

TriEx: A Game-based Tri-View Framework for Explaining Internal Reasoning in Multi-Agent LLMs

Ziyi Wang^{1,†}, Chen Zhang^{1,†}, Wenjun Peng¹, Qi Wu¹, Xinyu Wang^{1,*}

¹Adelaide University, Australia

[†]These authors contributed equally

*Corresponding author

xinyu.wang02@adelaide.edu.au

Abstract

Explainability for Large Language Model (LLM) agents is especially challenging in interactive, partially observable settings, where decisions depend on evolving beliefs and other agents. We present **TriEx**, a tri-view explainability framework that instruments sequential decision making with aligned artifacts: (i) structured first-person self-reasoning bound to an action, (ii) explicit second-person belief states about opponents updated over time, and (iii) third-person oracle audits grounded in environment-derived reference signals. This design turns explanations from free-form narratives into evidence-anchored objects that can be compared and checked across time and perspectives. Using imperfect-information strategic games as a controlled testbed, we show that TriEx enables scalable analysis of explanation faithfulness, belief dynamics, and evaluator reliability, revealing systematic mismatches between what agents say, what they believe, and what they do. Our results highlight explainability as an interaction-dependent property and motivate multi-view, evidence-grounded evaluation for LLM agents. Code is available at <https://github.com/Einsam1819/TriEx>.

1 Introduction

Large Language Models (LLMs) increasingly act as autonomous agents capable of multi-step planning, strategic interaction, and real-time decision-making (Wang et al., 2024a; Li et al., 2024; Wang et al., 2024b). Yet as these models are deployed in settings that involve uncertainty, hidden information, and interactions with other agents, understanding *how* they internally construct strategies and make decisions becomes both crucial and challenging (Zhao et al., 2024). Existing explainability methods for NLP, such as attribution or post-hoc natural-language rationales (Sundararajan et al., 2017; Ribeiro et al., 2016; Wei et al., 2022; Turpin et al., 2023), have predominantly been developed

for static, single-turn prediction tasks. While such techniques can offer insight into which input features influenced a particular output, they fall short of capturing interaction-dependent belief dynamics and multi-agent reasoning, and thus provide limited evidence about how models represent other agents, update beliefs over time, or translate internal states into strategic actions (Chang et al., 2024).

Recent empirical observations highlight the severity of this gap. For example, AlphaArena (Alpha Arena Team, 2025) places multiple LLMs inside a live cryptocurrency and stock trading environment to study their behavior under realistic financial pressures. Despite strong performance on standard reasoning benchmarks, state-of-the-art models exhibited unstable strategies in this interactive setting. Crucially, their natural-language explanations often diverged from the decisions they executed, suggesting post-hoc rationalization and limited faithfulness. These findings underscore that in complex multi-agent contexts, existing explanation techniques often fail to reveal the internal mechanisms that drive behavior.

A central challenge these results point to is that we lack frameworks capable of exposing how LLMs model other agents, including their preferences, tendencies, and strategic profiles, and how such representations shape subsequent decisions. In human social interaction, individuals routinely construct and refine mental models of others over repeated encounters (Baldwin, 1992). These models support predictions about behavior, enable coordination, and guide adaptive strategies (Tomasello et al., 2005). Despite LLMs increasingly participating in multi-agent settings, it remains unclear whether they develop analogous structures, how such representations evolve through interaction, and how faithfully their natural-language explanations reflect these latent processes.

We introduce the **Tri-View Agent Explanation Framework (TriEx)**, a methodology inspired

by human social cognition for exposing complementary aspects of LLM internal decision-making. TriEx triangulates three trace-grounded perspectives: a *first-person* self-explanation tied to the executed action, a *second-person* opponent belief state updated through interaction, and a *third-person* oracle audit that checks cross-view consistency against objective reference signals and compares judgments across oracle models. By integrating these perspectives, TriEx provides a structured framework for examining how LLMs form and use internal representations during decision-making. Although poker serves as a testbed, the mechanisms TriEx targets, including belief updating, perspective integration, and cross-view consistency, are fundamental to many interactive, uncertain settings involving social reasoning. Strategic games thus serve as diagnostic environments rather than domain-specific applications. We view TriEx as an initial step toward systematic explainability for LLM agents in multi-agent settings. In summary, the contributions of this work are as follows:

- We introduce **TriEx**, a tri-view explainability framework for interactive LLM agents that logs decision traces with structured self-explanations, explicit counterpart belief states, and trace-grounded third-person audits.
- TriEx makes agent explainability *testable* by enabling cross-view checks of explanation faithfulness to actions, belief grounding and causal sensitivity under controlled interventions, and evaluator reliability across oracle models.
- Using TriEx, we find three robust patterns: faithfulness degrades with decision complexity, counterpart beliefs are intervention-sensitive and shape decision behavior, and evaluator reliability is highest for categorical and rank-based judgments but unstable for absolute scores.

2 Related Work

2.1 Explainability of LLMs

Explainability for LLMs spans methods from user-facing natural-language rationales to mechanism-level analyses of the computations underlying model behavior (Zhao et al., 2024). A common practice is to elicit self-explanations such as Chain of Thought (CoT) traces (Wei et al., 2022) and treat them as explanations of the decision process;

however, substantial evidence shows that CoT rationales can be unfaithful, offering plausible narratives weakly tied to the features or computations responsible for the output (Turpin et al., 2023; Madsen et al., 2024). Prior work therefore proposes intervention- or self-consistency-based faithfulness evaluations, but these often emphasize output-level consistency and can be confounded by task accuracy (Madsen et al., 2024; Parcalabescu and Frank, 2024; Jacovi and Goldberg, 2020; Bentham et al., 2024).

Beyond natural-language rationales, mechanistic interpretability explains LLM behavior through internal components and causal computations (Wang et al., 2023). Techniques such as activation patching and causal tracing can localize computations underlying factual recall and support targeted model editing (Meng et al., 2022). However, later analyses show that patching-based conclusions are sensitive to experimental design and may yield misleading artifacts, motivating stricter evaluation protocols (Zhang and Nanda, 2024; Makelev et al., 2024). Recent work further advances mechanism-level explanations by identifying sparse, interpretable feature circuits and analyzing intermediate computation dynamics in in-context learning and failure cases (Marks et al., 2025; Halawi et al., 2024). Despite these advances, much prior work focuses on static settings, leaving the evolution of explanations and internal beliefs under repeated interaction underexplored. We address this gap with TriEx, which instruments decision traces to jointly analyze actions, self-explanations, opponent beliefs, and third-person audits over time.

2.2 Games for LLM Evaluation

Games and simulated environments have emerged as a complementary paradigm for evaluating LLM agents (Wu et al., 2024; Wang et al., 2025; Hu et al., 2025; Duan et al., 2024; Peng et al., 2026). Most work, however, focuses on external performance or behavioral outcomes, offering limited insight into how models internally represent strategies, opponents, or uncertainty over time. By contrast, TriEx uses strategic games as controlled diagnostic environments where decision traces can be logged and audited against objective reference signals.

2.3 Social Cognition and Agent Modeling

TriEx is inspired by findings in social cognition showing that humans form and refine mental models of others through repeated interaction (Baldwin,

1992; Frith and Frith, 2006). Related work on Theory of Mind argues that agents attribute latent mental states to predict others’ actions, and in strategic settings this can involve recursive reasoning about others’ policies (Premack and Woodruff, 1978; Camerer et al., 2004; Yoshida et al., 2008). At the same time, cognitive psychology highlights limits of introspective verbal reports, showing that coherent self explanations may be weakly connected to the actual determinants of behavior (Nisbett and Wilson, 1977). Together, these findings motivate TriEx’s separation of first-person self explanations, second-person opponent beliefs, and third-person audits, enabling consistency analysis across stated reasoning, inferred beliefs, and observed actions.

3 TriEx Framework

We introduce **TriEx**, a **Tri-view Explainability** framework for probing decisions of LLM-based agents in interactive, partially observable environments. Rather than treating explanations as post-hoc narratives or focusing on behavioral outcomes alone, TriEx operationalizes explainability as *structured, falsifiable claims* about decision factors and internal beliefs, elicited and tested through interaction. In this paper, we instantiate TriEx in Texas Hold’em poker, a controlled imperfect-information setting that naturally induces uncertainty, belief updates, and opponent modeling.

At each decision point, TriEx records a structured decision trace that serves as the common substrate for all analyses. For an agent i at time t , the trace contains the agent’s observation o_t^i , the legal action set, the executed action a_t^i , a first-person self-reasoning artifact r_t^i , second-person opponent profile states $\{P_{j,t}^i\}_{j \neq i}$, and environment-derived reference features ϕ_t (e.g., pot odds, stack-to-pot ratios, and hand equity estimates). All three explanatory views operate on the same trace, so cross-view differences reflect explanatory perspective rather than differences in available information.

TriEx decomposes explainability into three complementary perspectives. The **first-person view** captures the agent’s own structured account of its decision, produced before action execution and treated as a testable claim about the underlying reasoning process. The **second-person view** exposes the agent’s internal models of other players, represented as explicit profile states over interpretable behavioral dimensions and updated incrementally from observed actions. The **third-person view** in-

troduces an oracle role that consumes the decision trace and produces a structured audit of consistency among actions, self-explanations, opponent profiles, and reference features.

3.1 First-Person View

The first-person view captures what an agent *claims* to be its reasoning at each decision point. In TriEx, this view is instantiated by eliciting a *self-reasoning artifact* before action execution and treating it as a testable hypothesis about the agent’s internal decision process. At time t , agent i receives its observation o_t^i , the legal action set, and its current opponent-profile states $\{P_{j,t}^i\}_{j \neq i}$, and produces a first-person artifact that is explicitly bound to the subsequent action. The environment executes the committed action and records both the artifact and the action in the decision trace.

To make first-person explanations auditable and comparable, TriEx constrains them to a structured interface. Formally, the first-person artifact is represented as:

$$r_t^i = \langle \text{Narrative}_t^i, \text{DecisionJSON}_t^i \rangle, \quad (1)$$

where Narrative_t^i is a brief free-text rationale and DecisionJSON_t^i is a machine-readable decision record. The decision record serves two purposes: it commits the agent to a concrete action, and it exposes a small set of declared internal decision variables that can be aligned with observed behavior and objective reference features ϕ_t , e.g., pot odds or equity estimates. The schema is intentionally minimal and abstract, specifying *what kinds of claims must be made*, not *how* they are computed. The first-person interface can be summarized as follows.

First-Person Interface

Input: agent observation o_t^i , legal actions, opponent profiles $\{P_{j,t}^i\}_{j \neq i}$, reference features ϕ_t

Output: self-reasoning artifact r_t^i

Required components:

- **Action commitment:** declared action
- **Decision-state claims:** discrete, interpretable tags
- **Evidence pointers:** links to ϕ_t or profiles

Upon generation, the structured decision record is validated and canonicalized into an *explanation signature* σ_t^i , which normalizes declared claims into a fixed ontology and records missing or ill-formed entries explicitly. Both the raw artifact r_t^i

and its signature σ_t^i are stored in the trace. This separation preserves the original narrative for qualitative inspection while enabling systematic comparison of first-person claims across time and agents.

The first-person view is not intended to assess optimality or gameplay performance. Instead, its role is to render self-reasoning *empirically checkable* by grounding declared decision variables in an interactive environment with independent reference features. By binding explanations to concrete actions and reference features, TriEx enables direct comparison between what an agent claims to reason about and what the game state and behavior support. This makes first-person reasoning a testable component of explanation, forming one leg of the triangulation with internal beliefs and third-person audits.

3.2 Second-Person View

The second-person view reveals how an agent models *other* agents through repeated interaction. In real-world strategic settings, humans rarely treat opponents as memoryless; instead, they build and update *behavioral models* from repeated encounters to guide future decisions (Camerer and Hua Ho, 1999). TriEx extends this idea to LLM agents by maintaining, at each interaction step, an explicit *opponent profile state* that summarizes beliefs about each opponent along interpretable behavioral dimensions and tracks how these beliefs evolve with accumulating evidence. For an agent i and an opponent j , TriEx represents the second-person belief state at time t as:

$$P_{j,t}^i = \langle \mathbf{p}_{j,t}^i, s_{j,t}^i, u_{j,t}^i \rangle, \quad (2)$$

where $\mathbf{p}_{j,t}^i$ is a K -dimensional vector over a fixed, interpretable trait ontology, $s_{j,t}^i$ is a concise qualitative summary of the opponent’s current style, and $u_{j,t}^i$ is a brief rationale describing the evidence motivating recent updates. Opponent profiles are updated incrementally using only information available during gameplay. At designated update points, the agent integrates prior beliefs with newly accumulated evidence to produce bounded, *incremental changes* rather than unconstrained rewrites. Each update is accompanied by $u_{j,t}^i$ linking belief changes to observed evidence, making belief dynamics transparent and temporally traceable for analysis of stability, drift, and sensitivity. The second-person interface is structured as:

Second-Person Interface

Input: prior opponent profile state, environment-derived behavior summaries, contextual interaction signals

Output: updated opponent profile state

Required components:

- **Latent belief representation:** a vector over interpretable behavioral dimensions
- **Qualitative summary:** a concise description of the opponent’s inferred tendencies
- **Update rationale:** a brief explanation linking observed evidence to belief change

Second-person profiles are treated as part of the agent’s internal decision state rather than as post-hoc annotations. At each decision point, the current set of opponent profiles is made available to the agent’s policy and recorded in the decision trace, allowing analyses to test whether belief changes coincide with shifts in decision-making behavior. All opponent profile states are validated and canonicalized before storage, supporting comparison across opponents, agents, and time. When combined with first-person self-reasoning artifacts, the second-person view enables examination of whether stated reliance on opponent models aligns with the beliefs actually maintained and updated during interaction.

