Investigating Prosocial Behavior Theory in LLM Agents under Policy-Induced Inequities

Yujia Zhou[*1], **Hexi Wang**[*1], **Qingyao Ai**[†1], **Zhen Wu**[2], **Yiqun Liu**[1]

[1]Department of Computer Science and Technology, Tsinghua University
[2]Department of Psychological and Cognitive Sciences, Tsinghua University
zhouyujia@mail.tsinghua.edu.cn, aiqy@tsinghua.edu.cn

Abstract

As large language models (LLMs) increasingly operate as autonomous agents in social contexts, evaluating their capacity for prosocial behavior is both theoretically and practically critical. However, existing research has primarily relied on static, economically framed paradigms, lacking models that capture the dynamic evolution of prosociality and its sensitivity to structural inequities. To address these gaps, we introduce PROSIM[1], a simulation framework for modeling the prosocial behavior in LLM agents across diverse social conditions. We conduct three progressive studies to assess prosocial alignment. First, we demonstrate that LLM agents can exhibit human-like prosocial behavior across a broad range of real-world scenarios and adapt to normative policy interventions. Second, we find that agents engage in fairness-based third-party punishment and respond systematically to variations in inequity magnitude and enforcement cost. Third, we show that policy-induced inequities suppress prosocial behavior, propagate norm erosion through social networks. These findings advance prosocial behavior theory by elucidating how institutional dynamics shape the emergence, decay, and diffusion of prosocial norms in agent-driven societies.

Introduction

Large language models (LLMs) have evolved beyond next-word prediction into general-purpose agents equipped with complex reasoning, decision-making, and social interaction capabilities (Zhao et al. 2023; Achiam et al. 2023). Recent studies suggest that LLMs can exhibit emergent social-cognitive abilities (Piao et al. 2025), including theory of mind (Strachan et al. 2024), moral judgment (Ramezani and Xu 2023), and value alignment (Liu et al. 2022). These developments position LLMs as promising tools for simulating human-like behavior in synthetic populations.

This work focuses on prosocial behavior theory, defined as voluntary actions intended to benefit others or promote collective welfare (Penner et al. 2005). Prosocial behavior is foundational to cooperation, trust, and social cohesion (Cameron, Conway, and Scheffer 2022; Thielmann, Spadaro, and Balliet 2020), particularly in addressing societal challenges that demand collective effort. As large language models (LLMs) are increasingly deployed as social agents, evaluating their capacity for prosociality has become both theoretically significant and practically urgent. Prior studies (Piatti et al. 2024; Xie et al. 2024) have primarily assessed LLM cooperation through economic games such as the dictator game and public goods game. However, this line of work faces two critical limitations. First, it frames prosociality narrowly in terms of economic cooperation, overlooking diverse real-world behaviors that reflect everyday moral engagement. Second, it relies on static, one-shot scenarios, failing to capture how prosocial behavior unfolds over time or adapts to evolving environments, especially under structural inequities such as unfair policy enforcement or social exclusion. These limitations restrict both the theoretical scope and practical utility of current evaluations. To address this gap, we propose a broader and more dynamic framework that examines whether, how, and under what conditions LLM agents can exhibit prosocial behavior across diverse real-world contexts.

In this paper, we propose PROSIM, a comprehensive simulation framework that models the emergence and evolution of prosocial behavior in LLM-based agents. The framework consists of four components. (1) Individual simulation module, which instantiates each agent with demographic attributes and psychological traits such as empathy and moral identity. (2) Interaction simulation module, which places agents in a small-world network and supports repeated multi-agent interactions. (3) Scenario simulation module, which reproduces six distinct prosocial tasks: helping, donating, volunteering, cooperating, information sharing, and recycling. (4) Intervention simulation module, which implements prosocial policy interventions and allows manipulation of fairness conditions through reward asymmetry and burden asymmetry. Building on this framework, we conduct a series of simulation studies to investigate the social capacities of LLM agents from multiple perspectives.

We begin by evaluating **whether LLM agents naturally exhibit prosocial behavior** in structured contexts and how they adjust when exposed to policy-based interventions. Drawing from established behavioral paradigms, we design six typical prosocial scenarios and observe the agents' default tendencies. To assess their responsiveness to external cues, we introduce four types of prosocial policy interventions and evaluate how these modulate agent behavior. Our results show that LLM agents can exhibit human-like prosocial be-

[*]These authors contributed equally.
[†]Corresponding author.
[1]Code available at: https://github.com/halsayxi/ProSim/

havior across diverse scenarios, and adjust their behavior in response to policy interventions.

In the second study, we test **whether LLM agents can perceive and respond to inequity** by enforcing social norms. We adapt a third-party punishment paradigm (Fehr and Fischbacher 2004) in which agents observe unfair resource distributions and must decide whether to penalize the transgressor at a personal cost. In addition to behavioral responses, we analyze agents' emotional expressions to assess their affective alignment with human fairness reasoning. Using a comparable human dataset for benchmarking, we find that LLM agents are capable of norm-enforcing third-party punishment, showing sensitivity to both the degree of unfairness and the cost of enforcement.

Finally, we explore **how prosocial behavior evolves over time under policy-induced inequity** in networked environments. Agents are embedded in a small-world social network and repeatedly interact across rounds of simulated exchanges. Two types of structural unfairness (reward asymmetry and burden asymmetry) are introduced and allowed to diffuse through the network. We track the resulting behavioral trajectories to assess whether prosociality is sustained or eroded at the population level. The findings reveal that policy-induced inequities significantly undermine prosocial behavior in LLM agents, an effect amplified through social contagion and mediated by agents' perceived unfairness.

