# FROM SIMULATION TO EVIDENCE-GROUNDED DECISION ANALYSIS: A BENCHMARK AND DECISION-SUPPORT FRAMEWORK FOR MULTI-AGENT SOCIAL SYSTEMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Most LLM-based social-simulation studies evaluate realism, social intelligence, or task success within a simulated world. We ask two related questions: **whether agents exhibit adaptive social strategies beyond role-consistent responses**, and **how execution traces and simulation results can support disciplined decision analysis**. We present SOCIALSIM-BENCH, an executable benchmark with social-media and commercial-negotiation environments such as role profiles, private goals, visibility constraints, and deterministic state transitions. We evaluate agents with rule-based outcomes and five LLM-as-a-Judge dimensions, while also analyzing the strategies they employ, finding that even the excellent LLMs only perform simple strategies without applying complex methods. We further study a simulation-to-decision workflow that transforms traces into conditional reports with triggers, alternatives, trade-offs, and exit conditions. Our report ablations show that trajectory summaries are more useful than equally budgeted raw traces. Matched simulation evidence can provide additional decision-relevant content, but current results do not establish that simulation-grounded reports consistently outperform carefully structured scenario-only reports. Accordingly, SOCIALSIM-BENCH evaluates trace-grounded report properties rather than real-world decision outcomes.

## 1 INTRODUCTION

Large language models increasingly enable role enactment and social interaction through advances in instruction following, long-context reasoning, and tool use Li et al. (2026a); Zhao et al. (2026); Ouyang et al. (2022); Yao et al. (2022); Schick et al. (2023). Social simulation has evolved along two directions: large-scale agent populations for studying collective dynamics and emergent phenomena, and small-scale LLM agents with explicit roles and goals for modeling situated interaction, negotiation, and social reasoning Mou et al. (2026); Centola & Macy (2007); Epstein & Axtell (1996); Ashkinaze et al. (2025); Piao et al. (2025a); Li et al. (2023); Zhou et al. (2024); Chen et al. (2024; 2025). However, existing evaluations mainly focus on in-simulation behavior, such as consistency, interaction quality, or task success. This leaves two questions open: *can LLM agents reason about evolving social dynamics and develop adaptive strategies beyond role-playing?* And *how can plausible simulation trajectories be transformed into evidence for decision exploration?* Plausible simulation behavior alone therefore neither establishes substantive strategic reasoning nor provides actionable decision support.

For the first question, We present SOCIALSIM-BENCH, an executable benchmark spanning **social-media**, which models posts, comments, reactions, follows, and private messages around specific news events, and **commercial negotiation** environments, which models offers, sessions, contracts, resources, and external shocks in commercial negotiations. Its scenarios specify roles, private goals, information visibility, constraints, evolving events, and termination conditions. We assess trajectories with environment-specific rule metrics to measure the executable actions and environment-specific results, and a five-dimensional LLM-as-a-Judge evaluation of goal alignment, social rea-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

soning, persona consistency, communication, and risk-and-constraint compliance Li et al. (2026b). Together, these measures distinguish task outcomes from plausible or role-consistent behavior.

To study the second question, we also examine whether simulation traces can support structured decision exploration. Because observed choices in real world are not reliable labels of optimal decisions, we use traces to identify possible consequences, risks, alternatives, trade-offs, and conditions for changing course. This follows robust decision making and exploratory modeling, which reason across plausible futures rather than predict a single outcome Lempert (2002); Bankes (1993); Bryant & Lempert (2010); DeAngelis & Diaz (2019). We compare direct LLM reports with harness-based generation using Codex and Claude Code, and introduce a professional harness that summarizes trajectories into traceable, conditional recommendations. Report evaluation covers grounding, conditional decision value, uncertainty calibration, and related properties. Thus, SOCIALSIM-BENCH extends social simulation from generating plausible interactions to producing traceable evidence for decision analysis.

Our experiments compare various models using deterministic rule scores and five LLM-as-a-judge evaluation dimensions Li et al. (2026b); Gu et al. (2026). We find that model performance is highly dependent on the simulation domain, and models with strong coding capabilities do not necessarily perform well on these social interaction tasks. The human annotations prove the effectiveness of LLM evaluation. We further examine the interaction trajectories and find that even strong LLMs rarely demonstrate sophisticated social reasoning. Although they can exhibit recognizable strategies, their behavior is often limited to direct and locally responsive actions rather than complex, adaptive strategic reasoning. We further examine whether skills Yang et al. (2026) and refinement mechanism Shinn et al. (2023) can broaden or redirect the behaviors explored. For the downstream reporting task, we compare direct generation with simulation-grounded generation across multiple report settings, and include Codex and Claude Code as strong general-purpose harness baselines. We evaluate these reports using human and LLM evaluation. These results support the value of an explicit trace summarization and evidence-merging.

Our contributions are: (1) **an executable social simulation benchmark and dataset** for evaluating strategic behavior across negotiation and social-media scenarios; (2) **a controlled cross-domain analysis** of models, behavioral metrics, and skill and refinement interventions; and (3) **a traceable simulation-to-decision harness and evaluation standard** that separates evidence grounded in traces from simulation hypotheses and unvalidated operational safeguards.

## 2 RELATED WORK

### 2.1 SOCIAL SIMULATION AND SOCIAL-AGENT EVALUATION.

Social-simulation systems study decisions, interactions, and collective behavior in virtual environments Hu et al. (2026); Li & Tao (2026); Wan et al. (2026); Wang et al. (2025). Prior work spans believable agents, social goals, role-playing, large-scale societies, and social-media dynamics Park et al. (2023); Ma et al. (2026); Xue et al. (2026); Zhang et al. (2026); Wang et al. (2026); Zhou et al. (2024); Chen et al. (2024); Piao et al. (2025b); Ashkinaze et al. (2025); Touzel et al. (2025); Chen et al. (2025); Piao et al. (2025a); Wu et al. (2024); Qiao et al. (2025). SOCIALSIM-BENCH instead evaluates valid state transitions, adaptive strategies, and traceable evidence for downstream decision analysis.

### 2.2 SIMULATION-BASED DECISION SUPPORT.

Agent-based models have long represented heterogeneous decisions and their consequences DeAngelis & Diaz (2019). Instead of predicting a single fixed future, robust decision-making and exploratory modeling explore many possible future scenarios Lempert (2003). Recent LLM-social-simulation surveys organize this literature from individual and scenario simulation to society-scale simulation Mou et al. (2026). SOCIALSIM-BENCH adopts this possibility-oriented view for language-mediated interaction, where evidence is distributed across messages, action validity, and changing social state. It converts multiple trajectories into mechanisms, triggers, trade-offs, and exit conditions without treating any trajectory as a real-world forecast.

# 3 METHODOLOGY

This section introduces the **Data Generation**, **Evaluation Framework**, and **Decision Report Generation** components of SOCIALSIM-BENCH. The framework is in Figure 1.

