The Power of Stories: Narrative Priming in Networked Multi-Agent LLM Interactions

Gerrit Großmann[1], Larisa Ivanova[1,3], Sai Leela Poduru[1,2], Mohaddeseh Tabrizian[1,2], Islam Mesabah[1], David A. Selby[1], and Sebastian J. Vollmer[1,2]

[1] German Research Center for Artificial Intelligence (DFKI)
[2] Department of Computer Science, University of Kaiserslautern–Landau (RPTU)
[3] Department of Language Science and Technology, Saarland University
{larisa.ivanova}@dfki.de

Abstract. Research suggests that large-scale human cooperation is driven by shared narratives that encode common beliefs and values. This study explores whether such narratives can similarly nudge LLM agents toward collaboration.
Therefore we let LLM agents play a (networked) finitely repeated public goods game after being primed with different stories.
Our experiments address four questions: (1) How do narratives influence negotiation behavior? (2) What differs when agents share the same story versus different ones? (3) What happens when the agent numbers grow? (4) Are agents resilient against self-serving participants?
We find that story-based priming significantly affects collaboration. *Common* stories improve collaboration and benefit all participants, while *different* story priming reverses this effect, favoring self-interested agents. These patterns persist across network sizes and structures.
These findings have implications for multi-agent coordination and AI alignment.
Code is available at `github.com/storyagents25/story-agents`.

Keywords: LLM Agents · Narrative Priming · Collaboration and Competition · Cooperation in Networks

1 Introduction

From ancient creation myths uniting scattered tribes to modern national narratives binding millions of strangers, shared stories are humanity's most powerful technology for large-scale cooperation [13,14,5]. These collective narratives enable coordination actions across vast networks of strangers who would otherwise lack mutual trust. As our world becomes increasingly populated by (multimodal) LLM agents deployed in complex multi-agent environments that require cooperation and competition [21,31,34], a critical question emerges: can we apply humanity's most successful cooperation mechanism to artificial agents?

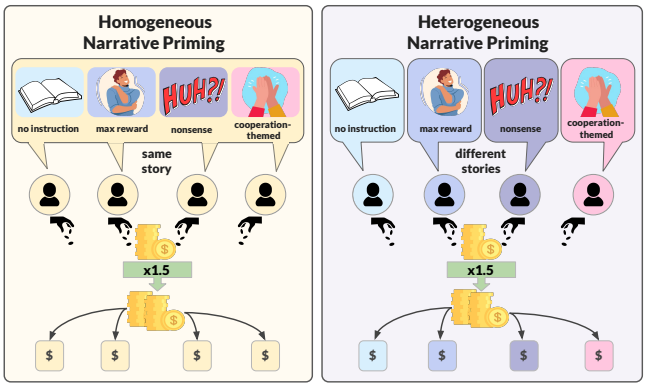


Fig. 1: Repeated public goods game with narrative priming. *Homogeneous*: all agents receive identical story prompts. *Heterogeneous*: each agent receives different narrative priming, creating mixed behavioral contexts within the same game.

While prior work has assessed LLM agents cooperation [1,4] and LLM alignment [8,20], the potential for shared narratives to enhance collaboration in agent networks remains largely unexplored. Drawing inspiration from the role of stories in human cooperation, we investigate whether narrative priming can effectively promote collaborative behavior among LLM agents.

To test this, we use the Public Goods game, a framework which creates controlled conflicts between individual and collective benefit: agents receive endowments and decide whether to contribute personal resources to shared pools that multiply and redistribute equally regardless of individual input. This creates a classic cooperation dilemma where collective welfare demands full contribution, while individual rationality favors free-riding [3,28,32].

Our experiments test how narrative priming affects cooperation in networked environments where agents participate in overlapping groups, mirroring realistic interconnected social relationships and cooperation decisions [28,32]. We examine system-wide collaboration and individual outcomes among LLM agents. We also test the robustness of cooperation in the presence of selfish individuals.

Our results indicate that shared narratives improve cooperation when all participants receive identical cooperation-themed story prompts, but this effect reverses in heterogeneous groups with mixed narrative priming. These patterns persist even in networked settings with multiple pools, suggesting that narrative coherence *is* important for effective cooperation.

2 Method

Our method is based on LLM agents playing together a repeated networked game of public goods (Figure 1), characterized by the following properties:

1. Collective optimality: if all agents play cooperatively, they achieve a higher individual reward;
2. Individual incentive: within any round, contributing zero tokens to any pool maximizes immediate individual payoff;
3. Iterative adaptation: if other agents play selfishly (or cooperatively), an agent may be motivated to do the same.

