STRATEGIC TRADEOFFS BETWEEN HUMANS AND AI IN MULTI-AGENT BARGAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Coordination tasks traditionally performed by humans are increasingly being delegated to autonomous agents. As this pattern progresses, it becomes critical to evaluate not only these agents' performance but also the processes through which they negotiate in dynamic, multi-agent environments. Furthermore, different agents exhibit distinct advantages: traditional statistical agents, such as Bayesian models, may excel under well-specified conditions, whereas large language models (LLMs) can generalize across contexts. In this work, we compare humans (N = 216), LLMs (`GPT-4o`, `Gemini 1.5 Pro`), and Bayesian agents in a dynamic negotiation setting that enables direct, identical-condition comparisons across populations, capturing both outcomes and behavioral dynamics. Bayesian agents extract the highest surplus through aggressive optimization, at the cost of frequent trade rejections. Humans and LLMs can achieve similar overall surplus, but through distinct behaviors: LLMs favor conservative, concessionary trades with few rejections, while humans employ more strategic, risk-taking, and fairness-oriented behaviors. Thus, we find that performance parity—a common benchmark in agent evaluation—can conceal fundamental differences in process and alignment, which are critical for practical deployment in real-world coordination tasks.

1 INTRODUCTION

Historically, bespoke statistical models have excelled in structured economic exchanges with well-defined action spaces (Sun and Müller, 2013; Dehghanpour et al., 2016). More recently, the generalizability and flexibility of large language models (LLMs) have enabled delegation in unstructured contexts: negotiating supply chain contracts (Van Hoek et al., 2022), coordinating international trading (Corvin, 2024), and facilitating online commerce transactions (PYMNTS, 2025; Gaarlandt et al., 2025). Unlike more structured models, LLMs can generalize under uncertainty and partial information, reducing the need for task-specific customization. As AI agents grow more sophisticated and are entrusted with higher-stakes decisions, understanding how different types of agents behave in dynamic, multi-agent settings becomes essential. In particular, navigating the trade-offs between agent types—not only in terms of negotiated outcomes, but also in their alignment with human strategies and norms—is critical for responsible delegation.

A barrier to understanding delegation tradeoffs has been the difficulty of evaluating disparate agent types in comparable interactive settings. Static QA-style benchmarks fail to capture the dynamic, multi-agent interactions that characterize real-world negotiation. Indeed, recent work has called for more realistic evaluation environments that reflect the social and strategic complexities faced by deployed AI agents (Raman et al., 2024; Goktas et al., 2025).

We address this challenge by introducing a novel multi-player bargaining game that enables direct comparison of populations (e.g., humans, LLMs, and statistical models) under identical conditions. In our game, players receive endowments of colored chips with private valuations and negotiate trades to maximize their individual surplus. Bargaining is an ideal testbed for dynamic, multi-agent evaluations, as it inherently exposes tensions between social norms, individual preferences, and strategic reasoning, reflecting the complexities of many real-world environments. Moreover, from a theoretical perspective, bargaining under partial information admits no Pareto efficient, incentive-compatible solutions (Myerson and Satterthwaite, 1983), forcing agents to navigate trade-offs between individual gain and mutually beneficial cooperation.

We conduct an empirical study of our game across three populations: (i) human participants (N = 216), (ii) LLM-based agents, and (iii) specialized Bayesian agents, all facing identical randomized endowments. The Bayesian agents achieved the highest surplus, proposing aggressive bargains at the cost of frequent rejections. LLMs matched human surplus in certain game conditions but exhibited fundamentally different behaviors: where humans made fairness-driven concessions, LLMs favored conservative, low-variance trades with high acceptance rates.

More broadly, these findings expose a limitation of outcome-based evaluation: aggregate surplus can mask important differences in behavior and alignment. As AI systems increasingly mediate real-world negotiations, it becomes essential to assess not just what agents achieve, but how they navigate trade-offs, adjust to strategic dynamics, and engage with social norms.

2 RELATED WORK

AI agents have excelled in well-defined tasks with explicit objectives, preferences, and constrained action spaces (Sun and Müller, 2013; Dehghanpour et al., 2016). Recent advances in LLMs have broadened the scope of delegation to include tasks requiring broader social reasoning, such as preference elicitation, strategic pricing, natural-language negotiation, and alliance formation in multi-agent games like Diplomacy (Fish et al., 2024; Soumalias et al., 2025; Tessler et al., 2024; Vaccaro et al., 2025; Lubars and Tan, 2019; Bakhtin et al., 2022; Kramár et al., 2022). Beyond social reasoning, LLMs exhibit emergent capabilities in economic computation and human alignment, replicating behaviors in bargaining, auctions, and other economic games (Horton, 2023; Aher et al., 2022; Argyle et al., 2022), and reflecting human subgroup-specific patterns when conditioned on demographic traits (Manning et al., 2024; Xie et al., 2024; Zhu et al., 2024; Park et al., 2023).

However, successful deployment depends not just on capabilities (Jahani et al., 2025), but also on user expectations, skills, and alignment with social norms; users may suboptimally delegate when system behavior violates these expectations (Mozannar et al., 2023; Agarwal et al., 2023; Mullainathan and Obermeyer, 2021; Palminteri et al., 2024; Kapania et al., 2022; Qian and Wexler, 2024). Addressing alignment gaps requires evaluating both capabilities and behavior in dynamic, social contexts—a need recognized by recent calls for rigorous methods to steer and assess agent behavior in strategic multi-agent settings (Raman et al., 2024; Goktas et al., 2025). To concretely examine these challenges, we study how humans, LLM-based agents, and Bayesian models behave in a dynamic bargaining task, offering a comparative lens on alignment in strategic social interactions.

3 THE BARGAINING GAME

We introduce a multi-player bargaining game which captures strategic trade-offs, social signaling, and coordination challenges found in real-world negotiations. Figure 1 presents an overview of the game, in which three players aim to maximize individual surplus over nine turns. Each begins with ten chips of each color. Green chips act as a shared numeraire valued at $0.50 by all; other chip colors (red, blue, purple) have private values, drawn uniformly between $0.10 and $1.00. Game complexity is varied by incrementally adding chip types: red, then blue, then purple, with green always present.[1]

Gameplay. In a fixed, random order, players propose trades by offering chip(s) of one color in exchange for another. The two non-proposers simultaneously decide whether to accept or decline the proposal; if both accept, one is chosen at random to clear the transaction. Players can only propose and accept offers that they have sufficient inventory to fulfill. All players observe transactions and chip holdings, but not others' chip valuations. At the end of the game, each player receives a payout based on the surplus generated from their final chip holdings.

4 A PARETO OPTIMAL UPPER BOUND

Before exploring empirical outcomes, we provide a theoretical upper bound analysis. Although our bargaining game lacks a tractable exact game-theoretic solution, we can compute a Pareto-optimal allocation via linear programming to benchmark agent performance in our empirical results.

---

[1]We show in Appendix A.2 that decision complexity grows with the number of chip colors.
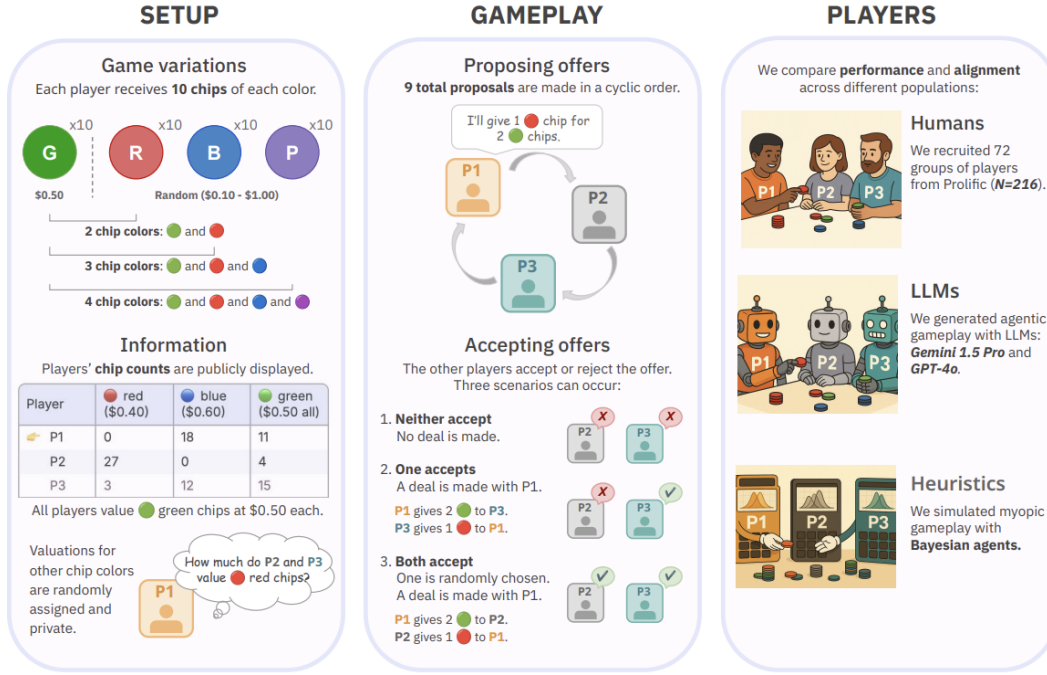
Figure 1: This provides a broad overview of the bargaining game. *Left:* Game setup. *Center:* Simplified version of gameplay. *Right:* The agent populations evaluated in this study.

4.1 NOTATION

Consider a game $\mathbf{M} = (\mathbf{I}, \mathbf{G}, \mathbf{v}, \mathbf{a})$ with a set of agents $\mathbf{I}$ and chips $\mathbf{G}$. Each agent i has a personal valuation v_{ig} for each good g. These valuations are static, independent across agents, and complete for each agent-good pair. All agents have an initial allocation a_{ig}^0 of each chip, which indicates the amount of chip g allocated to agent i. Given these allocations, an agent's **welfare** w_i is the sum of their valuations weighted by their allocations: $w_i = \sum_{g \in G} v_{ig} a_{ig}$. The total welfare $\mathbf{w}$ is thus the sum of all individual welfare: $\mathbf{w} = \sum_{i \in I} w_i = \sum_{i \in I} \sum_{g \in G} v_{ig} a_{ig}$. We will refer to the difference between initial and post-trade welfare as **surplus gain** in the following analysis.

4.2 EFFICIENT ALLOCATIONS

In our game, agents trade chips to improve their welfare. Given their initial endowments a_{ig}^0, mutually beneficial trades may exist that increase total welfare. We provide an upper bound to the obtainable total welfare, assuming the Pareto condition that no player ends with worse utility than they began.