## 3.1 SCENARIO AND ENVIRONMENT

We construct two evaluation environments: social-media interaction and commercial negotiation. Both environments specify role profiles, private goals, background story, information visibility, language style, and news.

**Social-media** episodes unfold like a shared online community. Agents take turns responding to what they see, choosing whether to post, comment, like, follow others, share content, or send private messages. Conversations can stay focused on the original topic or drift naturally into unrelated interests and disagreements. Over time, these interactions reveal how ideas spread, how discussions develop, who disengages, and how people respond to one another, while each agent retains its own goals, such as getting more likes.

**Commercial negotiation** episodes begin when participants decide whom to approach and which conversations to join. Once a discussion begins, they can negotiate openly or privately, revise offers, weigh competing terms, and decide whether to continue, walk away, or commit to an agreement. New information may arrive during the process and reshape what each party wants or is willing to accept. The full trajectory captures the evolving conversation, changing relationships, key decisions, and final outcome, whether that is a signed agreement, a stalled negotiation, or an early departure.

We represent each task with the tuple $s = (x, P, G, R, H, M, V, E)$, where $x$ denotes the public background, $P$ the participant profiles, $G$ their private goals, $R$ their relationships and resources, $H$ the interaction history, $M$ the action quotas, $V$ the visibility rules, and $E$ the schedule of external events. At turn $t$, the active agent $i$ selects a legal action based on the current state, its observable information, and its remaining quotas: $a_{i,t} \in \mathcal{A}(z_t, V_{i_t}, M_{i_t})$. For sequential interaction, we write $a_t^i = a_{i,t}$, and $a_t$ denotes the corresponding joint action. The environment then applies its deterministic transition rule: $z_{t+1} = F(z_t, a_t, E_t)$, and records the resulting observable event. An episode terminates when a success condition, departure decision, inactivity rule, or interaction horizon is reached.

## 3.2 DATA GENERATION

Specifically, large language models are used to generate diverse scenarios, participant backgrounds, roles, and objectives, while rule-based modules convert these descriptions into executable environments with explicit states, constraints, and outcome mechanisms. During evaluation, agents interact autonomously within identical scenarios, producing complete interaction trajectories for analysis.

**Social-media Scenario.** This scenario simulates open-ended multi-user public interactions on online platforms. It synthesizes diverse public events including viral news, influencer disputes, community discussions, and public opinion crises, and further generates heterogeneous user personas with distinct communication styles, interests, social influence, relational ties, and behavioral goals such as engagement growth, influence expansion, and reputation maintenance. During simulation, agents can perform platform-specific operations including posting, commenting, sharing, following, and private messaging. The environment deterministically updates content exposure, interaction networks, trending topics, and user influence throughout the interaction process.

**Commercial Negotiation Scenario.** This scenario models time-constrained multi-party commercial negotiation with competing objectives. It first instantiates the negotiation structure, timeline, and theme, then generates business contexts, participant goals, social relationships, and external events, covering typical negotiation patterns such as buyer–seller bargaining, market rivalry, and interest–relationship conflicts. Each scenario is standardized into a structured environment with fixed participant identities, initial resources, market conditions, negotiation phases, and contract rules. While language models assist contextual generation, key metrics including payoffs, price constraints, and state transitions follow deterministic rules, ensuring fair and consistent agent evaluation.

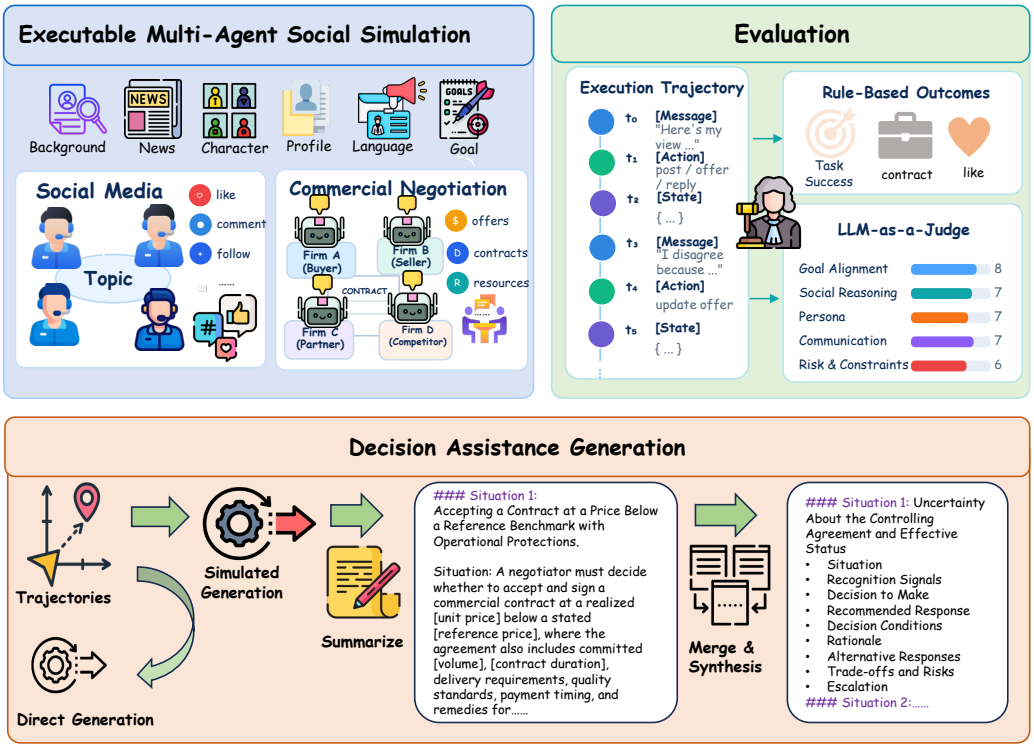



Figure 1: The overview of the simulation framework.

For both environments, the generation pipeline samples scenario configurations, compares new instances with existing data to reduce redundancy, and validates consistency among backgrounds, roles, goals, external events, and completion conditions. Validated scenarios are stored together with environment configurations, initial states, interaction constraints, available actions, and metadata to support reproducible evaluation. Each episode designates one role as the focal agent, while other participants are instantiated with fixed baseline models, enabling controlled comparison among different focal-agent under identical conditions. All generated scenarios are manually verified for quality.

We construct 96 commercial negotiation and 36 social-media scenarios. Beyond the released benchmark, the synthesis framework supports scalable scenario generation, enabling future extensions and large-scale data construction.

### 3.3 Evaluation

We evaluate each episode using two complementary metrics. A deterministic rule-based score quantifies objectively verifiable task outcomes, while an LLM-as-a-Judge assesses the quality of the focal agent's interaction process. This separation distinguishes between *what* an agent achieves and *how* it acts to reach its objectives.

