GAMBIT: Generalization Assessment via Modified Boards and Irregular Topologies for LLMs in Chess

Max Highsmith

April 2026

https://fae-chess.com

Abstract

We introduce GAMBIT, an evaluation framework that benchmarks large language model (LLM) strategic reasoning through 16 chess variants that are structurally absent from pretraining corpora. Despite impressive results on other metrics of intelligence, most general-purpose LLMs struggle to consistently beat random play on standard chess, with only specialized reasoning models achieving reliable above-random performance [11]. This suggests a fundamental deficit in strategic reasoning. Chess variants isolate this capability by preserving core strategic demands (planning, evaluation, adversarial decision-making) while eliminating memorization confounds. Our evaluation spans 20 models across 16 variants in 16,400 games, including, standard chess, 3D spatial variants (Raumschach, Projective 3D, Torus 3D), non-Euclidean topology (Torus, V-Cylinder, Cylinder, Klein, Möbius, Projective), rule modifications (Alice, Crazyhouse, Racing Kings), imperfect information (Fog of War, Kriegspiel), and alternative tile geometries (Hexagonal). Each model plays 30 games per variant against random baselines (with a reduced pilot of 5 for three expensive reasoning models), with complexity characterized through 500 self-play games per variant. Results reveal systematic performance patterns: the grand mean score across all models and variants is 53.8%, only marginally above the 50.0 random baseline. Most models cluster near random on the majority of variants, with only 6 (Grok 4 Fast, Gemini 3.1 Pro, GPT-4.1 Mini, GPT-5.4) consistently exceeding chance across multiple games. Strikingly, LLM-generated chess engines outperform the same models' prompt-based play in nearly every (model, variant) pair, suggesting that LLMs possess latent strategic knowledge they can express through code generation but not through direct move selection.

Contents

1 Introduction

1.1 Chess as a Mirror for Artificial Intelligence

Chess has served as a proving ground for artificial intelligence since before the field had a name. In 1770, Wolfgang von Kempelen's *Mechanical Turk* (a fraudulent chess-playing automaton concealing a human operator) captivated Enlightenment audiences and seeded the idea that machines might one day think strategically. The aspiration became scientific in the twentieth century. Alan Turing hand-simulated a chess program on paper in 1952, and Claude Shannon's foundational 1950 paper [14] framed chess programming as a lens for studying machine problem-solving, proposing the minimax search with evaluation functions that would dominate the field for decades.

The subsequent history of computer chess tracked, and often drove, the trajectory of AI itself. Early programs by Bernstein (1958) and Newell, Shaw, and Simon (1958) demonstrated that heuristic search could navigate combinatorial spaces. The iterative-deepening alpha-beta framework matured through the 1970s and 1980s, culminating in Deep Blue's 1997 victory over World Champion Garry Kasparov [2], a watershed moment that demonstrated brute-force search could surpass human intuition in a bounded domain. Two decades later, DeepMind's AlphaZero [16] achieved superhuman play through pure self-play reinforcement learning, mastering chess, shogi, and Go from tabula rasa in hours, signaling the power of learned representations over hand-crafted evaluation.

1.2 The Strategic Reasoning Gap in LLMs

Standard chess presents a revealing test case for language model reasoning. The domain is well-defined, requires multi-step planning and strategic evaluation. Yet most LLMs perform poorly at chess. Recent systematic evaluation [11] shows that the majority of language models cannot consistently beat random move selection, with only specialized reasoning models (o1, o3, o4-mini) achieving above-random performance.

This failure is instructive: despite excelling at tasks like code generation [41], mathematical reasoning [6], and natural language understanding [7], LLMs struggle with the kind of formal rule application and adversarial planning that chess demands. This suggests a specific deficit in strategic reasoning: the ability to internalize formal rules, evaluate positions, and plan multi-step sequences under adversarial conditions. Even among reasoning models that show above-chance chess performance, the widespread availability of chess data on the internet makes it difficult to determine whether these models are developing generalizable strategic reasoning or simply memorizing patterns.

1.3 Our Approach

GAMBIT addresses this ambiguity by evaluating LLMs on chess variants. These variants preserve the core cognitive demands of chess (tactical calculation, positional evaluation, adversarial planning), while their limited data availability reduces the confound of rote memorization.

Moves that are effective in standard chess may become liabilities on boards with additional dimensions or novel topologies. As LLMs continue to improve at chess, GAMBIT provides a framework for assessing the extent to which these gains reflect generalizable strategic reasoning.

In this paper we contribute:

1. **A suite of 16 chess variants** spanning distinct dimensions of game complexity and structural variation from standard chess: standard chess as a contaminated control, 3D spatial rea-

soning (Raumschach, Projective 3D, Torus 3D), non-Euclidean topology (Torus, V-Cylinder, Cylinder, Klein, Möbius, Projective), rule modifications (Alice, Crazyhouse, Racing Kings), imperfect information (Fog of War, Kriegspiel), and alternative tile geometries (Hexagonal).

2. **Systematic empirical evaluation** across 20 models and 16 variants, comprising 16,400 games with 30 trials per (variant, model) pair (5 for three expensive reasoning models). Including: LLM vs random baselines, LLM vs traditional AI baselines, and self-play random control data (500 games per variant) for complexity characterization.

3. **Complexity analysis** characterizing each variant's intrinsic difficulty using multiple complexity proxies, and examining model performance trends across these axes.

The framework, all game logic, and evaluation harness are open source at https://github.com/Max-Highsmith/faechess, with all variants playable by humans or agents at https://fae-chess.com.

2 GAMBIT LLM

2.1 Design

GAMBIT comprises a series of experiments involving 16 chess variants organized into six cognitive categories: *Standard* chess (included as a contamination-saturated control), *Topological* variants that alter board connectivity through edge identifications (Torus, Cylinder, V-Cylinder, Möbius, Klein, Projective), *3D* variants that extend spatial reasoning to cubic lattices (Raumschach, Torus 3D, Projective 3D), *Rule* variants that modify game mechanics (Alice Chess, Crazyhouse, Racing Kings, Five-Board), *Imperfect Information* variants that restrict board visibility (Fog of War, Kriegspiel), and a *Hexagonal* variant that replaces the square grid entirely. Each variant is implemented as a self-contained game engine with full move generation, legality checking, and board serialization, all sharing a common evaluation harness (Table 1; Figure 1). The variants were selected to span qualitatively distinct reasoning demands while remaining grounded in chess conventions, so that the rules can be communicated to an LLM in a single natural-language prompt. Before a game begins, the model receives a system prompt containing the complete rules of the variant: board geometry, piece movement, win conditions, and any special mechanic. If the model's response cannot be parsed after one retry, a random legal move is played on its behalf (recorded as a *fallback*). Games proceed until checkmate, stalemate, draw by repetition, or a configurable ply limit is reached. This design isolates strategic reasoning from ancillary capabilities: the model never needs to generate legal moves, track game state, or parse algebraic notation; it only needs to choose among explicitly enumerated options given full information (or, in imperfect-information variants, the information permitted by the rules).

2.2 Benchmarks

The core evaluation of GAMBIT pits two agents against one another. In additon to the tested models we introduce two non-LLM based agents: Random and Tree. In the random approach a list of valid moves is presented to the agent each turn and a random number generator chooses which move to select. The random opponent provides a floor baseline: any score found statistically significantly above 50.0 indicates the model has learned something beyond chance.

The "tree approach" pits each LLM against a local depth-1 minimax search engine using alpha-beta pruning [14]. The evaluation function is a weighted sum of material and positional terms.

Each piece type is assigned a standard material value (Pawn = 100, Knight = 300–320, Bishop = 330, Rook = 500, Queen = 900, King = 10,000), and a positional bonus rewards centrality on the board (computed as the Manhattan distance from the center, scaled by a small coefficient). Move ordering prioritizes captures by victim value to improve pruning efficiency. A small random jitter (± 10) is added to evaluations to prevent deterministic play. By providing a stronger calibrated baseline, the tree approach measures whether LLMs can beat a simple but non-trivial lookahead opponent. For move quality analysis (Section 3.3), we increase the search to depth 2 to rank all legal alternatives at each position.

2.3 Metrics

Our evaluation uses two families of metrics: *performance metrics* that quantify how well each model plays, and *complexity metrics* that characterize the intrinsic difficulty of each variant.

2.3.1 Performance Metrics

Score. Each game yields a per-game score $X \in \{0, 0.5, 1\}$ (loss, draw, win). The aggregate score for a (model, variant) cell is

$$\text{Score} = \frac{W + 0.5D}{N} \times 100$$

where W, D, and N are wins, draws, and total games respectively. A score of 50.0 indicates random-level play; values above 50.0 indicate above-random strategic capability. We also report **Win Rate** (W/N) and **Decisive Win Rate** ($W/(W + L)$, excluding draws) as supplementary metrics. Score is the primary metric because it is sensitive to the full outcome distribution: a model that draws frequently is meaningfully stronger than one that loses, but pure win rate treats both equally. Decisive win rate is useful for variants with high draw rates (e.g., Raumschach), where it separates models that survive to draws from those that win outright.

Move Quality (Top-k Rate). For each LLM move during a game, we retrospectively rank all legal moves using a depth-2 minimax search with alpha-beta pruning (Table 3). A move is considered *top-k* if it falls within the top k moves by minimax evaluation. We report the percentage of LLM moves that rank in the top 5, providing a measure of tactical quality independent of game outcome.

2.3.2 Complexity Metrics

To characterize the intrinsic difficulty of each variant independent of any particular model's capabilities, we simulate 500 random-vs-random self-play games per variant, where both sides select uniformly at random from their legal moves. This provides a model-free baseline that captures structural properties of the game itself: how wide the decision tree is, how quickly material changes hands, how often games end decisively, and how much tactical pressure the board geometry creates. All metrics below are computed from these random self-play trajectories (Table 6; see Table 5 for formal definitions):

- **Branching Factor:** Average number of legal moves per position. Measures the width of the game tree; standard chess averages ∼30, while 3D variants can exceed 150.
- **Game Tree Complexity** ($\log_{10}$)**:** Estimated as $\log_{10}(B^{D/2})$ where B is the branching factor and D is the average game depth. Captures the total search space size.

- **Threat Density:** Fraction of board squares under attack by the opponent, averaged across all positions. Higher values indicate more tactical pressure; wraparound topologies exceed 60% compared to standard chess at ∼35%.
- **King Mobility:** Average number of legal king moves per position. Lower values indicate more constrained, tactically pressured positions.
- **Irreversibility Rate:** Percentage of moves that are irreversible (pawn advances and captures). Higher rates indicate more committal, strategically consequential play.
- **Eval Volatility:** Mean absolute material swing between consecutive positions. Higher volatility indicates chaotic, swingy games; non-orientable surfaces average >1.1 compared to standard chess at 0.35.
- **Capture Rate:** Percentage of moves that capture opponent pieces. Correlates with tactical intensity and topology-induced complexity.
- **Draw Rate:** Percentage of games ending in stalemate or repetition. Higher draw rates indicate more defensive, positional dynamics (Table 11).
- **Average Plies:** Mean game length in half-moves. Shorter games may indicate faster material attrition or more decisive openings (Figure 3; Figure 4).

These metrics are analyzed jointly via PCA to identify the principal axes of variation across variants (Section 3.5).

3 Results

3.1 LLM vs Random

We run n=30 trials of LLM vs Random for 17 models over 16 game variants (Table 2). We additionally run n=5 trials for three high-cost frontier models (Gemini 3.1 Pro, Claude Opus 4.7, and GPT-5.4) over the same 16 variants; trial counts are limited to 5 due to API cost (Table 8).

Using Chatbot Arena Elo as a proxy for general intelligence (Table 10), we observe a positive but weak correlation between model Elo and average performance across games ($R^2 = 0.15$; Figure 5).

Across the full 320 model-variant cells, only 37 (11.6%) show statistically significant improvement over random (BH-corrected $p < 0.05$; Table 2). No model achieves significance on a majority of its tested variants. Grok 4 Fast leads with significant above-random scores on 9 of 16 variants, followed by GPT-4.1 Mini (3 of 16), GPT-4o Mini (3 of 16), and DeepSeek V3 (3 of 16). Meanwhile, 5 models show at least one statistically significant *below*-random cell, and models such as Claude Haiku (mean 49.0), ERNIE 4.5 (49.5), Nemotron Nano 9B (49.4), and Gemma 3 27B (49.3) score at or below random level when averaged across all 16 variants.

A few models show striking variant-specific strengths. Grok 4 Fast dominates topological variants, scoring 98.3 on Klein, 96.7 on Projective, and 90.0 on Möbius. Gemini 3.1 Pro achieves 100.0 on both Racing Kings and 100.0 on Fog of War. GPT-5.4 reaches 100.0 on Fog of War and 90.0 on Racing Kings. However, even these top performers regress toward random level on high-complexity variants: Grok 4 scores 55.0 on Standard and 50.0 on Raumschach, and Gemini 3.1 Pro scores 50.0 on Standard, Raumschach, and Kriegspiel.

Notably, performance varies dramatically across variant categories. Topological variants (Klein, Möbius, Projective, V-Cylinder) and Racing Kings exhibit the widest spread between models, while Standard and Raumschach produce uniformly near-random scores across all models.

3.2 LLM vs Tree

We evaluate n=5 trials of each of 20 LLMs against a depth-1 minimax search engine with alpha-beta pruning across 14 variants (Table 12).

Nearly all models fail to beat even this minimal lookahead opponent. On Torus, 16 of 20 tested models score 0.0 (all losses); on Raumschach, no model exceeds 50.0. The sole consistent exception is **Grok 4 Fast**, which is the only model to achieve statistically significant above-D1 scores across multiple variants: 50.0 on Raumschach (0 win, 5 draws in 5 games), 50.0 on Torus (1 wins, 3 draws, 1 losses), and 70.0 on Alice (2 wins, 3 draws, 0 loss). On Alice Chess, Nova Pro also reaches 50.0 (2 wins, 1 draw, 2 losses), but no other model wins more than a single game against the tree on any variant.