Definition 1. (Optimal Allocation of Chips). *An optimal allocation of chips* is any set of allocations $\mathbf{A}^* = \{a_{ig}^* \mid i \in I, g \in G\}$ that solves:

$$\arg\max_{\mathbf{A}} \sum_{i \in I} \sum_{g \in G} v_{ig} a_{ig}$$

subject to:

(i) **Conservation of goods**: $\sum_{i \in I} a_{ig} = \sum_{i \in I} a_{ig}^0, \forall\, g \in G$

(ii) **Pareto improvement**: $\sum_{g \in G} v_{ig} a_{ig} \geq \sum_{g \in G} v_{ig} a_{ig}^0, \forall\, i \in I.$

(iii) **Non-negativity**: $a_{ig} \geq 0 \forall\, i \in I, g \in G.$

Intuitively, in a centralized setting with full information, $\mathbf{A}^*$ represents a Pareto-efficient allocation that maximizes aggregate utility without making any agent worse off. While this optimal allocation

$\mathbf{A}^*$ may not be unique, the maximum total welfare $\mathbf{w}^*$ achieved at the optimum is unique. This optimization problem can be solved efficiently using interior point methods. Its computational complexity is bounded by $O\big((|I||G|)^{3.5}\big)$.

Note that this upper bound is unlikely to be achieved in practice, as no mechanism for decentralized bargaining with private information can guarantee both Pareto efficiency and incentive compatibility Myerson and Satterthwaite (1983).[2]

5 EXPERIMENTAL SETUP

In our design, human participants first completed games in groups of three. We then replicated each game (identical initial chip endowments) with LLM- and Bayesian- agent groups, enabling direct comparison across populations.[3]

5.1 HUMAN SUBJECTS

We recruited 216 U.S.-based participants from Prolific to complete the bargaining game on *Deliberate Lab* (Tsai et al., 2024).[4] Participants completed two games over a 30-minute session, which included time waiting for two others to join a live game. To incentivize strategic play, participants received the surplus value of their final chip holdings in addition to a \$4 base payment. The average total payout was \$12.24 ± \$6.12. Each participant completed two games with different chip valuations and profiles. For each of the 2-, 3-, and 4-chip game variants, 24 groups of three players played 2 games each, yielding $N = 3 \times 24 \times 3 \times 2 = 432$ effective participants across 144 total games.[5]

5.2 LLM AGENTS

We then replicated each human game with LLM agents. At each turn, an LLM agent received the same information as its human counterpart, including private chip valuations, public chip holdings, and trade history. One agent proposed a trade per round; the others evaluated it.

We created two types of agents. As our primary goal was to evaluate baseline performance under minimal optimization, we first constructed an *out-of-box* agent used simple prompts adapted from the human instructions. To assess the benefit of lightweight refinement, we also implemented a *refined* agent which generated candidate proposals and selected one using "type-2" reasoning, mimicking slower, more deliberate decision-making (Furniturewala et al., 2024).

To generalize our findings, we constructed LLM populations using two model families: OpenAI's `GPT-4o` (OpenAI et al., 2024) and Google's `Gemini 1.5 Pro` (Team et al., 2024), which were the latest full-capacity models from each family at time of writing.[6] We set prompt temperature to 0.5, following mid-range recommendations for structured reasoning tasks that require coherence and variability (Arora et al., 2024). Prompts were scaffolded using a chain-of-thought strategy through the open-source EDSL library (Wei et al., 2022; Horton et al., 2024).

5.3 BAYESIAN AGENTS

The Bayesian agents were programmed specifically to play this bargaining game. Their strategy is formalized as follows. Each agent i forms a joint belief about the valuations of all other players, $\mathbf{v}_{-i}$. When agent i proposes a trade (x_g, y_r) – offering x chips of type g in exchange for y chips of type r – they choose a trade to maximize their expected payoff:

---

[2]A proof that players of the bargaining game lack a dominant strategy is provided in AppendixA.1.

[3]For reproducibility, details on the human game interface are provided in Appendix B.1. Full LLM prompts are provided in Appendix G, and the Bayesian agent algorithms are detailed in Appendix A.3.

[4]IRB approval and informed consent were obtained; we applied no inclusion criteria for recruitment.

[5]The two games are evaluated independently, as ordering and learning effects were minimal.

[6]For robustness, we also benchmarked the latest lightweight models (`Gemini 2.5 Flash`, `GPT-o4-mini`), but they failed to execute basic surplus-maximizing trades. We therefore restrict our main analysis to full-capacity models; see Appendix F for additional context.

$$\max_{(\mathbf{x},\mathbf{y})} \sum_{\mathbf{v}_{-i}} \mathbf{1}\left\{\exists\, j \neq i : \mathrm{accept}\big(v_j, x_g, y_r\big)\right\} \,\times\, \Delta u_i(v_i, x_g, y_r) \,\times\, B_i\big(\mathbf{v}_{-i}\big), \tag{1}$$

where $\mathrm{accept}\big(v_j, x_g, y_r\big)$ indicates that player j with valuation v_j would accept the trade, $\Delta u_i(v_i, x_g, y_r) = v_{i,r} y_r - v_{i,g} x_g$ is player i's change in utility and $B_i(\mathbf{v}_{-i})$ is player i's belief (probability) that the other players' valuations are $\mathbf{v}_{-i}$. This agent assumes that their opponent's acceptance decision is *myopically rational*, i.e. a receiver with valuation v_j accepts if and only if they have sufficient chips for the trade and their expected utility is positive, i.e. $v_{j,g} x_g - v_{j,r} y_r > 0$.

All players update their priors after observing acceptance decisions. A player discards valuation states $\{v_j\}$ that contradict the observed event, assuming myopically rational opponents, and re-normalizes the remaining probabilities. When the trade is accepted, the proposer keeps only valuations that would be accepted in its distribution over the chosen receiver; the receiver keeps only valuations that would be accepted in its distribution over the proposer. Bystanders update priors for both accordingly. When there is no trade, all players discard valuations that conflict with the observed non-acceptance.
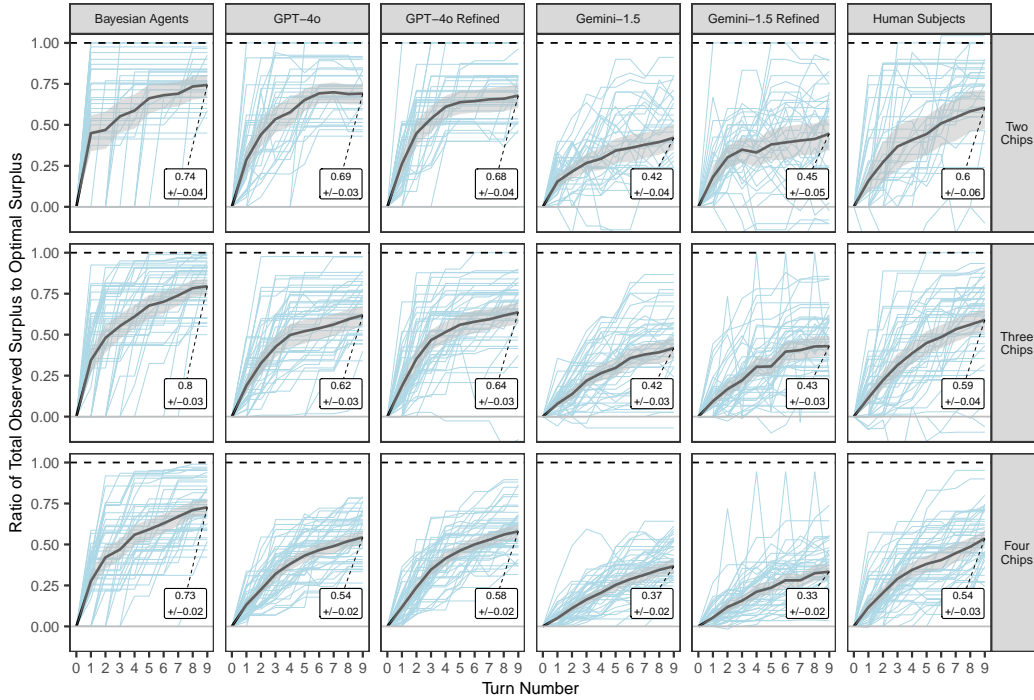
6 RESULTS

6.1 PERFORMANCE

Figure 2: Surplus trajectories for human, LLM, and Bayesian-learning agent simulations across game complexities. Each unique game yields a blue line corresponding to the ratio of surplus achieved relative to the computed optimal allocation. Means and 95% confidence intervals are highlighted and listed in Appendix 3. All populations generated positive aggregate surplus.

Figure 2 shows surplus trajectories by agent types and game variation. In the upper right panel, we see that Bayesian agents reached 74% of the Pareto-efficient optimum in the two-chip game. Across all variations, Bayesian agents achieved the highest surplus. GPT-4o performed comparably to human

participants; however, humans statistically outperformed `Gemini 1.5 Pro`.[7] Refined LLMs did not significantly impact surplus compared to out-of-the-box LLMs (statistical tests in Appendix E).

6.2 PROCEDURAL ALIGNMENT

In addition to performance, we analyze *procedural alignment*, how the agents engage in the mechanics of negotiation, through two lenses: 1) observed trading patterns, and 2) regret minimization strategies.

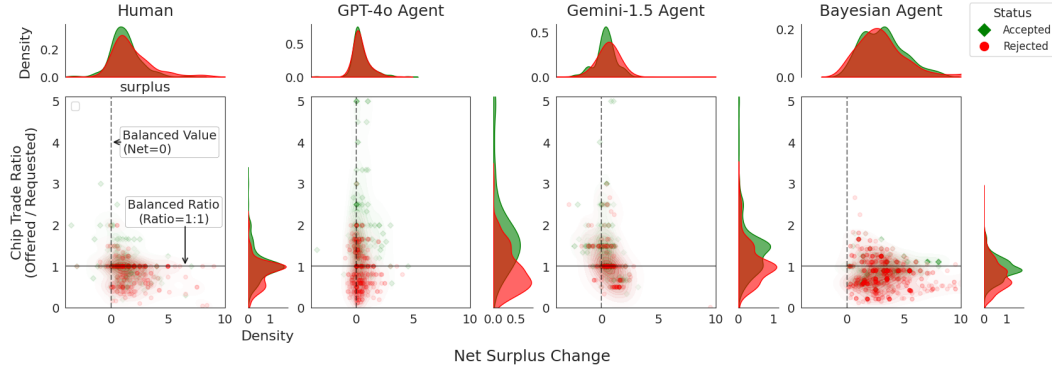### 6.2.1 TRADING PATTERNS AND VOLUME DYNAMICS

Figure 3: Trading patterns in the 3-chip game, visualizing (i) net surplus change for the proposer and (ii) trade ratio. Accepted trades in green, rejections in red. Marginal distributions across each axis are adjacent to the plots. The vertical dashed line marks **zero net surplus** (no net value created), and the horizontal solid line marks **balanced exchange** (1:1 ratio of chips). Figure values are in Table 5.

