When Reasoning Models Hurt Behavioral Simulation: A Solver-Sampler Mismatch in Multi-Agent LLM Negotiation

Sandro Andric
sandro.andric@nyu.edu

arXiv:2604.11840v1 [cs.LG] 12 Apr 2026

Abstract

Large language models are increasingly used as agents in social, economic, and policy simulations. A common assumption is that stronger reasoning should improve simulation fidelity. We argue that this assumption can fail when the objective is not to solve a strategic problem, but to sample plausible boundedly rational behavior. In such settings, reasoning-enhanced models can become better solvers and worse simulators: they can over-optimize for strategically dominant actions, collapse compromise-oriented terminal behavior, and sometimes exhibit a diversity-without-fidelity pattern in which local variation survives without outcome-level fidelity. We study this solver-sampler mismatch in three multi-agent negotiation environments adapted from earlier simulation work: an ambiguous fragmented-authority trading-limits scenario, an ambiguous unified-opposition trading-limits scenario, and a new-domain grid-curtailment case in emergency electricity management. We compare three reflection conditions, no reflection, bounded reflection, and native reasoning, across two primary model families and then extend the same protocol to direct OpenAI runs with GPT-4.1 and GPT-5.2. Across all three experiments, bounded reflection produces substantially more diverse and compromise-oriented trajectories than either no reflection or native reasoning. In the direct OpenAI extension, GPT-5.2 native ends in authority decisions in 45 of 45 runs across the three experiments, while GPT-5.2 bounded recovers compromise outcomes in every environment. The contribution is not a claim that reasoning is generally harmful. It is a methodological warning: model capability and simulation fidelity are different objectives, and behavioral simulation should qualify models as samplers, not only as solvers.

Introduction

Behavioral simulation and strategic problem solving are not the same task. A model that performs well on benchmarks rewarding correct reasoning, long-horizon planning, or strategic optimization may still be a poor instrument for simulating human behavior. This matters because large language models are increasingly deployed as agents in synthetic societies, negotiation simulators, policy-analysis tools, and economic games (Bonabeau 2002; Epstein and Axtell 1996; Aher, Arriaga, and Kalai 2023; Horton 2023). In these settings, the practical goal is often not to discover the optimal strategy from first principles. It is to sample a plausible distribution of boundedly rational trajectories under uncertainty (Simon 1955; Kahneman 2003; Rubinstein 1998).

That objective creates a tension that is easy to miss. If the target population is boundedly rational, then a model optimized for extended reasoning may become misspecified as a simulator. Instead of producing diverse, path-dependent trajectories shaped by incomplete information, satisficing, pressure, and timing, it may infer the strategic structure too cleanly and converge on a narrow policy. In a multi-agent setting, that shift can collapse the very variation that simulation is supposed to reveal.

We study this tension as a measurement problem. Recent work on generative social simulation argues that validation, rather than raw capability, is the central challenge for this entire research area (Larooij and Törnberg 2026; Collins, Koehler, and Lynch 2024; Wallach et al. 2025; Zhou et al. 2024; Hullman et al. 2026). We call the specific failure mode analyzed here the solver-sampler mismatch: the mismatch between the capabilities that make a model a stronger strategic solver and the properties required for it to act as a good behavioral sampler. A good solver seeks dominance, consistency, and efficiency. A good sampler, by contrast, must preserve bounded-rational variation. It must allow agents to concede late, misread leverage, settle on suboptimal but plausible arrangements, or fail to converge at all. When those behaviors disappear, simulation can become cleaner while becoming less faithful.

Our starting point comes from a broader line of multi-agent simulation work we refer to as Narrative Monte Carlo, in which LLM agents are treated as stochastic samplers over plausible negotiation trajectories. Within that work, a recurring empirical pattern emerged: models with stronger native reasoning frequently produced degenerate simulations. They favored a narrow subset of actions, rarely conceded, and often drove interactions to rigid terminal states. By contrast, bounded reflection mechanisms improved memory and convergence without inducing the same rigidity. This paper isolates that thread as a standalone methodological question rather than as a subcomponent of a larger simulation project.

This paper makes four contributions.

1. It formalizes the distinction between solver quality and sampler quality in multi-agent simulation.
2. It proposes a practical metric framework for behavioral

sampler fidelity.

3. It presents evidence from three scenario-based, institutionally structured multi-agent negotiation experiments spanning two coalition structures and two substantive domains, plus a direct OpenAI provider extension, showing that bounded reflection can improve simulation fidelity while native reasoning can degrade it.
4. It argues that model selection for behavioral simulation should optimize sampler fidelity rather than raw benchmark capability.

The claims here are intentionally narrow. We do not argue that reasoning is generally harmful or that reasoning models are universally inferior. We argue that, in bounded-rational multi-agent simulation settings, stronger native reasoning can degrade behavioral simulation fidelity under the conditions studied here.

Why Solvers Are Not Necessarily Samplers

The core mistake in many LLM simulation pipelines is objective slippage. Simulation asks one question: what range of plausible trajectories might agents produce under a structured set of conditions? Strategic problem solving asks another: what is the best action sequence under those conditions? These are related but not identical objectives.

If a model is rewarded or tuned for coherent long-horizon optimization, it may reinterpret a simulation prompt as a task to be solved. In multi-agent settings, this often manifests as:

- repeated use of the same dominant action
- suppressed concession behavior
- rapid convergence on a single normative interpretation of the game
- failure to preserve the frictions, misalignments, and partial commitments that create plausible bounded-rational trajectories

For social or policy simulation, that is an instrument-selection problem. The model is not failing in an absolute sense. It may be performing exactly as a strong strategic reasoner should. The failure is relative to the task of sampling boundedly rational behavior.

This framing matters because it shifts the question from ”which model is smartest?” to ”which model best matches the behavioral target?” A strong decision-support model may be the wrong simulation model. A weaker but more behaviorally diverse model may be the better simulator.

Behavioral Sampler Fidelity

We define behavioral sampler fidelity as the degree to which a model preserves the plausible variation, negotiation dynamics, and bounded-rational path dependence required by the target simulation task.

Outcome quality alone is insufficient for this purpose. A model can arrive at an apparently sensible final agreement while still producing a behaviorally implausible trajectory. For that reason, we evaluate sampler fidelity with trajectory-level diagnostics rather than endpoint accuracy alone.

These diagnostics should be read as necessary but not sufficient conditions for fidelity. High action entropy alone does not prove faithful simulation. Our claim is narrower and more defensible: near-zero diversity, zero concession, and universal turn-budget exhaustion are inconsistent with any plausible account of bounded-rational negotiation in the environments studied here. In that sense, the metrics are best understood as falsification tools for sampler failure rather than as complete validation criteria.

Primary Metrics

We use three primary metrics.

