GENSTRAT: TOWARD A SCIENCE OF STRATEGIC REASONING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed game suites for evaluating strategic reasoning in large language models (LLMs) saturate and are hard to keep free of contamination. We introduce GENSTRAT, a system for evaluating the strategic capabilities of LLMs driven by procedural generation of two-player zero-sum imperfect-information betting games, allowing an evaluator to draw novel, fresh games on demand. Each game is characterized on a scale of five complexity axes (state space, temporal contingency, opponent dependence, private information, precision) and measured by Monte Carlo simulation. Using this generator, we construct a 50-game benchmark and hold a tournament of twelve frontier and open-weight models, gathering data from over $40{,}000$ LLM vs. LLM matches. We further construct a composite complexity axis, and find that, on the eight hardest games, the gap between the best and worst model is $1.7\times$ its value across the whole set of games ($1.6\times$ with weights learned on held-out games). Ablations show that a higher thinking budget, disclosing the opponent's strategy, and higher reasoning matched to a game's specific demands (for private information and planning depth) were associated with higher margins for the LLM player.

1 INTRODUCTION

This paper introduces GENSTRAT, a framework for evaluating LLMs' strategic capabilities built upon a procedural generator of generalized betting games (GBGs), a family of two-player zero-sum imperfect-information card games. The generator can continually generate fresh imperfect-information games with novel features and structure. We propose five complexity axes—state space, temporal contingency, opponent dependence, private information, and precision—to characterize the space of games and study LLM performance across the games the generator can create. We then run a 12-model tournament on a benchmark of 50 generated games, consisting of over $40{,}000$ matches. We find that general strategic intelligence is a phenomenon in the space of GBGs: newer and larger models generally pull ahead in the leaderboard, especially compared to open-source models. The differences between models grow with the complexity of the games, most of all on games where the best action depends on how the rest of the deal unfolds, suggesting that as the current games saturate, we can continue to generate larger, more complex games. Furthermore, even within the GBG space, models exhibit significant heterogeneity in how their abilities vary over the space of games, suggesting that different models and training regimes may lead to different strategic capabilities.

Frontier LLMs are placed in economic-agent roles in controlled experiments, ranging from running small commerce operations (Anthropic, 2025) to marketplace simulations (Anthropic, 2026). Models have also exhibited algorithmic-collusion behavior in LLM-based pricing studies (Fish et al., 2024). How a model will perform once deployed is hard to anticipate, and performance on canonical games—especially ones for which a model has training data to fall back on—may not cleanly transfer to the environment a deployer cares about.

Existing strategic-reasoning benchmarks evaluate LLMs on fixed games or suites. These include poker settings (Guo et al., 2024; Huang et al., 2024), AvalonBench (Light et al., 2023), Diplomacy (FAIR et al., 2022), and broader suites such as GTBench and GameBench (Duan et al., 2024; Costarelli et al., 2024). These benchmarks share three limits. First, fixed suites may saturate as the frontier improves. Second, canonical games often exist in the training data of LLMs, which makes training-data contamination hard to rule out. Finally, benchmarks with fixed or canonical games do

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

not show how model performance shifts as the environment shifts, limiting the deployer's ability to understand how a model will generalize to novel environments, especially ones with adversarial elements.

GENSTRAT addresses these limits with a procedurally generated distribution of GBGs. Procedural generation has been shown to be productive for single-agent reinforcement learning (RL) (Cobbe et al., 2020; Chevalier-Boisvert et al., 2023). Multi-agent settings add an *amplification effect*, whereby small increases in environment complexity can produce large increases in the complexity of the strategic problem an agent must solve. Because a generator can draw freely from the same distribution, a fresh draw is very likely not one a model will have seen before or memorized. We then consider five complexity axes that describe what each game demands, computed by Monte Carlo simulation with randomly generated strategies, and provide a capability-profile decomposition for each LLM.

In the tournament we conduct, larger, more recent, reasoning-capable models emerge as the more competent players of GBGs. Significant heterogeneity emerges, however. `gemma-4-31b`, for example, gains on games with high private information ($+0.19$ chips/game per standard deviation of the axis, Section 7). We conduct ablations, including against fixed LLM-authored programs, both disclosed and hidden, and at different levels of thinking budget. Predictably, disclosing a program's strategy helps, by about 0.7 chips per game, and so does a higher thinking budget, in every model family tested (significant in one of four, Section 8). Moreover, we use an LLM judge to examine the reasoning traces of LLMs, finding that when LLMs use more of the types of reasoning associated with a game's particular complexities, they tend to earn higher margins on that game, for private information and planning depth. Finally, we report local jaggedness, a noise-corrected measure of how far an LLM's advantage departs from a smooth function of the games' properties.

2 Related work

Strategic games have a long history in AI research, from AlphaZero (Silver et al., 2018), Libratus and Pluribus (Brown & Sandholm, 2018; 2019), DeepNash (Perolat et al., 2022) and CICERO (FAIR et al., 2022), which are specialized systems for single games, to benchmarks of general-purpose LLMs on fixed games such as GTBench (Duan et al., 2024), GameBench (Costarelli et al., 2024) and AvalonBench (Light et al., 2023). Akata et al. (2025) study LLMs in repeated games, Lorè & Heydari (2024) disentangle game structure from framing, Collins et al. (2026) probe LLMs' ability to evaluate novel games, and Lin et al. (2026) study professional-poker-style tasks with tool use. The theory-of-mind literature is close (Strachan et al., 2024; Ullman, 2023), and the poker-ToM coding scheme of Lin & Hou (2026) reads strategic reasoning from traces, as our judge rubric does. Most closely related, gg-bench (Verma et al., 2025) generates games by LLM authoring and scores win rate against a self-play RL agent. In contrast, with GENSTRAT, game rules come from a parameterized, highly controllable generator. Games are also imperfect information games, and thus test a model's ability to perform under uncertainty and asymmetric information. Complexity is an input we set directly, and raising it requires no rebuilding. We decompose performance along measured axes, and a large-scale tournament maps model performance across them. Our concern with generalization off the evaluation distribution follows Vafa et al. (2024b;a; 2025). Procedural generation is standard in single-agent RL (Cobbe et al., 2020; Chevalier-Boisvert et al., 2023; Shaker et al., 2016).

3 Generalized betting games and GENSTRAT

We call the games we study *generalized betting games* (GBGs). A GBG is a two-player zero-sum extensive-form game of imperfect information. Kuhn poker (Kuhn, 1950) and Leduc poker (Southey et al., 2005) are both GBGs, and the family adds non-betting actions and rounds, other game-tree and information structures, and conditional branches. A game consists of a deck, private hands and other piles, a graph of phases (betting, simultaneous-move, auction and observation), and conditions on its branches. The generator randomizes structure, starting from the phase graph. Inside the graph it draws ranks, suits and hand sizes, betting order, and the auction and simultaneous rounds. It also draws observation triggers, position criteria, showdown metrics, side bets and branch predicates. A series of structural checks and repairs helps ensure games are coherent and playable (Appendix A). The integer seed and its stored generation configuration are sufficient to fully regenerate a game. To track the model frontier, complexity can be raised by relaxing the generator caps. A draw is rejected

if random play takes more than ten moves per player on average, if one seat wins more than 65% of the decided deals under random play, if more than 40% of its branch conditions hold almost always or almost never, if more than 20% of its phases fire in fewer than 5% of episodes, or if a branch can never fire. Of 12,351 sampled seeds, 2,000 were accepted.

A concrete game. Seed 4643 (full rulebook in Appendix B) uses a deck of eight cards (10, J, Q, K in two suits), with three private cards per player, 10 chips per player and an ante of 3. Hidden positions are assigned from the dealt hands, and a player who holds one peeks at one of the opponent's cards. Betting has up to two rounds, the first with any stake up to the player's stack and the second with fixed 1-chip bets. Between them each player may peek at an opponent card at a 50% risk of paying a chip. Both players then draw a card, and the chip leader gets a hidden position with a further peek. After betting comes a simultaneous round (Press, Hold, or Bluff) if either player holds a pair, and a card swap is offered if Alice holds a King. At showdown a hand scores the sum of its ranks, plus 3 for a pair of Js and minus 3 for each King. A player needs a plan for subsequent phases, and must engage in reasoning about opponent strategy and private information on both sides. Among the 50 games it sits at the 82nd percentile on state space and the 88th on temporal contingency (Section 3.1).

3.1 Five complexity axes

We describe each game along five axes, all measured through simulated play without using LLM results. The simulations use a population of simple policies: uniform random play and a subset of the 200 sampled policies, selected for performance. The axes are as follows (full definitions are in Appendix C).

- **State space.** This measures the combinatorial size of the game, as seen by a player. We estimate it by counting distinct information states in 3,000 deals of random play, adding the counts for the two seats and taking $\log_{10}$. A state includes the decision, the player's own cards and chips, its last three decisions, the events it has observed and the positions.
- **Temporal contingency.** In a game with several phases, a good decision now may depend on what the players will do later. We measure this dependence by simulating each available action with different policies for subsequent play, and with different draws of the unseen cards. Some of the payoff variation comes from actions being better on average, and some from favorable cards or play regardless of the action. Some variation comes from their interaction: how much the comparison between actions changes with the conditions of play. Temporal contingency is measured as this interaction's share of total variance.
- **Opponent dependence.** Opponent dependence measures how much the relative payoffs of actions depend on who the opponent is. We average over cards and our player's later choices, then measure variation in these payoff differences across opponents.
- **Information.** For this axis, we measure how much a player using limited-depth search, which chooses each action by simulating the rest of the deal (Appendix C), gains from planning around its own cards. We compare two such players. One searches over deals consistent with the cards it has seen. The other does not condition its simulated deals on its own hand. We measure the difference in their average chip leads per deal against the same opponents. The cards in the actual game are unchanged.
- **Precision.** The precision axis measures the cost of small departures from the best action. We compare the best action with its neighbors, using adjacent amounts of the wager grid for bets and bids (Appendix C), and all other actions for categorical choices. The loss is its payoff minus the average over those neighbors. To limit selection bias, the best action is chosen on one half of the simulations and its payoff evaluated on the other.

The temporal, opponent and precision scores are averages over decisions encountered during play, where each visit to a state counts once. For all four axes other than state space, we average over the two seats. Scores measured in chips are divided by the standard deviation of chip lead in population play so that different stakes are comparable. We standardize each axis across the 50 games for analysis (Section 7). Pairwise correlations are at most $|\rho| = 0.50$, and the largest variance inflation factor is 2.0. Appendix C reports the estimation checks and the checks from repeated runs.

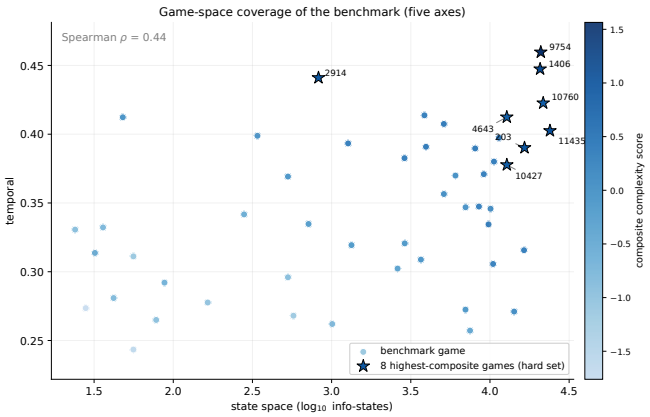

Figure 1: **The 50-game benchmark in two of the five axes** (state space vs. temporal), colored by composite complexity (Section 6). Stars mark the eight hardest games by composite complexity score (Table 4).

4 Benchmark construction

We selected 50 games from the accepted pool of 2,000 using farthest-point sampling on a sampling space composed of six preliminary complexity axes (Appendix C). Each axis was min-max normalized. In selecting games, each next game is picked to maximize its minimum distance to the games already selected. Then, we compute the five axes used in the paper on the set of 50 games. Figure 1 shows the 50 games in two of the five axes.

5 Tournament design

Models. Our tournament features twelve models. Nine models were played in an April–May 2026 tournament block: `gpt-5-4-high`, `gemini-3.1-pro-preview`, `claude-sonnet-4-6-max`, `gemini-2.5-pro`, `gemma-4-31b-it`, `deepseek-v3.1-together`, `gemini-3.1-flash-lite-preview`, `qwen-3.5-together`, and `llama-3.3-70b-together`. For those models, every model played every other model on a subset of games. Three more models were added in a September 2026 tournament block (`gpt-6-astra-high`, `claude-fable-5-1-high`, and `gemini-3-8-flash`). Each September model played three to five other models spread across all 50 games. Because the September models played fewer models each, they have wider confidence intervals, and we mark them † throughout. For every model, we record the provider's model identifier and all request settings, so that runs can be reproduced. In the rest of the paper, we use short model names (e.g., `gpt-5.4` for `gpt-5-4-high`).

Runs and single-deal matches. For each (model pair, game) matchup, a *run* (a paired deal) consists of two deals of the game, where each model plays once as Alice and once as Bob. The two deals share a play seed, so that each model in a seat faces the same conditions. Each matchup is played for up to 20 runs. Every match is a single deal, and models have no memory of earlier deals or identifying information about their opponent. Thus, each match tests planning under uncertainty with a fixed, unobservable opponent. Within a deal, a model sees only the opponent's (visible) actions so far, so opponent dependence is tested through a model's reasoning about likely opponent behavior, and the little evidence available within the deal.

Prompting and corpus. First, a model is provided with the rulebook as a system prompt. At each turn, a model is also given the information it would have been able to observe so far, and a legal action menu. It must then respond with an action, given in JSON format. A lenient parser attempts to fix minor formatting issues, and should it fail, falls back to a random legal action (in the worst case, this was 0.33% of moves, Appendix J). After removing time-outs and other failed or defective

Table 1: **Overall leaderboard** on April block and September block together (20,206 paired deals, 50 games), $\hat{\alpha}$ in chips/game, clustered bootstrap 95% CIs (B=2,000). † September models.

Model	$\hat{\alpha}$	95% CI
`gpt-6-astra`†	$+0.72$	$[+0.52, +0.94]$
`gpt-5.4`	$+0.64$	$[+0.55, +0.73]$
`gemini-3.8-flash`†	$+0.63$	$[+0.42, +0.84]$
`gemini-3.1-pro`	$+0.63$	$[+0.56, +0.70]$
`claude-sonnet-4.6`	$+0.45$	$[+0.34, +0.56]$
`claude-fable-5.1`†	$+0.45$	$[+0.24, +0.64]$
`gemini-2.5-pro`	$+0.16$	$[+0.08, +0.25]$
`gemma-4-31b`	$+0.05$	$[-0.03, +0.13]$
`deepseek-v3.1`	-0.14	$[-0.22, -0.06]$
`gemini-3.1-flash-lite`	-0.47	$[-0.56, -0.39]$
`qwen-3.5`	-0.55	$[-0.64, -0.47]$
`llama-3.3-70b`	-2.56	$[-2.67, -2.45]$

matches, and keeping only complete pairs (i.e., full runs), the full tournament has 40,412 matches. Counts of paired deals per model pair are in Appendix E.

6 Overall results

Estimator. For match s between models i and j, we regress the signed chip margin, the chips i gains minus the chips j gains (winning 3 chips is a margin of 6), on model indicators: $y_s = \alpha_i - \alpha_j + \varepsilon_s$ with $\sum_m \alpha_m = 0$. This is effectively Bradley–Terry (Bradley & Terry, 1952) with chip margins in place of win indicators, since models were instructed to maximize expected chips. $\hat{\alpha}_m$ is an estimate of a model's strength in chips per game. Standard errors come from a bootstrap over runs (B=2,000).

Leaderboard. About three chips per game separate the top and bottom of Table 1. The highest ranked models are `gpt-6-astra`, `gpt-5.4`, `gemini-3.8-flash` and `gemini-3.1-pro`. `claude-sonnet-4.6` and `claude-fable-5.1` follow, tied at $+0.45$. `gemini-2.5-pro`, `gemma-4-31b` and `deepseek-v3.1` form a middle tier, and `gemini-3.1-flash-lite`, `qwen-3.5` and `llama-3.3-70b` the tail. Refitting only the April round-robin gives the nine April models in this same order, Kendall τ=1.00 (Appendix E). The September models, whose intervals are widest, are the least settled: `gpt-6-astra` may or may not lead the next three.

Robustness. Leave-one-out refits with one game removed give mean Kendall τ=0.956 (min 0.879), with tiers intact every time, so models change places only within a tier. Dropping the bottom outlier (Appendix E) gives the same picture. We also refit the leaderboard on random subsets of 12 games, and find a median Spearman correlation of 0.94 with the full ranking (Appendix E). Of the 47 model pairs that met, 41 have a mean head-to-head margin in the direction of the leaderboard, and none of the six exceptions is significant (Appendix E). A win-indicator Bradley–Terry fit orders the models somewhat differently (ρ=0.77, Appendix E).

