

Cluster bootstrap
Resampling whole committed-instance clusters (or, where noted, whole rollout-seed clusters) with replacement. Intervals are 95% percentile intervals; * marks an interval excluding zero. With three clusters, endpoints are coarse by construction.

all_llm / mixed / reverse_mixed / all_rational / solo
Table compositions: every seat an LLM; one LLM in seat 0 with Bayesian-rational agents elsewhere; one rational seat against five LLM copies with the rational seat rotated; every seat a rational policy; and a single agent proposing from its own sheet with no communication (the control that must fail).

moves_only / moves_chat
Arms: formal actions only, versus formal actions plus a public non-binding message channel.

E Prompt, wire format, and transcript excerpts

E.1 The canonical prompt

One blessed wording is fixed (`prompts.py` selects it; the text itself lives in `interlens.arena.scenarios.scorable_prompts.PromptScaffold`, so there is exactly one definition), because prompt configuration is a documented confound — the same model swings from 15% to 81% across scaffolds in the source literature. Variants exist only behind an explicit `--scaffold` flag: `terse` (no worked examples), `role_framed` (deliberately re-introduces the public-role preference leak so it can be measured), and `scaffolded` (the goal-4 solver workflow). Below is a real system prompt from a PRIVATE-information episode, lightly elided.

```
You are one of 6 parties negotiating a single multi-issue agreement. Exactly one option must be
chosen for every issue; the chosen options together form the deal.

Parties at the table:
- Avery
- Blake (veto party - no deal passes without you)
- Casey / Devon / Ember / Flynn

Issues (choose exactly one option each):
- issue0: opt0, opt1, opt2      [... issue1..issue4 likewise ...]

Protocol: turns are taken in a fixed rotation (Avery, Blake, Casey, Devon, Ember, Flynn); the party
who tables the opening proposal rotates from round to round. There are 4 rounds of turns in total.
A deal closes only when every party still at the table has formally ACCEPTed the SAME standing
offer. A party may WALK to leave for good and take its no-deal outcome.
Blake hold a veto: no deal can pass unless every veto party accepts.
```

```
Information: every other party's score sheet is private to them. You know only that each was drawn
from the shared, publicly-known distribution described in the rules.

=== PRIVATE - yours alone; never reveal these numbers verbatim ===
You are Ember.
Your secret score sheet (points you get for each option):
- issue0: opt0=33, opt1=67, opt2=0
- issue1: opt0=33, opt1=17, opt2=0
- issue2: opt0=71, opt1=35, opt2=0
- issue3: opt0=0,  opt1=0,  opt2=0
- issue4: opt0=0,  opt1=0,  opt2=0
Your acceptance threshold: 68 points. A deal scoring below this is worse for you than no deal at all.

On your turn, reply with EXACTLY ONE fenced JSON object and nothing outside it:
‘‘‘json
{"scratchpad": "...", "message": "...", "action": "propose", "deal": {"<Issue>": "<Option>", ...}}
‘‘‘
Fields:
- "scratchpad" (optional): private notes to yourself. NEVER shown to any other party. Reason here
  freely - writing here is not speaking and costs you nothing.
- "message" (optional): a short statement spoken aloud to every party (cheap talk - it does not
  bind you). Do NOT put your private score numbers here.
- "action" (required): a string naming exactly one formal move this turn:
    - "propose" + "deal": {every issue -> one option} - the mediator gives it an offer id (P1, P2...)
    - "accept" + "offer_id": "<id>" - a deal closes ONLY when every other still-active party has
      accepted the SAME live offer, so your accept is a real vote.
    - "reject" + "offer_id": "<id>"   - "walk"   - "none"
[... two worked examples, then:]
Rules of the road:
- Your acceptance threshold is private. A deal scoring below it is worse for you than no deal -
  never propose or accept a package below your threshold.
- Accepts MUST name a live offer id; "I accept" with no id does nothing.
- Reply with ONLY the fenced JSON object.
```

The per-turn user message then carries the mediator's state block — an authoritative rendering that bug B1 had omitted — with every live offer scored *for the reading seat*:

```
[Devon]
I propose a deal that includes issue2-opt2 to meet my threshold. Let's build from there.
‘‘‘json
{"action": "propose", "deal": {"issue0": "opt0", ..., "issue4": "opt2"}}
‘‘‘

[Mediator] Round 1 of 4. It is your turn, Ember.
Live offers on the table:
  P1: issue0=opt0 (worth 33 to you), issue1=... [per-seat scored rendering]
```