While models were consistently defeated by the depth-1 tree, we observe that many achieve draw rates substantially above chance, particularly in draw-heavy variants like Raumschach. For example, GPT-4.1 Mini (30.0), Nova Pro (40.0), and GPT-4o Mini (40.0) all survive to draws against the tree in Raumschach at rates considerably above the zero-score losses suffered by weaker models such as Gemma (20.0) and Qwen (20.0). This suggests partial strategic capability insufficient for victory but adequate for survival.

3.3 Move Quality Analysis

Because the majority of models lost to the depth-1 tree, we use an extended depth-2 minimax search to evaluate move quality at a finer granularity. For each (model, variant) pair, we sample 10 games from the LLM vs Random experiments (Table 2) and replay every LLM move, ranking it among all legal alternatives using the depth-2 engine (Table 3).

When averaged across variants, every tested LLM achieves a higher mean top-5 move rate than the random baseline, though individual (model, variant) cells often fall at or below chance. In Standard chess, random play produces a 17.5% top-5 rate, while models range from 16.9% (Nemotron Nano 9B) to 72.1% (Grok 4 Fast). On Torus, the random baseline is 22.6%; DeepSeek V3 reaches 39.1% and Gemini Flash 33.8%. Even on Hexagonal, where the branching factor inflates the search space (random top-5 rate: 9.5%), DeepSeek V3 achieves 28.2% and Gemini Flash 32.8%.

This result is notable because many of these same models fail to achieve statistically significant increases in score or win rate. For instance, DeepSeek V3 scores 53.3 on Standard (near random by outcome) yet its top-5 move rate of 23.0% is well above the 17.5% random baseline. Similarly, Haiku scores 38.3 on Klein (below random by outcome) but achieves a 27.4% top-5 rate versus the 24.0% baseline, with a mean rank of 14.3 compared to 17.0 for random. The pattern indicates that LLMs do confer partial strategic signal: their move selections are measurably better than chance even when the cumulative advantage is insufficient to produce consistent victories.

3.4 LLM Engine Building

We prompt each model to generate a self-contained JavaScript function `chooseMove(input)` that receives the board state, legal moves, piece values, and a game clock, and returns a move index. The generated code runs in a sandboxed VM with no access to external libraries. The model is given up to five attempts to produce compilable code; if all five fail, the entry is marked as FAIL in Table 4 and the model plays random for that variant. If the code compiles but throws a runtime error on a particular move, that move falls back to random selection (tracked as "runtime fallbacks"). Each generated engine plays 30 games against a random opponent per variant. Strikingly, in nearly all (model, variant) pairs the LLM-generated chess engine outperforms the same model's prompt-based play (Table 4, Figure 8, Figure 9). The most dramatic cases include Claude Opus 4.7 on

Raumschach (codegen score 100.0 vs. prompt score 50.0), GPT-5.4 on Torus (codegen 100.0 vs. prompt average 52.7), and Haiku on Torus (codegen 86.7 vs. prompt 43.3). Across all 16 variants, codegen engines frequently achieve scores in the 60 to 100 range on variants where prompt-based play hovers near 50.

Notable exceptions where codegen performs at or below prompt-based play include Alice Chess, where several models' generated engines scored below baseline (Haiku 18.3, GPT-4.1 Mini 18.3); and Racing Kings, where Gemini Flash's codegen (33.3) underperformed its prompt score (96.7). Certain models also failed to produce compilable code within five attempts, falling back to random play: ERNIE 4.5 failed on the majority of variants, while DeepSeek V3, Nova Pro, GLM 4.7 Flash, and GPT-4.1 Mini each failed on two to three variants (Table 4).

3.5 Quantifying Game Complexity

To characterize each variant's intrinsic difficulty independent of LLM performance, we compute complexity metrics from 500 random self-play games per variant (80-ply cap). Table 5 defines each metric and the dimension of complexity it captures.

Principal component analysis on the 16 variants' complexity profiles reveals two dominant axes: PC1 (55.9% variance) separates variants by game scale and stability, with game-tree size, game length, draw rate, and king mobility loading positively against irreversibility, eval volatility, and capture rate loading negatively. PC2 (28.6% variance) captures tactical intensity, driven by threat density, capture rate, and eval volatility (Figure 2). To test whether models differ in their sensitivity to these complexity dimensions, we fit OLS regressions of score on PC1 and PC2 for each of the 20 models (Appendix E).

Of 20 models tested, 2 show a statistically significant relationship between PC1 position and score at $p < 0.05$ (nominal). Qwen 2.5 72B exhibits a positive slope ($\beta_1 = 1.54$, $p = 0.048$, $R^2 = 0.29$), performing better on high-PC1 variants (structured, longer games such as Racing Kings and Raumschach). Conversely, Phi-4 ($\beta_1 = -1.16$, $p = 0.021$, $R^2 = 0.39$) performs better on low-PC1 variants (chaotic, short topology games such as Klein and Projective). The β_2 coefficient (PC2) is not significant for any model, indicating that differentiation is entirely along the PC1 axis.

These nominal results do not survive Bonferroni correction for 20 comparisons (smallest corrected $p = 0.42$) or Benjamini-Hochberg FDR at $\alpha = 0.10$. We report this pattern as an exploratory, hypothesis-generating observation: some models may exhibit complementary complexity profiles along the dominant axis of variant difficulty, but confirmation requires evaluation on a larger set of variants.

4 Discussion

4.1 What GAMBIT Reveals About LLM Strategic Reasoning

Per-move quality vs. game-level outcomes. The move quality analysis reveals a persistent gap between per-move tactical quality and game-level performance. On average, models select moves ranked higher than random by the depth-2 evaluator (e.g., top-5 rates up to 72.1% vs. the 17.5% random baseline on Standard), though individual (model, variant) cells sometimes fall at or below chance. Yet even where per-move quality is measurably above random, this partial signal rarely translates into victories: most models score within a few points of 50.0. This gap between per-move quality and game-level outcomes suggests that incremental move-by-move advantages do not compound reliably over the course of a full game.

Raumschach as a specific deficit. Raumschach (5×5×5 3D chess) produces the most uniform near-random results of any variant. All 20 models score between 45.0 and 51.7 on Raumschach; no model achieves even a single percentage point of statistically significant lift. By contrast, 2D topological variants (which also demand novel spatial reasoning but in a familiar planar geometry) show far more variance: Grok 4 scores 98.3 on Klein, and GPT-4.1 Mini scores 66.7 on both Klein and Torus. Notably, Torus 3D (which also requires 3D spatial reasoning, but on a standard 8×8 board with toroidal wrap) shows meaningful model differentiation, with several models scoring above 60. This contrast suggests that the bottleneck is not 3D reasoning per se, but the representational challenge of Raumschach's 125-square cubic board with 26-directional attack vectors and a unique piece set (Unicorn).

The codegen paradox. Perhaps the most striking finding is that LLM-generated chess engines systematically outperform the same models' direct play. Claude Opus 4.7's codegen engine scores 100.0 on Raumschach (30/30 wins) while its prompt-based play scores exactly 50.0. GPT-5.4's codegen engine scores 100.0 on Torus (30/30 wins) while its prompt play achieves only 52.7. This dissociation suggests that LLMs possess strategic knowledge they can express through code (evaluation functions, search heuristics, material weighting) but cannot effectively deploy when forced to reason move-by-move in natural language. This extends the "Program of Thoughts" paradigm [42], which showed that delegating computation to generated code outperforms chain-of-thought reasoning on numerical tasks, to the domain of adversarial strategic planning: code generation provides a structured scaffold that elicits deeper reasoning than the sequential, token-by-token process of conversational move selection.

Model scaling and generalization. We observe a positive but modest correlation between Chatbot Arena Elo and GAMBIT performance ($R^2 = 0.15$; Figure 5). However, the relationship is far from monotonic. Grok 4 Fast (Elo 1419) is the top performer at 74.5 mean score, outperforming higher-ranked models like Gemini 3.1 Pro (Elo 1505, mean 63.1) and Claude Opus 4.7 (Elo 1503, mean 53.1). Conversely, Claude Haiku (Elo 1256) scores 49.0, essentially random. This suggests that general language model capability is a necessary but insufficient predictor of strategic reasoning in novel domains; architecture-specific or training-specific factors may play a larger role.

Standard chess performance does not predict variant performance. On Standard chess, models cluster near random: scores range from 48.3 to 55.0, with Grok 4 Fast and GPT-4o Mini tied at 55.0 and 15 of 20 models scoring between 48.3 and 50.0. This near-uniform result is notable because Standard chess is plausibly the most data-rich variant: opening theory, endgame tablebases, and annotated games are widely available in public corpora. Whatever chess knowledge models may have acquired from such data does not translate into above-random play in this task format. By contrast, the same models show wide variance on novel variants with little or no training-data presence, where top scores exceed 90. The variants most amenable to LLM play (Racing Kings, Fog of War) tend to have more decisive game dynamics, suggesting that variant-specific structural properties, rather than data exposure, drive the observed performance differences.

4.2 Implications for AI Evaluation

Contamination-resistant benchmarks. GAMBIT's design addresses a structural limitation of existing LLM benchmarks. Standard benchmarks like MMLU [7] and GPQA [13] face persistent contamination concerns as training corpora expand. GAMBIT sidesteps this problem by construction: chess variants like Klein Bottle chess and Projective 3D chess have no meaningful presence in any known training corpus. The Standard chess control provides supporting evidence: models score near random on Standard despite its abundant representation in public training data, suggesting that chess knowledge acquired from text does not transfer effectively to this evaluation format.

Since data exposure does not confer an advantage even for the most data-rich variant, the observed performance variance across novel variants is more parsimoniously attributed to variant-specific structural properties than to differential data contamination.

Game-based evaluation complements QA benchmarks. Unlike multiple-choice QA benchmarks, game-based evaluation measures sustained multi-step reasoning under adversarial conditions. A correct answer on a QA benchmark requires a single inference; winning a chess game requires 20 to 40 consecutive good decisions against an opponent that exploits mistakes. The observation that LLMs achieve measurably better move quality than random (Table 3) while still losing most games highlights this distinction: partial competence on individual decisions does not compound into reliable multi-step performance.

Transfer as the critical measure. GAMBIT's key contribution is measuring transfer: the ability to apply strategic reasoning to structurally novel domains. The wide variance in model performance across variants (from 20.0 to 100.0 within a single model) reveals that transfer is highly domain-dependent. Models that excel at Racing Kings (a low-branching, goal-directed variant) often fail on topological variants (which demand spatial reasoning about non-Euclidean connectivity). This suggests that "strategic reasoning" is not a monolithic capability but a family of skills that transfer selectively across task structures.

4.3 Limitations

- **Move selection vs. move generation.** By providing legal move lists, we reduce the task to selection rather than generation. This understates the difficulty of rule internalization.
- **Prompt sensitivity.** LLM performance may vary with how rules and board states are presented. We have not exhaustively ablated prompt design.
- **Move limit artifacts.** Games capped by ply limits may not reflect the full strategic arc (Figure 4). Future work should explore longer time horizons.
- **Positional evaluation depth.** The minimax baselines use simple material + position evaluation. Stronger engines would provide more discriminative baselines.

5 Conclusion

We have introduced GAMBIT, a contamination-resistant evaluation framework that isolates strategic reasoning from pattern retrieval by benchmarking LLMs on 16 chess variants structurally absent from pretraining corpora. Our evaluation of 20 models across 16,400 games yields four principal findings.

First, most LLMs play chess variants at or near random level. The grand mean score across all models and variants is 53.8%, with 14 of 20 models averaging below 55.0. This deficit persists even on Standard chess (where training data is abundant), confirming that poor performance reflects a genuine reasoning gap rather than data scarcity.

Second, the few models that do exceed random play show highly uneven transfer across variant categories. Grok 4 Fast (mean 74.5) dominates topological variants but scores 55.0 on Standard and 50.0 on Raumschach. Gemini 3.1 Pro (mean 63.1) excels at Racing Kings and Fog of War but matches random on Standard, Raumschach, and Kriegspiel. Strategic reasoning, as measured by GAMBIT, is not a single capability but a family of skills that transfer selectively.

Third, move quality analysis reveals that LLMs do learn partial strategic signal: their top-5 move rates consistently exceed random baselines, even when this advantage does not translate into

wins. The gap between per-move quality and game-level outcomes highlights the compounding difficulty of sustained multi-step reasoning.

Fourth, LLM-generated chess engines dramatically outperform the same models' prompt-based play, with codegen scores reaching 100.0 on variants where direct play hovers near 50.0. This codegen paradox suggests that LLMs possess latent strategic knowledge that is more effectively elicited through structured code generation than through sequential conversational reasoning.

As frontier models continue to advance, GAMBIT provides a durable framework for tracking genuine progress in novel-domain reasoning and strategic planning, with the guarantee that strong performance cannot be attributed to memorization.

References

[1] Bakhtin, A., Brown, N., Dinan, E., et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067–1074.

[2] Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, 134(1–2), 57–83.

[3] Carlini, N. (2024). Playing chess with large language models. *nicholas.carlini.com/writing*.

[4] Chollet, F. (2019). On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*.

[5] Chollet, F., Knoop, M., Kamradt, G., Landers, B., & Pinkard, H. (2025). ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2601.10904*.

[6] Glazer, E., et al. (2024). FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*.

[7] Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring massive multitask language understanding. *ICLR 2021*.

[8] Li, F., et al. (2024). Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the StepGame benchmark. *AAAI 2024*.

[9] Mirzaee, R. & Kordjamshidi, P. (2022). SpartQA: A textual question answering benchmark for spatial reasoning. *NAACL 2022*.

[10] Mirzaee, R. & Kordjamshidi, P. (2024). Reframing spatial reasoning evaluation in language models. *IJCAI 2024*.

[11] Montgomery, K., et al. (2025). LLM Chess: Benchmarking reasoning and instruction-following in LLMs through chess. *arXiv preprint arXiv:2512.01992*.

[12] Perolat, J., De Vylder, B., Hennes, D., et al. (2022). Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623), 990–996.

[13] Rein, D., et al. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*.

[14] Shannon, C. E. (1950). Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314), 256–275.