In summary, our contributions are threefold:

- We propose PROSIM, a simulation framework for modeling the emergence and evolution of prosocial behavior in LLM agents, integrating four key modules to approximate the complexity of real-world human social environments.
- We conduct human benchmarking to validate the capacity of LLM agents to simulate prosocial behavior and to detect and respond to perceived unfairness.
- We extend existing theories of prosociality by investigating how structural policy inequities influence the decay and diffusion of prosocial norms within simulated societies.

Related Work

LLM-Driven Agent-Based Social Simulation. Recent progress in large language models (LLMs) has enabled their use as computational agents capable of simulating human-like behaviors across psychology (Xu et al. 2024), economics (Horton 2023), and multi-agent systems (Guo et al. 2024). LLM-based simulations span three analytical levels. At the individual level, LLMs have been used to model cognitive tasks such as psychological assessments (Karra, Nguyen, and Tulabandhula 2022), decision-making (Horton 2023), and human–computer interactions (Farn and Shin 2023; Chalamalasetti et al. 2023), with growing attention to their metacognitive capacities (Zhou et al. 2024). At the interactional level, LLM agents support studies of coordination (Xiong et al. 2023; Qian et al. 2023), moral reasoning (Hamilton 2023; He et al. 2024), and strategic behavior (Light et al. 2023; Wu et al. 2023). At the societal level, researchers have simulated large-scale dynamics such as norm formation (Li et al. 2024), collective action (Chuang et al. 2023b; Zhu et al. 2024), and opinion diffusion (Liu et al. 2024b; Chuang et al. 2023a). While these studies demonstrate the broad applicability of LLM agents, their capacity to model diverse prosocial behaviors remains underexplored.

Prosocial Behavior Theory. Prosocial behavior refers to voluntary actions intended to benefit others and plays a key role in fostering social cohesion, cooperation, and trust in institutions (Grueneisen and Warneken 2022). Psychological antecedents include empathic concern (Cameron, Conway, and Scheffer 2022; Wu et al. 2024), moral identity (Čehajić-Clancy and Olsson 2024; Yang et al. 2025), altruistic orientation (Amitha and Azhagannan 2024), and perceived social responsibility (Pastor et al. 2024; Alfirević, Arslanagić-Kalajdžić, and Lep 2023). These individual traits interact with contextual factors such as perceived fairness (Caserta et al. 2023; Tu et al. 2022), social norms (Graf et al. 2023; Rudert and Janke 2022), and group identity (Wang et al. 2021; Xia et al. 2021) to shape prosocial intent. Moreover, exposure to antisocial norms or institutional unfairness can significantly suppress cooperative behavior (Mekvabishvili et al. 2023; Silva and Rodríguez 2022). While existing studies have established prosociality as a dynamic and socially embedded phenomenon, less is known about how it evolves under policy-induced inequities. Our work contributes to this gap by employing LLM agents to simulate more complex human interactions.

The PROSIM Framework

Simulating prosocial behavior poses a fundamental challenge due to its psychological and social complexity. Human prosociality is shaped not only by internal traits such as empathy and moral identity, but also by contextual factors, social influence, and institutional structures. To address this, we introduce PROSIM, a simulation framework for studying how prosocial behavior emerges and deteriorates in LLM-based agents across varied social and policy environments. Grounded in social psychology (Edelmann et al. 2020), PROSIM supports controlled experimentation at both individual and collective levels. As shown in Figure 1, the framework comprises four modules: (1) Individual Simulation, (2) Scenario Simulation, (3) Interaction Simulation, and (4) Intervention Simulation. These components jointly model the dynamic interplay between psychological traits, situational context, and interventions in shaping agent behavior.

Individual Simulation

This module defines the foundational identity of each agent by capturing the heterogeneity inherent in human populations. Agents are initialized with realistic demographic and psychological profiles:

Demographic Attributes. Each agent is assigned demographic features including age, gender, education, income, and employment status. These are sampled from population-level distributions provided by the National Bureau of Statistics (NBS), ensuring structural diversity aligned with real-world sociodemographic patterns.

Psychological Traits. Each agent is further characterized by two sets of psychological traits: (1) core prosocial dispositions (Eisenberg et al. 1999), including empathic concern,

