# From Win Rate to Worst Case: Auditing Language-Model Policies in Imperfect-Information Games

**Jiaxing Guo**
Imperial College London
henry.guo23@imperial.ac.uk

**Abstract**

Language models increasingly act strategically, yet evaluation against selected opponents can miss worst-case vulnerability. We elicit behavioral strategies in fixed finite-game schemas, lift them into sequence form, and compute exact best-response exploitability. We test 25 models in five scored games; Goofspiel is an exploratory boundary case. Every scored river and Leduc policy is exploitable despite secure-play prompts; river losses reach $3.1$ payoff units ($14\%$ of the terminal range; $71\times$ the value magnitude). We compare vulnerability with payoff against elicited opponent tables. In the primary draw, $88\%$ of policies are strictly *security-dominated*: a fully secure policy earns more against the stated opponent. At a $1\%$-range margin, $70\%$ remain dominated. Repeated forecasts, neutral wording, and a near-secure schema comparator preserve the result. Under two jointly conditioned draws, every policy has exploitability at least $0.10$, versus $3.3 \times 10^{-5}$ for the comparator; $31/50$ are strictly security-dominated; $49/50$ are practically suboptimal at their own exploitability budget. We define this cross-elicitation inconsistency between the two elicited tables as the *stated-opponent security gap*. State-wise action reconstruction preserves the result beyond matched sampling. In 200-big-blind four-action heads-up no-limit hold'em, nominal familywise-corrected lower bounds are positive for two of six policies and inconclusive for four. Auditing a fixed, elicited policy provides a game-specific worst-case bound unavailable from win rate or action agreement alone.

## 1 Introduction

Language models now negotiate, bid, and bargain, and are increasingly evaluated as strategic agents (Gandhi, Sadigh, and Goodman 2023; Fan et al. 2024; Duan et al. 2024; Akata et al. 2025; Chen et al. 2024). Common evaluations use proxies—win-rate against a fixed opponent, agreement with a solver's top action, or accuracy on curated scenarios. These measure performance against selected references, not how badly a policy can be beaten. In two-player zero-sum games, exploitability is the value an optimal adversary extracts; strong win-rate or action-match can therefore coexist with a vulnerable policy.

Performance against chosen opponents does not certify security. A complete fixed policy, however, can be audited directly. Each model is prompted to commit to a distribution over legal actions at every information set, lift that table into sequence form, and compute its exact security shortfall by linear programming. Here exploitability is an evaluation metric rather than a training objective.

We audit a 25-model panel across six imperfect-information testbeds spanning poker, simultaneous moves, negotiation, and dice bluffing. Every scored river and Leduc policy is substantially exploitable, while the small Kuhn game provides a boundary case: two models reach exact security. The losses are not an artifact of the output schema, and on the river they align with systematic over-aggression.

The second analysis tests a cross-elicitation consistency condition for a security–payoff trade-off. For each model we separately elicit an opponent table and compute the security–payoff frontier: the greatest value attainable against that stated opponent at each exploitability budget. Across five games, $73/83$ policies are strictly *security-dominated*: a fully secure policy would earn more against the very opponent the model describes. Even after requiring a shortfall greater than $1\%$ of the game's terminal payoff range, $58/83$ remain dominated. We define this cross-elicitation inconsistency as the *stated-opponent security gap*.

The paper makes three contributions:

- We instantiate and validate an exact audit for elicited policies complete within a predeclared finite schema. Sequence-form scoring removes rollout noise; equilibrium-through-schema controls isolate policy loss from representation loss.
- We use the security–payoff frontier to test whether vulnerability improves payoff against a stated opponent, including under practical margins and direct forecast conditioning. This separates exploitability magnitude from opponent-specific benefit.
- We corroborate the table audit with independently elicited state-wise actions: reconstructed policies track table scores and remain security-dominated in $68\%$ of cases. In 200-big-blind hold'em with four actions, lower bounds identify positive exploitability for two of six live policies.

## 2 Preliminaries and Audit Method

**Games and exploitability.** We study two-player zero-sum extensive-form games with perfect recall. In sequence form (Koller, Megiddo, and von Stengel 1996), a strategy for each player is a realization plan over action sequences, subject to linear flow constraints. The strategy spaces are

polytopes $\mathcal{X}$ and $\mathcal{Y}$, and player 1 earns $x^\top Ay$ from plans $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Let $v = \max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} x^\top Ay$. The exploitability of a player-1 strategy is its worst-case shortfall:

$$\mathrm{Expl}(x) \;=\; v - \min_{y\in\mathcal{Y}} x^\top Ay, \tag{1}$$

so $\mathrm{Expl}(x) = 0$ exactly for a maximin (equilibrium) strategy component. Unlike fixed-opponent win-rate, this quantity chooses the best response only after the policy is fixed. We use $|v|$ multiples only as a within-game descriptive scale. Cross-game figures use $\mathrm{Expl}(x)/r_g$, where $r_g = u_g^{\max} - u_g^{\min}$ is the terminal-payoff range, a ratio invariant to positive affine payoff transformations. Here "substantial" denotes at least $1\%$ on this scale. Game ranges and payoff conventions are in App. B. Bargaining uses a stated zero-sum reduction, Goofspiel is a value-zero boundary case, and the remaining games are natively zero-sum.

**Auditing a full policy.** Standard evaluations sample a model's moves; we instead freeze and score the object that determines exploitability. The box below specifies the audited object and the exact scoring procedure.

**The audit object.** *Input:* a prompt requests a complete strategy that minimizes exploitability (prompt excerpts, App. B; complete templates in the supplementary code). *Output:* strict JSON with one legal-action distribution per cell (54 on the river). *Lift:* cells broadcast to matching information sets (270 on the river), yielding a complete, class-measurable strategy. *Repair:* a fixed rule renormalizes legal mass or uses uniform play if none remains ($0.9\%$ of cells). *Exactness:* a linear program (LP) scores the lifted plan; equilibrium projected through the schema has exploitability $3.3 \times 10^{-5}$ (§3). *Correspondence check:* separately prompted actions reproduce the aggregate failure and moderately track table rankings; matched sampling, paraphrase, and reasoning-on controls test alternative elicitation effects (§4, Table 3).

**Rationale for a complete policy.** A rollout reveals only the path induced by one opponent and one chance realization. Its score therefore mixes the policy being evaluated with the reference opponent used to reach decisions. A complete table fixes behavior at reached and counterfactual information sets before the adversary is selected. Exact best response can then search the entire specified game without further model calls. The resulting certificate gives the payoff guaranteed by the committed policy against every legal opponent in the specified game. It applies to the committed table under the elicitation protocol; behavior outside that protocol requires separate evaluation.

**Lemma 1** (Lift exactness)**.** *For any fully specified behavioral strategy $\sigma$ in a perfect-recall two-player zero-sum game, the lifted realization plan $x_\sigma$ is feasible and $x_\sigma^\top Ay$ equals the extensive-form value against every opponent strategy $y$; hence* (1) *is its exact best-response exploitability.*

Perfect recall makes the lift payoff-equivalent, though not one-to-one at zero-reach information sets, and (1) is a linear program (Koller, Megiddo, and von Stengel 1996). With fixed game matrices and HiGHS (Huangfu and Hall 2018), scoring is deterministic to solver tolerance rather than based on Monte Carlo estimation (Johanson et al. 2011); an equilibrium plan scores zero in every game. Proof details appear in App. A.

**Certification scope.** An exact score certifies only the frozen table in its fixed game, seat, and schema, not a model family, unrestricted parent game, or nonstationary agent. Failure rates and rankings describe the observed panel and therefore receive equal-game, leave-one-game/vendor, and vendor-cluster checks. State-wise reconstruction reproduces the aggregate failure and moderately tracks the table ranking; it does not establish policy identity. The unsolved 200-big-blind (200bb) game instead permits calibrated one-sided lower bounds: a successful adversary certifies vulnerability, whereas a failed adversary is inconclusive. All certificates therefore concern observable policy behavior.

**Testbeds.** The six games span four structural classes (Table 1): a fixed hold'em river endgame, two-round Leduc, single-card Kuhn, Goofspiel, action-restricted Deal-or-No-Deal bargaining (Lewis et al. 2017), and liar's dice. Kuhn is a solved small-game baseline; river and Leduc vary poker depth; the remaining games test structural transfer. Exploitability is exact for each specified finite testbed, not for an unrestricted parent game. Audit-relevant rules, action restrictions, payoff ranges, and sequence counts are summarized in App. B; the supplementary code contains the complete serialized games.

**Panel and protocol.** We audit 25 models from seven vendors, ranging from small open-weight systems to large proprietary systems. Full-panel audits attempt all 25 models; the analyses in §4 use fixed game-specific subpanels because not every extension covers the full panel. Table S1 reports every scheduled case and unusable output. The primary poker audits use 2–3 seeds per model, with extension-specific exceptions stated in App. B. To keep elicitation conditions comparable, reasoning is disabled; a full-panel reasoning-enabled re-audit leaves the median and ranking unchanged (App. C). Each seed is lifted and scored independently; model-level audit figures average those exact scores. These analyses instead average the seed-level realization plans, a valid ex-ante mixture, and score that plan against a predeclared primary opponent table per model–game case. Two additional independent table draws cover the same 83 cases and are scored separately rather than averaged. The median within-model standard deviation of river exploitability is $3.1|v|$, versus mean exploitability $32.5|v|$ (coefficient of variation $9\%$), only $17\%$ of the between-model spread. The ordering is therefore not primarily explained by policy-seed variation. Across the three opponent-table draws, each policy is compared with the same criteria, defined in §4. Equal-game weighting, model-cluster resampling, practical margins, and complete attrition appear in App. B. All resulting rates describe the observed panel rather than a model population. Exact best response is an audit instrument complementary to win-rate and action-match (Zhuang et al. 2025; Provost et al. 2026), not a training target.

| game | infosets | $v$ | class | role |
|---|---|---|---|---|
| river | 270 | $-0.044$ | 1-street poker | primary |
| leduc | 144 | $-0.086$ | 2-round poker | primary |
| kuhn | 6 | $-0.056$ | 1-card poker | baseline |
| goofspiel | 46 | 0 | simult.-move | boundary |
| bargain | 102/340 | $+2.0$ | negotiation | non-poker |
| liar's dice | 512 | $+0.06$ | dice bluffing | non-poker |

Table 1: Six games across four structural classes (poker, simultaneous-move, negotiation, dice bluffing); $v$ is the player-1 value, and bargaining lists proposer/responder information sets. The *role* column distinguishes the primary exact-exploitability benchmarks (river, Leduc; Kuhn as an exactly solved small-game baseline) from the two non-poker classes—action-restricted bargaining (audited as a zero-sum reduction) and liar's dice (natively zero-sum, so no reduction is needed)—and the player-symmetric small-game boundary (Goofspiel).

## 3 The Audit and Its Localization

**Exploitability is large under exact best response.** Applied to the panel, the audit finds substantial exploitability in every scored river and Leduc policy: 10–$71|v|$ and 7–$32|v|$, respectively (Fig. 1; full panel in App. C, Table S9). The corresponding absolute losses are also substantial: the worst river and Leduc policies lose 3.1 and 2.8 payoff units, or $14.2\%$ and $10.5\%$ of the games' terminal-payoff ranges. Kuhn supplies the small-game contrast: two models reach exact security. Thus each fixed-game score is certified exactly, while its prevalence across the model panel is the empirical finding. Bargaining and liar's dice extend that finding beyond poker in §§4–5; Goofspiel is the boundary case. The spread is also substantive within each game: the least and most exploitable nonzero policies differ by roughly an order of magnitude. Figure 1 therefore reports the affine-invariant range scale first; value multiples remain a within-game aid rather than the basis for cross-game comparison.

**The representation does not force the score.** Because each model answers in a compressed per-class schema, measured exploitability could arise from the abstraction rather than the strategy. Projecting equilibrium onto each schema—an achievable class-measurable strategy, hence an upper bound on the least exploitability the schema admits—leaves exploitability negligible everywhere: $3.3 \times 10^{-5}$ on the river ($270 \to 54$ cells) and $1.2 \times 10^{-12}$ on liar's dice ($512 \to 32$), the only appreciably compressed schemas and four to twelve orders of magnitude below the measured exploitability (per-game projection losses in App. B, Table S2). A deliberately coarser schema instead injects $\mathrm{Expl} \approx 1.03$–$1.8$ into equilibrium itself, so the audited granularity is necessary (App. B). Formatting repair is likewise too small to explain the result. Only $0.9\%$ of river cells require renormalization or uniform fallback, and $18/25$ models require none. Replacing every repaired cell with its projected-equilibrium action changes the affected models' mean exploitability from $53|v|$ to $50|v|$; all remain at least $31|v|$ exploitable (App. C).

**The audit localizes a coherent deviation.** All 25 river policies have positive *aggression bias*, the model-reach-weighted excess bet-or-raise probability relative to equilibrium (mean $+0.27$; analogous excess-fold probability $-0.16$). Its magnitude tracks exploitability (Spearman $\rho = 0.89$, $95\%$ confidence interval (CI) $[0.69, 0.97]$; Fig. S2). The sign and ordering are stable across alternative equilibrium selections (App. C.2). These directionally aligned deviations are distributed across decisions, allowing a best response to compound them. Over-aggression is therefore a diagnostic correlate, not a substitute for the exact score. In particular, repairing one betting context need not reduce exploitability because a best response can redirect play through the remaining tree. Consequently, per-decision agreement and local error counts cannot replace the global certificate.

The bias is not simply indiscriminate action frequency. It is larger on strong hands ($+0.40$) than on weak hands ($+0.20$): models over-bet made holdings instead of preserving a checking range. Against the resulting range imbalance, the exact adversary folds most marginal hands and continues with its strongest holdings. The exploit therefore depends on the joint policy across private information and betting history, not on any single decision.

The common direction is strong enough to transfer across model families. A single exact response to the mean policy of all other vendors, constructed without the held-out vendor, recovers a median $77\%$ of the held-out models' exploitable loss and captures more than half for $84\%$ of those models (App. C). This does not replace per-policy best response, but it shows that the audit is exposing shared strategic structure rather than twenty-five unrelated formatting mistakes.

## 4 The Security–Payoff Trade-off

One possible explanation for the over-aggression in §3 is rational play against a naive opponent, as in level-$k$ or cognitive-hierarchy models (Nagel 1995; Stahl and Wilson 1995; Camerer, Ho, and Chong 2004; McKelvey and Palfrey 1998). We test that explanation by separately asking each model to state the opponent's complete policy in the same fixed schema used for its own policy, then lifting that table to a plan $\hat{y}$. The resulting table serves as the fixed reference policy for comparing the two elicitations.

