When Can Pairwise Routing Dynamics Be Tested in Mixture-of-Experts? Identifiability Conditions and Evidence from a Learned Router

Anonymous authors
Paper under double-blind review

Abstract

How can pairwise routing dynamics in a learned mixture-of-experts (MoE) model be tested against a reduced stochastic law? We provide a falsifiable measurement criterion for this question. The criterion is instantiated by a specified two-expert (E=2) score-gap chain with an exact endpoint probability and a central first-order blind spot. For the aggregate soft-routing schema in our historical eight-expert (E=8) records, we prove that the fixed hard count, one-selection transition, and specified endpoint are not identifiable; that is, they are not uniquely determined by the recorded quantities. This is a schema-specific boundary on a dynamics test, not a universal impossibility result for MoE-to-reduced-model transfer. We provide a logging contract for a finite-chain dynamics test and verify the resulting analysis procedure on an imposed E=2 subsystem (4,608 independently replayed trajectories). We then measure pairwise routing dynamics in a 17M-parameter, four-layer, character-level E=8 MoE that makes one hard expert choice per token during the forward pass. Training, expert-pair choice, and evaluation use disjoint token spans. Its predefined six-feature linear model did not improve mean leave-one-seed-out MSE over router learning rate alone (mean MSE difference -0.238; 5 of 8 seed folds favor it). A narrower local effect, *gain-dependent selected-pair soft-share movement*, replicates: movement is larger at $g = 50$ than at $g = 1$ for 15 of 16 paired seeds on Shakespeare and an independent enwik8 prefix. This effect is reproducible local routing dynamics, not a validated E=2 transfer law.

1 Introduction

Sparse mixture-of-experts (MoE) models send each token to a small subset of their component networks, called experts. Their routing patterns shape capacity use, stability, and the meaning of expert specialization (Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et al., 2021). Simple stochastic reductions are therefore attractive: they offer interpretable predictions for how competition between experts evolves. But an agreement between a reduced-model curve and a neural-router statistic is informative only when the two quantities describe the same state, time unit, and observation.

This paper studies pairwise routing dynamics in learned multi-expert routers. We ask when those dynamics can be tested against a two-expert stochastic model. We use E for the number of experts available to a router: E=2 is the deliberately specified two-expert reduced chain, whereas E=8 is the learned router with eight candidate experts per layer used in our experiments. They are not two experimental settings of the same system. Our answer is a concrete matching requirement, not a rule for fitting curves. The central theoretical result identifies a schema-specific measurement limit. The aggregate quantities saved by the historical logging interface cannot uniquely determine the hard state or the one-step transition required by the pairwise law. In this paper, "identify" means "determine uniquely from the recorded quantities." A different logging schema could retain the missing information and support a valid dynamics comparison. Figure 2 gives an explicit counterexample: the same saved aggregate record can arise from different hard states and therefore different endpoint predictions. Before making a numerical comparison, one must say (i) what is being counted, (ii) what counts as one step, (iii) how one step changes the score, and (iv) which held-out quantity that

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.

Table 1: **Claim boundary.** The table states the intended scope of each result; it does not weaken the positive evidence within that scope.

This paper establishes	This paper does not establish
The historical aggregate schema cannot identify the specified E=2 state, transition, or endpoint.	No richer logging schema can test a reduced MoE law.
The imposed E=2 analysis procedure recovers its specified chain.	A learned E=8 MoE follows that chain.
Gain-dependent selected-pair soft-share movement replicates under disjoint data roles.	An E=2-to-E=8 quantitative transfer or a universal MoE mechanism.

score predicts. Below, we call these the state, step, score-change, and measurement maps. A "map" here is simply an explicit rule connecting a quantity in the simple model to a quantity in the learned router. Without all four rules, a reproducible neural measurement may still be useful, but it cannot test the reduced law.

We make the requirement concrete with a defined two-expert (E=2) rich-get-richer score-gap chain. In this model, selecting an expert makes it more likely to be selected again. The chain has an exact formula for the chance of eventually reaching an endpoint. At the 50–50 starting state, a small change does not alter that chance to first order, even when the change is real. The theory therefore supplies both a prediction and a list of what an experiment would have to measure to test it.

We then apply the requirement to an eight-expert (E=8) learned router. Each complete token span is assigned, before training, to one of three disjoint roles: training the model, selecting one ordered expert pair, or evaluation. Evaluation uses no gradient or parameter update, and the same evaluation batch is measured before and after an update computed only from training data. The predefined aggregate-soft predictor does not improve held-out prediction over the router learning rate alone. Yet one local outcome reproduces on two character datasets. This contrast is the central empirical result: the gain-dependent selected-pair soft-share movement can replicate without establishing the state and transition maps required by the reduced dynamical model.

Our contributions are:

1. For a defined E=2 score-gap chain, we derive the exact endpoint probability and show why the 50–50 state hides the leading small effect. This supplies a concrete pairwise routing-dynamics hypothesis.
2. We prove that the historical aggregate soft-routing schema does not identify the fixed hard state, one-selection transition, or specified E=2 endpoint. We give a logging contract for a finite-chain dynamics test.
3. A 4,608-trajectory E=2 analysis control, in which the specified chain is imposed, validates the analysis procedure rather than a learned-router reduction.
4. We measure pairwise routing dynamics in an E=8 router using disjoint data roles. The predefined aggregate-soft predictor fails its held-out prediction test, while gain-dependent selected-pair soft-share movement replicates across two independent studies. Neither result establishes an E=2 step in the learned router.

2 Related Work

MoE routing research has developed ways to balance expert load, manage expert capacity, and route sparsely (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zhou et al., 2022; Lewis et al., 2021; Roller et al., 2021). Other work shows that routing behavior can change sharply with the model configuration (Zoph et al., 2022; Chi et al., 2022; Chen et al., 2022). Router-score regularization and auxiliary-loss-free load balancing illustrate further training-design choices (de Brébisson and Vincent, 2016; Wang et al., 2024). We ask a different question. Rather than proposing a new routing mechanism, we ask when measurements from an existing router are enough to test a simple stochastic model.

