Scaling Laws for Strategic Interactions

Joie Zhang JOIE@ALUMNI.PRINCETON.EDU
Princeton University

Danqi Chen DANQIC@CS.PRINCETON.EDU
Princeton University

Peter Henderson PETER.HENDERSON@PRINCETON.EDU
Princeton University

Lewis Hammond LEWIS.HAMMOND@COOPERATIVEAI.ORG
Cooperative AI Foundation

Abstract

LLM agents are becoming autonomous and ubiquitous, moving us toward a world where agent-agent interactions carry real consequences for the people and institutions they represent. These interactions will be strategic, mixed-motive, and structurally diverse, involving agents that differ in capability and number. We introduce three new multi-turn negotiation games grounded in representative real-world scenarios, designed such that we can run scaling law-style sweeps that systematically vary agent capability, number of agents, and degree of competition. Across 5712 runs and 30 models, we find that more capable agents (by LMArena Elo) reliably achieve higher payoff. When the game structure is fundamentally cooperative, this raises the payoff of other agents jointly, but in competitive game settings, these gains come at the expense of weaker agents. As we scale to settings with more agents, raising the capability of an entire group lifts all payoffs and lowers Gini inequality; however, raising the capability of only a single agent widens the gap between itself and the rest. Through qualitative analysis, we find that emotional persuasion tactics and ability to identify trade and compromise opportunities are the behaviours most strongly correlated with higher payoff, more so than logical persuasion and stubborn self-interest.

Keywords: scaling laws; strategic AI; LLM agents; negotiation benchmarks; multi-agent systems

1. Introduction

Suppose a small business owner deploys a free, consumer-grade LLM agent to procure cloud services from a major vendor whose AI agent runs on a frontier model with proprietary tool access. Both agents are cooperating to reach a deal that beats no-deal for both sides, but they are also competing to maximize their own principal's payoff. The vendor's agent is better at every cognitive sub-task within the negotiation, from identifying strategic advantages to employing persuasive language. Does this mean that a better deal for both parties can be found more efficiently, or does the stronger model manipulate the consumer's agent into accepting an unfair split of the surplus? Moreover, how does this change if the agents' objectives are more or less aligned, or if multiple consumer agents band together to seek a more favourable agreement?

Such asymmetric, mixed-motive interactions are no longer hypothetical. We already see agents strategically interacting not only with and on behalf of humans [2, 8, 22, 24] but with other agents, be

Code: https://github.com/joie-zhang/bargain.

© J. Zhang, D. Chen, P. Henderson & L. Hammond.

it on online marketplaces [19, 25], or as part of AI-based monitoring and oversight regimes [3]. This new landscape of agent interactions has two important features: (i) interactions are *strategic*, in the sense that agents possess different objectives due to being deployed by different actors or for different purposes; and (ii) interactions are between agents of *differing quality and quantity* because of the costs of training or deploying more capable models, or because of access to proprietary data or tools. These differences raise practical questions that loom large not only for individual actors deploying such agents in the economy and throughout society [7, 11, 23], but also in narrower applications critically important to AI safety, such as using weaker LLM agents to monitor stronger ones in scalable oversight regimes where the stronger agent may not be fully aligned [1, 6, 16].

In this paper, we answer these questions by searching for empirical *scaling laws for strategic interactions*. Unlike past works that have attempted to fit parametric power laws in training compute, we study the changes in individual and collective welfare when scaling agent *quality*, both in general (as measured by their LMArena Elo [5]) and more specifically when it comes to inference-time scaling (as measured by the number of reasoning tokens), or when scaling agent *quantity*. We also investigate how these trends shift as the degree of cooperation and competition of the underlying game structure varies. To do so, we develop three new multi-turn negotiation games that isolate different real-world strategic mechanisms: item allocation over rivalrous goods, diplomatic negotiation over continuous treaty terms, and participatory budgeting over threshold public projects. These environments are interpretable enough to smoothly vary the degree of competition and the number of agents, yet complex enough to pose a meaningful challenge to frontier models.

We introduce the benchmark in Section 3, analyze bilateral play in Section 4.1, and extend to $N \in \{2, 4, 6, 8, 10\}$ agents in Section 4.2.

2. Related Work

Scaling Laws for Strategic Behavior. Prior scaling work fits power laws in compute or parameters for single-agent regimes: vision [26], LLM pre-training [13, 15], and single-agent RL [12]. Within strategic interaction, scaling-law work has largely focused on *multi-agent reinforcement learning* in Hex, Connect Four, and Pentago [14, 21], while natural-language bargaining between LLM agents remains less studied. The closest works to ours are: Engels et al. [9] model oversight success as the capability gap between a weak and a strong agent widens; Zhu et al. [28] show how weaker agents concede and lose surplus to stronger ones in buyer-seller markets; and Liu et al. [18] improve a single agent's play through inference-time opponent simulation and best-response search. In our work, we aim for a full picture of LLM negotiations along its main dimensions, jointly varying capability, degree of competition, and the number of agents, across three structurally distinct games, alongside test-time-compute and qualitative analyses of the behaviours that drive these outcomes.

3. Methodology

Each environment is a repeated negotiation played for at most T rounds by N agents over m objects. Agents receive private preferences at setup, and in each round publicly discuss, privately think, submit structured proposals, and privately vote; a proposal is accepted if it receives a $\lceil 2N/3 \rceil$-supermajority (both agents for $N = 2$), with ties broken by vote count then at random. If no proposal passes by round T, all agents receive zero utility. Utilities are discounted by γ^{t-1} for agreements reached in round t. The full phase protocol and flow diagram are in Appendix C.1.

Item Allocation. Agents bargain over m indivisible items $\mathcal{M}$. Agent i has private values $\boldsymbol{v}^i$ with $v^i_j \geq 0$ and $\sum_j v^i_j = 100$, and an allocation $\boldsymbol{A} \in \{0,1\}^{N\times m}$ assigns each item to exactly one agent, yielding $u^i(\boldsymbol{A}, t) = \gamma^{t-1}\boldsymbol{A}[i] \cdot \boldsymbol{v}^i$. Competition between agents i and j is the cosine similarity α_{ij} between their value vectors, and we summarize collective conflict at $N \geq 2$ by $\bar{\alpha} = \frac{2}{N(N-1)} \sum_{i<j} \alpha_{ij}$: higher $\bar{\alpha}$ makes items rivalrous, while orthogonal $\bar{\alpha}$ values encourage cooperation by permitting conflict-free assignment.

Diplomatic Treaty Negotiation. Agents negotiate a continuous treaty vector $\boldsymbol{a} \in [0,1]^m$, where agent i has ideal positions $\boldsymbol{p}^i$ and simplex-weighted issue importances $\boldsymbol{w}^i$, receiving $u^i(\boldsymbol{a}, t) = \gamma^{t-1}\, 100 \sum_k w^i_k (1 - |p^i_k - a_k|)$. Two competition controls govern preferences: the correlation ρ over ideal positions and the interest overlap θ over issue-importance vectors. Full details in Appendix C.3.

Participatory Budgeting. Agents co-fund threshold public projects under budget scarcity. Agent i has private budget b^i and valuation $\boldsymbol{v}^i$ with $\sum_j v^i_j = 100$; a contribution matrix $\boldsymbol{X} \in \mathbb{R}^{N\times m}_{\geq 0}$ satisfies $\sum_j x_{ij} \leq b^i$, funds project j iff $\sum_i x_{ij} \geq c_j$, and returns contributions to unfunded projects, yielding $u^i(\boldsymbol{X}, t) = \gamma^{t-1} \sum_{j \in S(\boldsymbol{X})} (v^i_j - x_{ij})$ for funded set $S(\boldsymbol{X})$. The two competition parameters are value alignment α_{ij} (cosine similarity between valuations; unlike Game 1, higher alignment aids co-funding) and budget scarcity $\sigma = \sum_i b^i / \sum_j c_j$. More details in Appendix C.4.

Metrics and Experimental Setup. Our capability coordinate is LMArena Elo (March 31, 2026 snapshot; 30-model roster spanning Elo 1110–1504). For welfare and fairness in Games 1–2, we compare outcomes to the Nash Bargaining Solution $o^{\text{NBS}} = \arg\max_o \prod_i u^i(o)$ [20], where each agent maximizes utility while the max-product enforces an equitable split; for Game 3's public goods, the Lindahl analogue [10, 17] has each agent pay in proportion to what it gains. We report each agent's fair-share residual $u^i_{\text{actual}} - u^i_{\text{benchmark}}$ and, for N-player runs, the Gini coefficient over utility. Full experiment details are in Appendix C.

4. Results

4.1. Two-Player ($N = 2$) Games

Capability Scaling. The bilateral sweep contains 1,941 runs against a fixed GPT-5-nano baseline (Section 3). Figure 1 (left) shows that more capable adversaries achieve higher average payoff across the Elo range in all three games. To rule out baseline-specificity, we replicate the sweep with Llama as the baseline in Appendix E.4; Figure 12 shows positive Elo slopes in all three games (slope estimates in Table 3).

We next examine how these gains affect the baseline. Figure 1 (right) shows a bifurcation between cooperative and competitive settings. Cooperatively, a stronger adversary lifts the baseline's payoff too, in all three games. Competitively, baseline payoff is non-monotonic: it rises with adversary Elo, then declines past the inflection near the baseline's own Elo of 1337 — consistent with the stronger adversary extracting payoff from the weaker baseline once it exceeds baseline capability.

How do bilateral outcomes compare to NBS and Lindahl fairness? We compare bilateral outcomes to cooperative-game-theoretic fairness benchmarks: NBS for Games 1–2 and Lindahl-style cost sharing for Game 3 (Section 3). Each run's fair-share residual is realized utility minus benchmark utility; positive means above the reference share. Figure 8(a) shows the residual rising with adversary

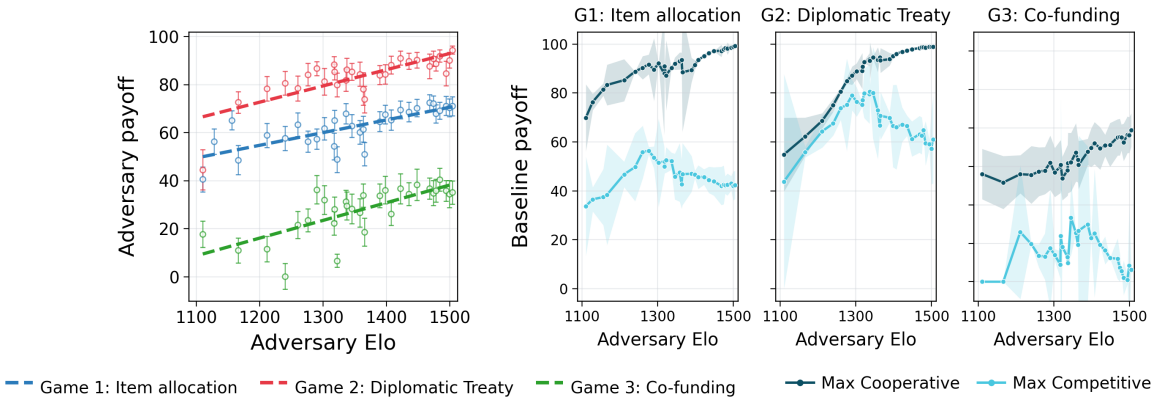

Figure 1: **Bilateral capability scaling against GPT-5-nano.** *(a)* Adversary payoff by Elo across the three games (model-level means with SEM; dashed lines are linear fits). *(b)* Baseline payoff at cooperative (dark teal) vs. competitive (cyan) extremes: cooperative rises monotonically; competitive peaks near Elo 1300–1350 then declines as the stronger adversary extracts.

Elo and crossing zero at Elo 1410–1461 depending on game — adversaries first close in on the benchmark, then extract above it.

How does this affect the baseline? Appendix Figure 11 plots both agents' fair-share gaps at the cooperative and competitive endpoints, showing the same bifurcation. Cooperatively, stronger adversaries pull both toward their fair shares. Competitively, the outcome still moves toward the benchmark, but zero-sum: value shifts from baseline to adversary. Game 3 shows an initial nonlinearity because maximally competitive public-good instances often start at zero payoff for both parties, and adversary capability must clear a threshold before meaningful payoff appears.

What underlying mechanisms drive capability scaling? To understand why stronger agents earn more, we ran a qualitative analysis on all 1,941 bilateral transcripts. GPT-5.5-xhigh judges first proposed 50 candidate labels from a broad sample; we consolidated these into a 23-label codebook and re-annotated every transcript with independent GPT-5.5-xhigh judges. The 23 labels group into six categories: pressure, emotional persuasion, logical persuasion, trade/compromise, self-interest/exploitation, and formalization. Prompts and codebook are in Appendix F.

Figure 2 (left): the strongest positive correlate of payoff is identifying trades and compromises — high-performing agents earn more by finding mutually acceptable agreements rather than applying pressure. Emotional and logical persuasion are also positive (emotional the stronger); self-interest and exploitation are negatively correlated, indicating overtly adversarial tactics reduce realized gains. Figure 2 (right): pressure, trade/compromise, and logical persuasion rise most clearly with adversary capability. Among these, pressure is weakly negative ($\rho = -0.05$), while trade/compromise ($\rho = +0.17$) and logical persuasion ($\rho = +0.12$) are substantial positive correlates. Stronger agents thus improve primarily by finding beneficial trades and compromises, and secondarily via more effective logical persuasion.

Does scaling test-time compute lead to higher payoffs? A second axis for scaling adversary capability is native test-time compute — increasingly important as frontier systems shift toward reasoning models. We run 216 games on GPT-5, Claude Sonnet-4.6, and Gemini-3 Flash, holding the

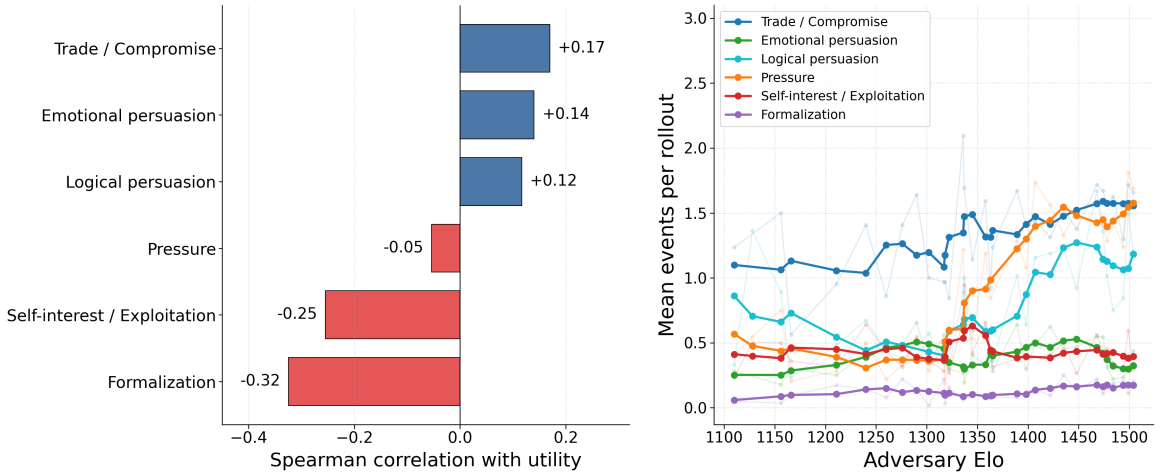

Figure 2: **Strategic-behavior analysis of bilateral play.** Left: Spearman ρ between per-category event count and adversary utility. Right: category event count vs. adversary Elo. The categories that scale with Elo (pressure, self-interest) are not those correlated with payoff (emotional persuasion, trade/compromise).