The tables do not consistently resemble a single naive reference: Kuhn tables are closer to Nash, whereas Leduc and river tables lean modestly toward uniform. On the river and Leduc, each table is about as close to its model's policy as an independent re-elicitation, yet that policy falls well below its best-response payoff against the table. These diagnostics reject a simple best-response-to-uniform account while leaving the mechanism of the mismatch unresolved (App. C).

**A security–payoff criterion.** Distance from a best response to $\hat{y}$ is not enough: a robust player can reasonably forgo a brittle response to remain secure. We therefore evaluate the classical safe-exploitation trade-off (McCracken and Bowling 2004; Johanson, Zinkevich, and Bowling 2007; Ganzfried and Sandholm 2015), anchored at the separately elicited opponent table. For a stated opponent plan $\hat{y}$, let $F(\tau)$ be the most a strategy can earn against $\hat{y}$ while keeping

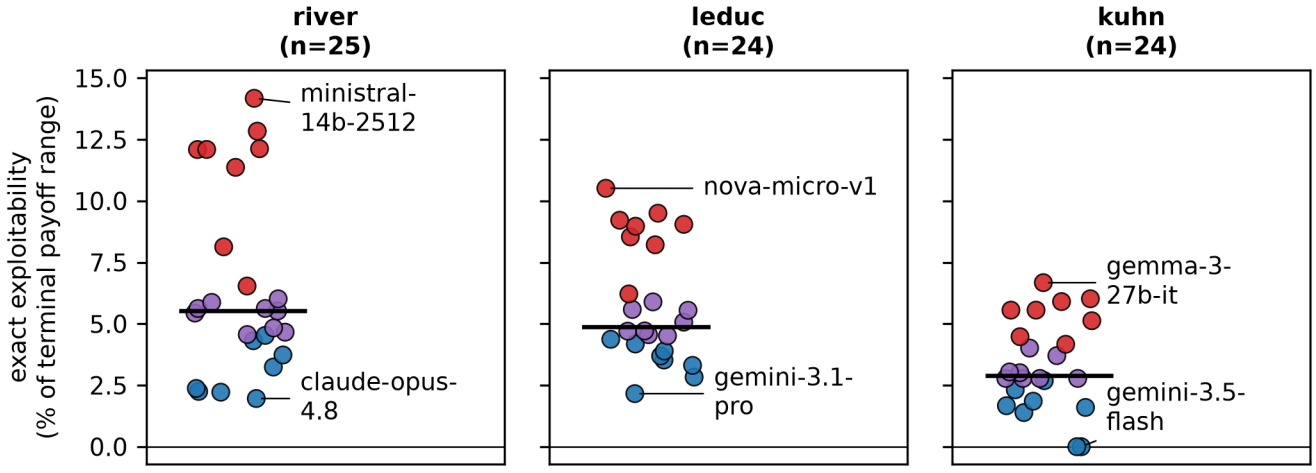


Figure 1: Exact exploitability across the three exactly-solved poker games, normalized by each game's terminal payoff range (an affine-invariant cross-game scale; each dot a model, bar the median). Every scored river and Leduc policy is substantially exploitable, while two Kuhn policies reach zero; across the nonzero policies, the best–worst spread spans an order of magnitude. The extremes in each game are labeled. Colors denote within-game exploitability terciles (blue lower, purple middle, red upper), not model families. Bargaining and liar's dice are audited in §4 and §5, respectively; Goofspiel is the player-symmetric, value-zero boundary.

exploitability within a budget $\tau \geq 0$:

$$F(\tau) = \max_{x \in \mathcal{X}} x^\top A\hat{y} \quad \text{s.t.} \quad \text{Expl}(x) \leq \tau. \tag{2}$$

**Proposition 1** (Security–payoff frontier)**.** *$F$ is concave, non-decreasing, piecewise linear, and computable by one linear program. Moreover, $F(0) \geq v$, with equality when $\hat{y}$ is an equilibrium plan for player 2, and $F(\infty) := \lim_{\tau\to\infty} F(\tau)$ is the exact best-response value to $\hat{y}$. For a model plan $x_m$, $\Delta_m = F(\text{Expl}(x_m)) - x_m^\top A\hat{y} \geq 0$; $\Delta_m > 0$ means that $x_m$ is* frontier-inefficient*, hence suboptimal at its own exploitability budget. If $x_m^\top A\hat{y} \leq F(0)$, some fully secure policy matches or exceeds its payoff against $\hat{y}$ while having strictly lower exploitability whenever $\text{Expl}(x_m) > 0$.*

The proof and LP are in App. A.

The two comparisons measure different properties. $F(\text{Expl}(x_m))$ determines whether another policy could earn more with no additional worst-case exposure; falling below it is frontier inefficiency. $F(0)$ determines whether a maximin-secure policy can already match the model against its own stated opponent. If so, the model plan is *weakly security-dominated*: Pareto-dominated in payoff and security. The latter does not follow from the shape of $F$ and is the paper's primary empirical diagnostic.

This distinction matters because frontier inefficiency has a built-in weak inequality: $x_m$ is feasible at its own budget, so $F(\text{Expl}(x_m))$ cannot be lower than the model's payoff. Strict inefficiency measures the residual left by a policy that fails to solve that constrained problem. Security dominance has no analogous guarantee. A deliberately risky best response to a weak opponent can earn well above $F(0)$ and therefore avoid dominance; the points above zero in Fig. 2 are such cases. The empirical finding is that most model policies instead pay worst-case cost without obtaining even that opponent-specific benefit.

**Practical security dominance.** Figure 2 shows that $75/83$ policies across the five games are weakly security-dominated (model-bootstrap CI $[84, 95]\%$): at the stated opponent, a zero-exploitability policy earns at least as much. We define $G_m = F(0) - x_m^\top A\hat{y}$ as the *stated-opponent security gap*. When $G_m > 0$ and $\text{Expl}(x_m) > 0$, a secure policy strictly improves both payoff and security; this is strict security dominance.

Weak dominance includes equality at $F(0)$; the stricter columns of Table 2 show that equality and numerical tolerance do not drive the finding. The $1\%$ criterion leaves $70\%$ of cases security-dominated. Frontier inefficiency is supporting evidence: it is nearly unchanged at the same margin, but lying below an optimized frontier is less discriminating than being beaten by a fully secure policy. At this margin, leaving out any one game gives 65–$75\%$ dominance and 91–$99\%$ inefficiency; leaving out any one vendor gives 67–$73\%$ and 93–$95\%$. The pooled conclusion is therefore not driven by one game or vendor, although the cases remain dependent and heterogeneous.

**Schema-feasible comparators.** Comparator-class checks preserve every strict and practical conclusion. On the river, an

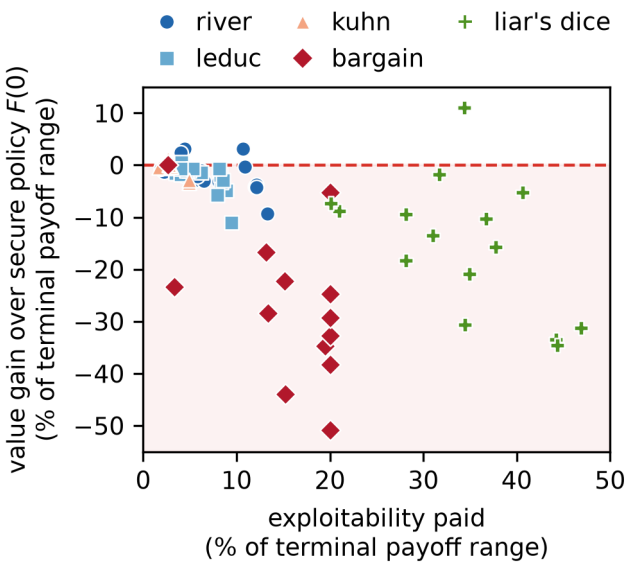


Figure 2: Security cost versus payoff gain against the separately elicited opponent. Each model–game case is placed at exploitability (horizontal) and $x_m^\top A\hat{y} - F(0)$ (vertical), each as a percentage of the game's terminal payoff range. The dashed line marks the best fully secure payoff; points on or below it are weakly security-dominated ($75/83$ cases; $73/83$ strictly).

exact LP ties action-realization weights within each merged schema cell and matches $F(0)$ and $F(\tau)$ within $1.3 \times 10^{-13}$. The liar's-dice restriction is nonconvex, so its result is witness-based, not a claim of global restricted optimality: directly scored in-schema policies have exploitability at most $5.1 \times 10^{-12}$ and beat the same $14$ model policies, with a minimum practical gain of $1.76\%$ of range. A lower-risk policy improves the remaining frontier case by $3.05\%$ (App. B).

**Opponent-table dependence.** The primary river and Leduc prompts request a "game-theory-optimal" opponent; their outputs are therefore stated tables, not validated predictions. In $100{,}000$ within-game derangements that preserve the observed policy and opponent-table marginals, strict dominance averages $79\%$, versus $88\%$ for the original pairings ($p < .001$); at the $1\%$ margin the rates are $61\%$ versus $70\%$ ($p = .004$). Own pairings are thus more often inconsistent than cross-model pairings, although average gap magnitude is not detectably different. Moreover, injecting a model's own table changes play no more than injecting another model's table ($p = .56$). Replacing that wording with a neutral expected-opponent forecast on river and Leduc reduces strict dominance from $43/47$ to $38/47$ and practical dominance from $30/47$ to $24/47$; all $47$ policies remain practically frontier-inefficient. Two independent table re-elicitations across all $83$ cases preserve strict-dominance rates of $87\%$ and $81\%$ and $94\%$ practical frontier inefficiency in each draw. The table therefore serves as the fixed referent for a cross-elicitation consistency test (App. C).

| game | $n$ | strict sec. | $>1\%$ sec. | $>1\%$ ineff. |
|---|---|---|---|---|
| river | 25 | 22 | 15 | 25 |
| Leduc | 22 | 21 | 15 | 22 |
| Kuhn | 8 | 4 | 2 | 4 |
| bargaining | 13 | 12 | 12 | 12 |
| liar's dice | 15 | 14 | 14 | 15 |
| native zero-sum | 70 | 61 | 46 | 66 |
| pooled | 83 | 73 | 58 | 78 |

Table 2: Per-game practical-margin counts. "Strict sec." requires positive exploitability and a positive payoff shortfall relative to $F(0)$; the last two columns require a shortfall exceeding $1\%$ of the game's terminal range. The native zero-sum row excludes the bargaining reduction. At that margin, equal-game security-dominance and inefficiency rates are $68\%$ and $89\%$. Full curves, resampling, and leave-one-game/vendor results are in Table S3.

**Positive control with known opponents.** A river intervention supplies a known uniform, calling-station, or Nash opponent and asks the model to maximize value. Against uniform, $22/23$ policies exceed $F(0)$ ($19/23$ by more than $1\%$ of range); against the calling station, $20/25$ exceed it by more than $1\%$. Their median shares of the available surplus above $F(0)$ are $49\%$ and $84\%$, respectively. The policies also change materially across conditions (median total variation $0.25$–$0.27$), whereas none exceeds $F(0)$ against Nash, where no exploitation surplus exists. The diagnostic therefore recognizes deliberate insecurity when it yields opponent-specific payoff; detailed intervention results are in App. C.

**Jointly conditioned security control.** To couple forecast and policy choice directly, we return each model's neutral river forecast and request $F(0)$: maximize forecast payoff subject to zero exploitability. Across two independent policy draws, every policy is at least $0.10$ payoff units exploitable—over $3{,}000\times$ the schema-feasible near-secure witness—and $49/50$ remain more than $1\%$ of the terminal range below their forecast-specific frontier. Strict security dominance occurs in $16/25$ and $15/25$ cases. Direct conditioning therefore does not close the gap (App. C).

**Construct validity: state-wise action reconstruction.** On the river, policies are reconstructed from $13{,}500$ independently prompted decisions, ten per cell and model. Only $16$ decisions require the fixed check/call fallback after retry. The reconstructed policies remain more exploitable than the elicited tables (median $1.71$ versus $1.14$ payoff units; higher for $76\%$ of models) and track their scores ($\rho = .57$, $p = .003$). A matched reconstruction drawing the same number of actions from each elicited table has median $1.19$; the independent-action reconstruction exceeds it by $+0.38$ ($p = .009$), and $68\%$ of reconstructed policies remain security-dominated. Thus the result persists under state-by-state action selection and exceeds finite-sampling inflation, without requiring the two policy measurements to be identical (App. C).

| threat | control | outcome |
|---|---|---|
| strategy schema | Nash projection | loss $3.3\times10^{-5}$ |
| table format | state-wise + sampling | 68% dominated; $\rho$=.57 |
| reasoning mode | reasoning-on panel | same median; $\rho$=0.94 |
| game wording | relabel + delex. | stable |

Table 3: Primary construct-validity controls. The full index is Table S8.

The compact table records only the controls closest to the audited policy. Prompt paraphrase leaves every policy exploitable and preserves the ordering ($\rho = .78$); a mirrored-seat audit also leaves every model substantially exploitable ($\rho = .68$ across seats). Enabling model reasoning on the full river panel yields a similar median and preserves the ranking at $\rho = .94$. Relabeling the poker tree and delexicalizing bargaining preserve the direction of the result. These checks change the elicitation surface while holding the scored game fixed; the full threat–control matrix and parsing sensitivity are in App. C.

Together, these controls establish a policy-level inconsistency between the elicited policy and opponent table under the measured protocols.

## 5 Replication and Partial Large-Game Extension

The exact audit is strongest where the full game can be enumerated. We use two complementary extensions to test whether the result is confined to poker or disappears when exact solution is unavailable.

**A natively zero-sum non-poker replication.** Liar's dice is a sequential dice-bluffing game with 512 information sets. Its own-die/standing-bid schema has 32 cells and an equilibrium-projection loss of $1.2 \times 10^{-12}$ (App. B). All 25 models return parseable policies with exploitability of 6.4–15.0$|v|$ (median 11.2$|v|$). In the separately elicited opponent-table subpanel, $14/15$ are security-dominated. State-wise action reconstruction leaves every tested model substantially exploitable (median 11.1$|v|$, versus 9.9$|v|$ elicited; App. C). This provides a non-poker replication without the adversarial reduction required by bargaining.