The absorption algebra is classical (Feller, 1968; Karlin and Taylor, 1975), and reinforced-urn, stochastic-approximation, and fixation models provide broader context for competition and reinforcement (Pemantle, 2007; Robbins and Monro, 1951; Launay and Limic, 2012; Gottfried and Grosskinsky, 2024; Collevecchio et al., 2013). Our contribution is not a new summation identity. It is a framework for testing pairwise routing dynamics: what must be measured before a neural update can be compared to such a law. This differs from probability calibration, which compares predicted probabilities with observed frequencies rather than asking whether the model and experiment use the same state, step, score change, and measurement (Guo et al., 2017; Naeini et al., 2015). It is also distinct from latent-model identifiability, which asks whether model parameters can be recovered from observed variables (Allman et al., 2009), and from Markov lumpability, which asks when an aggregated process remains Markov (Kemeny and Snell, 1976; Buchholz, 1994). We instead ask whether a saved MoE routing observation retains the specific state, transition, and endpoint needed to test a specified routing-dynamics law.

3 Pairwise Reduced Model

3.1 Score-gap chain and exact absorption

Consider two experts with router scores (logits) s_1, s_2. Their score gap is $\Delta = s_1 - s_2$: a positive gap favors expert 1. The logistic function $\pi(\Delta) = \sigma(\Delta)$ converts the gap into a selection probability. The reduced model assumes that one selection is also one score update: selecting expert 1 increases the gap by a fixed amount δ, while selecting expert 2 decreases it by the same amount:

$$\Delta \leftarrow \begin{cases} \Delta + \delta, & \text{with probability } \pi(\Delta), \\ \Delta - \delta, & \text{otherwise.} \end{cases} \tag{1}$$

For the finite chain, the possible gaps lie on a discrete grid: $\Delta_i = (i - N/2)\delta$ for $i \in \{0, \dots, N\}$. The two grid endpoints $0, N$ are absorbing, meaning that the process stops after reaching either one. Thus $\pi_i = \pi(\Delta_i)$ and $q_i = 1 - \pi_i$. This rich-get-richer update is a *model assumption*; an actual neural-router update can also depend on hidden activations, expert parameters, optimizer state, and all other experts.

What this model predicts. Let $h_i(\delta)$ be the probability of ending at the upper endpoint when the chain starts at grid position i. Under exactly this finite, centered grid, and without making any claim about a neural router, standard one-dimensional random-walk algebra gives

$$h_i(\delta) = \frac{\sum_{j=0}^{i-1} \rho_j}{\sum_{j=0}^{N-1} \rho_j}, \qquad \rho_j = \exp\left[-\delta \left\{ \frac{j(j+1)}{2} - \frac{jN}{2} \right\}\right]. \tag{2}$$

At $\delta = 0$, $h_i(0) = i/N$, the neutral gambler's-ruin result (Feller, 1968; Karlin and Taylor, 1975). Equation 2 is therefore an exact conditional prediction: it says what absorption would look like if a measured router admitted this state and transition map.

3.2 Analysis control with an imposed E=2 chain

We first verify the recording and endpoint-estimation procedure in a setting where all four maps are known. After E=8 pretraining, we keep all model and optimizer parameters fixed and impose the fixed-$N = 1024$ E=2 chain on held-out tokens. This is a simulation-based analysis control, not a learned-router update. Across 4,608 independently replayed trajectories, central upper-endpoint frequencies are 0.515625 versus 0.500000 for $\delta = 0.004$, and 0.480469 versus 0.500000 for $\delta = 0.010$, each from 768 trajectories. The analysis procedure therefore recovers the imposed chain, but provides none of the maps required for a learned E=8 router.

3.3 When a small effect can be seen

Expanding Equation 2 at $\delta = 0$ gives

$$h_i(\delta) = \frac{i}{N} + \delta C(i, N) + O(\delta^2), \qquad C(i, N) = \frac{iA_N - NA_i}{N^2}, \tag{3}$$

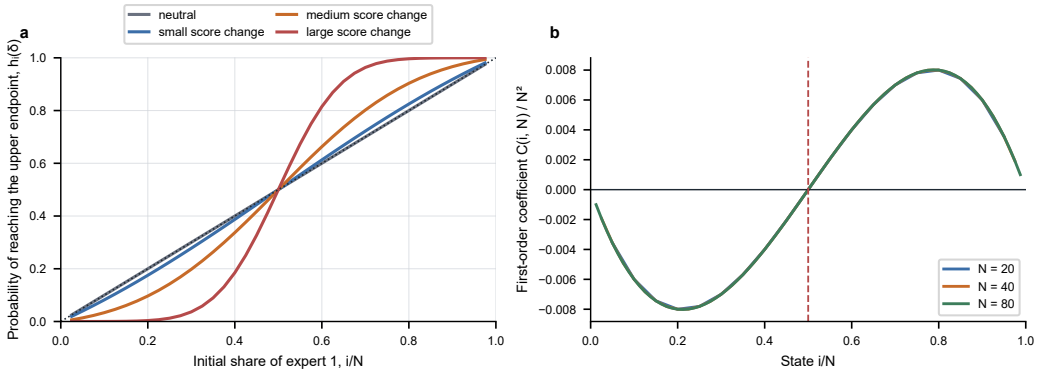


Figure 1: **Two-expert theory and when its small effect is visible. a**, Equation 2 is exact for the defined finite chain; increasing δ changes upper-wall absorption relative to the neutral reference. **b**, its leading small-effect coefficient is zero at the 50–50 state. Both panels describe the simple model, not fitted E=8 results.

where

$$A_i = \frac{1}{2}\left[\frac{(i-1)i(2i-1)}{6} + (1-N)\frac{(i-1)i}{2}\right]. \tag{4}$$

For an estimated share $\hat{s}$ with reference share s, assume a binomial standard error (the usual uncertainty estimate for independent yes/no draws) $\mathrm{SE} = \sqrt{s(1-s)/n}$ and define

$$\delta_{2\mathrm{SE}}(i, N, n) = \frac{2\mathrm{SE}}{|C(i, N)|}. \tag{5}$$

The coefficient $C(i, N)$ is the leading linear change in endpoint probability caused by a small δ. The scale $\delta_{2\mathrm{SE}}$ asks a limited question: assuming independent binary draws and a normal approximation, how large must that small change be to exceed twice this uncertainty estimate at a given state and token count? It is not a neural-network phase boundary and is not valid as an estimate for correlated tokens. Under a two-sided twice-the-standard-error rule, its rejection probability is $1/2 + \Phi(-4) \approx 0.50003$; if the direction is chosen in advance, the corresponding one-sided probability is exactly $1/2$. Crucially, $C(N/2, N) = 0$: at a 50–50 state, the leading linear signal is zero. This "blind spot" does not say that longer-run or nonlinear effects are absent.