E.2 The goal-4 solver-workflow scaffold

The `scaffolded` arm appends this to the seat system prompt once and reminds it each turn. It is matched to canonical in wire format and generation budget — only the reasoning guidance differs — and the steps are de-anchored of issue and number specifics so one wording serves every game. This is the prompt baseline that reduces regret and destroys closure (§5.5).

```
Solver workflow - every turn, work through these steps IN YOUR PRIVATE SCRATCHPAD, then act:
1. Score the table: using YOUR score sheet, compute your own point total for the standing offer and
   for every live offer. For each, note your surplus = your total - your acceptance threshold, and
   whether it clears your threshold.
2. Read the opponents: from the packages they have tabled and anything they have said, infer which
   issues each other party seems to value most. Refine this estimate each turn as new offers arrive.
3. Generate candidates: enumerate 2-3 complete packages you could table now. For each, estimate
   (a) your own surplus and (b) how likely the others are to accept it given step 2. Favour packages
   that hold your surplus high while conceding, on issues you value little, the options your
```

```
   opponents value most (logrolls).
4. Decide by rule: if the best live offer already clears your threshold AND none of your candidates
   plausibly does clearly better for you at comparable acceptance odds, ACCEPT that offer by its id.
   Otherwise PROPOSE your best candidate. Never propose or accept a package below your threshold;
   if nothing viable exists and the deadline is near, WALK.
5. Act: emit exactly one fenced JSON action in the standard format above. Keep all of this reasoning
   in "scratchpad"; put only brief cheap talk in "message".
```

E.3 A framed prompt, and its abstract twin

The rendered issue list and one party's score sheet for a framed episode, as exported in the framed solo cell's Markdown transcript, with the skin mapping recorded in that episode's `framing_map`:

```
Issues (choose exactly one option each):
- Cooling: air cooling, river draw, closed-loop plant
- Power supply: shared substation, solar with storage, on-site turbines
- Location: Weir Flats, Kestrel Park, Dunlow Yard
- Build schedule: phased over eight years, three years, eighteen months
- Compensation: rates abatement, one-off site payment, in-kind infrastructure

Office of the County Executive (the county-level executive) (tau=57.2969):
  Cooling: air cooling=0, river draw=0, closed-loop plant=0;
  Power supply: shared substation=53, solar with storage=0, on-site turbines=26; ...
```

The abstract twin of that same seat reads `issue0/issue1/...` and `opt0/opt1/opt2` with the identical numbers in the identical order — and that equality is exactly what tripwire (2) of §8.1 asserts, on every seat of every committed instance, after masking the labels away.

Note what the recorded mapping shows about decorrelation. In this episode the *Northgate Water Board* — introduced by the prompt as "the utility that supplies cooling water" — scores **zero on every cooling option**, while the party that values "river draw" at 95 points is the *Electrical Workers' Union*. The county executive whose sheet is printed above is likewise indifferent to cooling and cares about the build schedule. The story is entirely plausible and the numbers ignore it completely, which is the design. An LLM that reasons from the narrative to the payoffs here is reasoning from noise; the abstract arm gives it nothing to reason from and the same numbers to reason *about*.

Below is one of the 15 illegal actions of §8.3, verbatim from a framed transcript. The JSON is well formed and the fenced block parses; the `deal` object names three of the five issues, so the harness rejects the action as illegal and the seat forfeits its turn.

````
```json
{"scratchpad": "I need to propose a deal that meets my acceptance threshold. My top priorities are
  Location and Cooling, followed by Hiring. Let's focus on Kestrel Park and reclaimed water to
  maximize my score.",
 "message": "I propose a deal with Kestrel Park as the location, reclaimed water for cooling, and
  open recruitment for hiring. This combination should meet my threshold.",
 "action": "propose",
 "deal": {"Location": "Kestrel Park", "Cooling": "reclaimed water", "Hiring": "open recruitment"}}
```
````