[15] Shi, F., et al. (2022). StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. *AAAI 2022*.

[16] Silver, D., Hubert, T., Schrittwieser, J., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.

[17] Srivastava, A., Rastogi, A., Rao, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR 2023*.

[18] Suzgun, M., Scales, N., Schärli, N., et al. (2023). Challenging BIG-bench tasks and whether chain-of-thought can solve them. *ACL 2023 Findings*.

[19] Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2023). PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *NeurIPS 2023 Datasets and Benchmarks Track*.

[20] White, C., Dooley, S., Roberts, M., Pal, A., & Goldblum, M. (2024). LiveBench: A challenging, contamination-limited LLM benchmark. *arXiv preprint arXiv:2406.19314*.

[21] Anthropic. (2024). Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet. *Anthropic Model Card*, October 2024. https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf.

[22] Anthropic. (2025). System Card: Claude Opus 4 & Claude Sonnet 4. *Anthropic System Card*, May 2025. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

[23] OpenAI. (2024). GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

[24] OpenAI. (2025). Introducing GPT-4.1 in the API. *Official blog post*, April 2025. https://openai.com/index/gpt-4-1/.

[25] OpenAI. (2025). OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*.

[26] OpenAI. (2025). OpenAI o3 and o4-mini System Card. *OpenAI System Card*, April 2025. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[27] Gemini Team & Google DeepMind. (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

[28] Gemma Team & Google DeepMind. (2025). Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

[29] DeepSeek-AI. (2024). DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

[30] Meta AI. (2025). The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. *Official blog post*, April 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[31] Yang, A., Yang, B., Zhang, B., Hui, B., et al. (2024). Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

[32] Abdin, M., Aneja, J., Behl, H., Bubeck, S., et al. (2024). Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.

[33] xAI. (2025). Grok 3 Beta — The Age of Reasoning Agents. *Official announcement*, February 2025. https://x.ai/news/grok-3.

[34] xAI. (2025). Grok 4 Model Card. *xAI Model Card*, August 2025. https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

[35] Amazon. (2025). The Amazon Nova Family of Models: Technical Report and Model Card. *arXiv preprint arXiv:2506.12103*.

[36] NVIDIA. (2025). Llama-Nemotron: Efficient Reasoning Models. *arXiv preprint arXiv:2505.00949*.

[37] GLM Team, Zeng, A., Xu, B., et al. (2024). ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.

[38] ERNIE Team, Baidu. (2025). ERNIE 4.5 Technical Report. https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf.

[39] Chen, A., Li, A., Gong, B., et al. (2025). MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. *arXiv preprint arXiv:2506.13585*.

[40] Singh, V., Krauss, L., Jaghouar, S., et al. (2026). Arcee Trinity Large Technical Report. *arXiv preprint arXiv:2602.17004*.

[41] Chen, M., Tworek, J., Jun, H., Yuan, Q., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.

[42] Chen, W., Ma, X., Wang, X., & Cohen, W. W. (2023). Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research (TMLR)*, 2023. *arXiv preprint arXiv:2211.12588*.

[43] Berlekamp, E. R., Conway, J. H., & Guy, R. K. (2001). *Winning Ways for Your Mathematical Plays* (2nd ed., Vols. 1–4). A K Peters.

[44] Li, M. & Vitányi, P. (2019). *An Introduction to Kolmogorov Complexity and Its Applications* (4th ed.). Springer.

[45] Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

[46] Maack, F. (1907). *Raumschach*. Hamburg. Described in Pritchard, D. B. (2007), *The Classified Encyclopedia of Chess Variants*, John Beasley.

[47] Parton, V. R. (1953). Alice Chess. *Fairy Chess Review*, 8(5).

[48] Gliński, W. (1936). Hexagonal Chess. Popularized in Gliński, W. (1973), *First Book of Hexagonal Chess*, Hexagonal Chess Publications.

[49] Pritchard, D. B. (2007). *The Classified Encyclopedia of Chess Variants*. John Beasley.

A Board Serialization Format

Board positions are serialized as text for LLM consumption using variant-specific coordinate systems adapted to each board geometry (e.g., file-rank-level triples for 3D variants, axial coordinates for hexagonal, board-prefixed coordinates for multi-board variants):

```
White: Ba1B, Kc1A, Nc3A, Pa2A, Pa3B, Qc1B, Ra1A, Re1A, Ub1B, Ue1B
Black: Ba5D, Bd5D, Kc5E, Nc3E, Nd5E, Pa4D, Pa4E, Qc5D, Ra5E, Re5E,
       Ub5D, Ue5D
```

Legal moves are presented as a numbered list:

```
1. Nc3A->b5B
2. Nc3A->a4A
3. Pa3B->a4B
4. Qc1B->d2C
...
```

The model responds with a single integer selecting a move by number.

B Token Usage and Evaluation Cost

Table 7 reports the average number of prompt and completion tokens consumed per game for each (model, variant) pair. Input token counts reflect the cumulative prompt size across all LLM moves in a game (system prompt plus serialized board state and legal move list at each turn); output tokens are minimal since the model responds with a single move index.

Table 8 translates these token counts into USD costs using OpenRouter's listed per-token pricing at the time of evaluation.

C Statistical Power Analysis

The reliability of LLM performance measurements varies dramatically across variants due to differences in outcome distributions. Variants where most games end in draws (e.g., Standard, Raumschach) produce low per-game score variance and require few trials to detect skill differences. Variants with high decisiveness (e.g., V-Cylinder, Projective) produce noisy score distributions and require substantially more trials.

Table 9 quantifies this by computing the per-game score variance $\mathrm{Var}(X)$ for each variant, where the per-game score X takes values 1 (win), 0.5 (draw), or 0 (loss). The variance determines the standard error at any sample size via $\mathrm{SE} = \sqrt{\mathrm{Var}(X)/N}$, and the final column reports the N required to detect a 5 percentage-point lift from the 50.0 baseline with 80% power ($\alpha = 0.05$, two-sided).

D Model Characteristics vs Performance

To investigate what model properties predict chess variant performance, we correlate each model's average LLM-vs-random score (across all tested variants) with three external characteristics: general intelligence ranking (Figure 5), model size (Figure 6), and API cost (Figure 7).

E Performance Gradient Analysis in PCA Space

To quantify how each model's performance varies with game complexity, we fit an ordinary least squares (OLS) regression for each of the 20 models:

$$\text{Score}_v = \beta_0 + \beta_1 \cdot \text{PC1}_v + \beta_2 \cdot \text{PC2}_v + \varepsilon_v$$

where v indexes the 16 variants, Score_v is the model's LLM-vs-random score on variant v, and PC1_v, PC2_v are the variant's coordinates in the principal component space derived from the complexity metrics (Section 3.5). The coefficient β_1 captures the model's sensitivity to the dominant complexity axis: a positive β_1 indicates improving performance on high-PC1 variants (chaotic, short, capture-heavy), while a negative β_1 indicates improving performance on low-PC1 variants (structured, long, draw-prone).

Statistical testing. For each model, we test $H_0 : \beta_1 = 0$ using a two-sided t-test:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}, \quad \text{SE}(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_v \text{PC1}_v^2}}, \quad \text{MSE} = \frac{\text{SS}_{\text{res}}}{n-3}$$

where $n = 16$ and 3 is the number of estimated parameters (intercept, β_1, β_2). P-values are computed from the t-distribution with $n - 3 = 13$ degrees of freedom.

Multiple testing correction. Because we test 20 models, we apply both Bonferroni correction ($\alpha_{\text{adj}} = 0.05/20$) and Benjamini-Hochberg FDR control at $\alpha = 0.10$. The smallest Bonferroni-corrected p-value is 0.42; no model survives either correction.

Results. Figure 10 summarizes the β_1 coefficients for all 20 models, and Figure 11 visualizes the corresponding performance gradients in PCA space. At the nominal $p < 0.05$ threshold, 2 of 20 models show significant β_1 coefficients. These three models also have the highest R^2 values in the dataset (0.31 to 0.39), indicating that their scores track PC1 consistently across variants, not merely on average. Models with larger absolute β_1 values but lower R^2 (e.g., Grok 4 Fast: $\beta_1 = +3.68$, $R^2 = 0.16$) do not reach significance because their high residual variance inflates the standard error.

The β_2 coefficient is not significant for any model ($p > 0.10$ for all), indicating that model differentiation occurs along the PC1 axis only.

Interpretation. These results suggest a possible complementary structure: some models perform relatively better on chaotic, decisive variants while others perform relatively better on structured, positional variants (see Figures 12 and 13 for representative pairwise comparisons). However, because no result survives multiple testing correction, this pattern should be treated as hypothesis-generating rather than confirmatory. Evaluation on a larger variant set would be needed to confirm whether complexity-dependent performance profiles are a genuine property of LLM architectures.

F Code and Data Availability

All source code, game engines, evaluation harness, and analysis scripts are publicly available at https://github.com/Max-Highsmith/faechess. All 16 chess variants are playable in-browser (human vs. human, human vs. AI, or AI vs. AI) at https://fae-chess.com.

G Reproducibility

All code is open source. The evaluation framework supports any model available through OpenRouter. Full supplementary figures and tables are provided in Appendix H.

```
git clone https://github.com/Max-Highsmith/faechess.git && cd faechess
npm install
echo "OPENROUTER_API_KEY=your_key" > .env

# Run LLM vs Random evaluation (30 trials, two variants, three models)
npm run eval -- --variants raumschach,torus \
  --models haiku,gpt4o-mini,gemini-flash-20 --trials 30

# Run codegen evaluation
node experiments/run-codegen-eval.mjs \
  --models haiku,gpt4o-mini --variants raumschach,torus

# Run complexity metrics (500 random self-play games per variant)
npm run complexity
```

H Supplementary Figures and Tables

Variant	Origin	Geometry	State	Novel Pieces	Cognitive Demand
Standard					
Standard	∼600 AD	8×8 plane	64 sq.	None	Baseline
3D					
Raumschach	[46]	5×5×5 cube	125 sq.	Unicorn	3D spatial reasoning
Torus 3D	Original	5×5×5 torus	125 (wrap)	Unicorn	3D + toroidal topology
Projective 3D	Original	5×5×5 proj.	125 (ident.)	Unicorn	3D + non-orientable
Topology					
Torus	[49]	8×8 torus	64 (wrap)	None	Wrap-around reasoning
Cylinder	[49]	8×8 cyl.	64 (wrap)	None	Horizontal wrap
V-Cylinder	[49]	8×8 v-cyl.	64 (wrap)	None	Vertical wrap
Klein	Original	8×8 Klein	64 (ident.)	None	Non-orientable surface
Möbius	Original	8×8 Möbius	64 (ident.)	None	Non-orientable strip
Projective	Original	8×8 proj.	64 (ident.)	None	Double identification
Rule Variants					
Alice	[47]	2 × 8×8	128 sq.	None (teleport)	Cross-board state
Racing Kings	[49]	8×8 plane	64 sq.	None	Race without captures
Crazyhouse	[49]	8×8 plane	64 + reserve	None (drops)	Piece-drop tactics
Imperfect Information					
Fog of War	Original	8×8 plane	64 (hidden)	None	Partial observability
Kriegspiel	[49]	8×8 plane	64 (hidden)	None	Blind play, inference
Tile Geometry					
Hexagonal	[48]	91-hex board	91 hex	None	Hexagonal adjacency

Table 1: Summary of chess variants implemented in GAMBIT, grouped by category.

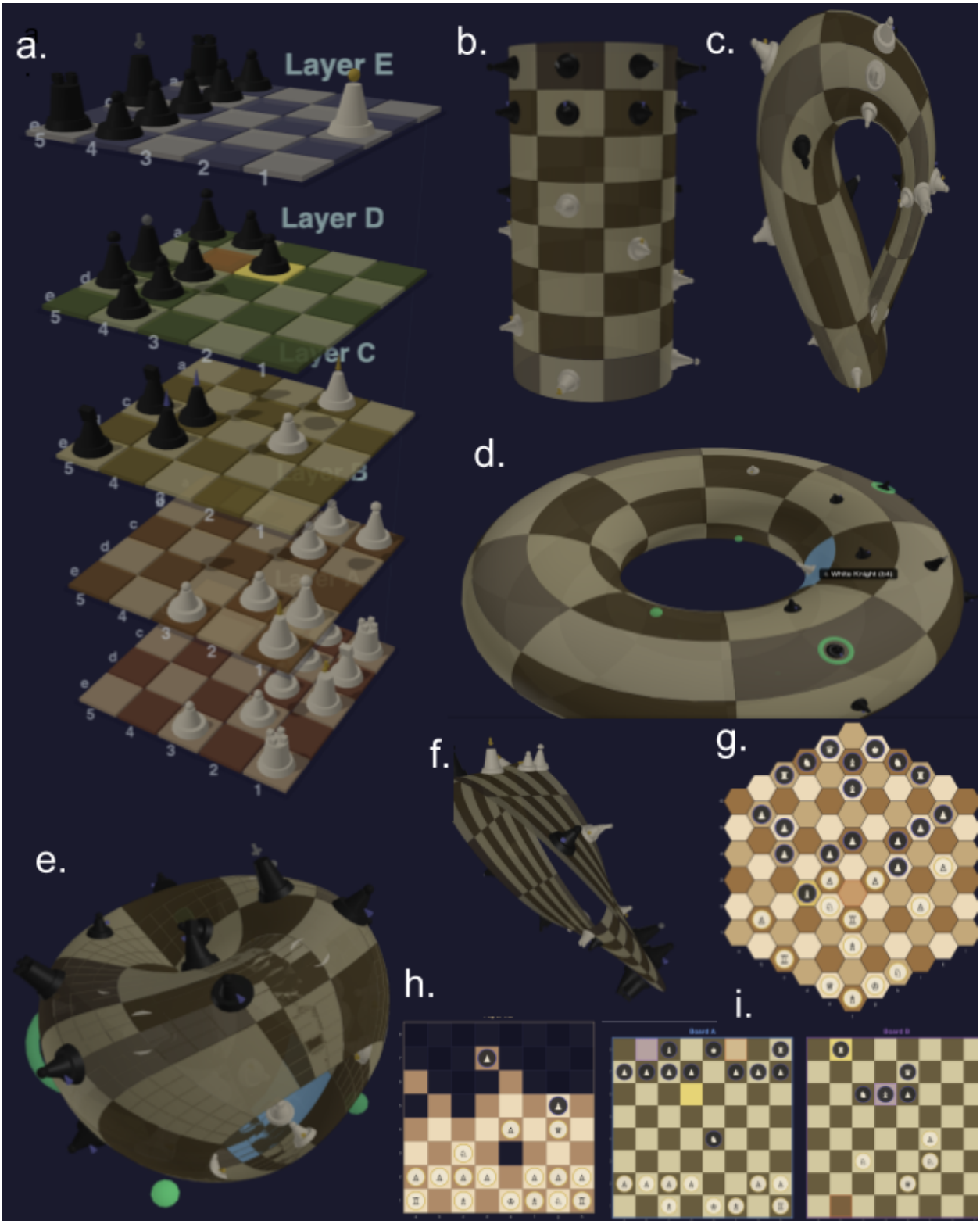

Figure 1: Representative board renderings from the GAMBIT variant suite. (a) Raumschach, (b) Cylinder, (c) Klein Bottle, (d) Torus, (e) Projective Plane, (f) Möbius, (g) Hexagonal, (h) Fog of War, (i) Alice Chess. Not shown: Torus 3D, Projective 3D, V-Cylinder, Standard, Racing Kings, Crazyhouse, Kriegspiel.

