# Iterative Inference in a Chess-Playing Neural Network

**Elias Sandmann**[*]
Fraunhofer HHI

**Sebastian Lapuschkin**[*]
Fraunhofer HHI
TU Dublin

**Wojciech Samek**[*]
Fraunhofer HHI
TU Berlin

## Abstract

Do neural networks build their representations through smooth, gradual refinement, or via more complex computational processes? We investigate this by extending the logit lens to analyze the policy network of Leela Chess Zero, a superhuman chess engine. Although playing strength and puzzle-solving ability improve consistently across layers, capability progression occurs in distinct computational phases with move preferences undergoing continuous reevaluation—move rankings remain poorly correlated with final outputs until late, and correct puzzle solutions found in middle layers are sometimes overridden. This late-layer reversal is accompanied by concept preference analyses showing final layers prioritize safety over aggression, suggesting a mechanism by which heuristic priors can override tactical solutions.

## 1 Introduction

How do neural networks progressively build understanding as information flows through their layers? Do they incrementally refine representations by gradually increasing confidence in the correct answer, or do they fundamentally recompute preferences at each layer? Theoretical work has formalized the view that residual networks perform iterative inference, empirically resulting in higher accuracy across layers when intermediate representations are decoded through the model's classifier (Jastrzębski et al., 2018). The logit lens (nostalgebraist, 2020) applies this approach to transformer-based language models, projecting hidden states through the unembedding matrix revealing similar dynamics with each layer achieving systematically lower perplexity (Belrose et al., 2023). However, the extent to which this iterative process involves gradual heuristic accumulation or discrete algorithmic computation—and how these mechanisms interact—remains an open question.

We examine these questions by extending the logit lens to analyze the policy network of Leela Chess Zero (Leela Chess Zero team), an open-source AlphaZero (Silver et al., 2018) reimplementation that achieves strong play even without external search (lepned, 2024). Leela's architecture provides distinct interpretability advantages: unlike decoder-only language models, it uses all tokens simultaneously for prediction, providing complete observability of activations causally affecting outputs and enabling faithful intermediate decoding. Prior work provided initial insights into this model's inference process by demonstrating that it internally implements learned look-ahead (Jenner et al., 2024), yet it remains unclear whether algorithmic structure alone fully characterizes its behavior. Our analysis provides additional evidence for algorithmic computation, showing that move preferences are repeatedly reorganized across layers rather than refined smoothly (see Figure 1 and Appendix G). However, correct puzzle solutions discovered in middle layers—including forced checkmates—are sometimes overridden by seemingly safer alternatives in final layers. Concept preference analyses indicate that later layers systematically favor safer positions, suggesting that safety-oriented heuristics may contribute to these reversals. Together, these findings reveal that Leela's inference process combines algorithmic computation with learned heuristic priors, offering a concrete case study for understanding iterative inference in a structured decision-making domain.

[*] `firstname.lastname@hhi.fraunhofer.de`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

Our contributions are: (1) We extend the logit lens technique to Post-LN transformer architectures. (2) We demonstrate that iterative inference in Leela combines algorithmic computation with learned heuristic priors with capability advancing through distinct phases. (3) We introduce concept preference analysis as an alternative to representation probing. Our code is available on GitHub.

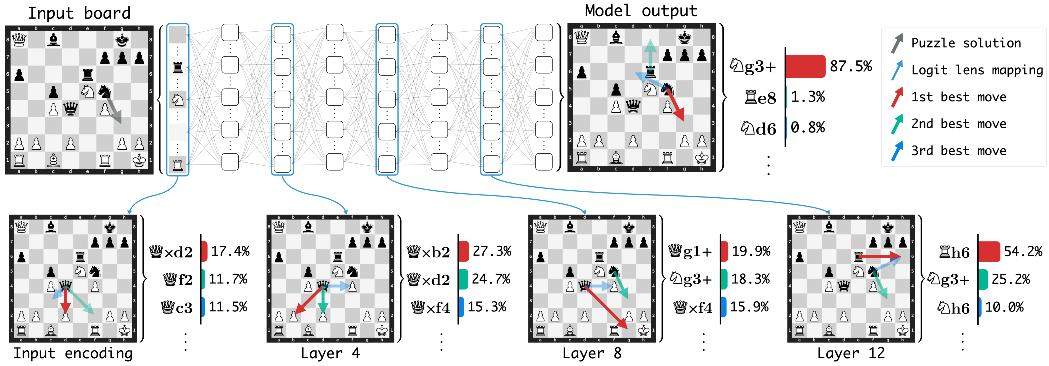


Figure 1: Our extended logit lens reveals progressive policy refinement across transformer layers in Leela Chess Zero. We map intermediate activations to policy distributions for a tactical puzzle. The model's top-ranked move changes at each stage, with the correct solution **♘g3+** only emerging as a plausible candidate in the middle layers before becoming the decisive top choice in the final output. Full probabilities are provided in Appendix F and additional examples in Appendix G.

## 2 Methodology

### 2.1 Model architecture

We analyze the `T82-768x15x24h` transformer model from Leela Chess Zero, the strongest neural chess engine available today (Jenner et al., 2024). This model uses a Post-LN architecture similar to the original transformer (Vaswani et al., 2017) with DeepNorm scaling (Wang et al., 2022), featuring a 15-layer transformer encoder with 768-dimensional embeddings and specialized output heads. Chess positions are encoded as $8 \times 8$ grids where each square corresponds to a token position.

Leela is trained using the AlphaZero paradigm and normally functions in tandem with MCTS as a chess engine. However, we focus solely on the policy network, which already demonstrates strong chess-playing ability even without external search. Architectural details are provided in Appendix C.

### 2.2 Encoder-only Post-LN logit lens

The logit lens projects intermediate activations after layer $\ell$ through the final layer normalization and unembedding matrix to obtain layer-wise predictions. This approach works seamlessly for Pre-LN transformers, where layer normalization precedes each sublayer and leaves the residual stream free of normalization operations. Post-LN architectures create a challenge by applying normalization after residual connections instead, transforming the residual stream at each layer and creating non-linear dependencies that prevent straightforward application of the logit lens.

Functionally, the Pre-LN logit lens is equivalent to applying zero ablation to all sublayer outputs beyond a given layer $\ell$ (Belrose et al., 2023). We extend this principle to Post-LN models by applying the same zero ablation while preserving the subsequent layer normalizations. We also ablate layer normalization biases for layers beyond $\ell$, with further justification provided in Appendix B.

Since Leela is an encoder that uses representations from all tokens simultaneously, we project the intermediate representations of all 64 squares through the policy head to obtain policies at each layer.

### 2.3 Performance evaluation

We follow Ruoss et al. (2024) for playing strength and puzzle-solving evaluation.

**Playing strength assessment** We conduct a round-robin tournament between policies derived from representations at the input and all 15 transformer layers. Each pairing plays 200 Encyclopedia of

Chess Openings (Matanović, 1978) positions, with one game per side using argmax move selection and five games per side using a temperature of $\tau = 1.0$. Elo ratings are computed using BayesElo (Coulom, 2008). We include their external Leela policy network as an anchor, using their computed Elo score. Additionally, we deploy the layer-wise policies as bots on Lichess across multiple time controls, using a temperature of 1.0 for the first five full-moves to introduce opening diversity.

**Puzzle-solving performance** We evaluate tactical understanding using their dataset of 10,000 Lichess (Lichess.org, 2025) puzzles, each constructed with a single clear winning line while all other moves are significantly inferior. For each layer's policy, we use argmax selection to predict moves and consider a puzzle solved if it reproduces the principal variation.

### 2.4 Representational analysis

**Intermediate policy dynamics** To characterize how policy distributions evolve across layers, we compute the Jensen-Shannon divergence between layer policies and the final model output, policy entropy at each layer, the probability assigned to the final model's top move by each layer, and Kendall's $\tau$ ranking correlation between intermediate and final move rankings.

**Layer-wise concept preferences** To examine which chess concepts each layer prioritizes, we analyze how layer-wise policies weight moves by their conceptual effects. Following McGrath et al. (2022), who trained linear probes on AlphaZero's intermediate representations, we use Stockfish 8's handcrafted continuous evaluation terms as human-interpretable concepts. Rather than probing for concept representation, we measure *concept preference* directly from layer-wise move probabilities.

For each move $m$ from position $s$ to resulting position $s'$, and each concept $c$, we compute $\Delta c_m = c(s') - c(s)$, representing the change in $c$ caused by $m$. All evaluations are from the perspective of the current player, so positive values indicate an improvement. At each layer $\ell$, we compute:

$$\Delta c_\ell = \sum_{m \in \text{legal}(s)} \pi_\ell(m) \cdot \Delta c_m = \mathbb{E}_{\pi_\ell}[\Delta c_m]$$

where $\pi_\ell(m)$ is the move probability assigned by layer $\ell$'s policy. This represents the expected concept change when sampling moves according to $\pi_\ell$. Full details and plots are provided in Appendix I.

## 3 Results

### 3.1 Phased capability progression

**Tournament strength** Table 1 reports Elo ratings across layers. Playing strength increases with depth but suggests a three-phase progression rather than uniform improvement. Early layers show rapid gains through layer 5, middle layers form a performance plateau through approximately layer 10, and late layers demonstrate sharp strengthening beginning around layer 11. This pattern holds consistently under both deterministic ($\tau = 0$) and stochastic ($\tau = 1$) move selection. Real-world Lichess deployment shows similar trends with clear late-layer strengthening, though with less pronounced separation between phases due to greater variability.

Table 1: Playing strength (Elo rating) across transformer layers and evaluation methods

| Evaluation | Input | L0 | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 | L13 | Full | Anchor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Internal Tournament ($\tau = 0$) | 443 | 650 | 699 | 790 | 871 | 962 | 1007 | 993 | 1014 | 1006 | 1042 | 1057 | 1083 | 1337 | 1681 | 2263 | 2292 |
| Internal Tournament ($\tau = 1$) | 369 | 701 | 708 | 813 | 911 | 1080 | 1098 | 1064 | 1068 | 1069 | 1110 | 1113 | 1151 | 1355 | 1394 | 1640 | 2292 |
| Lichess Blitz | 518 | 651 | 681 | 688 | 741 | 769 | 774 | 850 | 803 | 843 | 832 | 960 | 948 | 1252 | 1581 | 2274 | - |
| Lichess Bullet | 693 | 904 | 891 | 916 | 972 | 1009 | 926 | 984 | 1052 | 994 | 1061 | 1064 | 1129 | 1331 | 1659 | 2246 | - |
| Lichess Rapid | 558 | 816 | 709 | 697 | 915 | 717 | 939 | 1021 | 1018 | 1032 | 957 | 1107 | 1095 | 1290 | 1581 | 2253 | - |

**Puzzle solving capabilities** Figure 2 shows clear improvement in puzzle-solving ability across network layers within each Elo range, and consistent decrease in performance across increasing difficulty levels for all layers. Dashed lines and shading mark phase boundaries with annotated slope ratios indicating relative improvement rates between phases. While the early-to-middle phase distinction is less pronounced than in tournaments, the final-phase acceleration is clearly visible, particularly for harder puzzles where improvement rates exceed 60 times the middle phase.

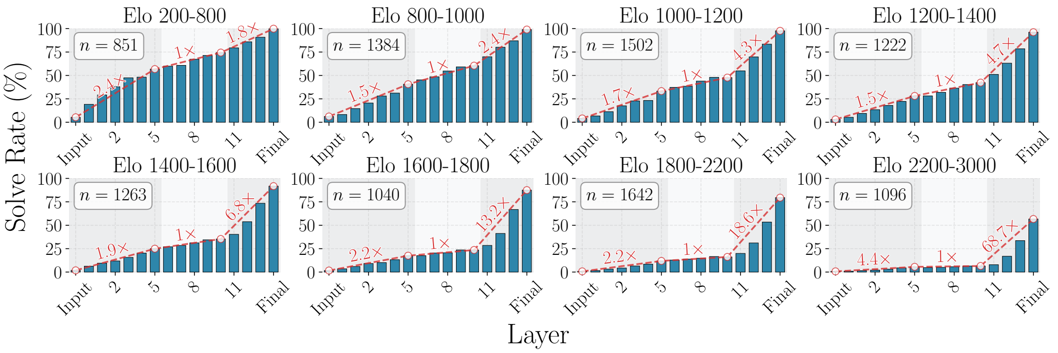


Figure 2: Puzzle-solving performance across layers, stratified by Elo rating. Red dashed lines mark phase boundaries derived from tournament analysis, indicating phase-specific improvement rates.

### 3.2 Representational dynamics

**Solution discovery and forgetting** Figure 3 tracks four metrics across layers: current solve rate, cumulative discoveries, new solutions, and median probability of principal variation moves. The gap between current and cumulative rates indicate that solutions are frequently discovered and subsequently discarded. Both the median-probability trajectory and new-solution discovery rate follow the three-phase pattern from playing strength, with steady improvement in early layers, a mid-layer plateau with stable probabilities and reduced discovery rates, and renewed acceleration in later layers where many puzzles are solved for the first time. The final cumulative solve rate exceeds the last layer's, implying that earlier layers solve puzzles later forgotten.

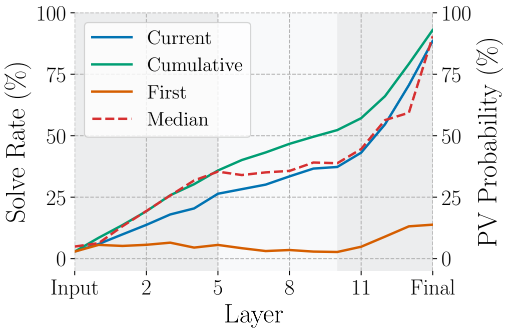


Figure 3: Layer-wise puzzle-solving performance across network depth showing current solve rate, cumulative and first discoveries, and median probability assigned to principal-variation (PV) moves. Background shading indicates the three computational phases identified in tournament analysis.

Figure 4 illustrates this phenomenon through an example where the correct solution maintains high probability through most layers before being overtaken by a seemingly safer but losing move in the final output. This late-layer reversal toward conservative alternatives occurs consistently across forgotten puzzles (Appendix H), suggesting final layers encode safety-oriented priors overriding earlier tactical solutions. The value head correctly evaluates both resulting positions, indicating the policy head's final selection contradicts the model's own assessment of the resulting position.

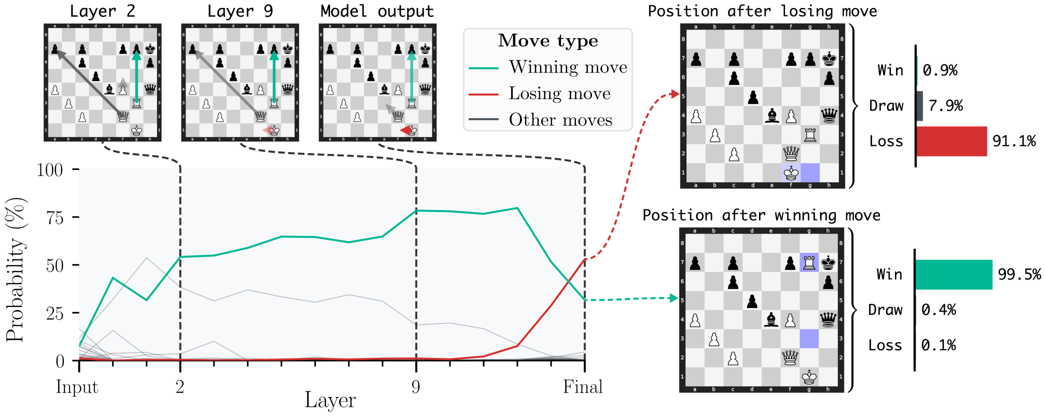


Figure 4: Representative example of solution forgetting. **Left:** The correct move **♖×g7+** maintains high probability through most layers before dropping, while the losing move **♔f1** rises from near-zero to become the top choice. **Right:** Value head evaluations correctly assess both resulting positions.

**Distributional analysis** We summarize how policy distributions evolve across layers (Appendix D). Kendall's $\tau$ correlation between intermediate and final move rankings is initially negative, stays low through the middle layers, and rises sharply in the final ones, indicating that the model reevaluates move preferences until very late in the network. Although individual positions exhibit varying entropy, the overall entropy distribution changes little with depth. Jensen–Shannon divergence varies substantially, with some positions aligning early while most remain divergent until late. The probability assigned to the final model's top move increases primarily in the final layers, though with high early-layer agreement for some positions. Overall, the results indicate that while some positions converge early across all metrics, others undergo substantial move reordering throughout depth, with stable entropy indicating genuine preference reevaluation rather than distributional sharpening.

**Concept preference of intermediate layers** The right side of Figure 5 shows material and total $\Delta c_\ell$ peak in early-to-middle layers before declining, consistent with McGrath et al. (2022)'s finding that these concepts are most strongly represented at intermediate depths in AZ. When controlling for material, total Stockfish evaluation increases through layer 12 before declining in final layers—uniquely among all performance metrics—indicating that Leela's final evaluation diverges from Stockfish's. The left side shows early and middle layers favoring aggressive over defensive concepts, with higher $\Delta c_\ell$ for opponent king vulnerability and own threats, while later layers shift toward a balanced evaluation, increasing own king safety and reducing opponent threats, with all four concepts converging to similar values. This late-layer shift toward conservative, balanced evaluation aligns with the forgotten puzzle phenomenon, where final layers favor safer alternatives over tactical solutions. Middle layers exhibit stable preferences across concepts, mirroring the performance plateau observed earlier.

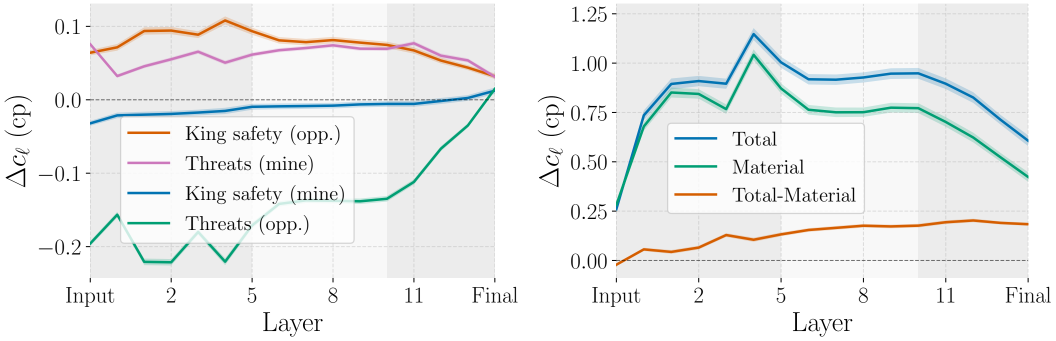


Figure 5: Mean of expected concept deltas ($\Delta c_\ell$) over positions across layers, measured in centipawns with 95% CI. **Left:** King-safety and threat concepts for the moving and opposing sides. **Right:** Total, material, and residual evaluations. Shaded regions indicate network phases.

## 4 Discussion

Our analysis of Leela's policy network provides insights into its iterative inference process by revealing how move preferences evolve across layers. While overall playing strength and puzzle-solving ability generally improve with depth, this progression occurs at varying rates. Leela appears to exhibit distinct computational stages similar to those proposed for LLMs (Lad et al., 2025): an early phase of rapid improvement, a middle phase of plateauing performance analogous to feature engineering, and a late phase of feature integration, with the final layer showing a sharp increase in MLP output norm consistent with feature consolidation (Appendix D.1). The onset of this final phase at layer 11 coincides with the emergence of several look-ahead heads in layers 11 to 13 (Jenner et al., 2024; Cruz, 2025) that relocate information from future-move squares to current candidate squares, allowing the policy head to integrate this information into its final move predictions. Across these phases, move preferences are repeatedly reevaluated rather than gradually refined, with probabilities fluctuating substantially across layers in a manner consistent with algorithmic recomputation. Building on McGrath et al. (2022)'s finding that concept representations vary with depth, we analyze how concept preferences evolve—examining which concepts each layer prioritizes when selecting moves. The shift from aggressive tactics in early layers to safety-oriented evaluation in final layers provides a potential mechanism for forgotten puzzles, where learned priors override algorithmically identified tactical solutions. Together, these results suggest that Leela's inference process integrates algorithmic computation with learned heuristic priors.

## Author contributions

Elias conceived, implemented, and conducted the study, performed all experiments, and wrote the manuscript. Sebastian and Wojciech provided supervision and feedback throughout the project.

## Acknowledgments and Disclosure of Funding

We gratefully acknowledge Jenner et al. (2024) for the codebase and inference framework on which this work is built. Without their open-source release, this research would not have been possible.

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR) as grants [BIFOLD (01IS18025A, 01IS180371I), xJuRAG (16IS25015B)]; the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant ACHILLES (101189689); and the German Research Foundation (DFG) as research unit DeSBi [KI-FOR 5363] (459422098).

## References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112.

Rémi Coulom. Whole-history rating: A bayesian rating system for players of time-varying strength. In *Computers and Games*, 2008.

Diogo Cruz. Understanding the learned look-ahead behavior of chess neural networks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=np4Bg2zIxL.

dje dev. Discussion on layer normalization in transformer architecture. Forum post on Lc0 community chat, 2025. URL https://lc0.org/chat. Accessed: 2025-08-17.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Stanisław Jastrzębski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. Residual connections encourage iterative inference, 2018. URL https://arxiv.org/abs/1710.04773.

Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart Russell. Evidence of learned look-ahead in a chess-playing neural network, 2024. URL https://arxiv.org/abs/2406.00877.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference?, 2025. URL https://arxiv.org/abs/2406.19384.

Leela Chess Zero team. Leela Chess Zero. URL https://lczero.org/.

Leela Chess Zero team. CCRL dataset, 2018. URL https://lczero.org/blog/2018/09/a-standard-dataset/. 2.5 million games from CCRL 40/40 and 40/4 tournaments.

lepned. How well do Lc0 networks compare to the greatest transformer network from DeepMind?, 2024. URL https://lczero.org/blog/2024/02/how-well-do-lc0-networks-compare-to-the-greatest-transformer-network-from-deepmind/.

Lichess Bot Devs. Lichess bot: A bridge between lichess api and chess engines. https://github.com/lichess-bot-devs/lichess-bot, 2025. Accessed: 2025-08-18.

Lichess.org. Lichess: Free online chess, 2025. URL https://lichess.org. Accessed: 2025-08-18.

Aleksandar Matanović. *Encyclopaedia of Chess Openings*. Batsford Limited, 1978.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47), November 2022. ISSN 1091-6490. doi: 10.1073/pnas.2206625119. URL http://dx.doi.org/10.1073/pnas.2206625119.

Timothee Mickus, Denis Paperno, and Mathieu Constant. How to dissect a muppet: The structure of transformer embedding spaces, 2022. URL https://arxiv.org/abs/2206.03529.

Diganta Misra. Mish: A self regularized non-monotonic activation function, 2020. URL https://arxiv.org/abs/1908.08681.

Daniel Monroe. Transformer progress, February 2024. URL https://lczero.org/blog/2024/02/transformer-progress/. Accessed: 2025-08-17.

nostalgebraist. Interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://openai.com/research/language-unsupervised, 2019. OpenAI Blog.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017. URL https://arxiv.org/abs/1710.05941.

Anian Ruoss, Grégoire Delétang, 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, 2024. URL https://arxiv.org/abs/2402.04494.

Matthew Sadler and Natasha Regan. *Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI*. New in Chess, Alkmaar, Netherlands, 2019. ISBN 978-90-5691-818-7.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Stockfish Developers. Stockfish 8. https://stockfishchess.org/blog/2016/stockfish-8/, 2016. Accessed: 2025-11-07.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008, 2017.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers, 2022. URL https://arxiv.org/abs/2203.00555.

## A Limitations and future directions

**Scope of analysis** Our analysis examines a single chess model with a Post-LN architecture. This makes it difficult to disentangle whether our findings reflect domain-specific characteristics of chess, architecture-specific properties of Post-LN transformers, or more general principles of neural network computation. Since no Pre-LN Leela models exist and the development team does not plan to train any due to Post-LN's superior performance in their training paradigm (dje dev, 2025), comparing logit lens behavior across Pre-LN and Post-LN language models would help isolate architectural effects from domain-specific phenomena. We plan to include such comparisons in the full paper.

**Mechanistic understanding** Our analysis is purely observational, identifying *what* changes across layers (forgotten puzzles, concept preference shifts) without determining *which components* cause these behaviors. We do not identify the specific attention heads, MLP sublayers, or neuron populations responsible for encoding safety priors or overriding tactical solutions. While prior work has identified specific look-ahead attention heads (Jenner et al., 2024; Cruz, 2025) that provide partial mechanistic explanations for capability improvements and information aggregation in the final layers, the components responsible for learned heuristic priors and their interaction with algorithmic computation remain unknown. Causal interventions such as activation patching, steering, or targeted ablations could establish the mechanistic link between concept preference evolution and behavioral phenomena like forgotten puzzles.

**Concept analysis scope** Our concept preference analysis is limited to Stockfish 8's handcrafted evaluation terms applied to one-step look-ahead positions. While Stockfish's concepts could be extended to multi-move tactical sequences, we restrict our analysis to immediate position evaluation. We do not analyze long-term strategic planning or chess concepts beyond Stockfish's vocabulary. Additionally, we do not examine how concept preferences vary by position type (tactical vs. positional, opening vs. endgame), which could reveal context-dependent prioritization strategies.

**Characterizing safety priors** While we observe that safety-oriented concept preferences in final layers coincide with forgotten puzzles where forced checkmates are abandoned, we do not establish a causal link or systematically characterize when such preferences improve versus harm play. The priors presumably provide net benefit across Leela's training distribution, but understanding their failure modes more precisely could enable targeted interventions. Future work could investigate whether these preferences can be selectively modulated—for example, reducing safety prioritization in tactical positions to improve puzzle performance, or adjusting them to control the engine's playing style (aggressive vs. conservative). However, such modifications may be overridden by the interplay with MCTS in practical deployment.

**Generalization and applications to language and reasoning models** While our analysis focuses on chess transformers, the core findings—that neural networks integrate algorithmic computation with learned heuristic priors through distinct computational phases—likely reflect general principles that extend beyond this domain. The architectural advantages we leverage (all-token prediction, complete observability) are specific to encoder models, but the fundamental tension between algorithmic reasoning and distributional priors may manifest similarly in language models.

Our methodology could inform interpretability research in language models in several ways. First, concept preference analysis could be adapted to track behavioral tendencies in language models across layers—such as prioritizing truthfulness vs. helpfulness, or specificity vs. safety—offering an alternative to representation probing that identifies whether concepts are merely encoded rather than which behaviors intermediate layers exhibit. E.g., observing when layers shift from non-deceptive to deceptive output tendencies could identify where such behaviors emerge, providing targets for mechanistic investigation or intervention through activation steering or fine-tuning. Second, our observation that safety priors may override algorithmic solutions suggests that in reasoning-intensive tasks, suppressing learned distributional biases—such as the tendency to broaden toward common tokens in final layers—might improve performance on tasks requiring precise logical reasoning.

However, the specifics of how algorithmic and heuristic computation interact will likely differ substantially across domains due to differences in training objectives, task structure, and architectural choices. Empirical investigation in each domain remains necessary to characterize these dynamics.

## B Post-LN Logit Lens Extension

The logit lens technique projects intermediate layer representations through the model's output head to examine what the network would predict at different depths. This approach relies on the assumption that representations across layers exist in a shared basis that allows meaningful projection to output space—an assumption enabled by residual connections that create additive pathways for information flow (Jastrzębski et al., 2018). In transformer architectures, architectural inductive biases further encourage information about specific tokens to remain localized to their corresponding positions throughout the network, creating a privileged basis that facilitates cross-layer interpretation (Jenner et al., 2024).

In Pre-LN transformers, this involves taking intermediate representations, applying the final layer normalization, and projecting through the unembedding matrix. This approach works because layer normalization precedes each sublayer in Pre-LN models, leaving the residual stream unchanged and maintaining representational consistency across layers.

Post-LN architectures complicate this process by applying layer normalization after each sublayer's output is added to the residual stream. Unlike Pre-LN models where only a final normalization is needed, Post-LN models have sequential normalization operations that directly transform the residual stream at each layer. These intermediate normalizations create dependencies between layers that prevent simply taking an intermediate representation and applying only the final layer normalization and output projection, as the intermediate representation has not undergone the normalization transformations it would experience in a complete forward pass.

Our goal is to develop an extension that maps intermediate layer representations to the representational basis expected by the policy head, accounting for the normalization transformations unique to Post-LN architectures.

### B.1 Pre- vs Post-LN architectures and DeepNorm

The key difference between Pre-LN and Post-LN architectures lies in when layer normalization is applied relative to the residual connections. This placement affects how representations evolve through the network and impacts the applicability of interpretability techniques.

**Pre-LN** In Pre-LN transformers, layer normalization is applied before each sublayer (attention and feed-forward), with the computation following the pattern:

$$\mathbf{h}'_\ell = \mathbf{h}_{\ell-1} + \mathrm{MHA}_\ell(\mathrm{LayerNorm}(\mathbf{h}_{\ell-1})) \tag{1}$$

$$\mathbf{h}_\ell = \mathbf{h}'_\ell + \mathrm{FFN}_\ell(\mathrm{LayerNorm}(\mathbf{h}'_\ell)) \tag{2}$$

where $\mathbf{h}_\ell$ denotes the hidden representation at layer $\ell$, $\mathrm{MHA}_\ell$ is the multi-head attention operation, and $\mathrm{FFN}_\ell$ is the feed-forward network at layer $\ell$. This design ensures that the residual stream itself is never directly transformed by normalization operations, allowing intermediate representations to be projected through the final layer normalization and output head in a straightforward manner.

**Post-LN** Post-LN models apply layer normalization after adding sublayer outputs to the residual stream:

$$\mathbf{h}'_\ell = \mathrm{LayerNorm}(\mathbf{h}_{\ell-1} + \mathrm{MHA}_\ell(\mathbf{h}_{\ell-1})) \tag{3}$$

$$\mathbf{h}_\ell = \mathrm{LayerNorm}(\mathbf{h}'_\ell + \mathrm{FFN}_\ell(\mathbf{h}'_\ell)) \tag{4}$$

Each layer normalization operation directly transforms the accumulated representation, creating a sequence of transformations that intermediate representations must undergo to reach the final output space.

**DeepNorm** Leela employs DeepNorm (Wang et al., 2022) scaling to stabilize Post-LN training, modifying the computation to include residual scaling factors:

$$\mathbf{h}'_\ell = \mathrm{LayerNorm}(\alpha \cdot \mathbf{h}_{\ell-1} + \mathrm{MHA}_\ell(\mathbf{h}_{\ell-1})) \tag{5}$$

$$\mathbf{h}_\ell = \mathrm{LayerNorm}(\alpha \cdot \mathbf{h}'_\ell + \mathrm{FFN}_\ell(\mathbf{h}'_\ell)) \tag{6}$$

where $\alpha = (2N)^{1/4} \approx 2.34$ for $N = 15$ layers. This scaling, combined with specialized weight initialization, enables stable training while preserving Post-LN's representational advantages.

### B.2 Zero ablation methodology for Post-LN architectures

The standard logit lens can be understood as performing zero ablation—setting all sublayer outputs to zero for layers beyond $\ell$, then applying the final layer normalization and projection to output space. We extend this principle to Post-LN models by applying the same zero ablation of sublayer outputs while preserving the normalization operations and $\alpha$ that would transform these representations.

Specifically, for examining layer $k$, we set $\mathrm{MHA}_\ell(\cdot) = 0$ and $\mathrm{FFN}_\ell(\cdot) = 0$ for all $\ell > k$ during a forward pass, while preserving the $\alpha$ and layer normalization scaling parameters ($\gamma$). We set layer normalization biases ($\beta$) to zero for layers beyond $k$ since we believe this maintains better consistency with the standard logit lens paradigm. The rationale for these choices will be explained in the subsection following the decomposition analysis B.4.

This approach ensures that representations from layers 1 through $k$ undergo the same sequence of transformations they would experience in the complete network, while removing contributions from later layers. Normalization statistics ($\mu$ and $\sigma$) are recomputed during this modified forward pass to reflect the actual distribution of the truncated representations, rather than using statistics computed on the full model output. This follows the same principle as Pre-LN logit lens implementations, which directly apply the final layer normalization to the intermediate activations being analyzed.

