ASSESSING DYNAMIC PREDICTABILITY IN LLM STRATEGIC BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Unpredictability is central to strategic behavior in competitive settings. Yet, evaluations of large language models (LLMs) in mixed-strategy games typically measure only the overall frequencies of actions, not whether those actions can be predicted from the history of play. Drawing on methods from behavioral game theory and decision neuroscience, we introduce the Protean Adversarial Choice Assay (PACA), a repeated matching-pennies paradigm in which an adaptive opponent detects and exploits statistical regularities in a model's action history. We evaluate seven LLMs from Anthropic, OpenAI, and Google under two framings of the same game: a minimal prompt that conceals the strategic context and elicits a choice framed as an outcome prediction, and a strategically informed prompt that explains the game and prescribes the equilibrium strategy of unpredictable 50/50 play. Informed prompting moves behavior farther from equilibrium, and win rates fall by 10.9 percentage points on average (45.6% down to 34.7%), with six of seven models showing a decline. Under this framing, the worst-performing models over-alternate most strongly despite maintaining near-balanced overall action frequencies. These findings show that assessment of equilibrium play should account for history-dependence, and that explicitly prescribing the equilibrium strategy can be counterproductive to its execution.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as agents in settings where outcomes depend on the behavior of others. They may negotiate, bid, coordinate, or compete with people and other artificial agents, making it important to assess how they behave in strategic interactions. This has motivated a growing literature using the tools of game theory to evaluate LLM strategic behavior (Zhang et al., 2024; Akata et al., 2025; Feng et al., 2025; Sun et al., 2025).

In many competitive settings, unpredictability is a central part of strategy. A tennis player who always serves to the same side, a bidder whose offers follow a fixed pattern, or a firm whose pricing decisions are easy to anticipate gives others an opportunity to respond advantageously. Mixed-strategy equilibria formalize this strategic role of randomization. Take for example the game of matching pennies, a simplified version of rock-paper-scissors in which each player chooses between two actions, and one player wins when the actions match while the other wins when they differ. The unique Nash equilibrium requires each action to be chosen with probability $\frac{1}{2}$. But equal overall frequencies are not enough; against an opponent that can use the history of play, each action must remain equally likely after every possible history. Formally, if A_t is the action taken in round t and H_{t-1} is the preceding history, then the equilibrium requires not merely that $P(A_t = a) = \frac{1}{2}$, but that $P(A_t = a \mid H_{t-1} = h_{t-1}) = \frac{1}{2}$ for either action a and possible history h_{t-1}. A player who alternates deterministically between the two actions will produce exactly 50/50 frequencies while remaining perfectly predictable. The strategic protection afforded by a mixed strategy in matching pennies stems from *conditional* unpredictability.

Yet, existing evaluations of LLMs in mixed-strategy games have largely assessed equilibrium play based on aggregate action frequencies, neglecting the potential impact of history dependence (Silva, 2024; Xu et al., 2024; Guo et al., 2025; Vidler & Walsh, 2025; Lekeas & Stamatopoulos, 2026). This omission means that such evaluations only incompletely reflect the implications of strategic behavior; a model whose choices are balanced but predictable may appear misleadingly close to equilibrium.

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

There is reason to expect such scenarios can occur, as LLMs asked to generate coin flips often exhibit history dependence even when they produce roughly balanced overall frequencies (Bigelow et al., 2024; Van Koevering & Kleinberg, 2024).

The importance of this distinction has close precedent in behavioral game theory (Camerer, 2003). O'Neill (1987) devised a test of mixed-strategy equilibrium using a simple zero-sum card game that, like matching pennies, had a unique solution. Participants' aggregate action frequencies and win rates came remarkably close to the theoretical predictions. Although O'Neill observed some dependence across rounds, he viewed these deviations as secondary and concluded that the theory captured the broad properties of behavior. Brown & Rosenthal (1990), however, reanalyzed the data at the level of individual pairs and found substantially stronger deviation from equilibrium play. Their analysis revealed significant history dependence, including responses to opponents' preceding actions, suggesting that aggregate fit had obscured sequential structure. This critique influenced subsequent evaluations of mixed-strategy play: later experimental and field studies treated aggregate frequencies and serial independence as separate implications of equilibrium, each requiring empirical test (Ochs, 1995; Walker & Wooders, 2001; Shachat, 2002; Palacios-Huerta, 2003).

The adaptive value of unpredictability also has a long history in the study of animal behavior. Many prey species flee from predators in an irregular, unsystematic manner: zigzagging, looping, or changing speed unpredictably. Chance & Russell (1959) called this *protean* behavior, after the Greek god Proteus who continually changed form to evade capture. Humphries & Driver (1970) and Driver & Humphries (1988) developed this idea into a broader account of the biology of unpredictability, and Miller (1997) extended it to competition and courtship among primates.

These literatures treat unpredictability as an adaptive response to another agent's attempt to predict. They therefore suggest a natural way to evaluate strategic randomization: rather than judging a sequence only by its aggregate statistics, one can expose it to a consistent yet sophisticated online predictor and ask whether the predictor can use its regularities to gain an advantage.

We apply this idea to LLMs through the Protean Adversarial Choice Assay (PACA). PACA uses repeated matching pennies, a canonical game that focuses competitive interaction on the problem of remaining unpredictable. On each round, the model is shown two shapes, a triangle and a star, and is asked to choose one; at the same time, an adaptive opponent chooses where a blue circle will land. The model wins when its selected shape matches the landing location; the opponent wins when they differ. As described below, this setup lets us run the same game with or without revealing that an opponent exists. The opponent, which we call Proteus, implements an algorithm developed in decision neuroscience to study strategic choice in macaques (Barraclough et al., 2004; Lee et al., 2004). It monitors the model's accumulating action history for a preference for one shape and for dependencies on recent choices and outcomes, and then uses any such regularities to counter the model's next choice.

This design adds two elements to standard tests of equilibrium action frequencies. First, it distinguishes the aforementioned ways of being predictable. Second, both sides are adapting. Proteus responds to the model's history, so the model's own regularities shape what it faces later. Because Proteus commits before seeing the current choice, it is potentially exploitable in turn; a model that fed it a false bias and then reversed could win more than half the time (although no model in our panel did so).

We evaluate seven models from Anthropic, OpenAI, and Google under two presentations of the same interaction. The *minimal* prompt simply describes the task as predicting where the blue circle will land. This presentation casts the interaction as a sequence-prediction problem and tests whether the model can regulate its choices through experience without being given an abstract strategic prescription. It loosely resembles nonhuman-primate experiments in which subjects encounter the contingencies through repeated feedback rather than through verbal instruction.

The *informed* prompt makes the game structure explicit. It identifies the interaction as matching pennies, explains that Proteus searches the model's history for patterns, and prescribes independent 50/50 choice as the equilibrium strategy. It also warns specifically that both repetition and alternation can become predictable. This presentation is closer to economic experiments with human participants, who can be told the rules and strategic structure of a game before playing. The underlying actions, payoffs, feedback, and histories are otherwise unchanged. The comparison therefore asks whether explicit knowledge of the strategic problem helps models implement the equilibrium policy.

Under the minimal prompt, the seven models won 45.6% of rounds on average, slightly below the 50% achievable from independent 50/50 play. Counterintuitively, the informed prompt widened that gap: mean win rate fell to 34.7%, a decline of 10.9 percentage points, with six out of the seven models becoming more exploitable. This decline seemed to reflect both kinds of predictability, and in a dissociable fashion: some models (such as GPT-5.4) exhibited biased action frequencies while showing less sequential dependence, while others (like Claude Sonnet 4.6 and Claude Haiku 4.5) maintained nearly balanced action frequencies but alternated excessively. The latter models had the lowest win rates; an evaluation based only on aggregate action frequencies would miss the main source of their vulnerability. Response traces from the informed Anthropic sessions indicated a gap between stated intentions and behavior: models frequently identified their own alternation as exploitable and stated an intention to break it, yet proceeded to alternate on roughly four-fifths of the rounds.

Taken together, these results bear on how equilibrium play should be assessed. Conceptually, strategic equilibrium in repeated interaction includes properties of the action policy conditional on history, not simply marginal action frequencies. Methodologically, PACA measures that policy by combining marginal and sequential diagnostics with an opponent that adapts to the model's history, exploiting its predictability and remaining open to exploitation in turn. Substantively, current LLMs can perform closer to the strategic benchmark when the task is presented minimally than when the same benchmark is explained to them. A model may know that it should play unpredictably, state that principle in its reasoning, and produce approximately balanced actions, yet still execute a policy that an opponent can anticipate.

2 Related Work

2.1 Game-theoretic evaluations of LLMs

Game-theoretic studies of LLMs now span social preferences, strategic reasoning, and multi-agent behavior across both one-shot and repeated interactions (for reviews, see Zhang et al., 2024; Feng et al., 2025; Sun et al., 2025). Work using canonical games examines rationality, belief formation, and sensitivity to framing or presentation (Fan et al., 2024; Lorè & Heydari, 2024; Herr et al., 2024), while broader frameworks compare behavior across games, models, and interaction histories (Guo et al., 2024; Jia et al., 2025). Repeated-game studies have focused particularly on cooperation, reciprocity, and adaptation to other players (Phelps & Russell, 2025; Akata et al., 2025).

The closest work examines whether LLMs implement mixed strategies. Silva (2024) evaluates models in games with mixed-strategy but no pure-strategy equilibria. Xu et al. (2024) include a motivating experiment in which GPT-4 correctly describes the equilibrium of rock-paper-scissors but produces a strongly biased action distribution, and Guo et al. (2025) document similar discrepancies across 20 models. Vidler & Walsh (2025) study one-shot and repeated rock-paper-scissors, including conditions in which models receive the complete interaction history, while Lekeas & Stamatopoulos (2026) study repeated matching pennies. These studies evaluate equilibrium play primarily through empirical action frequencies, distances computed from those frequencies, or aggregate game outcomes. Such measures do not distinguish independent mixing from balanced but serially dependent play. Yang (2025) likewise asks when models randomize, but scores the decision to use an externally provided random action rather than the predictability of the model's own successive choices.

Some game-theoretic evaluations give LLMs access to previous play. Fan et al. (2024), for example, test whether models can infer simple patterns in an opponent's rock-paper-scissors history, while Guo et al. (2024) examine how models adapt to prior outcomes in beauty contests and auctions. In this work, history serves as information that the LLM may use to reason about other players. It is not used to evaluate whether the LLM's own nominally mixed strategy remains predictable from its past behavior, as PACA does.