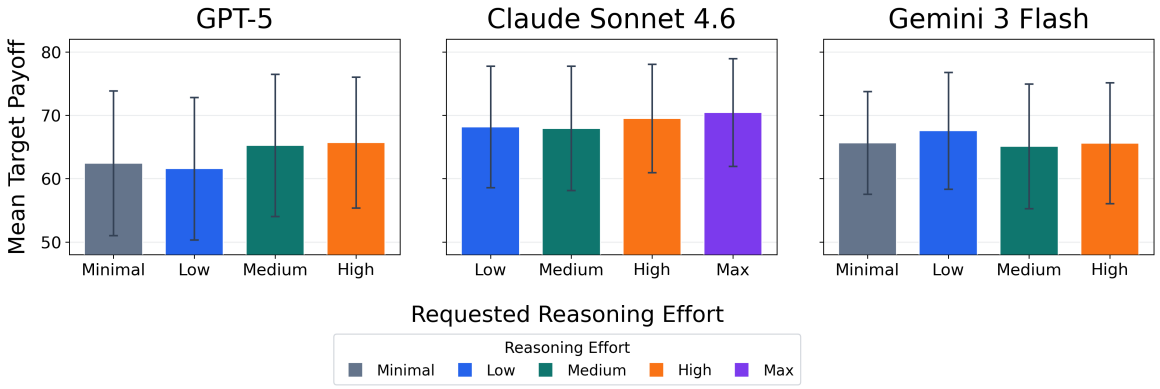

Figure 3: **TTC payoff does not reliably scale with requested reasoning effort.** Bars average both model orders and all nine game cells; error bars are SEM.

adversary model fixed and varying reasoning effort. Payoff does not reliably scale (Figure 3): GPT-5 climbs only modestly from 62.5 to 65.7 between minimal and high effort, Gemini is essentially flat (65.7 $\rightarrow$ 65.6) despite a sharp rise in observed tokens, and Claude's output-token proxy is non-monotone. Applying the 23-tag classification to the 216 TTC transcripts, the appendix behavior-intensity plot (Figure 16) shows how category frequencies shift with effort. For GPT-5, emotional and logical persuasion rise (positive correlates of payoff), but so does self-interested behavior. For Gemini, logical persuasion rises alongside formalization, pressure, and a drop in trade/compromise, which are all negatively associated with payoff. Test-time compute can thus intensify reasoning-model behaviors that do not translate into better negotiation outcomes.

4.2. N-Player Games

The multi-agent experiments span $N \in \{2, 4, 6, 8, 10\}$: 1,300 homogeneous, 1,300 heterogeneous, 130 homogeneous-control, and 325 homogeneous-control monoculture runs (Section 3).

Capability scaling in heterogeneous and homogeneous N-agent games. We first study capability scaling in the heterogeneous and homogeneous multi-agent settings. In heterogeneous games, we track each model's mean payoff over runs in which it was randomly sampled as one of the N agents; in homogeneous games, we vary a single adversary's capability while holding the $N - 1$ baseline agents fixed. Figure 23 shows that heterogeneous payoff rises with Arena Elo across all three games and all group sizes. Homogeneous-adversary experiments (Figure 24) show the same trend: adversary payoff rises with adversary Elo. Bilateral capability scaling thus generalizes to larger and more diverse negotiation groups.

Fairness in $N \geq 2$ multi-agent settings. We next examine outcomes relative to the NBS and Lindahl fairness benchmarks in the N-player setting. Figure 8 (right) plots the fair-share residual for the heterogeneous and homogeneous multi-agent runs; results mirror the $N = 2$ case. In both settings, increasing Elo moves utility toward the benchmark until a critical threshold, past which the focal model receives more than its benchmark share. In homogeneous-adversary runs, baseline payoff rises only mildly compared to the adversary's gains. The benchmark-relative patterns from bilateral play thus generalize to larger multi-agent settings.

Gini inequality in N-player multi-agent games. We ask whether heterogeneous groups of diverse-capability agents yield more unequal outcomes than homogeneous ones, comparing heterogeneous experiments against homogeneous-control monocultures from the same random model pool (Gini defined in Section 3). The appendix aggregate-Gini plot (Figure 28) shows no significant aggregate Gini difference, while a monoculture's own capability is the dominant determinant: scaling its Elo shifts within-run Gini below or above the heterogeneous reference.

Homogeneous-adversary experiments (varying only the inserted adversary) again show lower within-run Gini with stronger adversaries, but the appendix role-payoff plot (Figure 29) reveals a subtler dynamic: the reduction (left panel) largely comes from compressing the $N - 1$ baselines, while the right panel shows the adversary's own payoff rising and its gap over the baselines widening. Lower Gini thus does not mean a more equitable outcome for every agent. When all members' capabilities rise together, total payoff increases and inequality decreases; when only one does, inequality among the rest compresses while the stronger agent extracts more.

5. Conclusion

We study empirical scaling laws for LLM strategic interaction across three axes — model capability, native inference-time compute, and group size — in three controlled negotiation environments. The central lesson: capability alone does not determine whether a stronger agent *creates* value or *captures* it. Cooperative structure lifts both sides while competitive structure transfers value, native inference-time compute does not substitute for capability, and welfare depends on whether uplift is uniform or concentrated in one agent. As LLM agents are increasingly deployed as asymmetric-capability representatives in procurement, oversight, and multi-agent marketplaces, these distributional patterns matter as much as average performance, and we hope our benchmark supports future work beyond negotiation and into human-AI studies.

6. Camera-Ready Edits Summary

Reviewer ComG provided valuable feedback that our paper would benefit from deeper mechanistic explanations. Therefore, we deepened the qualitative analysis section of our paper by generating 50 labels describing different negotiation techniques across 6 different categories of strategies (emotional persuasion, trade/compromise, logical persuasion, pressure, self-interest/exploitation, and formalization). We then handpicked 23 of the 50 labels to minimize duplication and deployed GPT-5.5-xhigh agents to serve as LLM judges to annotate our 5,712 rollouts for every occurrence of each label during each model's conversation turns.

We ended up with 144,040 qualitative annotations (we manually verified a subset to ensure correctness) and then we analyzed these labels to draw correlations with model payoff and behaviors. We found that trade/compromise and emotional persuasion were positively correlated with payoff, while formalization and self-interest/exploitation were negatively correlated with payoff.

Additionally, we ran a new set of follow-up experiments (325 runs) to add more robustness to our current homogenous control experiments, which were previously only run on GPT-5-nano. These new experiments are what we described as the homogenous-control monoculture experiments.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv:1606.06565*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

[2] Anthropic. Project vend: Can claude run a small shop? (and why does that matter?). https://www.anthropic.com/research/project-vend-1, June 2025. Accessed: 2025-09-01.

[3] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[4] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069.

[5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[6] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv:1810.08575*, October 2018.

[7] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv:2012.08630*, December 2020.

[8] DoNotPay. The ai that fights for you. https://donotpay.com/about/, 2025. Accessed: 2025-09-01.

[9] Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark. Scaling laws for scalable oversight. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=u1j6RqH8nM. Spotlight.

[10] Duncan K. Foley. Lindahl's solution and the core of an economy with public goods. *Econometrica*, 38(1):66–72, 1970. doi: 10.2307/1909241. URL https://www.jstor.org/stable/1909241.

[11] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa

Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai. Technical Report 1, Cooperative AI Foundation, February 2025.

[12] Jacob Hilton, Jie Tang, and John Schulman. Scaling laws for single-agent reinforcement learning, 2023. URL http://arxiv.org/abs/2301.13442.

[13] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022. ISBN 978-1-7138-7108-8.

[14] Andy L. Jones. Scaling scaling laws with board games, 2021. URL http://arxiv.org/abs/2104.03113.

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL http://arxiv.org/abs/2001.08361.

[16] Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. On scalable oversight with weak llms judging strong llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 75229–75276. Curran Associates, Inc., 2024. URL https://openreview.net/forum?id=O1fp9nVraj.

[17] Erik Lindahl. Just taxation: A positive solution. In Richard A. Musgrave and Alan T. Peacock, editors, *Classics in the Theory of Public Finance*, pages 168–176. Macmillan, London, 1958. English translation of the 1919 original.

[18] Xiangyu Liu, Di Wang, Zhe Feng, and Aranyak Mehta. Scaling inference-time computation via opponent simulation: Enabling online strategic adaptation in repeated negotiation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.19309. To appear.

[19] Meta. Meet your business ai. https://www.facebook.com/business/ai/business-ai, September 2025. Accessed: 2025-09-01.

[20] John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: 10.2307/1907266. URL https://www.jstor.org/stable/1907266.

[21] Oren Neumann and Claudius Gros. Scaling Laws for a Multi-Agent Reinforcement Learning Model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZrEbzL9eQ3W.

[22] Pactum AI. Pactum ai: The leader in agentic ai for procurement for over half a decade. https://pactum.com/, 2025. Accessed: 2025-09-24.

[23] Matthew Sharp, Omer Bilgin, Iason Gabriel, and Lewis Hammond. Agentic inequality, 2025. URL https://arxiv.org/abs/2510.16853.

[24] Sierra. Meet sierra, the conversational ai platform for businesses. https://sierra.ai/blog/introducing-sierra, February 2024. Accessed: 2025-09-01.

[25] Kevin K. Troy, Dylan Shields, Keir Bradwell, and Peter McCrory. Project deal: Our Claude-run marketplace experiment. https://www.anthropic.com/features/project-deal, April 2026. Accessed: 2026-05-07.

[26] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022.

[27] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997. doi: 10.1145/279232.279236.

[28] Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 16–16, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.nllp-1.2. URL https://aclanthology.org/2025.nllp-1.2/.

Appendix A. Broader Impacts

This benchmark can help evaluate LLM agents before they are delegated economic, procurement, governance, or oversight responsibilities. By measuring welfare, fairness, agreement timing, and surplus capture together, the results can make it easier to identify settings where a more capable agent improves total surplus and settings where it mainly extracts surplus from weaker counterparts.

The same evidence may be dual use. A developer could use the benchmark to tune more exploitative bargaining agents, and strategic delegation could disadvantage users, firms, or communities with weaker or less expensive agents. In oversight settings, a capability gap between monitored and monitoring agents may also change both the efficiency and fairness of the interaction.

Responsible deployment should therefore evaluate welfare and fairness, monitor agent-agent negotiations after release, and avoid high-stakes autonomous bargaining without guardrails, audit logs, human escalation paths, and limits on the agent's authority to commit its principal. Benchmark results should not be interpreted as evidence that stronger agents are Pareto-safe across real-world strategic settings.

Appendix B. Declaration of LLM Usage

LLMs are the evaluated negotiation agents in this work, as described in Section 3. The authors also used general-purpose LLM assistants for editing, coding, and analysis support. The authors reviewed the experiment design, code changes, generated artifacts, analyses, and scientific claims.

Appendix C. Additional Experimental and Metric Details

C.1. Round Protocol

At initialization, each agent receives the game rules and its private preferences. Each subsequent round follows the same six-phase structure (Figure 4):

1. **Game setup:** one-time before round 1; agents receive rules and acknowledge private preferences.
2. **Discussion:** agents talk openly, revealing priorities, testing compromises, coordinating coalitions, or strategically withholding information.
3. **Private thinking:** each agent consolidates on a private scratchpad.
4. **Proposal:** each agent makes a formal proposal.
5. **Private voting:** each agent votes accept/reject on every proposal; a proposal passes with at least $\lceil 2N/3 \rceil$ approvals. If several pass, the top-vote proposal is selected; ties are broken uniformly at random. For $N = 2$, both agents must accept.
6. **Reflection:** if no proposal passes, agents privately reflect before the next round.

Utilities are discounted by γ^{t-1} for agreements at round t; if no agreement is reached by round T, all agents receive utility 0.

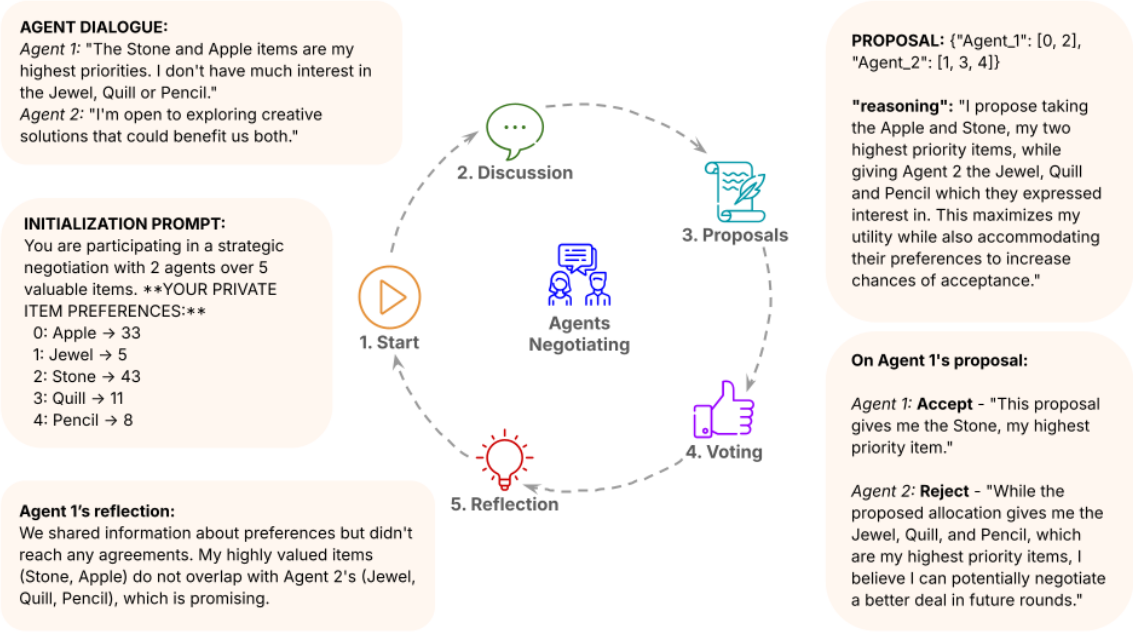

Figure 4: **Negotiation game dynamics.** Agents receive private preferences, discuss publicly, form private thoughts, submit structured proposals, and then vote privately. A proposal is implemented when it reaches the two-thirds supermajority threshold rounded up; for $N = 2$, this requires both agents to accept. If several proposals pass, the most-supported proposal is selected, with exact top-count ties broken randomly. Otherwise agents reflect and continue to the next round.

C.2. Full Experimental Inventory

Batch	Inventory
Bilateral GPT-5-nano	1,941 completed runs: 862 Game 1 runs over 31 adversary models, 540 Game 2 runs over 30 adversary models, and 539 Game 3 runs over 30 adversary models. All use $N = 2$, $T = 10$, $\gamma = 0.9$, two discussion turns, both model orders, and March 31, 2026 LMArena Elo.
Llama 3.3 70B baseline	500 completed runs: 140 Game 1, 180 Game 2, 180 Game 3. Ten adversary models span 1240–1504 Elo.
Multi-agent	2,730 completed runs: 1,430 homogeneous runs, 1,300 heterogeneous runs, and 16,380 agent-level observations. No missing or failed runs are imputed.
Random-monoculture control	325 completed runs: 125 Game 1, 100 Game 2, and 100 Game 3 configs. Each run clones one selected model into every seat, with 15 selected models sampled from the context-filtered heterogeneous pool across five Elo bands.
TTC stress test	216/216 completed jobs over GPT-5, Claude Sonnet 4.6, Gemini 3 Flash, four effort levels per family, nine game cells, two orders, and one seed.

Table 1: **Experiment inventory used in the revised paper.**

The bilateral GPT-5-nano sweep spans 30+ adversary models across the Elo range. Game 1 uses seven competition levels $\{0, 0.25, 0.5, 0.75, 0.9, 0.95, 1\}$. Game 2 uses nine cells sweeping

preference correlation $\rho \in \{-1, 0, 1\}$ and interest overlap $\theta \in \{0, 0.5, 1\}$. Game 3 uses nine cells sweeping value alignment $\alpha \in \{0, 0.5, 1\}$ and budget scarcity $\sigma \in \{0.2, 0.6, 1\}$. The Llama 3.3 70B replication (Appendix E.4) uses ten adversary models spanning the same Elo range.

The multi-agent generation grid uses $N \in \{2, 4, 6, 8, 10\}$. Game 1 sets $m = \lfloor 2.5N \rfloor$ items and competition levels $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. Game 2 sets 10 issues, $\theta \in \{0.2, 0.8\}$, and $\rho \in \{0.9, \rho_{\min}(N)\}$, with preference generation detailed in Appendix C.3. Game 3 sets $m = \lfloor 2.5N \rfloor$ projects, $\sigma \in \{0.2, 0.5\}$, and $\alpha \in \{0.2, 0.8\}$, with project costs, budgets, and valuations detailed in Appendix C.4. The original homogeneous controls use GPT-5-nano in every seat. Homogeneous-adversary runs use one focal adversary among GPT-5-nano agents, with focal position first or last and two seed replicates. Heterogeneous runs sample from a 25-model pool (filtered for all models to have a context length of at least 128K tokens) with equal-width Elo-standard-deviation strata. Random-monoculture controls mirror the same N and game-parameter grid but use N copies of one sampled model, with five selected monoculture models assigned to each game.

