Scaling Laws for Strategic Interactions

author names withheld

Under Review for NExT-Game 2026

Abstract

LLM agents increasingly negotiate on behalf of users and firms, often under asymmetries in capability or number. Does a more capable agent grow joint surplus, or extract an unfair share at weaker counterparts' expense? We run scaling-law-style sweeps along three axes—agent capability (LMArena Elo and reasoning-token budget), agent count (up to $N = 10$), and degree of competition—across three new multi-turn negotiation games: item allocation, treaty bargaining, and participatory budgeting. Across 5,000+ games and 30+ models, scaling capability simultaneously increases joint efficiency and the surplus utility stronger agents extract from weaker ones, with the crossover from fair to exploitative play within the current frontier. Game structure mediates these effects: treaty bargaining absorbs capability gaps more stably than the other two. Test-time-compute scaling does not reliably translate into bargaining gains. Selecting a more capable LLM agent does not guarantee a Pareto improvement, with implications for economic deployment and scalable oversight. Our code to reproduce our experiments can be found here: https://anonymous.4open.science/r/bargain/README.md.

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 in order to try 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 [4, 18, 40, 43] but with other agents, be it on online marketplaces [34, 45], or as part of AI-based monitoring and oversight regimes [5]. 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. As the example above illustrates, these differences raise a number of important and practical questions. Such questions loom large not only for individual actors deploying such agents in the

economy and throughout society [16, 23, 42], but also in more narrow (but critically important) applications, such as using weaker LLM agents to monitor stronger agents in scalable oversight regimes where the stronger agent may not be fully aligned [3, 15, 29].

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 [14]) 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 between agents changes. In order 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 both interpretable enough to smoothly vary the degree of competition and the number of agents, but also complex enough to pose a meaningful challenge to frontier models.

In Section 3.1 we introduce a unified three-game benchmark spanning rivalrous allocation, continuous treaty bargaining, and threshold public-goods co-funding, together with the metrics and experimental setup used throughout. In Section 4.1 we characterize how model capability and competition jointly shape utilities, welfare, and benchmark-relative exploitation in bilateral play, and Section 4.2 shows that native test-time reasoning effort is not a reliable substitute. Section 4.3 extends the analysis to $N \in \{2, 4, 6, 8, 10\}$ agents with homogeneous-control, one-focal-adversary, and heterogeneous-ecology designs.

2. Related Work

Scaling and strategic learning. Scaling-law work primarily studies single-agent prediction or control [24, 25, 28]. Strategic scaling has been explored in multi-agent reinforcement learning games such as Hex, Connect Four, and Pentago [27, 37], but natural-language bargaining between deployed LLM agents remains less understood. Concurrent work studies capability gaps in oversight [19], bilateral LLM negotiation losses for weaker models [50], and inference-time opponent modeling [32]. We differ by jointly varying model capability, group size, and degree of preference conflict under a single protocol.

LLM negotiation and cooperation. Recent benchmarks probe negotiation, resource exchange, and social interaction with LLM agents [1, 6, 7, 12, 17, 30, 38, 48]. Other work evaluates rationality and cooperation in matrix games, Diplomacy-like settings, and behavioral-economics tasks [2, 8, 10, 13, 21, 26, 33, 35, 46, 47]. Our benchmark separates two channels that these fixed-task evaluations often conflate: stronger agents may reduce irrationality and raise joint surplus, while also exploiting weaker counterparts in the surplus split.

Asymmetric capabilities in games. Formal models of bounded rationality and computational asymmetry study players with different computational power, costly strategies, or pre-computation budgets [11, 22, 39, 41, 44]. We use these ideas operationally by comparing realized LLM negotiations to cooperative-game-theoretic fairness references, asking whether empirical capability gains manifest as welfare creation, surplus extraction, or both.

3. Methodology

3.1. Environments

All three environments use the same repeated negotiation protocol. Agents receive private preferences, discuss publicly, write private scratchpads, submit structured proposals, and privately vote. A proposal is implemented when it receives at least a two-thirds supermajority, rounded up; for $N = 2$, both agents must accept. If no proposal passes after T rounds, all agents receive zero utility. Utilities are discounted by γ^{t-1} for agreements reached in round t.

Item allocation. Agents bargain over indivisible items $\mathcal{M}$. Agent i has private values $\boldsymbol{v}^i$ with $v^i_j \geq 0$ and $\sum_j v^i_j = 100$; an allocation $\boldsymbol{A}$ assigns each item to one agent and yields $u^i(\boldsymbol{A}, t) = \gamma^{t-1}\boldsymbol{A}[i] \cdot \boldsymbol{v}^i$. Competition is the cosine similarity α_{ij} between value vectors: high similarity makes items rivalrous, while orthogonal values permit efficient assignment to the highest valuers.

Diplomatic treaty negotiation. Agents negotiate a continuous treaty vector $\boldsymbol{a} \in [0, 1]^m$. Agent i has ideal positions $\boldsymbol{p}^i$ and issue weights $\boldsymbol{w}^i$ on the simplex, receiving $u^i(\boldsymbol{a}, t) = \gamma^{t-1}100 \sum_k w^i_k(1 - |p^i_k - a_k|)$. We vary preference correlation ρ over ideal positions and interest overlap θ over issue weights, separating agreement about desired positions from agreement about which issues matter.

Participatory budgeting. Agents co-fund threshold public projects. Agent i has budget b^i and values $\boldsymbol{v}^i$; contribution matrix $\boldsymbol{X}$ funds project j when $\sum_i x_{ij} \geq c_j$, returning contributions to unfunded projects. If $S(\boldsymbol{X})$ is funded in round t, $u^i(\boldsymbol{X}, t) = \gamma^{t-1} \sum_{j \in S(\boldsymbol{X})}(v^i_j - x_{ij})$. We vary value alignment α_{ij} and budget scarcity $\sigma = \sum_i b^i / \sum_j c_j$.

3.2. Metrics and Experimental Setup

Our main capability coordinate is LMArena Elo from the March 31, 2026 snapshot. The bilateral sweeps fix GPT-5-nano as the baseline because it sits near the middle of the adversary Elo range and reliably completes the protocol; a Llama 3.3 70B replication checks baseline robustness. We use raw utility, realized social welfare, agreement timing, and fairness residuals. For Games 1–2, the fairness reference is the Nash Bargaining Solution, $o^{\text{NBS}} = \arg\max_o \prod_i u^i(o)$ [36]. For Game 3, the reference is Lindahl-style benefit-proportional cost sharing [20, 31]. Across games, exploitation is $E_i = (u^i_{\text{actual}} - u^i_{\text{benchmark}})/|u^i_{\text{benchmark}}|$.

The bilateral GPT-5-nano sweeps use $N = 2$, $T = 10$, $\gamma = 0.9$, two discussion turns, both model orders, and 30+ adversary models. Game 1 uses seven competition levels; Games 2–3 use nine competition cells each. The $N > 2$ batches use $N \in \{2, 4, 6, 8, 10\}$ with homogeneous GPT-5-nano controls, one-focal-adversary groups, and heterogeneous rosters sampled from a 24-model pool. The test-time-compute stress test spans GPT-5, Claude Sonnet 4.6, and Gemini 3 Flash at four native reasoning-effort levels against GPT-5-nano. Appendix C gives the run inventory, grids, and reproducibility details.

4. Results

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

The bilateral GPT-5-nano sweep contains 862 item-allocation runs, 540 treaty runs, and 539 co-funding runs. Higher-Elo adversaries earn more utility in all three games: slopes are $+5.28$, $+6.76$,

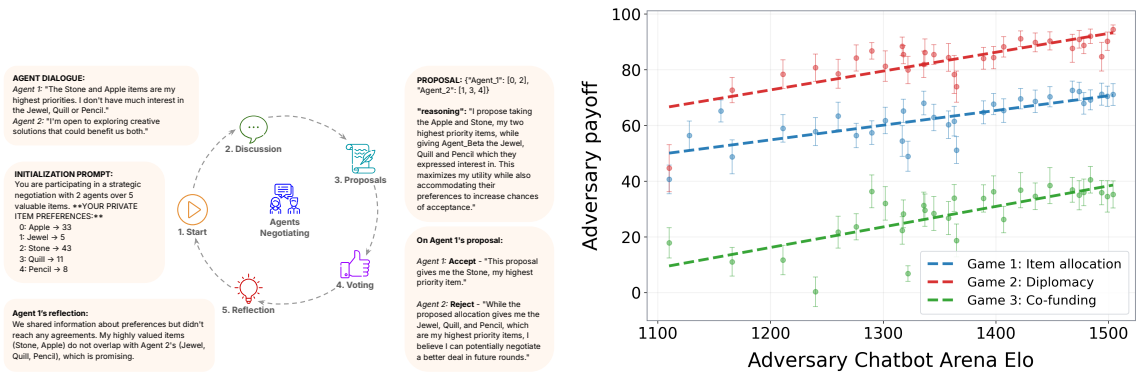

Figure 1: **Benchmark protocol and headline bilateral scaling.** Left: agents receive private preferences, negotiate over structured proposals, and privately vote until agreement or timeout. Right: adversary payoff against GPT-5-nano rises with Elo across all three games.

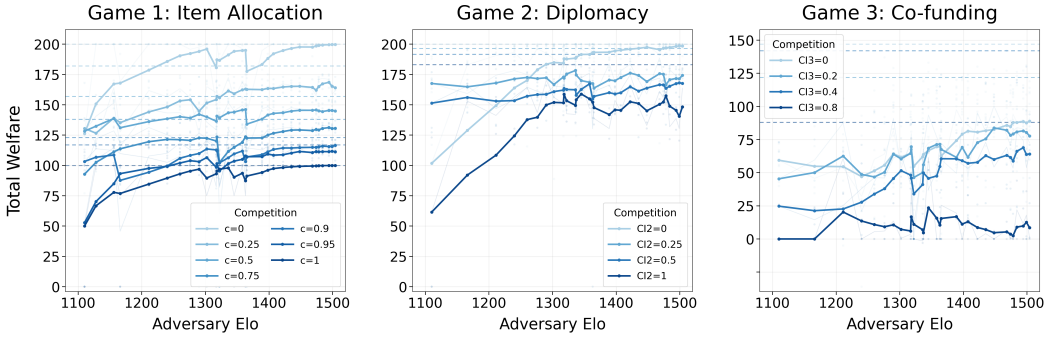

Figure 2: **Realized social welfare by competition stratum.** Solid marked curves are Elo-ordered moving averages of model-level means; dashed lines mark the maximum attainable welfare.

and $+7.37$ utility per 100 Elo for Games 1–3, respectively (Table 3). The Llama 3.3 70B replication preserves the positive adversary-utility slope, indicating that the trend is not specific to GPT-5-nano as the fixed opponent.

Capability is not merely making both parties better off. Higher Elo raises realized welfare toward the stratum optimum, especially in cooperative cells, but the welfare gain is not evenly shared. The welfare curves in Figure 2 show that stronger adversaries often find higher-surplus agreements, although Game 3's highest-competition cell remains bottlenecked by one-project scarcity and cost-sharing conflict. The NBS decomposition in Figure 3 shows that fair-share slopes are statistically indistinguishable from zero ($p > 0.19$), while the residual accounts for the full capability slope at $+4.98$, $+6.44$, and $+8.30$ points per 100 Elo (Table 4). Thus, stronger agents both improve feasibility and capture the surplus they help create.