**Rule-based Evaluation.** For social-media episodes, the deterministic score is computed solely from goals whose metrics are supported by platform states and whose progress values are observable. Let $\mathcal{S}$ denote this set. Unsupported or unobservable goals are excluded from calculation rather than treated as failures. For each supported goal $k$, the observed progress is mapped to a normalized score $q_k \in [0, 1]$. For higher-is-better metrics,

$$q_k = \begin{cases} 1, & t_k \le 0 \text{ and } x_k \ge t_k, \\ 0, & t_k \le 0 \text{ and } x_k < t_k, \\ \min\left(1, \dfrac{x_k}{t_k}\right), & t_k > 0, \end{cases} \tag{1}$$

where $x_k$ and $t_k$ correspond to the observed and target values, respectively. For lower-is-better metrics,

$$q_k = \begin{cases} 1, & x_k \le t_k, \\ 0, & x_k > t_k \text{ and } t_k \le 0, \\ \min\left(1, \dfrac{t_k}{x_k}\right), & x_k > t_k > 0. \end{cases} \tag{2}$$

The social-media rule score is defined as a priority-weighted average:

$$S_{\mathrm{SM}} = \frac{\sum_{k\in\mathcal{S}} w_k q_k}{\sum_{k\in\mathcal{S}} w_k}. \tag{3}$$

The social-media rule score is derived from observable platform metrics. Higher values correspond to better outcomes in engagement, content diffusion, and audience growth, whereas lower values indicate superior performance for violation-related metrics. We additionally report goal coverage and per-goal diagnostic statistics.

For buy–sell negotiation episodes, each contract is evaluated using the scenario-specific reference price $p_{\mathrm{ref}}$, cost price $c$, agreed unit price $p$, agreed transaction volume $v$, and number of resolved negotiation sessions $n_{\mathrm{sess}}$. A contract takes effect only when its status factor $\phi$ exceeds the threshold $\tau_{\mathrm{eff}}$, denoted as $I_{\mathrm{eff}} = \mathbb{I}[\phi \ge \tau_{\mathrm{eff}}]$. Buyer savings and seller earnings are calculated deterministically as $B_{\mathrm{total}} = v\max(0, p_{\mathrm{ref}} - p)I_{\mathrm{eff}}$ and $E_{\mathrm{total}} = v\max(0, p - c)I_{\mathrm{eff}}$, respectively.

The rule score integrates buyer savings, seller sustainability, and negotiation activity:

$$\begin{gathered} f_{\mathrm{save}} = \mathrm{clip}\left(\frac{\max(0, p_{\mathrm{ref}} - p)}{p_{\mathrm{ref}}} I_{\mathrm{eff}}, 0, 1\right), \quad f_{\mathrm{sustain}} = \mathbf{1}_{c>0}\,\mathrm{clip}\left(\frac{\max(0, (p-c)/c)}{\tau_{\mathrm{margin}}}, 0, 1\right), \\ S_{\mathrm{contract}} = (\lambda_{\mathrm{save}} f_{\mathrm{save}} + \lambda_{\mathrm{seller}} f_{\mathrm{sustain}} + \lambda_{\mathrm{activity}} f_{\mathrm{activity}}) I_{\mathrm{eff}}. \end{gathered} \tag{4}$$

where $\lambda_{\mathrm{save}} + \lambda_{\mathrm{seller}} + \lambda_{\mathrm{activity}} = 1$; non-positive realized prices receive zero scores.

**LLM-based Evaluation** . The LLM-as-a-Judge outputs averaged scores across five dimensions, each ranging from 0 to 10. **Goal alignment** quantifies progress toward the agent's stated goals while adhering to task constraints. **Social reasoning** measures whether the agent correctly interprets counterparties, contextual information, incentives, and interaction signals, and adapts its behavior accordingly. **Persona consistency** evaluates whether the agent's language, emotion, stance, and actions remain consistent with its assigned persona. **Communication quality** assesses the clarity, relevance, tone, responsiveness, and effectiveness of the agent's utterances. **Risk and constraint compliance** checks whether the agent avoids unsubstantiated commitments, unnecessary escalation, policy violations, and breaches of explicit scenario constraints.

Together, deterministic and LLM-based evaluation provide complementary validity evidence. The former grounds outcome assessment in reproducible environment states, while the latter captures interactive qualities that cannot be reduced to final-state statistics. Execution records preserve applicable rule parameters, contract information, state snapshots, and settlement events, rendering rule-based results auditable and reproducible.

## 4 Simulation-to-Decision Pipeline

We assess the contribution of simulation evidence across four report-generation conditions. **LLM direct** generates a report from the scenario description alone. **LLM simulation-grounded** additionally receives evidence organized by the benchmark pipeline, including trajectory summaries and recurring mechanisms identified across runs. **Harness direct** uses a general-purpose harness with only the scenario description. **Harness simulation-grounded** allows the harness to inspect simulation artifacts before writing its report. We evaluate both Codex and Claude Code in the two harness conditions, yielding six configurations in total.

We propose the professional harness. The simulation-grounded converts interaction trajectories into structured evidence. It first summarizes each trajectory, retaining key actions, goals, interactions, and outcomes. It then combines summaries across trajectories, grouping recurring patterns while preserving consequential exceptions and conflicting observations. Finally, it produces a report that

Table 1: Benchmark results in the social-media and commercial negotiation environments. Rule scores are multiplied by 10 for display; LLM Mean is the average of five LLM-as-a-Judge dimensions. Bold denotes the highest score in each domain and column.

| Domain | Model | Rule-based | LLM-as-a-Judge | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Rule | Goal | Reason | Persona | Comm. | Risk | Mean |
| **Social media** | GPT-5.6-Luna | 0.95 | 6.08 | 6.58 | 6.11 | 6.56 | 7.31 | 6.53 |
| | GPT-5.6-Terra | 1.03 | 6.14 | 6.78 | 6.39 | 6.92 | 7.56 | 6.76 |
| | Claude-Sonnet-4.6 | 1.09 | 6.47 | 6.72 | 6.42 | 6.86 | 7.28 | 6.75 |
| | Gemini-3.1-Flash-Lite | 1.02 | 5.19 | 5.64 | 5.75 | 5.53 | 6.14 | 5.65 |
| | GLM-5.2 | 1.69 | 6.19 | 6.69 | 6.53 | 6.61 | 6.69 | 6.54 |
| | DeepSeek-V4-Pro | **2.18** | **6.75** | **7.06** | **6.78** | **6.94** | **7.64** | **7.03** |
| **Negotiation** | GPT-5.6-Luna | 2.17 | 6.70 | 6.46 | 7.94 | 6.78 | 6.85 | 6.95 |
| | GPT-5.6-Terra | 2.35 | 6.94 | 6.60 | 7.92 | 6.94 | **6.98** | 7.08 |
| | Claude-Sonnet-4.6 | 2.41 | 6.89 | 6.99 | 8.22 | 7.23 | 6.67 | 7.20 |
| | Gemini-3.1-Flash-Lite | **2.48** | 5.78 | 5.88 | 7.46 | 5.85 | 5.73 | 6.14 |
| | GLM-5.2 | 2.40 | **7.33** | **7.10** | **8.58** | **7.41** | 6.93 | **7.47** |
| | DeepSeek-V4-Pro | 2.34 | 6.74 | 6.32 | 7.44 | 6.67 | 6.65 | 6.76 |

links plausible situations to conditional recommendations and their supporting evidence. This organization reduces the burden of interpreting long interaction histories while retaining the basis for each recommendation.