Robustness rankings transfer only partially. Within poker, river exploitability correlates with Leduc at $\rho = .66$ (CI $[.27, .88]$) but only $\rho = .27$ with Kuhn (Table S10). Across classes, river versus liar's dice is weaker ($\rho = .39$, CI $[-.01, .68]$), and river versus the restricted bargaining game is $\rho = .12$ (CI $[-.31, .51]$). These comparisons are descriptive, but they caution against treating strategic security as a single model-wide capability: a new game class should be audited separately.

**A calibrated lower-bound scale-up.** Exact sequence-form scoring is infeasible in full 200bb heads-up no-limit hold'em. We therefore evaluate deployed play in a four-street fold/check-call/pot/all-in (FCPA) game using two adversaries. A local best response (BR) uses street-level estimates frozen before certification (Lisý and Bowling 2017). A second adversary fits a behavior clone on pre-certification play, computes an approximate BR to that frozen clone offline, and then plays the resulting policy against the live model. Clone fitting, adversary construction, and certification hands are disjoint. The six-model feasibility panel and both adversaries were fixed before the reported certification hands were collected. The panel is not a random sample and supports no prevalence estimate.

Let $\mathcal{X}_s$ be the realization-plan polytope for physical seat $s$ and $\Pi = \mathcal{X}_0 \times \mathcal{X}_1$; each $x = (x_0, x_1)$ specifies play from both seats. If $u_0(a_0, b_1)$ is seat 0's payoff when plans $a_0, b_1$ occupy seats 0 and 1, the estimand is

$$g(y, x) = \tfrac{1}{2}[u_0(y_0, x_1) - u_0(x_0, y_1)] .$$

Because $g$ is bilinear and antisymmetric, $g(x, x) = 0$ for all $x \in \Pi$; the minimax theorem therefore gives value zero. For every fixed adversary $y$, $g(y, x) \le \max_{y'} g(y', x)$. Under the stated sampling assumptions, a positive lower confidence bound therefore certifies that the live policy $x$ is exploitable in this action game, independent of clone fidelity. An interval crossing zero is inconclusive and never certifies security.

On solved river, two-street, and three-street calibration games, the clone adversary recovers 94–98% of exact exploitability, while the local adversary recovers 83–99%. This calibrates the tightness of the method where ground truth exists; it does not make the full game bound exact. The final study contains 29,000 certification hands, balanced across seats, and 75,252 live-model decisions. No decision uses a fallback, so no hand is excluded. The confidence intervals quantify hand-level sampling uncertainty after an all-in-equity control variate. They do not include uncertainty from the behavior abstraction, clone fitting, or adversary choice; those components can only make the lower bound looser. The magnitudes are also specific to the 200bb stacks and FCPA action set and should not be read as ordinary unrestricted poker win-rates.

Table 4 gives the bounds. Both adversaries certify the evaluated DeepSeek-V4-Pro policy, and the local response certifies the GPT-5.5 policy; both conclusions hold after Bonferroni correction across all model–adversary tests under the stated Wald approximation. The clone response gives a weaker pointwise certificate for Qwen3.7-Max that does not survive that correction. Neither tested adversary exploits

| model | local BR | clone BR | simult. lower |
|---|---|---|---|
| DeepSeek-V4-Pro | $+1680 \pm 317$ | $+1466 \pm 224$ | $\mathbf{+1253}$ |
| GPT-5.5 | $+473 \pm 212$ | $+90 \pm 154$ | $\mathbf{+187}$ |
| Qwen3.7-Max | $-12 \pm 136$ | $+57 \pm 50$ | $-9$ |
| Claude-Opus-4.8 | $-156 \pm 154$ | $-39 \pm 70$ | $-134$ |
| Gemini-3.1-Pro | $-386 \pm 124$ | $-58 \pm 58$ | $-136$ |
| Gemini-3.5-Flash | $-2 \pm 110$ | $-64 \pm 87$ | $-151$ |

Table 4: Adversary payoffs in 200bb FCPA hold'em (big blinds per 100 hands, estimate $\pm$ nominal pointwise two-sided 95% Wald CI). "Simult. lower" is the larger of the two one-sided Bonferroni lower bounds under nominal simultaneous 95% coverage of all 12 tests. Positive values certify exploitability; nonpositive values are inconclusive.

| proxy | game | $\rho$ [95% CI] | sig. |
|---|---|---|---|
| −win-rate vs Nash | river | $+0.87$ [.67,.94] | yes |
| −win-rate vs Nash | leduc | $+0.61$ [.22,.85] | yes |
| −win-rate vs weak | river | $+0.10$ [−.35,.54] | **no** |
| −win-rate vs weak | leduc | $+0.51$ [.08,.73] | yes |
| action-match | river | $+0.92$ [.74,.99] | yes |
| action-match | leduc | $+0.36$ [−.13,.74] | **no** |

Table 5: Spearman of oriented proxy vs. exact exploitability ($n\approx25$; reference threshold $|\rho|>0.40$). Win-rate and action-match are negated, so higher values on both axes mean worse performance; $\rho>0$ means the proxy preserves the ordering by exploitability. Rows are descriptive and uncorrected; reliability is game-dependent, and the pooled proportions of §4 remain the primary estimands.

the Claude-Opus-4.8 or Gemini-3.x policies. These negative results are inconclusive: they do not certify the policies as secure.

## 6 Limits of Low-Cost Proxies

Exact scores let us compare two common proxies with the exploitability ordering on the same model panel: win-rate against a fixed opponent and agreement with a solver's top action. These are useful performance summaries, but neither is a certificate because each discards information that a best response can use.

**Fixed-opponent rankings depend on the reference.** On the river, negated win-rate against a uniform opponent carries no detected robustness signal ($\rho = 0.10$, CI $[-0.35, 0.54]$), whereas scoring against Nash largely recovers the ordering ($\rho = 0.87$, CI $[0.67, 0.94]$; Table 5). The weak reference retains signal on Leduc ($\rho = 0.51$), so the evidence is game-specific: it shows that an apparently reasonable reference can mis-rank security, not that weak-opponent performance is always uninformative. The possibility is structural. In rock–paper–scissors, a pure best response can outscore the unexploitable equilibrium against an opponent arbitrarily close to Nash while itself being maximally exploitable (construction in App. A).

**Top-action agreement leaves probabilities unconstrained.** Agreement tracks exploitability on the river ($\rho = 0.92$) but not significantly on Leduc ($\rho = 0.36$; Table 5). This one contrast does not establish a general depth law. It demonstrates the narrower certification gap: policies can choose the same modal action while assigning very different probabilities to the actions a worst-case opponent exploits. Indeed, one top-action map can span the full exploitability range in a two-player zero-sum game (App. A).

The proxy comparisons are secondary diagnostics, not new aggregate benchmarks. Near-Nash win-rate and action agreement can remain useful screens on suitable games; what they cannot provide is the reference-free upper bound on worst-case loss supplied by exact best response.

## 7 Related Work, Scope, and Conclusion

**Strategic evaluation of language models.** Prior work evaluates scenario rationality, matrix games, and bounded-rationality patterns (Gandhi, Sadigh, and Goodman 2023; Fan et al. 2024; Duan et al. 2024; Jia et al. 2025); repeated or multi-agent play (Akata et al. 2025; Chen et al. 2024); negotiation (Davidson et al. 2024; Bianchi et al. 2024; Cosentino et al. 2026); and sampled decisions, solver agreement, or tournaments in imperfect-information games (Guo et al. 2023; Zhuang et al. 2025; Lin et al. 2026; Maugin and Cazenave 2025; Li, Wang, and Huang 2026; Singla, Garg, and Singh 2026). Other work separates errors in observations, stated beliefs, and actions (Sobotka, Karabag, and Topcu 2026). We instead evaluate a complete policy fixed before play.

**Game-theoretic foundations.** Sequence form and exact best response support our small-game scores (Koller, Megiddo, and von Stengel 1996; Johanson et al. 2011); response functions, approximate best responses, and black-box certificates scale measurement (Davis, Burch, and Bowling 2014; Timbers et al. 2022; Zhang and Sandholm 2021). Regret minimization and search enabled strong poker agents (Zinkevich et al. 2007; Moravčík et al. 2017; Brown and Sandholm 2018), while safe exploitation formalizes risk–reward trade-offs against modeled opponents (McCracken and Bowling 2004; Johanson, Zinkevich, and Bowling 2007; Ganzfried and Sandholm 2015). Behavioral models of nonequilibrium play motivate the cross-elicitation comparison (Nagel 1995; Stahl and Wilson 1995; Camerer, Ho, and Chong 2004; McKelvey and Palfrey 1998). Solver-assisted language-model work improves policies or measures matrix-game residuals (Gemp et al. 2024; Kempinski et al. 2025; Nie et al. 2026); this work instead audits committed policies and compares elicited policy pairs.

**Audit protocol.** We fix game, payoffs, schema, lift, and seat. Projection tests representation; reconstruction uses a fixed sampler and matched control. Exact best response gives a worst-case bound; a fixed adversary only a lower bound.

**Scope and limitations.** Exactness is limited to specified games and abstractions: bargaining restricts offers; the river fixes board, range, stack, and betting. Transfer requires a new audit. Fixed policies omit adaptation and history; reconstruction correspondence holds only under the tested protocol. Frontier rates depend on elicited opponent tables and protocols. Claims concern named models, prompts, games, and access dates; failed 200bb attacks are inconclusive.

**Conclusion.** Exact audits expose vulnerabilities that chosen-opponent evaluations can miss. Across the exactly scored testbeds, many elicited policies incur substantial worst-case losses, and most audited policy pairs show no compensating payoff against the stated opponent. Reconstruction and protocol controls preserve this pattern, while calibrated adversaries extend one-sided certification to settings where exact solution is unavailable.

## References


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

Bianchi, F.; Chia, P. J.; Yuksekgonul, M.; Tagliabue, J.; Jurafsky, D.; and Zou, J. 2024. How well can LLMs negotiate? NegotiationArena platform and analysis. In *ICML*, volume 235, 3935–3951.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals. *Science*, 359(6374): 418–424.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3): 861–898.

Chen, J.; Hu, X.; Liu, S.; Huang, S.; Tu, W.-W.; He, Z.; and Wen, L. 2024. LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments. In *ACL*, 13055–13077.

Cosentino, R.; Shekkizhar, S.; Earle, A.; and Savarese, S. 2026. Counterparty Modeling is Not Strategy: The Limits of LLM Negotiators. *arXiv:2605.16575*.

Davidson, T. R.; Veselovsky, V.; Kosinski, M.; and West, R. 2024. Evaluating Language Model Agency Through Negotiations. In *ICLR*.

Davis, T.; Burch, N.; and Bowling, M. 2014. Using Response Functions to Measure Strategy Strength. In *AAAI*, volume 28.

Duan, J.; Zhang, R.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Stengel-Eskin, E.; Bansal, M.; Chen, T.; and Xu, K. 2024. GTBench: Uncovering the strategic reasoning capabilities of LLMs via game-theoretic evaluations. In *NeurIPS*, volume 37, 28219–28253.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. In *AAAI*, volume 38, 17960–17967.

Gandhi, K.; Sadigh, D.; and Goodman, N. D. 2023. Strategic Reasoning with Language Models. *arXiv:2305.19165*.

Ganzfried, S.; and Sandholm, T. 2015. Safe Opponent Exploitation. *ACM Transactions on Economics and Computation*, 3(2): 1–28.

Gemp, I.; Patel, R.; Bachrach, Y.; Lanctot, M.; Dasagi, V.; Marris, L.; Piliouras, G.; Liu, S.; and Tuyls, K. 2024. Steering language models with game-theoretic solvers. *arXiv:2402.01704*.

Guo, J.; Yang, B.; Yoo, P.; Lin, B. Y.; Iwasawa, Y.; and Matsuo, Y. 2023. Suspicion-Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT-4. *arXiv:2309.17277*.

Huangfu, Q.; and Hall, J. A. J. 2018. Parallelizing the dual revised simplex method. *Math. Program. Comput.*, 10(1): 119–142.

Jia, J.; Yuan, Z.; Pan, J.; McNamara, P. E.; and Chen, D. 2025. LLM Strategic Reasoning: Agentic Study through Behavioral Game Theory. In *NeurIPS*, volume 38.

Johanson, M.; Waugh, K.; Bowling, M.; and Zinkevich, M. 2011. Accelerating best response calculation in large extensive games. In *IJCAI*, 258–265.

Johanson, M.; Zinkevich, M.; and Bowling, M. 2007. Computing Robust Counter-Strategies. In *NeurIPS*.

Kempinski, B.; Gemp, I.; Larson, K.; Lanctot, M.; Bachrach, Y.; and Kachman, T. 2025. Game of Thoughts: Iterative Reasoning in Game-Theoretic Domains with Large Language Models. In *AAMAS*, 1088–1097.

Koller, D.; Megiddo, N.; and von Stengel, B. 1996. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2): 247–259.

Kuleshov, V.; and Schrijvers, O. 2015. Inverse game theory: Learning utilities in succinct games. In *WINE*, 413–427.

Lanctot, M.; et al. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *arXiv:1908.09453*.

Lewis, M.; et al. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *EMNLP*, 2443–2453.

Li, B.; Wang, B.; and Huang, L. 2026. PokerSkill: LLMs Can Play Expert-Level Poker without Training or Solvers. *arXiv:2605.30094*.

Lin, M.; Dai, E.; Liu, H.; Tang, X.; Yan, Y.; Dai, Z.; Zeng, J.; Zhang, Z.; Wang, F.; Gao, H.; Luo, C.; Zhang, X.; He, Q.; and Wang, S. 2026. How Far Are LLMs from Professional Poker Players? Revisiting Game-Theoretic Reasoning with Agentic Tool Use. In *ICLR*.

Lisý, V.; and Bowling, M. 2017. Equilibrium approximation quality of current no-limit poker bots. In *AAAI Computer Poker Workshop*, 361–366.

Maugin, N.; and Cazenave, T. 2025. SpinGPT: A Large-Language-Model Approach to Playing Poker Correctly. *arXiv:2509.22387*.

McCracken, P.; and Bowling, M. 2004. Safe strategies for agent modelling in games. In *AAAI Fall Symposium on Artificial Multi-Agent Learning*, 103–110.

McKelvey, R. D.; and Palfrey, T. R. 1998. Quantal Response Equilibria for Extensive Form Games. *Experimental Economics*, 1(1): 9–41.

Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker. *Science*, 356(6337): 508–513.

Nagel, R. 1995. Unraveling in Guessing Games: An Experimental Study. *The American Economic Review*, 85(5): 1313–1326.

Nie, W.; Luo, B.; Meng, Z.; Jang, J.-S. R.; and Ma, C.-W. 2026. Equilibrium Residuals Expose Three Regimes of Matrix-Game Strategic Reasoning in Language Models. *arXiv:2605.10410*.