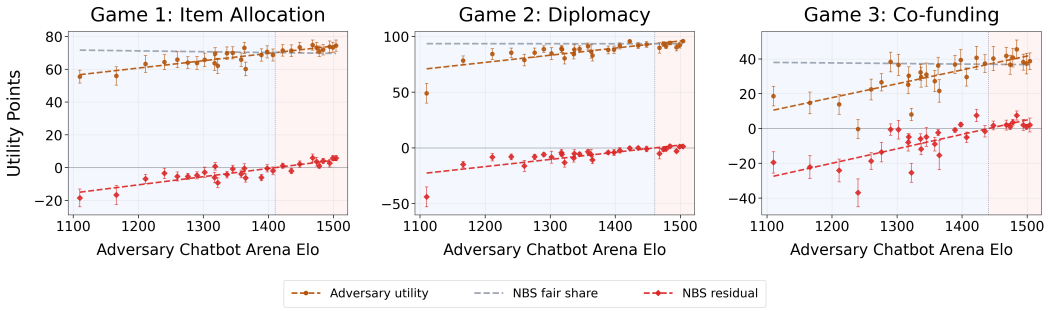

Figure 3: **Fair-share decomposition of adversary utility.** The NBS fair-share component is flat, while the residual rises sharply; above Elo 1410–1461, frontier models receive more than the fair-share benchmark.

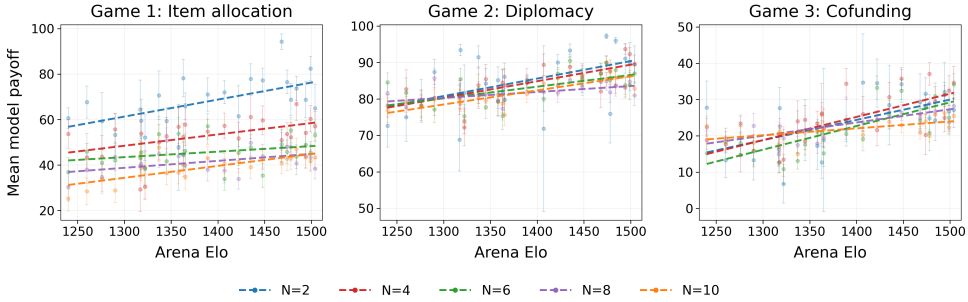

Figure 4: **Heterogeneous payoff scaling by Arena Elo and group size.** Each point averages a model's payoff over heterogeneous runs in which it appears. Capability remains predictive as group size grows.

4.2. Test-Time Compute

The native test-time-compute stress test completed 216 runs over GPT-5, Claude Sonnet 4.6, and Gemini 3 Flash at four effort levels each. Unlike model-capability scaling, requested reasoning effort does not reliably improve bargaining outcomes. Game-cell/order fixed-effect token–payoff slopes are small and statistically weak: GPT-5 $+0.71$ ($p = 0.667$), Claude $+1.02$ ($p = 0.473$), and Gemini -1.69 ($p = 0.064$) utility per 1k tokens. GPT-5 gains $+3.26$ utility from minimal to high effort on average but improves in only 6 of 18 matched situations (Table 6).

4.3. N-Player Games

The $N > 2$ experiments parse 1,430 homogeneous runs and 1,300 heterogeneous runs, totaling 16,380 agent-level observations for $N \in \{2, 4, 6, 8, 10\}$. In heterogeneous rosters, each model is scored whenever it appears in a random group. Mean payoff slopes remain positive at $+4.64$, $+3.60$, and $+4.81$ utility per 100 Elo in Games 1–3. Correlations are weaker than in bilateral play because roster composition, order, and competition cells vary, but the capability signal persists beyond dyads.

Adding weaker agents also does not reliably dilute a strong focal agent. In homogeneous-adversary runs, one non-nano model negotiates with $N - 1$ GPT-5-nano agents; payoff slopes

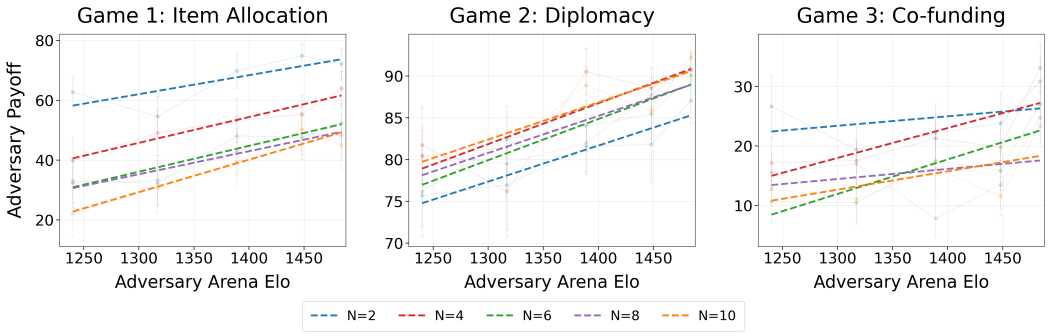

Figure 5: **Homogeneous-adversary payoff scaling for $N > 2$.** One focal adversary negotiates with $N - 1$ GPT-5-nano agents. Payoff slopes are positive for every game and group size.

are positive at every N, with average slopes of $+8.43$, $+4.59$, and $+3.44$ utility per 100 Elo in Games 1–3. From $N = 2$ to $N = 10$, the adversary-minus-fleet-mean advantage changes by $+3.03$, $+7.67$, and $+5.00$ utility, showing no consistent dilution (Section E.16). Larger groups add veto players, but they also create more opportunities for a stronger agent to identify proposals that weaker agents fail to coordinate around.

4.4. Synthesis

LMArena Elo predicts higher utility in bilateral play, inside larger GPT-5-nano fleets, and in random heterogeneous societies. The competition axis determines whether capability becomes coordination or extraction: cooperative cells raise welfare for both sides, while competitive cells let stronger agents capture above-fair surplus. Native test-time compute changes reasoning style but is not a reliable replacement for base-model capability.

5. Conclusion

We introduced three controllable negotiation environments for evaluating strategic AI agents and used them to measure how capability, group size, and preference conflict shape LLM bargaining outcomes. Higher-Elo agents often improve agreement quality, but the same capability advantage becomes above-fair surplus extraction in competitive settings and remains visible beyond two-player games. This makes model selection strategically consequential: deploying a more capable agent is not automatically Pareto-safe for weaker counterparts.

The main limitation is external validity. Our games abstract from live markets, legal negotiations, and human-agent interaction, and LMArena Elo is only a public proxy for capability. The benchmark is nevertheless useful precisely because it exposes mechanism-level variation: rivalrous goods, continuous compromise, and threshold public goods fail in different ways. Future work should extend these evaluations to richer institutions, adaptive opponents, and mechanism designs that make capability gains less extractive.

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 83548–83599. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/984dd3db213db2d1454a163b65b84d08-Paper-Datasets_and_Benchmarks_Track.pdf.

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 2025. doi: 10.1038/s41562-025-02172-y.

[3] 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.

[4] 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.

[5] 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.

[6] Leo Benac, Jonas Raedler, Zilin Ma, and Finale Doshi-Velez. A benchmark for multi-party negotiation games from real negotiation data. *arXiv preprint arXiv:2603.14066*, 2026. URL https://arxiv.org/abs/2603.14066.

[7] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NEGOTIATIONARENA platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[8] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), 2023. doi: 10.1073/pnas.2218523120.

[9] 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.

[10] Alan Chan, Maxime Riché, and Jesse Clifton. Towards the scalable evaluation of cooperativeness in language models, 2023. URL https://arxiv.org/abs/2303.13360.

[11] Ching-Lueh Chang. On the computational power of players in two-person strategic games. Master's thesis, National Taiwan University, 2006. URL http://ntur.lib.ntu.edu.tw//handle/246246/53652.

[12] Ainesh Chatterjee, Samuel Miller, and Nithin Parepally. AgreeMate: Teaching LLMs to haggle, 2024. URL http://arxiv.org/abs/2412.18690. version: 1.

[13] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120. URL https://doi.org/10.1073/pnas.2316205120.

[14] 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.

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

[16] 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.

[17] Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations, 2024. URL https://arxiv.org/abs/2401.04536. _eprint: 2401.04536.

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

[19] Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark. Scaling laws for scalable oversight, 2025. URL http://arxiv.org/abs/2504.18530.

[20] 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.

[21] Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. Strategic reasoning with language models. *arXiv:2305.19165*, May 2023. doi: 10.48550/ARXIV.2305.19165.

[22] Joseph Y. Halpern and Rafael Pass. Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory*, 156:246–268, March 2015. ISSN 00220531. doi: 10.1016/j.jet.2014.04.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0022053114000611.

[23] 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.

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

[25] 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.

[26] John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report 31122, National Bureau of Economic Research, 2023. URL https://www.nber.org/papers/w31122.

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

[28] 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.

[29] 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. *arXiv:2407.04622*, July 2024. doi: 10.48550/ARXIV.2407.04622.

[30] Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.310. URL https://aclanthology.org/2024.findings-emnlp.310/.

[31] 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.

[32] Xiangyu Liu, Di Wang, Zhe Feng, and Aranyak Mehta. Scaling inference-time computation via opponent simulation: Enabling online strategic adaptation in repeated negotiation. *arXiv preprint arXiv:2602.19309*, 2026. URL https://arxiv.org/abs/2602.19309.

[33] Olivia Macmillan-Scott and Mirco Musolesi. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6), 2024. doi: 10.1098/rsos.240255.

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

[35] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*, 2023. doi: 10.48550/ARXIV.2310.08901. URL https://openreview.net/forum?id=WnR5BCX8GS.

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

[37] Oren Neumann and Claudius Gros. Scaling Laws for a Multi-Agent Reinforcement Learning Model. In *Proceedings of the Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=ZrEbzL9eQ3W.

[38] Sean Noh and Ho-Chun Herbert Chang. LLMs with personalities in multi-issue negotiation games, 2024. URL http://arxiv.org/abs/2405.05248. version: 2.

[39] Thomas Orton. Modeling Precomputation In Games Played Under Computational Constraints. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2005–2011, Montreal, Canada, August 2021. International Joint Conferences on Artificial Intelligence Organization. ISBN 9780999241196. doi: 10.24963/ijcai.2021/276. URL https://www.ijcai.org/proceedings/2021/276.

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

[41] Christos H. Papadimitriou and Mihalis Yannakakis. On complexity as bounded rationality (extended abstract). In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pages 726–733. ACM, 1994. doi: 10.1145/195058.195445. URL https://doi.org/10.1145/195058.195445.

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

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

[44] Oliver Sourbut, Lewis Hammond, and Harriet Wood. Cooperation and Control in Delegation Games. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 229–237, Jeju, South Korea, August 2024. International Joint Conferences on Artificial Intelligence Organization. ISBN 9781956792041. doi: 10.24963/ijcai.2024/26. URL https://www.ijcai.org/proceedings/2024/26.

[45] 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.

