

Testing Behavioral Adaptation in Repeated Interactions Between LLM Agents

Lixin Liu¹ Yixuan Li¹ Xiankun Lin² Yutian Lu¹ Shiguang Ni^{1,*}

¹Tsinghua University

²Sun Yat-sen University

University Town, Xili, Nanshan District
Shenzhen, Guangdong 518055, China

66 Gongchang Road, Guangming District
Shenzhen, Guangdong 518107, China

{liulx25,yixuan-li25}@mails.tsinghua.edu.cn

linxk7@mail2.sysu.edu.cn

yutianlu@sz.tsinghua.edu.cn

*Corresponding author: ni.shiguang@sz.tsinghua.edu.cn

Abstract

High aggregate scores can make language agents appear socially responsive even when they merely repeat a prompt-conditioned action. We introduce an identifiability-first framework that separates observed outcomes from evidence of adaptation. Across 3,465 released iterated-prisoner’s-dilemma simulations, 94.5% never change joint regime, leaving reciprocity largely unidentifiable. Controlled histories across nine models reveal opponent-dependent responses that vary substantially across equivalent interfaces. The released checkpoints have the smallest opponent-recency effects. Only the intervention cohort shows reliable propagation after a scripted defection. The diagnostic transfers to longer histories, a multiplayer game, and released-interface wrappers, although model rankings do not. Outcome-only evaluation conflates regime selection with adaptive behavior. Separating opportunity, response, and robustness prevents stable outcomes from being mistaken for social dynamics.

1 Introduction

Language models are increasingly evaluated in task benchmarks (Liu et al., 2024) and social applications (Park et al., 2023; Aher et al., 2023; Horton et al., 2023; Argyle et al., 2023; Santurkar et al., 2023; Wang et al., 2024). Existing evaluations often summarize interaction quality with aggregate outcomes such as task success or cooperation rates. These scores can hide the process that produced them.

Consider two agents playing repeated Prisoner’s Dilemma. Each round, both choose cooperate (C) or defect (D) before seeing the other’s current choice. Our controlled prompts specify the rules, previous actions and payoffs, and the next choice. Mutual cooperation gives each agent 3 points; unilateral defection gives the defector 5 and the cooperator 0; mutual defection gives each 1.

The released logs also include payoff and framing variants. A ten-round trace of mutual cooperation cannot distinguish always cooperating from copying a cooperative opponent: both rules predict the same observed actions. Testing reciprocity requires a history on which those predictions differ.

Recent work examines robustness, reproducibility, and behavioral validity (Sarangi et al., 2026; Ye et al., 2026; Zhou et al., 2025; Pandey and Jajoo, 2026; Ziems et al., 2026), strategic behavior in games and negotiations (Fontana et al., 2025; Duan et al., 2024; Abdelnabi et al., 2024), payoff scaling (Huynh et al., 2026), and contextual framing (Robinson and Burden, 2025). Static evaluations also show prompt-format sensitivity (Sclar et al., 2024). We ask: *does the observed interaction test the proposed mechanism?* Candidates agreeing on every observed history are observationally equivalent; trajectory fit cannot distinguish them. This is a classical identification problem (Barrie and Törnberg, 2025; Pearl, 2009).

We organize evaluation around three questions (Fig. 1). First, *opportunity* asks whether logged decisions distinguish candidate behaviors. We call this audit the Cooperation Regime Profile (CRP). Second, *response* asks how a model acts when the missing histories are supplied. The Counterfactual History Cube (CHC) presents all three-round joint histories, including contexts that distinguish repetition from copying an opponent. Finally, *robustness* asks whether the response survives interface and decoding changes. Paired sequential tests then measure its persistence over later rounds.

The underlying identification principle is classical. Our contribution is to operationalize it for LLM-agent evaluation: we define claim-specific coverage measures for interaction logs, pair them with controlled history interventions, and require interface checks before interpreting a response as a model property. Table 1 separates the reused logs from the experiments newly run for this study. Re-

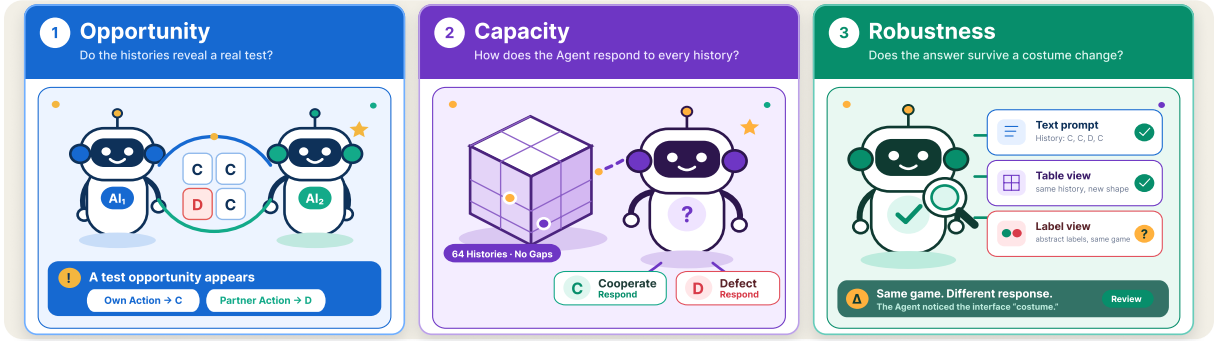


Figure 1: Three evaluation questions: did the logged histories test adaptation, can the model respond when relevant histories are supplied, and does that response survive interface changes? Capacity refers to response under controlled histories.

Table 1: Study map: released-log reanalysis and controlled experiments.

Setting	Material and procedure	Question answered
Released-log audit	Reused: 3,465 ten-round Social Sim'26 trajectories.	Did play contain opportunities to distinguish candidate behaviors?
Controlled histories	Supply all 64 three-round histories in three renderings.	Does changing opponent history change the next action?
Sequential shock	Paired ten-round runs with one scripted defection.	Does the response extend to later rounds?
Robustness and boundary tests	Vary seeds, payoff/label choices, history length, game, and prompt wrappers.	Which effects survive changes in sampling, interface, and game?

leased runs mostly settle into absorbing regimes and rarely expose adaptation. Under controlled variation, models do respond to their opponents, but the strength of that response depends on the checkpoint, game, and interface.

2 Identifiability-first evaluation

The three stages deliberately support different claims. The log audit measures whether adaptation was testable, controlled histories measure response when relevant variation is supplied, and robustness controls test whether that response is stable enough to interpret.

2.1 Opportunity: observational CRP

For binary actions $a_t^1, a_t^2 \in \{0, 1\}$, CRP maps each round to mutual cooperation (CC), one-sided defection (ASYM), or mutual defection (DD). It reports how often play occupies, enters, or leaves each regime. It also counts *identifying opportunities*: player-rounds in which copying the opponent and repeating one's own action predict different choices, alongside opportunities following an opponent action change. A conditional metric is undefined when its required event never occurs.

