FOS: A Multi-Agent LLM Simulation Platform for Social Science Experiments, Demonstrated on a 1,000-Agent Public Goods Game

Anonymous ACL submission

Abstract

Controlled social science experiments with LLM agents require the ability to run the same synthetic participants under different experimental conditions, use multiple models within a single study, and verify that agent outputs are valid and auditable—capabilities absent from existing multi-agent simulation frameworks. We introduce Future of Society (**FOS**), an open-source platform that supports exactly these requirements using locally runnable 3–4B parameter models, eliminating the need for costly API access. We demonstrate FOS through a large-scale Public Goods Game (PGG) with 1,000 LLM agents, five persona-generation contexts, three incentive conditions, and 29,590 analyzable decisions. Three findings emerge. First, small open-weight models (Gemma 3 4B, Qwen 3 4B, Phi4 Mini, Granite 4 3B, Ministral 3 3B) achieve 98.4% output compliance, confirming their viability for large-scale reproducible experiments. Second, which LLM makes each agent's decisions predicts contribution behavior far more strongly (partial $\eta^2 = 0.253$) than the incentive condition does (partial $\eta^2 = 0.006$), meaning single-model studies risk mistaking model-specific behavior for general findings. Third, the LLM used to write agent personas shapes downstream decisions substantially: the contribution gap between low- and high-income personas ranges from 0.91 to 8.72 tokens depending on which model authored them, showing that persona authorship is itself an experimental variable that must be reported.

1 Introduction

LLMs are increasingly used to construct synthetic agents for social-science-style experiments, but existing platforms leave three methodological gaps: reliance on frontier proprietary systems limits cost and reproducibility; the absence of treatment branching makes it difficult to vary a single condition across the same synthetic population; and single-model designs conflate substantive behavioral patterns with model-specific artifacts.

Future of Society (FOS) is an open-weight, locally runnable platform designed around these gaps. FOS uses structured prompts, strict JSON output requirements, and parseability checks so that 3–4B parameter models can be used in auditable experiments on consumer hardware. Its SimTree module clones a pre-branch simulation state, changes a specified treatment parameter, and reruns the same synthetic agents under alternative conditions. FOS also assigns decision-making models at the agent level, making model identity an explicit experimental factor rather than an unreported implementation detail. The goal is not to make LLM agents more autonomous, but to make their experimental conditions, model dependencies, and output failures measurable.

We demonstrate FOS through a large-scale Public Goods Game (PGG) with five profile-generation contexts $\times$ 200 agents $\times$ three SimTree multiplier branches $\times$ ten rounds. In each context, one LLM generates 200 profiles from the same prompt. The same agents are then evaluated across multiplier branches of 1.3$\times$, 2.0$\times$, and 5.0$\times$, with decision-making LLM assigned at the agent level across Gemma 3 4B (Gemma Team, 2025), Granite 4 3B (IBM Research, 2025), Ministral 3 3B (Liu et al., 2026), Phi4 Mini (Abouelenin et al., 2025), and Qwen 3 4B (Yang et al., 2025). This yields 30,000 planned decisions and 29,590 analyzable decisions after quality filtering. Our contributions are:

- We introduce FOS, an open-source platform for controlled LLM-agent experiments with SimTree treatment branching, structured-output auditing, and mixed-model agent populations.
- We show that small local models can support auditable experiments at scale, achieving 98.4% clean-response compliance / 98.6% analyzable-response coverage across 30,000 decisions.

- We show that decision-making LLM identity is strongly associated with contribution outputs, while multiplier-branch contrasts are small and non-monotonic in this large-group, low-MPCR configuration.
- We show descriptively that profile-generation contexts and demographic role descriptions are associated with downstream decision differences, demonstrating that persona construction should be documented as part of the experimental pipeline.

The experiment addresses four research questions: **RQ1** (parseable-response compliance), **RQ2** (decision-LLM identity effects), **RQ3** (multiplier effects across SimTree branches), and **RQ4** (persona and profile-generation context effects). RQ2 and RQ3 are the primary inferential questions; RQ1 and RQ4 are reported as diagnostic and descriptive analyses respectively.

2 Related Work

LLM agents and behavioral experiments. Recent work uses LLMs as synthetic agents in social-science-style research because they can be queried at scale and placed in tightly controlled prompt environments (Ziems et al., 2024; Manning et al., 2024; Thapa et al., 2025; Aher et al., 2023; Horton, 2023; Argyle et al., 2023). This control is useful only if researchers remain clear about what is being measured. Synthetic outputs may reflect culturally situated perspectives, but they can also reflect stereotypes, distributional skews, opinion and value biases, and task-specific prompt artifacts (Kozlowski and Evans, 2025; Veselovsky et al., 2023; Santurkar et al., 2023; Röttger et al., 2024; Liu et al., 2024). Model fidelity also varies across tasks, prompts, model families, and evaluation settings (Mu et al., 2024; Huang et al., 2024; Santurkar et al., 2023; Röttger et al., 2024). LLM agents are attractive precisely because they make large controlled experiments feasible, but their outputs are inseparable from the language model, prompt, and deployment context that produce them.

Multi-agent simulation platforms. Generative-agent and multi-agent frameworks have shown that LLMs can coordinate, remember, plan, and produce believable social behavior in rich environments (Park et al., 2023; Wu et al., 2023; Hong et al., 2023; Gao et al., 2024; Li et al., 2023; Qian et al., 2024). More recent large-scale social-simulation systems extend this line toward LLM-driven societies and computational social-experiment testbeds (Piao et al., 2025). Most of this infrastructure, however, is optimized for task completion, open-ended interaction, or broad societal simulation rather than controlled behavioral experiments with treatment branching over the same cloned population. For causal comparison, researchers need to clone the same synthetic population, alter one treatment, and audit every output. This is a different design problem from maximizing agent autonomy: experimental platforms often need less free-form interaction, not more, so that treatments and outputs remain comparable across runs. Concerns about sycophancy and prompt injection in multi-agent systems further reinforce the need for constrained, inspectable pipelines (Pitre et al., 2025; Shahroz et al., 2025).

Economic games, model heterogeneity, and persona effects. LLM agents have been studied in market experiments, public-goods-like settings, sustainability dilemmas, and game-theoretic benchmarks (del Rio-Chanona et al., 2025; Sreedhar et al., 2025; Piatti et al., 2024; Mao et al., 2025). These studies show both parallels with and divergences from human behavior, but many use a single underlying model, leaving model identity as an unmeasured confound. This is problematic because model performance, expressed opinions, and social-reasoning outputs vary across architectures, prompts, domains, and evaluation settings (Mu et al., 2024; Huang et al., 2024; Santurkar et al., 2023; Röttger et al., 2024). Persona prompting adds another layer: demographic attributes can systematically shift LLM outputs in ways that likely reflect learned associations, opinion-distribution mismatch, and role-play cues rather than genuine group behavior (Argyle et al., 2023; Santurkar et al., 2023; Levy et al., 2024; Liu et al., 2024). Rather than treating these sensitivities as peripheral limitations, we treat them as measurable features of the LLM-agent pipeline.

3 The Future of Society Platform

FOS is an open-source platform[1] for designing, running, and analyzing behavioral experiments with LLM-powered agents. The platform is organized around four requirements for controlled LLM behavioral assays.

