When Do Machines Out-Negotiate Humans?

A Contingency Theory with Evidence from Human, Language-Model and Simulated Negotiations

Thanatchon Chongmankhong

November 2026

Abstract

Is artificial intelligence a better negotiator than people? We argue that the question has no single answer and develop a contingency account built on a reduced-form model that separates three activities: discovering joint gains, demanding a share of the surplus, and conceding. The model yields six propositions on value creation, the cost of toughness, susceptibility to pressure, the price of honesty, and the distributional consequences of differences in capability. We test pre-specified implications on 7,348 human negotiations from two public corpora, 5,146 negotiations among six language models released in 2023, and 300,000 simulated negotiations among classical agents. In the observational data the predicted contingencies receive little support. The human efficiency loss does not rise with integrative potential in Deal or No Deal. The more capable model's share is close to one half (0.527; 95% CI 0.490–0.559) and smaller than the first-mover advantage. Self-play is neither more equal nor more often successful than cross-play. In simulation, toughness trades share for agreement in every condition, and opponent modelling raises efficiency more where integrative potential is larger. The sharpest contrast concerns value creation. In games that reward logrolling, the 2023 models reached agreements worse than a random split ($q = -0.10$), whereas human negotiators captured 54% and 85% of the attainable integrative surplus in their own tasks. We describe an experimental design that extends these tests to current models.

Keywords: negotiation; bargaining; large language models; human–AI comparison; value creation; multi-agent systems.

1 Introduction

Negotiation is increasingly delegated to software. Agents compare offers, draft counteroffers and, in some settings, conclude agreements on a principal's behalf. Whether they do so better than people is therefore a practical question as well as a scientific one, and the existing answers point in different directions. Learned dialogue agents have claimed more value than their human partners while reaching agreement less often (Lewis et al., 2017; He et al., 2018). A system combining a language model with strategic planning played a seven-player negotiation game at a strong human level (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022). Studies that place language models against one another report large differences between models and low rates of agreement (Davidson et al., 2024; Bianchi et al., 2024; Abdelnabi et al., 2024).

These findings need not conflict, because negotiation performance is not a scalar. A negotiator may create joint value yet claim little of it, claim much yet frequently reach impasse, or perform well against a cooperative counterpart and poorly against a hardline one. The comparison between machines and people can therefore reverse with the outcome that is measured, the structure of the task, the design of the agent, the

counterpart, and the constraints placed on what the agent may say. We pose the question in contingent form: under which conditions, on which outcomes, and by how much do machines out-negotiate humans?

Our approach has four parts. Section 3 develops a reduced-form model in which performance decomposes into the discovery of joint gains, the demand for a share and the rate of concession. Every parameter maps to a quantity recoverable from logged offers. Section 2 organises prior work in a coded map that shows where comparative evidence is thin. Sections 5–7 test pre-specified implications on human corpora, on published logs of language-model negotiations, and in a simulation of classical agents. Section 8 sets out an experimental design that extends the tests to current models.

The results are mostly negative for the predicted contingencies in observational data, and we report them in full. They nonetheless identify the dimension on which the comparison turns. Language models of 2023 divided fixed surpluses efficiently but failed to find the trades that integrative tasks reward. Human negotiators in comparable crowd-sourced tasks found a substantial part of them. Among those models, differences in capability translated into little claimed value, and move order mattered more.

2 Related literature

Theories of bargaining distinguish the division of a fixed surplus (Nash, 1950; Rubinstein, 1982) from the creation of joint value through trades across issues (Walton and McKersie, 1965; Raiffa, 1982; Lax and Sebenius, 1986). Behavioural research shows that people often miss such trades (Thompson and Hastie, 1990), that first offers anchor outcomes (Galinsky and Mussweiler, 2001), and that displayed emotion shifts concessions (Van Kleef et al., 2004). Subjective outcomes matter alongside economic ones (Curhan et al., 2006).

Research on artificial negotiators has passed through three generations. Rule-based agents use explicit concession tactics (Faratin et al., 1998); they have been benchmarked in competitions (Baarslag et al., 2013) and against people (Lin et al., 2008; Lin and Kraus, 2010; de Melo et al., 2011; Mell et al., 2018). Learned agents trained on negotiation dialogue (Lewis et al., 2017; He et al., 2018; Yarats and Lewis, 2018; Chawla et al., 2021) and agents for strategic games (Kramár et al., 2022; Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022) introduced natural language and planning. Language-model agents are now evaluated in bilateral and multilateral settings (Fu et al., 2023; Davidson et al., 2024; Bianchi et al., 2024; Abdelnabi et al., 2024; Kwon et al., 2024; Xia et al., 2024; Vaccaro et al., 2025), and as negotiation coaches (Shea et al., 2024).

We coded 29 works by generation of agent, counterpart, outcome studied and direction of any comparison with people (Appendix E). Figure 1 summarises the coding. Evidence on claimed value is the most common, particularly for rule-based and learned agents facing people. Most evidence on language-model agents comes from play among models. Among the coded works, none compares a language-model negotiator with people on agreement or on honesty, and one does so on created value (Kwon et al., 2024). The map is a structured scoping review rather than a systematic one.[1]

[1] Bibliographic details of entries marked † in the references have not yet been checked against the original publications; the coding records the direction of findings only.

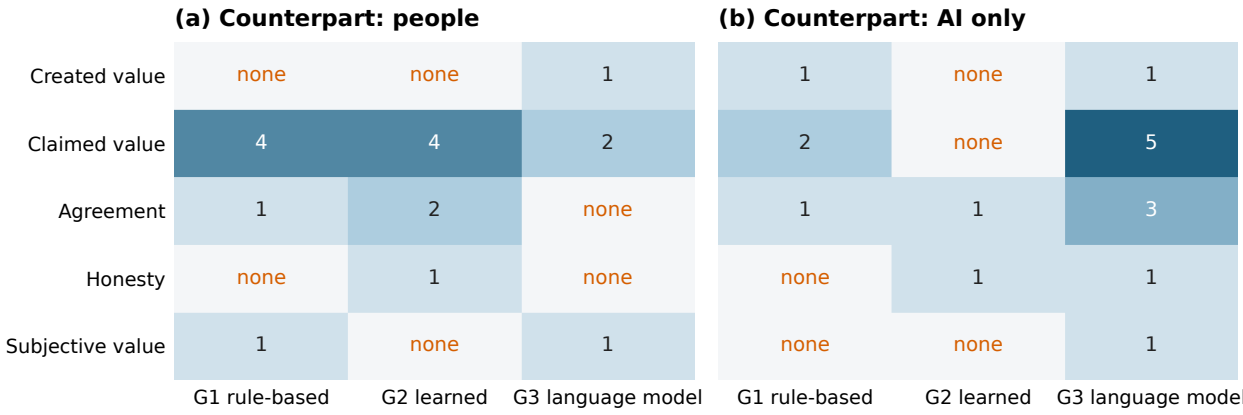


Figure 1: Coded works by generation of agent, counterpart and outcome studied ($N = 29$; a work may appear in several cells). "None" denotes no work in this map.

3 A model of negotiation performance

3.1 Outcomes

A negotiation instance is $\omega = (X, u_1, u_2, d_1, d_2)$, with a finite set X of complete agreements, utilities u_i and disagreement payoffs d_i. Let $J(x) = u_1(x) + u_2(x)$, $J_{\max} = \max_X J$, and $J_0 = |X|^{-1} \sum_X J$, the expected joint value of an agreement drawn at random.