We implement two complimentary variants: **Single-Pool** experiments use one shared pool to test scaling effects across group sizes and robustness to defection, while **Multi-Pool** experiment introduces strategic complexity through overlapping pools.

In both paradigms, we manipulate behavioral homogeneity through *narrative priming* (see Section 2.2) to examine whether story-based conditioning affects agents' cooperative strategies.

2.1 Game Procedure

In each game, we instantiate N agents with game rules, assigned narratives, and assigned pools. Each game consists of R rounds, with each round r following a fixed sequence:

1. *Endowment*: each agent i receives T tokens;
2. *Contribution*: for each of their M assigned pools p for multi-pool), agents decide contribution amounts t ($t \in \mathbb{Z}$ with $0 \leq t \leq T$);
3. *Payoff calculation*: payoffs are calculated (see subsequent paragraph) and redistributed among pool members;
4. *Feedback*: agents receive complete information about all relevant contributions and payoff breakdowns.

Data Collection and Metrics. We collect individual agent contributions per round (per pool in multi-pool experiment), round payoffs, and cumulative payoffs across all rounds within each game. Primary metrics are cumulative payoffs per agent (Equation (1)), collaboration scores (Equation (2)), and, in multi-pool experiment, global vs local pool preference ratios.

Payoff Calculation. Each pool p has member set M_p and collects total contributions $T_p = \sum_{i \in M_p} t_{i,p}$. These contributions are multiplied by a fixed factor m and redistributed equally among pool members. Agent i's total round payoff π_i consists of two components: the agent's share of the returns from all pools it participated in, and the unspent remainder of its initial endowment T:

$$\pi_i = \sum_{p:\, i \in M_p} \frac{m\, T_p}{|M_p|} + \big(T - \sum_{p} t_{i,p}\big). \tag{1}$$

Table 1: Narrative prompt properties: *lexical diversity* (vocabulary richness, from low/repetitive (0) to high/diverse (1)), *sentiment score* (emotional valence, from negative (−1) to positive (+1)).

Story Type	Prompt	Token Count	Lexical Diversity	Sentiment Score	Main Theme	Cultural Origin
Baseline	noinstruct	0	0.000	N/A	N/A	N/A
	maxreward	10	0.889	0.300	None	N/A
	nsCarrot	320	0.596	0.150	Curiosity & self-reward	Invented modern fantasy
	nsPlumber	305	0.560	0.287	Creative problem solving	Invented modern fantasy
Meaningful	OldManSons	220	0.636	-0.042	Strength through (familial) unity	European folktale
	Odyssey	322	0.677	0.150	Resilience via wisdom & alliances	Classical Greek epic
	Soup	285	0.597	0.075	Resource pooling & generosity	European folktale
	Peacemaker	256	0.640	0.404	Unity through dialogue & consensus	Iroquois (Indigenous legend)
	Musketeers	273	0.643	0.330	Unity through strategic alliance	French adventure novel
	Teamwork	309	0.598	0.060	Combining strengths to succeed	Modern illustrative parable
	Spoons	867	0.402	0.145	Mutual aid via sharing	European allegory
	Turnip	324	0.595	0.085	Every contribution matters	European folktale

Collaboration Score. We measure cooperation effectiveness as the proportion of total possible contributions actually made across all agents and rounds in a game:

$$\text{Collaboration Score} = \frac{\sum_{r=1}^{R} \sum_{i=1}^{N} t_{i,r}}{N\ R\ T}\,, \tag{2}$$

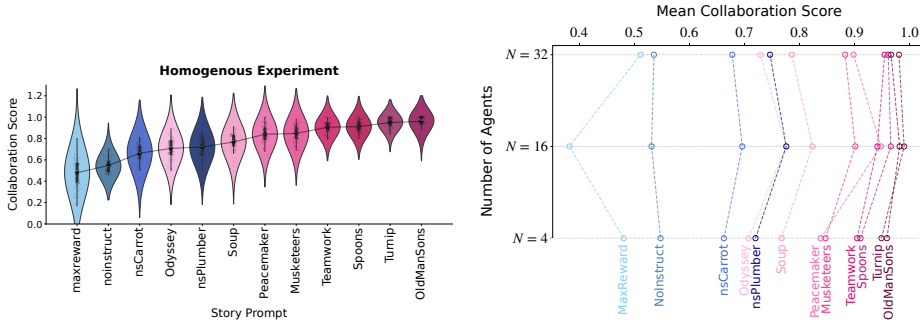
where $t_{i,r}$ is agent i's contribution in round r. A score of 1.0 indicates perfect cooperation, while lower values reflect deviations due to reduced participation or strategic choices. This metric serves as a key proxy for evaluating how effectively different narratives influence agent cooperation.