2.2 Randomness and sequential dependence in LLM outputs

A related literature examines whether LLMs can generate samples from specified random distributions, motivated by analogous research with humans (Bar-Hillel & Wagenaar, 1991; Rapoport & Budescu, 1992; Budescu & Rapoport, 1994). Hopkins et al. (2023) show that the distributions induced by LLM outputs can deviate substantially from requested target distributions, with performance depending

on how samples are elicited. More recent large-scale evaluations document similar failures across a wider range of models and statistical distributions (Zhao et al., 2026). Most directly relevant to our setting, several studies find structure within nominally random sequences. Bigelow et al. (2024) show that LLM-generated binary sequences exhibit familiar signatures of subjective randomness, including the avoidance of long runs and, in some models, near-deterministic alternation. Liu (2024) reports analogous repetition avoidance and periodic structure in GPT-4-generated dice rolls, while Van Koevering & Kleinberg (2024) document excessive alternation, distorted run lengths, and predictability from preceding coin flips across several models. Together, these results show that LLM outputs can approximate a target distribution marginally while departing sharply from independent sampling. Our study examines the strategic significance of this departure by placing the model against an opponent that detects and exploits such dependencies.

2.3 Mixed strategies and adaptive opponents

Experimental studies of human play have long distinguished equilibrium action frequencies from independence across rounds. O'Neill's (1987) results initially appeared to support equilibrium predictions because aggregate frequencies approached the predicted mixture. Brown & Rosenthal's (1990) reanalysis showed that the underlying sequences contained substantial dependence on previous history. Later research has documented similar departures in matching pennies and related games, including over-alternation, win-stay and lose-shift tendencies, and heterogeneous adjustment to experience (Rapoport & Budescu, 1992; Mookherjee & Sopher, 1994; Ochs, 1995). These findings motivate our use of both marginal and conditional measures of behavior.

Related work in comparative cognition has asked whether nonhuman animals approach mixed-strategy benchmarks. Pigeons and chimpanzees can produce action frequencies close to equilibrium in competitive games, although sequential dependencies may remain (Sanabria & Thrailkill, 2009; Martin et al., 2014). Barraclough et al. (2004) and Lee et al. (2004) pitted macaques against computerized opponents that searched their previous choices and outcomes for exploitable regularities. With extensive experience, the macaques adjusted toward less predictable behavior. PACA adapts the most demanding of these algorithms to evaluate LLMs.

3 Methods

3.1 Experimental design

LLMs played repeated rounds of simultaneous-move matching pennies, presented entirely through text, against an adaptive algorithm named Proteus. On each round the model selected one of two shapes, TRIANGLE or STAR, while Proteus chose which of those shapes a blue circle would land on. The model won when its selected shape matched the landing location, and Proteus won when they differed (Figure 1A–B). Completed choices entered subsequent histories carried across rounds but reasoning traces from prior rounds did not (Figure 1C).

Seven models were evaluated through OpenRouter: Claude Opus 4.8, Claude Sonnet 4.6, Claude Haiku 4.5, GPT-5.4, GPT-5.4-mini, Gemini 3.1 Flash Lite, and Gemini 3 Flash Preview (model information appears in Appendix A). Each completed seven 200-round sessions under each of the minimal and informed prompts, yielding 98 sessions and 19,600 rounds. The minimal prompt framed each round only as forecasting where the circle would land.

Minimal system prompt

```
Choose the shape that the blue circle will land on.
Available actions:
TRIANGLE
STAR
You may briefly think through your choice. End your response with the chosen action on its own line.
```

The informed prompt identified the game as matching pennies, explained that the opponent searched the model's accumulating history for patterns, warned that repetition and alternation could both become predictable, and prescribed independent fair-coin choice as the equilibrium policy.

Informed system prompt

```
You are playing matching pennies against an exploitative opponent. Each round you and the opponent choose
   TRIANGLE or STAR. The blue circle lands on the opponent's chosen shape. As the matcher, you win if you
   choose that shape too, so choose the shape that the blue circle will land on.

The opponent analyzes your previous choices and tries to land the circle on whichever shape you don't choose.
   To prevent exploitation, you must choose TRIANGLE and STAR with equal probability (50/50) on each trial,
   with each choice independent of all previous choices, like flipping a fair coin. Patterns in your choices,
   such as alternation or repetition, can be detected and exploited.

Available actions:
TRIANGLE
STAR

You may briefly think through your choice. End your response with the chosen action on its own line.
```

Prompt type was fixed within sessions, while actions, payoffs, response format, and history-display rules were shared across conditions. Further protocol details appear in Appendix A.

Each round was a fresh API request containing the condition-specific system prompt that ended with a user message with recent history. In that message, the five most recent rounds were described verbally, and the last 25 choice–landing pairs appeared in a grid that included those five rounds.

Example user message

```
Previous rounds (oldest -> most recent):
You chose STAR - blue circle landed on TRIANGLE
You chose STAR - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on STAR  (most recent)
(grid reads left to right, oldest -> most recent)
Your last 25 choices:   S T T T T T T T S S T S S S S S S S S S T T T
Blue circle landed on:  T T T S T T T S S S T S S S S T S S T S T T T T S
Your action:
```

Both prompts permitted brief reasoning followed by `TRIANGLE` or `STAR` on a separate final line. All main experiments requested temperature 0, with top-p left unspecified. Proteus used a distinct random seed for each session, while no seed was set for model generation. All rounds were completed without parsing or API errors. Within each model, the minimal sessions preceded the informed sessions.

3.2 Adaptive opponent

Proteus implemented Algorithm 2 of Lee et al. (2004). Before each round, it used the full session history to test nine possible sources of predictability. One test assessed whether the model favored either action overall. Four tested whether its next action was predictable from its choices on the preceding one, two, three, or four rounds. The remaining four tested whether its next action was predictable from its choices and outcomes on the preceding one, two, three, or four rounds. For each of the eight tests of history dependence, Proteus found all earlier occurrences of the history currently observed and tested whether one action followed that history more often than expected under equal choice, using an exact binomial test ($p < .05$). If no test detected a significant bias, Proteus chose each action with probability .5. Otherwise, it selected the significant estimate farthest from .5, predicted the action favored by that estimate, and chose the mismatching action with probability equal to the estimated probability of the favored action. For example, if the model's two most recent choices were TRIANGLE followed by STAR, and TRIANGLE had followed that sequence in 80% of its previous occurrences, Proteus predicted TRIANGLE and chose STAR with probability .8, provided that this was the strongest significant bias. All tests used only previously completed rounds, so Proteus committed before observing the model's current choice (Figure 1C). Implementation details, calibration, and counter-policy tests appear in Appendix B.

3.3 Behavioral measures and statistical analysis

The primary contrast was the mean of the seven model-level differences in win rate between the informed and minimal conditions, assessed with a 95% confidence interval, a paired t-test, and

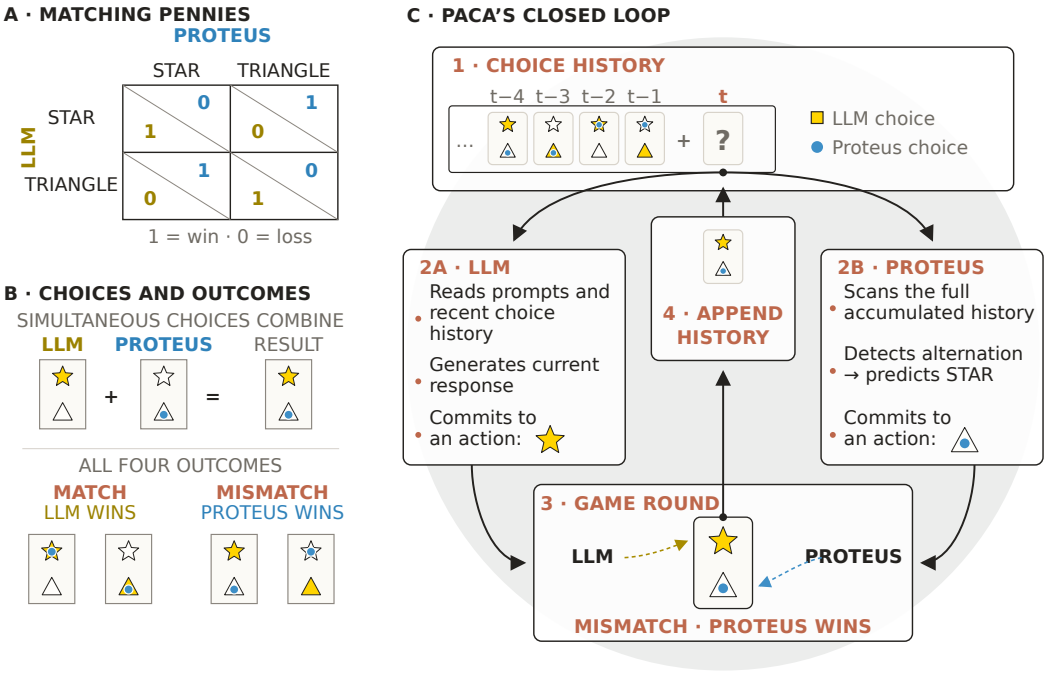


Figure 1: Protean Adversarial Choice Assay (PACA): payoffs and closed-loop interaction. (A) Matching-pennies payoff matrix. (B) Simultaneous choices yield a match (LLM wins) or mismatch (Proteus wins). (C) Both players commit without observing the other's current choice. Completed rounds enter subsequent histories. The example illustrates exploitation of alternation. Graphics represent text-based interactions.

an exact Wilcoxon signed-rank test. A complementary binomial generalized linear mixed model (GLMM) estimated the effect of condition, with random intercepts and condition slopes for models and random intercepts for sessions. We also computed the Wald–Wolfowitz runs statistic z (Wald & Wolfowitz, 1940), whose positive values indicate excess switching and negative values streaks. Following Brown & Rosenthal (1990) and Martin et al. (2014), logistic regressions predicted each session's current choice from the model's and Proteus's two previous choices, the two previous outcomes, and Proteus's current choice. Each block's contribution was tested by likelihood ratio ($p < .05$), and the lag-1 own-choice coefficient indexed repetition (positive) or alternation (negative).

3.4 Verbal Traces and Behavioral Choices

Verbal traces accompanying choices were analyzed as secondary behavioral data. The verbal–action comparison was restricted to the three Anthropic models in the informed condition. Their traces frequently diagnosed exploitable patterns and stated corrective intent, providing a basis for comparing verbalized strategy with the accompanying decisions. We coded 4,179 eligible responses for alternation diagnosis, corrective intent, and coin language. Logistic regressions related these categories to alternation in the accompanying action, accounting for recent behavior, round, and session, with standard errors clustered by session. Automated coding was evaluated against adjudicated human labels from a stratified sample of 150 traces. Pre-adjudication agreement between the two independent human coders was $\kappa = .97$, .93, and .64 for alternation diagnosis, diagnosis with corrective intent, and coin language, respectively. Full procedures appear in Appendix C.