### B.3 Transformer encoder decomposition

To provide intuition for why our proposed logit lens extension makes principled choices for Post-LN architectures, we present a decomposition of the transformer encoder output and reinterpret the logit lens in terms of this decomposition. Following Mickus et al. (2022), the final representation of the encoder at layer $L$ and token position $t$ can be decomposed into:

$$\mathbf{h}_{L,t} = \mathbf{i}_{L,t} + \mathbf{z}_{L,t}^{\mathrm{MHA}} + \mathbf{z}_{L,t}^{\mathrm{FFN}} + \mathbf{b}_{L,t} - \mathbf{m}_{L,t} \tag{7}$$

$$\mathbf{i}_{L,t} = \alpha^{2L} \cdot \frac{\bigodot_{\ell=1}^{L} \gamma_\ell^{\mathrm{MHA}} \odot \gamma_\ell^{\mathrm{FFN}}}{\prod_{\ell=1}^{L} \sigma_{\ell,t}^{\mathrm{MHA}} \sigma_{\ell,t}^{\mathrm{FFN}}} \odot \mathbf{h}_{0,t} \tag{8}$$

$$\mathbf{z}_{L,t}^{\mathrm{MHA}} = \sum_{\ell=1}^{L} \alpha^{2L-2\ell+1} \cdot \frac{\bigodot_{\ell'=\ell}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\prod_{\ell'=\ell}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot \tilde{\mathbf{z}}_{\ell,t}^{\mathrm{MHA}} \tag{9}$$

$$\mathbf{z}_{L,t}^{\mathrm{FFN}} = \sum_{\ell=1}^{L} \alpha^{2L-2\ell} \cdot \frac{\gamma_\ell^{\mathrm{FFN}} \bigodot_{\ell'=\ell+1}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\sigma_{\ell,t}^{\mathrm{FFN}} \prod_{\ell'=\ell+1}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot \tilde{\mathbf{z}}_{\ell,t}^{\mathrm{FFN}} \tag{10}$$

$$\begin{aligned}\mathbf{b}_{L,t} = &\sum_{\ell=1}^{L} \alpha^{2L-2\ell+1} \cdot \frac{\bigodot_{\ell'=\ell}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\prod_{\ell'=\ell}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot b_\ell^{\mathrm{MHA},V} W_\ell^{\mathrm{MHA},O} \\ &+ \sum_{\ell=1}^{L} \alpha^{2L-2\ell+1} \cdot \frac{\bigodot_{\ell'=\ell}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\prod_{\ell'=\ell}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot b_\ell^{\mathrm{MHA},O} \\ &+ \sum_{\ell=1}^{L} \alpha^{2L-2\ell} \cdot \frac{\gamma_\ell^{\mathrm{FFN}} \bigodot_{\ell'=\ell+1}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\sigma_{\ell,t}^{\mathrm{FFN}} \prod_{\ell'=\ell+1}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot b_\ell^{\mathrm{FFN},O} \\ &+ \sum_{\ell=1}^{L} \alpha^{2L-2\ell+1} \cdot \frac{\gamma_\ell^{\mathrm{FFN}} \bigodot_{\ell'=\ell+1}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\sigma_{\ell,t}^{\mathrm{FFN}} \prod_{\ell'=\ell+1}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot \beta_\ell^{\mathrm{MHA}} \\ &+ \sum_{\ell=1}^{L} \alpha^{2L-2\ell} \cdot \frac{\bigodot_{\ell'=\ell+1}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\prod_{\ell'=\ell+1}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \beta_\ell^{\mathrm{FFN}}\end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{m}_{L,t} = &\sum_{\ell=1}^{L} \alpha^{2L-2\ell+1} \cdot \frac{\bigodot_{\ell'=\ell}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\prod_{\ell'=\ell}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot \mu_{\ell,t}^{\mathrm{MHA}} \mathbf{1} \\ &+ \sum_{\ell=1}^{L} \alpha^{2L-2\ell} \cdot \frac{\gamma_\ell^{\mathrm{FFN}} \bigodot_{\ell'=\ell+1}^{L} \gamma_{\ell'}^{\mathrm{MHA}} \odot \gamma_{\ell'}^{\mathrm{FFN}}}{\sigma_{\ell,t}^{\mathrm{FFN}} \prod_{\ell'=\ell+1}^{L} \sigma_{\ell',t}^{\mathrm{MHA}} \sigma_{\ell',t}^{\mathrm{FFN}}} \odot \mu_{\ell,t}^{\mathrm{FFN}} \mathbf{1}\end{aligned} \tag{12}$$

where $\mathbf{h}_{0,t}$ is the initial input embedding, $\tilde{\mathbf{z}}_{\ell,t}^{\mathrm{MHA}}$ and $\tilde{\mathbf{z}}_{\ell,t}^{\mathrm{FFN}}$ are the raw unbiased outputs from multi-head attention and feed-forward networks at layer $\ell$, and $\alpha = (2N)^{1/4}$ is the DeepNorm scaling factor. The bias terms include $b_\ell^{\mathrm{MHA},V}$ (value projection bias), $W_\ell^{\mathrm{MHA},O}$ (output projection matrix), $b_\ell^{\mathrm{MHA},O}$ (attention output bias), and $b_\ell^{\mathrm{FFN},O}$ (feed-forward output bias). Layer normalization parameters are

$\gamma_\ell$ (learned scales), $\sigma_{\ell,t}$ (computed standard deviations), $\mu_{\ell,t}$ (computed means), and $\beta_\ell$ (learned biases), with superscripts indicating MHA or FFN sublayers. The notation $\bigodot$ denotes element-wise multiplication.

The decomposition separates the final representation into the transformed input embedding ($\mathbf{i}_{L,t}$), accumulated sublayer contributions ($\mathbf{z}_{L,t}^{\text{MHA}}$, $\mathbf{z}_{L,t}^{\text{FFN}}$), bias terms ($\mathbf{b}_{L,t}$), and mean centering effects ($\mathbf{m}_{L,t}$). Each component is scaled by normalization parameters from subsequent layers.

### B.4 Implications for the logit lens

The decomposition framework provides direct justification for our zero ablation approach. The Post-LN logit lens can be understood as truncating the summations in $\mathbf{z}_{L,t}^{\text{MHA}}$, $\mathbf{z}_{L,t}^{\text{FFN}}$, and $\mathbf{b}_{L,t}$ to include only layers $\ell \leq k$, while recomputing the normalization statistics ($\mu$ and $\sigma$) based on the modified representations. This mathematical perspective clarifies why our methodological choices are well-founded.

**Preservation of scaling parameters** The decomposition reveals why preserving the $\gamma$ and $\alpha$ scaling factors is essential: all terms in the truncated representation—including the input embedding $\mathbf{i}_{L,t}$ and retained sublayer contributions—are transformed by these parameters from subsequent layers. Removing these transformations would fundamentally alter how the truncated representation maps to the output space of the encoder, undermining the interpretability objective.

**Treatment of bias terms** The decomposition also illuminates our decision to ablate layer normalization biases ($\beta$). These terms appear in $\mathbf{b}_{L,t}$ alongside other bias components (attention and feed-forward biases), indicating they are structurally and functionally equivalent. For consistency with the standard logit lens principle of removing sublayer contributions beyond layer $k$, we ablate all of the different bias terms, including layer normalization biases. Empirically, we found that preserving versus ablating these biases produces qualitatively similar results with only nuanced differences, and we provide code to examine all configurations.

**Representational alignment versus learned priors** The treatment of bias terms reflects broader questions about their computational role in neural networks. Bias terms may serve as representational alignment mechanisms that bridge coordinate system differences between layers, or they may encode learned priors about the task domain. For example, the Tuned Lens approach learns affine transformations with bias terms to achieve better representational alignment between an intermediate and the final layer for Pre-LN architectures, motivated by biased outputs observed with the standard logit lens (Belrose et al., 2023). However, we hypothesize that these biased outputs could also be attributed to standard logit lens implementations preserving the final layer normalization bias.

Our approach faces similar limitations by using the policy head unchanged, which contains its own bias terms that may contribute to biased outputs if our hypothesis about layer normalization bias effects in Pre-LN models is correct. Moreover, the decomposition framework only captures directly accessible bias terms—those that can be linearly separated from the representation—while bias terms that interact with activation functions cannot be decomposed into constant additive components. Likely, bias terms serve both representational alignment and learned prior functions simultaneously and it's not clear how to disentangle these functions empirically given current interpretability techniques.

## C Complete model architecture

We analyze the `T82-768x15x24h` transformer model from Jenner et al. (2024), using their inference framework for our experiments. This model has 15 transformer layers, 768-dimensional representations, 24 attention heads, and approximately 109 million parameters. We use the original model that incorporates historical board information rather than their fine-tuned version to avoid potential artifacts from the fine-tuning process that could affect our interpretability analysis.

This Leela Chess Zero model employs a transformer-based architecture that processes chess positions through three main stages: input encoding transforms board states into token representations (with each of the 64 squares treated as a discrete token), a 15-layer transformer encoder with specialized attention mechanisms processes these representations, and task-specific output heads generate move probabilities, position evaluations, and game length predictions. The architecture employs various activation functions: Mish (Misra, 2020) for input processing and output heads, squared ReLU for feed-forward networks, and Swish (Ramachandran et al., 2017) within the Smolgen attention enhancement.

### C.1 Input encoding

Leela's input encoding transforms chess positions through binary feature planes, chess-specific positional encodings, and learned projections into the model's embedding space.

**Board representation** The board state is encoded using 112 binary feature planes of size $8 \times 8$, with the first 12 planes representing current piece positions for each piece type and color. Historical context is incorporated through 8 previous board positions (96 planes), plus auxiliary planes encoding castling rights (4 planes), side to move (1 plane), fifty-move clock (1 plane), and two constant planes (0s and 1s) that are architectural remnants from CNNs without functional meaning.

**Positional encoding** Leela employs chess-specific positional encodings that capture movement relationships between squares in a $64 \times 64$ matrix. Each square receives a 64-dimensional vector where position $(i, j)$ is set to 1 if any piece could legally move from square $i$ to square $j$ in one move regardless of current board state, 0 otherwise, and $-1$ for diagonal entries $(i = j)$ to distinguish self-reference.

**Input preparation** The 112 feature planes are reshaped from $112 \times 8 \times 8$ to $64 \times 112$, where each of the 64 entries corresponds to a board square with the 112-dimensional feature vector for that square concatenated with its corresponding 64-dimensional positional encoding vector, forming a $64 \times 176$ tensor. This combined representation undergoes a linear transformation with Mish activation, followed by elementwise scaling and shifting operations, producing the final $64 \times 768$ input for the transformer layers. This enriched preprocessing was motivated by observations that early attention layers contributed minimally to performance (Monroe, 2024). While no information mixing between tokens occurs here, this involves substantially more processing than typical language models (which simply use embedding matrices plus positional encodings), potentially explaining why our logit lens mappings yield meaningful results even before the first transformer layer.

### C.2 Transformer encoder

The transformer encoder consists of 15 identical layers with a model dimensionality of 768, processing 64 tokens (one per board square) through multi-head attention with Smolgen enhancement, feed-forward networks, and Post-LN normalization with DeepNorm scaling. Unlike autoregressive models, all tokens can attend to each other bidirectionally.

**Attention with Smolgen** Each layer uses 24 attention heads with 32-dimensional head size, applying scaled dot-product attention to the 64 board square tokens. The Smolgen module enhances standard self-attention by enabling attention scores to depend not only on individual square contents but also on the global board position. It compresses all square representations into a global vector, processes this through MLPs, then generates supplementary attention logits of shape $24 \times 64 \times 64$ that are added to the standard query-key dot products before softmax normalization.

**Feed-Forward** Each FFN layer expands from 768 to 1024 dimensions with squared ReLU activation. Unlike other domains that benefit from $4\times$ expansion ratios, chess models show little improvement from larger feed-forward networks (Monroe, 2024).

**Post-LN with DeepNorm** The model employs a Post-LN architecture, the original design used in early transformers (Vaswani et al., 2017), applying layer normalization after each sublayer (attention and feed-forward) within the residual stream rather than before sublayers as in modern Pre-LN variants, following empirical comparisons that demonstrated superior performance (dje dev, 2025). To mitigate vanishing gradients, the model uses DeepNorm scaling (Wang et al., 2022), which applies a constant upscaling factor to residual connections and uses specialized weight initialization. See Appendix B.1 for complete equations.

### C.3 Output heads

The transformer encoder's final $64 \times 768$ representations are processed by three specialized heads: the policy head generates move probability distributions, the value head predicts win/draw/loss outcomes, and the moves-left head estimates remaining game length.

**Policy head** The policy head first processes each square's 768-dimensional representation through a shared MLP with Mish activation. These processed representations are then transformed via two separate linear projections: one creating "source" representations for squares where moves originate, and another creating "target" representations for destination squares. The source and target representations are matrix-multiplied along the 768-dimensional axis, producing a $64 \times 64$ matrix where each entry contains a scalar logit representing the likelihood of a move from the corresponding source square to target square. Promotion moves require additional processing through two specialized branches that handle move selection and promotion-type preferences separately. The final output combines standard move logits with promotion logits, which are then filtered to extract only legal moves for the current position before applying softmax normalization.

**Value head** The value head processes each square's representation through an MLP with Mish activation, reducing them to 32 dimensions per square. These 64 representations are flattened into a single 2048-dimensional vector to enable global position assessment. A second MLP layer compresses this to 128 dimensions, followed by a linear projection to 3 logits, which are then converted to win, draw, and loss outcome probabilities via softmax normalization. Unlike AlphaZero's single scalar output, this three-way classification provides more nuanced position evaluation.

**Moves-left head** The moves-left head predicts the number of moves remaining until game termination. Each square's representation is processed through an MLP, then flattened into a global vector for position-level assessment. Two additional MLP layers progressively reduce the dimensionality to produce a final scalar output estimating remaining game length. The output layer uses Mish activation because ReLU would provide zero gradients when pre-activation values are negative during training, even though the target output is always positive.

## D Detailed policy dynamics analysis

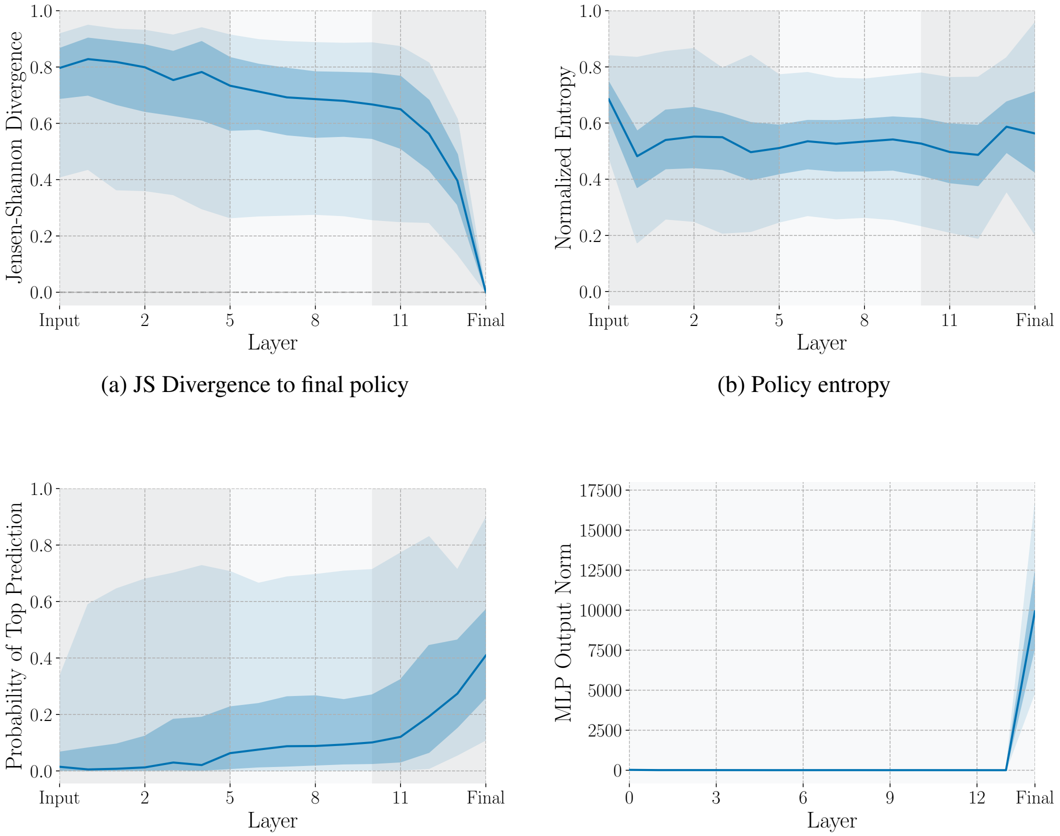
To provide a deeper, quantitative view of the model's inference process, we computed several policy metrics across the network's depth. The following analyses were conducted on a sample of 1000 positions from the CCRL dataset (Leela Chess Zero team, 2018). For the language model comparisons, we sampled prompts from the Pile dataset (Gao et al., 2020).

### D.1 Policy Metrics

Figure 6 presents four metrics that characterize the policy's evolution. We plot the median, the 25th-75th percentile range, and the 5th-95th percentile range and indicate the proposed phases as shaded background.

**Jensen-Shannon divergence** The Jensen-Shannon Divergence shows a median that remains high until the final layers, with the 5th percentile being much lower. This shows that a small subset of positions converge early to the final policy, while the vast majority have different move preferences than the final output throughout most layers. The phase structure is again apparent, with higher variability and erratic changes in early layers, a more stable plateau in the middle phase, and a sharp decline in the final phase. In this metric, the transition to the final phase appears to occur slightly later—around layer 12 rather than 11—though consistent with the overall three-phase progression described in the main text.

**Entropy** The policy entropy shows a relatively stable median with modest variation across layers, indicating that overall certainty changes only slightly with depth. The 95th and 5th percentiles are very high and very low respectively, indicating that for any layer there are positions where the layer is either very certain or very uncertain about what moves are good. The phase structure is again



(a) JS Divergence to final policy

(b) Policy entropy

(c) Probability of final top move

(d) Final MLP layer output norm

Figure 6: Metrics of policy dynamics for Leela Chess Zero. The central line shows the median over 1000 positions. The inner (darker) and outer (lighter) shaded regions represent the 25th-75th and 5th-95th percentile ranges, respectively. Shaded background indicated the proposed phases.

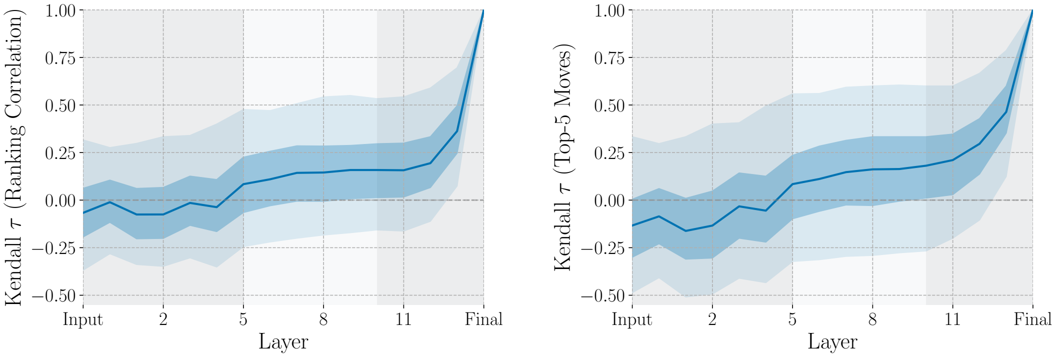
apparent, with more erratic fluctuations in the early layers, a stable middle phase, and increased variability toward the end.

**Probability of final top move** The median probability assigned to the final top move is near zero for the first layers, then grows approximately linearly with a steeper increase in the final layers. The outliers are very high (between 0.6 and 0.8), indicating that all layers correctly predict the top move for some positions with high probability, either having already converged or representing cases where solutions are later forgotten. The phase structure is again visible, with slightly more stable behavior in the middle layers and a pronounced acceleration in the final phase, which here again appears to begin around layer 12.

**MLP output norms** Following Lad et al. (2025), we examine the L2 norm of the final MLP layer's output to probe for their proposed residual sharpening mechanism. While the relatively constant policy entropy suggests no sharpening at the distributional level, the MLP output norms exhibit an extreme last-layer increase, suggesting a feature consolidation process analogous to that observed in language models.

### D.2 Ranking correlation (Kendall's $\tau$) analysis

**Chess models** To analyze the stability of move preferences, we compute Kendall's $\tau$ rank correlation between intermediate and final policies (Figure 7). We calculate this metric in two ways: first using all legal moves, and second, to mitigate noise from low-probability moves that are never seriously considered, only over moves that appear in the top five at any layer. Both approaches yield nearly identical results, with the correlation being negative in early layers and remaining low until around the 12th layer, where it increases sharply. The 5th and 95th percentiles in the top-5 moves analysis show that there are positions where intermediate layer move rankings are either very similar or very dissimilar to the final ranking until late in the network. The phase structure is again apparent, with high variability in early layers, relatively stable correlations across the middle phase up to layer 11, and a sharp rise beginning around layer 12, consistent with the transition to the late phase.

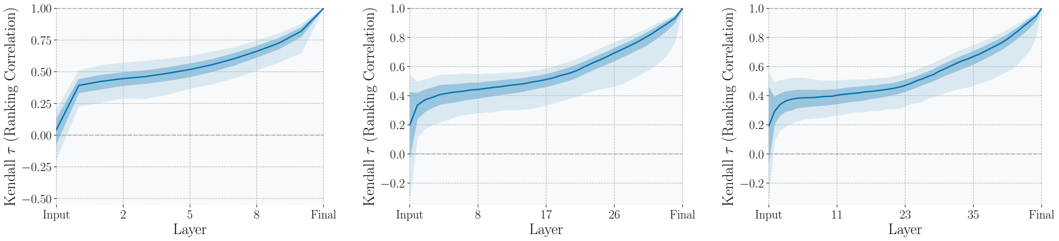


(a) Leela (All Moves)

(b) Leela (Top-5 Moves)

Figure 7: Kendall's $\tau$ move ranking correlation for Leela Chess Zero. The shaded regions represent the 5th-95th and 25th-75th percentile ranges. Shaded background indicated the proposed phases.

**Language models** Since we have not found analysis with this metric for language models, we conducted our own experiments for comparison using the GPT-2 series (small, large, XL) (Radford et al., 2019). The plots show a clearly monotonic increase in median correlation with a notable jump from the input embedding to the 0th layer, which is expected since the input embedding represents only the pure embedding of the previous token. While clear differences to Leela are observed, they may largely stem from domain- and architecture-specific factors—such as the vastly larger vocabulary or the limited last-token scope of the logit lens resulting from the autoregressive prediction setup discussed in the introduction.



(a) GPT-2 Small

(b) GPT-2 Large

(c) GPT-2 XL

Figure 8: Kendall's $\tau$ token ranking correlation for GPT-2 models. The shaded regions represent the 5th-95th and 25th-75th percentile ranges.

# E Complete tables for the tournament and Lichess play

## E.1 Internal tournament

We evaluate playing strength through internal round-robin tournaments processed with BayesElo (Coulom, 2008) using the default confidence parameter of $0.5$. Each model pairing played 200 distinct openings from the Encyclopedia of Chess Openings (Matanović, 1978), with one game per side at deterministic temperature $\tau = 0$ and five games per side at stochastic temperature $\tau = 1.0$. Tables 2 and 3 report the resulting Elo ratings, computed from all match outcomes. Each table lists the model's Elo estimate with asymmetric confidence bounds, total score percentage (wins plus half draws), average opponent rating, and draw rate. The results show consistent improvement in playing strength with network depth, with clear early- and late-layer acceleration in both deterministic and stochastic settings, while middle layers exhibit a prolonged performance plateau.

Table 2: Playing strength evaluation through internal tournament ($\tau = 0$)

| **Rank** | **Model** | **Elo** | + | − | **Games** | **Score (%)** | **Avg. Oppo.** | **Draws (%)** |
|---|---|---|---|---|---|---|---|---|
| 1 | Lc0 Policy Net (Anchor) | 2292 | 25 | 25 | 6400 | 97 | 1056 | 2 |
| 2 | Full Model | 2263 | 25 | 24 | 6400 | 96 | 1058 | 3 |
| 3 | Logit Lens Layer 13 | 1681 | 22 | 22 | 6400 | 86 | 1094 | 1 |
| 4 | Logit Lens Layer 12 | 1337 | 12 | 12 | 6400 | 74 | 1116 | 6 |
| 5 | Logit Lens Layer 11 | 1083 | 10 | 9 | 6400 | 55 | 1132 | 9 |
| 6 | Logit Lens Layer 10 | 1057 | 10 | 9 | 6400 | 53 | 1133 | 13 |
| 7 | Logit Lens Layer 9 | 1042 | 9 | 9 | 6400 | 52 | 1134 | 14 |
| 8 | Logit Lens Layer 7 | 1014 | 9 | 9 | 6400 | 49 | 1136 | 15 |
| 9 | Logit Lens Layer 5 | 1007 | 9 | 9 | 6400 | 49 | 1136 | 20 |
| 10 | Logit Lens Layer 8 | 1006 | 9 | 9 | 6400 | 48 | 1136 | 13 |
| 11 | Logit Lens Layer 6 | 993 | 9 | 9 | 6400 | 47 | 1137 | 15 |
| 12 | Logit Lens Layer 4 | 962 | 9 | 9 | 6400 | 44 | 1139 | 21 |
| 13 | Logit Lens Layer 3 | 871 | 9 | 9 | 6400 | 35 | 1145 | 18 |
| 14 | Logit Lens Layer 2 | 790 | 9 | 9 | 6400 | 27 | 1150 | 21 |
| 15 | Logit Lens Layer 1 | 699 | 10 | 10 | 6400 | 18 | 1156 | 21 |
| 16 | Logit Lens Layer 0 | 650 | 10 | 11 | 6400 | 16 | 1159 | 11 |
| 17 | Logit Lens Input | 443 | 14 | 15 | 6400 | 5 | 1172 | 6 |

Table 3: Playing strength evaluation through internal tournament ($\tau = 1$)

| **Rank** | **Model** | **Elo** | + | − | **Games** | **Score (%)** | **Avg. Oppo.** | **Draws (%)** |
|---|---|---|---|---|---|---|---|---|
| 1 | Lc0 Policy Net (Anchor) | 2292 | 32 | 28 | 32000 | 100 | 1040 | 0 |
| 2 | Full Model | 1640 | 8 | 8 | 32000 | 88 | 1081 | 1 |
| 3 | Logit Lens Layer 13 | 1394 | 6 | 5 | 32000 | 76 | 1096 | 3 |
| 4 | Logit Lens Layer 12 | 1355 | 6 | 5 | 32000 | 74 | 1099 | 4 |
| 5 | Logit Lens Layer 11 | 1151 | 5 | 5 | 32000 | 57 | 1112 | 6 |
| 6 | Logit Lens Layer 10 | 1113 | 5 | 5 | 32000 | 53 | 1114 | 7 |
| 7 | Logit Lens Layer 9 | 1110 | 5 | 5 | 32000 | 53 | 1114 | 7 |
| 8 | Logit Lens Layer 5 | 1098 | 5 | 5 | 32000 | 52 | 1115 | 8 |
| 9 | Logit Lens Layer 4 | 1080 | 5 | 5 | 32000 | 50 | 1116 | 8 |
| 10 | Logit Lens Layer 8 | 1069 | 5 | 5 | 32000 | 49 | 1117 | 7 |
| 11 | Logit Lens Layer 7 | 1068 | 5 | 5 | 32000 | 49 | 1117 | 7 |
| 12 | Logit Lens Layer 6 | 1064 | 5 | 5 | 32000 | 49 | 1117 | 7 |
| 13 | Logit Lens Layer 3 | 911 | 5 | 5 | 32000 | 34 | 1127 | 8 |
| 14 | Logit Lens Layer 2 | 813 | 5 | 6 | 32000 | 26 | 1133 | 8 |
| 15 | Logit Lens Layer 1 | 708 | 6 | 6 | 32000 | 18 | 1139 | 7 |
| 16 | Logit Lens Layer 0 | 701 | 6 | 6 | 32000 | 18 | 1140 | 6 |
| 17 | Logit Lens Input | 369 | 9 | 9 | 32000 | 4 | 1160 | 1 |

### E.2 Lichess bot performance

To validate our internal tournament findings in a real-world setting, we deployed layer-wise policies as bots on Lichess (Lichess.org, 2025) using the bot API framework (Lichess Bot Devs, 2025). Due to rate-limiting constraints, we distributed bots across multiple cloud instances and used policy sampling for the first five moves to introduce opening diversity. The bots participated in Bullet (1+0, 2+1), Blitz (3+0, 3+2, 5+0, 5+3), and Rapid (10+0, 10+5, 15+10) time controls until ratings stabilized. Several experimental constraints shaped the evaluation: (1) bots played exclusively against other bots per Lichess policy, (2) the limited pool of weak bots meant repeated matchups for early layers, and (3) deterministic play after the opening often led to repetitive game. Despite these limitations, the results largely corroborate our internal tournament findings, showing progressive improvement across layers with the most significant transitions occurring at similar points in the network depth. Table 4 presents the playing strength and performance statistics for our layer-wise Lichess bots, with bot names linked to their respective Lichess profiles.

Table 4: Performance of layer-wise policies on Lichess across different time controls

| **Bot** | **Rating** | | | **Total** | **Performance** | | |
|---|---|---|---|---|---|---|---|
| | **Bullet** | **Blitz** | **Rapid** | **Games** | **W** | **D** | **L** |
| LLLBot-In | 693 ± 54 | 518 ± 45 | 558 ± 45 | 437 | 2 | 150 | 285 |
| LLLBot-0 | 904 ± 54 | 651 ± 45 | 816 ± 50 | 419 | 2 | 85 | 332 |
| LLLBot-1 | 891 ± 54 | 681 ± 45 | 709 ± 46 | 431 | 5 | 92 | 334 |
| LLLBot-2 | 916 ± 50 | 688 ± 45 | 697 ± 45 | 453 | 13 | 134 | 306 |
| LLLBot-3 | 972 ± 55 | 741 ± 45 | 915 ± 50 | 463 | 9 | 80 | 374 |
| LLLBot-4 | 1009 ± 55 | 769 ± 45 | 717 ± 45 | 461 | 12 | 136 | 313 |
| LLLBot-5 | 926 ± 56 | 774 ± 45 | 939 ± 51 | 468 | 23 | 66 | 379 |
| LLLBot-6 | 984 ± 64 | 850 ± 45 | 1021 ± 46 | 446 | 22 | 62 | 362 |
| LLLBot-7 | 1052 ± 55 | 803 ± 45 | 1018 ± 49 | 452 | 23 | 50 | 379 |
| LLLBot-8 | 994 ± 63 | 843 ± 45 | 1032 ± 47 | 450 | 23 | 58 | 369 |
| LLLBot-9 | 1061 ± 58 | 832 ± 45 | 957 ± 54 | 445 | 32 | 47 | 366 |
| LLLBot-10 | 1064 ± 59 | 960 ± 45 | 1107 ± 45 | 494 | 39 | 73 | 382 |
| LLLBot-11 | 1129 ± 54 | 948 ± 45 | 1095 ± 45 | 436 | 34 | 46 | 356 |
| LLLBot-12 | 1331 ± 49 | 1252 ± 45 | 1290 ± 45 | 415 | 74 | 63 | 278 |
| LLLBot-13 | 1659 ± 48 | 1581 ± 45 | 1581 ± 45 | 368 | 124 | 39 | 205 |
| LLLBot-Full | 2246 ± 52 | 2274 ± 45 | 2253 ± 52 | 316 | 197 | 32 | 87 |

## F Complete probabilities for the example puzzle from Figure 1

The puzzle (Figure 9) features a knight sacrifice leading to mate (PV: 1... ♘**g3+** 2. **h×g3** ♖**h6#**). The input layer exhibits piece-specific bias toward queen moves, with multiple queen captures dominating early probabilities. The winning move ♘**g3+** first becomes the top choice at layer 5 (21.09%) but is then overtaken by ♕**g1+**, which dominates layers 6-10. This temporary preference for ♕**g1+** likely reflects a learned heuristic of the model that queen checks are tactically important, even though this particular check ultimately loses the queen without compensation. After layer 10, ♘**g3+** regains its position as the preferred move, with the exception of layer 12 where a rook lift ♖**h6** (54.15%) briefly becomes the top candidate, potentially pinning the h-pawn and threatening the king. The winning move's probability follows a non-monotonic trajectory, fluctuating throughout the layers before ultimately surging to 47.38% at layer 13 and 87.55% in the final output. Despite these fluctuations, this example demonstrates a relatively smooth progression compared to other cases in Appendix G, where the correct move is either identified immediately after early layers or remains unconsidered until the very late layers. Full probabilities are in Table 5.

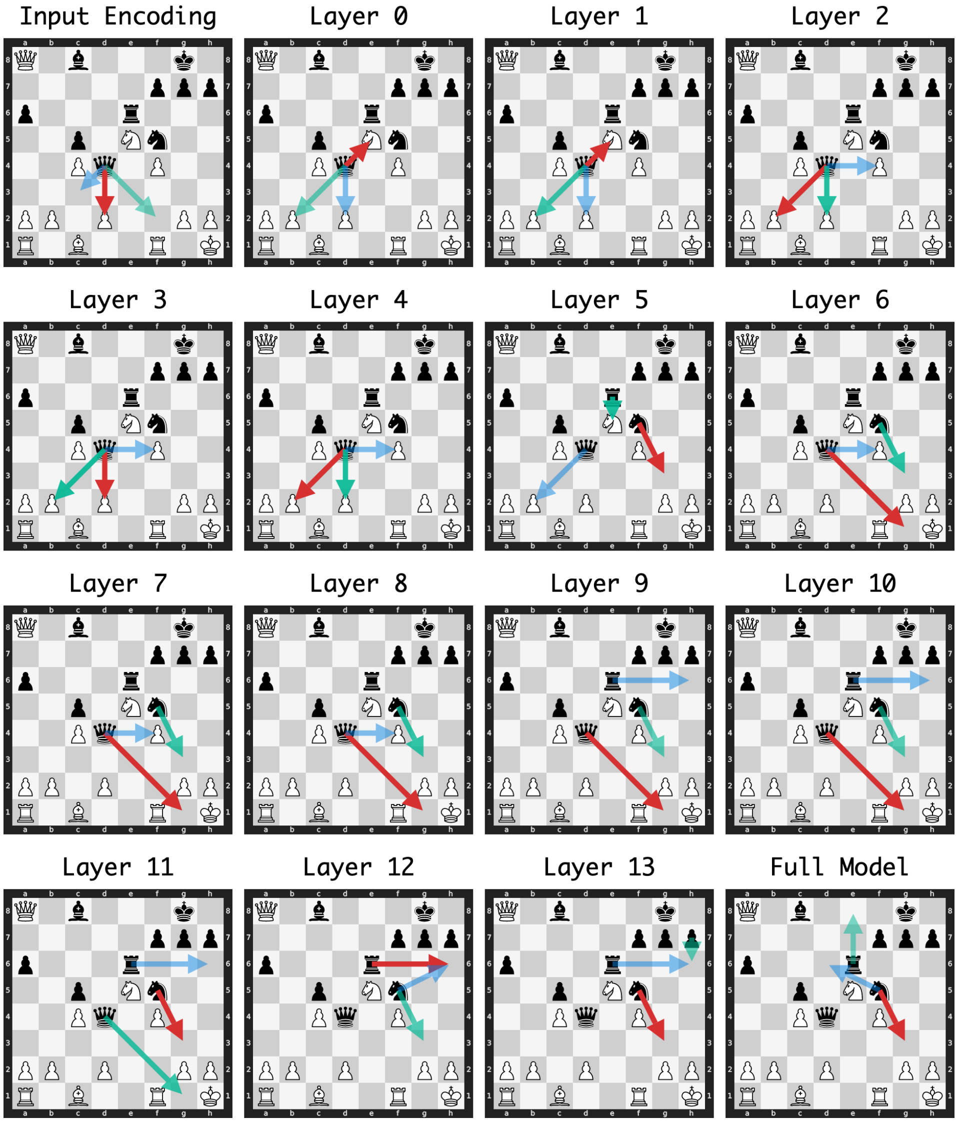


Figure 9: Layer-wise policy evolution for the puzzle in Figure 1.