Variant	Rand.	Gem. Pro N=5	GPT-5.4 N=5	Opus 4.7 N=5	Grok 4 N=30	GPT-4.1m N=30	GPT-4o Mini N=30	DS V3 N=30	Llama 4 N=30	Gem. Flash N=30	Phi-4 N=30	Arcee N=30	Scout N=30	GLM N=30	Nova N=30	Qwen 72B N=30	MiniMax N=30	ERNIE N=30	Nemotron N=30	Gemma N=30	Haiku 3.5 N=30
Standard																					
Standard	50.3	**50.0**	**50.0**	**50.0**	**55.0**	**50.0**	**55.0**	**53.3**	**50.0**	**50.0**	**50.0**	**48.3**	**48.3**	**50.0**	**53.3**	**48.3**	**48.3**	**50.0**	**48.3**	**50.0**	**50.0**
3D																					
Projective 3D	50.8	**50.0**	**60.0**	**60.0**	**88.3**	**51.7**	**58.3**	**60.0**	**50.0**	**55.0**	**51.7**	**53.3**	**53.3**	**45.0**	**60.0**	**56.7**	**48.3**	**51.7**	**45.0**	**55.0**	**51.7**
Raumschach	50.3	**50.0**	**50.0**	**50.0**	**50.0**	**48.3**	**48.3**	**50.0**	**50.0**	**50.0**	**50.0**	**48.3**	**51.7**	**51.7**	**50.0**	**50.0**	**50.0**	**48.3**	**50.0**	**45.0**	**48.3**
Torus 3D	50.1	**50.0**	**70.0**	**70.0**	**63.3**	**43.3**	**61.7**	**50.0**	**55.0**	**56.7**	**45.0**	**51.7**	**58.3**	**53.3**	**51.7**	**53.3**	**53.3**	**55.0**	**53.3**	**40.0**	**41.7**
Topology																					
Cylinder	50.4	**50.0**	**50.0**	**50.0**	**85.0**	**55.0**	**46.7**	**50.0**	**50.0**	**50.0**	**50.0**	**51.7**	**48.3**	**50.0**	**50.0**	**51.7**	**50.0**	**50.0**	**48.3**	**50.0**	**51.7**
Klein	52.1	**40.0**	**50.0**	**40.0**	**98.3**	**66.7**	**58.3**	**60.0**	**65.0**	**43.3**	**53.3**	**48.3**	**56.7**	**55.0**	**53.3**	**46.7**	**51.7**	**61.7**	**51.7**	**48.3**	**38.3**
Möbius	49.5	**60.0**	**40.0**	**60.0**	**90.0**	**58.3**	**66.7**	**53.3**	**61.7**	**51.7**	**48.3**	**50.0**	**66.7**	**51.7**	**48.3**	**60.0**	**46.7**	**53.3**	**43.3**	**48.3**	**48.3**
Projective	49.4	**60.0**	**20.0**	**30.0**	**96.7**	**61.7**	**66.7**	**48.3**	**65.0**	**36.7**	**60.0**	**55.0**	**53.3**	**65.0**	**35.0**	**35.0**	**45.0**	**33.3**	**50.0**	**48.3**	**45.0**
Torus	49.8	**60.0**	**60.0**	**70.0**	**50.0**	**66.7**	**70.0**	**50.0**	**50.0**	**50.0**	**53.3**	**55.0**	**48.3**	**56.7**	**53.3**	**50.0**	**60.0**	**48.3**	**43.3**	**48.3**	**43.3**
V-Cylinder	49.6	**80.0**	**50.0**	**20.0**	**83.3**	**51.7**	**61.7**	**40.0**	**50.0**	**41.7**	**60.0**	**56.7**	**46.7**	**45.0**	**43.3**	**41.7**	**45.0**	**48.3**	**55.0**	**41.7**	**43.3**
Rule Variant																					
Alice	48.2	**80.0**	**50.0**	**60.0**	**60.0**	**63.3**	**43.3**	**53.3**	**45.0**	**46.7**	**43.3**	**51.7**	**41.7**	**50.0**	**33.3**	**45.0**	**43.3**	**51.7**	**50.0**	**43.3**	**43.3**
Crazyhouse	49.3	**70.0**	**70.0**	**50.0**	**56.7**	**53.3**	**51.7**	**51.7**	**51.7**	**53.3**	**51.7**	**48.3**	**51.7**	**48.3**	**48.3**	**48.3**	**50.0**	**50.0**	**51.7**	**50.0**	**50.0**
Racing Kings	49.9	**100.0**	**90.0**	**50.0**	**53.3**	**83.3**	**55.0**	**80.0**	**58.3**	**96.7**	**55.0**	**50.0**	**55.0**	**50.0**	**71.7**	**61.7**	**50.0**	**51.7**	**50.0**	**68.3**	**71.7**
Imperfect Info																					
Fog of War	49.3	**100.0**	**100.0**	**90.0**	**100.0**	**61.7**	**45.0**	**81.7**	**53.3**	**58.3**	**58.3**	**53.3**	**45.0**	**50.0**	**60.0**	**60.0**	**60.0**	**40.0**	**50.0**	**51.7**	**61.7**
Kriegspiel	50.2	**50.0**	**50.0**	**50.0**	**88.3**	**53.3**	**48.3**	**50.0**	**50.0**	**51.7**	**51.7**	**53.3**	**50.0**	**48.3**	**51.7**	**48.3**	**50.0**	**48.3**	**50.0**	**50.0**	**45.0**
Tile Geometry																					
Hexagonal	50.0	**60.0**	**50.0**	**50.0**	**73.3**	**50.0**	**50.0**	**53.3**	**50.0**	**50.0**	**50.0**	**50.0**	**50.0**	**50.0**	**50.0**	**50.0**	**51.7**	**50.0**	**50.0**	**50.0**	**50.0**
Mean		**63.1**	**56.9**	**53.1**	**74.5**	**57.4**	**55.4**	**55.3**	**53.4**	**52.6**	**52.0**	**51.6**	**51.6**	**51.2**	**50.8**	**50.4**	**50.2**	**49.5**	**49.4**	**49.3**	**49.0**

Table 2: LLM vs Random performance (Score). Score = (W + D×0.5) / N × 100; 50.0 = random-level play. N (shown per model column) = games per cell, split evenly as white and black. Max plies = 80. Cell shading: sig. above 50 ($p < 0.05$, BH-corrected); above 50 ; sig. below 50 ; below 50 .

Variant	Rand.	Gem. Pro	GPT-5.4	Opus 4.7	Grok 4	DS V3	Haiku 3.5	Gem. Flash	GPT-4o Mini	GPT-4.1m	Nova	Qwen 72B	Phi-4	ERNIE	GLM	MiniMax	Gemma	Arcee	Llama 4	Scout	Nemotron	All
Standard																						
Standard	*17.5*	24.6	33.9	20	72.1	23	22.5	21.8	25	22.5	20	23.1	21.5	23	23.7	20.4	17.5	17.5	21	20	16.9	22.6
3D																						
Projective 3D (d=1)	*12.4*	21.5	16	18.9	40.8	13.3	17	22.5	18.3	12.5	17.8	16	15	17.9	10.8	17.4	18.3	16.2	16.1	16.9	11.1	17
Raumschach (d=1)	*9.7*	37	27	11.5	23.5	17.2	11.3	11.5	10.8	9.5	11.5	9.3	6.8	11.1	10.3	9.5	8	10	8.3	8.5	9.5	11.4
Torus 3D (d=1)	*12.2*	17	22.1	15.8	67.1	16.6	14.2	19.8	19.9	16.6	19	19.5	22.8	15.8	15	14.2	19.4	11.8	19.5	20.7	16.3	18.2
Topology																						
Cylinder	*18*	33	33.5	18.5	57.7	22	25.3	22	24.8	22.1	23	22.5	16.8	16.5	20.8	18.7	18.8	18.5	16	14.2	16.3	22
Klein	*24*	32.9	30.9	20.4	56.4	25.7	27.4	25.1	24.8	22.2	27.8	23.1	21	28.3	24.1	20.8	25.9	23.5	20.2	20.5	27.4	25.3
Möbius	*23*	28.8	27.1	22.1	65.2	24.2	26.3	18.5	30.5	26.3	20.5	26.6	25.3	27.5	24.2	26.3	17.5	27.7	21.1	21.7	22.7	25
Projective	*23.9*	33.6	28.4	24.7	54.3	25.1	31.1	24.4	22.9	23.9	26.9	14.7	25.4	20.2	29.7	22.3	23.9	19.5	19.8	23.1	21.8	24.7
Torus	*22.6*	46.4	37.1	22.4	33.3	39.1	29.8	33.8	25	24.1	22.6	25.3	22	22.6	21.3	21.5	24.6	27.2	22.6	23.3	21.7	26.1
V-Cylinder	*23.9*	34.8	28.9	20.3	62.2	25.2	29.3	25	21.9	23.8	24.6	23.5	18.9	25.3	22.9	25.9	27	28.7	20.1	20.3	25.3	25.7
Imperfect Info																						
Fog of War	*15.8*	31.4	45.3	24.8	61.6	28.4	25.5	17.2	17.1	22.2	24.9	26.3	19.8	14.7	16.7	17.5	15.3	16.2	17.7	15.5	14.7	20.1
Kriegspiel	*18.6*	27.5	30.5	28	67.8	27	26.8	21.8	22.3	24.8	20.4	24.8	25	17.1	21.2	23	15	17.5	20.3	16.3	19.3	24
Tile Geometry																						
Hexagonal	*9.5*	34.9	39.5	20	48.5	28.2	14.5	32.8	15	26.3	16	13.5	14.8	14.2	9.3	11	15.5	8.5	15.3	15.5	12	19.1
Average	*17.8*	**30.5**	**30.1**	**20.4**	**57.8**	**23.5**	**22.6**	**22.8**	**21**	**21.1**	**20.9**	**20.3**	**19.4**	**19.5**	**18.9**	**18.9**	**18.7**	**18.3**	**18**	**17.9**	**17.8**	**21.4**

Table 3: Move quality: percentage of LLM moves ranked in the top 5 by depth-2 minimax (higher = better). The "Rand." column shows the expected top-5 rate under uniform random play, computed as $\mathrm{E}[\min(5, B)/B]$. Racing Kings excluded (positional win condition invalidates material-based evaluator). Based on 2405 games across 20 models and 13 variants.