3.4 Finite-chain certificate for approximate transfer

An approximate transfer requires more than calling two updates "close." The processes must share a finite state space. The certificate requires the next-state discrepancy, one-step endpoint discrepancy, and error from observing tokens rather than the latent state. The standard finite-chain bound is in the supplement. The present E=8 study does not record these quantities.

4 Identifiability Conditions for Pairwise Dynamics

When a numerical comparison is meaningful. An E=2–to–E=8 comparison is meaningful only when four maps are specified: a state map ϕ, a step map u, a score-change map κ, and a measurement map ψ. These maps answer four concrete questions: what is counted, what constitutes one step, how that step changes the score, and what held-out quantity the score predicts. Table 2 applies this criterion to the direct E=8 comparison. Without all four maps, a neural-update gap and $\delta_{2\mathrm{SE}}$ are different objects and cannot test the same claim.

The table does not say that the direct comparison is false. It separates two questions: whether a local measurement repeats, and whether the simple model can be compared to it. The disjoint-role experiments below test one predefined aggregate-soft measurement and one local outcome; neither supplies the missing count, one-selection step, or per-step score change.

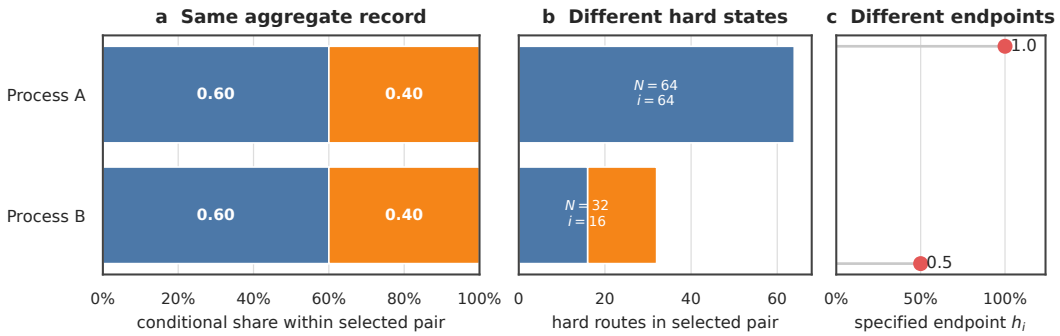


Figure 2: **An explicit aggregate-log non-identifiability construction. a**, Blue and orange segments denote experts a and b. Processes A and B have the same saved conditional share $s = .6$, pair mass $m = .4$, imbalance $|2s - 1| = .2$, and transformed mass $\mathrm{logit}(m)$. **b**, their unobserved hard states differ: $(N, i) = (64, 64)$ for A and $(N, i) = (32, 16)$ for B. **c**, the specified E=2 endpoint is consequently 1.0 for A and 0.5 for B. Thus the saved aggregate record does not identify the hard state or endpoint.

Table 2: **What the E=8 experiment must measure to test the E=2 model.** "Not measured" concerns this direct comparison only. It does not invalidate the E=2 theorem or every possible alternative comparison.

Role in the E=2 model	What must be defined	Candidate E=8 quantity	What the data show
State i: what is counted	a fixed count $\phi(R) \in \{0, \ldots, N\}$	tokens sent to the selected pair (not logged), or pair mass	the required count was not logged
One selection: what is one step	$u(R, R')$ is one E=2 step	one optimizer update	changes the full model, not one selected expert
Score change δ: how big is one step	$\kappa(R, R') = \Delta' - \Delta$ is stable	raw pair-score change (not logged)	no per-step score change was measured
Measurement: what the state predicts	$\psi(i, N)$ predicts a held-out statistic	soft conditional share (logged)	this chosen soft predictor did not help

Historical aggregate observation. For a fixed pair (a, b), let $p_t \in \Delta^7$ be token t's eight-expert softmax probabilities. The historical logging interface records four batch-level aggregates:

$$O(p) = (s, m, |2s - 1|, \mathrm{logit}(m)), \quad m = \frac{1}{n}\sum_t (p_{t,a} + p_{t,b}), \quad s = \frac{\sum_t p_{t,a}}{\sum_t (p_{t,a} + p_{t,b})},$$

before and after an optimizer update, together with the pair identity and data roles. These fields are recorded once per batch, not once per token. The scalar training loss and update identifier can be held fixed in the construction below. The final two fields are deterministic functions of s, m, so they add no independent token-level information. The schema does not record tokenwise hard routes or pair-score gaps.

Theorem 1 (Aggregate-log non-identifiability)**.** *For an eight-expert batch with n divisible by four, there are two token-level logit tensors with the same historical observation $O = (0.6, 0.4, 0.2, \mathrm{logit}(0.4))$. Their hard pair states are $(N, i) = (n, n)$ and $(n/2, n/4)$, respectively. Therefore the specified E=2 endpoint $h_i(\delta)$ is not identifiable from O: the two compatible states have endpoint probabilities 1 and $1/2$. Moreover, identical aggregate observations before and after an update can conceal a zero, unit, or multi-token hard-state transition.*

What the theorem rules out. This is a non-injectivity result for the observation operator O: different token-level configurations produce the same recorded aggregate. Therefore, no estimator whose input is only O can recover the target state, transition, or endpoint for every configuration compatible with the historical records. The result does not rely on an unrealistic probability parameterization, because every strictly positive probability vector is realized by logits. Holding the scalar training loss and update identity fixed does not add token-level route information. Additional assumptions—for example, tokenwise homogeneity or a known hard-route coverage rate—would change the theorem's domain, but the historical schema neither records nor validates them.

Proof sketch. The explicit probability construction is in Appendix A. Every strictly positive probability vector is realized by logits $z = \log p$. One process routes every token to a. The other routes one quarter to a, one quarter to b, and half outside the pair, while preserving the same aggregate pair mass and conditional share. A second construction preserves O while changing the hard count by one or eight tokens. The theorem is specific to the historical schema: no endpoint test based only on these aggregates can identify the specified E=2 object.

