EngineLab: Evaluating Strategic Generalization Under Rule Shifts

Tianyi Evans Gu TIANYIGU@UPENN.EDU

Lucas Yuan YUAN07@UPENN.EDU

Abstract

Strategies that work well in one game don't necessarily transfer to another. Yet we lack testbeds that isolate which rule changes break which strategies. We introduce *EngineLab*, a benchmark designed to study this question using chess variants as controlled perturbations of the traditional rule set. In *EngineLab*, we build agents from different combinations of strategic features and have them compete in round-robin tournaments, with each feature's importance measured via an empirical Banzhaf power index. Across seven variants, dominant strategies vary considerably: a feature that helps in one regime often hurts in another, and the top-ranked feature differs in 5 of 7 variants. An EXP3 bandit learner settles on the best strategy over $10\times$ faster in some variants than in others. To test transfer directly, we warm-start the learner on one variant and evaluate it on another. Across a full transfer matrix of 42 directed pairs with 5 seeds each, this helps between structurally similar regimes but hurts when the source has an inverted objective (e.g., Antichess, whose goal is to lose material). GPT-4o, when asked to predict feature rankings from rule descriptions, ranks them no better than chance (Kendall $\tau = +0.08$) by overgeneralizing from standard-game priors. These results suggest that strategic value is closely tied to a game's rules, making rule-shifted games a controlled benchmark for studying strategic transfer.

1. Introduction

In multi-agent systems, *strategic generalization* is the ability to keep playing well when the rules of the game change. We ask whether strategies optimized for one game become counterproductive under even modest rule perturbations. This is a question most benchmarks cannot address, since they fix the environment and measure performance within a single regime [9, 13].

We introduce *EngineLab*, a benchmark for evaluating whether interpretable strategic heuristics transfer across controlled rule perturbations. To study how optimal strategies shift, *EngineLab* uses chess variants, which are games that share chess's board and pieces but modify its rules or starting position. Programmatic agents constructed from modular evaluation features compete in round-robin tournaments, producing an empirical meta-game. An EXP3 bandit learner [1] discovers optimal strategies online, with convergence round reflecting how separated the top strategies' payoffs are in each variant.

Our contributions are:

1. An open and deterministic benchmark protocol with game-theoretic grounding (Banzhaf index and a second-order interaction decomposition) for studying strategic transfer.

2. Evidence that feature importance varies considerably across regimes: 6 of the 8 features present in multiple variants change sign, and the top feature differs in 5 of 7 variants.

© 2026 T.E. Gu & L. Yuan

3. A full transfer matrix (42 directed pairs, 5 seeds): warm-starting reliably hurts from inverted-objective variants but helps between similar ones (Appendix D).

4. An LLM strategic-reasoning probe: GPT-4o predicts feature rankings at near-chance accuracy ($\tau = +0.08$), overgeneralizing from standard-game priors (Appendix G).

Related work. EGTA evaluates agent populations via meta-game payoff matrices [17, 18]; PSRO [8] unifies it with deep RL, and Gatchel and Wiedenbeck [5] extend it to parameterized game families. We complement this line by studying heuristic transfer across discrete rule perturbations. Unlike methods that learn payoff models over continuous, parameterized game spaces, we use a fixed set of named chess variants, letting us read each feature's importance off directly and compare it across regimes. Omidshafiei et al. [10], Balduzzi et al. [2], and Czarnecki et al. [4] characterize multi-agent interaction geometry, while Shapley/Banzhaf and interaction indices [3, 6, 12, 16] connect cooperative game theory to interpretable attribution, underpinning our marginal and synergy analysis. We use EXP3 [1] as the meta-learner, and our setting is motivated by zero-shot generalization [7, 14], domain randomization [15], and chess variants as discrete, human-interpretable rule perturbations [11].

