SPECTRA: CONDITIONAL PREDICTION WITH SIDE-INPUT-INVARIANT STRUCTURAL STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting human chess moves calls for adapting to a player's strength. Can those predictions change while the model's board computation stays fixed? We study this question with SPECTRA. At each layer, board tokens read only board tokens; the rating-dependent side token reads the board tokens and itself; the prediction token reads all three types. Under the stated input and layer assumptions, changing the rating cannot alter any board state at fixed parameters. On the same 19,546 games, three seeds place SPECTRA 0.375 ± 0.159 top-1 percentage points above late fusion and 0.096 ± 0.085 points below unrestricted attention (mean $\pm$ seed SD). These comparisons evaluate complete architectures, not an isolated connection. Reusing board states makes 26 rating queries $5.12\times$ faster than ordinary SPECTRA execution at batch 64, while late fusion remains faster in absolute terms. Other input partitions preserve the same invariant but have different predictive costs, including a substantial loss on Waterbirds. Keeping board states fixed permits reuse, but does not by itself improve prediction; the measured tradeoffs depend on the task and input partition.

1 INTRODUCTION

Consider the same chess position for players of different strengths. A move predictor can use their Elo ratings, which summarize playing strength, because the players may choose differently on the same board with the same legal moves. We ask whether those predictions can change while sharing the same internal board computation.

Existing models already condition move predictions on skill. Our question is where the rating enters the computation. We treat it as a side input: context supplied alongside the board. Rating-dependent board states change the internal reference when predictions are compared and cannot simply be reused across ratings. Fixed states provide a common computation, without guaranteeing that it is correct, interpretable, or free of information correlated with skill.

Unrestricted attention allows the board tokens, side token, and prediction token to influence one another. Conventional late fusion instead computes board representations without ratings and introduces ratings at the prediction head. We study a particular one-way construction between these evaluated systems: board tokens read only board tokens; the side token reads the board tokens and itself; the prediction token reads all three types (Figure 1). This construction uses established directional-attention principles.

In SPECTRA, the mask enforces these reads at every layer. Provided that the board tokenizer and mask do not depend on the rating and the remaining operations act within each token, changing the rating cannot change any board state at fixed parameters. We call these states *protected*. This is inference-time noninterference for the declared input path (Goguen & Meseguer, 1982); side-dependent training losses can still affect the parameters that are learned.

We specify the three-stream construction and prove its declared-input invariant. Experiments check its implementation, whether ratings help predict recorded moves, and the cost of repeated queries. An exploratory change of uncertainty ranking tests the interpretation of where predictions change. These complete-architecture comparisons find a modest chess accuracy advantage over late fusion, which has lower cached latency. Three other input partitions retain state protection but incur different predictive costs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

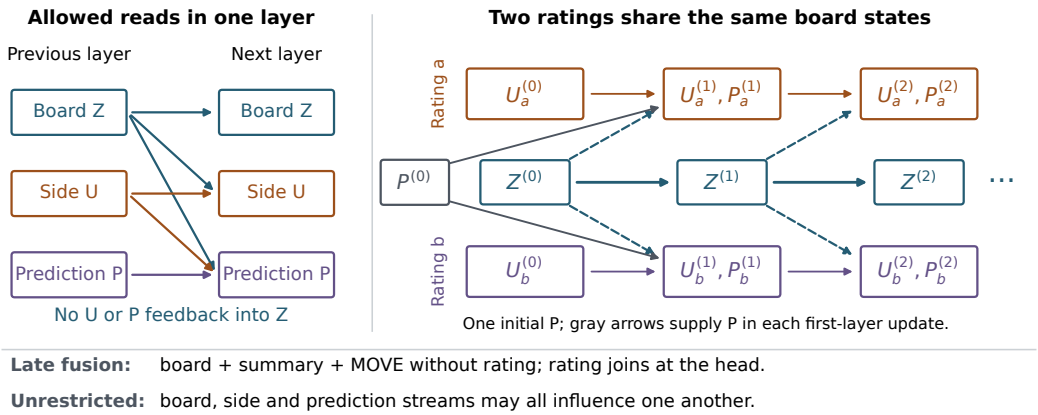


Figure 1: What the mask permits and what can be reused. Left: arrows carry information from the previous layer to the next; the prediction token remains a separate stream. Right: two ratings use the same board trajectory and separate side/prediction states. No conditional arrow returns to the board stream. The bottom contrasts the evaluated alternatives; Section 4 specifies their full forward paths.

2 Related Work

The closest comparisons concern where context enters, which states it can change, and how human moves are predicted.

Directional attention and information flow. Encoder–decoder attention lets decoder queries read encoder states without feeding back into them (Vaswani et al., 2017). UniLM's sequence-to-sequence mask likewise lets target tokens read source tokens, but not the reverse (Dong et al., 2019). Hasse Diagrams for Attention formalizes reachability under attention masks (Li & Guo, 2026). Following these precedents, we study structural, side, and prediction streams in a shared stack, including their prediction and reuse behavior. A protected encoder with a context decoder can satisfy compatible dependencies.

Side-information fusion. NOVA uses side information in attention queries and keys while retaining item-only values (Liu et al., 2021a). Its weights can therefore change under side-input substitution; excluding context from values does not make the resulting states invariant. ASIF, a WWW 2024 Companion paper, aligns item and attribute representations and combines them within layers, situating its design among early, late, and hybrid fusion (Wang et al., 2024). Our restriction instead requires designated structural states to remain identical when the side input changes.

Human-chess prediction. Maia uses rating-specific policies (McIlroy-Young et al., 2020); Maia-2 conditions attention queries on skill after a skill-independent ResNet encoder (Tang et al., 2024). Chessformer supplies board-square tokens, geometric attention, and a policy head (Monroe et al., 2026). Allie models game sequences and thinking time with adaptive search (Zhang et al., 2025); Otter adds move history and clock conditioning (Tarun Kumar S, 2026). These richer inputs address human behavior beyond our rating-substitution study. Different inputs, training data, and evaluation protocols preclude a direct ranking from published accuracy figures.

Frozen policies and conditional corrections. UniMaia injects prompt-conditioned residuals into a frozen chess backbone; later backbone layers receive the residual-updated states (Siu & Istead, 2026). Frozen weights therefore do not imply prompt-invariant activations. Matilda combines a frozen, rating-conditioned human policy with search-derived features and residual re-ranking (Carlson, 2026). Its preserved prior already reads ratings. Our guarantee concerns rating substitution in designated structural states, not parameter freezing or initial equivalence from zero-initialized residuals. This distinction does not establish a predictive advantage over these systems.

Other notions of invariance. Disentanglement and identifiability concern statistical structure (Locatello et al., 2019; Khemakhem et al., 2020); parametric information-flow control restricts training-data influence (Tiwari et al., 2024). IRM, GroupDRO, and Just Train Twice target predictive behavior across environments or groups (Arjovsky et al., 2019; Sagawa et al., 2020a;b; Liu et al., 2021b), while counterfactual fairness requires causal assumptions (Kusner et al., 2017). Our mask implies none of these properties. The non-chess experiments compare routes under selected partitions, not competitive group-robust training methods. Training and selection also affect comparisons (Gulrajani & Lopez-Paz, 2021); equal fitted capacity does not isolate a routing effect.

3 ASYMMETRIC TYPED-TOKEN ROUTING

Let X be the primary input and A the declared side input. Tokens are initialized as $Z^{(0)} = f_X(X)$, $U^{(0)} = f_A(A)$, and a learned prediction token $P^{(0)}$, then serialized as $[Z_1, \ldots, Z_m, U, P]$. Every layer uses the allowed-attention relation below, with queries indexing rows and keys indexing columns:

$$M = \begin{array}{c|ccc} & Z & U & P \\ \hline Z & 1 & 0 & 0 \\ U & 1 & 1 & 0 \\ P & 1 & 1 & 1 \end{array}.$$

An information-flow arrow $Z \to U$ means that a U query reads a Z key. Residuals, layer normalization, and feed-forward blocks operate within each token. The structural tokenizer and mask are independent of A; blocked reads receive zero attention weight before value aggregation. These assumptions exclude side-dependent preprocessing, masks, or cross-token normalization.

Proposition 1 (Declared-input noninterference)**.** *For fixed parameters and deterministic inference, any depth L, primary input X, and side inputs A, A', the structural output satisfies $Z^{(L)}(X, A) = Z^{(L)}(X, A')$. The same statement holds with coupled randomness in stochastic token-local operations.*

This proposition states the dependency restriction; its proof is a direct induction. Initially, structural tokens depend only on X. If this holds at layer ℓ, every key and value visible to a structural query is a function of $Z^{(\ell)}$ and fixed parameters. Attention and the token-local operations preserve the property at layer $\ell + 1$. Appendix A gives further details. The invariant holds for each fixed parameter setting, even when side-dependent objectives influence parameter updates during training.