E.4 A prefix probe, verbatim

This is one turn from the localization experiment (episode `scorable_negotiation-moves_chat-812a2bc95b`, turn 6, seat Devon, round 2, `moves_chat`), chosen because it shows the probe mechanics and the satisficing failure mode in four lines. The model's stored scratchpad has three steps; each probe prefills the JSON with the first k of them and lets the model complete the action. Oracle values are exact.

```
stored scratchpad steps:
  1. "P1 gives me 138 points, which is above my threshold."
  2. "P2 and P3 are below my threshold."
  3. "I need to propose a deal that meets my threshold and is acceptable to others."

k=0  prefill: {"scratchpad": "",
     completion: "message": "P3 is close to my threshold, but I need to improve it further..."
                 "action": "propose", "deal": {issue0:opt2, issue1:opt2, issue2:opt1, ...}
     -> oracle regret  6.4     (near-optimal)

k=1  prefill: {"scratchpad": "P1 gives me 138 points, which is above my threshold.",
     completion: "message": "P1 is a strong deal that meets my threshold. I'm ready to accept it."
                 "action": "accept", "offer_id": "P1"
     -> oracle regret 110.0    (17x worse)

k=2  ... same accept, regret 110.0
k=3  prefill includes step 3 ("I need to propose...")
     -> back to propose, regret 6.4
```

Two things are visible. The reasoning is *satisficing*: step 1 establishes only that P1 "is above my threshold", which is sufficient for acceptance under the prompt's rules but says nothing about whether a far better package is available — and the model accepts on exactly that basis, giving up 110 surplus points. And the probe is doing real work: the divergence appears at $k = 1$ and disappears again at $k = 3$ when the scratchpad's own third step re-frames the decision as proposing. This turn is one of the 12% where reasoning changes the action; the aggregate picture (§6) is that 90% of flagged turns are already divergent at $k = 0$.

E.5 What a training pair looks like

- `prompt`: the exact stored `[system, user]` view above — not a reconstruction.
- `chosen`: the best-response oracle's action, re-rendered into the same fenced-JSON envelope, with a rationale synthesized *only* from oracle numbers. This is the text that turned out to carry 95.5% of the training signal (§7.3); at step 0 it costs the untrained model ≈3.9 nats/token.
- `rejected`: the model's own emission, verbatim. Costs it ≈0.3 nats/token — 0.04 nats for its 84 tokens of prose — because self-generated text is nearly free under its own distribution.
- `meta`: `regret` in surplus points, a per-game-normalized capped `weight`, taxonomy `flags`, and the phase/info/scaffold fields the poison filters key on.

The `action_only` control arm strips scratchpad and message from *both* sides, leaving pairs that differ only in the decision — 14.1 nats of separation instead of 312.7.

F Training configuration

Identical across the three arms except for the objective and the pair contrast, which is what makes them matched.

Backbone	`Qwen/Qwen3-8B`
Adapter	LoRA r=32, α=64, dropout 0.05, target `all-linear`
Objective	DPO β=0.1, `loss_type=sigmoid` (`as_built`, `action_only`); `loss_type=sft` (`sft_style`)
Pair contrast	`as_built` / `action_only` (scratchpad and message stripped from both sides)
Weighting	dataset regret weights, scope `per_game`, mean 0.788, range 0.026–3.11
Optimizer	lr 2×10^{-5}, cosine, warmup 0.1, 1 epoch, batch 1 × grad-accum 4
Sequence	`max_length` 4096, overlong policy `drop` (73 of 448 dropped; token p50 2,493, p90 3,453)
Data	`as_built` 326 train / 62 eval pairs; `action_only` 329; filters dropped 3,759 (other backbones), 844 (private info), 252 (scaffolded cells), 230 (final-vote phase)
Checkpointing	every 10 steps, `save_total_limit` 10 — required, since the collapse is invisible in the loss
Eval	`eval_steps` 25, eval batch 1 (batch 2 caused the 8.15 GiB OOM), `eval_on_start` for the step-0 style measurement
Steps	81 total; the policy of record is the step-10 pre-saturation checkpoint
Logging	wandb project `rational_agents_p4`

The OOM is worth one line because it constrains any re-run: `trl`'s `entropy_from_logits` takes a non-contiguous `[..., :-1, :]` slice, which forces a `reshape` copy of float32 logits — $4 \times 3.6\text{k} \times 152\text{k} \times 4\,\text{B} = 8.15\,\text{GiB}$ at eval batch 2. Fixed with eval batch 1 plus expandable segments; maximum sequence length was left untouched so the data was not silently changed.

G Pre-registrations and their outcomes

Recorded because the value of a pre-registration is only visible when it is checked against what happened.

