EngineLab: Evaluating Strategic Generalization Under Rule Shifts

author names withheld

Under Review for NExT-Game 2026

Abstract

Strategies that work well in one game don't necessarily transfer to another. But, when and why they fail is poorly understood. We introduce EngineLab, a benchmark designed to study this question using chess variants as controlled perturbations of the traditional rule set. In EngineLab, we assign agents with different combinations of strategic weights and compete them in round-robin tournaments, with each feature's importance measured through its Banzhaf power index. Using seven chess variants as rule perturbations, we show that dominant strategies do vary across the games: 6 of 11 features change sign, and the top-ranked feature differs in 5 of 7 variants. An EXP3 bandit learner discovers optimal strategies at rates varying by $10\times$ across variants. A full cross-variant transfer matrix (42 directed pairs, 5 seeds each) shows that warm-starting reliably hurts when the source variant has an inverted objective, while it helps between structurally similar regimes. GPT-4o, when asked to predict feature rankings from rule descriptions, achieves near-random accuracy ($\tau = +0.08$), exhibiting the same overgeneralization from standard-game priors. These results suggest that rule-shifted games offer a useful benchmark for studying strategic transfer.

1. Introduction

A central challenge in multi-agent systems is *strategic generalization*: maintaining effective behavior when the rules of engagement change. Strategies optimized for one game may be counterproductive under even modest rule perturbations, yet most benchmarks 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. EngineLab uses chess variants, which are games sharing a common board and piece set but differing in one or two rules, to study how optimal strategies shift. 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 rate reflecting the effective payoff gap structure of each variant.

Our contributions are:

1. A **benchmark protocol** with game-theoretic grounding using Banzhaf index and ANOVA interaction decomposition.
2. Evidence that **strategic heuristics are non-transferable**: across 7 regimes, dominant features change in rank.
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 ($\tau = +0.08$), exhibiting the same overgeneralization pattern (Appendix G).

5. An **open, fully deterministic benchmark** for studying strategic transfer.

Related work. EGTA evaluates agent populations via meta-game payoff matrices [17, 18]; PSRO [8] unifies EGTA with deep RL; Gatchel and Wiedenbeck [5] extend EGTA to parameterized game families. We complement this line by studying heuristic transfer across discrete rule perturbations. Omidshafiei et al. [10], Balduzzi et al. [2], and Czarnecki et al. [4] characterize multi-agent interaction geometry. Feature attribution via Shapley/Banzhaf indices [3, 12, 16] connects cooperative game theory to interpretable analysis; Grabisch and Roubens [6] axiomatize interaction indices underpinning our synergy decomposition. We apply EXP3 [1] as a meta-learner over programmatic agents. Zero-shot generalization across environments [7, 14] and domain randomization [15] motivate our setting since chess variants are discrete, human-interpretable rule perturbations [11] which enables our controlled study of strategic transfer.

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. 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 $\sigma(a, b) = v(\{a, b\}) - v(\{a\}) - v(\{b\}) + v(\emptyset)$, which equals the two-way ANOVA interaction term [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, $\sigma = +0.09$).

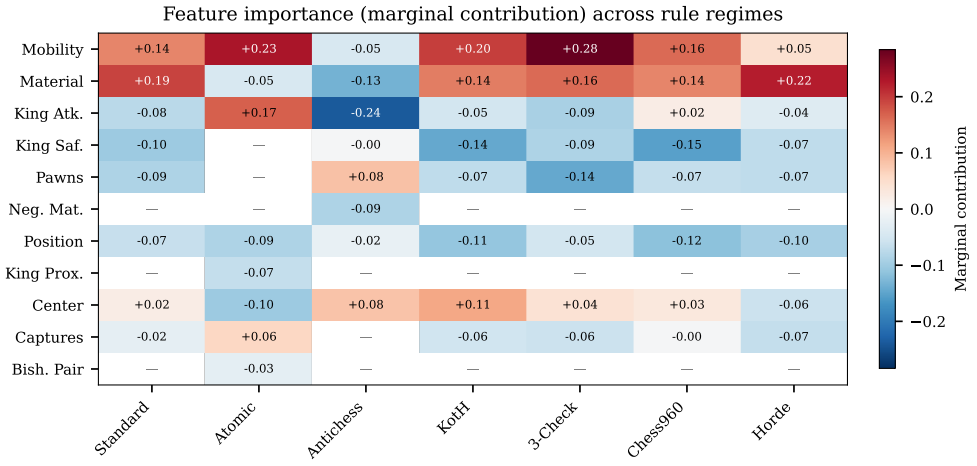

Figure 1: Feature marginal contributions across 7 rule regimes. Diverging colormap (blue = positive, red = negative). Dashes indicate features absent from a variant. Six features change sign across regimes; all sign changes are significant at 95% bootstrap CI.

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 use the *empirical convergence rate*, the round at which the top strategy stabilizes, as one measure of the effective payoff gap structure of each variant. 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 total features, generating 20 stratified-sample agents per variant. 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 (fixed seeds, deterministic move generation).