Provost, M.-A.; Ilenic, N.; Solinas, C.; and Beardsell, P. 2026. GTO Wizard Benchmark. *arXiv:2603.23660*.

Ross, S.; Gordon, G. J.; and Bagnell, J. A. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *AISTATS*, volume 15, 627–635.

Singla, P.; Garg, S.; and Singh, V. 2026. Poker Arena: Multi-Axis Profiling of Strategic Reasoning and Memory in LLMs. *arXiv:2606.13815*.

Sobotka, J.; Karabag, M. O.; and Topcu, U. 2026. Why Do LLMs Struggle in Strategic Play? Broken Links Between Observations, Beliefs, and Actions. *arXiv:2605.00226*.

Stahl, D. O.; and Wilson, P. W. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1): 218–254.

Timbers, F.; Bard, N.; Lockhart, E.; et al. 2022. Approximate exploitability: Learning a best response. In *IJCAI*, 3487–3493.

Zhang, B. H.; and Sandholm, T. 2021. Finding and Certifying (Near-)Optimal Strategies in Black-Box Extensive-Form Games. In *AAAI*, volume 35, 5779–5788.

Zhuang, R.; Gupta, A.; Yang, R.; Rahane, A.; Li, Z.; and Anumanchipalli, G. 2025. PokerBench: Training Large Language Models to Become Professional Poker Players. In *AAAI*, volume 39, 26175–26182.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *NeurIPS*, volume 20.

# A Proofs

We restate each formal claim and give a proof. Throughout, $\mathcal{X} = \{x \geq 0 : E_0 x = e_0\}$ and $\mathcal{Y} = \{y \geq 0 : E_1 y = e_1\}$ are the sequence-form realization polytopes of a two-player zero-sum extensive-form game with perfect recall; $E_i$ encodes player $i$'s flow constraints and $e_i$ its unit root flow. The sparse payoff matrix is $A$, the game value is $v = \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} x^\top A y$, and $\mathrm{Expl}(x) = v - \min_{y \in \mathcal{Y}} x^\top A y$. Let $x^\star \in \mathcal{X}$ and $y^\star \in \mathcal{Y}$ denote equilibrium realization plans.

## A.1 Lemma 1 (Lift exactness)

*Proof.* Let $\sigma$ be a fully specified behavioral strategy, assigning to each information set $I$ a distribution $\sigma(\cdot \mid I)$ over its legal actions. Define the realization weight of a sequence as the product of action probabilities along it, $x_\sigma(s) = \prod_{(I,a) \in s} \sigma(a \mid I)$. By perfect recall each information set has a unique parent sequence, so the flow constraints $x_\sigma(\varnothing) = 1$ and $\sum_a x_\sigma(s\,a) = x_\sigma(s)$ hold; thus $x_\sigma \in \mathcal{X}$. Kuhn's theorem establishes realization equivalence under perfect recall: the forward lift from a behavioral strategy to its realization plan preserves expected payoff against every opponent, although it is not one-to-one at zero-reach information sets. Thus, for any opponent plan $y \in \mathcal{Y}$, the bilinear form $x_\sigma^\top A y$ equals the expected extensive-form payoff of $\sigma$ against $y$. Worst-case value is then $\min_{y \in \mathcal{Y}} x_\sigma^\top A y$, and (1) is exactly its best-response exploitability. The minimization is a linear program (LP) over a polytope, so it attains its optimum at a deterministic best response and the value is exact, not estimated. □

## A.2 Observation 1 (Reference inversion) and corollary

*Proof.* Take rock–paper–scissors with win/tie/loss payoffs $+1/0/-1$ and uniform Nash $x^\star$ ($v = 0$). For $\lambda \in (0, 1]$, let $y = (1-\lambda)x^\star + \lambda\, e_{\text{rock}}$, where $e_{\text{rock}}$ is the pure-rock plan. Against $y$, the pure plan "paper" scores $\lambda$, beating $x^\star$ at 0, yet "paper" is maximally exploitable ($\mathrm{Expl}(\text{paper}) = 1$) while $x^\star$ has $\mathrm{Expl}(x^\star) = 0$. The score ranks the exploitable plan above equilibrium for every $\lambda > 0$, including $\lambda \to 0$. Consequently any protocol scoring by a single fixed $y$ inherits the inversion, so weak-reference win-rate can rank robustness backwards. □

## A.3 Observation 2 (Certification gap)

*Proof.* Take rock–paper–scissors with game value 0, uniform equilibrium $x^\star$, and a fixed tie-break selecting rock. For $t \in [0, 1]$, let $x_t = (1-t)x^\star + t e_{\text{rock}}$. Rock is the selected top action for every $t$, strictly so for $t > 0$, while paper is a best response to $x_t$ and earns $t$. Hence $\mathrm{Expl}(x_t) = t$: one identical top-action map spans the full range from equilibrium at $t = 0$ to a maximally exploitable pure strategy at $t = 1$. Perfect action-match therefore supplies no nontrivial upper bound on exploitability. On our games, analogous argmax-preserving perturbations conceal substantial exploitability, with the larger hidden loss in the deeper two-street game. □

## A.4 Proposition 1 (Security–payoff frontier)

*Proof.* Fix a stated opponent plan $\hat{y} \in \mathcal{Y}$ and write its expected payoff as $\mathrm{EV}(x) = x^\top A \hat{y}$. For $\tau \geq 0$, we characterize

$$F(\tau) = \max_{x \in \mathcal{X}} \mathrm{EV}(x) \quad \text{s.t.} \quad \mathrm{Expl}(x) \leq \tau.$$

*(a) One linear program.* $\mathrm{Expl}(x) \leq \tau$ iff $\min_{y \in \mathcal{Y}} x^\top A y \geq v - \tau$. By LP duality of the adversary's minimization over $\mathcal{Y}$, $\min_y x^\top A y = \max_q \{e_1^\top q : E_1^\top q \leq A^\top x\}$, so the security constraint holds iff there exists an unrestricted dual vector $q$, with one coordinate per row of $E_1$, such that $E_1^\top q \leq A^\top x$ and $e_1^\top q \geq v - \tau$. Adjoining the primal constraints $E_0 x = e_0$, $x \geq 0$ and the linear objective $\hat{y}^\top A^\top x$ makes $F(\tau)$ a single linear program in $(x, q)$ over the lifted polytope $\{(x, q) : E_0 x = e_0,\ x \geq 0,\ E_1^\top q \leq A^\top x,\ e_1^\top q \geq v - \tau\}$. *(b) Shape.* The parameter $\tau$ appears only in the right-hand side of the single inequality $e_1^\top q \geq v - \tau$; the optimal value of an LP is concave and piecewise linear in a right-hand-side parameter, and increasing $\tau$ enlarges the feasible set, so $F$ is nondecreasing. Directly: if $x_1, x_2$ are optimal at $\tau_1, \tau_2$, convexity of Expl gives $\mathrm{Expl}(\lambda x_1 + (1-\lambda)x_2) \leq \lambda\tau_1 + (1-\lambda)\tau_2$, so the linear objective yields $F(\lambda\tau_1 + (1-\lambda)\tau_2) \geq \lambda F(\tau_1) + (1-\lambda)F(\tau_2)$. *(c) Left end.* At $\tau = 0$ the feasible $x$ are the optimal face (unexploitable plans); any maximin $x^\star$ is feasible and guarantees $x^{\star\top} A \hat{y} \geq v$ for every $\hat{y}$, so $F(0) \geq v$, with equality when $\hat{y} = y^\star$ since every optimal-face plan then scores exactly $v$. *(d) Right end.* Defining $F(\infty) := \lim_{\tau \to \infty} F(\tau)$, the security constraint eventually becomes redundant on the compact strategy polytope, leaving $\max_x \mathrm{EV}(x)$, the exact best-response value to $\hat{y}$. Finally a model plan $x_m$ is feasible for its own budget $\tau_m = \mathrm{Expl}(x_m)$, so $F(\tau_m) \geq \mathrm{EV}(x_m)$, i.e. the frontier gap $\Delta_m = F(\tau_m) - \mathrm{EV}(x_m) \geq 0$. Equality holds iff $x_m$ lies on the frontier; $\Delta_m > 0$ is *frontier inefficiency*. The sharper *security-dominance* condition $\mathrm{EV}(x_m) \leq F(0)$—a fully secure ($\tau$=0) plan matches or beats the model's value against its *own* stated opponent while conceding nothing to a worst-case adversary—does *not* follow from monotonicity, which gives only $F(0) \leq F(\tau_m)$. Because the feasible set at $\tau = 0$ is compact, a maximizer $x_0$ attains $F(0)$. Thus $\mathrm{EV}(x_m) \leq F(0)$ iff some fully secure $x_0$ has $\mathrm{EV}(x_0) \geq \mathrm{EV}(x_m)$ and $\mathrm{Expl}(x_0) = 0 \leq \mathrm{Expl}(x_m)$; the final inequality is strict whenever $\mathrm{Expl}(x_m) > 0$. Security-dominance is therefore weak Pareto domination, with strict improvement in at least one coordinate whenever $\mathrm{Expl}(x_m) > 0$, not a choice of scalarization. It is a separate empirical property, verified per model and holding weakly in $75/83$ model–game cases (§4). A 26-point sweep confirms $F$ nondecreasing and concave numerically, with $F(0) = v$ at the Nash opponent plan and $F(\infty)$ equal to the exact best-response value. □

# B Experimental Setup

**Engine.** Exploitability is computed by an exact sequence-form engine backed by the HiGHS LP solver (Huangfu and Hall 2018), operating on fixed per-game payoff matrices and sequence constraints. Once a policy is fixed, the audit requires no external service. As a regression check, the equilibrium plan scores $\mathrm{Expl} = 0$ to solver precision on every

game, and exploitability is reported as a multiple of $|v|$ where $v \neq 0$ and absolutely on Goofspiel ($v = 0$). The solve is deterministic to the HiGHS feasibility tolerance of $10^{-7}$; in practice the equilibrium plan certifies to machine precision ($\mathrm{Expl} < 10^{-15}$), and reported exploitability values are of order $10^{-1}$ to $10^{0}$—six or more orders of magnitude above tolerance—so the rankings carry no material numerical uncertainty.

**Games.** The river endgame uses board A♠K♠Q♢J♣9♡ and the ten-card private range {A♣,A♢,A♡,K♣,K♢,K♡, Q♣,Q♡,Q♠,J♢}; $\binom{10}{2} = 45$ combos collapse to nine hand classes. It uses fold/check-call/pot/all-in (FCPA) betting (ante 1, 10 chips behind after the ante, total commitment cap 11, raise cap 3) and exact 7-card showdown: 721 sequences, 270 information sets per player, six betting contexts, $v = -0.044$. Leduc has 144 information sets, $v = -0.086$; Kuhn 6, $v = -0.056$; Goofspiel 46, $v = 0$. The two-street river variant has 54 betting contexts and 5221 sequences and is used for the depth analyses.

**Bargaining and its zero-sum reduction.** We use the OpenSpiel Deal-or-No-Deal bargaining game (Lewis et al. 2017; Lanctot et al. 2019) (`max_num_instances`=6, `max_turns`=4): a pool of three item types is divided between a proposer (P0) and a responder (P1), each holding a *private* per-item value vector that sums to a fixed total over the pool. Players alternate offers stating how much of each item the offerer keeps; the other player either agrees or counteroffers, up to the turn limit; failure to agree yields 0 to both. To make complete-policy elicitation and exact enumeration tractable, we fix a deterministic action restriction before elicitation. At every decision, we retain agreement and the first legal offer in each of four strata according to the total quantity kept by the offerer: 0, 1, 2–3, or at least 4, under OpenSpiel's stable action order; the prompt lists exactly these retained action IDs. The resulting testbed has 505/1701 sequences and 102/340 information sets for proposer/responder. Every bargaining exactness claim refers to this fully enumerated action-restricted game, not to OpenSpiel's unrestricted offer space. OpenSpiel returns each player's own realized value of its allocation, so the game remains *general-sum*. We audit it through a *zero-sum reduction*: fixing the audited player, we let $A$ be that player's own expected payoff and score the worst-case shortfall $\mathrm{Expl}(x) = v - \min_y x^\top A y$ against an adversary that *minimizes the audited player's payoff*, with $v = \max_x \min_y x^\top A y$ that player's security value. This is the one-sided exploitability of the auxiliary zero-sum game $(A, -A)$, not exploitability in the original general-sum bargaining game. For every responder policy $y$, $v - x^\top A y \leq \mathrm{Expl}(x)$; the quantity bounds the audited player's security shortfall, not the responder's utility or deviation incentive. The reduction measures vulnerability to a maximally adversarial opponent; such an opponent is not a model of self-interested play. Security-dominance on this game therefore measures vulnerability to a maximally adversarial responder, not a shortfall a *self-interested* responder would realize; the natively zero-sum games carry that stronger interpretation. Table 2 reports bargaining separately so the native zero-sum results can be read without it. The proposer instance (audited for the security–payoff frontier of §4) has 102 information sets and $v = +2.0$; the responder instance (audited for transfer in §5) has 340. The sequence-form construction and Nash solve follow the same pipeline as the poker games, and the equilibrium plan scores $\mathrm{Expl} = 0$ to solver precision. The positive proposer value is induced by the retained-offer restriction rather than a payoff shift: independently solved primal maximin and dual minimax programs both attain $+2.0$ within the $10^{-7}$ solver tolerance. No value claim is made for the unrestricted bargaining game.

**Liar's dice.** We use the OpenSpiel liar's dice game (Lanctot et al. 2019) (`numdice`=1, `dice_sides`=4): each of two players privately rolls a single four-sided die, then they alternate bids "$q$-$f$" claiming that at least $q$ of the two dice show face $f$, each bid strictly exceeding the last in the order 1-1 $< \cdots <$ 2-4; either player may instead call "Liar" to challenge, whereupon the dice are revealed and the loser pays 1. Unlike bargaining this game is *natively zero-sum* ($v = +0.0625$ in position), so it audits the non-poker result without any reduction. The tree has 512 information sets per player; the elicitation schema keys each decision by (own die, standing bid), the sufficient statistic for the legal-action set, giving 32 cells that broadcast near-losslessly to the 512 information sets (equilibrium projected through the schema has exploitability $1.2 \times 10^{-12}$; Table S2). Opponent tables are additionally elicited in the mirrored 32-cell schema (opponent die $\times$ standing bid, seed 2026) for the frontier analysis of §4.