Variant	Prompt Avg	Gem. Pro Codegen	GPT-5.4 Codegen	Opus 4.7 Codegen	Grok 4 Codegen	DS V3 Codegen	Llama 4 Codegen	ERNIE Codegen	Haiku 3.5 Codegen	MiniMax Codegen	GLM Codegen	Gem. Flash Codegen	GPT-4.1m Codegen	Qwen 72B Codegen	Nemotron Codegen	GPT-4o Mini Codegen	Scout Codegen	Arcee Codegen	Nova Codegen	Gemma Codegen	Phi-4 Codegen	N
Standard																						
Standard	50.5	**100.0**	**100.0**	**100.0**	**55.0**	**68.3**	**58.3**	**FAIL**	**63.3**	**51.7**	**50.0**	**51.7**	**55.0**	**53.3**	**70.0**	**51.7**	**56.7**	**51.7**	**53.3**	**51.7**	**55.0**	30
3D																						
Projective 3D	55.0	**100.0**	**85.0**	**86.7**	**80.0**	**FAIL**	**98.3**	**95.0**	**80.0**	**85.0**	**88.3**	**76.7**	**81.7**	**66.7**	**73.3**	**41.7**	**31.7**	**83.3**	**FAIL**	**75.0**	**45.0**	30
Raumschach	49.4	**98.3**	**100.0**	**100.0**	**98.3**	**51.7**	**53.3**	**48.3**	**48.3**	**61.7**	**50.0**	**51.7**	**53.3**	**53.3**	**53.3**	**48.3**	**48.3**	**55.0**	**51.7**	**53.3**	**51.7**	30
Torus 3D	52.2	**96.7**	**93.3**	**66.7**	**66.7**	**56.7**	**65.0**	**76.7**	**53.3**	**48.3**	**58.3**	**76.7**	**53.3**	**73.3**	**75.0**	**53.3**	**83.3**	**FAIL**	**75.0**	**71.7**	**60.0**	30
Topology																						
Cylinder	52.3	**100.0**	**96.7**	**100.0**	**65.0**	**65.0**	**55.0**	**FAIL**	**66.7**	**51.7**	**56.7**	**53.3**	**50.0**	**55.0**	**63.3**	**53.3**	**75.0**	**56.7**	**60.0**	**55.0**	**66.7**	30
Klein	56.3	**90.0**	**95.0**	**73.3**	**85.0**	**90.0**	**60.0**	**FAIL**	**78.3**	**75.0**	**66.7**	**83.3**	**83.3**	**70.0**	**73.3**	**71.7**	**46.7**	**51.7**	**45.0**	**45.0**	**63.3**	30
Möbius	55.7	**100.0**	**76.7**	**95.0**	**85.0**	**55.0**	**56.7**	**76.7**	**60.0**	**66.7**	**55.0**	**55.0**	**61.7**	**75.0**	**68.3**	**80.0**	**25.0**	**35.0**	**FAIL**	**61.7**	**53.3**	30
Projective	52.9	**93.3**	**93.3**	**100.0**	**70.0**	**75.0**	**78.3**	**51.7**	**81.7**	**60.0**	**70.0**	**76.7**	**78.3**	**70.0**	**45.0**	**73.3**	**85.0**	**81.7**	**75.0**	**76.7**	**63.3**	30
Torus	52.7	**100.0**	**100.0**	**100.0**	**73.3**	**70.0**	**86.7**	**FAIL**	**86.7**	**43.3**	**58.3**	**78.3**	**65.0**	**75.0**	**75.0**	**63.3**	**71.7**	**90.0**	**63.3**	**60.0**	**58.3**	30
V-Cylinder	50.3	**96.7**	**96.7**	**100.0**	**100.0**	**56.7**	**66.7**	**FAIL**	**76.7**	**FAIL**	**FAIL**	**76.7**	**68.3**	**86.7**	**55.0**	**51.7**	**85.0**	**61.7**	**61.7**	**66.7**	**63.3**	30
Rule Variant																						
Alice	47.5	**63.3**	**33.3**	**60.0**	**48.3**	**FAIL**	**61.7**	**FAIL**	**18.3**	**41.7**	**FAIL**	**60.0**	**18.3**	**48.3**	**60.0**	**65.0**	**45.0**	**53.3**	**55.0**	**43.3**	**51.7**	30
Crazyhouse	51.0	**100.0**	**100.0**	**100.0**	**66.7**	**75.0**	**73.3**	**FAIL**	**71.7**	**61.7**	**70.0**	**46.7**	**71.7**	**60.0**	**56.7**	**56.7**	**41.7**	**61.7**	**53.3**	**53.3**	**56.7**	30
Racing Kings	62.5	**98.3**	**96.7**	**100.0**	**63.3**	**100.0**	**100.0**	**56.7**	**48.3**	**100.0**	**FAIL**	**33.3**	**FAIL**	**38.3**	**41.7**	**56.7**	**45.0**	**40.0**	**35.0**	**31.7**	**43.3**	30
Imperfect Info																						
Fog of War	58.2	**100.0**	**100.0**	**100.0**	**56.7**	**100.0**	**76.7**	**FAIL**	**100.0**	**100.0**	**100.0**	**100.0**	**90.0**	**41.7**	**48.3**	**100.0**	**100.0**	**28.3**	**70.0**	**55.0**	**36.7**	30
Kriegspiel	52.3	**86.7**	**100.0**	**100.0**	**56.7**	**48.3**	**56.7**	**FAIL**	**55.0**	**48.3**	**48.3**	**38.3**	**50.0**	**56.7**	**60.0**	**53.3**	**51.7**	**55.0**	**46.7**	**51.7**	**51.7**	30
Tile Geometry																						
Hexagonal	51.7	**66.7**	**80.0**	**50.0**	**63.3**	**FAIL**	**53.3**	**FAIL**	**50.0**	**66.7**	**60.0**	**60.0**	**48.3**	**50.0**	**53.3**	**50.0**	**55.0**	**66.7**	**51.7**	**56.7**	**50.0**	30
Mean		**93.1**	**90.4**	**89.5**	**70.8**	**70.1**	**68.8**	**67.5**	**64.9**	**64.1**	**64.0**	**63.6**	**61.9**	**60.8**	**60.7**	**60.6**	**59.2**	**58.1**	**56.9**	**56.8**	**54.4**	8970

Table 4: Codegen vs Random performance. Each model writes a JavaScript chess engine which then plays 30 games per variant against a random opponent. Score = (W + D×0.5) / N × 100; 50.0 = random-level play. "Prompt Avg" shows the average prompt-based score across the same models for comparison. †Codegen failed (fell back to random). ‡Clock forfeits (engine too slow). *Runtime fallbacks occurred. Cell shading: sig. above 50 ($p < 0.05$, BH-corrected); above 50 ; sig. below 50 ; below 50 .

Metric	Dimension	Definition
Branching	Computational Complexity	Average number of legal moves available per position
$\log_{10}$(Space)	Computational Complexity	Logarithm (base 10) of estimated game tree size
ThreatKingMob	Tactical Complexity	Average number of legal king moves per position
IrrevEvalSwing	Game Dynamism	Average material swing between consecutive positions
CapDrawPlies	Game Pacing	Average game length in half-moves

Table 5: Complexity metrics used in the PCA (Section 3.5). All 9 metrics are computed from 500 random self-play games per variant and included in the principal component analysis.

Variant	Branching	Threat%	Irrev%	EvalSwing	Cap%	MQSpread	Plies
Alice	29.2	18.1	18.6	0.00	0.0	0.00	127
Crazyhouse	47.4	47.2	32.8	1.13	19.5	0.00	179
Cylinder	31.6	37.4	25.8	0.38	12.2	0.00	189
Fog of War	33.8	38.3	31.1	0.31	12.6	0.00	112
Hexagonal	52.2	41.6	19.4	0.32	11.7	0.00	196
Klein	31.1	43.6	39.6	1.07	25.4	0.00	104
Kriegspiel	27.5	34.6	26.4	0.37	11.8	0.00	189
Möbius	24.4	35.3	25.1	0.44	12.1	0.00	165
Projective	32.9	43.7	36.9	1.02	24.2	0.00	110
Projective 3D	157.2	58.9	19.2	0.34	12.3	0.00	177
Racing Kings	30.6	37.5	5.5	0.24	5.5	0.00	191
Raumschach	73.4	44.9	24.8	0.41	11.4	0.00	195
Standard	27.6	34.6	26.3	0.37	11.7	0.00	190
Torus	37.0	42.8	36.5	0.98	23.9	0.00	129
Torus 3D	135.5	58.3	18.0	0.46	13.6	0.00	184
V-Cylinder	31.4	40.4	36.9	0.97	23.2	0.00	127

Table 6: Complexity metrics for all variants. Generated from 500 random self-play games per variant, 200-ply cap.

Variant	Arcee	DS V3	ERNIE	Gem. Flash	Gem. Pro	Gemma	GLM	GPT-4.1m	GPT-4o Mini	GPT-5.4	Grok 4	Haiku 3.5	Llama 4	MiniMax	Nemotron	Nova	Opus 4.7	Phi-4	Qwen 72B	Scout	Avg
Standard																					
Standard	63.2K	34.2K	36.9K	35.9K	60.5K	36.9K	65.9K	33.6K	34.2K	30.0K	8.9K	37.7K	34.5K	65.6K	69.4K	36.1K	46.9K	33.9K	37.7K	35.1K	43.1K
3D																					
Projective 3D	183.3K	115.4K	95.7K	109.8K	152.8K	103.3K	206.3K	99.2K	100.3K	97.0K	80.7K	98.2K	98.6K	194.8K	208.3K	105.7K	108.9K	96.9K	111.9K	94.6K	124.7K
Raumschach	226.2K	122.0K	128.0K	115.5K	167.9K	113.5K	236.0K	113.5K	119.3K	110.7K	91.2K	123.6K	128.1K	230.9K	251.1K	122.3K	153.9K	122.0K	118.9K	116.5K	147.8K
Torus 3D	176.4K	96.2K	120.7K	101.2K	180.7K	76.6K	177.6K	75.2K	101.6K	70.6K	64.7K	77.4K	101.4K	179.9K	182.7K	100.2K	122.4K	95.5K	97.9K	104.5K	115.0K
Topology																					
Cylinder	64.8K	31.2K	36.5K	32.5K	46.0K	35.3K	65.5K	32.4K	32.7K	31.7K	25.1K	36.1K	35.2K	64.2K	69.0K	35.7K	46.8K	33.5K	31.6K	32.0K	40.8K
Klein	64.0K	27.5K	28.2K	30.1K	36.4K	30.7K	61.0K	31.3K	31.4K	27.6K	17.3K	32.5K	30.6K	60.0K	60.3K	30.7K	48.5K	32.3K	28.1K	27.8K	36.8K
Möbius	60.6K	24.6K	31.1K	32.1K	31.6K	33.6K	61.3K	24.8K	19.7K	26.4K	14.9K	27.7K	27.3K	63.1K	57.5K	32.8K	40.9K	29.8K	28.5K	24.7K	34.9K
Projective	49.7K	25.1K	28.5K	28.8K	31.0K	27.3K	54.5K	23.8K	27.9K	20.2K	13.9K	28.0K	28.5K	50.0K	63.5K	27.6K	41.2K	28.4K	25.2K	26.0K	32.7K
Torus	60.5K	31.7K	31.0K	31.5K	38.3K	30.5K	57.0K	30.2K	30.7K	28.5K	-	30.7K	32.2K	60.4K	59.8K	30.5K	46.0K	30.2K	30.1K	30.9K	38.1K
V-Cylinder	53.7K	26.0K	31.2K	27.7K	35.4K	26.5K	48.7K	27.6K	24.7K	22.1K	16.2K	26.3K	29.7K	57.5K	59.5K	29.7K	33.9K	30.3K	28.2K	26.3K	33.8K
Rule Variant																					
Alice	61.6K	29.9K	36.2K	26.7K	38.6K	38.4K	60.3K	22.1K	32.7K	28.3K	4.0K	41.1K	33.5K	62.3K	61.4K	24.5K	25.1K	39.1K	33.0K	33.2K	38.9K
Crazyhouse	83.6K	43.9K	46.0K	43.0K	69.5K	44.9K	83.2K	45.6K	43.6K	42.1K	5.8K	47.6K	51.9K	86.4K	90.1K	43.5K	72.1K	48.7K	43.8K	42.1K	55.4K
Racing Kings	84.6K	27.0K	45.9K	17.0K	21.0K	33.3K	87.8K	29.3K	42.4K	22.5K	9.1K	36.8K	42.0K	86.6K	87.8K	32.6K	58.6K	42.3K	34.0K	42.6K	47.4K
Imperfect Info																					
Fog of War	61.3K	23.5K	34.1K	27.5K	12.3K	34.0K	65.5K	27.2K	30.1K	8.6K	7.8K	34.4K	31.4K	51.8K	65.6K	32.3K	27.9K	31.2K	28.6K	31.1K	35.8K
Kriegspiel	59.2K	29.5K	31.4K	30.8K	39.6K	32.2K	60.1K	29.6K	29.0K	28.7K	21.8K	32.0K	30.6K	60.2K	64.1K	33.0K	41.9K	30.1K	31.1K	31.3K	37.5K
Tile Geometry																					
Hexagonal	92.1K	49.6K	54.6K	49.5K	59.5K	50.5K	97.0K	50.4K	49.8K	42.2K	43.5K	53.6K	50.8K	95.8K	102.6K	51.3K	64.6K	49.5K	51.9K	48.3K	61.1K
Average	**90.3K**	**46.1K**	**45.9K**	**46.2K**	**63.8K**	**46.7K**	**93.0K**	**43.5K**	**46.9K**	**39.8K**	**31.2K**	**47.7K**	**49.1K**	**92.4K**	**97.0K**	**48.0K**	**61.2K**	**48.4K**	**47.5K**	**46.7K**	**57.5K**

Table 7: Average prompt tokens per game by model and variant. Based on 8066 total games across 20 models and 16 variants. Output tokens are minimal (<200 per game) as the model responds with a single move index.

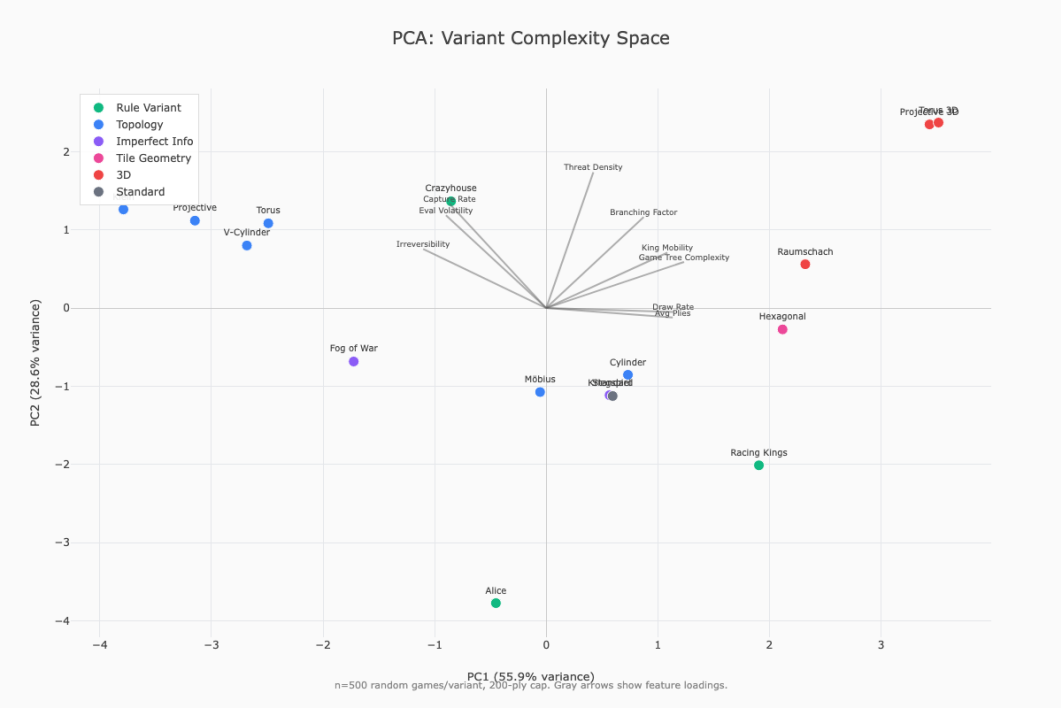

Figure 2: Principal Component Analysis of variant complexity metrics. The first two components explain 84.5% of variance. Gray arrows indicate feature loadings. Variants cluster by category, revealing structural similarities in complexity profiles. Based on 500 random games per variant, 200-ply cap.