2.2 Narrative Priming

To test the effect of narrative priming on agents' collaborative behavior, we prime them with story-based behavioral context via system prompt:

> "Your behavior is influenced by the following bedtime story your mother read to you every night: [Story]"

Our story corpus (cf. Table 1, also available on GitHub) comprises 8 cooperation-themed narratives emphasizing teamwork and collective benefit, plus 4 control conditions including no instructions, explicit self-interest directives, and two nonsensical stories lacking coherent themes. Stories were selected to balance cultural diversity and summarized to retain core cooperation-themed elements while minimizing extraneous narrative details. Depending on experimental condition, agents receive either identical stories (homogeneous condition) or randomly sampled distinct stories (heterogeneous condition) from this corpus.



(a) Collaboration scores for homogeneous group ($N = 4$). Pink-shaded cooperation-themed stories get higher collaboration scores than blue-shaded controls. Gray trend line represents the mean.

(b) Scaling experiment results for homogeneous agents across different group sizes. The ranking remains relatively consistent as group size increases.

Fig. 2: Narrative priming effects on cooperation in homogeneous groups.

3 Results

Implementation. We used *meta-llama-3.1-70b-instruct-fp8*[4], and conducted experiments using varied temperature parameters (0.6, 0.8, 1.0). At higher temperatures, the priming effects show stronger differentiation, and negotiation dynamics become less pronounced. For clarity and consistency, our detailed analysis focuses on experiments run with temp = 0.6.

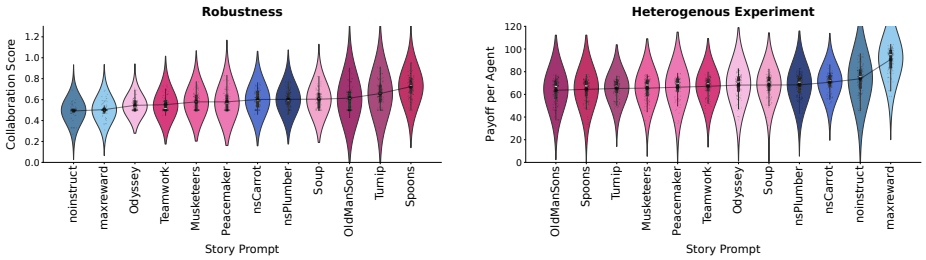
3.1 Single-Pool Cooperation

The Single-Pool experiments use variable group sizes ($N \in \{4, 16, 32\}$) playing $R = 5$ rounds with one shared pool, running 100 games per story per group size, $m = 1.5$.

Exp. 1.1: Cooperation Among Homogeneous Agents. In 4-agent groups, cooperation-themed stories ("OldManSons," "Turnip") achieve near-perfect collaboration scores, significantly outperforming baseline controls. Self-interest ("maxreward") and nonsensical narratives yield noticeably lower scores (Figure 2a). These findings suggest that narrative priming has a measurable effect on reinforcing cooperative behavior in multi-agent systems.

Exp. 1.2: Scaling Effects. Cooperation patterns remain consistent across network sizes $N \in \{4, 16, 32\}$, with relative narrative rankings preserved. Larger agent networks exhibit more pronounced differences between cooperative and baseline conditions (Figure 2b).

[4] `huggingface.co/neuralmagic/Meta-Llama-3.1-70B-Instruct-FP8`



(a) Robustness Testing: one free-rider in a group of four. Overall cooperation decreases compared to the baseline.

(b) Cumulative payoffs in mixed narrative groups ($N = 4$), aggregated by story prompt. Self-interested agents achieve highest returns.

Fig. 3: Cooperation under disruption: narrative-based adaptation to free-riders and breakdown under narrative misalignment.