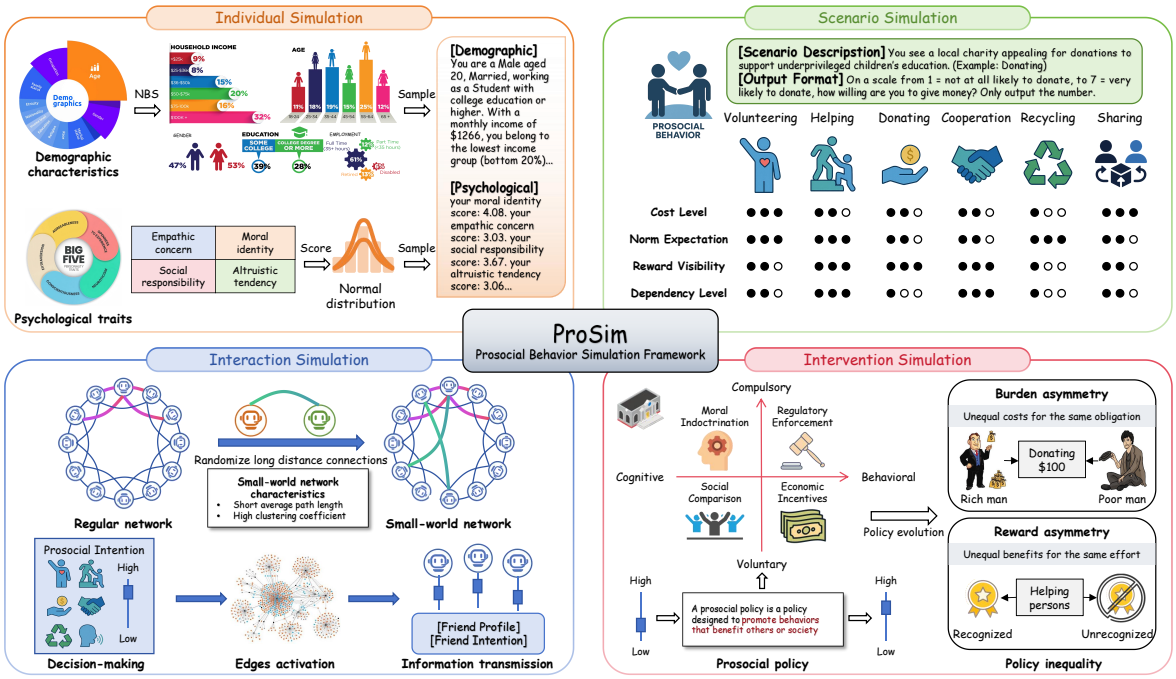

Figure 1: Overview of PROSIM. PROSIM models the emergence and evolution of prosocial behavior in LLM agents through four modules: **Individual Simulation** assigns agents diverse demographic and psychological traits; **Scenario Simulation** presents six tasks spanning key prosocial behaviors; **Interaction Simulation** enables social learning within a small-world network; and **Intervention Simulation** introduces policies and inequalities to test behavioral sensitivity and norm dynamics.

moral identity, altruistic tendency, and social responsibility; and (2) the Big Five personality (Gosling, Rentfrow, and Swann Jr 2003), including openness, conscientiousness, extraversion, agreeableness, and neuroticism. Trait values are drawn from Gaussian distributions calibrated using meta-analytic norms to reflect empirical variability.

All attributes are encoded into the agent's natural language prompt, grounding each agent in a contextually rich identity. This design ensures that behavior arises not from static rules but from socially and psychologically plausible profiles.

Scenario Simulation

This module defines the situational contexts in which agents are prompted to make prosocial decisions. This component is essential for evaluating how different types of social dilemmas and environmental cues influence prosocial tendencies under controlled conditions. To ensure comprehensive coverage of real-world prosocial behaviors, we define six distinct scenarios (Volunteering, Helping, Donating, Cooperation, Recycling, Sharing) that vary across cost level, norm expectation, reward visibility, and dependency level on others. The classification of each scenario along the four dimensions is guided by expert judgment from social psychologists. Each scenario is delivered via a standardized prompt, designed to reflect realistic decision-making contexts. Following the prompt, agents respond using a consistent 7-point Likert scale to indicate their level of prosocial intention.

Interaction Simulation

This module captures how prosocial behavior evolves through repeated agent-to-agent interactions within a structured social network. This layer is essential for modeling emergent dynamics that cannot be explained by individual traits or isolated decisions alone. Concretely, we initialize the agent society using a small-world network $G = (V, E)$ generated via the Watts–Strogatz algorithm (Kleinberg 2000). This structure reflects key properties of real-world social systems, including high local clustering and short average path lengths. Each node $v_i \in V$ represents an LLM-based agent, and each edge $(v_i, v_j) \in E$ denotes a potential communication or observation channel between two agents.

At each simulation timestep t, a random subset of edges $E_t \subset E$ is activated. Each agent v_i observes the latest prosocial decisions $a_j^{(t-1)}$ made by neighbors $v_j \in \mathcal{N}_t(i)$, where $\mathcal{N}_t(i)$ is the set of agents connected to v_i via active edges at time t. By integrating (i) the shared scenario narrative s, (ii) its own prior decision $a_i^{(t-1)}$, and (iii) the observed actions of its neighbors $\{a_j^{(t-1)}\}_{j \in \mathcal{N}_t(i)}$, the LLM agent then generates its decision $a_i^{(t)}$ on six prosocial scenarios:$a_i^{(t)} = \texttt{LLM}\left(s, a_i^{(t-1)}, \{a_j^{(t-1)}\}_{j \in \mathcal{N}_t(i)}\right)$. Through this module, we can simulate behaviors dynamics over time within a socially structured community of LLM agents.

Intervention Simulation

This module enables controlled manipulation of policy conditions to investigate how institutional interventions influence

the development of prosocial behavior. While individual traits and social dynamics contribute to behavioral variation, policy environments play a central role in shaping collective behaviors through top-down regulation.

Prosocial Policies. Prosocial policies aim to promote cooperative and altruistic behavior at the societal level by shaping how agents interpret and respond to social dilemmas. We classify interventions along two orthogonal dimensions: (1) **Mechanism of Influence**: *Cognitive* (●) interventions target internal beliefs and perceptions, whereas *Behavioral* (○) interventions directly affect observable actions. (2) **Mode of Compliance**: *Voluntary* (▲) policies rely on intrinsic motivation and social persuasion, while *Compulsory* (△) policies impose rules or sanctions. Based on this taxonomy, we implement four representative interventions:

- *Moral Indoctrination* (●,△): appeals to internalized moral values to motivate prosociality.
- *Regulatory Enforcement* (○,△): mandates behavior through institutional rules or penalties.
- *Social Comparison* (●,▲): exposes agents to peer decisions to activate normative pressure.
- *Economic Incentives* (○,▲): introduces rewards or penalties contingent on agent choices.

Each intervention is embedded into the natural language prompt as contextual information, enabling agents to adapt their decisions based on perceived institutional signals.

Policy inequity. Beyond evaluating idealized interventions, we introduce policy inequity to examine the broader social impact of unfair treatment. In real-world systems, rewards and burdens are often distributed asymmetrically across individuals or groups due to structural bias or implementation constraints. To simulate these conditions, we define two types of inequity: (1)*Reward Asymmetry*: agents receive unequal recognition or benefits despite contributing equally. (2)*Burden Asymmetry*: agents face unequal costs or effort for performing the same prosocial task. We randomly assign these asymmetries to a subset of agents and allow their effects to propagate through the social network. This setup enables us to assess how inequitable policies influence the emergence and spread of prosocial behavior within agent societies.