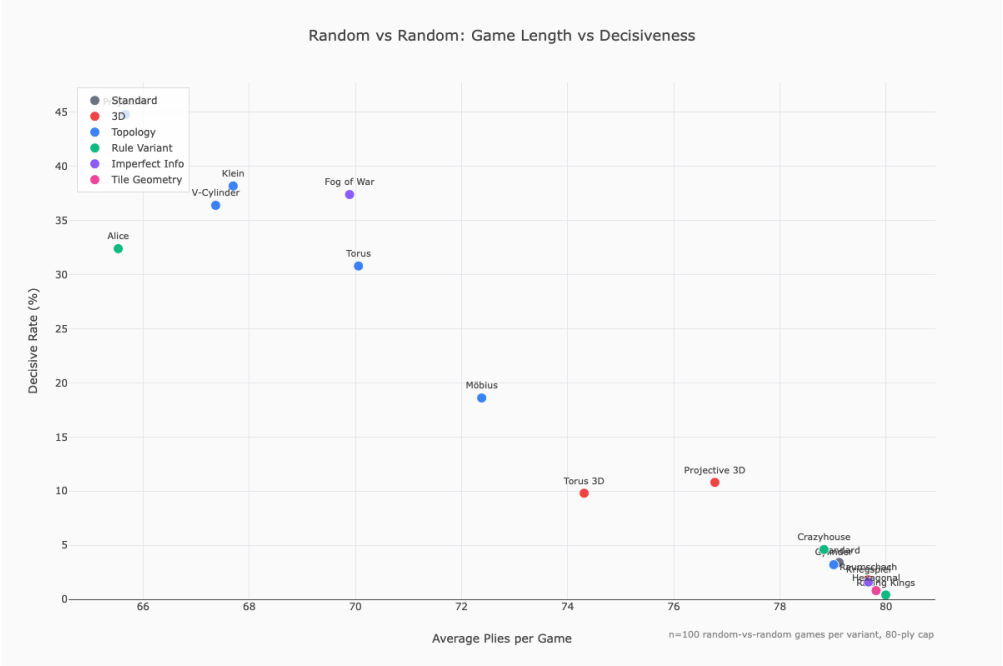

Figure 3: Game length vs decisiveness across variants. Each point represents one variant's average game length and decisive rate from 500 random-vs-random games (80-ply cap). Variants with shorter games tend to have higher decisive rates.

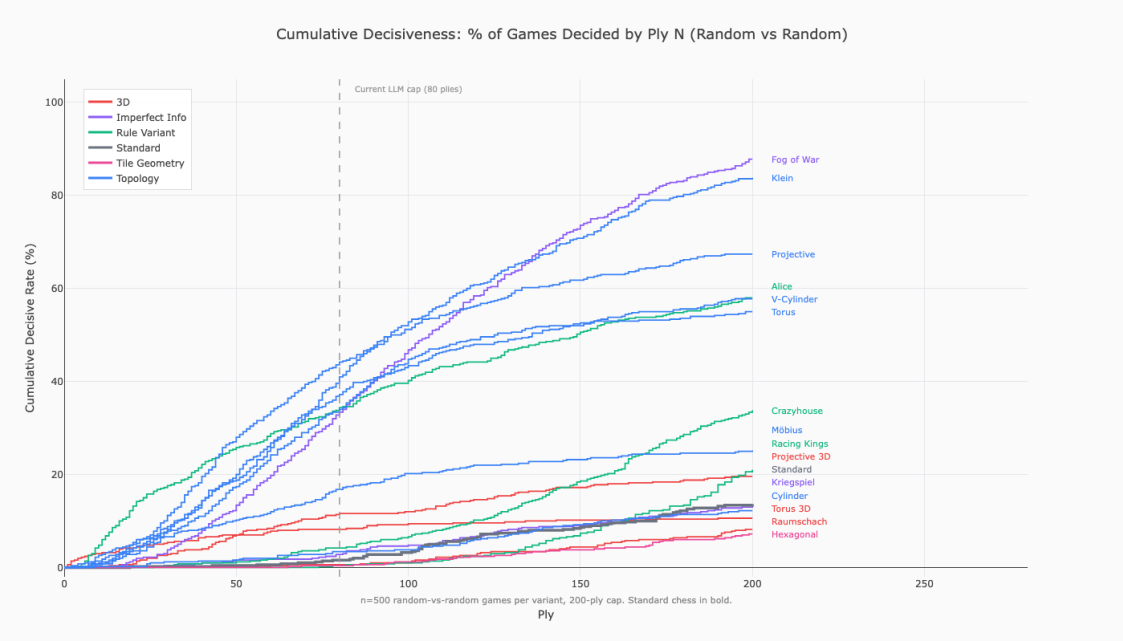

Figure 4: Cumulative decisiveness curves showing the percentage of games decided by ply N across variants. Each curve represents one variant from 500 random-vs-random games (200-ply cap). Steeper curves indicate faster game resolution. Standard chess (bold line) shows typical midgame decisiveness.

Variant	Arcee	DS V3	ERNIE	Gem. Flash	Gem. Pro	Gemma	GLM	GPT-4.1m	GPT-4o Mini	GPT-5.4	Grok 4	Haiku 3.5	Llama 4	MiniMax	Nemotron	Nova	Opus 4.7	Phi-4	Qwen 72B	Scout	Total
Standard																					
Standard	$0.0033	$0.0110	$0.0026	$0.0036	$0.1684	$0.0030	$0.0060	$0.0136	$0.0052	$0.0778	$0.0205	$0.0310	$0.0061	$0.0162	$0.0036	$0.0294	$0.2855	$0.0026	$0.0045	$0.0029	$7.04
3D																					
Projective 3D	$0.0081	$0.0370	$0.0067	$0.0110	$0.3437	$0.0083	$0.0144	$0.0398	$0.0151	$0.2455	$0.0837	$0.0794	$0.0161	$0.0281	$0.0091	$0.0850	$0.5521	$0.0071	$0.0135	$0.0083	$19.69
Raumschach	$0.0098	$0.0392	$0.0090	$0.0116	$0.3754	$0.0091	$0.0162	$0.0455	$0.0179	$0.2798	$0.1333	$0.0997	$0.0208	$0.0350	$0.0109	$0.0984	$0.8292	$0.0089	$0.0143	$0.0097	$23.12
Torus 3D	$0.0078	$0.0309	$0.0085	$0.0102	$0.4062	$0.0061	$0.0126	$0.0302	$0.0153	$0.1786	$0.1213	$0.0626	$0.0166	$0.0263	$0.0081	$0.0807	$0.6622	$0.0070	$0.0118	$0.0091	$18.35
Topology																					
Cylinder	$0.0034	$0.0103	$0.0026	$0.0033	$0.1306	$0.0028	$0.0060	$0.0131	$0.0049	$0.0823	$0.0895	$0.0297	$0.0068	$0.0142	$0.0036	$0.0350	$0.2952	$0.0027	$0.0038	$0.0032	$9.51
Klein	$0.0033	$0.0090	$0.0020	$0.0030	$0.1005	$0.0025	$0.0054	$0.0126	$0.0047	$0.0715	$0.0837	$0.0268	$0.0058	$0.0115	$0.0031	$0.0252	$0.3027	$0.0026	$0.0034	$0.0028	$8.58
Möbius	$0.0031	$0.0079	$0.0022	$0.0033	$0.0916	$0.0027	$0.0055	$0.0100	$0.0030	$0.0686	$0.0545	$0.0228	$0.0053	$0.0125	$0.0030	$0.0310	$0.2565	$0.0024	$0.0034	$0.0025	$7.32
Projective	$0.0026	$0.0081	$0.0020	$0.0029	$0.0874	$0.0022	$0.0050	$0.0096	$0.0042	$0.0524	$0.0508	$0.0230	$0.0055	$0.0099	$0.0033	$0.0234	$0.2587	$0.0023	$0.0030	$0.0026	$6.78
Torus	$0.0031	$0.0102	$0.0022	$0.0032	$0.1082	$0.0025	$0.0051	$0.0122	$0.0046	$0.0739	–	$0.0253	$0.0062	$0.0148	$0.0031	$0.0249	$0.2846	$0.0024	$0.0036	$0.0030	$6.11
V-Cylinder	$0.0028	$0.0085	$0.0022	$0.0028	$0.1015	$0.0021	$0.0045	$0.0111	$0.0037	$0.0576	$0.0903	$0.0217	$0.0058	$0.0115	$0.0031	$0.0254	$0.2154	$0.0024	$0.0034	$0.0026	$7.16
Rule Variant																					
Alice	$0.0031	$0.0097	$0.0026	$0.0027	$0.1063	$0.0031	$0.0052	$0.0089	$0.0049	$0.0731	$0.0062	$0.0337	$0.0064	$0.0140	$0.0031	$0.0200	$0.1501	$0.0031	$0.0040	$0.0030	$5.49
Crazyhouse	$0.0041	$0.0142	$0.0033	$0.0043	$0.1728	$0.0036	$0.0070	$0.0184	$0.0066	$0.1077	$0.0056	$0.0389	$0.0090	$0.0196	$0.0044	$0.0353	$0.4174	$0.0038	$0.0053	$0.0036	$8.91
Racing Kings	$0.0042	$0.0088	$0.0032	$0.0017	$0.0562	$0.0027	$0.0073	$0.0118	$0.0064	$0.0580	$0.0129	$0.0301	$0.0076	$0.0187	$0.0043	$0.0265	$0.3571	$0.0033	$0.0041	$0.0037	$6.60
Imperfect Info																					
Fog of War	$0.0031	$0.0076	$0.0024	$0.0028	$0.0345	$0.0027	$0.0059	$0.0110	$0.0046	$0.0222	$0.0171	$0.0283	$0.0061	$0.0134	$0.0034	$0.0279	$0.1685	$0.0025	$0.0035	$0.0031	$5.46
Kriegspiel	$0.0031	$0.0095	$0.0022	$0.0031	$0.1146	$0.0026	$0.0056	$0.0120	$0.0044	$0.0747	$0.0777	$0.0264	$0.0058	$0.0169	$0.0034	$0.0287	$0.2591	$0.0025	$0.0038	$0.0027	$8.52
Tile Geometry																					
Hexagonal	$0.0045	$0.0160	$0.0038	$0.0050	$0.1520	$0.0041	$0.0079	$0.0203	$0.0075	$0.1086	$0.1328	$0.0437	$0.0092	$0.0166	$0.0049	$0.0415	$0.3803	$0.0038	$0.0063	$0.0045	$13.12
Total	**$2.08**	**$7.13**	**$1.07**	**$2.24**	**$12.75**	**$1.80**	**$3.58**	**$8.40**	**$3.39**	**$8.16**	**$23.83**	**$18.69**	**$4.17**	**$8.17**	**$2.23**	**$19.15**	**$28.37**	**$1.78**	**$2.75**	**$2.02**	**$161.76**

Table 8: Average cost per game (USD) by model and variant, computed from OpenRouter token pricing. Total column shows cumulative spend per variant across all models.

Variant	W	D	L	N	Score	Var(X)	Decis.%	N_{5pp}	Uncorrected Min@30	Uncorrected Min@5	Bonferroni Min@30	Bonferroni Min@5
STANDARD												
Standard	9	485	8	500	50.3	0.0085	3%	27	53.3	58.1	55.0	62.2
3D												
Projective 3D	31	446	23	500	50.8	0.0269	11%	85	55.9	64.4	58.9	71.7
Raumschach	6	491	3	500	50.3	0.0045	2%	15	52.4	55.9	53.6	58.9
Torus 3D	25	451	24	500	50.1	0.0245	10%	77	55.6	63.7	58.4	70.7
TOPOLOGY												
Cylinder	10	484	6	500	50.4	0.0080	3%	26	53.2	57.8	54.8	61.8
Klein	106	309	85	500	52.1	0.0951	38%	299	61.0	77.0	66.6	90.8
Möbius	44	407	49	500	49.5	0.0465	19%	146	57.7	68.9	61.6	78.5
Projective	109	276	115	500	49.4	0.1120	45%	352	62.0	79.3	68.1	94.2
Torus	76	346	78	500	49.8	0.0770	31%	242	59.9	74.3	65.0	86.7
V-Cylinder	89	318	93	500	49.6	0.0910	36%	286	60.8	76.4	66.3	89.9
RULE VARIANT												
Alice	72	338	90	500	48.2	0.0807	32%	254	60.2	74.9	65.3	87.5
Crazyhouse	8	477	15	500	49.3	0.0115	5%	36	53.8	59.4	55.8	64.1
Racing Kings	1	499	1	500	50.1	0.0010	0%	4	51.1	52.8	51.7	54.2
IMPERFECT INFO												
Fog of War	90	313	97	500	49.3	0.0935	37%	294	60.9	76.8	66.5	90.4
Kriegspiel	5	492	3	500	50.2	0.0040	2%	13	52.3	55.5	53.4	58.4
TILE GEOMETRY												
Hexagonal	2	496	2	500	50.0	0.0020	1%	7	51.6	53.9	52.4	55.9

Table 9: Per-variant outcome variance and power analysis from 500 random-vs-random games per variant. Var(X): per-game score variance (win=1, draw=0.5, loss=0). N_{5pp}: sample size for 5pp effect at 80% power. Min@N: minimum score to reject H_0 (score=50) at sample size N. Uncorrected: α=0.05, two-sided. Bonferroni: α=0.0031 (0.05/16 variants). Decisiveness: fraction of non-draw games.