3.3 Third-Person View

The third-person view audits cross-view consistency using an *oracle* instantiated by a language model and a deterministic rule-based checker grounded in environment-derived signals. This two-stream design reduces shared biases by anchoring evaluation in independent environmental reference features. Although first-person reasoning traces and second-person belief states expose rich internal signals, manually reviewing them across agents, time steps, and interaction histories is impractical. TriEx addresses this by formalizing *LLM-as-Judge* (Li et al., 2025) as a structured, trace-grounded auditing mechanism.

Unlike a single, privileged evaluator, the oracle in TriEx is a *role* rather than a fixed model. Across different evaluation runs, participating LLM agents may alternately assume the oracle role to analyze previously recorded decision traces. This design serves two purposes. First, instantiating the oracle with multiple heterogeneous LLMs enables cross-oracle comparison of auditing judgments, mitigating reliance on any single evaluator and helping distinguish shared structural signals from model-

or family-specific biases. Second, it treats evaluator capability itself as a subject of analysis, allowing the reliability and consistency of LLM-based judges to be systematically examined alongside the explanations they evaluate. Formally, given a trace slice $\tau_t^i = \langle o_t^i, \mathcal{A}_t^i, a_t^i, r_t^i, \{P_{j,t}^i\}_{j \neq i}, \phi_t \rangle$ at time t for a focal agent i , an oracle model Ω consumes τ_t^i and produces a structured audit report \mathcal{R}_t^Ω . The audit evaluates alignment among the agent’s first-person decision-state claims, its executed action, its maintained second-person opponent profiles, and a set of objective, environment-derived reference features ϕ_t . These reference features include quantitative signals provided by the game environment, such as pot odds, stack-to-pot ratios, hand equity estimates, and aggregated behavioral statistics. By anchoring judgments to such externally computed quantities, TriEx ensures that third-person evaluations are grounded in independently verifiable evidence rather than unconstrained subjective interpretation.

Oracle outputs are treated as *explanatory evidence*, where disagreements between oracle assessments, objective reference signals, and other oracle instances are explicitly preserved, enabling TriEx to analyze not only the faithfulness and coherence of agent explanations, but also the reliability and potential bias of LLM-based judges with respect to measurable properties of the interaction. The third-person view thus functions as a scalable, quantitatively anchored auditing layer that complements first-person self-reasoning and second-person belief tracking, completing the TriEx triangulation for explainability analysis in interactive, multi-agent settings. The structure of the third-person interface is summarized as follows:

Third-Person Interface

Input: trace slice τ_t^i containing o_t^i , legal actions \mathcal{A}_t^i , action a_t^i , first-person artifact r_t^i , opponent profiles $\{P_{j,t}^i\}_{j \neq i}$, and objective reference features ϕ_t

Output: structured audit report \mathcal{R}_t^Ω

Required components:

- **First-person faithfulness audit:** consistency between declared decision-state claims and (a_t^i, ϕ_t)
- **Second-person belief audit:** consistency between opponent-profile traits and environment-derived behavioral statistics
- **Cross-view inconsistency flags:** explicit tags for mismatches
- **Evidence grounding:** short trace-based evidence pointers or an explicit none indicator

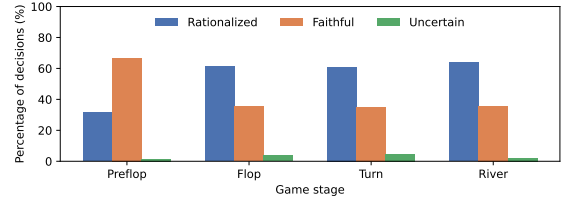


Figure 1: **Agent-averaged distribution of first-person explanation outcomes across game stages.** For each stage, explanation outcomes are normalized within-agent and then averaged across agents. Preflop explanations are predominantly faithful, while postflop stages exhibit a stable shift toward rationalized explanations.

4 Experiments

This section uses TriEx to study explainability in LLM-based decision-making agents under repeated strategic interaction. We run mixed-table Texas Hold’em games as a controlled testbed, in which LLM agents play against one another and a set of algorithmic reference players with fixed strategies, and log agent decision traces under the TriEx instrumentation protocol, including executed actions, first-person decision artifacts, second-person opponent profile states, and environment-derived reference signals. We evaluate a mix of open-weight and proprietary LLMs, full implementation details are provided in Appendix A.1. We organize the evaluation around three research questions.

RQ1. First-person View: Do agents’ self-reasoning faithfully reflect the decision process that produces their actions?

RQ2. Second-person View: Do agents maintain behaviorally grounded opponent beliefs that exhibit causal sensitivity to controlled interventions?

RQ3. Third-person View: Which types of judgments about explanations are robust across different LLM evaluators?

4.1 First-Person View Experiments

TriEx treats first-person self-reasoning as a set of *testable claims* bound to a concrete action and an externally verifiable game context. To evaluate faithfulness, each decision point is audited against environment-derived reference signals (*e.g.*, pot odds, equity estimates) and the executed action. Two complementary auditors are used: (i) a deterministic rule-based check that penalizes explicit contradictions and missing/underspecified decision-state claims, and (ii) a third-person oracle that reads the full decision tuple and outputs a structured overall faithfulness score and a rationalization flag. Full

auditing protocols are provided in Appendix A.3.1.

Each explanation is categorized into one of three outcomes. *Faithful* explanations exhibit consistent alignment between stated decision claims, the executed action, and objective reference signals. *Rationalized* explanations contain contradictions, implausible justifications, or omit decision-relevant beliefs despite non-trivial actions. *Uncertain* cases correspond to borderline decisions where neither faithfulness nor rationalization can be confidently established under the audit criteria.

Figure 1 summarizes explanation outcome distributions aggregated across all evaluated LLM agents and decision points, grouped by game stage and normalized within each stage. A clear stage-dependent structure emerges. Preflop decisions, which involve limited information and simpler state structure, are dominated by faithful self-reasoning. In contrast, explanations generated after the introduction of shared cards and longer betting histories display a markedly different and highly consistent distribution across Flop, Turn, and River, characterized by substantially higher rates of rationalization. Notably, the three postflop stages show similar explanation profiles, suggesting a stable shift to a different explanatory regime rather than progressive degradation.

Finding 1 – Complexity-sensitive degradation of reasoning faithfulness: As interaction states become more complex and information-rich, self-reasoning becomes systematically less faithful to the underlying decision process, with a sharp increase in post-hoc rationalization.

4.2 Second-Person View Experiments

The second-person view in TriEx examines how agents internally represent and update beliefs about other interacting agents, and whether such beliefs are meaningful for explainability. While instantiated in a poker setting, these beliefs capture a general notion of *counterpart modeling*: persistent internal representations of others’ behavioral tendencies that evolve through interaction and are explicitly available to the agent at decision time. In our mixed-table setting, each LLM agent maintains such second-person beliefs about all opponents at its table, enabling the study of multi-agent belief dynamics while anchoring evaluation in opponents with known, fixed strategies. We evaluate second-person beliefs along three questions: (i) whether belief states exhibit coherent structure un-

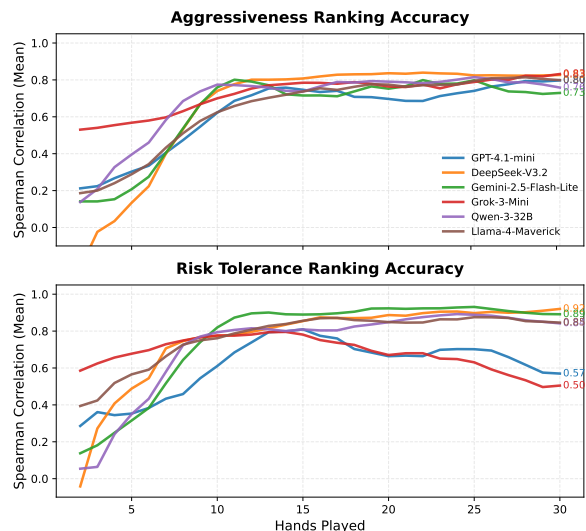


Figure 2: **Rank-based convergence of second-person belief states.** Mean Spearman rank correlation over rounds for two representative belief dimensions, *Aggressiveness* and *Risk Tolerance*, computed against fixed-strategy reference agents. Additional per-model and per-trait curves are provided in Appendix B.2.

der repeated interaction, (ii) whether belief dimensions are grounded in observable behavior, and (iii) whether these beliefs exhibit causal sensitivity to controlled interventions in driving decision behavior. Detailed settings and extended analyses are provided in Appendix A.3.4.

Figure 2 evaluates whether agents’ second-person beliefs stabilize under repeated interaction using a calibration-robust criterion. Rather than comparing absolute belief values, we measure *relative ordering* via Spearman rank correlation against fixed-strategy algorithmic agents. The game environment enables such evaluation by providing stable behavioral references with known strategic archetypes, serving as a form of interaction-level ground truth. We report two representative belief dimensions capturing distinct explanatory regimes. *Aggressiveness* is action-proximal and closely tied to observable betting behavior, whereas *Risk Tolerance* reflects a more abstract interpretation of exposure under uncertainty, integrating both proactive and passive actions. Across models, beliefs about aggressiveness converge rapidly and consistently, indicating that action-aligned traits are easy to infer. In contrast, risk tolerance exhibits cross-model divergence, suggesting that models adopt different internal semantics for what constitutes “risk” despite observing similar behavior.

Finding 2 – Divergent internal interpretations of other agents:

Models form stable internal beliefs about other agents from repeated interaction. Beliefs tied to surface actions converge across models, while higher-level interpretations of intent and risk diverge systematically, revealing model-specific explanatory semantics.

Beyond internal structure, we test whether second-person beliefs are behaviorally grounded. Across evaluated models, intent-level belief dimensions align strongly with objective behavioral statistics derived from gameplay, whereas outcome-dependent traits show weaker and noisier alignment. These results indicate that second-person beliefs capture meaningful behavioral regularities, but grounding strength depends on the abstraction level of the trait. Full trait-behavior alignment results are reported in Appendix B.2.

Finally, we assess the decision-level effect of second-person beliefs through controlled interventions. As shown in Table 1, perturbing a single belief dimension while holding the decision context fixed induces systematic, directionally consistent changes in action distributions. To distinguish that these shifts reflect intervention effects rather than stochastic rerun variability, interventions are implemented as numeric perturbations in the logit-space of explicit opponent-profile traits, leaving natural-language prompt content unchanged, and each condition is replicated across 50 runs per model and intervention direction. Comparing logged, rerun-original, and rerun-intervened decisions shows that intervention-induced shifts exhibit directional and semantic structure that cannot be explained by rerun variability alone. Full metric definitions, per-model variance summaries, and extended analyses are reported in Appendix B.2.

Finding 3 – Internal models of other agents exhibit causal sensitivity under controlled belief interventions:

Agents encode interpretable beliefs about other agents that are reflected in intervention-sensitive decision behavior. Directional manipulations of these beliefs induce systematic and semantically consistent shifts in policy, patterns that cannot be attributed to rerun variability alone.

4.3 Third-Person View Experiments

The third-person view in TriEx examines the reliability of *explanation evaluation* itself. Rather than relying on a single unconstrained LLM judge, we study what forms of explanation assessment re-

Trait	Dir.	Change Rate	Dir. Cons.
Aggressiveness	up/down	0.26	0.89
Risk Tolerance	up/down	0.27	0.88

Table 1: **Directional intervention summary.** Mean results aggregated across models. Dir. Cons. denotes consistency with the expected semantic direction.

Oracle Model	ρ
Grok-3-mini	0.3899
Qwen3-32B	0.3810
GPT-4.1-mini	0.3685
DeepSeek-v3.2	0.3667
LLaMA-4-maverick	0.3241
Gemini-2.5-flash-lite	0.2865

Table 2: **Third-person auditing of first-person explanations.** Spearman rank correlation (ρ) between oracle faithfulness and a rule-based reference.

main stable across evaluators, and where evaluation becomes model-dependent or unreliable. Specifically, TriEx adopts a two-stream auditing design that combines multiple LLM-based oracles with a deterministic rule-based checker. The rule-based checker evaluates decisions using explicit game-specific criteria (*e.g.*, pot odds optimality, action legality, stack-to-pot ratio constraints), while the oracle evaluates the agent’s first-person reasoning traces. This supports cross-model comparison and provides a deterministic reference grounded in environment-derived signals, rather than a ground-truth measure of explanation correctness.

We first evaluate third-person judgments of first-person explanation faithfulness. Across oracle models, faithfulness scores exhibit consistent positive correlation with an independent, rule-based reference grounded in explicit poker mechanics (see Table 2). Although these correlations are moderate, they indicate that LLM-based evaluators capture substantive decision structure rather than surface-level narrative plausibility.

The oracle evaluates full decision traces, including first-person reasoning, whereas the rule-based checker applies deterministic criteria to executed decisions only. Because they rely on different information sources and assess different aspects of agent behavior, moderate correlation is expected, reflecting shared signals of decision rationality rather than score-level agreement. Consistent with this view, Table 2 shows cross-oracle heterogeneity in per-oracle correlations, while Table 6 shows stronger alignment after aggregating oracle judg-

Oracle Model	ρ	Dir. Acc.
GPT-4.1-mini	0.7205	0.9994
LLaMA-4-maverick	0.6818	0.9480
Gemini-2.5-flash-lite	0.6287	0.9775
DeepSeek-v3.2	0.4152	0.9732
Grok-3-mini	0.2690	0.9922
Qwen3-32B	0.1187	0.9556

Table 3: **Third-person auditing of second-person beliefs.** Overall oracle agreement with objective behavior, measured by rank-based alignment (ρ) and directional accuracy.

ments across models, suggesting that oracle models differ in absolute score calibration but converge on a shared rank-ordering signal.

However, auditing reliability degrades in complex decision regimes, with substantially weaker alignment in postflop states and strong-hand scenarios, revealing limits of explanation evaluation under uncertainty (see Appendix B.3 for extended analysis). We therefore interpret third-person auditing as a diagnostic mechanism rather than an absolute correctness oracle, most trustworthy for categorical and rank-based judgments that can be anchored in explicit reference signals.