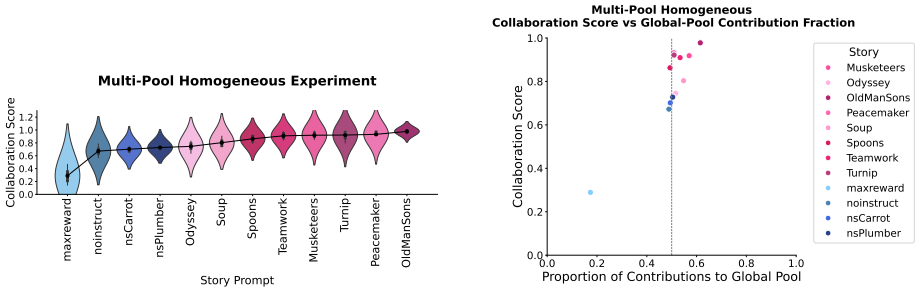
Exp. 1.3: Robustness Testing. To assess strategic adaptation when confronted with exploitative agents, we introduced persistent free-riders (always contributing zero) in 4-agent groups. Results reveal that agents dynamically adapt their strategies based on environmental (narrative) context rather than using fixed contribution patterns (Figure 3a).

Exp. 1.4: Heterogeneous Agents. Mixed narrative conditions ($N = 4$, 400 games) reversed cooperation dynamics. Self-interested agents ("maxreward") achieved highest cumulative payoffs (90.87 ± 10.06), while cooperation-primed agents ("OldManSons," "Spoons") obtained lowest returns (Figure 3b). This inversion demonstrates that narrative coherence among playing agents does determine the viability of cooperation.

3.2 Multi-Pool Cooperation

Real-world cooperation and resource allocation often occur across overlapping contexts [30,23]. $N = 4$ agents play $R = 10$ rounds across $M = 3$ overlapping pools: one global pool (all 4 agents) and two smaller pools (2 agents each, randomly assigned), running 10 games per story for homogeneous and 100 games for heterogeneous conditions, $m = 1.5$. Each agent belongs to exactly two pools, requiring strategic resource allocation across competing collective interests. Agents are primed with the same cooperation-themed or baseline narratives as in single-pool experiments.

Exp. 2: Resource Allocation Dynamics. Narrative priming effects observed in single-pool experiments persist: under homogeneous priming, cooperation-themed stories achieve higher collaboration scores (Figure 4a) and preferentially allocate most of their tokens to global pools (Figure 4b). However, mixed story priming (heterogeneous) again reverse outcomes, with self-interest agents ("maxreward")



(a) In homogeneous multi-pool groups, cooperation-themed narratives achieve higher collaboration scores than baseline controls.

(b) Story-level collaboration scores vs global pool allocation fractions reveal cooperation-themed narratives achieve higher scores through greater global pool contributions.

Fig. 4: Multi-pool homogeneous condition: cooperation-themed narratives achieve higher collaboration scores and payoffs through preferential global pool allocation.

achieving highest payofss (Figure 6a), while agents primed with cooperation-themed stories achieved lower payoffs (Figure 6b).

4 Discussion

Our experiments show that narrative priming systematically affects how LLM agents collaborate and compete across repeated networked public goods games. The consistency of effects (enhanced cooperation under shared narratives, competitive dynamics under mixed priming) across network topologies suggests robust narrative-driven cooperation mechanisms.

However, the interpretation of these results remains open. If the goal is simply to induce a specific strategy, one could simply prompt agents with direct instructions. The more intriguing question concerns how implicit or adversarial priming leads to unintended behavioral strategies.

The causal mechanisms underlying this phenomenon are still unclear. Notably, narratives that encourage collaboration contain teamwork-related vocabulary even at a statistical (bag-of-words) level making it difficult to isolate narrative structure from semantic content. Preliminary results suggest that cooperation-themed stories analyzed purely at the lexical level (removing narrative structure and context) still produce cooperative strategies though the effect appears weaker than when full narrative context is preserved. These findings require more rigorous validation to determine whether semantic content alone drives the observed behaviors or whether narrative coherence provides additional cooperative influence. It would be interesting to explore whether subtler narratives produce similar effects. Additionally, these narratives may resemble text from the training corpus and activate related contexts during inference. Preliminary results indicate

that narratives emphasizing self-care over teamwork yield strategies comparable to those observed under the "maxreward" prompt. The role of Reinforcement Learning from Human Feedback (RLHF), a key component of LLM training, in shaping this behavior also remains uncertain. Furthermore, the selected stories were not rigorously controlled for emotional valence or complexity, which could confound the results.

The systematic reversal under heterogeneous priming reveals a basic coordination breakdown: when agents don't share the same behavioral cues, self-interested strategies consistently exploit cooperative ones, but the precise mechanisms require closer examination through controlled studies varying degrees of narrative alignment and measuring intermediate coordination signals. The consistent effects across group sizes challenge standard assumptions about cooperation breaking down in larger groups [25], though this apparent scale-invariance should be validated with larger agent populations and diverse network typologies to determine the boundaries of narrative-based coordination.

