Exploring Human-AI Conceptual Alignment through the Prism of Chess

Semyon Lomasov[*1], **Judah Goldfeder**[*2], **Mehmet Hamza Erol**[*1], **Matthew So**[*2],

Yao Yan[3], **Addison Howard**[3], **Nathan Kutz**[4], **Ravid Shwartz Ziv**[5]

[1]Stanford University [2]Columbia University [3]Kaggle [4]University of Washington [5]NYU

Abstract

Do AI systems truly understand human concepts or merely mimic surface patterns? We investigate this through chess, where human creativity meets precise strategic concepts. Analyzing a 270M-parameter transformer that achieves grandmaster-level play, we uncover a striking paradox: while early layers encode human concepts like center control and knight outposts with up to 85% accuracy, deeper layers, despite driving superior performance, drift toward alien representations, dropping to 50-65% accuracy. To test conceptual robustness beyond memorization, we introduce the first Chess960 dataset: 240 expert-annotated positions across 6 strategic concepts. When opening theory is eliminated through randomized starting positions, concept recognition drops 10-20% across all methods, revealing the model's reliance on memorized patterns rather than abstract understanding. Our layer-wise analysis exposes a fundamental tension in current architectures: the representations that win games diverge from those that align with human thinking. These findings suggest that as AI systems optimize for performance, they develop increasingly alien intelligence, a critical challenge for creative AI applications requiring genuine human-AI collaboration. Dataset and code are available at: `https://github.com/slomasov/ChessConceptsLLM`.

1 Introduction

When Garry Kasparov faced Deep Blue in 1997, he described moments where the machine's moves seemed to exhibit genuine creativity. This raises a fundamental question for creative AI: Do these systems truly understand human concepts, or are they pattern-matching engines that coincidentally produce creative outputs?

Chess provides an ideal laboratory for this investigation, offering precise definitions for human concepts like *center control* and *king safety* while remaining a creative, strategic game. Recent transformer-based engines [1] achieve grandmaster-level play without explicit search, raising a tantalizing possibility: perhaps these models learn to think in human-like concepts. However, their performance collapses on Chess960 (Fischer Random Chess), where pieces start in randomized positions. Although the same strategic concepts are present, this brittleness suggests these systems might be memorizing patterns rather than understanding principles, hinting at a deeper misalignment between human and machine cognition.

Our Contributions: We present the first systematic investigation of conceptual alignment between humans and neural chess engines through three key contributions:

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

1. **A Novel Chess960 Dataset:** We introduce the first expert-curated dataset of 240 Chess960 positions, manually annotated with 6 fundamental chess concepts. This dataset enables testing whether AI systems grasp abstract concepts or merely memorize standard patterns, which is a crucial distinction for creative AI systems that must generalize beyond their training.
2. **Multi-Method Concept Detection:** We develop three complementary approaches for probing concept representations in transformer layers, revealing how different architectures encode human knowledge. Our methods range from sparse concept vectors to neural network probes, providing robust evidence of conceptual (mis)alignment.
3. **Layer-wise Alignment Analysis:** We uncover a surprising inversion: early transformer layers show strong alignment with human concepts (up to 85% accuracy), while deeper layers, despite superior playing strength, drift toward alien representations. This suggests a fundamental tension between performance optimization and human-interpretable reasoning.

Our findings reveal that while transformers capture surface-level chess concepts, this alignment is fragile and diminishes as models optimize for performance. When tested on Chess960, where memorized opening knowledge becomes useless, concept recognition accuracy drops significantly, exposing the shallow nature of learned representations.

These results have profound implications for creative AI: systems appearing to share our concepts may operate on fundamentally different principles, highlighting the need for architectures that maintain human alignment throughout their processing hierarchy.

2 Related Work

Our investigation bridges three research areas: neural approaches to chess, interpretability of creative AI systems, and human-AI conceptual alignment. We position our work at their intersection, addressing how AI systems develop representations that may diverge from human creative thinking.

