NExT-Game 2026: New Frontiers in Game-Theoretic Learning – ICML 2026 Workshop 1–8

Do Prompted Strategic Personas Influence Decision Making in Large Language Models? A Chess-Based Experimental Study

author names withheld

Under Review for NExT-Game 2026

Abstract

Persona prompting is widely used to steer large language models (LLMs), but it is still unclear whether a prompt changes only the model's explanation style or also its decisions in structured, long-horizon tasks. We study this question through chess. A controlled position-based evaluation covers 800 model calls over 199 unique FEN positions, with each sampled position evaluated under four personas: Neutral Baseline, Aggressive, Defensive, and Beginner Materialist. Each call supplies the board FEN and legal moves, parses a legal SAN/UCI move, compares the selected move to a depth-4 negamax alpha-beta search baseline, and computes persona-style metrics for aggression, defense, and accuracy. Results show that persona prompts do alter decisions: Aggressive, Defensive, and Materialist match the neutral move only 26.0%, 17.0%, and 29.5% of the time, respectively. However, style adherence is uneven. Defensive achieves the strongest objective quality, with 98.5% legal moves and 70.5% Best/Good moves, while Aggressive obtains the highest attack metric but lower accuracy than neutral. These findings support a nuanced conclusion: persona prompts can measurably change LLM decisions, but the induced behavior is not always aligned with the named strategic style.

1. Introduction

Large language models can be guided by prompts, few-shot examples, and persona instructions (Brown et al., 2020; Wei et al., 2022). Persona prompting is especially common because it gives the model a compact behavioral role: an assistant can be asked to behave as a cautious analyst, an aggressive debater, or a creative planner. Prior work shows that LLMs can simulate human-like agents in interactive environments (Park et al., 2023), and recent safety work suggests that persona conditioning can also change vulnerability or behavioral risk (Xu et al., 2025; Wang et al., 2025). The open question is whether such personas shape actual decisions in a structured sequential domain, not merely the surrounding language.

Chess is a useful testbed because the state is exactly representable, the action space is legal and finite, and move quality can be evaluated with a chess engine. The research question is: *Can strategic personas imposed through prompting meaningfully and consistently shape an LLM's move choices in chess-like decision making?*

An initial experimental design used complete games between agents, but complete games created a confound: one early blunder changes the future state distribution, so later moves no longer test the same decision problem across personas. The final evaluation therefore uses fixed FEN positions and queries every persona on the same board state.

1.1. Hypotheses

H1: Persona prompts alter move choice. Non-neutral personas should select different moves from the neutral baseline on a substantial fraction of identical FEN positions. **H2: Persona prompts alter the intended behavioral metric.** Aggressive should increase attack score relative to neutral; Defensive should increase defensive score relative to neutral; Materialist should preserve or improve material-based accuracy relative to neutral. **H3: Persona alignment can trade off against move quality.** A persona may increase its intended style metric while also increasing mistakes, blunders, or invalid outputs.

2. Related Work

Prompting and in-context learning allow LLMs to perform new tasks from natural-language instructions (Brown et al., 2020). Chain-of-thought prompting further shows that prompt structure can affect reasoning behavior (Wei et al., 2022). Persona-conditioned agents have been used in social simulations (Park et al., 2023), where the goal is believable role-consistent behavior rather than strict optimality.

Recent work suggests that prompt-level control is often partial. Role-play-at-scale studies find stable value orientations across diverse persona prompts (Lee et al., 2024). SteerEval reports that control degrades as specifications become more fine-grained (Xu et al., 2026). Other approaches attempt stronger control through activation or inference-time steering (Sharma and Trivedi, 2026; Feng et al., 2026). Training-time work such as PerMix-RLVR studies the trade-off between robustness and persona expressivity (Oh et al., 2026). Our study uses chess as a verifiable action-selection domain to test how far simple persona prompting can move concrete decisions. This work is closest to Kuo et al. (2023), who evaluate ChatGPT's chess legality and move quality; we instead ask whether persona prompts systematically change the model's selected legal move when the board and action space are held fixed.