1. Action entropy Measures behavioral diversity within trajectories. Low entropy indicates repetitive strategic play; higher entropy indicates a richer mix of negotiation actions.
2. Concession arc rate Measures whether a run contains meaningful concession behavior. Operationally, a run counts as containing a concession arc when the same agent first rejects or counters and later concedes or supports. This captures whether agents soften, adapt, or partially revise positions during negotiation rather than simply repeating hardline counters.
3. Max-turn exhaustion rate Measures how often interactions hit the turn budget rather than resolving earlier. High exhaustion is consistent with rigid play and poor adaptive behavior.

Secondary Metrics

We also report:

- Trajectory diversity
- Outcome entropy
- Parse success rate
- Average provider-error turns
- Average format-error turns

The last two are important because a model that cannot reliably stay within the simulation protocol is a weaker practical instrument even if some individual runs are behaviorally interesting.

Experimental Setting

Environments

The completed study contains three scenario-based multi-agent experiments. The first two vary structural conditions within the same trading-limits case family. The third preserves the institutional negotiation skeleton while transferring the experiment into a new emergency grid-curtailment domain.

Experiment 1 uses an ambiguous fragmented-authority trading-limits scenario designed to create nontrivial bargaining pressure across multiple parties.

This environment is well suited to the main claim because it requires:

- multi-party negotiation rather than bilateral optimization
- partial convergence rather than simple win-loss scoring

- concession and coalition dynamics rather than one-shot choice
- boundedly rational adaptation under ambiguity

Experiment 2 uses an ambiguous unified-opposition scenario from the same trading-limits case family. This preserves ambiguous authority while changing the coalition structure from fragmented opposition to unified opposition. It therefore tests whether the main result depends on the exact fragmentation structure of Experiment 1 while staying inside the same institutional negotiation regime.

Experiment 3 uses a new grid-curtailment case. This environment shifts the substantive domain from financial-market restrictions to emergency electricity curtailment while preserving the core experimental grammar:

- multiple institutional actors with heterogeneous mandates
- ambiguous authority and fallback decree pressure
- issue bundling rather than single-issue optimization
- path-dependent concessions under time pressure
- compromise, consensus, and authority-imposed terminal outcomes

This design tests whether the solver-sampler mismatch survives domain transfer when the negotiation architecture stays fixed.

Reflection Conditions

We compare three conditions.

1. No reflection No structured private reflection beyond the main prompt context.
2. Bounded reflection A short structured private ledger designed to mimic a limited cognitive horizon. In the actual experiment stack, that ledger contains five fields tracking one's own concessions, the other side's concessions, current state, opponent assessment, and open issues. This is not open-ended chain-of-thought. It is a constrained memory scaffold.
3. Native reasoning Provider or model native reasoning mode. In the completed experiments, Gemini native used medium reasoning effort, while DeepSeek native used provider-native reasoning explicitly enabled; the corresponding DeepSeek no-reflection and bounded-reflection conditions explicitly disabled that path.

This distinction is conceptually central. Our claim is not that all internal reflection harms simulation. Our claim is that bounded reflection and native long-horizon reasoning are different mechanisms with different behavioral consequences. Exact provider settings for Gemini, DeepSeek, and the OpenAI extension are listed in the protocol appendix so that `native` is inspected as a provider-specific treatment rather than treated as a single homogeneous mechanism.

Model Families

The primary experiment matrix uses two model families:

- Gemini 3.1 Flash Lite Preview
- DeepSeek V3.2

Each family was evaluated under no reflection, bounded reflection, and native reasoning, with 15 runs per condition.

After freezing that core matrix, we ran a direct-OpenAI extension through the same three experiments:

- GPT-4.1 under no reflection and bounded reflection
- GPT-5.2 under no reflection, bounded reflection, and native reasoning

We treat these OpenAI runs as a breadth extension rather than as the paper's primary identification layer. They were added after the primary matrix was complete, but they use the same environments, protocol, and aggregation schema.

For quick inspection, the provider-specific `native` settings used in the completed study are:

- Gemini 3.1 Flash Lite Preview: no explicit reasoning payload in `none` and `bounded`; provider-native reasoning with medium effort in `native`
- DeepSeek V3.2: provider reasoning explicitly disabled in `none` and `bounded`; provider reasoning explicitly enabled in `native`
- GPT-5.2: OpenAI Responses API with `reasoning.effort = "none"` in `none` and `bounded`; `reasoning.effort = "high"` in `native`

Token floors, timeout floors, and API-path details appear in the protocol appendix.

Analysis Procedure

All runs were aggregated through a common schema and summarized with bootstrap confidence intervals. Operational failures were separated into provider-error turns and format-error turns to avoid conflating instrumentation breakdowns with behavioral outcomes. An anonymized supplementary PDF bundles the exact prompts, reflection scaffolds, outcome taxonomy, transcript excerpts, and error-excluded robustness tables.

The full study scale is 2 primary model families x 3 reflection conditions x 3 experiments x 15 runs = 270 core runs, plus 225 direct-OpenAI extension runs, for 495 completed runs in total.

To make the study self-contained, we summarize the core design here. The experiments are an ambiguous fragmented-authority trading-limits scenario, an ambiguous unified-opposition trading-limits scenario, and a grid-curtailment case. The primary families are Gemini 3.1 Flash Lite Preview and DeepSeek V3.2, each evaluated under no reflection, bounded reflection, and native reasoning with 15 runs per cell. The direct-OpenAI extension adds GPT-4.1 under no reflection and bounded reflection and GPT-5.2 under no reflection, bounded reflection, and native reasoning, bringing the total completed run count to 495.

The public action schema is fixed across experiments:

- **Support**: endorse the current proposal or direction
- **Oppose**: reject the proposal without offering a concrete counter

- **Concede**: soften an earlier position or accept another actor's demand
- **Counter**: reject while proposing a modified alternative
- **Exit**: leave the negotiation or force escalation

The terminal outcome taxonomy is likewise fixed:

- **Compromise**: negotiated middle-ground resolution before the turn cap
- **Consensus**: broad agreement with no material unresolved dispute
- **Authority decision**: fallback institutional decision after unresolved issues remain
- **Deadlock**: no agreement and no constructive fallback resolution; in the completed matrix this category remained unused because unresolved runs flowed into the common authority-decision fallback

Deadlock remains part of the shared taxonomy, but no completed run in the 495-run matrix ended in deadlock; unresolved trajectories in these experiments instead flowed into authority decision under the common fallback rule.

That fallback is related to, but not identical with, max-turn exhaustion. Max-turn exhaustion is a process diagnostic that records whether a run reached the configured turn cap. Authority decision is the terminal outcome label applied when compromise or consensus has not been reached and the designated authority remains active under the shared classifier rule. Many rigid runs exhibit both, but they should not be read as the same variable.