We score five diagnostic rules on rounds $2:T$,

treating the first action as given. Repeat-first repeats that initial choice; inertia repeats the previous own choice; tit-for-tat (TFT) copies the opponent's previous choice. Grim defects after any opponent defection. Win-stay/lose-shift (WSLS) repeats the own action after a payoff of 3 or 5 and switches after 0 or 1 in the base game (Axelrod and Hamilton, 1981; Nowak and Sigmund, 1993).

2.2 Response under controlled histories

CHC presents each model with all $4^3 = 64$ three-round joint histories and asks for its round-four action. Write each round as (focal action, opponent action). After (C,C), (C,D), (D,C), TFT predicts cooperation because the opponent last cooperated; inertia predicts defection because the focal agent last defected. We render each history as lines, a table, and abstract GREEN/BLUE labels with identical payoffs. The primary probes use temperature 0.5, 20 fixed seeds, and constrained two-choice decoding.

The cube balances focal and opponent actions. *Balanced policy agreement* averages agreement separately on histories where a candidate predicts C and D, giving both groups equal weight. A constant agent scores 0.5. We also pair histories that

change one action: the opponent’s latest, the focal agent’s latest, or an earlier opponent action. Displayed payoffs are recomputed consistently. For opponent recency, $\mathcal{H}$ contains the 32 histories with the opponent’s third action left unspecified. Let $p_x(h)$ denote round-four cooperation probability when that action is x , averaged over the three formats:

$$\Delta_{\text{opp}} = |\mathcal{H}|^{-1} \sum_{h \in \mathcal{H}} [p_C(h) - p_D(h)]. \quad (1)$$

Thus $\Delta_{\text{opp}} > 0$ means more cooperation after opponent cooperation than defection. Own-recency and remote-history effects use analogous contrasts. The same seed is reused within pairs; effects measure response to the complete action–payoff record.

2.3 Robustness: interface and propagation controls

For each history, we subtract the lowest cooperation rate across formats from the highest. We compare this spread with the difference between two disjoint ten-seed estimates under the same format. Resampling histories summarizes variation across contexts in this exhaustive set; it does not estimate uncertainty over a deployment population. Separately, 2,000 resamples of seed clusters quantify decoding uncertainty while retaining all histories and formats. Greedy and alternative-temperature runs check decoding sensitivity.

We further apply five game-isomorphic transformations: positive payoff scaling, common payoff shift, reversed choice order, neutral labels, and label permutation. For each format, we generate 20 paired trajectories against an always-cooperative opponent or one that defects only in round three. Prefixes share focal draws until the shock is observable. Round four measures immediate response, and rounds 5:10 measure propagation. Seed-cluster intervals retain all formats and trajectories.

3 Observational audit of repeated play

We analyze all 3,465 released Social Sim’26 iterated-PD runs. The core grid contains 2,160 runs crossing instruction framing, payoff scale, vignette, model, persona, and five seeds. The remaining 1,305 runs isolate representation variations. Our analysis uses only canonical action events.

Table 2 shows that two constant paths cover 94.5% of runs. Only 174 runs expose an agent to any opponent action change. The reciprocal persona has no trial distinguishing reciprocity from

Table 2: Natural-log dynamics and identification coverage across 3,465 released runs.

Diagnostic	Result
No joint-regime switch	3,276/3,465 (94.5%)
Distinct exact 10-round paths	25
Runs with CC $\rightarrow$ $\neg$ CC collapse	124/3,465 (3.6%)
Recovery within two rounds	0/124
Reciprocal persona identifiable	0/540
Same path under all five seeds	674/693 (97.3%)

always-cooperate. Four candidates achieve 97.5–99.5% raw action accuracy, yet no run has a unique perfect explanation. All 3,276 constant runs admit multiple policies, and none of 189 dynamic runs perfectly matches a candidate.

Aggregate effects do not repair this support failure. In our reanalysis of the released logs, moralized instructions raise cooperation by 6.7 percentage points, and risk framing lowers it by 25.2 percentage points. Most differences are already present in round one. Across nine isolated representation arms, the moralized effect ranges from 0 to +43.3 points, and both payoff-scale conclusions reverse direction. The risk effect keeps its preregistered direction in all 10 arms. The other four effects do so in only 1–4 arms. The released logs therefore support prompt-conditioned regime selection, not reciprocity.

Observed responsiveness is localized. Across 328 player-rounds following an opponent action change, agents switch 114 times (34.8%). Across 55,112 no-change trials, they switch only 244 times (0.44%). Competitive personas account for 267/328 change trials, while cooperative and reciprocal personas generate none (Appendix B).

4 Counterfactual history intervention

We apply CHC to three Qwen2.5 scales (7B, 14B, 72B), Qwen3-8B, Mistral-7B, and Seed-OSS-36B. These six models form an intervention cohort rather than replacements for the four models in the released logs. Figure 2 summarizes the evidence chain. Appendix C provides the complete tables and prompts.

Identifying histories elicit response, but not one mechanism. Across the six intervention models, the opponent-recency effect is positive (8.6–44.7 points). Across all nine models, seed-cluster intervals exclude zero for seven, while the Qwen3.5-4B/9B intervals include zero (Fig. 6). Own-action

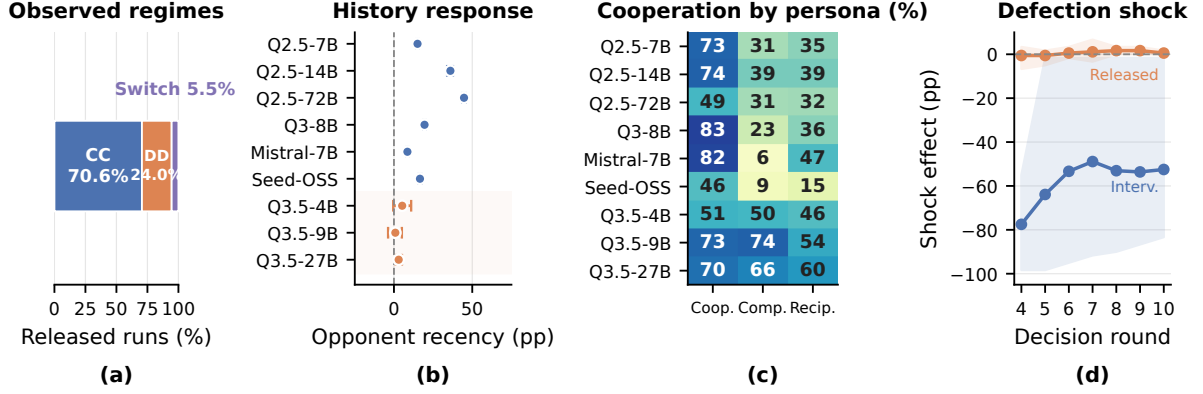


Figure 2: The evidence chain. (a) 94.5% of released runs never change joint regime. (b) The released checkpoints have the three smallest opponent-recency effects, and the Qwen3.5-4B/9B intervals include zero. (c) Cooperation rates under the three released personas. (d) The one-defection impulse propagates in the intervention cohort, while no released-cohort cumulative interval excludes zero. Bands span model ranges.

and remote-history effects vary in magnitude and sign, and the best candidate changes across models. Controlled histories therefore reveal response capacity that the logs cannot identify. They do not reveal one mechanism shared across models.