[1] Available at: Anonymous. URL omitted for review.

Table 1: Comparison between FOS and representative LLM-agent frameworks. The table summarizes the primary design emphasis of each framework rather than all possible user-implemented extensions.

Platform	Primary target	Treatment branching	Mixed decision LLMs	Structured audit logs	Behavioral-assay focus
AutoGen	Multi-agent conversation and task-solving workflows	Not a primary design goal	Possible through user configuration, but not central	User-defined or application-specific	Not primarily designed for controlled behavioral experiments
AgentScope	Flexible multi-agent application development	Not a primary design goal	Supported through framework configuration	User-defined or application-specific	Broad application framework rather than behavioral-assay infrastructure
CAMEL	Communicative role-play agents and collaborative task execution	Not a primary design goal	Not central to the framework design	Limited or user-defined	Focused on agent communication rather than controlled treatment comparison
ChatDev	Multi-agent software-development simulation	Not designed for experimental treatment branching	Not central to the framework design	Task-specific logs	Domain-specific software-development simulation
AgentSociety	Large-scale LLM-driven social simulation and computational social-experiment testbed	Not its primary design emphasis	Supported at the population or simulation-configuration level	Platform- and experiment-dependent logs	Focused on large-scale societal simulation rather than within-population treatment branching for controlled behavioral assays
FOS	Controlled behavioral experiments with LLM-agent populations	**Yes; SimTree clones pre-branch state and varies treatment parameters**	**Yes; decision-making LLM is recorded as an agent-level factor**	**Yes; raw responses, parser status, and failure categories are logged**	**Yes; designed for auditable behavioral assays**

System Architecture. FOS is organized as a controlled experiment pipeline rather than an open-ended agent interaction environment. The pipeline contains six core components. First, the scenario manager defines the task, available actions, treatment parameters, and observation structure. Second, the profile generator creates persona descriptions from researcher-specified demographic fields and trait values. Third, the model-assignment module assigns a decision-making LLM to each agent and records the model identifier as an experimental factor. Fourth, the SimTree state manager stores the pre-branch simulation state and creates treatment branches by modifying specified parameters while preserving the cloned agent population. Fifth, the LLM execution module sends fixed-format prompts to the assigned local model and records raw responses. Sixth, the parser and audit module classifies each response as clean, recoverable, or invalid and exports both behavioral outputs and quality metadata for analysis.

This architecture is designed to make model-conditioned behavior measurable. Each decision record includes the agent identifier, profile-generation context, decision-making LLM, branch condition, round number, prompt version, raw model response, parsed action, parsed amount when applicable, parser status, and failure category when parsing fails. The released repository contains the FOS platform code; the analysis-ready decision log is archived at Zenodo (Anonymous, 2026).

Small-model structured prompting. FOS uses short, fixed-format prompts designed for 3–4B local models. Each decision prompt contains five sections: a persona description, a researcher-written scenario, memory from prior rounds, available actions, and a strict JSON output requirement. To reduce structured-generation burden, FOS decomposes parameterized actions into two steps. The first prompt asks only for the action choice. If the agent chooses allocate, a follow-up prompt asks only for the contribution amount; if the agent chooses keep, no follow-up is needed. This decomposition separates action selection from numeric parameter generation, making failures easier to di-

agnose and allowing the audit module to identify whether a model failed at the action level or the amount level.

SimTree treatment branching. SimTree stores a pre-branch simulation state, including agent profiles, assigned decision-making models, memories, scene configuration, treatment parameters, and ordering policy. Researchers can clone that state, modify a specified treatment variable, and rerun the same agents under alternative conditions. In the PGG case study, the same 200 agents in each profile-generation context are evaluated under 1.3×, 2.0×, and 5.0× multiplier branches. Multiplier comparisons are therefore within-agent branch contrasts rather than comparisons between independently sampled populations. This matters because apparent treatment differences in agent simulations can be induced by regenerating profiles, changing context length, or altering model assignment. SimTree makes the counterfactual comparison explicit: what changes when the same synthetic population is placed under a different incentive parameter?

After branching, each condition generates its own payoff and memory trajectory. We treat this divergence as part of the treatment pathway in a repeated collective-action environment rather than as a design flaw. Accordingly, ten-round multiplier comparisons are interpreted as total branch contrasts. A Round-1-only diagnostic, reported in the robustness analysis, separates the initial response to the multiplier parameter from later branch-specific histories.

Mixed-model populations. FOS assigns decision-making LLMs at the agent level. A single experiment can therefore include agents powered by different models, making model identity a measurable experimental factor. The same infrastructure can also record the LLM used to generate profiles, allowing researchers to study profile generation as part of the pipeline rather than as neutral preprocessing.

Audit logs and scenario library. Every model response is stored before parsing and classified into one of three audit categories. A clean response satisfies the requested JSON schema and all action-specific constraints. A recoverable response violates a formatting requirement but can be deterministically mapped to a valid action or amount under pre-specified repair rules. An invalid response cannot be mapped to a valid decision without researcher judgment and is excluded from the main analyzable sample.Action variants are retained by mapping the response to a valid action deterministically; above-endowment amounts are flagged and excluded from the strict-clean specification but retained in the main analyzable sample. FOS also includes scenario templates for game theory, opinion dynamics, network diffusion, and social deduction, while allowing custom scenarios through the same Action and Scene interfaces. The purpose of the scenario library is not to maximize agent autonomy, but to provide reusable bounded-action environments for controlled behavioral comparison. A workflow diagram is provided in Appendix A.1.

4 Experiment Design

4.1 Public Goods Game Setup

We implement a repeated linear public goods game, also known in experimental economics as a voluntary contribution mechanism (Ledyard, 1995; Isaac and Walker, 1988; Zelmer, 2003; Chaudhuri, 2011). Each agent receives 20 tokens per round and decides whether to keep them or allocate some amount to a public pool. Total contributions are multiplied by $m \in \{1.3, 2.0, 5.0\}$ and redistributed equally among all 200 agents in the branch. Each branch runs for ten rounds. The resulting marginal per-capita return is low (MPCR $= m/N = 0.0065$ to 0.025), making this a large-group collective-action assay rather than a canonical small-group PGG. Large-group, low-MPCR designs are an established configuration in collective-action research: Weimann et al. (2019) ran groups of 60 and 100 subjects at MPCRs of 0.02–0.04 and found that substantial contributions persist and that small MPCR variations within large groups produce detectable but modest effects—precisely the pattern we observe. This design is appropriate for evaluating FOS at scale and for studying settings where each individual's contribution has little direct payoff effect; it should not be read as a test of canonical small-group human PGG behavior (Isaac and Walker, 1988; Peréda et al., 2019; Weimann et al., 2019; Burton-Chellew and West, 2021).