In compact form, the common fallback rule is:

1. no compromise or consensus has yet been detected
2. at least five turns have elapsed
3. the designated authority remains active in the interaction

If those conditions hold at termination, the run is classified as `authority_decision`.

Statistical Analysis

The unit of analysis is the run. Each condition contains 15 independent runs. We treat three endpoints as primary:

1. Action entropy
2. Concession arc rate
3. Max-turn exhaustion rate

Action entropy is the run-level Shannon entropy over the empirical action-type distribution within a run, `H = -sum_a p(a) log2 p(a)`, with no vocabulary-size normalization. Exact metric definitions for action entropy, concession arcs, max-turn exhaustion, and parse success appear in the statistical appendix. For each condition, we report bootstrap 95% confidence intervals over runs. For pairwise condition contrasts within a model family, we use two-sided permutation tests on run-level means with 10,000 random label shuffles. We also report Cliff's delta as a nonparametric effect-size measure. Secondary metrics, including parse success and protocol-error counts, are reported descriptively and used to interpret operational reliability rather than as the main basis of the behavioral claim.

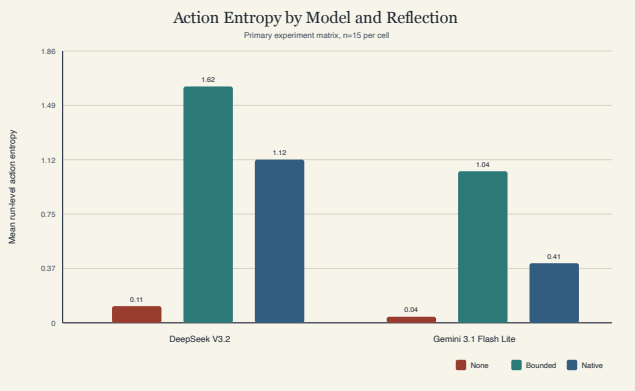

Figure 1: Action entropy by model family and reflection mode.

Because this manuscript is built from a structured experimental matrix rather than a preregistered confirmatory trial, the permutation-test results should be interpreted as inferential support for the observed pattern, not as the sole basis of the paper's conclusions. The strongest evidence remains the consistency of the directional pattern across primary endpoints and across model families. We also report an error-excluded robustness reanalysis that drops any run with at least one provider-error turn or format-error turn, plus two post hoc robustness checks asking whether temperature alone can recover sampler quality and whether the bounded-ledger effect survives prompt-structure changes without altering the action schema.

Results

The Main Pattern

The main pattern is clear across both model families:

- no reflection is rigid and collapses to authority decision
- bounded reflection produces substantially more diverse and compromise-oriented trajectories
- native reasoning does not recover bounded-sampler quality and remains rigid at the outcome level

This pattern is strongest and cleanest in the bounded condition.

Gemini Results

Gemini provides the clearest anchor result.

Under no reflection, Gemini is almost completely rigid:

- action entropy: 0.041 [0.000, 0.102]
- concession arc rate: 0.000 [0.000, 0.000]
- max-turn exhaustion rate: 1.000 [1.000, 1.000]
- outcome distribution: authority decision in 15 of 15 runs

Under bounded reflection, the same family behaves very differently:

- action entropy: 1.040 [0.955, 1.142]
- concession arc rate: 1.000 [1.000, 1.000]

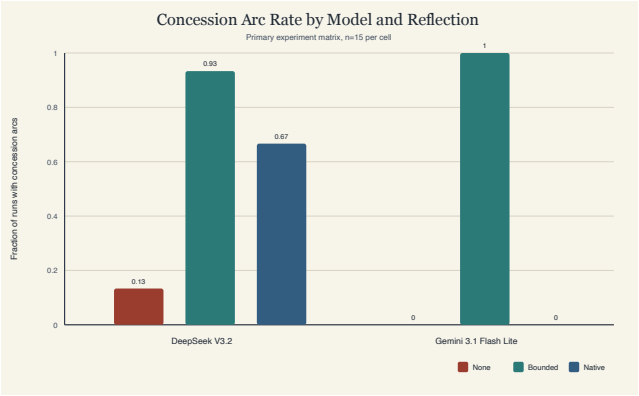

Figure 2: Concession-arc rate by model family and reflection mode.

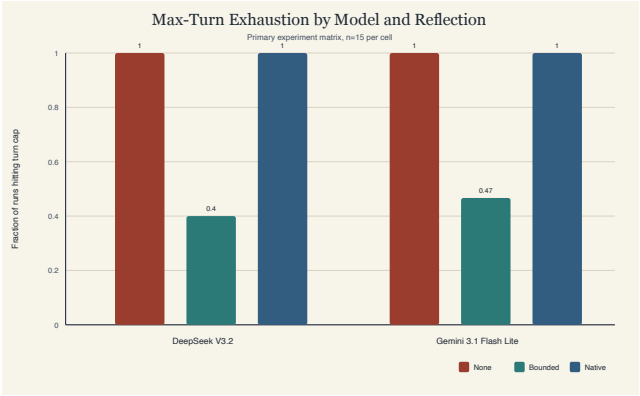

Figure 3: Max-turn exhaustion by model family and reflection mode.

- max-turn exhaustion rate: 0.467 [0.200, 0.733]
- parse success rate: 1.000 [1.000, 1.000]
- outcome distribution: compromise in 15 of 15 runs

Under native reasoning, Gemini returns to rigid outcomes and becomes noisier operationally:

- action entropy: 0.409 [0.282, 0.521]
- concession arc rate: 0.000 [0.000, 0.000]
- max-turn exhaustion rate: 1.000 [1.000, 1.000]
- parse success rate: 0.267 [0.067, 0.533]
- average provider-error turns: 0.867 [0.533, 1.200]
- outcome distribution: authority decision in 15 of 15 runs

The key point is not only that bounded performs better than native. It is that bounded is both more behaviorally plausible and more operationally stable as a sampler.

The pairwise contrasts support that reading. Relative to no reflection, bounded reflection increases mean action entropy by 0.968, increases concession-arc rate by 1.000, and reduces max-turn exhaustion by 0.533. All three primary contrasts are supported by permutation tests ($p <= 0.0019$) with large effect sizes. Relative to native reasoning, bounded reflection again improves all three primary endpoints ($p <= 0.0019$), while native reasoning also shows a large operational reliability penalty on parse success (0.267 vs 1.000, permutation $p = 0.0001$).

DeepSeek Results

DeepSeek replicates the same directional contrast.