C.3. Game 2 Preference Generation

Game 2 gives each agent two private preference objects: an ideal-position vector over treaty issues and an issue-importance vector on the simplex. Ideal positions are generated with a Gaussian copula so that each issue position has a uniform marginal distribution on $[0, 1]$ while preserving a controlled cross-agent correlation. For each issue k, we sample

$$\boldsymbol{z}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_z), \qquad (\Sigma_z)_{ii} = 1, \quad (\Sigma_z)_{ij} = \rho_z \text{ for } i \neq j,$$

and set $p_k^i = \Phi(z_k^i)$, where Φ is the standard normal CDF. The latent Gaussian correlation is chosen as

$$\rho_z = 2\sin(\pi\rho/6),$$

which calibrates the transformed positions to the target Pearson correlation ρ. For $N \geq 2$, the equicorrelation matrix must be positive semidefinite, requiring $\rho_z \geq -1/(N-1)$. The most negatively correlated feasible target for $N \geq 2$ is therefore

$$\rho_{\min}(N) = \frac{6}{\pi} \arcsin\left(-\frac{1}{2(N-1)}\right).$$

Thus, in the multi-agent experiments, we use $\rho = 0.9$ for highly aligned ideal positions and $\rho = \rho_{\min}(N)$ for the most opposed feasible condition.

Issue-importance vectors are generated separately. For a target interest overlap θ, we jointly optimize all N nonnegative weight vectors on the simplex to match the target pairwise cosine similarity up to numerical tolerance.

C.4. Game 3 Project, Budget, and Valuation Generation

Game 3 instantiates a threshold public-goods problem. Each project j has an integer cost c_j sampled from the fixed range $[10, 30]$. The budget-scarcity parameter is

$$\sigma = \frac{\sum_i b^i}{\sum_j c_j}.$$

For a realized project-cost vector, total budget is set to approximately $\sigma \sum_j c_j$ and then divided symmetrically across agents after integer rounding, so each agent has the same fixed budget for the whole game.

Each agent's project-valuation vector is nonnegative and sums to 100. For a target preference-alignment level α, we jointly optimize the N valuation vectors so that their pairwise cosine similarities match α up to numerical tolerance, then integerize while preserving each agent's total value of 100. Higher α therefore means agents tend to value the same projects; unlike Game 1, this creates more scope for mutually beneficial co-funding because funded projects are public goods.

In each round, every agent submits a contribution vector over projects. The environment aggregates these submissions into a single joint proposal for that round. The joint proposal is implemented only if it receives the required supermajority. If it fails, no money is spent, previous contributions do not carry over, and the next round starts from the same fixed budgets. In an accepted proposal, project j is funded if and only if aggregate contributions meet its threshold, $\sum_i x_{ij} \geq c_j$; projects below threshold give no value and their proposed contributions are returned.

C.5. Model Rosters

Table 2 lists the 30 adversary models used in the bilateral GPT-5-nano baseline sweeps, ordered by descending LMArena Elo. The separate low-reasoning GPT-5-nano condition serves as the fixed baseline; `gpt-5-nano-high` is included here as an adversary. Game 1 includes one additional legacy model not shown, bringing its adversary count to 31; Games 2 and 3 each use 30 of these adversaries.

$N \geq 2$ **heterogeneous pool (24 models).** The heterogeneous multi-agent pool is the subset of the bilateral roster whose usable context window is at least 128K tokens. This removes five models: DeepSeek R1 (64K realized), Qwen2.5 72B (32.7K realized), Llama 3.1 8B (16.4K realized), Llama 3.2 3B (80K realized), and Llama 3.2 1B (60K realized). We also removed QwQ 32B for de-duplication purposes because its Elo of 1336 was really close to GPT-5-nano-high's Elo of 1337, and GPT-5-nano-high was already included as an adversary as well. The remaining 24 models span the Elo range of 1240–1504 across eight providers.

Llama 3.3 70B baseline adversaries (10 models). The Llama baseline replication uses the following 10 adversary models chosen to be as evenly distributed as possible along the Elo range of 1240–1504: Amazon Nova Micro (1240), Claude 3 Haiku (1260), Amazon Nova Pro (1290), GPT-4o mini (1317), DeepSeek V3 (1358), Claude Sonnet 4 (1389), DeepSeek R1-0528 (1422), Gemini 2.5 Pro (1448), GPT-5.4 High (1484), and Opus 4.6 Thinking (1504).

TTC target models. The TTC stress test uses three model families at four provider-native effort levels each, with GPT-5-nano (low reasoning) as the fixed baseline:

- **GPT-5** (`gpt-5-2025-08-07`): reasoning effort {minimal, low, medium, high}.
- **Claude Sonnet 4.6** (`claude-sonnet-4-6`): thinking budget {low, medium, high, max}.
- **Gemini 3 Flash** (`gemini-3-flash-preview`): thinking budget {minimal, low, medium, high}.

#	Model	Provider	Short name	Elo
1	claude-opus-4-6-thinking	Anthropic	Opus 4.6 Thinking	1504
2	claude-opus-4-6	Anthropic	Opus 4.6	1499
3	gemini-3.1-pro	Google	Gemini 3.1 Pro	1494
4	gpt-5.4-high	OpenAI	GPT-5.4 High	1484
5	gpt-5.2-chat-latest-20260210	OpenAI	GPT-5.2 Chat	1478
6	claude-opus-4-5-20251101-thinking-32k	Anthropic	Opus 4.5 Thinking	1474
7	claude-opus-4-5-20251101	Anthropic	Opus 4.5	1468
8	gemini-2.5-pro	Google	Gemini 2.5 Pro	1448
9	qwen3-max-preview	Alibaba	Qwen3 Max	1435
10	deepseek-r1-0528	DeepSeek	DeepSeek R1-0528	1422
11	claude-haiku-4-5-20251001	Anthropic	Haiku 4.5	1407
12	deepseek-r1	DeepSeek	DeepSeek R1	1398
13	claude-sonnet-4-20250514	Anthropic	Sonnet 4	1389
14	gemma-3-27b-it	Google	Gemma 3 27B	1365
15	o3-mini-high	OpenAI	o3-mini-high	1363
16	deepseek-v3	DeepSeek	DeepSeek V3	1358
17	gpt-4o-2024-05-13	OpenAI	GPT-4o	1345
18	gpt-5-nano-high	OpenAI	GPT-5 nano high	1337
19	qwq-32b	Alibaba	QwQ-32B	1336
20	gpt-4.1-nano-2025-04-14	OpenAI	GPT-4.1 nano	1322
21	llama-3.3-70b-instruct	Meta	Llama 3.3 70B	1318
22	gpt-4o-mini-2024-07-18	OpenAI	GPT-4o mini	1317
23	qwen2.5-72b-instruct	Alibaba	Qwen2.5 72B	1302
24	amazon-nova-pro-v1.0	Amazon	Nova Pro	1290
25	command-r-plus-08-2024	Cohere	Command R+	1276
26	claude-3-haiku-20240307	Anthropic	Claude 3 Haiku	1260
27	amazon-nova-micro-v1.0	Amazon	Nova Micro	1240
28	llama-3.1-8b-instruct	Meta	Llama 3.1 8B	1211
29	llama-3.2-3b-instruct	Meta	Llama 3.2 3B	1166
30	llama-3.2-1b-instruct	Meta	Llama 3.2 1B	1110

Table 2: **Bilateral adversary roster (30 models shown).** Elo scores are from the March 31, 2026 LMArena snapshot. Models are accessed via native provider APIs or OpenRouter.

C.6. Reproducibility, Release, and Compute Resources

The GitHub includes the benchmark code, prompts, generation configs, analysis scripts, run artifacts, and reproduction instructions for the reported tables and figures. It records the model rosters and provider identifiers, the March 31, 2026 LMArena Elo snapshot used as the capability proxy, the exact grids in Section 3 and Appendix C.2, and the available seeds and model orders. The released run artifacts include structured result JSON, interaction logs, derived CSVs, and plotting scripts; the benchmark environments, prompts, and generated transcripts are new assets documented with the release. Existing assets are credited through citations or provider/model identifiers, including LMArena Elo, LLM provider APIs and models, numerical optimization libraries used for benchmark computations, and referenced benchmark environments. The artifact records available licenses and terms of use; closed-provider model weights are not redistributed.

All reported LLM inference uses off-the-shelf hosted or routed API models; no local model training or fine-tuning is performed. Local orchestration, JSON validation, benchmark computation, and plotting are CPU-only Python workloads. The Slurm multi-agent and TTC workers used one node, one task, 4 CPU cores, 16GB memory, and an 8 hour wall-clock limit per config. Status logs for the final multi-agent runs record about 201 worker-hours for the 1,430 homogeneous jobs and about 389 worker-hours for the 1,300 heterogeneous jobs, with most jobs taking minutes and the longest successful jobs taking roughly 2–3 hours. The TTC Slurm batch contains 216 successful

jobs under the same 4-core/16GB worker envelope; the monitoring report records 2,631 target-model LLM calls and about 3.1M target compute/output tokens, not counting the fixed GPT-5-nano baseline calls. Bilateral and Llama-baseline batches use the same protocol, and exact API-call counts can be recomputed from the released interaction logs. Closed-provider inference hardware, provider-side batching, and hidden reasoning-token accounting are not directly disclosed to the authors.

C.7. Behavioral and Distributional Metrics

For each run, we report agreement status, final round, each agent's final utility, mean run utility, social welfare, and where applicable benchmark-relative exploitation. For multi-agent runs, we also compute utility dispersion:

- **Utility standard deviation and range:** direct dispersion measures over final utilities within the run.
- **Gini coefficient:** computed on final utilities after shifting a run by subtracting its minimum only when at least one utility is negative; all-zero utility vectors have Gini 0. For small N, the analysis exports both raw Gini and a small-N corrected Gini. Formally, for utility vector $\boldsymbol{u} = (u_1, \ldots, u_N)$ set $\tilde{u}_i = u_i - \min_k u_k$ if any $u_i < 0$, else $\tilde{u}_i = u_i$; then

$$G(\boldsymbol{u}) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |\tilde{u}_i - \tilde{u}_j|}{2N \sum_{i=1}^{N} \tilde{u}_i}, \tag{1}$$

  with $G = 0$ when the shifted vector is all zero. The corrected value for small N is $G_{\text{corr}} = \min\{NG/(N-1), 1\}$.
- **Payoff-vs-Elo slope:** per-run OLS fit of final utility on Elo, exported as signed and absolute slope per 100 Elo points.
- **Elo dispersion:** within-run Elo variance and standard deviation, used to test whether heterogeneous capability spread predicts inequality.

For Game 3, the protocol generates both structured contribution vectors and natural-language transcripts, supporting an additional public-good metric:

- **Utilitarian efficiency:** $\eta = SW_{\text{actual}}/SW_{\text{optimal}}$, where the optimum is the best feasible funded set for that cost and budget draw.

C.8. Implementation Reliability Notes

All agent outputs are validated against structured schemas before tabulation; runs that fail validation after bounded recovery are recorded as failures and excluded from analysis. For larger N, remaining failures are concentrated in proposal formatting and vote parsing, with additional provider-quota failures in some Slurm runs. In the TTC batch, GPT-5 reports reasoning tokens directly, while Claude and Gemini do not expose hidden reasoning-token counts; their token axes are visible-output proxies.

Appendix D. Benchmark and Normalization Details

D.1. Utility Normalization Across Games

All three games are designed so that raw utilities have a comparable nominal scale. The achievable maximum still differs by game and parameter setting. In Games 1 and 2, the simplex constraint on valuations or importance weights implies $u^i \in [0, 100]$ before discounting. In Game 3, an agent can in principle receive up to 100 value from funded projects before costs. Achievable welfare shrinks as budget scarcity increases. For cross-scarcity comparisons in Game 3, efficiency metrics normalize by the welfare-maximizing funded set for the realized budget and cost draw.

Raw utility values are reported directly in the multi-agent results. Negative Game 3 utilities are retained, so an agent can record a negative final utility when it pays more than its value from funded projects.

D.2. Computing Social Optima

Game 1. The utilitarian optimum assigns each item j to an agent with maximal valuation v_j^i. At the bilateral scale, enumerating all $2^5 = 32$ allocations is also exact; for general N, the direct enumeration size is N^m, while the utilitarian optimum remains itemwise.

Game 2. The social welfare function decomposes by issue:

$$SW(\boldsymbol{a}) = \sum_k \sum_i w_k^i (1 - |p_k^i - a_k|). \tag{2}$$

For each issue, a weighted median of agent ideal points maximizes the sum of weighted absolute-loss utilities.

Game 3. The socially optimal funded set S^* maximizes

$$\sum_{j \in S} \left(\sum_i v_j^i - c_j \right) \tag{3}$$

subject to $\sum_{j \in S} c_j \leq B$. This is a 0-1 knapsack problem over projects. At the bilateral $m = 5$ scale, all 32 subsets can be enumerated exactly. For larger N in the multi-agent grid, we solve the same integer knapsack by dynamic programming over realized project costs and total group budget.

D.3. Fairness Benchmarks

Games 1 and 2: Nash Bargaining Solution. The NBS maximizes the product of agents' gains over the disagreement point:

$$o^{\text{NBS}} = \arg\max_o \prod_{i=1}^{N} u^i(o). \tag{4}$$

For the bilateral fair-share residual plots, Game 1 enumerates allocations and Game 2 solves the continuous problem with L-BFGS-B [4, 27].

Game 3: Lindahl-style cost sharing. For public goods, the fairness benchmark is benefit-proportional cost sharing. For each funded project j,

$$x_{ij}^{\text{Lindahl}} = c_j \cdot \frac{v_j^i}{\sum_k v_j^k}. \tag{5}$$

The main-text *fair-share residual* for both benchmarks is $u^i_{\text{actual}} - u^i_{\text{benchmark}}$, reported directly in utility units; positive values indicate that the agent receives more than its benchmark share.

Appendix E. Additional Result Tables

E.1. Bilateral Payoff Slopes

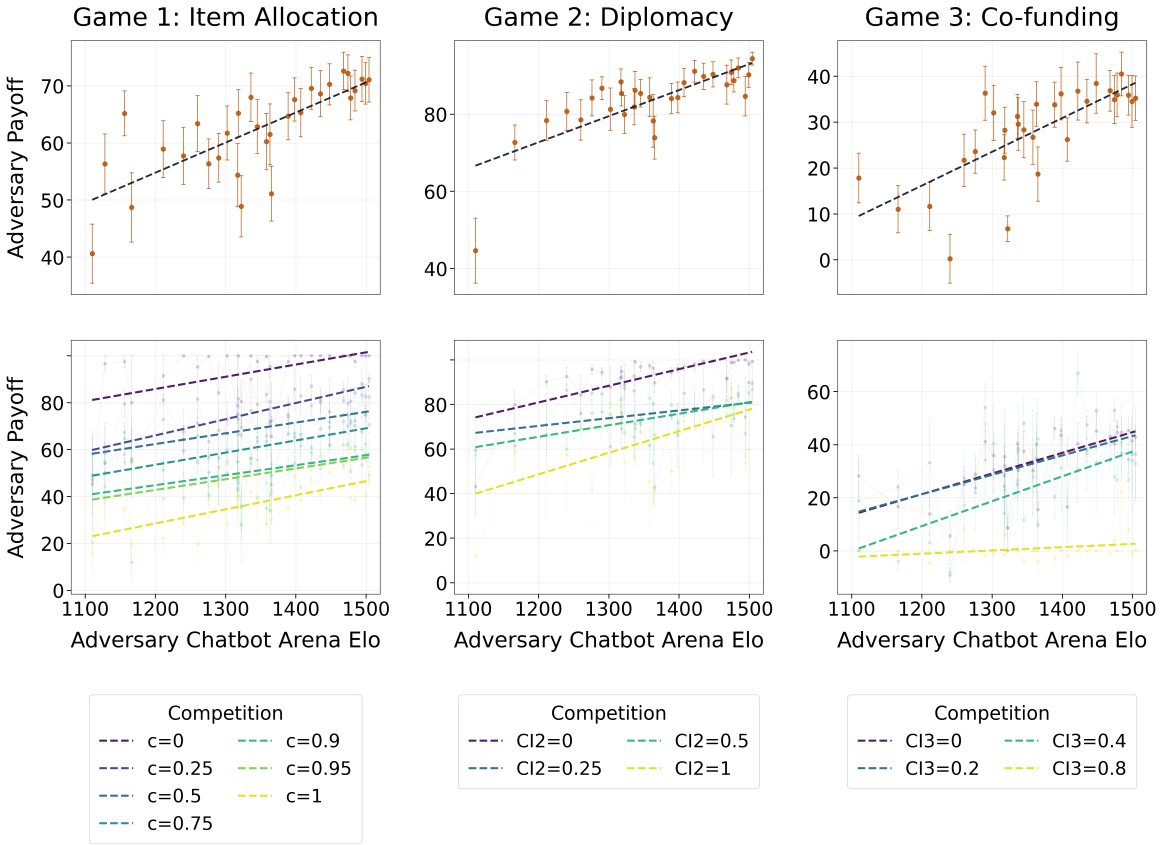

Figure 5: **Bilateral adversary payoff by game and competition.** Top row: overall adversary payoff against GPT-5-nano by Elo. Bottom row: the same relationship stratified by competition settings with per-stratum linear fits.