4.2 Agent Population and Design Structure

Agents are assigned one of nine demographic archetypes formed by crossing three age groups (18–30, 31–50, 51+) with three income levels (Low, Middle, High), plus Trust and Risk Tolerance

scores from 0 to 100. The experiment has a crossed-and-branched structure: five profile-generation contexts $\times$ 200 agents $\times$ three SimTree multiplier branches $\times$ ten rounds. In each profile-generation context, one LLM writes all 200 profiles from the same template. Agents are then assigned decision-making LLMs across Gemma 3 4B, Granite 4 3B, Ministral 3 3B, Phi4 Mini, and Qwen 3 4B. Thus, multiplier effects are identified within agents across SimTree branches; decision-LLM effects are identified between agents within profile-generation contexts; and profile-generation effects are descriptive because profile writer and run are confounded.

Appendix Table 3 summarises the design structure. The multiplier condition is the only treatment varied within the same agents through SimTree branching. Decision-making LLM is assigned at the agent level and is therefore compared between agents within profile-generation contexts. Profile-generation context is descriptive because each context corresponds to one profile-writing model and one generated profile set; the design does not separate the profile-writing model from run-specific generation variation. This distinction guides the interpretation of all results below.

4.3 Scale and Analysis

The design yields 30,000 planned decisions and 29,590 analyzable decisions after quality filtering.

We organize analysis by the four research questions introduced above. Because contributions are bimodal, we also report a two-part specification separating the allocation decision from contribution amount among allocators. Partial η^2 values are used as descriptive variance summaries rather than as the sole inferential basis for the paper.

Statistical specifications. All analyses use dependence-aware specifications detailed in Appendix A.6: within-agent fixed-effects models with agent-clustered standard errors for multiplier contrasts; agent-clustered models controlling for multiplier branch, round, profile-generation context, demographics, trust, and risk tolerance for decision-LLM and persona effects; and a two-part model separating the allocation decision from conditional contribution amount. Partial η^2 values are used only as descriptive variance summaries.

5 Results

5.1 Parseable-Response Compliance

Across 30,000 planned decisions, 29,528 responses were fully clean and 29,590 decisions were retained in the analyzable sample after quality filtering. We distinguish clean-response compliance from analyzable-response coverage. Clean-response compliance is defined as the share of planned decisions whose raw response satisfied the requested JSON schema and action-specific constraints without repair; this rate is 98.4% . Analyzable-response coverage is defined as the share of planned decisions retained after applying pre-specified recovery and filtering rules; this rate is 29,590 / 30,000 = 98.6%. The most common non-clean categories were unparseable actions (313), missing or invalid amounts (97), and contributions above the 20-token endowment (58).

Compliance varied by decision-making model. Gemma 3 4B and Qwen 3 4B were near-perfect under the structured-output format, while Ministral 3 3B produced the largest share of non-clean outputs at 96.8% clean or analyzable compliance. The dominant failure mode was unparseable action selection rather than invalid numeric amounts. This pattern suggests that the two-step decomposition reduced the burden of numeric parameter generation: when failures occurred, they tended to occur at the coarser action-selection step rather than at the follow-up amount step. Full data-quality summaries by model, failure category, and inclusion rule are reported in Appendix A.2.

The high overall compliance rate confirms that 3–4B local models can sustain auditable experiments at this scale without API access. Crucially, between-model variation in compliance is itself a measurable property of the decision model: a single-model design would treat compliance as a background constant, obscuring this dimension of model-conditioned behavior. Full quality breakdowns by model are reported in Appendix A.2.

Figure 1 shows raw contribution and allocation trajectories by multiplier branch. The ten-round SimTree branch contrasts are small and non-monotonic. In the within-agent SimTree model, which compares each agent against itself across branches and clusters standard errors by agent, the $2.0\times$ branch is associated with 0.42 fewer contribution tokens than the $1.3\times$ branch (95% CI: -0.62 to -0.23), while the $5.0\times$ branch is associated with 0.50 more tokens (95% CI: 0.30 to 0.71). These

branch contrasts are statistically detectable but substantively small: both are much smaller than the 8.6-token spread between the lowest- and highest-contributing decision LLMs. The non-monotonicity of the 2.0× branch also shows that the response does not follow a simple dose-response pattern with respect to the multiplier parameter.

The aggregate non-monotonicity masks heterogeneity across decision-making LLMs. A model including decision-LLM × multiplier interactions (Appendix A.5) shows that the 2.0× reduction is driven primarily by Gemma 3 4B, which serves as the reference category; other models show positive interaction terms that partially offset this reduction. This indicates the treatment is processed through model-specific response policies rather than producing a uniform behavioral effect across the population. The aggregate result is therefore best interpreted as a property of this particular five-model, 200-agent, low-MPCR assay: in a setting where each contributed token returns at most 0.025 tokens directly to the contributor, the multiplier manipulation produces muted and model-differentiated responses rather than a simple monotonic incentive response of the kind often examined in standard linear PGG and voluntary-contribution designs (Ledyard, 1995; Isaac and Walker, 1988; Zelmer, 2003; Chaudhuri, 2011). In methodological terms, the figure also demonstrates why branching is useful: the treatment can be evaluated without changing the profile texts, agent histories at branch start, or decision-model assignment.

5.2 Decision-Model and Profile-Generation-Context Differences

Figure 2 summarizes the two model-context results. With standard errors clustered by agent, decision-making LLM identity is strongly associated with contribution outputs. Relative to Gemma 3 4B, Phi4 Mini contributes 6.14 fewer tokens (95% CI: −6.85 to −5.44), Qwen 3 4B contributes 2.50 more tokens (95% CI: 1.82 to 3.18), and Granite 4 3B contributes 2.01 fewer tokens (95% CI: −2.60 to −1.41). The spread between Phi4 Mini and Qwen 3 4B is 8.64 tokens, far larger than the within-agent multiplier-branch contrasts. As a descriptive variance summary, partial η^2 is 0.253 for decision LLM and 0.006 for multiplier condition.

Profile-generation contexts are also associated with downstream behavioral differences. Because each run uses one LLM to write all 200 profiles, these differences are descriptive rather than causal.