Under no reflection, DeepSeek is again rigid:

- action entropy: 0.114 [0.000, 0.250]
- concession arc rate: 0.133 [0.000, 0.333]
- max-turn exhaustion rate: 1.000 [1.000, 1.000]
- parse success rate: 1.000 [1.000, 1.000]
- outcome distribution: authority decision in 15 of 15 runs

Under bounded reflection, DeepSeek becomes substantially more sampler-like:

- action entropy: 1.622 [1.425, 1.775]
- concession arc rate: 0.933 [0.800, 1.000]
- max-turn exhaustion rate: 0.400 [0.133, 0.600]
- parse success rate: 0.667 [0.400, 0.867]
- outcome distribution: 12 compromise, 2 consensus, and 1 authority decision

Under native reasoning, DeepSeek shows more internal behavioral variation than no reflection, but it does not translate that variation into better terminal simulation outcomes:

- action entropy: 1.120 [0.843, 1.306]
- concession arc rate: 0.667 [0.400, 0.933]
- max-turn exhaustion rate: 1.000 [1.000, 1.000]
- parse success rate: 0.467 [0.200, 0.733]
- outcome distribution: authority decision in 15 of 15 runs

This is a useful nuance. Native reasoning does not always collapse all intermediate variation. But in this environment it still fails to produce the compromise-oriented terminal behavior that bounded reflection supports, and it does so with materially worse protocol reliability.

The DeepSeek contrasts support the same broad conclusion. Relative to no reflection, bounded reflection increases mean action entropy by 1.265, increases concession-arc rate by 0.800, and reduces max-turn exhaustion by 0.600; each of these primary contrasts is supported by permutation tests ($p <= 0.0008$) with large effect sizes. Relative to native reasoning, bounded reflection still improves action entropy (+0.464, $p = 0.0036$) and max-turn exhaustion (-0.600, $p = 0.0008$). The concession-arc contrast remains directionally favorable to bounded reflection but is weaker in this sample (0.933 vs 0.667, permutation $p = 0.1703$). This is consistent with the qualitative read of DeepSeek native reasoning: it introduces more local variation than no reflection, but still fails to deliver compromise-oriented terminal behavior.

Cross-Condition Interpretation

The cross-condition story is therefore not simply ”reasoning reduces entropy.” That would be too coarse. The actual pattern is sharper:

- no reflection often produces clean but rigid hardline play

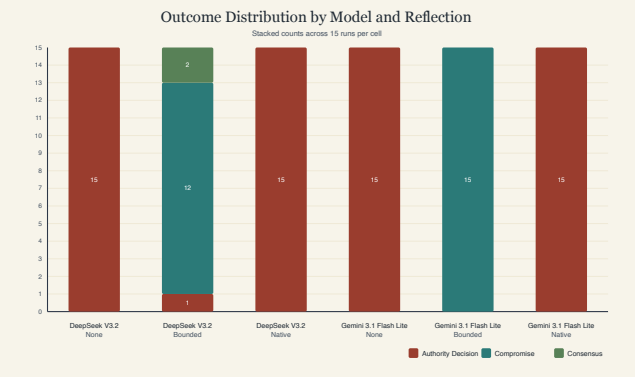

Figure 4: Outcome distribution across the six Experiment 1 cells.

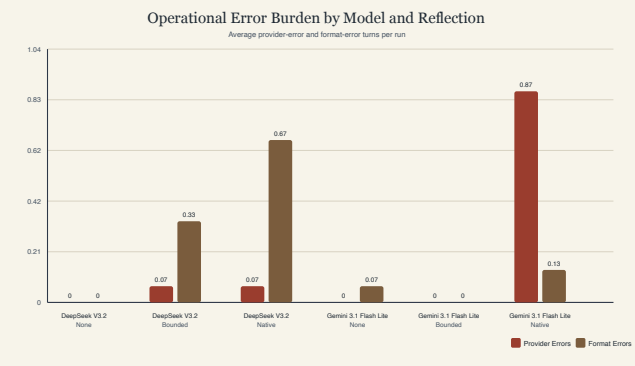

Figure 5: Operational error burden, showing why native reasoning is also a weaker practical instrument even when some internal variation remains.

- bounded reflection increases diversity, concession, and negotiated resolution
- native reasoning can reintroduce some internal variation but still behaves like a poor sampler at the outcome level and a noisy instrument operationally

This distinction is important because it suggests that the main problem is not reflection itself. The problem is a mismatch between open-ended strategic reasoning and the boundedly rational behavior we want to simulate.

Experiment 2: Ambiguous Unified Opposition

The second experiment shows that the Experiment 1 result is not confined to one exact scenario cell.

In the unified-opposition trading-limits experiment, the same cross-condition pattern reappears across both model families:

- no reflection remains rigid and ends in authority decision in 15 of 15 runs for both Gemini and DeepSeek
- bounded reflection again produces compromise-oriented behavior, with 14 compromise and 1 consensus run for Gemini and 13 compromise and 2 authority-decision runs for DeepSeek
- native reasoning again returns to authority decision in 15 of 15 runs for both families, while remaining operationally noisier than bounded

The metric profile is also consistent with the main experiment. For Gemini:

- no reflection: action entropy 0.222 [0.073, 0.353], concession arc rate 0.333 [0.067, 0.600], max-turn exhaustion 1.000 [1.000, 1.000]
- bounded reflection: action entropy 1.186 [1.043, 1.308], concession arc rate 1.000 [1.000, 1.000], max-turn exhaustion 0.667 [0.400, 0.867]
- native reasoning: action entropy 0.359 [0.222, 0.471], concession arc rate 0.067 [0.000, 0.200], max-turn exhaustion 1.000 [1.000, 1.000]

For DeepSeek:

- no reflection: action entropy 0.303 [0.000, 0.562], concession arc rate 0.067 [0.000, 0.200], max-turn exhaustion 1.000 [1.000, 1.000]
- bounded reflection: action entropy 1.518 [1.305, 1.652], concession arc rate 1.000 [1.000, 1.000], max-turn exhaustion 0.533 [0.267, 0.800]
- native reasoning: action entropy 1.275 [1.058, 1.435], concession arc rate 0.733 [0.467, 0.933], max-turn exhaustion 1.000 [1.000, 1.000]

Experiment 2 shows the same directional pattern under a different coalition structure. The core result is no longer confined to one scenario cell: bounded remains the strongest sampler condition, while native remains outcome-rigid and operationally noisier.

The run-level contrasts support the same interpretation. For Gemini in Experiment 2, bounded reflection outperforms no reflection on action entropy (1.081 vs 0.160, permutation $p < 0.0001$) and concession arcs (1.000 vs 0.333, $p = 0.0002$), and outperforms native reasoning on all three primary endpoints, including exhaustion (0.667 vs 1.000, $p = 0.0449$). For DeepSeek, bounded reflection again dominates no reflection on all three primary endpoints ($p <= 0.0058$) and improves over native reasoning most clearly on exhaustion (0.533 vs 1.000, $p = 0.0058$) and parse success (0.733 vs 0.133, $p = 0.0015$), while the action-entropy and concession contrasts are directionally favorable but weaker in this sample.