Table 5: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♘g3+** | 3.70% | 0.04% | 0.20% | 2.30% | 8.94% | 0.62% | 21.09% | 24.77% |
| **♖h6** | 0.01% | 0.01% | 0.03% | 0.10% | 0.11% | 0.04% | 0.03% | 3.50% |
| **♕g1+** | 6.51% | 1.65% | 5.34% | 8.90% | 8.76% | 1.06% | 2.94% | 29.18% |
| **♕×b2** | 5.96% | 17.19% | 20.98% | 24.84% | 22.25% | 27.29% | 13.75% | 3.22% |
| **♕×e5** | 1.26% | 27.24% | 21.76% | 3.97% | 2.18% | 2.71% | 2.28% | 0.85% |
| **♕×d2** | 17.39% | 15.37% | 20.20% | 22.17% | 24.44% | 24.74% | 12.49% | 2.41% |
| **♖×e5** | 0.19% | 13.28% | 13.17% | 8.52% | 5.92% | 11.72% | 19.41% | 8.12% |
| **h6** | 0.01% | 1.16% | 0.01% | 0.01% | 0.00% | 0.01% | 0.00% | 0.00% |
| **♕×f4** | 4.20% | 7.53% | 7.52% | 15.33% | 12.42% | 15.28% | 13.52% | 16.19% |
| **♕×c4** | 4.38% | 10.67% | 7.54% | 11.15% | 12.02% | 12.68% | 13.12% | 8.54% |
| **♕f2** | 11.68% | 0.11% | 0.69% | 0.08% | 0.16% | 0.06% | 0.02% | 0.09% |
| **♕c3** | 11.47% | 0.03% | 0.09% | 0.49% | 0.28% | 0.28% | 0.23% | 0.11% |
| **♘h6** | 0.09% | 0.05% | 0.01% | 0.03% | 0.00% | 0.02% | 0.03% | 1.90% |
| **♕d3** | 9.82% | 0.02% | 0.23% | 0.05% | 0.29% | 0.04% | 0.05% | 0.08% |
| **♕e3** | 9.55% | 0.06% | 0.07% | 0.17% | 0.06% | 0.07% | 0.06% | 0.12% |
| **♕e4** | 5.76% | 0.13% | 0.28% | 0.03% | 0.07% | 0.03% | 0.04% | 0.05% |
| **♘e3** | 4.74% | 0.05% | 0.10% | 0.29% | 0.10% | 0.11% | 0.04% | 0.02% |
| **h5** | 0.02% | 2.46% | 0.11% | 0.03% | 0.02% | 0.01% | 0.01% | 0.01% |
| **♘d6** | 0.07% | 0.01% | 0.02% | 0.05% | 0.16% | 0.03% | 0.04% | 0.05% |
| **♖e8** | 0.01% | 0.80% | 0.36% | 0.08% | 0.03% | 0.01% | 0.01% | 0.02% |
| **♔h8** | 0.39% | 0.03% | 0.02% | 0.11% | 0.56% | 1.32% | 0.16% | 0.14% |
| **f6** | 0.13% | 0.02% | 0.00% | 0.00% | 0.04% | 0.01% | 0.02% | 0.05% |
| **♖c6** | 0.20% | 0.84% | 0.47% | 0.41% | 0.32% | 0.66% | 0.21% | 0.16% |
| **♕d8** | 0.05% | 0.15% | 0.04% | 0.21% | 0.06% | 0.05% | 0.02% | 0.02% |
| **g6** | 0.10% | 0.05% | 0.01% | 0.04% | 0.03% | 0.02% | 0.02% | 0.01% |
| **♘e7** | 0.25% | 0.00% | 0.01% | 0.03% | 0.02% | 0.03% | 0.02% | 0.02% |
| **g5** | 0.78% | 0.18% | 0.19% | 0.04% | 0.02% | 0.02% | 0.01% | 0.01% |
| **a5** | 0.01% | 0.75% | 0.15% | 0.18% | 0.05% | 0.04% | 0.03% | 0.01% |
| **♕d6** | 0.05% | 0.01% | 0.02% | 0.03% | 0.02% | 0.04% | 0.02% | 0.02% |
| **♔f8** | 0.07% | 0.03% | 0.03% | 0.03% | 0.15% | 0.31% | 0.06% | 0.09% |
| **♕d7** | 0.06% | 0.01% | 0.01% | 0.03% | 0.05% | 0.10% | 0.02% | 0.02% |
| **♕d5** | 0.35% | 0.01% | 0.02% | 0.07% | 0.08% | 0.05% | 0.11% | 0.09% |
| **♖d6** | 0.11% | 0.01% | 0.03% | 0.04% | 0.03% | 0.04% | 0.01% | 0.02% |
| **♖b6** | 0.05% | 0.01% | 0.01% | 0.03% | 0.09% | 0.24% | 0.04% | 0.03% |
| **♖f6** | 0.16% | 0.01% | 0.17% | 0.08% | 0.05% | 0.03% | 0.01% | 0.02% |
| **♖g6** | 0.08% | 0.01% | 0.02% | 0.04% | 0.06% | 0.08% | 0.01% | 0.01% |
| **♘h4** | 0.13% | 0.01% | 0.01% | 0.03% | 0.08% | 0.01% | 0.04% | 0.07% |
| **♖e7** | 0.21% | 0.01% | 0.07% | 0.03% | 0.04% | 0.13% | 0.02% | 0.02% |

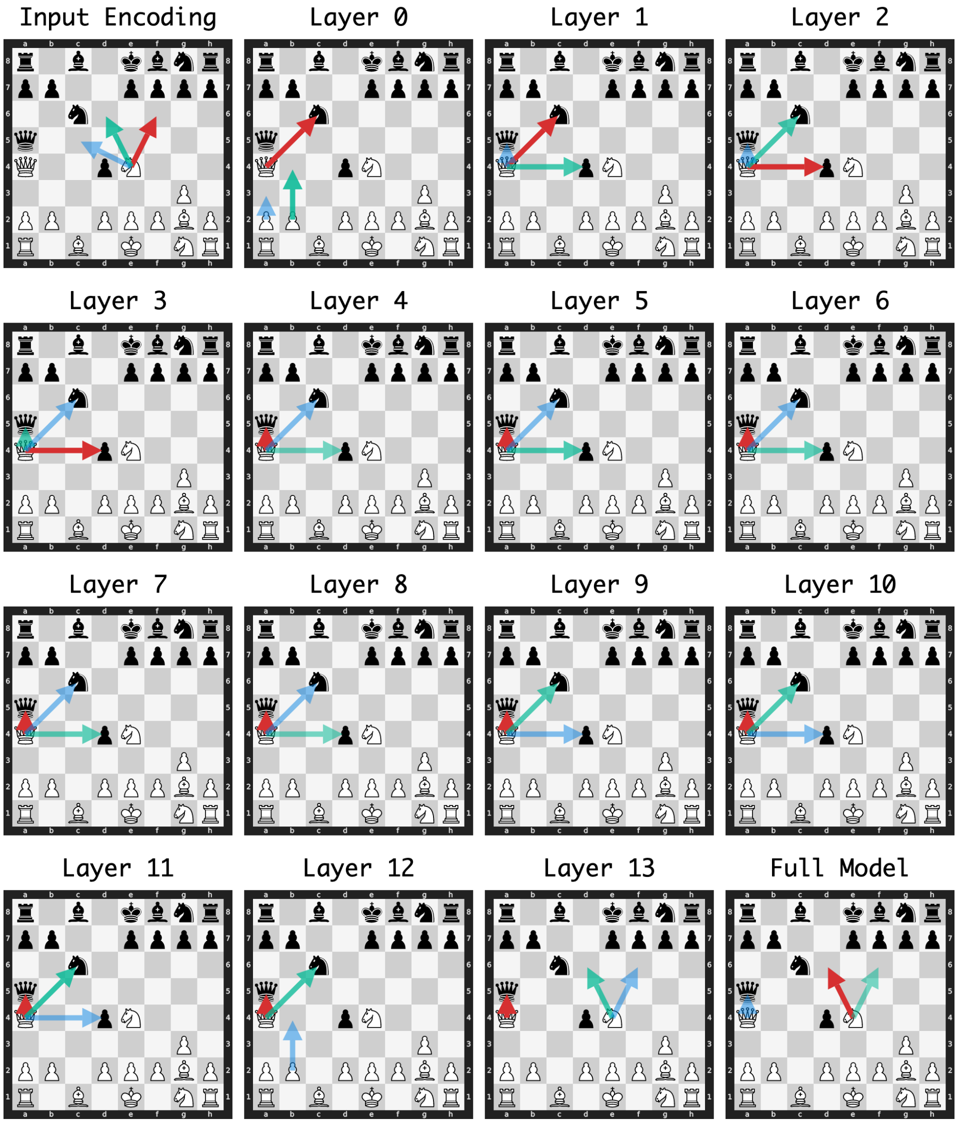
Table 6: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| Move | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 | Layer 13 | Final |
|---|---|---|---|---|---|---|---|---|
| ♘g3+ | 22.95% | 18.27% | 18.48% | 21.85% | 33.31% | 25.23% | 47.38% | 87.55% |
| ♖h6 | 7.29% | 14.22% | 9.13% | 10.10% | 25.73% | 54.15% | 10.59% | 0.24% |
| ♕g1+ | 31.08% | 19.92% | 48.52% | 49.22% | 30.02% | 3.93% | 2.22% | 0.28% |
| ♕×b2 | 0.96% | 4.92% | 0.59% | 0.47% | 0.14% | 0.06% | 0.85% | 0.25% |
| ♕×e5 | 0.52% | 0.71% | 0.99% | 0.30% | 0.09% | 0.04% | 0.58% | 0.37% |
| ♕×d2 | 0.78% | 3.62% | 0.50% | 0.43% | 0.07% | 0.04% | 0.58% | 0.24% |
| ♖×e5 | 7.27% | 5.84% | 5.52% | 3.57% | 2.38% | 1.35% | 0.70% | 0.37% |
| h6 | 0.05% | 0.38% | 0.30% | 1.07% | 2.08% | 3.49% | 17.35% | 0.25% |
| ♕×f4 | 13.18% | 15.89% | 6.21% | 6.16% | 0.78% | 0.17% | 0.48% | 0.27% |
| ♕×c4 | 9.54% | 10.13% | 4.60% | 1.64% | 0.36% | 0.14% | 0.64% | 0.29% |
| ♕f2 | 0.14% | 0.38% | 0.29% | 1.37% | 0.14% | 0.03% | 0.25% | 0.27% |
| ♕c3 | 0.22% | 0.09% | 0.05% | 0.02% | 0.04% | 0.01% | 0.46% | 0.22% |
| ♘h6 | 4.86% | 4.45% | 3.21% | 2.62% | 3.90% | 10.01% | 6.52% | 0.23% |
| ♕d3 | 0.05% | 0.09% | 0.11% | 0.06% | 0.11% | 0.05% | 0.40% | 0.26% |
| ♕e3 | 0.05% | 0.03% | 0.01% | 0.01% | 0.02% | 0.02% | 0.29% | 0.25% |
| ♕e4 | 0.04% | 0.03% | 0.02% | 0.01% | 0.02% | 0.01% | 0.24% | 0.26% |
| ♘e3 | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.18% | 0.21% |
| h5 | 0.09% | 0.04% | 0.07% | 0.12% | 0.10% | 0.06% | 0.72% | 0.25% |
| ♘d6 | 0.10% | 0.09% | 0.37% | 0.17% | 0.08% | 0.10% | 1.95% | 0.82% |
| ♖e8 | 0.10% | 0.16% | 0.20% | 0.09% | 0.08% | 0.29% | 0.94% | 1.35% |
| ♔h8 | 0.15% | 0.15% | 0.29% | 0.29% | 0.16% | 0.28% | 0.27% | 0.27% |
| f6 | 0.13% | 0.07% | 0.06% | 0.06% | 0.06% | 0.13% | 1.02% | 0.29% |
| ♖c6 | 0.09% | 0.05% | 0.04% | 0.01% | 0.02% | 0.02% | 0.18% | 0.34% |
| ♕d8 | 0.03% | 0.04% | 0.05% | 0.05% | 0.04% | 0.04% | 0.49% | 0.81% |
| g6 | 0.02% | 0.02% | 0.02% | 0.03% | 0.03% | 0.05% | 0.79% | 0.31% |
| ♘e7 | 0.01% | 0.01% | 0.02% | 0.01% | 0.01% | 0.01% | 0.79% | 0.39% |
| g5 | 0.01% | 0.01% | 0.02% | 0.02% | 0.01% | 0.01% | 0.50% | 0.29% |
| a5 | 0.02% | 0.01% | 0.01% | 0.00% | 0.00% | 0.01% | 0.14% | 0.23% |
| ♕d6 | 0.02% | 0.02% | 0.03% | 0.03% | 0.03% | 0.03% | 0.42% | 0.28% |
| ♔f8 | 0.06% | 0.10% | 0.09% | 0.05% | 0.02% | 0.04% | 0.19% | 0.35% |
| ♕d7 | 0.01% | 0.02% | 0.02% | 0.02% | 0.03% | 0.03% | 0.15% | 0.35% |
| ♕d5 | 0.09% | 0.10% | 0.08% | 0.04% | 0.06% | 0.02% | 0.34% | 0.27% |
| ♖d6 | 0.01% | 0.02% | 0.02% | 0.01% | 0.02% | 0.02% | 0.32% | 0.29% |
| ♖b6 | 0.02% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.20% | 0.28% |
| ♖f6 | 0.02% | 0.03% | 0.01% | 0.02% | 0.01% | 0.03% | 0.19% | 0.28% |
| ♖g6 | 0.01% | 0.02% | 0.01% | 0.02% | 0.03% | 0.04% | 0.17% | 0.28% |
| ♘h4 | 0.04% | 0.05% | 0.01% | 0.01% | 0.01% | 0.03% | 0.26% | 0.20% |
| ♖e7 | 0.01% | 0.01% | 0.01% | 0.00% | 0.01% | 0.01% | 0.26% | 0.26% |

# G Puzzle case studies

## G.1 Case study 1: knight fork and discovered attack

This puzzle (Figure 10) features a tactical sequence beginning with a knight check that sets up a discovered attack (PV: 1. ♘**d6+ e×d6** 2. ♗**×c6+**). After the knight is captured, the bishop delivers a check, forcing the black king to move and allowing White to capture the undefended queen. The input layer exhibits piece-specific bias toward knight moves, which are abandoned after layer 0 as the model shifts focus to queen captures. The queen trade ♛**×a5** dominates through most layers with the queen being protected by a knight. The winning move ♘**d6+** maintains low probability until layer 12 (2.77%), then suddenly jumps to 26.23% at layer 13 before becoming the decisive top choice in the final output. This sudden increase after layer 12—the same layer Jenner et al. (2024) identified as containing a "look-ahead" head—may indicate the model requires multi-move analysis rather than simple tactical heuristics for this position. Full probabilities are in Table 7.



Input Encoding

Layer 0

Layer 1

Layer 2

Layer 3

Layer 4

Layer 5

Layer 6

Layer 7

Layer 8

Layer 9

Layer 10

Layer 11

Layer 12

Layer 13

Full Model

Figure 10: Layer-wise policy evolution for Puzzle 12864.

Table 7: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♘d6+** | 21.53% | 0.02% | 1.32% | 4.01% | 0.10% | 0.04% | 2.37% | 3.13% |
| **♕×a5** | 0.04% | 6.74% | 7.68% | 10.13% | 25.17% | 48.88% | 44.79% | 51.85% |
| **♕×c6+** | 0.46% | 35.02% | 50.31% | 24.26% | 21.72% | 21.79% | 13.27% | 10.83% |
| **♕×d4** | 2.15% | 11.37% | 26.27% | 41.95% | 30.81% | 23.59% | 25.63% | 20.43% |
| **b4** | 2.28% | 27.81% | 0.56% | 1.23% | 1.40% | 0.01% | 0.02% | 0.35% |
| **♘f6+** | 23.64% | 0.01% | 0.15% | 0.82% | 0.03% | 0.03% | 0.01% | 0.06% |
| **♘c5** | 17.72% | 0.02% | 0.10% | 0.23% | 0.16% | 0.06% | 0.31% | 0.19% |
| **f4** | 17.23% | 0.22% | 0.03% | 0.04% | 0.03% | 0.07% | 0.04% | 0.19% |
| **a3** | 0.02% | 11.77% | 4.90% | 7.07% | 0.14% | 0.32% | 0.07% | 0.15% |
| **♘f3** | 0.22% | 0.01% | 0.05% | 0.36% | 5.93% | 0.12% | 5.70% | 4.50% |
| **g4** | 5.44% | 0.30% | 0.35% | 0.88% | 0.26% | 0.11% | 0.02% | 0.01% |
| **♕b4** | 0.09% | 0.21% | 0.61% | 0.93% | 4.29% | 0.28% | 0.11% | 0.43% |
| **♗f3** | 0.75% | 0.01% | 0.89% | 0.66% | 2.86% | 0.92% | 3.73% | 2.78% |
| **♕b3** | 0.09% | 0.04% | 0.20% | 1.03% | 2.62% | 0.44% | 1.40% | 2.02% |
| **h3** | 0.03% | 2.71% | 0.21% | 0.48% | 0.03% | 0.08% | 0.10% | 0.05% |
| **♘g5** | 2.46% | 0.02% | 0.02% | 0.14% | 0.06% | 0.10% | 0.03% | 0.02% |
| **♕a3** | 0.03% | 0.72% | 0.78% | 1.16% | 1.61% | 0.40% | 0.36% | 1.05% |
| **e3** | 1.36% | 0.55% | 0.07% | 0.32% | 0.17% | 0.40% | 0.22% | 0.18% |
| **♗f1** | 0.00% | 0.13% | 1.30% | 0.55% | 0.48% | 0.20% | 0.13% | 0.06% |
| **♘c3** | 1.28% | 0.01% | 0.93% | 0.43% | 0.07% | 0.07% | 0.03% | 0.07% |
| **♔f1** | 0.64% | 0.52% | 1.28% | 0.42% | 0.27% | 0.11% | 0.12% | 0.07% |
| **h4** | 0.06% | 0.88% | 1.27% | 0.77% | 0.15% | 0.13% | 0.09% | 0.09% |
| **b3** | 0.27% | 0.23% | 0.15% | 1.25% | 0.01% | 0.03% | 0.04% | 0.02% |
| **f3** | 0.80% | 0.00% | 0.00% | 0.00% | 0.01% | 0.09% | 0.06% | 0.13% |
| **♗h3** | 0.02% | 0.05% | 0.11% | 0.13% | 0.82% | 0.69% | 0.34% | 0.20% |
| **♕c4** | 0.60% | 0.04% | 0.02% | 0.07% | 0.19% | 0.08% | 0.17% | 0.22% |
| **♖b1** | 0.22% | 0.08% | 0.09% | 0.17% | 0.11% | 0.34% | 0.28% | 0.29% |
| **♔d1** | 0.20% | 0.08% | 0.07% | 0.09% | 0.18% | 0.29% | 0.14% | 0.16% |
| **♕d1** | 0.05% | 0.06% | 0.04% | 0.08% | 0.04% | 0.09% | 0.08% | 0.05% |
| **♕c2** | 0.09% | 0.04% | 0.04% | 0.09% | 0.04% | 0.10% | 0.11% | 0.16% |
| **♘h3** | 0.11% | 0.27% | 0.12% | 0.15% | 0.03% | 0.07% | 0.13% | 0.11% |
| **♕b5** | 0.11% | 0.05% | 0.08% | 0.08% | 0.23% | 0.08% | 0.10% | 0.13% |

Table 8: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♘d6+** | 4.07% | 3.57% | 4.96% | 6.54% | 3.94% | 2.77% | 26.23% | 65.34% |
| **♕×a5** | 50.76% | 45.54% | 45.91% | 51.70% | 48.54% | 50.24% | 41.97% | 12.18% |
| **♕×c6+** | 10.38% | 16.40% | 25.83% | 24.85% | 31.97% | 28.81% | 3.95% | 0.61% |
| **♕×d4** | 19.78% | 17.56% | 14.54% | 9.69% | 5.24% | 5.33% | 0.87% | 0.29% |
| **b4** | 1.03% | 3.59% | 1.10% | 1.88% | 4.79% | 8.06% | 4.34% | 0.29% |
| **♘f6+** | 0.01% | 0.02% | 0.04% | 0.02% | 0.01% | 0.01% | 12.46% | 13.67% |
| **♘c5** | 0.14% | 0.19% | 0.45% | 0.31% | 0.25% | 0.08% | 0.70% | 0.29% |
| **f4** | 0.75% | 0.93% | 0.66% | 0.98% | 2.89% | 1.73% | 1.54% | 0.26% |
| **a3** | 0.65% | 0.59% | 0.21% | 0.32% | 0.14% | 0.13% | 1.41% | 0.28% |
| **♘f3** | 2.91% | 2.80% | 1.57% | 1.31% | 0.60% | 0.17% | 0.86% | 0.28% |
| **g4** | 0.01% | 0.01% | 0.01% | 0.02% | 0.02% | 0.02% | 0.13% | 0.29% |
| **♕b4** | 1.85% | 1.26% | 0.58% | 0.17% | 0.11% | 0.06% | 0.06% | 0.27% |
| **♗f3** | 1.33% | 0.75% | 0.66% | 0.25% | 0.09% | 0.44% | 0.38% | 0.29% |
| **♕b3** | 3.44% | 3.32% | 1.26% | 0.54% | 0.18% | 0.46% | 0.28% | 0.29% |
| **h3** | 0.04% | 0.03% | 0.02% | 0.02% | 0.02% | 0.02% | 0.25% | 0.30% |
| **♘g5** | 0.06% | 0.17% | 0.01% | 0.02% | 0.01% | 0.10% | 0.26% | 0.27% |
| **♕a3** | 0.90% | 1.53% | 0.67% | 0.28% | 0.19% | 0.13% | 0.10% | 0.29% |
| **e3** | 0.25% | 0.23% | 0.12% | 0.09% | 0.10% | 0.05% | 0.43% | 0.28% |
| **♗f1** | 0.04% | 0.02% | 0.03% | 0.02% | 0.02% | 0.03% | 0.13% | 0.34% |
| **♘c3** | 0.08% | 0.08% | 0.06% | 0.05% | 0.03% | 0.09% | 0.72% | 0.29% |
| **♔f1** | 0.07% | 0.05% | 0.07% | 0.07% | 0.02% | 0.02% | 0.12% | 0.33% |
| **h4** | 0.14% | 0.09% | 0.04% | 0.03% | 0.03% | 0.02% | 0.15% | 0.33% |
| **b3** | 0.12% | 0.05% | 0.03% | 0.03% | 0.01% | 0.01% | 0.14% | 0.25% |
| **f3** | 0.05% | 0.07% | 0.03% | 0.02% | 0.01% | 0.07% | 1.11% | 0.26% |
| **♗h3** | 0.16% | 0.18% | 0.16% | 0.04% | 0.04% | 0.04% | 0.13% | 0.32% |
| **♕c4** | 0.23% | 0.40% | 0.31% | 0.30% | 0.32% | 0.62% | 0.31% | 0.28% |
| **♖b1** | 0.18% | 0.14% | 0.19% | 0.03% | 0.03% | 0.04% | 0.23% | 0.31% |
| **♔d1** | 0.15% | 0.11% | 0.13% | 0.10% | 0.13% | 0.13% | 0.16% | 0.32% |
| **♕d1** | 0.06% | 0.03% | 0.03% | 0.03% | 0.02% | 0.03% | 0.13% | 0.32% |
| **♕c2** | 0.14% | 0.08% | 0.09% | 0.11% | 0.14% | 0.09% | 0.16% | 0.29% |
| **♘h3** | 0.09% | 0.09% | 0.04% | 0.01% | 0.03% | 0.08% | 0.22% | 0.29% |
| **♕b5** | 0.11% | 0.14% | 0.18% | 0.14% | 0.08% | 0.12% | 0.07% | 0.29% |

### G.2 Case study 2: early and consistent solution lock-in

This puzzle (Figure 11) features a back-rank mate initiated by a queen sacrifice (PV: 1. **♕×e8+ ♖×e8** 2. **♖×e8#**). The winning move **♕×e8+** becomes the top candidate from layer 0 and maintains this position throughout all layers, with only brief competition from **b×c3** at layer 7 (50.8% vs 46.1%). The early convergence may result from the winning move combining both queen and capture biases observed in initial layers. This monotonic pattern contrasts with the delayed convergence seen in previous cases and represents one example where low entropy and KL divergence are maintained across layers from early stages. Full probabilities are in Table 9.

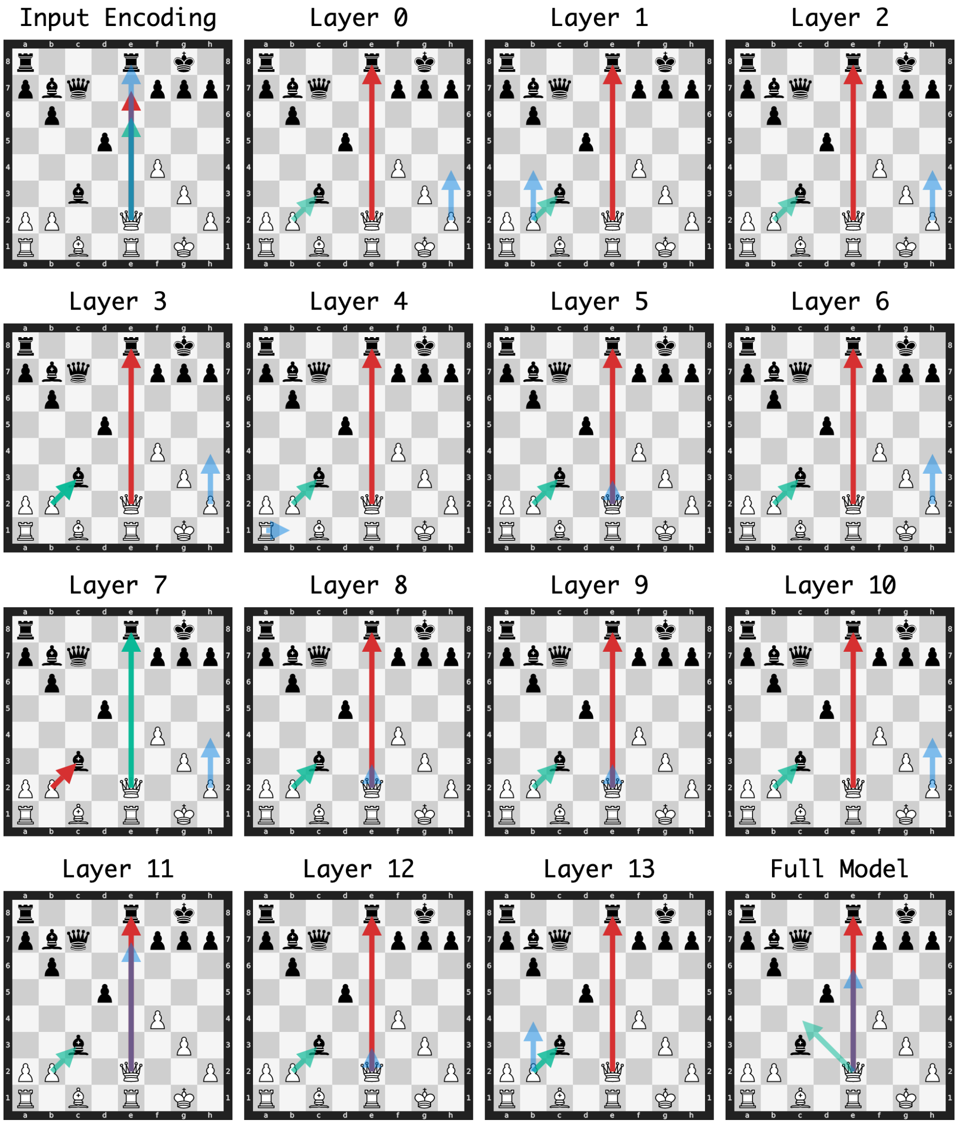


Figure 11: Layer-wise policy evolution for Puzzle 9745.

Table 9: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♕×e8+** | 10.74% | 51.58% | 74.47% | 82.24% | 51.25% | 68.41% | 64.05% | 66.06% |
| **b×c3** | 0.55% | 13.01% | 16.18% | 13.09% | 45.34% | 29.43% | 34.47% | 30.90% |
| **♕e7** | 35.40% | 0.37% | 0.26% | 0.07% | 0.10% | 0.05% | 0.03% | 0.06% |
| **♕e6** | 21.65% | 1.15% | 0.68% | 0.09% | 0.08% | 0.05% | 0.06% | 0.09% |
| **h4** | 0.05% | 10.18% | 2.06% | 0.54% | 0.52% | 0.07% | 0.04% | 0.26% |
| **b4** | 1.21% | 9.84% | 3.20% | 0.31% | 0.04% | 0.02% | 0.01% | 0.02% |
| **a4** | 0.05% | 6.55% | 0.43% | 0.09% | 0.01% | 0.01% | 0.04% | 0.11% |
| **♕e5** | 5.75% | 1.80% | 0.72% | 0.49% | 0.51% | 0.07% | 0.05% | 0.06% |
| **♕f2** | 4.09% | 0.32% | 0.08% | 0.18% | 0.16% | 0.04% | 0.05% | 0.18% |
| **g4** | 2.60% | 0.74% | 0.02% | 0.03% | 0.07% | 0.02% | 0.03% | 0.09% |
| **♕c4** | 2.50% | 0.11% | 0.56% | 0.39% | 0.12% | 0.08% | 0.09% | 0.15% |
| **♕e4** | 2.18% | 0.03% | 0.06% | 0.07% | 0.07% | 0.05% | 0.04% | 0.12% |
| **♕g2** | 2.08% | 0.02% | 0.03% | 0.10% | 0.04% | 0.03% | 0.03% | 0.06% |
| **♕b5** | 1.78% | 0.02% | 0.08% | 0.08% | 0.05% | 0.04% | 0.05% | 0.07% |
| **♕f3** | 1.44% | 0.04% | 0.02% | 0.06% | 0.03% | 0.04% | 0.04% | 0.05% |
| **a3** | 0.02% | 1.42% | 0.07% | 0.04% | 0.03% | 0.04% | 0.02% | 0.25% |
| **♔h1** | 1.23% | 0.04% | 0.06% | 0.06% | 0.05% | 0.06% | 0.03% | 0.07% |
| **♕g4** | 1.11% | 0.07% | 0.07% | 0.11% | 0.09% | 0.07% | 0.05% | 0.10% |
| **f5** | 1.07% | 0.93% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.08% |
| **♕e3** | 1.00% | 0.03% | 0.04% | 0.04% | 0.06% | 0.04% | 0.18% | 0.21% |
| **♔f2** | 0.40% | 0.28% | 0.14% | 0.51% | 0.47% | 0.13% | 0.05% | 0.09% |
| **♕d1** | 0.15% | 0.45% | 0.15% | 0.15% | 0.05% | 0.08% | 0.03% | 0.06% |
| **♗e3** | 0.13% | 0.00% | 0.01% | 0.06% | 0.02% | 0.03% | 0.08% | 0.14% |
| **b3** | 0.17% | 0.29% | 0.05% | 0.05% | 0.04% | 0.03% | 0.03% | 0.11% |
| **♖b1** | 0.14% | 0.02% | 0.03% | 0.09% | 0.08% | 0.39% | 0.03% | 0.08% |
| **♕a6** | 0.12% | 0.07% | 0.07% | 0.05% | 0.10% | 0.04% | 0.03% | 0.04% |
| **♕d3** | 0.38% | 0.02% | 0.04% | 0.10% | 0.06% | 0.05% | 0.05% | 0.07% |
| **♕h5** | 0.15% | 0.07% | 0.05% | 0.08% | 0.10% | 0.06% | 0.06% | 0.09% |
| **♕c2** | 0.33% | 0.03% | 0.03% | 0.06% | 0.04% | 0.07% | 0.04% | 0.08% |
| **♖f1** | 0.34% | 0.02% | 0.03% | 0.11% | 0.03% | 0.05% | 0.01% | 0.01% |
| **h3** | 0.02% | 0.26% | 0.06% | 0.06% | 0.04% | 0.01% | 0.01% | 0.04% |
| **♔g2** | 0.18% | 0.01% | 0.04% | 0.25% | 0.13% | 0.11% | 0.04% | 0.07% |
| **♖d1** | 0.21% | 0.15% | 0.10% | 0.04% | 0.03% | 0.16% | 0.03% | 0.04% |
| **♕d2** | 0.14% | 0.02% | 0.03% | 0.04% | 0.07% | 0.04% | 0.04% | 0.03% |
| **♔f1** | 0.20% | 0.02% | 0.03% | 0.13% | 0.07% | 0.07% | 0.02% | 0.01% |
| **♕f1** | 0.22% | 0.03% | 0.03% | 0.09% | 0.04% | 0.03% | 0.02% | 0.02% |
| **♗d2** | 0.21% | 0.00% | 0.01% | 0.01% | 0.01% | 0.02% | 0.03% | 0.04% |

Table 10: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| Move | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 | Layer 13 | Final |
|---|---|---|---|---|---|---|---|---|
| **♕×e8+** | 46.12% | 59.87% | 71.83% | 66.69% | 76.07% | 73.66% | 56.40% | 88.91% |
| **b×c3** | 50.83% | 37.08% | 25.57% | 29.89% | 20.90% | 23.58% | 34.74% | 0.28% |
| **♕e7** | 0.09% | 0.13% | 0.13% | 0.38% | 0.52% | 0.18% | 0.21% | 0.33% |
| **♕e6** | 0.07% | 0.03% | 0.03% | 0.03% | 0.08% | 0.07% | 0.23% | 0.41% |
| **h4** | 0.37% | 0.20% | 0.17% | 0.53% | 0.17% | 0.07% | 0.24% | 0.34% |
| **b4** | 0.02% | 0.04% | 0.13% | 0.17% | 0.06% | 0.15% | 0.94% | 0.33% |
| **a4** | 0.10% | 0.05% | 0.02% | 0.06% | 0.07% | 0.10% | 0.83% | 0.29% |
| **♕e5** | 0.07% | 0.06% | 0.07% | 0.06% | 0.16% | 0.18% | 0.18% | 0.37% |
| **♕f2** | 0.06% | 0.05% | 0.03% | 0.03% | 0.02% | 0.04% | 0.12% | 0.24% |
| **g4** | 0.10% | 0.09% | 0.15% | 0.25% | 0.14% | 0.12% | 0.58% | 0.30% |
| **♕c4** | 0.12% | 0.06% | 0.09% | 0.07% | 0.07% | 0.07% | 0.32% | 0.45% |
| **♕e4** | 0.09% | 0.05% | 0.07% | 0.08% | 0.17% | 0.15% | 0.18% | 0.40% |
| **♕g2** | 0.05% | 0.03% | 0.02% | 0.02% | 0.03% | 0.06% | 0.13% | 0.27% |
| **♕b5** | 0.06% | 0.10% | 0.07% | 0.05% | 0.06% | 0.07% | 0.17% | 0.21% |
| **♕f3** | 0.06% | 0.03% | 0.04% | 0.03% | 0.05% | 0.04% | 0.15% | 0.34% |
| **a3** | 0.18% | 0.09% | 0.07% | 0.09% | 0.05% | 0.04% | 0.42% | 0.32% |
| **♔h1** | 0.05% | 0.04% | 0.02% | 0.02% | 0.03% | 0.05% | 0.09% | 0.33% |
| **♕g4** | 0.07% | 0.04% | 0.05% | 0.04% | 0.05% | 0.04% | 0.16% | 0.35% |
| **f5** | 0.15% | 0.55% | 0.22% | 0.19% | 0.29% | 0.14% | 0.45% | 0.30% |
| **♕e3** | 0.29% | 0.62% | 0.26% | 0.07% | 0.07% | 0.20% | 0.22% | 0.21% |
| **♔f2** | 0.04% | 0.04% | 0.04% | 0.05% | 0.05% | 0.09% | 0.09% | 0.24% |
| **♕d1** | 0.05% | 0.02% | 0.04% | 0.04% | 0.06% | 0.04% | 0.14% | 0.29% |
| **♗e3** | 0.12% | 0.16% | 0.20% | 0.12% | 0.04% | 0.08% | 0.41% | 0.21% |
| **b3** | 0.12% | 0.06% | 0.10% | 0.07% | 0.06% | 0.04% | 0.30% | 0.41% |
| **♖b1** | 0.05% | 0.03% | 0.08% | 0.39% | 0.18% | 0.12% | 0.40% | 0.37% |
| **♕a6** | 0.05% | 0.05% | 0.04% | 0.04% | 0.05% | 0.04% | 0.14% | 0.38% |
| **♕d3** | 0.07% | 0.07% | 0.06% | 0.03% | 0.06% | 0.08% | 0.17% | 0.28% |
| **♕h5** | 0.07% | 0.05% | 0.04% | 0.06% | 0.06% | 0.04% | 0.24% | 0.37% |
| **♕c2** | 0.11% | 0.05% | 0.06% | 0.06% | 0.07% | 0.06% | 0.16% | 0.34% |
| **♖f1** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.03% | 0.17% | 0.25% |
| **h3** | 0.09% | 0.04% | 0.04% | 0.09% | 0.04% | 0.02% | 0.22% | 0.33% |
| **♔g2** | 0.04% | 0.04% | 0.03% | 0.05% | 0.09% | 0.12% | 0.10% | 0.27% |
| **♖d1** | 0.04% | 0.03% | 0.04% | 0.03% | 0.03% | 0.04% | 0.20% | 0.27% |
| **♕d2** | 0.03% | 0.03% | 0.04% | 0.02% | 0.01% | 0.03% | 0.07% | 0.27% |
| **♔f1** | 0.01% | 0.01% | 0.02% | 0.03% | 0.04% | 0.06% | 0.08% | 0.26% |
| **♕f1** | 0.03% | 0.02% | 0.02% | 0.02% | 0.05% | 0.06% | 0.14% | 0.26% |
| **♗d2** | 0.09% | 0.07% | 0.10% | 0.11% | 0.05% | 0.06% | 0.22% | 0.23% |

### G.3 Case study 3: knight sacrifice with rook mate

This puzzle (Figure 12) featured in Jenner et al. (2024) requires a knight sacrifice (PV: 1. ♘**g6+ h×g6** 2. ♖**h4#**). The input layer exhibits piece-specific bias toward knight moves. The winning move ♘**g6+** peaks as the top candidate at layer 3, then disappears as the model favors ♘**e6** (layers 4-7) and ♕**e6** (layers 7-9), the latter threatening the opposing queen while protected by the knight. The correct sacrifice re-emerges from layer 8, becoming the preferred move after layer 10. This pattern demonstrates the model's exploration of tactically sound alternatives before converging on the optimal sequence. Full probabilities are in Table 11.