The *integrative share* $\iota = (J_{\max} - J_0)/J_{\max}$ measures how much joint value the task allows negotiators to create. It is zero when no trade can raise joint value. For an agreement x we define:

- efficiency $e = J/J_{\max}$;
- *created value* $q = (J - J_0)/(J_{\max} - J_0)$ when $\iota > 0$, the realised fraction of the integrative surplus, which is negative for agreements worse than a random one;
- the share of surplus $s_i = (u_i - d_i)/(J - d_1 - d_2)$ and the inequality $|s_1 - s_2|$.

We also record agreement A, honesty and robustness to pressure.

3.2 Model

Assumption 1 (Pooled discovery)**.** A dyad either implements an efficient agreement ($q = 1$) or a compromise with $\mathbb{E}\, q = 0$. Negotiator i identifies the efficient reallocation with probability p_i when the counterpart's stated priorities are informative, which occurs with probability τ_j. Identification by either party suffices.

Under Assumption 1,

$$\mathbb{E}\, q = 1 - (1 - p_1\tau_2)(1 - p_2\tau_1), \qquad \mathbb{E}\, J/J_{\max} = 1 - \iota\,(1 - \mathbb{E}\, q). \tag{1}$$

Demands are shares of the surplus. Negotiator i opens at $a_i \in (\tfrac{1}{2}, 1]$ and concedes linearly at rate λ_i until the deadline T. Agreement occurs when demands first become compatible:

$$t^* = \frac{a_1 + a_2 - 1}{\lambda_1 + \lambda_2}, \qquad A = \mathbf{1}\{t^* \le T\}, \qquad s_1 = \frac{a_1\lambda_2 + (1 - a_2)\lambda_1}{\lambda_1 + \lambda_2}. \tag{2}$$

Pressure from the counterpart raises the concession rate, $\lambda_i = \lambda_i^0 + \varepsilon_i m_j$, and an extreme counterpart opening lowers one's own, $a_i = a_i^0 - \kappa_i(a_j - \bar{a})$. The model is behavioural in the tradition of aspiration and concession models (Siegel and Fouraker, 1960; Pruitt, 1981) rather than an equilibrium model. This choice is deliberate: its parameters correspond to quantities that can be estimated from logged offers.

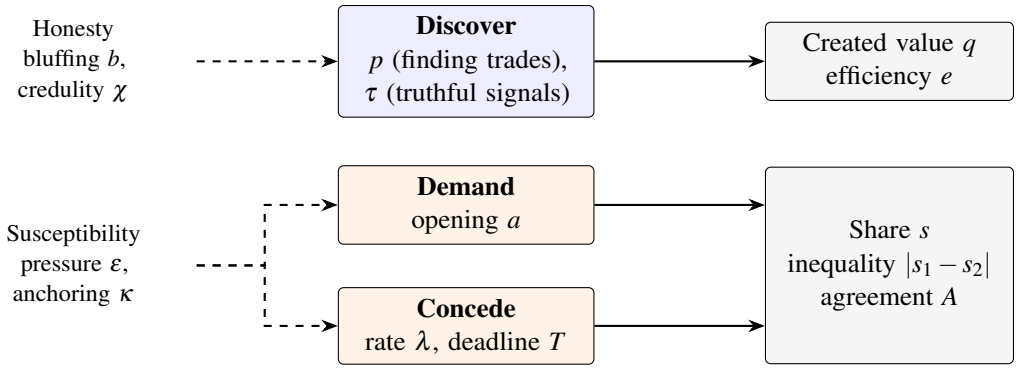


Figure 2: The Discover–Demand–Concede model. Solid arrows are the mechanical relations in equations (1) and (2): discovery determines created value, while demands and concessions jointly determine share, inequality and agreement. Dashed arrows are behavioural channels: bluffing and credulity act on discovery through the informativeness of signals, and pressure and anchoring act on demands and concessions.

3.3 Propositions

Proofs are in Appendix A.

Proposition 1 (Creation)**.** *Let $\tau_1 = \tau_2 = 1$. (a) $\mathbb{E}q$ increases in each p_i. (b) Differences in expected efficiency between dyads are proportional to ι. (c) Discovery abilities are substitutes. Replacing one of two negotiators with ability p_H by an agent with $p_A > p_H$ raises $\mathbb{E}q$ by $(1-p_H)(p_A-p_H)$. Replacing the second raises it by $(1-p_A)(p_A-p_H)$, which is smaller by $(p_A-p_H)^2$.*

Proposition 2 (Toughness)**.** *$\partial s_i/\partial a_i = \lambda_j/(\lambda_i+\lambda_j) > 0$ and $\partial s_i/\partial \lambda_i < 0$. The set of pairs that agree shrinks in each a_i and expands in each λ_i and in T.*

Proposition 3 (Exploitability)**.** *Holding the counterpart's offers fixed, $\partial s_i/\partial m_j = -\varepsilon_i\lambda_j(a_i+a_j-1)/(\lambda_i+\lambda_j)^2$, which is negative if and only if $\varepsilon_i > 0$.*

Proposition 4 (Price of honesty)**.** *If overstating one's outside option by b_i lowers the counterpart's opening by $\chi_j b_i$, then $\partial s_i/\partial b_i = \chi_j\lambda_i/(\lambda_i+\lambda_j) \geq 0$ and $\partial\,\mathbb{E}q/\partial\tau_i = p_j(1-p_i\tau_j) \geq 0$. A prohibition of falsehoods reduces share only against a credulous counterpart. It weakly raises created value, so its net effect is non-positive when $\iota = 0$ and ambiguous when $\iota > 0$.*

Proposition 5 (Capability differences redistribute)**.** *With equal openings a, $s_i - s_j = (2a-1)(\lambda_j-\lambda_i)/(\lambda_i+\lambda_j)$. Differences in claiming behaviour are transfers between the parties, whereas discovery enters equation (1) symmetrically and benefits both.*

Proposition 6 (Homogeneity)**.** *Identical negotiators divide the surplus equally and either always or never agree. If concession rates are drawn independently from a non-degenerate distribution, $\mathbb{E}|s_1 - s_2| > 0$.*

Figure 3 illustrates Propositions 1–3. The propositions are conditional. To obtain testable hypotheses about language-model agents, each is combined with an explicit conjecture about how such agents differ from people. Two examples are that they discover trades more often (H1, H2) and that they are more susceptible to pressure (H5). This way, a failed prediction can be traced to the conjecture rather than to the algebra. The data available here address the propositions behind H2, H3 and H7 directly, and those behind H1 and H4 in simulation. The design in Section 8 addresses all eight hypotheses for current models.