Experiment 3: New-Domain Grid Curtailment Transfer

The third experiment asks whether the effect survives a substantive domain transfer when the experimental structure remains the same.

It does. This is not merely another scenario variant. The grid-curtailment case replaces financial-market actors with electricity-system and public-interest actors, changes the substantive stakes from trading restrictions to emergency load shedding, and introduces a different time-pressure structure tied to near-term infrastructure reliability rather than market policy. The experimental grammar remains the

same, but the domain semantics and institutional roles are genuinely different.

In the emergency grid-curtailment environment, the same cross-condition pattern reappears across both model families:

- Gemini with no reflection: 15 of 15 runs end in authority decision
- Gemini with bounded reflection: 13 compromise runs and 2 authority-decision runs
- Gemini with native reasoning: 15 of 15 runs end in authority decision
- DeepSeek with no reflection: 15 of 15 runs end in authority decision
- DeepSeek with bounded reflection: 12 compromise runs, 2 consensus runs, and 1 authority-decision run
- DeepSeek with native reasoning: 15 of 15 runs end in authority decision

The metric profile again favors bounded reflection. For Gemini:

- no reflection: action entropy 0.260 [0.041, 0.516], concession arc rate 0.267 [0.067, 0.533], max-turn exhaustion 0.933 [0.800, 1.000]
- bounded reflection: action entropy 0.956 [0.829, 1.064], concession arc rate 1.000 [1.000, 1.000], max-turn exhaustion 0.800 [0.533, 1.000]
- native reasoning: action entropy 0.202 [0.082, 0.301], concession arc rate 0.267 [0.067, 0.533], max-turn exhaustion 1.000 [1.000, 1.000]

For DeepSeek:

- no reflection: action entropy 0.277 [0.073, 0.458], concession arc rate 0.200 [0.000, 0.400], max-turn exhaustion 1.000 [1.000, 1.000]
- bounded reflection: action entropy 1.809 [1.607, 1.901], concession arc rate 1.000 [1.000, 1.000], max-turn exhaustion 0.400 [0.133, 0.600]
- native reasoning: action entropy 1.476 [1.279, 1.612], concession arc rate 0.933 [0.800, 1.000], max-turn exhaustion 1.000 [1.000, 1.000]

The Diversity-Without-Fidelity Pattern Experiment 3 also contains the sharpest falsification of the idea that more internal variation automatically implies better simulation fidelity. DeepSeek native reasoning shows high action entropy (1.476 [1.279, 1.612]) and a very high concession-arc rate (0.933 [0.800, 1.000]), but still ends in authority decision in 15 of 15 runs. In other words, the local variation is present and the concession behavior is present, yet compromise-oriented terminal behavior never appears. Bounded reflection is the only condition that converts adaptation into negotiated resolution.

This is the strongest diversity-without-fidelity pattern in the paper. It rules out a simplistic reading in which native reasoning only fails because it suppresses all variation. In this experiment, native reasoning preserves substantial within-trajectory variation and still fails as a sampler at the outcome level. The contrast is statistically clear on the most consequential endpoint: for DeepSeek, bounded reflection reduces max-turn exhaustion from 1.000 to 0.400 relative to native reasoning ($p = 0.0006$, Cliff's delta 0.600) while shifting terminal outcomes from 15 of 15 authority decisions to 12 compromise runs, 2 consensus runs, and 1 authority decision. For Gemini, bounded reflection also strongly separates from native reasoning on action entropy (0.956 vs 0.202, permutation $p = 0.0001$) and concession arcs (1.000 vs 0.267, $p = 0.0002$).

That makes Experiment 3 especially valuable: it shows that the core claim is not confined to one case family cell, one coalition structure, or one substantive policy domain.

The error-excluded reanalysis leaves that directional conclusion intact. After dropping every run with at least one provider-error turn or format-error turn, retained bounded cells continue to concentrate on compromise-oriented outcomes, while retained no-reflection and native-reasoning cells continue to concentrate on authority decisions. The main change is effective sample size, especially for native conditions, because operational fragility is itself concentrated in native reasoning.

The retained primary-family tallies are compact enough to summarize directly:

- Experiment 1, Gemini: bounded `C15`; native `A4`
- Experiment 1, DeepSeek: bounded `A1/C7/S2`; native `A7`
- Experiment 2, Gemini: bounded `C14/S1`; native `A5`
- Experiment 2, DeepSeek: bounded `A2/C9`; native `A2`
- Experiment 3, Gemini: bounded `A2/C13`; native `A13`
- Experiment 3, DeepSeek: bounded `C3`; native `A5`

Retained no-reflection runs remain authority-dominated in every primary cell as well: `A14` or `A15` in Experiments 1 and 2, and `A14` or `A15` again in Experiment 3. So the clean-run result is not merely that native reasoning is error-prone. It is that even after excluding error-affected runs, bounded remains the only condition that consistently retains compromise-oriented outcomes.

Taken together, the transfer evidence is compact. Across Experiments 2 and 3, Gemini under no reflection and native reasoning remains authority-dominated while Gemini under bounded reflection remains compromise-oriented. DeepSeek under no reflection remains authority-dominated, DeepSeek under bounded reflection remains compromise-oriented, and DeepSeek under native reasoning shows the strongest diversity-without-fidelity pattern in Experiment 3. The OpenAI extension fits the same template: no reflection remains rigid, bounded reflection opens compromise outcomes, and GPT-5.2 native remains authority-dominated across all three experiments.

Targeted Robustness Checks

The first targeted robustness question after the three main experiments is whether bounded reflection only works because it injects more randomness. A compact temperature sweep on the Gemini anchor family argues against that reading. In Experiment 1, changing temperature from 0.3 to 0.7

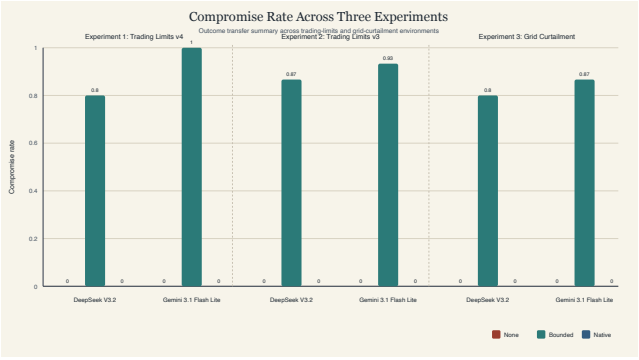

Figure 6: Compromise-rate transfer across all three experiments.

to 1.0 does not alter the outcome pattern: no reflection remains 10 of 10 authority decisions at all three temperatures, bounded reflection remains 10 of 10 compromises, and native reasoning remains 10 of 10 authority decisions. Action entropy shifts somewhat within conditions, but the terminal distribution does not.

