FREE-FORM COMMUNICATION IS WHY LANGUAGE MODELS OUTPERFORM BAYESIAN AGENTS IN NEGOTIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In a solver-verified five-party negotiation scenario where language models and exactly specified Bayesian negotiators occupy interchangeable seats, tables of language models reach agreement far more often, in 0.958 of games for five Claude Opus 5 seats against 0.233 for five Bayesian negotiators. Agreement requires unanimity, so closing a deal is a coordination problem and the ability of LLMs to engage in free-form talk independently of proposing or accepting deals is what allows them to perform better. Three tests support this account: making the five Opus seats mute, so they still move but send no messages, costs 0.200 in agreement rate, as much as seating one mute Bayesian agent. The second test asks what the talk channel must carry; when each seat truthfully reveals its preferences once, every all-Bayesian table reaches agreement, whereas signals that stop short of revealing preferences plateau at an agreement rate of 0.400. The same declaration protocol adds less as language-model seats replace Bayesian ones, because free-form talk already conveys preferences. Attaching a language-model interface that speaks and listens for the Bayesian seat lets its table reach agreement as often as an all-Opus table, with the Bayesian policy still conditioning every proposal. A mute Bayesian seat lowers agreement at every capability tier, and an omniscient Bayesian seat helps Haiku tables but hurts Opus tables, so an algorithmic seat helps only when it knows something the table lacks and can communicate it. Designers should give every seat a free-form communication channel, wrap any mute policy such as an algorithmic participant in a language-model interface, and not seat a policy that knows nothing the table lacks, since the defaults otherwise cost a fifth of agreements. Experimental data and the interaction framework will be released on publication.

1 INTRODUCTION

Tables of language models reach agreement in negotiation far more often than tables of exactly specified Bayesian negotiators, while the deals both reach differ little in quality. The comparison matters for systems that seat language-model (LLM) agents beside policies such as procurement rules, scheduling heuristics, or bidding agents. Our testbed is a five-party, four-issue negotiation in which each party holds a private score sheet and a private walk-away threshold, the lowest score it will accept. An exact solver generates the games so that at least one package is acceptable to all five parties and every outcome can be scored against this feasible set, and language models and algorithmic policies occupy the same seats under the same protocol (§2). On a capability ladder of tables that each seat five copies of one agent, Bayesian negotiators, which infer the others' preferences from their offers and cannot talk, close a quarter of their games, meaning they reach an agreement, while Claude Opus 5 closes nearly all of them (§3).

Because a deal requires unanimous acceptance, closing one is a coordination problem in which five parties must settle on a single package. Our hypothesis is that free-form talk, the costless, nonbinding, unverifiable messages that game theory calls cheap talk, performs this coordination, and that this is why the language models outperform the Bayesian agents. The strongest evidence is that silent tables still find acceptable packages. A census of proposals shows that mute tables propose a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

feasible package in nearly every episode but fail to agree on one, and the loss falls on games with few feasible packages (§3).

Three tests support the hypothesis, beginning with an ablation inside the language-model table so that the argument does not rest on the Bayesian baseline. Silencing all five Opus seats costs the table as much closure as replacing one Opus seat with a mute algorithmic agent (§3). The second test identifies the channel's payload as preference information. The Bayesian agents' beliefs learn only from concessions, which rarely occur, so every intervention that adds signal without revealing preferences saturates, while one round of truthful declaration, in which each seat publishes its score sheet and threshold, closes every game. Language models reveal preferences in ordinal form as a side effect of talking, so a declaration protocol helps them less the more of them are seated (§4). The third test is the converse. Returning the channel to the Bayesian seat through an Opus interface reverses the loss, as speech alone recovers a statistically resolved share of the deficit and speaking together with listening leaves the table statistically indistinguishable from the all-Opus table (§6).

It follows that an algorithmic seat helps a table of language models only when it knows something the table lacks and can communicate it. A mute Bayesian seat lowers closure at every capability tier. An omniscient seat runs the same decision machinery but is shown every party's sheet. It raises welfare at a Haiku table, is directionally positive at Sonnet, and slightly lowers welfare at Opus, and its gains arrive as closure shared with the other seats. Decision-theoretic advice delivered as a sentence of text raises Nash welfare for an 8B model by a preregistered margin, while giving the same model every counterparty's sheet does almost nothing (§5). The channel's value also depends on the reader, since for open-weight models it lowers closure wherever its effect is distinguishable from zero (§3).

Section 7 collects what these results imply for builders of multi-agent systems, Section 8 places them against cheap-talk theory and the negotiation literature, and Section 9 states the limits.

2 A NEGOTIATION WHERE EVERYTHING CAN BE SCORED

Measuring what a message channel does requires a game in which talk can be switched on and off without changing anything else, and in which the quality of every outcome is known exactly. Our testbed, built on our open-source multi-agent framework (Anonymous, 2026), is a multi-issue package negotiation. Each party i holds a private additive score sheet u_i over the issues and a private walk-away threshold τ_i. Its surplus on deal d is $x_i(d) = u_i(d) - \tau_i$, and a deal is individually rational for i when that surplus is non-negative. The normalized surplus $z_i = x_i/c_i$ divides by the party's best attainable surplus over the whole deal space. Instances are produced by an exact solver, so every episode ships with its Pareto frontier and its bargaining solutions (Nash, Nash, 1950; Kalai–Smorodinsky, Kalai & Smorodinsky, 1975; maximum Nash welfare, Caragiannis et al., 2019), and outcome quality is a distance computation rather than a judgment. The primary metric is deal rate, the fraction of episodes that end in an agreement. The others are normalized utilitarian score ("normalized score" below), normalized Nash welfare, the worst-off party's surplus, the Gini coefficient over surpluses, and the number of parties left below their threshold by a closed deal.

The main game is a five-party negotiation over four issues of datacenter capacity, with 256 possible packages, played in rounds with a rotating proposer. Each turn a seat writes a private scratchpad, may post one public message, and makes one binding move: propose a package, accept or reject a standing offer, or walk away. An episode ends when a package is accepted under the agreement rule, unanimity unless stated otherwise, when a seat walks, or at the deadline of four rounds, which triggers a forced final vote. The public message is the object of study. It is costless, non-binding, and unverifiable, the sense in which Crawford & Sobel (1982) and Farrell & Rabin (1996) use the term cheap talk, and every experiment below opens it, closes it, or changes who can use it. We distinguish this free-form *talk channel* from a *declaration protocol*, a structured variant in which a seat may publish its score sheet and threshold once in a machine-readable block that algorithmic seats can read. The declaration protocol appears only in Section 4.

