ALREADY SOLVED? WHEN TO STOP AND WHAT TO RETURN IN ITERATIVE REASONERS

Anonymous authors
Paper under double-blind review

ABSTRACT

An iterative reasoner can find a correct answer, keep computing, and ultimately return the wrong one. We study two coupled decisions: when to stop refining a state, and which generated answer to preserve. Our approach starts from a geometric observation: a hidden state can keep moving inside a region that decodes to the same answer. Margin-Motion Inclusion (MMI) compares the protection provided by the minimum logit margin with estimated remaining motion above an estimated noise floor, including a calibrated margin reserve against wrong early returns. We derive a decoding-ball guarantee without noise and conditions for recognizing the attractor's answer under persistent noise. Calibration then controls wrong early returns under exchangeability. The policy accepts and returns the first candidate that passes the test, potentially saving both iterations and restarts; otherwise, it returns the largest-margin candidate encountered. An oracle analysis separates missed solutions from delays in returning them. On recorded Sudoku and Maze trajectories from frozen Equilibrium Reasoners, fast MMI with maximum-margin fallback uses $34.8\times$ and $16.4\times$ fewer outer updates on average than terminal voting over 32 restarts of 64 updates. It also reduces errors by 20% and 75%, respectively. The gains come from recognizing and retaining answers, without changing the reasoner's learned dynamics.

1 INTRODUCTION

How much computation does a reasoning problem need? Iterative models offer an appealing response: return to the same representation, refine it, and spend more steps when the problem is difficult. Recurrent networks have learned to extend algorithmic computation beyond their training horizon (Schwarzschild et al., 2021), and recent latent reasoners bring this idea to shared neural updates (Geiping et al., 2025; Hao et al., 2025). Hierarchical Reasoning Models (HRM), Tiny Recursive Models (TRM), and Equilibrium Reasoners (EqR) apply related forms of refinement to structured problems such as Sudoku and mazes (Wang et al., 2025; Jolicoeur-Martineau, 2025; Huang et al., 2026). Instead of producing only one prediction, these models expose a succession of candidate answers.

That succession creates a decision problem of its own. Consider a solver that has already assembled a correct Sudoku board. Its internal representation may still change, even though no further change to the board is needed. If inference waits for a fixed budget, those extra iterations cost computation. If the board later changes, choosing only the final candidate can lose a solution that was already available. With several initial states, the policy must also decide which trajectory's answer to trust. Such overthinking and selection failures have precedents in early-exit and recurrent models (Kaya et al., 2019; Bansal et al., 2022); repeated-sampling studies likewise distinguish generating a correct candidate from recognizing it (Brown et al., 2024).

We study this problem in EqR. Its randomized trajectories give a particularly clear view of the opportunity: an answer can appear before a trajectory settles, and one successful trajectory can make later restarts unnecessary. The two questions are therefore linked: *which answer should be returned, and when has enough computation been spent?* A correctness oracle reveals the available headroom by identifying the first correct candidate. A practical policy must make its decision without that oracle.

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.

Our approach starts from the geometry of the learned dynamics. Without noise, a trajectory in a contractive basin approaches an attractor whose answer occupies a surrounding decoding region. A sufficiently large margin relative to a valid bound on distance to the attractor certifies entry into its decoding ball. When the attractor is correct, this is a *solution ball*: further refinement preserves the solution. This connects stopping to the destination learned during training.

Margin–Motion Inclusion (MMI) makes this picture into an observable stopping rule. It compares the answer's margin with estimated remaining movement above a measured noise floor, reserving additional margin against errors the estimate does not exclude. Noise-free contraction gives a decoding-ball guarantee; with noise, we derive conditions for agreement with the attractor's answer and calibrate observable wrong early returns. If no candidate passes, the policy returns the largest-margin answer encountered. Stopping and preservation work together: recognizing an answer can avoid both further iterations and later restarts, while preserving a strong candidate can prevent a poor final selection.

Without retraining EqR, fast MMI reaches $98.419 \pm 0.075\%$ on Sudoku-Extreme and $99.425 \pm 0.112\%$ on Maze across five inference seeds. It uses 58.96 and 125.18 mean outer updates, versus 2,048 for the full EqR computation grid. Relative to terminal voting, mean errors fall by 20% and 75%; compared with the stronger Sudoku convergence selector, the reduction is only 0.8%. The Maze reserve reaches $99.650 \pm 0.105\%$, within 0.125 percentage points of its oracle, at 633.38 updates. [repository link to be added].

2 From iterative trajectories to a returned answer

An *iterative reasoner* refines a persistent state and exposes candidate answers. For input x, restart b, and outer update t,

$$S_{b,0} \sim \mu_\theta(\cdot \mid x), \qquad S_{b,t+1} = F_\theta(S_{b,t}, x; \xi_{b,t+1}), \qquad \widehat{y}_{b,t+1} = \mathcal{D}_\theta(S_{b,t+1}, x). \tag{1}$$

Here θ is the fixed model, S contains all recurrent memory, μ_θ initializes it, ξ supplies randomness, and $\mathcal{D}_\theta$ decodes its answer. Deterministic (noise free) updates ignore ξ; fixed initialization makes μ_θ a point mass.

For the two-state models considered here, $S = (Z^H, Z^L)$. Dropping restart and outer-step indices, one internal cycle j performs K_L low updates and one high update:

$$Z^L_{j,k+1} = f_L(Z^L_{j,k}, Z^H_j + E(x); \epsilon^L_{j,k}), \qquad Z^H_{j+1} = f_H(Z^H_j, Z^L_{j,K_L}; \epsilon^H_j). \tag{2}$$

Here $E(x)$ embeds the input, $k = 0, \ldots, K_L - 1$, and $Z^L_{j+1,0} = Z^L_{j,K_L}$. One outer update F_θ composes the configured internal cycles; one function evaluation (NFE) counts this complete update. Figure 1 shows EqR's shared choice $f_L = f_H = f_\theta$:

$$f_\theta(v, u; \epsilon) = (1 - \lambda)v + \lambda R_\theta(v + u) + \beta\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

The refinement block R_θ updates state v conditioned on u; $\lambda \in (0, 1]$ is its interpolation weight and $\beta \geq 0$ its noise scale, with I the identity in the substate dimension.

Three instantiations. HRM uses separate low/high modules; TRM shares one. Both use fixed stored initial states and deterministic, undamped updates ($\lambda = 1, \beta = 0$) for a fixed encoded input (Wang et al., 2025; Jolicoeur-Martineau, 2025). EqR shares the block but draws fresh initial states and perturbs each damped internal call (Huang et al., 2026). Write $z_{b,t} = Z^H_{b,t}$ and $g(z_{b,t}) = \mathcal{D}_\theta(S_{b,t}, x)$. Its StableMax loss (Prieto et al., 2025) supervises successive candidates, while a separate halting head predicts complete-answer correctness (Huang et al., 2026). Implementation conventions are in Appendix E.

Fixed budgets and learned stopping. A budget (B, D) permits B restarts of at most D updates. Fixed-grid inference spends BD, then selects a terminal answer by vote or convergence. Voting uses complete-answer plurality, with ties broken by restart order. Adaptive computation time (ACT) learns when each trajectory should stop (Graves, 2016; Huang et al., 2026). In the fixed-breadth comparison, all B restarts run, costing $\sum_b T_b$ updates for stopping depths $T_b \leq D$; their returned candidates are then combined.

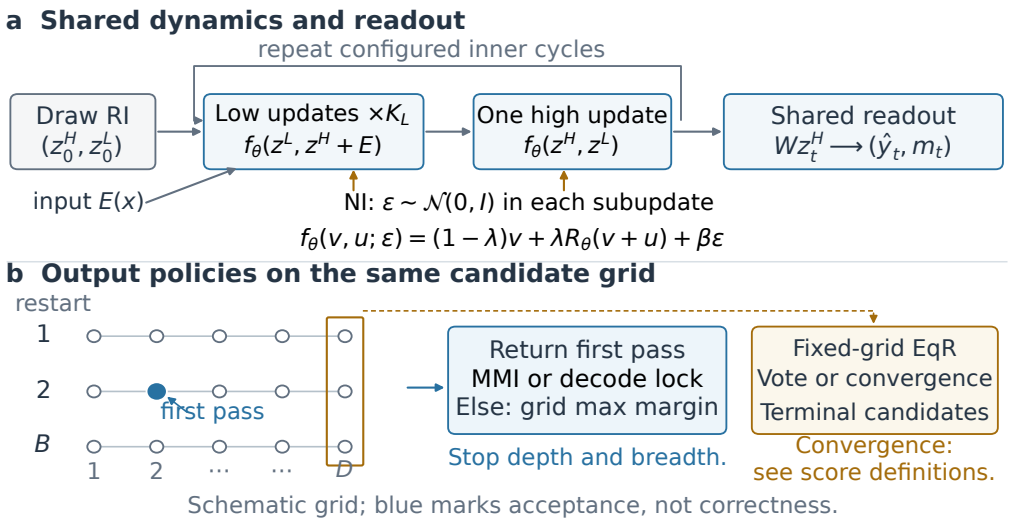

Figure 1: Shared dynamics, different return policies. The evaluated EqR checkpoints compose three internal cycles per outer update. The return policy selects the first candidate passing the chosen test (MMI or decode locking); otherwise it returns the largest-margin candidate visited. Acceptance can avoid both later updates and later restarts. RI denotes random initialization and NI noise injection. This is a computation schematic.

An oracle separates generation from recognition. For correct answer $y(x)$, define the first correct update on restart b as

$$\tau_b^*(x) = \min\{1 \le t \le D : g(z_{b,t}) = y(x)\},$$

with $\tau_b^* = \infty$ if none exists. The oracle succeeds when at least one restart contains a correct candidate. Its accuracy bounds selection from this same coupled grid; its first-correct times expose recognition delay. Following earlier oracle analyses (Schwartz et al., 2020; Brown et al., 2024), we measure both missed answers and the delay before returning a solution already generated.

3 MMI: RECOGNIZING ANSWERS FROM MARGIN AND MOTION

Our objective is to recognize a solution before refinement finishes. The argument connects three ideas: margin protects an answer, contraction relates it to an attractor, and supervision encourages that destination to be correct. We first establish the noise-free inclusion guarantee, then adapt the objective to persistent noise and obtain a measurable rule. Agreement with an attractor implies correctness only when that attractor is correct; Appendix B relates this requirement to supervised loss and coverage.