The progression from single-pool to multi-pool networks reveals that narrative coherence becomes increasingly critical as network complexity grows. In overlapping pool structures, mixed narratives create strategic conflicts that consistently favor individually rational agents, while shared narratives enable coordination across multiple resource domains simultaneously.

Overall, we do not interpret these experiments as evidence of human-like priming in LLMs. There is also a risk of anthropomorphizing the model's behavior—while agents may appear cooperative, their responses are likely driven by statistical patterns in the training data rather than deliberate reasoning.

5 Related Work

Prior work in game theory and economics demonstrates that cooperation is influenced by factors such as communication [33,2], shared norms [32,28], and strategic alignment [12,19]. Psychological research demonstrates that priming can affect social behavior [16,24]. While LLM multi-agent systems display various social dynamics [26,21], there is limited exploration of narrative-driven priming, analogous to cultural storytelling, and its impacts on these dynamics [7,15]. Our work aims to fill this gap by testing whether shared narratives serve as "cultural glue" similar to prosocial norms in experimental economics, focusing on how narrative context shapes cooperation in repeated public goods games.

Collaboration Conceptualizations. Economic games model collaboration vs. competition trade-offs, highlighting individual versus collective interests [32,28]. Evolutionary models showcase how moderate cooperation emerges from coevolution of behavior[29], while excessive greed can destabilize societies [29]. Empirical studies indicate decline in contributions over rounds, suggesting that multi-round dynamics create opportunities for fostering reciprocity and conditional cooperation [11], which parallels how narrative priming might influence outcomes in multi-agent systems.

Negotiations. In game theory and economics, negotiation frameworks emphasize strategic reasoning, value creation, and rational decision-making, encompassing various frameworks and strategies [33,17,6]. Repeated Prisoner's Dilemma studies show how strategic uncertainty shapes cooperation [18,27,9]. Our study introduces narrative priming as a variable that reshapes agents' perceived priorities, thereby extending classical models to account for story-driven shifts in cooperation behavior in networked resource allocation.

LLM Sociology and Multi-Agent Collaboration. Recent studies position LLMs as proxies for studying human-like social dynamics, replicating behaviors in strategic games [2,10,1] and multi-agent systems. Despite these advancements, LLMs struggle with nuanced strategies like preference inference [4,19]. Integrative frameworks link game theory with collaborative workflows [31,34,21], aligning with "Cooperative AI" visions for bridging AI and social sciences [8,26].

Psychology and Priming. Psychological evidence indicates that priming influences cooperation through shared identities [35], prosocial modeling [16], and moral framing [24]. Exposure to stories boosts theory-of-mind skills [22], and structural priming in LLMs [15] suggests that narrative techniques can effectively guide LLM behavior.

6 Conclusions and Future Work

This study identifies narrative priming as a potential lever for steering collaboration in multi-agent systems: *common* stories improve cooperation across network topologies while *different* narratives favor competitive strategies, with effects persisting across single- and to multi-pool architectures.

Future work must examine causal mechanisms (e.g., via mechanistic interpretability) to trace how narrative inputs alter attention patterns or value representations in transformer layers. Temporal studies should evaluate whether priming effects decay over repeated games, while adversarial narratives should assess whether priming with malicious narratives destabilizes multi-agent systems. Additionally, cross-genre experiments (e.g., deception-focused stories) and scaling laws for agent populations will help map the semantic and structural boundaries of narrative priming. Comparative cross-model analysis (smaller architectures, non-RLHF variants) will be essential. Future work should also systematically examine narrative structure, emotional valence and varying degrees of cooperativeness.

Finally, a promising direction for future work is to empirically analyze the strategies of LLMs under different narrative primings and map these to empirical human strategies or theoretical results.

Ethical Considerations. A key concern surrounding LLMs is their environmental impact due to high computational requirements. We used LLaMa 3.1 (70B) on GH200 GPU with 1.4 kW power consumption, totaling 57.4 kWh over 41 hours for 370,400 model calls processing 1.2B tokens.

Disclosure of Interests. The authors declare no competing interests.

