VERIFIABLE STRATEGIC REASONING IN ARTIFICIAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Strategic AI agents increasingly act in environments where decisions depend on anticipated responses of other agents, yet evaluation typically focuses on actions and outcomes rather than on whether the stated reasons for those actions can be independently examined. We introduce an Executable Strategic Rationale (ESR) representation that turns observable evidence, opponent-response predictions, counterfactual preferences, strategic intent, contingent plans, and the selected action into an external object that can be executed and audited without accessing private chains of thought. Candidate ESRs are evaluated from matched environment snapshots, allowing their claims and behavioral consequences to be checked under a common verification contract. In repeated auctions, outcome-oriented optimization improves utility most strongly, whereas verifier-guided optimization preserves the executable rationale interface and rationale-level properties do not improve uniformly with utility. Controlled perturbations of public strategic evidence systematically change regenerated rationales and downstream behavior, providing evidence of behavioral connectivity. The interface remains operational under out-of-distribution and external-auction transfer, although utility superiority does not transfer. These results separate behavioral performance from rationale auditability and provide a practical framework for externally testing strategic AI decisions.

1 INTRODUCTION

Artificial agents increasingly participate in auctions, negotiation, resource allocation, and other strategic environments where the value of an action depends on how other agents respond. Large language models (LLMs) are attractive in these settings because they can integrate heterogeneous interaction histories, condition decisions on natural-language context, and express strategic decisions in a form that humans can inspect. However, most evaluations ultimately reduce performance to action accuracy, reward, or regret. These quantities reveal whether a policy performs well, but not whether the rationale accompanying its action is meaningful, executable, or connected to the behavior that follows.

Free-form explanations are insufficient for this purpose. A policy may obtain high reward while producing an incomplete or post-hoc rationale, and a linguistically plausible explanation need not be compatible with the environment that actually executes the action. This distinction becomes especially important in strategic interaction: the quality of a decision can depend on an anticipated response, a counterfactual comparison, and a contingency that may only become relevant after the opponent reacts. A rationale that cannot be checked against those subsequent events provides little evidence beyond verbal plausibility.

We therefore study an *externalized strategic rationale*: an object whose claims are constrained by decision-time information and whose prescribed contingencies can be replayed and checked. Our goal is deliberately narrower than recovering private chain-of-thought or identifying a unique latent cognitive process. Instead, we ask whether an agent can be required to submit a strategic object whose evidence, predictions, preferences, and contingent actions admit independent environment-level tests.

We instantiate this idea with an *Executable Strategic Rationale* (ESR) for repeated auctions. An ESR links public evidence, action-conditional opponent-response predictions, a counterfactual pref-

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.

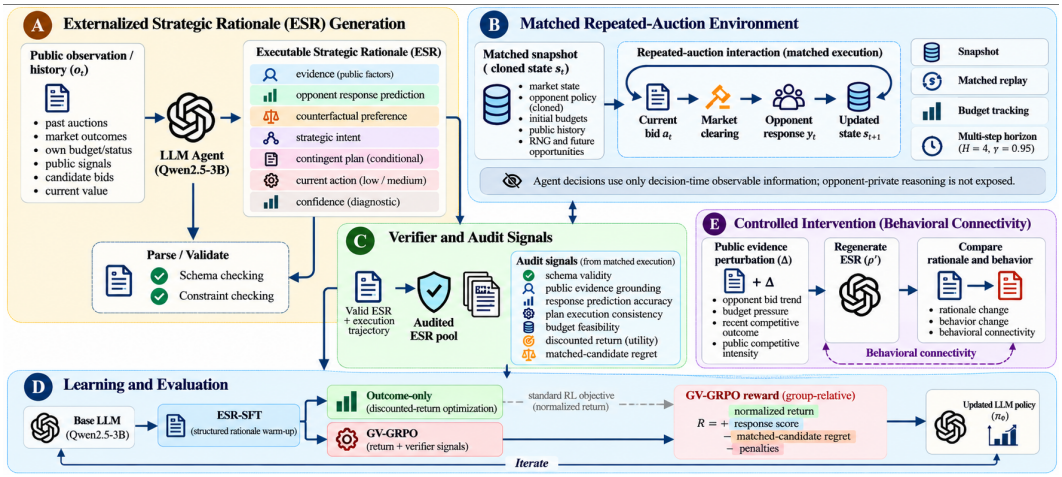


Figure 1: Overview of the proposed framework. (A) The LLM externalizes an Executable Strategic Rationale (ESR) from decision-time observable information. (B) Valid ESRs are executed from matched auction snapshots. (C) The verifier audits schema validity, public grounding, response prediction, execution feasibility, discounted return, and matched-candidate regret. (D) ESR-SFT initializes independent Outcome-only and GV-GRPO branches; only GV-GRPO uses verifier-guided reward. (E) Controlled public-evidence perturbations regenerate ESRs and test behavioral connectivity.

erence, strategic intent, a contingent plan, and a current action. Candidate ESRs are executed from matched simulator snapshots and checked for public grounding, structural validity, realized responses, execution consistency, budget feasibility, return, and regret. This converts a rationale from text that can merely be read into an interface that can also be executed and audited.