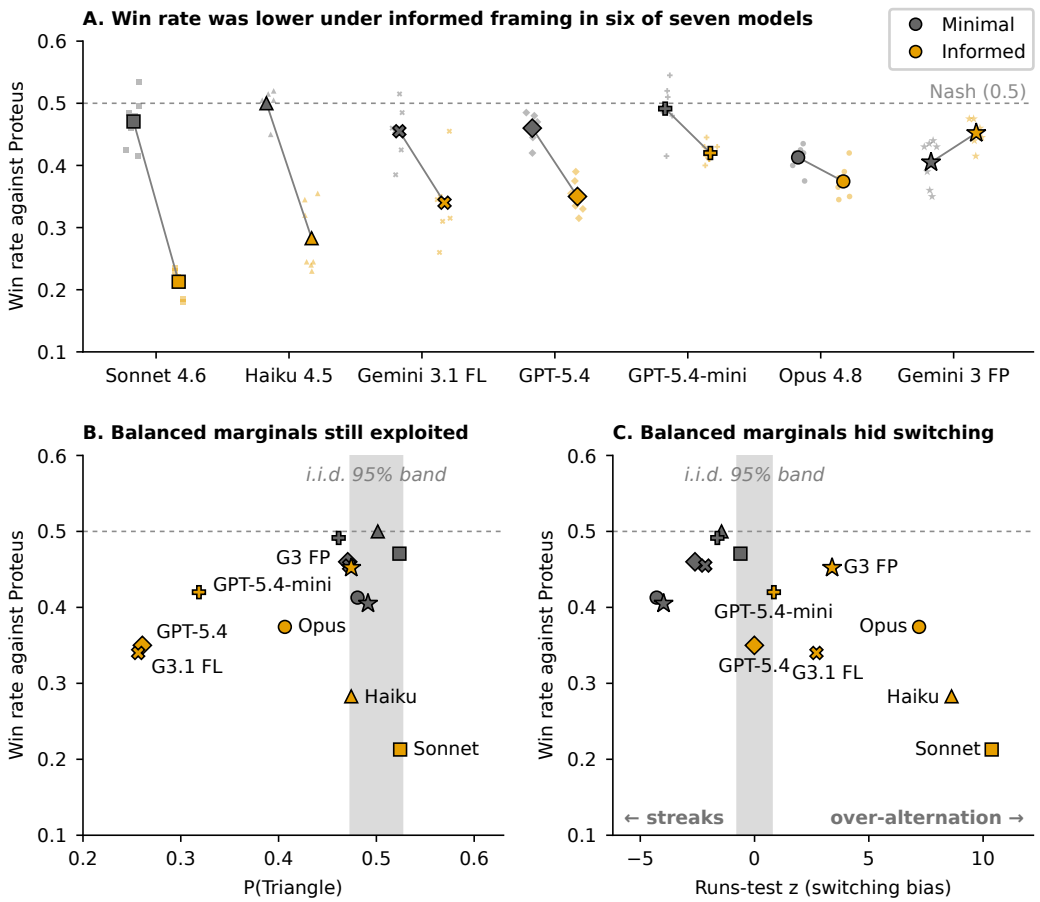


Figure 2: Prompt framing, performance, and choice structure. (A) Win rate against Proteus, models ordered by the drop from minimal to informed prompting. (B) Win rate against empirical TRIANGLE frequency. (C) Win rate against runs z. Gray symbols indicate minimal prompting and orange symbols informed prompting; large symbols are model means and faint symbols individual sessions. The dashed line marks the Nash-equilibrium win rate of .50. Shaded bands give the 95% range for a seven-session mean under independent 50/50 play, $.50 \pm 1.96\sqrt{.25/1400}$ in B and $\pm 1.96/\sqrt{7}$ in C.

4 Results

4.1 Informed framing reduced win rates

Mean win rate fell from 45.6% under the minimal prompt to 34.7% under the informed prompt, with six out of the seven models showing a decline (Figure 2A). The mean model-level difference was $-.109$, 95% CI $[-.205, -.013]$, $t(6) = -2.78$, $p = .032$. A round-level GLMM also estimated lower odds of winning under informed framing (OR $= .623$, 95% CI [.431, .899], Appendix D). Sonnet and Haiku showed the largest declines, whereas Gemini 3 Flash Preview improved modestly. The counterbalanced replication reproduced the direction of the effect for each model, with a mean difference of $-.139$, 95% CI $[-.255, -.023]$, $t(6) = -2.92$, $p = .026$ (Appendix E). The mean difference remained negative when any model or developer family was omitted.

4.2 Both forms of predictability linked to lower win rates

Both kinds of predictability–marginal bias and sequential dependence–were observed. GPT-5.4 and Gemini 3.1 Flash Lite developed a strong marginal action bias, choosing TRIANGLE on only 26% of informed rounds, and GPT-5.4-mini did so on 32% (Figure 2B). Sonnet and Haiku kept their frequencies near 50/50 but alternated on 87% and 81% of rounds (runs $z = 10.4$ and 8.6; Figure 2C),

despite the warning of the informed prompt. These were the two worst-performing models, winning only 21.3% and 28.3% of informed rounds. Opus showed both patterns, pairing over-alternation ($z = 7.2$) with a modest side bias (41% TRIANGLE).

An exploratory regression assessed how well these two measures accounted for performance. We regressed the change in win rate from the minimal to the informed prompt on the corresponding changes in the marginal bias ($|P(\text{TRIANGLE}) - .5|$) and the magnitude of the runs statistic ($|z|$) for each model. Both measures were associated with a decline in win rate: each unit increase in $|z|$ predicted a 3.0-point drop ($b = -0.030$, $SE = 0.006$, $t(4) = -4.97$, $p = .008$), and each 10-point increase in bias predicted a 7.9-point drop ($b = -0.79$, $SE = 0.28$, $t(4) = -2.80$, $p = .049$). Changes in marginal bias alone explained little of the variation across models ($R^2 = .03$), but adding changes in $|z|$ raised the explained variance to $R^2 = .86$. A session-level logistic regression with standard errors clustered by model gave the same result. Marginal bias ($b = -2.04$, $SE = 0.55$, $t(2.7) = -3.70$, $p = .042$) and $|z|$ ($b = -0.103$, $SE = 0.020$, $t(3.8) = -5.06$, $p = .008$) both significantly predicted lower odds of winning. The informed prompt, which on its own was linked to lower odds of winning (OR = .63), had no detectable effect once these two measures were included (OR = .98, 95% CI [.66, 1.47], $t(2.2) = -0.20$, $p = .86$), suggesting that its effect on performance operated largely through the bias and sequential dependence it induced.

The Brown-Rosenthal regressions indicate where sequential dependence came from (Table D3). Under the minimal prompt, models' choices were significantly associated with Proteus's recent choice history in all sessions for every model. Under the informed prompt, by contrast, this association was not significant in nearly all sessions for every model aside from Gemini 3.1 Flash Lite. Anthropic models' choices depended instead on their own history, with strongly negative lag-1 coefficients (medians -3.4 to -5.4) consistent with the alternation already described. Additional sequential diagnostics supported this characterization (Appendix D).

4.3 Models diagnosed their alternation but only partly corrected it

In the informed sessions, models from Anthropic often described the pattern they were producing. Of the 4,179 eligible responses, 44% identified the model's alternation as predictable or exploitable, almost always alongside a stated intention to break it. Yet responses that identified the pattern ended in an alternating choice as often as responses that did not, about 80% of the time in both cases. Sonnet, for example, noted, "My last choice was S, so the alternating pattern would predict T next." Despite then writing "I need to be unpredictable," it chose TRIANGLE, the very action the pattern predicted, before returning to STAR on the following round.

This raw comparison ignores the possibility that diagnoses could be triggered by a conspicuous alternating run, when continuing to alternate was most likely anyway. We thus compared responses that followed similar recent histories, using a logistic regression that controlled for the length of the current alternating run, the switch rate and pattern over recent rounds, the previous choice and outcome, round, and session (Appendix C). Relative to responses with similar histories, those that diagnosed the alternation were less likely to continue it (OR = .468, 95% CI [.323, .677]). Averaged across responses, the predicted probability of alternating was 73.0% with a diagnosis and 84.0% without one. Diagnosis thus accompanied only a partial correction. Mentions of flipping a coin showed the opposite association: responses that invoked a coin were more likely to alternate (82.7% versus 77.5%; OR = 1.442, 95% CI [1.228, 1.693]), although this category was coded less reliably.

4.4 Robustness checks and prompt ablations

Over-alternation did not depend on temperature or action labels. In supplemental informed-prompt sessions (Appendix F), it persisted at temperatures of 0.3, 0.7, and 1.0, with runs scores of roughly 6.5 to 10.2 in the Anthropic models, and across three action-label pairs, with scores of roughly 7 to 14 in the strongest over-alternators. Marginal bias was more sensitive to labels and option order, although GPT-5.4 kept a strong side bias across temperatures and label pairs. Both signatures also recurred in the counterbalanced replication (Appendix E).

Ablations of the informed prompt pointed to the fair-coin instruction as an important contributor. Removing it produced the clearest reduction or change in the principal signature in five of the six models whose performance had declined: Sonnet's runs score fell from 10.4 to 2.2, and GPT-5.4's

TRIANGLE frequency moved from .26 to .57. Removing the warning about repetition and alternation eliminated neither signature and strengthened over-alternation in Sonnet and Haiku (runs $z = 13.7$ for both). Gemini 3.1 Flash Lite was the exception, responding more to removal of the warning than to removal of the fair-coin instruction. The fair-coin instruction thus contributed to the signatures, but removing it did not eliminate them in every model.

5 Discussion

These results recall a lesson from the experimental study of mixed strategies. The participants of O'Neill (1987) appeared to play the equilibrium until Brown & Rosenthal (1990) examined the order of their choices, and several of our models likewise appear balanced until their choices are examined in sequence. Equilibrium in repeated matching pennies requires choices that are unpredictable given the history of play. The models that lost most under informed prompting chose each action about equally often, and across models, sequential structure accounted for more of the change in performance than marginal bias did. By pitting each model against an opponent that learns from its history, PACA measures what that predictability costs.

The prompt manipulation shows how such predictability can arise. The prompts reflect different ways that players might approach the game of matching pennies, harkening back to a distinction touched upon by von Neumann & Morgenstern (1944, p. 144): predicting the opponent versus playing irregularly to avoid being predicted. The minimal prompt, which cast each choice as a forecast, elicited the former approach, as the models tracked Proteus's recent choices and more closely approximated equilibrium play. In contrast, the informed prompt prescribed the latter approach, which counterintuitively caused the models to become more predictable. Despite an explicit warning, some models over-alternated, much as people and LLMs do when asked to produce random sequences (Bar-Hillel & Wagenaar, 1991; Bigelow et al., 2024; Van Koevering & Kleinberg, 2024).