Strength on each game. We also fit the estimator separately on each game. On each g the twelve $\hat{\alpha}_{m,g}$ sum to zero and, as in the overall fit, are net of the opponents' strength on g. No cell of the 12×50 grid is empty. Splitting its variance two ways (model main effect, model×game interaction) and correcting for estimation noise in each cell, the interaction is about a third of the main effect (ratio 0.34, bootstrap 95% CI $[0.20, 0.54]$), and comparable to it once the outlier `llama-3.3-70b` is excluded (0.65, $[0.14, 1.31]$, Appendix F). Half of the uncorrected ratio of 0.62 is noise.

The hardest games. We collapse the five axes into a composite complexity score, a learned discrimination score fit to the model outcomes: the first principal component of the models' slopes of $\hat{\alpha}_{m,g}$ on each axis separately (weights: temporal $+0.35$, state space $+0.29$, information $+0.09$, precision -0.19, opponent dependence -0.07, Appendix E). Refitting on the eight highest-scoring games (Table 4) widens the spread from 3.3 chips/game on all fifty to 5.6, 1.7 times. The eight games highest on the temporal axis alone, selected without any outcome, give 5.2 (Table 5), and by temporal

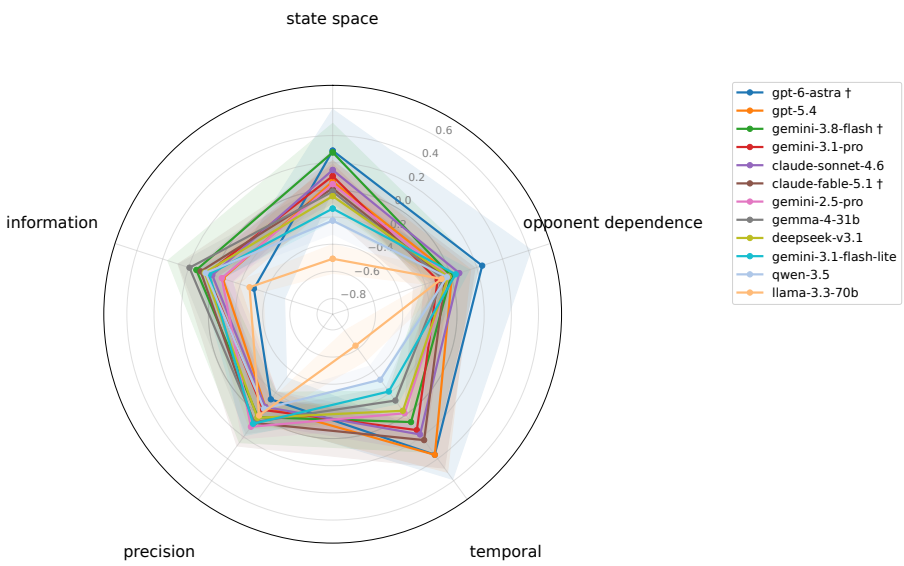

Figure 2: **Capability profile.** For each model, we regress its game strength $\hat{\alpha}_{m,g}$ on all five z-scored axes jointly, with an intercept, over the 50 games. Each spoke shows the coefficients for one axis, in chips/game per SD of the axis, with one polygon per model. A model lies outside the zero ring on an axis when its advantage over the average increases as a game increases in that axis value. Eleven of the 60 cells, six on temporal, are significant after applying Benjamini–Hochberg correction at $q < 0.05$ to clustered bootstrap p-values (B=2,000), and the shaded bands are 95% CIs. Figure 7 marks the significant cells, and the coefficients are in Appendix G.

tertile the leaderboard spans 1.8, 3.0 and 5.0 chips/game. The top six still overlap within the hard set, and we make no claim about their order.

7 Capability Profiles

Two models with similar average chip margins can derive their advantages from different parts of game space. This section asks, for each model, how its advantage changes along each of the five axes. For each model m across the 50 games, we fit the regression $\hat{\alpha}_{m,g} = \beta_{m,0} + \sum_{a=1}^{5} \beta_{m,a} z_a(g) + \varepsilon_{m,g}$. Here, $z_a(g) = (x_a(g) - \bar{x}_a)/s_a$, where $x_a(g)$ is game g's raw value on axis a and $\bar{x}_a$ and s_a are its mean and standard deviation respectively, across the 50 games. We note that $\sum_g z_a(g) = 0$ for every axis, so the intercept is m's average advantage, $\hat{\beta}_{m,0} = \frac{1}{50} \sum_g \hat{\alpha}_{m,g}$. The coefficient $\hat{\beta}_{m,a}$ denotes the predicted change in $\hat{\alpha}_{m,g}$ when $z_a(g)$ increases by one. Intervals use B=2,000 resamples (clustered at the run level), and we apply a Benjamini–Hochberg (BH) correction across the 60 model×axis entries (Figures 2 and 7). Eleven of the 60 slope coefficients are significant, of which six are on the temporal axis, suggesting that the temporal contingency axis is important in understanding the differential capabilities of the different models.

Factor analysis. We standardize each model's $\hat{\alpha}_{m,g}$ across games and take the first factor of the singular value decomposition (Appendix G.1). The higher a game's score on this factor, the more the stronger models gain and the weaker models lose. One factor accounts for 38% of the standardized variance. The score correlates with temporal contingency at $\rho = +0.73$. Its correlation with state space is $+0.65$. After we remove the factor, no residual correlation between two models is significant after Benjamini–Hochberg correction at $q < 0.05$ across the 66 model pairs.

Local jaggedness. A model's strength can vary smoothly with the axes, or it can jump between games that are alike. These jumps can make performance on a new game hard to anticipate. For each model we fit a smooth surface of its strength on each game over the five axes, where each game is predicted from the other 49. The local jaggedness J_m is the standard deviation of the residuals

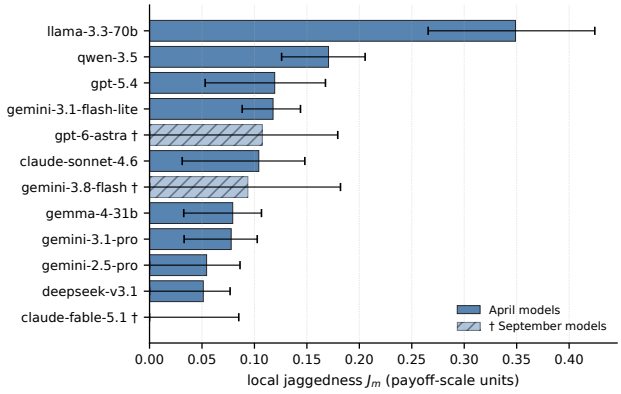

Figure 3: **Local jaggedness.** J_m is the spread of a model's advantage around a smooth function of the five axes, net of sampling noise. 95% CIs from a bootstrap over games (B=500). Hatched bars mark the September models, whose strength on most games rests on a single matchup.

Table 2: **Programmed opponent ablation.** 520 complete pairs on 20 games, with bootstrap 95% CIs over deals. Full table in Appendix I.

	disclosed	**hidden**
transfer slope λ of margin on tournament rating	$1.10\ [0.76, 1.41]$	$0.92\ [0.24, 1.58]$
disclosure effect, disclosed − hidden on the same deals (chips/game)	$+0.72\ [0.19, 1.30]$	
*Efficiency on the five solved games (*0 = *worst response,* 1 = *best response to the program)*		
`gpt-5.4`	$1.02\ [0.82, 1.28]$	$0.81\ [0.70, 0.90]$
`claude-sonnet-4.6`	$1.01\ [0.80, 1.25]$	$0.71\ [0.53, 0.84]$
`llama-3.3-70b`	$0.53\ [0.37, 0.72]$	$0.53\ [0.24, 0.79]$

after we subtract the sampling variance of the estimates for each game (Appendix H), in units of each game's payoff scale under random play. `llama-3.3-70b` (0.35) is by far the most jagged. `qwen-3.5` follows (0.17), and the April frontier models lie between 0.08 and 0.12 (Figure 3). For `gpt-6-astra` and `gpt-5.4` the surface explains about a third of the variation across games ($R^2 = 0.32$ and 0.35), mostly their gain on temporally contingent games. The September models met one opponent per game on most games, and their values are less precise.

8 Ablations and Trace Analysis

We run two ablations and one observational analysis of reasoning traces (Appendix I). They test whether improvements in strategic performance come from knowing the opponent, from more reasoning, or from reasoning that matches the specific challenges of a game.

Disclosed and hidden strategy programs. We study how tournament ratings fare against opponents that are not LLMs, and how a model responds to an opponent's strategy it is made aware of. Three author models (`claude-sonnet-4.6`, `gemini-3.1-pro` and `gemma-4-31b`) wrote complete strategy programs for both seats of 20 games. Each program is run by the engine and rendered into a procedural reading that specifies its moves. Five models played the programs on identical deals with the programmed opponent strategy appended to the rulebook (the *disclosed* arm), and three of them played the programmed opponent whose strategy was hidden (the *hidden* arm). We fit margin $= \lambda_{\text{arm}}\hat{\alpha}_{\text{player}} - \gamma_{\text{group,game,arm}}$, where λ_{arm} is the slope of margin on tournament rating and γ is a fixed effect for each author group, game and arm, excluding each model's own author's programs (Table 2 and Appendix I).

The leaderboard transfers to the case where opponent strategies are disclosed, but we do not find evidence that disclosure helps strong or weak models differentially. Disclosing the program helps, by $+0.72$ chips/game over hidden play on the same deals (95% CI $[0.19, 1.30]$), with

$+0.87$ $[+0.27, +1.53]$ for `gpt-5.4`, $+0.90$ $[-0.23, +2.08]$ for `claude-sonnet-4.6` and $+0.45$ $[-0.55, +1.53]$ for `llama-3.3-70b`, the weakest model. Only the benefit for `gpt-5.4` is individually significant. Against disclosed programs, a model's margin rises with its tournament rating at $\lambda_{\text{disclosed}}=1.10$ $[0.76, 1.41]$, which is consistent with one chip of margin for each chip of rating. Against hidden programs the slope is $\lambda_{\text{hidden}}=0.92$ $[0.24, 1.58]$, and the difference of $+0.18$ $[-0.40, +0.75]$ is consistent with disclosure helping strong and weak models equally. We do not test whether a model could work out its opponent over repeated play. Moreover, on five games we solve exactly for the best response to each program, and score efficiency from the worst response (0) to the best (1). The three frontier models score 0.98 to 1.02 (values above 1 are within sampling error of the best response), close to fully exploiting the programmed strategy. Without disclosure, frontier model efficiency drops to 0.81 $[0.70, 0.90]$ and 0.71 $[0.53, 0.84]$ (understandably, models cannot fully exploit a strategy they have not been explicitly told about). `llama-3.3-70b` scores 0.53 in both arms.

Thinking budget. We played low and high effort variants of four families (`gpt-5.4`, `claude-sonnet-4.6`, `gemini-3.1-pro`, `gemini-2.5-pro`) against the same opponents on identical deals. Within each of the 260 sibling pairs (the two variants of a family on the same deal and opponent), a difference in chip margin can be attributed to the thinking budget. In every family the high effort variant won more chips, from $+0.06$ chips/game for `gemini-2.5-pro` to $+0.53$ for `gemini-3.1-pro`, but only the `gemini-3.1-pro` gain is significant.

Reasoning traces. We use an LLM judge (`gemini-3-flash-preview`) to score $9{,}778$ move traces on three rubrics (opponent modeling, private information, planning depth). How much of each type of reasoning a trace contains barely tracks its matched axis. Only opponent modeling does, and weakly. Reasoning that matches a game's complexities is nonetheless associated with higher margins on that game, for private information and planning depth though not for opponent modeling. Reasoning about private information is associated with $+1.15$ chips per rubric point more on games in the top tertile of the information axis, and planning depth $+0.93$ more in the top tertile of the temporal axis.

9 Limitations and conclusion

GENSTRAT covers only zero-sum betting games for two players with imperfect information, leaving cooperative, multiplayer and agentic settings and play over several deals to future work. Within this scope, the axis results are robust at fifty games, although a larger benchmark would tighten them. The state space axis is moderately correlated with three of the other four. The three September models met three to five opponents each, giving them the widest intervals and, on most games, profiles resting on one matchup. Our audit of the served rulebooks against the engine found no engine defect. Some rulebooks leave out details the engine resolves itself, such as whether a prize card is random, and 12 misstate when a hidden cost is deducted. These defects are symmetric across seats, and dropping the affected games leaves the tiers of the leaderboard unchanged (Appendix J). Matches affected by two bugs, since fixed, were rerun where possible (Appendix J).

We introduced GENSTRAT, which makes evaluation renewable, and allows games of arbitrary complexity to be drawn on demand. On this distribution we found three main results. First, frontier models had a significant advantage in the tournament over older or open-source models. Second, similarly strong models differed in what kinds of games they excelled in, with certain kinds of model gaining more on games with more temporal contingency. Finally, models differed in local jaggedness, the part of a model's advantage that departs from its smooth trend over the axes, with `llama-3.3-70b` by far the most jagged. In ablations, disclosing an opponent's strategy raised margins by about 0.7 chips per game, a higher thinking budget raised them in all four families tested, significantly in one, and reasoning matched to the demands of a game was associated with higher margins on that game, for private information and planning depth. The generator makes the experiment repeatable on newly generated games.

AI USE STATEMENT

Twelve LLMs played the games in this study. We used LLM agents to develop features of the game generator. We used coding assistants to write statistical analysis code, and to draft manuscript text. Human auditors and LLM agents checked rulebooks for consistency, correctness, and clarity (Appendix J). Three LLMs wrote opponent strategy programs used in the program ablation (Section 8, Appendix I). LLM judges scored reasoning traces using a handwritten rubric (Section 8, Appendix I).

The authors developed research questions and designed the generator, tournament, and experimental axes. We reviewed all AI-assisted code and regenerated reported numbers. We reviewed and edited all text drafted with AI assistance. Authors take full responsibility for the final paper and its accompanying artifacts.

ETHICS STATEMENT

We evaluate commercial and open-source LLMs on generated card games. The study involves no human subjects or personal data. No real money was at stake.

REPRODUCIBILITY STATEMENT

Each game is determined by its integer seed and its stored generation configuration (Section 3). Sections 3 and 4 describe the evaluation framework in detail. Appendix C defines the five complexity axes. Tournament design details are provided in Section 5 and Appendix J. Estimators for statistical analyses are made available in Sections 6 and 7. Further details are provided in Appendices E, G, and H. Each model's provider identifier and request settings are recorded with the data.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Anthropic. Project Vend: Can Claude run a small shop? (And why does that matter?). Anthropic Research, 2025. URL `https://www.anthropic.com/research/project-vend-1`.

Anthropic. Project Deal: our Claude-run marketplace experiment. Anthropic, 2026. URL `https://www.anthropic.com/features/project-deal`.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Andreas Buja and Nermin Eyuboglu. Remarks on Parallel Analysis. *Multivariate Behavioral Research*, 27(4): 509–540, 1992.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J K Terry. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pp. 73383–73394, 2023.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056. PMLR, 2020.

Katherine M. Collins, Cedegao E. Zhang, Graham Todd, Lance Ying, Mauricio Barba da Costa, Ryan Liu, Prafull Sharma, Adrian Weller, Ionatan Kuperwajs, Lionel Wong, Joshua B. Tenenbaum, and Thomas L. Griffiths. Evaluating Language Models' Evaluations of Games. In *International Conference on Learning Representations (ICLR)*, 2026.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents, 2024. arXiv:2406.06613v2.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 28219–28253, 2024.

FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic Collusion by Large Language Models, 2024. arXiv:2404.00806v5, revised 2026.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. Suspicion Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT-4. In *First Conference on Language Modeling (COLM)*, 2024.

John L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2): 179–185, 1965.

Chenghao Huang, Yanbo Cao, Yinlong Wen, Tao Zhou, and Yanru Zhang. PokerGPT: An End-to-End Lightweight Solver for Multi-Player Texas Hold'em via Large Language Model, 2024. arXiv:2401.06781.

Harold W. Kuhn. A simplified two-person poker. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games, Vol. I*, number 24 in Annals of Mathematics Studies, pp. 97–103. Princeton University Press, 1950.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. AvalonBench: Evaluating LLMs Playing the Game of Avalon, 2023. arXiv:2310.05036v3.

Hsieh-Ting Lin and Tsung-Yu Hou. Readable Minds: Emergent Theory-of-Mind-Like Behavior in LLM Poker Agents, 2026. arXiv:2604.04157.