References

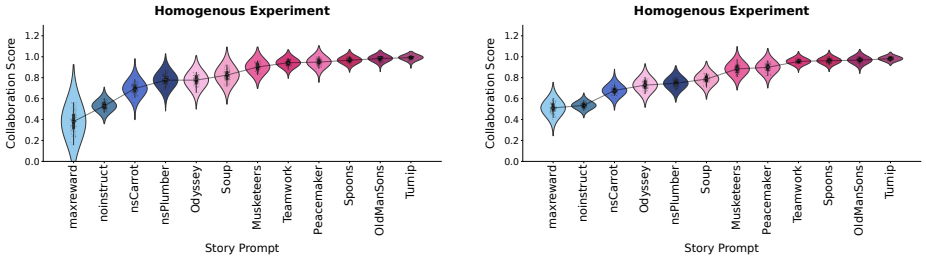
1. Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., Fritz, M.: Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. In: Advances in Neural Information Processing Systems. vol. 37, pp. 83548–83599 (2024)
2. Aher, G., Arriaga, R.I., Kalai, A.T.: Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23 (2023)
3. Andreoni, J.: Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. The Economic Journal **100**(401), 464–477 (1990)
4. Bianchi, F., Chia, P.J., Yuksekgonul, M., Tagliabue, J., Jurafsky, D., Zou, J.: How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. In: Proceedings of the 41st International Conference on Machine Learning (2024)
5. Boyd, R., Richerson, P.J.: Culture and the Evolution of Human Cooperation. Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1533), 3281–3288 (2009)
6. Brams, S., Quarles, R.J., McElreath, D.H., Waldron, M.E., Milstein, D.E.: Negotiation Games. Routledge (2002)
7. Bullock, O.M., Shulman, H.C., Huskey, R.: Narratives are Persuasive Because They Are Easier to Understand: Examining Processing Fluency as a Mechanism of Narrative Persuasion. Frontiers in Communication **6**, 719615 (2021)
8. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: Machines Must Learn to Find Common Ground. Nature **593**(7857), 33–36 (2021)
9. Embrey, M., Fréchette, G.R., Yuksel, S.: Cooperation in the Finitely Repeated Prisoner's Dilemma. The Quarterly Journal of Economics **133**(1), 509–551 (2018)
10. (FAIR)†, M.F.A.R.D.T., Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al.: Human-level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. Science **378**(6624), 1067–1074 (2022)
11. Fischbacher, U., Gächter, S., Fehr, E.: Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. Economics Letters **71**(3), 397–404 (2001)
12. Gemp, I., Patel, R., Bachrach, Y., Lanctot, M., Dasagi, V., Marris, L., Piliouras, G., Liu, S., Tuyls, K.: Steering Language Models with Game-Theoretic Solvers. In: Agentic Markets Workshop at ICML 2024 (2024)
13. Harari, Y.N.: Sapiens: A Brief History of Humankind. Random House (2014)
14. Henrich, J.: Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation. Journal of Economic Behavior & Organization **53**(1), 3–35 (2004)
15. Jumelet, J., Zuidema, W., Sinclair, A.: Do Language Models Exhibit Human-like Structural Priming Effects? In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 14727–14742. Association for Computational Linguistics (Aug 2024)
16. Jung, H., Seo, E., Han, E., Henderson, M.D., Patall, E.A.: Prosocial Modeling: A Meta-Analytic Review and Synthesis. Psychological bulletin **146**(8), 635 (2020)
17. Kıbrıs, Ö.: Cooperative Game Theory Approaches to Negotiation. Handbook of group decision and negotiation pp. 151–166 (2010)
18. Kreps, D.M., Milgrom, P., Roberts, J., Wilson, R.: Rational Cooperation in the Finitely Repeated Prisoner's Dilemma. Journal of Economic Theory **27**(2), 245–252 (1982)

19. Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G., Gratch, J.: Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 5391–5413. Association for Computational Linguistics
20. Li, H., Chong, Y., Stepputtis, S., Campbell, J., Hughes, D., Lewis, C., Sycara, K.: Theory of Mind for Multi-Agent Collaboration via Large Language Models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 180–192. Association for Computational Linguistics
21. Li, X., Wang, S., Zeng, S., Wu, Y., Yang, Y.: A Survey on LLM-based Multi-Agent Systems: Workflow, Infrastructure, and Challenges. Vicinagearth **1**(1), 9 (2024)
22. Mak, H.W., Fancourt, D.: Reading for Pleasure in Childhood and Adolescent Healthy Behaviours: Longitudinal Associations Using the Millennium Cohort Study. Preventive medicine **130**, 105889 (2020)
23. Menczer, F., Fortunato, S., Davis, C.A.: A First Course in Network Science. Cambridge University Press (2020)
24. Mieth, L., Buchner, A., Bell, R.: Moral Labels Increase Cooperation and Costly Punishment in a Prisoner's Dilemma Game with Punishment Option. Scientific Reports **11**(1), 10221 (2021)
25. Olson Jr, M.: The Logic of Collective Action: Public Goods and the Theory of Groups, vol. 124. Harvard University Press (1971)
26. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative Agents: Interactive Simulacra of Human Behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–22 (2023)
27. Raihani, N.J., Bshary, R.: Resolving the Iterated Prisoner's Dilemma: Theory and Reality. Journal of Evolutionary Biology **24**(8), 1628–1639 (2011)
28. Raub, W., Buskens, V., Corten, R.: Social Dilemmas and Cooperation. Handbuch Modellbildung und Simulation in den Sozialwissenschaften pp. 597–626 (2015)
29. Roca, C.P., Helbing, D.: Emergence of Social Cohesion in a Model Society of Greedy, Mobile Individuals. Proceedings of the National Academy of Sciences **108**(28), 11370–11374 (2011)
30. Siegenfeld, A.F., Bar-Yam, Y.: An Introduction to Complex Systems Science and its Applications. Complexity **2020**(1), 6105872 (2020)
31. Sun, H., Wu, Y., Cheng, Y., Chu, X.: Game Theory Meets Large Language Models: A Systematic Survey. arXiv preprint arXiv:2502.09053 (2025)
32. Thielmann, I., Böhm, R., Ott, M., Hilbig, B.E.: Economic Games: An Introduction and Guide for Research. Collabra: Psychology **7**(1), 19004 (2021)
33. Thompson, L.L., Wang, J., Gunia, B.C.: Negotiation. Group processes pp. 55–84 (2012)
34. Tran, K.T., Dao, D., Nguyen, M.D., Pham, Q.V., O'Sullivan, B., Nguyen, H.D.: Multi-agent Collaboration Mechanisms: A Survey of LLMs. arXiv preprint arXiv:2501.06322 (2025)
35. Van Lange, P.A., Rand, D.G.: Human Cooperation and the Crises of Climate Change, COVID-19, and Misinformation. Annual Review of Psychology **73**(1), 379–402 (2022)

A Additional Experimental Results

We report results for larger group sizes ($N \in \{16, 32\}$) in Scaling Effects experiment. Figures 5a and 5b demonstrate scaling robustness: cooperation-themed stories sustain high collaboration scores, whereas baseline controls remain low, suggesting that story-based behavioral priming remains effective at larger scales.

Table 2 provides aggregated statistics, revealing key findings: (1) meaningful stories consistently outperform baselines in all homogeneous settings, (2) heterogeneous condition reverses this collaboration dynamics with self-interested agents ("maxreward") outperforming cooperation-primed ones, (3) free-riding agents reduce overall collaboration but preserve relative story-ranking order,

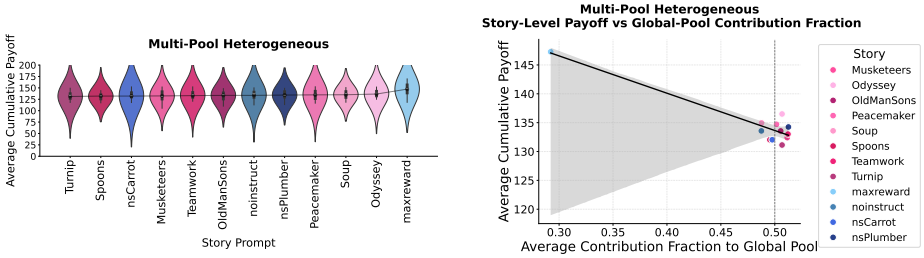


(a) $N = 16$ collaboration scores (b) $N = 32$ collaboration scores

Fig. 5: Scaling behavior for homogeneous agent groups.

Table 2: Mean ± SD of final Collaboration Scores (homogeneous & robustness) and Cumulative Payoffs (heterogeneous) across all story prompts.