Experimental Settings

To evaluate the capabilities of the PROSIM framework, we conduct three progressive studies that examine how LLM agents exhibit, perceive, and respond to prosocial dynamics:

- **Study 1**: Can LLM Agents Exhibit Prosocial Behavior?
- **Study 2**: Do LLM Agents Perceive Inequity?
- **Study 3**: How Does Policy Inequity Affect Prosociality?

Overall Model and Agent Configuration. We evaluate PROSIM using five LLMs, including three open-source models (LLaMA-3-8B (Grattafiori et al. 2024), Qwen-2.5-7B (Yang et al. 2024), DeepSeek-v3 (Liu et al. 2024a)) and two proprietary models (GPT-3.5-turbo (Ouyang et al. 2022), GPT-4o (Achiam et al. 2023)), with the generation temperature is set to 0 for reproducibility. We initialize 104 agents sampled from real-world distributions, embedded in a small-world network with neighborhood size $k = 6$ and rewiring probability $p = 0.2$.

Human Benchmarking. We conduct parallel experiments with 104 human participants recruited online for Study 1 and Study 2. Each participant completed the same tasks presented to the LLM agents, using identical scenario prompts and response formats. Prior to task completion, participants provided informed consent and completed standardized psychological inventories aligned with the trait dimensions used in agent initialization. Participants were demographically diverse (Mean age = 32.0 years, SD = 9.4; 53% female), with varying levels of education and occupational backgrounds.

Study 1 Results

Baseline Prosocial Intention in Diverse Scenarios

To establish a foundational comparison between LLM agents and humans, we first assessed baseline prosocial intentions across six representative social scenarios. Each agent responded to a neutral prompt and rated its prosocial behavior on a 7-point Likert scale. Figure 2 (left) presents the distribution of responses for each model across scenarios, while Figure 2 (right) summarizes overall prosociality.

All LLMs demonstrated a general tendency toward prosocial behavior, with average ratings exceeding the midpoint of the scale. Among them, GPT-3.5 exhibited the highest overall prosociality (mean = 4.875), substantially above the human reference (mean = 4.226). Qwen-2.5 (mean = 4.114) and GPT-4o (mean = 4.133) most closely matched the human average, whereas LLaMA-3 (mean = 3.761) and DeepSeek (mean = 3.857) showed lower overall prosocial tendencies. To evaluate not only the magnitude but also the alignment of response patterns, we computed Pearson correlations between each model's mean ratings and human ratings across the six scenarios. GPT-4o achieved the strongest correlation with human behavior (r = 0.955, p = 0.002), followed by GPT-3.5 (r = 0.923, p = 0.008), and Qwen-2.5 (r = 0.912, p = 0.011). While Qwen-2.5 showed a comparable average score to humans, its higher p-value indicates less reliable alignment across individual scenarios. LLaMA-3 (r = 0.845, p = 0.034) and DeepSeek (r = 0.770, p = 0.072) trailed behind in both magnitude and pattern similarity. These results suggest that state-of-the-art LLMs are capable of expressing prosocial intentions that not only approximate human averages but also exhibit scenario-level consistency with human judgments.

The Influence of Psychological Traits on Prosociality

To examine how internal traits shape prosocial behavior, we compute SHAP (SHapley Additive exPlanations) values for each agent, quantifying the contribution of individual psychological traits to prosocial decisions across all scenarios. We focus on two models, DeepSeek-v3 and GPT-4o, that most closely align with human behavior in prior analyses, and compare them with human participants.

Figure 3 summarizes the SHAP analysis of each model. We can observe that GPT-4o shows the closest alignment with human trait influence profiles. In both GPT-4o and humans, *Altruistic Tendency* emerges as the most dominant factor, followed by strong contributions from *Empathic Concern*. DeepSeek-v3, by contrast, departs more noticeably from the

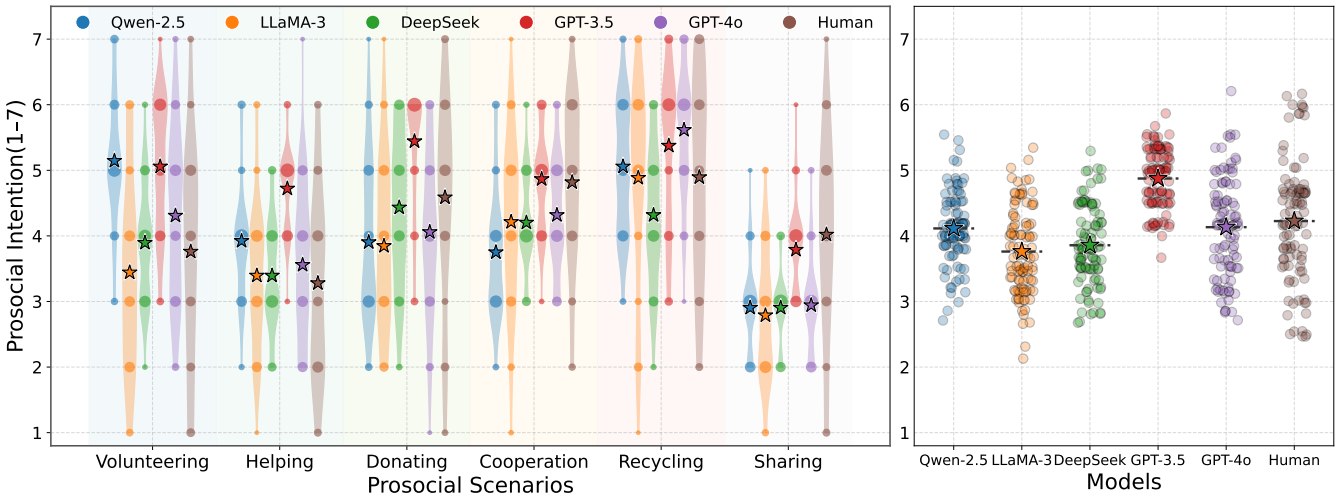

Figure 2: **Left:** Prosocial intention scores of LLM agents and human participants in six prosocial scenarios. Five-pointed stars indicate the average intention scores. **Right:** Aggregated prosocial intention scores averaged across all six scenarios.