2. The EngineLab Framework

2.1. Programmatic Agents and Meta-Game

EngineLab constructs agents from a set of n strategic evaluation features $F = \{f_1, \ldots, f_n\}$ (Table 2 in Appendix A). Each agent uses a subset $S \subseteq F$ with equal weights within a depth-2 alpha-beta search. Equal weights keep each agent defined purely by its feature set, so a feature's marginal contribution reflects the feature itself rather than weight tuning. With n=8 features per variant (2^8-1=255 possible subsets), we use stratified sampling: all 8 singletons, all $\binom{8}{2}$=28 pairs, the full set, and random subsets of sizes 3–7, yielding 20 agents. A round-robin tournament among all agents produces an empirical payoff matrix.

Definition 1 (Feature-Subset Meta-Game) *Let $F = \{f_1, \ldots, f_n\}$ be evaluation features. The feature-subset meta-game is $\mathcal{G} = (\mathcal{S}, u)$ where $\mathcal{S} \subseteq 2^F \setminus \emptyset$ is a set of feature subsets and $u(s_i, s_j) \in [0, 1]$ is the empirical win rate of the agent with features s_i against the agent with features s_j. In practice, $\mathcal{S}$ is a stratified sample of the $2^n - 1$ possible agents.*

Definition 2 (Feature Value Function) $v : 2^F \to \mathbb{R}$, *where $v(S)$ = mean win rate of all agents whose feature set is a superset of S.*

2.2. Formal Analysis

The *marginal contribution* of feature f is $m(f) = \mathbb{E}_S[v(S \cup \{f\}) - v(S)]$, averaged over subsets $S \not\ni f$.

Under exhaustive enumeration, $m(f)$ coincides with the Banzhaf power index [3]; with stratified sampling, it approximates $\beta(f)$. We also define *pairwise synergy* $\sigma(a, b) = v(\{a, b\}) - v(\{a\}) - v(\{b\}) + v(\emptyset)$, the extent to which two features' joint value departs from the sum of their separate contributions [6]. Together, marginals and synergies yield a second-order decomposition of the meta-game value function. Synergy structure also shifts across regimes (Appendix E): the most redundant pair in Standard (material + mobility, $\sigma = -0.78$) becomes synergistic in Atomic (capture threats + material, $\sigma = +0.09$).

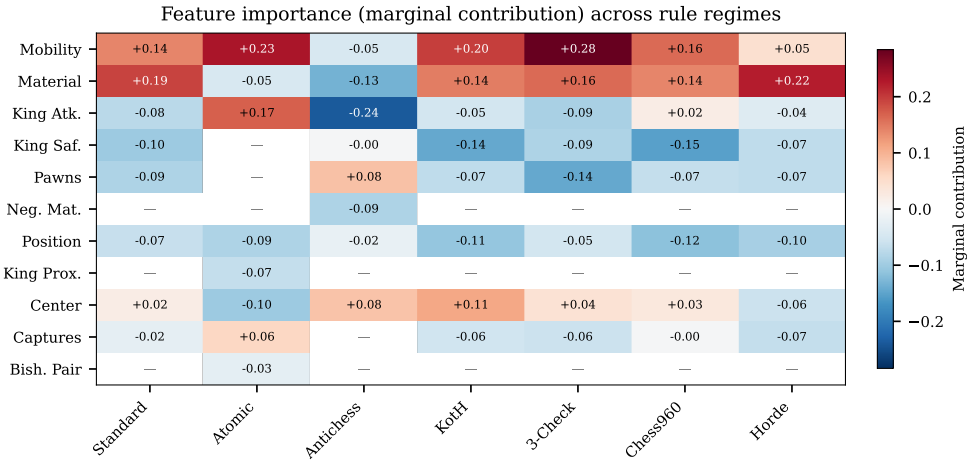


Figure 1: Feature marginal contributions across 7 rule regimes. Diverging colormap (red = positive, blue = negative). Dashes indicate features absent from a variant. Six of the eight features present in multiple variants change sign; all sign changes are significant at 95% bootstrap CI.

2.3. Online Strategy Learning

We deploy an EXP3 bandit learner [1] over the $K = |\mathcal{S}|$ feature-subset strategies, using the precomputed payoff matrix for $O(1)$ per-round updates. We track the *convergence round*: the first round after which the eventual best strategy holds the highest weight for a sustained window of checkpoints. To test transfer, we warm-start EXP3 with weights learned in one variant and run on another.