The released cohort responds least. The three Qwen3.5 checkpoints represented in the released logs have the smallest opponent-recency effects under the common CHC interface (+0.9–+5.3 points, Table 9). The seed intervals for the 4B and 9B checkpoints include zero. The 9B and 27B checkpoints also have the two lowest exact three-format agreements (18.4% and 11.2%). This pattern is consistent with the absorbing logs. It does not show that checkpoint identity caused those trajectories, since CHC does not reconstruct the released simulator (see Limitations).

Released personas shape behavior when history varies. Persona-conditioned probes prepend the three released texts and re-probe all nine models (108,864 decisions, Table 6). The cooperative instruction raises cooperation for every model by +0.5 to +62.6 points. The reciprocal instruction never lowers Δ_{opp} and raises it by up to +33.0 points. Balanced tit-for-tat agreement also rises for every model (Appendix E). These shifts indicate approximation when histories vary, not fixed-rule compliance. In reciprocal runs, opponents never change action. Reciprocity remains unidentifiable (0/540), and the mirror rule remains untested.

Representation effects exceed sampling noise. Across histories, mean format-induced cooperation ranges are 8.8–56.2 points. Independent 10-seed halves differ by only 1.1–4.8 points. All

format-range intervals exclude zero. Per-history format span exceeds split-half noise by 7.7–51.3 points, with every model’s interval excluding zero. This dominance holds in at least 17.2% of histories per model. Greedy agreement remains only 43.8–90.6%. Alternative temperatures keep cooperation and opponent-recency within four points. We report these measures together as the *interface-robustness diagnostic*.

Game-equivalent interfaces change conclusions.

Positive-affine payoff transformations preserve all greedy actions in only 70.3–96.9% of histories, despite preserving every trajectory ranking. Neutral-label permutations are less stable (7.8–53.1%). Across the six renderings, the opponent-recency effect can change sign for Qwen2.5-14B and Mistral. Evaluators should therefore report whether mechanism estimates are invariant, not only average cooperation.

A single defection can affect later rounds.

In the intervention cohort, the scripted defection lowers immediate cooperation by 55.0–98.3 points. Cumulative shock-minus-control cooperation over rounds 4:10 ranges from -6.43 to -0.65, with all 6 intervals excluding zero. Later-round divergence spans 1.7–98.3%. Qwen2.5-14B realigns, while Seed-OSS does not. Because Seed-OSS is effectively deterministic (11/192 CHC cells vary), we report its cluster estimate as a point. In the released cohort, no cumulative interval excludes zero (Fig. 7). CHC thus detects propagation only in the intervention cohort, although it does not reconstruct the original simulator.

Table 3: Evidence required for interaction claims.

Level	Question	Required report
Outcome	What occurred?	Occupancy and trajectory, not only a mean.
Opportunity	Was adaptation testable?	Identifying-event counts and undefined cells.
Response	Does behavior change under support?	Matched history effects and policy equivalence.
Robustness	Does the claim survive interfaces?	Model and representation ranges.

CHC diagnoses capacity, not the mechanism behind past runs. A positive Δ_{opp} in Equation (1) shows that a focal model can use opponent history under controlled support. It does not turn an always-CC simulation into evidence for TFT because the original agents never faced the counterfactual context. Sparse dynamics likewise do not prove incapability (Appendix E). The remaining explanations include checkpoint behavior, prompt representation, and endogenous history generation. This comparison cannot separate them.

5 What should an agent evaluation claim?

Occupancy records what happened. Identifying-event coverage tests whether a mechanism could have affected an observed decision. Controlled histories measure response when the needed variation is supplied, and interface controls test whether that response is stable.

Table 3 summarizes the required reports. Since 97.3% of conditions repeat one path, we recommend separate outcome metrics, opportunity denominators, intervention responses, and representation ranges.

Outcome claims require a fixed interaction distribution. Mechanism claims require identifying events and matched counterfactuals. Model rankings require the same diagnostic across interfaces. A claim should stop at the first failed link.

Extension beyond binary games. The framework does not require fitting the five policies used here. An evaluator can follow four steps: (1) state the behavioral claim and a competing explanation; (2) identify history features on which those explanations predict different next actions; (3) report how often those contrasts occur naturally, then construct matched histories when coverage is sparse; and (4) repeat the contrast under equivalent interfaces and in a short rollout. The public-goods experiment in

Appendix F illustrates the change from two-player actions to a multiplayer joint state: the relevant input is the coplayers’ contribution history.

A concrete dialogue example. To test whether an agent revises a plan after partner feedback, compare that claim with repeating the previous plan. Construct paired contexts in which a prerequisite either succeeds or fails, while holding the task, previous plan, and other information fixed. Update any downstream facts implied by the changed prerequisite. Define and validate a rubric for plan revision before scoring free-text responses, and report ambiguous cases. Equivalent phrasings test interface dependence; short rollouts test whether the revised plan persists and remains usable. This is guidance for applying the diagnostic, not an additional dialogue experiment or evidence that the tested game responses generalize to dialogue.

Coverage at larger scales. For n players, m actions per player, and L rounds, exhaustive enumeration requires m^{nL} histories before formats and seeds. Our binary dyad therefore grows from 64 histories at three rounds to 1,024 at five. Exhaustive coverage removes history-selection freedom in these small spaces. For larger spaces, sample complete contrast pairs from declared strata, such as feedback type and task difficulty, and report the untested strata. If contrasts are selected adaptively, use a separate fixed evaluation set. Responses then support claims about the sampled contexts; broader deployment claims additionally require natural-history coverage and validation in the target environment.

6 Conclusion

Repeated cooperation is not evidence of reciprocity unless the interaction gives reciprocity an opportunity to change a decision. The released logs mainly establish regime selection: 94.5% never change joint regime. Controlled histories show that models can respond to opponent behavior, but the magnitude depends strongly on checkpoint and interface, and a scripted response does not guarantee that the behavior will arise endogenously. Separating observed outcome, identifying opportunity, controlled response, and robustness yields claims that match the evidence. The diagnostic logic transfers across longer histories and a multiplayer game, while model rankings do not.

Limitations

Findings are limited to the tested models, prompts, and games. Controlled probes do not reconstruct the original simulator, establish adaptation in natural interactions, or reveal internal mechanisms.