Both side and prediction streams read intermediate structural states, which need not be recoverable from the final state. For the trajectory $\mathcal{T}_Z(X) = (Z^{(0)}, \ldots, Z^{(L)})$, the valid factorization is

$$Z^{(L)} = g(X), \qquad U^{(L)} = h(A, \mathcal{T}_Z(X)), \qquad \hat{Y} = q(Z^{(L)}, U^{(L)}, P^{(L)}).$$

Intermediate access distinguishes this implementation from conventional late fusion; we do not claim a function-class separation from an arbitrarily expressive fusion head.

The chess policy combines structural logits with a conditional adjustment:

$$\ell = \ell_Z(Z^{(L)}) + s\, g_P(P^{(L)})\, d(U^{(L)}, P^{(L)}), \qquad p = \mathrm{softmax}_{\mathrm{legal}}(\ell), \tag{1}$$

where $s \in [0.5, 2.5]$ is a clamped learned scalar and $g_P = 0.15 + 0.85\sigma(G(P^{(L)}))$. Here G is a gate multilayer perceptron (MLP; Appendix B). The structural head maps the 64 board states to 4,672 action logits; an MLP computes d from the concatenated side and prediction states. Structural response measures changes in ℓ_Z, and conditional response changes in $\ell - \ell_Z$. Appendix B specifies the controls' use of these modules.

Inspecting and reusing the same states. Changing the side input preserves every structural state. Where differentiable, $\partial Z^{(L)}/\partial U^{(0)} = 0$. A dependency check can therefore compare conditional predictions against the same structural computation. For several queries about the same input, the model can save the structural states once and supply each layer's saved state to the side and prediction streams for each new context. This cache is tied to the primary input, model parameters, and tokenizer; changing any of them requires rebuilding it. Saving the intermediate states avoids assuming that the final state can reconstruct them. The mask determines what stays fixed. Training determines whether the allowed changes improve prediction.

For multiple side inputs, protection requires the absence of every direct or indirect path to the structural stream, including paths through prediction tokens. Appendix A gives this extension.

4 EXPERIMENTAL DESIGN

The evaluation checks whether board states stay fixed, whether ratings improve prediction, and how much computation repeated queries can reuse.

Models and conditional inputs. We compare four-layer, width-384 transformers with $24,713,267$ fitted parameters for seeds 42, 1337, and 1842. Each uses 64 square tokens from 19 board planes and a 4,672-action policy. In SPECTRA and no-mask, the remaining tokens encode active/opponent Elo and prediction. Late fusion instead uses an Elo-independent summary slot and MOVE token in its board transformer; Elo joins at the adjustment head. Its full-attention board encoder and board-only gate differ from SPECTRA. No-mask opens all 129 reads blocked by SPECTRA, including prediction-token feedback. Equal parameter counts control fitted capacity, not every aspect of computation.

Training data and comparability. All nine chess fits use the same underlying 13.58-million-game Lichess Rapid corpus and preprocessing/filtering rules. Historical runs used different shard layouts; exact train/validation assignments and streamed example orders are not assumed identical across architectures. Architecture-specific forward paths share the reported optimization objectives and schedules, with definitions in Appendix C.

Training the conditional policy. Move cross-entropy is combined with rating-bucket contrastive learning, skill and auxiliary move prediction, and adjustment and gate regularization. Conditional residual and ranking losses are introduced after the initial 50,000 steps; they use structural logits and emphasize corrections where the structural policy is uncertain. During the initial phase, SPECTRA and no-mask still construct and process the Elo token: zeroing the effective side output does not remove Elo from attention or disable every auxiliary loss. Late fusion zeros its separate Elo input. The schedule increases conditional adjustment after the initial phase; it is not strict structure-only pretraining. Its advantage over one-stage training is untested, and these objectives can affect the observed uncertainty patterns.

Held-out evaluation. The Maia evaluation begins with 11.29 million rows. Rapid positions with a legal observed move, ply at least 10, adequate or missing clock information, and both ratings yield 1,091,559 eligible rows from 19,546 games. A deterministic hash selects one position per game. All fits use identical labels, game order, and Elo assignments. A retained identifier audit reports no overlap with the training corpus. Appendix D gives the construction.

Interventions and uncertainty. We sweep 26 Elo indices while holding the board and opposing rating fixed; behavioral averages use only the 16 supported buckets. Metrics include top-1 accuracy, negative log-likelihood (NLL), fixed-width 15-bin top-label expected calibration error (ECE), top-move changes, and Jensen–Shannon (JS) divergence from the observed-Elo policy. NLL penalizes low probability assigned to the recorded move. ECE is the frequency-weighted absolute confidence–accuracy gap. JS measures distributional change, not prediction quality.

Route tests use supported substitutions, an autograd projection of $Z^{(L)}$ with respect to literal $U^{(0)}$, and directional finite differences. Paired percentile intervals resample games conditional on each fit; seed summaries are means and sample SDs across three fits. Distinct games reduce within-game pseudoreplication but need not be independent across players or time. These intervals describe this sample conditional on the fitted models.

Games are ranked by a protected policy's top-two legal-move probability margin and split into quintiles. Q1 has the largest margins; Q5 has the smallest. The same indices apply to every model, first using SPECTRA's ranking and then, in an exploratory symmetric analysis, late fusion's ranking. The endpoint interaction is $(F_S - F_{LF})_{Q5} - (F_S - F_{LF})_{Q1}$, where F_S and F_{LF} are the fractions of supported alternatives changing the top move for SPECTRA and late fusion. Observed-versus-nearest-supported-Elo likelihood and accuracy provide separate tests of local usefulness.

Other domains. Adult (Becker & Kohavi, 1996), WILDS CivilComments (Koh et al., 2021), and Waterbirds (Sagawa et al., 2020a) test the route with tabular, text, and image tokenizers (Table 1). Their results compare SPECTRA with no-mask. Worst-group accuracy is the lowest accuracy across the specified evaluation groups. Full three-way controls are available only for chess.

Table 1: Task partitions. CivilComments indicators describe identity mentions, not verified author identity. Waterbirds derives both streams from the same image.

Task	Structural input	Side input	Target / evaluation groups
Chess	64 square tokens	Active/opponent Elo	Played move / Elo buckets
Adult	Ten feature tokens	Race, sex, marital status, native country	Income / label × sex × race
CivilComments	Eight DistilBERT projections	Eight mention indicators	Toxicity / label × identity mention
Waterbirds	49 spatial ResNet tokens	Pooled image context	Bird class / bird × background

5 RESULTS

5.1 DO BOARD STATES STAY FIXED, AND WHAT IS THE PREDICTION ACCURACY?

The first check is whether changing the rating leaves the board states fixed. All SPECTRA fits show exactly zero structural-token drift under the tested Elo substitutions. Autograd and finite differences also return zero at evaluator precision (Table 2). Late fusion has zero substitution drift in each seed; every no-mask partner responds. These measurements check the implementation on tested inputs, while the graph argument establishes the general property. The derivative measurements use a random projection and four finite-difference examples, not an exhaustive numerical Jacobian.

Table 2: Three checks of the declared route for SPECTRA and its no-mask control. Substitution reports the maximum change in post-transformer Z; autograd and central finite differences differentiate that same output with respect to the literal pre-transformer U input.

Architecture	Seed	max $\lvert\Delta Z\rvert$	max autograd	max finite diff.
SPECTRA	42	0	0	0
SPECTRA	1337	0	0	0
SPECTRA	1842	0	0	0
No-mask	42	58.6564	7.57825e-05	0.274658
No-mask	1337	109.111	0.000212527	1.544
No-mask	1842	93.0627	4.12973e-05	0.213623

Table 3: Parameter-matched chess comparison across three seeds and 19,546 shared games per seed. Top-1 reports mean ± seed SD; the remaining columns report three-seed means. Structural and conditional response measure ELO-sweep logit change attributed to the two pathways. Lower is better for NLL and ECE. The no-mask conditional response rounds to 0.000; it is not an exact zero.

Model	Top-1 (%)	NLL	ECE	Flip (%)	JS	Structural	Conditional
SPECTRA	49.920 ± 0.194	1.57490	0.01111	6.375	0.00357	0.000	0.192
Late fusion	49.545 ± 0.352	1.58421	0.02136	7.221	0.00415	0.000	0.180
No-mask	50.015 ± 0.116	1.56415	0.01457	10.887	0.01076	0.382	0.000

Table 3 compares prediction quality: SPECTRA reaches $49.920 \pm 0.194\%$ top-1 accuracy, compared with $49.545 \pm 0.352\%$ for late fusion and $50.015 \pm 0.116\%$ for no-mask. Its gains over late fusion correspond to 96, 86, and 38 additional correct moves out of 19,546. Two game-bootstrap intervals exclude zero; the third includes zero (Table 4). The gain is consistent across the retained fits but modest. No-mask's mean advantage is 0.096 points; this closeness is descriptive, not an equivalence result with a prespecified margin.

Table 4: Paired SPECTRA–late-fusion top-1 comparison on 19,546 shared games. Intervals resample games within each fitted seed.

Seed	SPECTRA (%)	Late fusion (%)	S−LF (pp)	95% CI
42	49.795	49.304	+0.491	[+0.056, +0.921]
1337	49.821	49.381	+0.440	[+0.010, +0.870]
1842	50.143	49.949	+0.194	[-0.297, +0.665]

No-mask has lower mean NLL, 1.56415 versus 1.57490, with paired intervals favoring it in two seeds. SPECTRA has the lowest measured ECE: 0.01111, compared with 0.02136 for late fusion and 0.01457 for no-mask. This ranking is specific to the recorded sample and binning; its stability under recalibration or distribution shift has not been established.