Humans and social norms. Figure 3 plots individual trade proposals for the three-chip game for each agent type. Results for the two- and four-chip games are qualitatively similar (see Appendix F.2). The vertical balanced value line indicates proposals that generated no net surplus, while the horizontal balanced ratio line corresponds to 1:1 exchanges of chips. In the left panel, we see that the human proposals cluster around the balanced ratio line: that is, humans often offer a similar number of chips as they are requesting. Most trade proposals lie to the right of the balanced value line (i.e., provide positive surplus to the proposer), though some reduce total surplus.

LLMs and concessionary exchanges. Both LLM populations propose trades near the vertical balanced net value line (yielding little total surplus), with no anchoring to balanced trade ratios. A noticeable vertical tail of proposals extends above the 1:1 balanced ratio line, indicating a tendency to offer more chips than requested (even up to 5:1). Many proposals also incur a net surplus loss.

Bayesian agents and rational surplus maximization. By construction, Bayesian agents strictly maximize surplus and avoid net-loss trades (no points left of the balanced value line). Their proposals reach deep into the positive surplus region. Rather than the 1:1 balanced ratio exhibited in human proposals, the Bayesian offers cluster below the parity line (ratio < 1), meaning they ask for more chips than they give. Proposal rejections are high.

In sum, these trading patterns reveal clear strategic differences: humans offer trades suggesting a consideration for balanced, fair exchanges (Danz et al., 2022), LLMs adopt a "concessionary posture" with high proposal acceptance rates, and Bayesian agents adopt a value-extractive strategy (Fisher et al., 2011), asking to receive more than they give and incurring higher rejection rates. To evaluate whether any of these strategies were more optimal ex-post, we next evaluate these trades using a regret-minimization framework.

---

[7]This may reflect the recency of such capabilities; `Gemini 1.5 Pro` was released three months prior to `GPT-4o` and represented an earlier generation of model development.

6.2.2 STRATEGIC ALIGNMENT AND REGRET MINIMIZATION

At each turn, a participant might take an action of *proposing*, *accepting*, or *declining* a trade. In **no regret** actions, no feasible later alternative would yield a better valuation for the target chip. **Forced regret** occurs when a higher-surplus option arises later, but the agent's earlier decisions—such as premature trades or overcommitting inventory—prevent acting on it. Finally, **unforced regret** captures scenarios where a better option becomes available, yet the agent fails to take it.

To avoid speculative modeling of unobserved strategies, we limit the scope of our analysis. First, we only consider myopically rational transactions (i.e., positive surplus) for proposers, acceptors, or decliners.[8] Interpreting irrational trades, which may be driven reciprocity, signaling, or cognitive error, would require assumptions we cannot credibly test. Second, when evaluating proposer behavior, we focus on accepted trades; rejected offers may result from strategic signaling by either party or proposer misjudgment of demand.[9]Third, "acceptors" include those who intended to accept a trade, not only those who were selected to trade.[10] This preserves the behavioral signal of willingness to accept, without conflating risk tolerance or random assignment.

Figure 4 shows an example of a player's valuations as they trade over the course of a single game.[11] On turn 1, Player 3 exhibited a **forced regret** action as an acceptor: they accepted a trade which resulted in their having insufficient chips to accept a more profitable offer in turn 2. On turn 7, Player 2 exhibited an **unforced regret** action as a decliner; they declined a profitable trade and did not recover the profits later. On Turn 8, Player 2 made a **no regret** proposal; this was the best deal they could have made, as there were no better future trades.[12]
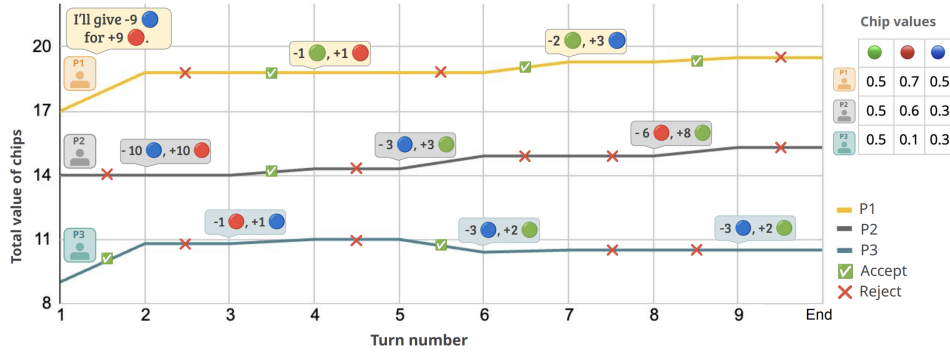

Figure 4: Example of player chip value trajectories over nine turns of a trading game (3-Chip). To avoid speculation, we do not analyze analyze Player 1's proposal on Turn 4 (an unaccepted trade), or Player 3's acceptance on Turn 5 (accepting a negative surplus offer).

Figure 5 shows the classifications of actions across games and populations.[13] The relative performance of Bayesian agents improves as game complexity increases. Combined with the surplus outcomes in Figure 2, this suggests that Bayesian agents are not necessarily improving, but rather that LLM and human performance degrades as the action space expands.

Bayesian agents generate a higher proportion of optimal (no regret) proposals and fewer regrettable ones compared to LLMs and humans, reflecting more effective surplus-maximizing calculations. However, Bayesian agents also register the most declines—both regrettable and non-regrettable. This indicates that their proposal strategies prioritize their own advantage, without necessarily considering whether those offers will be accepted by others.

In contrast, humans and LLMs exhibit more regrettable proposals and acceptances relative to Bayesian agents. They also exhibit fewer declines overall, indicating that they propose more agreeable offers at

---

[8]Bayesian agents satisfy this by design.

[9]This analysis is distinct from evaluations conducted from the perspective of decliners.

[10]This occurs when both accept the trade and one is randomly chosen.

[11]The sum of all players' surplus changes at each turn correspond to the blue lines in Figure 2.

[12]More detailed explanations are provided in Appendix D.

[13]We exclude refined LLMs as they behave similarly to out-of-the-box models; see Appendix F.1.
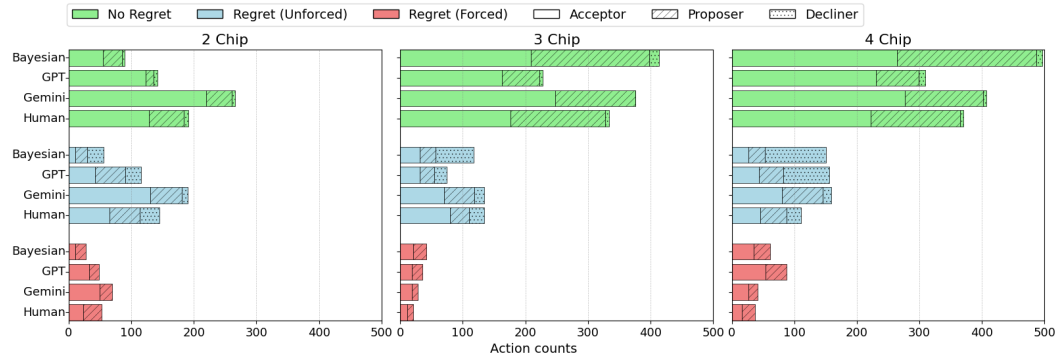
Figure 5: Counts of strategic actions across games. Bayesian agents performed more optimal, no-regret actions relative to other populations as game complexity increased. Declining a proposal would not cause forced regret, as this action does not commit any chips. The total number of actions naturally increases with game complexity, reflecting the larger solution space, not necessarily improved decision-making.

the cost of making less optimal actions. Differences in forced regret actions are relatively modest, though LLMs show slightly higher forced regret counts.

7 DISCUSSION

Humans reflect economic and social reasoning. Humans offered trades consistent with classical fairness norms, proposing a balanced trade ratio. While these proposals were less frequently optimal than that of Bayesian agents, they were also accepted more often. Post-game survey responses consistently emphasized fairness and cooperation as decision drivers, suggesting that humans prioritize social norms, even in one-shot interactions where long-term reputation effects would not affect payoff.[14] However, nuanced human social motivations are difficult to formalize into algorithms, posing a fundamental challenge when delegating coordination and negotiation to artificial agents.

LLMs can perform on-par, but may be concessionary by design. The latest model, GPT-4o, achieved human-level surplus with minimal prompting, demonstrating flexibility when applied to novel tasks. However, their behavior was compromising and reactionary: they favored value-balanced proposals which were less optimal than Bayesian proposals but more likely to be accepted, and exhibited high rates of forced regret trades, suggesting limited strategic foresight in dynamic scenarios. This conservative play style may arise from several factors: (i) training on cooperative, information-sharing dialogues, where minimizing friction is implicitly rewarded (Wei et al., 2022); (ii) a tendency to generate risk-averse, passive responses compounded by a lack of outcome-driven feedback (Ouyang et al., 2022); and/or (iii) a possible learned aversion to asymmetric outcomes, where generous offers serve as a coordination strategy to secure agreement (Fisher et al., 2011). In adversarial or zero-sum negotiations, this overly concessionary tendency could lead LLMs to systematically overcompromise.

Bayesian agents outperform, but lack social adaptability. Bayesian agents excelled at surplus-maximizing proposals, aggressively extracting value. This strong performance was constructed through a hand-crafted algorithm specifically aligned to the game's incentives. In contrast, humans and LLMs relied on their flexible "out of the box" capabilities, lacking such tailored inductive biases.

In practice, Bayesian agents' extractive, rejection-tolerant strategy might fail in real-world negotiations, especially in repeated interactions requiring trust and reciprocity. Extending Bayesian agents to more socially compatible trading would require richer models of preferences, fairness, and adaptive behavior, which could increase complexity and decrease robustness.

---

[14]Post-game survey questions and quantitative responses are in Appendix C.

Hybrid models offer promise, but need social reasoning.   We find that humans consider social factors, LLMs favor safe, consensus-seeking strategies, and Bayesian agents excel through narrow, task-specific optimization. These contrasting strengths suggest potential for complementary approaches, where LLMs are augmented with Bayesian tools or planning modules to improve foresight, such as inventory management and lookahead reasoning. However, negotiation is not just a planning problem; success also depends on modeling others' intentions, social norms, and the unwritten rules of cooperation. Without explicit social reasoning, even hybrid agents will struggle with the social nuances that humans navigate instinctively.

Procedural alignment matters for human-AI interaction.   Furthermore, how agents negotiate is as important as what they achieve. Despite humans and LLMs yielding similar aggregate surplus in some contexts, they exhibited distinct strategies. Crucially, these differences only became visible through a custom experimental design that enabled matched, one-to-one comparisons across populations; simple outcome-level aggregation would have obscured these divergences. As AI agents increasingly participate in human-facing decision-making, focusing solely on efficiency metrics risks missing misalignments in process, intent, and social compatibility.

Efficiency and scale advantages.   A final consideration beyond strategy is scale. LLMs and Bayesian models can resolve negotiations near-instantaneously, unaffected by cognitive load or decision fatigue. LLMs, in particular, are faster to deploy out-of-the-box compared to bespoke Bayesian models; in our case, our out-of-the-box agents required minimal prompt scaffolding around the human instructions, whereas we constructed a custom algorithm for the Bayesian agents. These efficiency gains make automation attractive for high-volume or time-sensitive negotiations, even if strategic or social limitations remain.