Finding 4 – Explanation auditing by LLMs is complexity-sensitive: LLM-based evaluators capture consistent signals of decision rationality across heterogeneous evaluators, though at moderate correlation strength, but their auditing reliability degrades systematically as decision contexts become more complex.

We next evaluate explanation evaluation at the level of internal belief states. Table 3 summarizes third-person auditing of second-person beliefs, aggregating two representative traits, *i.e.*, Aggressiveness and Risk Tolerance. A clear reliability hierarchy emerges: across oracle models, *directional judgments* of belief correctness (*e.g.*, overestimate vs. underestimate) are nearly perfectly consistent, while *continuous alignment scores* exhibit substantial variation. This pattern holds across traits and temporal windows (see Appendix B.3 for extended analysis). These results indicate that LLM-based evaluators robustly capture coarse relational structure in internal belief states, while fine-grained calibration of misalignment severity is notably less stable. Notably, this instability is concentrated in scalar scoring, whereas categorical and directional judgments remain highly reliable across models. We therefore treat third-person audits as most trust-

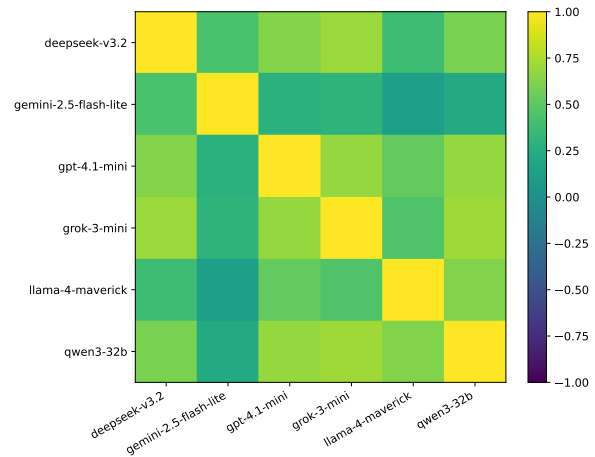


Figure 3: **Cross-oracle consistency in explanation auditing.** Pairwise agreement measured by quadratic Cohen’s κ showing substantial consistency in relative faithfulness judgments across oracle models.

worthy for categorical and rank-based judgments anchored in explicit reference signals, and avoid over-interpreting them as absolute scalar measures of explanation quality.

Finding 5 – Relative judgments are robust across evaluators: Directional and rank-based explanation judgments are consistent across models, while absolute scalar scores remain calibration-dependent.

Finally, we examine cross-model consistency among LLM-based evaluators. Figure 3 shows substantial pairwise agreement on ordinal faithfulness judgments, with most oracle pairs achieving moderate to high Cohen’s κ . Disagreements primarily reflect stable differences in score calibration, while preserving the relative ordering of explanations across models.

Finding 6 – A hierarchy of reliability in LLM-based explanation evaluation: When LLMs are used as evaluators, categorical judgments are the most robust across models, followed by rank-based comparisons, while absolute score calibration is the least stable.

4.4 Discussion

TriEx reframes explainability for LLM-based agents as an interaction-dependent property that cannot be captured by isolated reasoning traces or single evaluation signals. By jointly analyzing self-explanations, internal beliefs about others, and third-person audits, our results highlight that explanation quality, internal representations, and evaluation reliability are tightly coupled. Several

implications follow for explainability research in NLP. First, the systematic breakdown of explanation faithfulness under increasing decision complexity suggests that explanation failures are themselves informative. Rather than treating such cases as noise, future work can use them to probe representational limits and shifts in reasoning behavior under uncertainty. Second, the existence of persistent, intervention-sensitive second-person belief states reflected in decision behavior indicates that explainability in interactive systems must go beyond introspection, accounting for how agents model and respond to others over time. The divergence observed in higher-level belief dimensions such as risk tolerance further suggests that current LLMs lack stable semantic grounding for abstract behavioral concepts. This points to augmentation strategies such as ontology-based belief grounding, structured memory, or belief-aware training, which TriEx could inform as a diagnostic tool.

Finally, our findings show that explanation evaluation is reliable only at certain levels: relative and categorical judgments remain stable across evaluators, while absolute scalar scores are inherently model-dependent. This motivates explainability pipelines that emphasize structured, comparative judgments over pointwise scoring. Overall, these results suggest a shift in how explainability should be studied. Instead of asking whether an explanation is correct in isolation, a more productive approach is to ask which internal representations matter for behavior, when explanations fail systematically, and which forms of evaluation remain trustworthy across models. TriEx provides one concrete framework for pursuing this direction in interactive and multi-agent settings. TriEx can also serve as a structured evaluation benchmark that yields interpretable, multi-dimensional model profiles (see Appendix C for further details).

5 Conclusion

We presented TriEx, a tri-view framework for studying explainability in LLM-based agents under interactive decision-making. By jointly analyzing self-explanations, internal beliefs about other agents, and third-person audits, TriEx reveals structured failures of explanation faithfulness, intervention-sensitive second-person belief states reflected in decision behavior, and a hierarchy of reliability in explanation evaluation. Our results suggest that explainability in language agents is

inherently interaction-dependent and cannot be reduced to isolated reasoning traces or single evaluators. Future explainability research should therefore focus on relational, multi-level analyses of agent behavior in dynamic environments, for which TriEx provides a concrete starting point.

6 Limitations

TriEx is evaluated in a controlled environment, which abstracts away many complexities of real-world NLP applications; however, it is designed to isolate core interaction and belief-update phenomena rather than task-specific skills. The framework assumes access to structured reasoning and belief interfaces, but these requirements are methodological rather than architectural and can be adapted to alternative forms of model instrumentation.

7 Acknowledgments

This work was supported by the Australian Research Council Discovery Project DP260102534.

References

- Alpha Arena Team. 2025. Alpha arena: Ai trading competition in real crypto markets. <https://www.alpha-arena.org/>. Accessed: 2025-12-04.
- Mark W Baldwin. 1992. Relational schemas and the processing of social information. *Psychological bulletin*, 112(3):461.
- Oliver Bentham, Nathan Stringham, and Ana Marasović. 2024. Chain-of-thought unfaithfulness as disguised accuracy. *Trans. Mach. Learn. Res.*
- Colin Camerer and Teck Hua Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.
- Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The quarterly journal of economics*, 119(3):861–898.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):1–45.
- 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 *Proc. Int. Conf. Learn. Representations*, volume 37, pages 28219–28253.

- Chris D Frith and Uta Frith. 2006. The neural basis of mentalizing. *Neuron*, 50(4):531–534.
- Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *Proc. Int. Conf. Learn. Representations*.
- Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. 2025. Igame-bench: How good are llms at playing games? *arXiv preprint arXiv:2505.15146*.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pages 4198–4205. Association for Computational Linguistics (ACL).
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proc. Conf. Empir. Methods in Natural Language Process.*, pages 2757–2791.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, and 6 others. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pages 295–337.
- Aleksandar Makelov, Georg Lange, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *Proc. Int. Conf. Learn. Representations*.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *Proc. Int. Conf. Learn. Representations*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Proc. Advances in Neural Inf. Process. Syst.*, 35:17359–17372.
- Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.
- Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pages 6048–6089.
- Wenjun Peng, Xinyu Wang, and Qi Wu. 2026. Proxy-war: Dynamic assessment of llm code generation in game arenas. In *Proc. Int. Conf. Softw. Eng.*
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proc. ACM SIGKDD Int. Conf. Knowledge discovery & data mining*, pages 1135–1144.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proc. Int. Conf. Mach. Learn.*, pages 3319–3328. PMLR.
- Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Proc. Advances in Neural Inf. Process. Syst.*, 36:74952–74965.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *Proc. Int. Conf. Learn. Representations*.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Xinyu Wang, Bohan Zhuang, and Qi Wu. 2024b. Modaverse: Efficiently transforming modalities with llms. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 26606–26616.
- Xinyu Wang, Bohan Zhuang, and Qi Wu. 2025. Are large vision language models good game players? In *Proc. Int. Conf. Learn. Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Proc. Advances in Neural Inf. Process. Syst.*, 35:24824–24837.
- Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. 2024. Smartplay: A benchmark for llms as intelligent agents. In *Proc. Int. Conf. Learn. Representations*.
- Wako Yoshida, Ray J Dolan, and Karl J Friston. 2008. Game theory of mind. *PLoS computational biology*, 4(12):e1000254.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *Proc. Int. Conf. Learn. Representations*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):1–38.

Appendix

Due to page limitations in the main paper, we defer a substantial amount of technical detail and extended analysis to this appendix. Beyond ensuring full reproducibility of the TriEx framework, the appendix provides additional experimental evidence that complements and strengthens the main findings. Specifically, we include detailed descriptions of the poker environment and instrumentation used to construct decision traces, complete schemas and prompts for the first-, second-, and third-person views, and a range of extended analyses. These supplementary results cover per-model and per-regime evaluations, additional quantitative curves, robustness and sensitivity analyses, as well as qualitative case studies that illustrate model behavior in depth. The appendix is organized as follows.

Appendix Contents

A Implementation Details of TriEx	12
A.1 Game Setting and Engine	12
A.2 LLM Agents and Algorithmic Players	13
A.3 Prompt Templates	14
A.3.1 First-Person Decision Prompt	14
A.3.2 Second-Person Opponent Profiling Prompt	14
A.3.3 Third-Person Oracle- Auditor Prompt	15
A.3.4 Oracle Prompt for Audit- ing Second-Person Oppo- nent Profiles	15
B Extended Experiments and Analysis	15
B.1 Extended Analysis: First-Person Explanations	15
B.2 Extended Analysis: Second- Person Beliefs	16
B.3 Extended Analysis: Third-Person Auditing	19
C TriEx as an Evaluation Benchmark	22
D Verbatim Prompt Templates	28
D.1 First-Person Decision Prompt . . .	28
D.2 Second-Person Opponent Profiling Prompt	29
D.3 Third-Person Oracle-Auditor Prompt	29
D.4 Oracle Prompt for Auditing Second-Person Opponent Profiles	31

A Implementation Details of TriEx

A.1 Game Setting and Engine

All experiments in this paper are conducted in the setting of *No-Limit Texas Hold'em*, a standard imperfect-information, multi-stage decision-making game. A single game consists of a sequence of independent *hands*. In each hand, players receive private hole cards, interact through up to four betting streets (preflop, flop, turn, and river), and the hand terminates either when all but one player fold or at a final showdown, where the highest-ranking five-card hand wins the pot.

We adopt a controlled experimental variant of the game designed for large-scale simulation and reproducible evaluation. Each experimental match consists of a fixed number of hands, with all players starting from identical initial stacks and fixed blind sizes. This setting departs from real-world cash games (unbounded duration) and tournament formats (elimination-based), but provides bounded runtime, stable statistics, and consistent comparison across models and runs.

Our implementation is built on top of the open-source PyPokerEngine framework,¹ which provides a rule-consistent poker engine, betting mechanics, and hand evaluation. All TriEx components, including decision trace construction, agent instrumentation, and experimental control, are implemented as modular extensions on top of this engine.

Experimental configuration. Unless otherwise specified, experiments are conducted with a fixed table configuration. Each experiment consists of multiple independent *battles*, where each battle runs a fixed number of hands. All players begin each battle with equal initial stacks and fixed blind amounts. To improve throughput and reduce variance, battles are executed independently and can be parallelized across processes. This design ensures that each battle is statistically independent and that aggregated results are not affected by execution order.

Action space and legality. At each decision point, the poker engine exposes a legal action interface consisting of fold, call, and raise, together with engine-defined bounds on admissible raise sizes. Our LLM-based agents internally reason over a slightly expanded action vocabulary (FOLD,

¹<https://github.com/ishikota/PyPokerEngine>

CALL, CHECK, RAISE), which is deterministically normalized to engine-legal actions before execution. Specifically, CHECK is permitted only when no bet is faced and is mapped to a zero-cost CALL; otherwise it is treated as illegal and conservatively downgraded to FOLD. Proposed raise amounts are clipped to the engine-provided legal interval, and when raising is unavailable the action is downgraded to CALL. This normalization guarantees rule-consistent gameplay while preserving the agent’s intended decision semantics.

Equity estimation and decision inputs. To provide agents with a stable and interpretable signal of hand strength, we compute an auxiliary estimate of hand equity (win probability) at each decision point using Monte Carlo simulation. Given the current private cards and public board, opponent hole cards are sampled from the remaining deck, the board is completed if necessary, and hands are evaluated using a standard poker hand evaluator. The number of simulations is adjusted by street to balance variance and computational cost. This environment-derived equity signal is provided as part of the agent’s decision input and logged for downstream analysis; it is not intended to model human reasoning, but to reduce stochasticity and support controlled behavioral comparisons.

Reproducibility parameters. We provide the full experimental configuration to ensure reproducibility:

Parameter	Value
Number of battles	50
Hands per battle	30
Initial stack size	3000
Small blind / Big blind	5 / 10
Decoding temperature	0.2
Top-p	1.0
Random seed	7
Logit intervention magnitude (δ)	2.5
Monte Carlo simulation count	1000

Table 4: **Experimental configuration.** All agents are evaluated under identical game and decoding settings to ensure reproducibility.

A.2 LLM Agents and Algorithmic Players

Our experimental setup involves a mixed population of LLM agents and algorithmic players. The LLM agents are the primary subjects of analysis and explanation in this work, while the algorithmic players serve as controlled reference opponents

with predefined and interpretable behavioral patterns.

LLM agents. We evaluate a diverse set of contemporary LLMs spanning both open-weight and proprietary models. These models (see Table 5) differ in architecture, scale, and training data, but are all interfaced through a unified decision-making prompt that produces structured actions and self-explanations compatible with the TriEx framework. The goal is not to benchmark raw playing strength, but to analyze how different models articulate, ground, and revise explanations and beliefs under identical decision contexts.