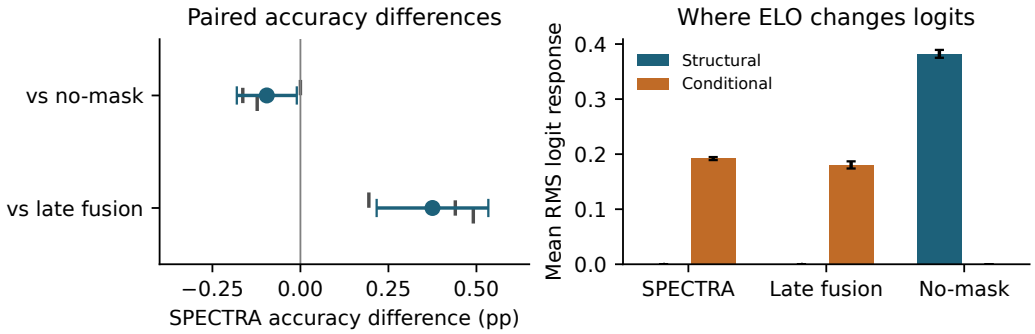


Figure 2: Accuracy differences and Elo pathways. Left: means and seed SDs of paired accuracy differences, with individual seed marks. Right: RMS logit response through the structural and conditional terms in Eq. 1, with error bars showing sample SD across three seed means. Response magnitude is not a quality score.

Figure 2 contrasts accuracy with computational dependence. Mean structural logit response is zero for both protected models and 0.382 for no-mask. Conditional response is 0.192 for SPECTRA, 0.180 for late fusion, and approximately 0.0001 for no-mask. No-mask instead uses Elo through its structural path. Similar accuracy can conceal different computational dependencies.

5.2 How do ratings affect predictions, and are the changes useful?

Keeping the board states fixed still allows the predicted move to change. We first examine where those changes occur, then test whether the observed rating improves prediction. Under SPECTRA's structural-margin ranking, its flip rate rises from 0.018% in Q1 to 24.304% in Q5. Late fusion rises from 0.175% to 16.234%, giving an interaction of 8.237 ± 0.297 points. No-mask's interaction relative to late fusion is 7.058 ± 0.841 points under the same ranking.

Table 5: Sensitivity of the SPECTRA–late-fusion flip interaction to the shared uncertainty ranking. Values are percentage points, mean ± seed SD. Both ranking models have ELO-independent structural policies. These additional analyses are exploratory.

Ranking model	Uncertainty measure	Q5–Q1 interaction
SPECTRA	Top-two margin	+8.237 ± 0.297
SPECTRA	Normalized entropy	-0.847 ± 0.348
Late fusion	Top-two margin	-12.457 ± 0.963
Late fusion	Normalized entropy	-2.981 ± 0.594

The exploratory symmetric check changes the interpretation (Figure 3). Late fusion's structural-margin ranking yields a SPECTRA–late-fusion interaction of -12.457 ± 0.963 points; normalized-entropy rankings also give negative interactions (Table 5). Shared indices make each panel internally

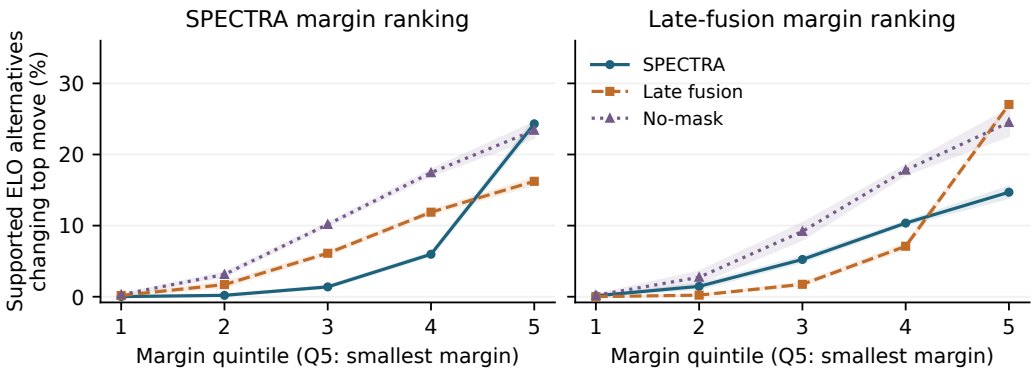


Figure 3: The frozen games stratified by two protected structural policies. Each panel uses a common ranking for all models; curves show seed means and bands seed SD. The apparent SPECTRA–late-fusion selectivity advantage changes direction with the ranking.

consistent. The ranking still depends on the model. In these sweeps, the larger response follows whichever model supplies the small-margin ranking.

Across supported alternatives, SPECTRA flips its top move in 6.375% of cases, late fusion in 7.221%, and no-mask in 10.887%. Mean JS divergences are 0.00357, 0.00415, and 0.01076. Lower sensitivity alone does not establish robustness: a prediction may need to change with context.

Neither the ranking-dependent response curves nor the overall flip rates establish predictive usefulness. We therefore compare the probability assigned to the recorded move under different ratings.

For local usefulness, we compare the recorded move's log-probability at the observed Elo with its value at the nearest supported alternative by embedding-index distance (Appendix F). Mean gains are 0.00064 nats for SPECTRA, 0.00103 for late fusion, and 0.00207 for no-mask. Paired intervals on all 19,546 games exclude zero for the likelihood gain in every fit. For SPECTRA, the top-1 gain averages only 0.0068 points, and each seed's interval includes zero. Appendix F reports the full intervals and margin-stratified analyses. The evidence supports a small local likelihood benefit; it does not establish accurate human behavioral counterfactuals.

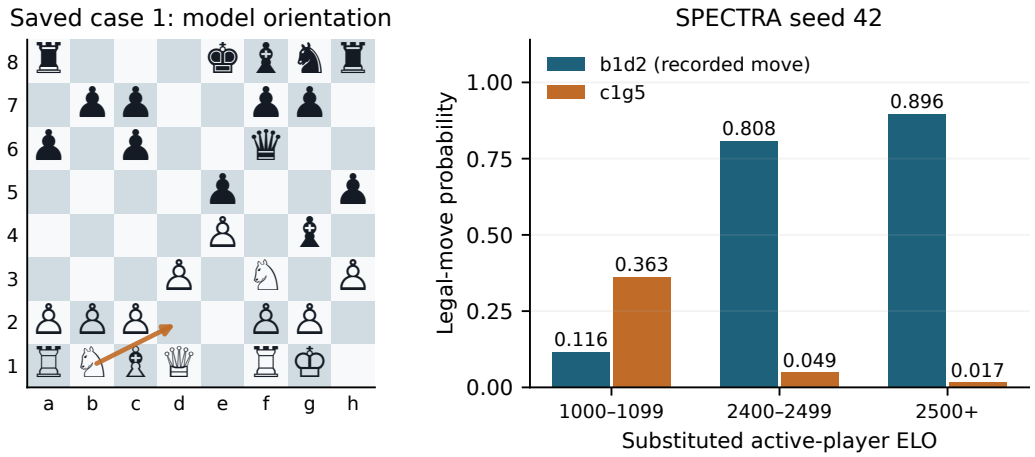


Figure 4: One board, several conditional predictions. The arrow marks the recorded move in normalized board orientation; bars show two legal moves at three supported active-player ratings, with the opponent fixed. This saved example was selected for a large policy shift, not as a typical outcome. It illustrates a permitted response while the structural states remain unchanged.

Figure 4 makes the distinction concrete. In this saved seed-42 example, the probability of the recorded move b1d2 falls from 0.896 at the observed rating bucket to 0.116 at 1000–1099, where c1g5 becomes the top prediction. The board computation is unchanged. The example illustrates conditional access to a common structural trajectory; the full-sample likelihood test, rather than this selected case, assesses predictive usefulness. Appendix F provides the selection rule and provenance.

5.3 How much computation can rating queries reuse?

The same fixed states also let us avoid repeating the board computation for every rating. The evaluator caches $\mathcal{K}_Z(X) = (\mathcal{T}_Z(X), \ell_Z)$, then computes each rating-specific output using the matching structural state at every layer. If C_Z is the cache-construction cost and C_U the cost per context, the idealized speedup for K contexts is

$$S_K = \frac{K(C_Z + C_U)}{C_Z + KC_U}.$$

Late fusion caches its board encoder's outputs and repeats only the side branch and head. SPECTRA repeats more computation because its side and prediction tokens read structure at each layer. No-mask can reuse board preprocessing, but its post-mixing structural states depend on Elo.

Table 6: Measured batch-64 cost for 26 ELO queries. Latencies are the mean ± SD of the three per-seed medians. Speedup is relative to the same architecture under ordinary execution.