The verbal traces show the gap between stated strategy and behavior most directly. This gap resembles the unfaithful reasoning documented in other settings (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025). Related work draws a similar distinction between knowing a strategy and executing it. Yang (2025) found that stronger models chose to randomize more often when given hints, in a setting where the system carried out the random choice, whereas PACA required models to generate each choice themselves. Lekeas & Stamatopoulos (2026) found that later processing within a model could override intermediate predictions favoring equilibrium play, and that chain-of-thought reasoning worsened equilibrium play in smaller models while yielding near-equilibrium play in models above 70B parameters. Whether access to an external source of randomness would close the gap in PACA remains to be tested.

The informed-prompt results also fit a broader observation about rationality and cognitive complexity. Stanovich (2013) argued that simpler cognitive architectures can satisfy the axioms of rational choice more easily, because more complex minds respond to context and symbolic meaning as well as to consequences. Analogous patterns seem to appear in competitive games: people generate more random sequences when competing than when instructed to randomize (Rapoport & Budescu, 1992), and chimpanzees came closer to equilibrium than humans in comparable games (Martin et al., 2014). The informed prompt may have played a similar role for the models, adding strategic context that interfered with the very policy it described.

6 Conclusion

We introduced PACA, an assay that evaluates mixed-strategy play in LLMs against an adaptive opponent that exploits regularities in the history of play. Across seven models, balanced action frequencies often coexisted with predictable sequential structure, and prescribing the equilibrium strategy made most models easier to exploit, by inducing either over-alternation or a fixed action bias. Models frequently recognized these patterns in their reasoning without fully correcting them. These results suggest that evaluations of strategic behavior in LLMs should test conditional structure rather than aggregate frequencies alone, and that a strategy stated in the prompt or in the model's own reasoning can be an unreliable guide to the policy it executes. Extending PACA beyond its present scope (Appendix G) to richer games, opponents with other learning rules, and agents with access to external randomness would test how far these findings generalize.

REPRODUCIBILITY STATEMENT

The appendices provide the opponent calibration, request protocol, model provenance, behavioral diagnostics, trace-coding procedures, and counterbalanced replication. The experiment's data-and-code will be released upon publication and will include the session logs, round-level records, configurations, and analysis scripts. The main study was not preregistered. The replication design, configuration, and analysis were fixed and version-tagged before collection.

AI USE STATEMENT

AI language models from Anthropic, OpenAI, xAI, and Google assisted with research ideation and literature discovery, drafting and editing, writing and reviewing experiment and analysis code, conducting and cross-checking statistical analyses, and simulation work.

REFERENCES

E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

M. Bar-Hillel and W. A. Wagenaar. The perception of randomness. *Advances in Applied Mathematics*, 12(4):428–454, 1991. doi: 10.1016/0196-8858(91)90029-I.

D. J. Barraclough, M. L. Conroy, and D. Lee. Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7(4):404–410, 2004.

E. Bigelow, E. S. Lubana, R. Dick, H. Tanaka, and T. Ullman. In-context learning dynamics with random binary sequences. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.17639.

J. N. Brown and R. W. Rosenthal. Testing the minimax hypothesis: A re-examination of O'Neill's game experiment. *Econometrica*, 58(5):1065–1081, 1990.

D. V. Budescu and A. Rapoport. Subjective randomization in one- and two-person games. *Journal of Behavioral Decision Making*, 7(4):261–278, 1994. doi: 10.1002/bdm.3960070404.

C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

M. R. A. Chance and W. M. S. Russell. Protean displays: A form of allaesthetic behaviour. *Proceedings of the Zoological Society of London*, 132(1):65–70, 1959.

Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, and E. Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

P. M. Driver and D. A. Humphries. *Protean Behaviour: The Biology of Unpredictability*. Clarendon Press, Oxford, 1988.

C. Fan, J. Chen, Y. Jin, and H. He. Can large language models serve as rational players in game theory? A systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (16):17960–17967, 2024. doi: 10.1609/aaai.v38i16.29751.

Xiachong Feng, Longxu Dou, Minzhi Li, Qinghao Wang, Yu Guo, Haochuan Wang, Chang Ma, and Lingpeng Kong. A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios. *Transactions on Machine Learning Research*, 2025. URL https://mlanthology.org/tmlr/2025/feng2025tmlr-survey/.

S. Guo, H. Bu, H. Wang, Y. Ren, D. Sui, Y. Shang, and S. Lu. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*, 2024. URL https://arxiv.org/abs/2401.01735.

Z. Guo, H. Lv, C. Zhang, Y. Zhao, Y. Zhang, and L. Cui. The illusion of randomness: How LLMs fail to emulate stochastic decision-making in rock-paper-scissors games? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8618–8637, 2025. doi: 10.18653/v1/2025.findings-emnlp.458.

N. Herr, F. Acero, R. Raileanu, M. Pérez-Ortiz, and Z. Li. Are large language models strategic decision makers? A study of performance and bias in two-player non-zero-sum games. *arXiv preprint arXiv:2407.04467*, 2024. URL https://arxiv.org/abs/2407.04467.

A. K. Hopkins, A. Renda, and M. Carbin. Can LLMs generate random numbers? Evaluating LLM sampling in controlled domains. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, 2023.

D. A. Humphries and P. M. Driver. Protean defence by prey animals. *Oecologia*, 5(4):285–302, 1970. doi: 10.1007/BF00815496.

J. Jia, Z. Yuan, J. Pan, P. E. McNamara, and D. Chen. LLM strategic reasoning: Agentic study through behavioral game theory. In *Advances in Neural Information Processing Systems*, 2025.

T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, K. Lukosiute, K. Nguyen, N. Cheng, N. Joseph, N. Schiefer, O. Rausch, R. Larson, S. McCandlish, S. Kundu, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

D. Lee, M. L. Conroy, B. P. McGreevy, and D. J. Barraclough. Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research*, 22(1):45–58, 2004.

P. V. Lekeas and G. Stamatopoulos. What suppresses Nash equilibrium play in large language models? Mechanistic evidence and causal control. *arXiv preprint arXiv:2604.27167*, 2026.

Q. Liu. Does GPT-4 play dice? *ChinaXiv preprint 202402.00204*, 2024. doi: 10.12074/202402.00204V1.

N. Lorè and B. Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490, 2024. doi: 10.1038/s41598-024-69032-z.

C. F. Martin, R. Bhui, P. Bossaerts, T. Matsuzawa, and C. Camerer. Chimpanzee choice rates in competitive games match equilibrium game theory predictions. *Scientific Reports*, 4:5182, 2014.

G. F. Miller. Protean primates: The evolution of adaptive unpredictability in competition and courtship. In A. Whiten and R. W. Byrne (eds.), *Machiavellian Intelligence II: Extensions and Evaluations*, pp. 312–340. Cambridge University Press, 1997.

D. Mookherjee and B. Sopher. Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, 7(1):62–91, 1994.

J. Ochs. Games with unique, mixed strategy equilibria: An experimental study. *Games and Economic Behavior*, 10(1):202–217, 1995.

B. O'Neill. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences*, 84(7):2106–2109, 1987.

I. Palacios-Huerta. Professionals play minimax. *Review of Economic Studies*, 70(2):395–415, 2003. doi: 10.1111/1467-937X.00249.

S. Phelps and Y. I. Russell. The machine psychology of cooperation: Can GPT models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 6(1):015018, 2025. doi: 10.1088/2632-072X/ada711.

A. Rapoport and D. V. Budescu. Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, 121(3):352–363, 1992.

F. Sanabria and E. Thrailkill. Pigeons (Columba livia) approach Nash equilibrium in experimental matching pennies competitions. *Journal of the Experimental Analysis of Behavior*, 91(2):169–183, 2009.

J. M. Shachat. Mixed strategy play and the minimax hypothesis. *Journal of Economic Theory*, 104 (1):189–226, 2002. doi: 10.1006/jeth.2001.2915.

A. Silva. Large language models playing mixed strategy Nash equilibrium games. *arXiv preprint arXiv:2406.10574*, 2024.

K. E. Stanovich. Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice. *Thinking & Reasoning*, 19(1):1–26, 2013. doi: 10.1080/13546783.2012.713178.

H. Sun, Y. Wu, Y. Cheng, and X. Chu. Game theory meets large language models: A systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10669–10677, 2025. doi: 10.24963/ijcai.2025/1184.

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

K. Van Koevering and J. Kleinberg. How random is random? Evaluating the randomness and humaness of LLMs' coin flips. *arXiv preprint arXiv:2406.00092*, 2024.

A. Vidler and T. Walsh. Playing games with large language models: Randomness and strategy. *arXiv preprint arXiv:2503.02582*, 2025.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11(2):147–162, 1940.

M. Walker and J. Wooders. Minimax play at Wimbledon. *American Economic Review*, 91(5): 1521–1538, 2001. doi: 10.1257/aer.91.5.1521.

Z. Xu, C. Yu, F. Fang, Y. Wang, and Y. Wu. Language agents with reinforcement learning for strategic play in the Werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 55434–55464, 2024. URL https://proceedings.mlr.press/v235/xu24ad.html.

L. Yang. Do LLMs know when to flip a coin? Strategic randomization through reasoning and experience. *arXiv preprint arXiv:2506.18928*, 2025.

Y. Zhang, S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024. URL https://arxiv.org/abs/2404.01230.

M. Zhao, Y. Du, and M. Wang. Large language models are bad dice players: LLMs struggle to generate random numbers from statistical distributions. *arXiv preprint arXiv:2601.05414*, 2026. URL https://arxiv.org/abs/2601.05414.

APPENDICES

A PROMPTS, PROCEDURE, AND MODEL PROVENANCE

A.1 MODELS AND EXECUTION RECORD

Main-study requests were made on July 2, 2026 through OpenRouter. The requested model, served-model identifier, response metadata, and UTC timestamp were recorded on every round. Each model returned a single served-model identifier across both main conditions. Table A1 reports the requested and served identifiers, and Table A2 reports access windows. Supplemental calls were made July 2–4, and the counterbalanced replication July 13–14, 2026.

Table A1: Requested and served model identifiers for the main experiment.

Model	Requested identifier	Served identifier
Opus 4.8	`anthropic/claude-opus-4.8`	`anthropic/claude-4.8-opus-20260528`
Sonnet 4.6	`anthropic/claude-sonnet-4.6`	`anthropic/claude-4.6-sonnet-20260217`
Haiku 4.5	`anthropic/claude-haiku-4.5`	`anthropic/claude-4.5-haiku-20251001`
GPT-5.4	`openai/gpt-5.4`	`openai/gpt-5.4-20260305`
GPT-5.4-mini	`openai/gpt-5.4-mini`	`openai/gpt-5.4-mini-20260317`
Gemini 3.1 FL	`google/gemini-3.1-flash-lite`	`google/gemini-3.1-flash-lite-20260507`
Gemini 3 FP	`google/gemini-3-flash-preview`	`google/gemini-3-flash-preview-20251217`

Table A2: Main-study access windows on July 2, 2026 (UTC).