Our experiments reveal a central distinction: stronger behavioral performance does not guarantee uniformly better rationale-level properties. Outcome-only optimization achieves the strongest utility, while verifier-guided optimization preserves an executable rationale interface and exposes failures that reward-only evaluation would hide. Controlled perturbations of public strategic evidence additionally produce systematic changes in regenerated ESRs and downstream trajectories, which we term *behavioral connectivity*. Out-of-distribution and AuctionNet transfer further show that preservation of the verification interface and transfer of utility are separate questions.

Our contributions are threefold. First, we formulate strategic rationales as executable external objects rather than unconstrained natural-language explanations. Second, we develop a matched-execution verification protocol that checks rationale claims and contingent behavior without requiring access to private reasoning traces. Third, we empirically separate strategic utility from rationale auditability through outcome-only, verifier-guided, controlled-intervention, and transfer evaluations.

2 Related Work

Strategic LLM agents. Existing benchmarks study social interaction, negotiation, auctions, and game-theoretic reasoning in LLM agents (Zhou et al., 2024; Bianchi et al., 2024; Duan et al., 2024; Chen et al., 2023). Repeated-game studies further show that strategic behavior depends on game structure and can benefit from explicit opponent prediction (Akata et al., 2025). These works establish that LLMs can exhibit non-trivial strategic behavior, but they primarily evaluate actions, rewards, coordination, or behavioral patterns. Our focus is complementary: we make the strategic rationale itself an executable evaluation object and test its relationship to downstream behavior.

Faithfulness and externalized reasoning. Natural-language reasoning traces need not faithfully explain model decisions (Turpin et al., 2023; Lanham et al., 2023), motivating approaches that make reasoning externally executable or verifiable (Lyu et al., 2023; Lightman et al., 2024). Our claim is deliberately narrower than private-reasoning faithfulness. ESR verification checks an external

strategic object against an interactive environment, rather than asserting access to the model's true internal reasoning.

Outcome and process verification. Process supervision distinguishes evaluation of intermediate structure from evaluation of final outcomes. Our setting applies the same high-level motivation to strategic interaction, where intermediate claims can be checked against public market events, contingent execution, and realized trajectories rather than only against a static textual target. This allows Outcome-only and GV-GRPO to share the same environment while differing in whether executable rationale properties enter the learning signal.

3 Problem Setup

We study a repeated auction between a focal agent i and one configured opponent j. The state distribution contains both first-price and second-price mechanisms, budget constraints, and history-conditioned opponent responses. At round t, the focal agent observes the mechanism, current value, its own budget, candidate bids, and public interaction history. It does not observe the opponent's internal reactive state, opponent budget, future opportunities, or replay RNG state. The executable action space is

$$\mathcal{A} = \{\texttt{low}, \texttt{medium}\},$$

where each label maps to a state-dependent legal numerical bid.

If the focal agent wins, its payment follows the active auction rule and its per-round utility is $r_t = v_t - p_t$; otherwise $r_t = 0$. Formal evaluation starts from a frozen intermediate snapshot and rolls forward for at most $H = 4$ steps with $\gamma = 0.95$,

$$G_{t:H} = \sum_{k=0}^{H-1} \gamma^k r_{t+k}.$$

We report an executable two-action regret by comparing the realized action with the better of the two actions available from the same frozen state. This quantity is an environment-level comparator for the implemented action set; it is not a Nash-, Bayes-, or globally optimal-policy regret. Full payment, budget, information-boundary, and replay definitions are given in Appendix A.

The information boundary is part of the evaluation target. Decision-time prompts may contain only information available to the focal agent, whereas the simulator necessarily maintains additional hidden state so that matched replay can reproduce the same strategic situation. We therefore distinguish *agent-visible evidence* from *simulator state used for execution*. The verifier may use the latter for deterministic replay, but hidden opponent state is never exposed as evidence to the language model.