3. Experiments

Setup. We evaluate 7 rule regimes: Standard, Atomic, Antichess, King of the Hill (KotH), ThreeCheck, Chess960, and Horde. Each variant uses 8 of 11 features, those relevant to its rules (some variant-specific, e.g., negative material in Antichess), yielding 20 stratified-sample agents. Round-robin tournaments produce 380 games per variant at search depth 2; bootstrap confidence intervals (Appendix C) confirm significance of reported marginals. All experiments are fully deterministic due to the fixed seeds and deterministic move generation.

3.1. Cross-Variant Strategy Profiles

Figure 1 shows marginal contributions across all 7 regimes. Non-transferability is pronounced: 6 of the 8 features present in multiple variants change sign (all sign changes are significant at 95% bootstrap CI; the remaining 3 features appear in only one variant each). Material, the most valued feature in standard chess ($m = +0.19$, CI $[+0.12, +0.27]$), becomes harmful in Antichess ($m = -0.13$), where losing pieces is the objective. Mobility is the most broadly useful feature, positive in 6 of 7 variants and top-ranked in 4, though it too turns negative in Antichess. King attack swings the most, from $+0.17$ in Atomic to -0.24 in Antichess. Table 1 summarizes per-variant profiles alongside EXP3 convergence data. The EXP3 learner's best strategy matches the tournament champion or runner-up in all 7 variants, confirming that the learner recovers the strongest strategies from the meta-game payoffs.

Table 1: Per-variant summary. *Top feature*: highest marginal. *EXP3 best*: strategy with highest final weight after T=5000 rounds. *Conv.*: round at which top strategy stabilizes. In Antichess, Pawns and Center control tie at the top ($+0.08$); we list Pawns.

Variant	Top feature	EXP3 best strategy	Conv.
Standard	Material (+.19)	Material + Center	2450
Atomic	Mobility (+.23)	Mobility + Captures	5000
Antichess	Pawns (+.08)	Neg. mat. + K. atk.	4100
KotH	Mobility (+.20)	Material + Center	2300
3-Check	Mobility (+.28)	Mobility	400
Chess960	Mobility (+.16)	Material + Center	2500
Horde	Material (+.22)	Material	1700

3.2. Learning Dynamics and Transfer

The convergence round varies by more than $10\times$ across variants, from 400 in Three-Check, where one strategy dominates, to the full T=5000 horizon in Atomic, where several strategies have similar payoffs (Figure 2, Appendix F).

Cross-variant transfer. For each ordered (source, target) pair, we warm-start the learner on the source and record its change in regret on the target versus a cold start; this *transfer matrix* spans all 42 directed pairs, 5 seeds each (Table 3 in Appendix D). Antichess, with its inverted objective, is the worst source: warm-starting from it hurts every target (mean regret increase $+5$ to $+65$). Conversely, Three-Check is the most beneficial source, reducing regret in 5 of 6 targets because its strong mobility prior transfers well. Across all 42 pairs, transfer helps in 28 and hurts in 14. Strategic knowledge transfers reliably between regimes with aligned objectives but breaks down when the source's win condition is inverted.

4. Discussion and Conclusion

Our results show that rule-shifted games provide a controlled benchmark for studying strategic transfer. Key findings: (1) strategic heuristics are largely non-transferable across regimes; (2) the EXP3 convergence round reflects how separated the top strategies' payoffs are, providing a learning-based complement to static feature rankings; (3) across the full 42-pair transfer matrix, whether transfer helps is governed by structural compatibility between source and target, not variant similarity alone, in line with known effects of domain priors under distribution shift [14].

Limitations. Agents use equal feature weights and shallow search (depth 2), which may not capture deeper strategic interactions; Appendix B shows $\tau = 0.52$ between depth-2 and depth-3 rankings for Atomic, indicating moderate robustness. EXP3 is a simple learner; more expressive methods may exhibit different transfer properties. The benchmark spans 7 regimes within one game family; extending to structurally distinct games would strengthen generality claims.

LLM strategic reasoning. Asked to predict feature rankings from rule descriptions alone, GPT-4o performs at chance (Kendall $\tau = +0.08$, 14% top-1; Appendix G), overgeneralizing from standard-game priors.