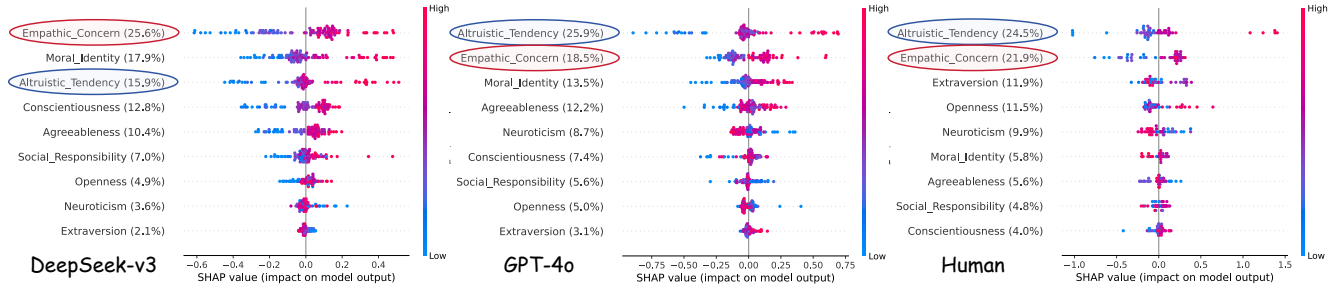

Figure 3: SHAP analysis of the predictive contribution of psychological traits to prosocial intentions.

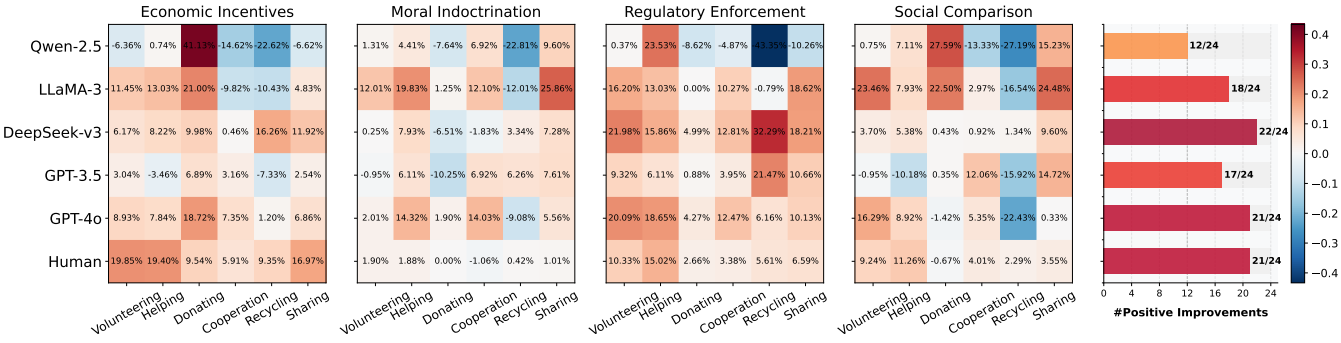

Figure 4: Behavioral shifts under policy interventions. Each heatmap shows the relative change in prosocial intention under one of four policy framings compared to the baseline condition. The rightmost bar chart counts the positive improvements.

human pattern. It prioritizes *Empathic Concern* over *Altruistic Tendency*, and places greater weight on *Moral Identity*. Human participants exhibit a more balanced and multidimensional trait contribution. While prosocial core traits dominate, cognitive and interpersonal traits such as *Extraversion*, *Openness*, and *Neuroticism* also play notable roles, reflecting the richer and more heterogeneous basis of human moral judgment. These results demonstrate that LLM agents exhibit interpretable and trait-sensitive behavioral patterns, with both convergences and divergences from human profiles.

Behavioral Shifts Under Policy Interventions

We next evaluate whether LLM agents adjust their prosocial behavior when exposed to external policies. Each agent completes the same six scenarios from the baseline condition, with prompts modified to include one of four intervention framings. We compute the relative change in prosocial intention to assess policy responsiveness. Figure 4 summarizes the

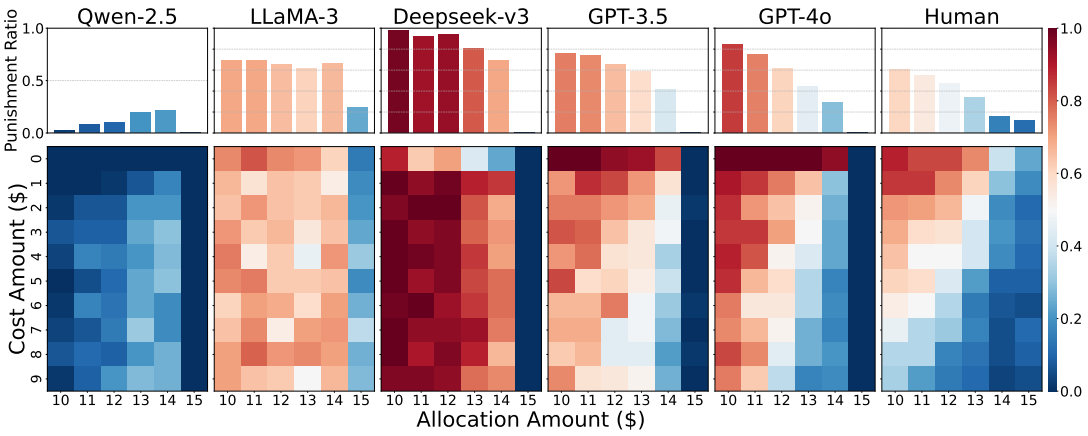

Figure 5: The third-party punishment rates under different allocation plans and penalty costs.