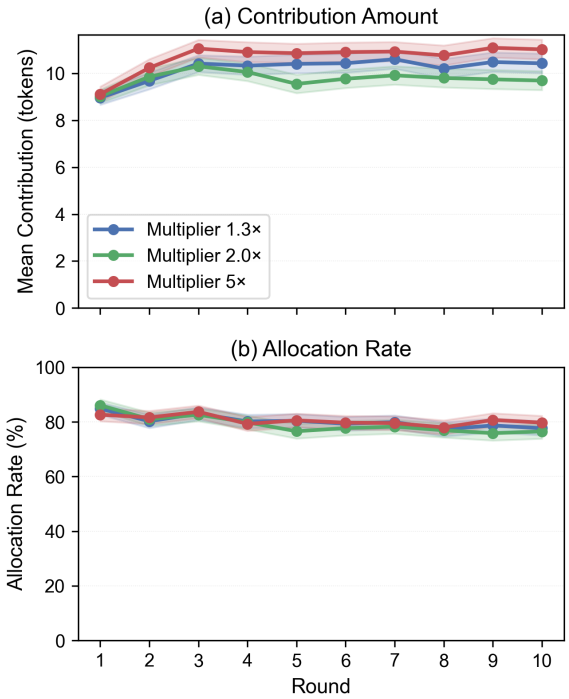
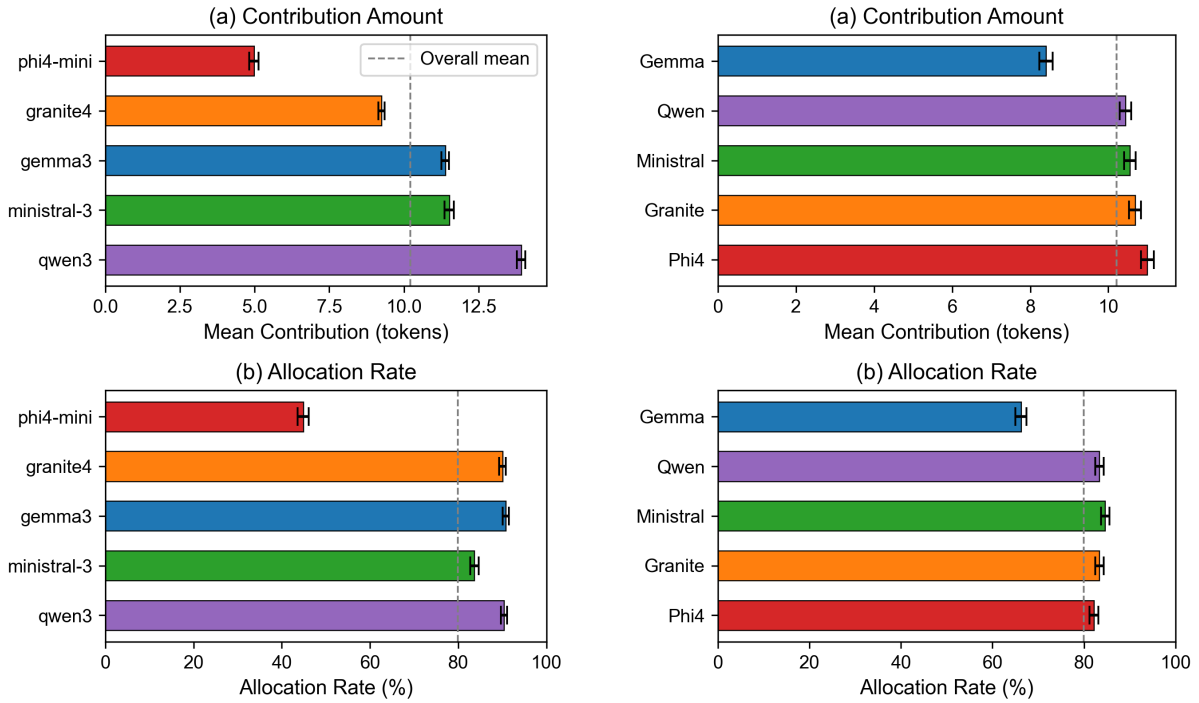


Figure 1: Raw contribution and allocation trajectories across rounds by multiplier condition. The branches are generated from the same agent populations using SimTree. In the dependence-aware within-agent model, the 2.0× branch is −0.42 tokens relative to 1.3× and the 5.0× branch is +0.50 tokens, indicating detectable but substantively small branch-level differences in this low-MPCR setting.

The descriptive low–high income gap is −8.72 tokens in the Gemma-written context, compared with only −0.91 tokens in the Granite-written context. FOS makes this step visible: when profile text is generated by an LLM, the profile writer should be documented as part of the experimental pipeline. That the same label produces such different decisions depending on which model textualized it confirms these are model-specific encodings, not stable demographic signals.

5.3 Profile Textualization and Persona-Conditioned Decision Differences

Demographic role descriptions are associated with systematic downstream decision differences, interpreted as measurements of what these models encode about demographic categories, a feature of the systems being studied, not a claim about real populations. Income is the largest persona-conditioned predictor, accounting for partial $\eta^2 = 0.128$, second only to decision LLM. Figure 3 shows that the low–high income gap ranges from −8.72 tokens in the Gemma-written context to −0.91 tokens in



(a) Decision-making LLM effects.

(b) Profile-generation context effects.

Figure 2: Model context effects on contribution and allocation. Panel (a) compares the LLM assigned to make each agent's decisions; these effects are estimated between agents within profile-generation contexts. Panel (b) compares profile-generation contexts. The profile-generation estimates are descriptive because profile writer and run are confounded.

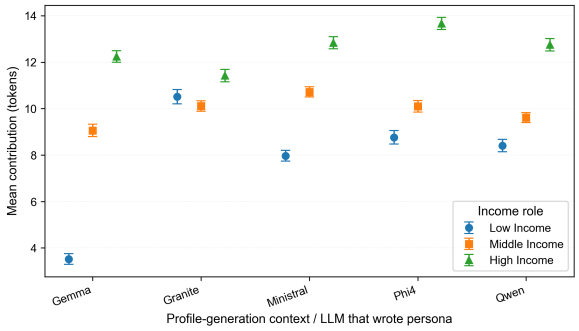


Figure 3: Income role-description effects by profile-generation context. The low–high income gap is largest in the Gemma-written context (-8.72 tokens) and smallest in the Granite-written context (-0.91 tokens).

Table 2: Two-part decomposition by decision-making LLM. Alloc.% = allocation rate; Mean = mean contribution (keep = 0); Cond. = conditional mean among allocators; Full% = full-endowment share.

Decision LLM	Alloc. %	Mean	Cond.	Zero %	Full %
Phi4 Mini	44.9	4.97	11.09	56.2	2.4
Granite 4 3B	90.1	9.25	10.26	10.1	2.0
Gemma 3 4B	90.8	11.38	12.54	9.4	1.3
Ministral 3 3B	83.7	11.52	13.77	16.5	7.1
Qwen 3 4B	90.4	13.92	15.40	9.7	24.6

the Granite-written context: the same demographic label produces substantially different downstream decisions depending on which model textualized it.

Table 2 decomposes the model hierarchy descriptively. Phi4 Mini's low mean contribution (4.97 tokens) is driven by its low allocation rate (44.9%); conditional on allocating, its mean of 11.09 tokens is comparable to other models. Qwen 3 4B combines a high allocation rate (90.4%) with the highest conditional mean (15.40 tokens). Ministral 3 3B shows the opposite pattern: a below-average allocation rate (83.7%) but a high conditional mean (13.77 tokens) and the highest full-endowment share (7.1%).

5.4 Robustness

Robustness checks support the same substantive interpretation. First, exclusion-criteria sensitivity checks yield stable coefficient signs and magnitudes across four datasets: the *full* specification retains clean responses, action variants, and above-endowment amounts; *no-cap* excludes

above-endowment amounts; *no-variant* excludes action variants; and *strict-clean* retains only responses satisfying the JSON schema without repair. Second, dependence-aware models address the repeated-decision structure directly: within-agent SimTree contrasts show small multiplier-branch differences, while agent-clustered decision-LLM models preserve the model hierarchy. Third, the two-part model shows that decision-LLM differences appear both in the decision to allocate and in the amount contributed among allocators. Ministral 3 3B shows the inverse pattern of Phi4 Mini: it allocates at roughly half the rate of Gemma (OR $= 0.498$, 95% CI $[0.318, 0.778]$) but contributes 1.19 more tokens conditional on allocating (95% CI $[0.818, 1.558]$), with the highest full-endowment share of any model. This split illustrates that extensive- and intensive-margin effects can run in opposite directions and would be masked by a combined contribution measure. Fourth, a Round-1-only diagnostic estimates multiplier contrasts before branch-specific memory histories accumulate; both CIs span zero, supporting the interpretation that ten-round estimates are total branch contrasts. Full coefficients are reported in Appendix A.5.