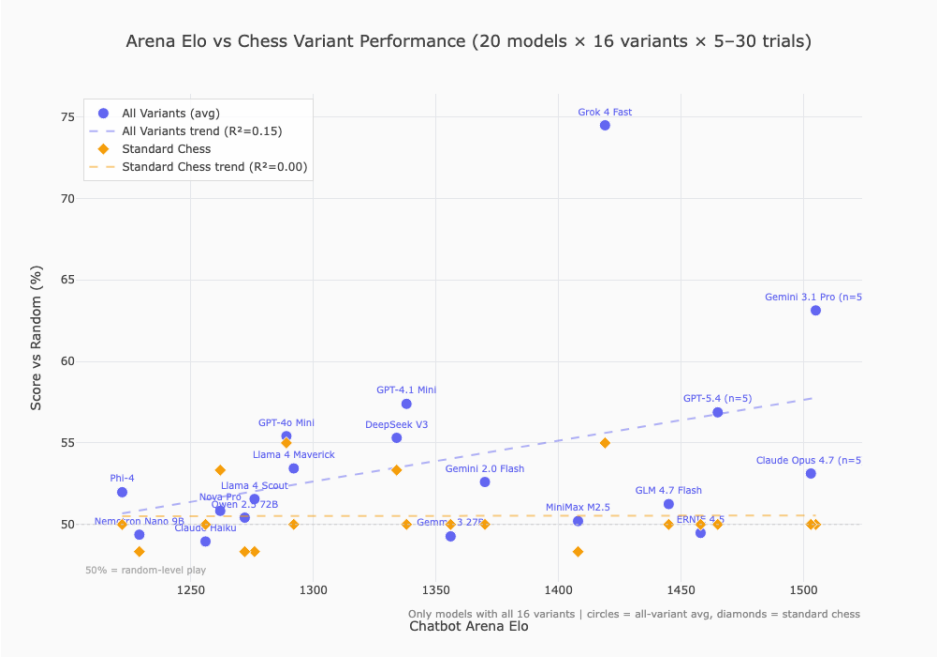

Figure 5: Chatbot Arena Elo vs LLM-vs-random score (19 models, each tested on all 16 variants with 5–30 trials per variant). Purple circles show the mean across all variants (R^2=0.15); amber diamonds show standard chess only (R^2=0.00). Dashed lines are linear trends; the dotted line at 50% marks random-level play.

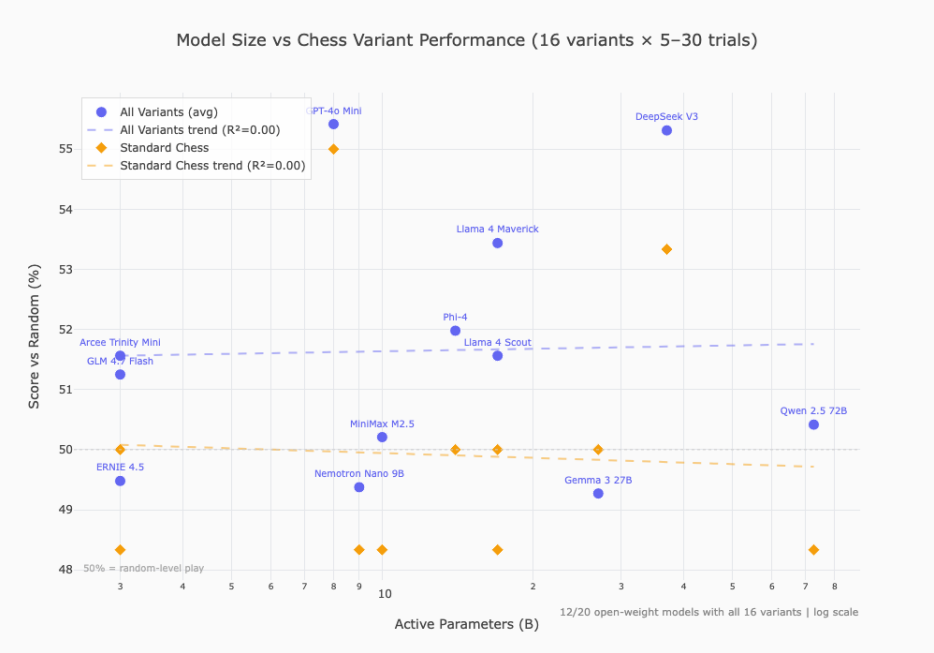

Figure 6: Active parameter count vs LLM-vs-random score (12 open-weight models, each on all 16 variants, 5–30 trials). Purple = all-variant mean (R^2=0.00); amber = standard chess (R^2=0.00). MoE models use per-token active count. Log scale.

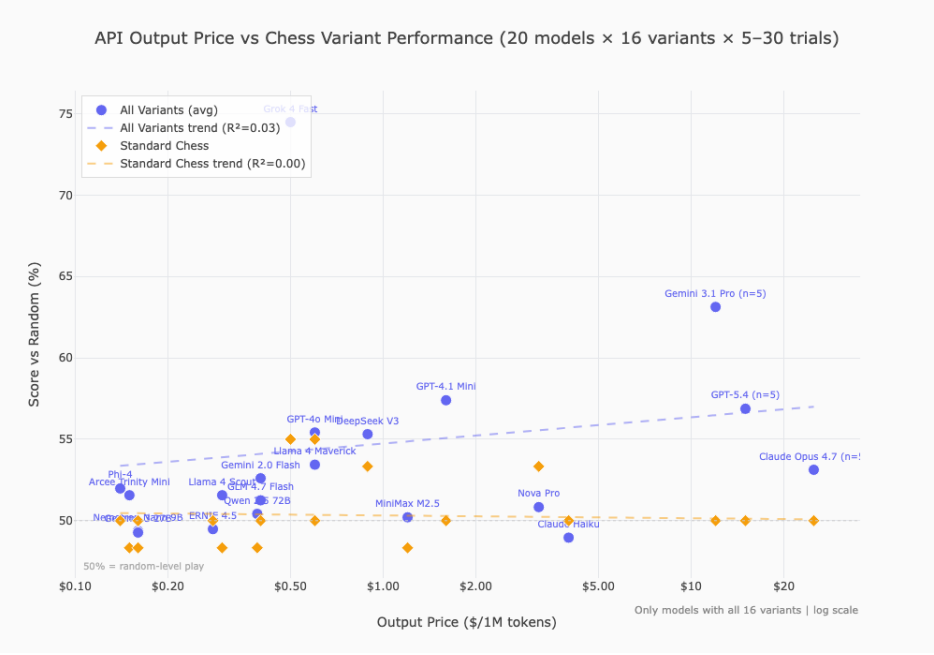

Figure 7: API output price vs LLM-vs-random score (20 models, each on all 16 variants, 5–30 trials). Purple = all-variant mean (R^2=0.03); amber = standard chess (R^2=0.00). Pricing from OpenRouter API. Log scale.

Model	Company	Params	Ctx	$/1M In	$/1M Out	Elo	AA
Gemini 3.1 Pro	Google DeepMind	Undisclosed	1.0M	$2.00	$12.00	1505	57
Claude Opus 4.7	Anthropic	Undisclosed	1.0M	$5.00	$25.00	1503	58
GPT-5.4	OpenAI	Undisclosed	1.1M	$2.50	$15.00	1465	35
ERNIE 4.5	Baidu	21B / 3B active	120K	$0.070	$0.280	1458	15
GLM 4.7 Flash	Zhipu AI	30B / 3B active	203K	$0.060	$0.400	1445	–
Grok 4 Fast	xAI	Undisclosed	2.0M	$0.200	$0.500	1419	24
MiniMax M2.5	MiniMax	230B / 10B active	197K	$0.150	$1.20	1408	–
Gemini 2.0 Flash	Google DeepMind	Undisclosed	1.0M	$0.100	$0.400	1370	–
Gemma 3 27B	Google DeepMind	27B	131K	$0.080	$0.160	1356	–
GPT-4.1 Mini	OpenAI	Undisclosed	1.0M	$0.400	$1.60	1338	–
DeepSeek V3	DeepSeek	671B / 37B active	164K	$0.320	$0.890	1334	–
Llama 4 Maverick	Meta	402B / 17B active	1.0M	$0.150	$0.600	1292	18
GPT-4o Mini	OpenAI	˜8B (est.)	128K	$0.150	$0.600	1289	–
Llama 4 Scout	Meta	109B / 17B active	328K	$0.080	$0.300	1276	14
Qwen 2.5 72B	Alibaba (Qwen)	72.7B	33K	$0.120	$0.390	1272	–
Nova Pro	Amazon	Undisclosed	300K	$0.800	$3.20	1262	–
Claude Haiku	Anthropic	Undisclosed	200K	$0.800	$4.00	1256	–
Nemotron Nano 9B	NVIDIA	9B	131K	$0.040	$0.160	1229	15
Phi-4	Microsoft	14B	16K	$0.065	$0.140	1222	10
Arcee Trinity Mini	Arcee AI	26B / 3B active	131K	$0.045	$0.150	–	–

Table 10: Characteristics of the 20 LLM models evaluated, sorted by Chatbot Arena Elo. Pricing from OpenRouter API. Arena Elo (19 models) from `openlm.ai/chatbot-arena`; Artificial Analysis Intelligence Index (9 models) from `artificialanalysis.ai`. – = not available.

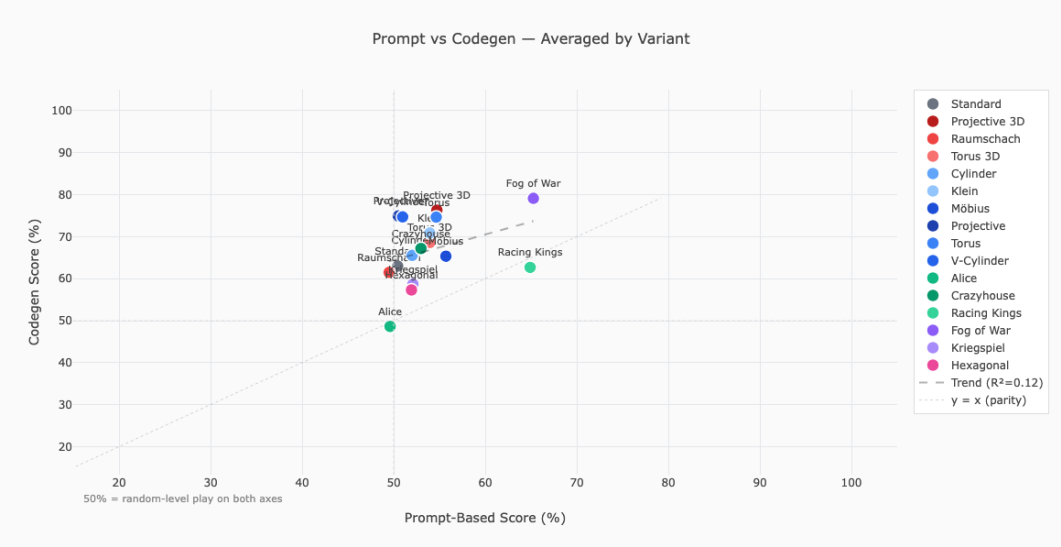

Figure 8: Prompt-based vs codegen score averaged by variant (16 variants, each averaged across 20 models). Points above the parity line indicate variants where codegen outperforms prompt-based play; points below indicate the reverse. The dashed line shows a linear trend (R^2=0.12). 50% = random-level play on both axes.

Variant	N	W+	D	B+	Draw%	W%	B%	Plies	Unadjusted n=30	Unadjusted n=5	BH-corrected n=30	BH-corrected n=5
STANDARD												
Standard	500	9	485	6	97.0%	1.8%	1.2%	79.1	2	1	4	2
3D												
Projective 3D	500	22	446	32	89.2%	4.4%	6.4%	76.8	4	2	7	3
Raumschach	500	2	491	7	98.2%	0.4%	1.4%	79.7	2	1	3	2
Torus 3D	500	22	451	27	90.2%	4.4%	5.4%	74.3	4	2	7	3
TOPOLOGY												
Cylinder	500	8	484	8	96.8%	1.6%	1.6%	79.0	2	1	4	2
Klein	500	91	309	100	61.8%	18.2%	20.0%	67.7	7	3	13	6
Möbius	500	48	407	45	81.4%	9.6%	9.0%	72.4	5	2	9	4
Projective	500	110	276	114	55.2%	22.0%	22.8%	65.7	8	3	14	6
Torus	500	70	346	84	69.2%	14.0%	16.8%	70.1	6	3	12	5
V-Cylinder	500	98	318	84	63.6%	19.6%	16.8%	67.4	7	3	13	6
RULE VARIANT												
Alice	500	86	338	76	67.6%	17.2%	15.2%	65.5	7	3	12	5
Crazyhouse	500	8	477	15	95.4%	1.6%	3.0%	78.8	3	1	5	2
Racing Kings	500	1	499	0	99.8%	0.2%	0.0%	80.0	1	1	1	1
IMPERFECT INFO												
Fog of War	500	96	313	91	62.6%	19.2%	18.2%	69.9	7	3	13	6
Kriegspiel	500	5	492	3	98.4%	1.0%	0.6%	79.7	2	1	3	2
TILE GEOMETRY												
Hexagonal	500	1	496	3	99.2%	0.2%	0.6%	79.8	1	1	2	1

Table 11: Random vs. random outcomes across all variants (80-ply cap). W^+/B^+ = white/black wins, D = draws. The "Min Wins" columns show the minimum decisive wins needed (assuming remaining games are draws) for statistical significance at α=0.05 (two-sided). Unadjusted: per-test α=0.05. BH-corrected: Benjamini–Hochberg FDR control at q=0.05 over k=320 tests (worst-case rank-1 threshold, α_{BH}=1.56e-4).