Measurement	SPECTRA	Late fusion
Cached latency (ms)	222.3 ± 15.4	49.2 ± 4.4
Ordinary latency (ms)	1140.6 ± 122.0	1080.1 ± 115.8
Speedup (×)	5.12 ± 0.20	21.92 ± 0.40
Cache memory (MiB)	31.14	7.33

At batch 64 and 26 Elo queries, SPECTRA is $5.12 \pm 0.20\times$ faster than its ordinary execution; late fusion is $21.92 \pm 0.40\times$ faster than its own. Across the three comparisons, late fusion has 4.4–4.7 times lower cached latency and uses less cache memory (Table 6). The measured benefit is reuse within SPECTRA; late fusion remains faster for this workload.

The workload matters. With one context, SPECTRA's speedup is $0.961\times$, a small overhead; with 16 contexts it is $4.537\times$. Timing 26 embedding indices includes indices absent from evaluation support and does not validate behavior at those ratings. Table 14 gives the batch-size sweep. The benchmark fixes GPU, FP32 dtype, warm-ups, ordering, and chunking across the protected models. On 16 boards per seed, cached execution preserves raw-logit top-1 indices and top-5 sets; legal-policy agreement was not separately verified (Appendix H).

For repeated rating queries, SPECTRA's small predictive advantage must be weighed against late fusion's lower latency. An optimized no-mask preprocessing cache and broader numerical-equivalence tests remain untested.

5.4 Does preserving a chosen partition help prediction?

Does state protection help under other input partitions (Table 1)? All nine routed fits across Adult, CivilComments, and Waterbirds have zero measured structural drift; their no-mask partners have nonzero drift. Adult and CivilComments have similar average accuracy across routes and favor SPECTRA in mean worst-group accuracy (Table 7).

Adult's gain comes from three additional correct predictions in one fit on a 76-example group; the other pairs tie (Appendix I). CivilComments reverses direction in one seed. Neither result establishes a general fairness benefit.

On Waterbirds, no-mask leads by 4.22 points in average accuracy and 13.33 points in worst-group accuracy, with the same direction in all seeds. Both streams originate in the image. Removing feedback may discard useful interactions, but the comparison does not uniquely identify that explanation over encoder or optimization effects. The evaluated routed models perform worse under this partition even though the invariant holds.

Table 7: Cross-domain accuracy (%; mean ± seed SD). The route holds in every domain; predictive costs vary across the evaluated task partitions.

Domain	SPECTRA avg.	No-mask avg.	SPECTRA worst-group	No-mask worst-group
Adult	80.19 ± 0.34	79.93 ± 1.27	70.18 ± 2.01	68.86 ± 3.31
CivilComments	87.78 ± 0.24	88.10 ± 0.46	61.91 ± 1.75	57.22 ± 5.62
Waterbirds	88.31 ± 1.28	92.53 ± 1.09	72.80 ± 7.49	86.13 ± 2.22

6 Discussion

State protection and predictive usefulness are separate questions. The local Elo test finds a small likelihood benefit without a reliable top-1 gain for SPECTRA. The exploratory ranking reversal prevents interpreting the original curve as an architecture-independent selectivity advantage.

Reuse depends on what each model must recompute. Late fusion repeats its side branch and head; SPECTRA also updates side and prediction tokens at each layer using saved board states. The fitted no-mask models change their board states with the rating. These comparisons do not isolate the side token's contribution to accuracy.

Partition choice matters too. Ratings can vary while the board stays fixed; Waterbirds derives both streams from one image. Its accuracy loss shows that the invariant can hold without improving prediction. The comparison does not uniquely explain the loss.

7 Scope and limitations

The guarantee holds for a declared side-input path at fixed parameters. Structural inputs can contain proxies, pretrained encoders can mix attributes, and side-dependent objectives can affect learned parameters. Forward noninterference permits these effects; it does not guarantee statistical independence, human counterfactuals, privacy, or fairness. Fine-tuning must preserve the mask and interfaces; new token-mixing operations require a new dependency argument.

The experiments compare complete architectures, not an isolated connection. An unperformed side-read ablation would block only the side token's structural reads, retaining the prediction token's access, other mask entries, heads, objectives, and training procedure. Three-way non-chess studies, more seeds, and an independent uncertainty signal would test generalization (Appendix L). The additional ranking and local-utility analyses reuse the frozen holdout; they are exploratory, not independent replication. Checkpoints will be released after acceptance and can be supplied during review on request.

8 Conclusion

SPECTRA lets rating-dependent side and prediction streams read successive board states while keeping those states unchanged under rating substitution. The guarantee concerns a declared input path at fixed parameters. In the evaluated chess systems, SPECTRA has a modest accuracy advantage over late fusion and benefits from reusing board states, but late fusion remains faster. Rating sensitivity is model-relative, and the other partitions show that state protection can coexist with predictive loss. The study establishes this construction's dependency property and characterizes the evaluated systems; it does not isolate the predictive contribution of the side token's intermediate reads.

ETHICS STATEMENT

Adult and CivilComments contain sensitive or identity-related attributes, and Waterbirds uses group labels for evaluation. We use these established research benchmarks to study computational dependencies. The invariant does not establish that a deployed system is fair, private, or safe. Deployment would require domain-specific assessment of proxies, data use, and downstream effects. Chess predictions describe aggregate move choice rather than personal traits.

REPRODUCIBILITY STATEMENT

The supplementary materials contain source code, checkpoint identity manifests, frozen evaluation identifiers, per-game policy arrays, and machine-readable results. The v2 analysis script recomputes the additional uncertainty rankings, local Elo effects, paired intervals, and figures from stored arrays without checkpoint inference. Checkpoints will be made available after acceptance and can be provided during review on request. The cross-domain command summarizes frozen audit results; its exact historical inference runner is not supplied. The appendix details objectives, checkpoint provenance, and reproduction limits.

AI USE STATEMENT

Generative AI tools assisted with literature discovery and synthesis, conceptual and mathematical critique, code generation and review, diagnostic checks, result interpretation, figure implementation, and editorial revision. This revision also used AI assistance to inspect the released training source and recompute analyses from stored predictions. AI output was not used as an independent source of evidence. The authors retain responsibility for the final claims, cited sources, experiments, and released code.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. Dataset.

Jason Carlson. Matilda: Engine-agnostic search with human policy guidance. *arXiv preprint arXiv:2606.25176*, 2026. URL https://arxiv.org/abs/2606.25176v3. Version 3.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Joseph A. Goguen and José Meseguer. Security policies and security models. In *1982 IEEE Symposium on Security and Privacy*, pp. 11–20. IEEE, 1982. doi: 10.1109/SP.1982.10014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang.

WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/265bee74aee86df77e8e36d25e786ab5-Abstract-Conference.html.

Chentao Li and Han Guo. Hasse diagrams for attention: A partial order framework for designing transformer masks. *arXiv preprint arXiv:2606.09951*, 2026.

Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4249–4256, 2021a. doi: 10.1609/aaai.v35i5.16549.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6781–6792. PMLR, 2021b. URL https://proceedings.mlr.press/v139/liu21f.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman AI with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

Daniel Monroe, George Eilender, Philip Chalmers, Zhenwei Tang, and Ashton Anderson. Chessformer: A unified architecture for chess modeling. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=2ltBRzEHyd.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28448–28467. PMLR, 2023. URL https://proceedings.mlr.press/v202/qiu23c.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations (ICLR)*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sherman Siu and Lesley Istead. UniMaia: Steering chess policies with language for human-like play. *arXiv preprint arXiv:2605.27767*, 2026. URL https://arxiv.org/abs/2605.27767.

Zhenwei Tang, Difan Jiao, Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, and Ashton Anderson. Maia-2: A unified model for human-AI alignment in chess. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tarun Kumar S. Otter: A time-aware, history-conditioned human chess AI. *arXiv preprint arXiv:2608.05206*, 2026. URL https://arxiv.org/abs/2608.05206.

Trishita Tiwari, Suchin Gururangan, Chuan Guo, Weizhe Hua, Sanjay Kariyappa, Udit Gupta, Wenjie Xiong, Kiwan Maeng, Hsien-Hsin S. Lee, and G. Edward Suh. Information flow control in machine learning through modular model architecture. In *33rd USENIX Security Symposium*, pp. 6921–6938. USENIX Association, 2024.

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 (NeurIPS)*, 2017.

Shuhan Wang, Bin Shen, Xu Min, Yong He, Xiaolu Zhang, Liang Zhang, Jun Zhou, and Linjian Mo. Aligned side information fusion method for sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 112–120. ACM, 2024. doi: 10.1145/3589335.3648308.

Yiming Zhang, Athul Jacob, Vivian Lai, Daniel Fried, and Daphne Ippolito. Human-aligned chess with a bit of search. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0ef1afa0daa888d695dcd5e9513bafa3-Abstract-Conference.html.

A Proof and extensions

Write the layer-ℓ state as $H^{(\ell)} = [Z^{(\ell)}, U^{(\ell)}, P^{(\ell)}]$. A structural query sees keys and values only from $Z^{(\ell)}$, so its multi-head attention output depends on $Z^{(\ell)}$ and the fixed parameters. Residual addition, layer normalization, and the position-wise feed-forward network act on that token alone; none introduces a cross-token edge. Since $Z^{(0)} = f_X(X)$, induction over $\ell = 0, \ldots, L-1$ gives $Z^{(L)} = g(X)$. Replacing $U^{(0)}$ or $P^{(0)}$ cannot change $Z^{(L)}$, and the corresponding derivative is zero wherever it is defined.