Social impacts and risks.   Agents that achieve surplus-yielding outcomes through fundamentally different negotiation styles pose risks for trust, fairness, and social acceptance. If deployed in real-world negotiations, models like LLMs may default to safe, concessionary behavior, leading to suboptimal deals and reinforcing status-quo power dynamics. Conversely, optimization-focused agents like Bayesians may extract disproportionate value, undermining perceptions of fairness and cooperation. These dynamics may lead to unexpected interaction effects in mixed-agent environments. Designing AI agents that are not only strategically capable but also socially compatible will be critical for their long-term acceptance and responsible deployment.

8 LIMITATIONS AND FUTURE WORK

Our study focuses on a single, stylized negotiation game with static valuations and single shot interactions, conditions that align well with Bayesian agents' task-specific optimization and may exaggerate their advantage. While humans and LLMs bring general-purpose reasoning to the task, their broader capabilities in dynamic, multi-round, or reputation-driven negotiations remain unexplored. Furthermore, our LLMs used minimal prompting, reflecting realistic deployment but limiting exploration of more strategic behaviors from alternative prompts or fine-tuning.

Beyond performance metrics, our analysis reveals distinct behavioral profiles but does not disentangle whether these arise from reasoning limitations, social norms, or inductive biases inherent to each system. Future work should investigate how humans perceive and trust AI negotiators with different procedural styles. Another promising avenue for future work is to allow agents to communicate with natural language. Finally, developing methods to steer LLM agents towards desired procedural norms explicitly represents a significant ongoing challenge and opportunity.

9 CONCLUSION

This study demonstrates an empirical approach for comparing agents in dynamic, social settings. By examining human, LLM, and Bayesian behavior in a structured bargaining game, we show how performance measures can conceal fundamentally different strategic approaches. As AI systems are increasingly delegated complex tasks, understanding how their decision-making processes align with human expectations and social values is critical to fostering trust, effective cooperation, and responsible deployment in real-world applications.

REFERENCES

N. Agarwal, A. Moehring, P. Rajpurkar, and T. Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research, July 2023. URL http://www.nber.org/papers/w31422.

G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022.

L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*, 2022.

C. Arora, A. I. Sayeed, S. Licorish, F. Wang, and C. Treude. Optimizing large language model hyperparameters for code generation, 2024. URL https://arxiv.org/abs/2408.10577.

A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

A.-M. Corvin. Web summit 2024: Alibaba conjures up global trade ai, portugal launches own llm. *TechInformed*, November 2024. URL https://www.techinformed.com/web-summit-2024-alibaba-conjures-up-global-trade-ai-portugal-launches-own-llm/.

D. Danz, L. Vesterlund, and A. J. Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883, 2022.

K. Dehghanpour, M. H. Nehrir, J. W. Sheppard, and N. C. Kelly. Agent-based modeling in electrical energy markets using dynamic bayesian networks. *IEEE Transactions on Power Systems*, 31(6):4744–4754, 2016. doi: 10.1109/TPWRS.2016.2524678.

S. Fish, Y. A. Gonczarowski, and R. I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7, 2024.

R. Fisher, W. L. Ury, and B. Patton. *Getting to yes: Negotiating agreement without giving in*. Penguin, 2011.

S. Furniturewala, S. Jandial, A. Java, P. Banerjee, S. Shahid, S. Bhatia, and K. Jaidka. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.13. URL https://aclanthology.org/2024.emnlp-main.13/.

J. Gaarlandt, W. Korver, N. Furr, and A. Shipilov. Ai agents are changing how people shop. here's what that means for brands. *Harvard Business Review*, 2025.

D. Goktas, A. Greenwald, T. Osogami, R. Patel, K. Leyton-Brown, G. Schoenebeck, D. Cornelisse, C. Daskalakis, I. Gemp, J. Horton, D. C. Parkes, D. M. Pennock, A. Prakash, S. S. Ravindranath, M. O. Smith, G. Swamy, E. Vinitsky, S. Wasserkrug, M. Wellman, J. Wu, H. Xu, J. Zhang, Y. Zhang, S. Zhao, and Q. Zhu. Strategic Foundation Models. working paper or preprint, Feb. 2025. URL https://hal.science/hal-04925309.

J. Horton, A. Filippas, and R. Horton. Edsl: Expected parrot domain specific language for ai powered social science. Whitepaper, Expected Parrot, 2024.

J. J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

E. Jahani, B. S. Manning, J. Zhang, H.-Y. TuYe, M. Alsobay, C. Nicolaides, S. Suri, and D. Holtz. Prompt adaptation as a dynamic complement in generative ai systems, 2025. URL https://arxiv.org/abs/2407.14333.

S. Kapania, O. Siy, G. Clapper, A. M. SP, and N. Sambasivan. "because ai is 100% right and safe": User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517533. URL https://doi.org/10.1145/3491102.3517533.

J. Kramár, T. Eccles, I. Gemp, A. Tacchetti, K. R. McKee, M. Malinowski, T. Graepel, and Y. Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications*, 13(1):7214, 2022.

N. Lambert and R. Calandra. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*, 2023.

B. Lubars and C. Tan. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d67d8ab4f4c10bf22aa353e27879133c-Paper.pdf.

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who should predict? exact algorithms for learning to defer to humans, 2023. URL https://arxiv.org/abs/2301.06197.

S. Mullainathan and Z. Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care*. *The Quarterly Journal of Economics*, 137(2):679–727, 12 2021. ISSN 0033-5533. doi: 10.1093/qje/qjab046. URL https://doi.org/10.1093/qje/qjab046.

R. B. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(83)90048-0. URL https://www.sciencedirect.com/science/article/pii/0022053183900480.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

S. Palminteri, B. Garcia, and C. Qian. The moral turing test: Evaluating human-llm alignment in moral decision-making, Oct 2024. URL osf.io/preprints/psyarxiv/ct6rx_v1.

J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

PYMNTS. Openai turns chatgpt into shopping assistant with new commerce features. *PYMNTS*, April 2025. URL https://www.pymnts.com/news/artificial-intelligence/2025/openai-turns-chatgpt-into-shopping-assistant-with-new-commerce-features/.

C. Qian and J. Wexler. Take it, leave it, or fix it: Measuring productivity and trust in human-ai collaboration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 370–384, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705083. doi: 10.1145/3640543.3645198. URL https://doi.org/10.1145/3640543.3645198.

N. Raman, T. Lundy, S. Amouyal, Y. Levine, K. Leyton-Brown, and M. Tennenholtz. Steer: Assessing the economic rationality of large language models. *arXiv preprint arXiv:2402.09552*, 2024.

E. Soumalias, Y. Jiang, K. Zhu, M. Curry, S. Seuken, and D. C. Parkes. Llm-powered preference elicitation in combinatorial assignment. *arXiv preprint arXiv:2502.10308*, 2025.

Z. Sun and D. Müller. A framework for modeling payments for ecosystem services with agent-based models, bayesian belief networks and opinion dynamics models. *Environmental modelling & software*, 45:15–28, 2013.

G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024.

V. Tsai, C. Qian, and D. L. community contributors. Deliberate Lab: Open-Source Platform for LLM-Powered Social Science, Sept. 2024. URL https://github.com/PAIR-code/deliberate-lab.

M. Vaccaro, M. Caoson, H. Ju, S. Aral, and J. R. Curhan. Advancing ai negotiations: New theory and evidence from a large-scale autonomous negotiations competition, 2025. URL https://arxiv.org/abs/2503.06416.

R. Van Hoek, M. DeWitt, M. Lacity, and T. Johnson. How walmart automated supplier negotiations. *Harvard Business Review*, 8(11):2022, 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022.

T. Xia, Z. He, T. Ren, Y. Miao, Z. Zhang, Y. Yang, and R. Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3579–3602, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.213. URL https://aclanthology.org/2024.findings-acl.213/.

C. Xie, C. Chen, F. Jia, Z. Ye, S. Lai, K. Shu, J. Gu, A. Bibi, Z. Hu, D. Jurgens, J. Evans, P. Torr, B. Ghanem, and G. Li. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=CeOwahuQic.

K. Zhu, J. J. Horton, Y. Jiang, D. C. Parkes, and A. V. Shah. Evidence from the synthetic laboratory: Language models as auction participants. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024. URL https://openreview.net/forum?id=FB9mTtJpJI.

A ALGORITHMS AND PROOFS

A.1 PROOF OF NO DOMINANT STRATEGY

Claim. *No agent $i \in \mathbf{I}$ in the game $\mathcal{M}$ has a dominant strategy.*

Setup. Let $\mathbf{I}$ be the set of agents, $\mathbf{G}$ be the set of goods. For each agent $i \in \mathbf{I}$ and good $g \in \mathbf{G}$, there is a valuation $v_{ig} \in \mathbb{R}_{\geq 0}$. The (feasible) allocations are given by a_{ig}, the amount of good g allocated to agent i.

The **welfare** of agent i is

$$w_i = \sum_{g \in \mathbf{G}} v_{ig} \, a_{ig}.$$

The **total social welfare** is

$$w = \sum_{i \in \mathbf{I}} w_i = \sum_{i \in \mathbf{I}} \sum_{g \in \mathbf{G}} v_{ig} \, a_{ig}.$$

Let each agent i have a strategy space $\mathcal{S}_i$. A (pure) strategy $s_i \in \mathcal{S}_i$ is a complete description of agent i's behavior in the market (e.g., how they propose, reject, accept, etc.). We write $\mathbf{s}_{-i}$ for a strategy profile of all agents other than i. Let $u_i(s_i, \mathbf{s}_{-i})$ denote agent i's resulting payoff (or final utility/welfare) under strategy s_i against opponents' strategies $\mathbf{s}_{-i}$. A strategy s_i^* is called *dominant* for agent i if, for **all** $\mathbf{s}_{-i} \in \mathcal{S}_{-i}$,

$$u_i\big(s_i^*, \mathbf{s}_{-i}\big) \; \geq \; u_i\big(s_i, \mathbf{s}_{-i}\big) \quad \text{for all } s_i \in \mathcal{S}_i.$$

In other words, s_i^* guarantees the highest possible utility for i, irrespective of what other players do.

Proof. Assume, by way of contradiction, that there exists a dominant strategy s_i^* for some agent i. By the definition of dominance, s_i^* must satisfy

$$u_i\big(s_i^*, \mathbf{s}_{-i}\big) \; \geq \; u_i\big(s_i, \mathbf{s}_{-i}\big) \quad \text{for all } s_i \in \mathcal{S}_i \text{ and for all } \mathbf{s}_{-i} \in \mathcal{S}_{-i}.$$

Construct two different scenarios of other agents' strategies and valuations, denoted by $\mathbf{s}_{-i}^A$ and $\mathbf{s}_{-i}^B$. These scenarios can differ in 1) how other agents value the goods $\{v_{jg} : j \neq i\}$, and 2) the demands or offers they make (i.e. how $\mathbf{s}_{-i}$ translates into allocations a_{jg}).