Variant	D1 vs Rnd	Grok 4	Gem. Pro	GPT-4o Mini	ERNIE	Gemma	GLM	GPT-5.4	Nemotron	Nova	Phi-4	Scout	GPT-4.1m	DS V3	Gem. Flash	MiniMax	Arcee	Opus 4.7	Qwen 72B	Llama 4	Haiku 3.5	N
STANDARD																						
Standard	-	80.0	30.0	20.0	30.0	40.0	40.0	20.0	30.0	40.0	10.0	30.0	40.0	20.0	30.0	30.0	30.0	40.0	30.0	30.0	20.0	5
3D																						
Projective 3D	-	60.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5
Raumschach	-	50.0	30.0	40.0	50.0	20.0	40.0	20.0	30.0	40.0	20.0	50.0	30.0	30.0	30.0	30.0	40.0	40.0	20.0	30.0	20.0	5
Torus 3D	-	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	5
TOPOLOGY																						
Cylinder	-	50.0	30.0	30.0	30.0	20.0	30.0	40.0	10.0	20.0	30.0	30.0	20.0	10.0	20.0	0.0	10.0	10.0	30.0	10.0	20.0	5
Klein	-	50.0	0.0	40.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	5
Möbius	-	60.0	0.0	60.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5
Projective	-	70.0	0.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	5
Torus	-	50.0	50.0	0.0	0.0	0.0	0.0	10.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5
V-Cylinder	-	50.0	0.0	0.0	10.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5
RULE VARIANT																						
Alice	-	70.0	30.0	10.0	30.0	30.0	20.0	20.0	60.0	50.0	0.0	0.0	0.0	10.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	5
Crazyhouse	-	50.0	50.0	0.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	0.0	50.0	50.0	50.0	50.0	50.0	10.0	5
Racing Kings	-	80.0	70.0	20.0	20.0	20.0	0.0	20.0	0.0	0.0	70.0	20.0	20.0	20.0	40.0	30.0	10.0	0.0	0.0	0.0	20.0	5
TILE GEOMETRY																						
Hexagonal	-	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	40.0	50.0	50.0	40.0	50.0	50.0	50.0	5
Mean		56.4	24.3	20.7	19.3	17.9	17.9	17.9	17.9	17.9	16.4	16.4	15.7	15.0	14.3	14.3	13.6	12.9	12.9	12.1	10.0	1400

Table 12: LLM vs Minimax D1 performance (Score). Score = (W + D×0.5) / N × 100; 50.0 = D1-level play. D1 vs Rnd shows tree baseline against random. Max plies = 80.

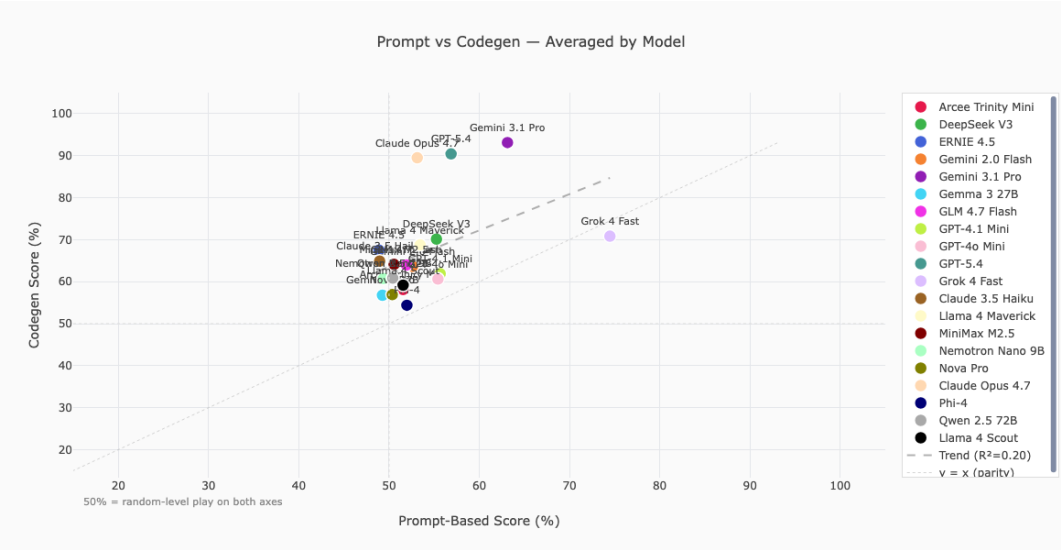

Figure 9: Prompt-based vs codegen score averaged by model (20 models, each averaged across 16 variants). Points above the parity line indicate models whose generated engines outperform their prompt-based play; points below indicate the reverse. The dashed line shows a linear trend (R^2=0.20). 50% = random-level play on both axes.

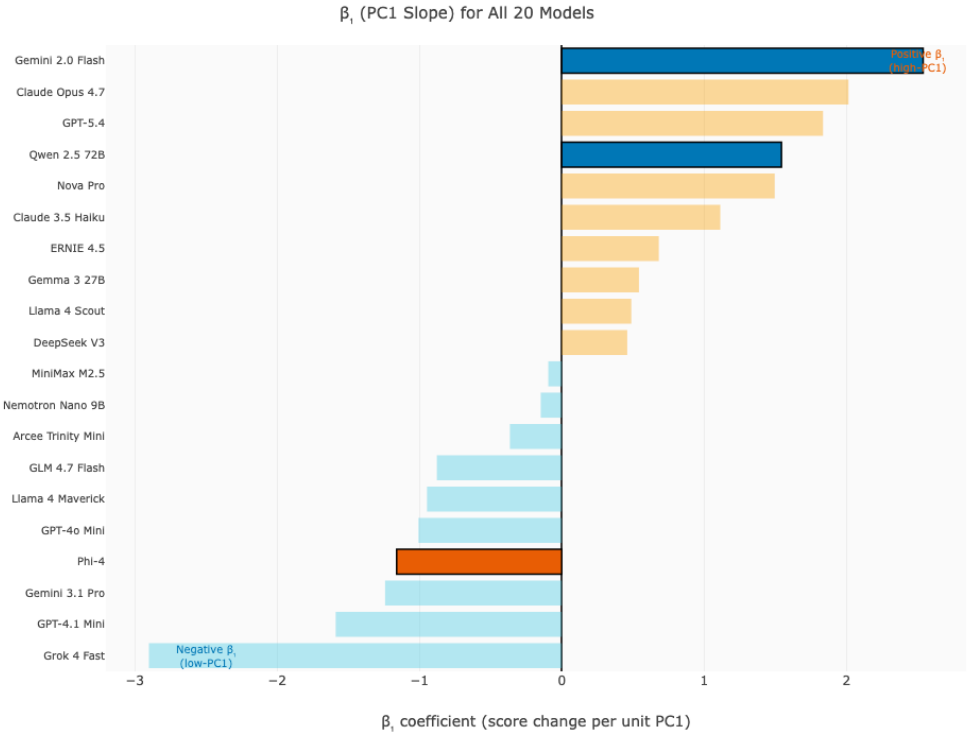

Figure 10: PC1 slope (β_1) for all 20 models. Three models reach $p < 0.05$ (nominal, outlined in black): Phi-4 ($\beta_1 = +-1.16$), Qwen 2.5 72B ($\beta_1 = 1.54$), and Gemini 2.0 Flash ($\beta_1 = 2.54$). Positive β_1 indicates better performance on high-PC1 (chaotic topology) variants; negative indicates better performance on low-PC1 (structured, positional) variants. No result survives Bonferroni correction ($\times 20$).

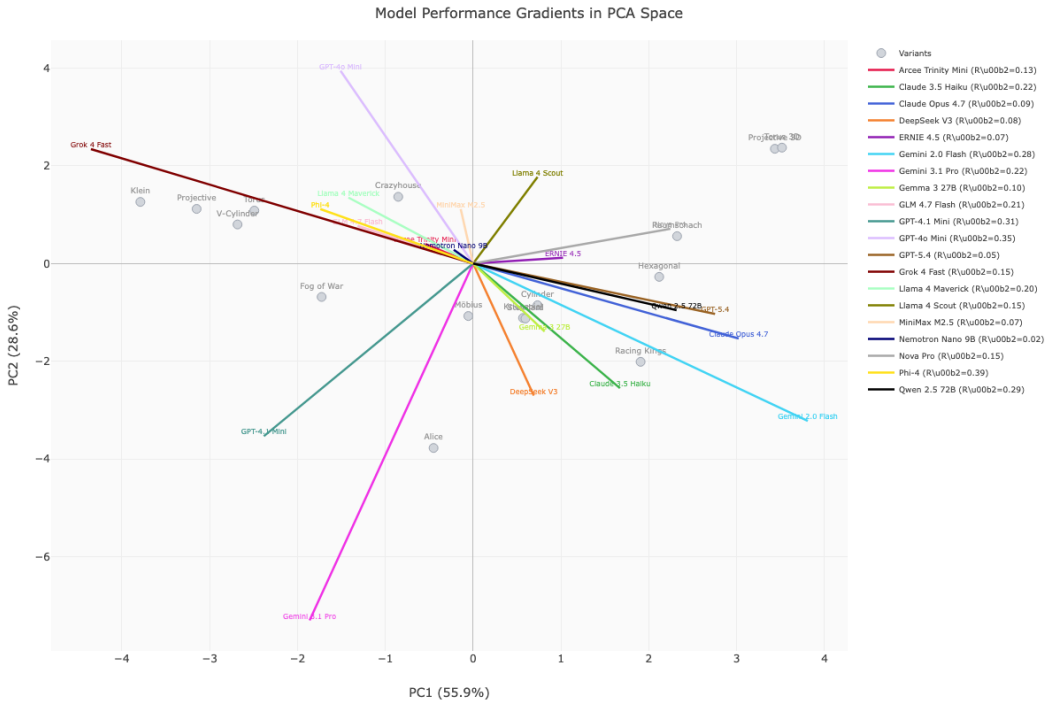

Figure 11: Model performance gradient vectors in PCA complexity space. Each arrow shows the OLS-fitted direction (score $\sim$ PC1 + PC2) in which a model's score increases across 16 variants. Arrow length is proportional to gradient magnitude. PC1 (55.9%) captures game length, branching, and volatility; PC2 (28.6%) captures tactical intensity.

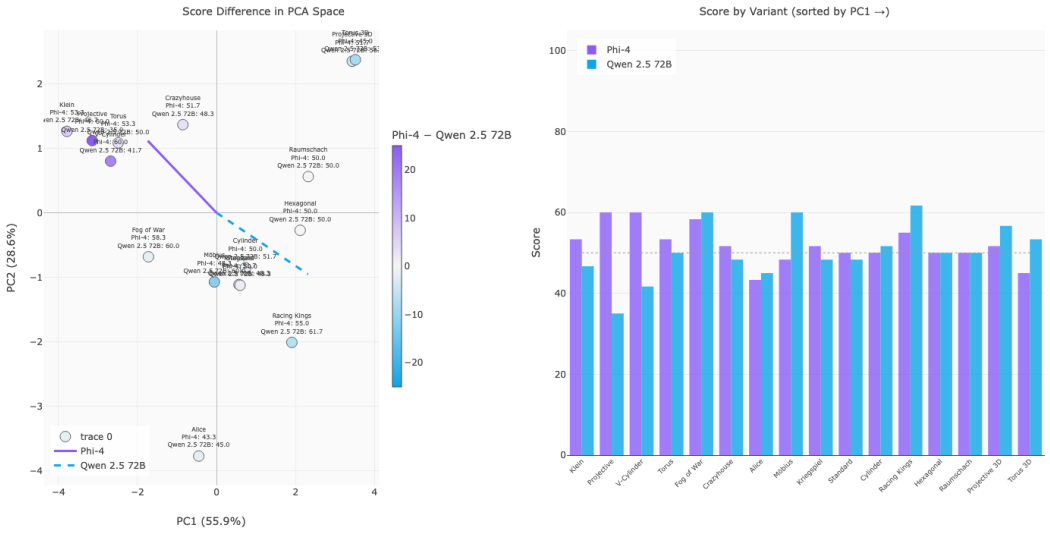

Figure 12: Opposing complexity preferences: Phi-4 vs. Qwen 2.5 72B. Left: variants in PCA space, colored by score difference (purple = Phi-4 advantage, blue = Qwen advantage). Arrows show each model's gradient vector. Right: scores by variant sorted by PC1. Phi-4 excels on high-PC1 variants (Klein, Projective); Qwen excels on low-PC1 variants (Raumschach, Hexagonal). Cosine similarity of gradient vectors: -0.986.

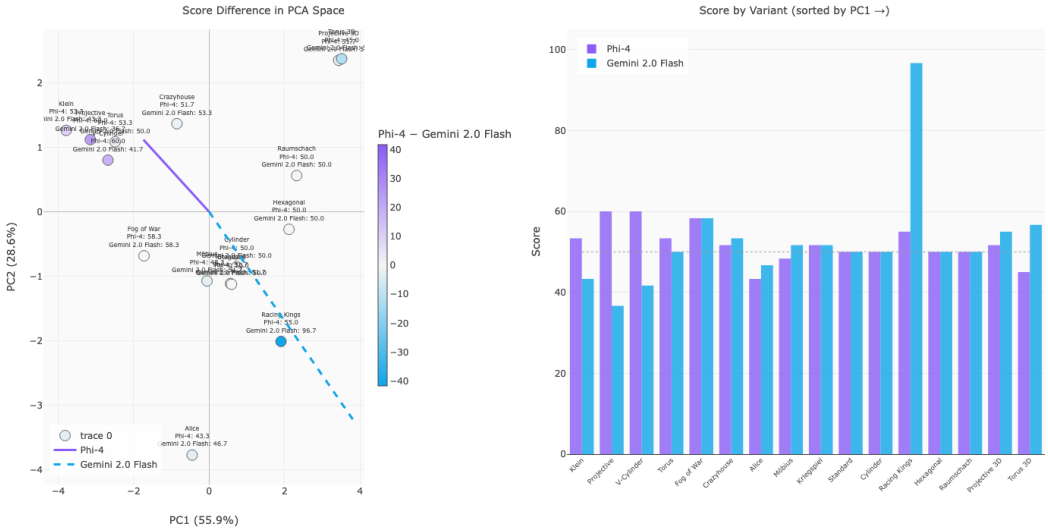

Figure 13: Opposing complexity preferences: Phi-4 vs. Gemini 2.0 Flash. Left: variants in PCA space, colored by score difference. Right: scores by variant sorted by PC1. Gemini Flash shows a steeper negative slope ($\beta_1 = 2.54$, $R^2 = 0.28$) than any other model, with its strongest performance on low-PC1 structured variants. Cosine similarity of gradient vectors: -0.998.