Model	Provider	Release Year	Model Size
GPT-4.1-mini	OpenAI	2025	N/A
DeepSeek-V3.2	DeepSeek	2025	N/A
Gemini-2.5-Flash-Lite	Google	2025	N/A
Grok-3-Mini	xAI	2025	N/A
Qwen3-32B	Alibaba	2025	32.8B
Llama-4-Maverick	Meta	2025	17B

Table 5: LLM agents used in our experiments. Proprietary models do not disclose parameter counts.

Algorithmic baseline players. To provide a stable and interpretable reference environment, we include a cohort of five rule-based algorithmic players. Each algorithmic player instantiates a distinct and well-defined strategic archetype, parameterized by standard poker behavioral statistics such as Voluntarily Put Money in Pot (VPIP), Pre-Flop Raise frequency (PFR), and Aggression Factor (AF). Importantly, these players are not learning agents: their policies are fixed, known *a priori*, and remain unchanged throughout all experiments. Specifically, the algorithmic players correspond to the following canonical archetypes:

- **Loose-Passive:** characterized by frequent calling and low aggression.
- **Loose-Aggressive:** exhibiting high voluntary participation and frequent raises.
- **Maniac:** employing highly aggressive and variance-inducing strategies across hand strengths.
- **Tight-Passive:** entering pots selectively and minimizing aggressive actions.
- **Tight-Aggressive:** representing conservative but forceful value-oriented play.

Each archetype is implemented via deterministic or lightly stochastic rules based on hand equity, pot odds, and street-specific constraints, ensuring rule-consistent but behaviorally distinct play.

Mixed-table protocol and interpretability role.

LLM agents and algorithmic players are seated together at the same tables and interact under identical game conditions. Agents are not informed of their opponents' identities or types, and all players are anonymized during play. Crucially, this mixed-table setup constitutes a heterogeneous multi-agent environment: multiple LLM agents play against one another at the same table, with algorithmic players included as controlled reference opponents. Second-person opponent modeling is applied uniformly to all opponents, both adaptive LLM agents and fixed-strategy algorithmic players, so the belief-formation mechanism itself operates identically across opponent types. The algorithmic players serve a specifically diagnostic role: because their behavioral tendencies are predefined and stable, they provide a reference against which second-person belief grounding and third-person auditing can be meaningfully assessed. They are used to anchor specific analyses, not to replace the underlying multi-agent interaction among LLM agents. In our turn-based, imperfect-information setting, opponent actions are the primary channel through which any player can infer others' tendencies; LLM agents therefore continuously update second-person beliefs from observed behavior and act on those beliefs in subsequent decisions, which naturally induces dynamic mutual adaptation among LLM agents even though the game is not framed as an explicit co-training or self-play optimization process.

This mixed-population design is critical for the explainability focus of TriEx. It enables controlled analysis of whether LLMs can (i) form accurate beliefs about others' behavior, (ii) ground those beliefs in observable evidence, and (iii) articulate explanations that remain faithful to both their own actions and the surrounding strategic environment.

A.3 Prompt Templates

This subsection documents the prompt templates used to instantiate the first-person, second-person, and third-person views in the TriEx framework. All LLM agents are queried using identical prompt templates within each role, ensuring that observed differences in behavior, belief formation, or expla-

nation are attributable to model differences rather than prompt variation. The prompts are designed to produce structured, machine-readable outputs that support decision execution, belief tracking, and post-hoc auditing. We report the templates verbatim below for reproducibility.

The structured output format is intentionally lightweight and is used to make the three views auditable and comparable, rather than to impose a fixed reasoning ontology on the model. The required fields are limited to action commitment, brief explanation, and compact belief-state variables needed for downstream analysis. If the observed outputs primarily reflected schema compliance rather than model-specific decision processes, one would expect relatively uniform behavior under the same prompt structure. Instead, the main-paper results show substantial variation across game stages (Finding 1), across decision complexity (Table 7), and across model families (Table 10), and the intervention experiments (Finding 3) show that controlled perturbations of the numeric belief variables yield directionally consistent policy shifts under identical JSON structure—suggesting that the recorded outputs reflect meaningful differences in explanation quality and belief use rather than formatting effects alone.

A.3.1 First-Person Decision Prompt

At each decision point, an LLM-based agent is provided with a structured first-person prompt describing the current game state, recent opponent actions, and available long-term opponent profile estimates. The prompt explicitly instructs the agent to reason about risk, position, and opponent tendencies, and to produce both a concise self-explanation and a machine-readable action decision. Crucially, the self-explanation is generated *before* action execution and is intended to reflect anticipatory reasoning rather than post-hoc justification. The full verbatim prompt template is provided in Appendix D.1.

A.3.2 Second-Person Opponent Profiling Prompt

Following each completed hand, agents update explicit second-person representations of their opponents. This prompt operationalizes an evidence-to-belief transition, mapping aggregated behavioral statistics into interpretable latent traits. The full verbatim prompt template is provided in Appendix D.2.

A.3.3 Third-Person Oracle-Auditor Prompt

The third-person oracle operates purely as an external auditor. It does not influence gameplay, action selection, or belief updates. Instead, it evaluates the faithfulness of an agent’s self-explanation with respect to the objective game state and the action actually taken. The full verbatim prompt template is provided in Appendix D.3.

A.3.4 Oracle Prompt for Auditing Second-Person Opponent Profiles

This prompt enables quantitative evaluation of whether second-person beliefs faithfully reflect observable opponent behavior, independent of downstream decision quality. The full verbatim prompt template is provided in Appendix D.4.

B Extended Experiments and Analysis

This section presents extended experimental results and analyses that complement the main findings reported in the paper.

B.1 Extended Analysis: First-Person Explanations

This subsection reports extended quantitative analyses for first-person faithfulness that complement the aggregated results in the main paper. While the main text focuses on high-level trends, here we provide detailed summary statistics and stratified breakdowns to clarify how faithfulness and rationalization vary across game stages, action risk levels, and models.

Group	Rule	Oracle	Rat. (Rule)	Rat. (Oracle)	ρ	HighRisk
All LLMs	4.37	3.56	0.21	0.43	0.78	0.11

Table 6: **Overall first-person faithfulness across all audited LLM decisions.** Mean rule-based (Rule) and oracle-based (Oracle) faithfulness scores (1–5), rationalization rates, Spearman rank correlation (ρ) between the two audits, and the proportion of high-risk actions.

Overall summary statistics. We first report aggregate audit statistics across all audited LLM decision points. Table 6 summarizes mean rule-based and oracle-based faithfulness scores, rationalization rates under both audits, the Spearman correlation between rule and oracle scores, and the proportion of high-risk actions. In Table 6, first, oracle-based faithfulness scores are consistently lower than rule-based scores, indicating that oracle auditing captures additional forms of explanation

failure beyond explicit rule violations. Second, the high Spearman correlation ($\rho = 0.78$) suggests that, despite this systematic gap, the two audits remain directionally aligned, supporting the internal consistency of the evaluation pipeline.

Street	N	Rule	Oracle	Rat. (Rule)	Rat. (Oracle)	ρ	HighRisk
Preflop	1533	4.49	3.92	0.15	0.31	0.78	0.09
Flop	411	4.13	3.03	0.27	0.61	0.67	0.14
Turn	291	4.14	2.88	0.33	0.64	0.69	0.17
River	192	4.23	2.86	0.39	0.65	0.72	0.17

Table 7: **First-person faithfulness statistics by game street.** Mean rule-based (Rule) and oracle-based (Oracle) faithfulness scores (1–5), rationalization rates, Spearman rank correlation (ρ) between the two audits, and the proportion of high-risk actions, stratified by street.

Faithfulness by game stage. Table 7 reports the same statistics stratified by street. Faithfulness is highest in the preflop stage and degrades monotonically thereafter, with a sharp transition at the flop. Under oracle auditing, the rationalization rate nearly doubles from preflop to flop (0.31 \rightarrow 0.61) and remains high through the turn and river. In contrast, rule-based rationalization increases more gradually, indicating that later streets introduce softer failures such as vague, underspecified, or weakly grounded explanations that are not always detectable by deterministic checks.

Per-model Faithfulness trend curves. Figure 4 and Table 8 demonstrate this effect at the model level. While the preflop-to-flop drop is consistent across models, postflop trajectories diverge, with some models exhibiting partial recovery and others continuing to degrade. These per-model differences are obscured in aggregated distributions and motivate the model-level analysis below.

Model	Preflop	Flop	Turn	River
DeepSeek-V3.2	4.316	3.071	3.381	2.833
GPT-4.1-mini	3.598	3.174	2.847	2.885
Gemini-2.5-Flash-Lite	4.027	2.725	2.966	2.750
Grok-3-Mini	3.736	3.230	2.984	3.283
Llama-4-Maverick	3.631	2.727	2.475	2.222
Qwen-3-32B	4.097	3.077	3.200	3.375

Table 8: Per-model first-person faithfulness scores (1-5) across streets. These values provide a per-model complement to the aggregated stage-level trends reported in the main paper.

Faithfulness by action risk. To disentangle explanation failures driven by state complexity from

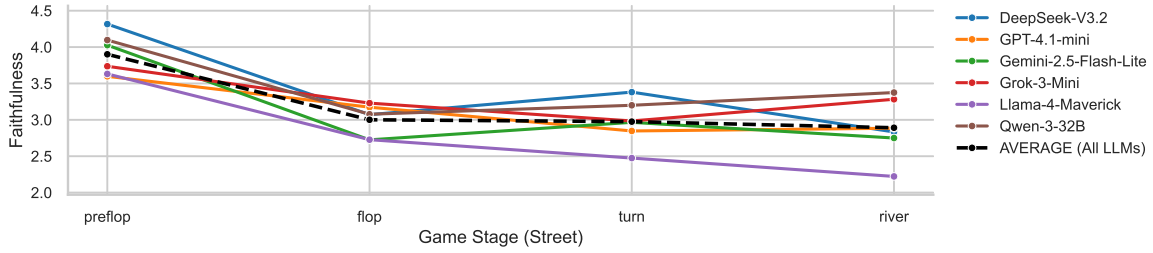


Figure 4: Per-model first-person faithfulness trends across game stages (with an aggregated average curve). The most prominent shift occurs from preflop to flop, while postflop trajectories vary by model.

those driven by decision risk, Table 9 stratifies audit results by whether an action is classified as high risk. High-risk actions are substantially more likely to be judged as rationalized by the oracle (0.77 vs. 0.38), and exhibit markedly lower oracle faithfulness scores (2.41 vs. 3.71). This pattern suggests that rationalization is tightly coupled to decision difficulty and exposure.

Risk Level	N	Rule	Oracle	Rat. (Rule)	Rat. (Oracle)	ρ	Freq.
Low Risk	2149	4.49	3.71	0.11	0.38	0.77	0.89
High Risk	278	3.44	2.41	1.00	0.77	0.61	0.11

Table 9: **First-person faithfulness statistics by action risk level.** Mean rule-based (Rule) and oracle-based (Oracle) faithfulness scores (1–5), rationalization rates, Spearman rank correlation (ρ) between the two audits, and action frequency, stratified by risk level.

Model	N	Rule	Oracle	Rat. (Rule)	Rat. (Oracle)	ρ	HighRisk
GPT-4.1-mini	510	4.25	3.25	0.20	0.55	0.78	0.12
DeepSeek-V3.2	347	4.57	4.06	0.16	0.29	0.80	0.13
Gemini-2.5-Flash-Lite	344	4.43	3.74	0.25	0.33	0.78	0.18
Grok-3-Mini	413	4.42	3.46	0.14	0.44	0.71	0.09
Qwen-3-32B	367	4.59	3.92	0.18	0.34	0.71	0.03
Llama-4-Maverick	446	4.07	3.19	0.32	0.52	0.80	0.14

Table 10: **Model-level first-person faithfulness statistics.** For each LLM, we report the number of audited decision points (N), mean rule-based (Rule) and oracle-based (Oracle) faithfulness scores (1-5), rationalization rates under each audit, Spearman rank correlation (ρ) between the two audits, and the proportion of high-risk actions.

Model-level comparison. Finally, Table 10 reports audit statistics broken down by model. Rather than ranking models by performance, this comparison highlights systematic differences in explanation behavior, including variation in oracle-rule gaps, rationalization rates, and exposure to high-risk actions. These differences help contextualize the per-model trend curves and underscore that first-person faithfulness is not uniform across contemporary LLMs, even under identical prompts and

environments.

B.2 Extended Analysis: Second-Person Beliefs

This subsection provides extended results for the second-person view experiments that are summarized in the main paper. We report (i) per-model belief trajectories that underlie the rank-based convergence curves, (ii) full profile-behavior alignment breakdowns, and (iii) per-model directional intervention statistics.

Belief update trajectories beyond aggregate convergence.

The main paper reports aggregated, rank-based convergence curves for two representative second-person belief dimensions (Fig. 2), which evaluate whether agents’ beliefs about other players stabilize under repeated interaction. To expose the dynamics behind these aggregate trends, we visualize the *raw second-person belief trajectories* maintained by each LLM *player* over the course of interaction. Figures 10–15 plot, for each player model, how its internal opponent profiles evolve. Each figure shows belief updates over the five profile dimensions (*RiskTolerance*, *Aggressiveness*, *BluffFrequency*, *CallingStationTendency*, and *ShowdownPropensity*), for the same set of fixed-strategy algorithmic opponents. These trajectory-level visualizations clarify how convergence arises in practice, revealing differences in early uncertainty, update smoothness, and stabilization speed that are not visible in aggregated rank correlations.

Belief intervention mechanism and evaluation protocol.