Logging contract for a specified test. For a fixed-pair finite-chain test, we require

$$\mathcal{L} = \{\text{pair/cohort IDs; tokenwise hard routes for every token; pre/post tokenwise pair gaps; event boundary and update identity; data role}\}.$$

The theorem shows why aggregate fields cannot replace hard routes or their fixed population. Without pre/post tokenwise gaps, the score-change map is also unobserved. If $\mathcal{L}$ records a fixed N, actual unit transitions, and separated calibration and held-out roles, it is sufficient to estimate the specified transition kernel, endpoint frequencies, and inputs to the finite-chain certificate. In particular, it makes the specified reduced-law hypothesis empirically falsifiable; it does *not* make the hypothesis true or prove that a natural MoE obeys the E=2 law.

Protocol 1: audit a reduced routing-dynamics claim. For a proposed reduced law, (1) state the finite state space, endpoint, and transition unit; (2) specify the state, step, score-change, and measurement maps; (3) list the logged fields, estimator, and failure condition for each map; (4) test whether those fields identify the target state and transition; and (5) only if the first four steps pass, compare transition kernels or endpoint predictions on held-out data. Otherwise, report the observation as a local routing phenomenon, not as a transfer test. This protocol is an audit and reporting checklist, not a benchmark.

5 Testing Pairwise Dynamics in an E=8 Router with Disjoint Data Roles

5.1 Separating measurement from update

The empirical question is deliberately narrower than a general MoE claim: does a held-out measurement from the learned eight-expert router support the proposed pairwise observation map, and what remains reproducible if it does not? We use a 17M-parameter, four-layer, character-level MoE. At each routed layer, its router assigns every token to one of eight experts (E=8). A *hard route* is the expert selected for that token. We train with a straight-through estimator (STE): the forward pass uses the hard route, while backpropagation uses a differentiable surrogate. This setting makes each data role and before/after router measurement explicit.

To avoid data reuse, we first divide the corpus into complete 65-token examples and then assign them to a deterministic 70/10/20 split. The training split updates parameters; the expert-pair-selection split chooses one ordered pair of experts; the evaluation split is held out. No input or target token appears in more than one role. We measure the same evaluation batch, without gradients or updates, before and after one update computed only from training data. Earlier analyses that reused overlapping windows are not used here.

For every layer, the released records contain the selected pair's soft-routing share, total soft probability, imbalance, and transformed total soft probability before and after the update. A *soft-routing share* is computed from router probabilities before the hard expert choice. The records do not contain the number of tokens hard-routed to the pair or the pair's raw score gap. We therefore cannot

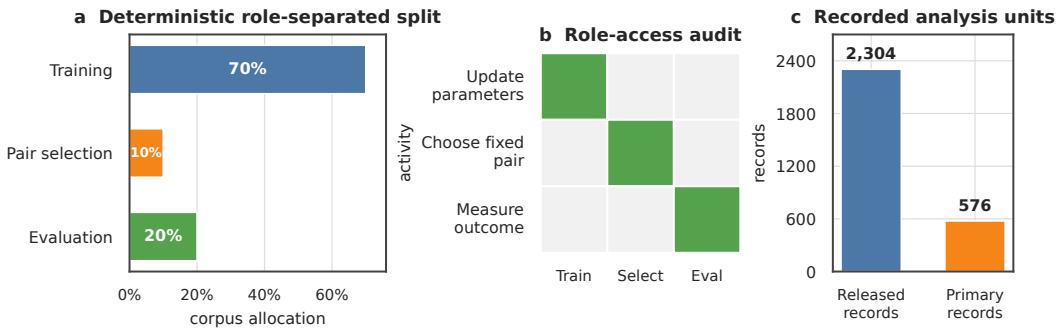


Figure 3: **Role separation and recorded units in the learned E=8 study. a**, Complete 65-token examples are assigned to training, pair-selection, and held-out evaluation roles in a deterministic 70/10/20 split. **b**, Green cells show the only corpus role used by each activity; gray cells are not accessed. **c**, The release has 2,304 layer-window records across four routed layers, of which 576 first-routed-layer records form the pre-specified primary prediction analysis. The same held-out evaluation batch is measured before and after a training-only update.

Table 3: **Evidence status and permitted interpretation.** Seeds, rather than windows or tokens, are the inferential unit. "Fixed before analysis" means chosen before the named analysis, not externally timestamped preregistration.

Evidence	Choice status	Reported unit	Permitted statement
E=2 analysis control	fixed chain and replay	4,608 trajectories	analysis recovers its specified law
Aggregate-soft prediction	predefined feature set	8 leave-one-seed-out folds	descriptive held-out prediction result
Shakespeare primary outcome	first routed layer selected before replication	16 paired seeds	local movement replication; 15/16 is descriptive
enwik8 primary outcome	independent protocol, no immutable prior timestamp	16 paired seeds	independent replication; sign calculation is post-hoc diagnostic

reconstruct those quantities. The records do establish disjoint data roles, a fixed pair identity, the same evaluation examples before and after an update, and no optimizer call on evaluation data. The prediction study uses eight matched random seeds (48–55), three router learning-rate multipliers $g \in \{1, 10, 50\}$, and 24 evaluation windows per run. The released data contain 2,304 layer-window records across four routed layers. The predefined prediction analysis uses the 576 records from the first routed layer only. Here g multiplies the router learning rate; it controls how far one optimizer update can move router parameters, not the number of experts.

5.2 Do aggregate soft measurements add predictive information?

For every evaluation batch, we average each expert's softmax probability over tokens. The *pair mass* is the combined average probability assigned to the two selected experts. The *conditional share* is expert 1's fraction of that pair mass. We test whether these aggregate soft measurements predict the magnitude of conditional-share change better than a baseline that uses only $[1, \log g]$, where g is the router learning-rate multiplier.