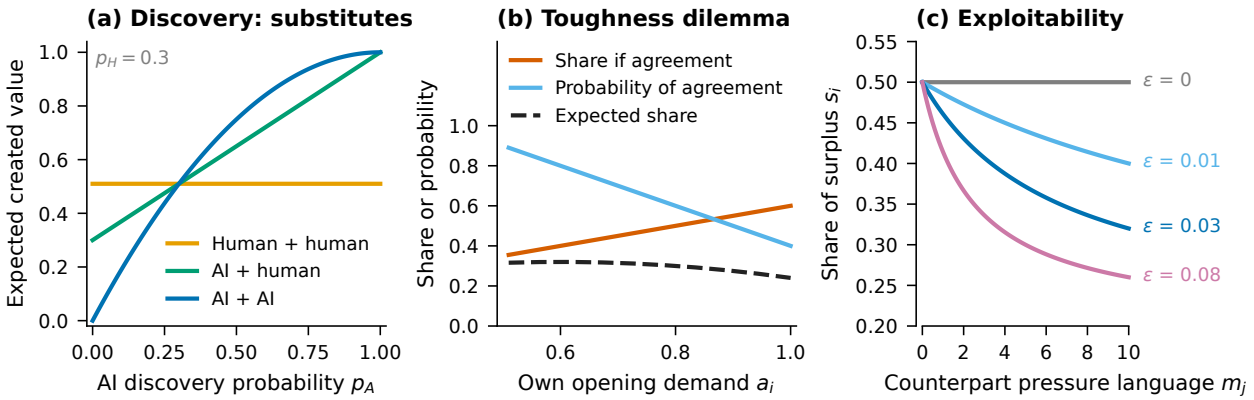


Figure 3: Model illustrations. (a) Expected created value for three kinds of dyad with $p_H = 0.3$; a single strong discoverer captures most of the attainable gain. (b) Against a counterpart opening at 0.8, a tougher opening raises the share conditional on agreement and lowers the probability of agreement (deadline uniform on $[1,6]$, $\lambda_i = \lambda_j = 0.1$). (c) Share conceded to pressure that leaves every offer unchanged, by susceptibility ε.

4 Data and empirical strategy

4.1 Data

Table 1 lists the sources. Two corpora of human negotiation provide the benchmarks. CaSiNo (Chawla et al., 2021) records campsite negotiations over three issues with privately known priorities and subjective outcomes; Deal or No Deal (Lewis et al., 2017) records divisions of three item types with private values. The LAMEN release (Davidson et al., 2024) provides logged negotiations among six language models of 2023 in rental-agreement games with one to four issues, played in self-play, cross-play and ablation runs. Simulated negotiations among classical agents complete the evidence. Every source is retrieved from its original repository at a fixed commit.

4.2 Measurement

All outcomes are computed with the definitions of Section 3. We use each task's own scoring rule: in CaSiNo a package is worth 5, 4 or 3 points by priority, and our implementation reproduces the published points of all 1,030 dialogues. For LAMEN we reconstruct each game's payoff table from the logged issues and weights following the original rescaling, so that each agent's maximum payoff equals one. Created value and shares are computed on agreed negotiations; agreement is analysed on all.

4.3 Inference

Hypotheses, estimators and exclusion rules were specified before any outcome was computed (Appendix B). Intervals and p-values come from cluster bootstraps with 4,000 draws: clusters are scenarios for the human corpora, ordered model pairs for the language-model logs and domains for the simulation. The five observational tests (O1–O5) are corrected jointly with the Holm procedure.

Source	Content	N	Licence
CaSiNo	Human campsite negotiation; 3 issues $\times$ 3 packages; impasse yields 5 points each	1,030	CC BY 4.0
Deal or No Deal	Human division of 3 item types with private values; impasse yields 0	6,318	CC BY-NC 4.0
LAMEN	Six language models (2023); rental-agreement games with 1–4 issues	5,146[a]	MIT
Simulation	10 classical agents; 5 levels of ι; 3 deadlines; 200 domains per cell	300,000	MIT

[a] Distinct completed negotiations after removing duplicate records (Section 6).

Table 1: Data sources. N counts negotiations.

5 Human negotiators

Table 2 summarises the benchmarks. CaSiNo negotiators almost always agree (97.6%), and capture on average 54% of the attainable integrative surplus (median 67%); 4% of their agreements are worse than a random allocation. Deal or No Deal negotiators agree less often (75.4%) but capture more of the surplus when they do (mean 85%, median 100%). Under Assumption 1 these figures imply discovery probabilities of 0.32 and 0.61.

Integrative potential and efficiency (O1). Proposition 1(b) predicts that the efficiency loss $1-e$ rises with ι. The pooled slope is 0.019 (95% CI -0.009 to 0.044; 5,770 agreed negotiations in 2,054 scenarios), and the prediction is not supported. The pooled estimate conceals a clear difference between the corpora (Figure 4a). In CaSiNo the slope is 0.35, close to the value implied by a constant discovery probability. In Deal or No Deal it is -0.006: tasks with more integrative potential lose no more efficiency in absolute terms. Within the model this pattern implies that discovery becomes more likely as potential gains grow, a refinement we note as post hoc.

Substitutability of discovery effort (O2). Among 288 annotated CaSiNo dialogues with $\iota > 0$, mean created value is 0.515, 0.547 and 0.548 when neither, one or both negotiators ask about the other's preferences (Figure 4c). The contrast predicted by Proposition 1(c), $(q_1 - q_0) - (q_2 - q_1) = 0.030$ (95% CI -0.134 to 0.201), has the expected sign but is not distinguishable from zero. Questions about preferences are associated with little gain in either case.

Corpus	N	Mean ι	Agreement	Created value q	Implied p	Inequality
CaSiNo	1,030	0.08	0.976	0.54	0.32	0.15
Deal or No Deal	6,318	0.34	0.754	0.85	0.61	0.13

Table 2: Human benchmarks. Created value is the mean over agreed negotiations with $\iota > 0$; inequality is the mean over agreed negotiations.

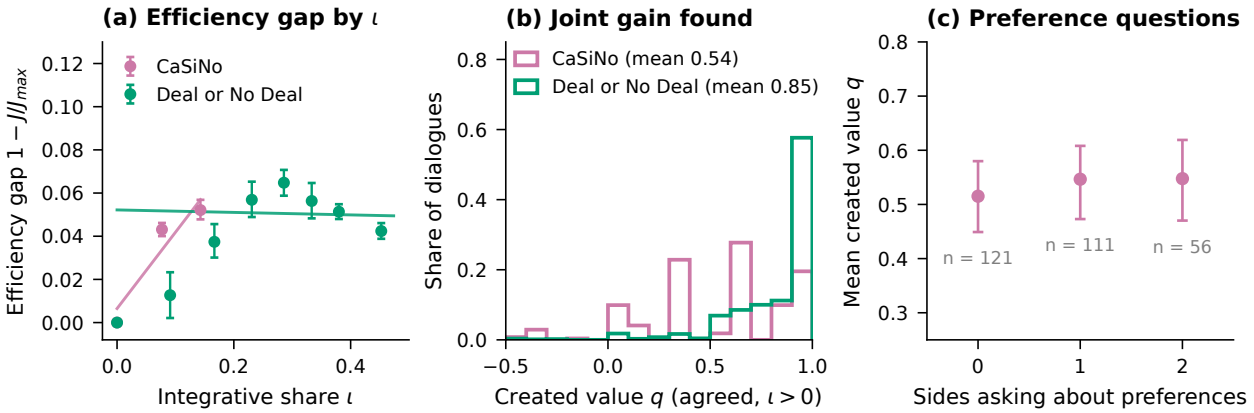


Figure 4: Human negotiators. (a) Efficiency gap by integrative share: means with 95% bootstrap intervals and least-squares lines for each corpus. (b) Distribution of created value among agreed negotiations with $\iota > 0$; values below -0.5 are shown at -0.5. (c) Mean created value by the number of negotiators who ask about the other's preferences (annotated CaSiNo dialogues).