References

- Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems*, volume 37, pages 83548–83599.
- Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.
- Robert Axelrod and William D. Hamilton. 1981. The evolution of cooperation. *Science*, 211(4489):1390–1396.
- Christopher Barrie and Petter Törnberg. 2025. Emergent LLM behaviors are observationally equivalent to data leakage. *arXiv preprint arXiv:2505.23796*.
- Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. GTBench: Uncovering the strategic reasoning capabilities of LLMs via game-theoretic evaluations. In *Advances in Neural Information Processing Systems*, volume 37, pages 28219–28253.
- Nicolò Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer than humans: How do large language models behave in the prisoner’s dilemma? *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):522–535.
- John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper 31122, National Bureau of Economic Research.
- Trung-Kiet Huynh, Dao-Sy Duy-Minh, Thanh-Bang Cao, Phong-Hao Le, Hong-Dan Nguyen, Phu-Quy Nguyen-Lam, Minh-Luan Nguyen-Vo, Hong-Phat Pham, Phu-Hoa Pham, Thien-Kim Than, Chi-Nguyen Tran, Huy Tran, Gia-Thoai Tran-Le, Alessio Buscemi, Le Hong Trang, and The Anh Han. 2026. Payoff scaling shapes cooperation in LLM agents across languages. *arXiv preprint arXiv:2601.19082*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*.
- Martin Nowak and Karl Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner’s dilemma game. *Nature*, 364(6432):56–58.
- Atharva Pandey and Gautam Jajoo. 2026. Reason-mediated behavioral models for auditing LLM social simulators. *arXiv preprint arXiv:2607.24649*.
- Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulators of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.
- Isaac Robinson and John Burden. 2025. Framing the game: How context shapes LLM decision-making. *arXiv preprint arXiv:2503.04840*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Sneheel Sarangi, Maximilian Puelma Touzel, Aurélien Bück-Kaeffer, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. 2026. EASE configuration facilitates a reproducible science of LLM social simulations. *arXiv preprint arXiv:2605.30258*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*.
- Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA- π : Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940.

Jinyi Ye, Lei Cao, Ding Chen, and Emilio Ferrara. 2026. Stop drawing scientific claims from LLM social simulations without robustness audits. *arXiv preprint arXiv:2605.18890*.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025. The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. *arXiv preprint arXiv:2509.18052*.

Caleb Ziems, William Held, Su Doga Karaca, David Grusky, Tatsunori Hashimoto, and Diyi Yang. 2026. Will scaling improve social simulation with LLMs? *arXiv preprint arXiv:2607.02464*.

A CRP definitions and coverage

Table 4 defines each CRP metric and the event required for its denominator.

CC persistence is defined for 2,570/3,465 runs (74.2%), recovery for 124 (3.6%), and reciprocity over inertia for 140 (4.0%). Only 174 runs contain an opponent-change opportunity. Of 693 conditions, 674 have one path across all seeds, 15 have two, four have three, and none has five.

B Observational mechanism audit

Treating the opening action as fixed, we score repeat-first, inertia, TFT, grim, and WSLS on 18 later player-rounds. All five score 100% on the 540 reciprocal-persona runs because those runs remain at CC. On the full data, raw accuracy is high for several candidates, while accuracy falls on the 189 dynamic runs and no candidate fits one exactly. This is why we use identifying coverage before policy fit.

The signal decomposition further separates initial selection from later dynamics. Moralized, risk, low-stakes, high-stakes, and fictional effects are respectively +6.7, -25.2, -7.9, -5.7, and -4.6 cooperation points overall, but their later-minus-initial components are only +1.0, +4.0, +1.2, +1.1, and -0.5. Competitive persona is the exception: its -32.2-point overall effect includes a -14.1-point later component.

C Counterfactual History Cube details

C.1 Protocol

CHC crosses all 4^3 ordered histories with three representations and sampling seeds 8001–8020. Each prompt specifies a fixed opponent, the standard $(T, R, P, S) = (5, 3, 1, 0)$ payoff matrix, all three completed rounds, and a request for the fourth action. The canonical-line and canonical-table forms

use COOPERATE/DEFECT. The abstract form replaces them with GREEN/BLUE everywhere while preserving payoff semantics.

vLLM uses a two-choice guided grammar. The Transformers fallback uses a token-prefix trie that permits only prefixes of the same two labels and then EOS. Normalization strips surrounding whitespace, normalizes case, and requires an exact legal label. Any other output aborts the run, with no retry or silent fallback. All 177,048 stored raw outputs, including appendix controls, are valid (0 invalid).

The model receives no policy name or recommendation. For example, a canonical history is rendered as:

```
Round 1: you=COOPERATE (3); oppo-
nent=COOPERATE (3)
Round 2: you=COOPERATE (0); oppo-
nent=DEFECT (5)
Round 3: you=DEFECT (5); oppo-
nent=COOPERATE (0)
Choose your action for round 4.
```

This case distinguishes static/TFT (cooperate) from inertia/grim/WSLS (defect). Exhaustive enumeration avoids selecting examples favorable to a preferred strategy.

C.2 Policy-score interpretation

We retain raw agreement as a diagnostic only. Grim predicts defection on 87.5% of the exhaustive histories. An always-defect agent would therefore achieve 87.5% raw grim agreement but exactly 50% balanced agreement. All policy comparisons in Table 5 use the latter.

C.3 Persona-conditioned CHC

Each of the three released persona instructions is prepended to the CHC user message exactly as the released runs rendered it (PERSONA: A game player with no prior information about the other player. *<text>*). The system role, rules, payoffs, history, and round-four request are unchanged. The COMPETITIVE persona in the released data is the strategic instruction. The probe contains 108,864 decisions ($9 \text{ models} \times 3 \text{ personas} \times 64 \text{ histories} \times 3 \text{ representations} \times 20 \text{ seeds}$, plus the greedy slice). Table 6 reports the cooperation and opponent-recency effects. Figure 3 shows the opponent-recency shift for each model and persona.

C.4 Robustness controls

The seed extension adds 15 seeds without changing histories, prompts, temperature, or decoding constraints. For each model and history, the format

Table 4: Definitions and coverage conditions for CRP metrics.

Metric	Definition
Occupancy	Fraction of rounds in CC, ASYM, or DD, summing to one.
Collapse	$CC \rightarrow \neg CC$ transitions divided by all transitions.
CC persistence	$CC \rightarrow CC$ divided by transitions originating in CC. Undefined without a CC origin.
Recovery within 2	Collapses followed by CC within two rounds, divided by collapses. Undefined without collapse.
Switch rate	Adjacent joint regimes that differ, divided by $T - 1$.
Endgame delta	Last-three minus first-three cooperation rate.
Reciprocity over inertia	Mean of +1 or -1 where copying the opponent and repeating oneself disagree.
Path fingerprint	Base-4 encoding of ordered joint actions.

Table 5: Twenty-seed CHC results. Rates and agreements are percentages. The Δ columns are paired cooperation effects in percentage points. Best/Bal. is the highest balanced candidate agreement. Format is exact three-format agreement.