The argument fixes parameters and assumes deterministic inference, or the same realization of token-local randomness in both executions. The structural tokenizer and attention mask must be independent of the side input. The mask is applied before softmax, and operations outside attention must not mix tokens. Layer normalization here normalizes features within a token. A side-dependent mask, tokenizer, or cross-token normalization would require a different argument. Side-dependent losses may affect fitted structural parameters during training; the proposition concerns forward dependence after those parameters are fixed. The released unit test checks the concrete PyTorch convention directly: rows are queries, columns are keys and values, and a Boolean `True` entry is blocked. It does not infer mask orientation from a diagram.

The proof also covers multiple side inputs. Let the side types be $U_1, \ldots, U_k$ and specify a directed visibility relation among token types. A structural type with no incoming path from a side type remains invariant to that input. Side tokens may be isolated, ordered, or mutually visible. The contract reduces to graph reachability; choosing a useful graph remains an empirical design problem.

B Architectures and token definitions

Table 8 lists the dimensions; the forward-path differences below explain why equal fitted capacity does not make the controls identical.

The mask has 129 blocked entries: 64 structural queries cannot read the side-information key, another 64 cannot read the prediction key, and the side-information query cannot read the prediction key. The prediction query sees all 66 tokens. Late fusion uses a 66-token transformer: 64 board tokens, an ELO-independent summary slot initialized to zero, and a learned MOVE token. Its full-attention trunk sees no ELO. The side branch encodes the two rating buckets separately and joins the board-only MOVE representation at the adjustment MLP. The gate is therefore board-only in late fusion, whereas it can depend on ELO through $P^{(L)}$ in Spectra and no-mask. The summary slot can acquire board information during attention even though its initial value is zero. All three conditions use the same 4,672-action output convention.

The gate network G in Eq. 1 has dimensions $384 \to 256 \to 1$, with GELU between its linear layers. For Spectra and no-mask, the conditional adjustment is computed from $[U^{(L)}; P^{(L)}]$, with the

Table 8: Chess architecture shared by SPECTRA and no-mask. Late fusion matches fitted capacity and trunk dimensions, but uses different transformer inputs and head conditioning.

Input / component	Value	Input / component	Value
Board planes	$19 \times 8 \times 8$	Structural tokens	64
Side / prediction tokens	1 / 1	Sequence length	66
Transformer layers	4	Width / heads	384 / 16
Feed-forward width	1,536	Dropout	0.1
Backbone channels / blocks	256 / 12	Move actions	4,672
ELO buckets	26	Fitted parameters	24,713,267
SPECTRA blocked reads	129	No-mask blocked reads	0

gate and clamped scalar specified in Eq. 1. No-mask also permits prediction-token feedback to the other tokens. Thus late fusion is a protected architectural baseline, and no-mask removes the whole routing mask; neither is a freshly trained single-edge ablation. A strict control for the side stream's access to structure would retain SPECTRA's tokens, gate, head, objectives, and other mask entries, and block only the 64 reads from structural keys to the side query. The prediction token would still read structural states, so this control would not remove all intermediate structural access. It has not been trained for this study.

The non-chess encoders use DistilBERT (Sanh et al., 2019) for CivilComments and ResNet (He et al., 2016) for Waterbirds. CivilComments indicators are {male, female, LGBTQ, Christian, Muslim, other religion, Black, White}. They encode mentions in the comment, not verified properties of its author. Waterbirds differs from the other tasks because A and X come from the same image. That experiment tests the usefulness of an image partition, not the routing of a separately observed metadata field.

C TRAINING AND EVALUATION COMPARISONS

Table 9 summarizes the optimization settings. The following schedules specify how the chess objectives enter the retained training runs.

Table 9: Optimization settings reported for the retained runs. The chess objective and its scheduled coefficients are specified below; domain implementations are included in the training source.

Setting	Chess	Adult	CivilComments	Waterbirds
Batch size	512 (4,096 effective)	1,024	64	256
Backbone learning rate	2×10^{-4}	5×10^{-5}	3×10^{-5}	3×10^{-5}
Head / side learning rate	see text	10^{-4}	10^{-4}	2×10^{-4}
Weight decay	0.01	0.05	0.10	0.20
Warm-up steps	2,000	500	300	300
Maximum recorded steps	300,000	25,000	60,000	8,000
Gradient clip	2.0	n/a	n/a	n/a

Chess uses AdamW with separate backbone and side-parameter learning-rate schedules. After a 2,000-step warm-up, the backbone rate is 2×10^{-4} until step 50,000, then follows cosine decay with floor 4×10^{-5}. At step 190,000 it restarts at 1.2×10^{-4} over an 80,000-step cycle.

The side group ramps to 10^{-3} over 2,000 steps after step 50,000 and then decays; at the restart it ramps from 2.1×10^{-4} to 7×10^{-4} over 2,000 steps before decaying. The rating embeddings, their projection, skill head, adjustment MLP, and learned scale use the side schedule; the gate uses the backbone schedule. Scheduler boundaries are converted to optimizer-update units using the accumulation factor. The source configures stochastic weight averaging (SWA) from recorded step 220,000 with a 5,000-step interval. However, its update test runs before the global-step increment, at steps congruent to 7 modulo 8, and cannot meet that interval with eight-step accumulation. The six hash-verified retained SPECTRA/no-mask exports each record `swa_n=1`: they contain one SWA snapshot, not a multi-snapshot average. Their export metadata records step 300,000; this does not

establish that the saved snapshot contains the last optimizer update. The reported scores evaluate these exact checkpoint files. Late-fusion records identify SWA exports, but their snapshot counts were not independently checked here.

During the first 50,000 recorded steps, the policy uses ℓ_Z and suppresses the explicit conditional adjustment. In the ordinary SPECTRA/no-mask forward path, the ELO token is nevertheless constructed and passed through the transformer before this output selection. In late fusion, the warm-up branch zeroes its separate ELO representation. These are architecture-specific warm-up paths; the common flag name `force_zero_u` does not imply that the literal token is replaced in every implementation. Auxiliary heads and adjustment-magnitude regularization remain active in both phases. In particular, the no-mask structural policy can already depend on ELO during warm-up.

After step 50,000, `player_scale` ramps the explicit adjustment from zero to full strength over 5,000 steps in the ordinary forward path. The reported schedules use the trainer's recorded global step. Policy cross-entropy has weight 1 and smooths over legal moves with coefficient 0.01, reduced to 0.005 at step 200,000. Table 10 gives the other active terms. No schedule ablation establishes that this two-phase procedure is necessary.

Table 10: Chess objective in the released training implementation. All ramps are linear in recorded step; k denotes 1,000 steps. Auxiliary prediction and adjustment regularization are active during warm-up.

Term	Coefficient / schedule	Definition
InfoNCE	0 before 50k; 0.10 to 0.30 at 50–90k; to 0.16 at 90–110k	Contrastive side representations from paired boards, with rating-bucket matches as positives.
Skill prediction	0 before 50k; 0.05 to 0.20 at 50–90k	Cross-entropy for the 26 active-player rating indices from the side representation.
Auxiliary prediction	0.02; decays to 0.01 at 70–90k	Mean of five MOVE-token losses: moved/captured piece, origin, destination, and check indicator.
Residual correction	0 to 0.08 at 55–75k	Legal-move cross-entropy of detached structural logits plus effective adjustment, weighted by $1 - p_Z(y \mid X)$.
Ranking correction	0 to 0.50 at 50–54k	0.8 times top-16 structural-candidate ranking loss plus 0.2 times full legal-move correction loss.
Adjustment penalty	10^{-5}, both phases	Mean squared raw adjustment d.
Gate entropy	-0.005, phase 2	Mean binary entropy of the effective gate, evaluated in FP32 with $\epsilon = 10^{-6}$.

For all three chess architectures, phase 2 also adds $-0.005 H_\epsilon(g)$, where

$$H_\epsilon(g) = -\operatorname{mean}\big[g \log(g+\epsilon) + (1-g)\log(1-g+\epsilon)\big], \qquad \epsilon = 10^{-6}.$$

Here g is the effective gate after the floor transformation, broadcast across move actions. The code evaluates this term in FP32; its negative coefficient encourages nonsaturated gates. It is inactive during phase 1.

The residual and top-16 ranking terms use the gated, ramped adjustment, before multiplication by the learned scalar s. The ranking fallback uses the raw adjustment. Structural logits are detached in these correction losses; this blocks their direct logit path but does not freeze the structural encoder, which the side and prediction streams can read. Ranking cross-entropy is averaged over examples whose recorded move lies in the structural top 16; the full-action fallback uses all examples. The ranking temperature is 1. InfoNCE normalizes side representations and uses a bucket-dependent temperature $0.06 + 0.08(1 - b/25)$, where b is the active-player bucket. Pairs match rating buckets rather than individual players. These objectives can encourage corrections on structurally uncertain examples, so the observed selectivity cannot be assigned to routing alone.