Minhua Lin, Enyan Dai, Hui Liu, Xianfeng Tang, Yuliang Yan, Zhenwei Dai, Jingying Zeng, Zhiwei Zhang, Fali Wang, Hongcheng Gao, Chen Luo, Xiang Zhang, Qi He, and Suhang Wang. How Far Are LLMs from Professional Poker Players? Revisiting Game-Theoretic Reasoning with Agentic Tool Use. In *International Conference on Learning Representations (ICLR)*, 2026.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Noor Shaker, Julian Togelius, and Mark J Nelson. *Procedural Content Generation in Games*. Springer, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' Bluff: Opponent Modelling in Poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 550–558, 2005.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

Tomer Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, 2023. arXiv:2302.08399v5.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the World Model Implicit in a Generative Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 26941–26975, 2024a.

Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48919–48937. PMLR, 2024b.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. In *International Conference on Machine Learning (ICML)*, 2025.

Vivek Verma, David Huang, William Chen, Dan Klein, and Nicholas Tomlin. Measuring General Intelligence with Generated Games, 2025. arXiv:2505.07215.

A Modular game design

Construction of GBGs proceeds via a GBG builder function. The core layer of this builder includes the underlying `GameState` (players, variables, piles), the `Rulebook` (containing the phase graph that describes the progression of play), and `Phase` (ordered lists of actions). A GameState evolves via game operations—events that happen in-game. These include `Deal`, `ChipTransfer`, `Shuffle`, `PeekAtHand`, `SwapWithOpponent`, `StealCardMove`, `Wager`, and `ScoreAdjustment`. Each module contains an NL description of its own logic, allowing for automated generation of rulebooks by stitching together individual NL components. Then, in the phase graph, transitions between phases can be conditioned on variables and game outcomes. As a result, games can flow non-linearly via branches, bounded loops, or conditional sub-phases. A complexity parameter $c \in [0, 1]$ determines how often complex structural features are drawn. As we increase c, the distribution changes gradually from simple games (like Kuhn poker) to more complex games with compound phase graphs. Given a game's seed s and its stored generation configuration, the builder function can build the game $G = \texttt{reconstruct}(s)$ deterministically, making games reproducible.

Natural language rulebooks (Appendix B) are generated from the phase graph and provided to the LLM as a system prompt. The natural language version communicates all rules and structure of the game.

Sampled features can combine to produce a game that plays poorly. We address this with deterministic construction rules and acceptance checks based on $2{,}000$ episodes of random play. We check that conditions trigger sufficiently often, that phases are generally reachable, and that the two seats are reasonably balanced under random play. A game is rejected if random play averages more than ten moves per player, if more than 40% of its conditions hold in fewer than 15% or more than 85% of their evaluations, if more than 20% of its phases are reached in fewer than 5% of episodes, if either seat wins more than 65% of the episodes that have a winner, or if a static check of the rulebook finds a branch that can never fire. Of $12{,}351$ seeds tried, $2{,}000$ were accepted.

B Sample game rulebook

Compared to the 50 benchmark games, seed 4643 scores high on state space (82nd percentile) and temporal contingency (88th), near the middle on precision (48th) and information (46th) and opponent dependence (44th). The rulebook below is the verbatim system prompt that the LLM agents receive at the start of play. It is generated deterministically from the phase graph of the game.

```
## Overview
**Players:** Alice, Bob
**Player order:** Player 1 = Alice (acts first in betting), Player 2 = Bob.
**Deck:** 8 cards (10, J, Q, K of clubs and hearts)
**Starting chips:** 10 each (visible to all). Chips move between players' stacks and the pot; the total
of all chips (stacks plus pot) is conserved across the game.
**Pot:** starts at 0
**General rules:**
- **Information:** The number of cards remaining in the Deck is public information. Hand sizes are
private. Players do not know how many cards opponents hold unless revealed by an action. If the Deck is
empty when a draw action occurs, no cards are drawn and the action has no effect. Public draws emit a
public skipped-draw notice; private draws emit that notice only to the drawing player - but per the
inferability principle, the Deck-size change from any successful draw is publicly visible.
- **Legal actions:** Players may choose any listed option. If its preconditions are unmet (empty hand),
the action has no effect.
- **Resolution:** In any multi-step effect, each step resolves in order; if one step's precondition is
unmet, that step is skipped and later steps still resolve.
- **Game end:** If the game ends during any phase, all remaining phases are skipped.
- **Costs & pot:** Costs are paid to the pot before the effect resolves, unless otherwise specified.
When an action lists a cost: if the player cannot afford the stated cost at the moment they select it,
the action is not offered as an option. If the cost is probabilistic (e.g. a 50% chance of paying 1
chip), the player commits to the action first and the cost is resolved afterwards; if they cannot afford
the resolved cost, the action is skipped and their choice slot is consumed (they do not get to pick
again).
- **Showdown & misc.:** There is no hand size limit; hands grow and shrink freely via draws. Unless
stated otherwise, all random selections are uniform. At Showdown, each player's hand is revealed as of
the moment the Showdown phase begins, including any cards swapped, drawn, or stolen during earlier
phases.
## Visibility
- **Hands**: Each player's cards are private to that player. Opponents cannot see hand contents unless
explicitly revealed.
- **Actions**: All player actions (bets, checks, folds) and their effects are publicly announced unless
stated otherwise. Actions marked as private are executed silently.
- **Inferability**: "Private" hides *which branch fired, what card was drawn/peeked, or who chose what* -
 not the observable side-effects. Deck sizes, pot totals, chip stacks, and position holdings are always
public and reconcile at settlement. If one private branch would change a public number (e.g. shrink the
Deck by 1) and another would not, the opponent can infer which branch ran from that number. Treat the
privacy label as concealing the *identity/choice*, not the *occurrence*.
- **Position ownership**: Visibility of position assignments is stated per position (public or hidden).
- **Condition checks**: Checks marked "(public)" are announced to all players when evaluated.
- **Branch taken**: Private conditional blocks do not announce which branch was taken, even if players
can infer it.
- **Private phases**: Phases marked *(private)* do not announce their start or end. Individual actions
and conditions within them still follow their own visibility rules.
- **Card identity**: Peeked or observed cards are private to the observer unless stated otherwise.
- **Hidden information**: All hidden information is tracked and enforced; players are told only what the
rules specify.

## Positions
A player may hold multiple positions simultaneously.
- **Sage** -- only one player can hold this position at a time. Assigning it to a new player removes it
from the previous holder. The identity of the holder is hidden from other players.
- **Scout** -- any number of players may hold this position. Holding it is boolean (receiving it again
has no additional effect). The identity of the holder is hidden from other players.
- **Sentinel** -- any number of players may hold this position. Holding it is boolean (receiving it
again has no additional effect). The identity of the holder is hidden from other players.
Positions are held until explicitly reassigned by a later assignment phase. Assignment predicates,
including predicates over mutable state such as chip counts or card counts, are evaluated only at the
moment that assignment phase runs, not on every reference. Once assigned, a position persists even if
the original predicate later becomes false.
- **Inferability of hidden positions:** effects gated by a hidden position (e.g. chip movements, card
draws) may themselves be publicly observable. The assignment is not announced, but opponents may infer
the holder from outcomes.

## Sequence of Play
1. **Setup**: shuffle deck, deal cards
2. **Determine Positions**
3. **Ante Phase**: each player antes into the pot
4. **Observations**
5. **Betting Phase**: players may bet, check, fold, or call
6. **Post-Betting Phase**
7. **Showdown**: reveal hands and settle the pot

## Phase Details

### Setup
- Shuffle the Deck.
- Deal 3 cards face-down to each player (Alice, Bob) (or as many as remain if the deck is short). Each
player sees only their own cards.

### Determine Positions
- If Alice's Hand has exactly 3 Clubs: Alice receives the position Sentinel. (Only the recipient is
notified.)
- If All cards in Bob's Hand are the same suit: Bob receives the position Sentinel. (Only the recipient
is notified.)
- If Alice's Hand has exactly 0 Clubs: Alice receives the position Scout. (Only the recipient is
notified.)
- If Bob's Hand has at least 2 Clubs: Bob receives the position Scout. (Only the recipient is notified.)

### Ante Phase
```

```
- Alice antes 3 chips (moved from Alice's chip stack into the pot).
- Bob antes 3 chips (moved from Bob's chip stack into the pot).

### Observations
- If Alice holds Sentinel: Alice peeks at 1 card in Bob's hand (selected uniformly at random). The card
remains in Bob's hand. The peek is private: only Alice sees the card. If Bob's hand is empty, Alice is
informed that it is empty.
- If Bob holds Sentinel: Bob peeks at 1 card in Alice's hand (selected uniformly at random). The card
remains in Alice's hand. The peek is private: only Bob sees the card. If Alice's hand is empty, Bob is
informed that it is empty.
- If Alice holds Scout: Alice peeks at 1 card in Bob's hand (selected uniformly at random). The card
remains in Bob's hand. The peek is private: only Alice sees the card. If Bob's hand is empty, Alice is
informed that it is empty.
- If Bob holds Scout: Bob peeks at 1 card in Alice's hand (selected uniformly at random). The card
remains in Alice's hand. The peek is private: only Bob sees the card. If Alice's hand is empty, Bob is
informed that it is empty.

### Betting Phase
- Betting proceeds for up to 2 rounds.
- Each round, Betting Round is played. This continues until a player has 3 or fewer chips, the round cap
(2) has been reached, or the game has ended.

#### Sub-phase: Betting Round *(called from Betting Phase)*
- «bet round» starts at 0 and increases by 1 after each round.
- If «bet round» is greater than 0 and the game is still in progress (public): Perform **Pre-Betting
Actions** steps (both players participate per this sub-phase's own rules), then continue.
- If «bet round» is greater than 0 and the game is still in progress (public): Perform **Inter-Round
Draw (All Players, inter-round event 1)** steps, then continue.
- If «bet round» is greater than 0 and the game is still in progress: Perform **Assign Position: Sage**
steps, then continue.
- If «bet round» is greater than 0 and the game is still in progress: Perform **Dynamic Role Effects**
steps, then continue.
- If «bet round» is equal to 0 (public):
    - If Alice has chips remaining and Bob has chips remaining (public):
        Round 1: variable betting, wager any amount from 1 chip up to your remaining stack.
        Then, perform **Variable Betting Round (Round 1)** steps, then continue.
    - Otherwise:
        Round 1: variable betting unaffordable, falling back to fixed 1-chip betting for this round.
        Then, Alice picks one of: bet or check. (choice is announced to all players)
        - If Alice chooses 'check' (public): Perform **Bob: Bet or Check (Round 1, Fallback)** steps,
then continue.
        - Otherwise, if Alice chooses 'bet':
            Alice bets 1 chip (moved from Alice's chip stack into the pot).
            Then, perform **Bob: Fold or Call (Round 1, Fallback)** steps, then continue.
  - Otherwise:
    Round 2: fixed betting, each bet or call is 1 chip.
    Then, Alice picks one of: bet or check. (choice is announced to all players)
    - If Alice chooses 'check' (public): Perform **Bob: Bet or Check (Round 2)** steps, then continue.
    - Otherwise, if Alice chooses 'bet':
        Alice bets 1 chip (moved from Alice's chip stack into the pot).
        Then, perform **Bob: Fold or Call (Round 2)** steps, then continue.
*(Play returns to **Betting Phase**.)*

#### Sub-phase: Pre-Betting Actions *(called from Betting Round)*
- Alice picks one of: Peek at opponent's hand (50% chance to cost 1 chip) or Pass. (the choice itself is
not announced; any public side-effect a branch happens to produce - e.g. chip or card movements - can
still reveal which branch ran)
  - If Alice chooses 'Pass': Alice does nothing.
  - Otherwise, if Alice chooses 'Peek at opponent's hand (50% chance to cost 1 chip)': Alice peeks at 1
card in Bob's hand (selected uniformly at random). The card remains in Bob's hand. The peek is private:
only Alice sees the card. If Bob's hand is empty, Alice is informed that it is empty. Has a 50% chance
of costing 1 chip (paid to the pot); otherwise free.
- Bob picks one of: Peek at opponent's hand (50% chance to cost 1 chip) or Pass. (the choice itself is
not announced; any public side-effect a branch happens to produce - e.g. chip or card movements - can
still reveal which branch ran)
  - If Bob chooses 'Pass': Bob does nothing.
  - Otherwise, if Bob chooses 'Peek at opponent's hand (50% chance to cost 1 chip)': Bob peeks at 1 card
in Alice's hand (selected uniformly at random). The card remains in Alice's hand. The peek is private:
only Bob sees the card. If Alice's hand is empty, Bob is informed that it is empty. Has a 50% chance of
costing 1 chip (paid to the pot); otherwise free.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Inter-Round Draw (All Players, inter-round event 1) *(called from Betting Round)*
- If the Deck has at least 2 cards (public): Inter-round draw: each player draws 1 card from the Deck
(private).
  - Otherwise: Inter-round draw: the Deck has fewer than 2 cards remaining; the both-players draw is
skipped.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Assign Position: Sage *(private)* *(called from Betting Round)*
*Players are not told when this phase begins or ends.*
- If Alice has more chips than Bob: Alice receives the position Sage. (Only the recipient is notified.)
  - Otherwise, if Bob has more chips than Alice: Bob receives the position Sage. (Only the recipient is
notified.)
  - Otherwise (tied): Coin flip -- heads: Alice receives the position Sage. (Only the recipient is
notified.); tails: Bob receives the position Sage. (Only the recipient is notified.).
    *(The branch condition is evaluated privately; which branch fires is not announced directly.)*
*(Play returns to **Betting Round**.)*

#### Sub-phase: Dynamic Role Effects *(private)* *(called from Betting Round)*
```

```
*Players are not told when this phase begins or ends.*
- If Alice holds Sage: Alice peeks at 1 card in Bob's hand (selected uniformly at random). The card
remains in Bob's hand. The peek is private: only Alice sees the card. If Bob's hand is empty, Alice is
informed that it is empty.
- If Bob holds Sage: Bob peeks at 1 card in Alice's hand (selected uniformly at random). The card
remains in Alice's hand. The peek is private: only Bob sees the card. If Alice's hand is empty, Bob is
informed that it is empty.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Bob: Fold or Call (Round 2) *(called from Betting Round)*
- Bob picks one of: fold or call. (choice is announced to all players)
  - If Bob chooses 'call' (public): Bob bets 1 chip (moved from Bob's chip stack into the pot).
  - Otherwise, if Bob chooses 'fold':
      The entire pot is given to Alice.
      Then, the game ends: Alice wins after Bob folds.

#### Sub-phase: Bob: Bet or Check (Round 2) *(called from Betting Round)*
- Bob picks one of: bet or check. (choice is announced to all players)
  - If Bob chooses 'check' (public): No action.
  - Otherwise, if Bob chooses 'bet':
      Bob bets 1 chip (moved from Bob's chip stack into the pot).
      Then, perform **Alice: Fold or Call (responding, Round 2)** steps, then continue.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Alice: Fold or Call (responding, Round 2) *(called from Bob: Bet or Check (Round 2))*
- Alice picks one of: fold or call. (choice is announced to all players)
  - If Alice chooses 'call' (public): Alice bets 1 chip (moved from Alice's chip stack into the pot).
  - Otherwise, if Alice chooses 'fold':
      The entire pot is given to Bob.
      Then, the game ends: Bob wins after Alice folds.

#### Sub-phase: Bob: Fold or Call (Round 1, Fallback) *(called from Betting Round)*
- Bob picks one of: fold or call. (choice is announced to all players)
  - If Bob chooses 'call' (public): Bob bets 1 chip (moved from Bob's chip stack into the pot).
  - Otherwise, if Bob chooses 'fold':
      The entire pot is given to Alice.
      Then, the game ends: Alice wins after Bob folds.

#### Sub-phase: Bob: Bet or Check (Round 1, Fallback) *(called from Betting Round)*
- Bob picks one of: bet or check. (choice is announced to all players)
  - If Bob chooses 'check' (public): No action.
  - Otherwise, if Bob chooses 'bet':
      Bob bets 1 chip (moved from Bob's chip stack into the pot).
      Then, perform **Alice: Fold or Call (responding, Round 1, Fallback)** steps, then continue.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Alice: Fold or Call (responding, Round 1, Fallback) *(called from Bob: Bet or Check
(Round 1, Fallback))*
- Alice picks one of: fold or call. (choice is announced to all players)
  - If Alice chooses 'call' (public): Alice bets 1 chip (moved from Alice's chip stack into the pot).
  - Otherwise, if Alice chooses 'fold':
      The entire pot is given to Bob.
      Then, the game ends: Bob wins after Alice folds.

#### Sub-phase: Variable Betting Round (Round 1) *(called from Betting Round)*
- Alice and Bob may each bet any whole number of chips (minimum 1). When a player bets, their chosen
amount is transferred from their chip stack into the pot immediately. When a player calls, their matched
amount is transferred to the pot; if they cannot match the full bet, they go all-in with their remaining
chips and the excess from the original bettor is returned to that bettor. If both players check in
sequence, the round ends with no additional chips moving.
- Alice picks one of: bet or check. (choice is announced to all players)
  - If Alice chooses 'check' (public): Perform **Bob: Bet or Check (Round 1)** steps, then continue.
  - Otherwise, if Alice chooses 'bet':
      Alice chooses a wager amount (minimum 1 chip, up to Alice's current chip count). When Alice
commits the chosen amount, it is transferred from Alice's chip stack into the pot via the normal public
chip announcement. If the submitted amount exceeds the legal maximum, the bid is clamped to the maximum
(all-in or declared cap). If the submitted amount is malformed or below the minimum, the engine
substitutes a uniform-random legal amount. In either case the submitting player is notified privately
and play continues with the resolved amount.
      Then, perform **Bob: Fold or Call (Round 1)** steps, then continue.
*(Play returns to **Betting Round**.)*

#### Sub-phase: Bob: Fold or Call (Round 1) *(called from Variable Betting Round (Round 1))*
- Bob picks one of: fold or call. (choice is announced to all players)
  - If Bob chooses 'call' (public): Bob matches the bet. If they cannot match the full amount, they go
all-in and only the matched portion from each player remains in the pot.
  - Otherwise, if Bob chooses 'fold':
      The entire pot is given to Alice.
      Then, the game ends: Alice wins after Bob folds.

#### Sub-phase: Bob: Bet or Check (Round 1) *(called from Variable Betting Round (Round 1))*
- Bob picks one of: bet or check. (choice is announced to all players)
  - If Bob chooses 'check' (public): No action.
  - Otherwise, if Bob chooses 'bet':
      Bob chooses a wager amount (minimum 1 chip, up to Bob's current chip count). When Bob commits the
chosen amount, it is transferred from Bob's chip stack into the pot via the normal public chip
announcement. If the submitted amount exceeds the legal maximum, the bid is clamped to the maximum
(all-in or declared cap). If the submitted amount is malformed or below the minimum, the engine
substitutes a uniform-random legal amount. In either case the submitting player is notified privately
and play continues with the resolved amount.
```

```
      Then, perform **Alice: Fold or Call (responding, Round 1)** steps, then continue.
*(Play returns to **Variable Betting Round (Round 1)**.)*

#### Sub-phase: Alice: Fold or Call (responding, Round 1) *(called from Bob: Bet or Check (Round 1))*
- Alice picks one of: fold or call. (choice is announced to all players)
  - If Alice chooses 'call' (public): Alice matches the bet. If they cannot match the full amount, they
go all-in and only the matched portion from each player remains in the pot.
  - Otherwise, if Alice chooses 'fold':
      The entire pot is given to Bob.
      Then, the game ends: Bob wins after Alice folds.

### Post-Betting Phase
- Post-betting events may fire based on game state. Each of the following conditions is evaluated
top-to-bottom at its own step; every condition whose check passes at that moment performs its associated
action. The conditions are not mutually exclusive: more than one may fire in the same phase.
- If the current pot is more than 3 chips (public): Perform **Observation Round (post-betting)** steps,
then continue.
- If at least one player has two cards of the same rank in their hand (public): Perform **Simultaneous
Round (post-betting)** steps (both players participate per this sub-phase's own rules), then continue.
- If Alice's hand contains a King (the deck's top rank) (public): Perform **Post-Betting Actions** steps
(both players participate per this sub-phase's own rules), then continue.

#### Sub-phase: Observation Round (post-betting) *(private)* *(called from Post-Betting Phase)*
*Players are not told when this phase begins or ends.*
- This observation round is private: non-observers are not notified that any observation occurs.
- Sample 1 card(s) from the current contents of the Deck with replacement (i.e., the Deck is unchanged).
If the Deck is empty, 0 cards are observed. The holder(s) of the Sentinel position (if any) privately
observe the sampled cards. Sampled cards are not kept. Only the observer sees which cards were sampled.
If no player holds the designated position(s) at this point, the observation is skipped.
- Sample 1 card(s) from the current contents of the Deck with replacement (i.e., the Deck is unchanged).
If the Deck is empty, 0 cards are observed. The holder(s) of the Sage position (if any) privately
observe the sampled cards. Sampled cards are not kept. Only the observer sees which cards were sampled.
If no player holds the designated position(s) at this point, the observation is skipped.
- If Alice holds Sentinel: Sample 1 card(s) from the current contents of Bob's Hand with replacement
(i.e., Bob's Hand is unchanged). If Bob's Hand is empty, 0 cards are observed. Alice privately observes
the sampled cards. Sampled cards are not kept. Only the observer sees which cards were sampled.
  - Otherwise: No effect.
    *(The branch condition is evaluated privately; which branch fires is not announced directly.)*
- If Bob holds Sentinel: Sample 1 card(s) from the current contents of Alice's Hand with replacement
(i.e., Alice's Hand is unchanged). If Alice's Hand is empty, 0 cards are observed. Bob privately
observes the sampled cards. Sampled cards are not kept. Only the observer sees which cards were sampled.
  - Otherwise: No effect.
    *(The branch condition is evaluated privately; which branch fires is not announced directly.)*
*(Play returns to **Post-Betting Phase**.)*

#### Sub-phase: Simultaneous Round (post-betting) *(called from Post-Betting Phase)*
- All players choose simultaneously and in secret. No player knows the others' choices until all are
locked in. Once all choices are submitted, each player's choice is publicly revealed, then the combined
outcome is resolved and announced publicly. If a player fails to choose, a uniformly random option is
selected for them.
- Alice chooses one of the following options: Press, Hold, or Bluff.
- Bob chooses one of the following options: Press, Hold, or Bluff.
- Outcomes based on choices:
  - If Alice chooses 'Press' AND Bob chooses 'Hold':
      - If Bob holds Sage: Nothing occurs.
      - Otherwise: Bob pays 1 chip to Alice. If Bob has fewer than 1 chip, only the available amount is
transferred.
          *(The branch condition is evaluated privately; which branch fires is not announced directly.)*.
          *(Note: different branches produce different observable chip effects; the branch taken may be
partially inferable from chip changes.)*.
  - If Alice chooses 'Hold' AND Bob chooses 'Hold':
      - If Bob holds Sage: Nothing occurs.
      - Otherwise: Bob pays 1 chip to Alice. If Bob has fewer than 1 chip, only the available amount is
transferred.
          *(The branch condition is evaluated privately; which branch fires is not announced directly.)*.
          *(Note: different branches produce different observable chip effects; the branch taken may be
partially inferable from chip changes.)*.
  - If Alice chooses 'Hold' AND Bob chooses 'Bluff':
      - If Bob holds Sage: Nothing occurs.
      - Otherwise: Bob pays 1 chip to Alice. If Bob has fewer than 1 chip, only the available amount is
transferred.
          *(The branch condition is evaluated privately; which branch fires is not announced directly.)*.
          *(Note: different branches produce different observable chip effects; the branch taken may be
partially inferable from chip changes.)*.
  - If Alice chooses 'Press' AND Bob chooses 'Bluff':
      - If Alice's highest-ranked card has a higher rank than Bob's highest-ranked card:
          Alice draws 1 card from the Deck. Alice sees the card identity; the Deck-size decrement is
public, signaling to both players that a draw occurred.
          Then, Bob pays 2 chips to Alice. If Bob has fewer than 2 chips, only the available amount is
transferred.
      - Otherwise:
          Alice draws 1 card from the Deck. Alice sees the card identity; the Deck-size decrement is
public, signaling to both players that a draw occurred.
          Then, Alice pays 2 chips to Bob. If Alice has fewer than 2 chips, only the available amount is
transferred.
          *(The branch condition is evaluated privately; which branch fires is not announced directly.)*.
          *(Note: different branches produce different observable chip effects; the branch taken may be
partially inferable from chip changes.)*.
  - If Alice chooses 'Hold' AND Bob chooses 'Press':
```

```
      Alice draws 1 card from the Deck. Alice sees the card identity; the Deck-size decrement is public,
signaling to both players that a draw occurred.
      Then, Alice pays 1 chip to Bob. If Alice has fewer than 1 chip, only the available amount is
transferred.
  - If Alice chooses 'Bluff' AND Bob chooses 'Hold':
      Alice draws 1 card from the Deck. Alice sees the card identity; the Deck-size decrement is public,
signaling to both players that a draw occurred.
      Then, Alice pays 1 chip to Bob. If Alice has fewer than 1 chip, only the available amount is
transferred.
  - For all other combinations (Press/Press, Bluff/Press, Bluff/Bluff):
      Alice pays 1 chip to Bob. If Alice has fewer than 1 chip, only the available amount is transferred.
      Then, Bob draws 1 card from the Deck. Bob sees the card identity; the Deck-size decrement is
public, signaling to both players that a draw occurred
*(Play returns to **Post-Betting Phase**.)*

#### Sub-phase: Post-Betting Actions *(called from Post-Betting Phase)*
- Alice picks one of: Swap with opponent or Pass. (the choice itself is not announced, but some branches
produce public side-effects - chip movements, card exchanges, or other observable state changes - that
reveal which branch ran; other branches leave no public trace)
  - If Alice chooses 'Pass': Alice does nothing.
  - Otherwise, if Alice chooses 'Swap with opponent': Alice swaps one random card from their hand with
one random card from Bob's hand. In a two-player game both participants learn each other's exchanged
card, so after the swap the identities of the two exchanged cards are known to every player. If either
hand is empty, nothing happens.
- Bob picks one of: Swap with opponent or Pass. (the choice itself is not announced, but some branches
produce public side-effects - chip movements, card exchanges, or other observable state changes - that
reveal which branch ran; other branches leave no public trace)
  - If Bob chooses 'Pass': Bob does nothing.
  - Otherwise, if Bob chooses 'Swap with opponent': Bob swaps one random card from their hand with one
random card from Alice's hand. In a two-player game both participants learn each other's exchanged card,
so after the swap the identities of the two exchanged cards are known to every player. If either hand is
empty, nothing happens.
*(Play returns to **Post-Betting Phase**.)*

### Showdown
- Both players' hands are revealed to all players. Showdown: each player's evaluated hand is scored. The
base score is the sum of the rank values of all cards in the evaluated hand. (An empty hand scores 0.)
GOOD PAIRS: A pair of Js adds 3 points. POISON CARDS: Ranks K subtract 3 points each. Higher adjusted
score wins. (Rank values: 10 is face value, J=11, Q=12, K=13.)
- The showdown winner takes the entire pot. On a tie, each player receives half the pot (rounded down);
any remaining odd chip is awarded uniformly at random to one tied player.
- The game ends: Showdown complete.
```