That result matters methodologically. If higher stochasticity alone were enough to rescue the rigid sampler conditions, then at least some no-reflection or native-reasoning cells should have crossed into compromise-oriented outcomes at temperature 1.0. They do not. The bounded-reflection condition therefore appears to be doing more than adding noise.

The bounded-ledger sensitivity check points in the same direction, but more specifically. We reran Gemini on Experiment 1 with three bounded-ledger variants that preserve the same five state fields and the same public action schema while varying instruction tightness. The current ledger remains compromise-oriented with 9 compromise runs and 1 consensus run, and a more permissive ledger remains compromise-oriented with 10 compromise runs. But a compact ledger that truncates the private state too aggressively collapses toward authority decision, with 9 authority-decision runs and only 1 compromise run.

This is a useful refinement of the paper's intervention claim. The bounded effect is not an artifact of one magical prompt. But neither is it produced by any arbitrary note-taking scaffold. The private state must remain bounded yet sufficiently expressive to track concessions, opponent flexibility, and open issues across turns. At the same time, this does not fully remove the normative-bias concern: a ledger that explicitly tracks concessions and opponent assessment may still privilege compromise-relevant state variables over more adversarial strategic bookkeeping. We therefore interpret the sensitivity result as evidence that the effect is not tied to one exact wording, not as proof that the current ledger family is normatively neutral.

Finally, we added one more trajectory-level diagnostic using the completed logs: first concession timing, defined as the earliest turn where any agent softens from rejection or countering into concession or support. This metric sharpens the distinction between genuine negotiation dynamics and late cosmetic movement. Across all three experiments, bounded-reflection conditions concede earlier and more consistently than the corresponding no-reflection and native-reasoning conditions. For example, in Experiment 2 the mean first concession turn is 8.067 for Gemini with bounded reflection versus 13.000 for Gemini with no reflection and 14.000 for Gemini with native reasoning; for DeepSeek it is 7.400 for bounded reflection versus 14.000 for no reflection and 9.818 for native reasoning.

Direct OpenAI Provider Extension

The remaining generality question after the three main experiments is whether the pattern survives a third provider family. To test that, we ran a direct-OpenAI extension with GPT-4.1 and GPT-5.2 on the same three experiments.

The extension preserves the main qualitative result.

- GPT-4.1 with no reflection is rigid or nearly rigid across all three experiments: 14 of 15 authority-decision runs in Experiment 1, 15 of 15 in Experiment 2, and 15 of 15 in Experiment 3.
- GPT-4.1 with bounded reflection materially improves compromise-oriented outcomes in all three experiments: 9 compromise and 6 authority-decision runs in Experiment 1, 10 compromise and 5 authority-decision runs in Experiment 2, and 13 compromise and 2 authority-decision runs in Experiment 3.
- GPT-5.2 with no reflection is fully rigid across all three experiments: 15 of 15 authority-decision runs in every case.
- GPT-5.2 with native reasoning is also fully rigid across all three experiments: 15 of 15 authority-decision runs in every case.
- GPT-5.2 with bounded reflection partly recovers compromise behavior in the two trading-limits experiments and strongly recovers it in the grid-curtailment transfer: 5 compromise and 10 authority-decision runs in Experiment 1, 7 compromise and 8 authority-decision runs in Experiment 2, and 13 compromise and 2 authority-decision runs in Experiment 3.

The OpenAI extension broadens provider coverage beyond the Gemini and DeepSeek families used in the primary matrix. It also shows that the solver-sampler mismatch is not just a story about one provider-specific native-reasoning implementation. Even with a different API surface and a separate direct-provider path, no reflection and native reasoning remain rigid, while bounded reflection remains the only condition that reliably opens compromise-oriented trajectories.

We also ran the missing crossed mechanism cell for `gpt-5.2` on Experiment 1 by combining the bounded ledger with provider-native reasoning. In compact form, the resulting 2x2 looks like this:

This sharpened the mechanism story rather than weakening it. The ledger can still induce some local variation under native reasoning, but in this family it does not recover concession behavior or terminal compromise once native reasoning is enabled.

The OpenAI results also add one useful nuance. The bounded benefit is not equally strong for every family in ev-

Cell	Entropy	Concession Arc	Parse Success	Outcomes
ledger off + native off	0.000	0.000	1.000	15 authority_decision
ledger on + native off	0.680	0.867	0.667	5 compromise, 10 authority_decision
ledger off + native on	0.000	0.000	1.000	15 authority_decision
ledger on + native on	0.871	0.000	0.000	15 authority_decision

Table 1: Experiment 1 GPT-5.2 mechanism check from the crossed ledger/native ablation.

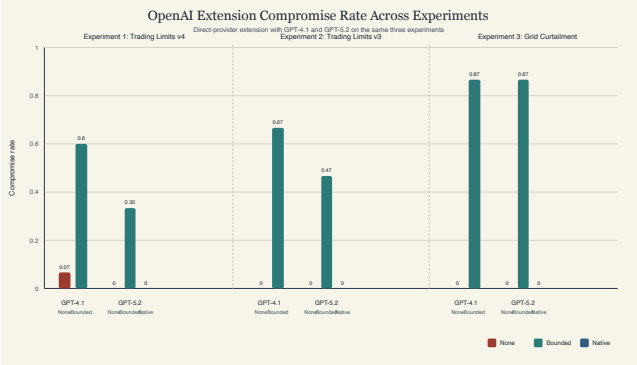

Figure 7: OpenAI extension compromise-rate summary across the three experiments.

ery environment. Relative to Gemini and DeepSeek, GPT-5.2 with bounded reflection is weaker in the first two experiments and cleaner in the grid-curtailment transfer. Concretely, compromise outcomes rise from 5 of 15 in Experiment 1 to 7 of 15 in Experiment 2 and then to 13 of 15 in Experiment 3. We do not have a definitive mechanistic account for that recovery pattern, but one plausible explanation is that the grid-curtailment case gives the bounded ledger a clearer mapping from private state to public action. Compared with the trading-limits environments, the grid case has more concrete operational triggers, more legible institutional roles, and fewer semantically overlapping issue bundles, so a constrained private note may be enough to stabilize compromise-oriented behavior without requiring the model to reason over as much coalition ambiguity. We therefore treat this as a boundary-condition result rather than as a solved mechanism: bounded reflection appears broadly helpful, but its payoff can still depend on how clearly the environment exposes role mandates, fallback pressure, and negotiable issue structure.

Because these runs were added after the core matrix was frozen, we treat them as a provider-breadth extension rather than as the primary identification layer. The detailed extension summaries appear in the statistical appendix.