Model	Access window
Opus 4.8	04:13–06:19
Sonnet 4.6	06:19–09:31
Haiku 4.5	09:32–11:54
GPT-5.4	11:54–12:52
GPT-5.4-mini	12:52–13:39
Gemini 3.1 FL	13:39–14:20
Gemini 3 FP	14:20–15:42

Main requests used temperature 0 with top-p unset and a 20,000-token response cap. The longest observed response used 829 completion tokens and 2,008 characters, so the cap was not approached. No inference seed was passed to the API. Main opponent seeds were derived from master seed 20260701, with a separate seed recorded per session. Supplemental opponent seeds used master seed 20260801. These seeds controlled Proteus's stochastic choices rather than hosted-model inference. Different opponent choices produced different later histories and therefore different model inputs even at the same requested decoding temperature.

Choices were parsed automatically from responses ending in the selected action. Response text, latency, served-model metadata, and complete Proteus channel state were logged on each round. There were no parse or API failures in the main experiment. Exact per-session seeds, run identifiers, prompt hashes, command lines, and software versions belong to the accompanying execution record. No previous assistant response or reasoning was appended to a later request.

A.2 HISTORY CONSTRUCTION

Prompt framing, minimal or informed, was fixed throughout a session. Each API request resent the relevant system prompt, followed by one user message containing recent play. The user message contained the five most recent completed rounds in descriptive form and the last 25 choice–landing pairs in a grid. The five descriptive rounds therefore also appeared in the grid. Both displays expanded from the available history until reaching their respective limits and then became rolling windows. On round 1, the message consisted of "No previous rounds yet," a blank line, and "Your action:". The following is the message at round 26 of the first main-study minimal session. T denotes TRIANGLE and S denotes STAR.

```
Previous rounds (oldest → most recent):
You chose STAR - blue circle landed on TRIANGLE
You chose STAR - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on TRIANGLE
You chose TRIANGLE - blue circle landed on STAR (most recent)
(grid reads left to right, oldest → most recent)
Your last 25 choices: S T T T T T T T T S S T S S S S S S S S S T T T
Blue circle landed on: T T T S T T T S S S T S S S S T S S T S T T T T S
Your action:
```

Wording, ordering, labels, and format were generated by the same template in both conditions. The histories themselves differed because each session generated its own choices and landings. The fixed windows limited context growth without requiring the model to preserve earlier responses. The illustration in Figure 1 depicts the game and information flow, not visual stimuli presented to the model.

A.3 Scope of the history window

The recent window kept the context load approximately constant over the session. It did not preclude the memoryless reference policy of independent equiprobable action selection. Proteus, however, accumulated evidence across the full session. That asymmetry could limit a model attempting to infer the opponent's statistics or counter-exploit a bias it had induced. The conclusions therefore concern the policy expressed under this calling protocol and information window, not the best performance achievable with complete history or an external random-action tool.

B Adaptive opponent and calibration

B.1 Detection, commitment, and updating

Proteus implemented the choice-plus-outcome algorithm developed by Lee et al. (2004). Its nine channels evaluated the same accumulating session record. One channel tested overall action preference. Four channels tested the action following each currently relevant sequence of one to four previous choices. Four further channels incorporated the corresponding choice–outcome histories. History length therefore defined the conditioning context, not a sliding window of observations used to estimate the bias. Earlier occurrences of a context throughout the session contributed to the estimate.

Each channel applied an exact binomial test against equal choice probability, with detection criterion $p < .05$. When one or more channels detected a bias, Proteus selected the strongest detected regularity and used it to predict the next action. It then chose the mismatching action with probability equal to the estimated bias. Without a detection, it selected each action with probability one-half. Proteus committed before observing the model's current choice. Only after both choices had been committed did the completed round enter the record used for subsequent decisions.

For illustration, suppose the current context consists of TRIANGLE followed by STAR. If 19 of 26 previous occurrences of that context were followed by TRIANGLE, the estimated conditional frequency is approximately .73. The two-sided binomial test gives $p = .029$. This channel would therefore predict TRIANGLE, and Proteus would choose STAR with probability .73. The example illustrates one channel rather than a separate opponent policy.

B.2 Null calibration

The nine channels reuse overlapping histories and are evaluated repeatedly. Their operating characteristics were therefore calibrated empirically rather than inferred from nine independent tests. In 1,000 sessions of 200 rounds, a synthetic player chose TRIANGLE or STAR independently with probability one-half on every round. Proteus used the same detection and commitment procedure as in the LLM experiment. Recorded quantities included channel activation, the selected channel, reward rates, and the duration and persistence of false-positive episodes.

At least one channel was active on 11.56% of null rounds, with a between-session standard deviation of 13.56 percentage points. Per-channel rates declined with history length, from 3.61% for marginal bias and 3.38% for the one-choice channel to 0.01% for the four-step choice-plus-outcome channel. The higher-order contexts occurred infrequently. False-positive activation episodes averaged 1.92 rounds and generally disappeared within a few further random choices. Proteus's mean win rate was .4991, consistent with the theoretical .500 reference for independent equiprobable play.

These checks distinguish detection from successful exploitation. Chance regularities can activate Proteus, but did not produce a systematic advantage against the null player. In the model analyses, channel activation is the fraction of rounds on which any channel crossed its detection criterion. It is an internal state of the opponent rather than an independent confirmatory test or a complete measure of loss severity.

B.3 Counter-policy checks

Additional simulations examined the limits of the .500 reference against this particular algorithm. Fixed side bias, deterministic alternation, noisy alternation, and win-stay/lose-shift policies did not outperform the reference. A myopic agent with the complete session history simulated both candidate choices and selected the action with the higher immediate win probability. Across 100 sessions it achieved .5017 (SE .0039).

A feed-and-harvest policy could outperform Proteus. It incurred early losses while establishing a false side bias and then responded to the strongest active channel. Across feeding windows, win rates were approximately .55 to .57, with a maximum of .574. The accumulated estimate persisted after the player reversed its behavior, allowing later gains to exceed the initial cost. Thus, independent 50/50 play is a minimax-safe reference, not a ceiling against this fixed opponent. None of the tested LLMs exhibited the corresponding feed-and-harvest profile.

Calibration seeds were pinned. Null sessions used base seed 7000, structured-policy checks 5000, the myopic agent 9000, and feed-and-harvest checks 9500, incremented across sessions. The underlying implementation and configurations are part of the experiment's data-and-code release.

C Verbal traces and behavioral choices

C.1 Coding and eligible responses

The secondary analysis used the informed sessions of Claude Opus 4.8, Sonnet 4.6, and Haiku 4.5. Each response was associated with the action emitted at its end. Excluding the first round of each of the 21 sessions left 4,179 responses with a defined preceding choice. Three nonexclusive categories were coded mechanically. An alternation diagnosis required the stem `alternat` together with `predict`, `exploit`, `pattern`, or `detect`. Diagnosis with stated corrective intent additionally required `break`, `independen`, or `random`. Coin language required `coin`.

These categories were reconstructed from the documented definitions because the original executable coder was not recovered. The reconstructed rules reproduced the reported category counts exactly. The validation below evaluates whether these string-defined categories corresponded to the intended meanings. The analyses used the coded categories for all eligible responses, not human judgments for the entire response set.

Table C1: Trace categories and unadjusted alternation in informed Anthropic sessions.

Category	Responses	Alternating actions
All eligible responses	4,179	3,349 (80.1%)
Alternation diagnosis	1,838 (44.0%)	1,472 (80.1%)
Diagnosis and corrective intent	1,834 (43.9%)	1,471 (80.2%)
Coin language	1,901 (45.5%)	1,529 (80.4%)

Categories were nonexclusive. The first round of each session was excluded because it had no preceding choice. The rates are descriptive, not a conditional comparison.

The unadjusted alternation rate was approximately 80% in every category and 80.1% overall. This comparison does not account for the behavioral histories that elicited the statements. A diagnosis can occur disproportionately during a conspicuous alternating run, so its association with the accompanying action was evaluated conditional on recent play.

C.2 Conditional analysis

Primary logistic regressions included rounds 12–200 so that every observation had a complete ten-transition history, yielding 3,969 responses. The outcome was whether the action emitted with a response differed from the previous choice. Separate models estimated the association for each trace category. Covariates included log prior alternation-run length and its square, switch rate over the preceding ten transitions, the exact preceding three-transition pattern, prior choice and outcome, linear and quadratic round, and session fixed effects. Confidence intervals used session-clustered robust standard errors, with critical values from a t-distribution with 20 degrees of freedom (the number of sessions minus one).

Adjusted probabilities were obtained by standardizing fitted probabilities with the category set absent or present while retaining the observed covariates. Alternation diagnosis was associated with an adjusted reduction of 11.0 percentage points, and diagnosis with corrective intent with a reduction of 10.4 points. The categories differed on only four eligible responses and should not be read as independent replications. Coin language was associated with a 5.3-point increase in alternation. The complete estimates appear in Table C2.

Table C2: Conditional associations between trace categories and alternation.

Category	OR [95% CI]	Absent / present	Difference [95% CI]
Diagnosis	.468 [.323, .677]	84.0% / 73.0%	-11.0 [-16.4, -5.7]
Diagnosis and intent	.490 [.342, .701]	83.8% / 73.5%	-10.4 [-15.6, -5.2]
Coin language	1.442 [1.228, 1.693]	77.5% / 82.7%	$+5.3$ [$+3.0$, $+7.5$]

Absent and present are adjusted alternation probabilities. Differences are percentage points. Primary models used rounds 12–200 ($n = 3{,}969$) and session-clustered uncertainty. Covariates are described in Appendix C.

The diagnosis results indicate partial correction relative to comparable measured histories, while continued alternation remained frequent. They do not identify a causal effect of the words on the action. The verbal statement and action were produced in the same response, and the analysis cannot establish the computational relation between them. A single action consistent with an alternating pattern is not itself inconsistent with random choice. The departure is the sustained conditional structure across responses.

C.3 Validation of the coding

Two coders independently applied the frozen codebook and its ten worked examples to a stratified sample of 150 traces. They had no calibration discussion or shared adjudication. A third blinded coder, provided with the trace, category, and codebook, resolved disagreements and uncertain cases. Pre-adjudication agreement between the two independent human coders was measured using unweighted Cohen's kappa, treating Uncertain as a third response category. Kappa was .97 for alternation diagnosis, .93 for diagnosis with corrective intent, and .64 for coin language.