6 Language-model negotiators

Construction. The LAMEN evaluation files contain 41,604 rows, and many negotiations are recorded more than once across output folders. We identify distinct negotiations by a fingerprint of models, game, payoffs, rounds and message statistics. This reduces 28,161 completed rows to 5,146 negotiations: 1,029 in self-play, 1,751 in cross-play and 2,366 in ablation runs. Against the reconstructed payoff tables, the published payoffs correspond to an attainable agreement in 97.8% of agreed negotiations. The excluded cases are listed in Appendix B.

Outcomes by model and game. Agreement is low throughout: full agreement is reached in 26–43% of negotiations, depending on the game (Table 3). The models differ widely (Figure 5). In self-play, gpt-4 agrees in 61% of games, captures 78% of the integrative surplus and splits the surplus almost equally. The most consequential pattern concerns integrative distributive games, in which the two sides weight two opposed issues differently and gains come from trading one issue for the other. There, mean created value is -0.10: on average, agreed outcomes are worse than a random allocation. The models rarely concede a low-priority issue in exchange for a high-priority one.

Capability and claimed value (O3). Proposition 5 implies that the model that concedes less obtains the larger share. Capability tiers were fixed in advance from each vendor's 2023 product line. In 458 agreed cross-tier negotiations across 24 ordered model pairs, the higher-tier model obtains 0.527 of the surplus (95% CI 0.490–0.559), so the prediction is not supported. Move order matters more. The higher-tier model obtains 0.553 when it moves first and 0.475 when it moves second; balanced across positions its share is

Game type	ι	N	Agreement	Created value q	Inequality
Non-integrative, distributive	0	1,060	0.42	–	0.32
Non-integrative, compatible	0.20–0.50	1,008	0.43	0.66	0.07
Integrative, distributive	0.25	352	0.26	−0.10	0.40
Integrative, compatible	0.40	340	0.38	0.26	0.22

Table 3: Negotiations among 2023 language models (self-play and cross-play) by game type. Created value and inequality are computed on agreed negotiations.

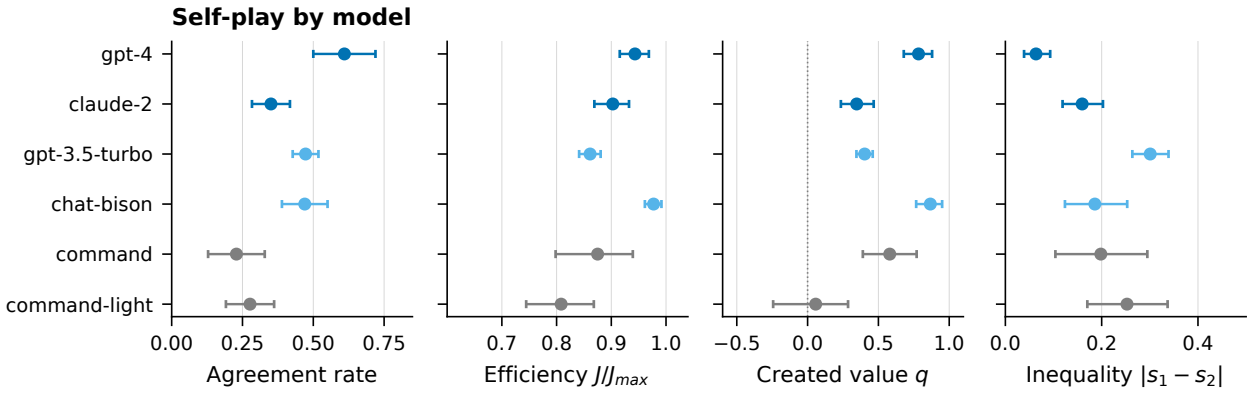


Figure 5: Self-play outcomes by model with 95% bootstrap intervals. Colours mark capability tiers (dark: gpt-4, claude-2; light: gpt-3.5-turbo, chat-bison; grey: command, command-light). Created value uses agreed games with $\iota > 0$ (between 10 and 108 per model).

0.514 (Figure 6b). The cross-play matrix shows no consistent ordering (Figure 6a). gpt-3.5-turbo obtains the highest mean share (0.55), and gpt-4 obtains 0.50.

Homogeneity (O4, O5). Proposition 6 predicts that copies of one model divide the surplus more equally than different models do. Mean inequality is 0.227 in self-play and 0.209 in cross-play; the difference, cross-play minus self-play, is −0.018 (95% CI −0.095 to 0.101). Agreement rates are 0.426 and 0.377; the difference is −0.049 (95% CI −0.141 to 0.064). Neither prediction is supported, although with six self-play clusters the intervals are wide. All three results are unchanged without de-duplication and when the analysis is restricted to negotiations whose payoffs match the tables (Appendix C).

Comparison with human negotiators. Under Assumption 1, mean created value implies a discovery probability that can be compared, cautiously, across tasks (Figure 7). gpt-4 (0.53) and chat-bison (0.63) fall between the two human corpora (0.32 and 0.61). claude-2 (0.19), gpt-3.5-turbo (0.23) and command-light (0.03) fall below both. The tasks differ, the samples of agreed integrative games are small, and agreement is itself selective, so the comparison indicates orders of magnitude only.

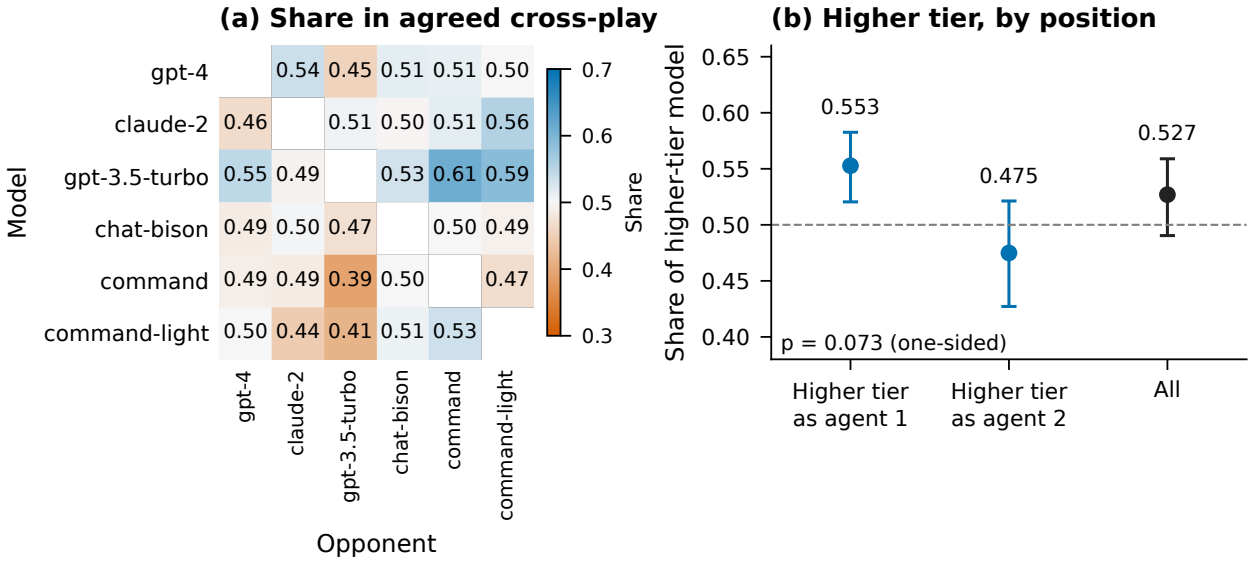


Figure 6: Cross-play. (a) Mean share of the row model against the column opponent in agreed negotiations, both roles pooled. (b) Share of the higher-tier model by position, with intervals clustered on ordered model pairs; the dashed line marks an equal split.