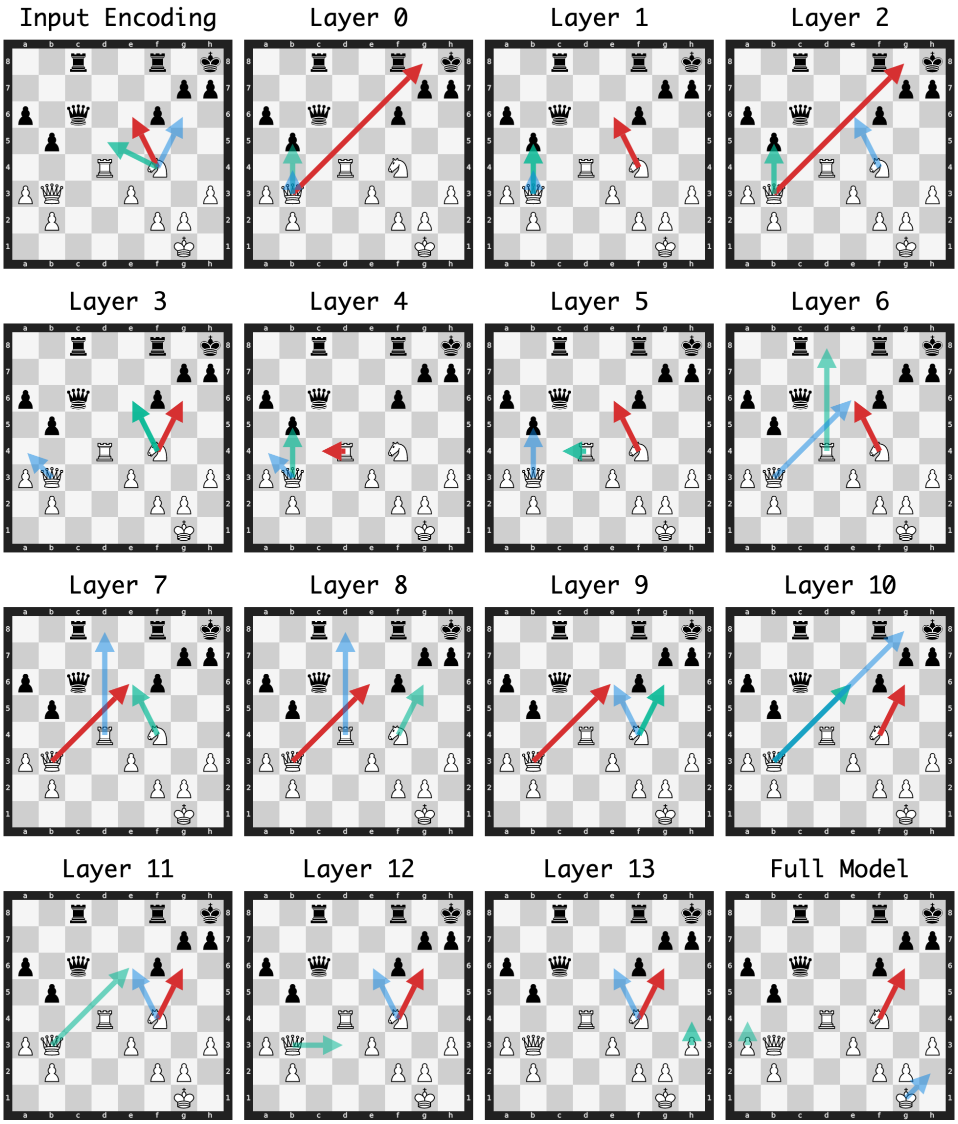


Figure 12: Layer-wise policy evolution for Puzzle 483.

Table 11: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♘g6+** | 11.14% | 0.37% | 0.64% | 0.19% | 14.28% | 0.64% | 6.00% | 3.24% |
| **♖c4** | 0.87% | 0.19% | 0.39% | 0.77% | 6.80% | 48.97% | 18.55% | 9.85% |
| **♕g8+** | 0.04% | 37.56% | 9.18% | 29.26% | 8.63% | 1.70% | 1.71% | 2.12% |
| **♘e6** | 20.13% | 3.37% | 20.25% | 9.54% | 13.41% | 2.89% | 26.73% | 23.12% |
| **♕e6** | 0.75% | 4.64% | 2.37% | 0.30% | 2.86% | 1.23% | 6.72% | 14.76% |
| **h4** | 0.07% | 7.82% | 0.18% | 0.47% | 0.43% | 0.14% | 0.34% | 0.12% |
| **♕×b5** | 5.37% | 16.97% | 17.42% | 16.55% | 8.24% | 17.00% | 8.75% | 3.83% |
| **♖d8** | 2.71% | 0.08% | 0.07% | 0.22% | 1.35% | 1.12% | 0.63% | 15.68% |
| **♘d5** | 15.13% | 0.07% | 0.23% | 3.24% | 5.28% | 1.14% | 2.13% | 0.74% |
| **♕b4** | 0.05% | 14.59% | 12.53% | 4.31% | 3.02% | 0.45% | 0.32% | 0.28% |
| **♔h2** | 0.15% | 0.54% | 0.15% | 1.23% | 0.60% | 1.00% | 2.85% | 7.21% |
| **♕d3** | 0.27% | 0.11% | 0.78% | 0.46% | 0.61% | 0.47% | 0.34% | 0.77% |
| **♖b4** | 0.10% | 3.95% | 9.58% | 1.28% | 1.03% | 0.10% | 0.03% | 0.03% |
| **♖d7** | 9.53% | 0.30% | 0.25% | 0.23% | 0.43% | 0.56% | 0.24% | 0.43% |
| **♖d6** | 8.93% | 0.14% | 0.19% | 0.34% | 0.67% | 0.78% | 0.21% | 0.27% |
| **♕a4** | 0.03% | 1.16% | 4.29% | 5.99% | 8.87% | 3.53% | 6.25% | 2.69% |
| **♘d3** | 0.85% | 0.09% | 0.08% | 1.15% | 2.33% | 0.51% | 3.51% | 2.67% |
| **♕c2** | 0.35% | 0.22% | 0.82% | 4.07% | 2.35% | 0.98% | 0.24% | 1.14% |
| **♖a4** | 0.12% | 0.31% | 0.90% | 1.27% | 2.02% | 1.44% | 7.49% | 2.75% |
| **a4** | 0.06% | 2.74% | 0.10% | 0.34% | 0.33% | 0.16% | 0.28% | 0.13% |
| **♖d5** | 5.56% | 0.17% | 3.27% | 1.48% | 0.25% | 0.08% | 0.14% | 0.34% |
| **♕f7** | 0.76% | 0.52% | 4.44% | 1.96% | 1.69% | 0.51% | 0.18% | 0.34% |
| **♕d1** | 0.04% | 0.14% | 0.22% | 0.38% | 0.67% | 0.95% | 0.28% | 0.59% |
| **e4** | 3.87% | 0.39% | 1.09% | 3.23% | 0.82% | 0.15% | 0.08% | 0.14% |
| **♘e2** | 2.93% | 0.06% | 0.42% | 0.09% | 0.05% | 0.11% | 0.22% | 0.11% |
| **g4** | 2.93% | 0.39% | 0.17% | 0.75% | 0.87% | 0.23% | 0.31% | 0.23% |
| **♕c4** | 0.87% | 0.58% | 2.68% | 2.76% | 2.12% | 1.16% | 1.68% | 1.66% |
| **♖d2** | 0.21% | 0.10% | 0.15% | 0.31% | 0.68% | 2.37% | 0.31% | 0.16% |
| **♔h1** | 1.61% | 0.31% | 0.68% | 1.08% | 0.30% | 2.28% | 0.66% | 0.47% |
| **♔f1** | 0.26% | 0.18% | 0.07% | 0.42% | 1.18% | 1.99% | 0.28% | 0.36% |
| **f3** | 0.66% | 0.09% | 0.31% | 0.61% | 1.97% | 0.25% | 0.11% | 0.07% |
| **♖e4** | 1.21% | 0.15% | 1.07% | 1.43% | 1.29% | 0.67% | 0.49% | 1.73% |
| **♖d3** | 0.30% | 0.10% | 1.65% | 0.88% | 0.18% | 0.09% | 0.16% | 0.06% |
| **♕d5** | 0.62% | 0.13% | 1.45% | 0.54% | 0.76% | 0.56% | 0.45% | 0.62% |
| **♕a2** | 0.02% | 0.57% | 1.14% | 1.06% | 1.43% | 1.42% | 0.56% | 0.49% |
| **♕c3** | 0.14% | 0.18% | 0.36% | 1.03% | 0.85% | 1.20% | 0.27% | 0.28% |
| **g3** | 0.45% | 0.41% | 0.17% | 0.38% | 1.00% | 0.18% | 0.14% | 0.23% |
| **♘h5** | 0.79% | 0.22% | 0.04% | 0.11% | 0.18% | 0.39% | 0.14% | 0.07% |
| **♖d1** | 0.11% | 0.08% | 0.20% | 0.28% | 0.19% | 0.58% | 0.25% | 0.22% |

Table 12: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| Move | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 | Layer 13 | Final |
|---|---|---|---|---|---|---|---|---|
| **♘g6+** | 8.80% | 15.41% | 17.79% | 22.84% | 38.02% | 60.31% | 34.33% | 70.65% |
| **♖c4** | 5.83% | 3.22% | 2.65% | 1.14% | 0.42% | 0.13% | 0.23% | 0.18% |
| **♕g8+** | 3.09% | 2.18% | 7.24% | 9.71% | 1.63% | 0.72% | 0.53% | 0.16% |
| **♘e6** | 14.17% | 13.05% | 11.89% | 9.52% | 9.15% | 7.89% | 11.68% | 1.46% |
| **♕e6** | 20.13% | 19.05% | 19.79% | 18.19% | 12.49% | 3.40% | 3.67% | 0.48% |
| **h4** | 0.08% | 0.19% | 0.07% | 0.14% | 2.00% | 2.11% | 17.94% | 4.76% |
| **♕×b5** | 4.14% | 5.70% | 3.74% | 2.16% | 0.52% | 0.58% | 1.43% | 0.16% |
| **♖d8** | 10.70% | 14.67% | 6.79% | 4.26% | 6.70% | 1.06% | 0.38% | 0.17% |
| **♘d5** | 0.60% | 0.59% | 0.40% | 0.20% | 0.10% | 0.10% | 0.54% | 0.15% |
| **♕b4** | 0.46% | 0.74% | 0.55% | 0.57% | 0.42% | 0.55% | 0.58% | 0.71% |
| **♔h2** | 10.63% | 5.59% | 4.27% | 3.32% | 1.08% | 3.43% | 3.55% | 5.54% |
| **♕d3** | 2.55% | 2.07% | 3.87% | 5.28% | 7.64% | 9.98% | 6.51% | 4.04% |
| **♖b4** | 0.02% | 0.02% | 0.01% | 0.02% | 0.04% | 0.05% | 0.04% | 0.15% |
| **♖d7** | 0.34% | 0.30% | 0.34% | 0.42% | 0.88% | 0.38% | 0.16% | 0.21% |
| **♖d6** | 0.16% | 0.16% | 0.15% | 0.20% | 0.23% | 0.13% | 0.09% | 0.18% |
| **♕a4** | 1.47% | 1.62% | 0.72% | 1.28% | 1.32% | 0.33% | 0.33% | 0.14% |
| **♘d3** | 3.70% | 3.23% | 4.31% | 2.44% | 3.75% | 3.09% | 8.82% | 0.49% |
| **♕c2** | 1.90% | 3.52% | 5.13% | 7.69% | 2.28% | 0.20% | 0.19% | 0.16% |
| **♖a4** | 1.44% | 1.05% | 0.10% | 0.41% | 0.49% | 0.13% | 0.11% | 0.15% |
| **a4** | 0.25% | 0.08% | 0.03% | 0.04% | 0.14% | 0.05% | 1.32% | 5.80% |
| **♖d5** | 0.27% | 0.23% | 0.11% | 0.08% | 0.16% | 0.19% | 0.34% | 0.22% |
| **♕f7** | 0.90% | 1.19% | 1.59% | 3.37% | 2.46% | 0.28% | 0.26% | 0.20% |
| **♕d1** | 2.18% | 1.44% | 3.39% | 2.26% | 3.90% | 1.11% | 0.65% | 0.46% |
| **e4** | 0.14% | 0.10% | 0.16% | 0.16% | 0.15% | 0.19% | 0.44% | 0.18% |
| **♘e2** | 0.19% | 0.12% | 0.40% | 0.29% | 0.50% | 0.54% | 1.50% | 0.18% |
| **g4** | 0.27% | 0.18% | 0.08% | 0.06% | 0.09% | 0.06% | 0.21% | 0.40% |
| **♕c4** | 1.13% | 0.79% | 0.65% | 0.45% | 0.19% | 0.13% | 0.17% | 0.17% |
| **♖d2** | 0.07% | 0.07% | 0.11% | 0.02% | 0.02% | 0.06% | 0.27% | 0.25% |
| **♔h1** | 0.45% | 0.48% | 0.63% | 0.90% | 0.23% | 0.27% | 0.38% | 0.19% |
| **♔f1** | 0.57% | 0.49% | 0.49% | 0.32% | 1.00% | 0.95% | 0.23% | 0.15% |
| **f3** | 0.18% | 0.08% | 0.08% | 0.06% | 0.06% | 0.05% | 0.40% | 0.18% |
| **♖e4** | 1.11% | 0.58% | 0.85% | 0.77% | 0.78% | 0.65% | 0.29% | 0.19% |
| **♖d3** | 0.06% | 0.05% | 0.04% | 0.05% | 0.10% | 0.10% | 0.23% | 0.19% |
| **♕d5** | 0.76% | 0.98% | 0.85% | 0.76% | 0.66% | 0.30% | 0.70% | 0.16% |
| **♕a2** | 0.61% | 0.33% | 0.29% | 0.19% | 0.13% | 0.07% | 0.33% | 0.27% |
| **♕c3** | 0.25% | 0.16% | 0.18% | 0.26% | 0.15% | 0.15% | 0.14% | 0.15% |
| **g3** | 0.11% | 0.06% | 0.04% | 0.04% | 0.01% | 0.03% | 0.20% | 0.29% |
| **♘h5** | 0.10% | 0.09% | 0.02% | 0.05% | 0.05% | 0.08% | 0.33% | 0.18% |
| **♖d1** | 0.17% | 0.13% | 0.19% | 0.05% | 0.08% | 0.17% | 0.52% | 0.25% |

### G.4 Case study 4: queen sacrifice for rook mate

This puzzle (Figure 13) involves a queen sacrifice to enable a rook mate (PV: 1. ♕**e8+** ♖×**e8** 2. ♖×**e8#**). The winning move ♕**e8+** receives minimal probability until late layers, then increases sharply to become the top choice at layer 13 and dominates the final output (58.9%). Throughout most of the middle to late layers, the model favors **g×f7+**, which delivers check but lacks the forced mate continuation. Early layers prefer immediate material captures (♕×**a3**, ♖×**a3**) that maintain substantial probability until late layers. Full probabilities are in Table 13.

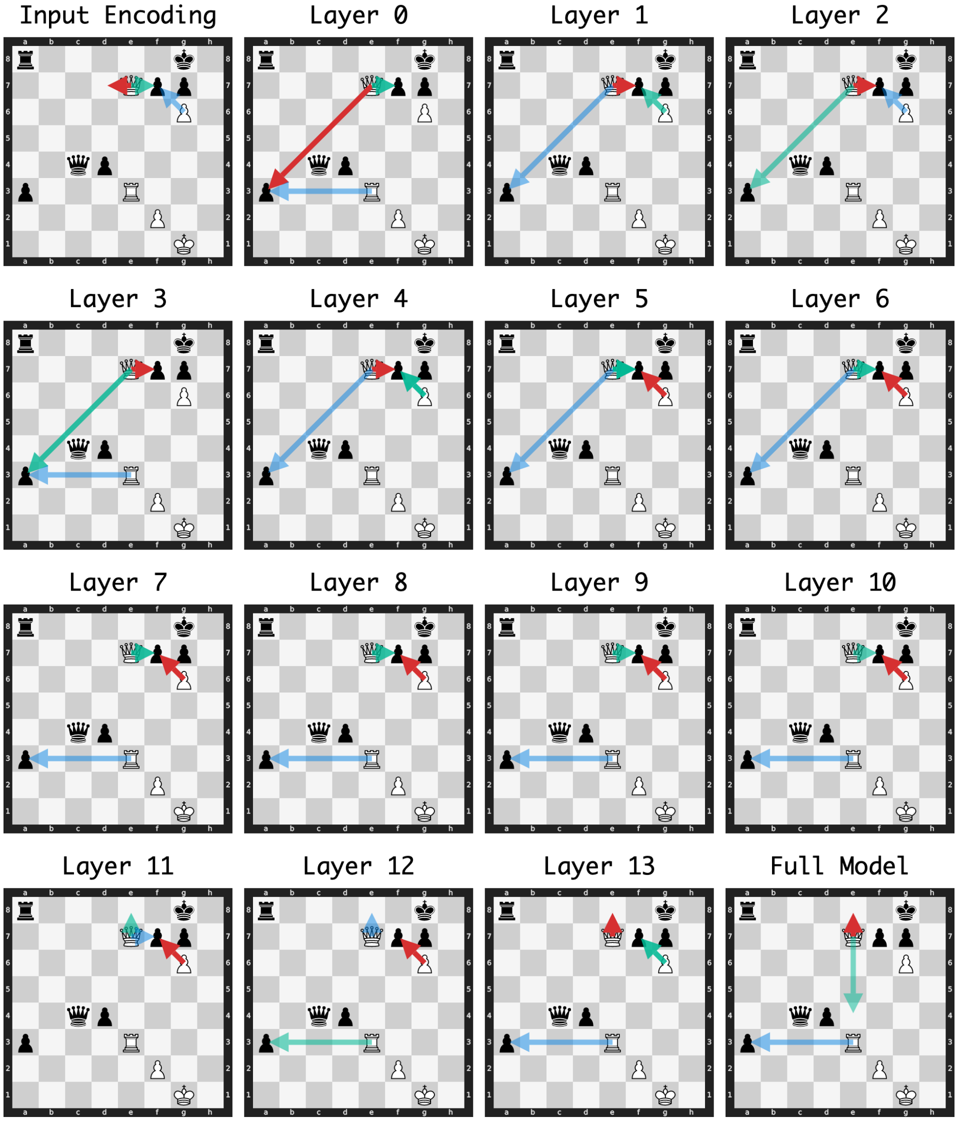


Figure 13: Layer-wise policy evolution for Puzzle 215.

Table 13: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| Move | Input | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
|---|---|---|---|---|---|---|---|---|
| **♕e8+** | 2.73% | 0.18% | 0.30% | 0.67% | 3.10% | 0.03% | 0.15% | 3.48% |
| **♕×f7+** | 13.47% | 30.73% | 38.98% | 32.49% | 33.46% | 37.33% | 33.63% | 31.74% |
| **g×f7+** | 12.07% | 3.16% | 26.81% | 27.44% | 14.66% | 34.00% | 33.96% | 38.81% |
| **♕×a3** | 0.62% | 34.10% | 18.65% | 28.12% | 26.36% | 15.37% | 15.63% | 10.99% |
| **♖×a3** | 0.05% | 22.90% | 8.54% | 7.36% | 16.37% | 12.10% | 13.80% | 10.10% |
| **♕d7** | 15.33% | 0.07% | 0.24% | 0.03% | 0.03% | 0.02% | 0.02% | 0.01% |
| **♕c7** | 11.27% | 0.04% | 0.04% | 0.03% | 0.02% | 0.02% | 0.01% | 0.01% |
| **♕e4** | 0.23% | 0.02% | 0.03% | 0.04% | 0.05% | 0.03% | 0.06% | 0.05% |
| **♕f6** | 7.92% | 0.04% | 0.24% | 0.06% | 0.06% | 0.02% | 0.05% | 0.03% |
| **♕g5** | 2.88% | 7.28% | 1.71% | 0.11% | 0.09% | 0.04% | 0.04% | 0.02% |
| **♖h3** | 0.01% | 0.04% | 0.05% | 0.26% | 0.34% | 0.03% | 0.70% | 1.98% |
| **♖f3** | 0.07% | 0.04% | 0.41% | 0.31% | 0.14% | 0.02% | 0.33% | 0.50% |
| **♕b7** | 2.30% | 0.06% | 0.05% | 0.02% | 0.02% | 0.02% | 0.02% | 0.02% |
| **♕d6** | 6.25% | 0.03% | 0.23% | 0.03% | 0.02% | 0.02% | 0.02% | 0.01% |
| **♕e6** | 5.98% | 0.05% | 0.08% | 0.08% | 0.04% | 0.02% | 0.02% | 0.02% |
| **♕c5** | 4.04% | 0.03% | 0.02% | 0.09% | 0.04% | 0.02% | 0.02% | 0.01% |
| **♕f8+** | 2.91% | 0.14% | 1.61% | 0.87% | 3.51% | 0.24% | 0.33% | 1.00% |
| **♕e5** | 3.49% | 0.03% | 0.09% | 0.08% | 0.03% | 0.02% | 0.02% | 0.01% |
| **♖e4** | 0.13% | 0.03% | 0.03% | 0.05% | 0.15% | 0.02% | 0.17% | 0.18% |
| **♕d8+** | 2.73% | 0.17% | 0.43% | 0.12% | 0.24% | 0.02% | 0.03% | 0.11% |
| **♕b4** | 1.54% | 0.07% | 0.10% | 0.21% | 0.09% | 0.03% | 0.01% | 0.01% |
| **f4** | 1.44% | 0.10% | 0.04% | 0.04% | 0.03% | 0.04% | 0.02% | 0.03% |
| **♖g3** | 0.03% | 0.03% | 0.10% | 0.17% | 0.12% | 0.02% | 0.33% | 0.30% |
| **♕a7** | 0.70% | 0.05% | 0.07% | 0.04% | 0.02% | 0.02% | 0.01% | 0.01% |
| **♕h4** | 0.68% | 0.11% | 0.10% | 0.18% | 0.04% | 0.02% | 0.02% | 0.03% |
| **♔g2** | 0.02% | 0.03% | 0.43% | 0.65% | 0.52% | 0.12% | 0.25% | 0.22% |
| **♖c3** | 0.03% | 0.03% | 0.03% | 0.02% | 0.06% | 0.07% | 0.02% | 0.02% |
| **♖e6** | 0.34% | 0.06% | 0.07% | 0.04% | 0.05% | 0.02% | 0.02% | 0.02% |
| **♖e5** | 0.27% | 0.03% | 0.16% | 0.06% | 0.04% | 0.02% | 0.05% | 0.04% |
| **♔h1** | 0.23% | 0.05% | 0.08% | 0.05% | 0.04% | 0.03% | 0.06% | 0.04% |
| **♔h2** | 0.02% | 0.04% | 0.09% | 0.14% | 0.12% | 0.06% | 0.12% | 0.13% |
| **♖b3** | 0.03% | 0.05% | 0.05% | 0.05% | 0.06% | 0.05% | 0.03% | 0.02% |
| **♖d3** | 0.05% | 0.03% | 0.05% | 0.05% | 0.03% | 0.03% | 0.03% | 0.03% |
| **♖e2** | 0.07% | 0.03% | 0.05% | 0.02% | 0.02% | 0.02% | 0.01% | 0.01% |
| **♖e1** | 0.01% | 0.12% | 0.04% | 0.02% | 0.01% | 0.05% | 0.01% | 0.01% |
| **f3** | 0.09% | 0.02% | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

Table 14: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♕e8+** | 1.73% | 7.55% | 7.69% | 10.74% | 18.80% | 17.46% | 21.69% | 58.86% |
| **♕×f7+** | 31.45% | 26.33% | 27.12% | 20.10% | 16.39% | 6.96% | 2.25% | 0.27% |
| **g×f7+** | 38.73% | 34.13% | 38.96% | 32.23% | 25.25% | 34.38% | 19.50% | 1.46% |
| **♕×a3** | 11.94% | 12.60% | 9.55% | 8.78% | 5.44% | 3.29% | 7.27% | 0.45% |
| **♖×a3** | 12.33% | 13.88% | 11.84% | 15.80% | 16.38% | 19.25% | 12.83% | 9.17% |
| **♕d7** | 0.01% | 0.01% | 0.01% | 0.01% | 0.03% | 0.02% | 0.18% | 0.24% |
| **♕c7** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.30% | 0.21% |
| **♕e4** | 0.06% | 0.15% | 0.39% | 0.57% | 0.91% | 1.17% | 8.04% | 9.86% |
| **♕f6** | 0.04% | 0.02% | 0.01% | 0.02% | 0.02% | 0.02% | 0.13% | 0.18% |
| **♕g5** | 0.02% | 0.02% | 0.01% | 0.01% | 0.01% | 0.01% | 0.06% | 0.20% |
| **♖h3** | 1.36% | 2.72% | 1.67% | 4.23% | 6.57% | 6.83% | 6.64% | 6.35% |
| **♖f3** | 0.35% | 0.46% | 0.27% | 2.00% | 5.91% | 6.43% | 5.91% | 1.98% |
| **♕b7** | 0.01% | 0.02% | 0.02% | 0.03% | 0.07% | 0.31% | 6.30% | 4.20% |
| **♕d6** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.00% | 0.06% | 0.18% |
| **♕e6** | 0.02% | 0.04% | 0.10% | 0.04% | 0.01% | 0.02% | 0.06% | 0.18% |
| **♕c5** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.06% | 0.22% |
| **♕f8+** | 0.75% | 0.60% | 0.97% | 3.51% | 0.60% | 0.04% | 0.46% | 0.22% |
| **♕e5** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.07% | 0.18% |
| **♖e4** | 0.10% | 0.18% | 0.10% | 0.41% | 2.02% | 2.08% | 3.10% | 0.75% |
| **♕d8+** | 0.06% | 0.16% | 0.13% | 0.24% | 0.40% | 0.01% | 0.07% | 0.25% |
| **♕b4** | 0.01% | 0.01% | 0.02% | 0.02% | 0.01% | 0.01% | 0.09% | 0.17% |
| **f4** | 0.06% | 0.12% | 0.17% | 0.27% | 0.16% | 0.24% | 0.93% | 0.16% |
| **♖g3** | 0.22% | 0.30% | 0.14% | 0.35% | 0.42% | 0.98% | 1.22% | 0.54% |
| **♕a7** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 1.06% | 0.77% |
| **♕h4** | 0.04% | 0.10% | 0.11% | 0.11% | 0.09% | 0.07% | 0.48% | 0.71% |
| **♔g2** | 0.29% | 0.19% | 0.27% | 0.15% | 0.05% | 0.01% | 0.09% | 0.18% |
| **♖c3** | 0.06% | 0.03% | 0.04% | 0.02% | 0.05% | 0.05% | 0.12% | 0.35% |
| **♖e6** | 0.02% | 0.03% | 0.03% | 0.03% | 0.04% | 0.03% | 0.08% | 0.24% |
| **♖e5** | 0.03% | 0.04% | 0.02% | 0.04% | 0.13% | 0.14% | 0.16% | 0.21% |
| **♔h1** | 0.03% | 0.03% | 0.03% | 0.03% | 0.04% | 0.05% | 0.11% | 0.21% |
| **♔h2** | 0.11% | 0.14% | 0.12% | 0.10% | 0.03% | 0.01% | 0.12% | 0.19% |
| **♖b3** | 0.03% | 0.02% | 0.03% | 0.03% | 0.04% | 0.04% | 0.18% | 0.19% |
| **♖d3** | 0.05% | 0.06% | 0.08% | 0.10% | 0.07% | 0.06% | 0.14% | 0.18% |
| **♖e2** | 0.02% | 0.01% | 0.02% | 0.01% | 0.01% | 0.01% | 0.08% | 0.17% |
| **♖e1** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.16% | 0.17% |
| **f3** | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.01% | 0.15% |

## G.5 Case study 5: bishop sacrifice for discovered attack

This puzzle (Figure 14) features a bishop sacrifice that wins material via a discovered attack (PV: 1. ♗**h2+** ♔×**h2** 2. ♖×**c6**). The model initially favors the immediate material gain ♖×**e3** through layer 11, while the winning move ♗**h2+** gains probability gradually from layer 5 and becomes the top choice after layer 12, demonstrating a late-stage override of a simple material-gain heuristic in favor of a deeper sacrifical sequence. Full probabilities are in Table 15.

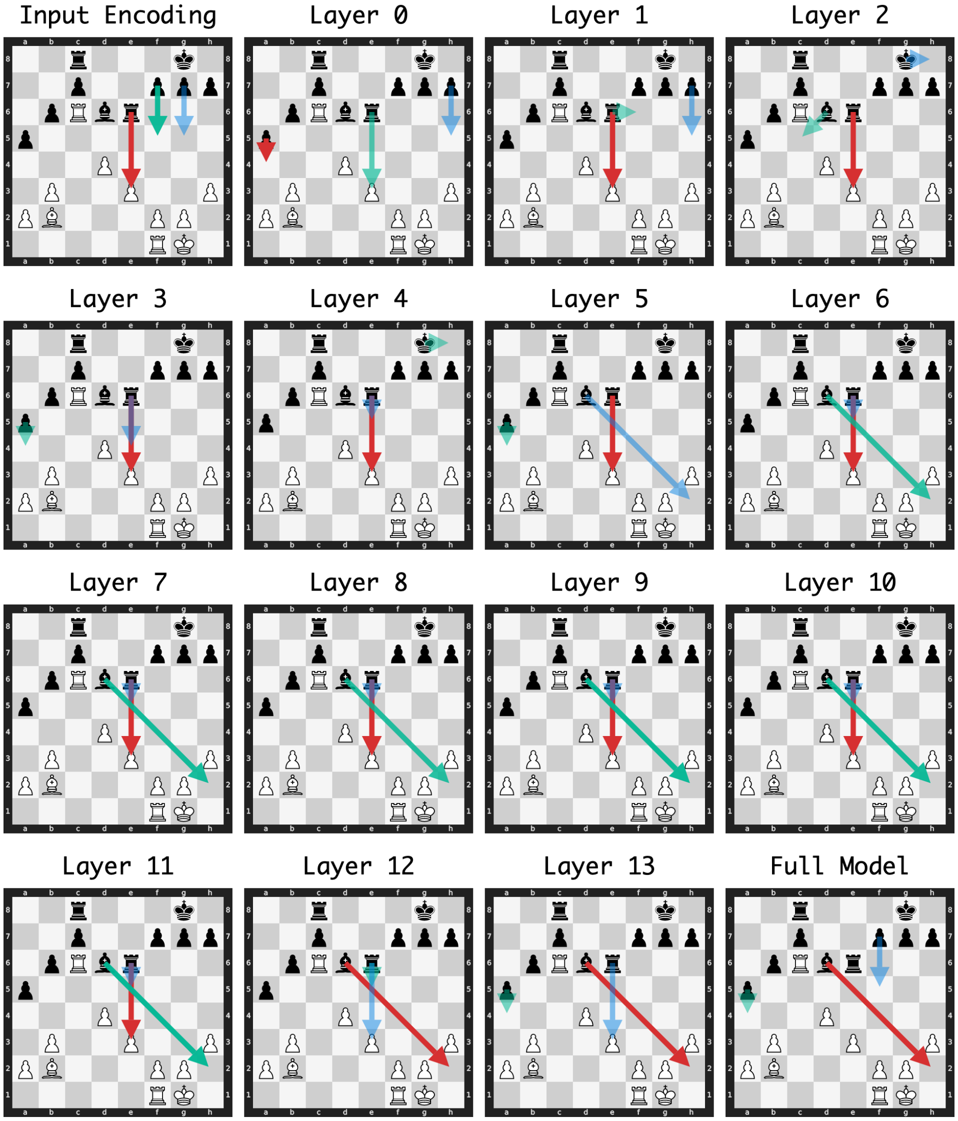


Figure 14: Layer-wise policy evolution for Puzzle 10363.