In addition to its action, the model emits an ESR

$$\rho_t = (e_t, \hat{y}_t, c_t, \iota_t, \pi_t, a_t, q_t),$$

containing stated public evidence e_t, action-conditional opponent-response predictions $\hat{y}_t$, a counterfactual preference c_t, strategic intent ι_t, contingent plan π_t, current action a_t, and reported confidence q_t. Only a_t and the realized-response branches of π_t directly control bids; the remaining fields are externally stated claims that can be grounded, checked for consistency, or scored against realized public behavior.

This separation yields two evaluation axes. *Behavioral performance* concerns realized return and executable regret. *Rationale auditability* concerns whether the submitted ESR is structurally valid, publicly grounded, executable, predictively consistent with realized responses, and behaviorally connected to the policy that follows. A central question of the paper is whether optimizing the first axis automatically improves the second.

4 Method

4.1 Executable rationale and matched verification

Given observation o_t, the model generates candidate text $y_k \sim \pi_\theta(\cdot \mid o_t)$, which is parsed and validated as $\rho_k = \mathrm{ParseValidate}(y_k)$. Parsing checks that required fields are present and values lie

in the legal schema; validation additionally rejects prohibited private-information references, invalid action–preference combinations, infeasible current bids, and malformed contingent plans. Public evidence is grounded against the observation and public history. Candidates that fail parsing or structural validation are not executed.

Each valid ESR is replayed from a clone of the same frozen snapshot, which fixes the starting public state, budgets, opponent implementation and internal state, future opportunities, and RNG stream. This matched construction keeps the strategic starting point and stochastic replay conditions common across candidates while isolating the policy induced by each ESR. The first rollout action is the ESR current action; subsequent actions are selected by the contingent plan from the previously realized opponent-response category,

$$a_t^{(k)} = a_k, \qquad a_{t+\ell}^{(k)} = \mathrm{Branch}(y_{t+\ell-1}^{j}, \pi_k),\ \ell \geq 1.$$

The branch dispatcher therefore reacts to realized opponent behavior rather than to the model's own predicted response.

The verifier checks parsing/schema validity, prohibited private-information references, public-evidence grounding, preference–action consistency, plan validity, budget feasibility, successful rollout, realized response prediction, return, and regret. Some checks are available before execution, whereas others require the matched trajectory. These are verification signals for the *externalized* object, not certificates of private reasoning faithfulness: the method evaluates what the agent commits to publicly, not whether the ESR is the unique internal reason for the action. Appendix B gives the full schema and verification contract.

4.2 Learning strategic rationales

We first supervised-fine-tune Qwen2.5-3B-Instruct (Qwen Team et al., 2024) on 3,000 verified ESR targets (500 validation examples), producing ESR-SFT. This stage teaches the canonical structured interface and provides a common initialization; it does not use game-level verifier reward.

From ESR-SFT we train two independent reinforcement-learning branches. *Outcome-only* optimizes group-normalized discounted return and executes only the current action throughout the four-step rollout. Its ESR fields are therefore not positively rewarded for evidence quality, response prediction, counterfactual preference, or contingent-plan execution. *GV-GRPO* instead executes the complete ESR and combines realized utility with verification-derived signals.

For valid candidate k, let z_k be the within-group standardized return, $s_k \in \{0, 1\}$ the first evaluable opponent-response score, and $\tilde{d}_k$ the max-normalized group-relative regret. The final reward is

$$R_k = z_k + 0.3s_k - 0.3\tilde{d}_k - P_k,$$

where P_k aggregates verifier penalties; invalid candidates receive only the negative penalty. The reward intentionally does not assign an independent positive term to every ESR field. Strategic intent and confidence, for example, remain diagnostic fields, while plan quality influences learning through execution validity and downstream consequences. Group-relative advantages are standardized final rewards and optimized with clipped GRPO (Shao et al., 2024).

The stable setting uses group size 4, horizon 4, $\gamma = 0.95$, clip range 0.2, KL coefficient 0.05, and 100 updates. Full penalty weights, return/regret normalization, the token-level objective, SFT hyperparameters, and checkpoint-selection details appear in Appendix C. The three training conditions therefore isolate increasingly strong forms of supervision: ESR-SFT learns the interface, Outcome-only emphasizes utility, and GV-GRPO couples utility to executable verification.

4.3 Controlled public-evidence interventions

Static validity alone cannot show that the externalized rationale is connected to behavior. We therefore perturb the public decision context rather than manually editing a generated ESR. Four interventions modify recent opponent bid trend, focal budget pressure, recent competitive outcomes, or public competitive intensity. For each frozen state, the model regenerates an ESR under the intervention and under an exact matched control, after which the resulting policy is replayed from the corresponding matched state.