- In scenario A, suppose the other agents are willing to trade or allocate a certain good g^* to agent i only if i offers a high price or concedes in some manner. A specialized alternative strategy s_i^A (instead of s_i^*) might yield a strictly higher payoff if i trades aggressively for g^*.
- In scenario B, suppose the others behave differently (e.g., they no longer place much value on g^* but highly value some other good). Now a different specialized strategy s_i^B might yield a strictly higher payoff (because trading strategy for g^* in scenario A is no longer beneficial here).

Contradiction of dominance: Because scenarios A and B are both feasible profiles of opponents' behavior ($\mathbf{s}_{-i}^A$ and $\mathbf{s}_{-i}^B$), there is no single strategy s_i^* that can simultaneously guarantee at least as high a payoff as both s_i^A and s_i^B across those distinct opponent profiles. Formally,

$$\exists \mathbf{s}_{-i}^A : u_i\big(s_i^A, \mathbf{s}_{-i}^A\big) > u_i\big(s_i^*, \mathbf{s}_{-i}^A\big), \quad \text{and} \quad \exists \mathbf{s}_{-i}^B : u_i\big(s_i^B, \mathbf{s}_{-i}^B\big) > u_i\big(s_i^*, \mathbf{s}_{-i}^B\big).$$

This contradicts the definition of s_i^* as a dominant strategy.

Because the above argument does not rely on any special assumption beyond the existence of at least two different plausible profiles $\mathbf{s}_{-i}^A$ and $\mathbf{s}_{-i}^B$, we conclude that no strategy for agent i can be dominant. Since i was arbitrary, no agent in the game has a dominant strategy. $\square$

A.2 Empirical computation of increasing game complexity

Claim. *The decision space increases quadratically as the number of chip colors increases.*

Proof. This is a simple consequence that in games with k different colors of chips, there are $\binom{k}{2}$ possible pairs of chips to transact. Since the distribution of opponents' valuations is symmetric across chips (except the green numeraire chip), each proposer must consider $O(k^2)$ different feasible trades. Indeed, if we empirically compute the expected number of myopically rational trades[15] in the starting configuration of the 2, 3, and 4-chip game, we arrive at 37.1, 120.7, 250.7 respectively. That is the 3-chip game has approximately $\binom{3}{2}$ as many myopically rational trades in expectation as the 2-chip game, and the 4-chip game has approximately $\binom{4}{2}$. This approximation becomes more exact as the number of chips increases due to the reduced relative impact of the numeraire.

A.3 Multi-agent trading algorithm with Bayesian learning

Algorithm 1 Multi-agent Trading Algorithm with Bayesian Learning

---

   Inputs:
     $\mathbf{I}$: set of players
     $\mathbf{G}$: set of goods
     $\mathbf{A^0}$: initial allocation $\{a_{ig}\}$ for all $i \in \mathbf{I}, g \in \mathbf{G}$
     $\{B_i(\mathbf{v}_{-i})\}_{i \in \mathbf{I}}$: each player's prior belief over other players' valuations

1: $\mathbf{A} \leftarrow \mathbf{A^0}$ ▷ Initialize current allocation
2: **for all** $i \in \mathbf{I}$ **do**
3:     **initialize** $B_i(\mathbf{v}_{-i})$ ▷ Each agent's prior over other agents' valuations
4: **end for**
5: $possible\_trades \leftarrow \texttt{true}$
6: **while** $possible\_trades$ **do**
7:     $trades\_found \leftarrow \texttt{false}$
8:     $I_{\text{shuffled}} \leftarrow \text{RandomShuffle}(\mathbf{I})$
9:     **for** $k = 1$ **to** $|\mathbf{I}_{\text{shuffled}}|$ **do**
10:       **for** $l = k + 1$ **to** $|\mathbf{I}_{\text{shuffled}}|$ **do**
11:         $(i, j) \leftarrow (I_{\text{shuffled}}[k], I_{\text{shuffled}}[l])$
12:         $(\mathbf{A_{ij}^*}, outcome, r) \leftarrow \text{SOLVEA}^*\Big(i,\ j,\ \mathbf{G},\ \{a_{ig}, a_{jg}\}_{g \in \mathbf{G}},\ B_i,\ B_j\Big)$ ▷ Proposer i finds best trade; r is the responding agent (who accepts/rejects)
13:         **if** $outcome = \text{ACCEPT}$ **then**   ▷ Trade executed
14:           $\{a_{ig}, a_{jg}\}_{g \in \mathbf{G}} \leftarrow \mathbf{A_{ij}^*}$
15:           $trades\_found \leftarrow \texttt{true}$
16:         **end if**
17:         $\text{BAYESIANUPDATE}\big(i,\ j,\ r,\ outcome,\ B_i,\ B_j,\ \{B_k\}_{k \neq i,j}\big)$ ▷ All players update beliefs based on accept/reject
18:       **end for**
19:     **end for**
20:     $possible\_trades \leftarrow trades\_found$
21: **end while**
22: **return** $\mathbf{A}$

Explanation of key subroutines:

- $\text{SOLVEA}^*(i, j, \dots)$:
  - **Proposer's optimization.** Agent i (the proposer) solves

  $$\max_{(\mathbf{x},\mathbf{y})} \sum_{\mathbf{v}_{-i}} \mathbf{1}\Big\{\exists\, \text{receiver } r \neq i : \text{ACCEPT}(v_r; \mathbf{x}, \mathbf{y})\Big\} \times \Big(u_i\big(h_i - \mathbf{x} + \mathbf{y}\big) - u_i\big(h_i\big)\Big) \times B_i(\mathbf{v}_{-i}),$$

---

[15]We define a myopically rational trade to be one where the proposer gains surplus and assigns some probability that at least one opponent would gain surplus by accepting the trade.

where $\mathbf{x}$ are the chips given up by i and $\mathbf{y}$ are the chips requested, h_i denotes i's current holdings, and B_i is i's belief over others' valuations.

- **Receiver's acceptance criterion.** A potential responder r accepts the trade $(\mathbf{x}, \mathbf{y})$ *myopically* if

$$u_r(h_r - \mathbf{y} + \mathbf{x}; v_r) > u_r(h_r; v_r).$$

The subroutine returns $\mathbf{A}_{ij}^*$ (the updated allocation for i and j), the *outcome* (ACCEPT or REJECT), and the identity of the actual *responder* r (in case multiple receivers are considered, or if j is always the designated responder, $r = j$).

- BAYESIANUPDATE $\big(i, j, r, outcome, B_i, B_j, \{B_k\}_{k \neq i,j}\big)$:

  - On ACCEPTANCE: All agents discard any valuation states inconsistent with r accepting $(\mathbf{x}, \mathbf{y})$. Specifically,

  $$v_r \text{ is retained only if } u_r(h_r - \mathbf{y} + \mathbf{x}; v_r) > u_r(h_r; v_r).$$

  - On REJECTION: All agents discard any valuation states for r that would have *accepted* $(\mathbf{x}, \mathbf{y})$, since that contradicts the observed rejection.

  - Then, each agent renormalizes its belief distributions B_k to sum to 1 over the remaining (still-plausible) valuations.

This procedure repeats until no further beneficial trades can be found. The Bayesian updates ensure that, over time, agents refine their beliefs about each other's valuations based on observed accept/reject decisions, thereby *learning* which trades are more likely to succeed.

B HUMAN DATA COLLECTION

B.1 IMPLEMENTATION ON DELIBERATE LAB

Platform. The bargaining game was implemented via Deliberate Lab (Tsai et al. (2024)), an open-source platform for conducting group human-LLM social science experiments. We will publish the bargaining game implementation within the platform upon acceptance.

Participant interface. Upon entering Deliberate Lab through a web link, participants proceed through a multi-stage experiment including informed consent, game instructions, comprehension checks, and payout information. Upon completing the final comprehension check, they wait in a "Lobby" stage for other participants. When three participants are in the lobby, they are sent an invitation to join a live bargaining game (Fig 6). Following the game, there is a post-game survey. For anonymity, we used a Deliberate Lab feature that assigns participants an anonymous animal avatar (e.g., "Bear") as they join the experiment.

Bargaining game user interface. To submit trade proposals, participants used text and selection fields to specify number and type of chip, respectively. The interface automatically calculated and displayed projected total chip value if the proposal were to be accepted (Fig 6); it also prevented players from submitting invalid offers by disabling the "submit" button.

Experimenter interface. To conduct experiments, the experimenter used Deliberate Lab's dashboard (Fig 8) to monitor participants as they progressed through the experiment's information and comprehension check stages, then manually transferred participants into groups of three to play the bargaining game once the player was ready to be transferred. The experimenter was also able to send attention checks to participants (e.g., if several minutes elapsed without any visible progress) and remove them from the experiment (e.g., if attention checks went unanswered).
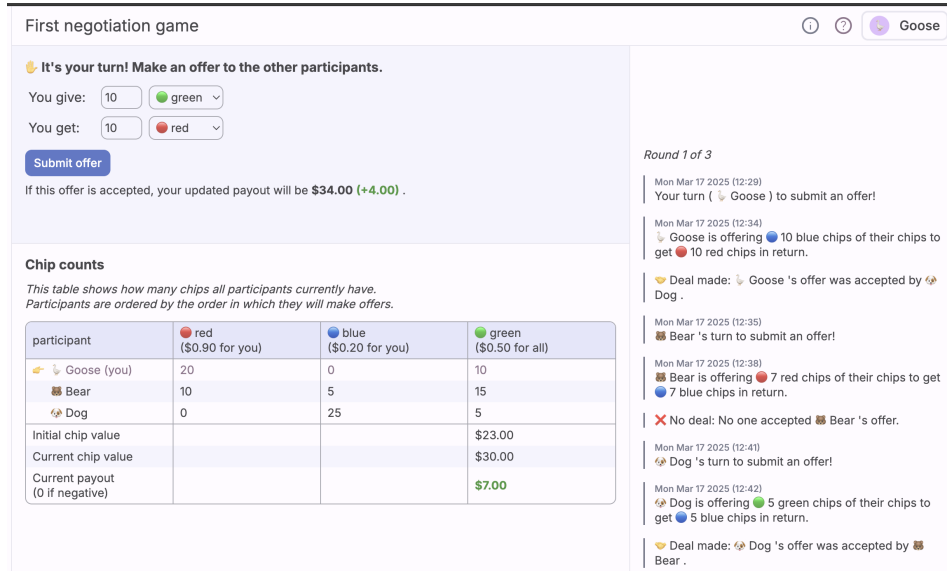
Figure 6: The game interface shown to participants. This shows a participant's game state. The top left indicates that they are currently the *Proposer*, offering an interface to propose an offer. A table below displays all players' chip volumes. To the right, the public ledger shows a history of game rounds, turns, and trade statuses.