Table 15: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

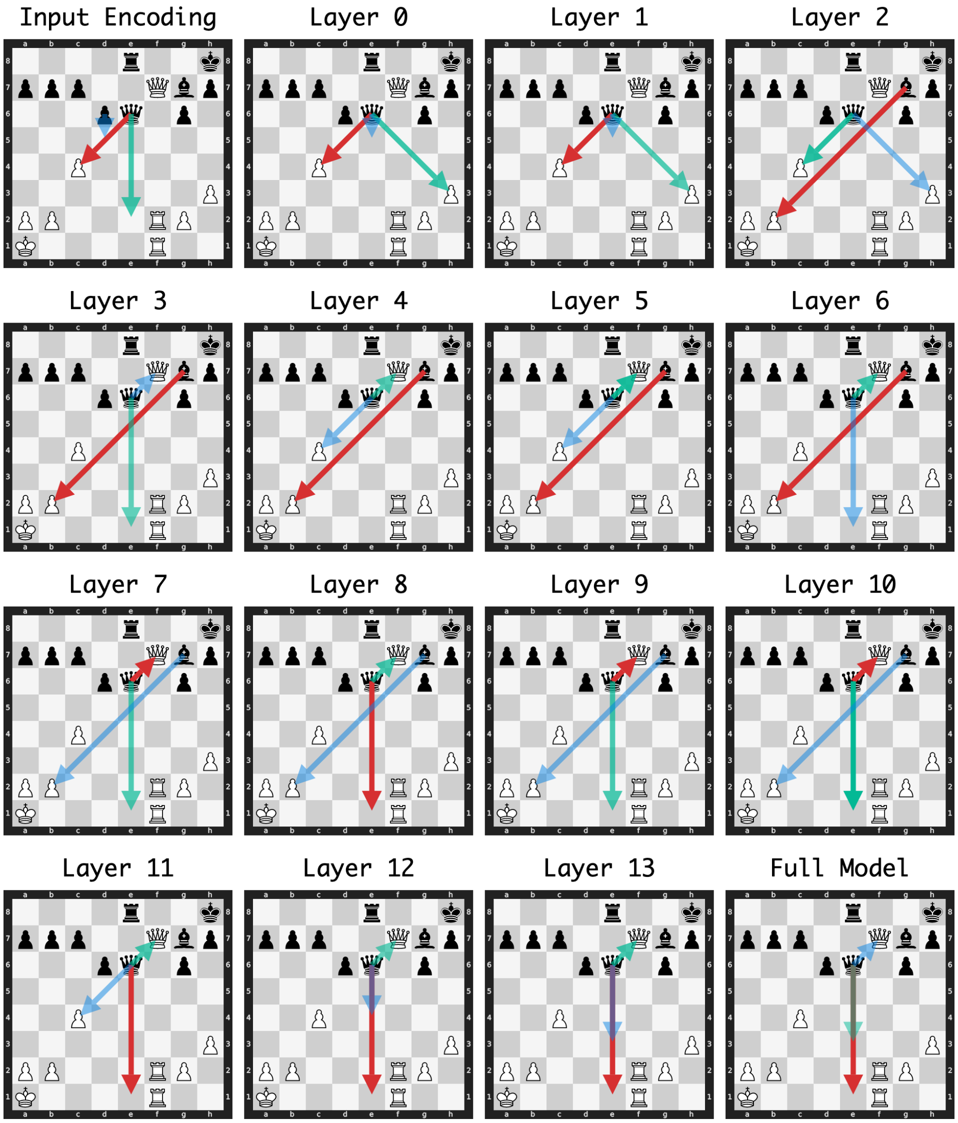
| Move | Input | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
|---|---|---|---|---|---|---|---|---|
| **♗h2+** | 0.06% | 0.01% | 0.65% | 0.23% | 0.45% | 0.21% | 3.92% | 31.39% |
| **♖×e3** | 25.43% | 24.14% | 74.33% | 83.14% | 79.68% | 92.11% | 74.68% | 46.36% |
| **a4** | 5.81% | 47.36% | 0.77% | 0.21% | 9.00% | 0.04% | 12.92% | 8.74% |
| **f5** | 22.75% | 0.42% | 0.11% | 0.02% | 0.05% | 0.01% | 0.01% | 0.17% |
| **♖e5** | 2.90% | 0.01% | 0.05% | 0.11% | 0.23% | 1.07% | 2.55% | 9.16% |
| **h5** | 0.15% | 15.19% | 3.82% | 0.17% | 0.03% | 0.00% | 0.01% | 0.01% |
| **b5** | 4.88% | 10.36% | 0.74% | 0.54% | 0.02% | 0.01% | 0.05% | 0.05% |
| **g5** | 6.84% | 0.35% | 0.19% | 0.04% | 0.01% | 0.01% | 0.02% | 0.03% |
| **♗c5** | 5.00% | 0.01% | 0.72% | 3.85% | 0.36% | 0.67% | 0.15% | 0.13% |
| **♖f6** | 1.68% | 0.02% | 4.36% | 0.39% | 0.27% | 0.23% | 0.14% | 0.39% |
| **♗e5** | 4.28% | 0.01% | 0.11% | 0.68% | 0.27% | 0.43% | 0.17% | 0.29% |
| **♖e4** | 3.88% | 0.01% | 3.38% | 1.21% | 2.38% | 0.17% | 0.14% | 0.15% |
| **♔h8** | 3.46% | 0.06% | 0.84% | 2.50% | 1.36% | 1.45% | 0.17% | 0.11% |
| **♖a8** | 2.22% | 0.05% | 0.15% | 0.61% | 0.14% | 0.17% | 0.13% | 0.09% |
| **♖g6** | 0.75% | 0.01% | 2.12% | 0.48% | 0.60% | 0.13% | 0.10% | 0.05% |
| **♖ee8** | 0.23% | 0.32% | 1.94% | 1.69% | 0.47% | 0.08% | 0.32% | 0.14% |
| **♖ce8** | 0.68% | 0.42% | 1.92% | 0.79% | 0.80% | 0.10% | 0.63% | 0.20% |
| **f6** | 0.69% | 0.10% | 0.03% | 0.07% | 0.00% | 0.01% | 0.01% | 0.14% |
| **♗a3** | 0.30% | 0.01% | 0.21% | 0.13% | 1.31% | 0.31% | 1.46% | 0.76% |
| **♗e7** | 1.44% | 0.01% | 0.08% | 0.15% | 0.19% | 0.13% | 0.16% | 0.05% |
| **♖h6** | 0.11% | 0.01% | 1.21% | 0.36% | 0.42% | 0.31% | 0.17% | 0.22% |
| **h6** | 0.06% | 0.85% | 0.16% | 0.09% | 0.01% | 0.02% | 0.02% | 0.05% |
| **♗g3** | 0.57% | 0.02% | 0.32% | 0.22% | 0.24% | 0.29% | 0.11% | 0.13% |
| **♖e7** | 1.04% | 0.01% | 0.62% | 0.15% | 0.16% | 0.16% | 0.12% | 0.07% |
| **♖b8** | 0.92% | 0.02% | 0.07% | 0.32% | 0.11% | 0.30% | 0.11% | 0.11% |
| **♗b4** | 0.25% | 0.01% | 0.18% | 0.39% | 0.46% | 0.25% | 0.91% | 0.33% |
| **♖f8** | 0.89% | 0.02% | 0.27% | 0.21% | 0.18% | 0.12% | 0.10% | 0.04% |
| **g6** | 0.88% | 0.15% | 0.02% | 0.03% | 0.00% | 0.00% | 0.00% | 0.00% |
| **♗f4** | 0.66% | 0.01% | 0.15% | 0.41% | 0.18% | 0.36% | 0.22% | 0.28% |
| **♖d8** | 0.62% | 0.02% | 0.18% | 0.31% | 0.13% | 0.18% | 0.12% | 0.07% |
| **♔f8** | 0.55% | 0.02% | 0.23% | 0.37% | 0.36% | 0.53% | 0.23% | 0.16% |
| **♗f8** | 0.01% | 0.01% | 0.08% | 0.15% | 0.15% | 0.13% | 0.16% | 0.11% |

Table 16: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| Move | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 | Layer 13 | Final |
|---|---|---|---|---|---|---|---|---|
| **♗h2+** | 36.07% | 36.91% | 41.18% | 39.83% | 35.44% | 67.90% | 68.57% | 92.87% |
| **♖×e3** | 40.66% | 52.04% | 44.63% | 45.31% | 36.86% | 6.02% | 3.89% | 0.17% |
| **a4** | 3.69% | 0.97% | 0.22% | 0.64% | 2.16% | 3.45% | 7.54% | 1.08% |
| **f5** | 0.71% | 1.52% | 0.28% | 0.20% | 0.83% | 0.37% | 1.00% | 0.71% |
| **♖e5** | 16.34% | 5.68% | 10.72% | 11.30% | 19.27% | 18.62% | 2.00% | 0.17% |
| **h5** | 0.02% | 0.02% | 0.01% | 0.01% | 0.03% | 0.02% | 0.43% | 0.26% |
| **b5** | 0.09% | 0.08% | 0.04% | 0.04% | 0.17% | 0.02% | 0.92% | 0.15% |
| **g5** | 0.04% | 0.03% | 0.03% | 0.02% | 0.09% | 0.04% | 0.49% | 0.16% |
| **♗c5** | 0.09% | 0.17% | 0.13% | 0.11% | 0.13% | 0.15% | 0.68% | 0.16% |
| **♖f6** | 0.18% | 0.10% | 0.16% | 0.10% | 0.92% | 0.22% | 0.28% | 0.16% |
| **♗e5** | 0.13% | 0.20% | 0.13% | 0.09% | 0.34% | 0.23% | 0.53% | 0.16% |
| **♖e4** | 0.06% | 0.06% | 0.07% | 0.06% | 0.17% | 0.10% | 0.64% | 0.16% |
| **♔h8** | 0.08% | 0.12% | 0.26% | 0.42% | 0.57% | 0.60% | 1.33% | 0.19% |
| **♖a8** | 0.07% | 0.05% | 0.05% | 0.05% | 0.05% | 0.05% | 0.30% | 0.30% |
| **♖g6** | 0.02% | 0.03% | 0.02% | 0.01% | 0.15% | 0.06% | 0.48% | 0.15% |
| **♖ee8** | 0.07% | 0.06% | 0.04% | 0.03% | 0.22% | 0.13% | 0.65% | 0.21% |
| **♖ce8** | 0.19% | 0.14% | 0.22% | 0.21% | 0.30% | 0.22% | 0.34% | 0.39% |
| **f6** | 0.09% | 0.05% | 0.02% | 0.03% | 0.06% | 0.03% | 1.66% | 0.18% |
| **♗a3** | 0.45% | 0.46% | 0.37% | 0.25% | 0.41% | 0.13% | 0.96% | 0.16% |
| **♗e7** | 0.04% | 0.08% | 0.08% | 0.09% | 0.07% | 0.16% | 0.44% | 0.16% |
| **♖h6** | 0.08% | 0.06% | 0.06% | 0.04% | 0.38% | 0.10% | 0.42% | 0.15% |
| **h6** | 0.02% | 0.00% | 0.01% | 0.01% | 0.06% | 0.02% | 1.19% | 0.14% |
| **♗g3** | 0.10% | 0.09% | 0.14% | 0.12% | 0.27% | 0.27% | 1.06% | 0.19% |
| **♖e7** | 0.01% | 0.04% | 0.03% | 0.02% | 0.09% | 0.06% | 0.42% | 0.17% |
| **♖b8** | 0.04% | 0.03% | 0.04% | 0.04% | 0.05% | 0.04% | 0.31% | 0.17% |
| **♗b4** | 0.15% | 0.27% | 0.28% | 0.12% | 0.11% | 0.20% | 0.61% | 0.15% |
| **♖f8** | 0.03% | 0.05% | 0.07% | 0.11% | 0.15% | 0.10% | 0.34% | 0.17% |
| **g6** | 0.00% | 0.00% | 0.00% | 0.00% | 0.01% | 0.00% | 0.61% | 0.14% |
| **♗f4** | 0.26% | 0.32% | 0.44% | 0.41% | 0.33% | 0.40% | 0.81% | 0.18% |
| **♖d8** | 0.03% | 0.04% | 0.06% | 0.06% | 0.07% | 0.07% | 0.29% | 0.27% |
| **♔f8** | 0.12% | 0.18% | 0.10% | 0.15% | 0.16% | 0.13% | 0.43% | 0.17% |
| **♗f8** | 0.08% | 0.18% | 0.12% | 0.13% | 0.06% | 0.07% | 0.37% | 0.14% |

### G.6 Case study 6: queen sacrifice to back rank mate

This puzzle (Figure 15) features a queen sacrifice leading to a back-rank mate (PV: 1. **♕e1+ ♖×e1** 2. **♖×e1#**). The winning move **♕e1+** exhibits non-monotonic behavior: receiving consideration in layers 1-3 (peak 17.9% at layer 3), disappearing at layer 4 (0.13%), then increasing to 79.6% in the final output. Competing moves in later layers include material captures **♕×f7** and **♗×b2+**, with the latter persisting under the top three moves until layer 10 despite leading to immediate recapture. Full probabilities are in Table 17.


Input Encoding

Layer 0

Layer 1

Layer 2

Layer 3

Layer 4

Layer 5

Layer 6

Layer 7

Layer 8

Layer 9

Layer 10

Layer 11

Layer 12

Layer 13

Full Model

Figure 15: Layer-wise policy evolution for Puzzle 945.

Table 17: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| Move | Input | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
|---|---|---|---|---|---|---|---|---|
| **♕e1+** | 0.59% | 0.08% | 3.28% | 4.64% | 17.91% | 0.13% | 1.31% | 24.94% |
| **♗×b2+** | 5.81% | 2.34% | 3.76% | 23.00% | 47.49% | 49.87% | 37.94% | 25.68% |
| **♕×f7** | 1.86% | 10.83% | 12.22% | 12.21% | 9.86% | 25.84% | 34.15% | 25.31% |
| **♕×c4** | 19.73% | 23.07% | 25.79% | 21.47% | 8.31% | 14.34% | 19.08% | 15.96% |
| **♕×h3** | 3.01% | 17.62% | 22.63% | 17.96% | 4.93% | 6.23% | 3.84% | 2.97% |
| **♕e5** | 0.69% | 11.59% | 20.98% | 2.26% | 0.31% | 0.12% | 1.11% | 1.78% |
| **♕e2** | 13.68% | 0.11% | 0.08% | 0.10% | 0.14% | 0.04% | 0.04% | 0.49% |
| **♕e4** | 4.04% | 0.08% | 0.14% | 0.51% | 0.24% | 0.09% | 0.33% | 0.89% |
| **d5** | 8.31% | 11.55% | 2.24% | 0.90% | 0.05% | 0.03% | 0.03% | 0.13% |
| **♕e3** | 5.69% | 0.06% | 0.38% | 0.43% | 0.20% | 0.09% | 0.19% | 0.24% |
| **c5** | 6.12% | 0.65% | 0.66% | 0.21% | 0.02% | 0.02% | 0.05% | 0.03% |
| **♕f5** | 5.80% | 0.08% | 0.11% | 0.53% | 0.40% | 0.20% | 0.05% | 0.04% |
| **g5** | 5.66% | 3.86% | 0.04% | 0.30% | 0.08% | 0.01% | 0.01% | 0.01% |
| **♖g8** | 0.61% | 0.05% | 0.21% | 5.26% | 0.88% | 0.06% | 0.05% | 0.07% |
| **b5** | 2.70% | 4.54% | 0.12% | 0.16% | 0.03% | 0.01% | 0.02% | 0.02% |
| **♕e7** | 0.93% | 0.03% | 0.21% | 0.29% | 0.26% | 0.10% | 0.06% | 0.11% |
| **♕d5** | 4.03% | 0.57% | 1.98% | 1.52% | 1.56% | 0.13% | 0.12% | 0.15% |
| **h5** | 0.11% | 3.69% | 0.81% | 1.05% | 0.29% | 0.05% | 0.07% | 0.09% |
| **a5** | 0.10% | 3.04% | 0.04% | 0.06% | 0.03% | 0.01% | 0.01% | 0.00% |
| **♗e5** | 0.11% | 2.53% | 1.90% | 0.74% | 0.30% | 0.10% | 0.47% | 0.33% |
| **h6** | 0.05% | 1.96% | 0.12% | 0.31% | 0.19% | 0.03% | 0.04% | 0.07% |
| **♖a8** | 1.67% | 0.09% | 0.23% | 0.34% | 0.67% | 0.25% | 0.07% | 0.03% |
| **♗c3** | 1.24% | 0.05% | 0.18% | 0.50% | 0.27% | 0.12% | 0.05% | 0.04% |
| **♕c8** | 0.09% | 0.05% | 0.07% | 0.24% | 0.22% | 0.11% | 0.04% | 0.03% |
| **♗h6** | 0.02% | 0.15% | 0.46% | 0.51% | 1.16% | 0.16% | 0.08% | 0.04% |
| **♕g4** | 0.75% | 0.20% | 0.37% | 0.87% | 0.96% | 0.16% | 0.11% | 0.06% |
| **♗f6** | 0.86% | 0.03% | 0.02% | 0.21% | 0.26% | 0.12% | 0.03% | 0.01% |
| **♖f8** | 0.68% | 0.06% | 0.10% | 0.84% | 0.51% | 0.54% | 0.10% | 0.14% |
| **♕f6** | 0.73% | 0.03% | 0.02% | 0.22% | 0.27% | 0.15% | 0.04% | 0.03% |
| **♖b8** | 0.70% | 0.06% | 0.15% | 0.24% | 0.38% | 0.13% | 0.04% | 0.02% |
| **♗d4** | 0.69% | 0.05% | 0.12% | 0.26% | 0.17% | 0.06% | 0.08% | 0.06% |
| **♕d7** | 0.62% | 0.03% | 0.08% | 0.33% | 0.41% | 0.10% | 0.02% | 0.02% |
| **c6** | 0.57% | 0.06% | 0.03% | 0.18% | 0.14% | 0.06% | 0.12% | 0.06% |
| **♖e7** | 0.48% | 0.04% | 0.18% | 0.29% | 0.25% | 0.12% | 0.03% | 0.04% |
| **♖d8** | 0.48% | 0.05% | 0.08% | 0.14% | 0.12% | 0.12% | 0.06% | 0.04% |
| **♗f8** | 0.00% | 0.05% | 0.10% | 0.40% | 0.48% | 0.17% | 0.07% | 0.04% |
| **a6** | 0.04% | 0.44% | 0.03% | 0.17% | 0.11% | 0.02% | 0.02% | 0.01% |
| **♖c8** | 0.39% | 0.05% | 0.07% | 0.12% | 0.08% | 0.06% | 0.03% | 0.02% |
| **b6** | 0.35% | 0.15% | 0.02% | 0.23% | 0.05% | 0.03% | 0.04% | 0.01% |

Table 18: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♕e1+** | 24.64% | 35.02% | 24.27% | 34.15% | 40.43% | 36.69% | 38.62% | 79.64% |
| **♗×b2+** | 21.70% | 15.85% | 11.51% | 9.60% | 3.70% | 1.89% | 2.67% | 0.20% |
| **♕×f7** | 29.86% | 24.28% | 41.84% | 34.92% | 21.48% | 17.89% | 23.77% | 4.48% |
| **♕×c4** | 14.16% | 12.55% | 11.00% | 9.50% | 8.00% | 5.18% | 0.56% | 0.10% |
| **♕×h3** | 2.74% | 3.40% | 2.12% | 1.45% | 0.88% | 1.46% | 2.05% | 0.15% |
| **♕e5** | 1.01% | 2.26% | 2.81% | 1.78% | 5.98% | 9.57% | 4.52% | 1.89% |
| **♕e2** | 1.44% | 1.22% | 0.23% | 1.14% | 0.42% | 0.14% | 0.36% | 0.12% |
| **♕e4** | 1.13% | 1.95% | 2.52% | 2.40% | 7.98% | 12.92% | 6.97% | 2.75% |
| **d5** | 0.12% | 0.07% | 0.10% | 0.07% | 0.39% | 0.02% | 0.30% | 0.27% |
| **♕e3** | 0.29% | 0.44% | 0.59% | 0.72% | 4.44% | 9.13% | 11.24% | 5.13% |
| **c5** | 0.02% | 0.03% | 0.02% | 0.04% | 0.02% | 0.01% | 0.12% | 0.13% |
| **♕f5** | 0.06% | 0.04% | 0.04% | 0.03% | 0.03% | 0.03% | 0.10% | 0.14% |
| **g5** | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.09% | 0.15% |
| **♖g8** | 0.07% | 0.06% | 0.04% | 0.06% | 0.03% | 0.02% | 0.08% | 0.12% |
| **b5** | 0.06% | 0.01% | 0.00% | 0.01% | 0.01% | 0.02% | 0.56% | 0.17% |
| **♕e7** | 0.28% | 0.63% | 0.85% | 1.73% | 4.44% | 3.46% | 3.03% | 0.33% |
| **♕d5** | 0.13% | 0.14% | 0.18% | 0.16% | 0.10% | 0.04% | 0.12% | 0.15% |
| **h5** | 0.97% | 0.80% | 0.49% | 0.41% | 0.46% | 0.22% | 0.50% | 0.30% |
| **a5** | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.18% | 0.13% |
| **♗e5** | 0.25% | 0.59% | 0.37% | 0.12% | 0.03% | 0.04% | 0.09% | 0.14% |
| **h6** | 0.16% | 0.10% | 0.11% | 0.15% | 0.05% | 0.04% | 0.37% | 0.13% |
| **♖a8** | 0.04% | 0.02% | 0.05% | 0.12% | 0.06% | 0.09% | 0.16% | 0.12% |
| **♗c3** | 0.03% | 0.02% | 0.02% | 0.02% | 0.03% | 0.03% | 0.11% | 0.13% |
| **♕c8** | 0.09% | 0.04% | 0.04% | 0.03% | 0.08% | 0.35% | 1.20% | 1.18% |
| **♗h6** | 0.02% | 0.02% | 0.01% | 0.01% | 0.00% | 0.01% | 0.08% | 0.11% |
| **♕g4** | 0.10% | 0.05% | 0.10% | 0.16% | 0.12% | 0.06% | 0.14% | 0.14% |
| **♗f6** | 0.01% | 0.02% | 0.01% | 0.01% | 0.00% | 0.01% | 0.06% | 0.14% |
| **♖f8** | 0.14% | 0.10% | 0.26% | 0.55% | 0.38% | 0.06% | 0.18% | 0.09% |
| **♕f6** | 0.04% | 0.04% | 0.05% | 0.04% | 0.03% | 0.03% | 0.06% | 0.12% |
| **♖b8** | 0.03% | 0.01% | 0.01% | 0.00% | 0.00% | 0.01% | 0.08% | 0.10% |
| **♗d4** | 0.04% | 0.04% | 0.03% | 0.02% | 0.03% | 0.03% | 0.33% | 0.31% |
| **♕d7** | 0.04% | 0.02% | 0.03% | 0.03% | 0.03% | 0.03% | 0.13% | 0.15% |
| **c6** | 0.07% | 0.03% | 0.03% | 0.05% | 0.02% | 0.01% | 0.14% | 0.11% |
| **♖e7** | 0.03% | 0.04% | 0.10% | 0.18% | 0.24% | 0.43% | 0.55% | 0.12% |
| **♖d8** | 0.06% | 0.02% | 0.02% | 0.01% | 0.01% | 0.01% | 0.09% | 0.11% |
| **♗f8** | 0.07% | 0.05% | 0.10% | 0.30% | 0.06% | 0.04% | 0.17% | 0.10% |
| **a6** | 0.03% | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.08% | 0.12% |
| **♖c8** | 0.02% | 0.02% | 0.03% | 0.02% | 0.01% | 0.02% | 0.03% | 0.09% |
| **b6** | 0.02% | 0.01% | 0.01% | 0.01% | 0.00% | 0.00% | 0.12% | 0.15% |

# H Forgotten puzzle solutions

## H.1 Forgotten solution 1: rook sacrifice for queen trade

This puzzle (Figure 16) presents a rook sacrifice forcing a queen trade (PV: 1. ♖×**g7+** ♔×**g7** 2. ♕×**d7+**). The winning move ♖×**g7+** dominates throughout nearly all layers, serving as the top candidate from layer 0 onwards (excluding layer 1) with probabilities exceeding 50% from layer 2 and peaking at 79.71% at layer 12. Its primary competitor, the materially conservative queen capture ♕×**a7**, maintains 30%-54% probability through layers 0-7 before declining steadily; however, this move leaves the rook hanging and gains insufficient compensation. The solution is abandoned in the final two layers in favor of ♔**f1**—a king move that receives negligible probability (< 1%) throughout layers 0-11, then surges to 7.63% (layer 12), 28.78% (layer 13), and 52.98% (final output). Notably, the model's value head evaluates the current position as unfavorable (71.8% loss probability), yet when performing a one-step lookahead by evaluating resulting positions after each legal move, correctly distinguishes the forcing sequence: it assigns near-certain victory (99.5% win) to the position after ♖×**g7+** while evaluating all alternatives as losing positions (91%-100% loss) (Table 19). Full probabilities are in Tables 20 and 21.

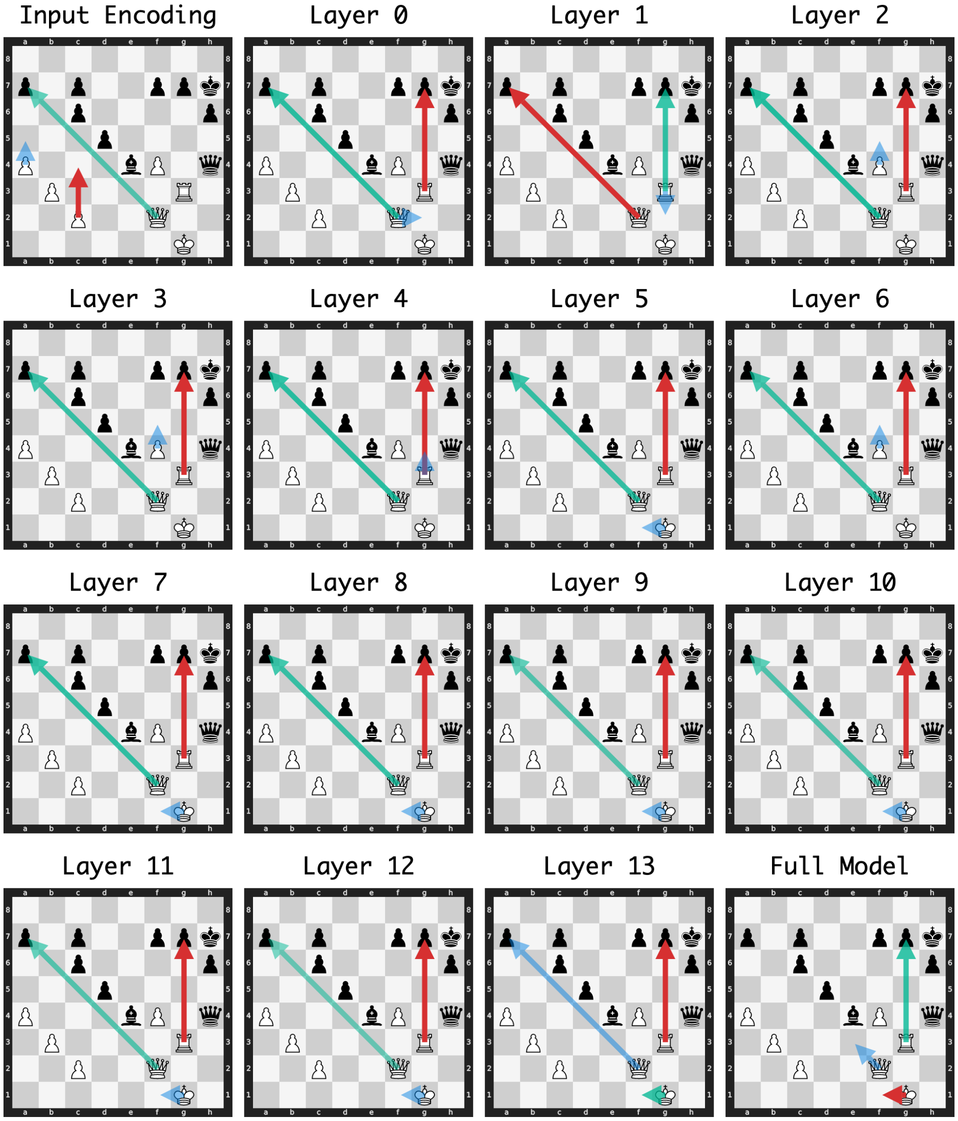


Figure 16: Layer-wise policy evolution for puzzle ID 58Ib0.

Table 19: Move evaluation for puzzle ID `58Ib0`: Stockfish evaluation at depth 20 and model WDL prediction for resulting positions

| **Move** | **Stockfish** | **Δ (cp)** | **Win** | **Draw** | **Loss** | **Δ Win** |
|---|---|---|---|---|---|---|
| *Current position* | +6.18 | — | 20.0% | 8.2% | 71.8% | — |
| **♖×g7+** ⋆ | +6.15 | −3 | 99.5% | 0.4% | 0.1% | +79.5% |
| **♔f1** | −2.47 | −865 | 0.9% | 7.9% | 91.1% | −19.1% |
| **♕e3** | −3.01 | −919 | 0.4% | 2.5% | 97.1% | −19.6% |
| **♕h2** | −3.29 | −947 | 0.5% | 2.9% | 96.6% | −19.5% |
| **♖g2** | −3.64 | −982 | 0.3% | 1.5% | 98.2% | −19.7% |
| **♕e1** | −4.53 | −1071 | 0.2% | 0.7% | 99.1% | −19.8% |
| **♖f3** | −7.32 | −1350 | 0.1% | 0.5% | 99.4% | −19.9% |
| **♕×a7** | −7.41 | −1359 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖h3** | −8.01 | −1419 | 0.0% | 0.1% | 99.9% | −20.0% |
| **♕g2** | −8.17 | −1435 | 0.0% | 0.1% | 99.9% | −20.0% |
| **♕c5** | −8.60 | −1478 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♕f3** | −8.60 | −1478 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♕d4** | −8.64 | −1482 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♕d2** | −8.87 | −1505 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♕e2** | −9.30 | −1548 | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖g6** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖g5** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖g4** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖e3** | −∞ | −∞ | 0.0% | 0.2% | 99.8% | −20.0% |
| **♖d3** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **♖c3** | −∞ | −∞ | 0.0% | 0.3% | 99.6% | −19.9% |
| **♕b6** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **♕f1** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −20.0% |
| **f5** | −∞ | −∞ | 0.0% | 0.3% | 99.6% | −19.9% |
| **a5** | −∞ | −∞ | 0.1% | 0.4% | 99.6% | −19.9% |
| **b4** | −∞ | −∞ | 0.0% | 0.3% | 99.6% | −19.9% |
| **c3** | −∞ | −∞ | 0.1% | 0.8% | 99.1% | −19.9% |
| **c4** | −∞ | −∞ | 0.0% | 0.2% | 99.8% | −20.0% |

Table 20: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♖×g7+** | 7.59% | 43.23% | 31.56% | 54.14% | 54.87% | 58.87% | 64.77% | 64.55% |
| **♕×a7** | 11.67% | 32.00% | 53.74% | 38.34% | 31.18% | 36.95% | 33.33% | 30.51% |
| **♔f1** | 1.18% | 0.05% | 0.42% | 0.36% | 0.14% | 0.23% | 0.57% | 0.95% |
| **c4** | 16.71% | 0.12% | 0.07% | 0.07% | 0.00% | 0.03% | 0.08% | 0.30% |
| **♕g2** | 1.09% | 15.74% | 2.47% | 0.16% | 0.06% | 0.07% | 0.03% | 0.03% |
| **a5** | 10.29% | 3.11% | 0.01% | 0.01% | 0.02% | 0.01% | 0.02% | 0.08% |
| **f5** | 2.50% | 0.87% | 3.04% | 3.50% | 10.06% | 0.08% | 0.20% | 1.90% |
| **b4** | 9.11% | 0.33% | 0.07% | 0.09% | 0.01% | 0.01% | 0.01% | 0.03% |
| **♕b6** | 7.73% | 0.09% | 0.15% | 0.72% | 1.39% | 0.53% | 0.11% | 0.09% |
| **♕c5** | 7.26% | 0.04% | 1.17% | 0.09% | 0.03% | 0.06% | 0.06% | 0.30% |
| **♕d4** | 5.44% | 0.03% | 0.19% | 0.08% | 0.03% | 0.06% | 0.03% | 0.16% |
| **♕e3** | 3.53% | 0.03% | 0.20% | 0.09% | 0.03% | 0.08% | 0.02% | 0.03% |
| **♖g2** | 0.25% | 3.78% | 4.31% | 0.10% | 0.06% | 0.08% | 0.03% | 0.05% |
| **♕f3** | 3.29% | 0.03% | 0.08% | 0.08% | 0.07% | 0.08% | 0.02% | 0.04% |
| **♕h2** | 1.06% | 0.04% | 0.17% | 0.05% | 0.01% | 0.03% | 0.03% | 0.03% |
| **c3** | 1.66% | 0.02% | 0.02% | 0.01% | 0.00% | 0.00% | 0.01% | 0.03% |
| **♖g6** | 2.05% | 0.07% | 0.78% | 0.78% | 0.77% | 0.09% | 0.21% | 0.22% |
| **♕e2** | 1.72% | 0.04% | 0.22% | 0.06% | 0.03% | 0.06% | 0.01% | 0.01% |
| **♕f1** | 0.02% | 0.05% | 0.33% | 0.06% | 0.02% | 0.06% | 0.06% | 0.03% |
| **♖g4** | 0.70% | 0.05% | 0.04% | 0.16% | 0.49% | 1.29% | 0.16% | 0.26% |
| **♕d2** | 1.21% | 0.03% | 0.15% | 0.06% | 0.03% | 0.09% | 0.01% | 0.01% |
| **♖f3** | 1.01% | 0.03% | 0.06% | 0.08% | 0.07% | 0.15% | 0.04% | 0.06% |
| **♖e3** | 0.83% | 0.03% | 0.22% | 0.09% | 0.04% | 0.09% | 0.03% | 0.03% |
| **♖c3** | 0.55% | 0.03% | 0.24% | 0.11% | 0.03% | 0.03% | 0.03% | 0.06% |
| **♖d3** | 0.67% | 0.06% | 0.05% | 0.14% | 0.09% | 0.23% | 0.03% | 0.03% |
| **♖g5** | 0.65% | 0.04% | 0.11% | 0.39% | 0.34% | 0.35% | 0.05% | 0.13% |
| **♕e1** | 0.14% | 0.03% | 0.10% | 0.06% | 0.03% | 0.06% | 0.02% | 0.02% |
| **♖h3** | 0.10% | 0.02% | 0.04% | 0.11% | 0.09% | 0.32% | 0.05% | 0.07% |

Table 21: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♖×g7+** | 61.81% | 64.85% | 78.35% | 78.00% | 76.68% | 79.71% | 51.59% | 31.61% |
| **♕×a7** | 34.41% | 31.04% | 18.61% | 19.62% | 16.66% | 8.60% | 2.48% | 0.29% |
| **♔f1** | 0.61% | 1.04% | 1.10% | 0.67% | 2.15% | 7.63% | 28.78% | 52.98% |
| **c4** | 0.26% | 0.13% | 0.03% | 0.03% | 0.08% | 0.06% | 0.88% | 0.24% |
| **♕g2** | 0.06% | 0.05% | 0.06% | 0.06% | 0.16% | 0.22% | 1.08% | 0.33% |
| **a5** | 0.20% | 0.17% | 0.23% | 0.13% | 0.48% | 0.08% | 0.67% | 0.26% |
| **f5** | 0.45% | 0.50% | 0.06% | 0.09% | 0.43% | 0.13% | 0.86% | 0.23% |
| **b4** | 0.14% | 0.16% | 0.05% | 0.04% | 0.16% | 0.05% | 0.92% | 0.24% |
| **♕b6** | 0.18% | 0.09% | 0.05% | 0.05% | 0.06% | 0.03% | 0.27% | 0.29% |
| **♕c5** | 0.41% | 0.23% | 0.15% | 0.11% | 0.29% | 0.12% | 0.66% | 0.30% |
| **♕d4** | 0.25% | 0.66% | 0.39% | 0.26% | 0.39% | 0.36% | 0.56% | 0.30% |
| **♕e3** | 0.11% | 0.07% | 0.07% | 0.08% | 0.20% | 0.32% | 1.56% | 4.46% |
| **♖g2** | 0.05% | 0.07% | 0.07% | 0.07% | 0.17% | 0.11% | 0.96% | 1.53% |
| **♕f3** | 0.03% | 0.03% | 0.04% | 0.04% | 0.13% | 0.16% | 0.24% | 0.26% |
| **♕h2** | 0.06% | 0.05% | 0.06% | 0.07% | 0.19% | 0.19% | 1.84% | 2.60% |
| **c3** | 0.05% | 0.03% | 0.03% | 0.05% | 0.26% | 0.35% | 2.11% | 0.19% |
| **♖g6** | 0.26% | 0.16% | 0.08% | 0.13% | 0.25% | 0.05% | 0.23% | 0.34% |
| **♕e2** | 0.02% | 0.02% | 0.02% | 0.02% | 0.04% | 0.08% | 0.31% | 0.29% |
| **♕f1** | 0.07% | 0.09% | 0.08% | 0.10% | 0.35% | 0.93% | 1.34% | 0.37% |
| **♖g4** | 0.16% | 0.10% | 0.09% | 0.04% | 0.06% | 0.05% | 0.21% | 0.28% |
| **♕d2** | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.04% | 0.22% | 0.29% |
| **♖f3** | 0.03% | 0.02% | 0.03% | 0.03% | 0.07% | 0.07% | 0.27% | 0.45% |
| **♖e3** | 0.04% | 0.07% | 0.08% | 0.06% | 0.11% | 0.04% | 0.18% | 0.37% |
| **♖c3** | 0.06% | 0.15% | 0.09% | 0.07% | 0.33% | 0.19% | 0.68% | 0.27% |
| **♖d3** | 0.02% | 0.03% | 0.02% | 0.02% | 0.05% | 0.08% | 0.22% | 0.26% |
| **♖g5** | 0.14% | 0.10% | 0.08% | 0.10% | 0.12% | 0.17% | 0.29% | 0.29% |
| **♕e1** | 0.03% | 0.03% | 0.02% | 0.02% | 0.04% | 0.09% | 0.37% | 0.37% |
| **♖h3** | 0.05% | 0.04% | 0.04% | 0.04% | 0.05% | 0.08% | 0.21% | 0.28% |

### H.2 Forgotten solution 2: queen sacrifice to back rank mate

This puzzle (Figure 17) presents a forced mate-in-two sequence requiring a queen sacrifice (PV: 1. **♕×c8+ ♖×c8** 2. **♖e8#**). The winning move **♕×c8+** dominates from layers 3-12, peaking at 65.4% (layer 5). However, this solution is abandoned in the final two layers, where **♕×a7**—a materially safe queen capture receiving only 5.6%-13.2% probability in layers 4-12—surges to 63.7% in the final output. The reversal may reflect a safety prior against queen sacrifices overriding mid-layer tactical calculations. The model's value head evaluates the current position as unfavorable (82.3% loss) despite the mate-in-two available, yet through one-step lookahead correctly assigns near-certain victory (100% win) to the position after **♕×c8+** while evaluating all alternatives as near-certain losses (99%-100% loss) (Table 22). This illustrates a failure mode where sound mid-layer analysis is overwritten by conservative final-layer adjustments despite accurate position evaluation capabilities. Full probabilities are in Tables 23 and 24.