For intervention I and control C, the primary estimand is

$$\Delta_{\mathrm{PRP}} = \Pr(\text{combined ESR-field change} \mid I) - \Pr(\text{combined ESR-field change} \mid C).$$

The combined event tracks registered changes in response prediction, counterfactual ranking, and contingent plan. We additionally measure directional agreement with the pre-specified perturbation direction and policy-associated trajectory divergence, which records changes in the current action or executed branch/trajectory. This design tests *behavioral connectivity*: whether controlled changes to public strategic evidence systematically propagate into regenerated rationales and executable behavior. It does not identify the causal effect of a private reasoning process or prove that one ESR field is individually necessary. Details are in Appendix D.

5 Experimental Protocol

All main policies use Qwen2.5-3B-Instruct. ESR-SFT is the supervised baseline and common initialization for independently trained Outcome-only and GV-GRPO branches. Main stable GV-GRPO and Outcome-only behavioral results use seeds 42, 43, and 44; ESR-SFT rationale metrics that originate from a single frozen evaluator are reported as such rather than being presented as three-seed aggregates.

The primary evaluation uses the same frozen two-bidder auction-state protocol described above. Behavioral metrics are discounted return and executable regret. Rationale-level metrics include strict ESR validity, successful plan execution, first evaluable response-prediction accuracy, counterfactual-preference accuracy, and applicable grounding/constraint checks. We keep these categories separate because successful execution does not imply that every semantic claim in the ESR is correct, and high return does not imply that the ESR remains auditable.

For intervention analysis, we use 64 unique states and 128 complete original/intervention/control triplets, totaling 768 generation/rollout records. This separate protocol uses horizon 4 and $\gamma = 0.5$, with 10,000 state-level paired bootstrap resamples for the primary PRP statistic. The control is constructed to preserve the original public information, while the intervention changes only the registered public factors and associated consistency fields.

Distributional transfer uses four OOD conditions with 64 states each, or 256 states per seed. External transfer uses AuctionNet (Su et al., 2024), with 48 complete episodes and 2,304 ticks. The OOD study focuses on whether the ESR interface remains structurally valid and executable under shifted conditions. Because execution coverage differs across methods, utility is interpreted conditionally rather than as an unconditional cross-method comparison. AuctionNet is likewise treated primarily as an external execution/verification transfer test rather than as evidence that GV-GRPO should dominate utility in a different simulator. Additional protocol details are given in Appendices D–E.

6 Results

6.1 Utility and rationale quality do not improve together

Both reinforcement-learning branches improve substantially over ESR-SFT in realized utility (Figure 2A). Outcome-only achieves the highest return (1.285 ± 0.182) and lowest regret (0.113 ± 0.167), whereas GV-GRPO occupies an intermediate behavioral point (0.984 ± 0.069 return; 0.599 ± 0.033 regret). This ordering is consistent with their objectives: Outcome-only directly optimizes realized discounted return, whereas GV-GRPO additionally allocates learning signal to response verification, matched-candidate regret, and constraint satisfaction.

Behavioral performance alone, however, does not determine the quality of the externalized rationale. GV-GRPO retains a complete executable interface, reaching 100% schema validity and plan execution and $70.8 \pm 0.7\%$ response-prediction accuracy (Figure 2B). At the same time, its counterfactual-preference accuracy is $45.7 \pm 1.3\%$, below the 69.0% frozen ESR-SFT evaluator. The important result is therefore not that one objective dominates every metric, but that utility and rationale-level properties can move differently under optimization.

This separation is useful diagnostically. A reward-only evaluation would highlight the strong Outcome-only return while revealing little about whether the policy continues to expose a struc-