For accepting or rejecting proposals, the interface calculated and displayed the projected total chip value if the player were to accept the proposal. It also disabled acceptance for offers the player could not fulfill (Fig 7, bottom). All participants were shown a table of all players' chip quantities and a log of actions (i.e., when trades were proposed, accepted, and rejected, and when each round and turn of the game changed). Turn order was determined randomly.
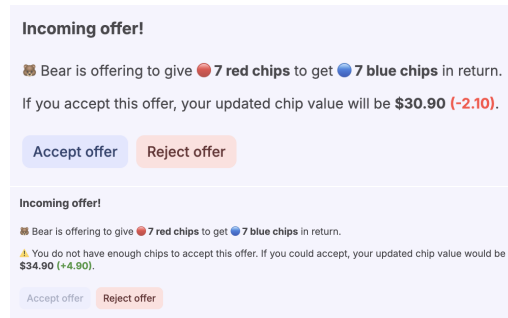

Figure 7: Accepting or rejecting a trade. The bottom image shows a disabled "accept" option.
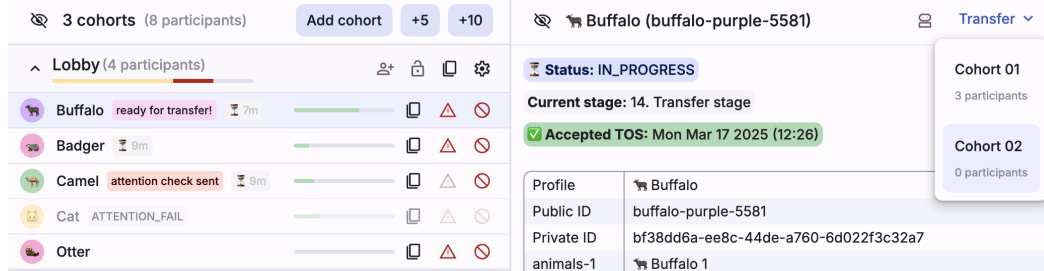

Figure 8: Experiment dashboard for managing participants. On the left, experimenters can view participants as they enter the experiment, send attention checks, and/or remove participants. On the right, experimenters can monitor the status of a selected participant and assign them to a cohort of three for gameplay.

C   POST-GAME PARTICIPANT SURVEY

C.1   SURVEY QUESTIONS

Table 1: Post-game survey questions.

	Survey question	Response format
1	Please describe your strategy in the game in a few sentences.	Freeform text
2	Please describe any experiences or background that may have influenced your performance in the game.	Freeform text
3	On a scale from 1 to 10, how would you rate your trading strategy, where 1 is highly **competitive** (focused mainly on your own gains) and 10 is highly **collaborative** (focused on other players's potential gains and mutual benefits?	1 through 10 scale; 1: "Not at all aggressive", 10: "Very aggressive"
4	On a scale from 1 to 10, how certain are you that you made the best possible trades during the experiment?	1 through 10 scale; 1: "Not at all confident", 10: "Very confident"
5	On a scale from 1 to 10, how satisfied are you with your final trading outcomes?	1 through 10 scale; 1: "Not at all satisfied", 10: "Very satisfied"
6	On a scale of 1 to 10, how would you rate the mental effort you put into today's trading games, where 1 means 'I barely engaged in any thinking' and 10 means 'I put in significant mental effort'?	1 through 10 scale; 1: "No mental exertion at all", 10: "Very high mental exertion"
7	Please provide any additional context on your answers above.	Freeform text
8	Please help us to improve this experiment. How was your experience today? Were there any elements of the instructions or gameplay that you found confusing?	Freeform text

C.2   QUANTITATIVE SURVEY RESPONSES

Table 2: Summary statistics of self-assessment variables (mean and standard error). There was no significant change in certainty or exertion as game complexity increased.

Game variation	Competitive	Certainty	Satisfaction	Exertion
2 chip	5.57 (0.34)	5.57 (0.33)	6.68 (0.31)	7.11 (0.24)
3 chip	5.32 (0.34)	5.92 (0.33)	7.07 (0.30)	7.10 (0.25)
4 chip	5.75 (0.34)	6.04 (0.33)	7.46 (0.26)	7.46 (0.19)

D   ADDITIONAL CLARIFICATION ON STRATEGIC REGRET CLASSIFICATIONS

Here, we utilized *counterfactual analysis* to visually dissect the three kinds of regret behavior previously discussed in Section 6.2.2, as illustrated in Fig 9. In these visualizations, the player's actual path of total chip value is depicted by a solid colored line, while the counterfactual path—representing the outcome had a different decision been made — is indicated by a gray dashed line.

Counterfactual definition: The counterfactual path is constructed by simulating an alternative decision at a critical juncture (e.g., accepting a declined offer, declining an accepted offer, or making a different proposal) and projecting the subsequent value trajectory. This allows for a direct comparison between the actual outcome and what could have been.

No regret scenarios.   This scenario is demonstrated by Player 2 (P2) in the role of a proposer, as shown in the upper-right panel of Fig 9, focusing on Turn 9. Player 2 makes a proposal that is accepted, leading to an actual chip value of 15.3. The counterfactual path, representing a scenario
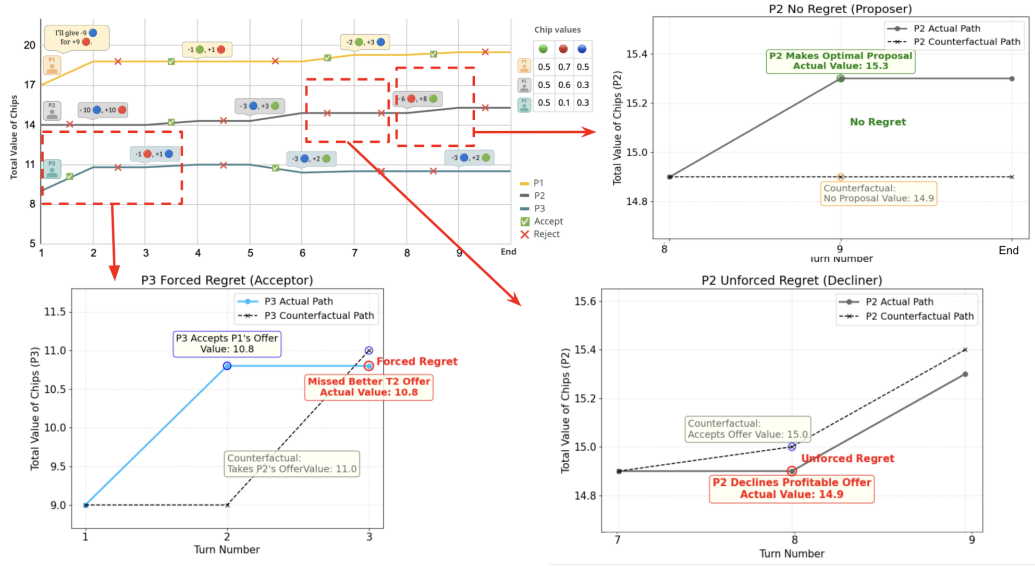
17

Figure 9: Regret behavior visualization: The top-left panel shows an overview trajectory in Fig 4. The top-right panel illustrates a "No Regret" scenario for P2 as a proposer. The bottom-left panel depicts "Forced Regret" for P3 as an acceptor. The bottom-right panel shows "Unforced Regret" for P2 as a decliner at Turn 8. Solid lines represent actual player paths, while dashed lines indicate counterfactual paths.

where P2 might not have made this proposal or made a less optimal one, results in a lower value of 14.9. In this "No Regret" instance, the player's actual total value path is superior to or equal to the evaluated counterfactual path, indicating that the decision made was optimal given the available information and subsequent events.

Forced regret scenarios. The lower-left panel of Fig 9 illustrates "Forced Regret" experienced by Player 3 (P3) as an acceptor from Turn 1 to Turn 2. P3 initially accepts an offer from P1 at Turn 1, resulting in an actual value of 10.8. However, a more lucrative offer from P2 becomes available in Turn 2. Due to the commitment made in Turn 1 (e.g., changed chip inventory), P3 is unable to capitalize on this subsequent, better opportunity. The counterfactual path shows that had P3 declined P1's initial offer, they could have potentially accepted P2's offer, leading to a higher value of 11.0.

Unforced regret scenarios. The lower-right panel of 9 demonstrates Player 2 (P2) in the role of a decliner from Turn 8 to 9. P2 is presented with a profitable trade offer but chooses to decline it. P2's actual chip value remains at 14.9 after this decision. The counterfactual path, however, demonstrates that had P2 accepted this profitable offer, their chip value would have increased to 15.0. This highlights an avoidable error: a superior option was available and actionable, but the agent failed to select it, leading to a sub-optimal outcome. Unlike forced regret, no prior decision prevented P2 from taking the better option; the regret stems directly from the choice made at that specific juncture.

E   SURPLUS VALUES AND SIGNIFICANCE

Table 3: Scaled final surplus gain (mean and standard error). This is the ratio of observed total surplus to optimal surplus.

Game variation	Human	GPT-4o	GPT-4o refined	Gemini	Gemini refined	Bayesian agent
2 chip	0.60 (0.06)	0.69 (0.03)	0.68 (0.04)	0.42 (0.04)	0.45 (0.05)	0.74 (0.04)
3 chip	0.59 (0.04)	0.62 (0.03)	0.64 (0.03)	0.42 (0.03)	0.43 (0.03)	0.80 (0.03)
4 chip	0.54 (0.03)	0.54 (0.02)	0.58 (0.02)	0.37 (0.02)	0.33 (0.02)	0.73 (0.02)

Table 4: Two-sided t-test p-values comparing surplus values across populations (N=144 per game). Arrows indicate whether the column's mean surplus was higher (↑) or lower (↓) than the row's.

	Human	GPT-4o	GPT-4o refined	Gemini	Gemini refined	Bayesian agent
2 chip						
Human	–	0.491 ↑	0.597 ↑	0.263 ↓	0.066 ↓	0.654 ↑
GPT-4o	–	–	0.302 ↓	0.000 ↓	0.000 ↓	0.141 ↑
GPT-4o refined	–	–	–	0.000 ↓	0.000 ↓	0.414 ↑
Gemini	–	–	–	–	0.359 ↑	0.000 ↑
Gemini refined	–	–	–	–	–	0.000 ↑
Bayesian agent	–	–	–	–	–	–
3 chip						
Human	–	0.525 ↑	0.021 ↑	0.001 ↓	0.000 ↓	0.000 ↑
GPT-4o	–	–	0.164 ↑	0.000 ↓	0.004 ↓	0.000 ↑
GPT-4o refined	–	–	–	0.000 ↓	0.000 ↓	0.001 ↑
Gemini	–	–	–	–	0.799 ↑	0.000 ↑
Gemini refined	–	–	–	–	–	0.000 ↑
Bayesian agent	–	–	–	–	–	–
4 chip						
Human	–	0.601	0.258 ↑	0.000 ↓	0.000 ↓	0.000 ↑
GPT-4o	–	–	0.270 ↑	0.000 ↓	0.000 ↓	0.000 ↑
GPT-4o refined	–	–	–	0.000 ↓	0.000 ↓	0.000 ↑
Gemini	–	–	–	–	0.003 ↓	0.000 ↑
Gemini refined	–	–	–	–	–	0.000 ↑
Bayesian agent	–	–	–	–	–	–

F ASIDE ON SMALLER MODEL PERFORMANCE

We tested smaller models from both the Gemini and GPT families— `Gemini-2.5-Flash` and `GPT-o4-mini`—and found that their overall performance falls short of both human participants and larger models such as GPT-4o, particularly in terms of surplus generation. Even with prompt refinement, the average surplus gain remains substantially lower.