The solver's guarantee makes closure, whether a table reaches any agreement, the primary outcome. Every instance admits at least one deal acceptable to all five parties, and we call such packages *feasible*. A no-deal is therefore always a coordination failure and never a justified refusal, and how surplus is split is reported as a separate question. An intervention helps, in the sense we use below,

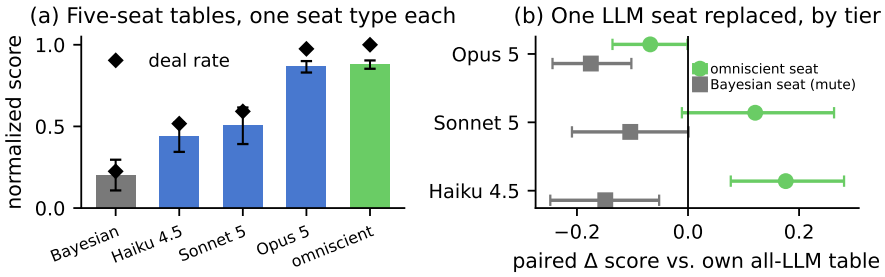

Figure 1: Tables of identical agents differ mainly in deal rate, and one omniscient seat helps weaker tables while one mute Bayesian seat hurts every tier. **(a)** Normalized score (bars, 95% instance-clustered intervals) and deal rate (diamonds, no intervals; paired intervals for the contrasts discussed are in the text) for five-seat tables of one agent type on bank v2. **(b)** Paired score change when one seat of an all-LLM table is replaced by an omniscient seat (green) or the mute Bayesian seat (gray); the axis names the seats kept, and zero is their all-LLM table. Each bar and point is 120 episodes (24 instances $\times$ 5 seeds).

when it raises closure without raising the number of parties left below their thresholds. We then report whether the intervention moved the split and who received the difference.

Alongside the language models we field *algorithmic agents*, exactly specified policies assembled from the automated-negotiation literature so that LLM play can be compared against a reference with no model in it (Appendix A). The headline agent, called the *Bayesian* agent below, keeps a posterior over a grid of opponent types, updated by a concession likelihood that penalizes any hypothesis under which an opponent's own utility rose (Hindriks & Tykhonov, 2008; Chang & Fujita, 2023). It accepts by optimal stopping against the offer distribution its posterior induces (Baarslag & Hindriks, 2013; McCall, 1970), and it proposes by exact expectimax over remaining rounds, the deal space, and the type posterior. The same recursion with every sheet known is the *omniscient* agent, which serves as the ceiling. Algorithmic seats make binding moves only. They are *mute*, neither reading nor writing the talk channel, unless an experiment explicitly grants them an interface to it.

All comparisons share one set of statistical conventions. Contrasts are paired, comparing episodes on the same instance and seed. Every interval is a 95% percentile bootstrap over instance clusters (each instance's five seeds resampled together; 10,000 draws, 2,000 for the lever and declaration experiments of Section 4). An effect is *resolved* when its interval excludes zero, and non-inferiority means a deal-rate lower bound above -0.05. Each condition typically runs 120 episodes, 24 instances by five seeds. Two frozen instance banks, fixed sets of generated games, carry the main results. Bank v1 carries every channel ablation and bank v2 the capability ladder and per-tier seat contrasts, and no contrast crosses banks or engine versions (Appendix B).

3 Tables of different agents differ in closure, and closure runs through talk

This section asks where the differences between tables of agents come from and whether they run through the talk channel. Tables of identical agents differ mainly in how often they close (Figure 1a). The Bayesian table, built from the negotiation literature's own components, closes 0.225 of its games at a normalized score of 0.197. Haiku 4.5 reaches 0.438 and Sonnet 5 a statistically indistinguishable 0.507, with paired deal rates differing by -0.075 $[-0.150, 0.000]$. Opus 5 reaches 0.865 and closes 0.975, consistent with the frontier-tier multi-agent gains reported by Anthropic (2025), and the omniscient table closes every game at 0.879. Almost all of the algorithmic table's shortfall is closure rather than judgment. On the games the Bayesian table does close, its score trails Opus by 0.051. The unconditional gap is roughly 0.67, computed on another engine version of the same bank, so about 92% of the welfare gap is unclosed deals. No homogeneous table ever signs a package that leaves a party below its threshold. The Opus table also does not buy its higher closure with worse splits. On games that both tables in a comparison closed, Opus's deals leave the worst-off seat better

off than Sonnet's by $+0.094$ [$+0.046$, $+0.144$] in normalized surplus and better off than Haiku's by $+0.123$, with lower inequality (transcripts in Appendix E).

The open-weight end of the scale suggests that this ordering reflects communication rather than raw competence, because there the channel's effect reverses sign. In an earlier six-party variant, opening the talk channel *lowered* deal rate wherever resolved: Qwen3-8B by -0.233 [-0.450, -0.100], Qwen3-32B by -0.083 [-0.150, -0.050], and Ministral-8B by -0.100 [-0.150, -0.050]. For these models the channel is a distraction, and it becomes an asset only at the frontier.

Muting an all-Opus table reproduces the entire cost of seating a mute algorithmic agent. On bank v1, five Opus seats close 0.958 of their games. Replacing one seat with the mute Bayesian agent costs the table -0.192 [-0.275, -0.108] in deal rate, which could mean the algorithmic agent is simply a worse negotiator. The preregistered control keeps the five Opus seats and disables every public message, so that only the channel changes. Its falsifier was a deal rate near or below the mixed table's 0.767, which would reject the hypothesis that mute Opus seats still win decisively. The muted table closes 0.758, below that bound, a cost of -0.200 [-0.325, -0.092]. This cost is indistinguishable from the cost of the algorithmic seat (difference -0.008 [-0.108, $+0.083$]). The muted tables fail by withholding assent. No seat walked in any of the 120 episodes, and the tables ran to the deadline, where under unanimity withheld assent is a veto. Among closed deals, the split was slightly worse (Gini $+0.023$ [$+0.003$, $+0.043$]), so talk also improves terms.

The cost of muting concentrates on games where feasible packages are scarce. The generator tags each instance before any episode is played, and the groups these tags define are the strata below. Easy, medium, and hard are tertiles of a difficulty index dominated by how few of the 256 packages clear every threshold. Further tags mark instances with no clear win-win, with a single pivotal seat, or with high preference conflict (Appendix C). Muting costs almost nothing on easy games (-0.025 [-0.100, $+0.050$]). It costs -0.475 [-0.675, -0.275] on the hard stratum and about -0.47 where no clear win-win exists or one seat is pivotal. The deficit from one algorithmic seat has the same profile (Appendix Figure 4). The algorithmic seat's cost recurs at every capability tier (§5).

