PEBBLEBED · TECHNICAL REPORT

Nous: Planning by Revisable Imagination in a Discrete Diffusion Chess Engine

Nous is a searchless chess engine that plans by generating and iteratively revising imagined trajectories. It doubles as an exact ground truth testbed for three questions in machine reasoning.

Balaji Daggupati Pebblebed July 2026 55M parameters · no search at play time
lichess.org/@/PebbleTraining

ABSTRACT

Nous (an old word for intuitive knowing) is a masked discrete diffusion model that plays chess by imagining a multistep continuation and revising it, rather than by tree search. A 55M parameter bidirectional transformer trained as an absorbing state diffusion over interleaved board and move trajectories, it reaches roughly 1600 blitz on Lichess with no search at play time. Because chess supplies exact ground truth (Stockfish, tablebases), Nous doubles as an instrument for measuring phenomena that are ordinarily unobservable. We report four results. First, imitation acquires positional structure but not terminal logic. Second, a **verifier resolution threshold**: a learned value model degrades move selection when used to rank but improves it when used only to veto, until its accuracy crosses a measurable point. Third, reinforcement learning with verifiable rewards is shown by decomposition to **amortize (reorder) rather than create** ability, 93 percent versus 7 percent. Fourth, the binding constraint on calculation is **imagination fidelity** (about 28 percent next move agreement), which improves with model scale. A controlled scaling study (55M versus 122M, matched) finds fidelity rises with scale while the test time compute ceiling does not. We close with a falsifiable scaling hypothesis.

A shorter, nontechnical essay, "Introducing Nous," covers the same three findings for a general reader. This document is the full account: the architecture, the measurements, and the exact conditions under which each claim would be considered false.

1 Motivation

Strong chess engines are search engines. Alpha beta engines evaluate on the order of ten million positions per move; AlphaZero style systems amortize search into a policy but retain Monte Carlo tree search at inference. Human play is different in kind. Protocol studies (de Groot) place expert search at tens of positions per move, organized as a few candidate lines explored in depth and *repaired* when their continuations prove unfavorable. That repair operation is not expressible in either dominant model family: autoregressive decoders cannot revise a committed prefix, and tree search discards lines rather than editing them in place.

Masked discrete diffusion is the exception. Because it materializes an entire future as a partially masked draft and denoises it over multiple passes, any span, including the opening move of a plan, can be remasked and regenerated conditioned on the remainder. This report asks whether that property can serve as the load bearing calculation mechanism of a chess engine. The resulting system, Nous, doubles

as an instrument for three questions that also arise in large model reasoning but lack exact ground truth there.

2 Architecture

2.1 Representation

A position and continuation window is tokenized as an interleaved trajectory of states and moves over a vocabulary of 202 symbols. Each board state occupies 71 tokens: 64 square occupancy tokens, the side to move, castling rights encoded natively in Chess960 form by rook file, and the en passant file. Each move is 3 tokens (origin, destination, promotion), so one ply is 74 tokens. A training window is STATE0 plus five repetitions of [MOVE, STATE], 441 tokens. Re encoding the full board at every ply is deliberate: ablating the intermediate states collapses next move accuracy from 30.2 to 21.2 percent, so the explicit states function as working memory rather than redundancy.

2.2 Backbone and heads

The trunk is a 55M parameter bidirectional transformer: d_model 512, 16 layers, 8 heads, RMSNorm, rotary positional embeddings, SwiGLU. Two heads share it. The actor (denoiser) predicts clean tokens at every future position. The critic mean pools the current position and maps it through a small MLP to a distribution over 128 value bins (HL Gauss). Against depth 16 Stockfish the critic attains Pearson 0.86 and mean absolute error 0.13 on a minus one to plus one scale; retraining on action values reduces the error to 0.061.

2.3 Objective

The forward process is absorbing state, in the MDLM family: at noise level t, each future token is independently replaced by a mask token with probability t, and the current position is never masked. Training minimizes a time weighted cross entropy over masked positions, normalized by sequence length. The normalization is load bearing: normalizing by masked token count and applying the 1/t ELBO weight yields an effective per token weight proportional to 1/t squared, which starves the high noise regime that generation begins from and produces a model with healthy validation loss that generates zero percent legal moves. The operative lesson, that token level loss is not a valid gate whereas generation time probes are, governs all subsequent evaluation.