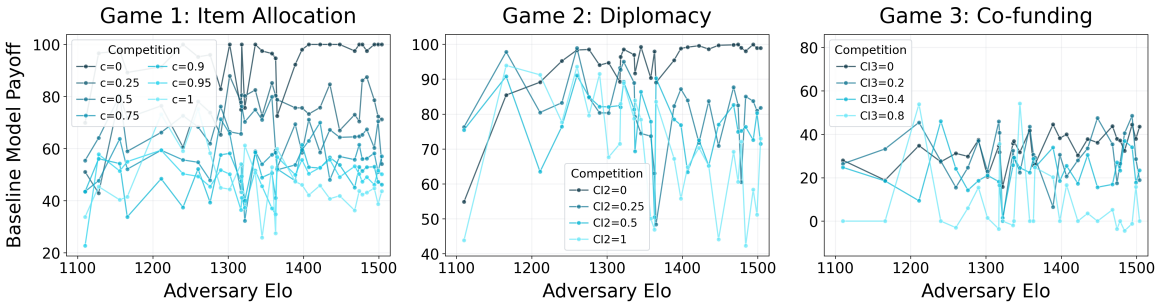

Figure 6: **Raw exact-competition version of the GPT-5-nano baseline payoff plot.** Each line shows raw per-Elo baseline-payoff means at an exact competition setting, without the competition-band averaging or moving-average smoothing used in the main-text Figure 1.

Game	GPT adv.	GPT base	Llama adv.	Llama base
Item allocation	+5.28	+1.14	+3.95	−0.65
Diplomatic treaty	+6.76	+1.66	+5.02	−0.24
Co-funding	+7.37	+2.42	+10.70	+2.19

Table 3: **Bilateral payoff slopes per 100 Elo.** "GPT" fixes GPT-5-nano as the baseline; "Llama" fixes Llama 3.3 70B as the baseline. The adversary payoff slope is positive in every game and under both baselines. Baseline payoff moves less and changes sign across baselines, which motivates the competition-stratified analysis.

	Game 1	Game 2	Game 3
Utility slope	+4.48***	+6.41***	+7.97***
NBS fair share slope	−0.51 (ns)	−0.03 (ns)	−0.33 (ns)
NBS residual slope	+4.98***	+6.44***	+8.30***
Zero-crossing Elo	1410	1461	1441

Table 4: **NBS decomposition slopes (per 100 Elo).** The fair-share slope is statistically indistinguishable from zero in all three games. The residual slope accounts for the full utility-Elo relationship. *** denotes $p < 0.001$; ns denotes $p > 0.05$.

E.2. Bilateral Social-Welfare Detail

Figure 7 reports realized total social welfare by competition stratum under the GPT-5-nano baseline. In cooperative Game 1 and Game 2 cells, agents allocate items to their highest valuers or set treaty dimensions near mutually preferred ideals. In the most competitive cells, the utilitarian optimum is easier to approximate even if the split is unequal. The middle cells are harder: preferences are neither identical nor orthogonal, leaving more room for inefficient compromise.

Game 3 is the important exception. When only one project can be funded and preferences are orthogonal, the setting becomes sharply distributional: if my project is funded and you pay, I win; if neither side accepts the burden, both receive zero. The highest Game 3 competition stratum has lower welfare capture than the analogous extremes in Games 1–2 because the public-good mechanism creates a harsher feasibility constraint than splitting items or choosing a continuous midpoint.

The highest-competition co-funding cell also explains why some bilateral points sit near zero despite a positive fitted slope. With total budget 22 and project costs $30, 16, 22, 30, 11$, even the two cheapest projects cannot both be funded. Agents have orthogonal preferences, so funding the other agent's project can leave an agent with zero value and a positive cost. Stronger models can move from slightly negative to slightly positive outcomes, but no model can create much surplus unless it finds a mutually acceptable one-project cost share.

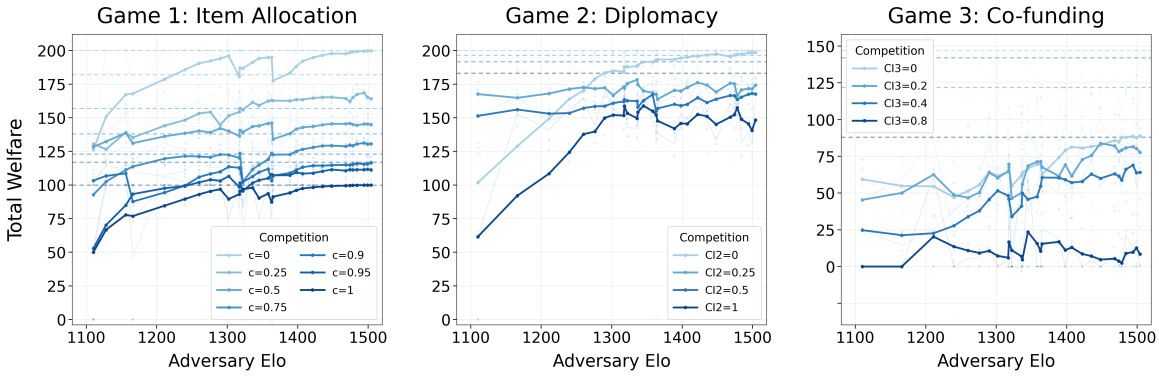

Figure 7: **Realized social welfare by competition stratum in bilateral play.** Solid marked curves are Elo-ordered moving averages of model-level means; dashed horizontal lines mark the maximum attainable welfare. Stronger adversaries move agreements toward higher-welfare outcomes; Game 3's highest-competition cell remains bottlenecked by one-project scarcity.

E.3. Bilateral Fairness Plots

Figure 8 shows the signed fair-share residual: whether the adversary is above or below its own benchmark share. The distance diagnostic below asks a different, unsigned question: how close the final outcome is to the benchmark point. These can move in opposite rhetorical directions. For example, an inefficient low-Elo outcome such as $(0, 0)$ can be very far from a benchmark split like $(50, 50)$, while a high-Elo outcome such as $(56, 48)$ is much closer to that benchmark point but still gives the adversary more than its fair share. Thus a declining benchmark distance should be read as improved benchmark proximity, not as evidence that unequal extraction disappears.

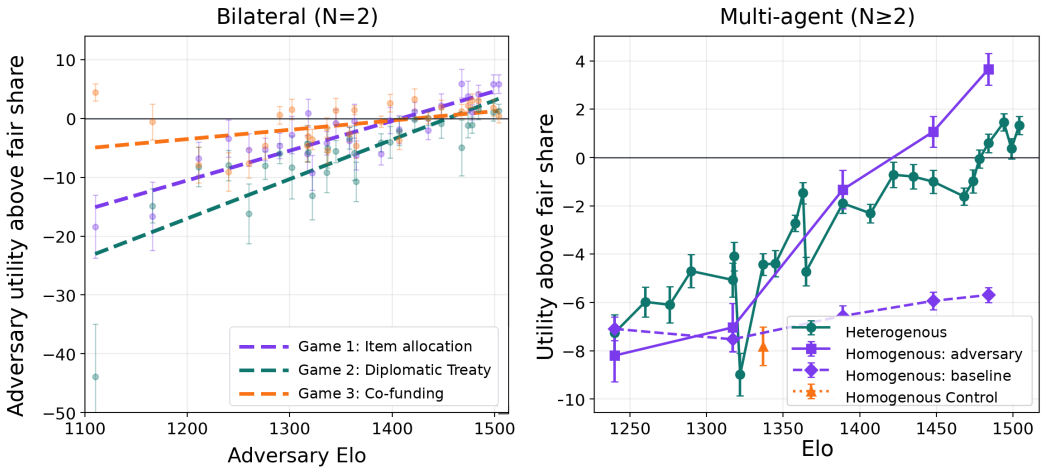

Figure 8: **Adversary fair-share residual (actual − NBS/Lindahl) rises with Elo at every group size.** *(a)* Bilateral $N = 2$ vs. GPT-5-nano; crosses zero at Elo 1410–1461 depending on game. *(b)* $N \geq 2$ residual vs. reference Elo, pooled across games and competition bands, for heterogeneous rosters, homogeneous-adversary runs, and the homogeneous GPT-5-nano control; every series crosses zero near Elo ~ 1450. Positive = above fair share.

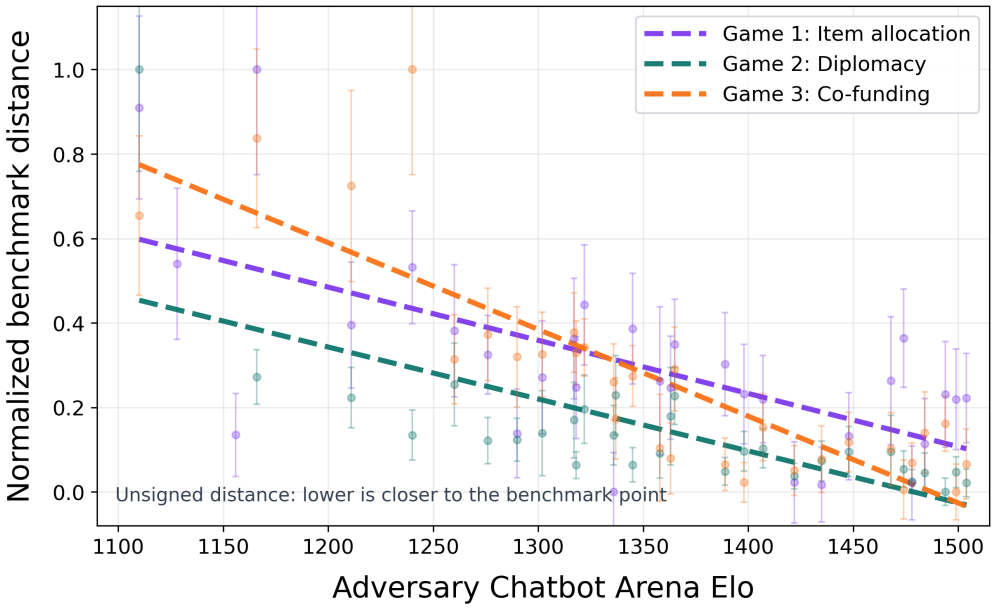

Figure 9: **Unsigned benchmark distance decreases with adversary Elo.** Points are model-level means with SEM error bars; distances are normalized within each game for display because Games 1–2 use NBS distance and Game 3 uses a Lindahl-style cost-sharing distance. This plot measures proximity to the benchmark point, not which player captures the surplus; Figure 8 shows that stronger adversaries increasingly sit above their own benchmark share.

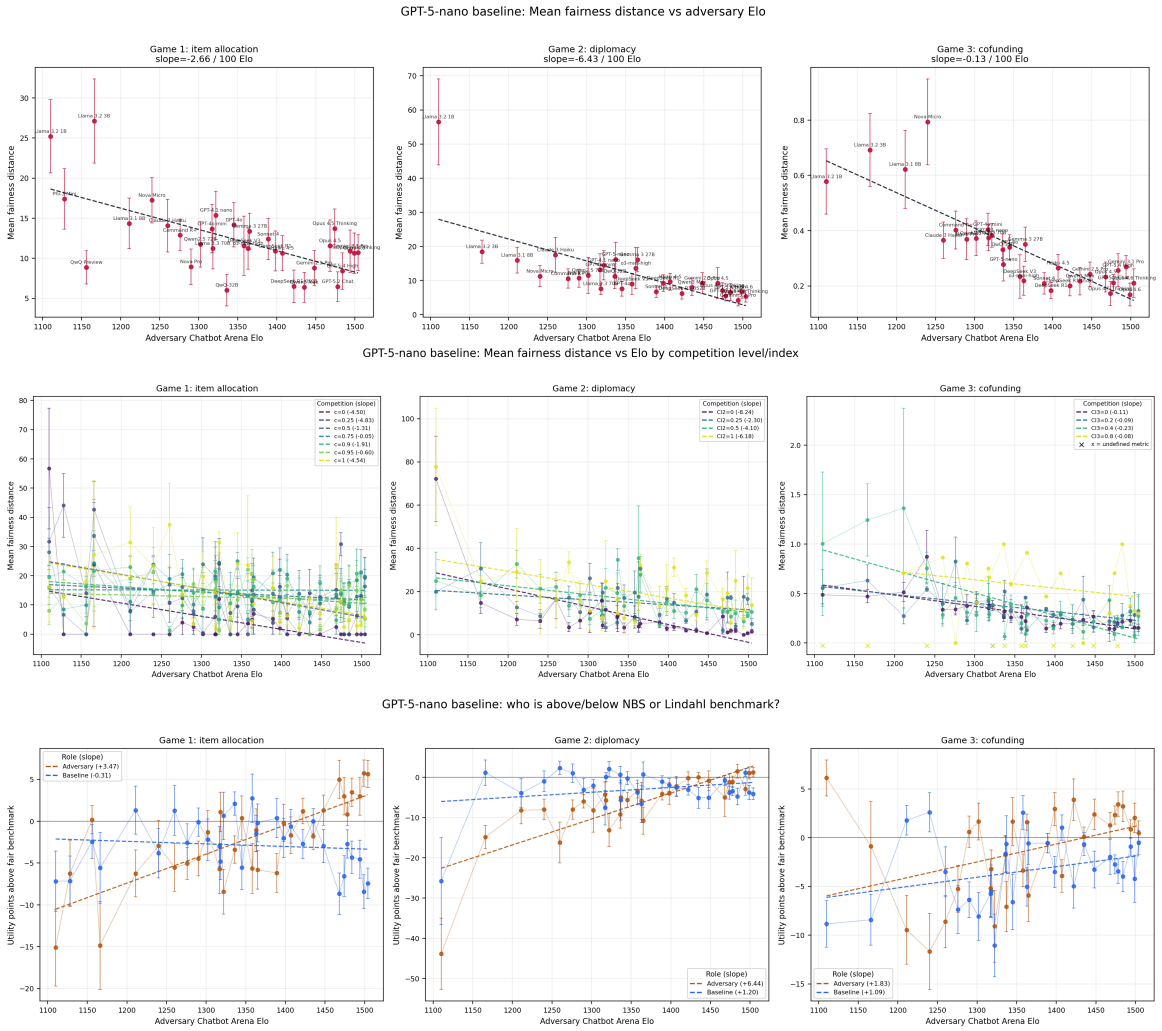

Figure 10: **Fairness and benchmark-relative extraction.** Stronger adversaries reduce distance from the NBS/Lindahl reference, especially in Games 1–2, but the role-specific residuals show that they sit further above their own benchmark share as Elo rises.