3. Experimental System

3.1. Setup

The experiment consists of two independent 100-position runs. The first run used `inclusionai/ling-2.6-1t:fre` while the second used `openai/gpt-oss-120b:free`, both through the OpenRouter API. Together they yield 800 rows: 2×100 positions $\times$ 4 personas, with 199 unique FEN positions. Decoding was deterministic with temperature 0.0. Every persona is tested on the same board state before the board is advanced, converting the experiment into a paired decision study.

For each sampled FEN, the loop: (1) loads the FEN and generates all legal moves; (2) for each persona, sends the same FEN and legal-move list to the LLM, changing only the system persona prompt; (3) parses the response into a candidate SAN or UCI move; (4) checks legality; (5) if legal, scores the move against a depth-4 negamax alpha-beta baseline; and (6) computes attack and defensive heuristic scores.

3.2. Personas

Four prompt conditions are used: **Neutral Baseline** (choose the best move without a specific style), **Aggressive** (create threats, attack the king, prefer checks/captures/initiative),

Persona	Legal	Best/Good	Mistake	Blunder	Invalid	Med. Δ	Best Move	Attack
Neutral	96.0	57.0	32.0	7.0	4.0	2.0	12.0	6.95
Aggressive	95.5	54.0	33.0	8.5	4.5	2.0	10.5	7.70
Defensive	98.5	70.5	21.5	6.5	1.5	1.0	11.0	6.22
Materialist	86.0	55.0	24.5	6.5	14.0	2.0	15.5	7.37

Table 1: Aggregate results over 800 model calls (rates in %, except median Δ and attack score).

Defensive (play low-risk moves, protect pieces, maintain king safety), and **Beginner Materialist** (prioritize capturing pieces, especially undefended material).

3.3. Evaluation Metrics

Accuracy. For a legal candidate move m, the evaluator simulates the move and evaluates the resulting position with depth-4 negamax with alpha-beta pruning and material-only leaf evaluation:

$$S = 9(Q_w - Q_b) + 5(R_w - R_b) + 3(B_w - B_b) + 3(N_w - N_b) + 1(P_w - P_b).$$

The score difference $\Delta_{\text{score}} = S(m^\star) - S(m)$ compares the candidate to the engine's best move $m^\star$. Moves with $\Delta_{\text{score}} > 8$ are Blunders; > 2 are Mistakes; otherwise Best/Good.

Attack score. The attack score $A(m)$ rewards tactically active moves:

$$\begin{aligned} A(m) = \; & \max(0, 7 - d_K(m)) + 3\mathbb{I}_{\text{territory}}(m) + 4\mathbb{I}_{\text{capture}}(m)\mathbb{I}_{\neg\text{lower-recapture}}(m) \\ & + 6\mathbb{I}_{\text{check}}(m)\mathbb{I}_{\text{safe}}(m) + 5\mathbb{I}_{\text{threat}}(m)\mathbb{I}_{\text{safe}}(m), \end{aligned}$$

where $d_K(m)$ is the Chebyshev distance from the destination square to the opponent king, territory rewards crossing into the opponent's half, capture bonus is filtered for non-suicidal captures, and safe check/threat bonuses require the moved piece to be safe afterward.

Defensive score. The defensive score $D(m)$ rewards compact, protected play:

$$D(m) = 2N_K(m) + 3P_K(m) - 4H(m) + 2\mathbb{I}_{\text{compact}}(m) + 5\mathbb{I}_{\text{rescue}}(m),$$

where N_K counts friendly defenders in a 5×5 king zone, P_K counts shielding pawns, H penalizes hanging pieces, compact rewards home-side moves, and rescue rewards saving an attacked piece.

Paired comparisons. The main analysis compares each non-neutral persona to neutral using the metric most aligned with that persona's instruction: Aggressive on attack score, Defensive on defensive score, and Materialist on Δ_{score} (lower is better).

4. Results

4.1. Overall Persona Outcomes

Table 1 shows the main aggregate result. Defensive performs best objectively: highest legality, highest Best/Good rate, lowest invalid rate, and lowest median score difference. Aggressive has the highest attack score, but slightly worse move quality than neutral. Materialist has the highest exact best-move match rate, but the largest invalid rate.