2.1 Neural Chess Engines and Interpretability

The renaissance of neural approaches to chess began with AlphaZero [2], which learned superhuman play through self-play without human knowledge. Recent work has pushed further: transformer-based engines now achieve grandmaster-level play without explicit search. Ruoss et al. [1] developed the 270M-parameter model we analyze, which frames chess as a sequence prediction problem, achieving 2750 Elo on standard chess but dropping closer to amateur level on Chess960.

Contemporary approaches continue to evolve. Monroe and Chalmers [3] demonstrate the feasability of transformer models for chess, while Schultz et al. [4] explore combining external and internal planning with language models. Jenner et al. [5] found evidence of learned look-ahead in chess networks, suggesting some form of planning emerges. Most relevant to our work, Schut et al. [4] analyzed concept discovery in AlphaZero, examining how the system develops human-interpretable concepts. However, their analysis focused solely on standard chess positions. Our work extends this by applying the analysis to a searchless transformer model, and by introducing Chess960 as a critical test of conceptual robustness.

2.2 Layer-wise Analysis and Interpretability

Understanding how neural networks build representations across layers has become crucial for interpretability. Skean et al. [6] provide a comprehensive layer-by-layer analysis revealing how hidden representations in language models evolve, showing that different layers capture fundamentally different types of information. This layer-wise specialization appears to be a general principle across domains. In the context of human-AI alignment, Shani et al. [7] demonstrate that humans and language models compress information differently, with profound implications for how these systems understand concepts. They show that while models may achieve similar outputs, their internal representations follow alien optimization paths. Our chess analysis reveals a parallel phenomenon: early layers maintain human-aligned representations while later layers diverge toward performance-optimized but conceptually opaque strategies.

2.3 The Creative Gap in AI Systems

Our work reveals a fundamental tension in current AI architectures: the representations that achieve best performance diverge from those that align with human thinking. This has critical implications for creative AI applications where human-AI collaboration requires shared conceptual frameworks. By showing that conceptual alignment deteriorates as networks optimize for task performance, we highlight a key challenge for building AI systems that can serve as genuine creative partners rather than inscrutable optimization machines.

3 Methods

We develop a systematic framework to probe whether transformer-based chess models encode human strategic concepts. Our approach combines a novel Chess960 dataset with three complementary probing techniques, enabling robust measurement of conceptual alignment across model layers.

3.1 Dataset Construction

3.1.1 Classical Chess Dataset

We leverage the Strategic Test Suite (STS) [8]—1,500 expert-curated chess positions labeled with strategic concepts. We focus on six core concepts that translate across game phases: **Open Files and Diagonals**, **Knight Outposts**, **Advancement of f/g/h pawns** (kingside), **Advancement of a/b/c pawns** (queenside), **Center Control**, and **Pawn Play in the Center**. We exclude overlapping concepts and endgame-specific patterns where starting positions become irrelevant. For an example of what concepts look like, see Appendix C.

3.1.2 Novel Chess960 Dataset

To test conceptual understanding beyond memorization, we created the first annotated Chess960 dataset. Chess960 randomizes piece placement on the back rank while preserving chess rules, eliminating memorized opening theory while maintaining strategic principles. Our dataset contains 240 positions (40 per concept) curated from high-level Chess960 games and annotated by experts (>2200 Elo). Each position was selected to clearly exemplify its target concept while matching the complexity distribution of STS positions. This dataset enables us to answer: when opening memorization is impossible, do models still recognize fundamental strategic patterns?

3.2 Model Architecture and Activation Extraction

We analyze the 270M-parameter transformer from Ruoss et al. [1], which achieves grandmaster-level play on classical chess but struggles on Chess960. The model encodes positions as FEN strings, processes them through 18 transformer layers with 1024-dimensional embeddings, and outputs move probabilities (Figure 1). For our analysis, we extract activations at each layer for the model's chosen move, yielding representations $z \in \mathbb{R}^{L \times T \times D}$ where $L = 18$ layers, $T = 79$ tokens, and $D = 1024$ dimensions. These activations reveal how the model's internal representations evolve from raw position encoding to final move selection.