Future directions. Extensions include learned feature weights, deeper search, non-chess rule-shifted games, and additional LLM evaluations across model families. We release *EngineLab* as an open, deterministic benchmark for studying strategic transfer.

References

[1] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[2] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Pérolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning*, pages 434–443, 2019.

[3] John F. Banzhaf. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[4] Wojciech Marian Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. Real world games look like spinning tops. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[5] Madelyn Gatchel and Bryce Wiedenbeck. Learning parameterized families of games. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pages 1044–1052, 2023.

[6] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565, 1999.

[7] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76: 201–264, 2023.

[8] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[10] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Pérolat, and Rémi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9:9937, 2019.

[11] David Pritchard. *The Classified Encyclopedia of Chess Variants*. John Beasley, 2007.

[12] Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953.

[13] 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.

[14] Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685, 2009.

[15] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[16] Jiachen T. Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[17] Michael P. Wellman. Methods for empirical game-theoretic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1552–1556, 2006.

[18] Michael P. Wellman, Karl Tuyls, and Amy Greenwald. Empirical game theoretic analysis: A survey. *Journal of Artificial Intelligence Research*, 82, 2025.

Appendix A. Feature Descriptions

Table 2: Strategic evaluation features. Each agent uses a subset as its evaluation function. Features marked with † are variant-specific.

Feature	Description
Material	Piece count difference
Piece position	Piece-square table bonus
Center control	Central square occupation
King safety	Pawn shield around own king
King attack	Threats near enemy king
Mobility	Legal move count difference
Pawn structure	Doubled/isolated pawn penalty
Bishop pair	Two-bishop bonus
Capture threats	Hanging piece detection
Neg. material†	Reward for losing pieces
King proximity†	King centralization bonus

Appendix B. Depth Sensitivity

To assess robustness to search depth, we compare feature marginals between depth-2 and depth-3 search for Atomic chess, the variant exhibiting the most distinctive strategic profile. Over 7 common features, the Kendall τ rank correlation between depths is $\tau = 0.52$, indicating moderate agreement. The largest shift occurs for capture threats ($\Delta = -0.11$), which decreases in importance at greater depth as the search horizon extends beyond immediate tactical threats. This suggests that while the qualitative pattern (mobility dominates) is robust, quantitative rankings are sensitive to search depth, a consideration for future extensions of the benchmark.

Appendix C. Bootstrap Confidence Intervals

We compute a 95% bootstrap confidence interval for each feature marginal by resampling game results with replacement (1000 resamples). As representative examples, mobility's marginal in Atomic has CI $[+0.15, +0.31]$ and material's in Standard has CI $[+0.12, +0.27]$, both excluding zero. More importantly, every one of the 6 sign changes reported in Section 3.1 has a confidence interval that excludes zero, so these reversals are not artifacts of sampling noise.

Appendix D. Cross-Variant Transfer Matrix

Table 3 shows the full 42-pair transfer matrix. Each cell reports the mean regret change $\Delta =$ warm $-$ cold over 5 seeds ($\pm$ std). Negative values (bold) indicate transfer helps; positive values indicate transfer hurts.

Key observations. Antichess (inverted win condition) is the only variant that always hurts as a source. Three-Check produces the largest regret reductions of any source, helping 5 of 6 targets; Standard, KotH, and Chess960 also help 5 of 6, but with smaller magnitudes. Transfer is approximately symmetric for structurally similar pairs (Standard↔Chess960: -35/-59) but strongly asymmetric

Table 3: Cross-variant transfer: mean regret change (Δ) from warm-starting, 5 seeds each. Negative = transfer helps (bold); positive = transfer hurts.