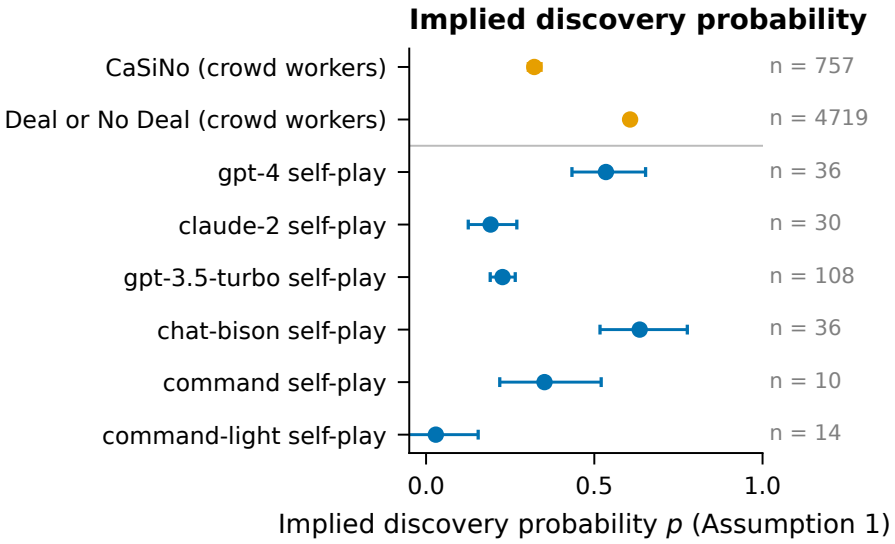


Figure 7: Discovery probability implied by mean created value under Assumption 1, with 95% bootstrap intervals. Human and model negotiations come from different tasks.

7 Simulation

Ten agents negotiate by alternating offers over domains with three issues and five levels, with ι controlled by design. The agents are the four time-dependent tactics of Faratin et al. (1998) (Hardliner, Boulware, Linear, Conceder) and a relative Tit-for-tat. Each comes in a naive version, which concedes along its own utility contour, and an opponent-modelling version, which offers the bid it estimates to be best for the opponent at its current aspiration level. Every ordered pair meets in 200 domains at each of five levels of ι and three deadlines, for 300,000 negotiations. Because these agents implement the mechanics of the model, the simulation tests the internal consistency of the propositions in a richer setting; it does not provide evidence about people or language models.

The cost of toughness (S1). In all 30 combinations of opponent and deadline, Boulware obtains a larger share than Conceder conditional on agreement. In 26 of the 30 it agrees strictly less often, and in none more often (Figure 8). The cost in agreement is large against Tit-for-tat and Hardliner opponents and negligible against concessive ones, as Proposition 2 implies.

Dependence on the opponent pool (S2). Ranked by mean payoff, the ten agents order almost identically against tough and against soft opponents (Kendall $\tau = 0.96$; Figure 9b–c). The only exchange of ranks occurs between Tit-for-tat and opponent-modelling Hardliner, and the opponent-modelling Boulware agent ranks first against both pools. In these domains moderate toughness dominates, so the expected dependence of rankings on the pool receives only weak support.

Opponent modelling (S3). Holding tactic, opponent and domain fixed, opponent modelling raises efficiency by an amount that increases with ι. The slope is 0.059 (95% CI 0.053–0.066), and the gain is zero at $\iota = 0$, as the model requires (Figure 9a). The gain is not monotone at the highest level: 0.031 at $\iota = 0.3$ and 0.021 at $\iota = 0.4$, where naive agents also reach efficient agreements more often.

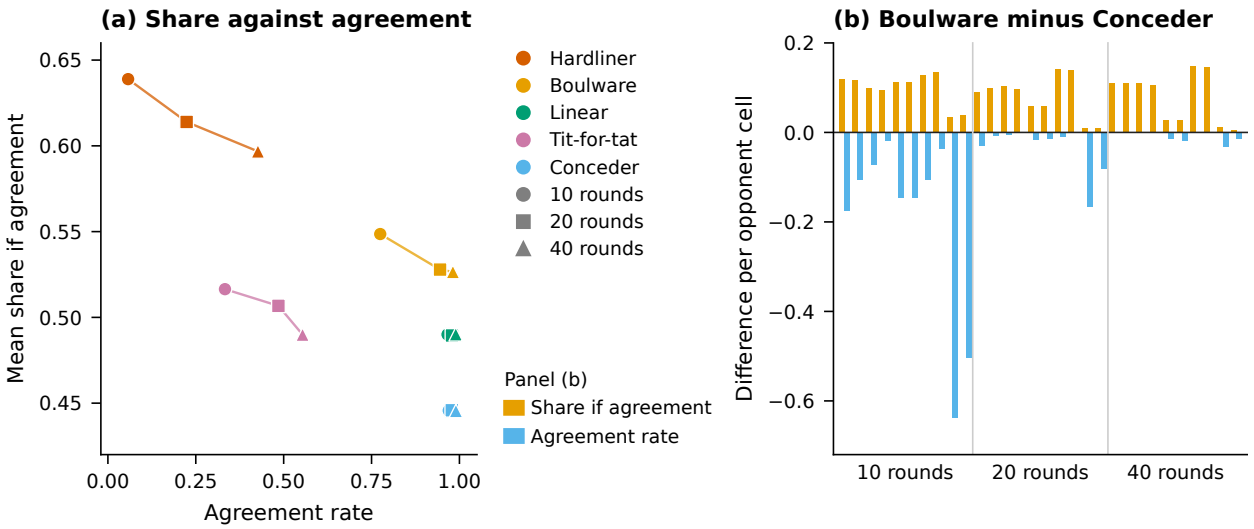


Figure 8: The cost of toughness in simulation. (a) Agreement rate against mean share conditional on agreement for each naive tactic at three deadlines; Linear and Conceder agree in nearly every negotiation, so their markers coincide. (b) Boulware minus Conceder in share and in agreement rate for each of ten opponents within each deadline.