# 5 Results

## 5.1 Experimental Settings

We ask whether rule-based and LLM-based evaluations agree, whether skills and refinement change focal-agent behavior, and whether simulation evidence improves decision reports. We test GPT-5.6-Luna, GLM-5.2, Claude-Sonnet-4.6, Gemini-3.1-Flash-Lite, GPT-5.6-Terra, and DeepSeek-V4-Pro as focal agents, with GPT-5.6-Luna controlling all other agents. Agent actions and refinement use temperature $0.3$; GPT-5.6-Terra judges both domains at temperature $0.0$.

## 5.2 Main Results

Table 1 shows that model rankings vary across environments and metrics. In social media, DeepSeek-V4-Pro achieves the strongest overall performance, leading both the rule score and LLM-as-a-Judge mean, as well as all five individual judge dimensions. In **commercial negotiation**, Gemini-3.1-Flash-Lite obtains the highest rule score, whereas GLM-5.2 achieves the highest judged mean. These differences indicate that task outcomes and behavioral quality capture distinct aspects of agent performance.

Several conclusions follow. First, **agent competence is multidimensional**: success on one metric does not imply superiority across others, as task achievement, communication, persona consistency, and risk compliance measure different capabilities. Second, **rule-based and LLM-based evaluations are complementary**: rule metrics capture observable outcomes and state transitions, while judge metrics assess behavioral qualities beyond final states. Finally, **model evaluation should be task- and metric-aware rather than based on a universal ranking**. Models strong on general agent benchmarks may not transfer directly to social simulation, where adaptive interaction and strategic consistency are essential. These findings motivate multi-dimensional evaluation for agentic systems. The validity of LLM evaluation and multi-run stability analysis are discussed in Appendix E and Appendix E.1, respectively.

## 5.3 Behavior Analysis

We classify focal-agent strategies based on direct evidence from actions and interactions, allowing multiple strategies within a trajectory. Results reveal a substantial gap between frequent local strate-

Table 2: Benchmark results for social-media and commercial negotiation environments. LLM Mean is the mean of the five LLM-as-a-Judge dimensions. $\Delta$ Rule and $\Delta$ LLM denote percentage changes in Rule and LLM Mean relative to the Basic setting for the same model and domain.

| Model | Treatment | Rule | LLM-as-a-Judge | | | | | LLM Mean | $\Delta$ Rule | $\Delta$ LLM |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Goal | Reason | Persona | Comm. | Risk | | | |
| **Social media** | | | | | | | | | | |
| GPT-5.6-Luna | Basic | 0.947 | 6.08 | 6.58 | 6.11 | 6.56 | 7.31 | 6.53 | – | – |
| | Skills | 0.842 | 6.00 | 6.58 | 6.50 | 6.81 | 7.28 | 6.63 | -11.1% | +1.5% |
| | Refine | 1.108 | 6.11 | 6.58 | 6.31 | 6.81 | 7.25 | 6.61 | +17.0% | +1.2% |
| Claude-Sonnet-4.6 | Basic | 1.092 | 6.47 | 6.72 | 6.42 | 6.86 | 7.28 | 6.75 | – | – |
| | Skills | 0.821 | 6.14 | 6.67 | 6.28 | 6.69 | 6.86 | 6.53 | -24.8% | -3.3% |
| | Refine | 0.825 | 6.56 | 6.86 | 7.06 | 7.14 | 7.42 | 7.01 | -24.5% | +3.9% |
| **Negotiation** | | | | | | | | | | |
| GPT-5.6-Luna | Basic | 2.24 | 6.78 | 6.42 | 7.91 | 6.78 | 6.90 | 6.96 | – | – |
| | Skills | 2.27 | 6.51 | 6.07 | 7.48 | 6.44 | 6.58 | 6.62 | +1.4% | -4.9% |
| | Refine | 1.91 | 5.25 | 4.39 | 4.58 | 3.30 | 3.24 | 4.15 | -14.7% | -40.5% |
| Claude-Sonnet-4.6 | Basic | 2.445 | 6.92 | 7.02 | 8.35 | 7.35 | 6.74 | 7.28 | – | – |
| | Skills | 2.445 | 7.01 | 6.96 | 8.30 | 7.29 | 7.01 | 7.31 | +0.0% | +0.5% |
| | Refine | 2.469 | 4.69 | 4.89 | 4.83 | 5.65 | 4.71 | 4.95 | +1.0% | -31.9% |

gies and rare adaptive behaviors. In negotiation, agents commonly rely on information gathering and competitor exploitation, while coalition-based strategies, information control, and negotiation-sequence control remain rare. Similarly, in social media, agents frequently perform topic selection and content positioning, but seldom exhibit feedback-driven adjustment, reputation–exposure trade-offs, or network-based strategies.

These findings suggest that agents primarily rely on immediate, locally responsive behaviors rather than sustained, multi-step, or structure-aware adaptation. Consequently, role consistency and task success may overestimate strategic social reasoning: agents can express strategic knowledge but often fail to retrieve and apply it during dynamic interactions. Detailed strategy definitions are provided in Appendix D.

Table 3: Strategy usage across the two environments. Difficulty levels indicate the complexity of strategy execution.

| Domain | Strategy | Difficulty | Episodes | Usage |
|---|---|---|---|---|
| Social media | Topic selection | Low | 375 | 86.8% |
| | Content positioning | Low | 327 | 75.7% |
| | Keyword/tag use | Low | 89 | 20.6% |
| | Feedback-driven adaptation | High | 37 | 8.6% |
| | No catalog strategy | Low | 27 | 6.2% |
| | Other strategy | High | 23 | 5.3% |
| | Sharing strategy | Low | 18 | 4.2% |
| | Controversy trade-off | High | 16 | 3.7% |
| | Reputation–exposure trade-off | High | 16 | 3.7% |
| | Content reposting | High | 14 | 3.2% |
| | Engagement maintenance | Low | 11 | 2.5% |
| | Audience segmentation | High | 7 | 1.6% |
| | Trend following | High | 2 | 0.5% |
| | Traffic-seeking interaction | High | 0 | 0.0% |
| | Network leverage | High | 0 | 0.0% |
| | Cross-user diffusion | High | 0 | 0.0% |
| Negotiation | Information acquisition | Low | 993 | 86.2% |
| | Competitor leverage | High | 913 | 79.3% |
| | Multi-party offer comparison | High | 897 | 77.9% |
| | Anchoring | Low | 817 | 70.9% |
| | Conditional exchange | High | 526 | 45.7% |
| | Dynamic adaptation | High | 440 | 38.2% |
| | Relationship maintenance | Low | 231 | 20.1% |
| | Issue separation | High | 224 | 19.4% |
| | Concession management | High | 176 | 15.3% |
| | Exit threat | High | 172 | 14.9% |
| | Negotiation sequence control | High | 81 | 7.0% |
| | Information control | High | 54 | 4.7% |
| | Coalition | High | 20 | 1.7% |
| | Other strategy | High | 16 | 1.4% |
| | Coalition division | High | 2 | 0.2% |
| | No catalog strategy | Low | 2 | 0.2% |