3.1 THE MARGIN MEASURES HOW MUCH MOVEMENT AN ANSWER CAN TOLERATE

For the geometric analysis, we assume a single linear decoder layer, with an optional bias, followed by a positionwise argmax. This is the readout used by the evaluated EqR checkpoints. There are P evaluated output positions and K symbols. With decoder matrix W, rows $w_a^\top$, and biases b_a, the logit at position p is $h_{p,a} = w_a^\top z_p + b_a$. The decoded answer is $g(z) = (c_p)_{p=1}^P$, where $c_p = \arg\max_a h_{p,a}$ with deterministic ties. The winner's lead over its nearest rival is its top-two margin. A complete Sudoku board can change when even one position changes, so we take the weakest position:

$$m(z) = \min_p \big[h_{p,c_p} - \max_{a \ne c_p} h_{p,a}\big], \qquad \Gamma = \max_{a \ne b} \|w_a - w_b\| > 0, \qquad d(z) = \frac{m(z)}{\Gamma}. \tag{4}$$

The constant Γ converts movement in representation space into possible change in a logit gap. We use Euclidean norms, equivalently Frobenius norms for state arrays; the complete state uses the Euclidean product norm, so projection onto its high component cannot increase distance. The quantity d is therefore a conservative *decoder radius*: how far the representation may move while protecting every position of the answer.

Lemma 1 (A margin protects an answer). *If $m(z) > 0$ and $\|z' - z\| < d(z)$, then $g(z') = g(z)$.*

To see why, hold each current winner fixed. For any rival a, its gap can decrease by at most $\left\|w_{c_p} - w_a\right\| \left\|z'_p - z_p\right\| \le \Gamma \|z' - z\|$. If that decrease is smaller than the weakest original gap, no rival can overtake a winner. This is the affine-head form of margin-based robustness (Hein and Andriushchenko, 2017; Tsuzuku et al., 2018); the full proof is in Appendix A.

Geometrically, the lemma places an open ball inside the current answer's decoding region. That region can extend much farther in some directions, so d need not be its exact boundary distance. The lemma also does not identify the correct answer: a wrong board can occupy a robust region. It tells us what must happen for an answer to *change*. Along a trajectory we write $m_t = m(z_t)$.

3.2 Contraction turns the last step into a bound on remaining travel

Fix the input and a model. Let F be the noise-free complete update, with fixed point S^* and high component z^*. Its attraction basin contains states converging to S^*; bounding their remaining travel requires more. Following EqR's residual analysis (Huang et al., 2026, Appendix A.1) and inexact fixed-point theory (Birken, 2015), assume a closed, forward-invariant set $\mathcal{K}$ containing S^*, where

$$\|F(S) - F(V)\| \le L \|S - V\|, \qquad S, V \in \mathcal{K}, \qquad 0 \le L < 1.$$

Forward invariance keeps updates in $\mathcal{K}$; contraction shrinks distances there. Within $\mathcal{K}$, if the entering step has length $r_t^S = \|S_t - S_{t-1}\|$, subsequent steps have lengths at most $Lr_t^S, L^2 r_t^S, \ldots$. Their geometric sum bounds all remaining travel.

Proposition 2 (Noise-free inclusion). *Suppose $S_{t-1}, S_t \in \mathcal{K}$, with $S_t = F(S_{t-1})$, and set $R_t = Lr_t^S/(1-L)$. If $m(z_t) > \Gamma R_t$, every future answer equals the current answer. The stronger condition $m(z_t) > 2\Gamma R_t$ gives $\|S_t - S^*\| < d(z^*)$, placing the complete state inside the attractor-centered decoding ball.*

Figure 2(a) shows a trajectory crossing decoding regions before entering the attractor's solution ball. The factor two pays for distance to that unknown center and possible loss of margin there (Appendix A). If the attractor is correct, its decoding ball is a solution ball: contraction preserves its answer before equilibrium is reached. The supervision argument in Appendix B explains when that destination is likely to be correct.

Following prediction patience (Zhou et al., 2020), *decode locking* accepts after w consecutive identical complete answers. A window wholly after solution-ball entry must be locked. Agreement can also precede a later change, so a lock alone does not prove entry. It nevertheless provides a simple stopping test alongside MMI (Table 1).

3.3 Noise changes the goal: recognizing the attractor's answer

Noise changes what convergence can promise. Even near an attractor, fresh perturbations can move a state outside its decoding ball. Requiring every subsequent state to remain inside may therefore prevent acceptance even when the current answer agrees with the attractor. Our objective becomes recognizing that answer despite persistent movement. Contraction controls the remaining transient; the margin distinguishes answers within the resulting noisy neighborhood.

What still contracts? Retain the noise-free map F, fixed point S^*, and L-contractive region $\mathcal{K}$ from Section 3.2. Over D updates, suppose $S_0, \ldots, S_D$ remain in $\mathcal{K}$ and satisfy $S_t = F(S_{t-1}) + \eta_t$, with $\|\eta_t\| \le \varepsilon$. The perturbation includes noise propagated through the entire outer update. Gaussian injections admit finite-horizon bounds with high probability when their amplification due to the dynamics is bounded (Appendix A.1). Our deterministic claims hold whenever these motion assumptions hold.

Choose a floor $\rho \ge 2\varepsilon/(1-L)$ and define

$$r_t^S = \|S_t - S_{t-1}\|, \qquad e_t^S = (r_t^S - \rho)_+, \qquad R_t^S = \frac{L}{1-L} e_t^S, \qquad (v)_+ = \max(v, 0).$$

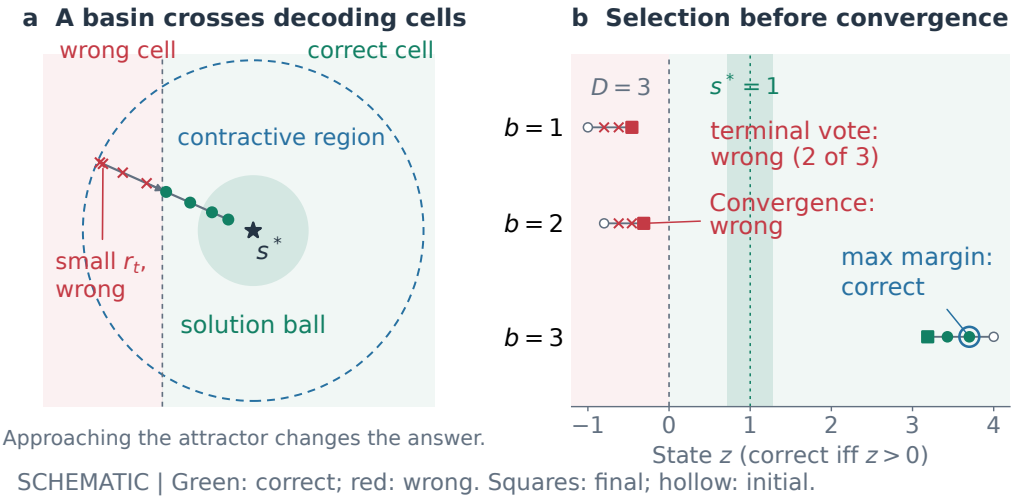

Figure 2: Geometry and selection. Left: one basin crosses decoding regions before reaching a solution ball. Right: terminal voting and minimum residual can miss a correct, larger-margin candidate. These exact affine-contraction schematics illustrate possibilities, not frequencies; all attractors shown are correct.

The removed movement is accounted for by a closed neighborhood around the attractor:

$$\mathcal{N}_\rho = \{S : \|S - S^*\| \le a\}, \qquad a = \frac{L\rho + \varepsilon}{1 - L}.$$

This conservative region need not equal the support of settled noisy states. Increasing the floor enlarges it. Here distance to a set means distance to its closest point:

$$\operatorname{dist}(S, \mathcal{N}_\rho) := \inf_{V \in \mathcal{N}_\rho} \|S - V\| = \big(\|S - S^*\| - a\big)_+.$$

Lemma 3 (Excess controls distance to the noisy neighborhood)**.** *Under the motion assumptions,* $e^S_{t+1} \le L e^S_t$ *for* $1 \le t < D$*, and*

$$\|S_t - S^*\| \le a + R^S_t, \qquad \operatorname{dist}(S_t, \mathcal{N}_\rho) \le R^S_t, \qquad 1 \le t \le D.$$

The proof combines $r^S_{t+1} \le L r^S_t + 2\varepsilon$ with the contraction bound on distance to the fixed point (Appendix A.2). Zero excess certifies neighborhood membership; positive excess only bounds a possible distance outside it.

What can the margin distinguish inside? Suppose $m^* = m(z^*) > 0$. If $a < d(z^*)$, the neighborhood fits inside the attractor's decoding ball. Otherwise competing answers may occur. How deep can a disagreement become?

Proposition 4 (Local disagreements have bounded margin)**.** *Every* $S \in \mathcal{N}_\rho$*, with high component* z*,*

$$g(z) \ne g(z^*) \quad \Longrightarrow \quad m(z) \le J_{\text{geom}}, \qquad J_{\text{geom}} = [\Gamma a - m^*]_+.$$

Moving within the neighborhood changes a winner–rival gap by at most Γa, while overturning the attractor's answer first consumes at least m^*. This is a demanding worst-case bound: it covers every point in the neighborhood, including points the dynamics never visit.

Deep supervision motivates a less demanding, probabilistic separation. Lemma 3 bounds distance to the noisy neighborhood by R^S_t, which contracts while the motion assumptions hold. A trajectory that enters stays there under these assumptions (Appendix A.2). When supervision covers these later states, training repeatedly encounters states in or near that neighborhood. Rewarding the target there encourages a correct winner to become more confident and an incorrect winner to lose its lead over the target. Confident errors are especially costly under the supervised loss. With a correct attractor

and suitable supervision coverage, this favors broad margins for correct states and shallow margins for disagreements. Shared updates can interfere, however; training does not guarantee this separation. Appendix A.2 gives the precise loss–margin argument and its conditions. We therefore assume rare deep disagreements explicitly, rather than identifying a fitted allowance with J_{geom}.

Theorem 5 (MMI recognizes the attractor's answer with high probability)**.** *Fix the checkpoint and $J \geq 0$ before drawing an input and its $B \times D$ trajectory grid. For restart b, write S_b^* for its attractor. Let $\mathcal{G}$ be the event that the motion assumptions hold for every restart, with $\Pr(\mathcal{G}) \geq 1 - \delta_{\mathrm{motion}}$. For each (b, t), let $\bar{S}_{b,t}$ be the closest point to $S_{b,t}$ in its attractor's neighborhood, and $\bar{z}_{b,t}$ its high component. Suppose*

$$\Pr\{\exists (b,t) : g(\bar{z}_{b,t}) \neq g(z_b^*),\ m(\bar{z}_{b,t}) > J \mid \mathcal{G}\} \leq \delta_{\mathrm{sep}}.$$

The probability that any candidate satisfying

$$\boxed{m_{b,t} > 2\Gamma R_{b,t}^S + J,}$$

disagrees with its attractor is at most $\delta_{\mathrm{motion}} + \delta_{\mathrm{sep}}$.