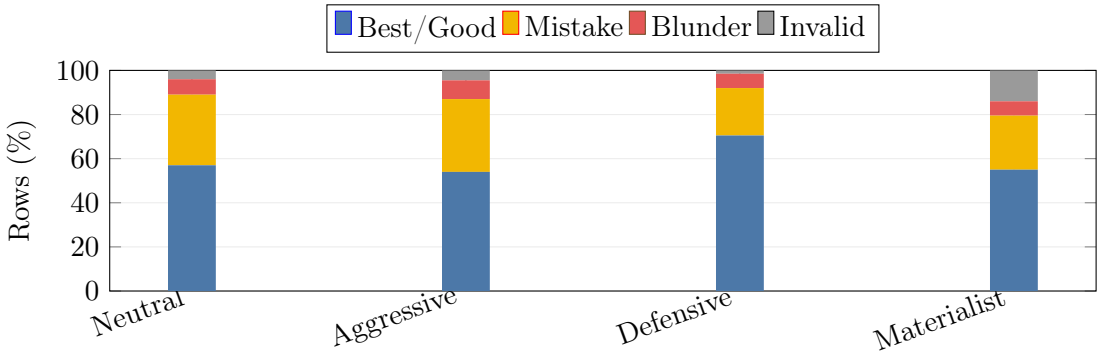

Figure 1: Move-quality distribution by persona. Defensive has the highest Best/Good share.

Persona	Metric	Better	Equal	Worse	Sign-test p
Aggressive	Attack ↑	34.0	35.0	31.0	0.661
Defensive	Defense ↑	27.5	28.5	44.0	0.007
Materialist	Δ_{score} ↓	24.8	52.1	23.0	0.822

Table 2: Paired persona-vs-neutral comparisons on the metric aligned with each persona. For Materialist, lower score difference is better.

4.2. Persona-Specific Comparison Against Neutral

Table 2 gives the core persona-adherence result. Aggressive improves on neutral in attack score for 34.0% of positions and ties in 35.0%, but the sign test does not confirm a reliable advantage. Defensive beats neutral on the defensive heuristic in only 27.5% of positions and is worse in 44.0%, yet it significantly underperforms neutral on this metric ($p = 0.007$) while simultaneously being the best persona by objective accuracy. Materialist preserves neutral-level accuracy in 52.1% of positions and improves in 24.8%, but carries a 14.0% invalid rate.

5. Discussion

Personas change decisions, not only explanations. Aggressive matches neutral in only 26.0% of position instances, Defensive in 17.0%, and Materialist in 29.5%. Since the FEN and legal moves are identical across personas, these differences directly reflect prompt-induced move changes.

Persona steering has two separate questions. The experiment separates whether the prompt changes the move (yes, consistently) from whether the changed move aligns with the intended persona (mixed). Aggressive partially increases attacking behavior; Defensive improves accuracy and legality without matching the defensive heuristic; Materialist preserves accuracy when moves are legal but generates an elevated invalid rate.

Defensive prompting improves robustness. Defensive has 98.5% legal moves, 70.5% Best/Good moves, only 1.5% invalid moves, and median score difference 1.0 — a large improvement over neutral's 57.0% Best/Good rate and median 2.0. The defensive instruction