Against the adjudicated labels, inverse-probability weighting accounted for the stratified sampling. Precision was 86.5% for alternation diagnosis, 86.5% for diagnosis with corrective intent, and 80.2% for coin language. Recall was 100.0%, 99.9%, and 93.8%, respectively. Confidence intervals were percentile intervals from 100,000 bootstrap samples of the seven whole sessions within each model, using seed 20260717. Table C3 gives the intervals.

The diagnosis categories were more reproducible than coin language. The latter association is therefore the most tentative, with greater uncertainty about whether a keyword match captured the intended meaning. The coding audit supports interpretation of the mechanically defined categories but does not remove measurement error from the full response set.

Table C3: Validation of the mechanically defined trace categories.

Category	κ	Precision [95% CI]	Recall
Diagnosis	.97	86.5 [76.3, 94.4]	100.0
Diagnosis and intent	.93	86.5 [76.2, 94.4]	99.9
Coin language	.64	80.2 [69.3, 90.1]	93.8

Precision and recall compare automated coding with the adjudicated human labels, with inverse-probability weighting. Kappa measures agreement between the two independent human coders before adjudication, treating Uncertain as a third category.

C.4 Representative traces

The following excerpts illustrate the coded language and accompanying choices. They supplement the aggregate analysis rather than establish the behavioral effect by themselves.

Sonnet 4.6, informed round 60. "Looking at my last 25 choices, I see a strong alternating pattern (S T S T S T. . .) with only minor deviations. An exploitative opponent would detect this and predict my next choice. My last choice was S, so the alternating pattern would predict T next. . . . I need to be unpredictable." The response ended with TRIANGLE, followed by STAR on the next round.

Opus 4.8, informed round 6. "I need to choose randomly, independent of past choices. My recent pattern shows alternation (S T S T S), which is exploitable. To stay unpredictable, I'll make an independent random choice. TRIANGLE" The preceding choice was STAR, and the next round returned to STAR.

Haiku 4.5, informed round 41. "The best strategy is to flip a mental coin: I'll choose TRIANGLE, but this should be a genuine 50/50 random choice, not based on pattern-breaking logic." No external coin was supplied. The statement is treated as observable response content, not evidence that an internal randomization operation occurred.

Sonnet 4.6, informed round 66. "The 'expected' next choice following this pattern would be T. . . I should choose randomly. Mentally flipping a coin: TRIANGLE. TRIANGLE" The following round selected STAR. References to a coin did not reliably correspond to independent action selection in the aggregate analysis.

D Behavioral diagnostics and statistical details

D.1 Session measures

Measures were computed separately for each 200-round session and summarized across the seven sessions within each model-by-prompt cell. Win rate is the proportion of matching choices. Empirical $P(\text{TRIANGLE})$ describes marginal action frequency. Activation is the fraction of rounds on which any of the nine Proteus channels crossed the detection threshold, and mean active channels counts the channels active on each round. These operating measures should not be equated with a monotonic scale of performance loss.

Alternation rate is the fraction of transitions on which the choice changes. Win-stay and lose-shift rates are conditional on the model's previous outcome and referenced to its own previous choice. Lose-stay, used in Figure D1, equals one minus lose-shift. The Wald–Wolfowitz runs statistic compares the number of runs with the distribution expected from random ordering of the observed action counts. Positive values indicate over-alternation. For comparison with Lee et al. (2004), we compute two information-theoretic diagnostics: the Miller–Madow-corrected entropy of overlapping choice triplets and mutual information between the previous two choices and the current choice. Table D1 reports the results.

D.2 Model-level contrast and complementary analyses

The primary effect was calculated by subtracting each model's mean minimal win rate from its mean informed win rate and averaging those seven differences. Its 95% confidence interval used

Table D1: Cell-mean behavioral measures by model and prompt.

Performance, marginal frequency, and switching

Model	Prompt	Win	P(T)	Act	Chan	Alt	Runs z
Opus 4.8	minimal	0.413	0.481	0.790	1.859	0.350	-4.284
Opus 4.8	informed	0.374	0.406	0.685	2.571	0.731	7.199
Sonnet 4.6	minimal	0.471	0.524	0.249	0.322	0.479	-0.618
Sonnet 4.6	informed	0.213	0.524	0.888	5.162	0.868	10.375
Haiku 4.5	minimal	0.500	0.501	0.354	0.594	0.451	-1.445
Haiku 4.5	informed	0.283	0.474	0.842	4.324	0.805	8.619
GPT-5.4	minimal	0.460	0.471	0.570	0.996	0.408	-2.606
GPT-5.4	informed	0.350	0.261	0.863	3.863	0.386	-0.008
GPT-5.4-mini	minimal	0.491	0.461	0.423	0.648	0.440	-1.629
GPT-5.4-mini	informed	0.420	0.319	0.834	2.731	0.462	0.845
Gemini 3.1 FL	minimal	0.455	0.472	0.369	0.731	0.424	-2.158
Gemini 3.1 FL	informed	0.340	0.256	0.854	4.006	0.455	2.709
Gemini 3 FP	minimal	0.405	0.491	0.774	1.821	0.361	-3.981
Gemini 3 FP	informed	0.452	0.474	0.314	0.782	0.620	3.392

Outcome-conditioned and information-theoretic measures

Model	Prompt	Win-stay	Lose-shift	H_3	MI
Opus 4.8	minimal	0.834	0.479	2.878	0.076
Opus 4.8	informed	0.171	0.672	2.496	0.237
Sonnet 4.6	minimal	0.408	0.381	3.004	0.012
Sonnet 4.6	informed	0.081	0.854	2.101	0.464
Haiku 4.5	minimal	0.637	0.538	2.988	0.024
Haiku 4.5	informed	0.200	0.808	2.399	0.311
GPT-5.4	minimal	0.715	0.513	2.956	0.033
GPT-5.4	informed	0.451	0.298	2.502	0.006
GPT-5.4-mini	minimal	0.699	0.573	2.974	0.017
GPT-5.4-mini	informed	0.478	0.417	2.700	0.018
Gemini 3.1 FL	minimal	0.554	0.405	2.968	0.027
Gemini 3.1 FL	informed	0.227	0.295	2.384	0.055
Gemini 3 FP	minimal	0.836	0.495	2.880	0.084
Gemini 3 FP	informed	0.327	0.577	2.911	0.057

Win is win rate against Proteus. P(T) is empirical TRIANGLE frequency. Act is any-channel activation, Chan is mean active channels, and Alt is alternation rate. Runs z is the mean session statistic. H_3 is Miller–Madow-corrected choice-triplet entropy in bits. MI is mutual information between the previous two choices and the current choice in bits. Lose-stay equals one minus lose-shift.

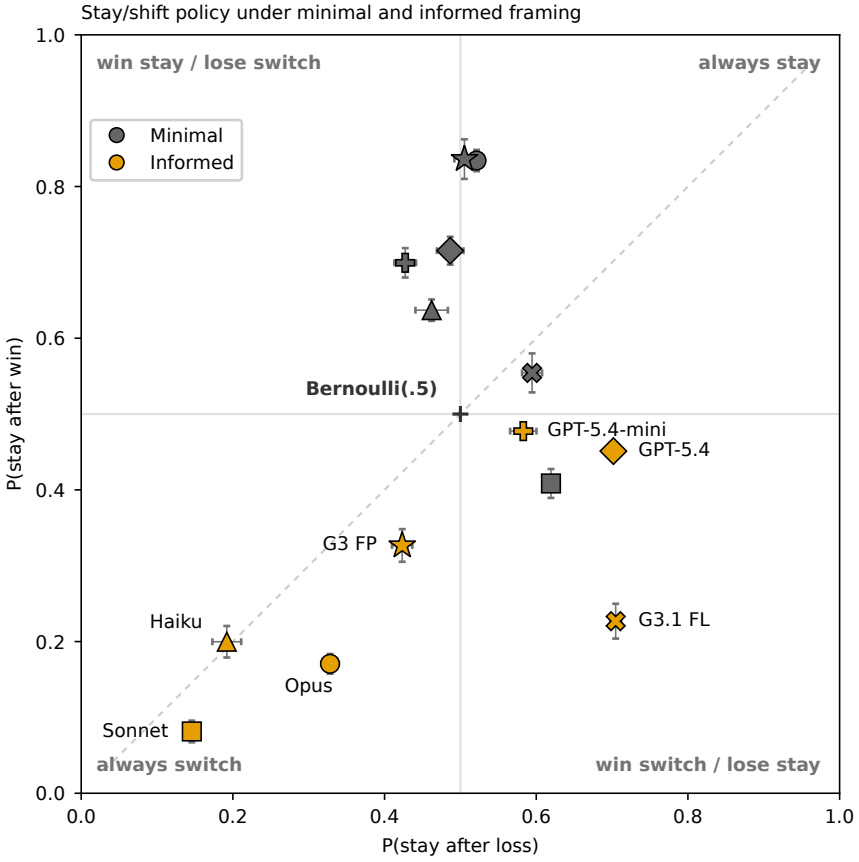


Figure D1: Outcome-conditioned stay probabilities under minimal and informed prompts. Points are model-by-prompt means across seven sessions, with SEM. Open symbols are minimal and filled symbols informed. The diagonal denotes equal stay probabilities after wins and losses. Independent Bernoulli(.5) play lies at the midpoint.

the model-level variation. The paired t-test gave $t(6) = -2.78$, $p = .032$, and the exact Wilcoxon signed-rank test gave $W = 2$, $p = .047$.

Leave-one-model-out estimates ranged from $-.135$ to $-.084$. The mean difference also remained negative when the three Anthropic models, two OpenAI models, or two Google models were omitted (Table D2). These checks describe stability within the tested panel and do not turn that panel into a representative sample of model families.

Table D2: Model and developer-family omission checks.

Omitted subset	n	Mean difference
Opus 4.8	6	$-.121$
Sonnet 4.6	6	$-.084$
Haiku 4.5	6	$-.091$
GPT-5.4	6	$-.109$
GPT-5.4-mini	6	$-.115$
Gemini 3.1 FL	6	$-.108$
Gemini 3 FP	6	$-.135$
Anthropic	4	$-.062$
Google	5	$-.139$
OpenAI	5	$-.116$

Differences are informed minus minimal win rate.

A complementary round-level binomial generalized linear mixed model specified framing as a fixed effect, a random intercept and framing slope for model, and a random intercept for session. It was fit by maximum marginal likelihood with a Laplace approximation. The informed coefficient was $-.474$, corresponding to an odds ratio of .623 and an approximate profile 95% interval of [.431, .899].