On $\mathcal{G}$, moving to the reference reduces each current winner's gap by at most $\Gamma R_{b,t}^S$. Acceptance therefore gives the reference the same answer and margin greater than J; disagreement requires one of the exceptional events. The simultaneous assumption covers selection across restarts and stopping times. Any policy selecting a passing candidate inherits this per-input bound. It does not bound error conditional on acceptance or cover fallback. Ground-truth correctness additionally requires a correct attractor. Full proofs appear in Appendix A.2.

The motion charge thus pays for a possible remaining transient, while the allowance excludes deep local disagreements with high probability. At zero excess, only the margin test remains. The guarantee concerns the answer returned now, not preservation under every future perturbation.

3.4 The observable MMI rule

Inspired by Theorem 5, we build an observable rule from the high-state residual $r_t = \|z_t - z_{t-1}\|$ and fitted excess $e_t = (r_t - \widehat{\rho})_+$. We estimate contraction from w recent transitions:

$$q_j = \begin{cases} e_j/e_{j-1}, & e_{j-1} > 0, \\ 0, & e_{j-1} = 0, \end{cases} \qquad \widehat{L}_t = \max_{j=t-w+1}^{t} q_j, \qquad \boxed{m_t > 2\Gamma \frac{\widehat{L}_t}{1 - \widehat{L}_t} e_t + J.} \tag{5}$$

This is Margin–Motion Inclusion (MMI), with motion charge $A_t = 2\Gamma \widehat{L}_t e_t / (1 - \widehat{L}_t)$. The theorem supplies the geometric motivation; Section 5 provides a direct calibration guarantee for this observable rule. Applying the attractor-agreement guarantee additionally requires the theorem's complete-state motion and separation assumptions. Figure 4 illustrates the margin budget.

The principal setting is $w = 3$, requiring $w + 1$ residual observations. Invalid measurements or $\widehat{L}_t \geq 1$ cause abstention. Every zero denominator gives ratio zero, including positive-over-zero; zero motion charge leaves $m_t > J$ (Appendix E).

4 When no test passes: keep the strongest answer

Stopping and fallback answer different questions. MMI and decode locking decide whether to return now. If the chosen test never accepts, we must still choose among the answers generated. Terminal voting favors common endpoints and convergence favors small motion; either can miss a stronger intermediate candidate (Figure 2(b)). We retain the largest minimum margin over *all visited states*, favoring the answer whose weakest position is best protected. With a frozen head, margin and decoder radius have identical rankings. Maximum-confidence intermediate selection has recurrent precedents (Bansal et al., 2022); our score needs no additional verifier.

One policy, either acceptance test. Certify-then-escalate (CTE) runs restarts in a fixed order, stores each new maximum-margin candidate, and checks the chosen test: MMI or decode locking. It returns the first passing candidate; if none passes within BD updates, it returns the stored candidate,

with earliest-visit ties (Algorithm 1). Acceptance can avoid the current trajectory's remaining depth and all later restarts. Fallback consumes the full budget.

The supervision argument in Section 3.3 also motivates this ranking. Appendix C derives its idealized training direction and shared-update conditions, drawing on neural-collapse geometry and gradient coupling (Papyan et al., 2020; Zhu et al., 2021; Jacot et al., 2018).

5 Measuring the floor and choosing the allowance

The objective is to measure persistent movement, then reserve margin against wrong answers that the motion charge does not exclude. Floor and allowance inputs are disjoint. We freeze the scoring rule before fitting either allowance; evaluation labels never set a threshold.

Floor: how much do repeated futures differ? Copy a complete high/low state into eight independent noisy continuations. At late lags, subtract their mean high-state increment and compute the sample variance of the vector deviations. Average, then take the square root. This removes shared drift and measures propagated RMS variation $\widehat{\rho}$, rather than a uniform movement bound (Appendix D.1).

Allowance: what would block a wrong early answer? A wrong candidate can pass only if its margin exceeds the motion charge by more than J. On each labeled calibration input, take the largest positive surplus $m - A$ among all eligible, valid, wrong candidates across the full trajectory grid, or zero if none exists. Include wrong candidates even when their answer is unchanged or they occur at the terminal update. Each input contributes one score, keeping its dependent states and restarts together.

Sort the n scores and take rank $k = \lceil (n+1)(1-\alpha) \rceil$, using $J = \infty$ if $k > n$. Fast uses $\alpha = .05$ and reserve .01; thus $J_{\text{reserve}} \geq J_{\text{fast}}$. Under exchangeability with the scoring rule fixed, this controls wrong early commitment on a new input at level α, marginally over calibration and new-input randomness (Lei et al., 2018; Angelopoulos and Bates, 2021). It excludes fallback errors and is not precision conditional on acceptance. The geometry motivates the surplus; this calibration controls its observable correctness target without verifying the unobserved attractor assumptions. Appendix D.2 gives the proof.

6 Experiments: fewer updates, fewer missed answers

Datasets and protocol. All Sudoku inputs come from the released *Sudoku-Extreme* test split. Maze uses the released 30×30 unique-path benchmark. Input-only hash ranking selects 4,000 of the 422,786 unique Sudoku test inputs; all 1,000 Maze test inputs enter its study pool. Each pool has a frozen 20% calibration / 80% evaluation split. Sudoku assigns 128 inputs to floor measurement, 672 to allowance fitting, and 3,200 to evaluation; Maze assigns 100, 100, and 800.

Five random seeds change the initial states and subsequent noise; inputs, checkpoints, and calibrated parameters stay fixed. Each seed uses a noise scale of .01. We sweep $B \in \{1, 2, 4, 8, 16, 32\}$ and $D \in \{4, 8, 16, 32, 64\}$. Each allowance is fitted once at $B = 32, D = 64$ and reused for all smaller prefixes. Results report the mean and standard deviation across these five runs; the $\pm$ values show variation due to randomness on the same test inputs. NFE counts outer updates on all inputs, including failures and full fallback. Update-reduction factors average per-seed cost ratios; they are not measured wall-clock speedups.

MMI and its ablations select from the same recorded trajectories, charging only executed updates. Convergence uses the final three logit-motion residuals. We evaluate EqR's ACT baseline using the released implementation (Huang et al., 2026). Replay holds trajectories fixed; it does not reproduce the original paper's batching or random-number sequence (Appendix E).

6.1 Early recognition saves computation, with task-dependent tradeoffs

Table 1 reports accuracy, unconditional NFE and wrong acceptances. Both MMI profiles match the Sudoku oracle in every seed, with no wrong early returns among 15,745 accepted seed–input

outcomes; full-budget fallbacks contribute much of the work. On Maze, fast MMI uses 16.4× fewer updates than terminal voting and reduces errors from 2.275% to 0.575%. Decode locking uses fewer updates, but is slightly less accurate.

Table 1: Results at $B = 32, D = 64$, noise .01: mean ± sample SD over five seeds, using 3,200 Sudoku and 800 Maze inputs per seed. Wrong accepts are incorrect early returns divided by all inputs. NFE includes failures and full fallback. Grid uses the better aggregate vote/convergence accuracy.

Sudoku-Extreme	Accuracy (%)	Mean NFE	Wrong accepts (%)
Oracle	98.419 ± 0.075	—	—
EqR (Grid)	98.406 ± 0.086	2048.00 ± 0.00	—
EqR (ACT)	86.587 ± 0.218	311.14 ± 0.47	—
Decode lock (3)	98.413 ± 0.081	58.27 ± 1.64	0.006 ± 0.014
MMI (fast)	98.419 ± 0.075	58.96 ± 1.66	0.000 ± 0.000
MMI (reserve)	98.419 ± 0.075	58.96 ± 1.66	0.000 ± 0.000
Maze	Accuracy (%)	Mean NFE	Wrong accepts (%)
Oracle	99.775 ± 0.056	—	—
EqR (Grid)	97.725 ± 0.137	2048.00 ± 0.00	—
EqR (ACT)	91.200 ± 0.473	394.49 ± 2.27	—
Decode lock (3)	99.350 ± 0.105	41.55 ± 2.12	0.425 ± 0.143
MMI (fast)	99.425 ± 0.112	125.18 ± 3.42	0.350 ± 0.105
MMI (reserve)	99.650 ± 0.105	633.38 ± 4.32	0.075 ± 0.068

Figure 3 compares the high-accuracy frontier. At each budget, Grid uses the better aggregate vote/convergence result. Table 2 and the supplementary results retain the components.

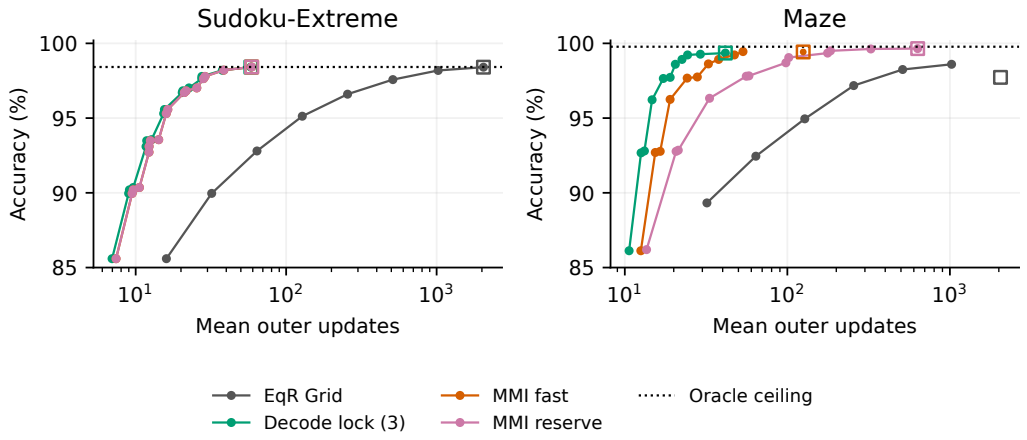

Figure 3: Five-seed mean accuracy versus outer updates. Hollow squares mark the corresponding $B = 32, D = 64$ results in Table 1, including dominated points. MMI uses the same full-grid allowances at every budget. The dotted line is the full-grid oracle ceiling.

One calibration, two operating points. The Maze reserve reaches 99.650% at 633.38 updates, reducing fast MMI's error rate by 39.1%, with a 0.125 percentage-point oracle gap. On Sudoku, all 672 calibration scores are zero: every eligible, valid wrong candidate already satisfies $m \leq A$. Both allowances therefore equal zero and give the same rule. The motion charge remains essential: a zero allowance means no additional margin reserve was required by these calibration scores, not that noise or future errors disappear. Given the sampled surplus Sudoku states, zero allowance then implies that the noisy neighborhood is probably included in the decoding ball.