Procedurally, these small models also behave quite differently from their larger counterparts. As shown in Figure 10, proposals by Gemini-2.5-Flash and GPT-4-mini are heavily concentrated around a narrow region: nearly zero net surplus gain and a 1:1 trade ratio. Roughly 90% of offers consist of exchanging one chip for exactly one of another type (e.g., 1 red for 1 green). Loss-incurring proposals are virtually nonexistent—but so are high-surplus trades (e.g., net surplus > 2). Prompt refinement slightly broadens the distribution but fails to shift the models away from this dominant 1:1 pattern.

This low-variance, risk-averse behavior is characteristic of the "lazy generation" effect reported in small language models (Lambert and Calandra, 2023), and reflects the narrow, template-driven action spaces observed in recent bargaining benchmarks for sub-10B parameter agents (Xia et al., 2024).
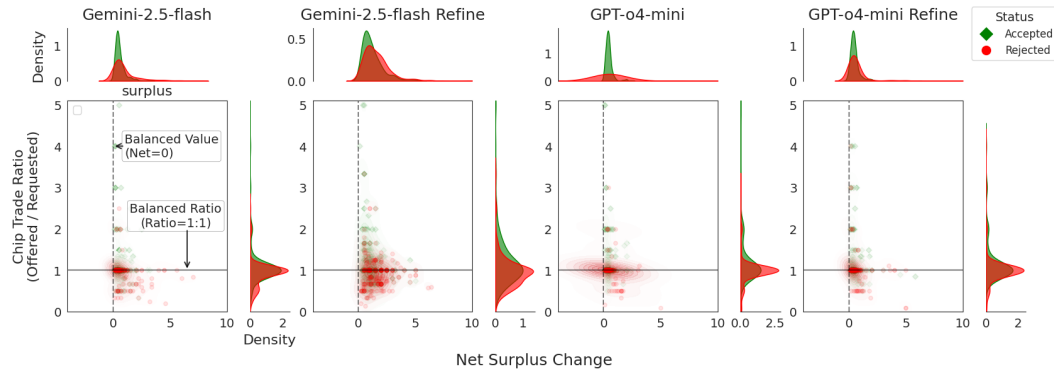

Figure 10: Comparison of the trade space (3-Chip) between small LLM agents (GPT-o4-mini and Gemini-2.5-flash).
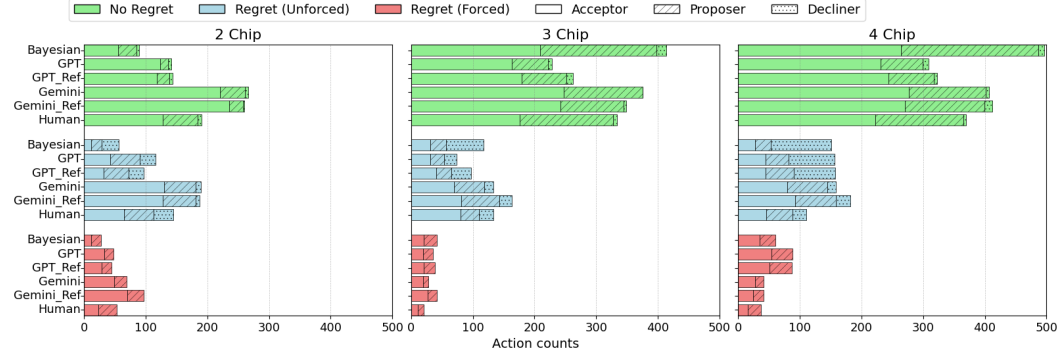
F.1 FULL STRATEGIC ANALYSIS VISUALIZATION

Figure 11: This visualization follows the same structure as Figure 5, with additional Refined LLMs. Strategic differences between refined and out-of-the-box LLMs are empirically similar.

F.2 SUPPLEMENTARY TRADE SPACE VISUALIZATIONS

Summary statistics of trade space values are in Table 5, with the 2-chip visualization in Figure 12 and 4-chip visualization in Figure 13. They exhibit similar behaviors as discussed in Section 6.2.1.
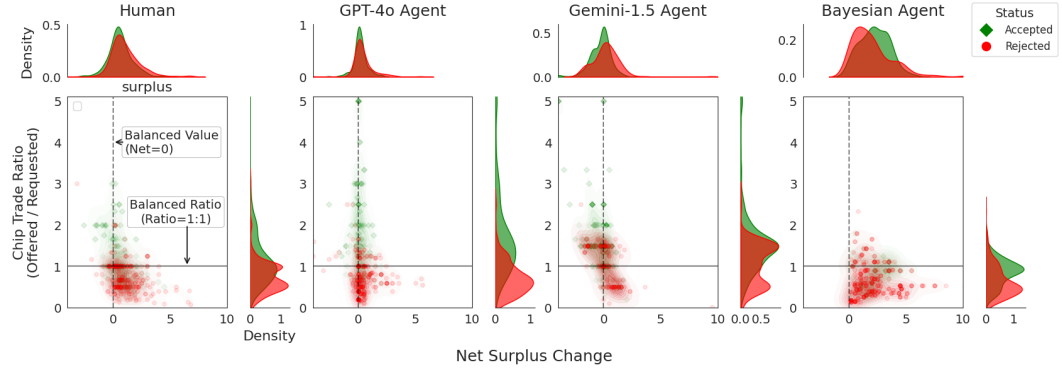

Figure 12: Comparison of the trade space (2-Chip) between Human participants, LLM agents (GPT-4o and Gemini-1.5 Pro), and Bayesian agents.
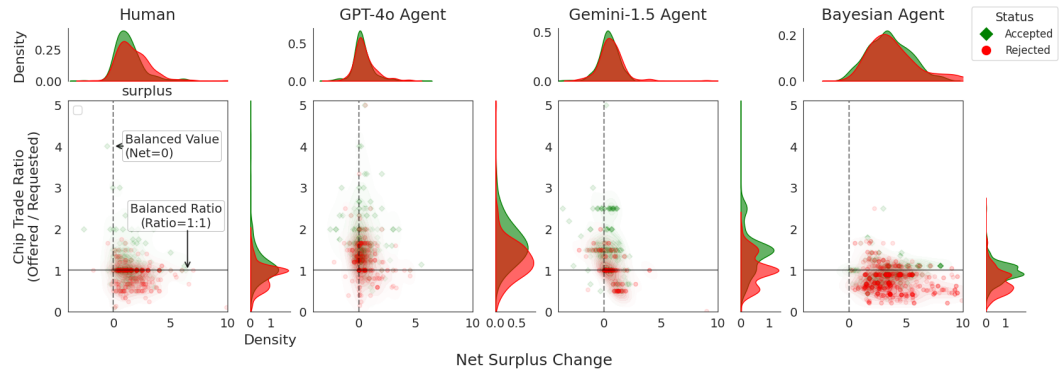

Figure 13: Comparison of the trade space (4-Chip) between Human participants, LLM agents (GPT-4o and Gemini-1.5 Pro), and Bayesian agents.

Table 5: Summary statistics from trade space visualizations.

Game	Player	Acceptance	Surplus Mean (SD)	Ratio Mean (SD)	Median [Surplus, Ratio]
2-chip	Human	Accepted	0.611 (1.031)	1.147 (0.652)	[0.500, 1.000]
		Rejected	1.088 (1.313)	0.735 (0.368)	[0.800, 0.667]
	GPT-4o	Accepted	0.148 (0.617)	1.711 (0.976)	[0.100, 1.500]
		Rejected	0.414 (1.031)	0.715 (0.393)	[0.100, 0.667]
	Gemini-1.5 Pro	Accepted	-0.305 (0.913)	1.508 (0.721)	[-0.100, 1.500]
		Rejected	0.230 (1.317)	1.048 (0.475)	[0.100, 1.000]
	Bayesian	Accepted	2.245 (1.252)	0.867 (0.315)	[2.300, 0.900]
		Rejected	2.058 (1.904)	0.636 (0.372)	[1.500, 0.526]
3-chip	Human	Accepted	1.211 (1.317)	1.039 (0.387)	[1.100, 1.000]
		Rejected	1.830 (1.803)	0.840 (0.325)	[1.300, 1.000]
	GPT-4o	Accepted	0.437 (0.838)	1.765 (0.945)	[0.200, 1.600]
		Rejected	0.345 (0.963)	0.918 (0.501)	[0.200, 0.800]
	Gemini-1.5 Pro	Accepted	0.329 (0.899)	1.394 (0.610)	[0.400, 1.500]
		Rejected	0.528 (1.819)	1.075 (0.497)	[0.600, 1.000]
	Bayesian	Accepted	3.173 (1.853)	0.944 (0.274)	[3.200, 0.900]
		Rejected	3.385 (2.725)	0.783 (0.355)	[2.800, 0.700]
4-chip	Human	Accepted	1.331 (1.214)	1.173 (0.562)	[1.200, 1.000]
		Rejected	1.757 (1.438)	0.863 (0.288)	[1.500, 1.000]
	GPT-4o	Accepted	0.369 (0.929)	1.745 (0.831)	[0.200, 1.600]
		Rejected	0.534 (0.981)	1.262 (0.518)	[0.300, 1.250]
	Gemini-1.5 Pro	Accepted	0.403 (0.858)	1.438 (0.540)	[0.400, 1.500]
		Rejected	0.650 (1.187)	1.022 (0.377)	[0.600, 1.000]
	Bayesian	Accepted	3.536 (1.793)	0.965 (0.251)	[3.400, 0.900]
		Rejected	4.251 (2.824)	0.738 (0.303)	[3.500, 0.700]