2.4 Decoding and move selection

Inference uses MaskGIT style decoding over K passes: token values are the argmax, and stochasticity is confined to the commit order through Gumbel perturbed confidences. Move selection follows a propose then veto scheme: factorized first move logits score all legal moves; the top candidates, plus an injection that guarantees every check and capture a hearing, are evaluated by the critic on their resulting positions; the policy preferred move stands unless a candidate exceeds it by a veto margin. A ledger of rulebook guards (mate, repetition, fifty move, stalemate, static exchange) supplies terminal logic the model has not internalized, and each guard carries a retirement test.

3 Training

A base model is pretrained on Lichess games rated at least 1800, then distilled on Stockfish principal variation continuations rather than individual state and move pairs, so the supervision is native to trajectories. Era 6 trains on 23.75M deep principal variation windows from the public Lichess evaluation database. Era 6.5 trains jointly on 51.5M windows: deep principal variations, 16M master game windows from a master game collection, opening windows, 8M critical positions mined from multi variation evaluation gaps, multistep tactics, and self play blunder corrections, with all sources present from the first step. The value head is then retrained, with trunk and actor frozen, on 8M action value records, isolating a judge upgrade while the player is held identical. Reinforcement learning (Era 6.7) applies GRPO with verifiable rewards on tactics: the policy generates several candidate lines seeded to distinct first moves, the reward is the fraction of the true solution line reproduced, the advantage is group relative, and a KL term anchors to the pre reinforcement model.

TABLE 1 · PLAYING STRENGTH BY GENERATION (WIN RATE VERSUS STOCKFISH, N = 150)

GENERATION	CHANGE	SKILL 0	SKILL 1
gen 5	imitation and PV distillation	65%	36%
6	deep PV corpus, 6.4x data	82%	61%
6.6	action value critic	**97%**	**88%**
6.7	verifiable reward RL	97%	88%

4 Results

4.1 Imitation acquires structure, not consequence

Trained by imitation, the model develops competent positional play (it outperforms higher rated opponents from the opening) but not consequential logic: it hangs material, misses mate in one, and repeats winning positions into draws. The illegal move taxonomy is diagnostic. All generation errors are geometric (sliding through occupancy), with zero ownership or occupancy violations, so the model learns the appearance of legal chess exactly while missing move dynamics. Corrective signal came only from oracle supervised data or rulebook guards, never from additional imitation.

4.2 The verifier resolution threshold

A learned value model reliably **vetoes** catastrophic candidates but degrades selection when used to **rank** them, until its accuracy crosses a threshold. Five configurations (greedy value argmax, sequential Monte Carlo particle guidance, adversarial worst reply reranking, value first calculation) each reduced strength under the r = 0.86 critic, down to minus 20 points per 75 games for adversarial reranking, whereas the same critic used only as a coarse veto added an estimated 100 to 150 Elo. Retraining the critic on action values (error 0.13 to 0.061) reversed the sign: adversarial reranking became plus 10.7 points per 150 games, and on a specific live blunder the earlier judge selected the losing move while the upgraded judge

selected the winning one, with the policy held fixed. This is the reward model, or best of N, selection problem from large model reasoning, measured against exact ground truth.

4.3 Imagination fidelity is the binding constraint

A calibration probe over the engine's own games finds that its imagined opponent reply matches the realized move about 28 percent of the time, and per ply consistency of the imagined line falls from 73 percent at one ply to 28 percent at five. This accounts for the failure of lookahead even under the improved judge: revision and chained calculation operate on the imagined line, so compounding a low fidelity world model degrades the estimate. Chained calculation scored 23 percent against the base engine's 42 percent on multistep tactics. The constraint is therefore neither decoding nor judge quality but the truthfulness of the imagination itself.

4.4 Reinforcement learning amortizes rather than creates

Verifiable reward reinforcement learning produced the first net positive mechanism result: held out multistep first move accuracy 66.6 to 73.6, with no whole game regression, deployed as Era 6.7. A decomposition against the base model candidate set is decisive.

TABLE 2 · RL EFFECT DECOMPOSITION AND REWARD SOURCE COMPARISON (HELD OUT)

MEASUREMENT	VALUE
Corrected tactics already in base top 5 (reordering)	**93%**
Corrected tactics outside base top 5 (creation)	7%
First move gain, exact or verifiable reward	**+7.0**
First move gain, weak learned critic (error 0.13)	+4.8
First move gain, the model's own critic (error 0.061)	+2.2

Reinforcement learning reorders within the base model candidate set and does not extend it; verifiable reward dominates any learned critic reward; and rewarding the model with its own value function yields the least gain, indicating the signal must carry external truth. A live play caveat: reinforcement learning raises tactical initiation without raising calculation depth, so the model enters tactics it cannot resolve, which is net positive but carries an unsound sacrifice tax; the puzzle benchmark, being tactic present by construction, overstates the live value.