High difficulty  Low difficulty

## 5.4 Metric Correlation

Figure 2 shows that deterministic task performance and LLM-based behavioral evaluation capture complementary aspects of agent capability. The Rule score exhibits negligible correlations with the five judge dimensions, indicating that successful environment-state transitions do not necessarily reflect stronger reasoning, goal alignment, persona consistency, communication, or risk awareness.

In contrast, the LLM-as-a-Judge dimensions form a coherent behavioral-quality cluster, cap-

Table 4: Report-generation quality. LLM Win is the pairwise win rate assigned by the LLM judge. Human Win is computed from human ranking as the fraction of the five lower-ranked reports defeated by each report. Fact., CF, Alt., Cond., and Uncert. are 0 to 10 quality scores. Scenarios denotes the mean number of scenarios covered. Bold values indicate the best performance.

| Domain | Condition | Report-generation quality | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LLM Win | Human Win | Fact. | CF | Alt. | Cond. | Uncert. | Scenarios |
| **Social media** | LLM direct | 40.0% | 0.0% | 7.75 | 7.75 | 8.30 | 8.05 | 5.55 | 14.0 |
| | Codex direct | 30.0% | 20.0% | 7.65 | 7.85 | 8.20 | 8.20 | 5.75 | **14.1** |
| | Claude Code direct | 40.0% | 40.0% | 7.80 | 7.90 | 8.25 | 8.20 | 6.45 | 12.4 |
| | Codex sim. | 30.0% | 80.0% | 7.65 | 8.10 | 7.80 | 8.25 | 6.85 | 9.1 |
| | Claude Code sim. | 60.0% | 60.0% | 7.50 | 8.40 | **8.55** | 8.65 | 7.30 | 9.6 |
| | LLM sim. | **100.0%** | **100.0%** | **8.45** | **8.40** | 8.50 | **8.85** | **8.90** | **15.0** |
| **Negotiation** | LLM direct | 70.0% | 0.0% | 7.80 | 7.75 | 8.05 | 8.30 | 7.15 | 14.0 |
| | Codex direct | 10.0% | 40.0% | 7.35 | 7.50 | 7.85 | 8.05 | 6.80 | **20.2** |
| | Claude Code direct | 40.0% | 20.0% | 7.65 | 7.80 | 8.15 | 8.15 | 6.90 | 16.3 |
| | Codex sim. | 30.0% | **100.0%** | 7.55 | 8.00 | 8.00 | 8.15 | 7.10 | 7.4 |
| | Claude Code sim. | 50.0% | 80.0% | 7.25 | 8.15 | 8.30 | 8.40 | 7.05 | 10.0 |
| | LLM sim. | **100.0%** | 60.0% | **8.80** | **8.45** | **8.50** | **8.80** | **9.05** | 17.0 |

turing the quality of social interaction and strategic behavior. These findings motivate our multi-perspective evaluation framework: rule-based metrics assess task outcomes and constraint satisfaction, while LLM-based evaluations characterize behavioral quality and adaptation, providing a more comprehensive assessment of LLM social agents.

### 5.5 Skills and Refine

Table 2 shows that neither skills nor refine consistently improves agent performance. Skills augment the focal agent's prompt with state- and role-aware guidance before action generation, whereas refine modifies the generated response after the base agent selects an action.

The effects differ substantially across models and domains. In social media, skills slightly improve GPT-5.6-Luna's judged performance, while refine improves its rule score by 17.0%. For Claude-Sonnet-4.6, skills reduce both rule-based and judged performance; refine improves the judged mean but reduces the rule score. In negotiation, skills slightly improve GPT-5.6-Luna's rule score while reducing its judged mean, whereas refine substantially degrades both measures. Neither intervention materially improves Claude-Sonnet-4.6 in negotiation.

These findings show that effective intervention depends on alignment with the base model and the environment. Additional guidance may conflict with an agent's existing decision process, while post-hoc rewriting may alter action semantics when language is coupled to structured contracts, commitments, and timing.

### 5.6 Report Generation Result

We evaluate decision reports using factual fidelity, counterfactual plausibility, alternative diversity, scenario coverage, conditional decision value, and uncertainty calibration. The first two criteria assess consistency with scenario conditions and agent incentives. Alternative diversity and scenario coverage measure the breadth of strategies and situations represented. Conditional decision value measures whether reports specify usable triggers and trade-offs, whereas uncertainty calibration measures whether they distinguish evidence, inference, and hypotheses. Scenario coverage is reported separately and excluded from the overall quality score. We additionally use blind pairwise win rates to compare reports.

We compare six report-generation conditions through blind pairwise evaluation. Direct reports use scenario descriptions only, whereas simulation-grounded reports additionally use simulated outcomes and execution traces. The LLM simulation condition achieves the highest LLM-judge win rate of 100.0% in both domains. Human evaluation shows the same advantage in social media, although negotiation results are more mixed.

The results indicate that access to raw simulation artifacts alone is insufficient. Harness-based methods with execution-trace access achieve inconsistent gains and lower win rates than the scenario-forecast pipeline. Effective simulation-grounded reporting requires an evidence-organization layer that summarizes trajectories, identifies recurring mechanisms and scenario patterns, and links them to conditional recommendations. These findings support a simulation-to-report harness within the evaluated settings.

### 5.7 Ablation and Attribution

We compare six report-generation conditions in matched social-media and negotiation ablations for our harness in Table 6. **Direct** receives only the scenario, whereas **Direct-Structured** receives the same conditional-report instructions as the evidence-bearing conditions without simulation evidence. **Raw-Traj** receives execution traces, **Per-Traj-Summary** receives structured summaries of individual trajectories, **Full** receives the matched simulation evidence used by the production workflow, and **Shuffled-Trace** receives evidence from a different report run.