Model	Coop.	Best	Bal.	Δ_{opp}	Δ_{own}	Δ_{remote}	Format
Qwen2.5-7B	36.8	WSLS	63.5	+15.1	-19.6	+21.7	56.7
Qwen2.5-14B	47.1	WSLS	70.3	+36.0	+25.3	+4.6	49.4
Qwen2.5-72B	38.1	TFT	72.4	+44.7	+21.8	-3.5	61.3
Qwen3-8B	20.6	INERTIA	63.2	+19.5	+26.3	+1.9	79.5
Mistral-7B	42.2	INERTIA	57.8	+8.6	+15.6	-5.8	43.3
Seed-OSS-36B	8.4	GRIM	60.0	+16.6	+16.6	+4.2	90.9

Table 6: Persona-conditioned CHC. Cooperation rates are percentages and opponent-recency effects are percentage points. Columns compare no persona (NONE) with the three released instructions.

Model	Cooperation rate				Opponent-recency Δ_{opp}			
	None	Coop.	Comp.	Recip.	None	Coop.	Comp.	Recip.
Qwen2.5-7B	36.8	73.0	30.7	35.5	+15.1	+21.7	+11.9	+44.6
Qwen2.5-14B	47.1	74.0	38.9	38.8	+36.0	+24.7	+34.8	+69.0
Qwen2.5-72B	38.1	48.9	31.2	31.9	+44.7	+43.6	+39.1	+63.8
Qwen3-8B	20.6	83.2	23.2	35.8	+19.5	+0.5	+21.5	+29.6
Mistral-7B	42.2	81.7	5.9	47.3	+8.6	-1.1	+3.6	+16.1
Seed-OSS-36B	8.4	46.3	8.8	15.2	+16.6	+32.7	+15.1	+30.3
Qwen3.5-4B	28.4	50.9	49.6	46.1	+5.3	+10.6	+5.6	+7.4
Qwen3.5-9B	51.7	73.0	74.1	54.3	+0.9	+8.9	+6.6	+2.6
Qwen3.5-27B	69.8	70.3	65.7	60.1	+3.0	+3.8	+3.6	+19.1

statistic is $\max_r p_{hr} - \min_r p_{hr}$. Seed noise is the mean absolute difference between estimates from seeds 8001–8010 and 8011–8020. Percentile intervals use 2,000 deterministic bootstrap samples of the 64 histories. Every model’s format-spread interval excludes zero. The separate greedy pass asks all 1,152 model–history–format cells once. Both passes cover the same 64 histories.

The metamorphic pass contains 384 greedy decisions per model. It crosses 64 histories with the base prompt and five game-isomorphic transformations. Payoff scale and shift are positive-affine. Neutral-label variants map `OPTION_A/OPTION_B` to actions in both orders. Invariance is exact normalized-action agreement across all members of a transformation family. The reversed neutral-label variant is the least base-consistent transformation for seven of nine models, with 44.8% mean

agreement across models. It reverses Δ_{opp} for Qwen2.5-14B (+43.8 $\rightarrow$ -9.4 points) and Mistral (+18.8 $\rightarrow$ -18.8).

The deterministic sequential slice contains six 10-round trajectories per model. The primary sequential experiment adds 20 sampled pairs per format and model, for 7,200 decisions. The opponent scripts differ only in round three, after the focal action is chosen. Identical prefixes reuse one sampled focal action. Percentile intervals use 2,000 resamples of the 20 seed clusters. Table 7 reports the aggregate results. Figures 4 and 5 break down the policy, format, invariance, and propagation diagnostics.

C.5 Paired effects with intervals

For each paired effect, 2,000 resamples of the 32 history pairs summarize context heterogeneity in

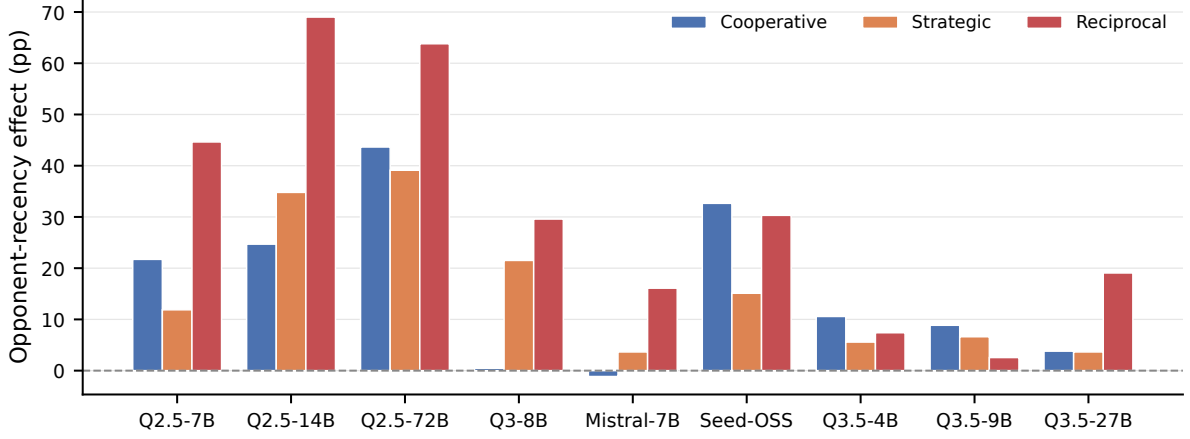


Figure 3: Opponent-recency effect under each released persona across the nine models, in percentage points.

Table 7: Robustness results. Format p span and Seed noise are percentage points. Greedy, Payoff, and Label are percentages. Shock area is the mean shock-minus-control sum over focal rounds 4–10.

Model	Format p span	Seed noise	Greedy	Payoff	Label	Shock area
Qwen2.5-7B	42.9	4.3	54.7	70.3	46.9	-3.90
Qwen2.5-14B	50.2	3.5	46.9	84.4	45.3	-0.65
Qwen2.5-72B	38.5	1.9	64.1	90.6	48.4	-4.77
Qwen3-8B	19.9	3.6	81.2	81.2	7.8	-3.50
Mistral-7B	56.2	4.8	43.8	87.5	53.1	-4.92
Seed-OSS-36B	8.8	1.1	90.6	96.9	42.2	-6.43