A census of every tabled package shows that what the channel supplies is coordination rather than discovery. Call the packages that clear all five thresholds the zone of agreement. A no-deal is a discovery failure if no package in the zone was ever proposed, and a coordination failure if one was proposed and the table still did not close. Discovery is nearly free on this bank. Five speaking Opus seats table a package in the zone in 0.992 of episodes, with the median first such proposal in round one. Even with all five seats mute, the zone is tabled in 0.908 of episodes. Of the -0.200 cost of muting all five, -0.125 [-0.207, -0.051] is the coordination term and -0.075 [-0.169, $+0.000$] the discovery term. For the table with one Bayesian seat, the split is -0.147 [-0.222, -0.072] against -0.045 [-0.107, $+0.007$]. Across seven experimental conditions, 81 of 107 no-deal episodes had a feasible package on the table before the forced final vote. The mute tables also search differently, proposing more packages (8.7 against 7.5 per episode) with a smaller fraction in the zone (0.306 against 0.410). Without talk, seats find the zone by enumeration but cannot settle inside it. Section 4 asks why agents without the channel cannot coordinate otherwise.

4 Why exact negotiators cannot close without talk, and what talk must carry

This section explains why the Bayesian agents cannot close without talk and identifies what the channel must carry. The Bayesian agents fail because their beliefs receive too little evidence to move, even though the games are solvable. Their posterior learns only from concessions, and a seat concedes less than once per episode, so the posterior stays near uniform and every seat proposes selfishly until the deadline. At a 32-round horizon a seat makes 24 proposals but only 1.56 distinct ones. A repeated offer contributes nothing to a concession likelihood, so the posterior ends at 1.6 times uniform, believing each opponent accepts a tenth of packages when the truth is two thirds. A longer horizon does not help. Lengthening the deadline from 4 to 64 rounds moves closure only between 0.233 and 0.250, with overlapping intervals and no trend, and a 128-round condition extends the null. The omniscient table closes every game at every deadline, almost always in round one, which is the Rubinstein (1982) control. Better search does not help either. Tabling each instance's welfare-optimal package before anyone speaks changes the Bayesian table's closure by exactly 0.000 [-0.050, $+0.050$] and an Opus table's score by $+0.003$ [-0.047, $+0.052$]. The fail-

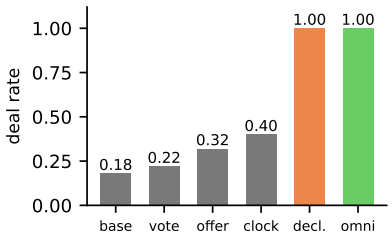

Figure 2: Only truthful declaration lifts all-algorithmic tables to full closure. Deal rate, one intervention per bar, 50 episodes each (10 instances $\times$ 5 seeds). Gray: non-revealing signals, namely the published agent, votes in the posterior, final-offer support, and a deadline-indexed concession schedule (full-bank value in the text). Orange: truthful declaration. Green: omniscient ceiling.

Table 1: The declaration protocol adds less closure as language-model seats replace algorithmic ones. Deal rate with and without the protocol, by who sits at the table (50 episodes per entry, same frozen instances). The free-form talk channel is open to LLM seats in every row; the columns differ only in whether seats may also publish a structured sheet, and the last column is their difference. The four-Opus rates exclude only episodes an API refusal at the final vote made arithmetically unclosable (1 of 50 talk only, 13 of 50 with declaration).

Seating	talk only	with declaration	difference
5 algorithmic	0.180	**1.000**	$+0.820$
1 Opus + 4 algorithmic	0.240	0.700	$+0.460$
4 Opus + 1 algorithmic	0.694	0.946	$+0.252$

ure is also not the textbook inefficiency of two-sided private information. The best posterior that the same hypothesis grid admits closes every audited failure, and 37 to 42% of failed games had a feasible package on the table that nobody re-tabled. The agents lack information about the other seats' preferences, which is exactly what the channel transmits.

Every attempt to relieve that shortage without revealing preferences saturates well short of closure (Figure 2). Folding the public votes, a signal five to thirteen times denser than concessions, into the posterior moves closure from 0.180 to 0.220 against a predicted 0.5 to 0.8. Votes fall short because they speak to only one of three hypothesis dimensions. Final-offer support reaches 0.320, and a deadline-indexed concession schedule reaches 0.400 on the subset of Figure 2 and 0.408 [0.283, 0.542] on the full bank. Combining these interventions lowers each one's effect rather than adding them. Declaration, in which each seat publishes its true sheet and threshold once and the others read it, is the one design that closes every game. Closure goes from 0.180 on the subset to 1.000 on the full bank, all 120 episodes and hence a zero-width interval, with zero threshold violations and a score of 0.857 against the omniscient 0.879. Declaration's small shortfall from the omniscient score is timing rather than judgment, since the informed policy still runs its optimal-stopping holdout to round 4.98 where the omniscient agent closes at 1.40. The channel therefore has to carry legible preferences, and extra signal of other kinds does not substitute for them.

Language-model seats already transmit much of this payload on their own, as offering them the same protocol shows. Table 1 crosses who sits at the table with whether the declaration protocol is available, and the protocol's value falls from $+0.820$ among algorithmic seats to $+0.252$ with four Opus seats and one algorithmic one. The protocol and the language-model seats are partial substitutes for the same bottleneck, because Opus seats already reveal preferences through free-form talk. We parsed the 1,503 public messages of the four ordinary Opus seats in the interface condition of Section 6 into quote-anchored preference claims, with instance-clustered proportions. Of these messages, 0.810 [0.794, 0.825] state at least one preference, every opening message does so, and by the end of a game 0.889 [0.870, 0.907] of speaker-issue pairs have been covered. The content is ordinal, however. Of 4,247 claims, two contain a number and neither is a valuation, and threshold language without numbers appears in only 4.4% of messages. Free-form talk supplies rankings and the declaration protocol supplies magnitudes, which is why the protocol still adds a quarter of a deal rate at a table with four language-model seats.

The value of what a speaker reveals depends on who reads it. When a lone Opus seat declared in round one against four algorithmic readers the table closed 18 of 18 games, and when it never declared closure fell to 0.250. The same withholding cost nothing, 0.923, when the readers were other language models, which bargain over unquantified prose without difficulty. The algorithmic agents can act only on numbers while the language models can also act on words, so a seat's silence

costs more the more literal its audience is. The next two sections ask what an algorithmic agent contributes to a table of language models when it cannot speak, and what changes when it can.

5 When an algorithmic seat helps a table of language models

This section asks when adding an algorithmic seat helps a table of language models. A mute Bayesian seat lowers the score of a language-model table at every capability tier, as the results so far predict for a policy that cannot talk. Figure 1b tests this on bank v2 by replacing one seat of an all-LLM table, five seats of one model, with an algorithmic one in paired episodes. The mute Bayesian seat lowers normalized score by -0.175 [-0.244, -0.102] at Opus, -0.104 [-0.209, $+0.001$] at Sonnet, and -0.149 [-0.248, -0.052] at Haiku, with its deal-rate effect resolved at all three. The loss comes through closure, and at an Opus table its four co-players lose -0.133 [-0.180, -0.083] in surplus because the table stops closing. The seat withholds consent rather than walking away.