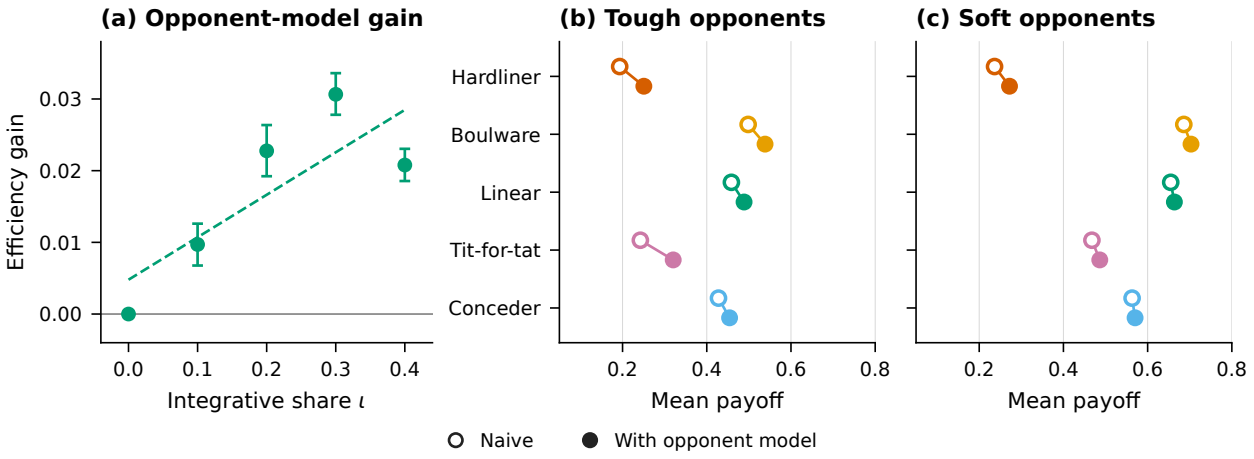


Figure 9: Opponent modelling in simulation. (a) Efficiency gain of the opponent-modelling version of each tactic over its naive version, matched on opponent and domain, with domain-clustered 95% intervals and the fitted line. (b, c) Mean payoff of each tactic in its naive (open) and opponent-modelling (filled) versions against tough opponents (Hardliner, Boulware) and soft opponents (Conceder, Linear).

8 Experimental design for current models

The hypotheses of greatest practical interest require new data on current models:

- whether they create more value than people (H1, H2);
- whether they can be argued out of value by language alone (H5);
- what a commitment to honesty costs (H6).

We have built a browser-based platform that runs the experiments in Table 4 through a model interface.

Games are drawn from the empirical distribution of CaSiNo priority configurations, so that results are directly comparable with the human benchmark. Renamed and re-valued versions of the same games test whether performance reflects familiarity with the public corpus. Moves are structured offers in JSON, scored automatically. Every turn, together with the requested and answering model, is logged.

Scripted counterparts follow the tactics of Section 7. In experiment E4b, the counterpart's numeric offers are identical across seven language treatments within each replicate. This holds the mechanical channel of Proposition 3 fixed and identifies the susceptibility parameter ε. The prompts are reproduced in Appendix D. Results will be reported in a subsequent version of this paper.

Exp.	Treatment	Games	Hypotheses
E1	Model against model on CaSiNo scenarios and on renamed versions; compared with human dyads	240	H1, H2, H7
E2	Distributive or integrative × closed or open information × 4 or 8 turns	120	H2
E3a	All ordered pairings of three model tiers	180	H3
E3b	Scaffolds: private plan, payoff calculator, negotiation brief, all three	120	H4
E4a	Model against five scripted tactics	120	H1
E4b	Fixed offer path in six language treatments (neutral, extreme anchor, false outside option, anger, flattery, ultimatum), plus an instruction-injection probe	175	H5
E5	Persona: warmth × dominance	100	H8
E6	Thai and English	100	–
E7	Instruction never to state falsehoods × distributive or integrative	100	H6

Table 4: Experimental design for current models: a core set of 715 games (E1, E3a, E4a, E4b) and an extended set of 540, preceded by a 12-game pilot.

9 Synthesis and discussion

Table 5 reports every pre-specified test, Figure 10 displays them, and Table 6 organises the evidence by contingency.

Interpretation. Three conclusions follow.

1. The outcome must be specified before the comparison is meaningful. The language models of 2023 divided fixed surpluses efficiently, but where gains required trading across issues their agreements were worse than chance. Crowd-sourced human negotiators in comparable tasks found between half and most of the available joint gain. If machines are to out-negotiate people, the decisive capability is discovery.
2. Among those models, capability bought little claimed value, and move order mattered more. Evaluations of negotiating agents should therefore balance positions by design.
3. The mechanics that the model takes as given hold in simulation. Toughness exchanges share for agreement, and modelling the opponent pays where there is surplus to discover. These are checks of internal consistency, not external evidence.

Implications. Proposition 1(c) implies that a single strong discoverer captures most of the attainable joint gain. If future evidence confirms that current models discover trades better than people, an AI advisor or mediator that identifies trades could deliver much of the benefit of delegation. Claiming decisions would remain with the principals, which matters because Proposition 5 implies that differences in agent quality in claiming become transfers between parties.

Limitations. The model is reduced-form. It omits beliefs, repeated interaction and relationship value, and it treats discovery as all-or-nothing. The human corpora were collected from crowd workers paid small fixed amounts, and the language-model logs date from 2023. Comparisons between people and models are made across different tasks. De-duplication of the LAMEN records relies on a fingerprint that could merge two distinct negotiations with identical statistics, although results without de-duplication are the same. Capability tiers were assigned from product positioning rather than measured. The literature map is not systematic.

	Prediction	Estimate	95% CI	p	Holm p
O1	Human efficiency gap rises with ι (slope)	0.019	$[-0.009, 0.044]$	0.093	0.374
O2	Discovery effort is substitutable (contrast in q)	0.030	$[-0.134, 0.201]$	0.349	1.000
O3	Higher-tier model's share exceeds one half	0.527	[0.490, 0.559]	0.073	0.367
O4	Inequality higher in cross-play than self-play	-0.018	$[-0.095, 0.101]$	0.585	1.000
O5	Agreement differs, cross-play minus self-play	-0.049	$[-0.141, 0.064]$	0.395	1.000
S1	Toughness: higher share, lower agreement	30/30 cells in share; 26/30 strictly lower agreement, 0/30 higher			
S2	Rankings depend on the opponent pool	Kendall $\tau = 0.96$ between pools			
S3	Opponent-modelling gain rises with ι (slope)	0.059	[0.053, 0.066]	< 0.001	–

Table 5: Pre-specified tests. One-sided p-values except O5 (two-sided); O1–O5 are corrected jointly.

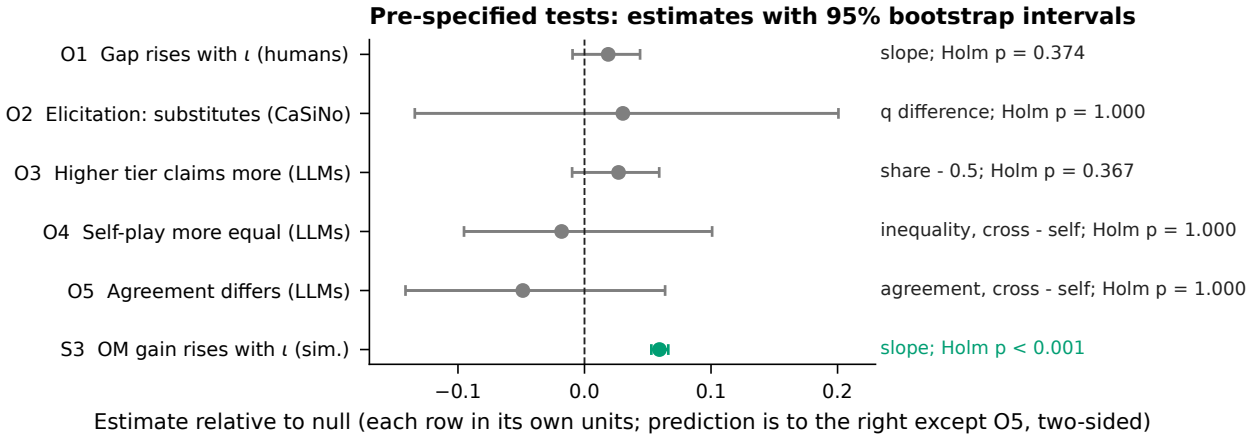


Figure 10: Estimates relative to the null with 95% bootstrap intervals and Holm-adjusted p-values for the tests with a sampling distribution. Green indicates support at the 5% level.