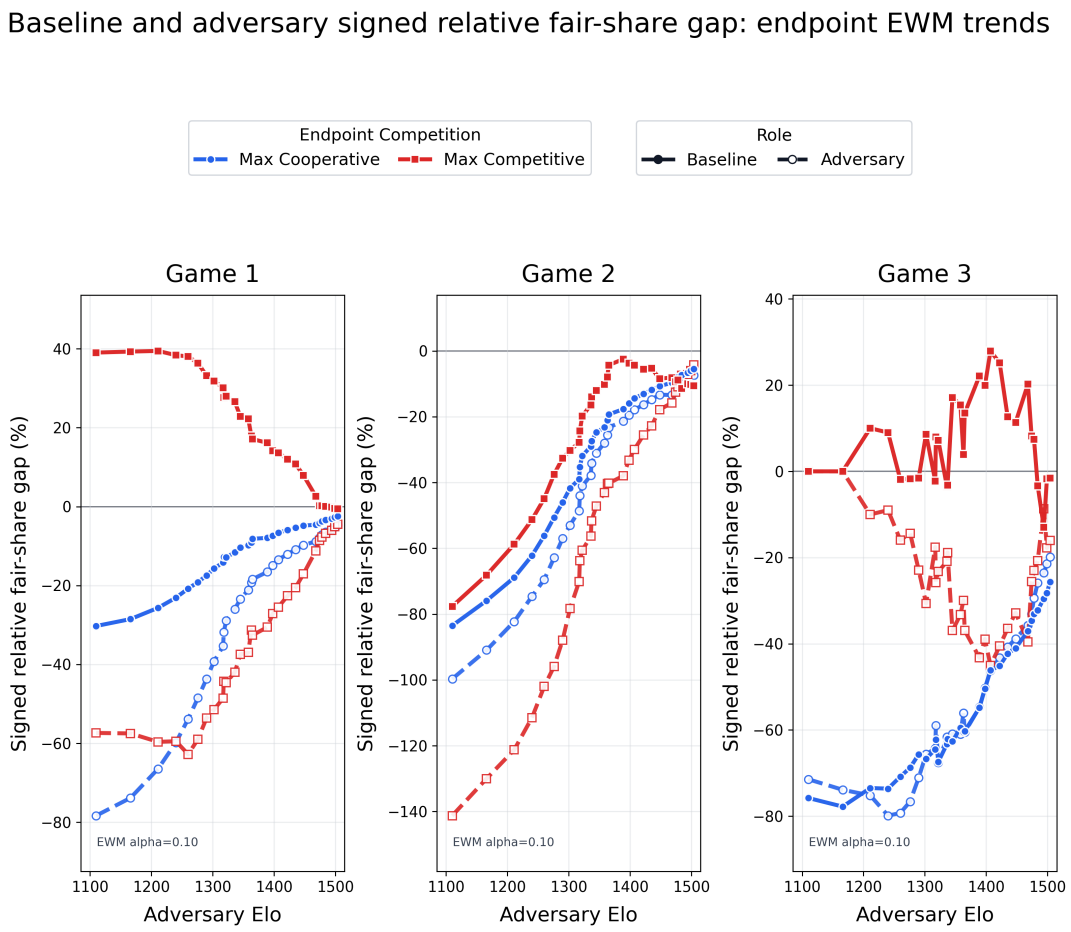

Figure 11: **Baseline and adversary fair-share gaps at cooperative and competitive endpoints.** Curves plot each role's signed relative gap from its fairness benchmark, smoothed over adversary Elo with an EWM coefficient of 0.10. Blue curves show the maximally cooperative endpoint and red curves show the maximally competitive endpoint; filled solid lines are the GPT-5-nano baseline and open dashed lines are the adversary. In cooperative settings, both roles generally move upward toward their benchmark shares as adversary capability increases. In competitive settings, especially Games 1 and 3, the adversary's fair-share gap improves as the baseline's worsens, indicating more redistributive benchmark convergence.

E.4. Llama 3.3 Baseline Details

The Llama baseline report exports all runs, per-model means, competition summaries, model-order summaries, and Elo-trend summaries. The main paper summarizes the replication; Table 5 gives the numerical slopes, Figure 12 gives the three overall plots, and Figure 13 gives the baseline-payoff breakdown moved out of the main text. Model-order effects are small to modest: the mean first-position advantage is +0.75 utility in Game 1, +0.72 in Game 2, and -1.04 in Game 3. Higher competition lowers social welfare in all three games, with welfare slopes of -91.58, -46.99, and -30.26 per unit of the respective competition index in Games 1–3.

Game	Runs	Adv. slope / 100 Elo	Adv. r	Base slope / 100 Elo	Gap slope / 100 Elo
Item allocation	140	+3.95	0.79	-0.65	+4.60
Diplomatic treaty	180	+5.02	0.88	-0.24	+5.26
Co-funding	180	+10.70	0.87	+2.19	+8.51

Table 5: **Llama 3.3 70B fixed-baseline replication.** Slopes are from the 500-run appendix analysis, with the adversary model varied and Llama 3.3 70B fixed as the baseline. The adversary-minus-baseline gap widens with Elo in all three games.

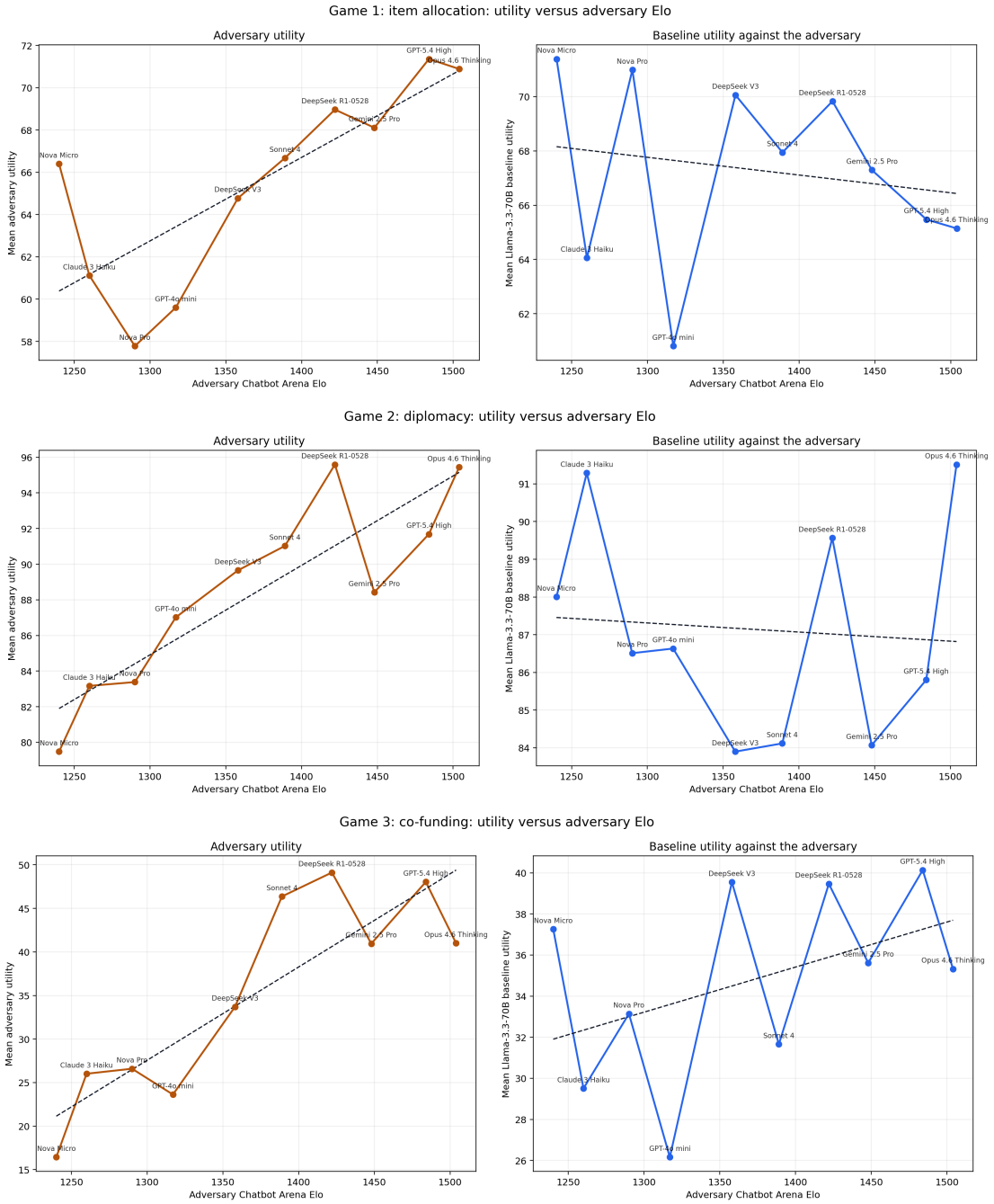

Figure 12: **Llama 3.3 70B baseline utility replication.** Top to bottom: item allocation, diplomatic treaty, and co-funding. Each panel varies the adversary model and holds the Llama baseline fixed.

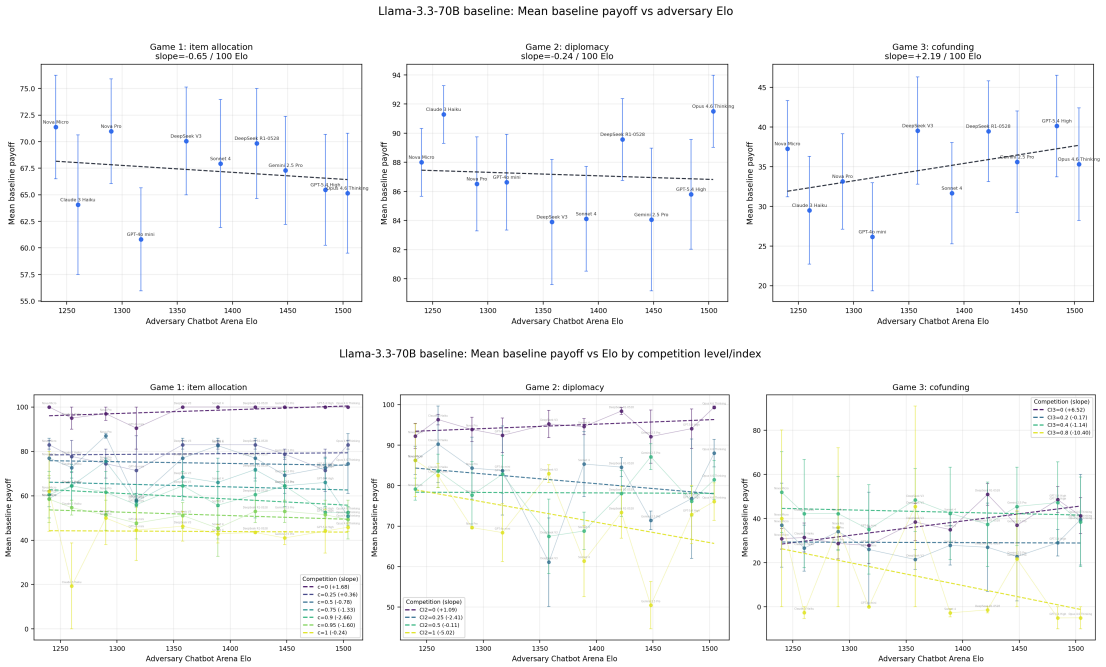

Figure 13: **Llama 3.3 baseline payoff against stronger adversaries.** Top: the three game-level Llama baseline payoff trends. Bottom: the same trends stratified by competition.

E.5. TTC Additional Readout

Family	Weak → strong	Utility delta	Improved	Worsened
GPT-5	minimal → high	+3.26	6	7
Claude Sonnet 4.6	low → max	+2.27	7	3
Gemini 3 Flash	minimal → high	-0.05	7	5

Figure 14: **Weak-to-strong TTC comparisons.** Counts are over 18 matched game-cell/order situations per family. The remaining cells are flat.

The cleanest monotone observed-token ladder is Gemini output tokens per call, whose payoff is flat from minimal to high effort. GPT-5 reports true reasoning tokens; measured medium effort uses more reasoning tokens per call than measured high effort in this one-seed run. Claude's output-token proxy is non-monotone in requested effort. For this reason, the paper treats TTC as a protocol stress test instead of a compute-scaling law.

The stage-token correlations are descriptive and confounded by game difficulty. Total target tokens in discussion, private thinking, proposal, and voting have correlations near -0.54 with target utility because longer and harder games consume more calls and larger contexts. The same token totals have near-zero correlations with target-minus-baseline utility gap, so the stage analysis does not support a direct token-to-advantage interpretation.

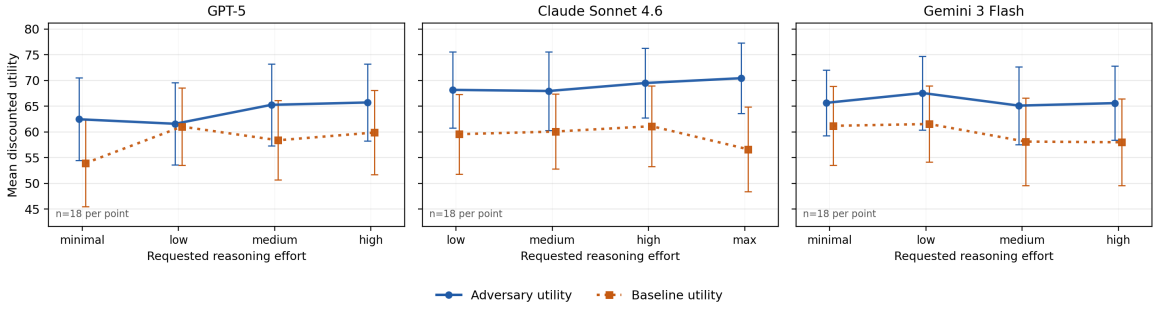

Figure 15: **Requested reasoning effort does not produce a monotone payoff curve.** Each panel uses the corresponding provider family's ordered effort labels on the x-axis; these are ordinal requested settings, not equal token budgets across providers. Solid lines show adversary utility, dotted lines show the GPT-5-nano baseline utility, points are means over 18 matched runs, and error bars are SEM.

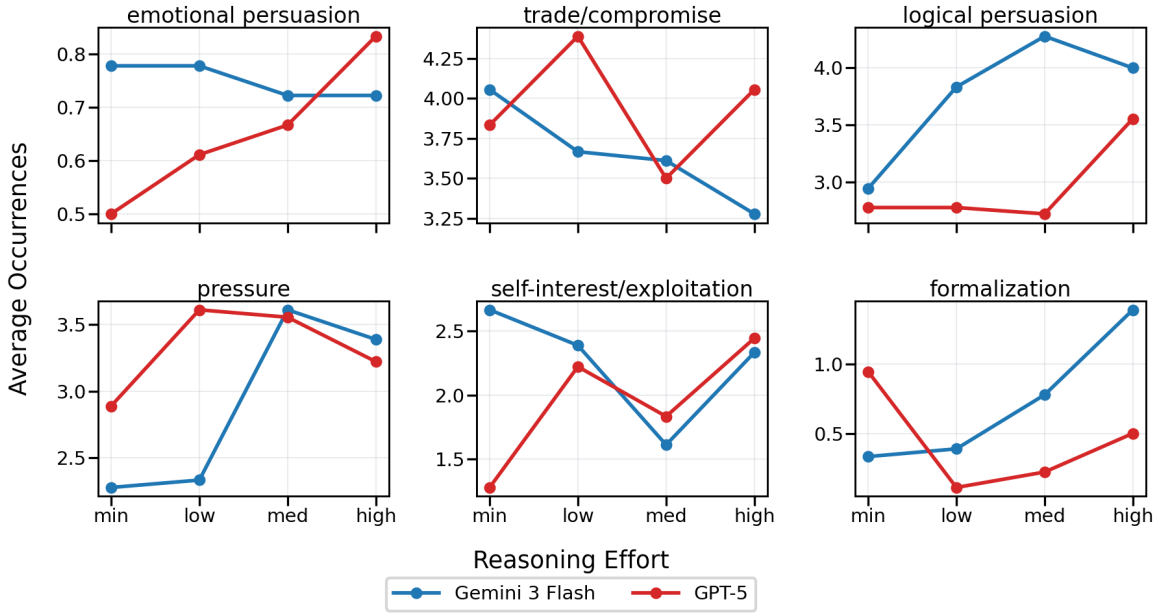

Figure 16: **Reasoning effort shifts strategic behavior without improving payoff.** GPT-5 ramps self-interest/exploitation with effort; Gemini 3 Flash drops trade/compromise. Both shifts land in categories Figure 2 marks as neutral or payoff-negative.

Family	Level	n	Target utility	Gap	Consensus
GPT-5	minimal	18	62.46	8.52	0.89
GPT-5	low	18	61.59	0.55	0.89
GPT-5	medium	18	65.27	6.88	0.89
GPT-5	high	18	65.73	5.84	0.94
Claude Sonnet 4.6	low	18	68.18	8.61	1.00
Claude Sonnet 4.6	medium	18	67.96	7.87	1.00
Claude Sonnet 4.6	high	18	69.50	8.38	1.00
Claude Sonnet 4.6	max	18	70.44	13.81	1.00
Gemini 3 Flash	minimal	18	65.66	4.47	1.00
Gemini 3 Flash	low	18	67.56	6.02	1.00
Gemini 3 Flash	medium	18	65.11	6.99	0.89
Gemini 3 Flash	high	18	65.60	7.61	0.94

Table 6: **Overall TTC summary by family and effort.** Gap is target utility minus GPT-5-nano baseline utility.