6 Discussion

The PGG case study shows why LLM-agent platforms need experimental control, multi-model populations, and auditable output pipelines. FOS makes model identity explicit at two stages of the pipeline: the LLM that writes profiles and the LLM that makes decisions. The 8.64-token spread between Phi4 Mini and Qwen 3 4B is a substantial fraction of the 20-token endowment and far larger than any multiplier-branch contrast, so single-model studies risk reporting a model-specific output pattern as a general behavioral property. The persona results extend this point to profile construction. Demographic role descriptions are associated with systematic downstream decision differences, and these associations vary across profile-generation contexts. For researchers using generated personas, the persona author is part of the measurement instrument: two studies can use the same demographic labels but produce different downstream agents if those labels are textualized by different models, prompts, or random seeds.

The small and non-monotonic multiplier-branch contrasts should be interpreted in the same controlled-assay frame. With MPCR ranging from 0.0065 to 0.025, direct payoff returns remain small under all multiplier conditions—consistent with Weimann et al. (2019), who find that small MPCR variations within large groups produce modest effects. Contribution outputs likely reflect model-specific interpretations of the shared-pool framing rather than payoff maximization alone; future work should test smaller-group designs with higher MPCRs.

More broadly, the case study suggests a reporting standard for LLM-agent experiments. Researchers should state which model writes the profiles, which model makes each decision, how models are assigned to agents, how outputs are parsed or repaired, and whether treatment comparisons are made between different synthetic populations or within cloned branches of the same population. Without this information, downstream readers cannot tell whether a reported behavioral pattern reflects a treatment condition, a persona template, a profile-generation context, a parsing rule, or the decision LLM itself.

7 Conclusion

FOS enables controlled behavioral assays of LLM-agent systems through SimTree treatment branching, mixed-model populations, and structured compliance auditing. In a large-scale PGG case study, we find that small local models can produce auditable outputs at scale; decision-making model identity is associated with substantially more modeled variance than the multiplier condition in this task configuration; demographic role descriptions are associated with persona-conditioned decision differences; and profile-generation contexts are associated with downstream behavioral variation. These findings support a practical methodological claim: LLM-agent experiments should vary and report model identity, profile-generation procedures, assignment mechanisms, output parsing rules, and compliance rates rather than treating them as hidden implementation details.

Limitations

This study is limited to five open-weight models in the 3–4B range. Whether model-identity effects of comparable magnitude appear at larger scales is an open empirical question; the present results establish their magnitude within this tier and that the 98.4% compliance rate demonstrates

small models are viable for auditable assays. The PGG is also a large-group, low-MPCR assay: it demonstrates FOS at scale and yields small, non-monotonic multiplier-branch contrasts, but it does not test canonical small-group human PGG behavior. The agents are LLM-driven systems, not human participants, so persona-conditioned differences should be read as measurements of how models operationalize demographic role descriptions rather than as evidence about real demographic groups.

The design also has two important identification limits. First, decision-making LLM is assigned between agents rather than crossed within the same profile texts. Although the analysis controls for profile-generation context, demographic fields, trust, risk tolerance, multiplier branch, and round, a fully crossed design in which every decision model evaluates the same profiles would provide a stronger estimate of decision-LLM differences. Second, profile-generating LLM and run are confounded because each profile-generation context uses one model to generate one 200-agent profile set. We therefore treat profile-generation-context differences as descriptive. Future work should generate multiple profile sets per profile-writing model, vary random seeds, and cross the same demographic skeletons with multiple textualization models.

Finally, the present experiment uses strict JSON-only decision prompts, which improve auditability but may constrain the kinds of reasoning that can be observed. Additional diagnostic experiments with payoff-comprehension checks, alternative prompt framings, and explicit reasoning traces could help separate payoff calculation, prosocial framing, persona role-play, and parser-induced behavior. Decoding parameters follow Ollama defaults (temperature 0.8, top_p 0.9, top_k 40) (Ollama, 2026) and no fixed random seed was used; this is a deliberate choice: fixing a single seed would only confirm reproducibility of one draw; running independent stochastic instances instead demonstrates that key findings are stable across stochastic variation rather than artefacts of a specific seed.

Ethical Considerations

This research involved no human subjects. The agents are computational systems powered by LLMs and have no sentience, consciousness, or welfare interests. The demographic role descriptions used in the experiment raise important ethical issues because LLMs may translate social categories into stereotyped or biased decision outputs. We therefore frame persona-conditioned differences as measurements of model social representations and prompt-conditioned decision outputs, not as claims about real demographic groups. We also recommend that users of FOS avoid interpreting synthetic persona outputs as evidence about protected or marginalized populations without external validation. All models used in this study—Gemma 3 4B, Granite 4 3B, Ministral 3 3B, Phi4 Mini, and Qwen 3 4B—are open-weight and publicly available.

References

Abdelrahman Abouelenin and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371.

Anonymous. 2026. Future of society — PGG case study: Analysis-ready decision log.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Maxwell N Burton-Chellew and Stuart A West. 2021. Payoff-based learning best explains the rate of decline in cooperation across 237 public-goods games. *Nature human behaviour*, 5(10):1330–1338.

Ananish Chaudhuri. 2011. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14(1):47–83.

R. Maria del Rio-Chanona, Marco Pangallo, and Cars Hommes. 2025. Can generative AI agents behave like humans? evidence from laboratory market experiments. *Preprint*, arXiv:2505.07457.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Zeyu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. AgentScope: A flexible yet robust multi-agent platform. *Preprint*, arXiv:2402.14034.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, and Xiangliang Zhang. 2024. Social science meets LLMs: How reliable are large language models in social simulations? *Preprint*, arXiv:2410.23426.

IBM Research. 2025. Granite 4.0 language models. https://github.com/ibm-granite/granite-4.0-language-models. Accessed: 2026-05-22.

R. Mark Isaac and James M. Walker. 1988. Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199.

Austin C. Kozlowski and James Evans. 2025. Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, 54:1017–1073.

John O. Ledyard. 1995. Public goods: A survey of experimental research. In John H. Kagel and Alvin E. Roth, editors, *The Handbook of Experimental Economics*, pages 111–194. Princeton University Press.

Sharon Levy, William D. Adler, Tahilin Sanchez Karver, Mark Dredze, and Michelle R. Kaufman. 2024. Gender bias in decision-making with large language models: A study of relationship conflicts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, and et al. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850.

Benjamin S. Manning, Kehang Zhu, and John Horton. 2024. Automated social science: Language models as scientist and subjects. *SSRN Electronic Journal*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYMPICS: LLM agents meet game theory. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2845–2866.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12074–12086.

Ollama. 2026. Ollama. https://github.com/ollama/ollama. Version 0.23.2; accessed 22 May 2026.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

María Peréda, Ignacio Tamarit, Alberto Antonioni, Jose A. Cuesta, Penélope Hernández, and Angel Sánchez. 2019. Large scale and information effects on cooperation in public good games. *Scientific Reports*, 9(1):15023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems*, volume 37.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15174–15186.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15295–15311.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004.