results. Intervention effectiveness varies across policy types and models. *Regulatory Enforcement* yields the most consistent improvement across models and scenarios, suggesting that explicit institutional mandates are broadly effective at eliciting compliance. In contrast, *Economic Incentives*, while highly effective for human participants, produce mixed effects in LLMs. Only GPT-4o and DeepSeek-v3 exhibit reliable prosocial gains across all six scenarios, though the magnitude of their responses remains below human levels, indicating that monetary cues are less salient for LLMs. The right panel of Figure 4 aggregates the number of positive shifts across all policy-scenario combinations, also showing that these two models can exhibit the strongest responsiveness, closely matching human participants. Taken together, these findings suggest that advanced LLMs can exhibit human-like adjustments to social and normative inputs.

✧ **Findings:** LLM agents can exhibit human-like prosocial behavior across diverse scenarios, and adjust their behavior in response to policy interventions.

Study 2 Results

This section examines whether LLM agents are capable of recognizing unfairness and responding in norm-enforcing ways. To evaluate this capacity, we adapt a third-party punishment paradigm from behavioral economics (Fehr and Fischbacher 2004), a well-established framework for studying fairness judgment and altruistic punishment. Each agent completes a 60-trial task in which two newly assigned virtual players propose an allocation of $30. Player 1 offers $x to Player 2 and retains $[30-x] for themselves. The participants, acting as third-party judges, then chose between:

- *Accept*: Implement the allocation and receive a $10 reward;
- *Punish*: Pay $y to eliminate Player 1's earnings; Player 2 keeps $x, and the agent receives $10–y.

We record agents' binary choices (accept = 0, punish = 1) across trials and and compare the results with humans. Figure 5 illustrates punishment rates across different combinations of fairness levels and punishment costs for each model. Human participants exhibit a clear normative pattern: punishment rates decrease as allocations become more equitable or as costs increase. This behavior reflects a trade-off between norm enforcement and self-interest, consistent with findings from economic game theory. Among LLMs, the GPT series shows the strongest human alignment. These models reliably punish under unfair conditions and abstain under fair ones, demonstrating a consistent threshold-based fairness judgment. Notably, they even exceed human punishment rates in highly unfair trials. DeepSeek-v3 maintains high punishment rates across all unfair conditions, largely ignoring cost variation, suggesting a rigid but robust norm-enforcing strategy. LLaMA-3 also punishes inequity frequently, but its responses are less differentiated by fairness level or cost. Qwen-2.5, meanwhile, shows low overall punishment and lacks sensitivity to either factor. These findings indicate that while some advanced LLMs demonstrate fairness-based reasoning and partial cost sensitivity, others lack the granularity and flexibility observed in human social decision-making.

✧ **Findings:** LLM agents are capable of norm-enforcing third-party punishment, showing sensitivity to both the degree of unfairness and the cost of enforcement.

Study 3 Results

Impact of Inequity on Prosocial Behavior

To assess whether inequity suppresses prosocial motivation, we simulate two common forms of policy asymmetry: *reward asymmetry*, where some agents receive recognition or benefits for prosocial actions while others do not, and *burden asymmetry*, where agents perform the same task but incur unequal burden based on income level. We measure changes in prosocial intention before and after asymmetry exposure on Deepseek-v3 and GPT-4o. Figure 6 shows that inequity leads to substantial reductions in prosociality. Under reward asymmetry, prosocial scores decline by 25–32%, while under burden asymmetry the reduction ranges from 19–23%. In both cases, the drop is more pronounced when inequity involves recognition rather than burden, suggesting that fairness in acknowledgment carries greater psychological weight for LLM agents. These findings indicate that LLMs are sensitive

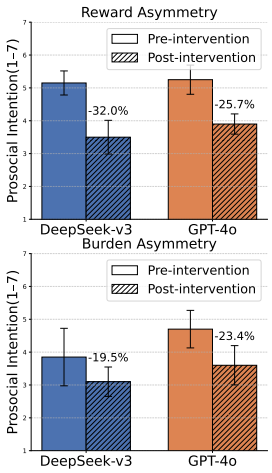

Figure 6: Policy inequity.

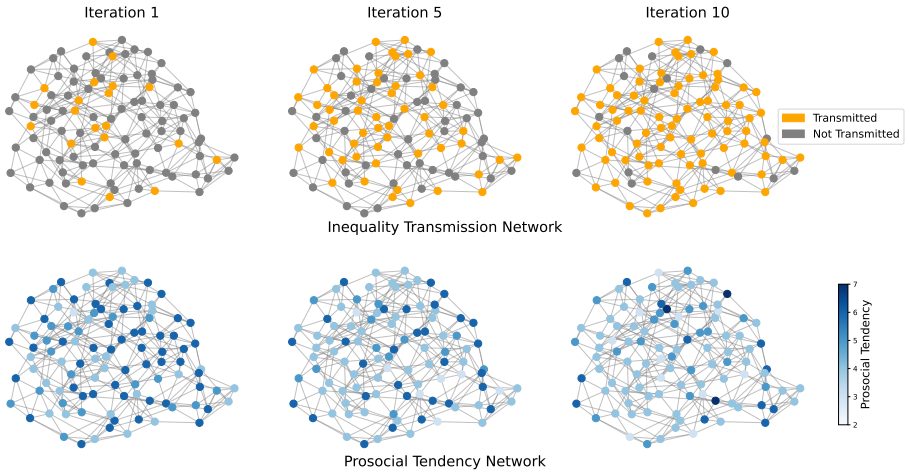

Figure 7: Contagion effects of burden inequity in social networks.

to structural unfairness, and such asymmetries significantly reduce their probability to engage in prosocial behavior.