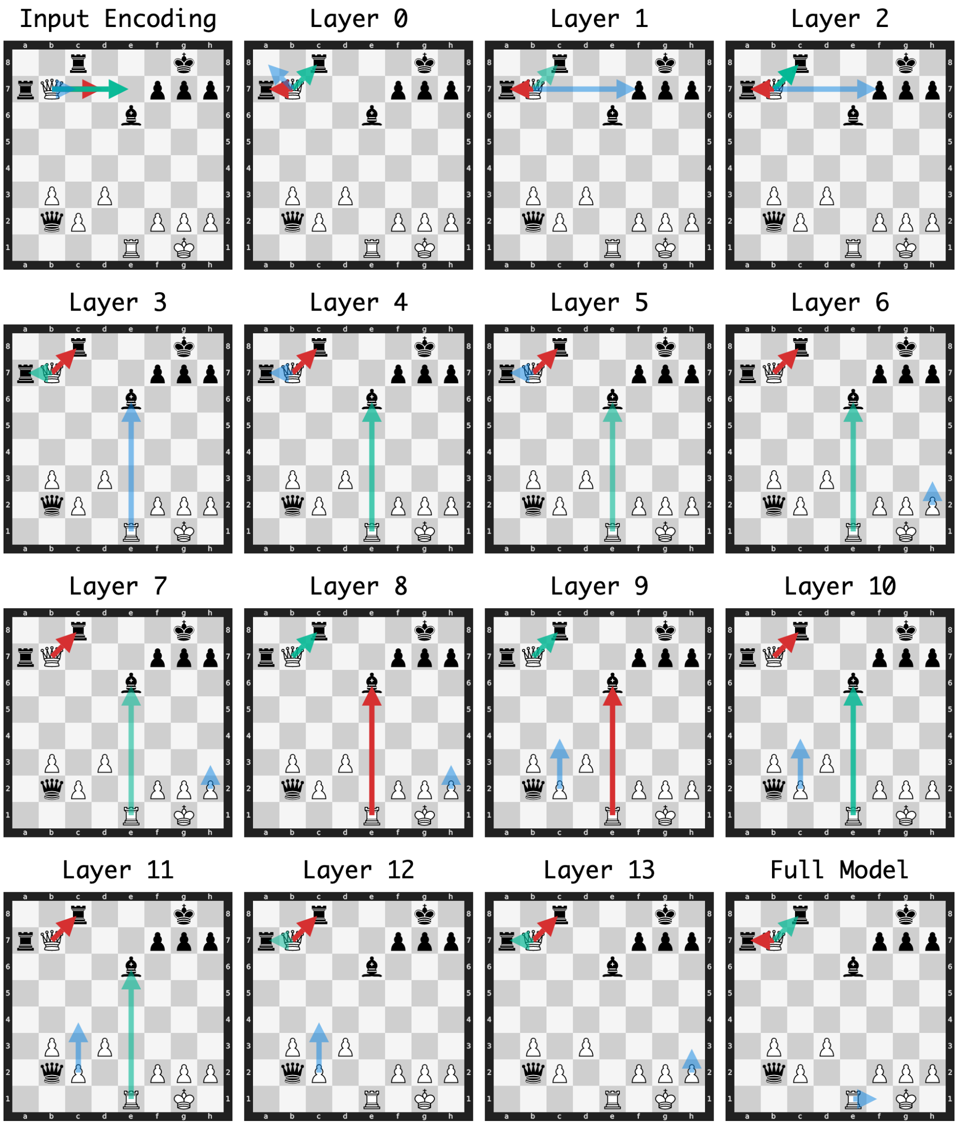


Figure 17: Layer-wise policy evolution for puzzle ID 1Egyn.

Table 22: Move evaluation for puzzle ID `1Egyn`: Stockfish evaluation at depth 20 and model WDL prediction for resulting positions

| **Move** | **Stockfish** | **Δ (cp)** | **Win** | **Draw** | **Loss** | **Δ Win** |
|---|---|---|---|---|---|---|
| *Current position* | $+\infty$ | — | 12.3% | 5.4% | 82.3% | — |
| **♕×c8+** ⋆ | $+\infty$ | — | 100.0% | 0.0% | 0.0% | +87.7% |
| **♕×a7** | −4.62 | $-\infty$ | 0.3% | 0.5% | 99.2% | −12.0% |
| **♕e4** | −7.11 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕b4** | −7.23 | $-\infty$ | 0.0% | 0.0% | 99.9% | −12.3% |
| **♕c6** | −7.39 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕b6** | −7.64 | $-\infty$ | 0.0% | 0.0% | 99.9% | −12.3% |
| **♕f3** | −7.67 | $-\infty$ | 0.0% | 0.0% | 99.9% | −12.3% |
| **♕b5** | −7.70 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **h4** | −8.96 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **h3** | −9.11 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕e7** | −9.17 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **d4** | −9.18 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **c4** | −9.22 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **b4** | −9.33 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **g3** | −9.33 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕c7** | −9.35 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖f1** | −9.44 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **g4** | −9.46 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♔h1** | −9.47 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **c3** | −9.52 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **f3** | −9.55 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖d1** | −9.66 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♔f1** | −9.68 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **f4** | −10.00 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕a8** | −10.26 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕b8** | −10.29 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕×f7+** | −11.09 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕a6** | −11.11 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕d7** | −12.92 | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♕d5** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖×e6** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖e5** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖e4** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖e3** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖e2** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖c1** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖b1** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |
| **♖a1** | $-\infty$ | $-\infty$ | 0.0% | 0.0% | 100.0% | −12.3% |

Table 23: Move probabilities by layer for puzzle ID `1Egyn` (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♕×a7** | 0.17% | 52.44% | 65.76% | 42.03% | 22.35% | 9.27% | 7.87% | 5.63% |
| **♕×c8+** | 4.64% | 36.25% | 14.07% | 40.27% | 51.99% | 51.44% | 65.38% | 45.93% |
| **♖×e6** | 4.69% | 0.29% | 0.43% | 1.09% | 7.39% | 26.31% | 16.77% | 22.38% |
| **c4** | 1.14% | 0.06% | 0.43% | 0.18% | 0.39% | 0.35% | 1.94% | 4.92% |
| **h3** | 0.01% | 0.05% | 0.19% | 0.38% | 1.28% | 0.18% | 0.67% | 11.68% |
| **♕d7** | 16.31% | 0.06% | 0.13% | 0.18% | 0.13% | 0.05% | 0.05% | 0.05% |
| **♕e7** | 15.69% | 0.08% | 0.58% | 0.20% | 0.20% | 0.10% | 0.05% | 0.07% |
| **♕c7** | 9.48% | 0.41% | 0.08% | 0.06% | 0.06% | 0.08% | 0.06% | 0.06% |
| **♕×f7+** | 8.99% | 3.99% | 5.86% | 7.18% | 5.30% | 5.94% | 3.80% | 1.96% |
| **♕c6** | 7.50% | 0.05% | 0.09% | 0.10% | 0.15% | 0.11% | 0.08% | 0.05% |
| **♕b8** | 5.81% | 0.25% | 0.36% | 0.24% | 1.11% | 0.28% | 0.14% | 1.11% |
| **♖e5** | 4.80% | 0.05% | 0.12% | 0.21% | 0.29% | 0.29% | 0.08% | 0.23% |
| **♕d5** | 4.66% | 0.04% | 0.81% | 0.12% | 0.13% | 0.06% | 0.04% | 0.02% |
| **♕a8** | 3.99% | 4.31% | 2.96% | 1.33% | 0.86% | 0.20% | 0.16% | 1.50% |
| **♕b6** | 3.08% | 0.04% | 0.14% | 0.13% | 0.12% | 0.11% | 0.12% | 0.08% |
| **f4** | 2.93% | 0.05% | 0.24% | 0.10% | 0.08% | 0.02% | 0.02% | 0.02% |
| **♖b1** | 0.12% | 0.05% | 0.25% | 0.90% | 1.87% | 0.84% | 0.55% | 1.00% |
| **♖c1** | 0.07% | 0.04% | 0.19% | 0.35% | 1.45% | 1.14% | 0.82% | 1.33% |
| **f3** | 0.18% | 0.06% | 0.38% | 0.15% | 0.12% | 0.04% | 0.04% | 0.06% |
| **♖a1** | 0.22% | 0.20% | 0.47% | 1.32% | 0.80% | 0.71% | 0.23% | 0.22% |
| **♔h1** | 0.46% | 0.23% | 1.04% | 0.20% | 0.39% | 0.23% | 0.08% | 0.18% |
| **h4** | 0.02% | 0.09% | 0.16% | 0.16% | 0.26% | 0.06% | 0.07% | 0.21% |
| **c3** | 0.11% | 0.10% | 0.86% | 0.27% | 0.17% | 0.03% | 0.01% | 0.01% |
| **g4** | 0.85% | 0.07% | 0.34% | 0.08% | 0.05% | 0.02% | 0.01% | 0.00% |
| **d4** | 0.76% | 0.06% | 0.78% | 0.67% | 0.36% | 0.05% | 0.05% | 0.05% |
| **♕b5** | 0.66% | 0.05% | 0.20% | 0.14% | 0.21% | 0.11% | 0.08% | 0.04% |
| **b4** | 0.62% | 0.08% | 0.35% | 0.29% | 0.19% | 0.04% | 0.04% | 0.07% |
| **♖d1** | 0.09% | 0.04% | 0.23% | 0.12% | 0.47% | 0.47% | 0.25% | 0.25% |
| **♕e4** | 0.52% | 0.04% | 0.32% | 0.25% | 0.13% | 0.07% | 0.05% | 0.05% |
| **♕b4** | 0.51% | 0.07% | 0.47% | 0.18% | 0.20% | 0.07% | 0.05% | 0.03% |
| **♖e4** | 0.16% | 0.05% | 0.41% | 0.13% | 0.11% | 0.21% | 0.03% | 0.04% |
| **♖e2** | 0.08% | 0.03% | 0.14% | 0.15% | 0.31% | 0.36% | 0.08% | 0.40% |
| **♕f3** | 0.32% | 0.05% | 0.39% | 0.21% | 0.23% | 0.07% | 0.05% | 0.05% |
| **♔f1** | 0.08% | 0.09% | 0.21% | 0.21% | 0.37% | 0.29% | 0.04% | 0.06% |
| **g3** | 0.12% | 0.04% | 0.23% | 0.09% | 0.12% | 0.05% | 0.11% | 0.09% |
| **♖f1** | 0.13% | 0.05% | 0.13% | 0.07% | 0.06% | 0.02% | 0.01% | 0.03% |
| **♕a6** | 0.03% | 0.05% | 0.08% | 0.08% | 0.15% | 0.18% | 0.13% | 0.12% |
| **♖e3** | 0.01% | 0.05% | 0.13% | 0.14% | 0.13% | 0.14% | 0.02% | 0.01% |

Table 24: Move probabilities by layer for puzzle ID `1Egyn` (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♕×a7** | 3.84% | 8.94% | 7.32% | 5.35% | 6.29% | 13.17% | 23.22% | 63.67% |
| **♕×c8+** | 42.17% | 25.30% | 27.56% | 34.34% | 49.62% | 58.19% | 53.17% | 28.66% |
| **♖×e6** | 19.71% | 29.14% | 32.92% | 30.31% | 23.06% | 10.02% | 3.74% | 0.16% |
| **c4** | 7.57% | 11.96% | 18.05% | 23.09% | 16.65% | 11.86% | 1.84% | 0.23% |
| **h3** | 16.71% | 13.75% | 5.01% | 2.96% | 1.91% | 4.74% | 8.27% | 0.23% |
| **♕d7** | 0.04% | 0.04% | 0.11% | 0.05% | 0.07% | 0.04% | 0.10% | 0.21% |
| **♕e7** | 0.04% | 0.04% | 0.08% | 0.03% | 0.04% | 0.04% | 0.16% | 0.22% |
| **♕c7** | 0.06% | 0.05% | 0.12% | 0.05% | 0.06% | 0.03% | 0.11% | 0.20% |
| **♕×f7+** | 2.49% | 6.63% | 3.37% | 1.58% | 0.53% | 0.34% | 1.84% | 0.19% |
| **♕c6** | 0.04% | 0.04% | 0.07% | 0.03% | 0.03% | 0.03% | 0.48% | 0.21% |
| **♕b8** | 1.63% | 0.57% | 0.56% | 0.23% | 0.30% | 0.15% | 0.14% | 0.19% |
| **♖e5** | 0.07% | 0.11% | 0.20% | 0.09% | 0.18% | 0.10% | 0.23% | 0.21% |
| **♕d5** | 0.01% | 0.02% | 0.04% | 0.02% | 0.01% | 0.02% | 0.12% | 0.21% |
| **♕a8** | 1.71% | 0.38% | 0.90% | 0.29% | 0.27% | 0.13% | 0.09% | 0.21% |
| **♕b6** | 0.08% | 0.07% | 0.11% | 0.04% | 0.03% | 0.02% | 0.13% | 0.23% |
| **f4** | 0.02% | 0.03% | 0.05% | 0.02% | 0.02% | 0.03% | 0.26% | 0.18% |
| **♖b1** | 1.08% | 0.65% | 0.73% | 0.12% | 0.18% | 0.16% | 0.17% | 0.20% |
| **♖c1** | 0.92% | 0.13% | 0.57% | 0.15% | 0.22% | 0.25% | 0.10% | 0.22% |
| **f3** | 0.22% | 0.15% | 0.09% | 0.01% | 0.01% | 0.02% | 1.33% | 0.20% |
| **♖a1** | 0.27% | 0.47% | 0.26% | 0.06% | 0.08% | 0.13% | 0.13% | 0.22% |
| **♔h1** | 0.12% | 0.15% | 0.43% | 0.83% | 0.12% | 0.10% | 0.29% | 0.18% |
| **h4** | 0.26% | 0.34% | 0.20% | 0.06% | 0.07% | 0.13% | 0.95% | 0.22% |
| **c3** | 0.01% | 0.03% | 0.06% | 0.04% | 0.03% | 0.01% | 0.12% | 0.24% |
| **g4** | 0.00% | 0.00% | 0.01% | 0.00% | 0.01% | 0.01% | 0.11% | 0.19% |
| **d4** | 0.05% | 0.13% | 0.07% | 0.03% | 0.02% | 0.03% | 0.36% | 0.21% |
| **♕b5** | 0.03% | 0.04% | 0.07% | 0.02% | 0.02% | 0.02% | 0.13% | 0.23% |
| **b4** | 0.08% | 0.18% | 0.08% | 0.02% | 0.01% | 0.01% | 0.50% | 0.22% |
| **♖d1** | 0.09% | 0.23% | 0.53% | 0.05% | 0.05% | 0.03% | 0.10% | 0.21% |
| **♕e4** | 0.04% | 0.06% | 0.03% | 0.02% | 0.03% | 0.02% | 0.18% | 0.24% |
| **♕b4** | 0.03% | 0.04% | 0.03% | 0.01% | 0.01% | 0.01% | 0.11% | 0.22% |
| **♖e4** | 0.02% | 0.01% | 0.02% | 0.01% | 0.01% | 0.02% | 0.37% | 0.21% |
| **♖e2** | 0.15% | 0.04% | 0.09% | 0.01% | 0.01% | 0.03% | 0.18% | 0.22% |
| **♕f3** | 0.08% | 0.14% | 0.06% | 0.02% | 0.02% | 0.01% | 0.16% | 0.23% |
| **♔f1** | 0.05% | 0.03% | 0.04% | 0.03% | 0.01% | 0.02% | 0.13% | 0.23% |
| **g3** | 0.06% | 0.05% | 0.02% | 0.01% | 0.01% | 0.02% | 0.31% | 0.19% |
| **♖f1** | 0.13% | 0.02% | 0.02% | 0.02% | 0.01% | 0.04% | 0.07% | 0.25% |
| **♕a6** | 0.11% | 0.05% | 0.12% | 0.02% | 0.01% | 0.01% | 0.25% | 0.23% |
| **♖e3** | 0.01% | 0.01% | 0.01% | 0.00% | 0.00% | 0.01% | 0.10% | 0.22% |

### H.3 Forgotten solution 3: rook sacrifice for material gain

This puzzle (Figure 18) presents a rook sacrifice forcing material advantage (PV: 1. **♖×g4+ ♔×g4** 2. **♗×d5**). The winning move **♖×g4+** dominates from layers 2-13, maintaining probabilities above 58% throughout and peaking at 85.88% at layer 12. However, the solution drops dramatically in the final layer to third place (12.14%), overtaken by **♖f1+** (40.43%) and **♖h1** (23.00%)—two materially conservative rook moves that receive minimal probability ($<$ 6%) through layers 0-12. Unlike previous cases, these alternative moves are strategically sound rather than clearly losing: Stockfish evaluates **♖f1+** at $-0.05$ pawns and **♖h1** at $-0.15$ pawns, both near-neutral positions. The model's value head reflects this nuance through one-step lookahead evaluation: while correctly assigning near-certain victory (97.9% win) to the position after **♖×g4+**, it evaluates the alternatives as roughly even positions (**♖f1+**: 73.6% loss, **♖h1**: 63.3% loss), rather than the near-certain losses (99%+) seen in previous puzzles (Table 25). This represents the most substantial final-layer reversal observed (85.88% to 12.14%), demonstrating that the safety prior against piece sacrifices can override even stronger mid-layer confidence than in previous cases. Full probabilities are in Tables 26 and 27.

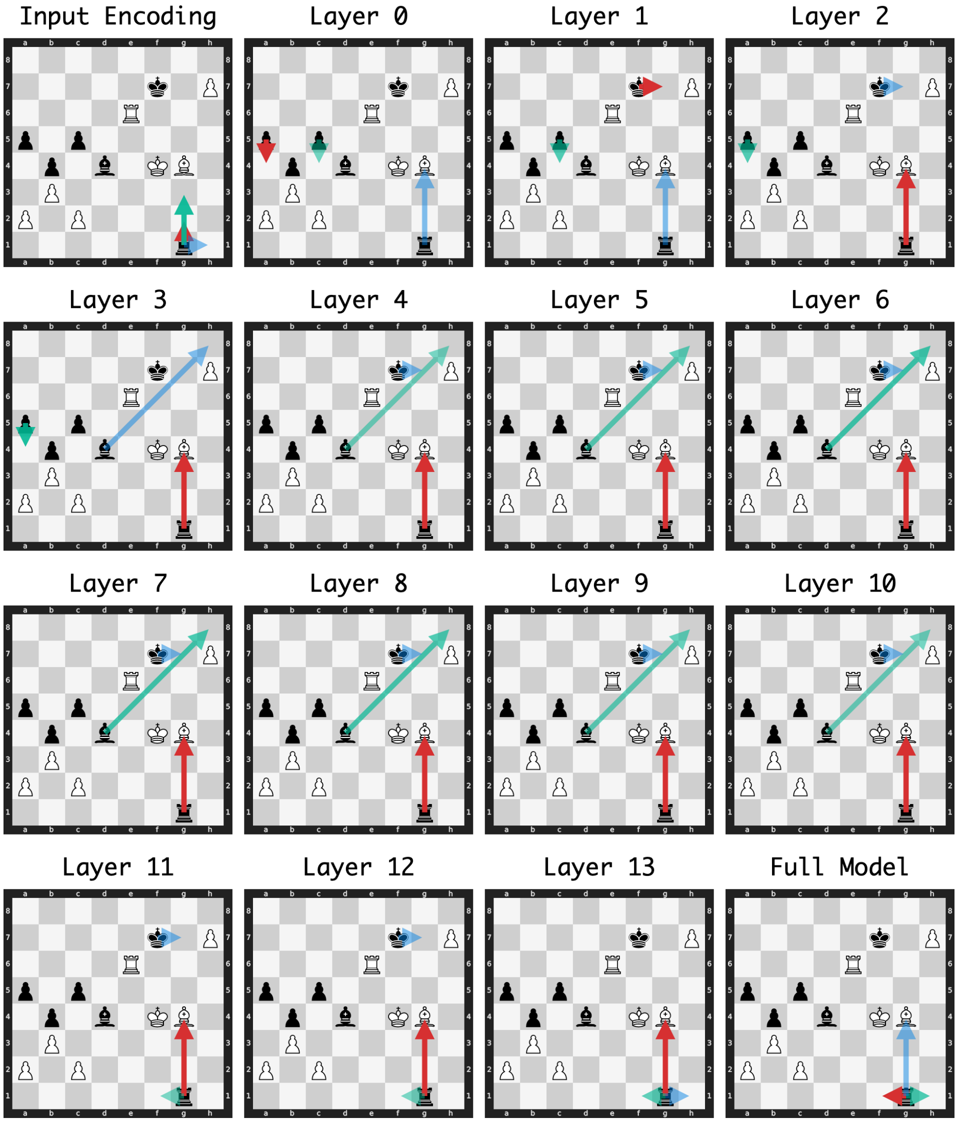


Figure 18: Layer-wise policy evolution for puzzle ID `DsaGi`.

Table 25: Move evaluation for puzzle ID `DsaGi`: Stockfish evaluation at depth 20 and model WDL prediction for resulting positions

| **Move** | **Stockfish** | **Δ (cp)** | **Win** | **Draw** | **Loss** | **Δ Win** |
|---|---:|---:|---:|---:|---:|---:|
| *Current position* | +5.76 | — | 4.5% | 23.1% | 72.4% | — |
| **♖×g4+** ⋆ | +6.10 | +34 | 97.9% | 1.9% | 0.2% | +93.4% |
| **♖f1+** | −0.05 | −581 | 0.8% | 25.6% | 73.6% | −3.7% |
| **♖h1** | −0.15 | −591 | 0.9% | 35.8% | 63.3% | −3.7% |
| **♔g7** | −2.46 | −822 | 0.6% | 8.9% | 90.5% | −4.0% |
| **♖c1** | −2.75 | −851 | 0.3% | 3.7% | 96.0% | −4.2% |
| **♖d1** | −2.77 | −853 | 0.1% | 0.9% | 99.0% | −4.4% |
| **♖a1** | −2.81 | −857 | 0.2% | 1.9% | 97.9% | −4.3% |
| **♖b1** | −2.86 | −862 | 0.2% | 1.3% | 98.5% | −4.3% |
| **♗c3** | −3.28 | −904 | 0.2% | 2.4% | 97.3% | −4.3% |
| **♗b2** | −3.48 | −924 | 0.3% | 4.1% | 95.6% | −4.2% |
| **♗g7** | −3.50 | −926 | 0.2% | 1.1% | 98.7% | −4.4% |
| **♗h8** | −3.56 | −932 | 0.1% | 0.9% | 98.9% | −4.4% |
| **♔f8** | −3.63 | −939 | 0.1% | 0.7% | 99.1% | −4.4% |
| **♖g2** | −3.64 | −940 | 0.3% | 5.0% | 94.7% | −4.2% |
| **♗a1** | −3.75 | −951 | 0.2% | 1.1% | 98.8% | −4.4% |
| **a4** | −3.75 | −951 | 0.5% | 13.1% | 86.4% | −4.1% |
| **c4** | −3.88 | −964 | 0.2% | 1.4% | 98.4% | −4.4% |
| **♗e5+** | −7.19 | −1295 | 0.0% | 0.0% | 99.9% | −4.5% |
| **♖g3** | −7.42 | −1318 | 0.0% | 0.1% | 99.9% | −4.5% |
| **♖e1** | −7.42 | −1318 | 0.0% | 0.1% | 99.9% | −4.5% |
| **♗f6** | −7.49 | −1325 | 0.0% | 0.1% | 99.8% | −4.5% |
| **♗e3+** | −7.88 | −1364 | 0.0% | 0.1% | 99.9% | −4.5% |
| **♗f2** | −9.17 | −1493 | 0.0% | 0.1% | 99.9% | −4.5% |

Table 26: Layer-wise policy probability evolution (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **a4** | 0.89% | 88.47% | 11.08% | 19.91% | 26.54% | 0.20% | 1.28% | 1.85% |
| **♖×g4+** | 2.52% | 1.09% | 11.98% | 30.36% | 28.68% | 85.97% | 71.28% | 61.70% |
| **♖f1+** | 5.69% | 0.14% | 1.58% | 8.96% | 5.76% | 0.20% | 0.24% | 1.07% |
| **♔g7** | 0.15% | 0.58% | 29.34% | 14.77% | 5.86% | 1.61% | 2.89% | 2.20% |
| **♗h8** | 0.15% | 0.03% | 0.61% | 0.82% | 18.77% | 4.10% | 16.51% | 28.35% |
| **♖h1** | 8.10% | 0.19% | 2.38% | 2.71% | 5.04% | 0.32% | 0.68% | 1.45% |
| **c4** | 1.00% | 7.93% | 18.03% | 3.85% | 0.99% | 0.03% | 0.44% | 0.46% |
| **♖g2** | 12.19% | 0.09% | 1.19% | 1.38% | 0.63% | 0.81% | 0.13% | 0.15% |
| **♖g3** | 11.46% | 0.04% | 2.14% | 3.13% | 2.10% | 1.57% | 0.13% | 0.08% |
| **♖b1** | 7.62% | 0.19% | 2.34% | 0.67% | 0.45% | 0.73% | 0.43% | 0.16% |
| **♗e3+** | 7.46% | 0.02% | 0.54% | 1.03% | 0.24% | 0.28% | 0.07% | 0.04% |
| **♗f2** | 7.44% | 0.05% | 0.61% | 1.76% | 0.28% | 0.55% | 0.17% | 0.15% |
| **♗c3** | 6.60% | 0.04% | 1.07% | 0.42% | 0.53% | 0.33% | 2.27% | 0.39% |
| **♖a1** | 6.35% | 0.08% | 5.28% | 2.84% | 0.58% | 0.22% | 0.09% | 0.06% |
| **♗g7** | 0.88% | 0.55% | 5.41% | 0.36% | 0.97% | 0.45% | 2.37% | 1.28% |
| **♖d1** | 5.39% | 0.16% | 0.82% | 0.65% | 0.23% | 0.45% | 0.11% | 0.05% |
| **♖e1** | 5.34% | 0.10% | 0.40% | 0.65% | 0.46% | 0.50% | 0.11% | 0.11% |
| **♖c1** | 3.03% | 0.08% | 1.92% | 1.05% | 0.46% | 0.33% | 0.11% | 0.07% |
| **♗f6** | 2.65% | 0.02% | 0.21% | 0.34% | 0.09% | 0.09% | 0.12% | 0.05% |
| **♗a1** | 2.49% | 0.05% | 0.83% | 0.79% | 0.65% | 0.38% | 0.25% | 0.13% |
| **♗b2** | 1.27% | 0.03% | 0.42% | 0.67% | 0.15% | 0.24% | 0.08% | 0.04% |
| **♗e5+** | 1.33% | 0.03% | 0.91% | 1.61% | 0.24% | 0.35% | 0.10% | 0.05% |
| **♔f8** | 0.00% | 0.03% | 0.91% | 1.29% | 0.29% | 0.28% | 0.15% | 0.13% |

Table 27: Layer-wise policy probability evolution (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **a4** | 1.31% | 0.59% | 0.06% | 0.20% | 0.56% | 0.38% | 0.83% | 0.72% |
| **♖×g4+** | 58.46% | 58.21% | 73.89% | 80.79% | 79.67% | 85.88% | 58.43% | 12.14% |
| **♖f1+** | 2.41% | 3.71% | 1.04% | 2.98% | 4.91% | 4.45% | 17.58% | 40.43% |
| **♔g7** | 5.83% | 6.04% | 4.91% | 3.80% | 4.79% | 4.22% | 5.51% | 8.01% |
| **♗h8** | 27.10% | 24.04% | 15.83% | 6.91% | 3.24% | 0.56% | 0.61% | 0.50% |
| **♖h1** | 1.36% | 1.72% | 0.53% | 1.30% | 2.02% | 1.68% | 5.93% | 23.00% |
| **c4** | 0.59% | 0.05% | 0.01% | 0.02% | 0.06% | 0.10% | 0.42% | 0.31% |
| **♖g2** | 0.17% | 0.20% | 0.13% | 0.31% | 0.26% | 0.14% | 0.67% | 5.38% |
| **♖g3** | 0.09% | 0.14% | 0.29% | 0.95% | 0.89% | 0.28% | 0.15% | 0.19% |
| **♖b1** | 0.25% | 0.30% | 0.08% | 0.56% | 0.72% | 0.34% | 3.57% | 0.36% |
| **♗e3+** | 0.03% | 0.05% | 0.03% | 0.02% | 0.02% | 0.00% | 0.12% | 0.27% |
| **♗f2** | 0.10% | 0.21% | 0.25% | 0.32% | 1.28% | 0.54% | 0.41% | 0.23% |
| **♗c3** | 0.12% | 0.26% | 0.09% | 0.04% | 0.02% | 0.02% | 0.13% | 1.32% |
| **♖a1** | 0.04% | 0.06% | 0.02% | 0.02% | 0.04% | 0.02% | 0.12% | 1.05% |
| **♗g7** | 1.61% | 3.74% | 2.40% | 1.15% | 0.76% | 0.62% | 1.79% | 0.50% |
| **♖d1** | 0.07% | 0.08% | 0.04% | 0.20% | 0.22% | 0.15% | 2.61% | 0.32% |
| **♖e1** | 0.07% | 0.08% | 0.09% | 0.08% | 0.16% | 0.28% | 0.11% | 0.20% |
| **♖c1** | 0.06% | 0.06% | 0.04% | 0.04% | 0.08% | 0.05% | 0.32% | 2.40% |
| **♗f6** | 0.04% | 0.04% | 0.02% | 0.01% | 0.01% | 0.01% | 0.14% | 0.27% |
| **♗a1** | 0.06% | 0.15% | 0.07% | 0.04% | 0.05% | 0.06% | 0.10% | 0.28% |
| **♗b2** | 0.03% | 0.03% | 0.02% | 0.02% | 0.01% | 0.02% | 0.15% | 1.62% |
| **♗e5+** | 0.04% | 0.08% | 0.05% | 0.03% | 0.04% | 0.01% | 0.10% | 0.24% |
| **♔f8** | 0.15% | 0.18% | 0.13% | 0.19% | 0.20% | 0.17% | 0.21% | 0.27% |

### H.4 Forgotten solution 4: queen sacrifice to back rank mate

This puzzle (Figure 19) presents a forced mate-in-two sequence requiring a queen sacrifice (PV: 1. ♕×**b1**+ ♖×**b1** 2. ♖×**b1**#). The winning move ♕×**b1**+ dominates through most layers, maintaining probabilities between 22.84% and 59.18% from layers 0-12 and peaking at 56.73% in layer 13. However, in the final layer, the materially conservative pawn advance **b5** marginally overtakes the solution (36.12% vs 35.99%), despite receiving only 9.49% probability in the preceding layer. This final-layer reversal demonstrates another instance of a safety prior against queen sacrifices overriding established tactical analysis, though the margin is narrower than in previous cases. Notably, the model's value head evaluates the current position as unfavorable (81.1% loss probability) and appears to base this evaluation on the inferior move **b5** rather than the forcing sequence, yet when performing a one-step lookahead by evaluating resulting positions after each legal move, correctly assigns near-certain victory (99.3% win) to the position after ♕×**b1**+ while evaluating all alternatives as losing positions (99.4%-100% loss) (Table 28). This pattern further illustrates the disconnect between mid-layer tactical understanding and final-layer output in positions involving material sacrifice. Full probabilities are in Tables 29 and 30.