What does the motion charge contribute? Omitting A_t leaves $m_t > J$. Keeping the allowances and MMI's other checks drops Sudoku accuracy from 98.419% to 75.269% for both profiles; Maze fast drops from 99.425% to 98.600%, and reserve from 99.650% to 99.625%. A threshold fitted for $m - A$ therefore need not suit m. Calibrating margin alone at the same reserve level, without residual-window checks (Appendix D.3), gives 98.381% at 55.77 updates on Sudoku and 99.525% at 753.17 on Maze. Motion is especially useful for the fast rule; its reserve benefit on Maze is modest.

6.2 Preserving the largest margin improves answer selection

Table 2 compares return rules on the complete grid. Maximum margin reaches the Sudoku oracle and improves Maze accuracy over both endpoint selectors, supporting the retention of intermediate answers for fallback.

Table 2: Complete-grid selection on Table 1's inputs and seeds (mean $\pm$ sample SD). Vote and convergence select endpoints; maximum margin accesses every update. The comparison therefore includes candidate availability.

Selector	Sudoku-Extreme (%)	Maze (%)
Vote	98.025 ± 0.041	97.725 ± 0.137
Convergence	98.406 ± 0.086	94.600 ± 0.347
Maximum margin	98.419 ± 0.075	99.500 ± 0.177

6.3 How much of the oracle gap remains?

Maximum-margin selection closes the Sudoku oracle gap (Tables 1–2); Maze reserve leaves 0.125 percentage points. Correct candidates can disappear before an endpoint, lose selection, or remain beyond an incorrect early commitment. Counterfactual grids expose these opportunities while charging stopped policies only for executed work.

For a correct return, local recognition delay is $T_A - \tau^*_{b_A}$, where b_A is the restart supplying the returned candidate and T_A is the acceptance update or D for fallback, including retrieval of an earlier candidate. Pairing each MMI profile with grid vote on inputs both answer correctly, fast and reserve have mean delays of 2.59 ± 0.01 updates on Sudoku, and 6.86 ± 0.19 and 19.06 ± 0.27 on Maze (five-seed mean $\pm$ sample SD). Grid-vote delays are 57.39 ± 0.10 on Sudoku and 54.74 ± 0.07 on Maze in these paired comparisons. Cohorts differ across comparisons; their sizes are in Appendix E. Unconditional costs and failures remain the primary comparison.

7 Discussion and relation to prior work

Overthinking, confidence exits, and prediction patience motivate our stopping and selection comparisons (Kaya et al., 2019; Bansal et al., 2022; Zhou et al., 2020). CALM calibrates exits, and repeated-sampling studies separate generation from selection (Schuster et al., 2022; Brown et al., 2024). We connect these decisions to decoder geometry and remaining motion (Appendix F).

Recognizing and retaining EqR's answers saves updates and improves accuracy, with strong simple competitors. Correctness also depends on learned destinations; wall-clock gains and transfer beyond these checkpoints remain open. Recognizing what a model has already solved complements improving what it can solve.

Reproducibility statement

The appendices provide complete proofs, calibration estimators, and policy definitions. The source package contains the inference and replay code, a pinned Apptainer environment, and instructions for generating trajectories, fitted allowances, per-input outcomes, tables, and figures from versioned public checkpoints and datasets. Appendix E identifies the cohorts and candidate pools.

AI use statement

In this work, we used generative AI tools to assist in writing and reviewing proofs, provide feedback on research methodology or experiments, implement methods, prepare data for analysis. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, generate synthetic datasets, translate research material and Interpret results; qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and revise scientific figures, develop software and research artifacts, search for and summarize relevant literature, draft and edit manuscript text, improve readability and structure, format references, and suggest titles and keywords. All AI-assisted work was thoroughly reviewed by the authors, including the mathematical arguments, implementations, experimental outputs, figures, and references. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. *International Conference on Learning Representations*, 2024.

Cem Anil, Ashwini Pokle, Kaiqu Liang, Johannes Treutlein, Yuhuai Wu, Shaojie Bai, J. Zico Kolter, and Roger B. Grosse. Path Independent Equilibrium Models Can Better Exploit Test-Time Computation. *Advances in Neural Information Processing Systems 35*, 2022.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. *Advances in Neural Information Processing Systems 32*, 2019.

Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing Equilibrium Models by Jacobian Regularization. *ICML, PMLR 139:554–565*, 2021.

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to Ponder. *8th ICML Workshop on Automated Machine Learning; arXiv:2107.05407*, 2021.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end Algorithm Synthesis with Recurrent Networks: Extrapolation without Overthinking. *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Philipp Birken. Termination criteria for inexact fixed-point schemes. *Numerical Linear Algebra with Applications 22(4):702–716*, 2015.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. *Advances in Neural Information Processing Systems, 32*, 2019.

C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory, 16(1):41–46*, 1970.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. *Advances in Neural Information Processing Systems 30*, 2017.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Alex Graves. Adaptive Computation Time for Recurrent Neural Networks. *arXiv preprint arXiv:1603.08983*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *ICML, PMLR 70:1321–1330*, 2017.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. *Conference on Language Modeling (COLM)*, 2025.

Matthias Hein and Maksym Andriushchenko. Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation. *Advances in Neural Information Processing Systems 30*, 2017.

Benhao Huang, Zhengyang Geng, and Zico Kolter. Equilibrium Reasoners: Learning Attractors Enables Scalable Reasoning. *International Conference on Machine Learning (ICML)*, 2026.

Roel Hulsman. Distribution-Free Finite-Sample Guarantees and Split Conformal Prediction. *arXiv:2210.14735*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in Neural Information Processing Systems, 31*, 2018.

Alexia Jolicoeur-Martineau. Less is More: Recursive Reasoning with Tiny Networks. *arXiv preprint arXiv:2510.04871*, 2025.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitraș. Shallow-Deep Networks: Understanding and Mitigating Network Overthinking. *ICML, PMLR 97:3301–3310*, 2019.

Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, PMLR 38:562–570*, 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association, 113(523):1094–1111*, 2018.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *International Conference on Learning Representations (ICLR)*, 2024.

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. *International Conference on Learning Representations*, 2020.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences, 117(40):24652–24663*, 2020.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the Edge of Numerical Stability. *International Conference on Learning Representations*, 2025.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident Adaptive Language Modeling. *Advances in Neural Information Processing Systems 35*, 2022.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The Right Tool for the Job: Matching Model and Instance Complexities. *ACL 2020, pp. 6640–6651*, 2020.

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks. *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Parameters for Reasoning. *International Conference on Learning Representations (ICLR)*, 2025.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research, 19(70):1–57*, 2018.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. *23rd International Conference on Pattern Recognition (ICPR), pp. 2464–2469*, 2016.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks. *Advances in Neural Information Processing Systems 31*, 2018.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations (ICLR)*, 2023.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical Reasoning Model. *arXiv preprint arXiv:2506.21734*, 2025.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2246–2251*, 2020.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. *Advances in Neural Information Processing Systems, 33*, 2020.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. BERT Loses Patience: Fast and Robust Inference with Early Exit. *Advances in Neural Information Processing Systems 33*, 2020.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A Geometric Analysis of Neural Collapse with Unconstrained Features. *Advances in Neural Information Processing Systems, 34:29820–29834*, 2021.

A DECODER GEOMETRY AND CONDITIONAL STOPPING GUARANTEES

We first establish the decoder geometry and noise-free inclusion, then prove recognition of the attractor's answer under persistent noise. Attractor correctness is a separate training question, developed in Appendix B.

Complete dynamics and the decoded state. Fix the input and parameters. The noise-free outer update is $S_{t+1} = F(S_t)$ for the complete carry $S_t = (Z_t^H, Z_t^L)$; the decoder reads $z_t = Z_t^H$. Euclidean product norms make this projection nonexpansive. With fixed affine logits $h_{p,a} = w_a^\top z_p + b_a$, write $c_p(z) = \arg\max_a h_{p,a}$ with deterministic ties, $g(z) = (c_p(z))_{p=1}^P$, and

$$m_p(z) = h_{p,c_p(z)} - \max_{a \neq c_p(z)} h_{p,a}, \qquad m(z) = \min_p m_p(z),$$
$$\Gamma = \max_{a \neq b} \|w_a - w_b\|_2 > 0, \qquad d(z) = m(z)/\Gamma.$$

There are P evaluated positions and K symbols. The radius d protects the entire decoded answer; it need not equal its exact boundary distance. Write $m_t = m(z_t)$ and let $y = (y_p)_{p=1}^P$ be the target.

Definition A1 (Basin and contractive neighborhood)**.** *For a fixed point S^* of F, its attraction basin is*

$$\mathcal{B}(S^*) = \{S : \lim_{k\to\infty} F^k(S) = S^*\}.$$

A contractive neighborhood used below is a closed set $\mathcal{K}$ containing S^*, with $F(\mathcal{K}) \subseteq \mathcal{K}$ and

$$\|F(S) - F(V)\| \le L\|S - V\| \quad (S, V \in \mathcal{K}), \qquad 0 \le L < 1.$$

Every such $\mathcal{K}$ lies in $\mathcal{B}(S^*)$, since $\|F^k(S) - S^*\| \le L^k\|S - S^*\|$. The converse fails: for example, $F(s) = s/(1 + |s|)$ attracts every real state to zero, but its derivative there is one. We do not assume all trajectories enter contractive neighborhoods.

The decoder bound below is the affine margin–Lipschitz principle (Tsuzuku et al., 2018, Proposition 2). Its short proof fixes the state norm and the whole-answer minimum needed here.

Lemma A2 (A margin protects a decoding ball)**.** *If $m(z) > 0$, then every z' satisfying $\|z' - z\| < d(z)$ has the same complete decode as z. Consequently, the complete-state ball $\{S : \|S - S^*\| < d(z^*)\}$ has the attractor's decode, where $z^* = Z^{H,*}$. It is a solution ball only if that decode equals the target answer.*

Proof. Keep the current winner $c_p = c_p(z)$ fixed. For each competitor $a \neq c_p$,

$$h_{p,c_p}(z') - h_{p,a}(z') = h_{p,c_p}(z) - h_{p,a}(z) + (w_{c_p} - w_a)^\top (z'_p - z_p)$$
$$\ge m(z) - \Gamma\|z' - z\| > 0.$$

The fixed biases cancel from the gap change. All named winner-versus-competitor gaps remain positive. Projection from the complete state cannot increase distance, giving the second statement. □

The intersection of this attractor-centered decoding ball with $\mathcal{K}$ is forward invariant: contraction decreases distance to its center. A geometric ball outside an invariant contractive set has no such dynamical guarantee merely because its initial decode is constant.