4.5 Test time compute, emergence, and scaffolding

Additional refinement passes raise tactical accuracy (36.8 to 41.5 percent up to K about 32) then saturate. Competencies emerge from the data distribution rather than from engineering: castling, lost when the deep principal variation corpus omitted opening positions, returned natively under joint training with no special casing, whereas sequential rehearsal traded it against roughly 10 points of tactics at any mixing ratio. A guards off retirement test shows the model has not internalized terminal logic at this scale: with selection and guards removed it scores 39 and 35 percent versus 97 and 88 percent for the full stack. By phase evaluation locates the residual weakness in the middlegame (average

centipawn loss 47, 193, and 149 for opening, middlegame, and endgame), consistent with calculation depth as the frontier.

5 Controlled scaling study

To test whether scale addresses the fidelity constraint and the compute ceiling, a 122M and a 55M model were trained from scratch on identical data and step budgets, so that size is the only variable.

TABLE 3 · MATCHED CONDITION SCALE COMPARISON

METRIC	55M	122M
Gauntlet, skill 0 and 1	81 / 47	55 / 34
Multistep tactics solve	29.4%	25.4%
Compute slope peak (saturation)	51.5% (K≈24)	47.3% (K≈32)
Imagination fidelity	12%	**20%**

Three results follow. First, at a fixed step budget the larger model is *undertrained* and weaker on strength metrics, so scaling parameters requires proportionally more optimization. Second, imagination fidelity nonetheless **rose with scale** (12 to 20 percent) despite undertraining, the strongest positive signal that the binding constraint responds to capacity. Third, the compute slope ceiling did *not* lift (both saturate at K about 24 to 32), indicating the ceiling is **bound by horizon, not by size**, which implicates the five ply imagination window. The indicated next system is larger, longer in horizon, and trained to convergence, with reinforcement learning as a finishing layer rather than a primary driver.

6 Discussion

The contributions are largely general; chess supplies the exact ground truth that the analogous settings lack. The verifier resolution threshold (4.2) is the reward model selection problem in reasoning models. Amortization versus creation under reinforcement learning (4.4) bears on the ceiling of reinforcement learning applied to a fixed base. Imagination fidelity (4.3) is a portable diagnostic for any model that plans within a learned world model. Each is answerable here directly and against a public, falsifiable scoreboard. The through line across all three is that each mechanism depends on the model's judgment staying tied to truth from outside itself.

7 Limitations and falsification

Absolute strength is modest (about 1600), and the reported mechanisms are, at 55M, either bounded (test time compute), net neutral pending scale (revision, chaining), or dependent on scaffolding (terminal logic). Puzzle benchmarks are tactic present and overstate live tactical value; one earlier benchmark contained train and evaluation leakage, since corrected with disjoint held out sets. The thesis is falsifiable: if, at scale, imagination fidelity does not rise enough to make the mechanisms net positive and the compute ceiling does not lift with a longer horizon, the result becomes a characterization of what imitation cannot learn, what a learned judge can and cannot adjudicate, and where searchless

planning by imagination fails, reported regardless. All negative results above were fixed in advance of the experiments that succeeded them.

References

1. Sahoo et al. Simple and Effective Masked Diffusion Language Models. NeurIPS, 2024.
2. Ye et al. Implicit Search via Discrete Diffusion: A Study on Chess. ICLR, 2025.
3. Janner et al. Planning with Diffusion for Flexible Behavior Synthesis. ICML, 2022.
4. Ruoss et al. Amortized Planning with Large-Scale Transformers: A Case Study on Chess. NeurIPS, 2024.
5. Shao et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2024.
6. Gao et al. Scaling Laws for Reward Model Overoptimization. 2022.
7. Yue et al. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? 2025.
8. Farebrother et al. Stop Regressing: Training Value Functions via Classification for Scalable Deep RL. 2024.

Instrument (live): lichess.org/@/PebbleTraining. Data: Lichess evaluation database (394M positions, CC0), DeepMind ChessBench (530M states, 15.3B action values), a master game collection, Lichess puzzle database, Syzygy tablebases.

Pebblebed · Technical Report · July 2026