Rana Shahroz, Zhen Tan, Sukwon Yun, Charles Fleming, and Tianlong Chen. 2025. Agents under siege: Breaking pragmatic multi-agent LLM systems with optimized prompt attacks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 9661–9674.

Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V. Nickerson, and Lydia B. Chilton. 2025. Simulating cooperative prosocial behavior with multi-agent LLMs: Evidence and mechanisms for AI agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining*, 15.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *Preprint*, arXiv:2305.15041.

Joachim Weimann, Jeannette Brosig-Koch, Timo Heinrich, Heike Hennig-Schmidt, and Claudia Keser. 2019. Public good provision by large groups – the logic of collective action revisited. *European Economic Review*, 118:348–363.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jennifer Zelmer. 2003. Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3):299–310.

Caleb Ziems, William B. Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Additional Results

A.1 Platform Workflow

A.2 Data Quality by Model

The structured-output audit by decision-making LLM is summarized in Figure 5

A.3 Persona-Conditioned Coefficients

Figure 6 shows the coefficient plot for persona-conditioned contribution shifts. These coefficients summarize how generated role descriptions and trait fields are translated into contribution outputs.

A.4 LLM × Multiplier Trajectories

Figure 7 reports LLM-specific trajectories within each multiplier branch. The plot is supplementary because the main decision-LLM differences are summarized in Figure 2, but it shows that the model hierarchy is stable across rounds and multiplier conditions.

A.5 Dependence-Aware Robustness Checks

Table 4 reports selected robustness coefficients with 95% CIs. Table 5 reports decision-LLM × multiplier interaction terms. Table 6 reports Round-1-only multiplier contrasts as a diagnostic for memory-divergence concerns. The within-agent SimTree model estimates multiplier effects by comparing each agent against itself across branches; the decision-LLM model clusters standard errors by agent; and the two-part model separates allocation from conditional contribution amount.

A.6 Regression Specifications

Let y_{ibr} denote the contribution of agent i in multiplier branch b and round r, with keep decisions coded as zero. Let a_{ibr} indicate whether the agent chose to allocate. Let M_b denote multiplier condition, L_i the decision-making LLM, P_i the profile-generation context, D_i demographic role-description fields, and X_i continuous persona traits.

The within-agent multiplier model is

$$y_{ibr} = \alpha_i + \beta M_b + \gamma_r + \epsilon_{ibr},$$

where α_i are agent fixed effects and γ_r are round fixed effects. Standard errors are clustered by agent.

Table 3: Experimental design structure of the PGG case study. The multiplier condition is varied within agents through SimTree branching, whereas decision-making LLM is assigned between agents. Profile-generation context is descriptive because each context corresponds to one profile-writing model and one generated profile set.

Factor	Levels	Unit of variation	Within/Between	Interpretation
Profile-generation context	5	200-agent profile set	Between context	Descriptive; profile-writing model and generation run are confounded.
Decision-making LLM	5	Agent	Between agent	Model-conditioned output differences, estimated between agents within profile-generation contexts.
Multiplier branch	3	SimTree branch	Within agent	Total branch contrast; the same agents are evaluated under $1.3\times$, $2.0\times$, and $5.0\times$ multiplier branches.
Round	10	Repeated decision	Within branch	Repeated decision process; later rounds include branch-specific payoff and memory histories.
Age role	3	Agent profile	Between agent	Descriptive persona-conditioned association; age groups are 18–30, 31–50, and 51+.
Income role	3	Agent profile	Between agent	Descriptive persona-conditioned association; income groups are Low, Middle, and High.
Trust	Continuous, 0–100	Agent profile	Between agent	Descriptive persona trait association.
Risk tolerance	Continuous, 0–100	Agent profile	Between agent	Descriptive persona trait association.

Model	Contrast	Est.	95% CI
Within-agent multiplier	2.0× vs. 1.3×	−0.423	[−0.620, −0.230]
Within-agent multiplier	5.0× vs. 1.3×	0.502	[0.300, 0.710]
Decision LLM	Granite vs. Gemma	−2.008	[−2.600, −1.410]
Decision LLM	Phi4 Mini vs. Gemma	−6.145	[−6.850, −5.440]
Decision LLM	Qwen 3 4B vs. Gemma	2.498	[1.820, 3.180]
Two-part allocation	Phi4 Mini vs. Gemma	0.064 OR	[0.042, 0.099]
Two-part allocation	Ministral 3 3B vs. Gemma	0.498 OR	[0.318, 0.778]
Two-part amount	Qwen 3 4B vs. Gemma	2.891	[2.465, 3.316]
Two-part amount	Ministral 3 3B vs. Gemma	1.188	[0.818, 1.558]
Income gap (Gemma-written)	Low–High	−8.720	descriptive
Income gap (Granite-written)	Low–High	−0.911	descriptive

Table 4: Selected dependence-aware robustness results. Token coefficients are in contribution tokens except where marked OR. Within-agent multiplier rows compare the same agents across SimTree branches; decision-LLM rows use agent-clustered SEs; two-part rows separate allocation from conditional contribution amount. Income-gap rows are descriptive means; profile-generation context and run are confounded so no clustered CIs are reported.

The decision-LLM model is

$$y_{ibr} = \beta L_i + \delta M_b + \gamma_r + \theta P_i + \lambda D_i + \rho X_i + \epsilon_{ibr},$$

with profile-generation-context, multiplier, and round fixed effects, demographic controls, trust and risk-tolerance controls, and agent-clustered standard errors.

The persona/profile-generation model is

$$y_{ibr} = \beta D_i + \theta P_i + \kappa(D_i \times P_i) + \delta M_b + \gamma_r + \phi L_i + \rho X_i + \epsilon_{ibr}.$$

These estimates are descriptive because profile-generation context and run are confounded.

For the two-part model, the extensive margin is estimated as

$$\Pr(a_{ibr} = 1) = \text{logit}^{-1}(\beta L_i + \delta M_b + \gamma_r + \theta P_i + \lambda D_i + \rho X_i),$$

and the intensive margin among allocators is

$$y_{ibr} \mid a_{ibr} = 1 = \beta L_i + \delta M_b + \gamma_r + \theta P_i + \lambda D_i + \rho X_i + \epsilon_{ibr}.$$

B Prompt Templates

Below are representative prompts sent to the decision-making LLM during our PGG experiment, extracted from the Granite run. All prompts follow the five-section structure (S1–S5) described in Section 3.