The soft-routing model uses the pair's imbalance $u = |2s - 1|$ and total mass $v = \mathrm{logit}(\mathrm{clip}(m, 10^{-6}, 1 - 10^{-6}))$, as well as terms that let their effect vary with $\log g$. The logit is a standard transform that expands probabilities near zero and one. The alternative is a six-feature linear least-squares model, $[1, \log g, u, v, (\log g)u, (\log g)v]$, for $\log(|\Delta s| + 10^{-8})$, where Δs is the conditional-share change across one optimizer update. It asks whether both the imbalance within the selected pair and the pair's total probability improve prediction beyond the learning-rate multiplier.

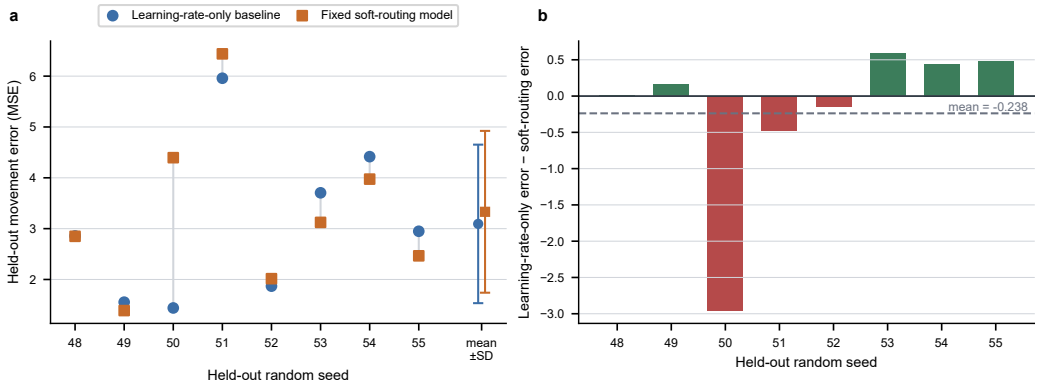


Figure 4: **The predefined aggregate-soft summary does not improve prediction.** **a**, each seed's leave-one-seed-out MSE for the learning-rate-only baseline and the fixed aggregate-soft model; error bars show the seed mean plus or minus one standard deviation. **b**, all eight paired MSE differences. The negative mean indicates no stable predictive improvement. This experiment tests one chosen measurement, not the endpoint formula.

This is the smallest alternative state constructible from the historical records: they retain batch-level pair mass and conditional share derived from full-softmax probabilities, but not a hard-pair count or raw pair-score gap. It is not the fixed-count E=2 state, and its performance does not estimate the performance of every possible measurement rule. For each leave-one-seed-out fold, we fit both linear models on the other seven seeds and standardize non-intercept inputs using only that training fold. The mean held-out MSE difference, baseline minus aggregate-soft model, is -0.238; only 5 of 8 held-out seed folds favor the aggregate-soft model. Thus the pre-specified six-feature model did not improve mean leave-one-seed-out MSE over the learning-rate-only baseline. We release the complete fold scores, fitted values, and source hashes; we do not change features, targets, or time horizons after seeing this result.

5.3 Gain-dependent selected-pair soft-share movement

Failure of the soft-routing measurement does not mean that every E=8 change is zero. We therefore name and report a separate local phenomenon: *gain-dependent selected-pair soft-share movement*. For each random seed, it is the mean absolute change in the selected pair's conditional share at the first routed layer (layer 0) over 24 evaluation windows at $g = 50$, minus the same value at $g = 1$. This outcome describes the effect of one optimizer update; it is not one E=2 selection. The first routed layer was the sole primary outcome selected before replication, so deeper layers remain descriptive. The Shakespeare study uses fresh matched seeds 56–71. All 48 runs and 4,608 layer-window records meet the disjoint-role, fixed-pair, evaluation-without-update, and completeness checks.

The matched mean difference is 1.466×10^{-3}; 15 of 16 seed differences are positive. The originally planned sign-flip calculation would require randomly assigning the two values of g, or an additional symmetry assumption. This design has neither, so it is not interpreted as a randomization test. As a post-hoc direction check, a binomial sign test asks whether a new independent seed is more likely than not to have a positive difference. If one additionally assumes independent seed draws and the superpopulation null $\Pr(D_r > 0) = 1/2$, it gives $p = 17/2^{16} = 2.59 \times 10^{-4}$. The primary descriptive evidence is instead the 15/16 direction count, paired mean difference, and a 10,000-resample percentile bootstrap interval over the 16 matched seed-pair differences, $[7.76 \times 10^{-4}, 2.29 \times 10^{-3}]$. This supports the local gain-dependent movement phenomenon, not an E=2 endpoint law, fixed-count step, $C(i, N)$, or δ_{2SE}.

We repeated the same outcome and robustness analysis on a one-million-character enwik8 prefix with new matched seeds 72–87. This is a new dataset, not pooled with Shakespeare and not run with the Shakespeare seeds. The analysis plan was written before this run, but its local timestamp cannot establish an immutable run-prior record. We therefore call it an independent replication,

not a preregistered confirmation. Its 48 runs also meet the disjoint-role, fixed-pair, and evaluation-without-update checks. The matched mean difference is 9.390×10^{-4}, with 15 of 16 positive seed differences. The same post-hoc direction diagnostic gives $p = 17/2^{16} = 2.59 \times 10^{-4}$ only under the same independent-draw and $\Pr(D_r > 0) = 1/2$ assumptions. Its primary descriptive evidence is 15/16 positive paired differences, the mean, and the paired bootstrap interval $[6.18 \times 10^{-4}, 1.29 \times 10^{-3}]$. The gain-dependent movement phenomenon therefore repeats across two character datasets. It answers the empirical question—the local measurement repeats—but not the transfer question: it does not identify an E=2 step or supply a quantitative E=2–to–E=8 comparison.

6 Conclusion

We provide a falsifiable criterion for testing a specified reduced routing-dynamics law in a learned MoE. For the defined E=2 score-gap chain, the endpoint probability and leading small-effect behavior are exact. A numerical E=2–to–E=8 comparison requires matching state, step, score-change, and measurement maps. Theorem 1 shows that the historical aggregate schema does not identify the state or transition needed for that comparison. Our disjoint-role experiments therefore keep two claims distinct. The pre-specified six-feature aggregate-soft model did not improve mean leave-one-seed-out MSE beyond the learning-rate-only baseline. In contrast, gain-dependent selected-pair soft-share movement is larger at $g = 50$ than at $g = 1$ in 15 of 16 seed comparisons on both Shakespeare and an independent enwik8 prefix. This replication concerns a local E=8 routing phenomenon; it does not recover a fixed count, one-selection step, or E=2 score change. A future transfer claim must follow Protocol 1 and measure the four maps in Table 2; an approximate claim must also measure the discrepancy between the two systems' transitions and observations. Until then, the E=2 law and local E=8 phenomenon are complementary results, not two estimates of the same quantity.