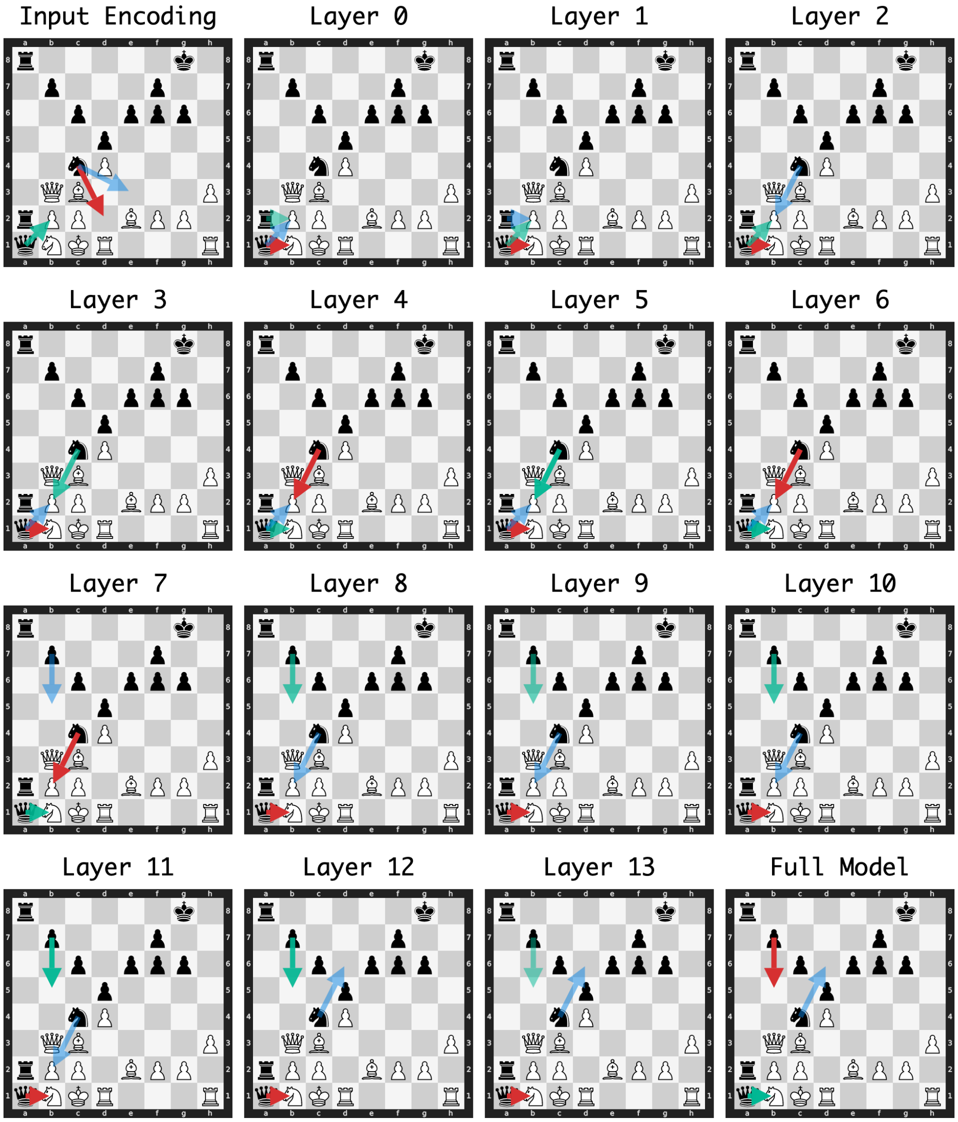


Figure 19: Layer-wise policy evolution for puzzle ID 00aDl.

Table 28: Move evaluation for puzzle ID `00aDl`: Stockfish evaluation at depth 20 and model WDL prediction for resulting positions

| **Move** | **Stockfish** | **Δ (cp)** | **Win** | **Draw** | **Loss** | **Δ Win** |
|---|---|---|---|---|---|---|
| *Current position* | +∞ | — | 8.3% | 10.6% | 81.1% | — |
| **♕×b1+** ⋆ | +∞ | — | 99.3% | 0.5% | 0.2% | +91.0% |
| **b5** | −4.12 | −∞ | 0.2% | 0.4% | 99.4% | −8.1% |
| **♖2a4** | −4.57 | −∞ | 0.2% | 0.3% | 99.5% | −8.1% |
| **♖2a7** | −4.72 | −∞ | 0.1% | 0.1% | 99.8% | −8.2% |
| **♘a5** | −4.74 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖2a6** | −4.77 | −∞ | 0.1% | 0.1% | 99.8% | −8.2% |
| **c5** | −4.78 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♘d6** | −5.07 | −∞ | 0.1% | 0.2% | 99.7% | −8.2% |
| **♔g7** | −5.40 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **b6** | −5.42 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **e5** | −5.45 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♔h8** | −5.49 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **f5** | −5.49 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **g5** | −5.53 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖2a5** | −5.55 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖8a4** | −5.56 | −∞ | 0.1% | 0.2% | 99.8% | −8.2% |
| **♖8a6** | −5.60 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖8a5** | −5.60 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖c8** | −5.63 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖d8** | −5.75 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖8a7** | −5.75 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♖b8** | −5.76 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♔f8** | −5.82 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♔h7** | −5.87 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♘e5** | −5.96 | −∞ | 0.0% | 0.0% | 99.9% | −8.3% |
| **♖f8** | −5.98 | −∞ | 0.0% | 0.0% | 99.9% | −8.3% |
| **♖e8** | −6.17 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♘e3** | −6.29 | −∞ | 0.0% | 0.0% | 99.9% | −8.3% |
| **♘b6** | −6.78 | −∞ | 0.0% | 0.0% | 99.9% | −8.3% |
| **♘×b2** | −7.09 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |
| **♕×b2+** | −7.11 | −∞ | 0.0% | 0.0% | 100.0% | −8.3% |
| **♖×b2** | −7.23 | −∞ | 0.0% | 0.1% | 99.8% | −8.3% |
| **♘a3** | −7.65 | −∞ | 0.0% | 0.0% | 99.9% | −8.3% |
| **♘d2** | −7.77 | −∞ | 0.0% | 0.0% | 100.0% | −8.3% |
| **♖2a3** | −8.16 | −∞ | 0.0% | 0.0% | 100.0% | −8.3% |
| **♖8a3** | −8.34 | −∞ | 0.0% | 0.1% | 99.9% | −8.3% |

Table 29: Move probabilities by layer for puzzle ID `00aDl` (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♕xb1+** | 9.73% | 59.18% | 45.52% | 31.92% | 27.26% | 26.44% | 28.89% | 26.24% |
| **b5** | 0.62% | 0.37% | 0.24% | 0.03% | 0.92% | 0.24% | 3.16% | 11.52% |
| **♘d2** | 28.22% | 0.29% | 0.66% | 0.20% | 3.58% | 0.04% | 2.79% | 2.42% |
| **♘xb2** | 3.17% | 1.86% | 1.48% | 22.14% | 24.65% | 27.68% | 28.02% | 27.10% |
| **♕xb2+** | 24.51% | 15.18% | 25.85% | 25.37% | 21.58% | 24.60% | 19.63% | 16.73% |
| **♖xb2** | 0.78% | 17.09% | 19.43% | 18.90% | 20.18% | 20.15% | 14.08% | 13.64% |
| **♘e3** | 18.31% | 0.08% | 0.22% | 0.01% | 0.03% | 0.02% | 0.01% | 0.01% |
| **♘d6** | 0.22% | 0.07% | 0.04% | 0.06% | 0.27% | 0.06% | 0.74% | 0.44% |
| **♘a5** | 0.47% | 0.23% | 0.33% | 0.03% | 0.33% | 0.09% | 2.02% | 0.35% |
| **f5** | 5.06% | 0.05% | 0.15% | 0.02% | 0.04% | 0.03% | 0.04% | 0.06% |
| **♘a3** | 0.19% | 0.16% | 0.43% | 0.03% | 0.20% | 0.02% | 0.05% | 0.25% |
| **b6** | 0.08% | 0.06% | 0.16% | 0.01% | 0.01% | 0.03% | 0.08% | 0.36% |
| **♖2a4** | 0.02% | 0.06% | 0.11% | 0.19% | 0.11% | 0.01% | 0.02% | 0.07% |
| **♔f8** | 0.18% | 2.27% | 0.29% | 0.05% | 0.06% | 0.05% | 0.01% | 0.01% |
| **♖8a4** | 0.08% | 0.10% | 0.53% | 0.04% | 0.02% | 0.01% | 0.02% | 0.04% |
| **c5** | 2.14% | 0.24% | 0.18% | 0.05% | 0.05% | 0.01% | 0.02% | 0.04% |
| **♖8a3** | 0.09% | 0.10% | 1.39% | 0.08% | 0.03% | 0.02% | 0.02% | 0.11% |
| **♖8a5** | 0.07% | 0.16% | 0.64% | 0.03% | 0.01% | 0.02% | 0.02% | 0.04% |
| **♘e5** | 1.33% | 0.03% | 0.15% | 0.01% | 0.01% | 0.02% | 0.02% | 0.02% |
| **e5** | 1.31% | 0.04% | 0.21% | 0.00% | 0.04% | 0.02% | 0.04% | 0.08% |
| **g5** | 1.27% | 0.42% | 0.44% | 0.01% | 0.03% | 0.01% | 0.04% | 0.04% |
| **♖2a6** | 0.10% | 0.09% | 0.11% | 0.03% | 0.03% | 0.03% | 0.01% | 0.01% |
| **♔g7** | 0.05% | 0.15% | 0.08% | 0.43% | 0.22% | 0.08% | 0.06% | 0.07% |
| **♖2a3** | 0.02% | 0.05% | 0.13% | 0.03% | 0.05% | 0.02% | 0.02% | 0.10% |
| **♖2a7** | 0.08% | 0.06% | 0.09% | 0.05% | 0.02% | 0.01% | 0.01% | 0.01% |
| **♖2a5** | 0.08% | 0.07% | 0.07% | 0.03% | 0.05% | 0.03% | 0.03% | 0.05% |
| **♖8a6** | 0.04% | 0.10% | 0.26% | 0.02% | 0.01% | 0.03% | 0.01% | 0.01% |
| **♖f8** | 0.27% | 0.61% | 0.13% | 0.03% | 0.00% | 0.01% | 0.00% | 0.01% |
| **♔h8** | 0.59% | 0.16% | 0.19% | 0.06% | 0.06% | 0.05% | 0.03% | 0.06% |
| **♔h7** | 0.05% | 0.38% | 0.03% | 0.06% | 0.10% | 0.07% | 0.02% | 0.02% |
| **♘b6** | 0.51% | 0.06% | 0.12% | 0.01% | 0.01% | 0.04% | 0.08% | 0.07% |
| **♖b8** | 0.07% | 0.04% | 0.06% | 0.01% | 0.01% | 0.01% | 0.00% | 0.01% |
| **♖8a7** | 0.06% | 0.07% | 0.12% | 0.02% | 0.01% | 0.01% | 0.01% | 0.01% |
| **♖d8** | 0.08% | 0.03% | 0.05% | 0.02% | 0.01% | 0.01% | 0.00% | 0.00% |
| **♖e8** | 0.10% | 0.04% | 0.05% | 0.02% | 0.01% | 0.01% | 0.00% | 0.00% |
| **♖c8** | 0.06% | 0.04% | 0.07% | 0.02% | 0.01% | 0.01% | 0.00% | 0.00% |

Table 30: Move probabilities by layer for puzzle ID `00aD1` (Part 2: Layer 7 to Final)

| **Move** | **Layer 7** | **Layer 8** | **Layer 9** | **Layer 10** | **Layer 11** | **Layer 12** | **Layer 13** | **Final** |
|---|---|---|---|---|---|---|---|---|
| **♕×b1+** | 25.22% | 26.13% | 22.84% | 30.07% | 37.85% | 41.36% | 56.73% | 35.99% |
| **b5** | 16.51% | 21.83% | 15.33% | 21.92% | 34.85% | 36.28% | 9.49% | 36.12% |
| **♘d2** | 1.73% | 0.64% | 10.39% | 4.54% | 0.42% | 0.22% | 0.21% | 0.32% |
| **♘×b2** | 27.04% | 18.68% | 13.53% | 13.65% | 6.67% | 0.67% | 0.78% | 0.35% |
| **♕×b2+** | 13.92% | 14.74% | 12.19% | 10.51% | 5.16% | 0.41% | 0.55% | 0.29% |
| **♖×b2** | 10.09% | 11.83% | 9.52% | 8.37% | 3.40% | 0.36% | 0.54% | 0.44% |
| **♘e3** | 0.01% | 0.01% | 0.03% | 0.01% | 0.02% | 0.02% | 0.18% | 0.32% |
| **♘d6** | 1.23% | 1.40% | 4.11% | 2.67% | 4.99% | 10.85% | 8.85% | 5.41% |
| **♘a5** | 1.17% | 1.14% | 4.14% | 2.16% | 1.24% | 2.00% | 5.97% | 1.85% |
| **f5** | 0.12% | 0.11% | 0.04% | 0.15% | 0.22% | 0.10% | 0.40% | 0.54% |
| **♘a3** | 0.87% | 0.71% | 3.20% | 1.05% | 0.58% | 0.76% | 1.88% | 0.94% |
| **b6** | 0.74% | 0.62% | 0.93% | 0.94% | 1.17% | 0.54% | 2.76% | 0.62% |
| **♖2a4** | 0.20% | 0.37% | 0.53% | 0.53% | 0.77% | 1.69% | 2.43% | 2.28% |
| **♔f8** | 0.01% | 0.01% | 0.02% | 0.05% | 0.06% | 0.09% | 0.13% | 0.40% |
| **♖8a4** | 0.10% | 0.29% | 0.36% | 0.47% | 0.50% | 1.08% | 1.17% | 2.18% |
| **c5** | 0.04% | 0.05% | 0.06% | 0.11% | 0.08% | 0.18% | 0.64% | 0.70% |
| **♖8a3** | 0.23% | 0.47% | 1.01% | 0.63% | 0.26% | 0.35% | 0.64% | 0.64% |
| **♖8a5** | 0.04% | 0.07% | 0.07% | 0.08% | 0.15% | 0.13% | 0.42% | 1.35% |
| **♘e5** | 0.02% | 0.05% | 0.04% | 0.06% | 0.11% | 0.19% | 0.19% | 0.34% |
| **e5** | 0.13% | 0.12% | 0.17% | 0.07% | 0.05% | 0.06% | 0.39% | 0.54% |
| **g5** | 0.08% | 0.08% | 0.06% | 0.17% | 0.20% | 0.25% | 0.32% | 0.41% |
| **♖2a6** | 0.01% | 0.01% | 0.02% | 0.03% | 0.05% | 0.14% | 0.74% | 0.93% |
| **♔g7** | 0.07% | 0.05% | 0.06% | 0.09% | 0.03% | 0.05% | 0.20% | 0.77% |
| **♖2a3** | 0.18% | 0.33% | 0.75% | 0.30% | 0.20% | 0.47% | 0.60% | 0.65% |
| **♖2a7** | 0.02% | 0.01% | 0.04% | 0.07% | 0.10% | 0.21% | 0.52% | 0.71% |
| **♖2a5** | 0.05% | 0.06% | 0.15% | 0.10% | 0.25% | 0.36% | 0.67% | 0.52% |
| **♖8a6** | 0.01% | 0.01% | 0.02% | 0.04% | 0.04% | 0.11% | 0.31% | 0.61% |
| **♖f8** | 0.01% | 0.01% | 0.01% | 0.04% | 0.01% | 0.05% | 0.16% | 0.40% |
| **♔h8** | 0.04% | 0.05% | 0.13% | 0.43% | 0.13% | 0.11% | 0.28% | 0.41% |
| **♔h7** | 0.02% | 0.02% | 0.06% | 0.17% | 0.11% | 0.07% | 0.20% | 0.51% |
| **♘b6** | 0.07% | 0.07% | 0.11% | 0.13% | 0.18% | 0.20% | 0.37% | 0.32% |
| **♖b8** | 0.01% | 0.01% | 0.01% | 0.03% | 0.01% | 0.07% | 0.16% | 0.47% |
| **♖8a7** | 0.01% | 0.01% | 0.03% | 0.07% | 0.07% | 0.23% | 0.28% | 0.47% |
| **♖d8** | 0.01% | 0.00% | 0.01% | 0.06% | 0.01% | 0.07% | 0.29% | 0.42% |
| **♖e8** | 0.01% | 0.00% | 0.01% | 0.11% | 0.02% | 0.11% | 0.24% | 0.39% |
| **♖c8** | 0.01% | 0.01% | 0.01% | 0.13% | 0.02% | 0.14% | 0.30% | 0.39% |

### H.5 Forgotten solution 5: pawn sacrifice for queen capture

This puzzle (Figure 20) presents a pawn sacrifice forcing a queen capture (PV: 1. **g5+** ♔×**g5** 2. ♕**g7#**). The winning move **g5+** exhibits highly non-monotonic behavior with substantial probability jumps between consecutive layers: 0.52% to 31.71% (layers 4-5), 1.60% to 18.16% (layers 8-9), and 1.45% to 40.26% (layers 12-13). The solution briefly emerges as the top candidate at layer 7 (38.73%) and again at layer 13 (40.26%), but drops to fourth place in the final output (11.17%). Through most middle layers (2-11), the model strongly favors the materially conservative queen trade ♕×**e5**, peaking at 90.95% (layer 4). The final layer instead prioritizes queen retreats to safety: ♕**d3** (29.14%), ♕**c4** (25.06%), and ♕**d1** (14.27%)—all receiving minimal probability through layers 0-11. The sharp inter-layer transitions suggest algorithmic computation of forcing sequences rather than gradual accumulation of tactical heuristics. The model's value head evaluates the current position as unfavorable (75.9% loss probability), yet through one-step lookahead correctly assigns near-certain victory (98.3% win) to the position after **g5+** while evaluating all alternatives as losing positions (89.1%-100% loss) (Table 31). Full probabilities are in Tables 32 and 33.

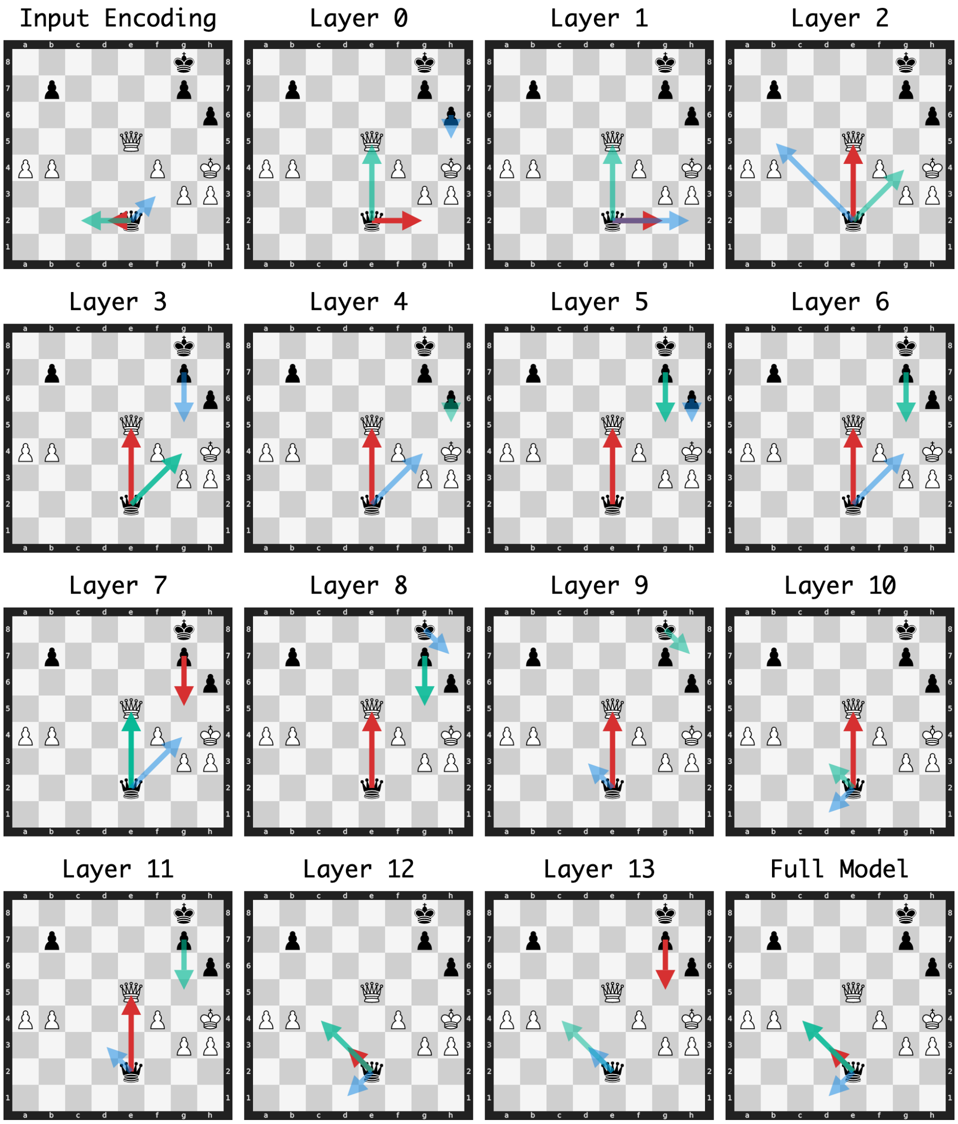


Figure 20: Layer-wise policy evolution for puzzle ID `6PIrs`.

Table 31: Move evaluation for puzzle ID `6PIrs`: Stockfish evaluation at depth 20 and model WDL prediction for resulting positions

| Move | Stockfish | Δ (cp) | Win | Draw | Loss | Δ Win |
|---|---|---|---|---|---|---|
| *Current position* | +7.17 | — | 1.8% | 22.3% | 75.9% | — |
| **g5+** ⋆ | +8.54 | +137 | 98.3% | 1.5% | 0.2% | +96.5% |
| **♕d3** | −2.70 | −987 | 0.8% | 9.7% | 89.5% | −1.0% |
| **♕c4** | −3.09 | −1026 | 0.8% | 10.1% | 89.1% | −1.0% |
| **♕a2** | −3.09 | −1026 | 0.5% | 3.6% | 95.9% | −1.3% |
| **♕f3** | −3.19 | −1036 | 0.5% | 4.9% | 94.6% | −1.3% |
| **♕d2** | −3.24 | −1041 | 0.8% | 8.3% | 90.9% | −1.0% |
| **♕a6** | −3.26 | −1043 | 0.5% | 3.8% | 95.7% | −1.3% |
| **♕c2** | −3.45 | −1062 | 0.6% | 5.1% | 94.3% | −1.2% |
| **♕d1** | −3.70 | −1087 | 0.7% | 8.0% | 91.4% | −1.1% |
| **♕h2** | −3.72 | −1089 | 0.2% | 0.9% | 98.8% | −1.6% |
| **♕f1** | −3.84 | −1101 | 0.3% | 2.0% | 97.6% | −1.5% |
| **♕f2** | −4.05 | −1122 | 0.3% | 1.5% | 98.3% | −1.5% |
| **♕g2** | −4.09 | −1126 | 0.4% | 2.8% | 96.8% | −1.4% |
| **♕×e5** | −6.70 | −1387 | 0.1% | 0.1% | 99.8% | −1.7% |
| **♔f8** | −12.22 | −1939 | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕b2** | −12.29 | −1946 | 0.0% | 0.0% | 100.0% | −1.8% |
| **g6** | −12.97 | −2014 | 0.0% | 0.0% | 100.0% | −1.8% |
| **♔h8** | −∞ | −∞ | 0.0% | 0.1% | 99.9% | −1.8% |
| **♔h7** | −∞ | −∞ | 0.0% | 0.2% | 99.7% | −1.8% |
| **♔f7** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕h5+** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕b5** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕g4+** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕e4** | −∞ | −∞ | 0.0% | 0.1% | 99.9% | −1.8% |
| **♕e3** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **♕e1** | −∞ | −∞ | 0.0% | 0.1% | 99.9% | −1.8% |
| **b6** | −∞ | −∞ | 0.0% | 0.1% | 99.9% | −1.8% |
| **h5** | −∞ | −∞ | 0.0% | 0.0% | 100.0% | −1.8% |
| **b5** | −∞ | −∞ | 0.0% | 0.1% | 99.9% | −1.8% |

Table 32: Move probabilities by layer for puzzle ID `6PIrs` (Part 1: Input to Layer 6)

| **Move** | **Input** | **Layer 0** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** |
|---|---|---|---|---|---|---|---|---|
| **♕×e5** | 0.45% | 22.15% | 17.94% | 40.14% | 35.67% | 90.95% | 45.36% | 34.28% |
| **♕g2** | 6.14% | 51.31% | 58.98% | 2.42% | 1.60% | 0.10% | 0.20% | 0.10% |
| **g5+** | 0.43% | 0.49% | 0.10% | 6.47% | 10.02% | 0.52% | 31.71% | 25.56% |
| **♕d3** | 7.43% | 0.06% | 0.11% | 0.20% | 0.16% | 0.12% | 0.35% | 0.49% |
| **♕g4+** | 1.18% | 0.15% | 3.21% | 11.73% | 27.60% | 1.41% | 4.40% | 15.12% |
| **♕c4** | 4.00% | 0.05% | 0.11% | 0.37% | 0.17% | 0.10% | 0.40% | 0.64% |
| **♕d2** | 15.94% | 0.10% | 0.36% | 0.32% | 0.11% | 0.07% | 0.08% | 0.11% |
| **♕d1** | 5.44% | 0.16% | 0.35% | 0.50% | 0.35% | 0.13% | 0.73% | 1.41% |
| **♕f3** | 8.72% | 0.33% | 4.48% | 2.81% | 0.83% | 0.15% | 1.18% | 1.41% |
| **♕c2** | 10.60% | 0.05% | 0.11% | 0.22% | 0.18% | 0.12% | 0.26% | 0.22% |
| **♕h5+** | 0.41% | 0.29% | 1.36% | 3.19% | 3.65% | 1.21% | 4.90% | 10.38% |
| **h5** | 0.01% | 10.01% | 0.56% | 6.59% | 9.11% | 2.07% | 6.07% | 2.14% |
| **♔h7** | 0.02% | 0.05% | 0.11% | 2.38% | 1.65% | 0.63% | 1.66% | 3.81% |
| **♕b5** | 1.48% | 0.11% | 0.37% | 7.88% | 4.30% | 0.24% | 0.36% | 0.32% |
| **♕f2** | 7.56% | 0.13% | 0.45% | 0.87% | 0.14% | 0.07% | 0.17% | 0.12% |
| **♕h2** | 3.91% | 7.51% | 6.78% | 1.66% | 0.50% | 0.09% | 0.09% | 0.05% |
| **♕e3** | 7.38% | 0.15% | 0.12% | 0.27% | 0.10% | 0.08% | 0.06% | 0.04% |
| **b5** | 0.23% | 2.29% | 0.22% | 5.87% | 1.98% | 0.09% | 0.37% | 1.22% |
| **♕f1** | 5.75% | 2.99% | 2.89% | 1.09% | 0.21% | 0.09% | 0.11% | 0.07% |
| **♕e1** | 5.71% | 0.32% | 0.32% | 0.30% | 0.16% | 0.09% | 0.12% | 0.08% |
| **♕e4** | 3.97% | 0.06% | 0.06% | 0.31% | 0.12% | 0.09% | 0.12% | 0.54% |
| **♕b2** | 2.26% | 0.10% | 0.14% | 0.21% | 0.08% | 0.07% | 0.06% | 0.04% |
| **♔f7** | 0.02% | 0.04% | 0.32% | 1.96% | 0.60% | 0.60% | 0.28% | 0.27% |
| **g6** | 0.06% | 0.26% | 0.03% | 0.38% | 0.07% | 0.10% | 0.44% | 1.13% |
| **♔f8** | 0.03% | 0.04% | 0.16% | 0.81% | 0.28% | 0.36% | 0.13% | 0.11% |
| **♕a6** | 0.10% | 0.12% | 0.11% | 0.43% | 0.10% | 0.10% | 0.17% | 0.08% |
| **b6** | 0.03% | 0.57% | 0.01% | 0.04% | 0.00% | 0.00% | 0.00% | 0.00% |
| **♔h8** | 0.22% | 0.06% | 0.07% | 0.32% | 0.13% | 0.27% | 0.12% | 0.16% |
| **♕a2** | 0.52% | 0.05% | 0.16% | 0.25% | 0.11% | 0.07% | 0.11% | 0.09% |

Table 33: Move probabilities by layer for puzzle ID `6PIrs` (Part 2: Layer 7 to Final)

| Move | Layer 7 | Layer 8 | Layer 9 | Layer 10 | Layer 11 | Layer 12 | Layer 13 | Final |
|---|---|---|---|---|---|---|---|---|
| **♕×e5** | 36.37% | 42.62% | 64.73% | 43.80% | 41.22% | 8.38% | 2.71% | 0.49% |
| **♕g2** | 0.08% | 0.04% | 0.05% | 0.06% | 0.04% | 0.06% | 0.19% | 0.36% |
| **g5+** | 38.73% | 33.36% | 1.60% | 0.93% | 18.16% | 1.45% | 40.26% | 11.17% |
| **♕d3** | 0.97% | 1.30% | 4.60% | 12.16% | 9.91% | 30.67% | 13.84% | 29.14% |
| **♕g4+** | 5.74% | 2.57% | 1.51% | 2.01% | 1.72% | 0.56% | 0.17% | 0.17% |
| **♕c4** | 1.15% | 2.18% | 2.49% | 6.23% | 5.38% | 21.09% | 14.71% | 25.06% |
| **♕d2** | 0.13% | 0.16% | 0.49% | 2.15% | 1.54% | 4.11% | 5.55% | 9.92% |
| **♕d1** | 1.35% | 1.37% | 2.93% | 7.26% | 5.74% | 12.16% | 7.89% | 14.27% |
| **♕f3** | 1.45% | 1.33% | 1.42% | 3.33% | 2.74% | 10.75% | 3.56% | 2.98% |
| **♕c2** | 0.21% | 0.13% | 0.18% | 0.40% | 0.34% | 0.97% | 1.19% | 1.99% |
| **♕h5+** | 3.40% | 2.57% | 3.61% | 7.02% | 3.03% | 0.56% | 0.14% | 0.17% |
| **h5** | 0.48% | 0.36% | 0.41% | 0.45% | 0.22% | 0.23% | 0.23% | 0.18% |
| **♔h7** | 4.90% | 5.66% | 7.91% | 5.91% | 3.09% | 5.70% | 3.03% | 0.18% |
| **♕b5** | 0.25% | 0.69% | 0.75% | 0.92% | 0.38% | 0.38% | 0.22% | 0.17% |
| **♕f2** | 0.07% | 0.03% | 0.04% | 0.05% | 0.04% | 0.05% | 0.18% | 0.21% |
| **♕h2** | 0.04% | 0.04% | 0.06% | 0.06% | 0.04% | 0.05% | 0.18% | 0.20% |
| **♕e3** | 0.03% | 0.03% | 0.04% | 0.06% | 0.08% | 0.08% | 0.27% | 0.18% |
| **b5** | 1.74% | 2.94% | 4.55% | 3.68% | 3.75% | 0.71% | 1.45% | 0.19% |
| **♕f1** | 0.07% | 0.04% | 0.05% | 0.07% | 0.04% | 0.05% | 0.17% | 0.25% |
| **♕e1** | 0.07% | 0.04% | 0.06% | 0.11% | 0.08% | 0.08% | 0.13% | 0.18% |
| **♕e4** | 0.53% | 0.49% | 0.66% | 1.36% | 1.21% | 0.14% | 0.36% | 0.18% |
| **♕b2** | 0.04% | 0.03% | 0.05% | 0.07% | 0.06% | 0.06% | 0.26% | 0.18% |
| **♔f7** | 0.33% | 0.39% | 0.42% | 0.47% | 0.48% | 0.60% | 0.27% | 0.18% |
| **g6** | 1.43% | 1.01% | 0.59% | 0.39% | 0.09% | 0.16% | 1.03% | 0.20% |
| **♔f8** | 0.10% | 0.12% | 0.17% | 0.19% | 0.23% | 0.33% | 0.16% | 0.17% |
| **♕a6** | 0.07% | 0.12% | 0.14% | 0.16% | 0.10% | 0.21% | 0.57% | 0.81% |
| **b6** | 0.01% | 0.01% | 0.00% | 0.01% | 0.01% | 0.01% | 0.67% | 0.17% |
| **♔h8** | 0.22% | 0.33% | 0.44% | 0.61% | 0.21% | 0.31% | 0.30% | 0.18% |
| **♕a2** | 0.06% | 0.05% | 0.06% | 0.08% | 0.05% | 0.08% | 0.31% | 0.48% |

# I Stockfish concept preference analysis

We analyze how different network layers prioritize chess concepts by measuring the expected change in concept values induced by each layer's move distribution. Unlike McGrath et al. (2022), who probed for concept *representations* using linear classifiers on intermediate activations, we directly measure concept *preferences* from policy outputs. This approach reveals what concepts each layer prioritizes when selecting moves, complementing representation-based analyses.

**Chess concepts** We use Stockfish 8's (Stockfish Developers, 2016) handcrafted evaluation function as our source of chess concepts, following McGrath et al. (2022) to enable direct comparison with prior work from Sadler and Regan (2019). Stockfish decomposes position evaluation into interpretable components including material balance, piece-specific features, king safety, threats, mobility, passed pawns, and spatial control. Each concept is evaluated separately for midgame (mg), endgame (eg), and produces a phase-interpolated value (ph) computed as a weighted sum of midgame and endgame values based on the game phase. All concepts are represented as continuous values. Table 34 summarizes the main concept categories we analyze.