Per-turn observation prompt. Before each turn, an agent receives a message from the engine. The game state is presented first and covers the current phase, the agent's own hand, public cards on the board, each player's chip stack, the current pot, role or position assignments, and who the opponents are. The engine also sends public events and private observations that, by the rules, should be visible to the agent. The final part of the message specifies the decision to be made. The agent must reply with a JSON-formatted response of the form `{"action": <chosen_option>}`. If a response cannot be successfully parsed, the engine plays a uniformly random legal action.

C COMPLEXITY AXES (FORMAL)

We compute five axis scores for each of the 50 benchmark games without any LLM play. Four of them are *value axes*, which are expectations over the decisions a player faces when simple policies play the game. The fifth counts the information states a player can reach.

Objects. Each axis is computed for one seat i (Alice or Bob) at a time and then averaged over both. A *deal* is one play of the game from the shuffle to settlement, and the *chip lead* of a seat is its change in chips minus that of its opponent. In seed 4643 both seats ante 3 of their 10 chips. If Alice bets at her first decision and Bob folds, she ends 3 chips up and Bob 3 chips down, and her chip lead is 6. Three axes are divided by STD_g, the standard deviation of the chip lead in a deal between two members of the population defined below, averaged over random pairs of members. At a decision, seat i observes its own cards, the public cards, the chip stacks, the pot, the phase, and the history of announced actions. We call this its *information state* I. A *policy* maps each I to a distribution over the set $\mathcal{A}(I)$ of legal actions at I, and $A = |\mathcal{A}(I)|$. When a player chooses an amount from more than six values, $\mathcal{A}(I)$ keeps at most six of them: the four smallest, the midpoint and the largest.

Population. We drew 200 random behavior strategies. Strategy j has one mixture over the legal actions for each information state, drawn once from a Dirichlet(1) distribution. We ranked the 200 strategies by mean chip lead against L_0, the uniform random policy, over 40 deals, and the better half (100 strategies) became *tier 1*. Seat i faces an opponent drawn from P_Θ, with weight $1/3$ on L_0

and $2/3$ on tier 1. For its own later decisions we draw a policy from P_K, with $1/4$ on L_0 and $3/4$ on tier 1.

Decision distribution. We let both seats play policies drawn from the population and record every decision of seat i with two or more legal actions. μ_i is the distribution of the information states at these decisions, with each visit counted once. A state that play reaches twice as often has twice the weight, and forced moves have none.

Action values. The temporal and opponent axes come from one decomposition of action values. How the deal ends after an action a at an information state I of seat i depends on a *context* $c = (k, \theta, w)$. Here the continuation $k \sim P_K$ makes the later decisions of seat i, and $\theta \sim P_\Theta$ is the opponent in the other seat. The world w assigns every card seat i has not seen (the opponent's hand and the undealt deck), uniformly over the assignments consistent with I. We write $Q(I, a, c)$ for the expected chip lead of seat i at the end of the deal when it takes a at I and play continues under c. Its means are $Q(I, a) = \mathbb{E}_c\, Q(I, a, c)$, $\bar{Q}(I, c) = \frac{1}{A}\sum_a Q(I, a, c)$, and the grand mean $\bar{Q}(I)$. Over a uniformly chosen action and a random context, the variance of $Q(I, a, c)$ splits exactly into three terms,

$$V_{\text{total}}(I) \;=\; \underbrace{\tfrac{1}{A}\textstyle\sum_a \big(Q(I,a) - \bar{Q}(I)\big)^2}_{V_a(I)} \;+\; \underbrace{\mathbb{E}_c\big(\bar{Q}(I,c) - \bar{Q}(I)\big)^2}_{V_c(I)}$$
$$+\; \underbrace{\mathbb{E}_c\, \tfrac{1}{A}\textstyle\sum_a \big(Q(I,a,c) - Q(I,a) - \bar{Q}(I,c) + \bar{Q}(I)\big)^2}_{V_{a\times c}(I)}.$$

V_c is mostly the luck of the cards. An action that is simply better on average raises $Q(I, a)$, which enters V_a. Note that with two actions and gap $D(c) = Q(I, 1, c) - Q(I, 2, c)$, the interaction is one quarter of the variance of the gap across futures. For the opponent, with $Q(I, a, \theta) = \mathbb{E}_{k,w}\, Q(I, a, c)$,

$$V_{a\times\theta}(I) \;=\; \mathbb{E}_\theta\, \tfrac{1}{A}\textstyle\sum_a \big(Q(I,a,\theta) - Q(I,a) - \bar{Q}(I,\theta) + \bar{Q}(I)\big)^2,$$

and $V_{a\times k}$ and $V_{a\times w}$ are defined in the same way, each after Q is averaged over the other two parts of the context.

State space. $\log_{10}\big(\sum_p |\mathcal{I}_p|\big)$, where $\mathcal{I}_p$ is the set of information states that player p reaches in 3,000 deals in which both players move uniformly at random over all legal actions, every legal amount included. A state is identified by the decision, the player's own hand, its last three decisions, its chips (rounded), everything it has been shown (the events, the messages, the board and the pot) and the positions. The count is a lower bound on the number of reachable states.

Temporal contingency ("temporal").

$$\mathrm{Temp}(g) \;=\; \tfrac{1}{2}\textstyle\sum_i\; \mathbb{E}_{I\sim\mu_i}\left[\dfrac{V_{a\times c}(I)}{V_{\text{total}}(I)}\right]$$

averages, over the decisions of each seat, the fraction of $V_{\text{total}}(I)$ that is driven by interaction between actions and context. The fraction is 0 when every difference between action values is the same in every context, and 1 when all of $V_{\text{total}}(I)$ lies in the interaction. The world w in the context includes the opponent's current hand. As a result, a high score can come from the unseen cards as well as from later play under k and θ. We call this temporal contingency, although it does not separate planning from these other sources. For example, a game of one move with actions ± 1, a fair coin $w = \pm 1$ that the player cannot see, and payoff aw has $V_a = V_c = 0$ and scores one.

Opponent dependence.

$$\mathrm{Opp}(g) \;=\; \tfrac{1}{2}\textstyle\sum_i\; \mathbb{E}_{I\sim\mu_i}\Big[\sqrt{V_{a\times\theta}(I)}\Big]\Big/\mathrm{STD}_g.$$

Here $\sqrt{V_{a\times\theta}(I)}$ is the typical size of the action×opponent interaction, and it is zero when the gap between two actions is the same against every opponent. We estimate $V_{a\times\theta}(I)$ as the mean product of the interaction residuals from the two halves of the sampled worlds, clipped at zero. We use its mean in units of STD_g as the measure of how much the comparison between actions depends on

the opponent. Note that it does not measure the value of adapting to a known opponent. We also computed that value, as the gain from the action best against θ over the action best against the whole population, averaged over θ, and it lies within $\pm 0.05\,\mathrm{STD}_g$ on every game.