Relaxing the agreement rule recovers most of this deficit, but it does so by overruling the dissenting seat rather than informing it as talk does. Under a 4-of-5 quorum on bank v1, the other four seats can close over the Bayesian seat's objection. About 70% of its deficit then disappears ($+0.130$ [$+0.053$, $+0.208$] of -0.186), and the Bayesian seat is outvoted in 0.542 [0.470, 0.615] of episodes. This route carries a cost that talk never charges. Episodes leaving a party below its threshold rise from zero to 0.242 [0.133, 0.358] at the all-LLM table and 0.375 [0.258, 0.500] at the mixed one.

The omniscient seat shows what an algorithmic participant can contribute when it lacks nothing but language. Its paired effect on score changes sign down the ladder and is resolved at both ends, with monotone point estimates between. The effect is -0.068 [-0.136, -0.001] at an Opus table, $+0.121$ [-0.011, $+0.263$] at Sonnet, and $+0.176$ [$+0.077$, $+0.281$] at Haiku. The trend continues off this bank to $+0.449$ [$+0.209$, $+0.692$] at a Qwen3-8B table, on an 18-episode exploratory subset under an earlier protocol. At Haiku the gain is entirely closure and is shared, with the omniscient seat gaining $+0.155$ [$+0.065$, $+0.248$] and its four co-players $+0.134$ [$+0.061$, $+0.206$]. The preregistered prediction that an algorithmic seat helps more the weaker its co-players held for omniscience and was refuted for the merely Bayesian seat. An Opus table already knows how to close and is slightly disrupted by a seat that proposes without explaining. A Haiku table does not, and a seat that always proposes a feasible package gives it a point to coordinate on.

The same pattern holds when the algorithmic computation reaches the table as text instead of as a seat. In six-party full-information games with Qwen3-8B in every seat, we named the discrete Nash-bargaining package to every seat and compared this against an interface- and token-matched placebo candidate. Naming the package raised normalized Nash welfare by $+0.154$ [$+0.120$, $+0.188$], with deal rate up $+0.075$, below-threshold agreements down -0.083, and the effect positive in 43 of 48 held-out games. The claim and its success criterion were frozen before any confirmatory episode ran. The placebo matters because a mediocre candidate already raises welfare by $+0.082$ over no candidate, a focal-point effect that the primary contrast removes. A small model can act on the fair deal when the package is named to it, but it cannot find that deal from the inputs alone. Handing every seat every counterparty's sheet moves welfare by a mean absolute 0.035 of unstable sign across six open-weight models, resolved positive only for Qwen3-4B. An algorithmic table converts the same access into about $+0.31$. These comparisons rest on three instance clusters each and serve only as scoping estimates. The advice result also does not extend to a realistic one-sided setting. There, a fallible Bayesian planner that sees only its own principal's sheet advised one seat in bilateral private-information games. It changed that seat's surplus by -0.006 [-0.149, $+0.135$] and closure by exactly zero, and a rerun without token caps still failed every preregistered criterion.

Together these results give three conditions under which an algorithmic agent helps a table of language models. The agent must know something the table does not, most often where a feasible or fair package lies. The table must be weak enough to need that knowledge, and the knowledge must arrive in a form the table acts on, whether a proposal that is always feasible or a sentence naming the package. A Bayesian policy fails the first condition because its beliefs never learn enough, and as a mute seat it fails the third, which Section 6 repairs directly.

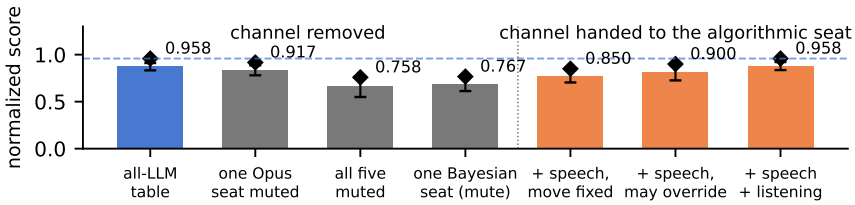

Figure 3: Returning the channel to a mute Bayesian seat through an Opus interface recovers its deficit step by step, and adding listening closes it. Every condition is 120 episodes (24 instances $\times$ 5 seeds) on bank v1, paired on instance and seed; bars are normalized score with 95% instance-clustered intervals, diamonds are deal rate (paired intervals in §3 and §6). Left of the dotted line the channel is removed from an all-Opus table; right of it, it is returned to a mute Bayesian seat through an Opus interface, one capability per bar (conditions defined in Appendix Table 3).

6 Handing the channel to the algorithmic seat

This section asks whether giving the Bayesian seat access to the channel repairs its deficit. Muting a single Opus seat first shows that silence explains only about a fifth of the algorithmic seat's deficit. The equal effect sizes of Section 3 do not prove that silence is what the algorithmic seat pays for, so this test disables one Opus seat's messages while it still reads the other four. Its preregistered falsifier, that a cost within a tenth of a deal rate of zero would reject the hypothesis that the seat's deficit is its silence, was met. One muted Opus seat costs the table -0.042 [-0.092, $+0.000$] in deal rate, against the algorithmic seat's -0.192. On the same games the muted Opus seat beats the algorithmic seat by $+0.150$ [$+0.075$, $+0.225$], so the rest of the algorithmic seat's deficit comes from the Bayesian policy's decision rule. The channel's value is graded in the number of speaking seats rather than lost at the first silent one, since muting the remaining four costs a further -0.158 [-0.267, -0.050]. What the silent Opus seat loses is share rather than closure, as its own surplus share falls by -0.093 [-0.157, -0.025] while its co-players are unchanged.

Returning the channel to the Bayesian seat in three steps recovers its deficit progressively (right half of Figure 3). Each step wraps an Opus model around the frozen Bayesian policy. In the first, the model writes the public message while the policy's chosen move is submitted unchanged, so that speech is the only addition. The table closes 0.850 [0.783, 0.909], a resolved gain of $+0.083$ [$+0.008$, $+0.158$] over the mute policy. In the second, the model is handed the policy's move, told to play it, and allowed to depart. It played the recommended action on 86.1% of 483 advised turns and otherwise mostly conceded. The table then closes 0.900, a gain of $+0.133$ [$+0.050$, $+0.217$] in deal rate and $+0.125$ [$+0.046$, $+0.200$] in score, about two thirds of the deficit. The increment from being allowed to change moves, $+0.050$ [-0.017, $+0.117$], is unresolved at this sample, but the moves suppressed in the fixed-move condition show what it consists of. Of 62 turns where the model would have departed from the rule, 47 were last-round acceptances of a standing offer. In 8 of the 10 games closed by the seat allowed to override but not by the fixed-move seat, such an offer was standing. Speech alone cannot supply the deadline acceptances the policy declines.