B.1 Profile-Generation Prompt Template

The profile-generation LLM receives the system message *"Return only valid* JSON*"* and the following user prompt once per demographic archetype:

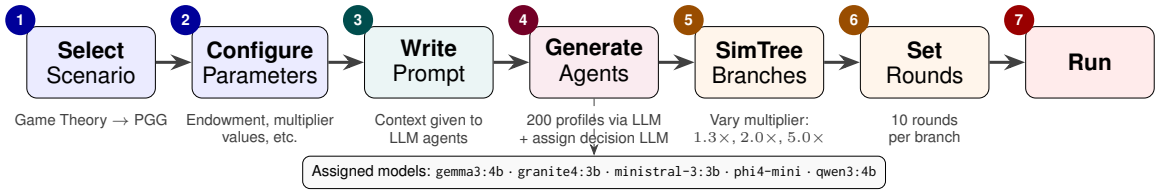


Figure 4: Appendix workflow diagram for Future of Society (FOS), illustrated with our PGG experiment. **(1)** Select a scenario from the library (e.g., PGG from Game Theory). **(2)** Configure parameters. **(3)** Write the scenario prompt that frames the game for agents. **(4)** Generate agents—the demographic generator creates profiles via LLM and assigns each agent a decision-making model. **(5)** Create SimTree branches to vary the multiplier across three conditions. **(6)** Set rounds (10 per branch). **(7)** Run.

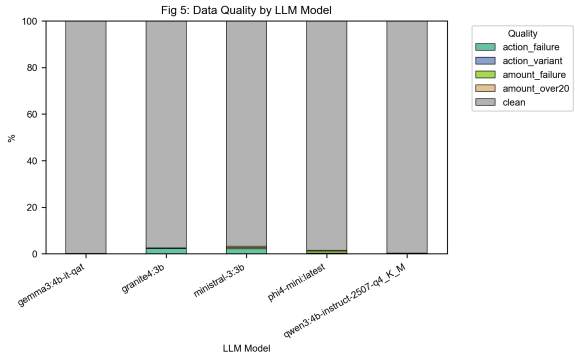


Figure 5: Structured-output audit by decision-making LLM. Clean responses dominate across models; non-clean categories are rare and are retained or excluded according to the robustness specification. This is a parseability measure, not an estimate of economic understanding.

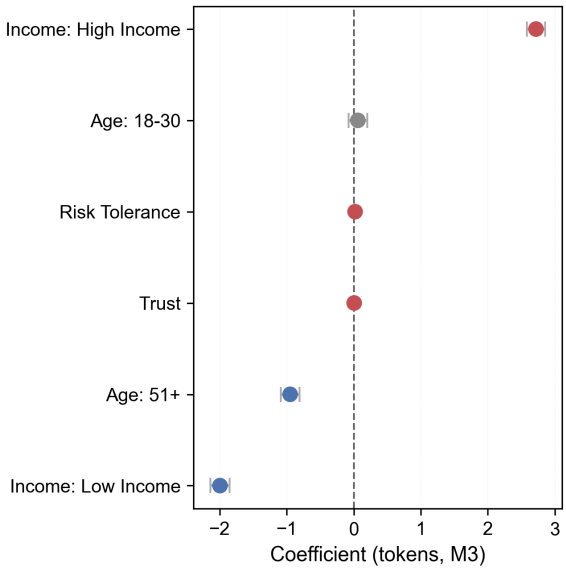


Figure 6: Persona-conditioned contribution shifts from the combined contribution model. Coefficients summarize how role-description fields and trait values are associated with contribution outputs after controlling for model and experimental context. Red indicates positive associations, blue indicates negative associations, and grey indicates confidence intervals crossing zero.

```
Create a one sentence
first-person bio for an agent
with these attributes:
Age: [18-30|31-50|51+], Income:
[Low|Middle|High Income]
Return JSON: {"description": "I
am...", "roles": ["Job Title 1",
..., "Job Title 5"]}
```

The LLM is called once per archetype (9 calls total). Trust and Risk Tolerance are appended after generation as Gaussian draws (μ=50, σ=15) and are not visible to the profile-generation LLM. Any downstream income effects therefore reflect model encodings of bare categorical labels, not rich narrative stereotypes.

B.2 Action Selection Prompt (Round 1)

```
=== EMBODY THIS PERSON ===
You are a Part-Time Student. I am resourceful and
determined to create opportunities for myself despite
the financial challenges I face. When responding,
think and react as this person would - not as a
neutral assistant.
Age: 18-30 | Income: Low Income | Trust: 60 (0-100)
| Risk Tolerance: 39 (0-100)

=== SECTION 2: SCENARIO ===
In this experiment, you receive 20 tokens each round.
Each person decides how much to contribute to a shared
pool. The pool is multiplied by 1.3 and distributed
equally among all 200 members. You keep any tokens
you do not allocate.

=== SECTION 3: AVAILABLE ACTIONS ===
- allocate: Allocate resources to the group account
- keep: Keep all your resources this round

=== SECTION 4: CONTEXT ===
This is the first round.

=== SECTION 5: JSON OUTPUT REQUIREMENT ===
Valid actions: "allocate", "keep"
Respond with ONLY JSON: {"action": "<action>"}
No markdown. No explanation. Only JSON.
```

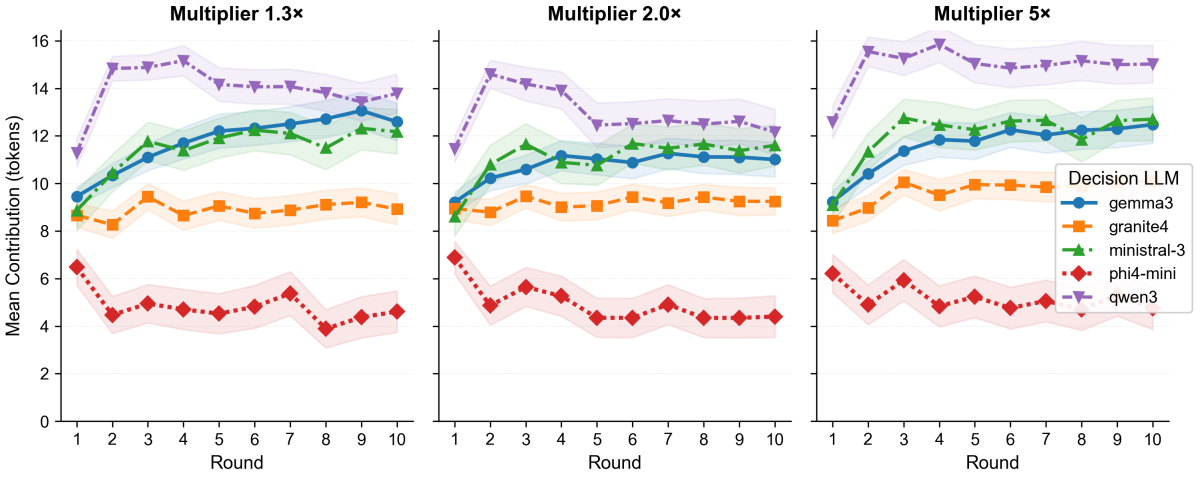


Figure 7: Contribution trajectories by decision-making LLM and multiplier condition. Each panel shows one multiplier branch, and each trajectory shows mean contribution across ten rounds for one decision-making LLM. The figure illustrates that Qwen 3 4B remains high-contributing and Phi4 Mini remains low-contributing across branches, while multiplier-specific shifts are smaller than the between-model differences.