Reproducibility and Ethics Statement

The accompanying code and supplement document the exact formulas, experiment settings, data split, raw summaries, and scripts that generate the tables. The release inventory gives a content hash for each raw result, analysis, and manifest, and distinguishes the historical measurement interface from updated measurement code for future runs. Statistical comparisons treat the random seed, not individual tokens or windows, as the independent unit. Experiments use public text corpora and small research models. We study potentially unstable routing behavior but make no deployment, safety, or downstream-performance claim.

AI ASSISTANCE DISCLOSURE

Generative AI tools were used solely to polish the manuscript's language and presentation. They were not used for research ideation, related-work identification, mathematical development, experimental design, data generation, code, analysis, or figure creation. The authors reviewed the final manuscript and take full responsibility for all scientific claims and released artifacts.

REFERENCES

Jiro Akahori, Andrea Collevecchio, Timothy Garoni, and Kais Hamza. Attraction properties for general urn processes and applications to a class of interacting reinforced particle systems. *arXiv preprint arXiv:1602.05677*, 2016.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

Tibor Antal and István Scheuring. Fixation of strategies for an evolutionary game in finite populations. *Bulletin of Mathematical Biology*, 68(8):1923–1944, 2006. doi: 10.1007/s11538-006-9061-4.

Peter Buchholz. Exact and ordinary lumpability in finite markov chains. *Journal of Applied Probability*, 31(1):59–75, 1994. doi: 10.2307/3215235.

Siyu Chen, Yuntian Deng, Qian Wu, Guangyong Gu, and Yu Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Ta-Chung Chi, Hsiang-Fu Fan, Peter J. Ramadge, and Alexander I. Rudnicky. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 2022. arXiv:2204.09179.

Andrea Collevecchio, Codina Cotar, and Marco LiCalzi. On a preferential attachment and generalized Pólya's urn. *The Annals of Applied Probability*, 23(3):1219–1253, 2013. doi: 10.1214/12-AAP869. arXiv:1203.5758.

Alexandre de Brébisson and Pascal Vincent. The Z-loss: a shift and scale invariant classification loss belonging to the spherical family. *arXiv preprint arXiv:1604.08859*, 2016.

Warren J. Ewens. *Mathematical Population Genetics*. Springer, 2004.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. arXiv:2101.03961.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1*. Wiley, 3rd edition, 1968.

Thomas Gottfried and Stefan Grosskinsky. Asymptotics of generalized Pólya urns with nonlinear feedback. *Electronic Journal of Probability*, 29:1–56, 2024. doi: 10.1214/24-EJP1157. arXiv:2303.01210.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. arXiv:1706.04599.

Daniel Hennes, James Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paulius Parmas, Edgar Duéñez-Guzman, and Karl Tuyls. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Springer-Verlag, New York, 2 edition, 1976.

Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6): 713–719, 1962.

Motoo Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1:177–232, 1964.

Mickaël Launay and Vlada Limic. Interacting urn models. *arXiv preprint arXiv:1207.5635*, 2012.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. arXiv:2006.16668.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, 2021. arXiv:2103.16716.

Xuesong Liu, Qiuhui Pan, Yibin Kang, and Mingfeng He. Fixation probabilities in evolutionary games with the Moran and Fermi processes. *Journal of Theoretical Biology*, 364:242–248, 2015. doi: 10.1016/j.jtbi.2014.08.047.

Patrick A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, pages 2901–2907, 2015.

Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007. doi: 10.1214/07-PS094.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.04426.

Hassan Shafiey and David Waxman. Exact results for the probability and stochastic dynamics of fixation in the Wright–Fisher model. *Journal of Theoretical Biology*, 430:64–77, 2017. doi: 10.1016/j.jtbi.2017.06.026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. arXiv:1701.06538.

Christine Taylor, Drew Fudenberg, Akira Sasaki, and Martin A. Nowak. Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology*, 66(6):1621–1644, 2004. doi: 10.1016/j.bulm.2004.03.004.

Arne Traulsen, Jorge M. Pacheco, and Martin A. Nowak. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, 246(3):522–529, 2007. doi: 10.1016/j.jtbi.2007.01.002.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024.

Bin Wu, Philipp M. Altrock, Long Wang, and Arne Traulsen. Universality of weak selection. *Physical Review E*, 82:046106, 2010. doi: 10.1103/PhysRevE.82.046106. arXiv:1010.2105.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2202.09368.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

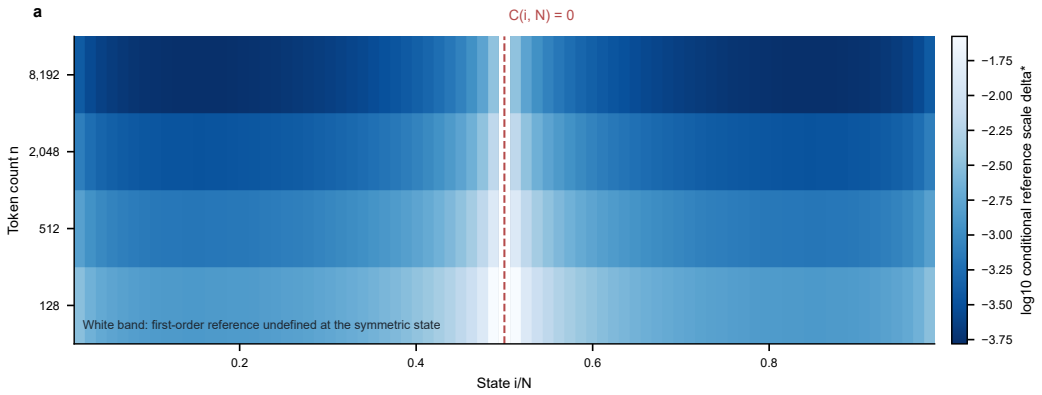


Figure 5: **The twice-standard-error reference depends on state and token count.** The analytical scale $\delta_{2\mathrm{SE}}(i, N, n)$ decreases with available tokens away from the 50–50 state, but is undefined at the central blind spot where $C(i, N) = 0$. This is a property of the defined E=2 measurement model, not an E=8 estimate.