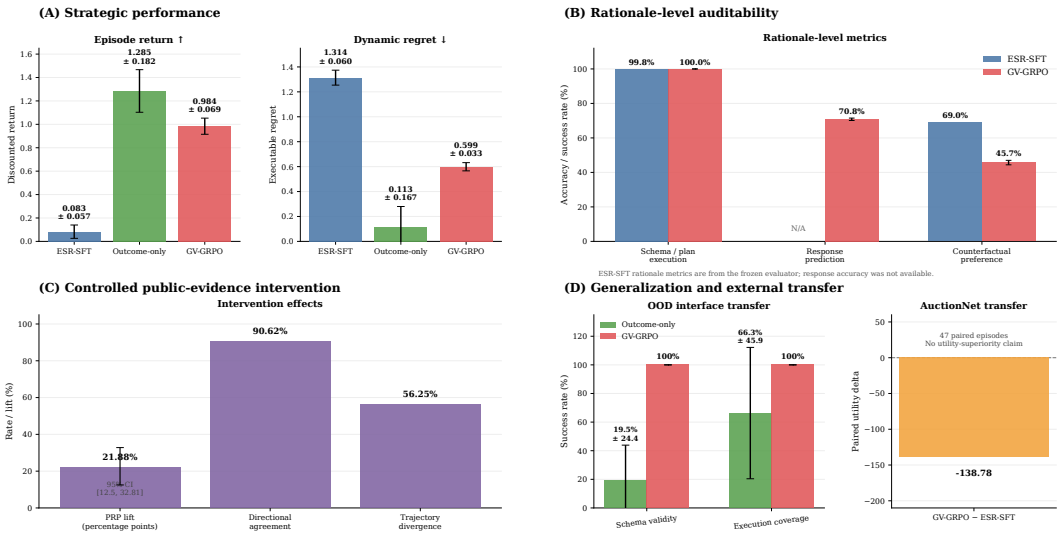


Figure 2: Main empirical results. (A) Behavioral return and executable regret. (B) Rationale-level auditability; ESR-SFT response-prediction accuracy was not available in the frozen evaluator. (C) Controlled-intervention effects; the PRP-lift error bar is the 95% state-level paired-bootstrap interval. (D) OOD interface transfer and the paired AuctionNet utility difference. Other error bars reproduce the dispersion reported by the corresponding frozen evaluator.

tured rationale. Conversely, perfect schema and plan execution in GV-GRPO should not be read as evidence that every semantic component improves. The ESR interface makes these otherwise hidden distinctions observable.

6.2 Controlled interventions reveal behavioral connectivity

Across 64 unique states and 128 complete matched triplets, public interventions increase the probability of a combined prediction–ranking–plan change by

$$\Delta_{\mathrm{PRP}} = 0.21875, \qquad 95\%\ \mathrm{CI} = [0.125, 0.328125].$$

Thus, modifying public strategic evidence produces an approximately 21.9-percentage-point increase in registered ESR-field changes beyond the exact matched-control baseline. Directional agreement is 90.6%, indicating that the observed rationale change usually follows the pre-specified direction of the perturbation.

Behavioral effects are also substantial. In 56.25% of complete triplets, the intervention changes either the current action or the subsequently executed contingent trajectory. Hence, controlled changes in public strategic evidence do not merely alter descriptive text; they are frequently accompanied by changes in executable policy. The result supports behavioral connectivity under controlled replay, while remaining weaker than causal identification of a unique private reasoning process.

6.3 The verification interface transfers beyond the training distribution

Under four OOD conditions, GV-GRPO retains $100.0 \pm 0.0\%$ strict schema validity and $100.0 \pm 0.0\%$ execution coverage across three seeds, compared with $19.5 \pm 24.4\%$ and $66.3 \pm 45.9\%$ for Outcome-only. These results indicate that outcome optimization alone does not guarantee preservation of the structured execution contract when strategic conditions shift.

We do not treat OOD utility as an unconditional head-to-head comparison because the effective executable-path denominators differ across methods. The transfer question is instead whether the ESR representation and execution pipeline remain operational under distribution shift. By this criterion, GV-GRPO preserves the interface more consistently.

The same distinction appears in zero-shot AuctionNet transfer. Both ESR-SFT and GV-GRPO retain the rationale/execution pipeline, but the paired GV-minus-ESR utility delta over 47 paired episodes

is -138.78. The external environment therefore provides no evidence of GV-GRPO utility superiority. This negative result sharpens the paper's claim: transfer of an auditable strategic interface is not equivalent to transfer of policy optimality.

7 Discussion and Limitations

Our results support a separation between *what an agent achieves* and *what strategic rationale it externalizes*. This distinction is particularly important in interactive environments, where an apparently successful action can coexist with a rationale that is incomplete, poorly calibrated, or inconsistent with future contingencies.

Utility is not auditability. Outcome-only is the strongest behavioral policy in the primary environment, while GV-GRPO explicitly trades return against response verification, regret, and invalidity constraints. The performance gap should therefore not be interpreted as evidence that verifier-guided training is a better utility optimizer. Its purpose is different: to retain a structured object whose claims and contingent actions remain externally inspectable.

The rationale metrics reinforce this interpretation. GV-GRPO preserves schema validity and plan execution, yet counterfactual-preference accuracy deteriorates relative to the supervised evaluator. Maintaining an executable rationale does not imply that every semantic component improves under reinforcement learning. This is a useful failure mode because a reward-only evaluation would not expose it.