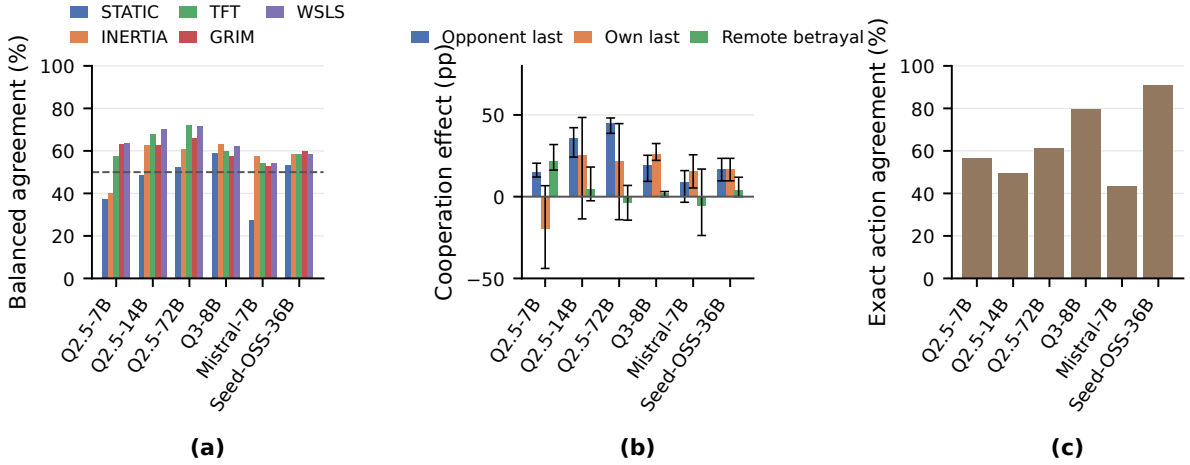


Figure 4: Detailed intervention panels. (a) Balanced candidate agreement. (b) Paired cooperation effects. Error bars span the three representations. (c) Exact sampled action agreement across formats.

the exhaustive cube. Separately, 2,000 cluster re-samples of the 20 decoding seeds quantify Monte Carlo uncertainty while retaining all histories and formats. Figure 6 shows the latter.

C.6 Same-cohort sequential shock

We apply the identical 20-seed paired shock protocol to the three released checkpoints. In contrast to the six-model cohort, all three cumulative shock-

minus-control intervals include zero (Fig. 7).

C.7 Temperature sensitivity

Table 8 compares five-seed CHC subsets at temperatures 0.1 and 1.0 with the declared 0.5 protocol. Constrained decoding is identical across temperatures.

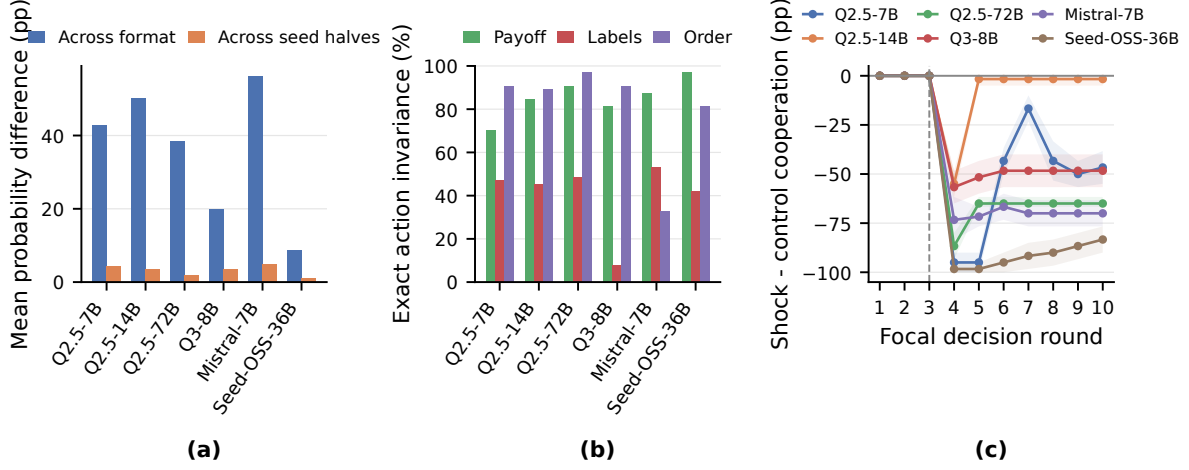


Figure 5: Robustness: (a) format and seed noise, (b) isomorph invariance, (c) shock propagation with 95% CIs.

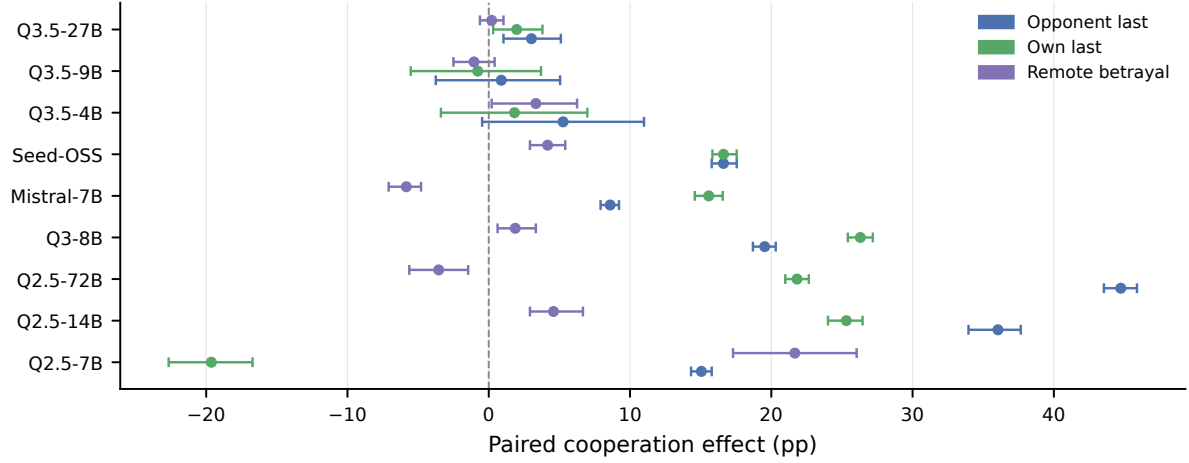


Figure 6: Paired CHC effects with 95% seed-cluster intervals, in percentage points. Released-cohort checkpoints occupy the top three rows.

Table 8: Cooperation rate (%) and opponent-recency effect (pp) across sampling temperatures.

Model	Cooperation rate			Δ_{opp}		
	0.5	0.1	1.0	0.5	0.1	1.0
Qwen2.5-7b	38.4	37.2	41.8	+14.4	+14.0	+15.2
Qwen2.5-14b	47.7	46.2	48.4	+37.9	+40.4	+35.6
Qwen2.5-72b	39.5	36.7	40.5	+44.8	+45.4	+44.8
Qwen3-8b	21.2	18.6	24.4	+20.0	+18.5	+18.8
Mistral-7B	42.3	42.5	44.7	+8.8	+10.0	+7.7
Seed-OSS-36B	8.6	8.1	9.3	+16.9	+16.2	+16.0

C.8 Same-cohort CHC

The three released-cohort checkpoints (Qwen3.5-4B/9B/27B) use the same protocol, prompts, seeds, and decoding constraints as the six-model CHC. Table 9 reports 11,520 decisions. The 27B checkpoint is the BF16 release, while the released simulations

served an fp8 build of the same checkpoint. The greedy metamorphic pass is stable under positive-affine payoff transformations, which preserve all actions (100%). Neutral-label permutations preserve only 29.7–46.9%.