Mechanism and Interpretation

The data support a mechanism we can describe cautiously as game-solving pressure. Native reasoning appears to push agents toward coherent strategic resolution of the situation rather than toward plausible boundedly rational enactment. In Experiment 1, this pressure shows up as:

- concentration on authority-driven endings
- suppressed negotiated compromise
- frequent exhaustion of the available interaction horizon
- elevated protocol noise relative to bounded reflection

A short qualitative contrast makes the mismatch concrete. In Experiment 1 under no reflection, Gemini repeatedly emits counter-actions across the full horizon and falls into the authority fallback. Under bounded reflection, the same family shifts into a concession cascade by the middle turns and reaches compromise. In Experiment 3 under DeepSeek native reasoning, the trace contains both countering and support moves and a high concession-arc rate, yet still ends in authority decision in every run. That is the central empirical distinction in this paper: local variation can survive while terminal simulation fidelity still collapses.

Bounded reflection behaves differently. It gives agents enough private state to track concessions, open issues, and opponent assessments without pushing them into unrestricted strategic optimization. In effect, bounded reflection appears closer to a cognitively limited private notebook than to an internal theorem prover.

That distinction matters for simulation design. If the target agents are not perfectly optimizing actors, then the simulation mechanism should not silently replace them with optimizer-like proxies. Native reasoning can do exactly that.

We do not claim direct access to provider internals, and we do not attempt a mechanistic claim about hidden chain-of-thought itself. The inference here is behavioral: under the conditions studied, native reasoning changes the observable distribution of trajectories in ways that reduce simulation fidelity.

Experiment 3 makes that point especially clear. There, DeepSeek native reasoning preserves high entropy and frequent concession arcs while still producing authority decision in every run. The OpenAI extension then broadens the same lesson across provider families: GPT-5.2 with native reasoning remains fully rigid across all three experiments, and GPT-4.1 with no reflection remains near-rigid unless the bounded scaffold is added. A model can look strategically strong or dynamically busy and still fail to sample compromise-oriented outcomes.

Implications

The results imply that model qualification for simulation should differ from model qualification for decision support.

For decision support, one may reasonably want:

- stronger strategic coherence
- better optimization
- tighter reasoning consistency

For behavioral simulation, one may instead want:

- higher trajectory diversity
- concession dynamics consistent with bounded-rational negotiation
- bounded-rational inconsistency
- lower protocol fragility

These are not the same objective. A model that is excellent for advising a policymaker or negotiator may be inappropriate for simulating the negotiators themselves.

The practical lesson is straightforward: simulation pipelines should qualify models as samplers, not assume that stronger reasoning automatically improves fidelity.

Related Work

This paper sits at the intersection of four adjacent literatures.

First, classic simulation and agent-based-modeling work treats validation as a core scientific obligation rather than as a post hoc add-on (Bonabeau 2002; Epstein and Axtell 1996; Collins, Koehler, and Lynch 2024). Recent reviews of LLM-based social simulation sharpen that point further: validation is now the central bottleneck for generative social simulation, and apparent behavioral realism can be misleading when the measurement target is underspecified (Larooij and Törnberg 2026; Wallach et al. 2025; Zhou et al. 2024; Hullman et al. 2026). Our paper builds directly on that validation framing by proposing trajectory-level diagnostics for sampler failure.

Second, work on LLM-based synthetic societies and simulated humans argues that language models can generate plausible interactive social behavior at scale (Park et al. 2023; Aher, Arriaga, and Kalai 2023; Horton 2023). Our paper does not reject that direction. It identifies a specific failure mode within it: stronger reasoning can make agents less faithful as samplers of boundedly rational behavior even when the resulting trajectories appear coherent.

Third, work on strategic games, bargaining, and negotiation with LLMs shows both promise and instability in interactive strategic settings (Zheng, Zhou, and Wang 2025; Wang et al. 2026; Fan et al. 2024; Akata et al. 2025). Negotiation-specific studies likewise show that LLM agents can deliberate, haggle, and coordinate without necessarily producing stable or human-like interaction patterns (Abdelnabi et al. 2024; Chatterjee, Miller, and Parepally 2024; Liu, Gu, and Song 2026; Chen et al. 2023; Lorè and Heydari 2023). Some of that work is explicitly interventionist: structured bargaining protocols, agent engineering, and negotiation-specific prompting are used to elicit concessions or improve deal completion. Our results complement that literature by shifting the target of evaluation. We ask not whether a scaffold or prompt helps an agent close deals, but whether it preserves the distributional properties required for behavioral simulation.

Fourth, the paper's bounded-rationality and negotiation framing is grounded in a longer behavioral tradition. Bounded-rationality theory emphasizes satisficing, procedural limits, and context-sensitive choice rather than full-game optimization (Simon 1955; Kahneman 2003; Rubinstein 1998). Negotiation theory and lab bargaining research add a more specific vocabulary around concession behavior, issue structure, and compromise formation (Pruitt 1981; Walton and McKersie 1965; Raiffa 1982). Recent LLM evaluation work points in a complementary direction: SPABCBench shows that stronger or newer models do not reliably improve end-user behavioral fidelity and that bounded-rational prompting can sometimes help (Li et al. 2026), CRAFT shows that stronger reasoning does not reliably improve multi-agent coordination under partial information (Nath, VanderHoeven, and Krishnaswamy 2026), and persona/payoff studies show that representational choices can override payoff-sensitive strategic behavior (Manoranjan and Gaikwad 2026). Our findings suggest that these adjacent observations can be understood through the same lens: in bounded-rational settings, more reasoning can mean less realism.

Broader Impact and Ethics

The main practical risk raised by this paper is misuse of LLM simulations in high-stakes settings. If policymakers, firms, or researchers treat stronger reasoning models as automatically better social simulators, they may over-trust clean but behaviorally distorted outputs. The result can be false confidence in narrow equilibrium-like trajectories, underestimation of compromise paths, or systematic removal of bounded-rational frictions that matter in real institutions. Our recommendation is therefore conservative: use simulation outputs as exploratory evidence, report model and prompt conditions explicitly, and treat sampler validation as a prerequisite for downstream policy or organizational use.

Limitations

This paper has important limitations.

First, the empirical base is still narrower than a full benchmark suite. The paper now contains three scenario-based multi-agent experiments, including a genuine domain transfer into emergency grid curtailment. That is a meaningful base, but a stronger archival version would still extend the analysis to additional domains and non-negotiation settings.

Second, provider-native reasoning is only partially observable. We measure its consequences at the behavioral and operational levels, not its internal mechanism.

Relatedly, `native reasoning` is provider-specific rather than a single standardized treatment. Gemini, DeepSeek, and the OpenAI extension expose different reasoning controls and runtime behavior, so the paper's causal claim is behavioral and comparative, not architectural: under the provider-native reasoning settings available in these families, the resulting trajectory and outcome distributions differ from the bounded condition.

Third, protocol reliability is itself part of the result but also complicates interpretation. A model that produces more provider or format errors is a weaker practical simulator, but some of the native/noise effect may reflect provider-side tooling constraints in addition to behavioral mismatch.