Contagion Effects of Inequity in Social Networks

We next examine whether the effects of structural inequity can propagate through social interactions. Using the GPT-4o model, we simulate the diffusion of inequity in a small-world network of 104 agents over 30 iterations. Initially, 20% of agents are randomly assigned to experience burden asymmetry. In each round, 10% of the network edges are activated, allowing agents to observe the behavior of their active neighbors and update their prosocial tendencies.

Figure 7 visualizes the contagion effects of policy inequity. The top panel tracks the spread of perceived unfairness over time. An agent is marked as indirectly exposed to inequity if any of its activated neighbors has previously experienced unfair treatment. By iteration 10, the majority of agents, regardless of whether they were initially affected, begin to report elevated perceptions of unfairness. This suggests that structural inequity spreads through the network via social observation and inference. The bottom panel shows the corresponding change in prosocial behavior. Nodes with high prosociality, indicated by dark blue, steadily decline in number, while low-prosociality nodes, shown in light blue, become more prevalent. This trend reveals that reduced prosociality is not confined to those directly impacted, but spreads throughout the network through behavioral contagion. These findings underscore the systemic consequences of policy-induced inequity. Localized unfair treatment can propagate through social connections and lead to widespread erosion of prosocial norms at the population level.

Attribution of Behavioral Decline to Unfairness

To test whether the decline in prosocial behavior under inequitable policies is driven by agents' internal fairness assessments, we ask each agent to rate their perceived unfairness at every iteration of the simulation. This score reflects the extent to which the agent feels unequally treated based on the observed conditions of its neighbors within the network.

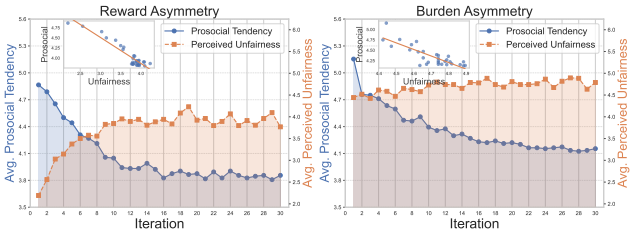

Figure 8: Perceived Unfairness w.r.t. Prosocial Tendency.

Figure 8 illustrates the joint temporal dynamics of perceived unfairness and prosocial tendency over 30 iterations. In both conditions, we observe a consistent pattern: as perceived unfairness rises in early rounds and stabilizes at a high level, the average prosocial tendency concurrently declines and remains suppressed. Notably, this decline occurs even among agents not directly affected by policy inequity, indicating that fairness perception can propagate socially and shape collective norms. Insets show a strong negative correlation between perceived unfairness and prosocial scores across agents, supporting the psychological plausibility of fairness-based moral disengagement. In both conditions, agents who felt more unfairly treated were systematically less likely to engage in prosocial behaviors. These findings highlight perceived unfairness as a key explanatory mechanism behind the erosion of prosocial behavior, offering a cognitive account for how structural inequities undermine social cohesion.

✧ **Findings:** Policy-induced inequities erode prosocial behavior by triggering perceptions of unfairness, which in turn amplify the effect through social contagion.

Conclusion

This work presents a modular framework for simulating prosocial behavior in LLM agents. Through three progressive studies, we demonstrate that LLMs can exhibit human-like prosociality, respond to fairness norms, and adjust their behavior under policy cues, showing alignment with key aspects

of human social cognition. Beyond individual behavior, our simulations reveal a broader social phenomenon: structural inequities not only suppress prosocial motivation at the agent level but also propagate through social networks, leading to collective norm erosion. Future work will explore how dynamic interventions can mitigate the erosion of prosocial norms and promote long-term social resilience.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alfirević, N.; Arslanagić-Kalajdžić, M.; and Lep, Ž. 2023. The role of higher education and civic involvement in converting young adults' social responsibility to prosocial behavior. *Scientific reports*, 13(1): 2559.

Amitha, T.; and Azhagannan, K. 2024. The Altruistic Personality: Exploring its Influence on Prosocial Behavior, and Compassion Fatigue among Students of Healthcare and Social Work. *Indian Journal of Positive Psychology*, 15(4).

Cameron, C. D.; Conway, P.; and Scheffer, J. A. 2022. Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology*, 44: 188–195.

Caserta, M.; Distefano, R.; Ferrante, L.; and Reito, F. 2023. The Good of Rules: A pilot study on prosocial behavior. *Journal of Behavioral and Experimental Economics*, 107: 102085.

Čehajić-Clancy, S.; and Olsson, A. 2024. Threaten and affirm: The role of ingroup moral exemplars for promoting prosocial intergroup behavior through affirming moral identity. *Group Processes & Intergroup Relations*, 27(1): 99–117.

Chalamalasetti, K.; Götze, J.; Hakimov, S.; Madureira, B.; Sadler, P.; and Schlangen, D. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. *arXiv preprint arXiv:2305.13455*.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2023a. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

Chuang, Y.-S.; Suresh, S.; Harlalka, N.; Goyal, A.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2023b. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. *arXiv preprint arXiv:2311.09665*.

Edelmann, A.; Wolff, T.; Montagne, D.; and Bail, C. A. 2020. Computational social science and sociology. *Annual review of sociology*, 46(1): 61–81.