D Identifiability of constant trajectories

Let $h_{1:t}$ be an observed joint-history prefix and π and π' deterministic candidate rules. If $\pi(h_{1:t}) = \pi'(h_{1:t})$ for every scored prefix, their action-agreement scores on those prefixes are identical. They may still disagree on unobserved histories. A constant focal path makes repeat-first and inertia indistinguishable. An all-CC path also admits TFT and grim. Constant paths can reject incompatible rules: WSLS, for example, predicts a switch after DD. Constancy does not make every policy equivalent.

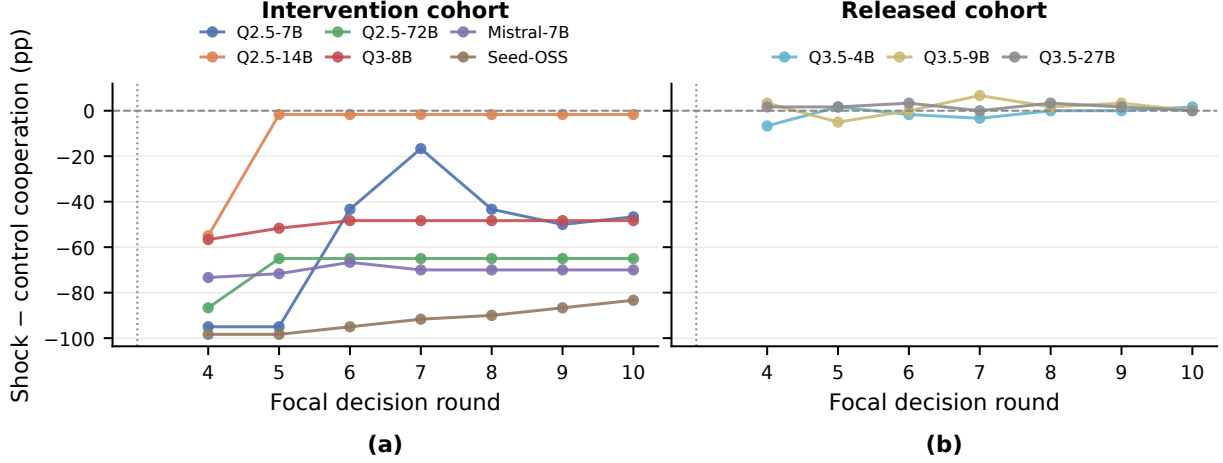


Figure 7: Shock-minus-control cooperation by round for every model. (a) Intervention cohort. (b) Released cohort.

Table 9: Twenty-seed CHC on the released-cohort Qwen3.5 checkpoints. Columns as in Table 5.

Model	Coop.	Best	Bal.	Δ_{opp}	Δ_{own}	Δ_{remote}	Format
Qwen3.5-4B	28.4	TFT	52.6	+5.3	+1.8	+3.3	55.9
Qwen3.5-9B	51.7	STATIC	52.6	+0.9	-0.8	-1.0	18.4
Qwen3.5-27B	69.8	WSLS	52.1	+3.0	+2.0	+0.2	11.2

Identifying-event counts refer to specified candidate contrasts. A zero count means that recorded decisions never distinguish those candidates. A positive count provides an opportunity whose value depends on the type, dependence, and statistical strength of the observed events.

The equivalence argument applies beyond the five candidate policies, but informative histories depend on the claim and its competing explanation. CHC supplies contrasts involving opponent recency, own recency, and remote defection.

In the released data, 3,276 of 3,465 runs keep both players’ actions constant and admit multiple candidate explanations. Only 174 runs contain an opponent action change, and 140 contain a trial distinguishing TFT from inertia. These counts establish missing coverage for the specified reciprocity contrast, without implying that every behavioral distinction is unidentifiable.

E Window, mixture, and ceiling controls

Memory window. A two-round variant of the CHC (16 exhaustive joint histories, three representations, 20 seeds) leaves the opponent-recency conclusions unchanged. Effects remain positive: Qwen2.5-14B changes from +36.0 to +37.3, Qwen2.5-72B from +44.7 to +47.7, and Mistral-7B from +8.6 to +5.2. The conclusions are robust to this window change.

Mixture mechanisms. Allowing convex mixtures of candidate policies ($\alpha \cdot \text{inertia} + (1 - \alpha) \cdot \text{TFT}$ and related pairs, $\alpha \in [0, 1]$) does not improve balanced agreement over the best single candidate for any of the nine models: the fitted α is always 0 or 1. This comparison covers only fixed-weight mixtures and does not rule out context-dependent policies.

Ceiling effects. The released cohort’s absent shock propagation is not a ceiling artifact. Its baselines are not always-cooperate (Qwen3.5-4B 28.4%, 9B 51.7%, 27B 69.8%). Within the same baseline interval, the intervention cohort propagates strongly. Shock areas are -4.92 for Mistral-7B, -3.90 for Qwen2.5-7B, and -4.77 for Qwen2.5-72B. Qwen3.5-4B has a similar 28.4% baseline but a shock area of only -0.08 . The cohort difference in dynamic sensitivity is therefore not attributable to baseline cooperation.

Label associations. The abstract GREEN/BLUE rendering may carry unintended color–action associations (“green = go”). We test this possibility with neutral-label metamorphic variants that use OPTION_A/OPTION_B in both orders. Label invariance remains low at 7.8–53.1% across models, so the observed format sensitivity is not specific to color semantics. Order sensitivity remains a separate concern.

Persona and mechanism alignment. Under the reciprocal persona, balanced tit-for-tat agreement rises for every model by 3.7 to 16.5 points. Qwen2.5-14B rises from 68.0% to 84.5%, Qwen2.5-7B from 57.5% to 72.3%, and Seed-OSS from 58.3% to 65.2%. This pattern is consistent with the mirror instruction shifting behavior toward tit-for-tat. It is directional evidence of approximation, not proof of rule-following.

F Boundary tests: what transfers?

Our claim is not that a checkpoint has an immutable responsiveness rank. It is that interaction claims require an evidence chain whose links must be re-measured when the available history, game, or interface changes. Three extensions test the tempting shortcuts that low response is caused only by a short window, that the diagnostic is confined to two-player IPD, or that CHC simply omits the released scaffold.

Memory windows. Four- and five-round exhaustive CHC variants (256 and 1,024 joint histories, 76,800 decisions per model) do not uncover a hidden large recency effect in the released 4B/9B checkpoints. Both intervals include zero at both lengths. Qwen3.5-9B remains the smallest responder (+0.8/+1.4). The 9B/27B checkpoints also retain the largest format spans (73.9–83.3%). Mistral enters the low-effect group at five rounds, so the substantive pattern survives additional history even though the exact rank does not (Table 10).