Under contraction, successive step lengths are bounded by a geometric series. Comparing that series with the margin proves the first stopping condition; a stronger comparison locates the state in the attractor's ball.

Proposition A3 (Noise-free residual inclusion)**.** *Suppose $S_{t-1}, S_t \in \mathcal{K}$ and $S_t = F(S_{t-1})$. Define the* complete-state *entering residual and its geometric tail by*

$$r_t^S = \|S_t - S_{t-1}\|, \qquad R_t = \frac{L}{1 - L} r_t^S.$$

Then $\|S_{t+k} - S_t\| \le R_t$ for all $k \ge 1$, and $\|S_t - S^*\| \le R_t$. In particular:

1. *If $m(z_t) > \Gamma R_t$, every future decode equals the current decode.*

2. *If $m(z_t) > 2\Gamma R_t$, then S_t lies inside the attractor-centered decoding ball of Lemma A2.*

Either statement implies correctness only if the attractor's answer is correct.

Proof. Invariance and contraction give $r^S_{t+j} \le L^j r^S_t$ for $j \ge 1$. Therefore

$$\|S_{t+k} - S_t\| \le \sum_{j=1}^{k} r^S_{t+j} \le r^S_t \sum_{j=1}^{k} L^j \le R_t.$$

Taking $k \to \infty$ proves the distance-to-attractor bound. The first part follows directly from Lemma A2. For the second, keep the current winners fixed at S^*. Their gaps are at least $m(z_t) - \Gamma R_t > \Gamma R_t \ge 0$, so they are also the attractor's winners, and $m(z^*) \ge m(z_t) - \Gamma R_t > \Gamma R_t$. Thus $\|S_t - S^*\| \le R_t < m(z^*)/\Gamma$. □

The factor two pays for locating the current state inside a ball centered at the unknown attractor. It is not needed for the sharper direct future-path test in part 1, and is not another factor for bounding two logits: Γ already bounds their difference. Replacing r^S_t by the recorded high-only residual requires an additional valid relation between those quantities.

A.1 From internal noise to an outer residual floor

Our objective is to turn bounds on internal noise injections into the complete-state perturbation bound used by Lemma 3. Its proof in Appendix A.2 derives excess contraction; the factor two in its residual floor bounds the difference between successive outer perturbations. Here we justify those outer bounds, including finite-budget Gaussian noise. All norms concern the complete recurrent state.

Bounding noise through internal updates. Write $F = G_q \circ \cdots \circ G_1$ for operations on the complete carry, including both changed and retained blocks. Their Lipschitz bounds must include cross-dependence. Split at noise-injection locations so execution is $\widetilde{V}_i = G_i(\widetilde{V}_{i-1}) + \xi_i$, with $\widetilde{V}_0 = S$. Noise before a refinement is represented by an earlier injection followed by that refinement. The comparison path is $V_i = G_i(V_{i-1})$, $V_0 = S$.

Let $\mathcal{K}_0, \ldots, \mathcal{K}_q$ contain these intermediate carries. Assume G_i is C_i-Lipschitz on $\mathcal{K}_{i-1}$ and, for chosen nonnegative caps ϵ_i,

$$G_i(\mathcal{K}_{i-1}) + \overline{B}(0, \epsilon_i) \subseteq \mathcal{K}_i, \qquad \mathcal{K}_q \subseteq \mathcal{K}_0 = \mathcal{K}. \tag{6}$$

Here $\overline{B}$ is a closed Euclidean ball, restricted to perturbed coordinates if applicable. These inclusions retain both paths in the Lipschitz domains and permit repeated updates. With global bounds the domains can be the whole state space; noiseless retention alone is insufficient.

Lemma A4 (Propagation through an outer update)**.** *For $S \in \mathcal{K}$, under Equation* (6)*, if $\|\xi_i\| \le \epsilon_i$, then the effective outer perturbation satisfies*

$$\|\widetilde{V}_q - F(S)\| \le \varepsilon, \qquad \varepsilon = \sum_{i=1}^{q} \epsilon_i \prod_{j=i+1}^{q} C_j. \tag{7}$$

An empty product equals one.

Proof. Set $a_i = \|\widetilde{V}_i - V_i\|$. Since both preceding states lie in $\mathcal{K}_{i-1}$, $a_i \le C_i a_{i-1} + \epsilon_i$, with $a_0 = 0$. Expanding this recurrence gives Equation (7). □

Internal noise can be amplified even when the outer map contracts. For $G_1(s) = 0$ and $G_2(v) = Av$, $A > 0$, the composition $F(s) = 0$ has contraction factor zero, yet noise ξ between the maps produces error $A\xi$.

Why Gaussian noise admits finite-budget bounds. Gaussian noise has no deterministic cap. Over a finite budget, suitable caps hold with high probability. Consider at most $M \geq 1$ outer updates, each with at most $q \geq 1$ injection locations ($M = BD$ for a full grid). Conditional on the history before injection, suppose the noise at location i is $\xi = \sigma_i E_i g$, where $g \sim \mathcal{N}(0, I_{d_i})$, $\|E_i\|_{\mathrm{op}} \leq 1$, with fixed scale bounds $\sigma_i \geq 0$ and positive integer dimensions d_i. The embedding E_i may depend on that history. Choose $0 < \delta < 1$ and set

$$u = \log(Mq/\delta), \qquad \epsilon_i = \sigma_i\big(\sqrt{d_i} + \sqrt{2u}\big). \tag{8}$$

With probability at least $1 - \delta$, every injection respects its cap. Indeed, for $X = \|g\|^2$, its Gaussian moment-generating function gives

$$\log \mathbb{E}e^{\lambda(X - d_i)} = -\lambda d_i - \tfrac{d_i}{2}\log(1 - 2\lambda) \leq \frac{d_i \lambda^2}{1 - 2\lambda}, \qquad 0 < \lambda < \tfrac{1}{2}.$$

Chernoff's inequality with $\lambda = \sqrt{u}/(\sqrt{d_i} + 2\sqrt{u})$ yields $\Pr\{X > d_i + 2\sqrt{d_i u} + 2u\} \leq e^{-u}$, and hence $\Pr\{\|g\| > \sqrt{d_i} + \sqrt{2u}\} \leq e^{-u}$. The same bound holds conditionally at each call. Taking expectations and a union bound over at most Mq calls proves the claim; independence across calls is unnecessary under these conditional assumptions.

If Equation (6) holds for these caps, the same event supplies Equation (7) throughout the chosen horizon, provided all initial carries lie in $\mathcal{K}$. If the outer map is also L-contractive on $\mathcal{K}$ with $L < 1$, this event supplies Lemma 3's residual floor. Random initialization outside $\mathcal{K}$ needs an additional failure term. Without domain retention, the result holds only up to exit; noise caps alone do not prevent exit. Unbounded Gaussian noise cannot supply a fixed deterministic cap for an infinite run.

What the floor controls. Suppose also that F is L-contractive on $\mathcal{K}$, $0 \leq L < 1$. For a noisy continuation from S_t and its noiseless reference $F^h(S_t)$, the same event implies, for every covered h,

$$\|S_{t+h} - F^h(S_t)\| \leq \varepsilon \frac{1 - L^h}{1 - L}, \qquad \|S_{t+h} - S^*\| \leq L^h \|S_t - S^*\| + \varepsilon \frac{1 - L^h}{1 - L}, \tag{9}$$

when $S^* \in \mathcal{K}$ is a fixed point. Iterate $a_{h+1} \leq L a_h + \varepsilon$ from $a_0 = 0$ for the first bound, and use contraction toward S^* for the second. The displacement radius $\varepsilon/(1 - L)$ differs from the residual floor $2\varepsilon/(1 - L)$. These bounds imply neither noisy convergence nor attractor correctness; the fitted high-state RMS floor measures typical variation, not these bounds.

A.2 Recognizing the attractor's answer under persistent noise

Our objective is to justify the two parts of MMI's test: the motion charge pays for distance to a noisy neighborhood, and the allowance separates answers within it. The conclusion is agreement with the attractor at the accepted state. It does not require every future noisy state to preserve that answer, or place the current state inside the attractor's decoding ball. We prove each statement of Section 3.3 below.

Assumptions and the meaning of distance. Fix an input and checkpoint. The complete state is $S = (Z^H, Z^L)$, its decoded component is $z = Z^H$, and $F(S^*) = S^*$. Use the Euclidean product norm; hence $\|z - v\| \leq \|S - V\|$ for the corresponding high components. For a finite horizon T, assume $S_0, \ldots, S_T, S^* \in \mathcal{K}$, F is L-contractive on $\mathcal{K}$, $0 \leq L < 1$, and

$$S_t = F(S_{t-1}) + \eta_t, \qquad \|\eta_t\| \leq \varepsilon \quad (1 \leq t \leq T).$$

Retention of the actual noisy path in $\mathcal{K}$ is an assumption; invariance of the noise-free map alone does not supply it. The affine decoder and $\Gamma > 0$ are those of Equation (4). For a floor $\rho \geq 0$, write

$$r_t^S = \|S_t - S_{t-1}\|, \qquad e_t^S = (r_t^S - \rho)_+, \qquad R_t^S = \frac{L}{1 - L} e_t^S,$$

$$a = \frac{L\rho + \varepsilon}{1 - L}, \qquad \mathcal{N}_\rho = \{V : \|V - S^*\| \leq a\}.$$

For any nonempty set $\mathcal{N}$, its distance from a point is

$$\operatorname{dist}(S, \mathcal{N}) := \inf_{V \in \mathcal{N}} \|S - V\|.$$

Thus distance measures the shortest movement needed to enter the set, not distance to its center. For our closed ball,

$$\operatorname{dist}(S, \mathcal{N}_\rho) = (\|S - S^*\| - a)_+. \tag{10}$$

Indeed, if S is in the ball, choose $V = S$. Otherwise the triangle inequality gives $\|S - V\| \geq \|S - S^*\| - a$ for every V in the ball; equality is attained at $\bar{S} = S^* + a(S - S^*)/\|S - S^*\|$. We call this closest point the *reference state*; inside the ball we set $\bar{S} = S$. This definition also covers $a = 0$.