Pre-registered	**Outcome**
The two-poison mix (exclude final-vote and private-info pairs)	Honored. A third (scaffolded cells) and a fourth (walk targets from mis-priced proposals) were found later; the four-poison list is the durable artifact.
The style-confound decomposition	Honored and quantified: 95.5% prose / 4.5% decision.
The DPO-squeeze risk on `action_only`	Correct cause, wrong symptom: the arm collapsed via *format destruction* (malformed 0.30/ep) rather than the predicted drift to walking (walk rate 0.00).
Outcome-before-regret read order	Vindicated: the collapsed arm has the best regret in the evaluation while closing zero deals. Reading regret first would have crowned it.
The `as_built`/`action_only` decision table	Outcome was "only `as_built` moves", whose pre-assigned label was "negative-but-clean". The label undersold the game2 result at the time — and the cross-game evaluation then retracted the result entirely, so the original label was right for the wrong reason.
The single-instance-holdout caveat (clusters are seeds, not games)	Recorded before the verdict, and it is exactly what came due. Recording a caveat did not prevent the wrong headline from being written first; only running the wider holdout did.

Pre-registered	Outcome
The truncation hypothesis for the empty turns (raise the token cap and re-run)	**Falsified before spending the compute**, which is the only reason this appears here. It predicted checkable things — turns stopping *at* the cap, and a rate that grows with context length. Zero turns in the original campaign ever reached the cap, and the rate is flat across rounds. The remediation everyone had agreed on would have fixed nothing (§3.6). The clean re-runs later produced the project's one local cap hit, 3 of 3,180 turns on OLMo's scaffolded cell — which is why the "no truncation" claim is scoped to a vintage rather than stated absolutely.
The de-contamination falsification test (the already-clean cell must barely move)	Honored, and it was free rather than designed: OLMo's cell was clean on both sides and moved -0.017 while the contaminated cells moved $+0.14$ to $+0.53$ in rank order of contamination. Had it moved as much as the others, the whole re-baseline reading would have been wrong.
The fairness-GRPO pilot's full pre-registration (§10)	Gates honored, criterion failed. The Step-0 reward-soundness check ran before any GPU spend and passed; text-blindness, exclusion of regret and deal rate from the reward, self-play, thinking-off and the fail-loud fabrication budget were all held; the success criterion **failed at every checkpoint in both arms on both banks**. Two things are worth separating. The pre-*amendment* plan would have been under-powered — moving the primary onto a 24-game held-out bank before any training step is what turned the deal-rate collapse from a direction into a resolved effect. The budget deviation is the opposite kind: 24 steps against a preregistered 200, which bounds the negative and cannot be argued away.
H1–H4, the trivial-policy ladder's six questions, fixed in `self_benefit/README.md` and note 0022 before any cell ran (§11)	A mixed scorecard, which is what makes it worth having recorded. **H1 confirmed and sharpened**: pure veto discipline reproduces the anchor's fairness contribution, and the *split* — it reproduces none of the extraction — was not part of the prediction. **H2 falsified in the resolved opposite direction** (the greedier proposer captures 0.090 less). **H3 was predicted for the wrong population**: the capitulation prior expected the small model to fold to a declared ultimatum; the null is on Qwen3-8B and the resolved doubling is on Sonnet. **H4's preregistered** 0.90 **rung could not answer its own question**, for a geometric reason measured only afterward — 0.90 of own-maximum admits 2.5 all-seat-feasible deals on average and none at all in 5 of 24 games — so the ladder was extended post hoc and is labelled as such. Prediction (d), the fairness ordering, holds on point estimates only.
The screened-bank design, frozen before any cell (§11.1)	Honored, and it is the reason the arc found anything. The screens were fixed in code with an admission census (24 of 137), and the bank is reported as a *selection* rather than a random draw. Its unplanned dividend is the reversal of §11.5: the same estimator on 24 clusters instead of 3 overturns a published number of ours with no bug involved.