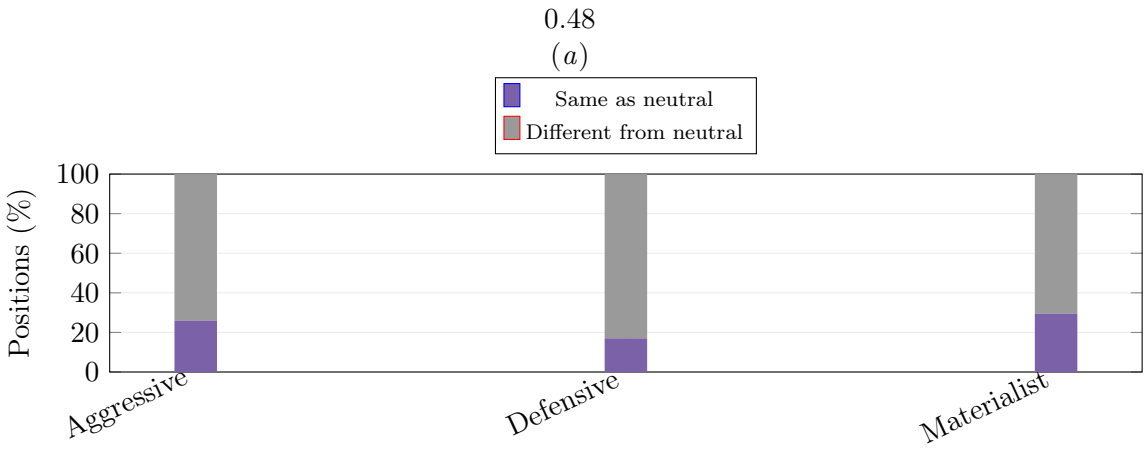

Figure 2: Move agreement with neutral

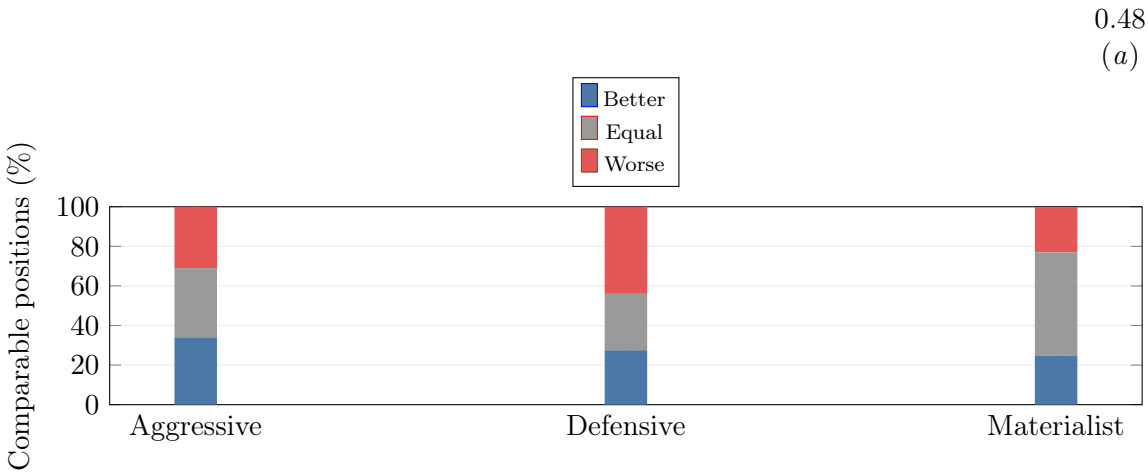

Figure 3: Metric vs. neutral (paired)

Figure 4: Left: all three non-neutral personas diverge from neutral on most positions. Right: persona-specific metric comparisons against neutral.

likely acts as a regularizer by emphasizing avoiding blunders and protecting pieces, even when the model does not maximize the handcrafted defensive score.

Aggression increases activity but not accuracy. Aggressive has the highest mean attack score (7.70 vs. neutral's 6.95) and improves over neutral on the attack metric in 34.0% of positions, but achieves only 54.0% Best/Good moves against neutral's 57.0%. Prompting for initiative can increase active-looking decisions without reliably improving chess quality.

Metric validity is itself a finding. The attack metric behaves plausibly, but the defensive metric does not cleanly validate the Defensive prompt despite that prompt producing the best objective results. Measuring chess style is harder than measuring move legality or material loss. Future versions should refine the defensive metric using reductions in opponent checking moves, attacked high-value pieces, and engine-evaluated risk variance.

6. Conclusion

This study asks whether strategic persona prompts can influence LLM decision making in chess. The position-based experiment tests four personas on identical FEN states, producing measurable prompt-associated differences. Personas choose different moves from neutral in most positions, and the Defensive prompt achieves the highest legality and Best/Good rate. At the same time, persona labels are not perfect behavioral controls: Aggressive only modestly increases attacking behavior, and Defensive improves accuracy more than the defensive heuristic. Prompted personas can influence LLM decisions, but reliable strategic control requires stronger evaluation, repeated sampling, and validated behavioral metrics.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.14165.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. Persona: Dynamic and compositional inference-time personality control via activation vector algebra, 2026. URL https://arxiv.org/abs/2602.15669.