A version of the axis defined on whole strategies gives the same result. For each game we played L_0 and 40 tier-1 members of each seat against each other on 60 common deals. Opponent dependence is then measured as the share of outcome variance in the interaction between the strategies of the two seats, cross-fitted over two halves of the deals. Its median over games is 0.06%, and it correlates with the axis at $\rho = 0.74$. With this measure in place of the axis, the capability profile of Section 7 has eight significant cells, five of them on temporal and none on opponent dependence.

Precision.

$$\mathrm{Prec}(g) \;=\; \tfrac{1}{2}\textstyle\sum_i \;\mathbb{E}_{I\sim\mu_i}\Big[\,Q(I,a^*) - \frac{1}{|N|}\sum_{a'\in N} Q(I,a')\Big]\Big/\mathrm{STD}_g,$$

where $a^*(I) = \arg\max_a Q(I,a)$ is the best action and $N(I)$ is the set of its neighbors. On ordered menus, such as wagers and bids, the neighbors are the adjacent amounts. On categorical menus they are all other actions. The value of a mixed strategy is linear in its probabilities, so the gap is the loss per unit of probability moved from a^* to its neighbors. On a menu of two actions it is the margin of the runner-up.

Information. $\mathrm{Lead}_i(\pi,\theta)$ is the expected chip lead of seat i at the end of a deal when it plays policy π against opponent θ. $\mathrm{BR}(\theta)$ is a limited-depth search policy. At each decision it samples worlds consistent with its observations, values every legal action by replays in which the opponent plays θ, and takes the action with the best mean. Inside a replay, each of its later decisions is chosen by a one-level search that values every legal action by two replays in which it then plays uniformly, while the opponent keeps playing θ. For $\mathrm{BR}_{\mathrm{pub}}(\theta)$ the sampled worlds also redraw its own hand, and everything else it has observed stays as it was. The axis is

$$\mathrm{Info}(g) \;=\; \tfrac{1}{2}\textstyle\sum_i \;\mathbb{E}_\theta\Big[\mathrm{Lead}_i\big(\mathrm{BR}(\theta),\theta\big) - \mathrm{Lead}_i\big(\mathrm{BR}_{\mathrm{pub}}(\theta),\theta\big)\Big]\Big/\mathrm{STD}_g,$$

with θ drawn with equal weight from L_0 and four tier-1 members. It is the loss per deal, in units of STD_g, of a search policy whose sampled worlds do not keep its own cards.

Estimation. We estimate every expectation by simulating play. Up to 90 decisions per seat are taken from population self-play. At each decision, all legal actions are replayed under the same 360 sampled contexts with common random numbers (1,080 replays for three legal actions). If an axis needs a best action, we choose it on one half of the worlds and value it on the other half, which mitigates the effect of selection bias (an action chosen because it did well on some draws should not be evaluated on those same draws). The information axis uses 40 deals against each opponent, and a three-hour cap on computation left fewer opponents for some seats and games. The 360 contexts are all combinations of 4 continuations, 9 opponents, and 10 worlds, and each is replayed once for every action. Table 3 reports, for the temporal, opponent and precision axes, the sampling error, the seat agreement, and the correlation with a rerun at about two-thirds of the budget on an independent population draw. The variance inflation factors of all five axes are in Appendix D.

Table 3: Diagnostics of the temporal, opponent and precision axes on the 50 benchmark games.

axis	SE / between-game SD	seat agreement (ρ)	reproduces across draws
temporal	0.21	$+0.69$	$\rho = 0.92$
opponent dependence	0.27	$+0.66$	0.88
precision	0.31	$+0.44$	0.86

Selection basis. Before computing the five axes, we selected the 50 games by farthest-point sampling on six preliminary axes, which a cheaper screening run computed for all 2,000 games in the pool. These axes are state space, temporal depth, information sensitivity, opponent modeling, risk, and brittleness. They are different quantities from our five axes even where the names are similar (the state space of the screening run did not count decisions in which a player chooses an amount, and it

did not tell apart the cards a player had seen), and thus were only used for selecting games. The two sets of axes do agree on which games are hard. Their composite complexity scores (Appendix E) correlate at $\rho = 0.79$, and seven of the eight hardest games are the same.

D BENCHMARK CHARACTERISTICS

Figure 4 shows the 2,000-game pool and the 50 selected games in two axes of the selection basis. Figure 5 shows the distribution of the 50 games on each of the five axes. The axes are moderately correlated (Figure 6): the largest pairwise Spearman correlation in absolute value is state space $\times$ opponent dependence (-0.50), and state space also correlates with temporal ($+0.44$) and precision (-0.45). Opponent dependence $\times$ precision is 0.48. The two chip-valued per-decision axes are smaller on long games, where a single decision carries a smaller share of the stakes. The largest variance inflation factor, $1/(1 - R^2)$ from regressing an axis on the other four, is 2.0 (opponent dependence):

Axis	**VIF (50-game benchmark)**
State space ($\log_{10}$)	1.70
Opponent dependence	2.00
Temporal	1.48
Precision	1.66
Information	1.30

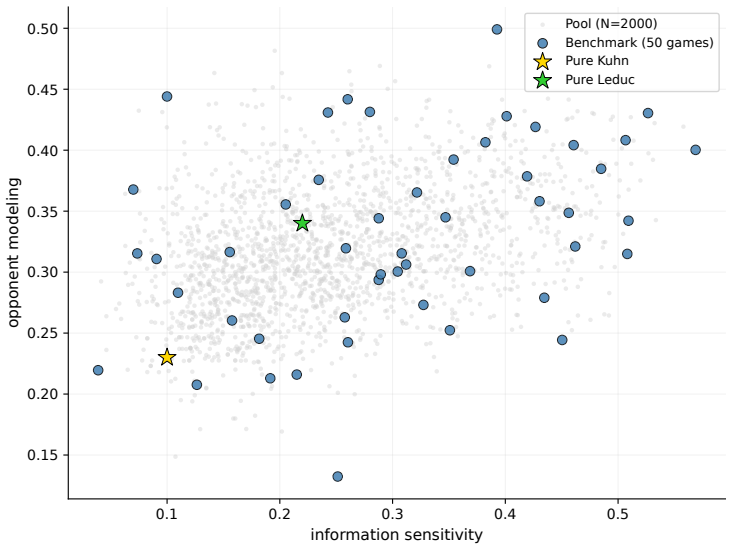

Figure 4: The 2,000-game accepted pool vs. the 50 benchmark games selected by farthest-point sampling (FPS), in two axes of the selection basis (information sensitivity $\times$ opponent modeling, screening-run values on the pool). Gold and green stars mark Kuhn and Leduc poker.

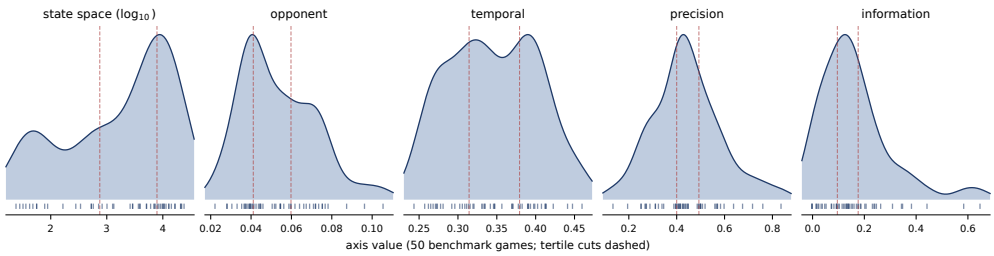

Figure 5: Per-axis distributions of the 50 benchmark games: kernel-density estimates with tick-marked observations and tertile cuts (dashed).

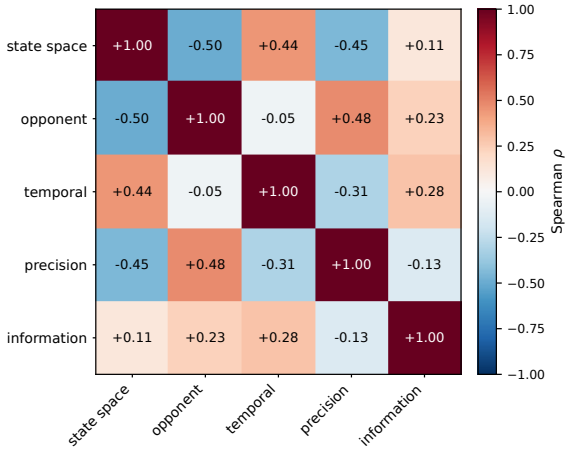

Figure 6: **Pairwise Spearman correlation of the five complexity axes across the 50 benchmark games.** The largest $|r|$ is 0.50 (state space $\times$ opponent dependence), under the conventional $|r| \approx 0.7$ threshold above which joint-regression standard errors inflate substantially. The largest per-axis VIF, listed earlier in this appendix, is 2.0.

E LEADERBOARDS

All leaderboards in the paper result from fitting a linear model on signed chip margins, $y_s = \alpha_i - \alpha_j + \varepsilon_s$ with $\sum_m \alpha_m = 0$. We are thus able to estimate a strength parameter for each model, $\hat{\alpha}_m$, which we use to rank models. Confidence intervals are estimated using a cluster bootstrap in which the two seats of a run form a cluster.

Composite complexity. We construct a composite complexity score for each game, which is a linear combination of its five axis scores. Weights are given by the first principal component of the matrix where element i, j is the slope of model i on axis j. Each slope comes from a regression of $\hat{\alpha}_{m,g}$, the strength of model m estimated from the deals of game g alone, on the z-scored axis. This principal component is the linear combination of axis scores along which the sensitivities of the models differ the most, and it explains 93% of the variance of the matrix after each column is centered over models. We rescale it to unit absolute sum, with the sign chosen to make the temporal weight positive. As a result, the composite score is a learned discrimination score. The weights are $+0.35$ for temporal, $+0.29$ for state space, $+0.09$ for information, -0.19 for precision, and -0.07 for opponent dependence. Refitting on the eight games with the highest composite score widens the leaderboard spread from 3.3 chips/game on all fifty to 5.6 (Table 4).

Learning the weights on a random half of the games and scoring the other half gives 5.2 ± 0.2 chips/game on the eight hardest games (four from each half) over twenty splits. The temporal axis alone uses no outcome information. By its tertiles, the leaderboard spans 1.83, 3.00, and 4.96

chips/game from easiest to hardest. An unweighted mean of the five z-scored axes gives 1.91, 3.69, and 4.36, and its eight hardest games widen the spread only to 3.85, as precision and opponent dependence are often higher on smaller games. Using the multivariate coefficients of Table 13 in place of the slopes gives a score that correlates with ours at $\rho = 0.95$ and widens the spread 1.80 times on its own top eight. Table 6 refits $\hat{\alpha}$ within tertiles of the composite score.

Table 4: **Leaderboard on the eight hardest games** by composite complexity (seeds 9754, 1406, 11435, 2914, 203, 10760, 10427, 4643), sum-to-zero $\hat{\alpha}$ on 3,411 paired deals, paired-cluster bootstrap 95% CIs (B=400). † September models.

Model	$\hat{\alpha}$ (hard 8)	**95% CI**	**deals**
`gpt-6-astra`†	$+1.52$	$[+0.95, +2.09]$	270
`gemini-3.1-pro`	$+1.28$	$[+0.96, +1.59]$	847
`gpt-5.4`	$+1.22$	$[+0.88, +1.57]$	457
`claude-fable-5.1`†	$+1.10$	$[+0.58, +1.64]$	260
`gemini-3.8-flash`†	$+0.74$	$[-0.23, +1.55]$	80
`claude-sonnet-4.6`	$+0.72$	$[+0.28, +1.20]$	346
`gemini-2.5-pro`	$+0.11$	$[-0.23, +0.45]$	720
`deepseek-v3.1`	-0.10	$[-0.41, +0.24]$	768
`gemma-4-31b`	-0.26	$[-0.61, +0.07]$	738
`gemini-3.1-flash-lite`	-1.07	$[-1.41, -0.74]$	788
`qwen-3.5`	-1.22	$[-1.62, -0.83]$	690
`llama-3.3-70b`	-4.05	$[-4.46, -3.61]$	858

Table 5: **Leaderboard on the eight most temporal games**: sum-to-zero $\hat{\alpha}$ on 3,358 paired deals, paired-cluster bootstrap 95% CIs (B=400). † September models.

Model	$\hat{\alpha}$ (temporal top 8)	**95% CI**	**deals**
`gpt-6-astra`†	$+1.73$	$[+1.19, +2.33]$	198
`gpt-5.4`	$+1.39$	$[+1.08, +1.74]$	406
`claude-fable-5.1`†	$+1.22$	$[+0.64, +1.80]$	188
`gemini-3.1-pro`	$+1.00$	$[+0.73, +1.25]$	857
`gemini-3.8-flash`†	$+0.80$	$[-0.19, +1.67]$	80
`claude-sonnet-4.6`	$+0.54$	$[+0.18, +0.93]$	398
`gemini-2.5-pro`	$+0.14$	$[-0.16, +0.47]$	720
`gemma-4-31b`	-0.26	$[-0.58, +0.06]$	750
`deepseek-v3.1`	-0.46	$[-0.76, -0.14]$	774
`gemini-3.1-flash-lite`	-1.15	$[-1.45, -0.87]$	799
`qwen-3.5`	-1.43	$[-1.74, -1.10]$	718
`llama-3.3-70b`	-3.52	$[-3.94, -3.11]$	828

Table 6: **Per-tertile additive paired-comparison leaderboards.**

Per-tertile leaderboards — $\hat{\alpha}$ by composite-complexity tertile

Model	$\hat{\alpha}_{T1}$	$\hat{\alpha}_{T2}$	$\hat{\alpha}_{T3}$
gpt-6-astra†	−0.03	+0.77	+1.29
gpt-5.4	+0.21	+0.63	+1.05
gemini-3.8-flash†	+0.25	+0.52	+1.04
gemini-3.1-pro	+0.29	+0.59	+0.96
claude-sonnet-4.6	+0.23	+0.51	+0.60
claude-fable-5.1†	+0.28	+0.28	+0.86
gemini-2.5-pro	+0.21	+0.14	+0.13
gemma-4-31b	+0.16	+0.09	−0.12
deepseek-v3.1	−0.06	−0.05	−0.22
gemini-3.1-flash-lite	−0.08	−0.43	−0.90
qwen-3.5	−0.15	−0.51	−1.01
llama-3.3-70b	−1.32	−2.53	−3.68

T1 = easiest 16 games, T2 = middle 17, T3 = hardest 17, by the five-axis composite complexity. Paired-cluster bootstrap 95% CI widths (B=400) reach at most 1.04 chips/game across all 36 cells (widest on gemini-3.8-flash, T3), with a median width of 0.29. Top-3 by tertile: T1 gemini-3.1-pro, claude-fable-5.1, gemini-3.8-flash; T2 gpt-6-astra, gpt-5.4, gemini-3.1-pro; T3 gpt-6-astra, gpt-5.4, gemini-3.8-flash. Bottom-3 by tertile: T1 llama-3.3-70b, qwen-3.5, gemini-3.1-flash-lite; T2 llama-3.3-70b, qwen-3.5, gemini-3.1-flash-lite; T3 llama-3.3-70b, qwen-3.5, gemini-3.1-flash-lite.

E.1 Robustness

Dropping each game in turn and refitting, we get a mean Kendall τ of 0.956 with the full ranking (minimum 0.879). None of these refits breaks the three tiers, and where the order does change, models move only within their tier, by at most three places (gemini-3.1-pro when seed 1406 is dropped). We also refit the leaderboard on random subsets of games, drawing 1,000 subsets at each of four sizes. The median Spearman correlation with the full ranking is 0.89 for 5 games, 0.92 for 8, 0.94 for 12 and 0.97 for 25, with 5th percentiles of 0.68, 0.76, 0.78 and 0.87. Refitting the April round-robin alone gives the nine April models in exactly their twelve-model order (Table 7, Kendall $\tau = 1.00$). With the bottom outlier llama-3.3-70b (−2.56 in the full fit) dropped, the refit on the other eleven (15,228 paired deals) keeps their order apart from claude-sonnet-4.6 and claude-fable-5.1, tied at +0.45 in the full fit, which swap (Table 8). The leaderboard is fit on margins since the models were told to maximize expected chips. A Bradley–Terry fit on the win indicator (ties dropped) orders the models at Spearman 0.77 with the margin fit, favoring frequent small-margin winners. Of the 47 model pairs that met, 41 have a mean head-to-head margin in the direction of the leaderboard, and none of the six exceptions is significant (Tables 9 and 10).

Table 7: **Leaderboard refit on the April round-robin alone** (nine models, paired-cluster bootstrap 95% CIs, B=2,000).