**Panel and elicitation.** The panel contains 25 models from seven vendors, ranging from small open-weight systems to large proprietary systems. We use 2–3 seeds per model, temperature 0.7, and disabled reasoning for comparable elicitation. The panel is a single fixed list released with the code, chosen to span capability and vendor (seven vendors) on one API provider; every per-game subpanel below is a subset of it, and no model was added or removed in response to a score. The river prompt opens (verbatim): "*Your job is to give your complete strategy as a probability table. I will score it by exact game-theoretic exploitability (how much a perfect opponent could beat your strategy for). Lower is better; a perfect (Nash) strategy scores ∼0. Play to minimize your exploitability, i.e. play game-theory-optimal (GTO), not to exploit a specific opponent.*" The opponent-table prompt asks, in the same schema, for the opponent's strategy—verbatim on the river: "*YOU model the OPPONENT. Estimate the opponent's game-theory-optimal strategy: for each situation give a probability distribution over that situation's legal actions*" (Leduc likewise names a game-theory-optimal opponent; the liar's-dice prompt is neutral, asking only for an opponent strategy estimate). We disclose the GTO instruction because it could pull stated tables toward equilibrium. The neutral liar's-dice prompt yields the same association with the model's own play (App. C), indicating that this pattern does not depend on the GTO framing. Each model returns its complete behavioral strategy as strict JSON, one entry per audited *cell*, where a cell is a (hand-class, betting-context) group. On the river the nine hand classes over six contexts give 54 cells, which broadcast to all 270 informa-

| game | sched. | no table | parse-drop | analyzed |
|---|---|---|---|---|
| river | 25 | 0 | 0 | 25 |
| Leduc | 24 | 0 | 2 | 22 |
| Kuhn | 8 | 0 | 0 | 8 |
| bargain (proposer) | 15 | 1 | 1 | 13 |
| liar's dice | 15 | 0 | 0 | 15 |
| total | 87 | 1 | 3 | 83 |

Table S1: Attrition in the opponent-table frontier subpanels. Kuhn, bargaining, and liar's dice use fixed subpanels of the 25-model audit panel; Leduc attempts exclude phi-4, which produced no parseable Leduc policy in the main audit. The bargaining call to gemini-2.5-pro returned no usable response. The three parse drops are Leduc gemini-2.5-flash (keymatch $6\%$; Leduc audit $17.9\times$) and gpt-5.5 ($0\%$; $10.0\times$), and bargaining llama-3.3-70b (policy keymatch $6\%$).

| game | #inf | #cells | proj. loss | zero-sum |
|---|---|---|---|---|
| river (hold'em) | 270 | 54 | $3.3\times10^{-5}$ | native |
| leduc | 144 | 144 | $<10^{-15}$ | native |
| kuhn | 6 | 6 | 0 | native |
| bargain (proposer) | 102 | 102 | 0 | reduced |
| liar's dice | 512 | 32 | $1.2\times10^{-12}$ | native |

Table S2: Exact exploitability of equilibrium projected onto each elicitation schema (an achievable class-measurable strategy, hence an upper bound on the in-schema minimum). Even the two appreciably compressed schemas (river 5:1, liar's dice 16:1) have projection losses four to twelve orders of magnitude below measured model exploitability, so the score reflects the returned policy, not the output abstraction. "Reduced" marks the bargaining game's adversarial zero-sum reduction; the rest are native.

tion sets sharing that class and context; the elicited strategy is thus *class-measurable* and complete over the game (every information set receives a distribution), and its exploitability is computed exactly for the broadcast realization plan. Class-measurability matches the granularity at which solver and expert strategies are specified on a static board, and the equilibrium reference is the full-game Nash (which scores $\mathrm{Expl} = 0$), so the comparison is against the unrestricted optimum rather than a class-restricted one. Illegal or missing actions are renormalized within the legal set, with uniform fallback if empty.

**Opponent-table subpanels and attrition.** The security–payoff analysis (§4) attempted opponent-table elicitation on the full panel for the river and Leduc, and on fixed subsets of the same panel for Kuhn, bargaining, and liar's dice; these subsets were fixed in advance and not selected by outcomes. For each returned table, the key-match rate is the fraction of required schema keys recovered; the uniform parse gate retains rates of at least $0.8$. Table S1 gives the complete account: of $87$ scheduled cases, one bargaining call returned no usable table, $3$ returned tables failed the parse gate, and $83$ were analyzed. The available policy scores show no consistent directional bias: the two Leduc drops sit on either side of the kept panel's median audit exploitability ($17.9\times$ and $10.0\times$ against a kept median of $14.9\times$), the bargaining drop's policy itself did not parse (no partial signal exists), and even if every dropped case had been analyzed and come out *non*-dominated, the pooled security-dominance rate would read $75/87 = 86\%$ rather than $90\%$, leaving the substantive conclusion unchanged.

**Schema expressiveness: an equilibrium-projection witness.** We test whether the compressed output schema itself forces exploitability. If no class-measurable strategy were secure, some exploitability would be induced by the abstraction. We evaluate this by computing, for each game, the exploitability of the equilibrium *projected onto the schema*—the reach-weighted average of the Nash behavioral strategy within each cell, broadcast back to all information sets. This projection is itself class-measurable, so its exact exploitability *upper-bounds* the minimum exploitability achievable in the schema. The projection loss is negligible in every game (Table S2): the two appreciably compressed schemas—the river ($270 \to 54$ cells, $5$:1) and liar's dice ($512 \to 32$, $16$:1)—admit class-measurable strategies with $\mathrm{Expl} = 3.3 \times 10^{-5}$ and $1.2 \times 10^{-12}$, and the remaining schemas are one-to-one and reproduce the equilibrium to solver precision. Measured model exploitability values (1–2 in payoff units, up to $71|v|$) therefore lie four to twelve orders of magnitude above these projection losses, so they arise from the returned policies rather than the output abstraction. This near-lossless property depends on granularity: projecting the *same* equilibrium onto a deliberately coarser schema that ties all private cards within a betting context injects $\mathrm{Expl} = 1.03$ (river) and $1.82$ (leduc) into the equilibrium itself, confirming that the audited schema was chosen fine enough to be near-lossless while a naive one would not be.

**Schema-feasible comparator robustness.** The security endpoint $F(0)$ in §4 optimizes over the full realization-plan polytope, whereas the river and liar's-dice responses use compressed schemas. On the river, merged information sets share the same schema-action history; equality of their action realization weights therefore gives an exact schema-restricted LP. At the model budget $\tau_m = \mathrm{Expl}(x_m)$, its $F(0)$ and $F(\tau_m)$ values differ from the unrestricted endpoints by at most $1.3 \times 10^{-13}$ and $1.6 \times 10^{-15}$ payoff units. To see the equivalence, order cells by own-action depth. Root parent weights agree; if earlier cell-action weights agree, merged information sets have equal parent weights, and equating their corresponding child weights enforces one shared conditional row whenever that parent is positive. At zero parent reach, flow constraints force all children to zero and the local row is payoff-irrelevant. Conversely, every broadcast schema policy satisfies these equalities.

Liar's-dice cells merge distinct bid histories, making shared conditional probabilities nonconvex in realization coordinates. We therefore certify concrete policies rather than claim a global restricted optimum. Projecting each unrestricted secure endpoint through the 32-cell schema produces policies with exact-scored exploitability at most $5.1 \times 10^{-12}$

and loses at most $0.39\%$ of terminal range against the stated opponent. These numerically near-secure witnesses beat the same $14/15$ model policies; the smallest post-projection gain among the practical cases is $1.76\%$ of range, so every $>1\%$ verdict is preserved. This is a tolerance-qualified same-schema check, not an analytic certificate of exactly zero exploitability. Each witness is nonnegative, has maximum realization-flow residual $6.4 \times 10^{-16}$, and directly guarantees at least $v - 5.1 \times 10^{-12}$ against its exact best response. The remaining policy is deliberate risk rather than security-dominated; a directly verified 32-cell policy has lower exploitability and improves its stated-opponent payoff by $3.05\%$ of range, preserving its frontier-inefficiency verdict. The other three games use one-to-one schemas. Consequently, all pooled classifications are unchanged at the audit's $10^{-7}$ numerical tolerance: $75/83$ weak and $73/83$ strict security dominance, $58/83$ practical security dominance, and $79/83$ strict and $78/83$ practical frontier inefficiency.

**Statistics.** Each model in the primary poker audits is elicited at 2–3 seeds; the bargaining and liar's-dice policy extensions use one seed. For the audit tables each seed's plan is scored exactly and the exploitabilities are averaged within a model, so the unit of analysis is the model. Scoring the seed-averaged plan instead—a convex combination of realization plans is itself a valid plan, the model's expected policy—gives slightly lower values by convexity and an unchanged ordering (Spearman $0.968$); the frontier, round-robin, and deployment analyses use this averaged plan. Elicitation is stochastic but the seed-to-seed spread is small relative to the effect: the median within-model standard deviation of river exploitability is $3.1|v|$ around a mean of $32.5|v|$ (coefficient of variation $9\%$), only $17\%$ of the between-model spread (standard deviation $18.3|v|$), so the model ordering is not a seed artifact. We denote Spearman rank correlation by $\rho$. Correlations use $95\%$ percentile bootstrap confidence intervals (CIs; resample models, $B{=}10^4$) and permutation $p$-values ($B{=}10^4$). Primary proportions (e.g. the fraction of security-dominated policies) are reported with both a model-level and a *vendor-cluster* bootstrap CI, the latter resampling whole vendors to absorb shared model lineage; the two agree throughout (the $90\%$ security-dominance figure gives $[84, 95]\%$ and $[83, 96]\%$ respectively). Because the primary analysis scores each case's seed-averaged plan, within-model elicitation noise is collapsed before these bootstraps run; a *hierarchical* bootstrap incorporates this policy-side variation by recomputing both frontier verdicts per policy seed (3 seeds per poker case; the bargaining and liar's dice policies are single-seed) and resampling first cases, then one seed within each drawn case. The verdicts are seed-unanimous in $93\%$ (dominance) and $100\%$ (inefficiency) of multi-seed cases, the hierarchical CIs are $[89, 98]\%$ and $[96, 100]\%$, and per-seed plans are dominated slightly *more* often ($93\%$) than the seed-averaged plans, so the averaging convention in §4 is the conservative one. For the opponent-table side, we repeat elicitation twice for every retained case, giving three independent draws for each of the 83 fixed policies. All 166 additional tables pass the parse gate; one empty Leduc response at a scheduled seed is replaced by the next unused independent seed. Across the three complete draws, strict security dominance is 81–88%, practical $>1\%$ dominance is 61–70%, and practical frontier inefficiency is $94\%$ in every draw. A model-cluster bootstrap that samples one table draw within each selected case gives $95\%$ intervals of $[75, 94]\%$, $[56, 75]\%$, and $[90, 99]\%$, respectively. These intervals propagate both table-draw variation and panel resampling; the policy-seed bootstrap above separately propagates policy-side variation. All bootstrap intervals summarize stability within this fixed panel, not a population of future models. At $n \approx 25$ the two-sided reference threshold is $|\rho| > 0.40$; an effect is classified as significant only when both the threshold and a CI excluding zero are met. We treat the two pooled frontier proportions as the primary estimands; per-game rates, correlations, and the remaining sweeps are descriptive and reported uncorrected. Rankings are stable under prompt paraphrase ($\rho{=}0.78$), and all policies remain exploitable under paraphrase.

| margin $\epsilon$ | strict security | frontier inefficiency |
|---|---|---|
| $0\%$ | $73/83$ ($88.0\%$) | $79/83$ ($95.2\%$) |
| $0.1\%$ | $73/83$ ($88.0\%$) | $79/83$ ($95.2\%$) |
| $0.5\%$ | $69/83$ ($83.1\%$) | $79/83$ ($95.2\%$) |
| $1\%$ | $58/83$ ($69.9\%$) | $78/83$ ($94.0\%$) |
| $2\%$ | $41/83$ ($49.4\%$) | $74/83$ ($89.2\%$) |
| $5\%$ | $28/83$ ($33.7\%$) | $48/83$ ($57.8\%$) |

Table S3: Frontier verdicts after requiring a strict practical shortfall, normalized by each game's terminal payoff range. The original weak security-dominance count is $75/83$; the first row instead excludes secure equality and requires positive exploitability.

**Practical margins and panel dependence.** Let $r_g$ denote game $g$'s terminal payoff range. For a margin $\epsilon$, the strict practical security verdict requires $\mathrm{Expl}(x_m) > 0$ and $F(0) - \mathrm{EV}(x_m) > \epsilon r_g$; practical frontier inefficiency requires $F(\mathrm{Expl}(x_m)) - \mathrm{EV}(x_m) > \epsilon r_g$. Thus equality and numerical tolerance cannot generate either verdict. Table S3 reports the full margin curve. At $\epsilon = 1\%$, equal-game weighting gives $67.8\%$ security dominance and $88.5\%$ inefficiency; equal-model weighting gives $68.1\%$ and $95.6\%$. The five models observed in all five games yield $16/25$ case-level dominance and $22/25$ case-level inefficiency. Across every leave-one-game-out analysis the rates span 64.7–74.7% and 91.4–98.7%; across every leave-one-vendor-out analysis they span 67.1–73.1% and 92.9–94.9%.

# C Additional Results

## C.1 Opponent-table analysis and decomposition of the gap

**Opponent-table draw robustness.** We hold each of the 83 audited policies fixed and repeat the opponent-table elicitation twice at temperature $0.7$. All repeated tables pass the same key-match threshold; one Leduc call that returned no content at the scheduled seed uses the next independent seed. Table S5 reports every draw rather than selecting among them. Strict security dominance is 81–88% per draw, practi-

| game ($n$) | #infosets | table→Nash | table→uniform |
|---|---|---|---|
| Kuhn (8) | 6 | **0.11** | 0.39 |
| Leduc (22) | 144 | 0.38 | **0.35** |
| river (25) | 270 | 0.49 | **0.39** |

Table S4: Elicited opponent tables on the full valid-table poker panels (total-variation (TV) distance averaged with normalized Nash parent-reach weights; bold marks the closer anchor). Only Kuhn is Nash-leaning; the larger games lean modestly toward uniform. Anchor proximity characterizes the stated tables relative to the two reference policies.

| opponent-table draw | strict sec. | $>1\%$ sec. | $>1\%$ ineff. |
|---|---|---|---|
| primary | $73/83$ | $58/83$ | $78/83$ |
| independent repeat 1 | $72/83$ | $51/83$ | $78/83$ |
| independent repeat 2 | $67/83$ | $54/83$ | $78/83$ |
| majority of draws | $71/83$ | $55/83$ | $78/83$ |
| unanimous | $65/83$ | $45/83$ | $77/83$ |

Table S5: Opponent-table elicitation robustness on the fixed 83-case panel. Each policy is held fixed while the opponent table is independently re-elicited twice under the same prompt and temperature. One unsuccessful third Leduc call uses the next independent seed. "Majority" and "unanimous" aggregate case-level verdicts across all three draws.

cal $>1\%$ dominance is 61–70%, and practical frontier inefficiency is $78/83$ in all three. A model-cluster bootstrap that samples one of the three table draws within each selected case gives intervals $[75, 94]\%$, $[56, 75]\%$, and $[90, 99]\%$, respectively. These intervals propagate table-draw variation and resampling of the observed model panel; they do not turn the panel into a random sample of games or models.