All nine chess checkpoints use the same underlying 13.58-million-game Lichess Rapid corpus and preprocessing/filtering rules. Historical runs used 28-shard and 200-file layouts of that corpus. Exact train/validation assignments and streamed example orders are not asserted to be identical across runs. Seed labels identify the retained fits; pairing refers to their evaluation on the same frozen 19,546 games, with identical labels, game order, and ELO assignments. These comparisons characterize complete trained architectures rather than an isolated mask-only intervention. Corpus identity is confirmed by the training provenance; historical partition and stream identity have not been independently replayed.

Table 11: Top-1 accuracy by seed label. Each row uses the same frozen evaluation tensor and game order across all three models; identical training streams are not asserted.

Seed	SPECTRA (%)	Late fusion (%)	No-mask (%)	S−LF (pp)	NM−S (pp)
42	49.795	49.304	49.918	+0.491	+0.123
1337	49.821	49.381	49.985	+0.440	+0.164
1842	50.143	49.949	50.143	+0.194	+0.000

Table 11 reports the individual fitted models underlying the aggregate chess comparison.

The fitted checkpoint state contains 24,713,267 parameters in every condition and 236 state keys. The general chess audit loader can instantiate an additional unused diagnostic policy head, producing a larger in-memory module count; that helper is absent from the serialized checkpoints and is not counted as fitted capacity. Checkpoint hashes and state-key counts are validated before evaluation.

D MAIA EVALUATION CONSTRUCTION

The Maia test file contains 11,288,275 rows from 159,921 games. We retain Rapid rows with a FEN, a legal observed UCI move, ply at least 10, a missing clock or at least 30 seconds remaining, and both active and opponent ELO. These rules leave 1,091,559 eligible rows from 19,546 games. One row per game is selected by hashing the evaluator seed, game identifier, and ply, then taking the minimum unsigned priority. The evaluator subsequently checks FEN parsing, action mapping, legal-mask membership, ground-truth legality, and active-color consistency.

The frozen evaluation tensor has SHA-256 prefix `225abbee973ddaf6`; the manifest records the complete digest. All nine architecture/seed evaluations use identical labels, public game hashes, ELO buckets, and game order. A read-only audit compares every Maia game identifier with all 13,583,266 training rows. It finds no overlap, including among the 19,546 selected games.

ELO bucket 0 contains ratings below 1000. Buckets 1 through 15 are the 100-point intervals 1000–1099 through 2400–2499; bucket 25 contains ratings of 2500 or above. The intermediate implementation reserves 26 indices, although the selected Maia sample supports only buckets $\{1, \ldots, 15, 25\}$. Behavioral summaries average only over supported alternatives unless a table explicitly states otherwise. "Supported" means observed at least once, not precisely estimated. The selected-game counts for buckets 1–15 and 25 are, respectively, 1,657, 1,792, 1,912, 1,977, 2,125, 2,131, 2,157, 1,782, 1,611, 1,193, 690, 370, 105, 23, 17, and 4 (total 19,546).

E STATISTICAL AND INTERVENTION PROTOCOLS

Each primary chess example comes from a different game. The evaluator stores correctness, target log-probabilities, confidence, entropy, move flips, JS divergence, and pathway responses. Aggregate metrics include NLL, ECE, and Brier score. The independent route audit uses 10,000 percentile game resamples and the three-way policy suite uses 5,000. The additional full-sample local-utility intervals use 5,000 paired game resamples, as do the new within-quintile utility intervals. The new ranking-interaction table reports seed SD rather than a game-bootstrap interval. Across seeds, we report the mean and sample standard deviation of three differences between fits sharing seed labels. A seed-level t-test has only two degrees of freedom and is not used as primary evidence. Because the retained cross-domain results store aggregate metrics, those sections report per-seed pairs and seed SD without claiming an example-paired test.

The literal intervention replaces $U^{(0)}$ after its input encoder and before token concatenation. Direct substitution compares two supported ELO tokens for the same board. For the autograd check, we draw a Gaussian tensor with the shape of $Z^{(L)}$, compute the mean-scaled random projection $\langle Z^{(L)}, R\rangle/|R|$, and differentiate it with respect to $U^{(0)}$. The finite-difference check draws a Gaussian direction for each of four examples, normalizes it to unit ℓ_2 norm, and evaluates

$$\frac{Z^{(L)}(U^{(0)} + 10^{-3}d) - Z^{(L)}(U^{(0)} - 10^{-3}d)}{2 \times 10^{-3}}.$$

Both tests differentiate the post-transformer structural tokens, not a composite policy head.

Late fusion is evaluated with the same substitution arrays. Its maximum structural-token drift is also exactly zero for all seeds; its conditional head remains responsive. The no-mask condition provides a negative control in each intervention family. The v2 analysis checks exact equality of game hashes, labels, and rating assignments across all nine policy arrays before computing summaries. Games within each seed use the same bootstrap indices across architectures and paired ELO conditions. The script records all random seeds. These additional analyses reuse the frozen evaluation sample and are exploratory; their pointwise 95% intervals have no multiplicity adjustment. They do not constitute an independent replication or an estimate of uncertainty over future training runs.

F ELO SUPPORT AND LOCAL USEFULNESS

For each selected game, the evaluator scores the observed legal move under all 26 ELO inputs. The main target-ELO effect compares the log-probability at the observed bucket b with alternatives minimizing $|b' - b|$ over the supported embedding indices $b' \neq b$. Equally near alternatives are averaged. This is index distance, not numerical Elo distance or adjacency of rating intervals: indices 16–24 are unused, so bucket 15 selects bucket 14 rather than tying buckets 14 and 25. The nearest-supported terminology in the tables refers to this implemented rule. An adjacent-rating-interval sensitivity analysis has not been performed. This comparison avoids treating an unobserved rating range as an ordinary test condition. A secondary frequency-weighted comparison samples from supported alternatives after excluding the observed bucket; the full 25-mismatch mean is retained only as a sensitivity analysis.

Table 12: Nearest-supported target-ELO consistency on $n = 4{,}096$ stratified distinct games per checkpoint. All six point estimates favor the observed bucket; intervals resample games.

Condition	Seed	Δ log-prob.	95% CI
SPECTRA	42	+0.00079	[-0.00002, +0.00161]
SPECTRA	1337	+0.00054	[-0.00028, +0.00135]
SPECTRA	1842	+0.00098	[+0.00018, +0.00181]
No-mask	42	+0.00176	[+0.00086, +0.00271]
No-mask	1337	+0.00128	[+0.00049, +0.00208]
No-mask	1842	+0.00242	[+0.00142, +0.00342]

Table 12 comes from the earlier six-checkpoint SPECTRA/no-mask audit on 4,096 stratified games per checkpoint. Two of its three SPECTRA likelihood intervals include zero. Table 13 reports a separate reanalysis of the complete nine-checkpoint sweep on 19,546 games. Its mean observed-versus-nearest likelihood gains are 0.00064 nats for SPECTRA, 0.00103 for late fusion, and 0.00207 for no-mask; each fit's paired likelihood interval excludes zero. All three SPECTRA top-1 intervals include zero. The distinction is between a small improvement in the recorded move's probability and a reliable gain in discrete move prediction.

Table 13: Observed ELO minus nearest-supported alternative on all 19,546 games. Each interval is a 5,000-resample paired game percentile interval conditional on one fit. Nats measure the log-probability gain for the recorded move.

Model	Seed	Log-likelihood gain [95% CI]	Top-1 gain, pp [95% CI]
SPECTRA	42	+0.00070 [+0.00033, +0.00105]	-0.054 [-0.153, +0.046]
SPECTRA	1337	+0.00048 [+0.00011, +0.00085]	+0.023 [-0.077, +0.120]
SPECTRA	1842	+0.00073 [+0.00037, +0.00109]	+0.051 [-0.054, +0.159]
Late fusion	42	+0.00107 [+0.00070, +0.00145]	-0.072 [-0.179, +0.036]
Late fusion	1337	+0.00092 [+0.00050, +0.00133]	-0.069 [-0.179, +0.036]
Late fusion	1842	+0.00108 [+0.00068, +0.00149]	-0.123 [-0.233, -0.013]
No-mask	42	+0.00193 [+0.00149, +0.00236]	-0.074 [-0.205, +0.054]
No-mask	1337	+0.00191 [+0.00150, +0.00230]	+0.125 [+0.010, +0.240]
No-mask	1842	+0.00239 [+0.00194, +0.00285]	+0.128 [+0.002, +0.253]

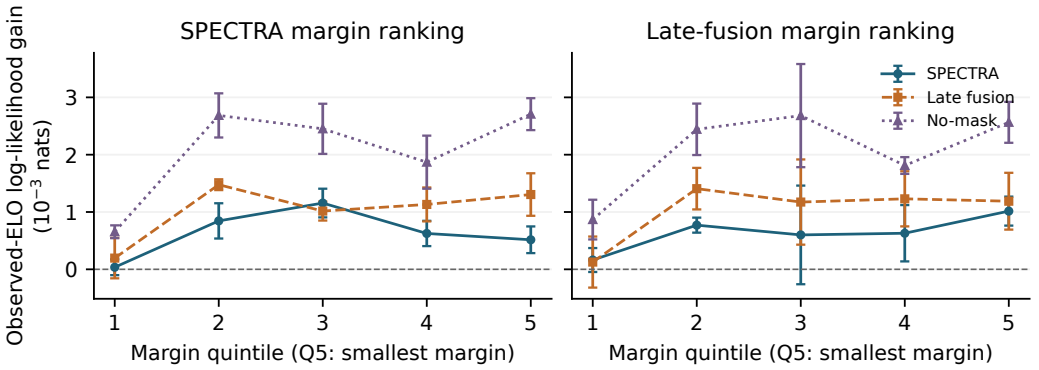


Figure 5: Observed-versus-nearest-supported-ELO log-likelihood gain, stratified by each protected model's structural margin. Points are three-seed means and error bars are seed SD, not game-bootstrap intervals. The v2 analysis records the corresponding paired intervals and top-1 effects for every seed, model, and quintile. These are exploratory subdivisions of the same frozen sample.