Model	$\hat{\alpha}$ (April nine)	95% CI
gpt-5.4	$+0.86$	$[+0.76, +0.97]$
gemini-3.1-pro	$+0.83$	$[+0.75, +0.90]$
claude-sonnet-4.6	$+0.64$	$[+0.53, +0.76]$
gemini-2.5-pro	$+0.36$	$[+0.29, +0.43]$
gemma-4-31b	$+0.25$	$[+0.17, +0.32]$
deepseek-v3.1	$+0.06$	$[-0.01, +0.13]$
gemini-3.1-flash-lite	-0.27	$[-0.35, -0.20]$
qwen-3.5	-0.35	$[-0.44, -0.27]$
llama-3.3-70b	-2.37	$[-2.47, -2.27]$

Table 8: **Llama-excluded leaderboard.**

Llama-excluded refit — $\hat{\alpha}$ (chips/game)

Model	$\hat{\alpha}$ (no llama)	95% CI
gpt-6-astra†	$+0.50$	$[+0.30, +0.70]$
gpt-5.4	$+0.41$	$[+0.32, +0.50]$
gemini-3.8-flash†	$+0.40$	$[+0.20, +0.61]$
gemini-3.1-pro	$+0.35$	$[+0.28, +0.43]$
claude-fable-5.1†	$+0.20$	$[+0.00, +0.39]$
claude-sonnet-4.6	$+0.18$	$[+0.07, +0.29]$
gemini-2.5-pro	-0.06	$[-0.14, +0.01]$
gemma-4-31b	-0.17	$[-0.25, -0.09]$
deepseek-v3.1	-0.35	$[-0.44, -0.27]$
gemini-3.1-flash-lite	-0.66	$[-0.74, -0.58]$
qwen-3.5	-0.79	$[-0.88, -0.70]$

Paired-cluster bootstrap 95% CI (B=1,000), sum-to-zero contrast. Refit on the 11 non-llama models (15,228 paired deals).

Table 9: Pairwise mean chip margin (row − column) over the benchmark games on which the pair played, both seats balanced. Bold = 95% paired-cluster bootstrap CI excludes 0 (B=500). Rows and columns sorted by leaderboard rank. Each of the three September models played a subset of the field, and "–" marks pairs that never met.

	gpt-6-astra†	gpt-5.4	gemini-3.8-flash†	gemini-3.1-pro	claude-sonnet-4.6	claude-fable-5.1†	gemini-2.5-pro	gemma-4-31b	deepseek-v3.1	gemini-3.1-flash-lite	qwen-3.5	llama-3.3-70b
gpt-6-astra†	–	+0.02	–	−0.05	–	**+0.84**	–	–	–	–	–	**+3.20**
gpt-5.4	−0.02	–	−0.04	+0.20	**+0.45**	−0.44	**+0.34**	**+0.60**	**+1.02**	**+0.91**	**+0.95**	**+3.22**
gemini-3.8-flash†	–	+0.04	–	−0.04	–	–	–	–	–	**+1.11**	–	–
gemini-3.1-pro	+0.05	−0.20	+0.04	–	+0.17	−0.06	**+0.42**	**+0.58**	**+0.71**	**+0.83**	**+1.29**	**+3.54**
claude-sonnet-4.6	–	**−0.45**	–	−0.17	–	+0.08	−0.01	+0.39	**+0.65**	**+1.25**	**+0.81**	**+3.26**
claude-fable-5.1†	**−0.84**	+0.44	–	+0.06	−0.08	–	–	**+0.25**	–	–	–	–
gemini-2.5-pro	–	**−0.34**	–	**−0.42**	+0.01	–	–	+0.07	+0.17	**+0.76**	**+0.60**	**+2.71**
gemma-4-31b	–	**−0.60**	–	**−0.58**	−0.39	**−0.25**	−0.07	–	+0.13	**+0.45**	**+0.74**	**+2.54**
deepseek-v3.1	–	**−1.02**	–	**−0.71**	**−0.65**	–	−0.17	−0.13	–	**+0.48**	**+0.22**	**+2.31**
gemini-3.1-flash-lite	–	**−0.91**	**−1.11**	**−0.83**	**−1.25**	–	**−0.76**	**−0.45**	**−0.48**	–	+0.33	**+1.79**
qwen-3.5	–	**−0.95**	–	**−1.29**	**−0.81**	–	**−0.60**	**−0.74**	**−0.22**	**−0.33**	–	**+2.08**
llama-3.3-70b	**−3.20**	**−3.22**	–	**−3.54**	**−3.26**	–	**−2.71**	**−2.54**	**−2.31**	**−1.79**	**−2.08**	–

Table 10: Paired deals per model pair in the unified corpus (20,206 in all). Rows and columns in leaderboard order, and "–" marks pairs that never met.

	gpt-6-astra†	gpt-5.4	gemini-3.8-flash†	gemini-3.1-pro	claude-sonnet-4.6	claude-fable-5.1†	gemini-2.5-pro	gemma-4-31b	deepseek-v3.1	gemini-3.1-flash-lite	qwen-3.5	llama-3.3-70b
gpt-6-astra†	–	260	–	170	–	100	–	–	–	–	–	160
gpt-5.4	260	–	160	220	240	80	200	260	216	220	210	260
gemini-3.8-flash†	–	160	–	170	–	–	–	–	–	170	–	–
gemini-3.1-pro	170	220	170	–	240	170	700	720	716	720	699	720
claude-sonnet-4.6	–	240	–	240	–	170	220	199	217	239	234	280
claude-fable-5.1†	100	80	–	170	170	–	–	160	–	–	–	–
gemini-2.5-pro	–	200	–	700	220	–	–	720	700	720	720	720
gemma-4-31b	–	260	–	720	199	160	720	–	719	720	719	700
deepseek-v3.1	–	216	–	716	217	–	700	719	–	698	714	719
gemini-3.1-flash-lite	–	220	170	720	239	–	720	720	698	–	718	720
qwen-3.5	–	210	–	699	234	–	720	719	714	718	–	699
llama-3.3-70b	160	260	–	720	280	–	720	700	719	720	699	–

F PER-GAME RESULTS

Table 11 gives each model's raw mean margin on each game and Table 12 the per-game strengths $\hat{\alpha}_{m,g}$, the leaderboard fit restricted to one game's deals, which corrects for the opponents a model happened to face there. Rows of both are sorted by the mean of the five z-scored axes.

Variance decomposition. The noise-corrected model×game interaction σ^2_{MG} of the 12×50 grid of $\hat{\alpha}_{m,g}$ is about a third of the model main effect σ^2_M on the full data (ratio 0.34, 95% CI $[0.20, 0.54]$) and comparable to it without `llama-3.3-70b` (0.65, $[0.14, 1.31]$). Here the main effect is the variance of the per-model means and the interaction is the variance of the residuals after removing model and game means. Each cell is estimated from a finite number of deals, and estimation noise accounts for roughly half of the raw interaction. We subtract from each component the noise variance estimated for it by the paired-cluster bootstrap.

Quantity	observed	noise	corrected	95% CI
Full 12 models				
σ^2_M	0.77	0.00	0.76	$[0.70, 0.82]$
σ^2_{MG}	0.48	0.22	0.26	$[0.15, 0.41]$
σ^2_{MG}/σ^2_M	0.62	—	0.34	$[0.20, 0.54]$
Llama-excluded				
σ^2_M	0.20	0.00	0.19	$[0.16, 0.24]$
σ^2_{MG}	0.31	0.19	0.13	$[0.03, 0.27]$
σ^2_{MG}/σ^2_M	1.58	—	0.65	$[0.14, 1.31]$

Table 11: Per-(model, game) pair-mean chip margin on the 50-game benchmark, computed as an exact mean over the rows played and sorted ascending by `cmplx`, the mean of the five z-scored axes. No cell is missing (a missing cell would print as "–"). Entries of $+0.00$ are mean margins of exactly zero. In 59 of these 68 cells the two deals of every run cancelled, and in the other 9 the runs net to zero.

seed	cmplx	gpt-6-astra†	gpt-5.4	gemini-3.8-flash†	gemini-3.1-pro	claude-sonnet-4.6	claude-fable-5.1†	gemini-2.5-pro	gemma-4-31b	deepseek-v3.1	gemini-3.1-flash-lite	qwen-3.5	llama-3.3-70b
2385	-0.82	+0.30	+0.16	+0.30	+0.33	+0.24	+0.00	+0.25	-0.13	+0.04	+0.23	-0.11	-0.81
6903	-0.80	+0.10	+0.17	-0.20	+0.85	-0.35	-0.50	+0.30	+0.36	-0.07	-0.14	-0.37	-1.20
10173	-0.80	+1.50	+0.18	+0.00	+0.73	-0.30	+0.30	+0.47	-0.03	+0.07	-0.12	+0.56	-1.94
7052	-0.70	+1.00	-0.25	+0.20	+1.09	+0.70	-0.40	+0.61	+0.89	-0.19	-0.75	-0.17	-2.11
2495	-0.60	-0.40	+0.28	+0.10	+0.08	-0.07	+0.20	+0.38	-0.08	+0.06	-0.17	-0.42	+0.06
11784	-0.55	+0.00	+0.55	+0.20	+0.72	+0.68	+0.00	+0.30	+0.29	+0.40	-0.20	-0.19	-1.93
5149	-0.54	+0.10	+0.00	+0.00	+0.12	+0.25	-0.10	+0.11	+0.46	+0.31	+0.01	-0.29	-0.95
9548	-0.50	+0.00	+0.46	+0.80	+1.24	+0.40	+0.00	+0.27	+0.25	+0.74	-0.33	-0.63	-2.17
9398	-0.44	-0.20	+1.01	-0.20	+1.59	+2.86	-0.90	+0.81	+0.09	+0.55	-0.66	-1.23	-2.81
963	-0.38	+0.00	+0.00	+0.30	+0.45	+0.00	+0.60	+0.49	+0.01	+0.21	+0.05	+0.03	-1.64
8489	-0.37	+0.50	+0.96	+0.10	+0.76	-0.15	-0.50	+0.78	+0.38	-0.58	+0.57	+0.34	-3.00
4968	-0.36	+0.90	+2.30	+1.30	+0.94	+0.40	+0.10	+0.88	+0.66	-0.42	-0.16	-0.31	-1.86
2914	-0.32	+0.50	+1.67	+1.00	+2.31	-0.37	+0.70	-0.34	-0.67	+0.44	-1.41	-1.25	+0.19
2345	-0.29	+0.00	+0.53	+0.00	+0.48	+0.45	+0.00	+0.21	+0.35	-0.14	+0.12	-0.30	-1.19
4426	-0.27	+0.20	+0.15	+0.00	+0.11	+0.04	+0.00	+0.14	+0.04	+0.10	+0.00	-0.05	-0.32
5470	-0.24	+0.50	+0.28	-0.20	+0.22	+0.10	-0.20	+0.23	-0.06	+0.03	-0.11	-0.13	-0.50
10738	-0.21	+1.00	-0.18	-0.10	+0.83	-0.10	-0.20	+0.52	+0.16	-0.35	+0.20	-0.51	-1.04
862	-0.20	+0.00	+0.00	+0.00	+0.54	+0.00	+0.00	+0.47	+0.00	+0.28	+0.43	-0.01	-2.05
2601	-0.19	+2.70	+0.08	+0.30	+0.69	-0.15	-0.30	+0.62	+0.28	-0.19	+0.59	+0.39	-2.39
10137	-0.15	-0.20	+0.65	+0.60	+1.61	+0.53	+0.00	+0.22	+1.19	+0.66	+0.09	-0.50	-3.83
6586	-0.14	+2.50	+0.30	+0.00	+0.21	+0.80	+0.40	+0.57	-0.07	+0.18	-0.06	+0.20	-1.50
1247	-0.13	+4.80	+0.10	+0.00	+1.38	+0.01	+0.00	+0.35	+0.65	+0.03	-0.54	+0.44	-2.70
10786	-0.08	-0.10	+0.62	+0.00	+0.88	+0.20	-0.50	+0.75	+0.10	+0.33	+0.12	-0.23	-2.71
11517	-0.06	+0.00	+0.00	+0.00	+0.50	+0.88	+0.00	+0.62	+0.00	+0.51	+0.00	+0.67	-2.66
10790	-0.04	-1.00	+0.91	+0.00	+0.48	+0.00	+0.30	+0.84	+0.55	-0.31	-0.11	-0.01	-2.83
3483	+0.05	+0.00	+0.65	-0.30	+1.16	+1.00	+0.10	+0.55	+1.36	+0.34	-0.31	-1.10	-2.61
7142	+0.06	+0.00	+0.50	+0.40	+0.37	-0.05	+0.00	+0.15	+0.54	-0.24	-0.14	+0.05	-1.00
8297	+0.06	+0.00	+0.00	+1.50	+0.64	+0.62	+0.30	-0.19	+1.33	+0.02	-0.35	-0.06	-2.31
10427	+0.08	+0.74	+0.29	+0.50	+1.24	+3.79	-0.22	+0.27	+0.95	-0.32	-0.16	+0.56	-4.04
9754	+0.12	-0.43	-0.27	-0.50	+2.52	+2.15	+0.63	+0.21	+1.27	-0.56	-1.04	-0.12	-3.95
11435	+0.16	+1.54	-0.01	+1.10	+2.06	+6.05	-0.09	+1.05	+0.31	+1.90	-0.52	+0.60	-6.14
10760	+0.16	+2.15	+0.97	-1.90	+1.78	+0.25	-0.52	+2.47	-0.33	+0.66	-1.58	+0.25	-5.13
6520	+0.16	+0.30	+0.24	-0.60	+0.73	+1.36	+0.70	+0.30	+0.87	+0.02	-0.53	-0.03	-2.50
6635	+0.16	+0.40	+1.89	-1.00	+1.23	+1.77	-0.40	+0.96	+0.15	+0.77	+0.01	-0.78	-3.06
7830	+0.18	+4.00	+0.30	+0.00	+0.90	-0.50	-0.90	+0.83	+0.79	+0.05	-0.11	-0.89	-2.54
7379	+0.21	-0.20	+1.23	+0.50	+1.42	+0.05	+1.00	+0.96	+0.44	+0.87	-0.64	-0.12	-3.48
5368	+0.26	+0.50	+0.40	+0.50	+0.45	+0.10	+0.00	+0.38	+0.45	+1.08	+0.00	+0.04	-3.15
4643	+0.27	+0.30	+2.15	+0.30	+2.14	+1.43	-1.40	+1.20	-0.10	-1.07	+1.38	-1.29	-3.22
6644	+0.30	+0.00	+1.50	+4.30	+2.02	+1.20	+1.20	+1.64	+0.20	+0.33	-0.98	-0.97	-3.58
933	+0.32	+0.00	+2.37	+2.20	+1.11	+0.70	-0.70	+0.74	-0.03	+0.55	-0.41	-0.61	-2.48
1406	+0.39	+2.35	+1.07	+3.00	+1.66	+0.64	-0.04	+1.83	+0.82	+0.52	-0.68	-0.47	-5.31
3651	+0.42	+3.30	+0.65	+1.00	+1.71	+0.70	-0.20	+0.89	-0.06	+0.60	+0.66	-0.66	-3.79
11520	+0.46	+0.00	+0.04	+0.00	+0.14	+0.00	+0.00	+0.00	+0.15	+0.20	+0.20	-0.02	-0.78
12140	+0.52	+0.00	+2.74	+0.00	+2.02	+2.62	+0.60	+0.99	-0.27	+0.55	-0.38	-0.61	-4.28
4059	+0.56	+4.50	+0.25	-0.70	+0.68	+1.10	-0.20	+0.69	+0.26	-0.01	+0.25	-0.17	-2.06
203	+0.58	+0.00	+2.60	-0.70	+2.13	+0.45	-0.10	+1.09	+1.93	+1.24	-0.80	-1.38	-4.66
1608	+0.64	+0.00	+1.00	+0.00	+0.14	-0.15	+0.00	+0.48	+0.48	+0.12	+0.17	+0.24	-1.99
1949	+0.92	+0.00	+1.28	+0.00	+1.01	+0.00	+0.00	+0.00	+0.56	+0.47	+0.82	-0.02	-3.39
2526	+1.01	+0.00	+2.98	+4.60	+2.84	+1.36	+1.00	+1.59	+2.79	+0.06	-1.39	-2.16	-4.76
2522	+1.13	+1.90	+0.00	+0.00	+0.00	+3.00	+0.00	+0.00	+0.59	+0.53	+0.61	+0.56	-2.54

Table 12: Per-game strength estimates $\hat{\alpha}_{m,g}$ (chips/game), per model. Rows are the 50 benchmark games sorted ascending by the mean of the five z-scored axes, and columns are the models in overall leaderboard order.