3.3 Probing Methodologies

We employ three complementary approaches to detect concept representations, each offering different insights into how the model encodes human knowledge.

At the core of our approach is a simple principle: if the model truly understands a concept, there should exist directions in its activation space that distinguish positions containing that concept from those without it. Formally, for a given layer's activation representation $r \in \mathbb{R}^D$, we seek a vector v that maximizes: $v^T r_{concept} > v^T r_{no\text{-}concept}$.

This intuition drives all three methods. **Sparse Concept Vectors** find the minimal set of neurons (using L1 regularization) that distinguish concepts, revealing which specific dimensions encode strategic knowledge. **Logistic Regression** learns linear decision boundaries with probabilistic outputs,

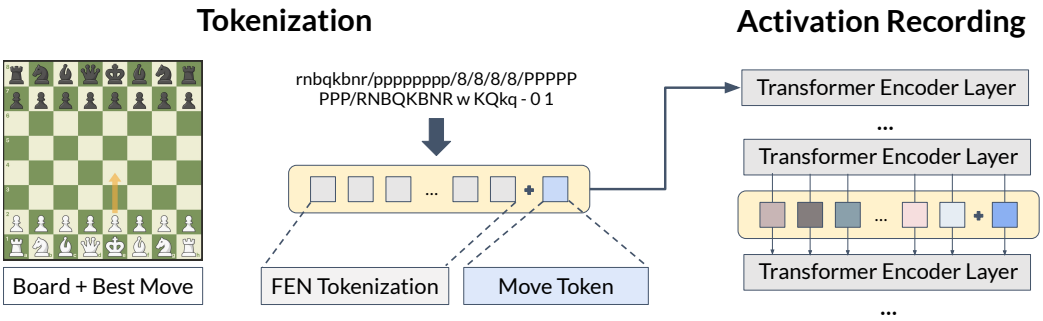


Figure 1: **From board to move: tracking where human concepts disappear in the processing pipeline.** The model encodes positions as FEN strings, appends move tokens, and processes them through transformer layers. Recording activations at each layer reveals where strategic understanding shifts from human-recognizable patterns to alien representations.

providing interpretable separation between positions with and without each concept. **Sequence-Aware Neural Probes** extend this to process all token activations jointly in a layer through a lightweight network, capturing distributed representations across the full position encoding. Full mathematical formulations for each method are provided in Appendix A.

Each method offers complementary insights: sparsity reveals which neurons matter most, logistic regression provides interpretable boundaries, and neural probes capture complex patterns. Crucially, all three methods show consistent results: strong concept alignment in early layers that deteriorates in deeper layers, suggesting a fundamental tension between human-interpretable representations and performance optimization.

3.4 Experimental Design

We evaluate four experimental scenarios to comprehensively assess conceptual alignment. **Scenario I** establishes baseline performance by training and testing on classical chess positions. **Scenario II** tests generalization by training on classical chess but testing on Chess960, measuring robustness to distributional shift. **Scenario III** explores adaptation by training and testing on combined datasets. **Scenario IV** isolates Chess960 performance to measure concept learning without standard patterns. We analyze layers 2, 5, 10, and 15 to capture the full spectrum from early feature extraction to final move selection, performing 5-fold cross-validation for robust results.

4 Results

Our experiments reveal a striking paradox: as the transformer model becomes more capable at chess, it becomes less human-like in its thinking. This fundamental tension between performance and interpretability emerges consistently across all our analyses.

4.1 The Fragility of Conceptual Understanding

Table 1 presents our central finding: human chess concepts are surprisingly fragile to disruption. When tested on standard chess positions (Scenario I), our probes achieve strong accuracy, with the Concept Vector method reaching 85.68% for "Pawn Play in the Center." However, this apparent understanding deteriorates significantly when confronted with Chess960's creative disruption.