A Proof of Aggregate-Log Non-identifiability

We give the construction for a 64-token batch. Repeating every token type equally often gives the corresponding construction for the 1,024-token historical evaluation batches. Let $c \notin \{a, b\}$ be a third expert. All unspecified remaining probability is split uniformly over the indicated outside experts, so every row is strictly positive and is realized by logits $z = \log p$.

In Process A, every token has $(p_a, p_b) = (0.24, 0.16)$, and each of the other six experts has probability 0.10. Thus every hard route is a, so $(N, i) = (64, 64)$. In Process B, use the following token types:

Type	Fraction	p_a	p_b	p_c	Remaining outside mass
A	$1/4$	.32	.12	included in $.56/6$	$.56/6$ over six experts
B	$1/4$	.16	.28	included in $.56/6$	$.56/6$ over six experts
C	$1/2$	.24	.12	.30	$.34/5$ over five experts

The hard routes of types A, B, and C are a, b, c, respectively, so $(N, i) = (32, 16)$. Yet both processes have $\bar{p}_a = .24$, $\bar{p}_b = .16$, and hence identical $(s, m, |2s - 1|, \mathrm{logit}(m)) = (.6, .4, .2, \mathrm{logit}(.4))$. The endpoint values differ because $h_{64}(\delta) = 1$, while symmetry of the centered specified chain gives $h_{16}^{(32)}(\delta) = 1/2$.

The same construction also proves the step statement. Keep the A/B type probabilities fixed, use $17/64, 15/64, 32/64$ fractions of A/B/C, and set the C type to $(p_a, p_b, p_c) = (.235, .125, .30)$. The aggregate observation is unchanged, but the hard state is $(32, 17)$: an unobserved unit move from $(32, 16)$. With fractions $24/64, 8/64, 32/64$ and C type $(.20, .16, .30)$, the same observation instead has hard state $(32, 24)$, an unobserved eight-token move. Reusing the original Process-B tensor gives an unobserved zero move. Thus identical logged observations before and after cannot distinguish these three step maps.

B Measurement Objects and Audit Boundaries

The E=2 quantities in Section 3 and the fixed selected-pair share in Section 5 are different measurements and are never combined. The E=8 study assigns complete supervised examples to 70% model training, 10% expert-pair selection, and 20% evaluation. Expert-pair selection uses only its own data, and evaluation batches never update the model. Historical experiments with overlapping windows are retained as exploratory context, not used as evidence in this paper.

Table 4: **All pre-specified leave-one-seed-out prediction folds.** Both models are trained on the other seven seeds using only their fold-specific standardization. These are descriptive calibration scores only.

Held-out seed	n	Baseline MSE	Aggregate-soft MSE	Difference
48	72	2.856	2.849	0.007
49	72	1.554	1.388	0.165
50	72	1.440	4.396	-2.956
51	72	5.959	6.438	-0.480
52	72	1.869	2.016	-0.147
53	72	3.705	3.123	0.582
54	72	4.416	3.974	0.442
55	72	2.948	2.466	0.483
Mean	–	3.093	3.331	-0.238

Pairwise measurements. For readers who want the exact definitions, let $z_{x,t,e}$ be the router score for expert e at token t in example x, and let B be an evaluation batch. Write $\pi_{x,t,e} = \mathrm{softmax}(z_{x,t})_e$ and $\bar{\pi}_e(B) = \frac{1}{|B|T}\sum_{x,t}\pi_{x,t,e}$. For its fixed, ordered selected pair (a, b), the reported soft pair share is

$$S_a(B; a, b) = \frac{\bar{\pi}_a(B)}{\bar{\pi}_a(B) + \bar{\pi}_b(B)}.$$

The hard pair share—the fraction after choosing only the highest-scoring expert for each token—is

$$H_a(B; a, b) = \frac{\sum_{x,t} \mathbb{1}\{\arg\max_e z_{x,t,e} = a\}}{\sum_{x,t} \mathbb{1}\{\arg\max_e z_{x,t,e} \in \{a, b\}\}},$$

and would be undefined, rather than imputed, when its denominator is zero. The released role-separated records do not contain H_a or its denominator, so this defines a future measurement rather than a reconstructed historical statistic. The soft-routing movement is $|S_a(B_{\text{after}}; a, b) - S_a(B_{\text{before}}; a, b)|$; its unit is one optimizer update, not one E=2 selection.

C Pre-specified Aggregate-Soft Prediction Analysis

Table 4 reports every descriptive leave-one-seed-out fold from the pre-specified calibration analysis. Each held-out seed contributes 72 first-routed-layer records (three values of g times 24 evaluation windows); the 576 records are repeated measurements, not inferentially independent tokens. The signed quantity is baseline MSE minus aggregate-soft-model MSE, so a positive number favors the pre-specified aggregate-soft predictor. Its across-seed mean is -0.238 with SD 1.155.

The pre-specified feature order is $[1, \log g, u, v, (\log g)u, (\log g)v]$, with fitted coefficients $(-8.756, 0.274, -0.053, 0.678, 1.003, -1.000)$; the dose-only coefficients are $(-8.756, 0.820)$. The release includes fold-specific normalizers, source and corpus hashes, and the deterministic data-split identifier. It records 2,304 finite layer-window rows with disjoint data roles, train-only updates, forward-only evaluation, the same evaluation batch before and after each update, and fixed expert pairs. It does not establish a historical parameter-integrity checksum or cross-forward equivalence assertion. Soft shares are always defined by full-softmax pair mass. The historical role-separated records contain no hard shares, hard-token counts, or pair-logit gaps, so neither the hard-route coverage rate nor a gap-based coupling can be inferred from them. The current measurement interface records these fields for future studies; those future records must be analyzed separately and cannot alter the reported historical results.