Term	Est.	95% CI
Multiplier 2.0×	−1.037	[−1.456, −0.617]
Multiplier 5.0×	−0.204	[−0.633, 0.225]
Granite × 2.0×	1.305	[0.739, 1.871]
Ministral × 2.0×	0.625	[0.032, 1.219]
Phi4 Mini × 2.0×	1.147	[0.607, 1.688]
Qwen × 2.0×	−0.007	[−0.678, 0.664]
Granite × 5.0×	0.975	[0.360, 1.590]
Ministral × 5.0×	0.781	[0.173, 1.389]
Phi4 Mini × 5.0×	0.556	[−0.034, 1.146]
Qwen × 5.0×	1.182	[0.521, 1.842]

Table 5: Decision-LLM × multiplier interaction terms. All coefficients in contribution tokens; standard errors clustered by agent. Reference categories: Gemma 3 4B and multiplier 1.3×. The negative multiplier main effects reflect Gemma's response to the 2.0× branch; positive interaction terms indicate that other models partially offset this reduction.

Contrast	Est.	95% CI
2.0× vs. 1.3×	0.078	[−0.231, 0.387]
5.0× vs. 1.3×	0.153	[−0.198, 0.505]

Table 6: Round-1-only within-agent multiplier diagnostic. Estimates are contribution-token differences before branch-specific memory histories accumulate. Both CIs span zero, consistent with the full ten-round result. This diagnostic supports the interpretation that post-branch memory divergence is part of the treatment pathway rather than a confound driving the ten-round estimates.

B.3 Follow-Up Amount Prompt

If the agent chooses `allocate`, a second prompt repeats S1, S2, and S4 and replaces S3 with:

```
=== FOLLOW-UP PROMPT ===
You chose to allocate. This action requires parameters.
CRITICAL: amount must be integer between 0 and 20.
Respond ONLY with valid JSON:
{"action": "allocate", "amount": <0-20>}
No markdown. No explanation. Only JSON.
```

B.4 Action Prompt with Context Memory (Round 4)

From Round 2 onward, S4 is populated with the agent's memory window (Round 1 always included for primacy; last 3 rounds for recency):

```
=== SECTION 4: CONTEXT ===
Round 1: I allocate 10, Average contribution from 199
  other agents: 11.3. Total contribution: 1739.
  My payoff: 21.53.
Round 2: I allocate 15, Average contribution from 199
  other agents: 11.3. Total contribution: 1886.
  My payoff: 17.38.
Round 3: I allocate 20, Average contribution from 199
  other agents: 11.3. Total contribution: 2079.
  My payoff: 13.72.
My score: 52.63
```

B.5 Platform Walkthrough

To illustrate how FOS is used in practice, we walk through the setup for the PGG case study step by step. The full workflow is summarized in Figure 4; this section describes each stage concretely.

Scenario and parameter setup. The researcher opens the experiment editor and selects a scenario from the built-in library (Figure 8). For the PGG case study, the Public Goods Game is selected from the Game Theory category. The researcher then

Table 7: Local hardware and runtime configuration.

Item	Value
Execution backend	Ollama 0.23.2
CPU	Intel Core i9-14900HX
RAM	32GB
GPU	NVIDIA RTX 5070 Laptop GPU
Operating system	Windows 11
Total planned decisions	30,000
Total analyzable decisions	29,590
Total runtime	23.3 hours
API calls	0
Parallelism / concurrency	Sequential (single agent per call)
Repository URL	`Anonymous.URLomittedforreview.`
Processed data	`Anonymous.URLomittedforreview.`

configures parameters: resource name (*tokens*), endowment per round (20), pool multiplier (1.3 for the baseline branch), and round visibility (Simultaneous). The scenario description entered here is the text agents will see in their prompts.

Actions, agents, and network. The researcher selects which actions are available to agents (*allocate*, *keep*) and toggles off any actions not relevant to the design. Agent generation uses the Demographic Generation mode: the researcher specifies demographic dimensions (Age: 18–30, 31–50, 51+; Income: Low, Middle, High) and continuous traits (Trust and Risk Tolerance, drawn from $\mathcal{N}(50, 15)$, clipped to $[0, 100]$). This yields nine archetypes. Decision-making LLMs are assigned via the LLM Distribution panel, with each of the five models receiving 20% of the 200 agents. Clicking *Start Generating* calls the profile-generation LLM once per archetype and populates the agent list. The network is then set to Fully Connected, placing all 200 agents in a shared pool.

SimTree branching and execution. After the baseline is configured, the researcher opens the Causal Intervention Experiment Design panel, defines treatment variants by modifying the pool multiplier for each branch ($2.0\times$ for Group B, $5.0\times$ for Group C), and clicks *Start Batch Run (3 branches)*. FOS clones the pre-branch agent population — including profiles, assigned decision-making models, and configuration — across all three conditions. Within the SimTree view (Figure 9), the branching structure is displayed as a tree of nodes; the researcher sets the number of rounds to 10 and clicks *Auto-advance* to run all branches. Each agent decision is logged in real time with its raw response, parsed action, parser status, and branch identifier.

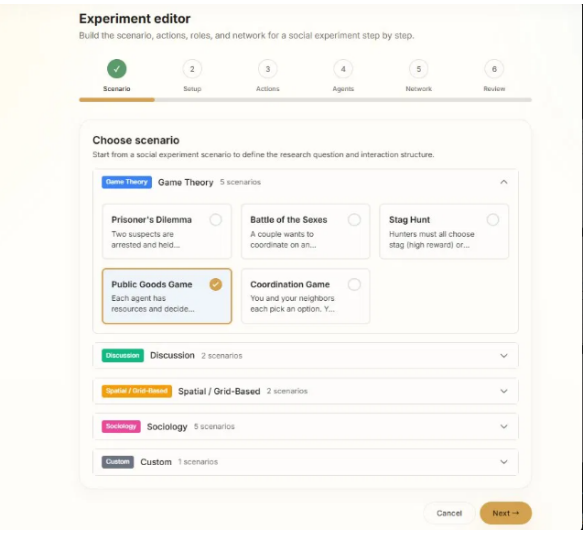


Figure 8: FOS experiment editor: scenario selection. The researcher selects a scenario from the built-in library; here the Public Goods Game is selected from the Game Theory category. Other categories include Discussion, Spatial/Grid-Based, Sociology, and Custom.

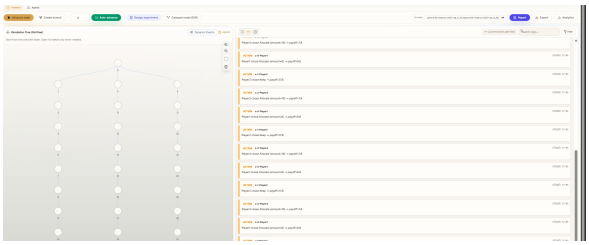


Figure 9: FOS SimTree view during execution. The left panel shows the branching structure: one pre-branch population (root node) cloned across three multiplier conditions, each advancing through ten rounds. The right panel shows the real-time audit log, recording each agent's action, contribution amount, and resulting payoff. (Shown with a small test population for illustrative clarity; the PGG case study used 200 agents.)

C Use of AI Writing and Coding Assistants

The authors used large language model assistants (including Claude and ChatGPT) during the preparation of this manuscript. AI assistance was used for assisting in writing code, spelling and grammar checking, and to help improve the readability of written sections. All substantive intellectual contributions, including research design, experimental methodology, analysis, and conclusions, are solely

the work of the authors. All AI-generated suggestions were reviewed and verified by the authors.