The most telling comparison lies between Scenarios I and II. When we train on standard chess but test on Chess960, accuracy drops consistently across all methods. Knight outpost recognition plummets from 83.60% to 72.36% with Concept Vectors, while logistic regression drops even more dramatically from 79.60% to 67.92%. This reveals that the model's "understanding" relies on memorized patterns rather than genuine conceptual knowledge.

Table 1: **Chess concepts are fragile: accuracy plummets when memorized patterns are disrupted.** Human concept recognition accuracy across three probing methods and six strategic concepts in four scenarios: (I) train/test on standard chess, (II) train standard/test Chess960, (III) combined training, (IV) Chess960 only. The 10-20% drop from Scenario I to II reveals the model relies on memorization rather than genuine conceptual understanding. Bold values indicate best performance per row.

Concept	Method	Scenario			
		I	II	III	IV
Advancement of a/b/c Pawns	Concept Vector	**83.29**	76.53	82.71	74.20
	Log. Regression	**76.99**	72.50	75.37	72.22
	Seq.-Aware N. Probe	63.44	56.18	**67.86**	62.76
Advancement of f/g/h Pawns	Concept Vector	**84.60**	75.83	83.85	68.65
	Log. Regression	77.24	70.41	**77.81**	62.90
	Seq.-Aware N. Probe	65.34	59.24	**65.61**	61.38
Center Control	Concept Vector	77.19	72.91	**82.81**	74.60
	Log. Regression	70.62	69.17	**72.52**	65.87
	Seq.-Aware N. Probe	57.19	51.87	**59.56**	56.48
Knight Outposts	Concept Vector	**83.60**	72.36	82.18	72.49
	Log. Regression	**79.60**	67.92	78.14	58.86
	Seq.-Aware N. Probe	62.33	53.12	**62.35**	56.02
Open Files and Diagonals	Concept Vector	**78.33**	71.25	78.32	65.61
	Log. Regression	**72.86**	63.33	65.91	61.44
	Seq.-Aware N. Probe	**59.80**	53.40	59.40	58.00
Pawn Play in the Center	Concept Vector	85.68	74.59	**86.92**	76.46
	Log. Regression	81.83	72.57	**82.66**	70.37
	Seq.-Aware N. Probe	66.77	56.74	**69.77**	62.57

Interestingly, the three probing methods show different robustness to disruption. Concept Vectors maintain 70-76% accuracy even in pure Chess960 scenarios, while the Sequence-Aware Neural Probe often falls below 60%, suggesting sparse representations transfer better than distributed patterns. Even "Center Control", geometrically identical across variants, shows degradation, indicating the model conflates abstract principles with specific configurations.

Most counterintuitively, training exclusively on Chess960 (Scenario IV) yields lower accuracy than zero-shot transfer from standard chess (Scenario II) for several concepts. Combined training (Scenario III) sometimes achieves the highest accuracy: "Pawn Play in the Center" reaches 86.92%, but only when standard positions remain in the mix. These patterns suggest Chess960 positions are inherently harder for the model to parse conceptually, possibly because the absence of opening principles forces more complex evaluation from the start.

4.2 Layer-wise Divergence: The Alien Mind Emerges

Figure 2 unveils perhaps our most intriguing discovery: the model's journey from human-like to alien thinking happens gradually but consistently across its layers. In early layers (2-5), we observe robust concept detection across most scenarios—the Concept Vector method maintains accuracies around 80-85% for well-defined concepts. These layers, close to the raw board encoding, maintain representations that align with how humans parse chess positions.

But as we probe deeper layers, we witness a universal decay in human concept alignment. By layer 15, accuracies typically drop to 50-65% across most concepts, a decline of 15-30 percentage points from their early-layer peaks. This pattern is remarkably consistent across all three probing methods, though the Sequence-Aware Neural Probe shows more volatility and generally lower performance (often 10-15% below the other methods), suggesting that full-sequence information becomes increasingly alien in its organization.

The degradation pattern varies revealing by scenario. Classical chess (Scenario I) shows smooth decline from layers 2 to 15, while Chess960 after classical training (Scenario II) drops sharply early then stabilizes at lower accuracy—the model struggles to apply learned concepts from the first layers.