[46] Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon M. Stewart, Jonathan K. Kummerfeld, Denis Peskoff, and Jordan Boyd-Graber. More victories, less cooperation: Assessing Cicero's Diplomacy play. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[47] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv:2404.01230*, April 2024. doi: 10.48550/ARXIV.2404.01230.

[48] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

[49] 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.

[50] 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. *arXiv:2506.00073*, May 2025. doi: 10.48550/ARXIV.2506.00073.

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 Sections 3.1 and 3.2. 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. 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.
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 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 $m = \lfloor 2.5N \rfloor$ issues, $\theta \in \{0.2, 0.8\}$, and $\rho \in \{0.9, \rho_{\min}(N)\}$, where $\rho_{\min}(N)$ is the feasible negative equicorrelation lower bound after the Gaussian-copula transform. Game 3 sets $m = \lfloor 2.5N \rfloor$ projects, $\sigma \in \{0.2, 0.5\}$, and $\alpha \in \{0.2, 0.8\}$. 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 24-model pool (filtered for all models to have a context length of at least 128K tokens) with equal-width Elo-standard-deviation strata.

C.2. Model Rosters

Table 2 lists the 30 adversary models used in the bilateral GPT-5-nano baseline sweeps, ordered by descending LMArena Elo. GPT-5-nano (Elo 1337) serves as the fixed baseline and is not counted among the adversaries. 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.

#	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	`qwq-32b`	Alibaba	QwQ-32B	1336
19	`gpt-4.1-nano-2025-04-14`	OpenAI	GPT-4.1 nano	1322
20	`llama-3.3-70b-instruct`	Meta	Llama 3.3 70B	1318
21	`gpt-4o-mini-2024-07-18`	OpenAI	GPT-4o mini	1317
22	`qwen2.5-72b-instruct`	Alibaba	Qwen2.5 72B	1302
23	`amazon-nova-pro-v1.0`	Amazon	Nova Pro	1290
24	`command-r-plus-08-2024`	Cohere	Command R+	1276
25	`claude-3-haiku-20240307`	Anthropic	Claude 3 Haiku	1260
26	`amazon-nova-micro-v1.0`	Amazon	Nova Micro	1240
27	`llama-3.1-8b-instruct`	Meta	Llama 3.1 8B	1211
28	`llama-3.2-3b-instruct`	Meta	Llama 3.2 3B	1166
29	`llama-3.2-1b-instruct`	Meta	Llama 3.2 1B	1110

Table 2: **Bilateral adversary roster (29 models shown).** All Elo scores are from the March 31, 2026 LMArena snapshot. The fixed baseline GPT-5-nano (Elo 1337) is not listed. Models are accessed via native provider APIs or OpenRouter.

$N \geq 2$ 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, excluding the baseline GPT-5-nano. This removes six 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), plus the baseline. The remaining 24 models span Elo 1240–1504 across eight providers.

Llama 3.3 70B baseline adversaries (10 models). The Llama baseline replication uses the following 10 adversary models, spanning Elo 1240–1504: Nova Micro (1240), Claude 3 Haiku (1260), Nova Pro (1290), GPT-4o mini (1317), DeepSeek V3 (1358), 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}.

C.3. Reproducibility, Release, and Compute Resources

The anonymous 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 coordinate, the exact grids in Sections C.1 and 3.2, 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.4. 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.

- **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. This supports additional public-good metrics:

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

$$F_{ij} = \frac{v_j^i / \sum_k v_j^k}{x_{ij} / \sum_k x_{kj}}, \tag{1}$$

  with $F_{ij} > 1$ indicating that the agent receives a larger value share than cost share.
- **Underfunding rate:** the fraction of positive-surplus projects that remain unfunded despite being feasible under some budget reallocation.
- **Commitment and adaptation checks:** transcript-extracted contribution commitments can be compared with subsequent pledge vectors, and round-to-round pledge changes give an adaptation score.

C.5. 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.

For multi-agent comparisons we use

$$\tilde{U}_i = \frac{NU_i}{SW^\star}, \tag{2}$$

with $SW^\star$ recomputed for each generated instance. This metric has a direct equal-share interpretation: $\tilde{U}_i = 1$ means agent i received the same utility as an equal split of optimal social welfare. When

no agreement is reached, utilities are zero and $\tilde{U}_i = 0$ whenever $SW^\star > 0$. We verify that every completed multi-agent row used in the paper has $SW^\star > 0$; for no-agreement Game 1 runs that omitted preferences from the result payload, the analysis recovers the setup preferences from the saved interaction prompts before computing $SW^\star$. Negative Game 3 utilities are retained, so $\tilde{U}_i$ can be negative when an agent 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:

$$SW(\boldsymbol{a}) = \sum_k \sum_i w_k^i (1 - |p_k^i - a_k|). \tag{3}$$

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

$$\sum_{j \in S} \left(\sum_i v_j^i - c_j \right) \tag{4}$$

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. Solution Concept Benchmarks

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

$$o^{\text{NBS}} = \arg\max_o \prod_{i=1}^{N} u^i(o). \tag{5}$$

For the bilateral exploitation plots, Game 1 enumerates allocations and Game 2 solves the continuous problem with L-BFGS-B [9, 49]. The reported exploitation index is

$$E_i = \frac{u^i_{\text{actual}} - u^i_{\text{NBS}}}{|u^i_{\text{NBS}}|}. \tag{6}$$

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

$$x_{ij}^{\text{Lindahl}} = c_j \cdot \frac{v_j^i}{\sum_k v_j^k}. \tag{7}$$

The corresponding exploitation index is

$$E_i = \frac{u^i_{\text{actual}} - u^i_{\text{Lindahl}}}{|u^i_{\text{Lindahl}}|}. \tag{8}$$

Appendix E. Additional Result Tables

E.1. Bilateral Payoff Slopes

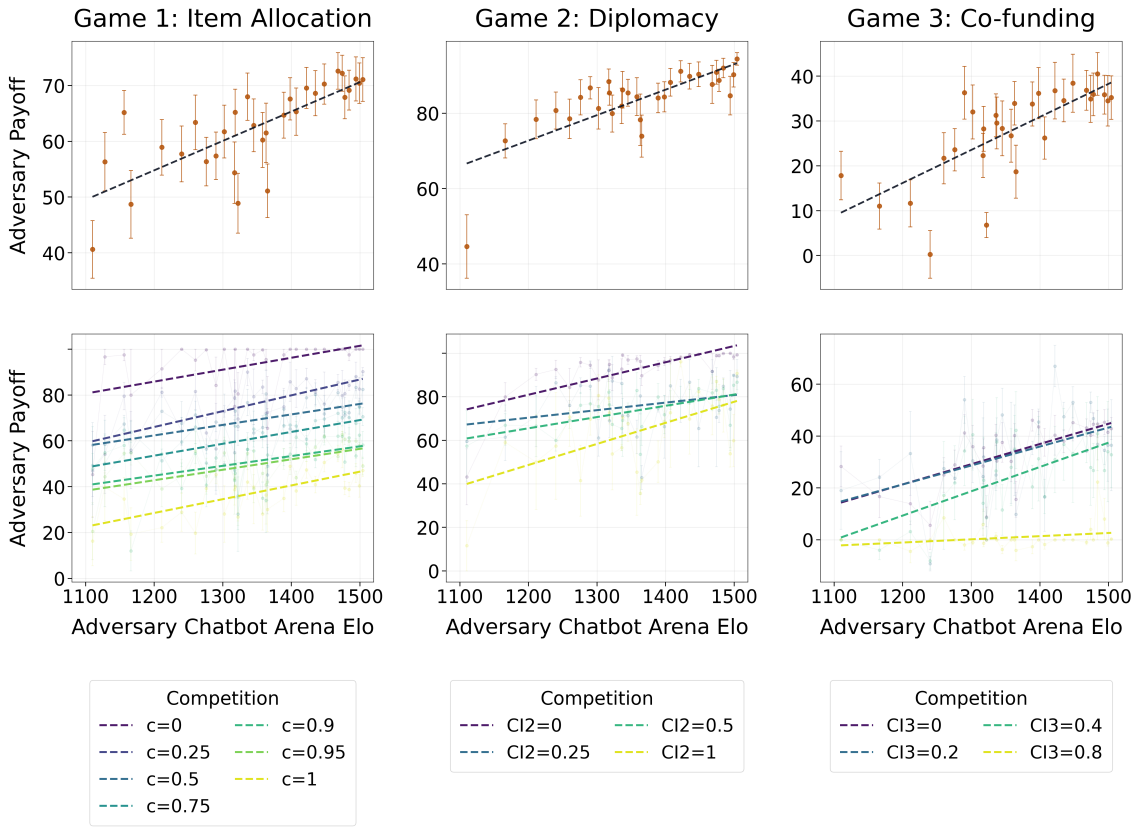

Figure 6: **Bilateral adversary payoff by game and competition.** Top row: overall adversary payoff against GPT-5-nano by Elo. Bottom row: the same relationship stratified by competition settings with per-stratum linear fits.

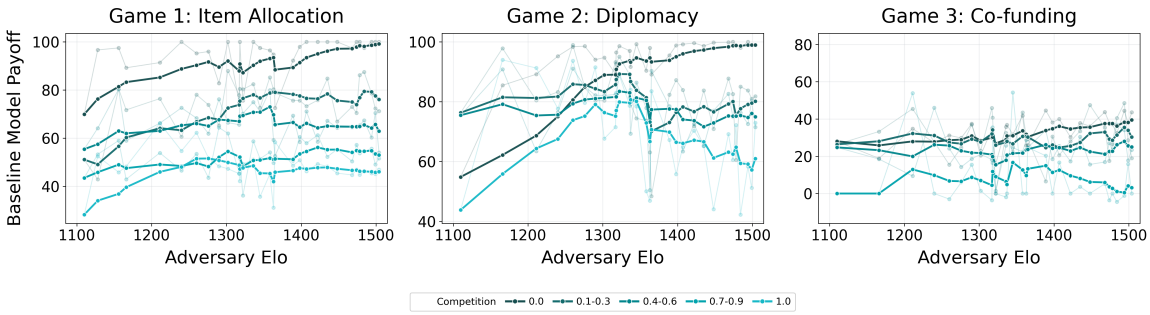

Figure 7: **GPT-5-nano baseline payoff against stronger adversaries, by competition.** Faint curves show raw per-Elo means; bold teal curves show Elo-ordered exponentially weighted moving averages within each competition stratum.

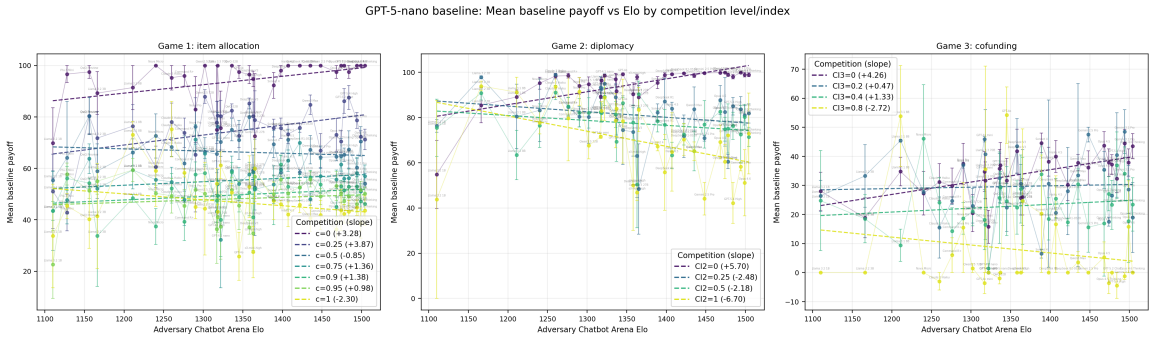

Figure 8: **Fitted-line version of GPT-5-nano baseline payoff by competition.** This is the earlier linear-fit view of the main-text baseline-payoff figure, with the baseline payoff plotted against adversary Elo separately within each competition stratum.