D.3 Brown–Rosenthal conditional regressions

Following Brown & Rosenthal (1990) and Martin et al. (2014), we fit a logistic regression for each session that predicted the model's current choice from four sets of predictors: the model's choices on the two previous rounds, Proteus's choices on those rounds, the outcomes of those rounds, and Proteus's current choice. To test whether a set of predictors mattered, we removed it from the full model and compared the fit using a likelihood-ratio test ($p < .05$). Because the other predictors stayed in the model, each test asked whether that set added information beyond the rest. We also report the coefficient for the model's previous choice: positive values mean the model tended to repeat its last choice, and negative values mean it tended to switch. Table D3 shows, for each model and prompt, how many sessions had a significant effect for each set of predictors and the median coefficient for the previous choice.

A significant effect of Proteus's current choice suggests that both players depended on history beyond the two previous rounds included in the regression. This effect was significant in nine of 98 sessions, close to the roughly five expected by chance. Removing Proteus's current choice from the regression had little effect on the other results: the Own, Opponent, and Outcome counts in Table D3 were unchanged in nine of its 14 rows (seven models, each under two prompts), and no count changed by more than two sessions.

E Condition order and counterbalanced replication

E.1 Original order and descriptive checks

In the original experiment, all seven minimal sessions preceded all seven informed sessions within each model. If anything, this maximizes the informativeness of the latter, but the prompt comparison was therefore not separated from order by design. Served-model metadata showed one identifier across both conditions for each model. Calls spanned approximately 41 minutes to 3 hours 12 minutes within model. Correlations between session order and win rate within condition were $r = -.095$, $p = .52$ under the minimal prompt and $r = .172$, $p = .24$ under the informed prompt. The mean

Table D3: Brown–Rosenthal profiles by model and prompt.

Model	Prompt	Own	Opponent	Outcome	Current	Lag-1
Opus 4.8	minimal	7/7	7/7	2/7	1/7	$+1.26$
Opus 4.8	informed	7/7	1/7	1/7	1/7	-3.68
Sonnet 4.6	minimal	2/7	7/7	3/7	0/7	$+0.38$
Sonnet 4.6	informed	7/7	1/7	1/7	1/7	-5.35
Haiku 4.5	minimal	5/7	7/7	0/7	0/7	$+0.08$
Haiku 4.5	informed	7/7	2/7	0/7	0/7	-3.40
GPT-5.4	minimal	5/7	7/7	0/7	1/7	$+0.66$
GPT-5.4	informed	0/7	0/7	0/7	0/7	$+0.04$
GPT-5.4-mini	minimal	4/7	7/7	1/7	1/7	$+0.34$
GPT-5.4-mini	informed	1/7	1/7	2/7	0/7	-0.31
Gemini 3.1 FL	minimal	3/7	7/7	0/7	2/7	$+0.71$
Gemini 3.1 FL	informed	6/7	4/7	3/7	1/7	-1.79
Gemini 3 FP	minimal	6/7	7/7	0/7	1/7	$+1.41$
Gemini 3 FP	informed	5/7	1/7	0/7	0/7	-1.06

Each count gives the number of sessions, out of seven, in which a set of predictors significantly improved the prediction of the model's current choice ($p < .05$). Own: the model's choices on the two previous rounds. Opponent: Proteus's choices on the two previous rounds. Outcome: whether the model won or lost the two previous rounds. Current: Proteus's choice on the current round. Lag-1: the median coefficient for the model's previous choice; positive values mean the model tended to repeat its last choice, and negative values mean it tended to switch.

absolute difference between the first and last sessions within a condition was .035, compared with a mean framing difference of $-.109$.

These observations did not reveal a simple within-condition trend, but they do not resolve the design confound. The counterbalanced experiment addressed it directly.

E.2 Replication design

The design and analysis of the replication were fixed and version-tagged before data collection began. Each model completed eight additional 200-round sessions, run as four adjacent pairs of one minimal and one informed session. Two pairs began with each condition, in the order minimal–informed, informed–minimal, informed–minimal, minimal–informed, or its mirror image. The two conditions therefore occupied serial positions with the same mean and the same spread, so any linear or quadratic trend in performance across sessions affected both conditions equally. Which order a model received was not aligned with its developer.

Both sessions in each pair used the same Proteus seed. Each model's effect was the mean of its four within-pair differences in win rate. The replication comprised 56 sessions and 11,200 rounds, collected on July 13–14, 2026, between 23:58 and 05:51 UTC. Each model returned a single served-model identifier, matching its requested checkpoint, across all eight sessions. Because the replication used the same seven models, it tests the robustness of the main result within this panel rather than its generality to other models.

E.3 Replication outcome and position checks

Win rate was again lower under the informed prompt in six of seven models. The mean model-level difference was $-.139$, 95% CI $[-.255, -.023]$, $t(6) = -2.92$, $p = .026$, compared with $-.109$ in the main study. Table E1 gives the model-level effects. Per-model intervals are unadjusted and based on four pairs per model. After Holm correction across the seven models, the differences for Opus, Sonnet, and GPT-5.4 remained significant.

Gemini 3 Flash Preview was again the exception, improving under the informed prompt. All four of its paired differences were positive, with a mean improvement of .069, compared with .047 in the main study. Its 95% CI, $[-.024, .161]$, was wide, however, so the consistent direction does not establish the size of the improvement.

Session order did not detectably alter the framing effect. The framing effect differed by $-.017$, 95% CI $[-.115, .080]$, $p = .68$, between pairs in which the informed session ran first and pairs in which it ran second. The change in the framing effect per session position, estimated within model, was .008, 95% CI $[-.017, .032]$, $p = .55$. These intervals do not rule out modest order effects, but neither analysis detected a reliable interaction. The replication therefore provides no evidence that running the minimal sessions first in the main study either produced or weakened the framing effect.

Table E1: Model-specific informed-minus-minimal win-rate differences.

Model	Main study	Replication	95% CI	p
Opus 4.8	$-.039$	$-.120$	$[-.178, -.062]$	.007
Sonnet 4.6	$-.258$	$-.275$	$[-.311, -.239]$	$< .001$
Haiku 4.5	$-.217$	$-.275$	$[-.497, -.053]$	.029
GPT-5.4	$-.110$	$-.213$	$[-.282, -.143]$	.002
GPT-5.4-mini	$-.071$	$-.056$	$[-.090, -.022]$	.013
Gemini 3.1 FL	$-.115$	$-.100$	$[-.195, -.005]$	.044
Gemini 3 FP	$+.047$	$+.069$	$[-.024, +.161]$	.099

Replication effects average four within-pair differences per model. Per-model intervals use a $t(3)$ reference and are unadjusted. The seven-model contrast is reported in the text.

E.4 Recurrence of the behavioral signatures

The two principal signatures also recurred. Sonnet and Haiku again paired near-balanced frequencies with strong over-alternation. Opus again combined over-alternation with a modest side bias. The GPT models and Gemini 3.1 Flash Lite showed their principal departures in overall action frequency. Table E2 compares informed-condition measures across the main and replication runs. These recurrent patterns support the behavioral characterization under counterbalanced order rather than only reproducing an aggregate win-rate difference.

Table E2: Recurrence of informed-condition behavioral signatures.

Model	P(TRIANGLE)	Runs z	Activation
Opus 4.8	$.406/.391$	$+7.2/+7.3$	$.685/.753$
Sonnet 4.6	$.524/.531$	$+10.4/+10.5$	$.888/.906$
Haiku 4.5	$.474/.476$	$+8.6/+9.8$	$.842/.874$
GPT-5.4	$.261/.169$	$0.0/+0.1$	$.863/.953$
GPT-5.4-mini	$.319/.334$	$+0.8/+0.9$	$.834/.849$
Gemini 3.1 FL	$.256/.283$	$+2.7/+2.5$	$.854/.831$
Gemini 3 FP	$.474/.471$	$+3.4/+2.8$	$.314/.280$

Cells give main-study / replication means under informed prompting. Activation is the fraction of rounds with at least one active channel.

F Supplemental robustness checks

F.1 Design and scope

The set of supplemental checks comprised 238 sessions across the same seven models. There were four informed-frame sessions per model at each of temperatures 0.3, 0.7, and 1.0 (84 sessions), four sessions per model for each of three component ablations (84), four sessions per model for each of two action-label substitutions (56), and single-session reversed-menu and minimal DIAMOND/SQUARE controls for each model (14). All completed without parse failures, API failures, or truncations.

Baseline prompt construction and parsing in the supplemental versions were identical to the main procedure. Each ablation removed one component of the informed prompt. Behavioral persistence was assessed through marginal frequency, runs scores, and Proteus activation (meaning how often Proteus identified any kind of significant bias).

F.2 PROMPT-COMPONENT ABLATIONS

The three ablations removed the sentence naming repetition and alternation, the fair-coin instruction, or the opponent description. The informed comparison came from the main study with seven sessions per model, whereas each ablation used four. Table F1 reports runs z, empirical TRIANGLE frequency, and channel activation for each condition.

Table F1: Component ablations of the informed prompt.

Model	Full informed	No warning	No fair coin	No opponent
Opus 4.8	$+7.2/.41/.69$	$+7.9/.46/.87$	$+6.0/.41/.70$	$+6.7/.38/.74$
Sonnet 4.6	$+10.4/.52/.89$	$+13.7/.50/.91$	$+2.2/.48/.28$	$+13.5/.50/.90$
Haiku 4.5	$+8.6/.47/.84$	$+13.7/.50/.91$	$+5.0/.49/.65$	$+10.8/.50/.83$
GPT-5.4	$0.0/.26/.86$	$+0.1/.12/.94$	$+1.2/.57/.50$	$-2.5/.18/.93$
GPT-5.4-mini	$+0.8/.32/.83$	$0.0/.40/.54$	$+3.1/.49/.41$	$+1.8/.21/.91$
Gemini 3.1 FL	$+2.7/.26/.85$	$+0.7/.63/.57$	$+6.0/.54/.81$	$-2.0/.35/.80$
Gemini 3 FP	$+3.4/.47/.31$	$-0.3/.51/.10$	$+4.9/.51/.66$	$-1.8/.44/.24$

Each cell gives runs z / TRIANGLE frequency / channel activation. The full informed baseline has seven sessions per model. Each ablation has four.

Removing the pattern warning did not remove the principal signatures. Over-alternation intensified in Sonnet and Haiku, while GPT-5.4's side bias remained. Removing the fair-coin instruction produced the clearest reduction or change in the principal signature in five of the six models whose main-study win rates had declined. Gemini 3.1 Flash Lite responded more strongly to removal of the pattern warning. Gemini 3 Flash Preview, which improved under the full informed prompt, became more vulnerable when the fair-coin instruction was removed. The effects of the prompt components were therefore heterogeneous. They implicate the randomization prescription without establishing it as the only cause.