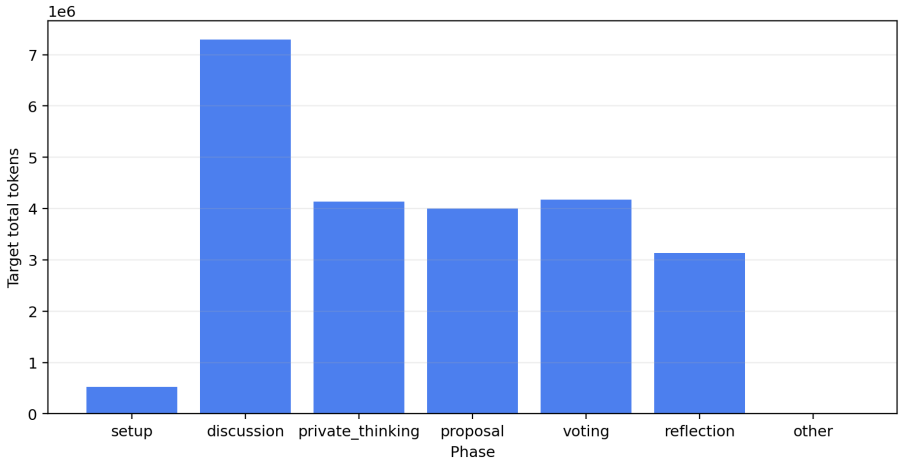

Figure 17: **TTC target tokens by phase.** The stage-token audit groups target calls into setup, discussion, private thinking, proposal, voting, reflection, and other phases. Most target tokens are spent after setup, especially in discussion and decision phases.

E.6. $N = 2$ Random-Pairing Check

The main bilateral design fixes GPT-5-nano as the primary baseline so that the opponent is controlled while the adversary varies. Figure 18 compares this controlled design against $N = 2$ heterogeneous random pairings. The slope signs and magnitudes are similar: fixed-baseline slopes are +5.28, +6.76, and +7.37 utility per 100 Elo in Games 1–3, while heterogeneous-pairing slopes are +7.49, +4.79, and +5.59. The fixed-baseline estimates are therefore not an artifact of comparing only against GPT-5-nano, although the random-pairing fits are naturally noisier because both sides vary.

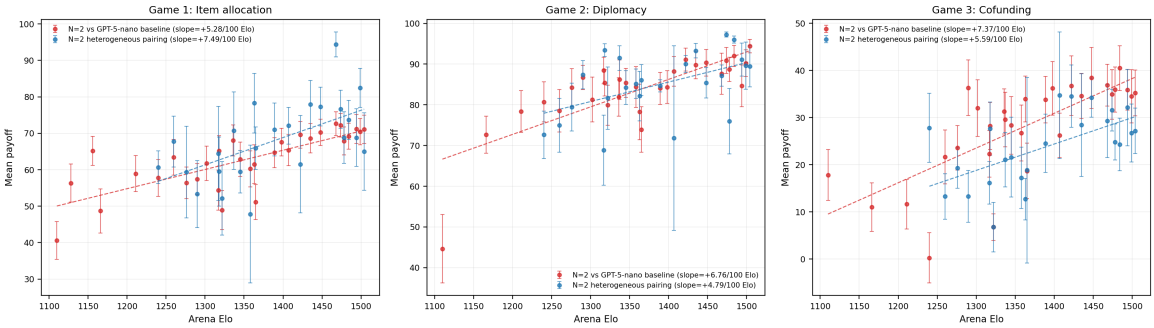

Figure 18: **Fixed baseline versus $N = 2$ heterogeneous pairings.** Capability scaling appears under both evaluation designs.

E.7. $N = 2$ Rounds to Consensus

Figure 19 shows rounds-to-consensus trends in the GPT-5-nano bilateral sweep. Games 1–2 have negative overall slopes: rounds fall by 0.15 and 0.09 per 100 Elo, ruling out the simplest story that stronger models earn more by prolonging bargaining. The Game 3 overall slope is flat. The competition breakdown is more informative: in the highest-scarcity cells, stronger models are sometimes more willing to reject negative-utility funding vectors, so deliberation can continue or end in no agreement—a qualitatively different failure mode from low-Elo format or arithmetic errors.

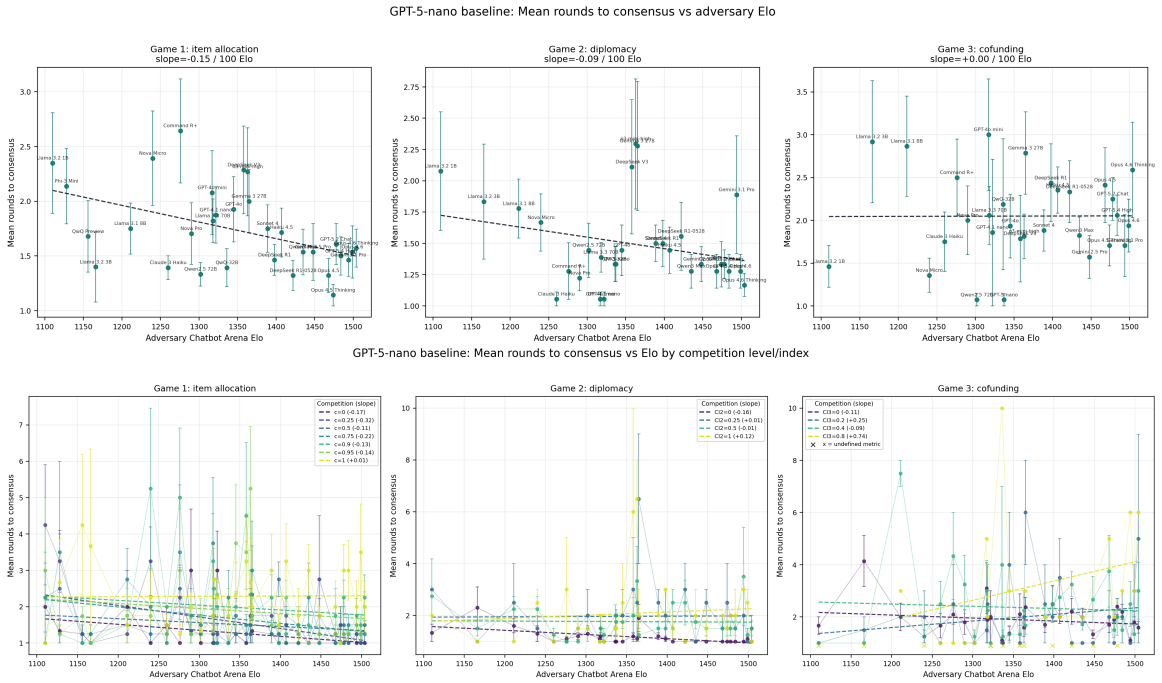

Figure 19: **Rounds to consensus in the GPT-5-nano bilateral sweep.** Top: the three game-level trends. Bottom: the same trends stratified by competition. Stronger adversaries usually do not need more rounds to earn more utility. The exception is scarce co-funding, where additional reasoning can produce better rejection of bad deals rather than faster agreement.

E.8. $N = 2$ Order Diagnostics

Figure 20 reports the adversary-payoff order diagnostics for the GPT-5-nano and Llama 3.3 baselines. The effect is not a stable law of first-mover advantage. Order interacts with game geometry: first proposals can establish the focal bundle in item allocation, but second movers can exploit revealed priorities; in co-funding, the first contribution vector can either coordinate a threshold or expose the proposer to bad cost sharing.

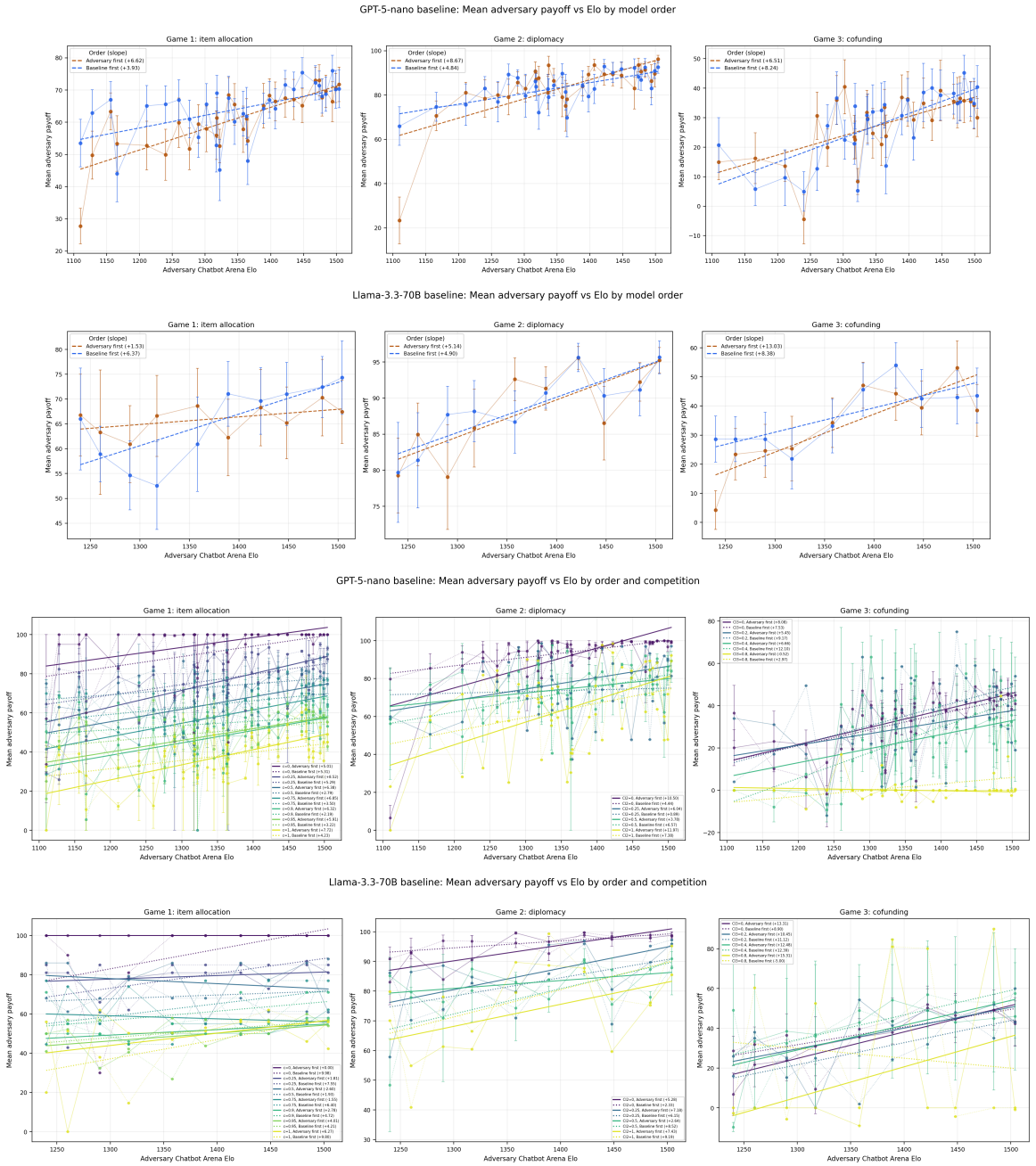

Figure 20: **Bilateral order diagnostics.** Rows 1–2 show adversary-payoff order trends for the GPT-5-nano and Llama 3.3 baselines. Rows 3–4 show the corresponding competition-stratified order trends.

E.9. Multi-Agent Competition Breakdowns

Figure 21 and Figure 22 give the competition-stratified versions of the main multi-agent ($N \geq 2$) capability plots. They support the main text's interpretation: capability advantages generally survive within game-specific competition strata, with the noisiest deviations in Game 3 public-good cells where feasibility and cost-sharing bind.

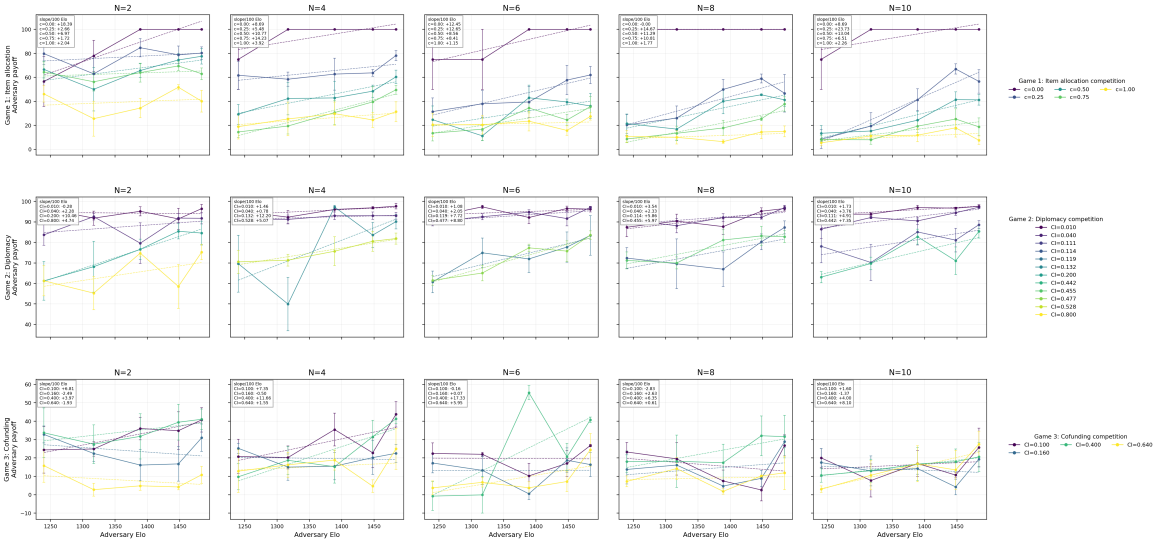

Figure 21: **Homogeneous-adversary payoff by competition.** Rows show item allocation, diplomatic treaty, and co-funding; columns show $N = 2, 4, 6, 8, 10$.

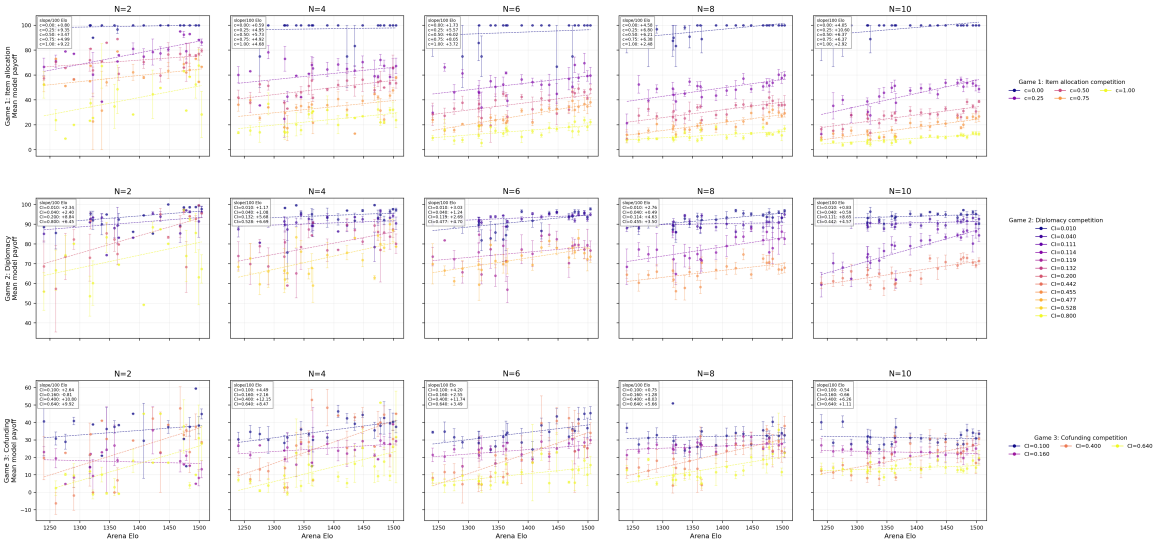

Figure 22: **Heterogeneous payoff by Arena Elo and competition.** Rows show item allocation, diplomatic treaty, and co-funding; columns show $N = 2, 4, 6, 8, 10$.

E.10. Heterogeneous Payoff Scaling

Figure 23 reports the full three-game heterogeneous payoff-vs-Elo result, and competition-stratified breakdowns appear in Figure 22 above. Figure 25 provides an Elo-bucket view that complements the continuous-Elo plot: the rank ordering by Elo holds at every group size, while the absolute spread compresses with N.

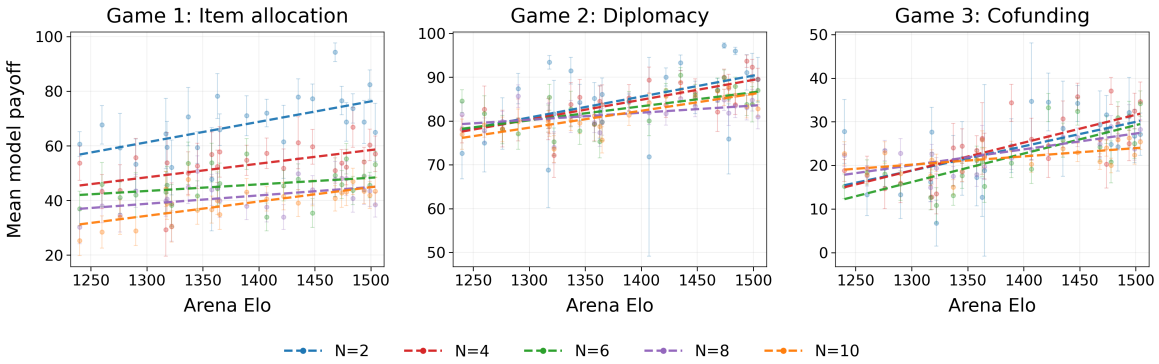

Figure 23: **Heterogeneous payoff scaling by Arena Elo and group size, all three games.** Each point averages a model's payoff over heterogeneous runs in which it appears. Capability remains predictive across N and games.