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. $N = 2$ Social-Welfare Optimality Detail

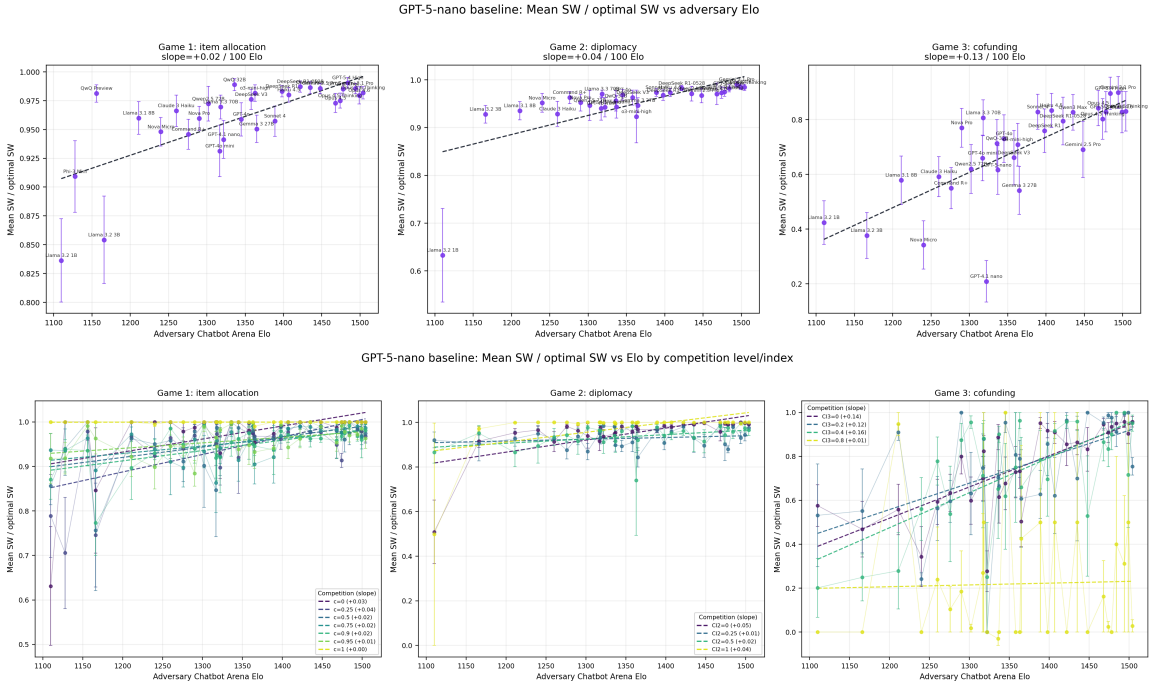

Figure 9: **Social-welfare optimality ratio in bilateral play.** Top: optimality ratio by adversary Elo. Bottom: the same metric stratified by competition. Stronger adversaries capture a larger share of the realized instance optimum, but Game 3's highest-competition cell remains bottlenecked by one-project scarcity and cost-sharing conflict.

E.3. Bilateral Social-Welfare Detail

Figure 9 shows realized social-welfare optimality 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.

E.4. Bilateral Fairness Plots

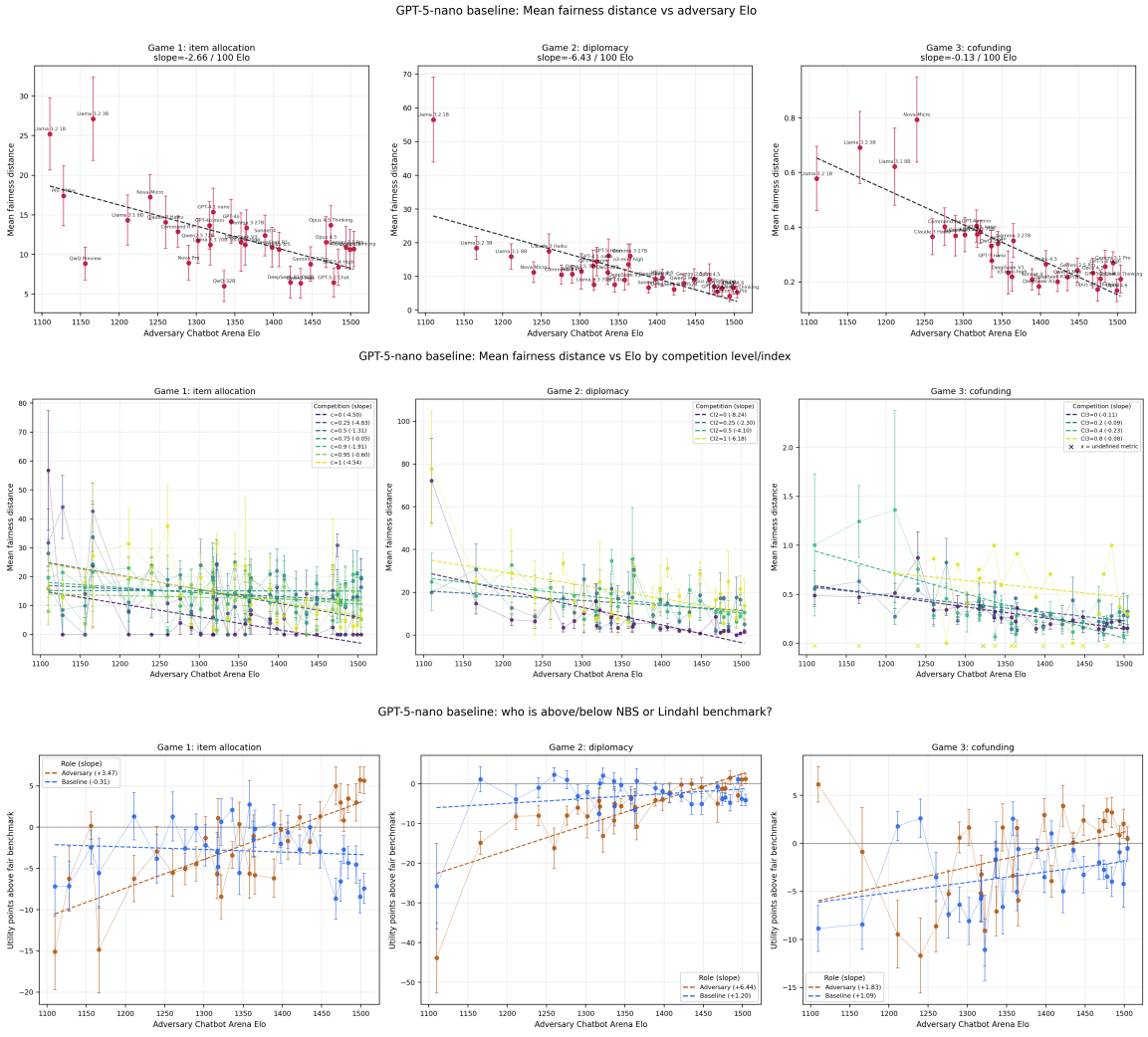

Figure 10: **Fairness and benchmark-relative extraction.** Stronger adversaries reduce distance from the NBS/Lindahl reference, especially in Games 1–2, but the role-specific residuals show that they sit further above their own benchmark share as Elo rises.

E.5. 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 11 gives the three overall plots, and Figure 12 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.

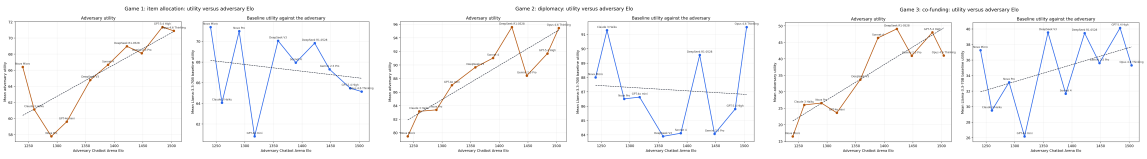

Figure 11: **Llama 3.3 70B baseline utility replication.** Left to right: item allocation, diplomatic treaty, and co-funding. Each panel varies the adversary model and holds the Llama baseline fixed.

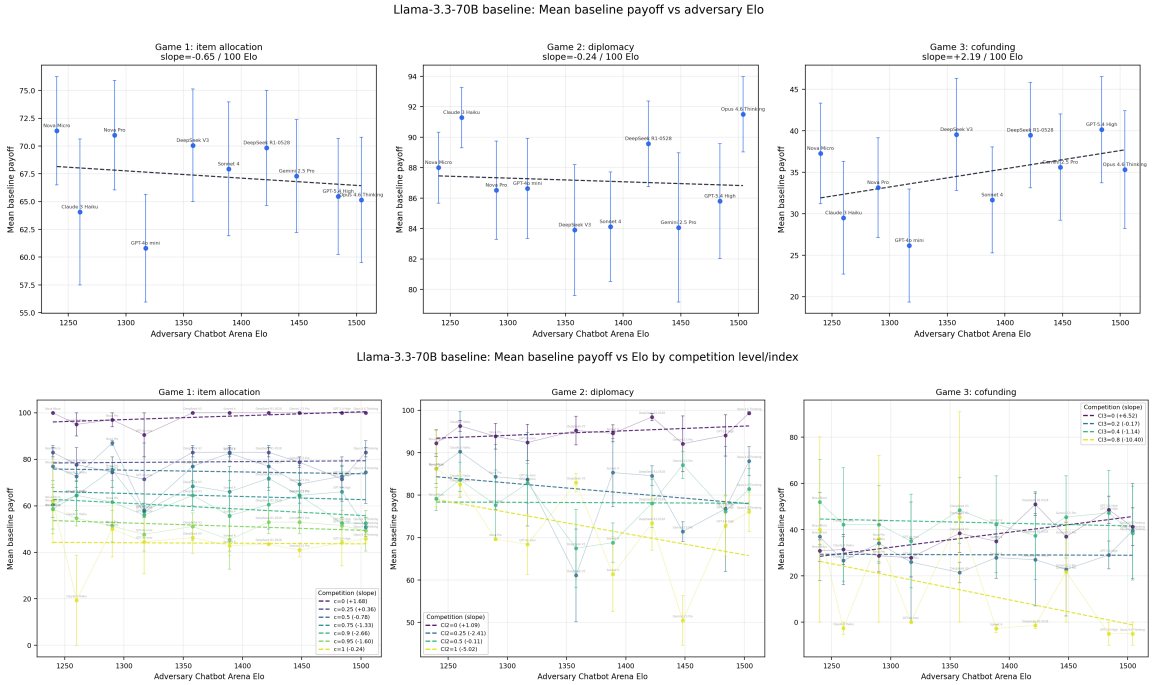

Figure 12: **Llama 3.3 baseline payoff against stronger adversaries.** Top: the three game-level Llama baseline payoff trends. Bottom: the same trends stratified by competition.