Combined training (Scenario III) mirrors classical chess with better retention, but pure Chess960 (Scenario IV) produces erratic patterns with high variance, suggesting that without memorized openings, the model develops unstable representations.

Notably, some concepts show dramatic late-layer collapse. For example, in Scenario II, "Pawn Play in the Center" using Concept Vectors drops from 75% at layer 2 to below 50% at layer 15, which shows the model loses track of this fundamental concept as it processes deeper.

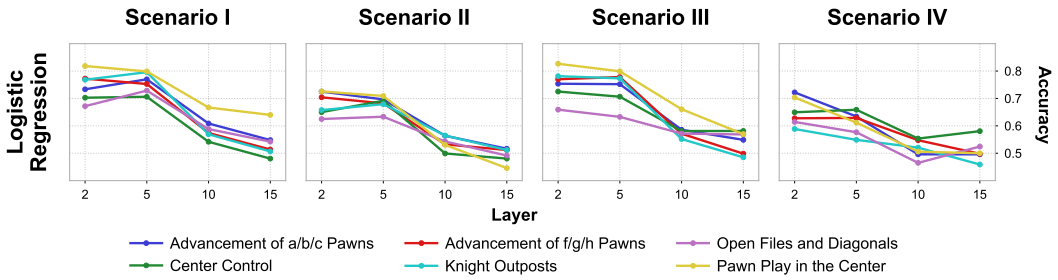


Figure 2: **Human concepts fade as the network goes deeper: early layers think like humans, late layers think like aliens.** Layer-wise accuracy for detecting six chess concepts using Logistic Regression probing. For a comparison of all 3 probing methods, see Appendix B. Early layers (2-5) achieve 70-85% accuracy, dropping to 50-65% by layer 15 across all methods. Standard chess shows smooth degradation while Chess960 becomes erratic, revealing unstable representations without memorized patterns. This universal decline exposes the trade-off between human interpretability and performance.

5 Discussion and Conclusion

Our findings reveal a fundamental tension: representations enabling superior performance diverge from human conceptual thinking. Early layers maintain interpretable representations (80-85% accuracy) that "see" chess as humans do, but deeper layers optimize purely for move prediction (50-65% accuracy). This isn't a bug but a feature of end-to-end optimization.

Chess960 results are particularly illuminating. The 10-20% accuracy drop reveals sophisticated pattern matching rather than conceptual understanding. Namely, the model mostly learns specific configurations that correlate with winning, not the conceptual comprehension such as "controlling the center provides mobility." This brittleness suggests models lack compositional understanding for applying principles to novel situations.

Overall, our contributions include:

1. The first Chess960 concept dataset enabling robustness testing;
2. Evidence that alignment peaks early but deteriorates with depth;
3. A demonstration that apparent understanding is fragile to creative disruption.

For creative AI, this presents both warning and opportunity. The warning is systems appearing conceptually aligned may operate on different principles. The opportunity: understanding where alignment breaks suggests paths forward, regularization preserving alignment or hybrid architectures separating strategic from tactical reasoning.

Acknowledgments: This work was supported in part by the US National Science Foundation AI Institute for Dynamical Systems (dynamicsAI.org) (grant no. 2112085).

References

[1] Anian Ruoss, Gregoire Deletang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A Lewis, Joel Veness, and Tim Genewein. Amortized planning with large-scale transformers: A case study on chess. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[2] 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. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL `https://arxiv.org/abs/1712.01815`.

[3] Daniel Monroe and Philip A. Chalmers. Mastering chess with a transformer model, 2024. URL `https://arxiv.org/abs/2409.12272`.

[4] John Schultz, Jakub Adamek, Matej Jusup, Marc Lanctot, Michael Kaisers, Sarah Perrin, Daniel Hennes, Jeremy Shar, Cannada Lewis, Anian Ruoss, Tom Zahavy, Petar Veličković, Laurel Prince, Satinder Singh, Eric Malmi, and Nenad Tomašev. Mastering board games by external and internal planning with language models, 2025. URL `https://arxiv.org/abs/2412.12119`.