Verification is not private-reasoning faithfulness. The verifier establishes properties of an externalized object: public grounding, executable contingencies, realized responses, budget feasibility, and downstream outcomes. It does not establish that the ESR is the model's unique or true internal reason for acting. The distinction matters because claims about private-reasoning faithfulness would require access to or identification of latent model mechanisms that are outside the scope of the current design.

At the same time, executable verification is stronger than linguistic plausibility alone. An ESR must survive schema checks, respect the decision-time information boundary, and remain compatible with an environment that can execute its prescribed actions. Controlled interventions add a behavioral test by showing that changes in public strategic conditions systematically alter regenerated rationales and, frequently, the executed policy.

Behavioral connectivity is controlled but not fully causal. The intervention study perturbs public evidence and regenerates the entire ESR. This design avoids manually editing one rationale field after generation, but it also means that several ESR fields may change together. We therefore interpret PRP lift, directional agreement, and trajectory divergence as evidence of controlled behavioral connectivity rather than as identification of the causal effect of a single field or of a private reasoning process.

Interface transfer is not utility transfer. The OOD and AuctionNet results provide a similar boundary. GV-GRPO preserves schema and execution coverage under self-built distribution shifts, while zero-shot AuctionNet does not produce a paired utility advantage. A verification framework should not be expected to guarantee payoff improvements under arbitrary changes in mechanism, value scale, or strategic dynamics. Its practical contribution is that external failures can be decomposed into failures of the rationale interface and failures of the underlying policy instead of being conflated.

Scope. The current study uses one focal LLM against configured reactive opponents in a two-bidder environment, a binary action space, a four-step formal horizon, and one base model family. The verifier can access full simulator state for deterministic replay even though the agent's evidence is restricted to decision-time public information. The response-prediction target is a coarse realized response category rather than a recovered latent opponent type or belief.

These choices make matched execution and auditing tractable but limit the generality of the conclusions. Stronger claims about verifiable strategic reasoning would require richer action spaces,

learned or jointly adapting opponents, multiple model families, longer horizons, and direct tests of rationale necessity and economic completeness. Further implementation and provenance limitations are listed in Appendix G.

8 CONCLUSION

We introduced ESR, an external interface for strategic rationales that can be executed, audited, and behaviorally tested without reading private reasoning traces. The framework couples a structured rationale representation with matched replay, environment-level verification, and verifier-guided policy optimization.

Repeated-auction experiments show that utility and rationale auditability do not improve monotonically together: Outcome-only is the strongest utility optimizer, while GV-GRPO preserves an executable rationale interface yet still exhibits imperfect counterfactual preference. Controlled public-evidence perturbations are systematically connected to regenerated rationales and downstream trajectories, and the verification interface remains operational under OOD and external-auction transfer even when utility superiority does not. These results motivate evaluating strategic AI not only by what it achieves, but also by whether the rationale it presents can be independently checked against the environment in which it acts.

AI USE STATEMENT

In this work, generative AI tools were used to assist with literature organization, manuscript drafting and editing, and code/result auditing. OpenAI Codex was used to inspect local source code, frozen experiment outputs, metric definitions, and claim provenance; ChatGPT was used to assist with restructuring the manuscript and editing technical prose. No experimental result was accepted solely from an AI-generated summary: reported claims and numerical results were checked against frozen artifacts, configuration files, evaluation scripts, and provenance records by the authors. The authors reviewed all AI-assisted text and analysis and take responsibility for the final content of the submission.

ETHICS STATEMENT

This work studies artificial agents in simulated auction environments and does not involve human subjects or private personal data. The framework is intended to improve the auditability of strategic agents rather than to facilitate deceptive market behavior. Deployment of autonomous bidding systems in real markets may raise concerns related to manipulation, fairness, and regulatory compliance; our experiments are limited to controlled simulation and transfer benchmarks, and we do not claim readiness for autonomous deployment in real markets.

REPRODUCIBILITY STATEMENT

The main text specifies the interaction model, ESR interface, matched-execution protocol, verifier-guided reward, intervention design, evaluation horizons, seeds, and headline metric definitions. Appendices provide the detailed environment, schema, reward, training, intervention, transfer, and provenance information required to interpret the frozen experimental artifacts.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3935–3951. PMLR, 2024.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of LLM agents in an auction arena. *arXiv preprint arXiv:2310.05746*, 2023.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 305–329, 2023. doi: 10.18653/v1/2023.ijcnlp-main.20.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

K. Su, Y. Huo, Z. Zhang, S. Dou, C. Yu, J. Xu, Z. Lu, and B. Zheng. Auctionnet: A novel benchmark for decision-making in large-scale games. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024.