E.6. Multi-Agent Completion and Reliability

Non-finished runs were concentrated in heterogeneous Game 3 high-N tail cells and in Game 1 proposal parsing at larger N. All reported results use the final completed-run data; intermediate completion snapshots are retained in the artifact for failure-mode interpretation only.

E.7. Normalized Payoff, Rounds, and Discounting

The supplemental artifact exports run-level and agent-level CSVs with $SW^{\star}$, normalized utilities, and per-cell aggregations. Bootstrap intervals for small homogeneous-control cells resample completed agent observations within a game–N–competition cell and report percentile 95% intervals for raw utility and $\tilde{U}_i$. These intervals are descriptive because some cells contain only two completed run replicates.

The rounds analysis compares early $N \in \{2, 4\}$ against later $N \in \{6, 8, 10\}$ and also fits

$$\text{final\_round} \sim N + \text{competition} + \text{game} + \text{family}. \tag{9}$$

The pooled coefficient on N is -0.020 rounds per added agent with $p = 0.094$. The strongest positive late-N effect is homogeneous-control co-funding, where later N values take +2.56 rounds relative to $N = 2, 4$ with a bootstrap 95% interval [0.44, 4.54].

For agreed runs, the undiscounted utility check uses

$$U_i^{\text{undisc}} = \frac{U_i}{0.9^{t-1}}, \tag{10}$$

where t is the final agreement round. No-agreement utilities are not reversed because there is no meaningful agreement round. The Elo coefficient remains positive under this check: +4.68 utility per 100 Elo with discounted utilities and +4.87 after reversing discounting for agreed runs.

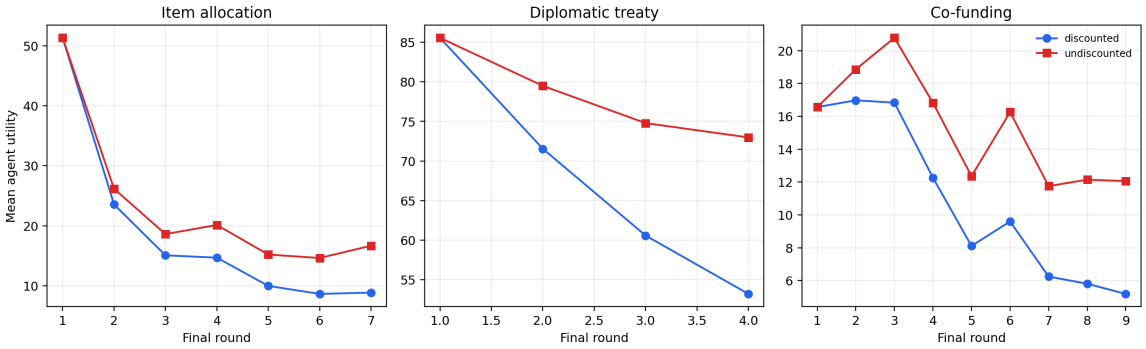

Figure 13: **Discounted and undiscounted utility by final round for $N > 2$.** Reversing the $\gamma = 0.9$ discount reduces the utility-round slope. Later agreements still have lower mean utility, indicating that hard instances take longer and settle worse.

E.8. Parser-Clean Robustness

Parser-clean robustness uses rows with successful status, strict voting-clean diagnostics, no token-limit markers, no vote-fallback markers, no synthetic vote or proposal markers, and no provider-degradation markers. This subset is most conservative in heterogeneous Game 1 and Game 3, where parser and provider failures are concentrated. The focal homogeneous-adversary slopes remain similar in parser-clean rows for Games 1 and 2: +10.30 and +4.15 utility per 100 Elo. Heterogeneous parser-clean slopes are positive in all three games (+5.74, +5.50, +3.01), though the Game 3 parser-clean subset is small.

Figure 14 and Figure 15 show the group-size diagnostics behind the main-paper caution about inequality. The Gini relationship remains weak overall, while the payoff-slope panels show that capability advantage varies substantially with N and game mechanics.

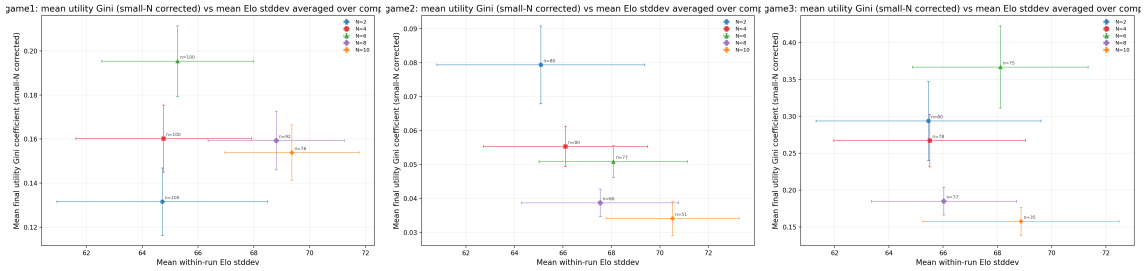

Figure 14: **Heterogeneous utility inequality by group size.** Each panel plots mean small-N-corrected utility Gini against within-run Elo standard deviation, averaging over competition settings. Left to right: item allocation, diplomatic treaty, and co-funding.

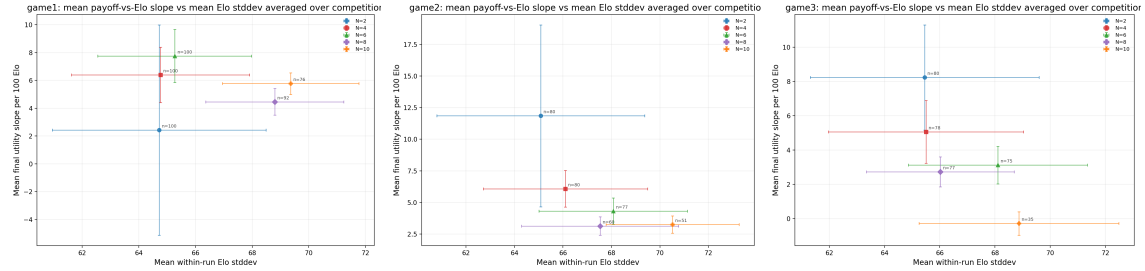

Figure 15: **Per-run payoff-vs-Elo slopes by group size in heterogeneous runs.** Each panel plots the signed OLS slope of final utility on Elo against within-run Elo standard deviation, averaging over competition settings.

E.9. 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

Table 6: **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

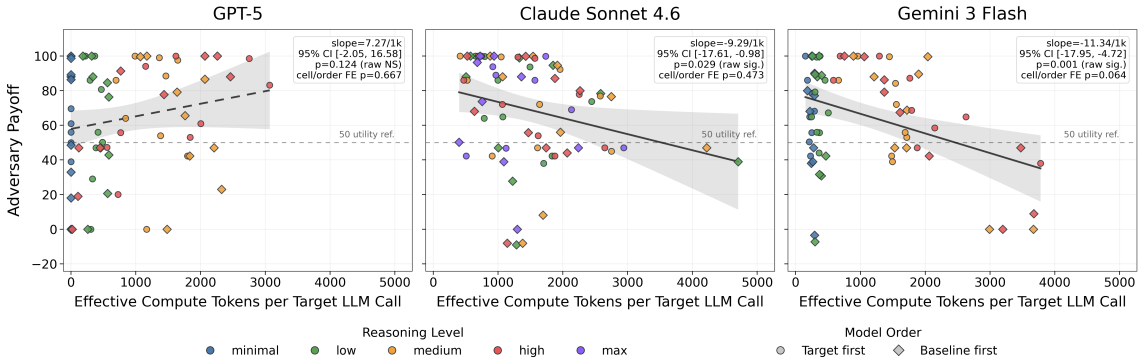

Figure 16: **Token–payoff scatter for TTC families.** Colors denote requested reasoning level; markers denote model order. Black lines are raw OLS fits with 95% confidence bands; dashed fits have $p \geq 0.05$.

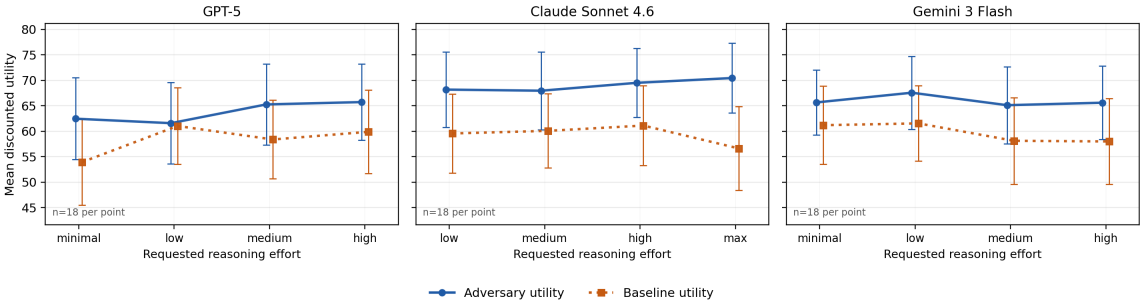

Figure 17: **Requested reasoning effort does not produce a monotone payoff curve.** Each panel uses the corresponding provider family's ordered effort labels on the x-axis; these are ordinal requested settings, not equal token budgets across providers. Solid lines show adversary utility, dotted lines show the GPT-5-nano baseline utility, points are means over 18 matched runs, and error bars are SEM.

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.

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

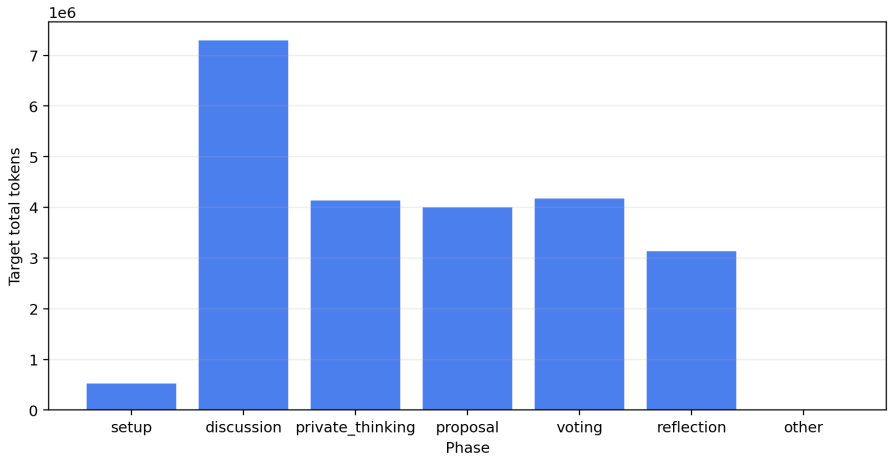

Figure 18: **TTC target tokens by phase.** The stage-token audit groups target calls into setup, discussion, private thinking, proposal, voting, reflection, and other phases. Most target tokens are spent after setup, especially in discussion and decision phases.

E.10. $N = 2$ 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 19 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.