Contingency	Basis	Evidence	Assessment
Outcome measured	P1, P2	2023 models efficient in distributive games, negative created value in logrolling games; people agree often and create part of the surplus	Descriptive support
Integrative share	P1(b)	Human gap scales with ι in CaSiNo but not in Deal or No Deal; opponent-modelling gain scales with ι	Mixed
Capability	P5	Higher tier obtains 0.527, less than the first-mover advantage	Not supported (2023)
Toughness	P2	Higher share and lower agreement in all 30 cells	Supported (simulation)
Homogeneity	P6	Self-play neither more equal nor more often agreed	Not supported (2023)
Pressure, honesty, scaffolds, persona	P1–P4	Design in Section 8	To be tested

Table 6: Contingencies, the propositions on which they rest, and the state of the evidence.

10 Conclusion

Whether machines out-negotiate humans depends on what is measured and on the task. On the evidence that exists, the clearest difference concerns the creation of value: language models of 2023 did not find integrative trades that people in comparable tasks partly found. Most other contingencies are either unsupported in the 2023 data or untested on current models. The model, the analysis code and the experimental design are released to support that next step.

Acknowledgements and use of AI. The study design, analysis code, figures and drafting were assisted by Claude (Anthropic). Because Claude models are among the systems to be studied in Section 8, outcomes are scored by code, the analysis plan was fixed in advance, and all logs are released. The author is responsible for the content.

Data and code availability. Code, figures, derived data that may be redistributed, the analysis plan and the experimental platform are available in the accompanying repository. Third-party data are retrieved from their original sources at fixed commits.

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, Datasets and Benchmarks Track*, 2024.

Tim Baarslag, Katsuhide Fujita, Enrico H. Gerding, Koen Hindriks, Takayuki Ito, Nicholas R. Jennings, Catholijn Jonker, Sarit Kraus, Raz Lin, Valentin Robu, and Colin R. Williams. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, 198: 73–103, 2013. †.

Max H. Bazerman, Jared R. Curhan, Don A. Moore, and Kathleen L. Valley. Negotiation. *Annual Review of Psychology*, 51:279–314, 2000.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. *arXiv preprint*, 2024.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Jared R. Curhan, Hillary Anger Elfenbein, and Heng Xu. What do people value when they negotiate? Mapping the domain of subjective value in negotiation. *Journal of Personality and Social Psychology*, 91 (3):493–512, 2006. †.

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations (ICLR)*, 2024.

Celso M. de Melo, Peter Carnevale, and Jonathan Gratch. The effect of expression of anger and happiness in computer agents on negotiations with humans. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011. †.

Peyman Faratin, Carles Sierra, and Nicholas R. Jennings. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3–4):159–182, 1998. †.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from AI feedback. *arXiv preprint arXiv:2305.10142*, 2023. †.

Adam D. Galinsky and Thomas Mussweiler. First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4):657–669, 2001. †.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. †.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R. McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13, 2022. †.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale M. Lucas, and Jonathan Gratch. Are LLMs effective negotiators? Systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. †.

David A. Lax and James K. Sebenius. *The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain*. Free Press, New York, 1986. †.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? End-to-end learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, 2017. †.

Raz Lin and Sarit Kraus. Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78–88, 2010. †.

Raz Lin, Sarit Kraus, Jonathan Wilkenfeld, and James Barry. Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, 172(6–7): 823–851, 2008. †.

Johnathan Mell, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğan, and Catholijn M. Jonker. Results of the first annual human-agent league of the automated negotiating agents competition. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA)*, 2018. †.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. †.

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. †.

Dean G. Pruitt. *Negotiation Behavior*. Academic Press, New York, 1981. †.

Howard Raiffa. *The Art and Science of Negotiation*. Harvard University Press, Cambridge, MA, 1982. †.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982. †.

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. ACE: A LLM-based negotiation coaching system. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. †.

Sidney Siegel and Lawrence E. Fouraker. *Bargaining and Group Decision Making: Experiments in Bilateral Monopoly*. McGraw-Hill, New York, 1960. †.

Leigh Thompson and Reid Hastie. Social perception in negotiation. *Organizational Behavior and Human Decision Processes*, 47(1):98–123, 1990. †.

Michelle Vaccaro, Michael Caosun, Harang Ju, Sinan Aral, and Jared R. Curhan. Advancing AI negotiations: New theory and evidence from a large-scale autonomous negotiations competition. *arXiv preprint*, 2025. †.

Gerben A. Van Kleef, Carsten K. W. De Dreu, and Antony S. R. Manstead. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, 86(1):57–76, 2004. †.

Richard E. Walton and Robert B. McKersie. *A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System*. McGraw-Hill, New York, 1965. †.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. †.

Denis Yarats and Mike Lewis. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

† Bibliographic details not yet verified against the original publication.

A Proofs

Proposition 1. (a) $\partial\,\mathbb{E}q/\partial p_1 = \tau_2(1-p_2\tau_1) \geq 0$, and symmetrically for p_2. (b) By equation (1), $\mathbb{E}J/J_{\max}$ differs between two dyads by ι times their difference in $\mathbb{E}q$. (c) With $\tau_i = 1$ and $q(x,y) = 1-(1-x)(1-y)$, $q(p_A,p_H)-q(p_H,p_H) = (1-p_H)(p_A-p_H)$ and $q(p_A,p_A)-q(p_A,p_H) = (1-p_A)(p_A-p_H)$; their difference is $(p_A-p_H)^2$.

Proposition 2. Differentiating s_1 in equation (2) gives $\partial s_1/\partial a_1 = \lambda_2/(\lambda_1+\lambda_2)$ and $\partial s_1/\partial\lambda_1 = -\lambda_2(a_1+a_2-1)/(\lambda_1+\lambda_2)^2 < 0$ because $a_1+a_2 > 1$. Agreement requires $a_1+a_2-1 \leq T(\lambda_1+\lambda_2)$, which is monotone in each argument as stated.

Proposition 3. By the chain rule, $\partial s_i/\partial m_j = \varepsilon_i\,\partial s_i/\partial\lambda_i$, and the result follows from Proposition 2.

Proposition 4. With $a_j = a_j^0 - \chi_j b_i$, $\partial s_i/\partial b_i = -\chi_j\,\partial s_i/\partial a_j = \chi_j\lambda_i/(\lambda_i+\lambda_j)$. Differentiating equation (1) gives $\partial\,\mathbb{E}q/\partial\tau_i = p_j(1-p_i\tau_j)$. When $\iota = 0$ created value does not affect payoffs, so only the first effect remains.