seed	cmplx	gpt-6-astra†	gpt-5.4	gemini-3.8-flash†	gemini-3.1-pro	claude-sonnet-4.6	claude-fable-5.1†	gemini-2.5-pro	gemma-4-31b	deepseek-v3.1	gemini-3.1-flash-lite	qwen-3.5	llama-3.3-70b
2385	−0.82	+0.50	−0.01	+0.29	+0.20	+0.01	+0.01	+0.06	−0.11	−0.12	+0.08	−0.14	−0.77
6903	−0.80	+0.25	+0.15	+0.43	+0.63	−0.20	+0.13	+0.23	+0.22	−0.22	−0.26	−0.24	−1.11
10173	−0.80	−0.33	+0.41	+0.41	+0.43	+0.13	+0.38	+0.30	+0.08	−0.16	+0.07	+0.10	−1.83
7052	−0.70	+1.95	+0.34	+1.15	+0.95	+0.19	−0.21	+0.23	+0.11	−0.25	−1.52	−0.66	−2.29
2495	−0.60	−0.50	+0.33	+0.02	+0.12	+0.02	+0.22	+0.32	−0.02	+0.02	−0.08	−0.35	−0.10
11784	−0.55	+0.30	+0.52	−0.15	+0.30	+0.60	+0.20	+0.26	+0.20	+0.04	−0.35	−0.03	−1.88
5149	−0.54	+0.29	+0.11	+0.19	+0.19	+0.31	+0.12	+0.11	+0.22	−0.04	−0.08	−0.40	−1.02
9548	−0.50	+0.36	+0.36	+1.41	+0.61	+0.61	−0.10	+0.21	−0.10	+0.07	−0.65	−0.60	−2.19
9398	−0.44	+0.69	+0.89	+0.69	+1.05	+1.53	+0.63	+0.45	−0.38	−0.24	−0.56	−1.67	−3.07
963	−0.38	+0.25	+0.18	+0.42	+0.25	−0.06	+0.66	+0.18	+0.06	−0.06	+0.12	−0.36	−1.63
8489	−0.37	+0.81	+1.30	+1.40	+0.31	+0.25	−0.93	+0.59	−0.43	−0.28	−0.30	+0.20	−2.93
4968	−0.36	−0.95	+0.45	+1.17	+0.58	−0.21	+0.68	+0.48	+0.64	−0.64	−0.13	−0.22	−1.85
2914	−0.32	+3.13	+2.63	−0.66	+1.30	−0.12	−0.38	−0.77	−1.08	−0.19	−1.66	−1.37	−0.81
2345	−0.29	+0.48	+0.33	+0.33	+0.48	+0.30	+0.48	+0.15	+0.03	−0.47	−0.20	−0.46	−1.46
4426	−0.27	−0.11	−0.16	+0.09	+0.09	+0.06	+0.06	+0.09	+0.09	+0.07	+0.01	+0.01	−0.31
5470	−0.24	−0.04	+0.20	−0.00	+0.24	+0.17	+0.04	+0.15	−0.08	+0.10	−0.13	−0.11	−0.54
10738	−0.21	+0.84	−0.16	−0.26	+0.60	+0.23	+0.40	+0.33	−0.03	−0.17	−0.19	−0.58	−1.01
862	−0.20	+0.22	+0.13	+0.24	+0.22	+0.20	+0.16	+0.24	+0.16	−0.07	+0.24	+0.13	−1.89
2601	−0.19	+0.54	+0.16	+0.51	+0.38	−0.13	+0.08	+0.22	+0.02	+0.02	+0.21	+0.15	−2.16
10137	−0.15	+0.37	+1.31	+0.67	+0.57	+0.55	+0.66	−0.24	+0.66	+0.29	+0.07	−1.12	−3.80
6586	−0.14	+1.06	+0.39	+0.09	+0.09	+0.12	+0.49	+0.14	−0.22	−0.23	−0.17	−0.31	−1.44
1247	−0.13	+2.26	+0.31	+0.31	+0.66	+0.16	+0.16	−0.19	+0.25	−0.48	−0.54	−0.39	−2.54
10786	−0.08	+0.48	+0.66	+0.58	+0.58	+0.63	+0.13	+0.03	+0.21	−0.19	−0.28	−0.14	−2.68
11517	−0.06	+0.29	+0.29	+0.09	+0.09	+0.29	+0.23	+0.15	+0.23	+0.20	+0.21	+0.37	−2.44
10790	−0.04	+0.10	+1.10	−0.36	+0.70	+0.51	+1.00	+0.37	+0.02	−0.15	−0.36	−0.08	−2.84
3483	+0.05	+0.37	+0.74	+0.07	+0.37	+1.28	+1.38	+0.29	+0.19	+0.09	−0.29	−1.72	−2.77
7142	+0.06	+0.26	+0.10	+0.50	+0.26	−0.32	+0.26	+0.10	+0.36	−0.35	−0.27	+0.07	−0.98
8297	+0.06	+0.55	+0.55	+1.01	+0.55	+0.28	+1.04	−0.65	+0.74	−0.20	−0.49	−0.79	−2.61
10427	+0.08	+0.22	+0.54	+1.04	+1.04	+1.95	+0.29	−0.29	−0.06	−0.21	−0.93	+0.26	−3.85
9754	+0.12	+1.39	+1.51	+1.20	+1.70	+1.82	+2.13	−1.02	−0.07	−0.49	−2.02	−1.60	−4.55
11435	+0.16	+0.79	+0.65	+1.12	+1.26	+0.36	+0.59	+0.07	+0.49	+0.98	+0.02	−0.61	−5.69
10760	+0.16	+2.31	+1.75	−0.66	+1.24	+1.94	+1.36	+1.69	−0.89	−0.85	−1.38	−1.73	−4.80
6520	+0.16	+1.35	+1.05	−0.28	+0.32	+1.72	+0.93	−0.04	+0.23	−0.36	−1.30	−0.60	−3.01
6635	+0.16	+1.17	+0.77	−0.15	+0.85	+0.78	+0.45	+0.50	−0.48	+0.34	−0.20	−1.21	−2.84
7830	+0.18	+1.44	+0.58	+0.58	+0.76	+0.08	−0.14	+0.50	+0.35	+0.08	−0.46	−1.19	−2.56
7379	+0.21	+0.73	+0.93	+1.43	+0.83	+0.13	+0.80	+0.20	−0.20	+0.22	−0.77	−0.99	−3.32
5368	+0.26	+1.54	+1.04	−0.17	+0.42	+0.85	+0.85	+0.21	−0.02	−0.05	−0.67	−0.54	−3.47
4643	+0.27	+1.74	+1.44	+1.45	+1.15	+0.96	−0.44	+0.06	−0.49	−0.69	−0.17	−1.82	−3.19
6644	+0.30	+0.73	+2.21	+2.86	+0.73	−0.45	+0.75	+0.71	−0.22	−0.70	−1.44	−1.39	−3.81
933	+0.32	+0.92	+0.92	+2.73	+0.53	+0.71	+0.01	−0.05	−0.51	−0.48	−1.19	−0.95	−2.63
1406	+0.39	+2.37	+1.46	+2.26	−0.07	+0.47	+1.60	+0.54	−0.51	−0.69	−0.74	−1.24	−5.44
3651	+0.42	−0.29	+1.00	+2.29	+1.29	+0.35	+0.15	−0.21	−0.03	−0.18	−0.00	−0.77	−3.59
11520	+0.46	+0.03	+0.12	+0.12	+0.03	+0.05	+0.05	+0.06	+0.09	+0.07	+0.06	+0.05	−0.71
12140	+0.52	+1.75	+1.75	+1.75	+1.05	+1.17	−0.01	−0.59	−0.61	−0.54	−0.92	−0.83	−3.96
4059	+0.56	+2.57	+0.20	−0.45	+0.25	+0.15	+0.05	+0.30	−0.25	+0.03	−0.41	−0.49	−1.93
203	+0.58	+0.47	+0.47	−0.23	+1.38	+1.69	+1.59	+0.31	+0.92	+0.72	−0.84	−2.11	−4.38
1608	+0.64	+0.40	+0.40	−0.11	+0.25	+0.16	+0.25	+0.23	+0.22	+0.20	−0.11	−0.05	−1.85
1949	+0.92	+0.57	+0.20	+0.20	+0.57	+0.16	+0.30	+0.21	+0.30	+0.03	+0.31	+0.25	−3.11
2526	+1.01	+1.08	+1.08	+3.21	+0.82	+0.65	+1.65	+0.12	+1.00	−0.47	−1.39	−2.91	−4.82
2522	+1.13	−0.39	+0.22	+0.26	+0.23	+0.71	+0.23	+0.21	+0.19	+0.13	+0.26	+0.25	−2.29

G CAPABILITY PROFILES

For each model m we fit, over the 50 games and with all five z-scored axes entered jointly, the profile regression $\hat{\alpha}_{m,g} = \beta_{m,0} + \sum_{a=1}^{5} \beta_{m,a} z_a(g) + \varepsilon_{m,g}$. Table 13 has the point estimates and Figure 7 plots them with their paired-cluster bootstrap intervals (B=2,000). Benjamini–Hochberg correction at $q < 0.05$ across the 60 coefficients leaves eleven, which are in bold. Figure 8 shows the same profiles in chips per game, on a game at the extreme of one axis.

Table 13: Per-model OLS slopes of $\hat{\alpha}_{m,g}$ on the five z-scored axes, fitted jointly. Bold = BH-adjusted $q < 0.05$ across the 60 coefficients.

	state-sp.	opponent	temporal	precision	information
gpt-6-astra†	+0.29	+0.24	**+0.36**	−0.14	−0.31
gpt-5.4	+0.06	+0.01	**+0.36**	−0.07	−0.06
gemini-3.8-flash†	+0.28	+0.00	+0.07	+0.02	+0.14
gemini-3.1-pro	+0.10	−0.10	**+0.14**	−0.04	+0.06
claude-sonnet-4.6	+0.15	+0.06	+0.18	−0.07	+0.02
claude-fable-5.1†	+0.01	−0.05	+0.23	+0.07	+0.11
gemini-2.5-pro	+0.04	−0.02	−0.01	+0.11	−0.06
gemma-4-31b	−0.01	−0.06	−0.13	+0.07	**+0.19**
deepseek-v3.1	−0.05	−0.02	−0.04	+0.03	+0.08
gemini-3.1-flash-lite	**−0.14**	+0.04	**−0.21**	+0.08	+0.04
qwen-3.5	**−0.23**	−0.03	**−0.32**	−0.06	+0.07
llama-3.3-70b	**−0.51**	−0.07	**−0.63**	+0.01	**−0.27**

In Figures 2 and 7 the per-(model, axis) slope $\hat{\beta}_{m,a}$ is in chips/game per σ of axis, the local sensitivity of model m's edge to axis a. The companion plot below, with each spoke in chips/game on a benchmark-extreme game for that axis instead of a per-σ slope, comes from the same fitted regression.

We evaluate each model's predicted per-game strength on a synthetic game where axis a is at its observed maximum z-score and the other four axes are at their observed medians,

$$\hat{\alpha}_{m,g_a^*} \;=\; \hat{\beta}_{m,0} + \hat{\beta}_{m,a}\, z_a^{\max} + \sum_{a' \neq a} \hat{\beta}_{m,a'}\, z_{a'}^{\text{med}}.$$

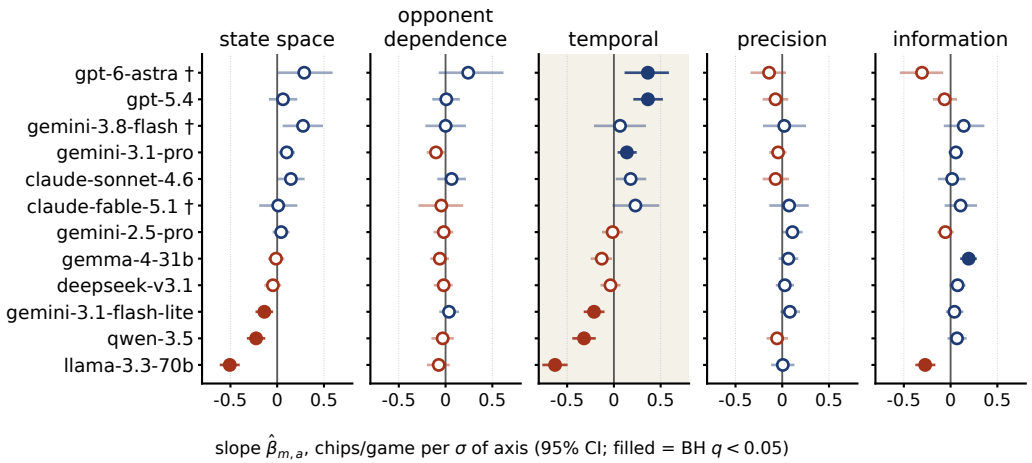

Figure 7: **Capability profile with intervals.** For each model, we regress its game strength $\hat{\alpha}_{m,g}$ on all five z-scored axes jointly, with an intercept, over the 50 games. Each panel shows the coefficients for one axis, with models in leaderboard order. Coefficients (chips/game per SD of the axis) are positive for models whose advantage over the average increases as a game increases in that axis value. Bars are 95% clustered bootstrap CIs (B=2,000). Filled markers (11 of 60 cells, 6 on the shaded temporal panel) are significant after applying Benjamini–Hochberg correction at $q < 0.05$. Blue markers are positive coefficients and red markers negative. † September models.

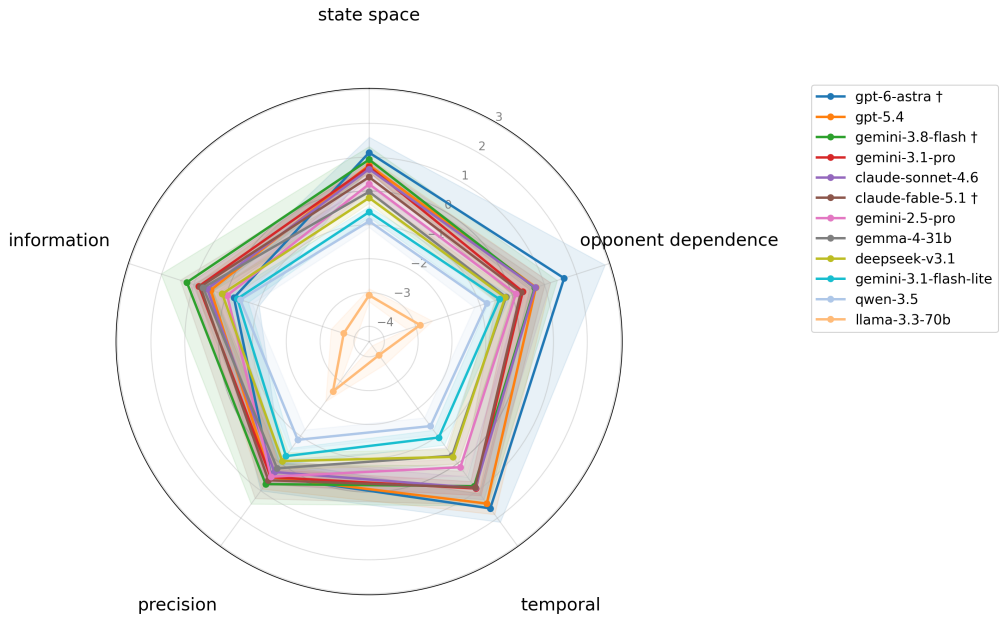

Figure 8: **Capability profile in absolute units.** The profile regression of Figure 2 is used to predict each model's per-game strength $\hat{\alpha}_{m,g_a^*}$ on a synthetic game. On that game axis a takes its 50-game maximum z-score and the other four axes take their 50-game medians. Units are chips/game relative to the across-model mean. The shaded bands are 95% paired-cluster bootstrap CIs (B=2,000) propagated through the linear predictor.

Robustness to rulebook-based confounds. Verbosity is the most direct way a rulebook could in principle confound the slopes. Rulebooks run from roughly 8,600 characters to just under 39,000. Adding the logarithm (base 10) of the character count as a regressor to the regression of Table 13, we find that the slopes are robust to rulebook verbosity. All eleven cells significant after BH correction in Figure 7 keep their sign, and the temporal slopes change by at most 0.09 chips/game. The added term mainly moves the state space slopes, with which log length correlates at $\rho = 0.82$ (no other slope moves by more than 0.09). After BH correction across the twelve models it is significant for two of them. With a coefficient of -0.41 per standard deviation of log length for `llama-3.3-70b`, it moves the state space deficit of `llama-3.3-70b` from -0.51 to -0.20, so part of what this axis captures for that model is rulebook length. With $+0.46$ for `qwen-3.5`, it deepens the state space deficit of `qwen-3.5` from -0.23 to -0.57.