3.1. Cross-Variant Strategy Profiles

Figure 1 shows marginal contributions across all 7 regimes, revealing sharp non-transferability: **6 of 11 features change sign** (all sign changes are significant at 95% bootstrap CI). 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. 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, providing mutual validation.

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.

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

EXP3 convergence varies by $>10\times$ across variants (Figure 2 in Appendix F): Three-Check converges in 400 rounds (single dominant strategy), while Atomic requires the full T=5000 horizon (multiple competitive strategies with smaller gaps).

Cross-variant transfer. We run a full transfer matrix: all 42 directed source→target variant pairs, each with 5 seeds (Table 3 in Appendix D). The results follow a pattern: Antichess, with its inverted objective, is the worst source; warm-starting from Antichess *hurts* every target (mean regret increase $+28$ to $+65$). Conversely, Three-Check is the best source, reducing regret in 5 of 6 targets (up to -117) because its strong mobility prior transfers well. Across all 42 pairs, transfer helps in 28 and hurts in 14. Strategic knowledge transfers reliably only between regimes with compatible strategic structure.

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 sharply non-transferable across regimes; (2) EXP3 convergence rate reflects the effective payoff gap structure of each variant, providing a learning-based complement to static feature rankings; (3) the full 42-pair transfer matrix reveals that transfer success depends on structural compatibility between source and target regimes, not just variant similarity, echoing broader observations about 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. We evaluate whether GPT-4o can predict feature rankings from rule descriptions alone (Appendix G). Overall Kendall $\tau = +0.08$ (near-random) with 14% top-1 accuracy (random: 12.5%), suggesting that LLMs exhibit the same overgeneralization from standard-game priors that the transfer matrix predicts.

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*, 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 1552–1555, 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
Rook activity	Open file and rank bonuses
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 95% bootstrap confidence intervals for feature marginals (1000 resamples of game results with replacement). In Atomic chess, mobility's CI is entirely above zero ($[+0.15, +0.31]$), confirming its significance. In Standard chess, material's CI is $[+0.12, +0.27]$. All 6 sign changes reported in Section 3.1 have CIs that exclude zero, confirming they 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 Δ = 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 and KotH are the best sources, helping 5 of 6 targets each. Transfer is approximately symmetric for structurally similar pairs (Standard↔Chess960: $-35/-59$) but strongly

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**	—

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

Appendix E. Pairwise Feature Synergy

We compute pairwise synergies $\sigma(a, b)$ 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

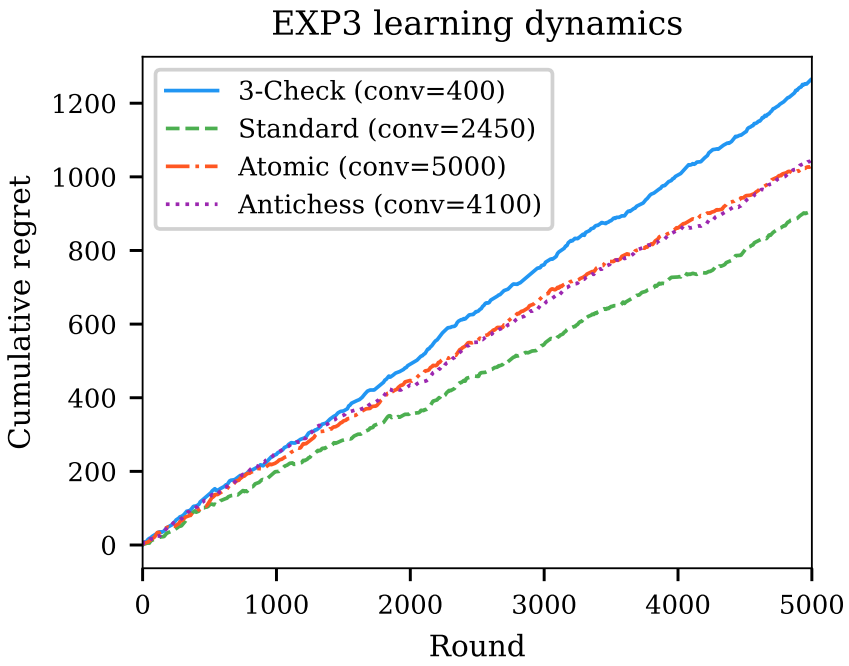

Figure 2: EXP3 cumulative regret for four rule regimes (T=5000). Three-Check converges rapidly (round 400); Atomic does not converge within the horizon. Full convergence ordering: Three-Check (400) $<$ Horde (1700) $<$ KotH (2300) $<$ Standard (2450) $<$ Chess960 (2500) $<$ Antichess (4100) $<$ Atomic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

GPT-4o correctly identifies material as the top feature in Standard chess (the only variant where its training-data priors align with ground truth), but fails in 6 of 7 variants. Its predictions reveal 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%). This mirrors the transfer matrix finding: strategic knowledge, whether from another variant's learned weights or from an LLM's training corpus, does not reliably transfer under rule shifts.