Three-player public goods. A two-round, three-player public-goods cube enumerates all 64 joint histories (3,840 decisions per model). The controlled contrast transfers: Qwen3.5-4B (+1.6 [−0.6, +4.0]) and 9B (−0.2 [−2.3, +1.8]) remain indistinguishable from zero, and 9B has a 90.8% format span. The model hierarchy does not transfer: Mistral is lower (−1.8) and Qwen3.5-27B is positive (+4.1). Thus CHC’s diagnostic logic generalizes beyond dyads, while responsiveness should not be narrated as a game-invariant model trait (Table 11).

Interface scaffold. Five conditions comprise the base CHC and four wrappers using the exact released Goal/Style text, Memory framing, neutral persona wrapper, and their combination (20,160 decisions per model). Stability is checkpoint-specific: Qwen3.5-9B remains weak (+1.5 to +6.1) and 27B is tightly bounded (+2.8 to +4.5), whereas

4B moves from +4.6 to +11.1. The overall ordering changes under the system and full wrappers. The ablation therefore does not “control away” the interface. It localizes which checkpoint claims survive it, exactly the role of the interface-robustness diagnostic (Table 12).

Together, these boundary tests sharpen the paper’s claim. The evaluation logic survives longer support and a multiplayer game, while individual effects and ranks can depend on the game and scaffold. A single outcome score would hide that distinction. The evidence chain makes it reportable.

G Controlled evaluator validation

The complete noiseless candidate confusion matrix contains one expected alias. Repeat-first and inertia both generate a constant focal action without an external perturbation. TFT, grim, and WSLS are uniquely recovered. At execution-error rates of 0, 5, 10, and 20%, the validation uses 500 deterministic-seed replicates for every policy–noise cell. Even with 20% execution error, the planted policy remains top-scoring in at least 99.6% of replicates per cell. Equal-cooperation but dynamically distinct trajectories receive distinct CRPs, and 100 log perturbations produce zero metric mismatches. These checks establish sensitivity within the tested candidate class; identifying-event counts establish the coverage missing from the released logs.

H Reproducibility

The released-log CRP audit operates on the independently released Social Sim’26 trajectories. The supplementary artifact contains the CHC, cohort, and persona prompt generators, constrained model runners, Ray dispatcher, deterministic aggregate scripts, intervention raw decisions, CSV tables, figures, and tests. The code declares every intervention variable, including history space, representations, temperature, seeds, transformations, opponent scripts, payoff matrix, and policy definitions. None is inferred from result labels.

Five models use vLLM 0.10.0 and Transformers 4.55.4. Seed-OSS uses the Transformers revision pinned by its model card because the installed vLLM does not yet register the architecture. The Qwen3.5 checkpoints use the Transformers release required by their model cards because neither installed vLLM nor the pinned revision registers the architecture. All fallbacks use the same chat messages and prefix-constrained two-choice decoding.

Table 10: Four- and five-round CHC windows: opponent-recency effects and format spans in percentage points.

Model	Δ_{opp} w4	95% CI	Δ_{opp} w5	95% CI	Fmt span w4/w5
Qwen2.5-7B	+17.2	[+16.3, +18.1]	+10.5	[+10.0, +11.0]	44.9/47.5
Qwen2.5-14B	+33.7	[+33.0, +34.3]	+34.9	[+34.4, +35.4]	42.0/34.7
Qwen2.5-72B	+33.0	[+31.4, +34.6]	+32.5	[+31.2, +33.8]	45.9/47.6
Qwen3-8B	+17.0	[+15.8, +18.3]	+14.4	[+13.9, +14.8]	37.9/37.7
Mistral-7B	+6.0	[+5.1, +6.8]	+3.1	[+2.5, +3.8]	44.0/59.9
Seed-OSS-36B	+20.4	[+19.7, +21.1]	+16.9	[+16.4, +17.4]	9.7/8.7
Qwen3.5-4B	+4.1	[-1.4, +9.5]	+5.6	[-0.2, +11.6]	29.3/36.1
Qwen3.5-9B	+0.8	[-3.8, +5.1]	+1.4	[-3.4, +5.6]	80.1/74.9
Qwen3.5-27B	+3.4	[+1.4, +5.3]	+5.3	[+3.2, +7.4]	83.3/73.9

Table 11: Three-player public-goods cube. Contribution and format agreement are percentages. Coplayer effects and format spans are percentage points.

Model	Contrib.	Coplayer Δ	95% CI	Fmt agree	Fmt span
Qwen2.5-7B	70.4	+9.6	[+8.4, +10.9]	51.2	48.1
Qwen2.5-14B	50.2	+16.8	[+16.0, +17.9]	70.0	29.5
Qwen2.5-72B	28.9	+27.3	[+26.5, +28.2]	79.1	20.2
Qwen3-8B	73.9	+6.4	[+6.1, +6.6]	80.5	19.5
Mistral-7B	91.1	-1.8	[-2.2, -1.5]	79.8	20.1
Seed-OSS-36B	39.6	+17.9	[+16.5, +19.2]	36.4	63.6
Qwen3.5-4B	13.5	+1.6	[-0.6, +4.0]	63.4	31.9
Qwen3.5-9B	53.0	-0.2	[-2.3, +1.8]	9.0	90.8
Qwen3.5-27B	75.5	+4.1	[+2.6, +5.5]	27.6	72.4

Table 12: Scaffold ablation: opponent-recency effects under the CHC base prompt and four released-scaffold wrappers, in percentage points.

Model	Base	+Sys	+Mem	+Pers	+Full
Qwen2.5-7B	+14.4	+5.8	+22.6	+19.3	+9.2
Qwen2.5-14B	+36.3	+39.6	+24.4	+38.8	+39.0
Qwen2.5-72B	+44.8	+37.2	+42.4	+43.3	+39.6
Qwen3-8B	+20.0	+17.0	+13.1	+14.7	+19.2
Mistral-7B	+8.2	+5.0	+13.7	+8.6	+0.8
Seed-OSS-36B	+16.4	+17.1	+14.4	+19.3	+15.7
Qwen3.5-4B	+4.6	+9.9	+7.9	+7.7	+11.1
Qwen3.5-9B	+2.2	+1.5	+4.2	+4.2	+6.1
Qwen3.5-27B	+3.3	+2.8	+4.5	+2.9	+3.7

Thinking is disabled for Qwen3 and set to budget zero for Seed.

The stored non-boundary data contain 34,056 core, 16,848 same-cohort, 108,864 persona, 11,520 temperature, and 5,760 two-round-window decisions (177,048 total). The boundary archive contains 691,200 long-history, 34,560 public-goods, and 181,440 scaffold decisions (907,200 total). Strict checks or SHA-256 manifests cover each file. Code, data, tests, and exact tables are included.