Fourth, this paper does not claim that reasoning is universally harmful. In tasks where the objective is optimization, planning, or advice, native reasoning may be preferable. Our

claim is only about simulation fidelity in bounded-rational multi-agent settings under the conditions studied here.

Fifth, the current external-validity story is strongest for institutional negotiation environments that share the structured multi-actor grammar used here. Experiment 3 demonstrates a genuine domain transfer, and the OpenAI extension broadens provider coverage, but future work should still test whether the same mismatch appears as strongly in qualitatively different multi-agent task families.

Sixth, the instrumentation concern is mitigated but not removed entirely. We report provider-error turns, format-error turns, and an error-excluded robustness reanalysis precisely because parser and transport failures can otherwise masquerade as behavioral evidence. At the same time, the clean-run pattern remains strong: after excluding every run with any provider-error or format-error turn, retained native runs across the two primary families and three experiments still end in `authority_decision` in 36 of 36 cases. This makes it unlikely that the native collapse is reducible to parser failure alone, but it does not eliminate all provider-side confounding.

Seventh, human-grounded calibration would be required for claims of external behavioral realism or downstream policy prediction. That is outside the scope of this paper. The present contribution is a comparative methodological result about sampler distortion within a fixed simulation framework.

Eighth, each cell contains 15 runs, which is enough to expose stable directional patterns but still leaves a nontrivial multiple-comparison burden given the number of endpoints, families, conditions, and environments reported. We therefore treat the strongest evidence as the consistency of the directional pattern across experiments and endpoints, with permutation tests and effect sizes serving as supporting inferential summaries rather than as the sole basis of the claim.

Ninth, the bounded ledger is intentionally compact rather than normatively neutral. Its fields were chosen to preserve boundedly rational negotiation state, and that means they may also foreground compromise-relevant variables such as concessions, opponent flexibility, and open issues. The current ledger-sensitivity ablation shows that the result is not reducible to one exact prompt wording, but a broader neutral-scaffold comparison remains future work.

Conclusion

The central result of this paper is simple: more reasoning does not necessarily mean more simulation fidelity. Across three scenario-based multi-agent negotiation experiments, bounded reflection consistently produces more diverse, concession-rich, and compromise-oriented trajectories than either no reflection or native reasoning. Native reasoning does not recover bounded-sampler quality and is noisier operationally. The sharpest version of that claim appears in the grid-curtailment transfer, where DeepSeek native shows high within-trajectory diversity and concession behavior yet still fails to produce a single compromise outcome.

The broader implication is methodological. Behavioral simulation should not choose models solely by benchmark strength or reasoning prestige. It should qualify them for the actual job of simulation. For bounded-rational agent modeling, the right question is not "which model is smartest?" but "which model is the best sampler for the target population?"

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games. In *The Twelfth International Conference on Learning Representations*.

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 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*, 337–371.

Akata, E.; Coda-Forno, J.; Oh, S. J.; Bethge, M.; and Schulz, E. 2025. Playing Repeated Games with Large Language Models. *Nature Human Behaviour*, 9: 1380–1390.

Bonabeau, E. 2002. Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. *Proceedings of the National Academy of Sciences*, 99(suppl_3): 7280–7287.

Chatterjee, A.; Miller, S.; and Parepally, N. 2024. AgreeMate: Teaching LLMs to Haggle. *arXiv preprint arXiv:2412.18690*.

Chen, H.; Ji, W.; Xu, L.; and Zhao, S. 2023. Multi-Agent Consensus Seeking via Large Language Models. *arXiv preprint arXiv:2310.20151*.

Collins, A. J.; Koehler, M.; and Lynch, C. J. 2024. Methods That Support the Validation of Agent-Based Models: An Overview and Discussion. *Journal of Artificial Societies and Social Simulation*, 27(1): 11.

Epstein, J. M.; and Axtell, R. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press and MIT Press.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17960–17967.

Horton, J. J. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Technical Report 31122, National Bureau of Economic Research.

Hullman, J.; Broska, D.; Sun, H.; and Shaw, A. 2026. This Human Study Did Not Involve Human Subjects: Validating LLM Simulations as Behavioral Evidence. *arXiv preprint arXiv:2602.15785*.

Kahneman, D. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, 93(5): 1449–1475.

Larooij, M.; and Törnberg, P. 2026. Validation Is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling. *Artificial Intelligence Review*, 59: 15.

Li, Y.; Li, L.; Lee, H.-P.; and Das, S. 2026. How Well Can LLM Agents Simulate End-User Security and Privacy Attitudes and Behaviors? *arXiv preprint arXiv:2602.18464*.

Liu, X.; Gu, S.; and Song, D. 2026. AgenticPay: A Multi-Agent LLM Negotiation System for Buyer-Seller Transactions. *arXiv preprint arXiv:2602.06008*.

Lorè, N.; and Heydari, B. 2023. Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. *arXiv preprint arXiv:2309.05898*.

Manoranjan, V.; and Gaikwad, S. N. 2026. When Identity Overrides Incentives: Representational Choices as Governance Decisions in Multi-Agent LLM Systems. *arXiv preprint arXiv:2601.10102*.

Nath, A.; VanderHoeven, H.; and Krishnaswamy, N. 2026. CRAFT: Grounded Multi-Agent Coordination Under Partial Information. *arXiv preprint arXiv:2603.25268*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Ringel Morris, M.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2:1–2:22.

Pruitt, D. G. 1981. *Negotiation Behavior*. Academic Press.

Raiffa, H. 1982. *The Art and Science of Negotiation*. Harvard University Press.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. MIT Press.

Simon, H. A. 1955. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1): 99–118.

Wallach, H.; Desai, M.; Cooper, A. A.; Garcia-Gathright, J.; Olteanu, A.; Pangakis, N. J.; Reed, S.; Sheng, E.; Vann, D.; Vaughan, J. W.; Vogel, M.; Washington, H.; and Jacobs, A. Z. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 82232–82251.

Walton, R. E.; and McKersie, R. B. 1965. *A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System*. McGraw-Hill.

Wang, C.; Kasenberg, D.; Stachenfeld, K.; and Castro, P. S. 2026. Discovering Differences in Strategic Behavior Between Humans and LLMs. *arXiv preprint arXiv:2602.10324*.

Zheng, K.; Zhou, J.; and Wang, H. 2025. Beyond Nash Equilibrium: Bounded Rationality of LLMs and Humans in Strategic Decision-making. *arXiv preprint arXiv:2506.09390*.

Zhou, X.; Su, Z.; Eisape, T.; Kim, H.; and Sap, M. 2024. Is This the Real Life? Is This Just Fantasy? The Misleading Success of Simulating Social Interactions With LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21692–21714. Association for Computational Linguistics.