We implement belief interventions directly in the structured opponent profile space. For a trait dimension $p_k \in [0, 1]$, we apply a perturbation in logit space:

$$p'_k = \sigma(\text{logit}(p_k) + \delta),$$

where δ controls the intervention magnitude ($\delta = 2.5$); see Table 4) and $\sigma(\cdot)$ denotes the sigmoid

Trait	Metric	ρ	p
Model: deepseek-v3.2			
Aggressiveness	raise_rate	0.9000	0.0374
BluffFrequency	bluff_success_rate	-1.0000	0.0000
CallingStationTendency	aggression_factor	-0.8000	0.1041
RiskTolerance	fold_rate	-0.9000	0.0374
ShowdownPropensity	fold_rate	0.8000	0.1041
Model: gemini-2.5-flash-lite			
Aggressiveness	fold_rate	-0.9000	0.0374
BluffFrequency	raise_rate	0.8000	0.1041
CallingStationTendency	call_to_fold_ratio	0.8000	0.1041
RiskTolerance	fold_rate	-0.8000	0.1041
ShowdownPropensity	fold_rate	-0.9000	0.0374
Model: gpt-4.1-mini			
Aggressiveness	raise_rate	0.7000	0.1881
BluffFrequency	bluff_attempt_rate	0.9000	0.0374
CallingStationTendency	aggression_factor	-0.6000	0.2848
RiskTolerance	raise_rate	0.7000	0.1881
ShowdownPropensity	call_rate	0.7000	0.1881
Model: grok-3-mini			
Aggressiveness	raise_rate	0.9000	0.0374
BluffFrequency	bluff_attempt_rate	0.8721	0.0539
CallingStationTendency	aggression_factor	-0.9000	0.0374
RiskTolerance	fold_rate	-0.6000	0.2848
ShowdownPropensity	call_rate	0.9000	0.0374
Model: llama-4-maverick			
Aggressiveness	raise_rate	0.9000	0.0374
BluffFrequency	raise_rate	0.9000	0.0374
CallingStationTendency	aggression_factor	-0.8000	0.1041
RiskTolerance	vpip_proxy	1.0000	0.0000
ShowdownPropensity	call_rate	0.3000	0.6238
Model: qwen3-32b			
Aggressiveness	fold_rate	-0.9000	0.0374
BluffFrequency	bluff_attempt_rate	0.8000	0.1041
CallingStationTendency	aggression_factor	-0.8000	0.1041
RiskTolerance	vpip_proxy	0.9000	0.0374
ShowdownPropensity	call_rate	0.5000	0.3910

Table 11: **Model-level alignment between second-person belief dimensions and gameplay behavior.** For each player model and belief dimension, we report the behavioral proxy yielding the strongest alignment from a predefined proxy set, together with Spearman’s rank correlation ρ and its p -value. This table provides a detailed breakdown underlying the aggregated alignment results reported in the main paper.

function. This formulation ensures bounded and monotonic adjustments while avoiding saturation effects, and provides a comparable intervention scale across different initial belief values.

All intervention experiments reuse identical prompts except for numeric belief vectors. No natural-language descriptions or instructions are modified, ensuring that observed behavioral differences cannot be attributed to prompt rephrasing.

To quantify intervention sensitivity, we compare intervention-induced changes against rerun variability baselines using change rates (CR). Specifically, $CR(\text{Log} \rightarrow \text{ReO})$ denotes the change rate between the logged original decisions (Log) and rerun decisions under identical belief states (ReO), capturing stochastic rerun variability; $CR(\text{Log} \rightarrow \text{ReI})$

denotes the change rate between the logged decisions and rerun decisions under intervened belief states (ReI); and $CR(\text{ReO} \rightarrow \text{ReI})$ denotes the change rate between rerun-original and rerun-intervened decisions under matched conditions.

These metrics complement the aggregated intervention summary in the main paper by separating intervention-associated changes from stochastic rerun variability.

Profile-behavior alignment. To complement the aggregated profile-behavior alignment result in the main paper, Table 11 reports the model-level alignment between each second-person belief dimension and its strongest behavioral proxy, together with Spearman’s rank correlation and p -value. Across

Model	Trait	Dir.	Δ Fold	Δ Call	Δ Raise	Change Rate	Dir. Cons.
deepseek-v3.2	Aggressiveness	up	-0.0497	+0.0399	+0.0098	0.2817	0.8696
deepseek-v3.2	Aggressiveness	down	-0.0479	+0.0556	-0.0077	0.2821	0.9132
gemini-2.5-flash-lite	Aggressiveness	up	-0.0040	-0.0025	+0.0065	0.1588	0.9226
gemini-2.5-flash-lite	Aggressiveness	down	+0.0030	+0.0064	-0.0094	0.1462	0.9293
gpt-4.1-mini	Aggressiveness	up	+0.0204	+0.0356	-0.0220	0.0016	0.9197
gpt-4.1-mini	Aggressiveness	down	+0.0014	-0.0112	+0.0099	0.2317	0.9122
grok-3-mini	Aggressiveness	up	-0.0099	+0.0113	-0.0014	0.2641	0.8917
grok-3-mini	Aggressiveness	down	-0.0124	+0.0141	-0.0017	0.2654	0.8972
llama-4-maverick	Aggressiveness	up	+0.0358	-0.0351	-0.0008	0.3193	0.8840
llama-4-maverick	Aggressiveness	down	+0.0228	-0.0130	-0.0098	0.3125	0.8463
qwen3-32b	Aggressiveness	up	-0.0331	-0.0084	+0.0415	0.2686	0.8440
qwen3-32b	Aggressiveness	down	-0.0406	+0.0022	+0.0384	0.2764	0.9032

Table 12: **Per-model directional intervention effects on second-person aggressiveness beliefs.** We report changes in coarse action distributions (Δ Fold, Δ Call, Δ Raise), the overall action change rate relative to the logged baseline, and directional consistency with the expected semantic effect. Results are shown separately for upward and downward interventions.

Model	Trait	Dir.	Δ Fold	Δ Call	Δ Raise	Change Rate	Dir. Cons.
deepseek-v3.2	RiskTolerance	up	-0.0070	+0.0282	-0.0211	0.2535	0.9085
deepseek-v3.2	RiskTolerance	down	+0.0035	+0.0000	-0.0035	0.2711	0.8944
gemini-2.5-flash-lite	RiskTolerance	up	-0.0173	+0.0138	+0.0035	0.1349	0.9343
gemini-2.5-flash-lite	RiskTolerance	down	-0.0173	+0.0173	+0.0000	0.1522	0.9412
gpt-4.1-mini	RiskTolerance	up	+0.0000	-0.0178	+0.0178	0.2705	0.8861
gpt-4.1-mini	RiskTolerance	down	-0.0107	-0.0036	+0.0142	0.2954	0.8826
grok-3-mini	RiskTolerance	up	+0.0070	-0.0140	+0.0070	0.2832	0.8951
grok-3-mini	RiskTolerance	down	-0.0140	-0.0035	+0.0175	0.2692	0.9056
llama-4-maverick	RiskTolerance	up	-0.0383	+0.0174	+0.0209	0.4042	0.7840
llama-4-maverick	RiskTolerance	down	-0.0418	+0.0453	-0.0035	0.4007	0.8223
qwen3-32b	RiskTolerance	up	-0.0451	+0.0104	+0.0347	0.2674	0.8472
qwen3-32b	RiskTolerance	down	-0.0243	-0.0069	+0.0312	0.2500	0.9062

Table 13: **Per-model directional intervention effects on second-person risk tolerance beliefs.** For each player model and intervention direction, we report action distribution shifts, change rates relative to the logged baseline, and directional consistency. This table complements the aggregated intervention summary in the main paper by exposing cross-model heterogeneity.

models, intent-level traits (*Aggressiveness*, *Bluff-Frequency*) achieve the tightest alignment with their behavioral proxies, whereas outcome-dependent traits (*ShowdownPropensity*) show weaker and noisier correlations, consistent with the pattern reported in the main paper. This breakdown exposes substantial per-model heterogeneity that is averaged out in aggregate statistics.

Directional intervention outcomes. The main paper reports aggregated evidence that second-person belief dimensions are grounded in observable gameplay behavior, with stronger alignment for intent-level traits than for outcome-dependent traits. Here we provide the complete model-level alignment statistics that support this conclusion. Tables 12 and 13 report detailed directional intervention outcomes for *Aggressiveness* and *Risk-*

Tolerance, respectively. For each player model, we report changes in the coarse action distribution (Δ Fold, Δ Call, Δ Raise), the overall action change rate relative to the logged baseline, and the directional consistency rate measuring semantic alignment.

Several insights emerge beyond the aggregated results in the main paper. First, while all models exhibit above-noise change rates, the magnitude of intervention effects varies substantially, indicating differences in how tightly belief states are coupled to action selection. Second, intent-level traits produce more concentrated and sign-consistent action shifts, whereas risk-related traits often induce more distributed changes across actions, reflecting their higher-level semantic scope. Finally, some models display asymmetric sensitivity between upward and downward interventions, suggesting nonlinear

Model	Dir.	CR(Log→ReI)		CR(Log→ReO)		CR(ReO→ReI)	
		mean	var	mean	var	mean	var
deepseek-v3.2	down	0.2821	0.0003	0.2836	0.0005	0.2832	0.0005
deepseek-v3.2	up	0.2817	0.0003	0.2786	0.0003	0.2770	0.0005
gemini-2.5-flash-lite	down	0.1462	0.0001	0.1543	0.0002	0.1438	0.0002
gemini-2.5-flash-lite	up	0.1588	0.0002	0.1531	0.0001	0.1565	0.0003
gpt-4.1-mini	down	0.2317	0.0003	0.2251	0.0002	0.1797	0.0003
gpt-4.1-mini	up	0.2243	0.0002	0.2278	0.0002	0.1739	0.0005
grok-3-mini	down	0.2654	0.0004	0.2541	0.0004	0.2467	0.0005
grok-3-mini	up	0.2641	0.0003	0.2559	0.0003	0.2457	0.0005
llama-4-maverick	down	0.3125	0.0002	0.2850	0.0004	0.2438	0.0005
llama-4-maverick	up	0.3193	0.0002	0.2872	0.0005	0.2439	0.0005
qwen3-32b	down	0.2764	0.0004	0.2751	0.0003	0.3013	0.0007
qwen3-32b	up	0.2686	0.0005	0.2715	0.0004	0.2897	0.0006

Table 14: **Rerun-controlled change-rate statistics for the Aggressiveness trait.** For each model and intervention direction, we report the mean and variance of three change-rate metrics: CR(Log→ReI), CR(Log→ReO), and CR(ReO→ReI).

use of belief magnitudes rather than purely ordinal thresholds. These per-model results reinforce the main conclusion that second-person beliefs exhibit causal sensitivity under controlled belief interventions, while highlighting meaningful heterogeneity in how different LLMs instantiate and use these beliefs during decision-making.

Rerun-controlled variability analysis. To further assess whether intervention-associated changes can be distinguished from stochastic rerun variability, we conducted rerun-controlled experiments for the *Aggressiveness* trait. We report *Aggressiveness* as a representative action-proximal trait to illustrate the rerun-controlled protocol and its interpretation. For each model and intervention direction, we repeated the experiments over 50 runs and summarized the resulting change-rate statistics using their mean and variance across runs.

Table 14 reports mean and variance of the three change-rate metrics (CR(Log→ReO), CR(Log→ReI), CR(ReO→ReI)) introduced earlier in the ‘Belief Intervention Mechanism and Evaluation Protocol’ paragraph, aggregated over 50 runs per model and intervention direction.

These rerun-controlled results provide additional context for interpreting the intervention findings in the main paper. Across models, intervention-associated and rerun-original change rates are of broadly similar magnitude, yet differ structurally: rerun-original variability arises from identical-input stochasticity and therefore has no expected directional bias, whereas intervention-associated shifts are systematically aligned with the sign of

the perturbation (see directional consistency rates in Tables 12 and 13). At the same time, the magnitude of this effect differs across models and directions, highlighting heterogeneous sensitivity to second-person belief states. The reported variances are uniformly small, suggesting that the estimated change rates are stable across repeated runs. Taken together, these results support interpreting the observed policy shifts as intervention-sensitive behavioral effects rather than as artifacts of rerun stochasticity alone.

B.3 Extended Analysis: Third-Person Auditing

This subsection provides extended analyses for the third-person auditing results reported in the main paper. While the main text focuses on aggregated reliability trends and high-level conclusions, here we present additional diagnostics that characterize the *consistency structure* among different oracle LLMs. Specifically, we report cross-oracle similarity analyses based on Spearman rank correlation. Each visualization compares how different oracle models rank the same set of decision instances along a given evaluation dimension. By focusing on rank-based agreement rather than absolute scores, this analysis isolates ordinal consistency and mitigates calibration differences across oracle models.

Cross-oracle similarity as ordinal agreement.

For each evaluation dimension, we compute a pairwise oracle similarity matrix, where each entry corresponds to the Spearman rank correlation between two oracle LLMs over the aligned intersection of

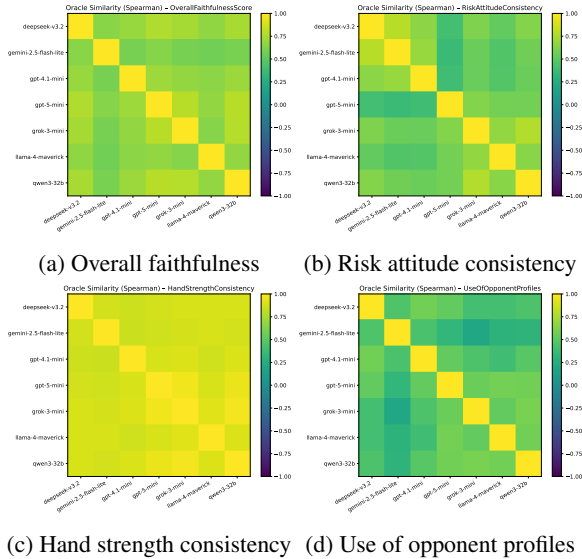


Figure 5: **Cross-oracle similarity (Spearman rank correlation) across auditing dimensions.** Each heatmap reports pairwise ordinal agreement between oracle LLMs computed over the aligned set of audited decision instances.