G LLM SIMULATION PROMPTS

G.1 RULE EXPLANATION

```
How the game works:
The game consists of 3 rounds of trading. During each round, each player will have a turn to
    propose 1 trade. These turns are pre-determined in a random order and the order stays
    the same in each round.

Trade proposals:
To propose a trade, a player must:
1. Request a certain quantity of chips of a single color from any other player to get.
2. Specify a certain quantity of chips of a different color to give in return.

Trade rules:
Players cannot offer more chips than they currently hold. For example, if you only have 5 red
    chips, you cannot offer 6 red chips.
Players cannot trade chips of the same color. For example, you cannot trade red chips for red
    chips.

Trade completion:
When an offer is presented, all other active participants get a chance to accept or decline.
    Note: Active participants are those not currently making the offer.

Participants make their decisions simultaneously and privately. The participant who receives
    the offer is not dependent on who accepts the trade first. Some possible outcomes:
If no one accepts, the trade does not happen, and the turn ends.
If multiple participants accept, one accepting participant is chosen at random to complete
    the trade with the offering participant. This means that participants cannot choose who
    they trade with.
If only one participant accepts, the trade will happen.

Key points to remember:
In each round, each player gets to propose one trade and respond to other player's trades
You can only propose trades between different colored chips, and cannot offer to give a chip
    amount that you do not have
When multiple players accept a trade, the trading partner is randomly selected
"""
```

G.2 PROPOSER'S PROMPT

We used a Chain-of-Thought approach to elicit LLM agent's trade proposals:

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational - do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips
    you receive (quantity x your valuation) minus the total value of chips you give up
    (quantity x your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your trade offer, and your reasoning. The text between tags should be
    concise.
```
<REASONING>
[Provide your concise reasoning in a few sentences, e.g. To gain more surplus, I want more
    xxx chips]
</REASONING>

[check if you have sufficient chips to trade. If you have n green chips, you can at most give
    n green chips. If you don't want to trade, you can ask for a large amount of chips that
    no one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
```
```

G.3 RECEIVER'S PROMPT

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

You have an offer. {{proposer}} is offering to give {{give}} and get {{get}} in return.
If you make this trade, your total wealth will change by: {{delta_surplus}}

Now, you need to decide whether to accept or decline.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your choice to accept or decline, and your reasoning. The text between
    tags should be concise.
```

<CHOICE>Yes or No </CHOICE>
```
```

G.4 REFINED PROPOSING PROMPT

The refined prompt first generates multiple trades proposals and uses another LLM query to choose
the best one among them.

G.4.1 MULTIPLE PROPOSAL GENERATION PROMPT

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.
```

```
REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational – do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips
    you receive (quantity $\times$ your valuation) minus the total value of chips you give
    up (quantity $\times$ your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
You can trade as many chips as you want in a single turn, assuming you have that many. Do not
    feel constrained to only trade a single chip at a time.
Propose 3 different good trade ideas, so a next step can decide on the best trade of the ones
    you propose here.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your trade offer, and your reasoning. The text between tags should be
    concise.
Repeat the below tags once for each of the trade ideas you propose.
```

[check if you have sufficient chips to trade. If you have n green chips, you can at most give
    n green chips. If you don't want to trade, you can ask for a large amount of chips that
    no one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
```
```

G.4.2 CHOOSING FROM MULTIPLE POSSIBLE PROPOSALS PROMPT

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational – do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips
    you receive (quantity $\times$ your valuation) minus the total value of chips you give
    up (quantity $\times$ your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
You can trade as many chips as you want in a single turn, assuming you have that many. Do not
    feel constrained to only trade a single chip at a time.

Below are three trade ideas you have proposed. Please pick the best trade proposal from the
    ones below that you will propose to the group.
Do not change anything about the trade you have selected from the ideas below.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags and content of your selected trade offer including its reasoning.
```

[check if you have sufficient chips to trade. If you have n green chips, you can at most give
    n green chips. If you don't want to trade, you can ask for a large amount of chips that
    no one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
```

```
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
```

Proposed trade ideas to choose from:
{{proposed}}

H REPRODUCIBILITY STATEMENT

We've provided screenshots of the interface for human experiments and cited all used tools (Appendix B.1), and will release this game for public usage upon acceptance. The algorithm for the Bayesian agent is provided in Appendix A.3, with code provided as supplemental material. The prompts for the LLM agents are provided in Appendix G, with model specifications in Section 5.2.

For reproducibility, the Bayesian agent algorithms are detailed in Appendix A.3, and full LLM prompts are provided in Appendix G with supplemental code. The platform used for human experiments with the bargaining game implementation is open-sourced and provided in Appendix B.1.

I ETHICS STATEMENT

All human participant research adhered to IRB-approved protocols, with informed consent provided and fair compensation above regional minimum wage (Section 5.1) of over $24 USD per hour. We used no personally identifiable data and anonymized participant references with a cohort hash following data collection. The work poses minimal foreseeable risk of harm or misuse. Potential safety implications of the work are discussed in Section 7.

A detailed description of the custom interface is provided in Appendix B.1, with the full text of instructions for Deliberate Lab instructions provided as supplemental material. IRB approval and informed consent by all participants were received prior to the study. Compensation is directly addressed in Section 5.1.

REFERENCES

N. Agarwal, A. Moehring, P. Rajpurkar, and T. Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research, July 2023. URL http://www.nber.org/papers/w31422.

G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022.

L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*, 2022.

C. Arora, A. I. Sayeed, S. Licorish, F. Wang, and C. Treude. Optimizing large language model hyperparameters for code generation, 2024. URL https://arxiv.org/abs/2408.10577.

A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

A.-M. Corvin. Web summit 2024: Alibaba conjures up global trade ai, portugal launches own llm. *TechInformed*, November 2024. URL https://www.techinformed.com/web-summit-2024-alibaba-conjures-up-global-trade-ai-portugal-launches-own-llm/.

D. Danz, L. Vesterlund, and A. J. Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883, 2022.

K. Dehghanpour, M. H. Nehrir, J. W. Sheppard, and N. C. Kelly. Agent-based modeling in electrical energy markets using dynamic bayesian networks. *IEEE Transactions on Power Systems*, 31(6): 4744–4754, 2016. doi: 10.1109/TPWRS.2016.2524678.

S. Fish, Y. A. Gonczarowski, and R. I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7, 2024.

R. Fisher, W. L. Ury, and B. Patton. *Getting to yes: Negotiating agreement without giving in*. Penguin, 2011.

S. Furniturewala, S. Jandial, A. Java, P. Banerjee, S. Shahid, S. Bhatia, and K. Jaidka. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.13. URL https://aclanthology.org/2024.emnlp-main.13/.

J. Gaarlandt, W. Korver, N. Furr, and A. Shipilov. Ai agents are changing how people shop. here's what that means for brands. *Harvard Business Review*, 2025.

D. Goktas, A. Greenwald, T. Osogami, R. Patel, K. Leyton-Brown, G. Schoenebeck, D. Cornelisse, C. Daskalakis, I. Gemp, J. Horton, D. C. Parkes, D. M. Pennock, A. Prakash, S. S. Ravindranath, M. O. Smith, G. Swamy, E. Vinitsky, S. Wasserkrug, M. Wellman, J. Wu, H. Xu, J. Zhang, Y. Zhang, S. Zhao, and Q. Zhu. Strategic Foundation Models. working paper or preprint, Feb. 2025. URL https://hal.science/hal-04925309.

J. Horton, A. Filippas, and R. Horton. Edsl: Expected parrot domain specific language for ai powered social science. Whitepaper, Expected Parrot, 2024.

J. J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

E. Jahani, B. S. Manning, J. Zhang, H.-Y. TuYe, M. Alsobay, C. Nicolaides, S. Suri, and D. Holtz. Prompt adaptation as a dynamic complement in generative ai systems, 2025. URL https://arxiv.org/abs/2407.14333.

S. Kapania, O. Siy, G. Clapper, A. M. SP, and N. Sambasivan. "because ai is 100% right and safe": User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517533. URL https://doi.org/10.1145/3491102.3517533.

J. Kramár, T. Eccles, I. Gemp, A. Tacchetti, K. R. McKee, M. Malinowski, T. Graepel, and Y. Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications*, 13(1):7214, 2022.

N. Lambert and R. Calandra. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*, 2023.

B. Lubars and C. Tan. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d67d8ab4f4c10bf22aa353e27879133c-Paper.pdf.

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who should predict? exact algorithms for learning to defer to humans, 2023. URL https://arxiv.org/abs/2301.06197.

S. Mullainathan and Z. Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care*. *The Quarterly Journal of Economics*, 137(2):679–727, 12 2021. ISSN 0033-5533. doi: 10.1093/qje/qjab046. URL https://doi.org/10.1093/qje/qjab046.

R. B. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(83)90048-0. URL https://www.sciencedirect.com/science/article/pii/0022053183900480.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

S. Palminteri, B. Garcia, and C. Qian. The moral turing test: Evaluating human-llm alignment in moral decision-making, Oct 2024. URL osf.io/preprints/psyarxiv/ct6rx_v1.

J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

PYMNTS. Openai turns chatgpt into shopping assistant with new commerce features. *PYMNTS*, April 2025. URL https://www.pymnts.com/news/artificial-intelligence/2025/openai-turns-chatgpt-into-shopping-assistant-with-new-commerce-features/.

C. Qian and J. Wexler. Take it, leave it, or fix it: Measuring productivity and trust in human-ai collaboration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 370–384, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705083. doi: 10.1145/3640543.3645198. URL https://doi.org/10.1145/3640543.3645198.

N. Raman, T. Lundy, S. Amouyal, Y. Levine, K. Leyton-Brown, and M. Tennenholtz. Steer: Assessing the economic rationality of large language models. *arXiv preprint arXiv:2402.09552*, 2024.

E. Soumalias, Y. Jiang, K. Zhu, M. Curry, S. Seuken, and D. C. Parkes. Llm-powered preference elicitation in combinatorial assignment. *arXiv preprint arXiv:2502.10308*, 2025.

Z. Sun and D. Müller. A framework for modeling payments for ecosystem services with agent-based models, bayesian belief networks and opinion dynamics models. *Environmental modelling & software*, 45:15–28, 2013.

G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024.

V. Tsai, C. Qian, and D. L. community contributors. Deliberate Lab: Open-Source Platform for LLM-Powered Social Science, Sept. 2024. URL `https://github.com/PAIR-code/deliberate-lab`.

M. Vaccaro, M. Caoson, H. Ju, S. Aral, and J. R. Curhan. Advancing ai negotiations: New theory and evidence from a large-scale autonomous negotiations competition, 2025. URL `https://arxiv.org/abs/2503.06416`.

R. Van Hoek, M. DeWitt, M. Lacity, and T. Johnson. How walmart automated supplier negotiations. *Harvard Business Review*, 8(11):2022, 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

T. Xia, Z. He, T. Ren, Y. Miao, Z. Zhang, Y. Yang, and R. Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3579–3602, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-acl.213. URL `https://aclanthology.org/2024.findings-acl.213/`.

C. Xie, C. Chen, F. Jia, Z. Ye, S. Lai, K. Shu, J. Gu, A. Bibi, Z. Hu, D. Jurgens, J. Evans, P. Torr, B. Ghanem, and G. Li. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=CeOwahuQic`.

K. Zhu, J. J. Horton, Y. Jiang, D. C. Parkes, and A. V. Shah. Evidence from the synthetic laboratory: Language models as auction participants. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024. URL `https://openreview.net/forum?id=FB9mTtJpJI`.