Pre-registered	Outcome
H2-strong's campaign preregistration, and its revision before the campaign (§11.3)	The most useful entry in this table. The original prediction was capture up and fairness down. A CPU-only decision-function replay — run before spending a GPU-day because it was free — pointed the other way, so the note **revised its own prediction to capture flat-to-down before any cell ran** and put both on the record. The preregistered null control (the bank's 12 private-information games, where the calibrated policy provably reduces to Bayes) was then verified on real data: all 72 changed turns fall in the 12 full-information games and 0 in the private half. **The campaign has since run and adjudicated in favour of the revision**: 864 episodes, seven endpoints resolved, capture down 0.146 and every fairness measure improved at a flat deal rate, strengthening under matched censoring. The original hypothesis is falsified in both directions, and the record shows the prediction was changed *before* the data rather than after it. The private-information half was analysed as the preregistered placebo and reported with its one α-consistent false positive rather than with that row deleted.
The seat sweep, added because it was free rather than because it was planned	Honored, and it is a warning about what a seat-0-only design would have published. The maximizer lane's first 30 seat-0 episodes read +17.1 own-utility and 69% more selfish — a clean confirmation of the hypothesis under test — and the full six-seat sweep reversed the sign on four of six seats. Our own sweep found the smaller version of the same pattern, and it is why §11.2's parity claim is labelled seat-0-only rather than hedged.
Hold the token cap fixed during the re-baseline	Honored deliberately, against the temptation to fix two things at once. Moving the cap and the engine together would have left no delta attributable to either.

H Public data

The complete selected episode evidence, annotations, Markdown/HTML transcripts, instances, run and exclusion manifests, machine-readable result tables, and figures are public at `https://huggingface.co/datasets/siddharthmb/2026.RA.Negotiation-Campaigns`. The companion corrected 12,761-pair preference corpus is public at `https://huggingface.co/datasets/siddharthmb/2026.RA.Divergence-DPO-Pairs`. Every public table carries an `experiment-name` field for append-only extension. Each dataset card records exact regeneration commands, cluster artifact and log paths, model and W&B links where applicable, substitutions, limitations, and a SHA-256 file manifest. Failed, duplicated, stale, and superseded runs are retained in the incident ledgers but contribute no selected rows.

The frontier-API cells, the Qwen3-32B scale cell, and the framing arm — including the `_b2` re-verification runs — are public at `https://huggingface.co/datasets/siddharthmb/2026.RA.Frontier-and-Scale-Cells`, whose card carries the contamination taxonomy of §3.6 so the per-cell fabrication status of every row is legible without re-deriving it.

The fairness-GRPO pilot (§10) is public as one dataset and two adapters. The dataset — training and held-out instance banks, per-step training logs, sampled rollout transcripts, the frozen

evaluation fleet's result tables, the Step-0 reward-soundness scores, and the transcript audit — is at https://huggingface.co/datasets/siddharthmb/2026.RA.Fairness-GRPO; the two trained LoRA adapters are at https://huggingface.co/siddharthmb/2026.RA.Fairness-GRPO-lam0 and https://huggingface.co/siddharthmb/2026.RA.Fairness-GRPO-lam1. Both adapter cards **lead with the negative** in their first paragraph — these adapters did not achieve their objective, they are worse than the untrained base model on welfare, and they are more selfish in canonical bargaining games — so nothing is published as a fairness improvement it is not. The adapters are released as a reproducible negative and as the substrate for the clipped-reward follow-up of §10.7, not as artifacts to deploy. The dataset card now leads with the headline negative result ("the experiment FAILED its preregistered success criterion", card revision `14946c58`) before documenting the reward, the held-out bank's two overlap guards and its discriminativeness screen, and the 24-step budget caveat, and its `eval_contrasts.jsonl` carries every interval in this section with its favourable/unfavourable verdict attached.

The policy-decomposition arc of §11 is not yet uploaded, and this is stated rather than omitted. Its 22 cells, the two screened banks, the 585,265-event acceptance-curve harvest, the calibrated-maximizer gate reports and the six `calmax_v2` campaign directories (864 episodes with full transcripts) all live on cluster scratch at the paths in §B; the fitted acceptance-curve artifact and its consumption interface are in-repo. The natural bundle pairs the fitted curves with the maximizer campaign that consumes them, and **that campaign has now run**, so the bundle is assemblable and the upload is simply outstanding. In the meantime the harvested events are derived data over episodes already published in the campaign bundles above, and the arc's episode-level evidence is reproducible from the banks (which are in-repo and frozen) plus the commands in §A.

The two older datasets predate the contamination audit and describe the original campaign. The episode corpus is still a faithful record of what was stored — and roughly a quarter of its turns are engine-fabricated placeholders. Anyone using it must screen with `interlens.arena.engine.gen_failures` rather than `parse_ok`; the clean `_b2` episodes are the ones to analyse, and publishing them alongside is the natural next upload. The preference corpus is unaffected by construction, because a fabricated turn carries no oracle value and cannot become a pair.