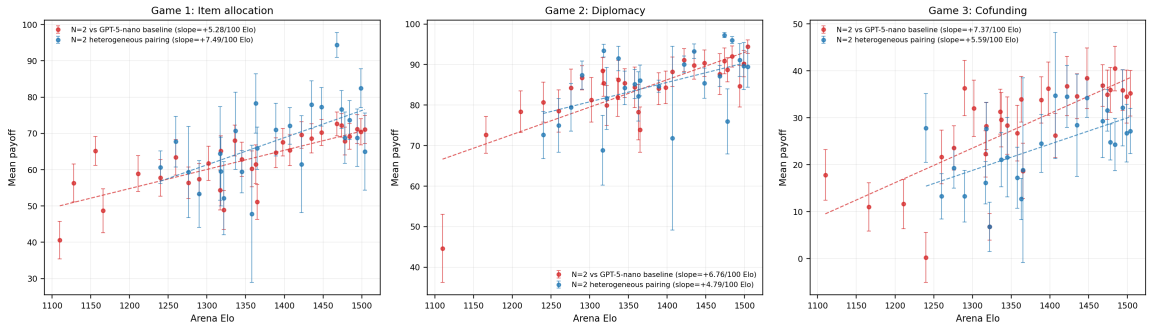

Figure 19: **Fixed baseline versus $N = 2$ heterogeneous pairings.** Capability scaling appears under both evaluation designs.

E.11. $N = 2$ Rounds to Consensus

Figure 20 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.

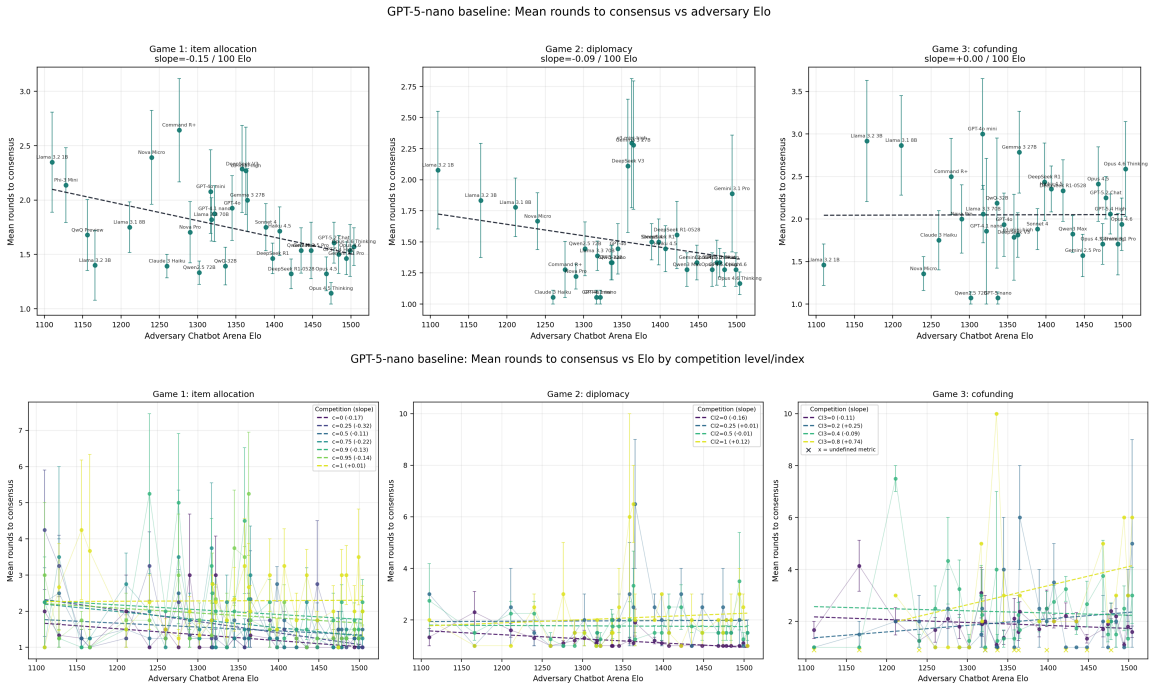

Figure 20: **Rounds to consensus in the GPT-5-nano bilateral sweep.** Top: the three game-level trends. Bottom: the same trends stratified by competition. Stronger adversaries usually do not need more rounds to earn more utility. The exception is scarce co-funding, where additional reasoning can produce better rejection of bad deals rather than faster agreement.

E.12. $N = 2$ Order Diagnostics

Figure 21 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.

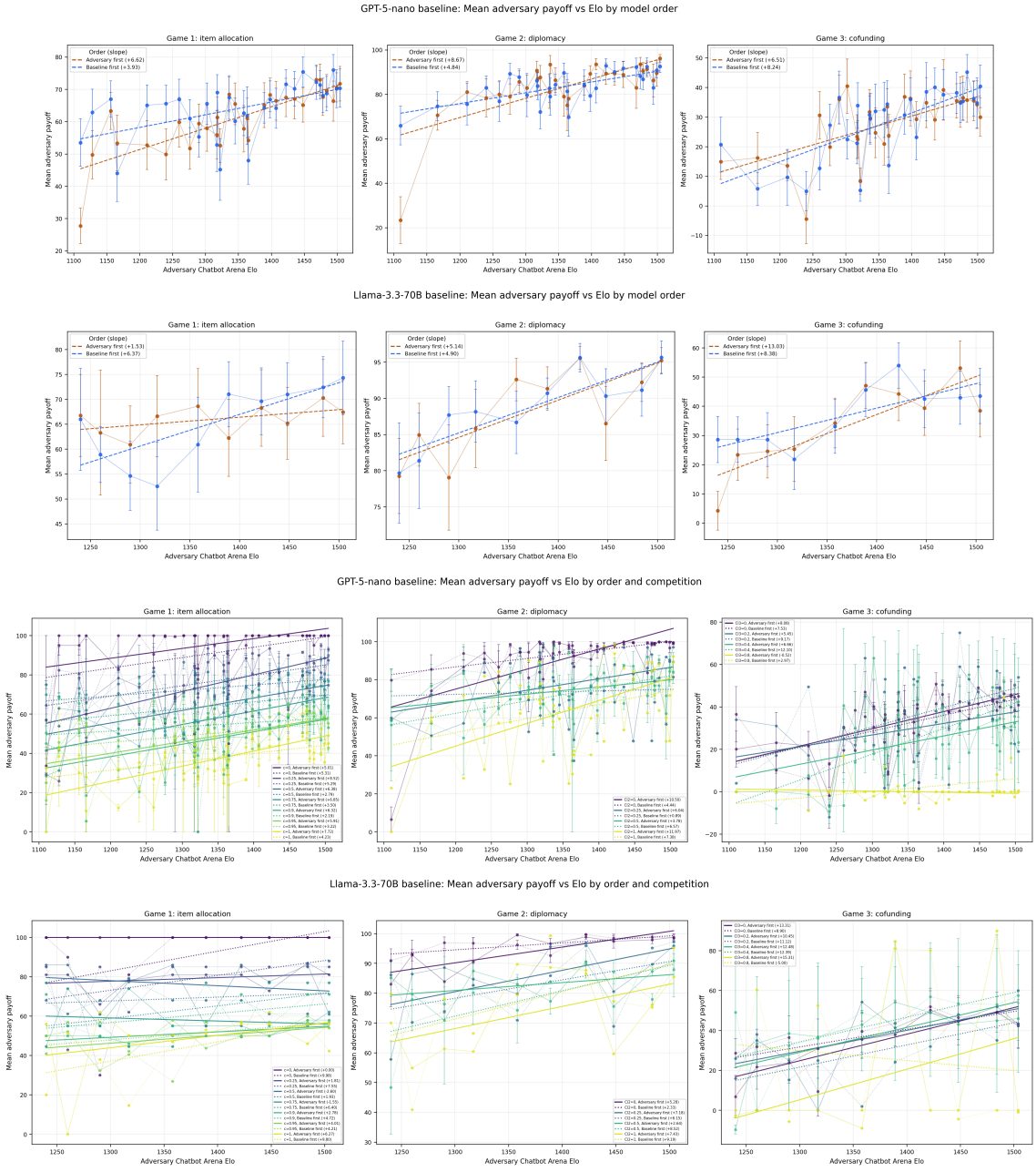

Figure 21: **Bilateral order diagnostics.** Rows 1–2 show adversary-payoff order trends for the GPT-5-nano and Llama 3.3 baselines. Rows 3–4 show the corresponding competition-stratified order trends.

E.13. Multi-Agent Competition Breakdowns

Figure 22 and Figure 23 give the competition-stratified versions of the main $N > 2$ 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.

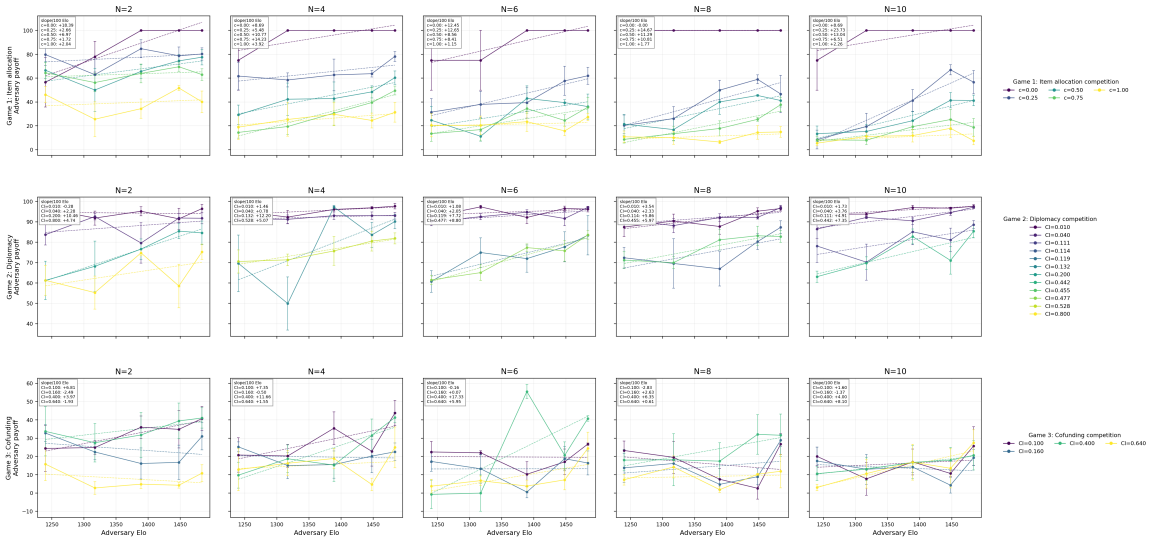

Figure 22: **Homogeneous-adversary payoff by competition.** Rows show item allocation, diplomatic treaty, and co-funding; columns show $N = 2, 4, 6, 8, 10$.