Figure 5 separates local predictive usefulness from the larger flip rates in Figure 3. A model can change its top move more often near a small-margin boundary without improving the probability of the observed human move. Moreover, a game supplies a move at the observed rating only: the alternative-rating predictions are not validated against human moves from those counterfactual contexts.

A saved board example. Figure 4 uses the first case in the earlier evaluator's frozen illustrative set, selected for a large cross-seed policy shift. It is an output-selected example, not a representative estimate or a new experiment. The public game hash is `f6f290f9cda6c95bb905f56c`; the active and opposing ratings were 2508 and 2534. The model-oriented board and opponent bucket remain fixed. Under the seed-42 SPECTRA checkpoint, the recorded move b1d2 receives probability 0.896 at the observed bucket, 0.808 at 2400–2499, and 0.116 at 1000–1099. At the last setting, c1g5 becomes the top move with probability 0.363. These are model responses to input substitution, not observed behavior of other rating groups.

G AMBIGUITY, CALIBRATION, AND REPRESENTATIONAL PATHWAYS

Structural ambiguity is defined from SPECTRA's board-only top-two legal-move probability margin, $m_Z(x) = p_Z^{(1)}(x) - p_Z^{(2)}(x)$, where $p_Z = \mathrm{softmax}(\ell_Z)$ over legal moves. Games are sorted in descending margin and split into approximately equal-sized quintiles, with Q1 least ambiguous and Q5 most ambiguous. Within a seed, all three architectures use the same indices. This gives a common comparison set for that ranking. We repeat the construction using late fusion's protected structural policy, and use stable sorting to resolve ties. Each ranking has quintile sizes 3,910, 3,909, 3,909, 3,909, and 3,909. Normalized-entropy rankings sort $\widetilde{H}_Z(x) = H(p_Z(x))/\log(\max\{2, |\mathcal{L}(x)|\})$ in ascending order, where $\mathcal{L}(x)$ is the legal-move set and $H(p) = -\sum_a p_a \log p_a$. Q5 has the greatest normalized entropy. Normalizing by legal-move count can change the ordering across boards.

The endpoint flip rate for SPECTRA rises from 0.018% in Q1 to 24.304% in Q5. Late fusion rises from 0.175% to 16.234%, and no-mask from 0.292% to 23.399%. Both protected architectures have context-invariant structural policies, which makes their rankings available before a sweep. Under late fusion's ranking, the SPECTRA–late-fusion Q5–Q1 interaction changes sign to -12.457 ± 0.963 points. Under SPECTRA and late-fusion normalized-entropy rankings it is -0.847 ± 0.348 and -2.981 ± 0.594 points, respectively. Thus a strong endpoint interaction under one model's margin does not define an architecture-independent ordering. Head differences and correction objectives also prevent single-edge attribution.

Native NLL is 1.57490, 1.58421, and 1.56415 for SPECTRA, late fusion, and no-mask. Fixed-width 15-bin top-label ECE is 0.01111, 0.02136, and 0.01457. SPECTRA has the lowest measured ECE in this evaluation, although these measurements do not establish calibration under distribution shift.

Pathway response is the root-mean-square change over legal-move logits relative to the observed-ELO context, averaged over supported alternative ELOs and games. Structural response applies this statistic to ℓ_Z; conditional response applies it to $\ell - \ell_Z$. Mean supported structural response is 0 for SPECTRA and late fusion and 0.382 for no-mask. Conditional response is 0.192 for SPECTRA, 0.180 for late fusion, and approximately 0.0001 for no-mask. For SPECTRA, the measured dependence lies in the permitted conditional term. In no-mask, a near-zero explicit adjustment can coexist with substantial ELO dependence in the structural logits.

H MEASURED STRUCTURAL REUSE

Caching follows the layerwise computation implied by the route. For each board, the evaluator stores

$$\mathcal{K}_Z(X) = \big(Z^{(0)}, \ldots, Z^{(L)}, \ell_Z\big),$$

where the sequence contains the structural input required by every routed layer, the final structural state, and the structural policy logits ℓ_Z. Each context-specific pass reads the matching cached structural state at each layer and computes $\hat{Y}(A) = q_{\mathrm{cache}}(A, \mathcal{K}_Z(X))$. Late fusion instead caches its board encoder's outputs and reruns only the ELO branch and policy head.

Before timing, we compare raw policy logits from cached and ordinary execution on 16 positions under all 26 contexts for each seed. Both protected architectures attain 100% raw-logit top-1-index and top-5-set agreement.

These checks precede legal-move masking and do not establish agreement of legal top moves, probabilities, or target log-likelihoods, including near ties. No trained-checkpoint legal-policy validation is added in this revision. The maximum absolute FP32 logit differences for SPECTRA are 0.01812, 0.00965, and 0.00940 in seeds 42, 1337, and 1842. These arise between the cached and ordinary implementations; the tested raw-logit top-1-index and top-5-set agreement does not imply bitwise logit equality or agreement on every possible input.

The timing benchmark holds GPU, dtype, batch ordering, 26-policy sequence, and chunking fixed across both architectures. It uses FP32 inference with TF32 disabled, five warm-ups, 30 timed calls, and CUDA synchronization before and after each call. Each seed contributes a median latency; reported SD is across these three medians. Cache memory counts the stored tensors, not total peak device memory. Late fusion is faster because strict tower separation leaves less conditional computation to rerun. SPECTRA reruns both side and prediction streams, including their reads of intermediate structural states. It attains a five-fold batch-64 speedup relative to its own ordinary execution; the absolute cached time remains higher than late fusion's. The no-mask reference is ordinary full execution; an optimized implementation may cache preprocessing and initial tokens, but not structural states after they mix with ELO.

Table 14: Model-execution speedup for 26 ELO settings, relative to each architecture's own ordinary execution (mean $\pm$ seed SD). SPECTRA and late fusion use measured reuse implementations. No-mask is its ordinary-execution reference (1.00); reusable preprocessing was not optimized. These ratios do not compare absolute latency across architectures.

Batch	SPECTRA	Late fusion	No-mask
1	2.70 ± 0.24	3.85 ± 0.40	1.00 (full run)
4	4.31 ± 0.28	11.50 ± 0.71	1.00 (full run)
16	4.99 ± 0.30	18.39 ± 0.90	1.00 (full run)
64	5.12 ± 0.20	21.92 ± 0.40	1.00 (full run)

I CROSS-DOMAIN RUNS AND SEED IDENTITIES

Table 15 shows the seed-level differences behind the cross-domain averages, including reversals and the consistent Waterbirds loss.

Table 15: Per-seed no-mask minus SPECTRA differences. The listed seeds are the actual retained run identities; they are not relabeled to match the chess seeds. Positive values favor no-mask.

Domain	Seed	Average accuracy (pp)	Worst-group accuracy (pp)
Adult	1254662429	-1.597	+0.000
Adult	2001813243	+0.172	-3.947
Adult	112651974	+0.633	+0.000
CivilComments	42	+0.854	-4.416
CivilComments	123	+0.519	-12.162
CivilComments	456	-0.403	+2.510
Waterbirds	42	+4.039	+12.150
Waterbirds	123	+4.453	+20.216
Waterbirds	456	+4.177	+7.620

Worst-group denominators are reported in Table 16. Adult's minimum is group 4 (income label 1, sex code 0, non-White race) in every fit. In seed 2001813243, SPECTRA scores 53/76 and no-mask 50/76; the other two pairs tie. The 1.316-point mean advantage therefore reflects three additional correct predictions in one retained fit. CivilComments minima can refer to different overlapping identity–label groups across fits, rather than the same subgroup.

Table 16: Worst-group identities and denominators in the retained cross-domain audits. Each cell gives group / number of examples, not the number correct.