Simulation evidence has domain-dependent effects. In social media, Direct-Structured beats Full in all six judgments, while Full wins only three against Shuffled-Trace, with one tie. In negotiation, Full beats Shuffled-Trace in all six judgments (8.60 vs. 6.45), but wins only four against Direct-Structured, with equal mean scores (8.32). Thus, matched evidence helps relative to mismatched evidence in negotiation. One possible explanation is that negotiation depends more on run-specific offers, commitments, and outcomes, whereas social-media advice is more readily generated from scenario descriptions. This suggests that the marginal value of simulation depends on how much decision-relevant information emerges only during execution. The Full–Shuffled-Trace gap in negotiation further indicates that evidence relevance matters, rather than simply providing additional context. Per-Traj-Summary beats Raw-Traj in all six judgments in each domain, providing consistent support for structured evidence organization. This result suggests that summarizing salient behaviors, state transitions, and outcomes can make trajectory evidence easier to use than exposing the report generator to raw interaction logs. These judgments repeat report orders and evaluation rounds, rather than independent scenarios.

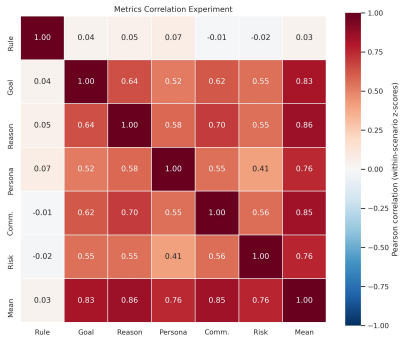



Figure 2: Scenario-normalized metric correlations across all paired model–scenarios. The Rule score shows negligible correlations with LLM-as-a-Judge dimensions, while judge-based behavioral metrics exhibit strong positive associations with each other.

The LLM classifies claims by their post-hoc relationship to reference traces and the information about the table is in Appendix B. Table 7 shows that Per-Traj-Summary has the highest trace-support rate, reaching 80.7% in social media and 57.9% in negotiation. However, scenario-only Direct reports also receive substantial support, including 74.3% in social media. These rates therefore measure post-hoc evidential support, not trace use during generation or information gain beyond scenario descriptions and general knowledge. This overlap indicates that some report claims are consistent with the simulated evidence because they follow from the scenario or general domain knowledge.

## 6 Conclusion

SocialSim-Bench frames multi-agent social simulation as an evidence-generation workflow for decision exploration. It instantiates negotiation and social-media environments, evaluates executable behavior using deterministic rule scores and LLM-as-a-Judge assessments, and transforms multiple execution traces through summarization, mechanism aggregation, and conditional report synthesis. Across the evaluated settings, even capable LLMs tend to rely on simple strategies rather than sustained complex strategies. Relative to direct advice, simulation can surface a broader range of plausible interaction pathways and provide more explicit, trace-grounded accounts of how outcomes emerge. Finally, the analysis of the LLM-as-a-Judge system proves the effectiveness of our datasets.

## AI USE STATEMENT

Generative AI tools were used for manuscript editing, literature search, scenario and persona and other generation, and the construction of synthetic interaction data used in the benchmark. Testing and analysis were also performed with assistance from LLMs. All simulation data, prompts, evaluation procedures, and retrieved references were reviewed and verified by the authors.

### ETHICS STATEMENT

This work does not involve human subjects, private personal data, or real-world interventions. All experiments use synthetic scenarios, simulated agents, and artificial interaction trajectories. The reported simulation outcomes are used for research analysis and are not presented as historical facts, causal estimates, or operational instructions for real-world decisions. We do not claim that the simulated strategies or recommendations are universally valid. The authors reviewed the scenario-generation process, evaluation procedures, and generated artifacts for privacy, safety, fairness, and potential misuse concerns. No additional ethical approval was required for the experiments reported in this paper.

## REFERENCES

Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. Plurals: A system for guiding llms via simulated social ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2025.

Steve Bankes. Exploratory modeling for policy analysis. *Operations research*, 41(3):435–449, 1993.

Benjamin P Bryant and Robert J Lempert. Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. *Technological Forecasting and Social Change*, 77(1): 34–49, 2010.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, 2024.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1274–1282, 2025.

Donald L DeAngelis and Stephanie G Diaz. Decision-making in agent-based modeling: A current review and future prospectus. *Frontiers in Ecology and Evolution*, 6:237, 2019.

Joshua M Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, volume 2026, pp. 28290–28341, 2026.

Robert Lempert. Agent-based modeling as organizational and public policy simulators. *Proceedings of the national academy of sciences*, 99(suppl_3):7195–7196, 2002.

Robert J Lempert. Shaping the next one hundred years: New methods for quantitative, long-term policy analysis. 2003.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Jiachun Li, Zhuoran Jin, Tianyi Men, Yupu Hao, Kejian Zhu, Lingshuai Wang, Dongqi Huang, Longxiang Wang, Shengjia Hua, Lu Wang, et al. Agentic environment engineering for large language models: A survey of environment modeling, synthesis, evaluation, and application. *arXiv preprint arXiv:2606.12191*, 2026a.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. Evaluating scoring bias in llm-as-a-judge. In *International Conference on Database Systems for Advanced Applications*, pp. 19–34. Springer, 2026b.

Yiming Li and Dacheng Tao. Ai agents alone are not (yet) sufficient for social simulation. *arXiv preprint arXiv:2603.00113*, 2026.

Bolei Ma, Yong Cao, Indira Sen, Anna-Carolina Haensch, Frauke Kreuter, Barbara Plank, and Daniel Hershcovich. Too open for opinion? embracing open-endedness in large language models for social simulation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1998–2016, 2026.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Huang Xuanjing, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Computing Surveys*, 58(11):1–41, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents. *arXiv preprint arXiv:2501.05171*, 2025a.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. 2025b.

Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. Botsim: Llm-powered malicious social botnet simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14377–14385, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, et al. Sandboxsocial: A sandbox for social media using multimodal ai agents. In *IJCAI*, pp. 11100–11103, 2025.

Shijun Wan, Xuehai Wu, Jiwen Zhang, Siyuan Wang, and Zhongyu Wei. Can agents read the room? benchmarking visual social intelligence in multimodal simulation. *arXiv preprint arXiv:2606.15152*, 2026.

Xintao Wang, Sirui Zheng, Hongqiu Wu, Weiyuan Li, Jen-tse Huang, Minghao Zhu, Can Zu, Qi Deng, Jiawei Wang, Qianyu He, et al. Agentopia: Long-term life simulation and learning in agent societies. *arXiv preprint arXiv:2606.07513*, 2026.

Zixu Wang, Bin Xie, Bingbing Xu, Shengmao Zhu, Yige Yuan, Liang Pang, LD Su, Zixuan Li, Huawei Shen, and Xueqi Cheng. A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications. *arXiv preprint*, 2025.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing llm agents. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 5163–5186, 2024.