D Supplementary Finite-Chain Derivations

Derivation of Equations 2–4. For $1 \le i \le N-1$, let h_i be the probability of hitting N before 0. The first-step recurrence is

$$h_i = \pi_i h_{i+1} + q_i h_{i-1}, \qquad h_0 = 0, \quad h_N = 1.$$

Writing $d_i = h_i - h_{i-1}$ gives $\pi_i d_{i+1} = q_i d_i$. Since $q_i/\pi_i = \exp[-(i - N/2)\delta]$, induction yields

$$d_{j+1} = \rho_j d_1, \qquad \rho_j = \prod_{m=1}^{j} \frac{q_m}{\pi_m} = \exp\left[-\delta \left\{ \frac{j(j+1)}{2} - \frac{jN}{2} \right\}\right].$$

Summing $d_1, \ldots, d_i$ and imposing $h_N = 1$ gives Equation 2. At $\delta = 0$, all $\rho_j = 1$, hence $h_i = i/N$.

Let $a_j = j(j+1)/2 - jN/2$. The finite expansion $\rho_j = 1 - \delta a_j + O(\delta^2)$ gives

$$h_i(\delta) = \frac{i - \delta A_i + O(\delta^2)}{N - \delta A_N + O(\delta^2)} = \frac{i}{N} + \delta \frac{iA_N - NA_i}{N^2} + O(\delta^2),$$

where $A_i = \sum_{j=0}^{i-1} a_j$. Summing the quadratic and linear terms in a_j gives Equation 4. These are statements about the stipulated finite centered lattice; they do not imply that a complete E=8 neural-router update has this transition kernel. □

Plain-language summary. The next result is a standard finite-process perturbation identity: two finite processes have similar endpoint probabilities only when their one-step behavior and their one-step endpoint chances are both close. It is included for readers who need the exact mathematical condition for an approximate comparison; the present E=8 experiments do not measure its inputs.

Theorem 2 (Finite-process endpoint bound)**.** *Let P and $\widetilde{P}$ be transition matrices on the same finite set of non-endpoint states. Their rows sum to at most one, because probability can move to an endpoint, and each process eventually leaves these states; these are the usual "transient substochastic" conditions. Let b and $\widetilde{b}$ be their one-step upper-endpoint probability vectors. If $\|P - \widetilde{P}\|_\infty \le \epsilon_P$ and $\|b - \widetilde{b}\|_\infty \le \epsilon_b$, then their absorption vectors satisfy*

$$\|h - \widetilde{h}\|_\infty \le \|(I - P)^{-1}\|_\infty (\epsilon_P + \epsilon_b).$$

Proof. The first-step equations are $h = Ph + b$ and $\widetilde{h} = \widetilde{P}\widetilde{h} + \widetilde{b}$. Since P is transient, $I - P$ is invertible. Subtracting the equations and multiplying by $(I - P)^{-1}$ gives

$$h - \widetilde{h} = (I - P)^{-1}\{(P - \widetilde{P})\widetilde{h} + (b - \widetilde{b})\}.$$

An absorption-probability vector satisfies $\|\widetilde{h}\|_\infty \le 1$. The induced infinity norm and the two assumed discrepancy bounds yield the stated inequality. □

E Extended Context for the Reduced Chain

The E=2 chain is closest, mathematically, to finite-population absorption models. The Moran process and early fixation calculations study the probability that one of two competing types reaches an absorbing boundary (Moran, 1958; Kimura, 1962). Diffusion approximations and population-genetics treatments analyze related state-dependent stochastic processes (Kimura, 1964; Ewens, 2004). Work on finite evolutionary populations studies fixation under different update and selection rules (Antal and Scheuring, 2006; Taylor et al., 2004; Traulsen et al., 2007; Liu et al., 2015; Shafiey and Waxman, 2017; Wu et al., 2010). These sources explain why an exact endpoint formula is standard once a finite state and a one-step rule are specified. They do not supply such a rule for a learned router.

Reinforcement models study how repeated choices can amplify early differences, including interacting-urn systems (Akahori et al., 2016). Neural replicator dynamics likewise provides a differentiable model of population change (Hennes et al., 2019). These are useful neighboring perspectives on competition and reinforcement, but they use their own state variables and update mechanisms. Our measurement criterion is needed precisely because a similar qualitative mechanism does not make a neural-router optimizer update an E=2 selection event.

F Position Relative to Cited Work

Table 5 compares the measurement objects of the cited work families with ours. It is a positioning aid, not an exhaustive survey or a claim that no other work studies related questions.

Table 5: **Position relative to cited work.** The distinction is the object being measured and the claimed connection to learned routing.

Work family	Object and relation to this paper
Finite birth–death processes (Feller, 1968; Karlin and Taylor, 1975)	Exact endpoint probabilities for defined Markov chains. We use this classical algebra for the centered E=2 score-gap chain; it does not by itself show that a neural-router update follows the chain.
Reinforced urns and fixation (Pemantle, 2007; Launay and Limic, 2012; Gottfried and Grosskinsky, 2024; Collevecchio et al., 2013)	Multi-type reinforcement, stochastic approximation, and eventual single-type dominance. Our distinct question is whether a measured E=8 pair has matching state, step, and measurement rules before numerical comparison.
Sparse MoE routing (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zhou et al., 2022; Lewis et al., 2021; Roller et al., 2021)	Training, load balancing, assignment, and scalable conditional computation. We do not propose a routing mechanism; we measure a selected pair and state exactly what prevents transfer to the simple model.
Probability calibration (Guo et al., 2017; Naeini et al., 2015)	Agreement between predicted probabilities and observed frequencies. Our comparison instead asks whether the theory and experiment use matching states, steps, score changes, and measurements.
This work	Exact conditional E=2 endpoint probability, plus separated-data E=8 pairwise measurement. The local E=8 outcome is evidence about one measurement; it is not an identified E=2 step or a closed-form E=8 law.

G Two-Sided Detectability Calculation

Under the leading-order normal approximation, write the centered share-error as $Z\mathrm{SE}$ with $Z \sim \mathcal{N}(0,1)$. At $|\delta C| = 2\mathrm{SE}$, a two-sided rule rejects when $|Z+2| > 2$. Its probability is

$$\Pr(Z > 0) + \Pr(Z < -4) = \frac{1}{2} + \Phi(-4).$$

Thus Equation 5 is a nominal 50%-power reference under a two-sided rule. It is exactly 50%-power only for a one-sided rule whose direction was specified before observing the data.