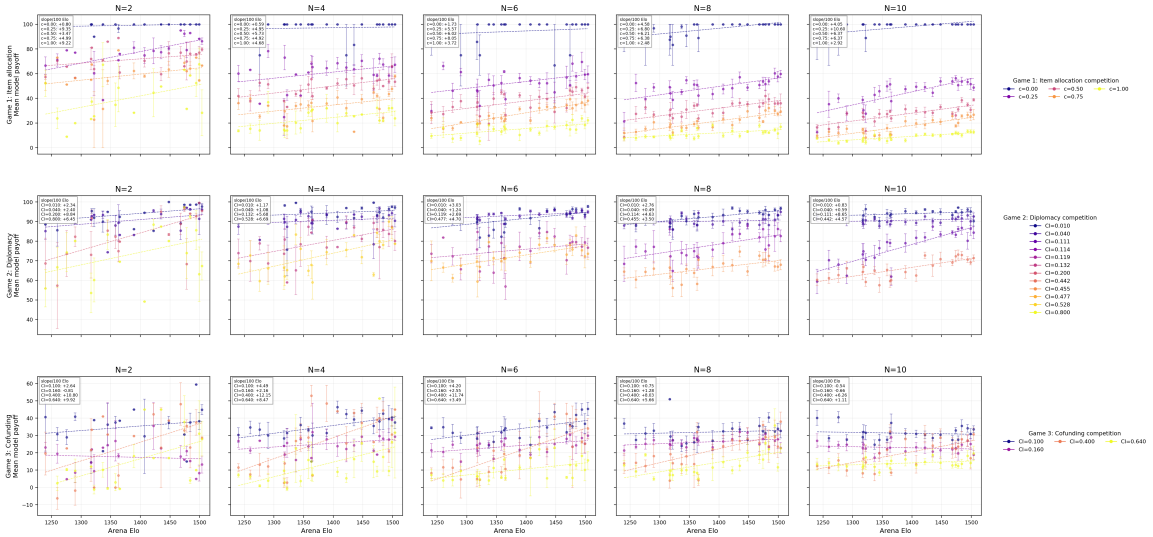

Figure 23: **Heterogeneous payoff by Arena Elo and competition.** Rows show item allocation, diplomatic treaty, and co-funding; columns show $N = 2, 4, 6, 8, 10$.

E.14. Heterogeneous Payoff Scaling

The heterogeneous payoff-vs-Elo figure is now Figure 4 in the main text. Competition-stratified breakdowns appear in Figure 23 above.

E.15. 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.

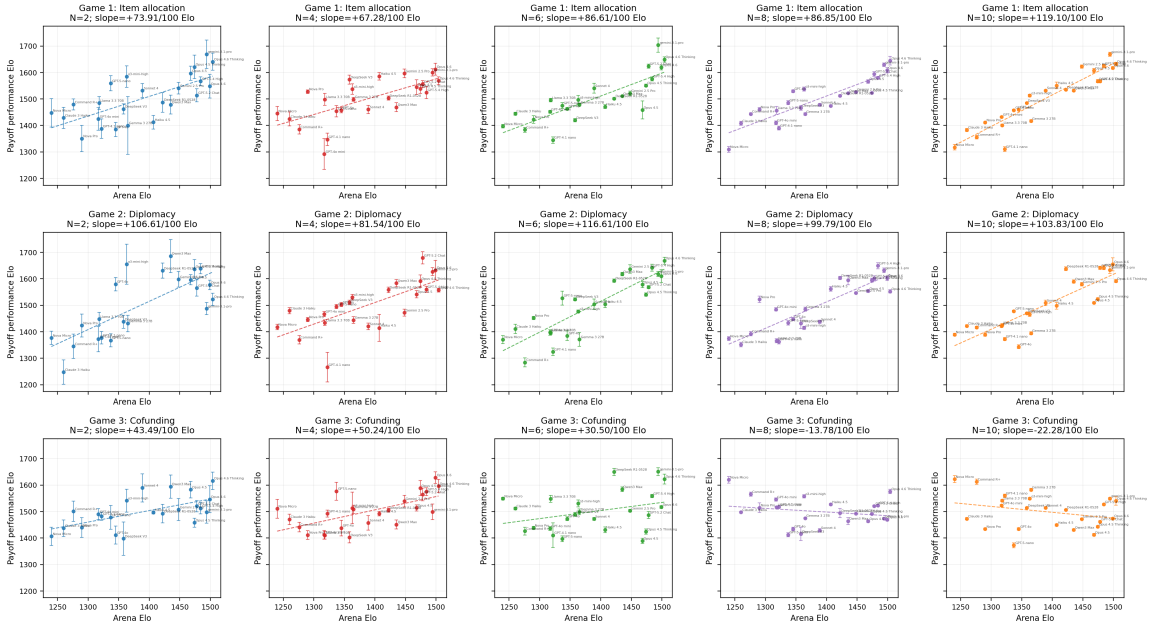

Figure 24: **Heterogeneous payoff-performance Elo.** Performance Elo is inferred from within-run utility rankings in heterogeneous rosters. It usually increases with Arena Elo, but Game 3 reverses at $N = 8$ and $N = 10$, where public-good spillovers and ties weaken the link between absolute payoff and beating one's roster-mates.

E.16. Multi-Agent Dilution Diagnostics

Figure 25 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.

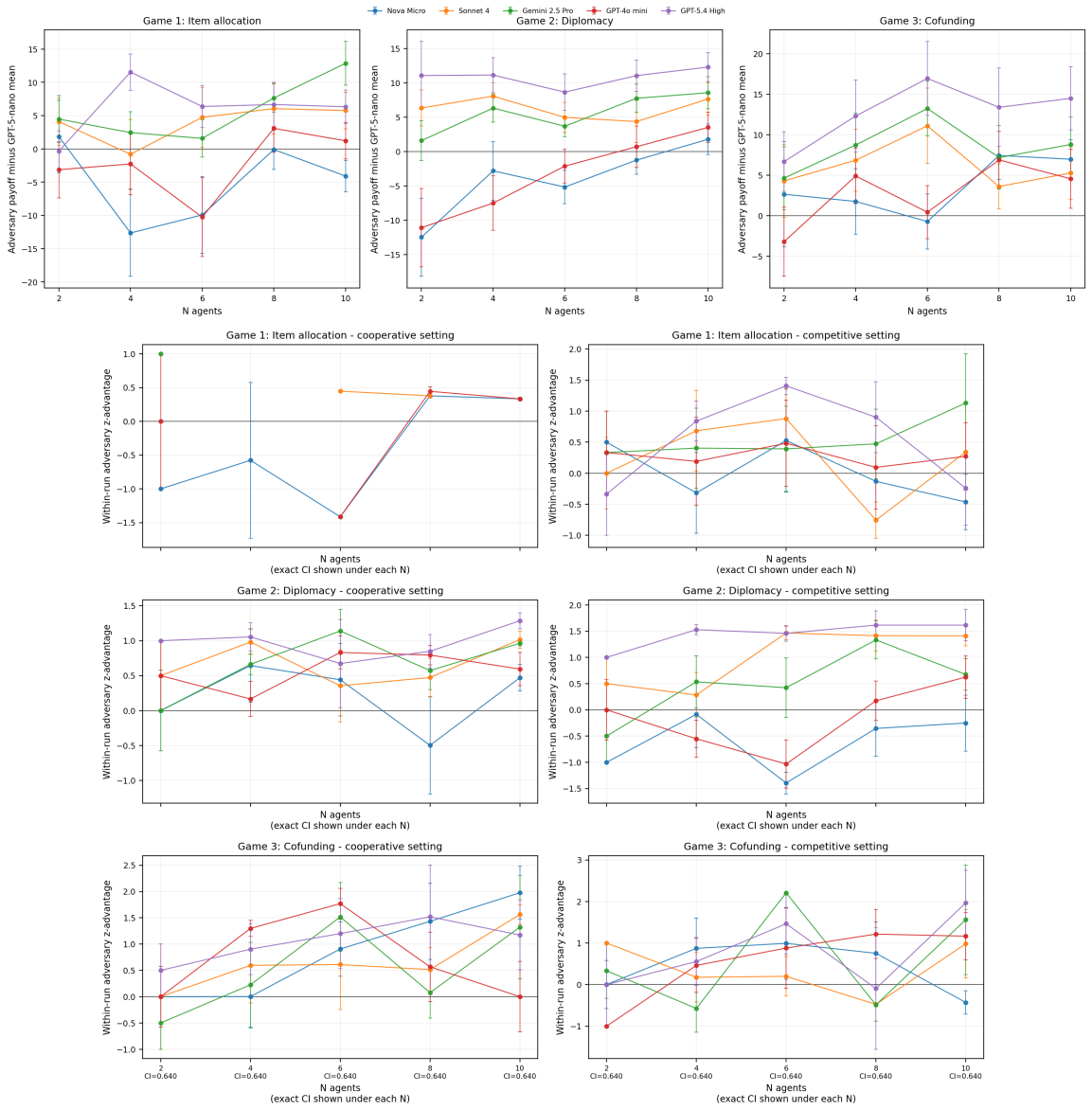

Figure 25: **Does a larger baseline fleet dilute a focal adversary?** Top: adversary payoff advantage over the GPT-5-nano fleet mean. Bottom: within-run z-advantage by competition band.

E.17. Multi-Agent Inequality by Group Size

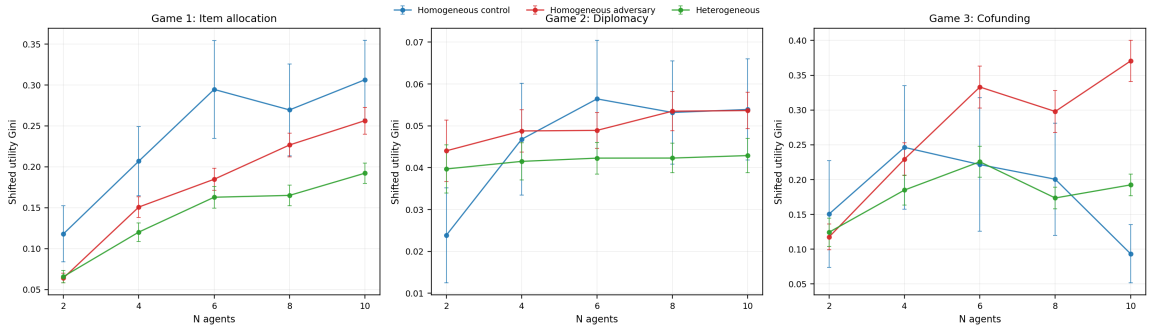

Figure 26: **Payoff inequality grows with group size in item allocation but not in treaty or public-good games.** Shifted utility Gini by N for homogeneous-control, homogeneous-adversary, and heterogeneous conditions. Game 1 shows the sharpest increase; Game 2 stays stable; Game 3 is dominated by feasibility rather than distributional competition.