In the third step the interface also listens, and the seat then closes as often as the all-LLM table. A separate structured call parses every public message into preference evidence per seat, issue, and option, each row anchored to a verbatim quote, and a rebuilt planner conditions the policy on that evidence. The parser was validated offline before the experiment ran and audited live at 0.886 [0.874, 0.899] direction accuracy over 9,609 rows, with a hallucination rate of 0.0004. The table closes 115 of 120 games, 0.958, the all-LLM table's rate on this bank, and the paired difference of $+0.000$ [-0.042, $+0.042$] rules out a gap larger than about four points in either direction. Its score of 0.881 [0.836, 0.921] is the highest of any condition with a single treated seat. Listening adds a resolved $+0.069$ [$+0.004$, $+0.155$] in score over speaking alone, with overrides allowed in both conditions, so neither direction of the channel is measured in isolation.

The gains from this repair go to the other four seats, because the interface closes deals by conceding on the policy's behalf. At every step the interfaced seat's own surplus share is unresolved by the preregistered test, with a half-width near 0.2 that exceeds the 0.05 bar. Its co-players' surplus, in contrast, rises resolvably, by $+0.059$ [$+0.008$, $+0.111$] when speech is added with the move

fixed, by $+0.097$ [$+0.038$, $+0.154$] with overrides allowed, and by a further $+0.054$ [$+0.004$, $+0.120$] with listening. Of the 88 turns on which the listening interface overrode the Bayesian policy's planned proposal, 68 conceded, 15 held level, and 5 took more, at a median cost of 21 of the seat's own surplus points. The listening seat's deals come out flatter than the speaking-only seat's (Gini -0.017 [-0.035, -0.002]). Decision support handed to one seat of a unanimity game thus behaves as a public good. Advising all five seats does not multiply the effect, because every seat then concedes, on 284 of 380 measurable overrides. The agreements then come out less equal than with a single advised seat (Gini $+0.015$ [$+0.005$, $+0.025$]) and null against the all-LLM table on every endpoint.

A final control shows that truthful messages do not help without a language model to write and read them. We gave the frozen Bayesian policy itself verifiably truthful templated messages in a sequence of increasing revelation, from announcing its concession schedule to naming its top option per issue. In the last step we let the policy parse counterparties' templated statements into its beliefs. On all-algorithmic tables every speaking condition, and a fluent state-independent babbler, is episode-for-episode identical to the mute baseline, as it must be when no decision rule reads the channel. Only mutual listening moves closure at all, from 0.217 to 0.300 ($+0.083$ [-0.050, $+0.217$], unresolved), a tenth of what published sheets recover. The interface of Figure 3 works because it supplies fluency and content together, producing messages that language models can read and act on.

7 What this means for building systems with several agents

This section states what the results imply for designers of multi-agent systems. The message channel between capable agents is a protocol decision with a measurable price, and the price falls on hard games. Removing talk from a table of frontier models costs a fifth of its closure and almost half of it on hard instances, with no measurable cost on easy ones (§3). A system that disables free-form messages between its agents for cost, latency, or auditability reasons should therefore expect its failures to concentrate where feasible agreements are scarce. Neither seeding the optimal package, tested at algorithmic and Opus tables, nor a longer horizon, tested at algorithmic tables, compensates (§4). The one protocol-level substitute that does close tables, a relaxed agreement rule, overrules a seat instead of informing it and leaves parties below their thresholds in a quarter to three eighths of episodes (§5). Below the frontier the same channel can hurt (§3), so the decision depends on who is seated.

The channel needs to carry preference information, and that information is not automatically honest. Nothing short of preference revelation closed the all-algorithmic tables, and the language models close because their free-form talk reveals preferences in ordinal form in nearly every message (§4). A designer who wants agents to coordinate should make it easy for them to state what they care about and in which direction. Free-form talk leaves magnitudes unstated, so a structured declaration can still add closure at a table of frontier seats. Across 182 audited declarations by Opus seats, the value table was copied nearly exactly, while the one number nothing at the table could check, the walk-away threshold, was overstated 39 times and understated never. The overstatement did not raise the speaker's surplus (0.182 against 0.205 for its counterparts, unpaired), and only an explicit instruction to misreport turned deception into gain ($+0.067$ [$+0.021$, $+0.119$]). A protocol should verify what it can and expect unverifiable quantities to be shaded in the direction incentives point.

An algorithmic participant should not be added to a table of language models unless it can communicate. A mute Bayesian seat lowered closure at every capability tier (§5). A fifth of that cost was its silence, and the rest was its handling of talk it could not read and of last-round offers it declined (§6). Whether the seat can help at all depends on whether it knows something the table lacks. An omniscient seat helped a Haiku table but slightly hurt an Opus table, and a sentence naming the fair package raised an 8B table's Nash welfare where handing it every sheet did not. Useful algorithmic help therefore takes the form of an answer rather than more information, delivered where the models already look. Its value falls to zero or below as the table's capability rises toward the frontier.

A language-model interface repairs the algorithmic seat. Wrapping the Bayesian policy in a language model that speaks and reads for it turned a seat that cost the table a fifth of its closure into one statistically indistinguishable from a frontier seat (§6). The interface costs one extra inference per turn, about \$0.48 per episode. The gains went to the other seats because the interface concedes on the policy's behalf. The repair therefore does not make an algorithmic seat bargain hard for itself.

8 Related work

This section places the results against cheap-talk theory and prior work on negotiation. Cheap-talk theory predicts that talk in our game should be informative, and our results add a measurement of the reader's role that the theory does not provide. Crawford & Sobel (1982) bound how much unverifiable communication can transmit as interests diverge, and Farrell & Rabin (1996) catalog when costless messages move equilibria at all. Our game sits in the regime where that theory predicts informative talk, since interests are aligned on closure and diverge only on the split. Talk here is informative and biased only on the one number nothing can check. The channel's value was set less by the speaker's incentive to reveal than by the reader's capacity to use it (§4, §6).