**Neutral-forecast prompt control.** The primary river and Leduc prompts request a game-theory-optimal opponent, so we re-elicit all 47 tables with a neutral instruction to forecast the opponent expected in play and explicitly state that neither Nash/GTO nor the model's own policy is requested. Strict security dominance falls from $43/47$ under the primary prompt to $38/47$ under the neutral prompt, and practical $>1\%$ dominance falls from $30/47$ to $24/47$. All 47 policies remain practically frontier-inefficient under both prompts. By game, neutral-prompt strict dominance is $16/25$ on the river and $22/22$ on Leduc; practical dominance is $11/25$ and $13/22$. Thus prompt framing changes the size of the security-dominance result, but neither eliminates the gap nor restores frontier efficiency.

**Jointly conditioned security control.** The control first freezes each model's neutral river forecast, renders the normalized table back to that model, and requests a complete policy that maximizes payoff against the forecast subject to exact maximin security. This requests $F(0)$ directly rather than comparing separately elicited, unconditioned objects. We independently elicit two policies per model. Neither draw contains an exactly secure policy; $25/25$ and $24/25$ fall more than $1\%$ below $F(\mathrm{Expl}(x_m))$, while $16/25$ and $15/25$ are strictly security-dominated. Median exploitability is $1.47$ and $1.31$ payoff units, compared with $1.14$ for the primary river policies. Exploitability rankings correlate across draws ($\rho = .83$), and $24/25$ models are practically frontier-inefficient in both. Restricting each draw to policies requiring no cell-level normalization warnings leaves $16/16$ and $20/21$ practically frontier-inefficient, $0/16$ and $0/21$ exactly secure, and $11/16$ and $11/21$ strictly security-dominated. Across the two draws, three initial calls failed before yielding a valid normalized policy (one malformed response, one route token-limit error, and one provider generation error). Each was retried with the same prompt and seed, and the first parseable response was retained without reference to its score. This control directly couples the supplied forecast, policy choice, and security budget within the same prompted task. Every policy is at least $0.0996$ payoff units exploitable, over $3{,}000\times$ the schema-feasible near-secure witness of Table S2. Against the same frozen forecasts, that witness strictly outperforms $31/50$ policies and exceeds $18/50$ by more than $1\%$ of the terminal range, compared with $31/50$ and $20/50$ for unrestricted $F(0)$.

| known opponent | $n$ | $>F(0)$ | $>1\%$ | median surplus |
|---|---|---|---|---|
| uniform | 23 | 22 | 19 | 49% |
| calling station | 25 | 20 | 20 | 84% |
| Nash | 24 | 0 | 0 | N/A |

Table S6: Positive control for the frontier diagnostic. Policies are elicited to maximize payoff against the named opponent. "Median surplus" is the median fraction of $F(\infty) - F(0)$ captured; it is undefined against Nash because the headroom is zero. The $>1\%$ column requires payoff above $F(0)$ by more than $1\%$ of terminal range.

**Pairing and positive controls.** To test whether dominance would arise against any observed table, we generate $100{,}000$ random within-game derangements, preserving both policy and opponent-table marginals while excluding own pairings. Their mean strict-dominance rate is $78.8\%$, below the observed $88.0\%$ ($p < .001$); the corresponding $>1\%$ rates are $61.3\%$ and $69.9\%$ ($p = .004$). The own pairings are more often dominated, but their mean gap magnitude is not detectably different ($p = .14$); the elevated own-pair rate shows pairing-specific structure in the dominance result. Table S6 supplies the complementary positive control. When models are explicitly asked to exploit a known weak opponent, most policies exceed the secure endpoint, confirming that the diagnostic distinguishes payoff-producing deliberate insecurity.

**Relation between the stated table and the model's own play.** Under these prompts, the stated opponent table closely resembles the model's own play. We compare each table with every panel opponent, the Nash opponent, and the model's *own* mirrored-seat policy using reach-weighted excess Brier distance $\sum_i w_i \|\hat{\sigma}_i - \sigma_i\|_2^2$, where $\hat{\sigma}$ is the stated table, $\sigma$ the comparison policy, and $w_i$ the normalized Nash parent-reach weights of Table S4. On the river,

the median table-to-own-policy distance (0.10) is $1.1\times$ the model's own seed-to-seed reproducibility floor ($68\%$ of models within $2\times$), and the match is model-identifiable—$24/25$ tables sit closer to their own policy than to any other model's (Wilcoxon $p < 10^{-5}$); Leduc replicates both facts (ratio $1.4\times$, $65\%$ within $2\times$; identity 75–$85\%$, $p \leq 0.0014$). Distance from Nash also covaries between the stated table and the model's own play: TV(table, Nash) tracks TV(own, Nash) at $\rho = 0.66$ on the river and 0.63 on Leduc. This association accounts for the size trend in Table S4. As predictions of the observed panel, the tables carry modest signal ($56\%$ beat the uniform anchor, a difference that is not statistically significant; calibration correlates with security, $\rho = 0.50$, $p = 0.012$; two models state exactly uniform tables), while the Nash opponent is a poor predictor of the panel (excess Brier 0.50 vs. uniform's 0.26).

**Liar's dice shows the same anchor pattern.** Eliciting opponent tables on liar's dice (mirrored 32-cell schema, 15 models, all keymatch $100\%$) extends the larger-game pattern of Table S4: stated tables sit *closer to the uniform anchor* than to Nash (mean reach-weighted TV 0.216 vs. 0.387; $13/15$ scored models closer to uniform). This pattern is consistent with the own-play association because the models' liar's-dice play is itself uniform-leaning (TV to uniform 0.352 vs. to Nash 0.439; $64\%$ of models closer to uniform). The support here is indirect—this game has no mirrored-seat panel, so we compare anchor leans rather than the direct table-to-own-policy distance used on the river and Leduc—and the uniform-leaning table does *not* restore support for the naive-opponent account of play: a best response to a near-uniform opponent would obtain large positive value, whereas the models earn near-zero or negative value against the very tables they state: five of 15 earn negative payoff, and all fall below the frontier. At these stated tables, $14/15$ policies are security-dominated and all are frontier-inefficient; the median dominance gap is $+0.53$ on a game value of $+0.0625$. The high dominance rate despite the tables' distance from Nash shows that near-Nash stated opponents do not explain this result.

**Custom-game control.** The GTO-targeted opponent tables of §4 could in principle reflect memorized solutions to famous games rather than a task-specific opponent estimate. We test this alternative on a custom-designed two-street bluffing game whose equilibrium was neither supplied in the prompt nor used as a published benchmark. Eliciting each model's estimate of the opponent's calling frequency, the stated table is again closer to the equilibrium opponent than to the naive call-everything anchor (mean reach-weighted TV 0.32 vs. 0.58 over eight models), and the models still over-commit and remain exploitable. The observed pattern therefore does not require recall of a published equilibrium.

**Association between the stated table and policy choice.** A table-injection intervention tests whether the elicited table influences policy choice differently from a same-format table from another model. We intervene on the river, holding the elicitation frame fixed and varying only which opponent table is injected: the model's *own* elicited table, a *different* model's table, and the own table with its hand-class labels permuted within each context. Injecting a model's own table shifts its policy toward exploiting that table no more than injecting a foreign table does. Here captured value is the fraction of payoff headroom from a fixed Nash policy to a best response achieved against the injected table. Its median is $-16\%$ under either own- or foreign-table injection (difference $-2\%$, Wilcoxon $p = 0.56$, $n$=21). The own-table and foreign-table effects are statistically indistinguishable. The label-permuted control is worse ($-39\%$), indicating sensitivity to the table's class structure. For the security diagnosis, exploitability stays high under every injected table—median 1.19 in payoff units with no injection, rising to 1.38 (own), 1.67 (other), and 2.13 (scrambled). The stated-policy inconsistency therefore persists when the opponent table is supplied directly during policy elicitation. The frontier result in §4 concerns the policy relative to the opponent the model describes.

**Inverse-best-response rationalizers.** The intervention above varies the opponent table given to the model. A complementary test, in the spirit of inverse game theory (Kuleshov and Schrijvers 2015), constructs opponent policies against which the observed policy has minimum regret. These policies characterize opponents that are observationally compatible with the returned policy. For a player 0 plan $x$ and opponent plan $y$, define the best-response payoff $\mathrm{BR}_0(y) = \max_{x' \in \mathcal{X}} x'^{\top} A y$ and regret $\mathcal{R}(x, y) = \mathrm{BR}_0(y) - x^{\top} A y \geq 0$. A rationalizing opponent minimizes this regret; by LP duality on $\mathrm{BR}_0$, the minimization is a single linear program, as is the partial-identification bound below. Write $\nu(y) = \mathrm{BR}_0(y) - v \geq 0$ for the opponent's *exploitability* (0 at equilibrium). We report, in units of $|v|$, over the river, Leduc, and Kuhn deployment policies ($n$=60, seed-averaged). *(i) Rationalizability.* The minimum regret $\mathcal{R}^{\star}(x) = \min_{y \in \mathcal{Y}} \mathcal{R}(x, y)$ is near zero for 57 of the 60 policies (median $\mathcal{R}^{\star}(x) \leq 0.1$), whereas uniform-random play has $\mathcal{R}^{\star}(x) = 0.7$–$1.3$: the observed policies can be rationalized as approximate best responses to *some* opponent, unlike the uniform-random comparison. The minimizing opponent can therefore be set-valued. *(ii) Equilibrium rationalizers are insufficient.* Because this minimizing set can be large—poker best responses are robust to a range of opponent frequencies—we do not point-identify $y$; instead we compute the *least-exploitable* rationalizer $\min_y \nu(y)$ subject to $\mathcal{R}(x, y) \leq \mathcal{R}^{\star}(x)$, which lower-bounds how far any minimum-regret rationalizer must sit from equilibrium. On Leduc (144 information sets), this bound is bounded away from 0 for *all* 24 models (median $\approx 4$, on a scale where the uniform opponent is 25 and the equilibrium opponent is 0). Thus no equilibrium opponent attains minimum regret for these policies, despite the request for a GTO table. *(iii) Identification varies by game.* The test is conservative: it recovers the ground-truth best response to uniform play within the numerical floor. The least-exploitable bound is forced above the floor for $24/24$ models on Leduc, $6/12$ on the river, and $4/24$ on Kuhn, where models already play near-optimally ($2.1\times$ median) and the test has less separation. These rationalizers quantify the observable policy mismatch while leaving its mechanism unresolved.

**The model's own policy does not rationalize its play.** We test whether play best-responds to the self-like opponent implied by the stated table. The model's own policy never meets the prespecified approximate-rationalizer tolerance $\mathcal{R}(x, y) \leq \mathcal{R}^\star(x) + \frac{1}{2}|v|$. No own policy meets it on river elicited, river deployed, or Leduc ($n$=25/25/21), and own-policy regret is not lower than regret against other model policies ($p = 0.30/0.56/0.39$). Because regret against an exploitable opponent is large for *any* non-exploiter, we normalize by the secure baseline $\mathcal{R}(x^\star, \text{own})$: models respond to their own mirror *worse* than a maximin policy would, by $+3.9|v|$ (river elicited, $p = .030$), $+7.8|v|$ (deployed, $p = .002$), and $+3.8|v|$ (Leduc, $p = .0006$); on the river $18/25$ models earn *below the game value* against a copy of their own policy (median shortfall $1.9 \times |v|$), so the observed policy does not exploit its own mirror. Among named references, Nash gives the smallest median regret (3.5–4.2×) in each analysis. Together with the own-table injection, these results show that the stated table is reproducible and self-similar but is not distinguished as a rationalizer of play. The frontier of §4 is accordingly anchored at an opponent table similar to the model's own policy: dominance indicates that greater exploitability yields no payoff gain even against such an opponent.

Table S7 assembles the decomposition of §4 per model: table accuracy, table–play alignment, best-response fidelity, and frontier optimality are separate measurements, and the failure is confined to the last two.

## C.2 Audit validity: threats and controls

Table S8 summarizes the validity threats considered in the audit, the control that tests it, and the outcome; the paragraphs below give details.

**Robustness.** Rankings hold under prompt paraphrase ($\rho$=0.78, all still exploitable). Table S9 gives exact exploitability for all 25 models across the three exactly-solved poker games; Fig. S1 places each river policy on the noisy-Nash ladder of §3; Table S10 gives the within-poker transfer matrix of §5.

**Sensitivity to format repair.** Illegal or missing actions in an elicited table are repaired by the deterministic rule of §2. These repairs are too rare to explain the measured losses: across the river panel they touch 0.9% of cells, 18 of 25 models need none, and the events are 113 dropped illegal actions (the model over-specified an out-of-turn action) against 66 missing-cell uniform fallbacks. Repair rate correlates with exploitability ($\rho$=0.69); the smallest models (3–8B) are weak at both valid formatting and strategy. This association does not show that repairs cause the loss. As a favorable sensitivity check, we replace every warned cell by its projected-equilibrium counterpart. Among the seven affected models, mean exploitability changes from 53× to 50× and every policy remains at least 31× exploitable (e.g. ministral-14b 71× → 71×, llama-3.1-8b 57× → 49×). The model's specified play, rather than the fallback, accounts for the exploitability.