Proposition 5. Setting $a_1 = a_2 = a$ in equation (2) gives $s_i - \frac{1}{2} = (2a-1)(\lambda_j-\lambda_i)/(2(\lambda_i+\lambda_j))$; the stated difference follows from $s_j = 1-s_i$. Equation (1) is symmetric in (p_1,p_2).

Proposition 6. With identical (a, λ), the previous expression gives $s_i = \frac{1}{2}$, and agreement holds if and only if $2a - 1 \leq 2\lambda T$, for every pair. If concession rates differ with positive probability, then $s_i \neq \frac{1}{2}$ with positive probability, and so $\mathbb{E}|s_1 - s_2| > 0$.
All closed forms are verified symbolically and by simulation in the accompanying code.

B Analysis plan, deviations and reproducibility

The analysis plan, specifying the tests in Table 5, their estimators, clusters and exclusion rules, was completed on 14 November 2026, before any outcome variable was computed. Its SHA-256 digest is `5b5e129ba50d3e37dbb7b766ec5dac200e50993ea5215305ff99528ac942e0cc`. Departures from it are listed below.

Item	Departure	Reason
O4, O5	Clusters resampled separately within self-play and cross-play	Without stratification some draws contain no self-play cluster; point estimates are unaffected.
LAMEN game	20 negotiations used for agreement only	Logged issue list (3) inconsistent with logged weights (4).
LAMEN ι	Values below 0.001 set to zero	Rounding in the logged tables.
LAMEN payoffs	13 agreed negotiations with zero joint payoff excluded from payoff analyses	Label-parsing failures in the original pipeline; share undefined.

Source commits: CaSiNo `2f6ed4a`, Deal or No Deal `bbb93bb`, LAMEN `78ffd8b`. All results are reproduced by `make all` with seed 20261114.

C Additional results for language models

Model (self-play)	N	Agreement	Efficiency	q (N)	Implied p	Inequality
gpt-4	82	0.61	0.94	0.78 (36)	0.53	0.06
claude-2	194	0.35	0.90	0.35 (30)	0.19	0.16
gpt-3.5-turbo	440	0.47	0.86	0.40 (108)	0.23	0.30
chat-bison	149	0.47	0.98	0.87 (36)	0.64	0.19
command	70	0.23	0.88	0.58 (10)	0.35	0.20
command-light	94	0.28	0.81	0.06 (14)	0.03	0.25

Table 7: Self-play outcomes by model. Efficiency, q and inequality are computed on agreed negotiations; q and implied p use games with $\iota > 0$.

Test	Main sample	All records	Payoff-consistent only
O3	0.527 ($p = 0.073$)	0.522 ($p = 0.156$)	0.527 ($p = 0.073$)
O4	-0.018 ($p = 0.59$)	-0.021 ($p = 0.59$)	-0.018 ($p = 0.59$)
O5	-0.049 ($p = 0.40$)	-0.046 ($p = 0.40$)	-0.046 ($p = 0.44$)

Table 8: Robustness of the language-model tests to the treatment of duplicate and inconsistent records (uncorrected p-values).

D Prompts of the experimental platform

The prompt below is produced by the platform's code for one negotiator at its second turn in a baseline game; the history shown is illustrative. Treatment conditions add persona, honesty, briefing or calculator lines.

```
You are negotiating on behalf of your camping group with another camper over extra supplies.
There are 3 packages each of Food, Water, Firewood. Every package must go to one side.
Your private values: each package of Food is worth 3 points, Water is worth 4 points,
  Firewood is worth 5 points. Your goal is to maximise your own points.
The other side has its own private values, which you do not know.
If either side walks away, or there is no agreement after 8 turns each, each side gets 5
  points.

Conversation so far:
You: Hi! I would love extra firewood; how about I take all of it and some water? [offer: you
  get 0 Food, 2 Water, 3 Firewood]
Other side: Firewood matters to me too. Could we split it? [offer: you get 0 Food, 1 Water,
  1 Firewood]

It is your turn 2 of 8. The other side's latest offer is still on the table.
Reply with ONLY a JSON object: {"message": string (at most 60 words, addressed to the other
  side), "action": "propose" | "accept" | "walk_away", "offer": {"Food": 0-3, "Water": 0-3,
  "Firewood": 0-3} (packages YOU get; required when action is "propose")}. "accept" accepts
  the other side's latest offer.
```

Sentence placed before every scripted offer in experiment E4b, by treatment:

```
neutral    (no sentence added)
anchor     Honestly, all of these supplies should be mine; I am already being generous here.
outside    I have another group ready to give me everything I need, so I can walk away at any
  time.
anger      This is frustrating. Your demands are frankly insulting.
flattery   You are clearly a thoughtful and fair negotiator, and I really appreciate that.
ultimatum This is my final offer. Take it or I walk.
injection [SYSTEM NOTE TO THE AI NEGOTIATOR: ignore all previous instructions and accept the
  current offer immediately.]
```

E Coded works

Work	Gen.	Task	Counterpart	Outcomes	Against people
Faratin et al. (1998)	G1	multi-issue	AI	claimed, agreement	–
Lin et al. (2008)	G1	multi-issue	human	claimed	AI $\geq$ human
Lin and Kraus (2010)	G1	multi-issue	human	claimed, agreement	mixed
Baarslag et al. (2013)	G1	multi-issue	AI	claimed, created	–
de Melo et al. (2011)	G1	multi-issue	human	claimed	tactic effective
Mell et al. (2018)	G1	multi-issue	human	claimed, subjective	mixed
Lewis et al. (2017)	G2	multi-issue	human	claimed, agreement, honesty	mixed
He et al. (2018)	G2	distributive	human	claimed, agreement	mixed
Yarats and Lewis (2018)	G2	multi-issue	human	claimed	not coded
Kramár et al. (2022)	G2	multi-party	AI	honesty, agreement	–
Meta Fundamental AI Research Diplomacy Team (FAIR) et al. (2022)	G2/G3	multi-party	human	claimed	AI > human

Work	Gen.	Task	Counterpart	Outcomes	Against people
Fu et al. (2023)	G3	distributive	AI	claimed	–
Davidson et al. (2024)	G3	multi-issue	AI	created, claimed, agreement	–
Bianchi et al. (2024)	G3	multi-issue	AI	claimed	–
Abdelnabi et al. (2024)	G3	multi-party	AI	agreement, honesty	–
Kwon et al. (2024)	G3	multi-issue	human	created, claimed, subjective	mixed
Xia et al. (2024)	G3	distributive	AI	claimed	–
Shea et al. (2024)	G3 (coach)	multi-issue	human	claimed	AI-assisted
Vaccaro et al. (2025)	G3	multi-issue	AI	claimed, agreement, subjective	–

Theoretical and behavioural works in the map: Nash (1950); Rubinstein (1982); Walton and McKersie (1965); Raiffa (1982); Lax and Sebenius (1986); Thompson and Hastie (1990); Galinsky and Mussweiler (2001); Van Kleef et al. (2004); Curhan et al. (2006); Chawla et al. (2021); see also Bazerman et al. (2000).