Proof of Lemma 3: excess and distance. First compare consecutive noisy updates:

$$\begin{aligned} r_{t+1}^S &= \|F(S_t) - F(S_{t-1}) + \eta_{t+1} - \eta_t\| \\ &\leq L r_t^S + 2\varepsilon. \end{aligned}$$

If $\rho \geq 2\varepsilon/(1 - L)$, then

$$e_{t+1}^S \leq [L(r_t^S - \rho) + 2\varepsilon - (1 - L)\rho]_+ \leq [L(r_t^S - \rho)]_+ = L e_t^S.$$

This proves excess contraction for $1 \leq t < T$, including zero excess. Next use the fixed-point identity and contraction:

$$\|S_t - S^*\| \leq \|S_t - F(S_t)\| + \|F(S_t) - F(S^*)\| \leq \|S_t - F(S_t)\| + L\|S_t - S^*\|.$$

Rearranging and substituting the noisy update yields

$$\begin{aligned} (1 - L)\|S_t - S^*\| &\leq \|S_t - F(S_t)\| \\ &= \|F(S_{t-1}) - F(S_t) + \eta_t\| \\ &\leq L r_t^S + \varepsilon \leq L e_t^S + L\rho + \varepsilon. \end{aligned}$$

Dividing by $1 - L$ gives $\|S_t - S^*\| \leq R_t^S + a$. Equation (10) now gives $\operatorname{dist}(S_t, \mathcal{N}_\rho) \leq R_t^S$. This distance conclusion only needs $\rho \geq 0$; the larger lower bound on ρ is needed for excess contraction. Once $S_t \in \mathcal{N}_\rho$, the same motion assumptions give

$$\|S_{t+1} - S^*\| \leq L\|S_t - S^*\| + \varepsilon \leq La + \varepsilon \leq a,$$

because $(1 - L)a = L\rho + \varepsilon \geq \varepsilon$. Thus a trajectory that enters the neighborhood stays there through the assumed horizon. A shrinking distance bound alone need not imply entry at a finite update. □

These are pathwise statements. They need no independence, centering, Gaussianity, or stationarity once the bounds hold. Finite-horizon noise bounds from Appendix A.1 can supply a perturbation event, but do not themselves establish contraction or path retention. The neighborhood is a conservative geometric region, not a claim about stationary support. At the smallest contraction-valid floor its radius is $\varepsilon(1 + L)/(1 - L)^2$, which can exceed $\varepsilon/(1 - L)$. Increasing ρ removes more movement from the excess while enlarging the region for which the allowance must account. Zero excess certifies membership; positive excess does not prove nonmembership.

The reference need not be visited by the dynamics or lie in $\mathcal{K}$. The distance calculation uses the ambient ball, where the affine decoder is still defined. A training-based separation assumption must cover these reference states too. If one instead intersects the ball with $\mathcal{K}$, an additional condition such as star-shapedness of $\mathcal{K}$ about S^* ensures that radial projection remains in the intersection.

Proof of Proposition 4: the geometric allowance. Suppose $m^* = m(z^*) > 0$, $S \in \mathcal{N}_\rho$, and $g(z) \neq g(z^*)$. Choose a position p where the winners differ, and denote them by $c = c_p(z)$ and $c^* = c_p(z^*)$. At the attractor, $h_{p,c^*}(z^*) - h_{p,c}(z^*) \geq m^*$. The change in their gap is at most $\Gamma\|z_p - z_p^*\| \leq \Gamma a$, so

$$h_{p,c}(z) - h_{p,c^*}(z) \leq -m^* + \Gamma a.$$

Because c^* is one of the rivals of c, the whole-answer minimum margin satisfies

$$0 \leq m(z) \leq m_p(z) \leq h_{p,c}(z) - h_{p,c^*}(z) \leq \Gamma a - m^*.$$

When $\Gamma a - m^* < 0$, this inequality rules out disagreement entirely. Otherwise it bounds its margin. Both cases give $m(z) \leq J_{\text{geom}} := [\Gamma a - m^*]_+$. □

This is a strong requirement on an allowance because it covers *every* point of the neighborhood, even points never reached at inference. It uses the worst decoder direction and the weakest attractor

margin, so it can be conservative. It is not an estimate of the typical wrong margin. Holding a and Γ fixed, increasing m^* reduces J_{geom}. If $a < d(z^*)$, all neighborhood states agree with the attractor; at $a = d(z^*)$, disagreement requires a zero-margin tie. If $a > d(z^*)$, disagreement is possible but not necessary, since the decoding region can extend beyond its guaranteed ball. Literal membership of S_t in the open attractor-centered ball follows from $a + R_t^S < d(z^*)$, which is a different condition from answer agreement.

Why the motion charge transports a margin test. Let $\bar{S}_t$ be the reference state from the lemma, and let $\bar{z}_t$ be its high component. For each current winner $c_p = c_p(z_t)$ and each rival $c \neq c_p$, linearity gives

$$\begin{aligned} h_{p,c_p}(\bar{z}_t) - h_{p,c}(\bar{z}_t) &\geq h_{p,c_p}(z_t) - h_{p,c}(z_t) - \Gamma\|\bar{S}_t - S_t\| \\ &\geq m_t - \Gamma R_t^S. \end{aligned}$$

Consequently, if $m_t > \Gamma R_t^S + J$ with $J \geq 0$, every one of these gaps exceeds J. Thus

$$g(\bar{z}_t) = g(z_t), \qquad m(\bar{z}_t) > J. \tag{11}$$

This is the link between the observed candidate and separation inside the neighborhood. In particular, MMI's stronger condition $m_t > 2\Gamma R_t^S + J$ implies it, including when $R_t^S = 0$. The factor two preserves the noiseless comparison in the paper; it is not necessary for the present answer-agreement conclusion.

The probabilistic separation assumption. Freeze the checkpoint and the allowance before drawing an evaluation input and its complete $B \times D$ grid. Probability includes this input, its initial states, and its injected noise. Conditioning on a fixed input instead gives a within-input version of the same result. For restart b, use its associated fixed point S_b^*, region $\mathcal{K}_b$, and ball $\mathcal{N}_{\rho,b}$; common constants L, ε, ρ suffice for the statement. Let $\mathcal{G}$ be the event that the dynamics assumptions above hold for all restarts through D, and suppose $\Pr(\mathcal{G}^c) \leq \delta_{\mathrm{motion}}$. On $\mathcal{G}$ define every reference $\bar{S}_{b,t}$ as above. The separation assumption is

$$\Pr\{\exists\, b, t : g(\bar{z}_{b,t}) \neq g(z_b^*),\ m(\bar{z}_{b,t}) > J \mid \mathcal{G}\} \leq \delta_{\mathrm{sep}}. \tag{12}$$

Assume $\Pr(\mathcal{G}) > 0$ when using this conditional notation. Both δ_{motion} and δ_{sep} lie in $[0, 1]$. One may restrict the maximum to a predeclared set of eligible candidates.

In words, with high probability there is no deep disagreement among the reference states available to this input's stopping policy. This is the direction needed for acceptance: *a state that disagrees rarely has margin above J*. The converse, that a small margin implies disagreement, is neither assumed nor needed: a correct answer can also have a small margin. The assumption permits rare violations and concerns the distribution of relevant references, rather than every point in the ambient balls.

Proof of Theorem 5: combining motion and separation. Consider an arbitrary realized grid for which $\mathcal{G}$ holds and the event inside Equation (12) does not occur. If any candidate (b, t) passes $m_{b,t} > 2\Gamma R_{b,t}^S + J$, Equation (11) gives the same answer at its reference, with reference margin above J. The absence of a deep reference disagreement forces that answer to equal $g(z_b^*)$. This argument holds simultaneously for every passing candidate. Therefore the event of any passing candidate disagreeing with its attractor is contained in the union of $\mathcal{G}^c$ and a separation failure on $\mathcal{G}$.

Set $p = \Pr(\mathcal{G}^c)$. Taking probabilities gives

$$\begin{aligned} \Pr\{\text{some passing candidate disagrees with its attractor}\} &\leq p + (1-p)\delta_{\mathrm{sep}} \\ &\leq \delta_{\mathrm{motion}} + (1 - \delta_{\mathrm{motion}})\delta_{\mathrm{sep}} \\ &\leq \delta_{\mathrm{motion}} + \delta_{\mathrm{sep}}. \end{aligned}$$

The second inequality uses $p \leq \delta_{\mathrm{motion}}$ and the fact that $p + (1-p)\delta_{\mathrm{sep}}$ is nondecreasing in p. Any policy returning a passing candidate inherits this bound, regardless of dependence between steps and restarts. □

If J dominates $J_{\mathrm{geom},b}$ for every relevant restart on $\mathcal{G}$, Proposition 4 makes $\delta_{\mathrm{sep}} = 0$. A smaller calibrated allowance needs its own separation bound; the geometric proposition does not assign it one. For ground-truth correctness, also require the associated attractors to be correct. If that condition

can fail, the probability of an accepted wrong attractor contributes an additional failure term. The theorem excludes fallback returns, which have not passed the test.

The bound is a probability of *making an erroneous early commitment on a random input*, not the error fraction conditional on accepting. If early acceptance has probability $q > 0$, then

$$\Pr\{\text{attractor disagreement} \mid \text{early acceptance}\} \leq \frac{\delta_{\text{motion}} + \delta_{\text{sep}}}{q}.$$

Here q denotes only acceptance probability. High precision among accepted inputs therefore also requires adequate acceptance coverage. Likewise, if every prespecified reference has conditional failure probability at most β given $\mathcal{G}$, a union bound gives $BD\beta$ for Equation (12), unless a stronger simultaneous argument is supplied. Neither a randomly sampled-state guarantee nor an unadjusted per-state probability can be substituted for the grid-level premise.

Why supervision motivates separation. The objective here is to explain both sides of the assumption: why correct states can acquire wide margins, and why deep wrong states are expensive under supervision. These are target-label statements. They apply to agreement with an attractor when its answer is correct. They do not show that every learned attractor is correct.

The isolated-supervision result in Appendix C shows the direction supplied by training: target–rival gaps increase, widening correct margins and reducing a wrong winner's lead over the target. Deep supervision applies this pressure to intermediate states; the transfer conditions specify when this direction survives shared updates. Here we establish the complementary loss–margin relations, which require no assumptions about the optimizer.

At one supervised position, let y be the target symbol and

$$s(h) = \begin{cases} 1 + h, & h \geq 0, \\ (1 - h)^{-1}, & h < 0, \end{cases} \quad Z = \sum_c s(h_c), \quad \ell_p = -\log \frac{s(h_y)}{Z}.$$

These are the StableMax transformation, normalizer, and target loss. Let $\ell(z) = \sum_{p=1}^{P} \ell_p(z)$, the *sum*, rather than the average, over evaluated positions. Set $f(h) = \log s(h)$. Then $f'(h) = 1/(1 + |h|)$, including at zero. If $|h_{p,c}| \leq M$ for all relevant logits and the answer is wrong, choose a wrong position with winner $c \neq y_p$. Its lead over the target is at least $m(z)$, so integration gives

$$f(h_{p,c}) - f(h_{p,y_p}) \geq \frac{h_{p,c} - h_{p,y_p}}{1 + M} \geq \frac{m(z)}{1 + M}.$$

Keeping just this rival in the loss denominator proves

$$g(z) \neq y \quad \Longrightarrow \quad \ell(z) \geq \log\left(1 + \exp \frac{m(z)}{1 + M}\right). \tag{13}$$

Thus a deep wrong answer incurs a large supervised loss at bounded logit levels. The logit qualification matters for StableMax: its log transformation compresses differences at large magnitudes.