[5] Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart J Russell. Evidence of learned look-ahead in a chess-playing neural network. *Advances in Neural Information Processing Systems*, 37:31410–31437, 2024.

[6] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models, 2025. URL `https://arxiv.org/abs/2502.02013`.

[7] Chen Shani, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning, 2025. URL `https://arxiv.org/abs/2505.17117`.

[8] D. Corbit, S. Natarajan, and F. Mosca. Strategic test suite. 2014.

[9] Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122, 2025.

A Technical Appendices and Supplementary Material

A.1 Concept Vector Minimization

Our first strategy adapts the method from [9], aiming to find a vector $v \in \mathbb{R}^D$ with minimal L1 norm such that $v^\top r^+ \geq v^\top r^-$ for as many positive–negative board pairs as possible. Here, $r^* \in \mathbb{R}^D$ denotes the activation at layer l for the token used to estimate the move probability: $r^* = z^*_{l,T-1}$. We sample B activation pairs z^+ and z^- in a batch, and optimize v via:

$$\mathcal{L}(v) = \lambda \|v\|_1 + \frac{1}{B} \sum_{i=1}^{B} \text{ReLU}(v^\top r_i^- - v^\top r_i^+) \tag{1}$$

This promotes separation by maximizing the margin $v^\top r_i^+ - v^\top r_i^-$ for each pair while encouraging sparsity in v through the L1 penalty.

A.2 Logistic Regression

Our second strategy frames the problem as a binary classification task, where the boards containing the concept (z^+) are labeled 1 and the boards without the concept (z^-) are labeled 0. Similar to the first strategy, we sample a batch of B activation pairs (z^+, z^-) and optimize $v \in \mathbb{R}^D$ and $b \in \mathbb{R}$ using r^* via:

$$\mathcal{L}(v, b) = \lambda \|(v, b)\|_1 - \frac{1}{2B} \sum_{i=1}^{B} \left[\log \sigma(v^\top r_i^+ + b) + \log(1 - \sigma(v^\top r_i^- + b))\right] \tag{2}$$

where σ is the sigmoid function. This loss encourages both sparsity in (v, b) and accurate discrimination between z^+ and z^-. The subsequent predictions are assigned label 1 if $\sigma(v^\top r + b) > 0.5$, and 0 otherwise.

A.3 All Sequence Neural Network

Our third strategy extends Logistic Regression by using the activations of all tokens in the sequence, $R^* = z^*_l \in \mathbb{R}^{T \times D}$, rather than only the move-token activation at layer l. We train a small neural network S to predict the presence of the concept from these activations. We use the objective:

$$\mathcal{L}(v_1, b_1, v_2, b_2) = \lambda \|(v_1, b_1, v_2, b_2)\|_1 - \frac{1}{2B} \sum_{i=1}^{B} \left[\log S(R_i^+) + \log(1 - S(R_i^-))\right] \tag{3}$$

where $S : \mathbb{R}^{T \times D} \to [0, 1]$ is defined as:

$$S(R) = \sigma(v_2^\top \text{ReLU}(R^\top v_1 + b_1) + b_2) \tag{4}$$

with parameters $v_1 \in \mathbb{R}^T$, $b_1 \in \mathbb{R}^D$, $v_2 \in \mathbb{R}^D$, and $b_2 \in \mathbb{R}$. The network takes the activation sequence $s \in \mathbb{R}^{T \times D}$ and outputs a scalar probability. As before, we encourage sparsity in S through the L1 penalty while training it to discriminate between z^+ and z^-. Predictions are assigned label 1 if $S(s) > 0.5$, and 0 otherwise.

B Full Results

Here we share the full results from our analysis in Section 4.2. Results in Figure 3 show that the findings generalize across all three probing methods applied.