Source ↓ / Target →	Std	Ato	Anti	KotH	3Ch	960	Hor
Standard	—	**−3**	+28	**−84**	**−93**	**−35**	**−51**
Atomic	**−19**	—	+5	**−19**	**−32**	**−26**	+7
Antichess	+26	+5	—	+29	+36	+41	+65
KotH	**−86**	**−15**	+18	—	**−111**	**−57**	**−36**
3-Check	**−101**	**−33**	+30	**−116**	—	**−60**	**−54**
Chess960	**−59**	**−31**	+28	**−89**	**−90**	—	**−49**
Horde	**−63**	+10	+61	**−49**	**−67**	**−33**	—

when objectives diverge (Standard→Antichess: $+28$; Antichess→Standard: $+26$). The smallest harm ($+5$) is shared by Antichess↔Atomic in both directions, plausibly because both regimes make material advantage less reliable.

Appendix E. Pairwise Feature Synergy

Using the value function v of Definition 2, we compute pairwise synergies $\sigma(a, b) = v(\{a, b\}) - v(\{a\}) - v(\{b\}) + v(\emptyset)$ for three representative variants:

Table 4: Top synergistic (+) and most redundant ($-$) feature pairs by variant.

Variant	Feature pair	σ
Standard	King safety + Capture threats	$+0.05$
	Material + Mobility	-0.78
Atomic	Capture threats + Material	$+0.09$
	Mobility + King attack	-0.84
Antichess	Center + King attack	$+0.33$
	Pawns + King safety	-0.36

Note: synergy magnitudes can exceed typical marginal sizes because σ measures departure from additivity. Two individually strong features that provide overlapping information yield large negative synergy. The most redundant pair in Standard (material + mobility, $\sigma = -0.78$) involves two individually strong features that provide overlapping strategic information. In Antichess, a different pair (pawn structure + king safety, $\sigma = -0.36$) is most redundant, reflecting the entirely different strategic structure. The most synergistic pair also differs: in Atomic, capture threats synergize with material ($+0.09$) because captures trigger explosions. Here, knowing which pieces to capture is more valuable than material alone.

Appendix F. EXP3 Learning Dynamics

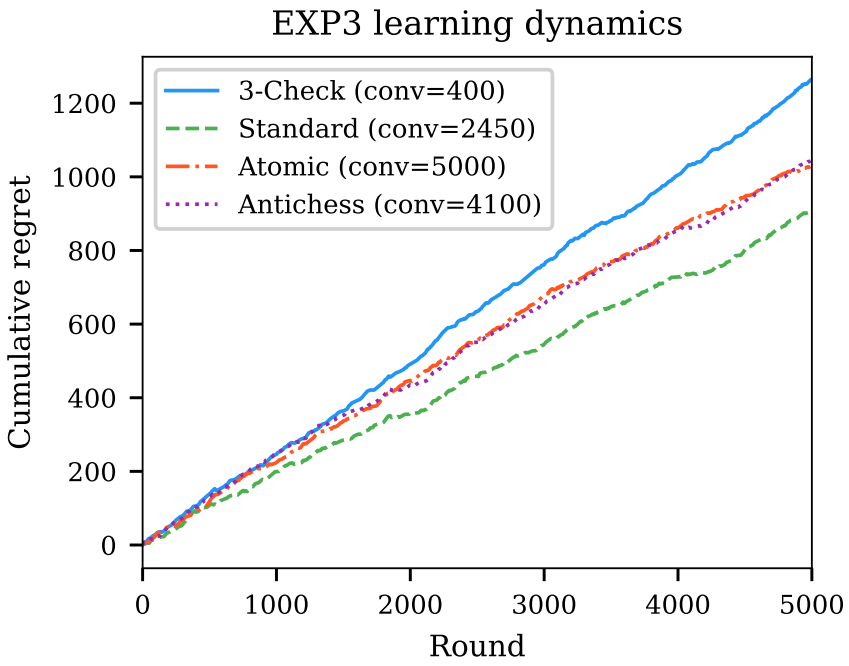


Figure 2: EXP3 cumulative regret for four rule regimes (T=5000). Three-Check converges rapidly (round 400); Atomic does not converge within the horizon. Full convergence ordering: Three-Check (400) $<$ Horde (1700) $<$ KotH (2300) $<$ Standard (2450) $<$ Chess960 (2500) $<$ Antichess (4100) $<$ Atomic (5000).

Appendix G. LLM Strategic Reasoning

We test whether GPT-4o can predict feature importance rankings from variant rule descriptions alone. For each of 7 variants, we provide the rule description and list of available features, then ask the model to rank features by predicted importance (3 independent runs per variant, 21 total).