Conversely, suppose $0 < \ell(z) \leq \lambda < \log 2$. For every position and every rival $c \neq y_p$,

$$\frac{s(h_{p,c})}{s(h_{p,y_p})} \leq e^{\ell_p} - 1 \leq e^{\lambda} - 1 < 1.$$

Monotonicity of s makes every target the winner. Taking logarithms and using $f' \leq 1$ then gives

$$h_{p,y_p} - h_{p,c} \geq f(h_{p,y_p}) - f(h_{p,c}) \geq -\log(e^{\lambda} - 1), \qquad m(z) \geq -\log(e^{\lambda} - 1). \tag{14}$$

A sufficiently small loss therefore guarantees a wide correct margin, while Equation (13) makes deep errors costly. These implications concern specified states, not monotonic improvement of every state during training.

When the loss argument supplies a probability bound. Suppose all associated attractors are correct on $\mathcal{G}$, and a fixed M bounds the logits of all its reference states. Each reference that violates separation has loss exceeding $\log(1 + \exp(J/(1 + M)))$. Therefore

$$\Pr\{\text{separation failure} \mid \mathcal{G}\} \leq \frac{\mathbb{E}[\sum_{b=1}^{B} \sum_{t=1}^{D} \ell(\bar{z}_{b,t}) \mid \mathcal{G}]}{\log(1 + \exp(J/(1 + M)))}. \tag{15}$$

To prove it, bound each failure indicator by its nonnegative loss divided by the denominator, sum across the grid, and take conditional expectation. Independence is unnecessary. The sum explicitly pays for selecting a candidate from many states; this bound can be loose. For fixed B, D, J and a valid M, the bound tends to zero as the expected reference-grid loss tends to zero. Training does not automatically guarantee that population-loss limit. A per-neighborhood logit bound is $\max_{p,c}(|h_{p,c}(z_b^*)| + \|w_c\|a)$; using one M in Equation (15) requires it to dominate those bounds throughout the distribution being considered.

The expectation is over deployment reference states, not the empirical training set. Low training loss transfers only with appropriate coverage and generalization control, especially because a radial reference can be absent from training. Appendix B explains this distinction for attractor correctness and selection. No claim that supervision alone verifies the separation assumption is needed here. The deployed allowances use the wrong-surplus calibration in Appendix D.2. Under its exchangeability conditions, it controls early commitment errors directly, using labels at recorded candidates. It does not estimate δ_{sep}: the radial reference states in this theorem need not occur in those recordings. This distinction keeps the geometric explanation separate from the statistical guarantee for the observable rule.

B When supervision makes wrong attractors rare

Our objective is to connect errors penalized during training to the answers reached at inference. The necessary link is *coverage*: a wrong destination forces supervised loss only if its error appears at supervised states. We derive that connection, then account for selecting one answer from several restarts.

A wrong supervised answer costs loss. Fix the checkpoint. Let $\mathbb{Q}$ be a diagnostic law over inputs, initial states and trajectory randomness. An input has target $Y = (Y_p)_{p=1}^{n(x)}$, with $1 \le n(x) \le P$ evaluated positions. For each restart $b \le B$, specify a reference answer $A_b^\star$ before applying a stopping policy: a terminal answer or a limiting attractor's answer, with $A_b^\star = \perp$ when the specified limit does not exist. Draw a uniform restart R, independently of the complete grid, and an index U according to a stated supervision rule. Define

$$W_U = \{g(z_{R,U}) \neq Y\}, \qquad W_b^\star = \{A_b^\star \neq Y,\ A_b^\star \neq \perp\}, \qquad p_{\text{bad}} = \mathbb{Q}(W_R^\star).$$

Thus p_{bad} measures wrong-reference probability for a uniform restart, rather than the number or volume of attraction basins. Let π_{p,Y_p} be the target probability under softmax or StableMax at the sampled state, and $\ell_U = n(x)^{-1} \sum_p -\log \pi_{p,Y_p}$ the mean supervised loss on exactly those positions, and $\varepsilon_{\text{sup}} = \mathbb{E}_Q \ell_U$. Softmax and StableMax preserve logit ordering. At a wrong position a rival has at least the target's probability, so the target probability is at most $1/2$. Its loss is therefore at least $\log 2$, giving

$$\mathbf{1}_{W_U} \le \frac{n(x)\ell_U}{\log 2}, \qquad \mathbb{Q}(W_U) \le \frac{P\varepsilon_{\text{sup}}}{\log 2}. \tag{16}$$

Proposition B1 (Supervision coverage controls wrong-attractor mass)**.** *Suppose* $\mathbb{Q}(W_U \mid W_R^\star) \ge \kappa > 0$ *whenever* $p_{\text{bad}} > 0$*. Then*

$$p_{\text{bad}} \le \frac{P\varepsilon_{\text{sup}}}{\kappa \log 2}. \tag{17}$$

Proof. If $p_{\text{bad}} > 0$, then $\mathbb{Q}(W_U) \ge \mathbb{Q}(W_U \cap W_R^\star) = p_{\text{bad}} \mathbb{Q}(W_U \mid W_R^\star) \ge \kappa p_{\text{bad}}$. Apply Equation (16). The case $p_{\text{bad}} = 0$ is immediate. □

Coverage has a direct temporal interpretation. Suppose U is uniform on $\{1, \dots, T_{\text{sup}}\}$ independently of the grid and restart. Conditional on a wrong reference, suppose its answer persists from update h through T_{sup} with probability at least $1 - \delta_{\text{cov}}$, where $1 \le h \le T_{\text{sup}}$ and $0 \le \delta_{\text{cov}} < 1$. Sampling within this wrong suffix gives

$$\kappa = (1 - \delta_{\text{cov}}) \frac{T_{\text{sup}} - h + 1}{T_{\text{sup}}}. \tag{18}$$

Convergence to a positive-margin attractor supplies eventual agreement, but not entry within the supervision horizon. Random starts and noise broaden the states seen during training; they do not by themselves supply a numerical coverage bound.

Transfer to the inputs and trajectories used at inference. Let $\mathbb{P}$ be the deployment law at the same checkpoint, including calibration randomness when applicable. Define

$$\delta = \mathrm{TV}(\mathcal{L}_Q(Y, A_1^\star, \dots, A_B^\star), \mathcal{L}_P(Y, A_1^\star, \dots, A_B^\star)) .$$

Total variation bounds the change in each event of this tuple and in the expectation of its average wrong-reference indicator. Consequently,

$$p_{\mathrm{bad}}^{\mathrm{deploy}} := \frac{1}{B} \sum_b \mathbb{P}(W_b^\star) \le \frac{P\varepsilon_{\sup}}{\kappa \log 2} + \delta. \tag{19}$$

Across training checkpoints, vanishing diagnostic population loss implies vanishing wrong-reference mass if coverage stays bounded away from zero and the distribution discrepancy also vanishes. This bounds population error; it does not assert monotonic improvement at every optimizer step.

Selecting a restart changes the relevant risk. A stopping policy selects after observing trajectories. Let $\mathcal{A}$ be early acceptance, $c = \mathbb{P}(\mathcal{A}) > 0$, and b_A, T_A the accepted restart and update, with answer $\widehat{Y} = g(z_{b_A, T_A})$. Charge an undefined reference or disagreement with it to

$$G = \mathcal{A} \cap \{A_{b_A}^\star = \perp \text{ or } \widehat{Y} \neq A_{b_A}^\star\}, \qquad \mathbb{P}(G) \le \gamma.$$

Proposition B2 (From covered references to correct returns)**.** *Under the preceding conditions,*

$$\mathbb{P}(\widehat{Y} \neq Y \mid \mathcal{A}) \le \frac{Bp_{\mathrm{bad}} + \delta + \gamma}{c} \le \frac{BP\varepsilon_{\sup}/(\kappa \log 2) + \delta + \gamma}{c}.$$

No independence between restarts or acceptance decisions is required.

Proof. Outside G, an accepted wrong answer equals a defined wrong reference. Hence $\mathcal{A} \cap \{\widehat{Y} \neq Y\} \subseteq G \cup \bigcup_b W_b^\star$. A union bound under $\mathbb{Q}$ gives $\mathbb{Q}(\bigcup_b W_b^\star) \le Bp_{\mathrm{bad}}$. Transferring this single event to $\mathbb{P}$ costs at most δ. Add γ, divide by c, and apply Equation (17). □

For attractor references, Theorem 5 bounds disagreement under its motion and separation assumptions. The wrong-surplus calibration instead controls wrong acceptance directly (Proposition D1); it need not establish agreement with an attractor. Neither result controls fallback, whose contribution is $\mathbb{P}(\mathcal{A}^c, \text{fallback wrong})$.

Why coverage and selection cannot be omitted. For coverage, take $F(s) = a + \frac{1}{2}(s - a)$, $a > 0$, and $s_0 = -M$ with $M > a(2^{T_{\sup}} - 1)$. Then $s_t = a - (M + a)2^{-t} < 0$ at every supervised time. With logits $(\lambda s, 0)$ and target class two, increasing $\lambda > 0$ drives every supervised softmax or StableMax loss to zero. Yet the attractor $s^* = a$ remains wrong: its error is never supervised, so $\kappa = 0$. For selection, if exactly one of B references is wrong and always has the largest margin, choosing it has error one although a uniform restart has error $1/B$. A policy accepting only rare wrong references likewise has conditional error one. The factors B and $1/c$ are therefore necessary without further assumptions.

EqR's training recipe motivates coverage (Huang et al., 2026, Appendix A.1), but does not verify these population conditions. The argument explains what successful supervision must cover; the experiments measure returned correctness and wrong early commitments directly.

C WHY SUPERVISION MOTIVATES MARGIN SELECTION

The maximum-margin fallback favors answers whose weakest position is best protected. Here we explain how supervision can create that preference: it raises the target above its competitors. The result motivates the ranking; it does not certify the selected answer.

Following one supervised position. Freeze a bias-free head $h = Wz$ and its normalizer $\Gamma = \max_{a \neq b} \|w_a - w_b\| > 0$. For target y, let m be the top-two margin, $d = m/\Gamma$ its decoder radius, and $q = h_y - \max_{a \neq y} h_a$ the target's signed lead. If correct, $d = q/\Gamma$; if wrong, $d \le -q/\Gamma$. Thus increasing q widens a correct position's protected radius and lowers an upper bound on wrong radius.