Table 34: Summary of chess concepts from Stockfish 8's evaluation function taken from (McGrath et al., 2022). Concepts are enumerated as `<concept_name> <side> <game_phase>`, where side is `[mine|opponent|t]` for current player, opponent, or total (difference), and game phase is `[mg|eg|ph]` for midgame, endgame, or phase-interpolated value.

| Concept | Description |
|---|---|
| `material t [mg\|eg\|ph]` | Material score, where each piece on the board has a predefined value that changes depending on the phase of the game. |
| `imbalance t [mg\|eg\|ph]` | Imbalance score that compares the piece count of each piece type for both colours. E.g., it awards having a pair of bishops vs a bishop and a knight. |
| `pawns t [mg\|eg\|ph]` | Evaluation of the pawn structure. E.g., the evaluation considers isolated double, connected, backward, blocked, weak, etc. pawns. |
| `knights [mine\|opponent\|t] [mg\|eg\|ph]` | valuation of knights. E.g., extra points are given to knights that occupy outposts protected by pawns. |
| `bishops [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of bishops. E.g., bishops that occupy the same color squares as pawns are penalised. |
| `rooks [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of rooks. E.g., rooks that occupy open or semi-open files have higher valuation. |
| `queens [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of queens. E.g., queens that have relative pin or discovered attack against them are penalized. |
| `mobility [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of piece mobility score. It depends on the number of squares attacked by the pieces. |
| `king safety [mine\|opponent\|t] [mg\|eg\|ph]` | A complex concept related to king safety. It depends on the number and type of pieces that attack squares around the king, shelter strength, number of pawns around the king, penalties for being on pawnless flank, etc. |
| `threats [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of threats to pieces, such as whether a pawn can safely advance and attack an opponent's higher value piece, hanging pieces, possible xray attacks by rooks, etc. |
| `passed pawns [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluates bonuses for passed pawns. The closer a pawn is to the promotion rank, the higher is the bonus. |
| `space [mine\|opponent\|t] [mg\|eg\|ph]` | Evaluation of the space. It depends on the number of safe squares available for minor pieces on the central four files on ranks 2 to 4. |
| `total t [mg\|eg\|ph]` | The total evaluation of a given position. It encapsulates all the above concepts. |

**Perspective normalization** Stockfish evaluates all positions from White's perspective, with positive values favoring White. However, Leela's policy network evaluates from the current player's perspective. To ensure consistency with Leela's perspective and following McGrath et al. (2022)'s approach, we convert Stockfish's evaluations to player-relative coordinates.

For concepts with side-specific values (knights, bishops, rooks, queens, mobility, king safety, threats, passed pawns, space), we transform White/Black labels to Mine/Opponent labels based on who is to move:

- If White to move:
  - White concepts → Mine
  - Black concepts → Opponent
  - Concept values unchanged
- If Black to move:
  - Black concepts → Mine
  - White concepts → Opponent
  - Concept Values negated

For aggregate concepts (material, imbalance, pawns, total), we negate values when Black is to move to convert from White's perspective to the current player's perspective. This ensures that positive concept values always represent advantages for the player to move.

**Concept delta calculation** For each position $s$ with player $p$ to move, we compute concept preferences as follows:

1. **Evaluate initial position**: Use Stockfish 8 to obtain concept values $c(s)$ for all concepts $c$, normalized to player $p$'s perspective.
2. **Generate and evaluate legal moves**: For each legal move $m$ leading to position $s'$, evaluate $c(s')$ using Stockfish and normalize to player $p$'s perspective (accounting for the perspective flip after the move).
3. **Calculate concept deltas**: For each move-concept pair, compute:

$$\Delta c_m = c(s') - c(s) \tag{13}$$

   where both $c(s')$ and $c(s)$ are evaluated from player $p$'s perspective, ensuring $\Delta c_m$ consistently represents concept change from the moving player's viewpoint.

4. **Compute layer-wise preferences**: For each layer $\ell$, obtain the move probability distribution $\pi_\ell$ using the logit lens and calculate the expected concept delta:

$$\Delta c_\ell = \sum_{m \in \text{legal moves}} \pi_\ell(m) \cdot \Delta c_m = \mathbb{E}_{\pi_\ell}[\Delta c_m] \tag{14}$$

This weighted average represents the expected change in concept $c$ when sampling moves according to layer $\ell$'s policy.

**Dataset and sampling** We sample $10{,}000$ positions from the CCRL dataset (Leela Chess Zero team, 2018), a standard benchmark consisting of $2.5$ million computer games from CCRL $40/40$ and $40/4$ tournaments. The dataset contains games between strong chess engines and provides diverse positions across different game phases and strategic themes.

Positions are sampled using a two-stage hierarchical process:

1. Sample 10% of games uniformly from the dataset
2. From each selected game, sample 5% of positions uniformly
3. Filter duplicate positions to ensure uniqueness

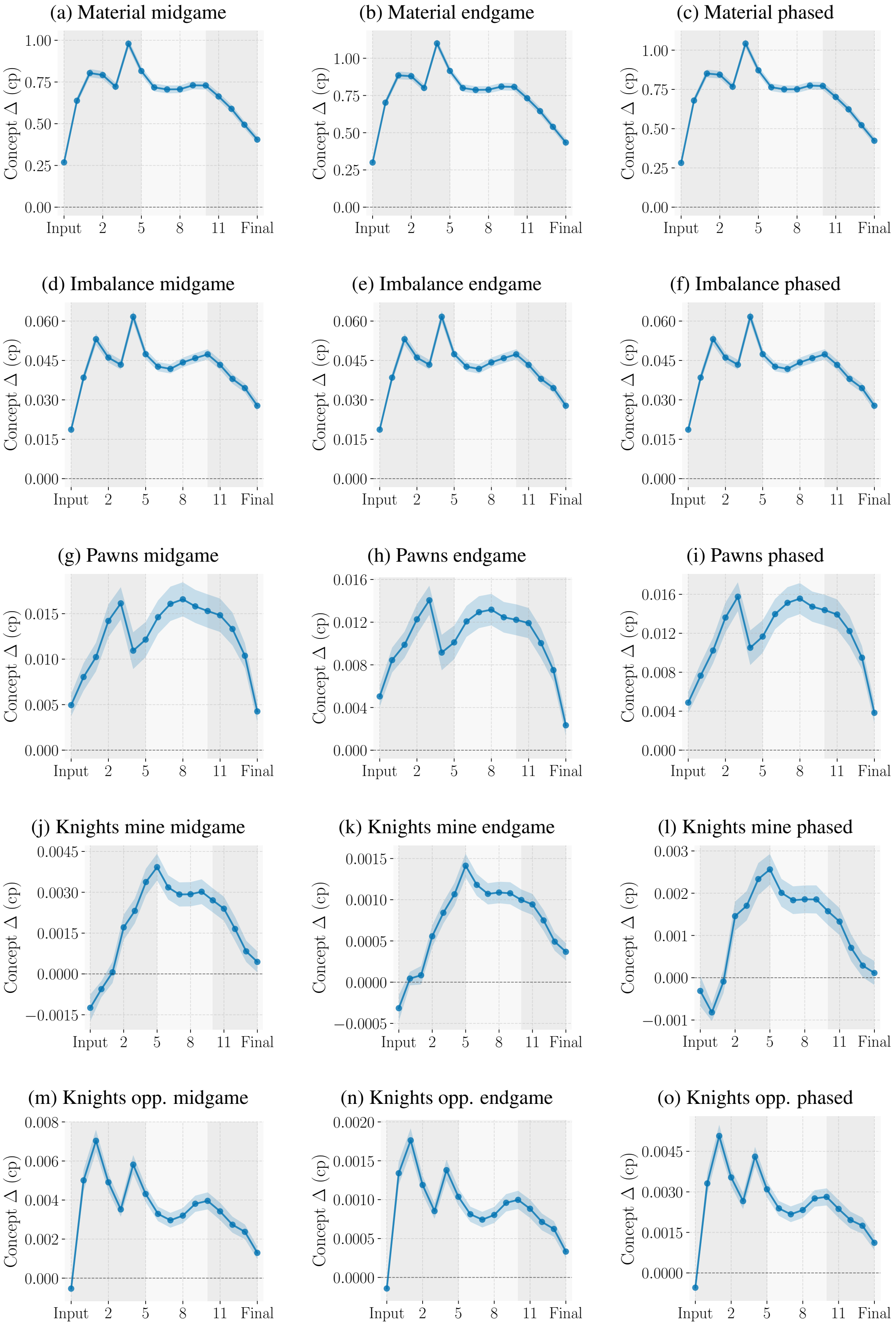
**Implementation details** For each sampled position, we:

- Extract move policies $\pi_\ell$ for the input embedding layer plus all 15 transformer layers, where the last layer corresponds to the final model output
- Evaluate the initial position and all legal move outcomes (typically 30-40 moves per position) using Stockfish 8
- Calculate concept deltas and probability-weighted averages for all 93 Stockfish concepts across midgame, endgame, and phase-interpolated variants
- Store results for analysis and visualization

Figures in the main paper present phase-interpolated (ph) concept values, as these represent Stockfish's actual evaluation for a given position. Comprehensive results showing all concept variants (mg, eg, ph) across all layers are provided in this appendix (Figures 21 to 27).

**Statistical analysis** We report mean $\Delta c_\ell$ values across all $10{,}000$ positions with $95\%$ confidence intervals computed via the $t$-distribution. Many concepts have values of zero—and therefore zero deltas—in positions lacking relevant pieces (e.g., bishop concepts when no bishops are present, or passed-pawn bonuses without passed pawns) or where the concept is undefined. This produces zero-inflated distributions that bias means toward zero. We do not control for this effect, which makes absolute $\Delta c_\ell$ values across concepts skewed and not directly comparable. Nevertheless, relative trends across layers remain informative for understanding how concept preferences evolve with depth.

**Interpretation of concept evolution** Analysis of concept preferences across layers reveals three distinct patterns corresponding to the computational phases identified in playing strength progression. In the early phase (layers Input to 5), most concepts exhibit erratic, volatile shifts with substantial fluctuations in preference values. During the middle phase (layers 5 to 10), concept preferences stabilize significantly, with most concepts maintaining relatively constant values across this range. This plateau mirrors the performance stagnation observed in tournament play and puzzle solving, suggesting a period of feature engineering. In the late phase (layers 11 to Final), concept preferences show consistent, smooth trends—typically monotonic increases or decreases for each concept, with minimal erratic behavior. This systematic evolution coincides with the sharp capability improvements and emergence of look-ahead mechanisms in final layers. The shift from aggressive to safety-oriented concepts (increased king safety preference, reduced threat preference) occurs primarily in this phase. These patterns hold across most concepts, though some show different dynamics requiring deeper chess expertise to interpret. The consistency of trends in middle and late phases is particularly striking given the volatility of early layers, suggesting fundamentally different computational regimes across network depth.



(a) Material midgame

(b) Material endgame

(c) Material phased

(d) Imbalance midgame

(e) Imbalance endgame

(f) Imbalance phased

(g) Pawns midgame

(h) Pawns endgame

(i) Pawns phased

(j) Knights mine midgame

(k) Knights mine endgame

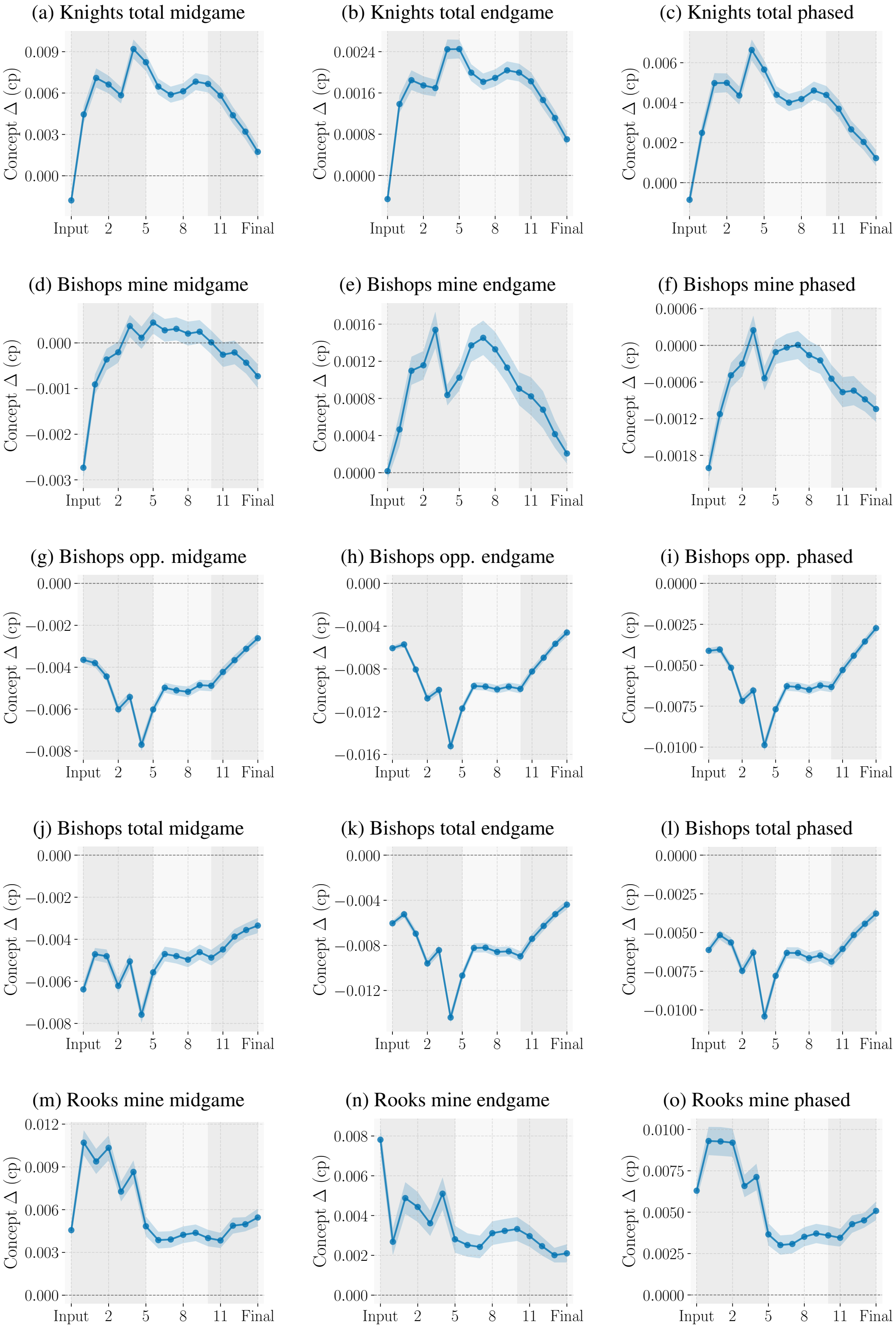
(l) Knights mine phased

(m) Knights opp. midgame

(n) Knights opp. endgame

(o) Knights opp. phased

Figure 21: Stockfish evaluation concepts across layers (Part 1 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.



(a) Knights total midgame

(b) Knights total endgame

(c) Knights total phased

(d) Bishops mine midgame

(e) Bishops mine endgame

(f) Bishops mine phased

(g) Bishops opp. midgame

(h) Bishops opp. endgame

(i) Bishops opp. phased

(j) Bishops total midgame

(k) Bishops total endgame

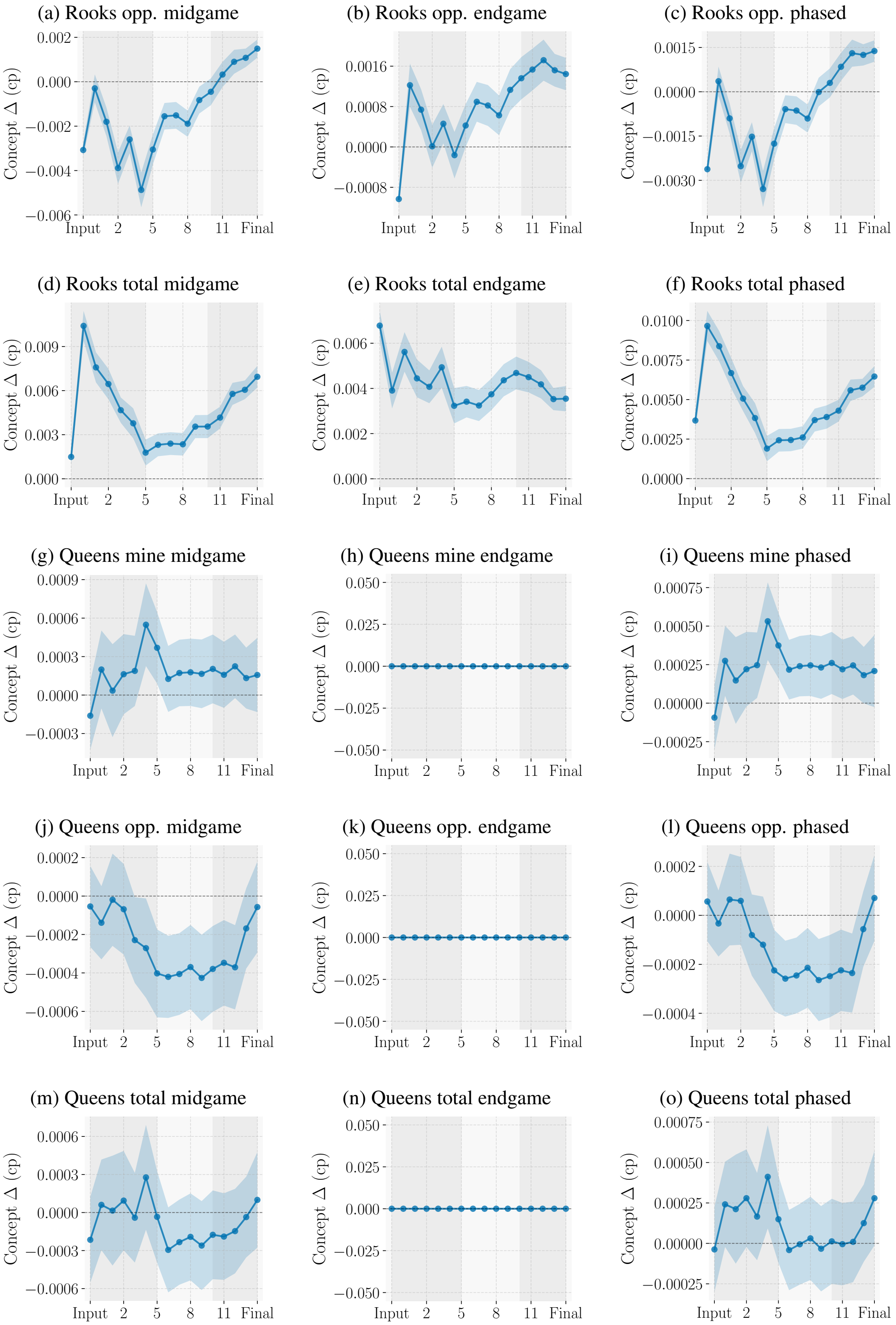
(l) Bishops total phased

(m) Rooks mine midgame

(n) Rooks mine endgame

(o) Rooks mine phased

Figure 22: Stockfish evaluation concepts across layers (Part 2 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.



(a) Rooks opp. midgame

(b) Rooks opp. endgame

(c) Rooks opp. phased

(d) Rooks total midgame

(e) Rooks total endgame

(f) Rooks total phased

(g) Queens mine midgame

(h) Queens mine endgame

(i) Queens mine phased

(j) Queens opp. midgame

(k) Queens opp. endgame

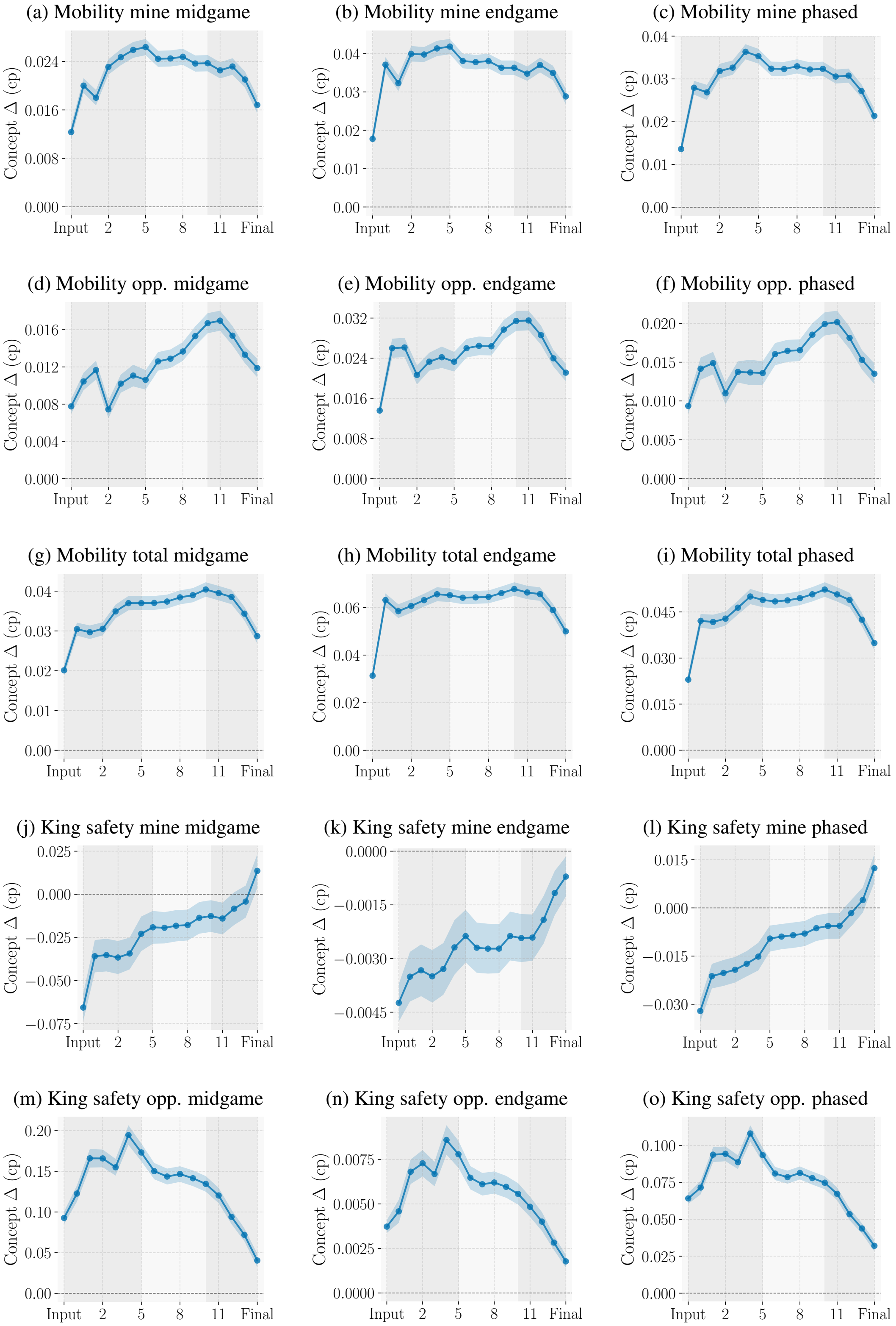
(l) Queens opp. phased

(m) Queens total midgame

(n) Queens total endgame

(o) Queens total phased

Figure 23: Stockfish evaluation concepts across layers (Part 3 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.



(a) Mobility mine midgame

(b) Mobility mine endgame

(c) Mobility mine phased

(d) Mobility opp. midgame

(e) Mobility opp. endgame

(f) Mobility opp. phased

(g) Mobility total midgame

(h) Mobility total endgame

(i) Mobility total phased

(j) King safety mine midgame

(k) King safety mine endgame

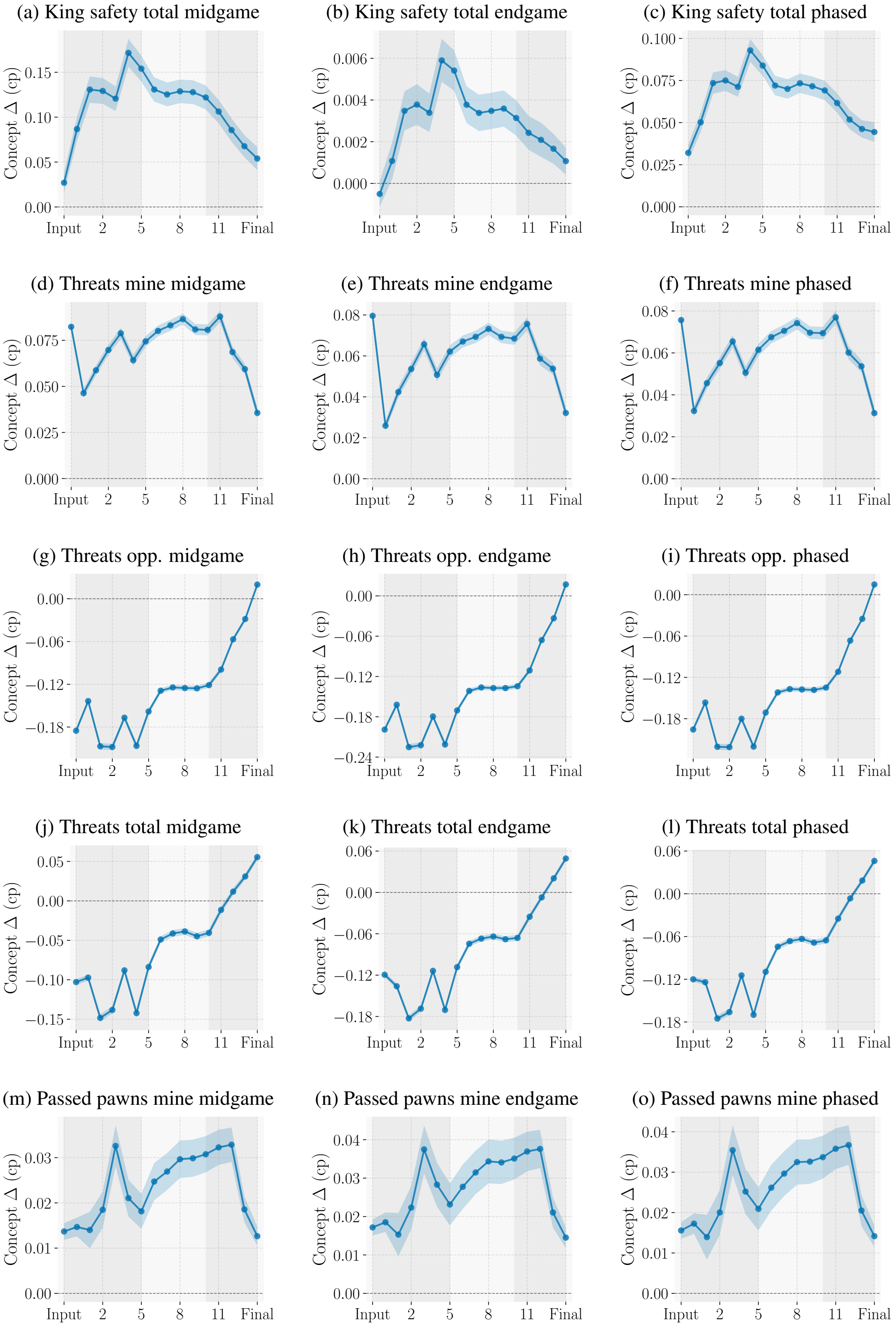
(l) King safety mine phased

(m) King safety opp. midgame

(n) King safety opp. endgame

(o) King safety opp. phased

Figure 24: Stockfish evaluation concepts across layers (Part 4 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.



(a) King safety total midgame

(b) King safety total endgame

(c) King safety total phased

(d) Threats mine midgame

(e) Threats mine endgame

(f) Threats mine phased

(g) Threats opp. midgame

(h) Threats opp. endgame

(i) Threats opp. phased

(j) Threats total midgame

(k) Threats total endgame

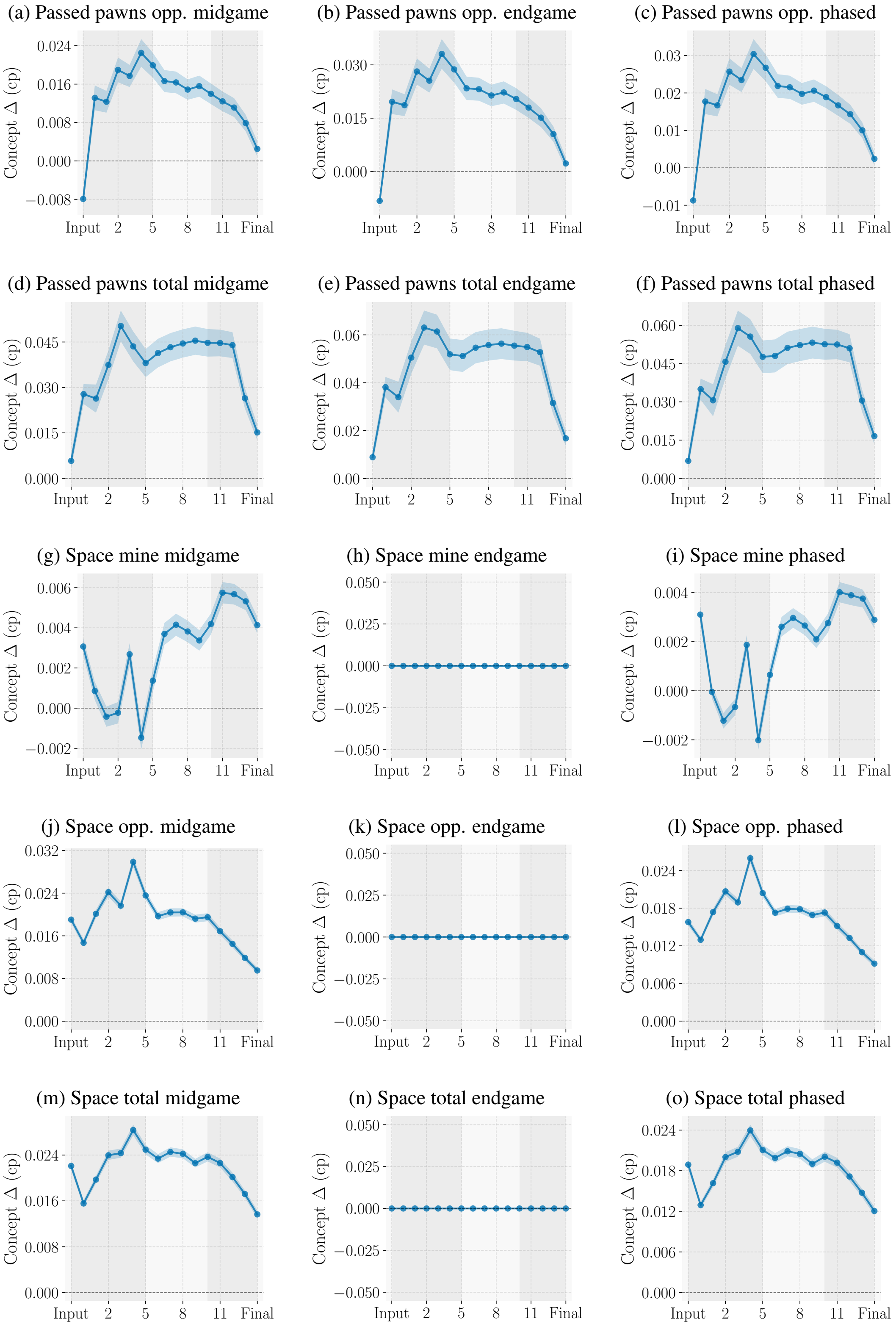
(l) Threats total phased

(m) Passed pawns mine midgame

(n) Passed pawns mine endgame

(o) Passed pawns mine phased

Figure 25: Stockfish evaluation concepts across layers (Part 5 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.



(a) Passed pawns opp. midgame

(b) Passed pawns opp. endgame

(c) Passed pawns opp. phased

(d) Passed pawns total midgame

(e) Passed pawns total endgame

(f) Passed pawns total phased

(g) Space mine midgame

(h) Space mine endgame

(i) Space mine phased

(j) Space opp. midgame

(k) Space opp. endgame

(l) Space opp. phased

(m) Space total midgame

(n) Space total endgame

(o) Space total phased

Figure 26: Stockfish evaluation concepts across layers (Part 6 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.

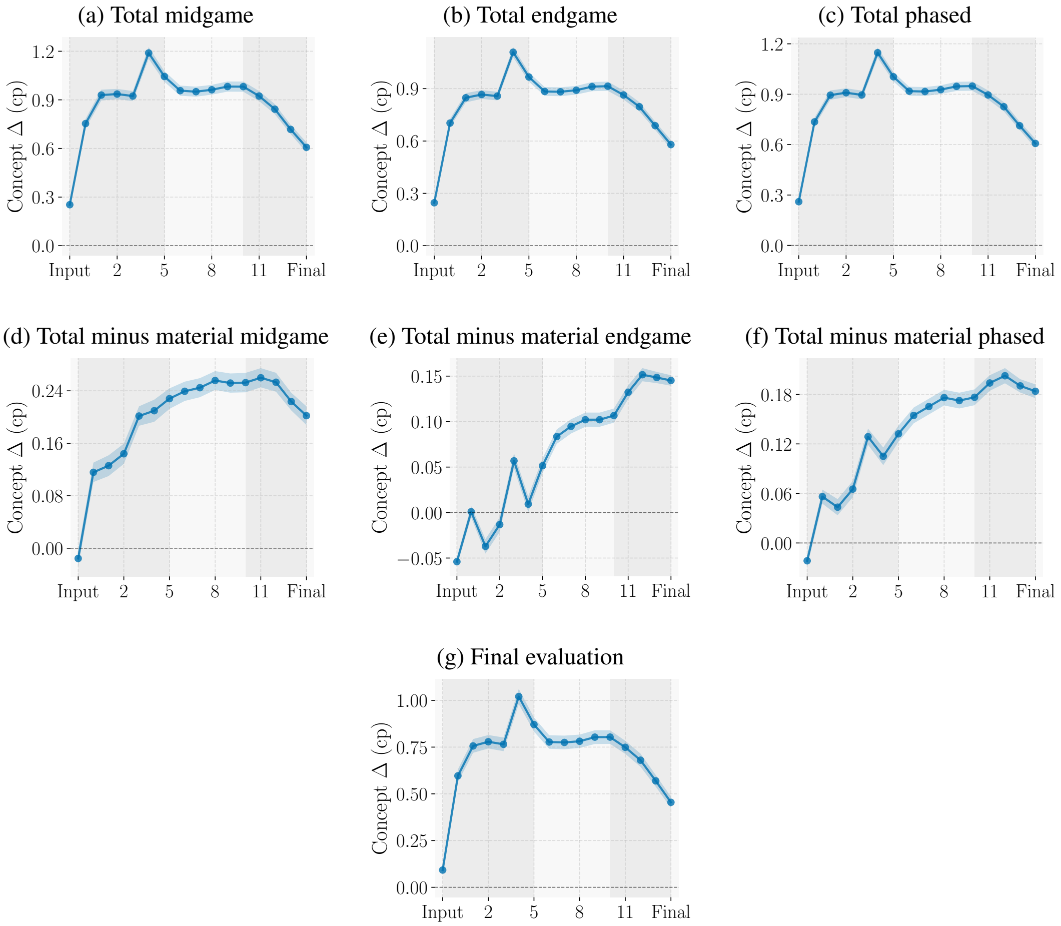


Figure 27: Stockfish evaluation concepts across layers (Part 7 of 7). Lines show mean probability-weighted concept delta across positions; shaded regions show 95% confidence intervals. All concepts evaluated from current player's perspective. Shaded regions indicate network phases.