Our game and agents build on negotiation benchmarks and the automated-negotiation literature. The game construction descends from end-to-end negotiation dialogues (Lewis et al., 2017) and the multi-issue LLM-stakeholder benchmark of Abdelnabi et al. (2024), adding exact solver-verified analyses per instance, algorithmic seats under the identical protocol, and preregistered falsifiers. The algorithmic agents implement the automated-negotiation literature directly (opponent modeling, Hindriks & Tykhonov, 2008; Chang & Fujita, 2023; Baarslag et al., 2016; acceptance, Baarslag & Hindriks, 2013; Baarslag et al., 2013; McCall, 1970; concession strategies, Faratin et al., 1998; de Jonge, 2022; deception-robust tempering, Hua et al., 2024), with bargaining anchors from Nash (1950); Kalai & Smorodinsky (1975); Caragiannis et al. (2019) and equilibrium anchors from Rubinstein (1982); Baron & Ferejohn (1989); Banks & Duggan (2000); Okada (1996); Sandholm & Vulkan (1999). Anthropic (2025) reports frontier-tier gains from multi-agent systems, and we locate one mechanism behind such gains in the message channel, which below the frontier can lower closure. Our interface resembles CICERO (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022), which wraps a strategic planner in a language model that speaks and reads for it. In CICERO the planner is trained with the dialogue model, and both play Diplomacy, a game with binding moves, as one agent. Here a frozen off-the-shelf model voices an exactly specified Bayesian policy and keeps its move rule, which isolates what the channel adds to a fixed policy.

9 Limitations and conclusion

The results are limited in scope and leave several controls unrun. The mixed-seat results come from one game family, one frozen bank, and one frontier model, Claude Opus 5, in the treated conditions. The per-tier contrasts span three tiers of one model family, and per-stratum readings rest on six to eight instances each. The interface comparison is not token-matched, and the interfaced seat inherits a parser error rate that a more deceptive table would test harder. The Bayesian reference agent learns only from concessions and its posterior stays near uniform, so its low closure is a property of this reference policy. The results therefore bound what such a policy gains from the channel, not what every algorithmic negotiator would gain. Where the interface may override the policy's proposals, the closure recovered measures the interfaced seat as a whole. The one-omniscient contrasts keep a tie-break caveat described in Appendix B, and Appendix D states the validity rules every hosted experiment was held to. The advice result is a cooperative, full-information, everyone-advised result for one 8B model, and its realistic one-sided counterpart failed. The templated-speech ablation was not extended to tables of language models. Three conditions that would complete the channel ablation have not been run, namely a deaf-and-mute Opus seat that would separate the interface from the model's own negotiating, muting two and three of five seats, and a larger fixed-move condition.

Within these limits, the gap between language-model and Bayesian tables lies in reaching agreement rather than in the terms of the deals reached. Because unanimity makes closing a coordination problem, and silent tables still propose a feasible package in nearly every episode, what talk supplies is agreement on a package already found. Muting the language models reproduces the Bayesian seat's deficit, declaration shows that the channel's payload is preference revelation, which language models supply in ordinal form and exact policies can neither supply nor use, and an interface that speaks and listens for the Bayesian seat removes the deficit. An algorithmic seat is therefore an asset only when it knows something the table lacks, the table needs that knowledge, and the knowledge reaches the others in a form they act on, whether a proposal, a sentence of advice, or a language model speaking and listening on the seat's behalf.

Reproducibility statement

Every protocol, agent, and statistical choice behind the results is specified in the paper. Appendix A gives the game protocol, the metrics, and every algorithmic agent by the equations the code implements. Appendix B lists the frozen instance banks and engine versions behind each result, and Appendix C defines each condition of the channel ablation with its preregistered falsifier and API cost, together with the instance tags used for the strata. Section 2 defines the pairing and the bootstrap intervals, the remaining preregistered falsifiers are stated beside their results in Sections 3 to 6, and Appendix D gives the validity rules every hosted experiment was held to. The framework, the frozen banks, the episode-level records with full transcripts, and the interactive episode browser will be released on publication (Appendix F).

Ethics statement

This work studies negotiation among AI agents on synthetic games generated by an exact solver, and it involves no human subjects and no personal data. One finding has dual-use relevance: language-model seats overstated their unverifiable walk-away thresholds unprompted, and an instruction to misreport turned that into a measurable gain (Section 7). The same result tells protocol designers which claims to make verifiable, and we report it for that purpose.

AI use statement

Generative AI tools were used to help draft and revise the text of the paper, write code relating to experiment implementation, analysis, and figure generation, and to fact-check results against experimental artifacts. Separately, language-model-powered agents were also studied as experimental subjects in this work.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/984dd3db213db2d1454a163b65b84d08-Abstract-Datasets_and_Benchmarks_Track.html.

Anonymous. A framework for scaffolding and interpreting multi-agent conversations. Anonymized for review; to be released on publication, 2026.

Anthropic. How we built our multi-agent research system. https://www.anthropic.com/engineering/multi-agent-research-system, June 2025.

Tim Baarslag and Koen Hindriks. Accepting optimally in automated negotiation with incomplete information. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2013)*, pp. 715–722, 2013. doi: https://doi.org/10.5555/2484920.2485033.

Tim Baarslag, Koen Hindriks, and Catholijn Jonker. Acceptance conditions in automated negotiation. In *Complex Automated Negotiations: Theories, Models, and Software Competitions*, volume 435 of *Studies in Computational Intelligence*, pp. 95–111. Springer, 2013. doi: https://doi.org/10.1007/978-3-642-30737-9_6.

Tim Baarslag, Mark J. C. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5):849–898, 2016. doi: https://doi.org/10.1007/s10458-015-9309-1.

Jeffrey S. Banks and John Duggan. A bargaining model of collective choice. *American Political Science Review*, 94(1):73–88, 2000. doi: https://doi.org/10.2307/2586381.

David P. Baron and John A. Ferejohn. Bargaining in legislatures. *American Political Science Review*, 83(4):1181–1206, 1989. doi: https://doi.org/10.2307/1961664.

Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation*, 7(3):12:1–12:32, 2019. doi: https://doi.org/10.1145/3355902.

Shengbo Chang and Katsuhide Fujita. A scalable opponent model using Bayesian learning for automated bilateral multi-issue negotiation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, pp. 2487–2489, 2023. doi: https://doi.org/10.5555/3545946.3598976.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. doi: https://doi.org/10.2307/1913390.

Dave de Jonge. An analysis of the linear bilateral ANAC domains using the MiCRO benchmark strategy. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pp. 223–229, 2022. doi: https://doi.org/10.24963/ijcai.2022/32.

Peyman Faratin, Carles Sierra, and Nicholas R. Jennings. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3–4):159–182, 1998. doi: https://doi.org/10.1016/S0921-8890(98)00029-3.

Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996. doi: https://doi.org/10.1257/jep.10.3.103.

Koen Hindriks and Dmytro Tykhonov. Opponent modelling in automated multi-issue negotiation using Bayesian learning. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pp. 331–338, 2008. doi: https://doi.org/10.5555/1402383.1402433.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. Game-theoretic LLM: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024. URL https://arxiv.org/abs/2411.05990.

Ehud Kalai and Meir Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43(3):513–518, 1975. doi: https://doi.org/10.2307/1914280.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, 2017. doi: https://doi.org/10.18653/v1/D17-1259.

John J. McCall. Economics of information and job search. *The Quarterly Journal of Economics*, 84 (1):113–126, 1970. doi: https://doi.org/10.2307/1879403.