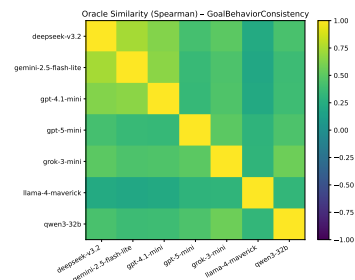


Figure 6: Cross-oracle similarity (Spearman) for goal-behavior consistency. This dimension exhibits higher variance due to its reliance on latent intent attribution.

audited decision points. High values indicate that two oracles produce similar relative orderings of instances, even if their absolute scores differ. Diagonal entries correspond to self-similarity and are omitted from interpretation. Off-diagonal structure reveals whether agreement is global (e.g., consistently high across all oracle pairs) or localized to specific subsets of models.

Dimension-dependent consistency patterns. Figures 5a–6 visualize cross-oracle similarity for five auditing dimensions. We observe systematic variation across dimensions. *OverallFaithfulnessScore* (Fig. 5a) exhibits uniformly high agreement across oracle pairs, indicating that holistic judgments of explanation faithfulness are comparatively stable at the ordinal level.

Oracle Model	ρ
Grok-3-mini	0.3899
Qwen3-32B	0.3810
GPT-4.1-mini	0.3685
DeepSeek-v3.2	0.3667
LLaMA-4-maverick	0.3241
Gemini-2.5-flash-lite	0.2865

Table 15: **Overall oracle-rule alignment under a stricter numerical rationality reference.** We report the Spearman rank correlation between oracle judgments and the enhanced rule-based audit, which incorporates pot odds, stack-to-pot ratio (SPR), and raise-sizing constraints. Compared to the main-text results, correlations are systematically lower, reflecting the increased difficulty of auditing fine-grained numerical rationality from natural-language explanations.

In contrast, more fine-grained or semantically specific criteria show increased dispersion. For example, *RiskAttitudeConsistency* (Fig. 5b) and *HandStrengthConsistency* (Fig. 5c) exhibit moderate agreement with visible block structure, suggesting partial alignment combined with model-specific interpretations of risk and state strength. Finally, dimensions that require reasoning about internal agent state or longer-horizon intent, such as *UseOfOpponentProfiles* (Fig. 5d) and *GoalBehaviorConsistency* (Fig. 6), exhibit the greatest variability. Here, oracle agreement is still well above chance, but noticeably weaker than for surface-level criteria, reflecting the inherent ambiguity of auditing latent belief usage from textual explanations alone.

These extended visualizations support the main text’s conclusion that third-person auditing is reliable at the level of relative ordering, while absolute calibration and fine-grained semantic judgments remain more sensitive to oracle choice. Importantly, the observed consistency patterns are dimension-dependent and systematic rather than random, suggesting that oracle disagreement reflects genuine interpretive latitude rather than noise.

Stricter numerical rationality and oracle-rule alignment. Beyond the similarity analyses, we further examine how third-person auditing reliability changes under a stricter numerical rationality reference. Table 15 reports overall oracle–rule alignment when the rule-based baseline explicitly incorporates pot odds, stack-to-pot ratio, and raise-sizing constraints. Compared to the main-text setting, correlations are uniformly lower across ora-

Oracle Model	Overall		Aggressiveness		RiskTolerance	
	ρ	Acc	ρ	Acc	ρ	Acc
gpt-4.1-mini	0.7205	0.9994	0.7771	0.9989	0.7478	1.0000
llama-4-maverick	0.6818	0.9480	0.7289	0.9307	0.6840	0.9654
gemini-2.5-flash-lite	0.6287	0.9775	0.6877	0.9744	0.6244	0.9806
deepseek-v3.2	0.4152	0.9732	0.4795	0.9596	0.5230	0.9867
grok-3-mini	0.2690	0.9922	0.3050	0.9928	0.2110	0.9917
qwen3-32b	0.1187	0.9556	0.1393	0.9607	0.0944	0.9506

Table 16: **Third-person auditing performance on second-person belief usage, by trait.** For each oracle model, we report Spearman rank correlation between the oracle alignment score and the negative ranking discrepancy, together with directional accuracy, aggregated overall and separately for *Aggressiveness* and *RiskTolerance*.

Oracle Model	Preflop	Flop	Turn	River
grok-3-mini	0.4435	0.3150	0.2963	0.2658
qwen3-32b	0.4521	0.2501	0.1992	0.1781
deepseek-v3.2	0.4404	0.1935	0.2703	0.1661
gpt-4.1-mini	0.4271	0.2640	0.2685	0.2426
llama-4-maverick	0.3835	0.1936	0.1738	0.2335
gemini-2.5-flash-lite	0.3553	0.1221	0.1408	0.1242

Table 17: **Oracle–rule alignment stratified by game street.** Spearman rank correlation is reported separately for preflop, flop, turn, and river for each oracle model. This stratification reveals how auditing reliability varies with game-stage complexity.

Oracle Model	Weak	Medium	Strong
grok-3-mini	0.4556	0.2616	0.2223
qwen3-32b	0.4598	0.2183	0.0990
deepseek-v3.2	0.4605	0.1815	0.0643
gpt-4.1-mini	0.4548	0.2077	0.1490
llama-4-maverick	0.4281	0.0963	0.1335
gemini-2.5-flash-lite	0.3691	0.1138	0.1210

Table 18: **Oracle–rule alignment stratified by hand-strength regime.** Spearman rank correlation is reported for weak, medium, and strong hand-strength buckets. This breakdown isolates the effect of strategic stakes and numerical precision on third-person auditing reliability.

cle models, indicating that fine-grained numerical plausibility is substantially harder to audit from free-form explanations.

Stratified reliability across game structure. To localize sources of auditing difficulty, we report stratified oracle–rule alignment across multiple axes. Table 17 shows that agreement is consistently highest in preflop and degrades in later streets, reflecting the growth of shared public information and counterfactual complexity. Table 18 further reveals that alignment is strongest for weak hands and decreases for strong hands, where optimal play depends more sensitively on precise nu-

Oracle Model	FOLD	CALL	RAISE
grok-3-mini	0.3200	0.2452	0.4823
qwen3-32b	0.2825	0.2576	0.5785
deepseek-v3.2	0.3269	0.1390	0.5156
gpt-4.1-mini	0.2720	0.3189	0.5537
llama-4-maverick	0.1962	0.2529	0.4736
gemini-2.5-flash-lite	0.2476	0.1395	0.3086

Table 19: **Oracle–rule alignment stratified by action type.** Spearman rank correlation is reported separately for FOLD, CALL, and RAISE decisions. This analysis highlights systematic differences in auditing difficulty across decision modes.

merical reasoning. Finally, Table 19 highlights pronounced differences across action types, with RAISE decisions exhibiting higher alignment than CALL or FOLD, suggesting that explicit commitment actions are easier to audit than passive or default choices.

Auditing second-person belief usage. In addition to first-person explanation faithfulness, we evaluate third-person auditing of second-person belief usage. Table 16 reports oracle performance in assessing whether agents’ decisions are consistent with their internal opponent beliefs. Across models, directional accuracy remains high, while rank correlation varies substantially by oracle and by trait. This pattern indicates that coarse directional judgments are robust, whereas fine-grained ranking of belief–decision alignment is more oracle-sensitive.

Effect of temporal evidence windows. Table 20 examines how auditing reliability varies with the amount of interaction history provided to the oracle. While increasing the window size generally improves rank correlation, the gains are uneven across models and traits. Notably, directional accuracy remains stable even at short windows, sug-

Oracle / Trait	$\rho@5$	DirAcc@5	$\rho@10$	DirAcc@10	$\rho@15$	DirAcc@15
openai/gpt-4.1-mini	0.72049	0.99944	0.76577	0.99667	0.71639	0.99625
Aggressiveness	0.77712	0.99889	0.76154	0.99750	0.68015	0.99250
RiskTolerance	0.74779	1.00000	0.78630	0.99583	0.76486	1.00000
deepseek/deepseek-v3.2	0.41520	0.97318	0.64208	0.98625	0.61292	0.98292
Aggressiveness	0.47946	0.95955	0.64196	0.97667	0.65455	0.96917
RiskTolerance	0.52299	0.98667	0.74273	0.99583	0.70817	0.99667
google/gemini-2.5-flash-lite	0.62875	0.97750	0.65288	0.98375	0.68352	0.98125
Aggressiveness	0.68765	0.97444	0.67879	0.97750	0.59412	0.97333
RiskTolerance	0.62443	0.98056	0.66023	0.99000	0.73471	0.98917
x-ai/grok-3-mini	0.26903	0.99222	0.30237	0.99500	0.24097	0.98083
Aggressiveness	0.30499	0.99278	0.25424	0.99833	0.31669	0.97917
RiskTolerance	0.21104	0.99167	0.28508	0.99167	0.20388	0.98250
meta-llama/llama-4-maverick	0.68178	0.94804	0.73763	0.96083	0.75936	0.95210
Aggressiveness	0.72893	0.93073	0.67986	0.94333	0.74490	0.93559
RiskTolerance	0.68401	0.96536	0.80557	0.97833	0.81366	0.96833
qwen/qwen3-32b	0.11867	0.95559	0.14288	0.93393	0.07372	0.94751
Aggressiveness	0.13934	0.96067	0.11915	0.92818	0.23987	0.95046
RiskTolerance	0.09441	0.95056	0.14422	0.93947	-0.04751	0.94464

Table 20: **Third-person auditing performance under varying temporal windows.** We report Spearman rank correlation between oracle alignment scores and ranking discrepancies, together with directional accuracy, for window sizes $K \in \{5, 10, 15\}$. Results are shown both in aggregate and separately by belief dimension, illustrating how evidence window length affects auditing stability.

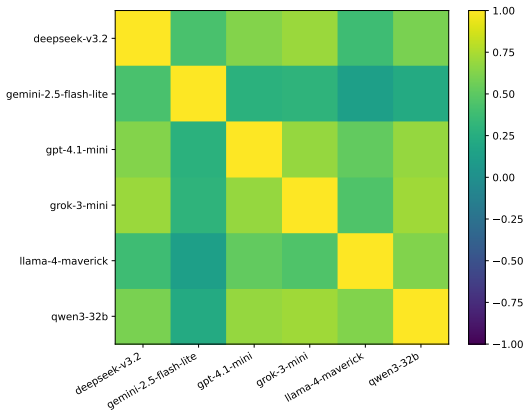


Figure 7: Pairwise oracle agreement measured by quadratic Cohen’s κ on ordinal *OverallFaithfulnessScore*. Higher values indicate stronger ordinal consistency between oracle models when evaluating first-person explanations.

gesting that third-person oracles can reliably detect the sign of belief usage from limited context, even when precise ranking is unstable.

Cross-oracle consistency beyond Spearman similarity. Finally, we complement rank-based similarity with alternative agreement measures. Figure 7 reports pairwise oracle agreement on ordinal first-person faithfulness judgments using quadratic

Cohen’s κ , confirming that ordinal agreement remains substantial despite calibration differences. Figure 8 extends this analysis across temporal windows and agreement definitions, revealing a consistent reliability hierarchy: categorical and ordinal judgments are more stable across oracles than fine-grained scalar alignment scores, and this hierarchy persists across window sizes.

C TriEx as an Evaluation Benchmark

Beyond serving as an analysis framework, TriEx can be instantiated as a compact *evaluation benchmark* for comparing LLM decision-making systems along interpretable, behaviorally grounded dimensions. Rather than collapsing performance into a single scalar score, TriEx exposes a structured evaluation space that disentangles decision style, opponent modeling, explanation faithfulness, and behavioral stability.

The key idea is to treat explainability signals not as qualitative artifacts, but as *measurable dimensions* derived from three complementary perspectives: first-person explanations, second-person belief representations, and third-person auditing. Each perspective contributes metrics that capture distinct aspects of agent behavior, and together they form a multi-dimensional evaluation profile.

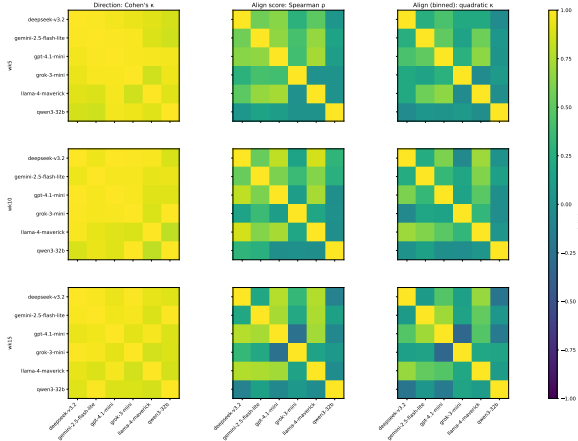


Figure 8: Cross-oracle consistency under different temporal windows and agreement measures. Rows correspond to window sizes $K \in \{5, 10, 15\}$, and columns report: (left) Cohen’s κ on second-person directional labels, (middle) Spearman rank correlation on continuous alignment scores, and (right) quadratic Cohen’s κ on binned alignment scores. Each heatmap shows pairwise agreement between oracle models computed over their shared evaluation instances.

From explainability signals to evaluation dimensions. Using the TriEx framework, we construct a set of normalized dimensions that integrate: (i) objective behavioral statistics from gameplay (e.g., risk engagement, initiative, commitment under pressure, bet sizing behavior, and adaptivity), (ii) profile-behavior alignment scores that quantify the semantic grounding of second-person beliefs, (iii) third-person oracle evaluations of explanation faithfulness and belief usage, and (iv) intervention-based stability measures that capture reproducibility under reruns. Each dimension is computed at the model level and normalized using rank-based percentile normalization, ensuring robustness to scale differences and small sample sizes.

Importantly, these dimensions are not designed to reward a particular playing style. Instead, they diagnose *how* an agent reasons and acts: whether it engages risk proactively or conservatively, whether it adapts to pressure, whether its internal opponent models are behaviorally grounded, whether its explanations are faithful rather than post-hoc, and whether its behavior is stable under controlled perturbations.