Dizhan Xue, Jing Cui, Shengsheng Qian, Chuanrui Hu, and Changsheng Xu. Some: A realistic benchmark for llm-based social media agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1391–1399, 2026.

Cehao Yang, Xiaojun Wu, Honghao Liu, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Yuanliang Sun, Wenjie Zhang, Zhichao Shi, Yijie Xu, et al. A survey of agent skills: Toward procedural infrastructure for llm agents. 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, et al. Silo-bench: A scalable environment for evaluating distributed coordination in multi-agent llm systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29379–29398, 2026.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024.

Table 5: Environment parameters and action spaces.

| Environment | Participants and horizon | Main action space | Termination and outcome fields |
|---|---|---|---|
| Commercial negotiation | 3–4 agents; invitation phase plus up to 16 sessions; timeline with news events | invite, accept/decline, join/leave, public message, private message, offer, amend, accept, reject, sign | signed contract, failed/expired negotiation, maximum sessions, departure; price, margin, savings, solvency |
| Social media | 6–12 agents; up to 200 global turns; per-agent action quotas | view, post, comment, like, bookmark, follow, share, private message, leave | maximum turns or voluntary departure; posts, replies, reactions, reach, discussion depth, follows, goal coverage |

## A Environment Parameters and Rule Details

Table 5 summarizes the main configuration dimensions. Exact action legality remains phase- and role-dependent.

**Hyperparameter setting.** We set the effectiveness threshold $\tau_{\text{eff}}$, seller-margin normalization threshold $\tau_{\text{margin}}$, activity saturation threshold $\tau_{\text{sess}}$, and component weights $\lambda_{\text{save}}$, $\lambda_{\text{seller}}$, and $\lambda_{\text{activity}}$ before evaluation and keep them fixed across all compared agents. This ensures that agents are evaluated under identical economic constraints and that performance differences arise from their decisions and interaction behavior rather than changes in the scoring rule.

## B The Claim Results

The table reports an LLM-assisted claim-level attribution analysis for reports generated in the social-media and negotiation domains. Domain identifies the simulation environment, while Condition denotes the report-generation input, including Direct, Direct-Structured, Raw-Traj, Per-Traj-Summary, Full, and Shuffled-Trace. Claims is the number of verifiable claims extracted from each report condition. Trace-supported is the proportion of claims that are either directly supported by a specific simulation event or reasonably inferred from repeated trace patterns. Knowledge denotes plausible recommendations or statements that are not supported by the supplied simulation evidence and are likely elicited from general model knowledge. Unsupported denotes claims that lack adequate evidence or conflict with the available traces. Not verifiable denotes claims for which the available evidence is insufficient to determine whether they are supported. The trace-supported rate combines explicit trace grounding and inference from trace patterns, and the four attribution categories together describe how report claims relate to the available simulation evidence.

## C Why Real-World Logs Are Not Direct Ground Truth

Real-world interaction data remain important, but an observed log is generally not a ground-truth label for a decision. For a given unit, we typically observe only one action and its realized outcome, while the outcomes under alternative actions remain unobserved. The observed action may also reflect incomplete information, selection effects, institutional constraints, changing incentives, or objectives that differ from those used in evaluation. Thus, neither a historical decision nor the outcome of a single model simulation should be treated as the uniquely optimal decision. Real-world actors can make mistakes, and simulated agents can likewise produce suboptimal decisions. Consequently, real-world observations cannot, by themselves, provide a definitive ground-truth benchmark for the quality of a decision report.

This limitation also means that a simulated outcome should not be interpreted as a prediction that will necessarily occur in the real world. Rather than treating simulation traces as evidence of what will happen, our reports treat the mechanisms identified in simulation as hypotheses about how an interaction may unfold under particular conditions. The role of real-world data is therefore com-

Table 6: Report-generation ablations in both domains. Wins counts judgments favoring the first method out of six, evaluated in both report orders across three rounds. Mean scores follow the comparison order. The social-media Full vs. Shuffled-Trace comparison includes one tie.

| Domain | Comparison | Wins | Mean score |
|---|---|---|---|
| Social media | Direct-Structured vs. Full | 6 | 8.33 vs. 7.63 |
| | Full vs. Shuffled-Trace | 3 | 7.82 vs. 7.88 |
| | Per-Traj-Summary vs. Raw-Traj | 6 | 8.37 vs. 8.08 |
| Negotiation | Full vs. Direct-Structured | 4 | 8.32 vs. 8.32 |
| | Full vs. Shuffled-Trace | 6 | 8.60 vs. 6.45 |
| | Per-Traj-Summary vs. Raw-Traj | 6 | 8.58 vs. 7.97 |

Table 7: LLM-assisted claim-level attribution in the social-media and commercial-negotiation report conditions. Post-hoc support combines claims classified as explicitly grounded in evidence and inferred from trace patterns.

| Domain | Condition | Claims | Trace-supported | Knowledge | Unsupported | Not verifiable |
|---|---|---|---|---|---|---|
| Social media | Direct | 35 | 74.3% | 25.7% | 0.0% | 0.0% |
| | Direct-Structured | 250 | 67.6% | 20.8% | 6.4% | 5.2% |
| | Raw-Traj | 199 | 64.3% | 29.1% | 1.5% | 5.0% |
| | Per-Traj-Summary | 306 | **80.7%** | 12.4% | 1.3% | 5.6% |
| | Full | 315 | 78.7% | 15.6% | 1.3% | 4.4% |
| | Shuffled-Trace | 257 | 72.4% | 20.6% | 2.7% | 4.3% |
| Negotiation | Direct | 132 | 18.2% | 60.6% | 17.4% | 3.8% |
| | Direct-Structured | 182 | 28.0% | 60.4% | 4.9% | 6.6% |
| | Raw-Traj | 294 | 20.4% | 58.2% | 16.7% | 4.8% |
| | Per-Traj-Summary | 361 | **57.9%** | 36.6% | 2.2% | 3.3% |
| | Full | 365 | 50.1% | 37.5% | 9.9% | 2.5% |
| | Shuffled-Trace | 91 | 22.0% | 49.5% | 23.1% | 5.5% |

plementary: they can be used to calibrate scenario distributions and constraints, falsify implausible mechanisms, and externally validate candidate recommendations through prospective or controlled studies.

Accordingly, an effective decision report should not simply prescribe a single action as universally optimal. A report generally cannot observe all relevant information in the current real-world environment, and its simulated evidence cannot guarantee that a particular recommendation will succeed. Instead, the report should identify distinct decision scenarios, characterize the conditions and trade-offs associated with each scenario, and provide corresponding courses of action. Each proposed course of action should be accompanied by its triggering conditions, uncertainty, and, where possible, low-risk tests that can be used to assess whether the hypothesized mechanism holds in the target environment. In this way, the report serves as a conditional decision aid rather than an oracle: it organizes plausible response strategies while leaving the final decision to be informed by current real-world evidence.