F.3 TEMPERATURE

The informed prompt was tested at temperatures of 0.3, 0.7, and 1.0, with four sessions per model at each setting. The principal signatures persisted. Strong over-alternation remained in the Anthropic models, with runs scores approximately +6.5 to +10.2 across the tested temperatures, and GPT-5.4 retained its marginal bias. Gemini 3 Flash Preview did not show the same sustained signature. Figure F2 displays channel activation across decoding settings.

Persistence under these settings argues against the temperature-0 setting as a sufficient explanation of the observed signatures. It does not establish that the difference between minimal and informed prompting has the same magnitude at every temperature, since the minimal condition was not crossed with the temperature manipulation.

F.4 ACTION LABELS AND MENU ORDER

The informed action menu was reversed, and the original TRIANGLE/STAR pair was replaced with DIAMOND/SQUARE and TRIANGLE/SQUARE. A minimal DIAMOND/SQUARE run provided a descriptive comparison. Over-alternation persisted across the three label pairs, with runs scores approximately +7 to +14 in the strongest over-alternators. Marginal preference was more sensitive to labels and order and did not follow a single common direction across models.

For GPT-5.4, the side bias persisted when STAR was removed. It favored STAR in the original pair and SQUARE in both substitutions, so the preference was not unique to STAR. The favored action was named second in all three pairs, but the single reversed-menu session, which listed STAR first in the action menu, still favored STAR ($P(\mathrm{TRIANGLE}) = .18$), so these data do not separate label from order. This qualifies the interpretation of that cell without eliminating the measured vulnerability. An opponent can exploit a systematic preference regardless of the label that elicits it. Table F2 reports signed differences between first- and second-listed action frequencies alongside runs scores.

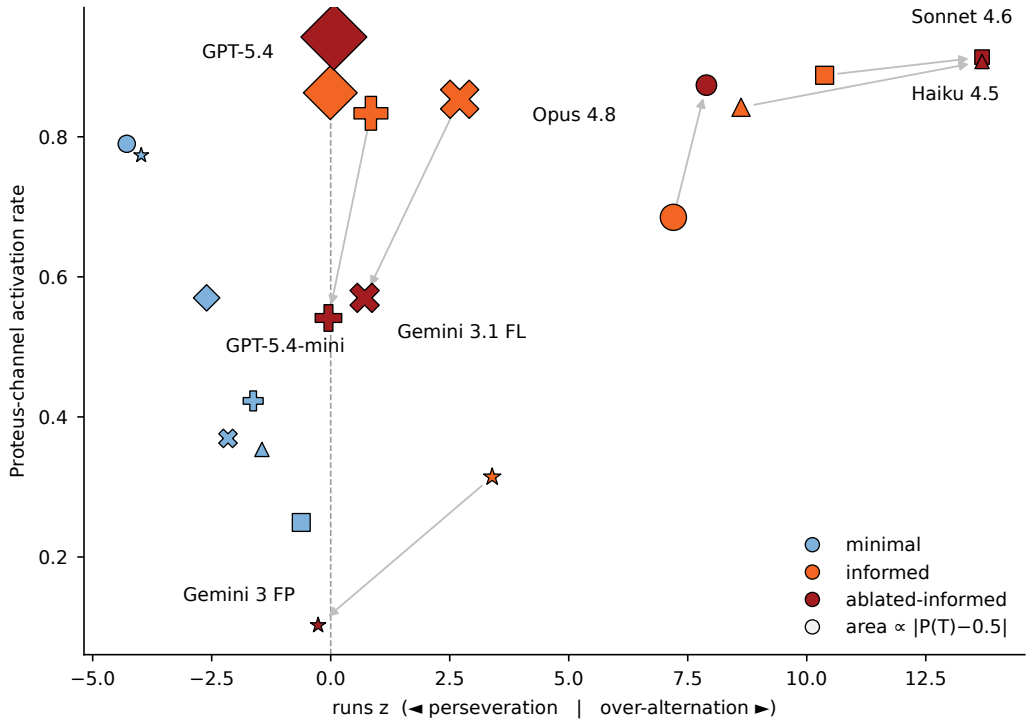


Figure F1: Effects of removing the pattern warning. Horizontal position is runs z, vertical position is Proteus activation, and marker area represents marginal bias. Colors distinguish minimal, full informed, and ablated-informed conditions. Each shape denotes one model. Removing the warning did not eliminate the principal informed signatures.

Table F2: Action-label comparisons under informed prompting.

Model	TRI/STAR	DIA/SQR	TRI/SQR	Runs z, three pairs
Haiku 4.5	STAR $-.05$	Near balance	TRI $+.12$	8.6 / 13.8 / 7.6
Sonnet 4.6	TRI $+.05$	DIA $+.11$	TRI $+.12$	10.4 / 9.4 / 9.7
Opus 4.8	STAR $-.19$	DIA $+.11$	Near balance	7.2 / 9.6 / 9.0
GPT-5.4	STAR $-.48$	SQR $-.36$	SQR $-.68$	0.0 / -1.8 / -3.5
GPT-5.4-mini	STAR $-.36$	SQR $-.10$	TRI $+.26$	0.8 / 0.9 / 1.8
Gemini 3.1 FL	STAR $-.49$	DIA $+.45$	TRI $+.10$	2.7 / 3.8 / 3.4
Gemini 3 FP	STAR $-.05$	DIA $+.12$	TRI $+.35$	3.4 / 1.9 / 1.1

Signed marginal differences are first-listed minus second-listed action frequency. The named token is the favored action. Runs scores follow the column order of the three action pairs.

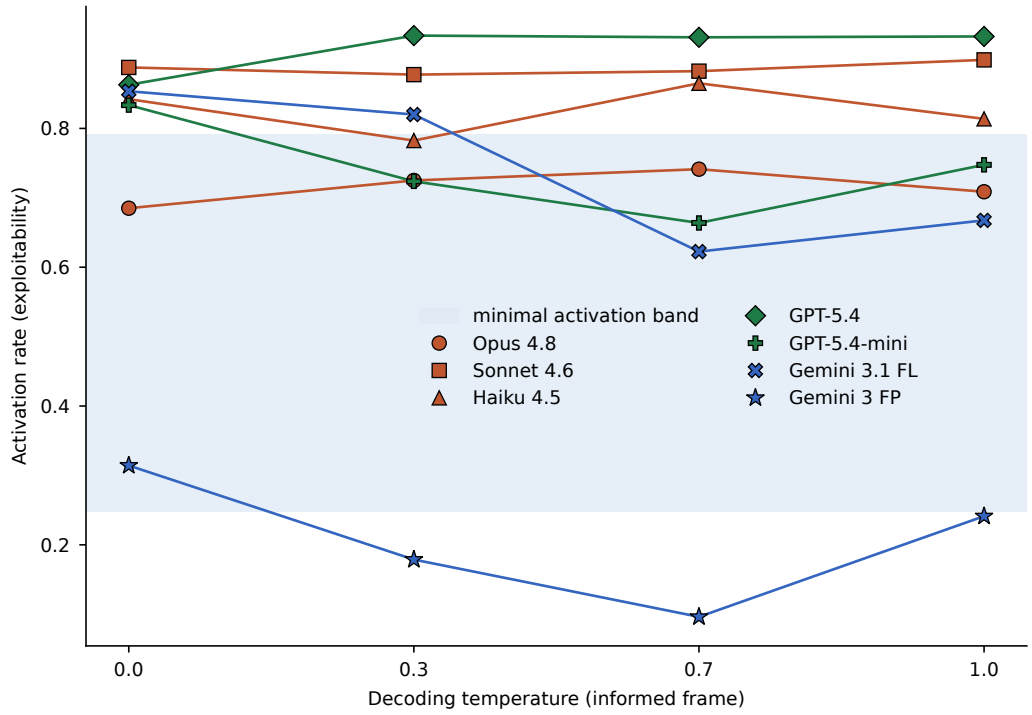


Figure F2: Proteus activation under informed prompting across temperatures. The shaded band spans the main-study minimal activation rates. The baseline uses the main experiment, and each nonzero-temperature cell has four sessions.

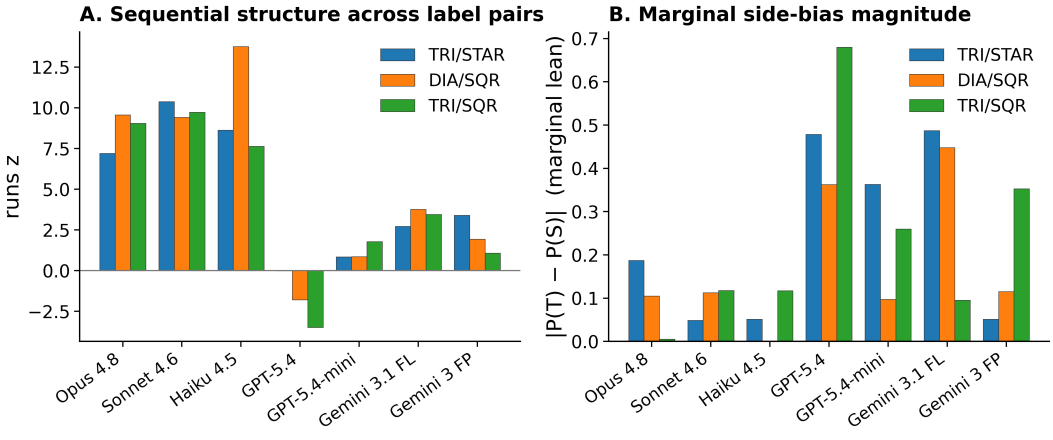


Figure F3: Sequential structure and marginal bias across action-label pairs. (A) Runs scores. (B) Absolute marginal frequency difference. Over-alternation persisted in Opus, Sonnet, and Haiku while marginal bias varied by model and label pair. TRI/STAR bars use the seven main-study informed sessions per model; DIA/SQR and TRI/SQR bars have four sessions each.

The single-session reversed-menu and minimal-label controls should be distinguished from the replicated main experiment. Matching pennies depends on payoff symmetry, which was retained, rather than an assumption that every action token has equal prior attraction.

G LIMITATIONS

We studied one game, matching pennies, because it isolates the problem of remaining unpredictable, and one adaptive opponent, an established algorithm from decision neuroscience; whether the same effects hold in other games and against other opponents remains to be tested. The panel of seven models from three developers, together with a counterbalanced replication, supports conclusions about these models under the two prompts; broader claims about LLMs would require a larger and more varied sample. The main experiment used fixed condition order, addressed by a separate prespecified counterbalanced replication. Temperature checks established persistence of the informed signatures but did not cross both prompts with every temperature. Proteus retained more history than the LLM, which may constrain counter-exploitation even though memoryless equilibrium play remains available. The trace findings are restricted to informed Anthropic sessions and do not establish causal effects of verbalization.