E.18. Multi-Agent Fairness, Inequality, and Efficiency

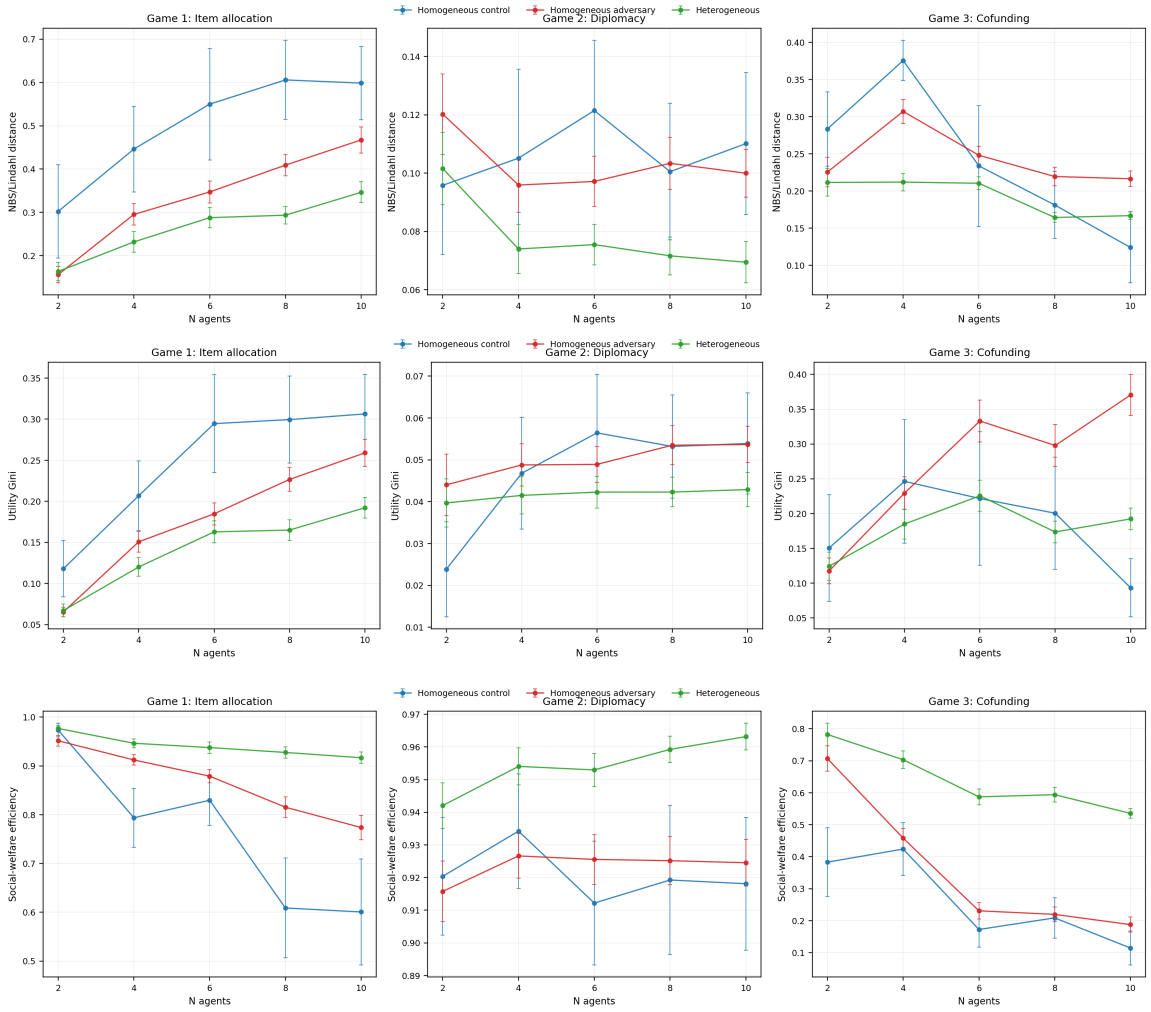

Figure 27: **Fairness, inequality, and welfare efficiency as group size increases.** Game 1 shows the clearest fairness degradation with N. Game 2 remains stable because continuous treaty issues allow compromise. Game 3 is dominated by public-good feasibility, cost shares, and no-consensus/all-zero outcomes.

E.19. Multi-Agent Fairness Details

Figure 28 shows agent-level residuals relative to the NBS or Lindahl benchmark. In Games 1–2, the residual rises with Elo in heterogeneous runs, indicating that higher-Elo agents tend to land above their benchmark share even as agreements become more efficient. Game 3 is flatter because the Lindahl residual is dominated by which projects are funded and who pays for them.

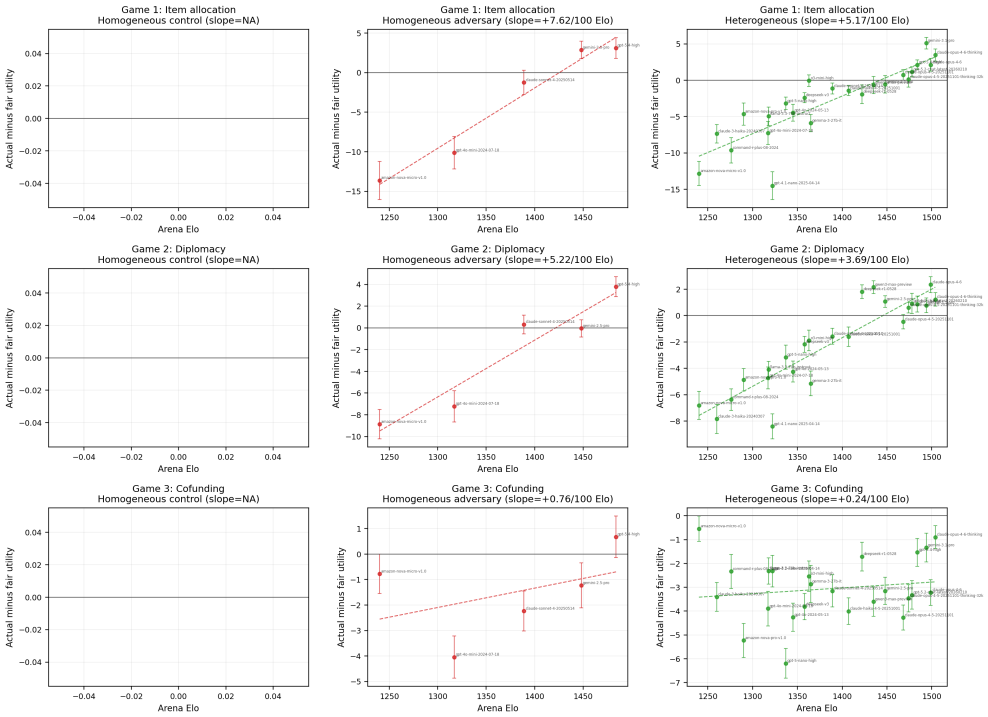

Figure 28: **Agent-level fair-benchmark residuals.** Positive residuals mean the agent receives more than its NBS share in Games 1–2 or more than the Lindahl-style benchmark in Game 3.

Figure 29 shows the competition-band version of the heterogeneous performance-Elo analysis. The Game 3 reversal in high-N all-band plots is a metric warning, not a strong causal claim that Arena Elo harms public-good play: the fits are weak and the ordinal metric is sensitive to ties, all-zero outcomes, and free-riding on funded projects.

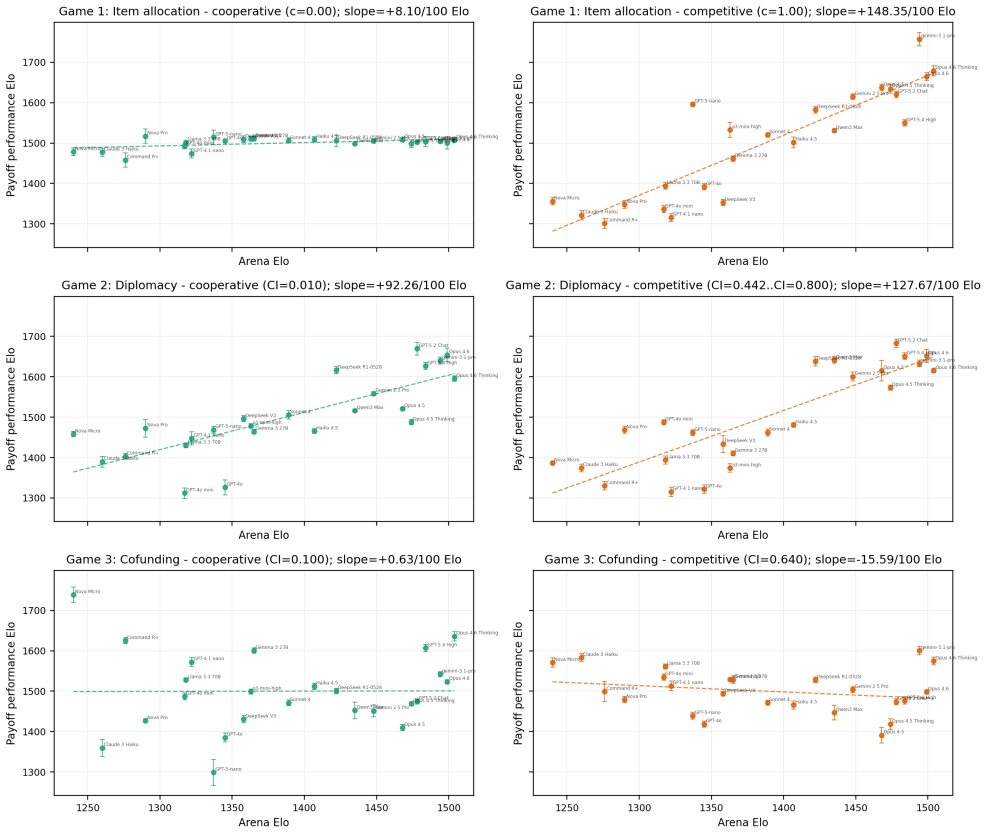

Figure 29: **Heterogeneous payoff-performance Elo by competition band.** Performance Elo is fitted from within-run payoff orderings and centered within each subset.

E.20. Elo Dispersion and Inequality

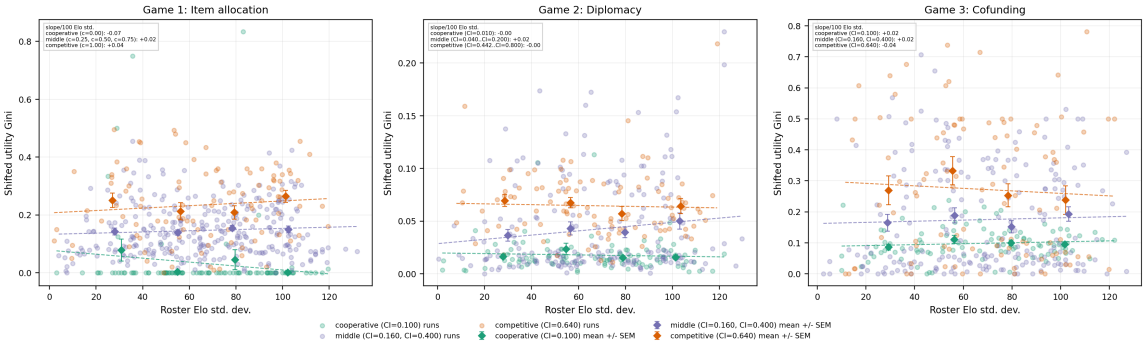

Figure 30: **Within-run Elo dispersion weakly predicts utility Gini.** Larger capability spread does not mechanically imply larger payoff inequality. Game geometry and the implemented proposal dominate the inequality signal.

E.21. 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. Prompt Templates

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

F.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 (Game 3 only):** If the joint proposal is rejected, agents see the counterfactual outcome and are asked to consider adjustments. Games 1 and 2 proceed directly to the next round's discussion.

F.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.

F.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.