### C.1 Skills and Refine

The **skills** intervention is implemented as an additional prompt-injection layer for the focal agent. At the beginning of an episode, the system loads a predefined skill pool from the evolution module. For each decision turn, the agent observes the current role, goals, visible history, available actions, and environment state. A lightweight activation procedure then matches the current observation against skill keywords and selects the relevant skills. Each activated skill is rendered as a structured prompt block containing its name, trigger condition, and behavioral guidance, and this block is appended to the focal agent's decision prompt before the language model generates an action. The intervention can be configured to target only the focal agent, the other agents, or all agents, and can be combined with either plain or summarizing memory. The environment and action parser remain unchanged, so skills affect the agent's decision context without modifying the transition rules.

Table 8: Correlation Between Human and LLM Scores Across Evaluation Dimensions.

| Scope | $n$ | Human Mean | LLM Mean | Pearson $r$ | $p$-value | Bias | MAE |
|---|---|---|---|---|---|---|---|
| Combined (domain mean) | $2 \times 18$ | 6.39 | 6.83 | 0.60 | — | 0.44 | 0.661 |
| Commercial Negotiation | 18 | 6.24 | 6.92 | 0.37 | 0.13 | 0.68 | 1.06 |
| Social Media | 18 | 6.54 | 6.74 | 0.83 | 0.00002 | 0.200 | 0.27 |

Table 9: Benchmark results across the social-media and commercial negotiation environments. For GPT-5.6-Luna, we report the mean and standard deviation over three independent runs. LLM Mean denotes the mean of the five LLM-as-a-Judge dimensions. We multiply the Rule scores by 10 for fair comparison.

| Domain | Model | Rule | Goal | Reason | Persona | Comm. | Risk | LLM Mean |
|---|---|---|---|---|---|---|---|---|
| **Social media** | GPT-5.6-Luna | $0.89 \pm 0.05$ | $6.06 \pm 0.03$ | $6.56 \pm 0.06$ | $6.31 \pm 0.18$ | $6.69 \pm 0.13$ | $7.34 \pm 0.09$ | $6.60 \pm 0.07$ |
| **Commercial negotiation** | GPT-5.6-Luna | $2.21 \pm 0.04$ | $6.77 \pm 0.06$ | $6.47 \pm 0.06$ | $7.94 \pm 0.03$ | $6.81 \pm 0.05$ | $6.86 \pm 0.04$ | $6.97 \pm 0.03$ |

The **refine** intervention is implemented as a post-generation response wrapper around the focal agent. The base agent first produces a normalized action and its natural-language realization according to the regular action protocol. A separate polishing module then receives the agent profile, current observation, recent interaction history, generated action, and the designated polishing model. It is instructed to improve clarity, tone, and persona consistency while preserving the original action type, recipients, contract identifiers, numerical terms, commitments, and other state-changing fields. The refined response replaces the original natural-language output only after parsing and consistency checks, while the environment continues to execute the normalized action through the original controller. Thus, refine changes the presentation and response strategy of the focal agent, whereas skills change the information and guidance available before action generation.

## D STRATEGIES

We use the following definitions for the strategies analyzed:

In the negotiation setting:

- **competitor_leverage**: Actively introduces, compares, or leverages multiple suppliers, buyers, or alternatives.
- **mult_party_offer_comparison**: Obtains or compares offers and terms from multiple participants.
- **anchoring**: Makes an early favorable price or term proposal to establish a reference point.
- **coalition**: Builds a cooperative relationship with selected participants to change bargaining power.
- **coalition_division**: Uses conflicting interests among participants to prevent a stable opposing coalition.
- **information_acquisition**: Seeks prices, requirements, constraints, or reservation information from participants.
- **information_control**: Selectively discloses resources, requirements, or alternatives.
- **conditional_exchange**: Trades price, volume, payment, delivery, or other terms conditionally.
- **issue_separation**: Separates issues rather than binding every issue into one negotiation move.
- **concession_management**: Adjusts concessions according to participant reactions without exposing a reservation point too early.
- **negotiation_sequence_control**: Controls whom to engage first, later, or again to exploit multi-party competition.
- **exit_threat**: Maintains alternatives and uses credible pause or exit conditions.

- **relationship_maintenance**: Preserves cooperation with participants who do not reach agreement.
- **dynamic_adaptation**: Adjusts strategy after new offers, events, or participant behavior.
- **other_strategy**: Uses a deliberate focal-agent negotiation strategy not covered by the listed catalog.
- **no_catalog_strategy**: Uses none of the listed strategies; select only when no catalog strategy has direct evidence.

In the social-media scenario:

- **content_positioning**: Selects topics and expression for a target audience.
- **topic_selection**: Selects a current relevant or trending topic.
- **keyword_tag_use**: Uses relevant keywords or tags to expand potential audience.
- **content_reposting**: Reposts or republishes after adapting to early feedback.
- **traffic_seeking_interaction**: Comments on high-visibility content to reach an existing audience.
- **network_leverage**: Interacts, follows, or collaborates with high-influence accounts.
- **sharing_strategy**: Selects content to share or redistribute with deliberate framing.
- **engagement_maintenance**: Replies and sustains discussion to convert exposure into attention.
- **controversy_tradeoff**: Uses or avoids contentious topics while weighing reach and risk.
- **audience_segmentation**: Adapts content for different audience groups.
- **cross_user_diffusion**: Uses multiple relationship nodes to broaden diffusion.
- **trend_following**: Participates while event attention is rising.
- **reputation_exposure_tradeoff**: Balances immediate exposure against account reputation or trust.
- **feedback_driven_adaptation**: Changes later behavior based on likes, comments, shares, or other feedback.
- **other_strategy**: Uses a deliberate focal-agent social-media strategy not covered by the listed catalog.
- **no_catalog_strategy**: Uses none of the listed strategies; select only when no catalog strategy has direct evidence.

# E Correlation Between Human and LLM Scores

As shown in Table 8, across all 36 episodes, human scores exhibit a moderate and statistically significant positive correlation with LLM-based scores Pearson's $r = 0.60$. The mean LLM score is 6.83, which is higher than the mean human score of 6.39, yielding a mean bias of 0.439 and a mean absolute error of 0.661.

## E.1 Stability

Across three independent runs, GPT-5.6-Luna achieves relatively stable performance in both simulation environments in Table 9. In the commercial negotiation setting, it obtains a Rule score of $2.21 \pm 0.04$ and an LLM Mean of $6.97 \pm 0.03$, with particularly consistent persona alignment and communication quality. In the social-media setting, GPT-5.6-Luna achieves a Rule score of $0.89 \pm 0.05$ and an LLM Mean of $6.60 \pm 0.07$. Its strongest dimension is risk and constraint compliance, while persona consistency shows comparatively larger variation. Overall, the small standard deviations across most dimensions indicate that the model exhibits reasonably consistent behavior across repeated simulations.