A Full Auction Environment and Information Boundary

The formal environment contains one focal bidder and one configured opponent. A complete simulator state can be written conceptually as

$$s_t = (m_t, t, B_t^i, h_t, v_t, \xi_t^j, B_t^j, R_t),$$

where m_t is the auction mechanism, h_t is public interaction history, v_t is the current market value, ξ_t^j and B_t^j are opponent internal state and budget, and R_t denotes frozen stochastic quantities needed for replay. The focal prompt contains the mechanism, round and maximum round, current value, initial/remaining focal budget, public history, recent market summary, legal bid range, and numeric candidate bids. It excludes opponent type parameters, internal reactive state, opponent budget, future opportunities, RNG state, and any private chain-of-thought.

For focal and opponent bids b_t^i, b_t^j, allocation is

$$x_t = \mathbf{1}[b_t^i \geq b_t^j \wedge B_t^i > 0],$$

with ties resolved in favor of the focal bidder. Conditional on winning, pre-budget payment is $\tilde{p}_t = b_t^i$ in first-price auctions and $\tilde{p}_t = b_t^j$ in second-price auctions. Realized payment and budget dynamics are

$$p_t = x_t \min(\tilde{p}_t, B_t^i), \qquad r_t = x_t(v_t - p_t), \qquad B_{t+1}^i = B_t^i - p_t.$$

The simulator cap is 24 rounds, but stable formal evaluation uses $H = 4$ and $\gamma = 0.95$. Environment-level regret is

$$\mathrm{Regret}(s_t) = \max_{a \in \{\texttt{low},\texttt{medium}\}} G(s_t, a) - G(s_t, a_{\mathrm{executed}}),$$

clipped at zero. This is a two-action executable comparator, not Nash or Bayes regret.

Matched replay clones the same mechanism, history, focal/opponent budgets, serialized opponent state, future opportunities, and Python/NumPy RNG state. This ensures common starting conditions but does not by itself identify a natural causal effect of rationale text.

B Executable Strategic Rationale Schema and Verifier

The stable ESR schema contains the following fields. The response label set is $\{\texttt{raise}, \texttt{maintain}, \texttt{reduce}, \texttt{exhausted}, \texttt{switch_strategy}\}$, although canonical

Field	Value/type	Direct execution role
Evidence	non-empty list of strings	grounding only
Opponent response	if-low/if-medium response labels	scoring only
Counterfactual preference	low/medium/indifferent	consistency metric
Strategic intent	free-form non-empty string	none
Contingent plan	four low/medium branches	later actions
Current action	low/medium	first action
Confidence	$[0, 1]$	none

numeric-bid classification typically yields raise/reduce/maintain. The prediction does not alter the opponent; it is compared with the first evaluable realized response.

Parsing and validation reject missing/invalid fields, preference–action conflicts, prohibited opponent-private references, invalid plans, illegal actions, and budget-infeasible current bids. Public evidence must be grounded in the focal observation/history. Valid candidates are replayed through the contingent-plan executor, which maps realized responses to branches. The verifier may hold the full snapshot for replay, but opponent private state and private reasoning are not exposed as evidence to the focal model.

C Learning and Reward Details

ESR-SFT uses Qwen2.5-3B-Instruct with 3,000 training and 500 validation examples. The target is compact canonical ESR JSON and the objective is assistant-only causal-LM cross-entropy. The frozen configuration uses one epoch, learning rate 2×10^{-4}, per-device batch size 2, gradient accumulation 8, sequence length 1024, warmup ratio .03, and LoRA $r = 8$, $\alpha = 16$, dropout .05 on q/v projections.

Outcome-only and GV-GRPO start independently from ESR-SFT. Outcome-only executes the current action throughout the four-step horizon and standardizes valid candidate returns within each group; invalid candidates receive a fixed negative penalty. It has no positive reward for response prediction, evidence, counterfactual preference, intent, or plan quality.

For GV-GRPO, valid-candidate return is standardized as

$$z_k = \frac{G_k - \bar{G}}{\sigma_G + \epsilon},$$

and non-negative group-relative regret d_k is normalized by

$$\tilde{d}_k = \begin{cases} d_k / \max_j d_j, & \max_j d_j > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

The formal reward is

$$R_k = z_k + 0.3 s_k - 0.3 \tilde{d}_k - P_k,$$

with return weight 1, response weight .3, regret weight .3, and penalties for invalid schema (2), private-information leakage (2), ungrounded evidence (1), invalid plan (2), action–preference conflict (1), illegal action (2), budget infeasibility (2), and rollout failure (2). Invalid candidates receive $R_k = -P_k$.