Where the ordering departs from the leaderboard. On seed 4968, in the easiest tertile (Appendix E), astra has a strength of -0.95 chips/game (Table 12), although it ranks first overall. Games whose ranking departs from the leaderboard in this way are mostly small, and most of the departure on them is compression of the differences between models. On the five games where the departure exceeds sampling noise (862, 1949, 2522, 11517 and 11520), opponent dependence is highest (mean z of $+1.76$), and precision and information are also above average. Seed 4968 is not one of them, but it shows the mechanism clearly. In this game of three cards a position publicly reveals the holder of Q, so a bet into Q comes from J or K. The traces of `gpt-5.4` and astra both derive the sequential equilibrium, in which K is bet for value, J is bluffed with probability $9/13$, and Q is exactly indifferent. In the tournament the bettor held K almost always, and `gpt-5.4` folds Q, which happens to be right against a field that never bluffs. Astra bluffs against llama, which calls with Q every time, and loses six chips per bluff. `deepseek-v3.1` also calls every time, and its trace puts probability $1/2$ on each of J and K and computes a positive expected value. In a single deal (Section 5) what separates these models is the prior in each trace about the opponent, and whether a prior is a failure depends on the field.

G.1 How many patterns the profiles share

We choose the number of common factors in the profiles of Section 7 by parallel analysis (Horn, 1965; Buja & Eyuboglu, 1992) of $\hat{A}$, the 12×50 matrix of the strength $\hat{\alpha}_{m,g}$ of each model on each game. Table 14 reports the analysis for all twelve models (20,206 paired deals). It also covers the nine April models, estimated from April deals only (18,436), and the eleven models other than `llama-3.3-70b` (15,228). With all twelve models, the first factor accounts for 38% of the standardized variance and the 95th percentile of the null with no common factor is 33% (p=0.003). The second factor accounts for 16%, compared with 18% under the null of exactly one factor (p=0.19). The April subset gives p<0.004 (none of 300 null draws) and p=0.10. Without `llama-3.3-70b` the two p-values are 0.013 and 0.37. The game scores of the first factor are highly correlated across the three sets of models ($|\rho| \geq 0.93$), and their correlation with the temporal axis lies between 0.65 and 0.73 (with state space, $+0.65$ in the full set). After the first factor is removed, no pairwise residual correlation between models is significant under Benjamini–Hochberg correction (Benjamini & Hochberg, 1995). The test has limited power, and weaker patterns may go undetected at fifty games.

Horn's procedure compares the observed eigenvalues with those of random normal matrices. We use bootstrap replicates of the estimated strengths for the reference distribution. Let $\tilde{A}$ be $\hat{A}$ with each row centered and scaled to unit variance across games, with singular value decomposition $\tilde{A} = \sum_k s_k u_k v_k^\top$. Factor k accounts for a share $s_k^2 / \sum_j s_j^2$ of the standardized variance. The entries of v_k are its game scores, and u_k holds its model loadings. We draw B=300 bootstrap replicates of $\hat{A}$, in which the two seats of a run form one cluster, and subtract the bootstrap mean from each. The resulting 300 matrices of estimation noise are added to the row means of $\hat{A}$ under the null of no common factor, and to the rank one approximation of $\hat{A}$ in chips under the null of exactly one factor. Each null matrix is then standardized and decomposed in the same way as $\tilde{A}$. The p-value for factor k is the fraction of null matrices in which the kth factor explains at least as large a share of the variance as the kth factor of $\tilde{A}$. Factors are retained in order until the first one that is not significant at the 5% level.

Table 14: **Parallel analysis of the standardized model×game strength matrix.** Columns give the share of variance explained by each of the first three factors, the p-values p_1 against the null of no common factor and p_2, p_3 against the null of exactly one factor, and the Spearman correlation between the first factor's game scores and temporal contingency. The nulls use B=300 bootstrap replicates, clustered by run.

Subset	deals	variance share factors 1 / 2 / 3	p_1	p_2	p_3	ρ_{temporal}
full twelve	20,206	0.38 / 0.16 / 0.11	0.003	0.19	0.35	+0.73
April nine, April deals	18,436	0.43 / 0.18 / 0.12	< 0.004	0.10	0.37	+0.72
llama excluded	15,228	0.39 / 0.17 / 0.14	0.013	0.37	0.11	+0.65

H LOCAL JAGGEDNESS

For model m on game g, set $y_m(g) = (\hat{\alpha}_{m,g} - \bar{\alpha}_{m\cdot})/\sigma_U(g)$, subtracting the model's mean strength over games and dividing by the standard deviation of the final payoff when both seats move uniformly at random, as in the state-space count of Appendix C. Denote the bootstrap standard error on this scale by $\text{se}_m(g)$. A Gaussian kernel ridge fit over the five z-scored axes gives the smooth surface f_m. Both its length scale and ridge penalty are chosen by leave-one-game-out cross-validation. We use exact residuals $r_m(g) = y_m(g) - f_{m,-g}(x_g)$, with game g omitted from its comparison surface, to calculate

$$J_m \;=\; \sqrt{\max\Big(0,\; \text{var}_g\, r_m(g) - \text{mean}_g\, \text{se}_m(g)^2\Big)},$$

and refit on $B = 500$ bootstrap samples of games for the 95% intervals. Table 15 pairs J_m with the surface's leave-one-game-out R^2. Both can be small when a model varies little across games, as for `deepseek-v3.1` (0.05, −0.02).

Validation. Independent jumps of size 0.20 give $J_m = 0.20$ in synthetic data. A trend mixed with jumps of size 0.14 gives 0.15. We used the 50 benchmark games and 200 draws per case with `gpt-5.4`'s sampling noise. The flat truth has zero variation, with an estimated J_m of 0.03. For the other cases we set the true SD across games at about 0.20. Linear dependence on one axis or two gives 0.05 in either case. With curvature in the temporal axis, the estimate is 0.11, since the surface removes only part of this trend. Across draws, the SD of J_m is about 0.04, or 0.06 with a September model's sampling noise.

Resolution. The September models had only one opponent on most games. The widest intervals belong to `gpt-6-astra` and `gemini-3.8-flash`, both $[0.00, 0.18]$. Among the nine April models, the orderings from random halves of each game's paired deals have a Spearman correlation of 0.57. We cannot resolve differences below about 0.1. Among the April models, both halves rank `llama-3.3-70b` first (0.34 and 0.36) and `qwen-3.5` second (0.18 and 0.16).

Table 15: Local jaggedness J_m with 95% game-bootstrap intervals. We also report R^2 of a model's smooth surface over the five axes (LOO CV). † denotes the September models.

Model	local jaggedness J_m [95% CI]	smooth-surface R^2
`llama-3.3-70b`	0.35 [0.27, 0.42]	+0.44
`qwen-3.5`	0.17 [0.13, 0.21]	+0.24
`gpt-5.4`	0.12 [0.05, 0.17]	+0.35
`gemini-3.1-flash-lite`	0.12 [0.09, 0.14]	+0.33
`gpt-6-astra`†	0.11 [0.00, 0.18]	+0.32
`claude-sonnet-4.6`	0.10 [0.03, 0.15]	+0.14
`gemini-3.8-flash`†	0.09 [0.00, 0.18]	+0.26
`gemma-4-31b`	0.08 [0.03, 0.11]	+0.11
`gemini-3.1-pro`	0.08 [0.03, 0.10]	+0.38
`gemini-2.5-pro`	0.05 [0.00, 0.09]	+0.10
`deepseek-v3.1`	0.05 [0.00, 0.08]	−0.02
`claude-fable-5.1`†	0.00 [0.00, 0.09]	+0.01

I ABLATIONS AND TRACE ANALYSIS

I.1 STRATEGY PROGRAMS

Programs. From the readings alone, independent readers predicted the move in all 755 sampled situations of the 100 programs written by `claude-sonnet-4.6`, and in all 45 of a re-test of six revised readings. We generate each reading as a procedural description of the program in natural language. In disclosed play, LLMs receive both this reading and the program in its strategy language. The strategy language allows deterministic and randomized branches, using only the observable state of the game. Four of the 40 programs of `gemma-4-31b` are the default program, which always takes the first option, used after two failed attempts to parse its output (Bob in 10427, both seats in 10760, Alice in 10790).

Design. A pair uses one play seed in both seats. We take the pair's average chip margin as a single observation. In disclosed play, the program accompanies the rulebook in the LLM's prompt. In hidden play, the program and its reading are omitted. The efficiency reported in Table 16 uses the worst response as 0 and the best as 1, so 0.5 is their midpoint and is not neutral performance.

Table 16: **Programmed opponent ablation.** 520 complete pairs. Slope coefficients relate players' chip margins against the programs to the tournament ratings of Table 1. We resample deals, B=300 times for slopes and B=2,000 for disclosure effects and efficiencies.

Quantity	**Estimate**	**95% CI**
$\lambda_{\text{disclosed}}$ (five players)	$+1.10$	$[+0.76, +1.41]$
λ_{hidden} (three players)	$+0.92$	$[+0.24, +1.58]$
$\lambda_{\text{disclosed}} - \lambda_{\text{hidden}}$	$+0.18$	$[-0.40, +0.75]$
Disclosure effect, chips/game (160 deals)	$+0.72$	$[+0.19, +1.30]$
`gpt-5.4`	$+0.87$	$[+0.27, +1.53]$
`claude-sonnet-4.6`	$+0.90$	$[-0.23, +2.08]$
`llama-3.3-70b`	$+0.45$	$[-0.55, +1.53]$
*Play efficiency on the five exactly solved games (*0 = *worst,* 1 = *best response)*		
	disclosed	**hidden**
`gpt-5.4`	1.02 [0.82, 1.28]	0.81 [0.70, 0.90]
`claude-sonnet-4.6`	1.01 [0.80, 1.25]	0.71 [0.53, 0.84]
`gemini-3.1-pro`	0.98 [0.81, 1.15]	not played
`gemma-4-31b`	0.97 [0.84, 1.09]	not played
`llama-3.3-70b`	0.53 [0.37, 0.72]	0.53 [0.24, 0.79]

I.2 THINKING BUDGET

We estimate the expected gain in chip margin within a pair from low to high effort, $\Delta = \mathbb{E}[\text{chip margin}_{\text{high}} - \text{chip margin}_{\text{low}}]$. Our estimate $\hat{\Delta}$ uses the 260 sibling pairs from seven (family, anchor) cells that were played on the benchmark games. For 9 of its seeds the block served the default-parameter variant of the seed from a cache, not the benchmark game, and we leave those matches out. Confidence intervals (95%) are estimated using a cluster bootstrap on game seed. The two anchors are `gemini-3.1-pro` (a frontier model near the top of the leaderboard) and `gemma-4-31b` (an open model from the middle of the pack).

Table 17: **Thinking ablation.** Paired gains in chip margin by family. Significant estimates are starred.

Model family	$\hat{\Delta}$	**95% CI**	**pairs**
`claude-sonnet-4.6`	$+0.37$	$[-0.31, +1.21]$	104
`gemini-2.5-pro`	$+0.06$	$[-1.83, +2.00]$	32
`gemini-3.1-pro`*	$+0.53$	$[+0.40, +0.60]$	30
`gpt-5.4`	$+0.19$	$[+0.00, +0.36]$	94

* 95% CI excludes 0.

All four estimates in Table 17 are positive, from $+0.06$ for `gemini-2.5-pro` to $+0.53$ for `gemini-3.1-pro`. Only the estimate for `gemini-3.1-pro` is significant, and it rests on three games. The interval for `gpt-5.4` touches zero.

I.3 Reasoning traces

Reasoning about the hand is worth $+1.15$ chips per rubric point more on games in the top tertile of the information axis than on other games (SE 0.38, $p = 0.002$), and planning depth $+0.93$ more in the top tertile of the temporal axis (SE 0.34, $p = 0.007$). Here the focal chip edge of a move, the final margin of the model that made it in that match, is regressed on the rubric scores defined below, each interacted with an indicator for the top tertile of its matched axis. Like every regression in this section, it has model fixed effects and standard errors clustered by game. A control for the same interaction with state space gives $+1.06$ ($p = 0.001$) and $+0.71$ ($p = 0.046$). For the earlier six axes neither interaction is significant ($+0.69$, $p = 0.10$, and $+0.40$, $p = 0.30$). Comparing top and bottom tertiles gives the same answer, $+0.59$ (SE 0.20) and $+0.69$ (SE 0.18), against $+0.08$ (SE 0.21) for opponent modeling. Across games, only the opponent axis predicts its matched rubric ($+0.04$ points per σ, $q = 0.03$). The rubric scores are lower on larger games, by 0.07 points per σ of state space for opponent modeling and 0.10 for planning depth (both $q < 0.001$). Note that all of these estimates are associations. For example, a trace that reasons about the hand often comes from a move where the hand is decisive.

Rubric. For each sampled move, the judge (`gemini-3-flash-preview`) is given the game's rulebook, the agent's prompt, the agent's full response, and the parsed action. The prompt is the game context, rendered deterministically, and we treat it as ground truth. The judge scores the reasoning visible in the response on three ordinal 0–3 dimensions, each matched to one value axis. Move quality is not scored. The judge prompt also contained a fourth dimension, risk reasoning, which matched the risk axis of the earlier selection basis, and we do not use its scores. Anchors:

Dimension	Anchors (0 / 1 / 2 / 3)
Opponent-modeling engagement	ignores the opponent / opponent mentioned but treated as fixed / forms an explicit model of the opponent's likely action / recursive: anticipates the opponent's anticipation
Private-information reasoning	does not use own hand as evidence / hand named, possibly with a brief strength label / analysis along one line: combines two private sources, or reasons about win rate or expected value / both lines, multi-source and quantitative
Forward-planning depth	myopic / mentions future turns without conditions / one conditional fork about a future state / two or more chained or parallel conditionals

The sample covers thirteen model variants, among them the variants with low reasoning effort from the thinking ablation. The three dimensions are matched to opponent dependence, information, and temporal contingency, in the order listed.

The anchors follow how traces are coded in the poker and game reasoning literature, including the theory of mind coding of Lin & Hou (2026). We drew 10,740 moves from the tournament and the thinking ablation, stratified by model variant, capability tier and quadrant of the earlier opponent and temporal axes. Of these, $10{,}721$ were parsed. We analyse the $9{,}778$ whose game is the benchmark game (101 to 372 per game). The 943 moves from ablation matches played on a default-parameter variant are left out (Appendix I.2). A subset was also scored by a second judge, `gemini-3.1-pro`, which gives the same signs for the private-information and planning interactions on the five axes.

J Data notes

Parser fallbacks. We did not need a fallback for seven of the twelve models. The highest rate among the others was 0.33% of moves. If neither the strict nor the lenient parser recovers a legal action, the engine plays a uniformly random legal action and records the fallback.

Model	matches	moves	fallbacks	move rate	match rate
gemini-3.1-flash-lite	9,850	23,734	79	0.33%	0.8%
gpt-5.4	4,652	11,021	13	0.12%	0.3%
gemini-3.1-pro	10,490	24,520	9	0.04%	0.1%
gemini-2.5-pro	9,400	21,975	9	0.04%	0.1%
deepseek-v3.1	9,398	22,682	8	0.04%	0.1%
gpt-6-astra	1,380	3,809	0	0.00%	0.0%
gemini-3.8-flash	1,000	2,194	0	0.00%	0.0%
claude-fable-5.1	1,360	3,628	0	0.00%	0.0%
claude-sonnet-4.6	4,078	9,272	0	0.00%	0.0%
gemma-4-31b	9,834	23,497	0	0.00%	0.0%
qwen-3.5	9,426	21,559	0	0.00%	0.0%
llama-3.3-70b	9,956	25,945	0	0.00%	0.0%

Rulebook fidelity audit. Eight independent LLM auditors and the authors checked all 50 served rulebooks against the engine and found no engine defect. In 32 rulebooks the text leaves out a detail that the engine resolves itself: that an auction prize swaps or peeks at a random card (14 and 12 games), that a bonus prize is capped at the losing bidder's stack (9), or that a forced discard or a forced play takes the card received first (9 and 3). In 12 rulebooks it also says that chip costs in hidden menus are settled at the end of the game, whereas the engine deducts them at once and the opponent sees the new stack. In all, 34 of the 50 rulebooks are affected. Every defect is symmetric across seats, since both players read the same rulebook and each model plays both seats of every deal. We refit the leaderboard without the games of each class in turn. The three tiers of Table 1 are unchanged, models change places only within their tier, and no rating moves by more than 0.28 chips/game. Because each defect goes with one mechanic, this test cannot separate an effect of the text from an effect of the mechanic.

Rerun and retained matches. During the project we found and fixed two bugs: a rulebook rendering defect in one April tournament session, and a parser bug that could turn a legal answer into a random fallback. We ran 45 affected matches again on the same play seeds and replaced them. The rendering defect also altered play on 21 matches, which are excluded from every analysis. Another 58 could not be run again and were retained as played (0.14% of the corpus). They occur in 54 paired deals. If we drop all 54, the largest change to a rating in Table 1 is about 0.010 chips/game. Both repair logs are distributed with the data, along with the lists of excluded and retained matches.