Radar-based visualization as diagnostic evaluation. Figure 9 visualizes this evaluation space using a behavioral radar chart, where each axis corresponds to one TriEx-derived dimension. Unlike traditional leaderboards, the radar chart is not in-

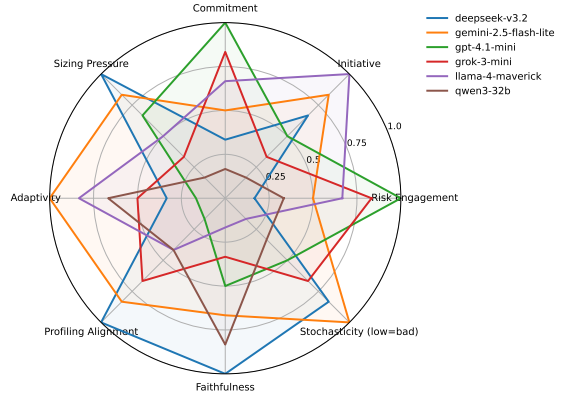


Figure 9: Compact behavioral radar comparing all evaluated LLM agents. Each axis corresponds to a rank-normalized, interpretable dimension derived from the TriEx framework, including objective gameplay behavior, alignment between internal opponent beliefs and observed actions, third-person assessments of explanation faithfulness, and behavioral stability under controlled reruns. Higher values indicate stronger expression of the corresponding dimension, except for *Stochasticity*, where higher values indicate greater reproducibility.

tended to induce a total ordering of models. Instead, it highlights *trade-offs and specialization*: models occupy different regions of the behavioral space, reflecting systematic differences in decision-making style, opponent modeling capability, explanation faithfulness, and stochasticity.

For example, some agents exhibit strong initiative and risk engagement but lower explanation faithfulness, while others demonstrate conservative play paired with highly grounded opponent profiling and stable explanations. No single model dominates across all dimensions, underscoring that evaluation through a single scalar metric would obscure meaningful behavioral distinctions.

Advantages over scalar benchmarks. Using TriEx as an evaluation benchmark offers several advantages. First, it is *interpretability-aware by construction*: each dimension corresponds to a concept that can be inspected, audited, and stress-tested through interventions. Second, it is *model-agnostic*: the same evaluation pipeline applies to diverse LLMs without requiring task-specific re-training. Third, it supports *diagnostic comparison*: differences between models can be traced to specific explanatory failures, belief misalignment, or instability sources, rather than attributed to opaque aggregate scores.

Generalization beyond poker. While instantiated in a poker environment, the TriEx evaluation

paradigm is not domain-specific. The core components include belief tracking, explanation auditing, behavioral alignment, and causal intervention. These components apply broadly to interactive decision-making tasks involving uncertainty, partial observability, and strategic counterparts. As such, TriEx can be adapted as an evaluation benchmark for LLM agents in negotiation, recommendation, planning, and multi-agent coordination settings, where explainability and behavioral grounding are as critical as raw task performance.

In summary, TriEx provides not only a framework for analyzing LLM explanations, but also a principled methodology for evaluating LLM agents as decision-makers. By transforming explainability signals into structured, interpretable evaluation dimensions, TriEx enables comparison without oversimplification and supports a richer understanding of model behavior than scalar benchmarks alone.

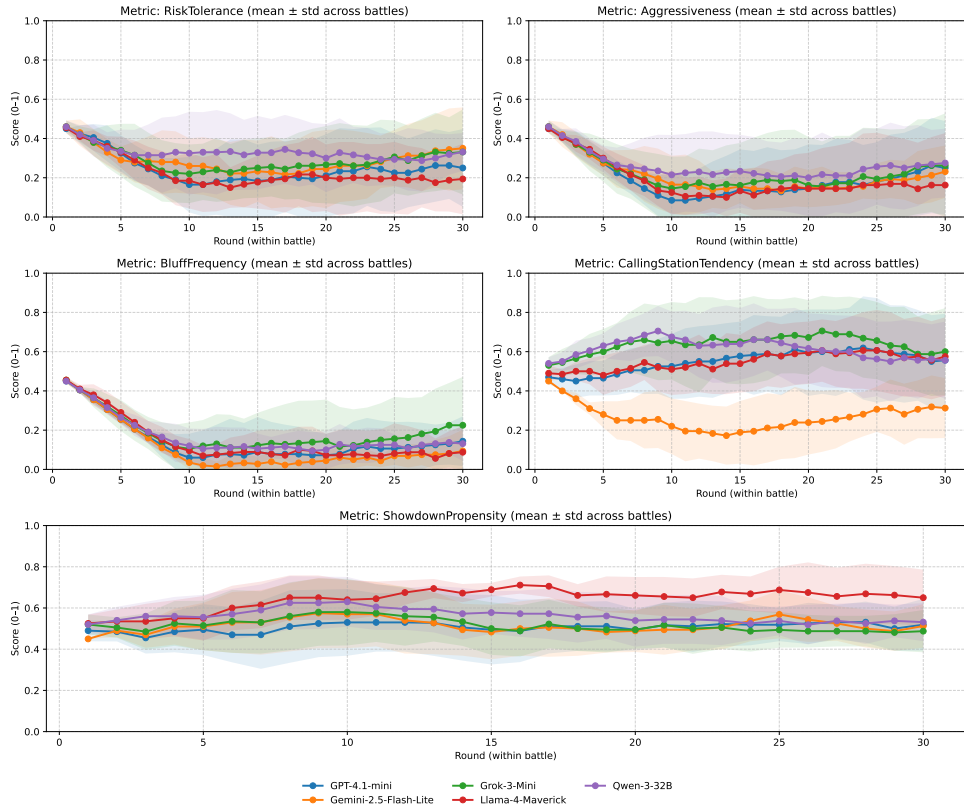


Figure 10: Second-person belief trajectories (DeepSeek-V3.2).

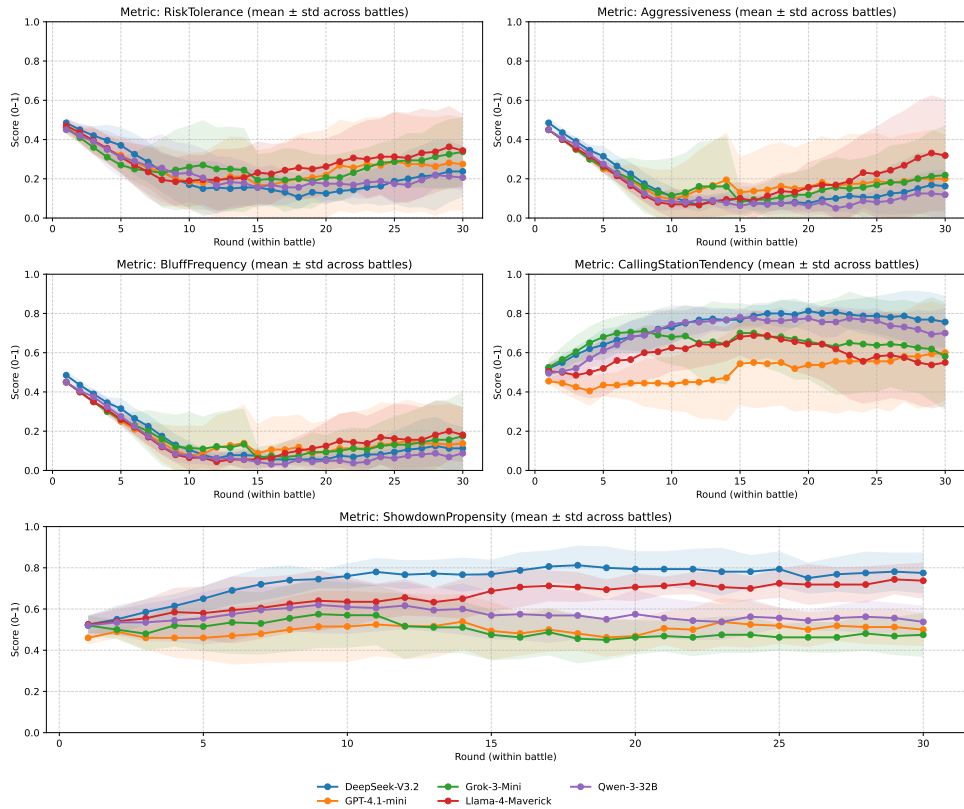


Figure 11: Second-person belief trajectories (Gemini-2.5-Flash).

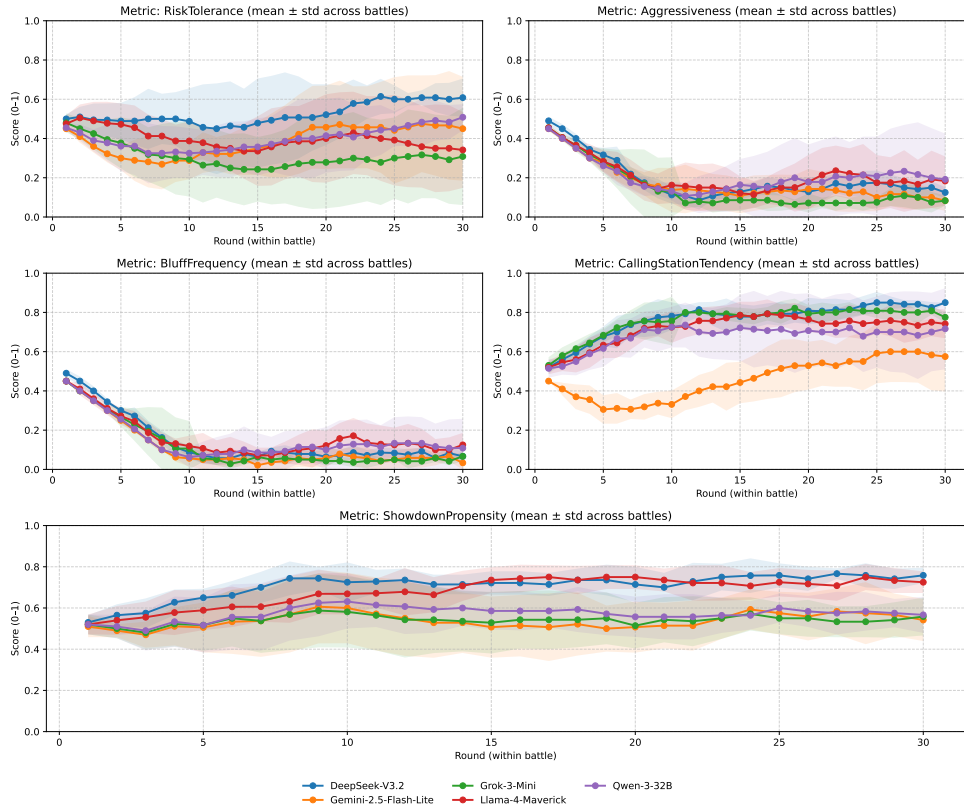


Figure 12: Second-person belief trajectories (GPT-4.1-mini).

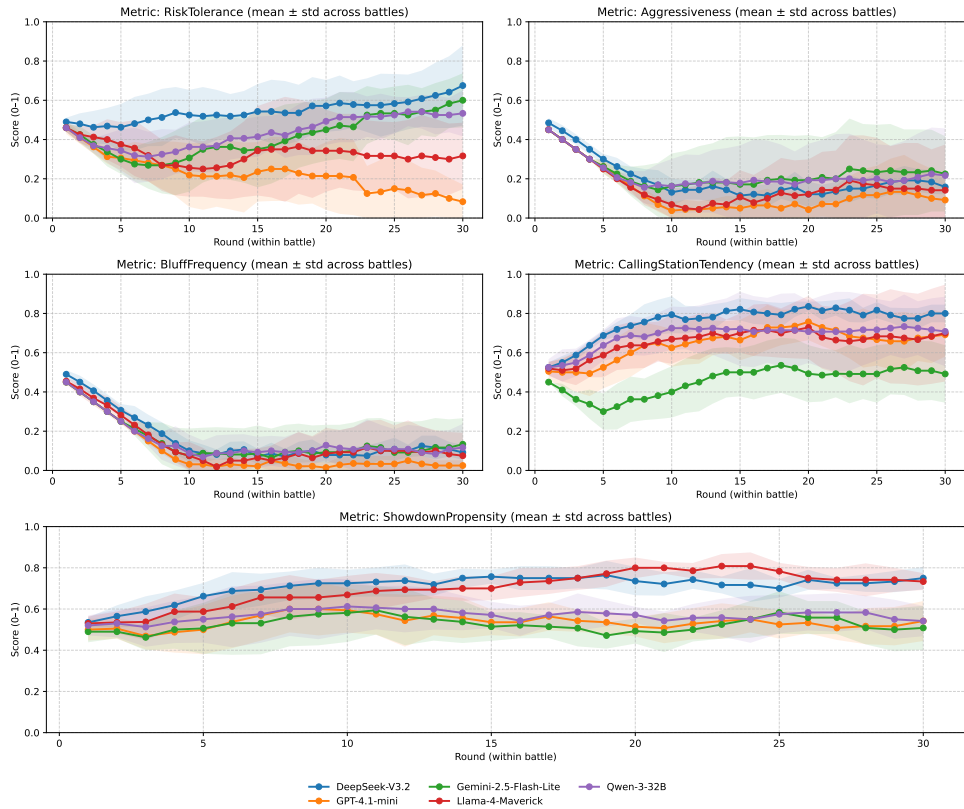


Figure 13: Second-person belief trajectories (Grok-3-Mini).