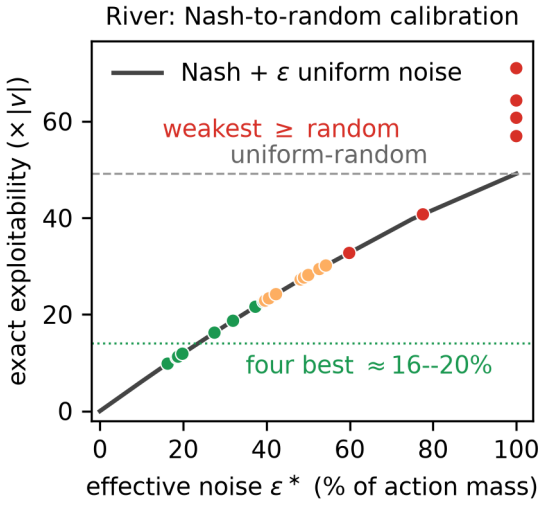


Figure S1: Calibrating the scale on the river: the noisy-Nash ladder mixes equilibrium with uniform play, and each model is placed at its *effective noise* $\epsilon^*$—the perturbation of equilibrium that matches its exact exploitability. The most secure model has the exploitability of a 16%-noise equilibrium, the four most secure policies 16–20%, and the median ≈ 49%; the six weakest policies (top) are at or beyond uniform-random. A secure solver would sit at $\epsilon^*$=0.

**Model-family dependence.** Many models share a vendor or training lineage, so the panel may provide fewer than 25 independent points. The two primary pooled results survive vendor clustering. The aggression–exploitability correlation ($\rho$=0.89) stays in $[0.87, 0.91]$ when any single vendor's models are removed, and a bootstrap that resamples *vendors* rather than models gives mean 0.89, CI $[0.81, 0.95]$. A *universal counter*—the exact best response to the panel's mean policy, one fixed counter-strategy for all models—is tested leave-one-vendor-out: building it on the other vendors and evaluating on a held-out vendor's models—which the counter does not use—recovers a median 77% of their exploitable loss (84% of held-out models above 50%). The shared pattern is therefore not explained solely by model lineage.

**Equilibrium-selection robustness.** The exploitability certificate is independent of the equilibrium representative, whereas diagnostics measured relative to equilibrium need not be. We therefore solve the sequence-form equilibrium-face LP under 64 reproducible random linear objectives, producing 64 alternative extreme equilibria in addition to the reference solution; every candidate has exploitability below $5 \times 10^{-13}$. Across all 65 representatives, all 25 river policies remain over-aggressive, the panel-mean signed bias ranges only from $+0.243$ to $+0.278$, and its rank correlation with exploitability remains $\rho \in [0.875, 0.906]$. The panel's highest-harm betting context is unchanged, and the per-model highest-harm context is unchanged for $24/25$ models. The proxy conclusion is likewise stable: action-match correlation ranges from 0.902 to 0.925 on the river but only 0.262 to 0.401 on Leduc. Thus the signed aggression pattern

| model | $d_{\mathrm{N}}$ | align | $r_{\mathrm{tab}}$ | $\Delta_m$ | model | $d_{\mathrm{N}}$ | align | $r_{\mathrm{tab}}$ | $\Delta_m$ |
|---|---|---|---|---|---|---|---|---|---|
| gpt-5.5 | 0.46 | 0.9 | 22.9 | 0.94 | nova-pro-v1 | 0.31 | 4.1 | 24.9 | 1.09 |
| claude-opus-4.8 | 0.35 | 0.8 | 10.7 | 0.45 | mistral-large-2512 | 0.46 | 0.5 | 22.9 | 1.01 |
| gemini-3.5-flash | 0.36 | 3.7 | 10.6 | 0.43 | mistral-medium-3.1 | 0.51 | 1.7 | 33.8 | 1.48 |
| gemini-3.1-pro-preview | 0.28 | 0.3 | 9.5 | 0.42 | deepseek-v3.2 | 0.49 | 1.2 | 35.1 | 1.54 |
| claude-opus-4.5 | 0.47 | 9.7 | 18.3 | 0.79 | llama-3.3-70b-instruct | 0.39 | 5.3 | 26.2 | 1.15 |
| llama-4-maverick | 0.41 | 1.1 | 19.2 | 0.84 | command-a | 0.41 | 0.6 | 29.1 | 1.28 |
| phi-4 | 0.39 | 3.9 | 27.2 | 0.64 | gemini-2.5-flash | 0.54 | 0.9 | 20.9 | 0.92 |
| qwen3.7-max | 0.49 | 0.5 | 19.7 | 0.85 | llama-3.1-8b-instruct | 0.54 | 0.5 | 49.0 | 2.15 |
| claude-sonnet-4 | 0.47 | 0.4 | 24.3 | 1.07 | nova-micro-v1 | 0.79 | 1.1 | 98.7 | 4.33 |
| claude-sonnet-4.6 | 0.48 | 1.7 | 23.3 | 1.02 | ministral-3b-2512 | 0.82 | 1.1 | 83.1 | 3.65 |
| gemma-3-27b-it | 0.54 | 2.5 | 20.9 | 0.83 | gemma-3-4b-it | 0.57 | 9.3 | 49.2 | 2.16 |
| gemini-2.5-pro | 0.41 | 0.5 | 32.5 | 1.32 | ministral-14b-2512 | 0.77 | 2.0 | 89.8 | 3.94 |
| deepseek-v4-pro | 0.46 | 0.4 | 21.9 | 0.96 | | | | | |
| panel median | 0.47 | 1.1 | 24.3 | 1.02 | | | | | |

Table S7: The stated-opponent/security inconsistency decomposed, per river model (ordered by exploitability): Nash distance $d_{\mathrm{N}}$ (reach-weighted TV from the stated table to the Nash opponent; 0 = equilibrium-accurate); table–play *alignment* (excess Brier from the stated table to the model's own seat-2 policy, as a multiple of the model's own seed-to-seed noise floor; $\approx 1$ = as close to own play as repeat elicitations are to one another); best-response *fidelity* $r_{\mathrm{tab}}$ (regret of the model's policy against its stated table, $\times|v|$; 0 = plays a best response to what it states); and frontier *optimality* $\Delta_m$ (dominated gap in payoff units; 0 = efficient at its own exploitability budget). Medians in the last row; Table S4 reports the full-panel *mean* of $d_{\mathrm{N}}$ (0.49). The pattern is consistent across models: opponent tables resemble the corresponding own policies at roughly the repeat-elicitation noise floor, yet those policies have regret against the tables and fall below the security–payoff frontier.

and shallow-versus-deeper proxy contrast do not depend on the solver-returned equilibrium representative; per-context attributions remain diagnostic rather than additive.

**Reasoning-enabled elicitation leaves the audit conclusions unchanged.** The main audit elicits with reasoning disabled for comparable, non-truncated output. Re-eliciting the *full* river panel with reasoning *enabled* (a 40k-token budget to prevent the reasoning allocation from truncating the JSON; 3 seeds; 24 of 25 models returned parseable policies) leaves the audit unchanged: the panel median is similar ($27.9\times$ off versus $27.3\times$ on; paired Wilcoxon $p = 0.47$), $58\%$ of models improve while the rest worsen (gemini-3.1-pro $11.9 \to 7.3\times$, but command-a $40.8 \to 53.5\times$), and the ranking is essentially preserved (Spearman $0.94$, CI $[0.83, 0.97]$). The four most secure models stay at 7.3–$9.9\times$ and no model reaches below $7.3\times$. Reasoning changes individual policies without eliminating exploitability, so the audit is not an artifact of suppressed deliberation.

**Mirrored seat audit.** Re-eliciting the panel from the opposite seat confirms the audit is not a first-seat artifact (Table S11): seat rankings correlate at $\rho = 0.68$ ($p < 0.001$) and every model stays substantially exploitable, mean $29\times$.

**Elicited versus deployed play.** For the full 25-model river panel, we retain ten successful raw responses per cell at temperature $0.7$ without asking for a complete strategy. Every prompt states the shared ten-card range, unconditional equity, betting history, current pot and call, commitments, remaining stacks, and exact action meanings. API failures are retried; every illegal or unparsable response remains in the archive and receives a fixed check/call fallback. The deterministic parser accepts a lone letter or an explicit final-answer letter; the wrapper repairs $16/13{,}500 = 0.12\%$ of responses after 54 API retries.

Deployed exploitability remains substantial (median $1.71$ vs. $1.14$ elicited; higher for $76\%$ of models) and tracks elicited exploitability at Spearman $0.57$ ($p = 0.003$). A matched-sampling control draws the same $N{=}10$ actions per cell from each elicited policy ($R{=}500$ reconstructions): its median is $1.19$, a $+0.05$ reconstruction inflation. The deployed-minus-matched median is $+0.38$ (Wilcoxon $p = 0.009$; $68\%$ above matched), so the increase exceeds reconstruction noise. Rerunning the security–payoff frontier leaves $68\%$ security-dominated (median gap $+1.24$). The validated full panel also reproduces the observable direction. Commitment assigns scores $0/1/2/3$ to fold/check-call/pot/all-in, averages them with equilibrium parent-reach weights, and subtracts equilibrium commitment. This bias is positive for every deployed policy, with mean $+0.53$ versus $+0.68$ elicited. Thus the stated-table audit captures a deployed ordering and failure pattern without requiring equality between the two policy measurements. The deployment analysis also extends beyond poker to the natively zero-sum liar's dice: reconstructing each policy from single-action sampling (ten models, $N{=}10$ per cell) leaves *every* model substantially exploitable, with a deployed median of $11.1\times$ versus $9.9\times$ elicited and all models above $8\times$; at this sampling depth the individual ranking is noisier than on the river ($n{=}9$), but the worst-case exploitability also persists under move-by-move play outside poker.

**Relabel control: structure, not vocabulary.** To test whether the aggression bias—the reach-weighted excess bet-or-raise probability defined in §3—is a surface poker prior, we re-elicit the river under isomorphic neutral labels on the

| threat | control | result |
|---|---|---|
| output schema too coarse | equilibrium projected through schema | loss $\leq 3.3 \times 10^{-5}$ |
| malformed cells | deterministic repair; equilibrium-cell sensitivity | $53\times \to 50\times$ |
| prompt wording | paraphrase re-elicitation | $\rho$=0.78, all exploitable |
| seat choice | mirrored-seat audit | $\rho$=0.68, mean $29\times$ |
| poker vocabulary | isomorphic relabel | bias $+0.68 \to +1.14$ |
| negotiation script | delexicalized bargaining | median 2.00 either way |
| reasoning disabled | full-panel reasoning-on | similar median, $\rho$=0.94 |
| stated-table format | state-wise action reconstruction | 68% dominated; $\rho$=.57 |
| sampling noise | matched-sampling reconstruction | $+0.38$, $p$=.009 |
| vendor lineage | leave-one-vendor-out + clustered bootstrap | $\rho \in [0.87, 0.91]$ |
| memorized equilibria | custom game with unpublished solution | table remains nonequilibrium |
| table unrelated to own play | calibration vs. own play | own-play match at noise floor |
| own-table specificity | table injection; rationalizer | no distinct own-table response |
| prize-order ambiguity | explicit-order Goofspiel | anti-transfer signal removed |

Table S8: Validity threats, the control that tests each, and the outcome. Details appear in the paragraphs of this appendix and §4; the primary audit conclusions remain unchanged across these controls.

identical tree—abstract states, types, and actions in place of cards and bets—and score with the same engine. The signed aggression bias remains positive and increases on average (mean $+0.68 \to +1.14$ over six models, all 20–$54\times$), so poker vocabulary does not explain the error in this control. It persists on the identical tree under neutral labels. We exclude Goofspiel from this comparison because its worded and neutral prompts share the same unstated prize order, so neither isolates structure from vocabulary there.

**Delexicalized bargaining: structure, not a negotiation script.** The relabel control above swaps vocabulary while preserving a game frame; for the bargaining proposer we remove domain vocabulary to test whether the high demand in §4 depends on negotiation-specific wording. We re-elicit the proposer policy on an *opaque* presentation of the identical tree: no negotiation vocabulary (no "offer," "keep," "accept," "proposer," or "bargain"), item names replaced by anonymous resource tokens $R_1, R_2, R_3$, each action written as a raw claim vector $\mathrm{take}[R_1{:}x, R_2{:}y, R_3{:}z]$, and acceptance rendered as a neutral CONFIRM; the tree, payoffs, hidden-type prior, and exact-exploitability engine are untouched. We measure demand intensity by the reach-weighted fraction of its own private valuation the proposer claims when it makes a claim—anchored by the equilibrium (16%) and a value-maximizing reference (77%). On the eleven panel models that return a well-formed policy under both presentations, delexicalization leaves the failure intact: median exploitability is 2.00 under either wording (the maximum, surrendering the proposer's entire equilibrium edge), and median demand is 56% worded versus 60% opaque—both far above equilibrium and near the value-maximizing reference—with no significant shift (paired Wilcoxon $p = 0.57$). This control provides no evidence that negotiation vocabulary generates the proposer's high demand.

**Goofspiel prize-order control.** Goofspiel's standard imperfect-information form fixes a descending prize order, but our prompt left the order unstated, so a model that read it as ascending would be scored against the wrong target. Re-eliciting the full panel with the descending order stated explicitly removes the ambiguity: the negative river-to-Goofspiel rank association under the ambiguous prompt becomes small, positive, and nonsignificant, while mean exploitability remains roughly unchanged. The ambiguous prompt therefore cannot support an anti-transfer claim. We treat Goofspiel only as the player-symmetric boundary game.

## C.3 Proxy analyses

**Opponent-strength degradation.** The weak-opponent signal degradation of §6 is continuous, not a single point. For $\lambda \in [0, 1]$, sweep the reference opponent $y_\lambda = (1-\lambda)\,\mathrm{Nash} + \lambda\,\mathrm{Uniform}$. The Spearman correlation of negated win-rate with exact exploitability decreases smoothly from 0.87 to 0.10 (Table S12). A ladder dominated by similarly weak opponents corresponds to large $\lambda$, where this ranking becomes uninformative.