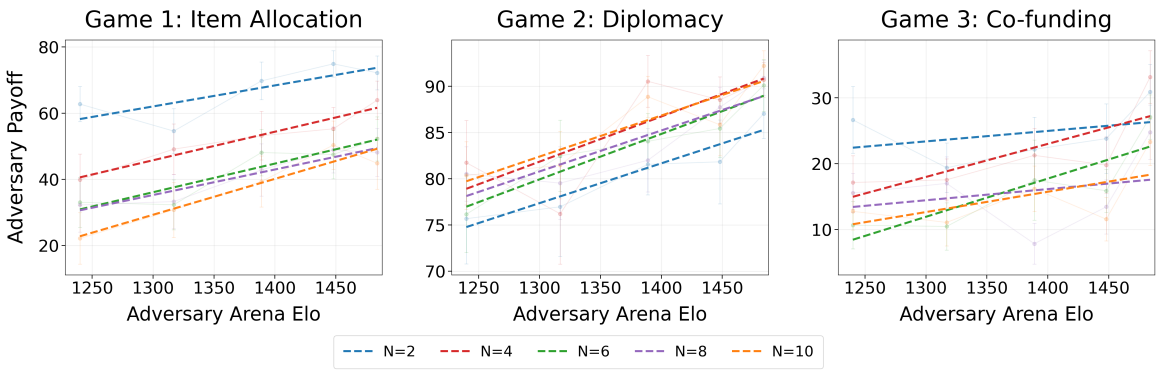

Figure 24: **Homogeneous-adversary payoff scaling by Arena Elo and group size, all three games.** Each point averages the inserted adversary's payoff over homogeneous-adversary runs with $N - 1$ GPT-5-nano agents. Capability remains predictive across N and games.

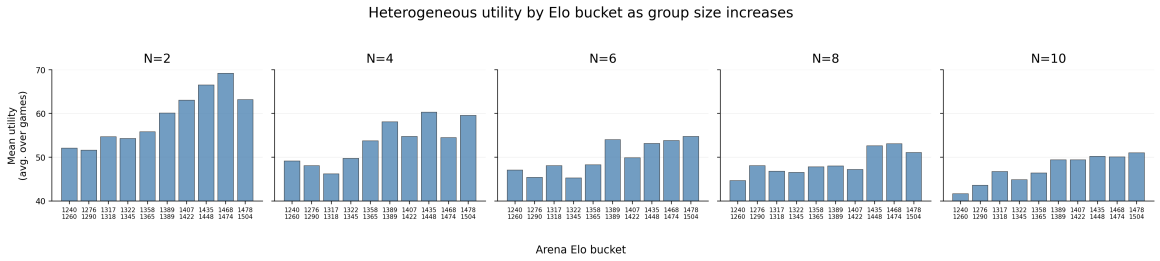

Figure 25: **Heterogeneous utility by Elo bucket as N grows.** Mean utility per Elo bucket at each group size. The Elo–utility ordering survives at every N; the absolute spread between top and bottom buckets compresses with N from ~ 18 utility points at $N = 2$ to ~ 9 at $N = 10$, but does not invert.

E.11. Multi-Agent Performance Elo

We convert every within-run payoff ordering into pairwise comparisons and fit a Bradley-Terry model, $P(i \succ j) = 1/(1 + 10^{(R_j - R_i)/400})$, then center the fitted ratings to mean 1500 within each subset. This payoff-performance Elo asks whether a model tends to beat its roster-mates, rather than whether its absolute utility is high.

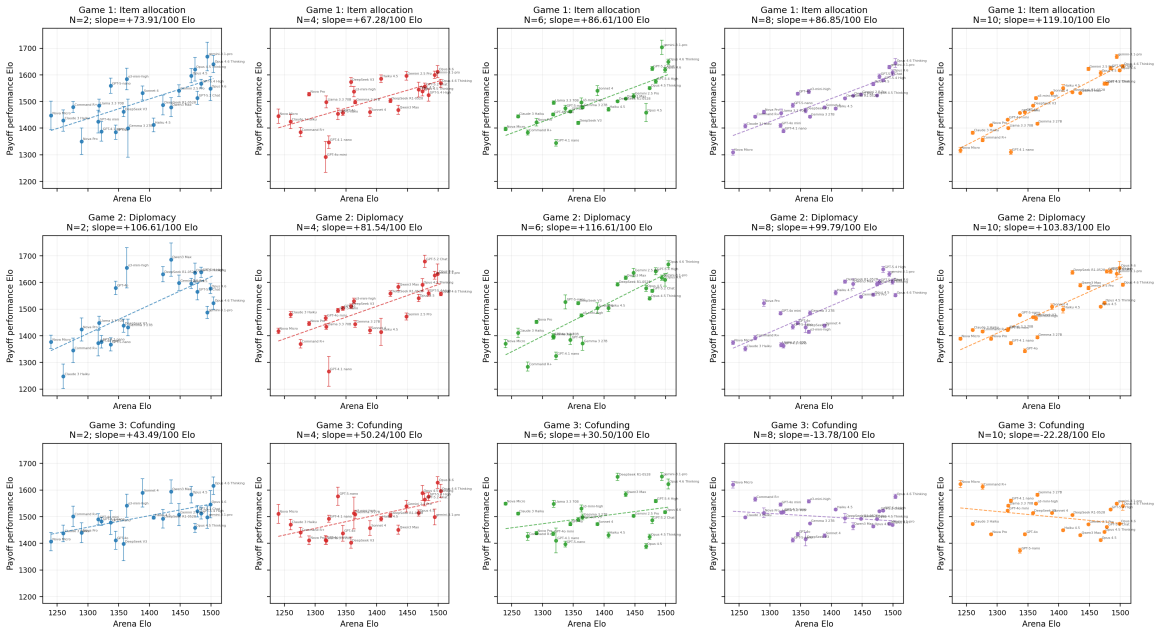

Figure 26: **Heterogeneous payoff-performance Elo.** Performance Elo is inferred from within-run utility rankings in heterogeneous rosters. It usually increases with Arena Elo, but Game 3 reverses at $N = 8$ and $N = 10$, where public-good spillovers and ties weaken the link between absolute payoff and beating one's roster-mates.

E.12. Multi-Agent Dilution Diagnostics

Figure 27 asks whether a larger GPT-5-nano fleet protects itself against an inserted adversary. The answer is mixed rather than protective. From $N = 2$ to $N = 10$, the average adversary-minus-fleet-mean advantage changes by +3.03 in Game 1, +7.67 in Game 2, and +5.00 in Game 3, although individual adversary models vary around those means.

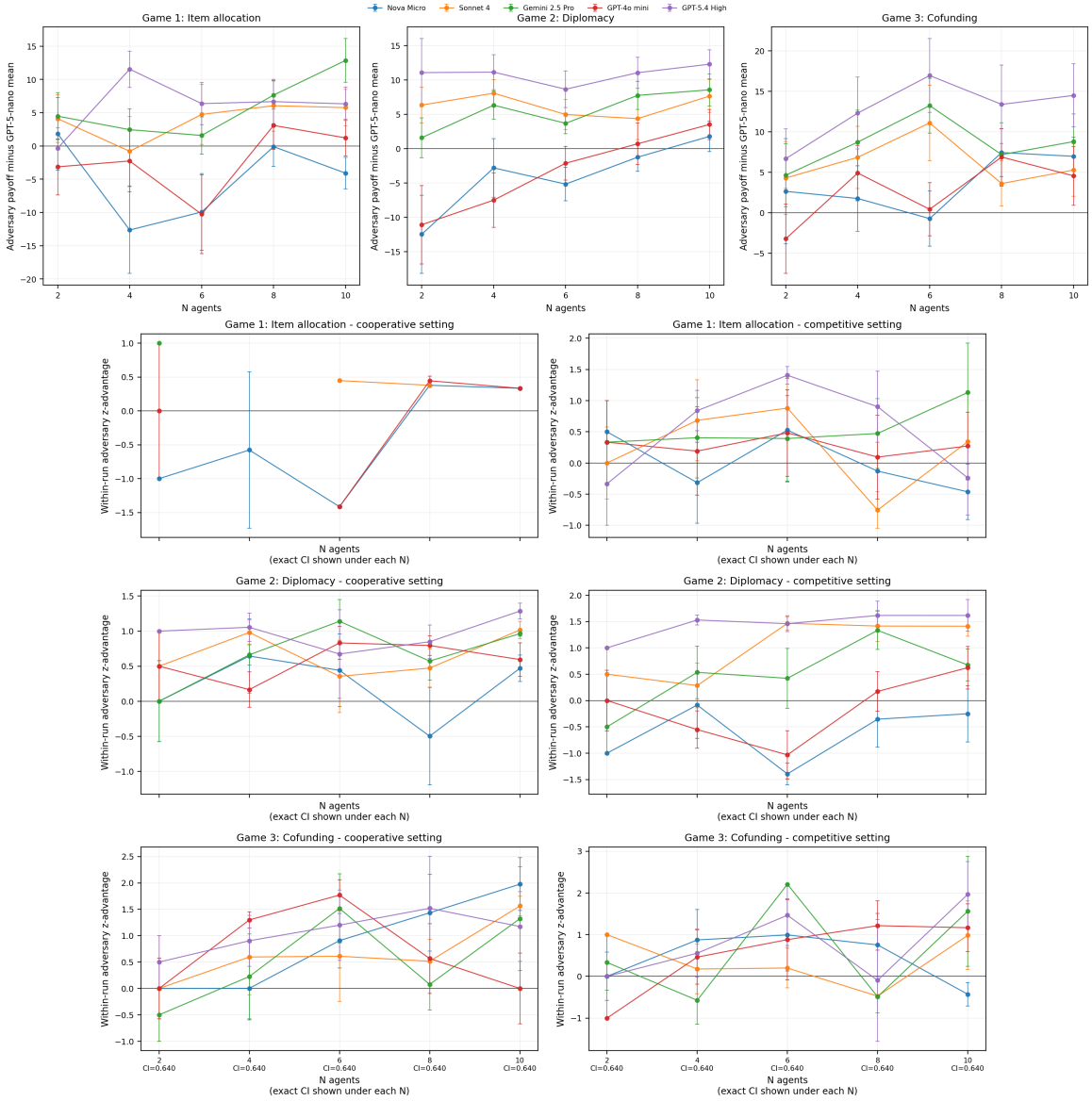

Figure 27: **Does a larger baseline fleet dilute a focal adversary?** Top: adversary payoff advantage over the GPT-5-nano fleet mean. Bottom: within-run z-advantage by competition band.

E.13. Multi-Agent Inequality by Group Size

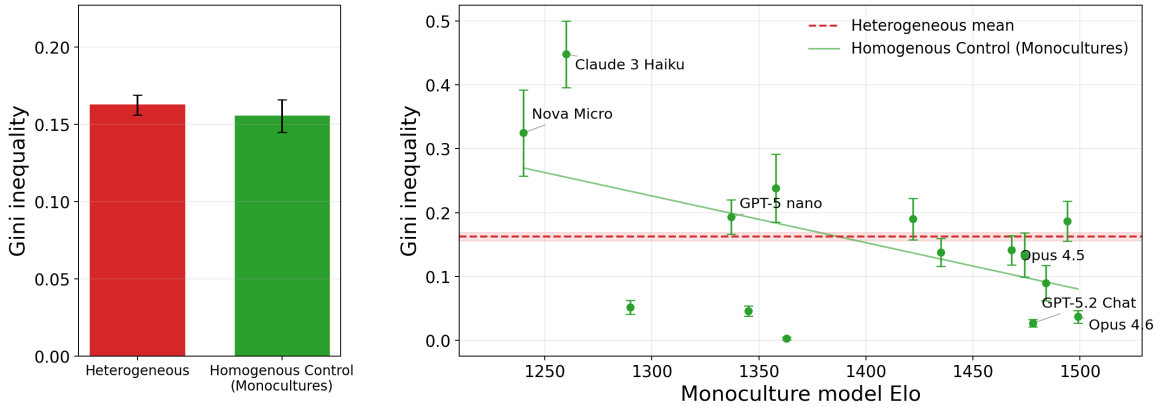

Figure 28: **Monoculture capability, not heterogeneity, drives Gini inequality.** *(a)* Heterogeneous rosters and pooled monoculture controls have equal aggregate corrected Gini. *(b)* Per-monoculture Gini vs. Arena Elo: weak monocultures sit above the heterogeneous reference (dashed red), capable ones below.

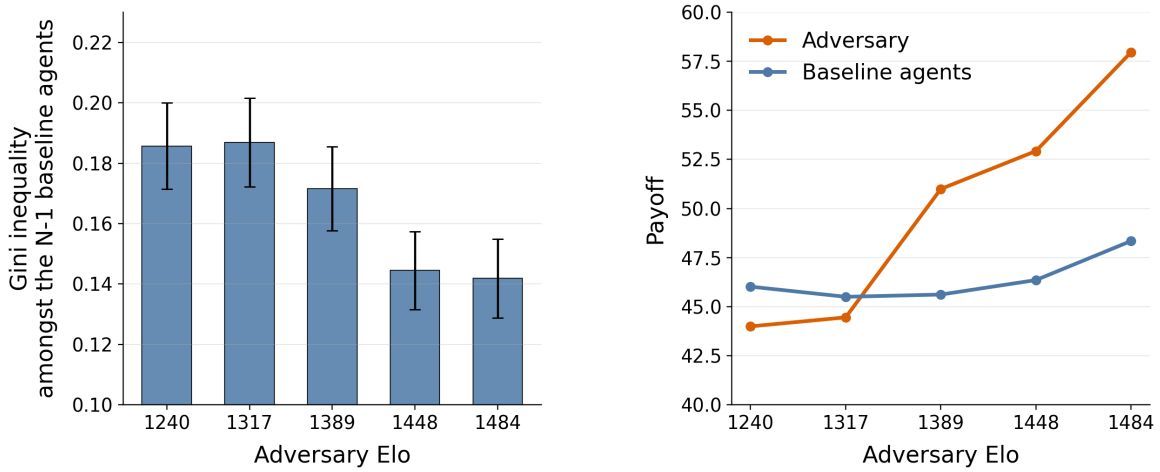

Figure 29: **Homogeneous-adversary inequality and role payoffs, by adversary Elo quartile.** Within-run Gini falls faster among the $N-1$ baselines than across all agents: the strong adversary equalizes the weak fleet while its own payoff rises sharply and per-baseline payoff stays nearly flat.

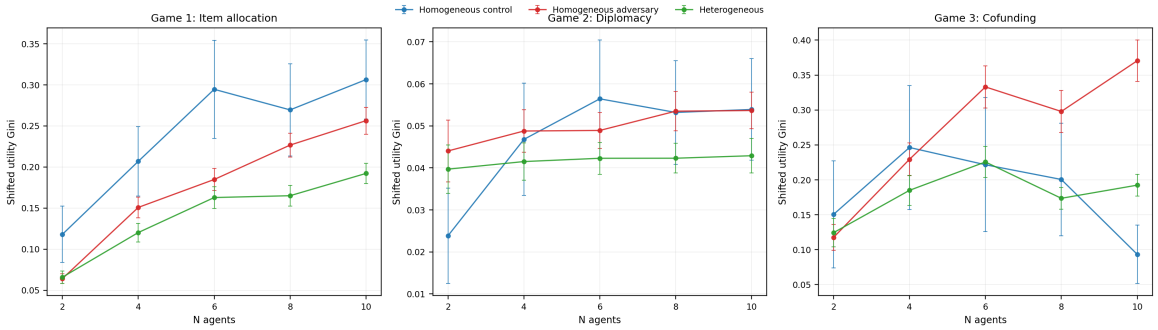

Figure 30: **Payoff inequality grows with group size in item allocation but not in treaty or public-good games.** Shifted utility Gini by N for homogeneous-control, homogeneous-adversary, and heterogeneous conditions. Game 1 shows the sharpest increase; Game 2 stays stable; Game 3 is dominated by feasibility rather than distributional competition.