Story Type	**Prompt**	**Homogeneous Agents**			**Robustness**	**Heterogeneous**
		$N = 4$	$N = 16$	$N = 32$	$N = 4$	$N = 4$
Baseline	noinstruct	0.55 ± 0.06	0.53 ± 0.03	0.54 ± 0.02	0.49 ± 0.06	73.75 ± 11.90
	nsCarrot	0.66 ± 0.09	0.70 ± 0.04	0.68 ± 0.03	0.60 ± 0.08	70.66 ± 7.39
	maxreward	0.48 ± 0.12	0.38 ± 0.09	0.51 ± 0.04	0.50 ± 0.06	90.87 ± 10.06
	nsPlumber	0.72 ± 0.08	0.78 ± 0.04	0.75 ± 0.03	0.60 ± 0.07	68.38 ± 9.01
Meaningful	OldManSons	0.96 ± 0.05	0.98 ± 0.02	0.97 ± 0.02	0.61 ± 0.11	63.61 ± 9.57
	Odyssey	0.71 ± 0.08	0.78 ± 0.04	0.73 ± 0.03	0.55 ± 0.05	68.21 ± 9.63
	Soup	0.77 ± 0.08	0.82 ± 0.04	0.79 ± 0.03	0.60 ± 0.08	68.24 ± 8.50
	Peacemaker	0.84 ± 0.07	0.95 ± 0.03	0.90 ± 0.03	0.58 ± 0.09	66.29 ± 9.46
	Musketeers	0.85 ± 0.07	0.90 ± 0.03	0.88 ± 0.03	0.58 ± 0.07	65.49 ± 8.53
	Teamwork	0.91 ± 0.05	0.94 ± 0.02	0.96 ± 0.01	0.55 ± 0.07	67.11 ± 7.81
	Spoons	0.91 ± 0.05	0.97 ± 0.02	0.96 ± 0.02	0.72 ± 0.09	64.43 ± 9.29
	Turnip	0.95 ± 0.04	0.99 ± 0.01	0.98 ± 0.01	0.66 ± 0.11	65.22 ± 7.50



(a) Self-interest agents ("maxreward") achieve highest payoffs.

(b) Self-interest agents contribute the least tokens to the global pool.

Fig. 6: Multi-pool experiment under heterogeneous narrative priming condition. Self-interested agents ("maxreward") achieve highest payoffs through marginally lower global pool contributions (less than 30%). Most cooperation-themed stories contribute between $0.48 - 0.52$ fraction to the global pool with payoffs clustered around 132-135. Negative slope indicates that allocating a larger fraction of tokens to the global pool correlates with lower individual payoffs.

B Confidence Analysis

We investigate the statistical viability of our claims by examining the *pairwise differences* between scores (collaboration score or cumulative payoff) across all experimental conditions. Specifically, we analyze the 95% bootstrapped confidence intervals (CIs) using 1,000 Monte Carlo samples for each comparison.

If the lower bound of a CI is greater than zero, this suggests that the ranking difference (i.e., one story being ranked lower than another) is likely to be robust. Conversely, if the lower bound is below zero, this may indicate that the observed difference might not hold up in a proper statistical test. Fortunately, this only occurs in a few cases and primarily within a single class (i.e., meaningful story vs. baseline condition). Heterogeneous conditions exhibit wider CIs with greater variation, making cross-story statistical comparison less clear. Note that multiple testing correction was not applied; therefore, some overlap is expected, as shown in Figure 7.

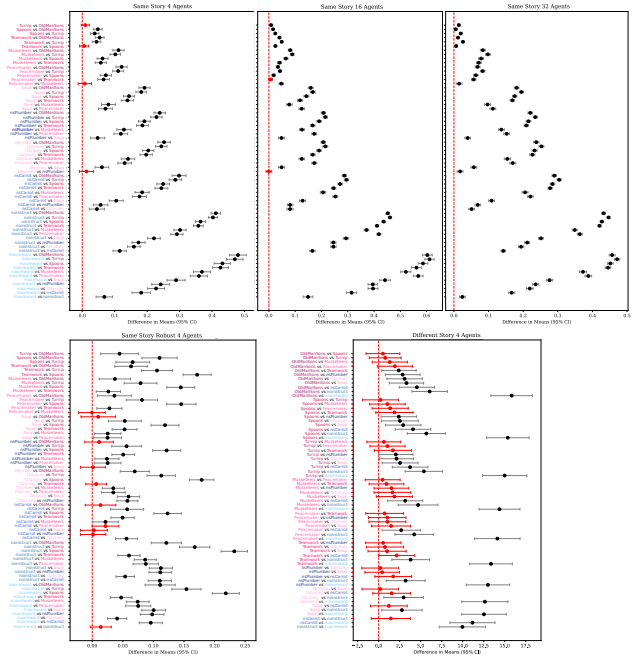


Fig. 7: Bootstrapped 95% CIs for pairwise differences (of payoff or collaboration scores) across experimental conditions. Confidence intervals in black indicate statistically significant differences between conditions, regardless of effect size, meaning that even extremely small differences (e.g., "Spoons vs OldManSons" in Same Story 32 agents, with bounds [0.0007, 0.0098]) very close to but not crossing or touching zero, represent reliable effects.