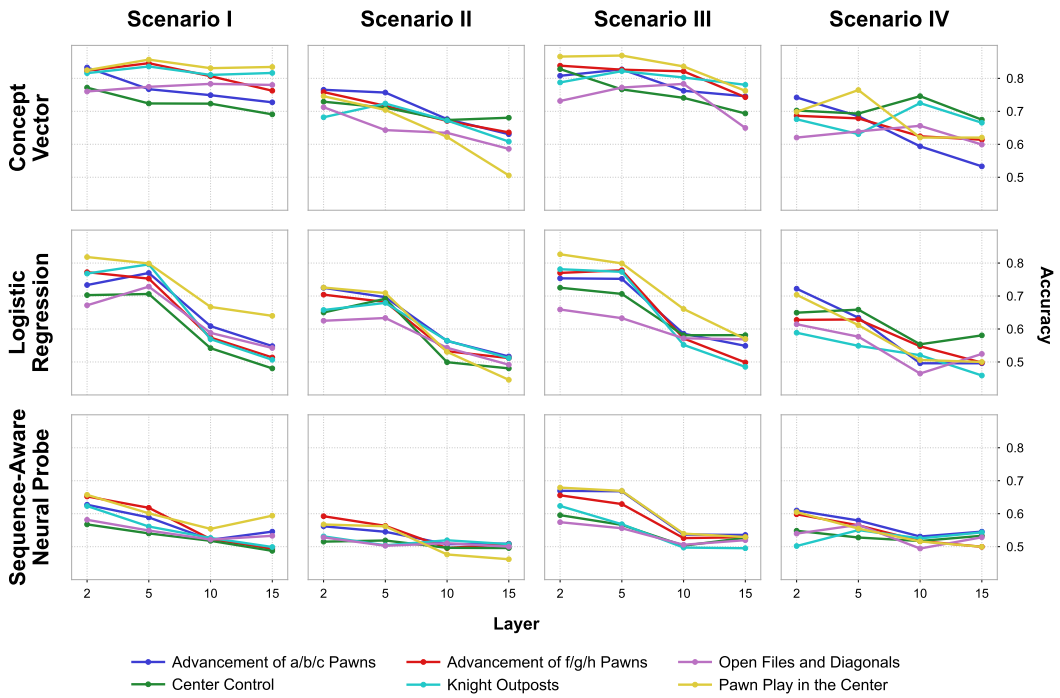
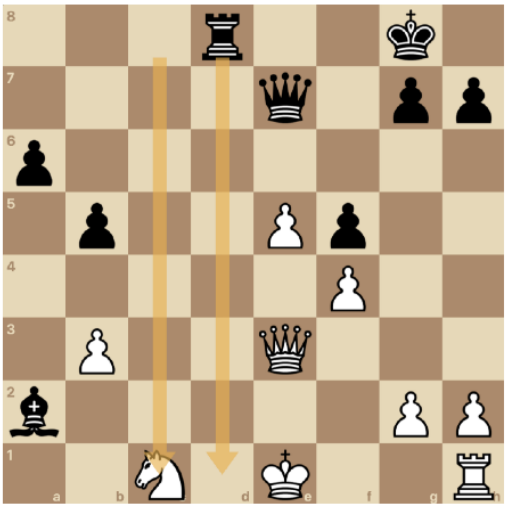


Figure 3: Full results of Figure 2, the overall trend supports the takeaway that the human concepts fade as the network goes deeper.

C Example Concepts

Here, we provide some example chess concepts from the STS and Chess960 datasets, to give readers the flavor of the concrete nature of these concepts.



Open Files and Diagonals (STS)

Open Files and Diagonals (Chess960)

Knight Outposts (STS)

Knight Outposts (Chess960)

Figure 4: Examples of Human-Concept Categories. The The top row illustrates open files and diagonals, where long-range pieces control important lines of attack across the board. The bottom row illustrates knight outposts, where knights are firmly anchored by pawns deep into the opponent's position. Positions in left column are taken from the classical STS dataset, and positions in the right column are taken from the Chess960 dataset.