Table 5: GPT-4o feature ranking prediction vs. empirical ground truth. τ: mean Kendall rank correlation over 3 runs. Top-1: fraction of runs whose predicted top feature matches the empirical #1. Random baseline: $\tau \approx 0.00$, top-1 $= 12.5\%$.

Variant	Empirical #1	GPT-4o #1	τ	Top-1
Standard	Material	Material	-0.05	3/3
Atomic	Mobility	King attack	$+0.33$	0/3
Antichess	Pawns	Neg. material	-0.12	0/3
KotH	Mobility	Center	$+0.21$	0/3
3-Check	Mobility	King attack	-0.10	0/3
Chess960	Mobility	Material	$+0.33$	0/3
Horde	Material	Pawns	-0.05	0/3
Overall			$\mathbf{+0.08}$	**14%**

Standard chess is the one variant where GPT-4o correctly identifies the top feature, as standard-chess heuristics happen to match the empirical ranking. It fails in the other 6. Its predictions hint at systematic overgeneralization. For Chess960, it predicts material (as in Standard) when mobility empirically dominates; for Antichess, it predicts negative material (plausible surface-level reasoning about the inverted objective) when pawn structure is actually most important. The positive τ for Atomic ($+0.33$) and Chess960 ($+0.33$) suggests partial understanding of some regimes, but top-1 accuracy remains 0/3 in both. The model captures rough ordinal trends without identifying the dominant feature. The overall $\tau = +0.08$ is barely above the random baseline (≈ 0.00), and top-1 accuracy (14%) is indistinguishable from chance (12.5%). Strategic knowledge does not carry over reliably under rule shifts, whether it was learned from another variant or absorbed from an LLM's training corpus, just as the transfer matrix showed.

Camera-Ready Edits Summary

Author information. De-anonymized the paper (added author names and contact emails) and added the copyright line.

Corrections.

- **Figure 1 caption:** the colormap legend was reversed in submission; corrected to *red = positive, blue = negative* to match the figure (and the colorbar).
- **Sign-change statistic:** reported as 6 of the *8* features present in multiple variants (rather than 6 of 11), since the 3 single-variant features cannot change sign; updated consistently in the abstract, contributions, and Section 3.1.
- **Section 3.2 transfer ranges:** corrected the Antichess-source regret increase to $+5$ to $+65$ (consistent with Table 3) and removed an inaccurate "up to -117" (the largest reduction in Table 3 is -116).
- **Appendix D:** corrected the set of strongest transfer sources (Three-Check, Standard, KotH, and Chess960 each help 5 of 6 targets).
- **Disclosed two ties:** the Antichess top feature (Pawns and Center control both $+0.08$; Table 1) and the smallest transfer harm (Antichess↔Atomic, $+5$ in both directions; Appendix D).
- **References:** corrected the Wellman (2006) page range to 1552–1556 and added page numbers (1044–1052) to Gatchel and Wiedenbeck (2023).
- **Feature table (Appendix A):** removed an unused "Rook activity" row so the table lists the 11 features actually used across variants.
- **Abstract wording:** agents are built "from different combinations of strategic features" (not weights) and "compete" is used correctly.

Clarifications (no change to results).

- Defined the online-learning metric precisely and renamed it the *convergence round* (Section 3.2).
- Corrected the variant description: variants modify the rules *or starting position* (not only "one or two rules"), since e.g. Horde and Chess960 change the starting setup.
- Described feature importance as an *empirical* (sampled) Banzhaf power index, and described pairwise synergy in plain terms (departure from additivity), with its defining formula now shown in Appendix E.
- Added the rationale for equal feature weights (Section 2) and rewrote the bootstrap-CI appendix (Appendix C) for clarity.
- Reframed the introduction's opening as the paper's research question and softened overstated language throughout.

Added discussion (no new experiments).

- Two additional read-offs from Figure 1 (mobility is broadly useful; king attack has the largest swing).
- A related-work sentence positioning the benchmark against methods that learn payoff models over continuous, parameterized game spaces.