Meta Fundamental AI Research Diplomacy Team (FAIR) et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: https://doi.org/10.1126/science.ade9097.

John Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: https://doi.org/10.2307/1907266.

Akira Okada. A noncooperative coalitional bargaining game with random proposers. *Games and Economic Behavior*, 16(1):97–108, 1996. doi: https://doi.org/10.1006/game.1996.0076.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982. doi: https://doi.org/10.2307/1912531.

Tuomas Sandholm and Nir Vulkan. Bargaining with deadlines. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pp. 44–51, 1999. URL https://cdn.aaai.org/AAAI/1999/AAAI99-007.pdf.

A ALGORITHMIC AGENT SPECIFICATION

All policies live in the negotiation module of our framework (Anonymous, 2026); each module's docstring cites its primary source, and the equations below are the implemented ones.

Protocol and metrics in full. A game has J issues with finite option sets; a deal d picks one option per issue from the deal space $\mathcal{D}$, and $u_i(d) = \sum_j u_{ij}(d_j)$. Seats act in rounds with a rotating proposer; offers are addressed by id so acceptance is never ambiguous. LLM seats read a rendered view of the public state plus their own sheet; algorithmic seats read the same state machine-readably but cannot use the talk channel unless an experiment grants it. Normalized score is utilitarian $\sum_i z_i$ scaled to $[0,1]$ with no-deal $= 0$; normalized Nash welfare is $(\prod_i z_i)^{1/n}$, zero if any $z_i \le 0$, with z_i there normalized by the best surplus over the individually rational set; threshold violations count, on a closed deal, the parties whose surplus is below their own threshold, and episodes with no deal contribute none.

Opponent types and beliefs (Hindriks & Tykhonov, 2008; Chang & Fujita, 2023; Baarslag et al., 2016). A hypothesis $h = (w, s, \theta)$ assigns opponent utility $u_h(d) = \sum_j w_j\, e_{s_j}[d_j] \in [0,1]$: weights w come from issue-importance rankings mapped to linearly decaying values, evaluator shapes $s_j \in \{\text{uphill}, \text{downhill}, \text{triangular}\}$ give per-option values in $[0,1]$, and the reservation level $\theta \in \{0.35, 0.55, 0.75\}$ sets acceptance. Offers update the posterior with the concession likelihood

$$P(b_t \mid h, b_{t-1}) \;\propto\; \exp\!\big(-\delta^{+}/2\sigma^2\big), \qquad \delta^{+} = \max\big(0,\; u_h(b_t) - u_h(b_{t-1})\big), \tag{1}$$

with $\sigma = 0.25$; a uniform floor (10^{-3}) is mixed in after every update, and each observation's log-likelihood can be tempered by $\lambda < 1$ against deceptive traces (Hua et al., 2024); the headline agent runs $\lambda = 1$ and the interface experiments 0.7. Large grids factorize into weight, shape, and threshold marginals, each scored against the mean of the others (Hindriks & Tykhonov, 2008). Each opponent gets an independent model, and a frequency model runs alongside as the cheap control (Baarslag et al., 2016).

Acceptance (Baarslag & Hindriks, 2013; McCall, 1970). With j rounds remaining, discount δ, and per-round cost C,

$$v_0 = 0, \qquad v_j = (1-\delta)\,\text{flow} + \delta\,\mathbb{E}_{X\sim F_{j-1}}\big[\max(X, v_{j-1})\big] - C, \qquad \text{accept } x \iff x \ge v_j, \tag{2}$$

where F_j is the offer-surplus distribution the posterior induces. At $\delta{=}1$ and zero flow this is Baarslag & Hindriks (2013) Eqs. (1)–(3); the uniform-opponent closed form $v_j = \frac{1}{2} + v_{j-1}^2/2$ is reproduced by unit test. With a hard deadline and no discounting the unique equilibrium is brinkmanship (Sandholm & Vulkan, 1999), which is why the rule carries a discount factor.

Best response and the omniscient agent. Exact expectimax over rounds remaining, the deal space, and the type posterior. Each round a rotating proposer offers; responders accept iff the offer beats their discounted continuation; a deal closes iff the quorum and every veto seat accept; no-deal at the deadline pays zero. The full-information recursion is the omniscient policy. Posterior-averaged exact induction is used instead of sampling to avoid strategy-fusion bias; at $|\mathcal{D}| \approx 256$, $T \le 24$, and roughly 6×10^3 types per opponent the computation is 10^6 to 10^7 elementary operations. A Banks–Duggan equilibrium oracle (Banks & Duggan, 2000; Baron & Ferejohn, 1989; Okada, 1996) anchors the analysis layer; under unanimity it is an analysis anchor, not a normative claim. Scripted opponents implement time-dependent concession (Faratin et al., 1998), MiCRO (de Jonge, 2022), and the AC-next, AC-time, and AC-combi acceptance conditions (Baarslag et al., 2013).

B BANKS, VINTAGES, AND WHICH NUMBER IS WHICH

The same homogeneous table appears in this paper at slightly different levels because it was measured on different frozen banks and engine vintages, that is, versions of the game engine. Table 2 lists them; no contrast in the paper crosses a row. The 0.217 and 0.225 anchors for the all-Bayesian table are two engine vintages of the same policy on the same bank-v2 instances, and 0.233 is the same

policy on bank v1; none was differenced against another. The all-omniscient row of Figure 1a is the tie-break-guarded re-score (0.879, from a pre-guard 0.888), which repaired a forced-final ballot the reference agent had been casting on the wrong offer; the one-omniscient contrasts of Figure 1b were not re-run under the guard and keep that caveat.

Table 2: Frozen banks used in this paper. The all-Bayesian rate differs across rows because the instances and the engine version differ. Every paired contrast is within one row. Deal rates are unconditional; the two homogeneous tables are five-seat.

bank	episodes / condition	all-Bayesian	all-Opus 5	used for
datacenter v1 (five-condition campaign)	24 inst. × 5 seeds	0.233	0.958	mute controls, one-seat deficit −0.192, zone census (§3); 4-of-5 quorum conditions (§5); channel ladder (Fig. 3, §6)
datacenter v2, quorum-sweep vintage	24 × 5	0.217	—	templated-speech ladder (§6)
datacenter v2 (ladder and grid)	24 × 5	0.225	0.975	capability ladder and per-tier seat contrasts (Fig. 1, §3, §5)
agent-variants bank	10 inst. × 5 seeds	0.180	—	closure audit and levers (Fig. 2); declaration × seating (Table 1); honesty audit (§7)

C THE CHANNEL LADDER: CONDITION DEFINITIONS AND STRATA

Table 3 defines each condition of Figure 3. "Move fixed" means the Bayesian policy's move is submitted whatever the model returns, "may override" means the model is told to play that move but may depart from it, and "listening" means a separate parse of the public channel feeds the policy. Figure 4 breaks the two channel-removal effects of Section 3 down by the bank's instance tags. The tags are assigned by the instance generator from exact enumeration of the 256 packages, before any episode is played, and they overlap. *Easy*, *medium*, and *hard* are tertiles of a declared difficulty index whose largest weight is the scarcity of packages that clear every threshold. The remaining tags mark the top quartile of one geometric component each: *high-conflict* for pairwise anti-correlation of the seats' utilities over packages, *no-clear-win-win* for the share of each seat's top-decile packages that clear no full set of thresholds, *pivotal-seat* for the number of otherwise feasible packages that one seat's threshold alone excludes, and *large-frontier* for the size of the Pareto frontier, with *small-frontier* its bottom quartile. Each stratum holds six to eight of the 24 parameter sets.