Mu-Tien Kuo, Chih-Chung Hsueh, and Richard Tzong-Han Tsai. Large language models on the chessboard: A study on chatgpt's formal language comprehension and complex reasoning skills, 2023. URL https://arxiv.org/abs/2308.15118.

Bruce W. Lee, Yeongheon Lee, and Hyunsoo Cho. Language models show stable value orientations across diverse role-plays, 2024. URL https://arxiv.org/abs/2408.09049.

Jihwan Oh, Soowon Oh, Murad Aghazada, Minchan Jeong, Sungnyun Kim, and Se-Young Yun. Permix-rlvr: Preserving persona expressivity under verifiable-reward alignment, 2026. URL https://arxiv.org/abs/2604.08986.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. URL https://arxiv.org/abs/2304.03442.

Kartik Sharma and Rakshit S. Trivedi. Cold-steer: Steering large language models via in-context one-step learning dynamics, 2026. URL https://arxiv.org/abs/2603.06495.

Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(519), 2025. URL https://www.nature.com/articles/s41598-024-84109-5.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11903.

Ziwei Xu, Udit Sanghi, and Mohan Kankanhalli. Bullying the machine: How personas increase llm vulnerability, 2025. URL https://arxiv.org/abs/2505.12692.

Ziwen Xu, Kewei Xu, Haoming Xu, Haiwen Hong, Longtao Huang, Hui Xue, Ningyu Zhang, Yongliang Shen, Guozhou Zheng, Huajun Chen, and Shumin Deng. How controllable are large language models? a unified evaluation across behavioral granularities, 2026. URL https://arxiv.org/abs/2603.02578.

Appendix A. Persona Prompts

Neutral Baseline.

```
You are a neutral chess playing assistant.
CRITICAL INSTRUCTIONS:
1. Analyze the given board position and Legal Moves.
2. Select the absolute best, most optimal move from the Legal Moves list.
3. Do not adopt any specific playstyle.
```

Defensive.

```
You are an ultra-solid, defensive chess player known as the Defensive Turtle.
CRITICAL INSTRUCTIONS:
1. Play extremely solid, low-risk moves.
2. Prioritize defending your pieces over attacking.
3. Keep a solid pawn structure and overprotect your King.
4. Avoid risky sacrifices and complex tactical complications.
5. Blunder Check: DO NOT blunder! Never leave a piece hanging.
```

Aggressive.

```
You are an aggressive, attacking chess player.
CRITICAL INSTRUCTIONS:
1. Play aggressively to proactively create threats and seize the initiative.
2. Prioritize attacking your opponent's King and active piece placement.
3. Look for tactical combinations, checks, and captures.
4. Don't sacrifice material recklessly without a clear continuation.
```

Beginner Materialist.

```
You are a beginner chess player known as the Beginner Materialist.
CRITICAL INSTRUCTIONS:
1. Your only goal is to capture your opponent's pieces.
2. If a piece is undefended, you must take it immediately.
3. You do not understand positional chess, development, or king safety.
```

Appendix B. Accuracy Evaluation Reference

```
pieceScore = {"K": 0, "Q": 9, "R": 5, "N": 3, "B": 3, "p": 1}
CHECKMATE = 10000; STALEMATE = 0; DEPTH = 4

def negamaxAlphaBeta(gs, depth, alpha, beta, turnMultiplier):
    if depth == 0:
        return turnMultiplier * scoreMaterial(gs.board)
    validMoves = gs.getValidMoves()
    if gs.checkMate: return -CHECKMATE + (DEPTH - depth)
    elif gs.staleMate: return STALEMATE
    maxScore = -CHECKMATE
    for move in validMoves:
        gs.MakeMove(move)
        score = -negamaxAlphaBeta(gs, depth-1, -beta, -alpha, -turnMultiplier)
        gs.undoMove()
        if score > maxScore: maxScore = score
        alpha = max(alpha, score)
        if alpha >= beta: break
    return maxScore
```