E.14. Multi-Agent Fairness, Inequality, and Efficiency

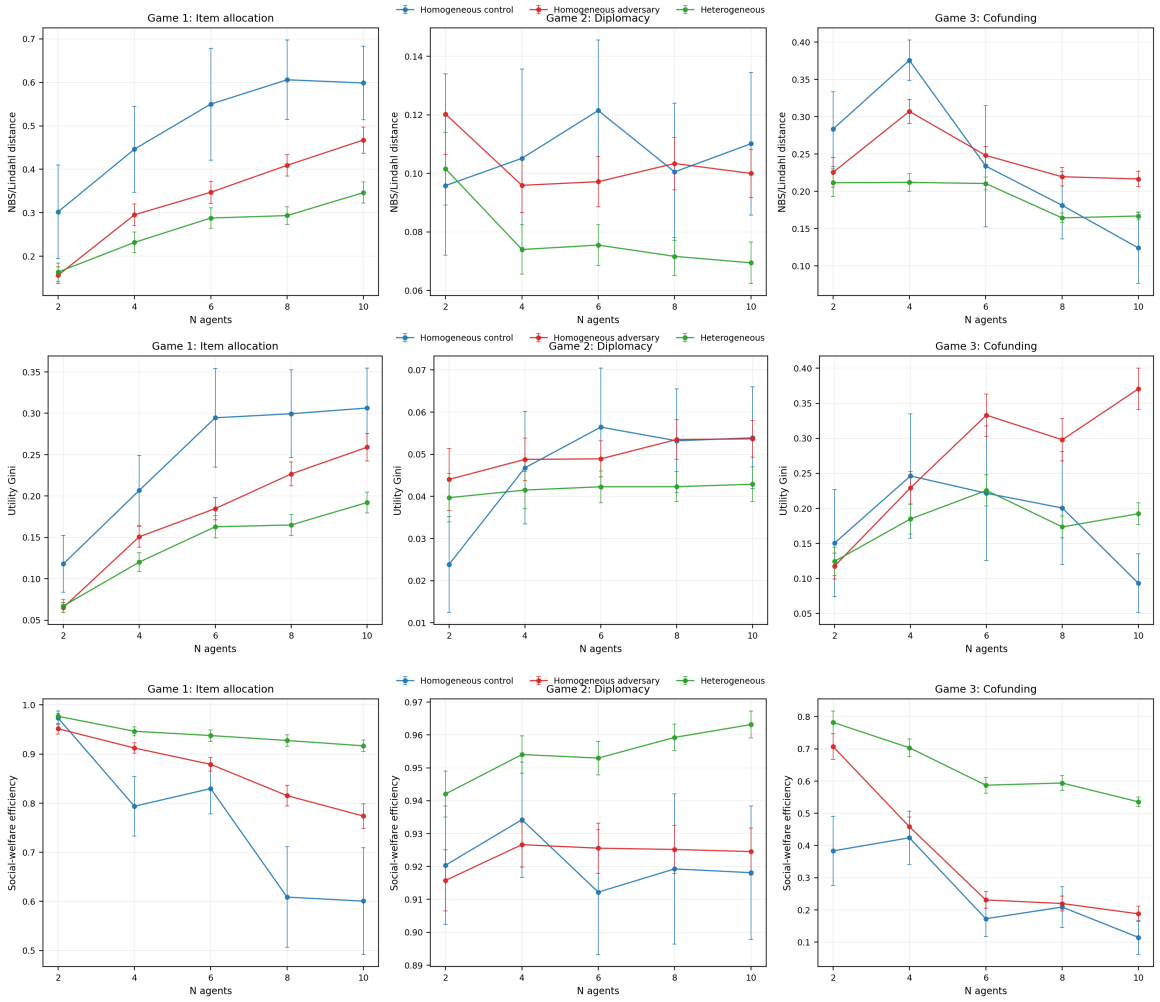

Figure 31: **Fairness, inequality, and welfare efficiency as group size increases.** Game 1 shows the clearest fairness degradation with N. Game 2 remains stable because continuous treaty issues allow compromise. Game 3 is dominated by public-good feasibility, cost shares, and no-consensus/all-zero outcomes.

E.15. Limitations

The empirical capability axis is LMArena Elo, which is measured in open-ended preference battles outside the bargaining games studied here. It is useful as a public ordering variable, although it does not isolate strategic reasoning from instruction following, formatting reliability, safety tuning, or provider-specific behavior. The multi-agent results are completed-run analyses; missing and failed high-N cells are excluded. This makes the reported estimates conditional on operational success. The TTC batch uses one seed per game cell and order, and only GPT-5 exposes provider-reported reasoning tokens in the same form as the analysis axis. Finally, the games are deliberately controlled abstractions. They isolate allocation, compromise, and public-good funding mechanisms while omitting durable commitments, long-horizon reputation, external tools, private side payments, and human-agent interaction.

Appendix F. Qualitative Strategic-Behavior Analysis

F.1. Annotation Pipeline

The qualitative analysis in Section 4.1 is based on an event-level LLM-judge annotation of the bilateral N=2 GPT-5-nano sweep and the test-time-compute (TTC) batch. We first sampled 2,730 N=2 rollouts and asked GPT-5.5-xhigh to brainstorm a candidate codebook of strategic-behavior labels, yielding 50 candidate labels. We then hand-pruned this set to 23 labels grouped into six categories: pressure, emotional persuasion, logical persuasion, trade/compromise, self-interest/exploitation, and formalization; Table 7 lists every chosen tag with its description and category.

Each rollout was independently re-annotated by GPT-5.5-xhigh judges in a second pass. The judge prompt explicitly instructs the model to label *every* conversation turn at which a tag's definition applies, not only the first, so a single rollout can contribute multiple events for the same tag. To avoid double-counting overlapping same-category tags within one turn, we deduplicate events at the (rollout, turn, category) level before any per-category aggregation; this removes 19% of raw N=2 events and 16–24% of TTC events across families. All category-level numbers in Section 4.1 and in the tables below use this turn-deduplicated count.

F.2. Codebook

Category	Tag	Description
pressure	Ultimatum language	The speaker frames acceptance as the only viable option, a final offer, or a take-it-or-leave-it demand.
	Conditional veto threat	The speaker ties acceptance or rejection to a concrete package condition.
	Fairness accusation pressure	The speaker pressures others by calling a proposal unfair, unreasonable, greedy, imbalanced, or one-sided.
	Frustration or disappointment display	The speaker explicitly displays frustration, disappointment, irritation, or exasperation.
	Adversarial callout	The speaker directly calls out another agent's behavior as evasive, nonconceding, noisy, or strategically unhelpful.
emotional persuasion	Rapport before pressure	The speaker uses warmth, appreciation, or praise immediately before disagreement, firmness, or a strategic ask.
	Empathy-then-pivot	The speaker acknowledges another agent's position, then pivots to their own ask or constraint.
logical persuasion	Utility arithmetic receipts	The speaker uses explicit arithmetic over values, utilities, costs, or totals to justify the deal.
	Agent-specific payoff accounting	The speaker calculates or narrates what named agents get under a proposal.
	Threshold-gap calculation	The speaker identifies exactly how far a vote/project/deal is from passing.
	Fairness ledger argument	The speaker itemizes balance, proportionality, equal burden, or equal gains in a ledger-like way.
trade/compromise	Conditional quid pro quo	The speaker offers an explicit if-you-do-X-I-will-do-Y exchange.
	Conditional support ledger	Agents maintain an explicit I-back-yours-if-you-back-mine support ledger.
	Concession laddering	The speaker marks a movement from an initial position toward a less demanding one.
	Low-weight concession leverage	The speaker explicitly turns a low-priority issue/item/project into a chip for gains on core priorities.
	Vote-history diagnostics	Agents use prior proposal vote counts or failed-round evidence to infer which package can pass.
self-interest/exploitation	Self-advocacy/value maximization	The speaker explicitly argues for maximizing their own payoff or preserving their own best outcome.
	Leverage preservation	The speaker refuses to give up a valuable asset/position because it preserves bargaining leverage.
	Zero-value subsidy	A Game 3 agent pays positive cost for a funded project they privately value at zero.
	Silent free beneficiary	A Game 3 funded project gives positive private value to an agent who contributes nothing to it.
	Accepted-loss capitulation	The rollout reaches consensus even though at least one agent ends with negative final utility.
formalization	Budget carryover hallucination	Agents talk as if failed-round spending, rejected proposals, or unfunded partial contributions carry forward.
	Counter-anchor cost policing	A speaker corrects an impossible or confused anchor by re-centering on actual costs, budgets, or constraints.

Table 7: **The 23 selected strategic-behavior tags, grouped by category.** Each tag was selected from a 50-label candidate pool generated by GPT-5.5-xhigh over 2,730 N=2 rollouts, hand-pruned for non-overlap and informativeness, then re-annotated event-by-event by GPT-5.5-xhigh judges.

F.3. Category-Level Payoff Correlations

Table 8 reports the Spearman correlation between turn-deduplicated event count and adversary final utility for each category, pooling over all 1,912 adversary speaker-rollouts in the bilateral sweep. This is the table behind the left panel of Figure 2.

Category	Spearman r	p	n rollouts used
trade/compromise	$+0.170$	$5.2e-14$	1418
emotional persuasion	$+0.140$	$5.9e-10$	727
logical persuasion	$+0.117$	$2.6e-07$	1042
pressure	-0.054	0.017	978
self-interest/exploitation	-0.255	$4.1e-30$	636
formalization	-0.324	$8.1e-49$	182

Table 8: **Category-level Spearman correlation between tag-event count and adversary final utility** (N=2 bilateral vs. GPT-5-nano, 1,912 adversary speaker-rollouts). Counts are deduplicated at the (rollout, turn, category) level so that overlapping same-category tags on a single conversation turn count once. Trade/compromise, emotional persuasion, and logical persuasion are payoff-positive; formalization, self-interest/exploitation, and pressure are payoff-negative.

F.4. Per-Tag Mechanism Evidence

Table 9 reports, for each of the 23 chosen tags, (i) the Spearman rank correlation between model Elo and the adversary's per-rollout tag intensity, (ii) the slope of intensity in events per 100 Elo points, and (iii) the Spearman rank correlation between per-model tag rate and per-model mean adversary utility. The deal-engineering signature of high-Elo play identified in Section 4.1—counter-anchor cost policing, conditional veto threats, threshold-gap calculations, and agent-specific payoff accounting—are the tags with simultaneously high ρ(Elo) and high ρ(payoff).

Category	Tag	ρ(Elo)	slope/100 Elo	ρ(payoff)
pressure	Adversarial callout	$+0.76$	$+0.11$	$+0.68$
	Conditional veto threat	$+0.73$	$+0.15$	$+0.67$
	Fairness accusation pressure	$+0.66$	$+0.07$	$+0.46$
	Frustration or disappointment display	$+0.61$	$+0.02$	$+0.57$
	Ultimatum language	$+0.57$	$+0.06$	$+0.54$
emotional persuasion	Rapport before pressure	$+0.19$	$+0.03$	$+0.13$
	Empathy-then-pivot	-0.58	-0.04	-0.51
logical persuasion	Threshold-gap calculation	$+0.67$	$+0.02$	$+0.41$
	Agent-specific payoff accounting	$+0.48$	$+0.05$	$+0.57$
	Fairness ledger argument	$+0.39$	$+0.03$	$+0.21$
	Utility arithmetic receipts	$+0.35$	$+0.08$	$+0.40$
trade/compromise	Low-weight concession leverage	$+0.33$	$+0.02$	$+0.34$
	Conditional quid pro quo	$+0.32$	$+0.04$	$+0.13$
	Vote-history diagnostics	$+0.27$	$+0.02$	$+0.11$
	Conditional support ledger	$+0.08$	$+0.00$	$+0.16$
	Concession laddering	$+0.02$	-0.00	-0.09
self-interest/exploitation	Silent free beneficiary	$+0.47$	$+0.02$	$+0.30$
	Zero-value subsidy	$+0.03$	$+0.00$	-0.27
	Leverage preservation	-0.04	-0.00	-0.15
	Self-advocacy/value maximization	-0.06	$+0.00$	-0.17
	Accepted-loss capitulation	-0.31	-0.01	-0.48
formalization	Counter-anchor cost policing	$+0.81$	$+0.05$	$+0.80$
	Budget carryover hallucination	-0.44	-0.03	-0.66

Table 9: **Per-tag mechanism evidence for the bilateral N=2 sweep against GPT-5-nano.** ρ(Elo) is the Spearman rank correlation between model Elo and the adversary's mean per-rollout tag intensity (turn-deduplicated). Slope is intensity per 100 Elo points. ρ(payoff) is the Spearman rank correlation, at the model level, between mean tag rate and mean adversary utility. Positive ρ(Elo) means the tag is used more often by stronger adversaries; positive ρ(payoff) means adversaries that use the tag more often earn more.

Appendix G. Prompt Templates

Full verbatim prompts for all games and phases are included in the Github. This section summarizes the prompt structure and key design choices.

G.1. Protocol Phases

Each round proceeds through a fixed sequence of prompted phases. The setup phase runs once at the start; all subsequent phases repeat each round.

1. **Setup (once):** Game rules, voting mechanics, discount factor, and private preferences are delivered in a single combined prompt. Agents acknowledge the rules before play begins.
2. **Discussion:** Agents take turns in a public channel. The first speaker receives a context-setting prompt; subsequent speakers see the accumulated conversation. From round 2 onward, the prompt references prior proposals and votes.
3. **Private Thinking:** Each agent receives its full preference reminder and is asked to produce a JSON object with fields for reasoning, strategy, key priorities, and potential concessions. This output is never shown to other agents.

4. **Proposal:** Each agent submits a structured JSON proposal (an allocation in Game 1, a treaty vector in Game 2, or a contribution vector in Game 3). The prompt specifies exact schema constraints and includes a second valid-JSON example.
5. **Voting:** All proposals from the round are shown together. Each agent votes accept or reject on every proposal independently, with a reminder of the utility formula and discount schedule.
6. **Reflection:** If the joint proposal is rejected, agents see the counterfactual outcome and are asked to consider adjustments.

G.2. Game-Specific Setup Prompts

Each game's setup prompt contains (a) a rules block shared with all agents and (b) a private preferences block unique to each agent.

Game 1: Item Allocation. The rules block states that N agents negotiate over m named items for up to T rounds. It specifies the two-thirds supermajority voting rule, the per-round discount factor γ, and the zero-utility disagreement payoff. The private block lists each item's name and the agent's secret valuation, notes the theoretical maximum, and reminds the agent that preferences may be revealed truthfully or misleadingly. Proposals are JSON objects mapping each agent to a list of item indices; the prompt enforces a complete-ownership invariant requiring every item to appear exactly once.

Game 2: Diplomatic Treaty. The rules block frames the negotiation as a multi-issue diplomatic accord over m continuous policy rates on a 0–100% scale. Each issue has a named policy dimension with a concrete minimum-to-maximum interpretation. The prompt explicitly distinguishes preferred rates from importance weights and provides worked utility examples. A simulation-boundary disclaimer restricts discussion to policy percentages and bargaining strategy. The private block lists each issue's secret ideal position and importance weight. Proposals are JSON vectors of m integer percentages.

Game 3: Participatory Budgeting. The rules block frames the game as a co-funding exercise over m named projects, each with a public cost. It emphasizes the all-or-nothing funding rule: a project is funded only if total contributions meet or exceed its cost, and contributions to unfunded projects are returned. The utility formula (valuation minus contribution, summed over funded projects) is stated with an explicit warning that over-contributing produces negative per-project utility. The private block lists each project's cost and the agent's secret valuation, the agent's private budget, total group budget, and collective coverage ratio. Each agent's proposal is a contribution vector; all vectors are combined into one joint proposal before voting.

G.3. Key Prompt Design Choices

- **No strategic coaching:** Prompts describe rules and utility formulas but do not suggest negotiation strategies, bluffing, or exploitation tactics. Agents are told they *may* reveal preferences truthfully or misleadingly; the choice is left to the model.
- **JSON schema enforcement:** Every structured output (thinking, proposal, vote) specifies an exact JSON schema and includes a second valid example to reduce formatting failures.
- **Urgency signals:** In the final two rounds, discussion and thinking prompts include a time-pressure warning. This is an operational reliability measure, not a strategic manipulation.

- **Round-over-round context:** From round 2 onward, discussion prompts reference prior proposals and vote outcomes so agents can adapt. Private thinking prompts include a full preference reminder to counteract context-window decay.
- **Voting independence:** The voting prompt presents all proposals simultaneously and asks for independent accept/reject decisions on each, with an explicit note that accepting one proposal does not preclude accepting another.