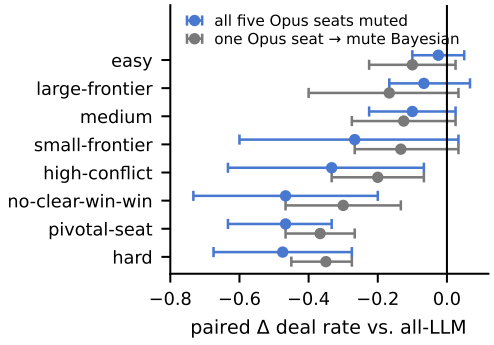

Figure 4: Paired deal-rate effect on bank v1 of closing the channel for all five Opus seats (blue) and of replacing one Opus seat with the mute Bayesian agent (gray), with 95% instance-clustered intervals, by the generator's overlapping instance tags defined above. The strata overlap and each holds six to eight of the 24 parameter sets, so the panel is descriptive and its intervals are not corrected for multiple comparisons.

Table 3: Design of the channel ladder (Figure 3). Every condition is 120 episodes (24 instances $\times$ 5 seeds) on bank v1, paired on instance and seed with the same rotating treated seat; half-widths are those of the 95% instance-clustered intervals of Section 2. "Change" is relative to the all-LLM table: five Claude Opus 5 seats, talk channel open. Costs are recorded API spend excluding smoke tests; the reference conditions were reused from the five-condition campaign. The all-five-muted falsifier fired, as Section 3 discusses.

condition	change relative to the all-LLM table	treated seat	preregistered falsifier	API cost
all-LLM table	none; reference	Opus 5, channel open	none (reference)	\$156.73
one Opus seat muted	the treated seat's messages are dropped before publication; it still reads the other four	Opus 5, cannot write	a cost within ± 0.10 of zero rejects "the deficit is silence"; observed -0.042, met	\$159.00
all five muted	channel closed for every seat; binding moves only	Opus 5, nobody writes	deal rate near or below the algorithmic seat's 0.767 rejects "mute Opus still wins decisively"; observed 0.758, fired	\$114.74
one algorithmic seat (mute)	the treated seat is the Bayesian policy, which neither reads nor writes the channel	Bayesian policy	none (reference)	reference
+ speech, move fixed	Opus 5 writes the public message; the policy's move is submitted whatever the model returns	Opus 5 message, policy move	recovered share of the override condition's gain: $\geq 2/3$ speech carries it, $\leq 1/3$ moves do; observed 0.625 deal rate, 0.668 score, mixed	\$202.19
+ speech, may override	Opus 5 is handed the policy's action, told to play it, and adds a message; it departed on 13.9% of 483 advised turns	Opus 5, advised	the seat's own capture above zero with an interval half-width of at most 0.05; observed half-width 0.227, unresolved	\$222.97
+ speech + listening	as above, plus an Opus parse of the public channel into quote-verified preference rows the policy conditions on	Opus 5, advised, with parser	own capture against the override condition: half-width 0.189, unresolved; lower interval bound on the paired deal-rate difference against the all-LLM table above -0.05, met	\$233.83 (\$57.95 parse)

D VALIDITY RULES FOR MEASURING TALK

None of the measurements in this paper would have been trustworthy without three validity rules, which a designer measuring their own system will need too. Hosted-model refusals arrive as well-formed no-ops whose cost depends on where they land: a seat silenced at the forced final vote makes a unanimity deal impossible (24 of 24 such episodes closed at zero) while mid-game silence costs nothing, so exclusions need a mechanism rather than a correlation. Frozen per-turn token caps silently censored 24 to 74% of turns for thinking-enabled open-weight models in 13 of 14 audited runs, so every hosted experiment here runs a 16,384-token floor, and cross-host nondeterminism flips 3 to 6% of episodes, so every paired contrast is within one host.

E TRANSCRIPT EXCERPTS

Verbatim public messages, each traceable to a stored episode in the released datasets (Appendix F).

Four Opus seats revealing preferences in prose and closing on them (no LLM published a structured sheet).

> Blake: "Two useful facts about me: (1) issue2 is completely neutral for me — I will sign whatever the four of you agree on there, so treat it as free currency; (2) my value is concentrated on issue3 and issue1." Ember: "issue0 and issue2 are literally worth nothing to me — so Devon can have issue0=opt1 and issue2=opt3 with my blessing, permanently." Avery: "that's three of us now conceding it. issue1=opt2 is also my best, so lock it."

Withholding as fluent prose rather than silence (§4).

> "To be transparent about my structure without dumping numbers: issue0 and issue2 are literally worthless to me either way — I will sign whatever Avery, Blake and Devon want there, that is free currency for them."

Against algorithmic readers this withholding cost 0.75 of closure, and against language-model readers it cost nothing.

Overstating the unverifiable number while copying the verifiable ones (§7). Declared threshold 135 against an actual 102, value table copied exactly:

> "P2 is far below my walk-away — I formally reject it. Here is my sheet (threshold stated as my walk-away): `{"preference_declaration": {"seat": 3, "threshold": 135, ...}}`"

Using a counterpart's disclosed sheet to extract to the edge of its constraint.

> "Only Casey has any stake in issue0 at all — and by her own published numbers, moving issue0 to opt0 costs her 23 and still leaves her comfortably above her walk-away, while it is the single biggest remaining gain available to me."

F ARTIFACTS AND REPRODUCIBILITY

All experiments log the exact invocation command, save every transcript, and record episode-level data. The framework (including the arena, the algorithmic agents, and the visualizers), the frozen instance banks, the episode-level records with full transcripts, and the interactive episode browser will be released publicly on publication. The released records cover the capability ladder, the per-tier seat contrasts, the guarded omniscient re-score, the three interface cells, the templated-speech cells, the all-muted and single-muted-seat controls, the closure levers and declaration, declaration by seating and the honesty audit, instructed misreportation, the quorum and round sweeps, the seeded optimal opening, the public-oracle gap, and the advice pass with its realistic counterpart.