Final group advantage is

$$A_k = \frac{R_k - \bar{R}}{\mathrm{std}_{\mathrm{pop}}(R) + 10^{-8}},$$

with zero advantages for constant groups. On assistant completion tokens, the clipped GRPO objective uses clip range .2 and KL coefficient .05. If $\Delta = \ell^{\mathrm{current}} - \ell^{\mathrm{ref}}$ and $r = \exp(\ell^{\mathrm{current}} - \ell^{\mathrm{old}})$, then

$$\mathcal{L} = \mathrm{mean}\left[-\min(rA, \mathrm{clip}(r, 0.8, 1.2)A) + 0.05(e^{-\Delta} + \Delta - 1) \right].$$

The stable formal setting uses group size 4, horizon 4, $\gamma = .95$, and 100 updates. Stable checkpoint selection is validation-based; older unstable checkpoints are excluded from main results.

D CONTROLLED INTERVENTION PROTOCOL

The formal intervention benchmark contains 64 unique states, 128 complete triplets, and 768 generation/rollout records. Each triplet contains the original state, a public-evidence intervention, and an exact matched control. Original and control prompts are byte-identical; the intervention prompt differs only through a registered public perturbation and required consistency fields. The model regenerates ESR outputs for each condition; ESR fields are not hand-edited.

The four intervention families are: (i) opponent bid trend, modifying the last three public competitor bids toward increasing or stable/decreasing patterns; (ii) focal budget pressure, modifying public budget/remaining-budget consistency toward high or low pressure; (iii) recent competitive outcome, modifying recent public outcomes/bids toward sustained loss or mixed win/loss; and (iv) public competitive intensity, modifying recent public competitor bids toward escalating or low-stable intensity.

The RQ3 manifest uses horizon 4 and $\gamma = .5$. PRP lift is the intervention-minus-control combined prediction/ranking/plan field-change rate on complete matched triplets. Uncertainty uses 10,000 state-level paired bootstrap resamples. Directional agreement measures whether the observed change direction matches the preregistered builder direction. Policy-associated trajectory divergence records current-action change or changed executed branch/trajectory; environment-only divergence is kept separate.

E OOD AND AUCTIONNET TRANSFER PROTOCOLS

OOD evaluation spans four held-out conditions with 64 states per condition and 256 states per seed for seeds 42, 43, and 44. Structural validity and execution coverage are directly comparable; utility is conditional on the relevant executable path and therefore may use method-dependent denominators.

The final zero-shot AuctionNet transfer contains 48 complete intention episodes and 2,304 auction ticks. Final ESR-SFT and GV-GRPO pipelines both retain complete generation/executable-structure coverage. Paired utility comparison is available for 47 episodes; the mean GV-minus-ESR difference is -138.7813, so the experiment is used as an external execution/verification transfer test rather than evidence of utility superiority.

F DETAILED PRIMARY RESULTS

Method	Return ↑	Regret ↓	Schema/Plan ↑	Resp. ↑	CF pref. ↑
ESR-SFT	0.083 ± 0.057	1.314 ± 0.060	99.8%	–	69.0%
Outcome-only	1.285 ± 0.182	0.113 ± 0.167	–	–	–
GV-GRPO	0.984 ± 0.069	0.599 ± 0.033	100.0%	$70.8 \pm 0.7\%$	$45.7 \pm 1.3\%$

Table 1: Detailed primary behavioral and rationale results corresponding to Figure 2. Outcome-only and GV-GRPO behavioral results are three-seed aggregates; ESR-SFT rationale metrics come from a frozen evaluator.

G ADDITIONAL LIMITATIONS AND PROVENANCE

The formal study is limited to one focal LLM and one configured opponent, a binary executable action space, reactive rather than unrestricted LLM opponents, a four-step formal horizon, and a single base model family. The environment has hidden opponent-side state but no independent hidden opponent valuation draw. Response prediction therefore denotes a coarse realized response category rather than latent-type or belief recovery. The verifier has access to full simulator state for deterministic replay, while evidence grounding is restricted to public decision-time information.

ESR-SFT rationale metrics originate from a frozen evaluator rather than the same three-seed run family as stable GV-GRPO. Outcome-only main behavioral values use the canonical three-seed table. Intervention headline values come from the fresh formal RQ3 protocol and are not pooled

with alternate checkpoints. OOD utilities are conditional on executable-path coverage. AuctionNet paired utility is computed on 47 paired episodes. Untraced historical percentages and superseded table values are excluded from main-paper claims.

Stronger verifiable strategic reasoning would additionally require evidence that stated reasons are necessary for the chosen action, economically complete relative to plausible alternatives, and stable across richer classes of agents, actions, horizons, and model families.