Domain	Seed	SPECTRA group / count	No-mask group / count
Adult	1254662429	4 / 76	4 / 76
Adult	2001813243	4 / 76	4 / 76
Adult	112651974	4 / 76	4 / 76
CivilComments	42	(5,1) / 74	(2,1) / 142
CivilComments	123	(5,1) / 74	(5,1) / 74
CivilComments	456	(2,0) / 423	(7,0) / 657
Waterbirds	42	2 / 642	2 / 642
Waterbirds	123	2 / 642	1 / 2,255
Waterbirds	456	1 / 2,255	2 / 642

Adult encodes groups as $4y + 2\,\mathrm{sex} + \mathbb{1}[\mathrm{White}]$. CivilComments pairs are (identity index, toxicity label): indices 2, 5, and 7 denote LGBTQ, other religion, and White mentions, respectively; comments may belong to multiple groups. Waterbirds uses $2y + \mathrm{background}$, with binary bird and background labels.

Adult evaluates 16,281 examples per checkpoint. CivilComments uses the same 16,384-example subset for each routed/no-mask evaluation comparison, while its separate mechanism audit covers 133,782 examples. Waterbirds evaluates all 5,794 examples. Within every domain and seed, routed and no-mask models have equal parameter counts. Routed runs have exact zero U-to-Z change, and all no-mask partners have a nonzero change.

J COMPLETE MASK AND ADDITIONAL RESULT TABLES

Figure 6 expands the typed relation to all chess-token pairs. Tables 17 and 18 give the retained SPECTRA/no-mask accuracy and likelihood comparisons with paired intervals.

Table 17: Paired comparison on $n = 19{,}546$ Maia test games, with one position per game. Differences are no-mask minus SPECTRA; intervals resample games.

Seed	SPECTRA top-1 (%)	No-mask top-1 (%)	Difference (pp)	95% CI
42	49.795	49.918	+0.123	[-0.322, +0.583]
1337	49.821	49.985	+0.164	[-0.292, +0.619]
1842	50.143	50.143	+0.000	[-0.466, +0.496]

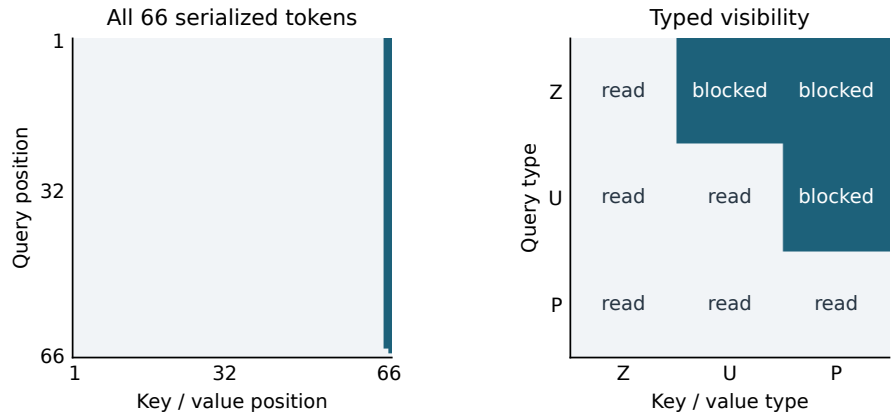


Figure 6: Full 66×66 chess mask and its typed summary. Dark entries are blocked in both panels; rows are queries and columns are keys and values.

Table 18: Paired negative log-likelihood comparison on the Maia test sample. Differences are no-mask minus SPECTRA; negative values favor no-mask. Intervals resample games.

Seed	SPECTRA NLL	No-mask NLL	Difference	95% CI
42	1.57666	1.56057	-0.01608	[-0.02175, -0.01050]
1337	1.58267	1.56487	-0.01780	[-0.02381, -0.01184]
1842	1.56535	1.56701	+0.00165	[-0.00450, +0.00808]

Table 19: Literal pre-transformer side-input interventions on $n = 19{,}546$ distinct games. Native–zero is native top-1 minus literal-zero top-1, so positive values favor the native side input. The ordinary prediction head remains active; the structural-only head is not used here.

Condition	Seed	Native–zero top-1 (pp)	Swap change (%)	max $\lvert\Delta Z\rvert$
SPECTRA	42	+0.240	1.770	0
SPECTRA	1337	+0.251	1.683	0
SPECTRA	1842	+0.046	1.852	0
No-mask	42	+0.292	2.420	58.6564
No-mask	1337	+0.266	2.174	109.111
No-mask	1842	+0.686	2.589	93.0627

In Table 19, "swap change" is the top-1 prediction-change rate after rebuilding $U^{(0)}$ from the nearest supported alternative player-ELO bucket while holding the board and opponent-ELO bucket fixed. Native–zero compares the ordinary observed-ELO representation with $U^{(0)} = \mathbf{0}$.

K ARTIFACT ORGANIZATION AND VALIDATION

The supplement contains chess inference evaluators and stored non-chess audit results. Its cross-domain command decodes a frozen historical record and recomputes summaries and figures; it does not load domain checkpoints or repeat their interventions. The exact historical cross-domain inference runner is not supplied, so those audit results cannot be independently regenerated from checkpoints using this package alone. Manifests identify checkpoints and evaluation inputs; JSON and NPZ files retain the results summarized in Table 20.

The v2 script verifies input-file hashes and matched game identities before adding uncertainty rankings, local-utility intervals, and absolute cache summaries. It performs no new training or checkpoint inference. The inherited evaluators check model identity, fitted capacity, route role, and evaluation completion as summarized below. The chess audit used Python 3.12.13, PyTorch 2.10.0 with CUDA 12.8, NumPy 2.0.2, and two Tesla T4 GPUs. These are the recorded inference environment details, not the environment of the subsequent v2 array analysis. Appendix E distinguishes inherited and newly computed intervals.

Table 20: Artifact lineage. The principal empirical results have traceable stored inputs and validation checks.

Component	Persisted evidence	Acceptance check and role
Checkpoint manifest	SHA-256, seed, route role, parameter and state-key counts	Rejects identity or capacity mismatches before model loading.
Frozen chess tensor	Public game hashes, order, labels, ELO buckets, tensor digest	Requires the same 19,546 games and targets for every evaluated architecture.
Per-game policy record	Correctness, log loss, flips, JS divergence, and structural/conditional responses	Supplies the paired estimates, bootstrap intervals, and pathway figures without re-reading log text.
Route validation	Substitution, autograd, finite-difference, and no-mask controls	Requires exact routed zeroes together with nonzero unrestricted controls.
Reuse benchmark	Hardware, dtype, batch order, warm-up, cached and uncached outputs, timings	Records numerical agreement before any speed measurement is reported.
Cross-domain record	Seed, sample count, aggregate and group metrics, drift check	Preserves actual seed identities and reports route enforcement separately from predictive variation.

Validated checkpoints and frozen identifiers feed per-example records, paired summaries, and paper assets. Stored arrays permit the v2 statistical reanalysis without training infrastructure or checkpoint inference. Fresh forward evaluation and retraining additionally require the model binaries and underlying datasets. The checkpoints are absent from the current release; they will be made available after acceptance and can be supplied during review on request.

The preserved historical README mentions a test directory that is not included; the supplied evaluator self-test is the available miniature routing/cache check, not a trained-model replication. The six-checkpoint evaluator is configured for Kaggle once the required model files and Maia CSV are mounted. It checks their hashes before generating the audit outputs. The v2 companion provides a separate stored-array analysis command and manifest, so readers can distinguish recomputation of reported statistics from fresh model evaluation.

L FURTHER COMPARISONS AND EVIDENCE BOUNDARIES

The following checks would strengthen the study but have not been performed. The side-read ablation described in Section 7 would keep the prediction token's structural access, heads, objectives, and training procedure fixed. It would test that connection within this procedure; it would not establish superiority over every protected encoder or frozen-policy correction system.

Conditional performance and calibration. Per-rating and game-phase analyses could reveal variation concealed by aggregate accuracy and expected calibration error. Such analyses need

subgroup counts and uncertainty intervals. Temperature scaling or isotonic recalibration would require fitting on a separate validation set before evaluating the final holdout. They must not be fitted on that holdout and reported as independent test improvements.

Independent rankings. An engine-based score would provide a ranking independent of the evaluated policies, but not ground-truth human difficulty. Think time also depends on skill, clock pressure, and familiarity. A future analysis should define the ranking before examining its comparative outcome, retain the observed reversal, and separate sensitivity from recorded-move likelihood.

Systems and normalization. The existing measurements cover the reported context counts and batch sizes on the recorded hardware. Per-layer cost profiles and additional devices would extend this evidence. Token-local normalization is compatible with the proof; statistics aggregated across structural and conditional tokens can introduce a forbidden path. For example, RMS normalization over each token's feature dimension differs from normalization over both token and feature dimensions. Changing the normalization axes, tokenizer, or cache inputs requires checking the dependency contract again.

Group-robust learning. Just Train Twice selects examples using an initial model and upweights them in a second training stage (Liu et al., 2021b). Automatic Feature Reweighting retrains a final layer with a weighted loss (Qiu et al., 2023); selective last-layer finetuning selects reweighting data using errors or disagreement (LaBonte et al., 2023). These methods address predictive group robustness rather than the declared forward dependency. Comparing against empirical risk minimization and these procedures would require aligned encoders, splits, annotation access, selection rules, and compute budgets. The current routed/no-mask comparisons do not establish such a ranking.