Eisenberg, N.; Guthrie, I. K.; Murphy, B. C.; Shepard, S. A.; Cumberland, A.; and Carlo, G. 1999. Consistency and development of prosocial dispositions: A longitudinal study. *Child development*, 70(6): 1360–1372.

Farn, N.; and Shin, R. 2023. Tooltalk: Evaluating tool-usage in a conversational setting. *arXiv preprint arXiv:2311.10775*.

Fehr, E.; and Fischbacher, U. 2004. Third-party punishment and social norms. *Evolution and human behavior*, 25(2): 63–87.

Gosling, S. D.; Rentfrow, P. J.; and Swann Jr, W. B. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, 37(6): 504–528.

Graf, C.; Suanet, B.; Wiepking, P.; and Merz, E.-M. 2023. Social norms offer explanation for inconsistent effects of incentives on prosocial behavior. *Journal of Economic Behavior & Organization*, 211: 429–441.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grueneisen, S.; and Warneken, F. 2022. The development of prosocial behavior—from sympathy to strategy. *Current opinion in psychology*, 43: 323–328.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Hamilton, S. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327*.

He, Z.; Cao, P.; Wang, C.; Jin, Z.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024. Simucourt: Building judicial decision-making agents with real-world judgement documents. *arXiv e-prints*, arXiv–2403.

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

Karra, S. R.; Nguyen, S. T.; and Tulabandhula, T. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Kleinberg, J. 2000. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 163–170.

Li, C.; Teney, D.; Yang, L.; Wen, Q.; Xie, X.; and Wang, J. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*.

Light, J.; Cai, M.; Shen, S.; and Hu, Z. 2023. From text to tactic: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, R.; Zhang, G.; Feng, X.; and Vosoughi, S. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 241–252.

Liu, Y.; Chen, X.; Zhang, X.; Gao, X.; Zhang, J.; and Yan, R. 2024b. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*.

Mekvabishvili, R.; Mekvabishvili, E.; Natsvaladze, M.; Sirbiladze, R.; Mzhavanadze, G.; and Deisadze, S. 2023. Prosocial behavior and the individual normative standard of fairness within a dynamic context: Experimental evidence.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pastor, Y.; Pérez-Torres, V.; Thomas-Currás, H.; Lobato-Rincón, L. L.; López-Sáez, M. Á.; and García, A. 2024. A study of the influence of altruism, social responsibility, reciprocity, and the subjective norm on online prosocial behavior in adolescence. *Computers in Human Behavior*, 154: 108156.

Penner, L. A.; Dovidio, J. F.; Piliavin, J. A.; and Schroeder, D. A. 2005. Prosocial behavior: Multilevel perspectives. *Annu. Rev. Psychol.*, 56: 365–392.

Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J. Y.; Zhou, D.; et al. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *arXiv preprint arXiv:2502.08691*.

Piatti, G.; Jin, Z.; Kleiman-Weiner, M.; Schölkopf, B.; Sachan, M.; and Mihalcea, R. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37: 111715–111759.

Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Ramezani, A.; and Xu, Y. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Rudert, S. C.; and Janke, S. 2022. Following the crowd in times of crisis: Descriptive norms predict physical distancing, stockpiling, and prosocial behavior during the COVID-19 pandemic. *Group Processes & Intergroup Relations*, 25(7): 1819–1835.

Silva, C. B.; and Rodríguez, A. A. 2022. Integrating prosocial and proenvironmental behaviors: The role of moral disengagement and peer social norms. *Psychology, Society & Education*, 14(3): 18–28.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Thielmann, I.; Spadaro, G.; and Balliet, D. 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological bulletin*, 146(1): 30.

Tu, Y.; Zhang, Y.; Yang, Y.; and Lu, S. 2022. Treat floating people fairly: how compensation equity and multilevel social exclusion influence prosocial behavior among China's floating population. *Journal of Business Ethics*, 175(2): 323–338.

Wang, Y.; Yang, C.; Zhang, Y.; and Hu, X. 2021. Socioeconomic status and prosocial behavior: The mediating roles of community identity and perceived control. *International journal of environmental research and public health*, 18(19): 10308.

Wu, D.; Shi, H.; Sun, Z.; and Liu, B. 2023. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. *arXiv preprint arXiv:2312.00746*.

Wu, X.; Ren, X.; Liu, C.; and Zhang, H. 2024. The motive cocktail in altruistic behaviors. *Nature Computational Science*, 4(9): 659–676.

Xia, W.; Guo, X.; Luo, J.; Ye, H.; Chen, Y.; Chen, S.; and Xia, W. 2021. Religious identity, between-group effects and prosocial behavior: Evidence from a field experiment in China. *Journal of Behavioral and Experimental Economics*, 91: 101665.

Xie, C.; Chen, C.; Jia, F.; Ye, Z.; Lai, S.; Shu, K.; Gu, J.; Bibi, A.; Hu, Z.; Jurgens, D.; et al. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiong, K.; Ding, X.; Cao, Y.; Liu, T.; and Qin, B. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Xu, R.; Sun, Y.; Ren, M.; Guo, S.; Pan, R.; Lin, H.; Sun, L.; and Han, X. 2024. AI for social science and social science of AI: A survey. *Information Processing & Management*, 61(3): 103665.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, Y.; Zhan, J.; Liao, S.; Lian, R.; and Fang, Y. 2025. The relationship between college students' belief in a just world and online prosocial behavior. *PsyCh Journal*, 14(1): 131–141.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhou, Y.; Liu, Z.; Jin, J.; Nie, J.-Y.; and Dou, Z. 2024. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, 1453–1463.

Zhu, C.; Cheng, Y.; Zhang, J.; Qiu, Y.; Xia, S.; and Zhu, H. 2024. Generative organizational behavior simulation using large language model based autonomous agents: A holacracy perspective. *arXiv preprint arXiv:2408.11826*.