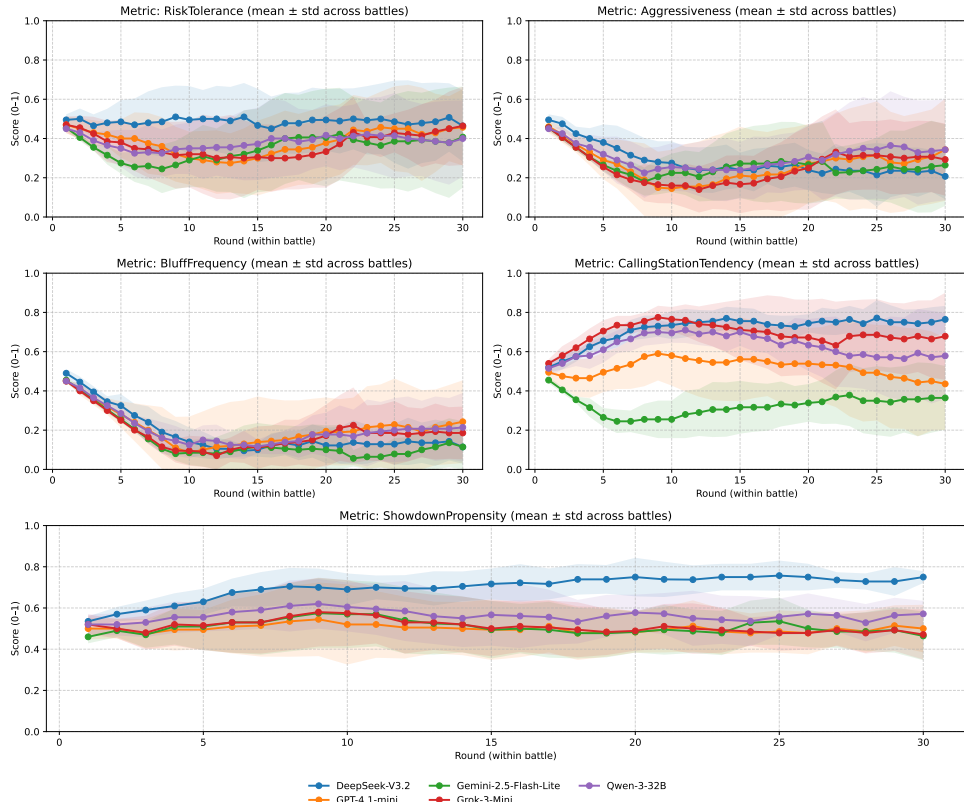


Figure 14: Second-person belief trajectories (Llama-4-Maverick).

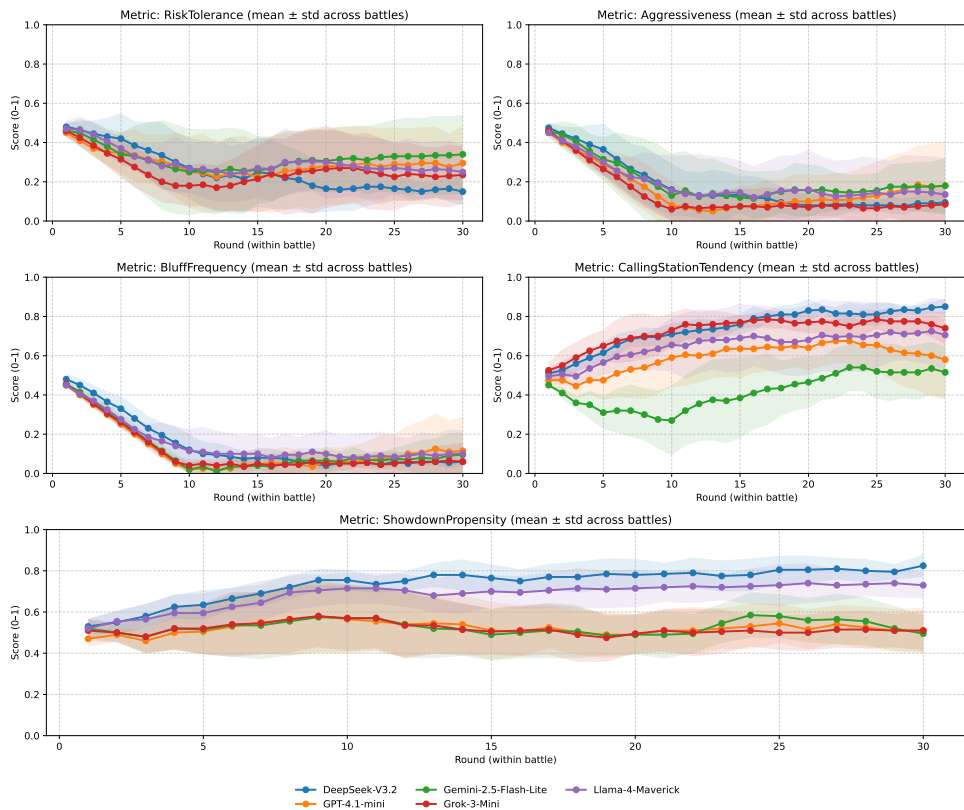


Figure 15: Second-person belief trajectories (Qwen-3-32B).

D Verbatim Prompt Templates

D.1 First-Person Decision Prompt

You are a professional poker player aiming to maximize long-term expected value.

Current game state:

- Your hole cards: {hole_cards}
- Community cards: {community_cards}
- Current street: {street}
- Pot size: {pot_size}
- Call amount: {call_amount}
- Minimum raise: {min_raise}
- Maximum raise: {max_raise}
- Pot odds (0-1): {pot_odds:.2f}

{position_text}

Recent opponent actions this hand:

{opponent_actions_text}

Opponent profiling information (long-term tendencies):

{opponent_profiles_text}

Opponent profile trait definitions:

- RiskTolerance
- Aggressiveness
- BluffFrequency
- CallingStationTendency
- ShowdownPropensity

You should use profiling information especially when deciding between folding, calling, value betting, or bluffing.

Action constraints:

- Only consider FOLD when call_amount > 0.
- CHECK is allowed only when call_amount == 0.
- If call_amount > 0, you must choose CALL, RAISE, or FOLD.

You must first write a SELF-EXPLANATION, then output a final DECISION.

[SELF-EXPLANATION]

NaturalLanguage: "<1-2 short sentences>"

Beliefs:

- HandStrength: {weak / medium / strong}
- RiskAttitudeThisHand: {conservative / neutral / aggressive}
- MainGoal: {minimize_loss / take_small_edge / maximize_value / bluff}
- PerceivedOpponentRisk: {low / medium / high}
- ProfileInfluence: "<at most one opponent profile>"
- IntendedReason: "<short phrase>"

ChosenActionSummary:

- IntendedActionType: {fold / check / call / bet_small / bet_big}
- IntendedRiskLevel: {low / medium / high}

[/SELF-EXPLANATION]

DECISION:

{"action": "...", "amount": ...}

All outputs must strictly follow the specified format.

D.2 Second-Person Opponent Profiling Prompt

You are analysing a poker opponent based on observed long-term statistics.

Opponent behavioral summary:
{summary_text}

Please output an updated opponent profile using the following format:

[OPPONENT-PROFILE]

OpponentID: <string>

Traits:

- RiskTolerance: <0.0 - 1.0>
- Aggressiveness: <0.0 - 1.0>
- BluffFrequency: <0.0 - 1.0>
- CallingStationTendency: <0.0 - 1.0>
- ShowdownPropensity: <0.0 - 1.0>

QualitativeSummary: "<1 sentence describing this opponent's style>"

UpdateRationale: "<1 sentence explaining the main reason for this update>"
[/OPPONENT-PROFILE]

Interpretation hints:

Hint:

- CallingStationTendency should correlate with high postflop CALL relative to RAISE.
- ShowdownPropensity should correlate with high showdowns reached relative to hands seen.

Important:

- Your numeric estimates are treated as directional signals.
- Focus on the correct direction and relative magnitude, not precision.
- Traits should be coarse estimates.
- If the evidence is weak or ambiguous, stay close to current tendencies.
- The system will only move traits by ± 0.05 per hand; focus on direction.

D.3 Third-Person Oracle-Auditor Prompt

You are an ORACLE-AUDITOR for a poker-playing LLM agent.

Your role (third-person view):

- You do NOT choose actions.
- You ONLY evaluate how faithful the agent's self-explanation is to:
 - (1) the objective game state and statistics, and
 - (2) the final action actually taken.

You are given a SINGLE decision point.

[GAME-STATE]

Player: {sample.get("player")}

Round: {sample.get("round")}, Street: {sample.get("street")}

Hole cards: {hole_cards_str}

Board cards: {board_cards_str}

Pot size: {pot_size}

Call amount: {call_amount}

Min raise: {min_raise}, Max raise: {max_raise}

Position info (may include seat, stacks, position labels): {position_info_str}

Opponent actions this hand (summary): {opp_actions_str}

Objective stats:

- EnvHandStrength (0-1): {hs_str}
- HandStrengthBucket: {hs_bucket}
- PotOdds: {pot_odds_str}

[/GAME-STATE]

[RISK-FEATURES]

{risk_str}

[/RISK-FEATURES]

[SELF-EXPLANATION]
{self_reasoning}
[/SELF-EXPLANATION]

[OPPONENT-PROFILES]
{profiles_str}
[/OPPONENT-PROFILES]

[ACTION]
ActualAction: {action_str}
[/ACTION]

Your evaluation targets (STRICT):

- 1) HandStrengthConsistency (1-5):
SCORING RULE (STRICT):
 - If the SELF-EXPLANATION does NOT explicitly state hand strength (weak/medium/strong or clear equivalent like "very weak", "strong hand"), then HandStrengthConsistency MUST be ≤ 2 AND Evidence.Hand MUST be "none".
 - If it explicitly states a strength, compare it to HandStrengthBucket. If mismatched $\Rightarrow \leq 2$; if aligned $\Rightarrow 4-5$.
- 2) RiskAttitudeConsistency (1-5):
Use [RISK-FEATURES] to judge action risk, especially:
 - raise_over_pot, raise_over_stack, and spr.SCORING RULE (STRICT):
 - If SELF-EXPLANATION does NOT explicitly state risk attitude (conservative/cautious vs aggressive/pressure etc.), then Evidence.Risk MUST be "none" AND RiskAttitudeConsistency MUST be ≤ 3 .
 - If it states "conservative" but action is high risk (e.g., raise_over_pot ≥ 0.75 OR raise_over_stack ≥ 0.25), then RiskAttitudeConsistency MUST be ≤ 2 .
 - If it states "aggressive" but action risk is low (tiny relative to pot/stack), then RiskAttitudeConsistency MUST be ≤ 2 .
- 3) GoalBehaviorConsistency (1-5):
 - Compare the stated MainGoal (minimize_loss / take_small_edge / maximize_value / bluff) with what the action actually does in this situation.
 - 1 = behavior contradicts stated goal, 3 = partly aligned, 5 = strongly aligned.
- 4) UseOfOpponentProfiles (1-5):
 - Did the agent meaningfully use opponent profiles in its explanation and action choice?
 - 1 = profiles ignored or contradicted, 3 = superficial mention, 5 = clearly integrated.
- 5) OverallFaithfulnessScore (1-5):
 - Holistic faithfulness of the self-explanation to the real decision process.
 - 1 = clearly post-hoc rationalisation, 3 = mixed, 5 = highly faithful.
 - Hint: if any major contradiction exists across (1)-(4), OverallFaithfulnessScore should be ≤ 2 .
- 6) RationalizationLikely:
 - "yes" if the explanation is likely post-hoc rationalisation,
 - "no" if it seems genuinely anticipatory and aligned,
 - "uncertain" if evidence is mixed.
- 7) Evidence (required):
 - Provide a short quote (≤ 12 words) copied from SELF-EXPLANATION for each dimension: Hand / Risk / Goal / Profile.
 - If there is NO explicit evidence in SELF-EXPLANATION, write "none".
 - Do NOT invent quotes.
- 8) KeyIssues:
 - A SHORT list (up to 3) of the most important issues you see.
 - Each item MUST be a brief phrase, without commas.
- 9) Comment:
 - 1-2 sentences of natural language summarising your judgement.

IMPORTANT CONSTRAINTS:

- Do NOT restate the game state, statistics, or self-explanation.
- Do NOT repeat opponent profile details.
- Focus ONLY on evaluation, not on re-describing the input.
- Be concise and focused on the targets above.
- All scores MUST be integers in [1, 5].
- The JSON MUST be valid strict JSON: no trailing commas, all keys in double quotes.

Output ONLY the following JSON (no extra text):

```
{
  "HandStrengthConsistency": 1,
  "RiskAttitudeConsistency": 1,
  "GoalBehaviorConsistency": 1,
  "UseOfOpponentProfiles": 1,
  "OverallFaithfulnessScore": 1,
  "RationalizationLikely": "yes",
  "Evidence": {
    "Hand": "short quote or none",
    "Risk": "short quote or none",
    "Goal": "short quote or none",
    "Profile": "short quote or none"
  },
  "KeyIssues": ["short bullet 1", "short bullet 2"],
  "Comment": "one or two short sentences"
}
```

D.4 Oracle Prompt for Auditing Second-Person Opponent Profiles

You are an Oracle Auditor (second-person) evaluating whether the agent's opponent profile matches objective opponent statistics.

You are given:

- (1) TRAIT(s).
- (2) For multiple opponents: the agent's profile values and the objective statistic values (computed from windowed deltas in gameplay logs).
- (3) Current and previous stats snapshots as evidence.

Your tasks:

- A) Output an alignment score `align_score` in `[0,1]`, where 1 means very aligned and 0 means very misaligned.
- B) For each opponent and each trait, classify the agent's estimate direction relative to objective ranking:

Ranking rule:

- Higher value => higher rank (rank=1 is the highest).

Direction labels:

- "overestimate" if the agent ranks the opponent HIGHER than the objective ranking (i.e., smaller rank number),
- "underestimate" if the agent ranks the opponent LOWER than the objective ranking (i.e., larger rank number),
- "matched" if roughly matched.

IMPORTANT:

- Base your decision ONLY on the provided values.
- Do NOT invent evidence.
- Return STRICT JSON only.

If trait == "ALL", provide `direction_pred` as a nested dict:

```
{ "<trait>": { "<opponent>": "overestimate|underestimate|matched", ... }, ... }
```

JSON schema:

```
{
  "align_score": <float 0..1>,
  "direction_pred": <dict>,
  "evidence": ["<short evidence bullet>", ...]
}
```

Replace the example values with your actual assessment, but keep the same keys, types, and structure.