**Model-vs-model round-robin.** The crossover sweep varies a synthetic reference; arena-style evaluations instead play models against one another. We play every elicited river policy against every other (both seats, exact bilinear payoffs, no sampling) and rank by mean two-seat payoff. The ladder ranking tracks exact exploitability at $\rho = 0.59$ (field-aware bootstrap CI $[0.41, 0.77]$, resampling models and rerunning the ladder within each resampled field; permutation $p = 0.002$), and a Copeland win–loss ladder at $\rho = 0.79$ ($[0.58, 0.90]$). The signal is driven by the least secure models: within the more secure half of the panel the mean-payoff ladder is uninformative ($\rho = -0.16$, $n$=12), the most secure mean policy finishes 8th of 25 (balanced play can forgo the exploitative value conceded by less secure models), and a $57\times$-exploitable model finishes 9th. The arena's mean opponent policy has exploitability 0.85 in payoff units ($19 \times |v|$)—between the Nash and uniform references of Table S12—so the round-robin behaves like a mid-$\lambda$ reference: enough to separate the least secure models from the remainder, but not to rank the least-exploitable policies.

| model | river | Leduc | Kuhn |
|---|---|---|---|
| claude-opus-4.8 | 9.8× | 10.7× | 1.2× |
| gpt-5.5 | 11.1× | 10.0× | 2.0× |
| gemini-3.5-flash | 11.3× | 8.6× | 0.0× |
| gemini-3.1-pro | 11.9× | 6.6× | 0.0× |
| claude-opus-4.5 | 16.2× | 11.2× | 2.0× |
| llama-4-maverick | 18.7× | 27.5× | 2.9× |
| qwen3.7-max | 21.6× | 12.7× | 2.0× |
| phi-4 | 22.7× | – | 3.7× |
| claude-sonnet-4.6 | 22.9× | 13.7× | 2.7× |
| gemma-3-27b-it | 23.4× | 28.8× | 4.8× |
| claude-sonnet-4 | 24.2× | 11.8× | 2.0× |
| deepseek-v3.2 | 27.2× | 13.8× | 3.2× |
| mistral-large-2512 | 27.7× | 15.4× | 4.3× |
| nova-pro-v1 | 28.2× | 18.9× | 3.0× |
| gemini-2.5-pro | 28.2× | 13.3× | 1.3× |
| mistral-medium-3.1 | 29.4× | 17.0× | 1.7× |
| deepseek-v4-pro | 30.2× | 14.3× | 1.9× |
| llama-3.3-70b | 32.8× | 28.0× | 1.0× |
| command-a | 40.8× | 14.3× | 4.0× |
| llama-3.1-8b | 57.0× | 27.2× | 4.3× |
| nova-micro-v1 | 60.6× | 31.9× | 1.2× |
| ministral-3b-2512 | 60.6× | 24.9× | – |
| gemma-3-4b-it | 60.8× | 25.9× | 2.2× |
| gemini-2.5-flash | 64.3× | 17.9× | 2.2× |
| ministral-14b-2512 | 71.0× | 16.9× | 4.0× |
| $\|v\|$ | 0.044 | 0.086 | 0.056 |

Table S9: Full panel: exact exploitability for all 25 models, ordered by river exploitability (multiple of game value $|v|$). Dashes are models that did not return a parseable policy for that game. Every model is highly exploitable except on the small Kuhn game. Goofspiel (player-symmetric, value zero) is treated as a boundary case in §5 and is excluded from the primary exploitability comparisons; its initial per-model scores were prize-order-ambiguous, and the explicit-order re-elicitation is summarized in the "Goofspiel prize-order control" below. The two non-poker games—action-restricted alternating-offer bargaining (audited as a zero-sum reduction, App. B) and natively zero-sum liar's dice—are reported in §4 and §5.

**Action-match depth contrast.** Frequency perturbations that preserve the modal action conceal substantial exploitability in both the river and two-street games. This illustrates the certification gap, not a general depth law.

| $\rho$ | river | leduc | kuhn |
|---|---|---|---|
| river | 1.00 | $+.66^{*}$ | $+.27$ |
| leduc | | 1.00 | $+.39$ |
| kuhn | | | 1.00 |

Table S10: Within-poker exploitability-ranking Spearman ($n$=23–24; $^{*}$ sig. with CI excl. 0). Robustness is partially stable within poker: river and Leduc rankings agree, whereas the correlations with Kuhn are smaller. Cross-class comparisons and Goofspiel's boundary-case role appear in §5.

| model | seat 0 | seat 1 |
|---|---|---|
| claude-opus-4.8 | 10× | 12× |
| gemini-3.1-pro | 12× | 10× |
| gpt-5.5 | 11× | 20× |
| claude-opus-4.5 | 16× | 12× |
| llama-4-maverick | 19× | 16× |
| mistral-large-2512 | 28× | 21× |
| deepseek-v4-pro | 30× | 20× |
| command-a | 41× | 18× |
| ... | ... | ... |
| llama-3.1-8b | 57× | 84× |
| nova-micro-v1 | 61× | 84× |
| ministral-3b-2512 | 61× | 73× |
| ministral-14b-2512 | 71× | 57× |
| seat $\rho = 0.68$ [0.34,0.86], $p < .001$; seat-1 mean 29×, all $> 0$ | | |

Table S11: Mirrored seat audit ($n$=25, abridged). Every model is substantially exploitable in both seats and rankings agree, so the measured exploitability is not specific to acting first.

| $\lambda$ | 0.0 | 0.2 | 0.3 | 0.4 | 0.5 | 0.7 | 1.0 |
|---|---|---|---|---|---|---|---|
| $\rho$ | $+.87$ | $+.61$ | $+.49$ | $+.41$ | $+.33$ | $+.16$ | $+.10$ |

Table S12: River: Spearman(exploitability, $-$win-rate vs. $y_\lambda$). Nash tracks robustness; the signal weakens as the reference approaches uniform.

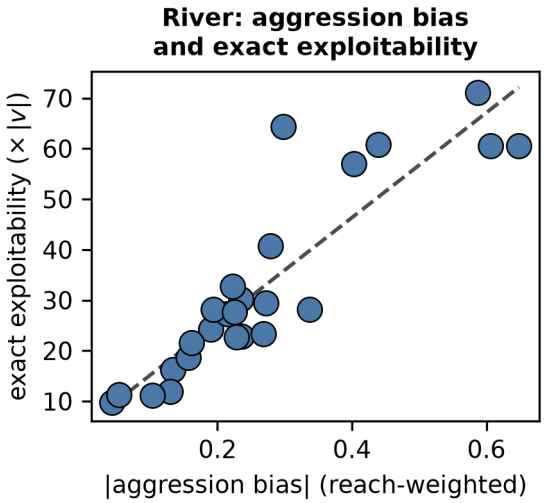


Figure S2: River exploitability tracks one interpretable axis: |aggression bias| from equilibrium rank-correlates with exact exploitability at $\rho = 0.89$ across the panel.

### C.4 Scaling to 200-Big-Blind Heads-Up No-Limit Hold'em

This study uses 200-big-blind stacks and fold/check-call/pot/all-in (FCPA) actions. Payoffs are reported in big blinds per 100 hands (bb/100). The local best response (LBR) plays the live model; "distill" is a best response to a frozen behavior clone, subsequently certified against the live model. The simultaneous lower bound is the larger Bonferroni-adjusted one-sided lower bound across those two adversaries. Both use an all-in-equity control variate, which replaces an all-in outcome by its conditional expectation over the remaining cards.

Full sequence form is unavailable at this scale, so we estimate a *lower bound* on deployed exploitability in four-street OpenSpiel `universal_poker`. The distill-then-exploit pipeline (i) fits an equity-bucketed behavior clone to live-model self-play, (ii) computes an offline Monte-Carlo best response to the frozen clone, and (iii) certifies that fixed adversary against the live model. Certification uses an all-in-equity control variate. A separate local best response queries the live model for actions but uses street-level fold rates estimated from the frozen pre-certification clone; neither adversary updates during evaluated hands. Clone fitting, adversary construction, and certification use disjoint hands. A predeclared rule excludes a hand only if all API or parsing retries are exhausted and a check/call fallback is required. Across both adversaries and all six models, all 29,000 hands and 75,252 live-model decisions were retained, with zero fallbacks and four transient errors recovered by retry. The six-model panel and two adversaries were fixed before certification, defining 12 confirmatory tests; later data-biased and on-policy adversaries are exploratory and enter neither positive certificates nor multiplicity correction.

The exploiter alternates physical seats by hand. Every reported run has even size and is exactly balanced: 1,000 hands per seat for each local-best-response run, and either 1,500 or 1,250 per seat for each distilled-adversary run. Let $\mathcal{X}_s$ denote the realization-plan polytope for physical seat $s$, and let $\Pi = \mathcal{X}_0 \times \mathcal{X}_1$. For policies $x = (x_0, x_1)$ and $y = (y_0, y_1)$, the reported estimand, with $u_0(a_0, b_1)$ denoting seat 0's payoff when those plans occupy seats 0 and 1, is $g(y, x) = \frac{1}{2}[u_0(y_0, x_1) - u_0(x_0, y_1)]$. The product polytope $\Pi$ is compact and convex, and $g$ is bilinear and antisymmetric. For any policy $x$, choosing $y = x$ gives $g(x, x) = 0$, so $\max_x \min_y g(x, y) \le 0$ and $\min_y \max_x g(x, y) \ge 0$; the minimax theorem therefore gives value 0. Thus the reported adversary payoff is a lower bound on the audited policy's exploitability, independent of how well the clone or abstraction fits: $\mathrm{Expl}_{\mathrm{sym}}(x) = \max_{y'} g(y', x) \ge g(y, x)$ for every fixed adversary $y$. The confidence interval is the hand-level mean plus or minus 1.96 standard errors after the all-in-equity replacement and quantifies only sampling uncertainty, not clone quality. If an all-in fixes both private hands before the river, a hand with contribution $c$ is scored as $c(2p-1)$, where $p$ is win probability plus half the tie probability over the remaining runouts. We enumerate all runouts when at most 1,000 remain and otherwise sample 1,000 uniformly without replacement. This conditional-expectation replacement preserves the mean by the tower property; all other hands retain their realized return. For retained hand scores $z_1, \ldots, z_n$, let $\bar{z}$ and $s_z$ be their sample mean and sample standard deviation; the reported half-width is $1.96 s_z/\sqrt{n}$. These are nominal two-sided Wald intervals under independent-hand sampling and a stationary model endpoint within each run. Family-wise certification uses the one-sided Bonferroni lower bound $\bar{z} - z_{1-0.05/12} s_z/\sqrt{n}$, where $z_q$ is the standard-normal $q$-quantile. Under that correction across all 12 model–adversary tests, the DeepSeek-V4-Pro and GPT-5.5 certificates remain positive, whereas the pointwise Qwen3.7-Max result does not. Here "certificate" means a confidence-qualified lower bound in the symmetrized FCPA game. As a distributional diagnostic, 200,000 nonparametric resamples of adjacent seat-balanced hand pairs give one-sided $0.05/12$ lower quantiles of $+1252$ and $+183$ for the DeepSeek-V4-Pro and GPT-5.5 local responses, respectively, nearly matching the Wald bounds; this bootstrap check is not a finite-sample coverage guarantee.

**Calibration.** On solved river, two-street, and three-street games, the behavior-clone exploiter recovers 94–98% of exact exploitability and the local best response recovers 83–99%. This supports tightness empirically but gives no formal guarantee in the full game.

**Results (Table S13).** The local best response (Lisý and Bowling 2017) and distill-then-exploit adversary each certify DeepSeek-V4-Pro, at $+1680$ and $+1466$ bb/100, respec-

| model | LBR | distill | simult. lower | status |
|---|---|---|---|---|
| DeepSeek-V4-Pro | $+1680 \pm 317$ | $+1466 \pm 224$ | $\mathbf{+1253}$ | exploitable |
| GPT-5.5 | $+473 \pm 212$ | $+90 \pm 154$ | $\mathbf{+187}$ | exploitable |
| Qwen3.7-Max | $-12 \pm 136$ | $+57 \pm 50$ | $-9$ | inconclusive |
| Claude-Opus-4.8 | $-156 \pm 154$ | $-39 \pm 70$ | $-134$ | inconclusive |
| Gemini-3.1-Pro | $-386 \pm 124$ | $-58 \pm 58$ | $-136$ | inconclusive |
| Gemini-3.5-Flash | $-2 \pm 110$ | $-64 \pm 87$ | $-151$ | inconclusive |

Table S13: Adversary payoffs and simultaneous exploitability lower bounds in the FCPA game (bb/100, estimate $\pm$ nominal pointwise $95\%$ Wald CI; $2{,}000$ LBR hands per model and $3{,}000$ distill hands, except $2{,}500$ per Gemini). All $29{,}000$ hands were retained with no fallbacks. "Simult. lower" has nominal simultaneous $95\%$ coverage across the 12 tests. Inconclusive rows do not certify security (§5).

tively. The local response also certifies GPT-5.5 at $+473$ bb/100, while its distill interval crosses zero. Qwen3.7-Max's pointwise distill bound does not survive correction. For three models, neither primary adversary has a positive point estimate. A matched-safety data-biased response (Johanson, Zinkevich, and Bowling 2007) and four rounds of on-policy retraining (Ross, Gordon, and Bagnell 2011) likewise produce no positive point estimate (e.g., Opus-4.8 moves $-113 \to -3 \pm 55$). These failures do not distinguish security from adversary weakness (Lisý and Bowling 2017; Timbers et al. 2022); the exact audit remains primary.

**Descriptive within-vendor comparisons.** River exploitability falls from Opus-4.5 to 4.8 ($16|v| \to 10|v|$) and from Gemini-2.5 to 3.x (Flash $64|v| \to 11|v|$; Pro $28|v| \to 12|v|$), but is nonmonotone across five DeepSeek generations (21–$35|v|$, with V4-Pro at $30|v|$). Thus newer models are not uniformly safer. With only 2–4 generations per lineage, these comparisons support directional observations within this panel, not a general temporal trend.

## D Computational Environment

Experiments used Python 3.14.3, NumPy 2.5.1, SciPy 1.18.0, OpenSpiel 2.0.1, and Matplotlib 3.11.1 in a version-pinned environment. HiGHS solved the exact programs on CPU to tolerance $10^{-7}$; no GPU was required. Computations ran on macOS 15.7.3 on a MacBook Pro with an Apple M3 Max (14 CPU cores) and 36 GB unified memory.

Model elicitation used OpenRouter chat completions through `openai` 2.46.0 at temperature $0.7$, with reasoning-disabled generation requested and replicate seeds beginning at $2026$. Calls were collected in June–July 2026. Retained materials include exact model routes, prompts, normalized outputs, and available raw responses. Only this stage required an external service.

Once a normalized policy is fixed, its exact audit score is deterministic; the provider response that produced it need not be reproducible bit for bit, even with a repeated seed. Seeds therefore index independent elicitation attempts rather than certify identical replay. Replication across seeds and the prompt, seat, and reasoning controls quantify sensitivity to that collection stage, while all best-response, frontier, permutation, and bootstrap computations operate on frozen records.

Primary poker audits average 2–3 seeds; single-seed extensions are identified in App. B. Game, schema, lift, repair, and scoring rules were fixed before evaluation. Opponent-table repeats, neutral wording, pairing permutations, known-opponent interventions, and joint conditioning are robustness analyses rather than preregistered primary tests. In the joint control, three failed calls were rerun with the same prompt and seed, retaining the first parseable response. Temperature, seeds, solver tolerance, and practical margins use the reported singleton settings rather than outcome-selected sweeps.