Proposition C1 (The direction supplied by isolated supervision)**.** *Suppose the complete K-symbol head satisfies*

$$WW^\top = \alpha(I - \mathbf{1}\mathbf{1}^\top/K) + \beta\mathbf{1}\mathbf{1}^\top, \qquad \alpha > 0, \quad \beta \geq 0,$$

where I is the identity and $\mathbf{1}$ the all-ones vector. Train only this feature by $\dot{z} = -W^\top g$, where $g = \nabla_h \ell$ is the softmax or StableMax loss gradient and the dot denotes training time u. Every target–rival gap increases. Correct positions remain correct and deepen; a wrong position's radius is nonincreasing until correction.

Proof. Both losses have $\ell = \log Z - \log s(h_y)$, $Z = \sum_a s(h_a)$. Softmax uses $s(v) = e^v$; StableMax uses $1 + v$ for $v \geq 0$ and $(1 - v)^{-1}$ otherwise (Prieto et al., 2025). Differentiation gives $g_a = s'(h_a)/Z > 0$ for $a \neq y$ and $g_y = s'(h_y)/Z - s'(h_y)/s(h_y) < 0$. The head assumption cancels common-logit motion:

$$\frac{d}{du}(h_a - h_b) = -\alpha(g_a - g_b).$$

Every target–rival gap therefore increases. Since s' is nondecreasing, an ordered pair of non-target logits has a nonincreasing gap. Equal non-target logits have equal derivatives, and uniqueness of the flow preserves their order. Until correction, the wrong margin is the smaller of the winner's leads over the target and the next wrong symbol; both decrease or stay flat. In the binary case, only the target competes. Dividing by fixed Γ proves the claim. □

Two nonnegative wrong StableMax logits have equal gradients, so their margin can plateau. Neural collapse motivates the symmetric-head idealization (Papyan et al., 2020; Zhu et al., 2021); it remains an assumption on *all* output symbols.

What shared training must preserve. For a finite update, write $\Delta h = -\eta\alpha g + b\mathbf{1} + E$ with $\eta, \alpha > 0$. Here b is a common shift and E collects departures from isolated descent, including other examples, optimizer effects and trajectory changes. Let $\epsilon = \max_a E_a - \min_a E_a$. Subtracting each updated rival from the target gives

$$q' \geq q + \eta\alpha \min_{a \neq y}(g_a - g_y) - \epsilon.$$

The target's lead improves when the ideal gain exceeds this discrepancy. Kernel coupling (Jacot et al., 2018) and conflicting gradients (Yu et al., 2020) motivate retaining this discrepancy; updating the head also changes Γ.

Under this isolated flow, supervision widens a fully correct answer's minimum margin. While any position is wrong, its margin upper-bounds the answer's minimum; that minimum can still rise if attained at a correct position. At a frozen checkpoint, margin and radius rank candidates identically; the experiments test how useful this ranking is. Correct-attractor frequency requires the separate coverage argument in Appendix B.

D CALIBRATING MMI: FROM THE NOISE FLOOR TO A STOPPING ALLOWANCE

Our objective is to measure persistent movement, then choose an allowance that blocks wrong answers left unprotected by the motion charge. Floor and allowance inputs are disjoint. We fix the scoring rule before allowance fitting and freeze both allowances before evaluation replay.

D.1 FLOOR: SEPARATE SHARED PROGRESS FROM VARIATION BETWEEN FUTURES

To measure persistent variation, compare independent futures from the *same complete state*. Their shared movement cancels, leaving the variation between noisy continuations.

For each of N_0 floor-calibration inputs, retain the high/low state after 63 outer updates and copy it into $M = 8$ independent continuations. Each continuation runs 32 further updates at noise scale .01. We use $N_0 = 128$ Sudoku anchors and $N_0 = 100$ Maze anchors. For input i, branch b, and continuation

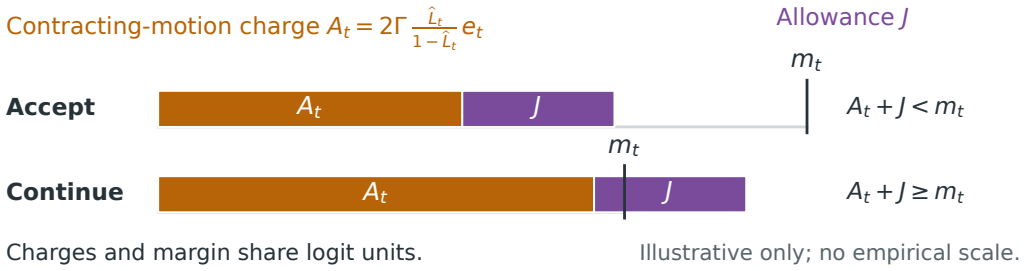

Figure 4: MMI as a margin budget. The motion charge A_t (Equation (5)) and allowance J must fit below the margin m_t. A positive remainder permits acceptance at eligible, valid windows. All terms are in logit-gap units. This is an algebraic schematic, not an empirical noise bound.

lag k, write $\delta_{i,b,k} = z_{i,b,k} - z_{i,b,k-1}$ for the high-state increment and $\bar{\delta}_{i,k} = M^{-1} \sum_b \delta_{i,b,k}$ for its mean across branches. With late lags $\mathcal{T} = \{25, \ldots, 32\}$, fit

$$\widehat{\rho}^2 = \frac{1}{N_0 |\mathcal{T}|} \sum_{i=1}^{N_0} \sum_{k \in \mathcal{T}} \frac{1}{M-1} \sum_{b=1}^{M} \left\| \delta_{i,b,k} - \bar{\delta}_{i,k} \right\|^2 . \tag{20}$$

Thus, center the eight increment vectors at each lag, sum their squared Frobenius deviations, and divide by seven. Average over anchors and lags, then take the square root. Late lags are chosen to reduce the influence of the shared starting state.

Centering vectors removes their common increment; centering residual *norms* would measure something different. Conditional on the anchor, the sample variance is unbiased for the increment covariance trace. The fitted $\widehat{\rho}$ measures propagated RMS variation, including noise amplification and answer switching. It is not a uniform movement bound or an unbiased estimator of a universal floor, and correlated lags do not supply independent calibration inputs.

D.2 Allowance: reserve against wrong margin surplus

The allowance should block wrong answers that the motion charge does not exclude, including unchanged answers and terminal candidates. Fitting uses labeled calibration inputs; deployment needs no labels.

Freeze the checkpoint, floor, decoder, window, arithmetic, candidate eligibility, maximum budget (B, D), and risk level $0 < \alpha < 1$. For input i with answer y_i, compute one score from its full grid:

$$Z_i^{\text{wrong}} = \max\left(\{0\} \cup \left\{ [m_{i,b,t} - A_{i,b,t}]_+ : \begin{array}{l} (b,t) \text{ eligible and valid,} \\ g(z_{i,b,t}) \neq y_i \end{array} \right\} \right). \tag{21}$$

Here $[v]_+ = \max(v, 0)$; eligibility and validity match deployed MMI: $w + 1$ residual observations, finite measurements, and $0 \leq \widehat{L}_t < 1$. Zero-excess and terminal candidates are included. Each input contributes one score, equal to zero when no valid wrong candidate has positive surplus. Grouping states and restarts by input preserves their dependence.

Sort the n input scores. Choose the allowance by

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad J(\alpha) = Z_{(k)}^{\text{wrong}}, \tag{22}$$

using infinity when $k > n$. Runtime uses the identical computed surplus and the strict comparison $m - A > J(\alpha)$.

Proposition D1 (Calibrated wrong-commitment control)**.** *Condition on the scoring ingredients fixed above. If the n calibration inputs and a new input, including their labels and random trajectory grids, are exchangeable, then*

$$\Pr\{\textit{CTE makes a wrong early commitment on the new input}\} \leq \alpha.$$

Proof. A wrong candidate passing the strict test implies $Z_{\text{new}}^{\text{wrong}} > J(\alpha) \geq 0$. For $k \leq n$, random tie breaking makes the new score's rank uniform among the $n+1$ exchangeable scores. Strict exceedance of the kth calibration score requires one of the largest $n+1-k$ ranks. Thus

$$\Pr\{Z_{\text{new}}^{\text{wrong}} > J(\alpha)\} \leq \frac{n+1-k}{n+1} \leq \alpha.$$

Ties only reduce strict exceedance. For $k > n$, the infinite threshold never accepts. CTE selects from these candidates, so its wrong early commitment is contained in the controlled event. □

Probability averages over allowance calibration and the new input; it excludes fallback and is not conditional on acceptance. This split-conformal rank argument (Lei et al., 2018; Angelopoulos and Bates, 2021) also has a monotone-loss interpretation (Angelopoulos et al., 2024). It controls whether *any* wrong candidate passes in the grid, which includes wrong early returns. It is not conditional on the realized allowance (Hulsman, 2022): a particular fitted threshold need not achieve an observed error rate below α.

One procedure, two risk settings. Choose $\alpha_{\text{fast}} > \alpha_{\text{reserve}}$, giving $J(\alpha_{\text{fast}}) \leq J(\alpha_{\text{reserve}})$. We use .05 and .01, respectively. These are declared risk tolerances, not predicted error rates or theoretically optimal choices. For $n = 672$ the ranks are 640 and 667; for $n = 100$, 96 and 100. A finite rank is unavailable when $\alpha < 1/(n+1)$. On the same grid, reserve cannot stop earlier than fast. Actual accuracy need not be monotone in the threshold: rejecting a correct early candidate can expose a later wrong one.

We fit once with $B = 32, D = 64, w = 3$ and reuse each allowance at all 30 evaluated breadth/depth prefixes. The score of any such prefix is at most its full-grid score, because the measurements and eligibility are causal and unchanged. Thus the same bound covers every prefix with the same allowance. Increasing the maximum budget or changing the scoring rule requires a new justification or recalibration. Repeated evaluation seeds do not increase n: a guarantee for several seeds jointly would require including that fixed collection in each calibration input's maximum.

The fitted $(\widehat{\rho}, \Gamma, J_{\text{fast}}, J_{\text{reserve}})$ values are $(15.976491, 15.939190, 0, 0)$ for Sudoku and $(26.268031, 12.331263, 3.593750, 9.453125)$ for Maze, with Γ defined in Equation (4). All 672 Sudoku calibration scores are zero: every eligible, valid wrong candidate already satisfies $m \leq A$. The resulting zero allowances retain the motion test; they do not establish that no future input can have a wrong passing candidate.

Relation to the neighborhood theorem. The allowance measures wrong margins remaining after the motion charge; at zero charge it is a wrong-margin threshold. Changing the floor changes this score and requires recalibration. Under exchangeability, the rank bound controls wrong early returns without requiring exact contraction estimates or correct attractors. It does not verify the separation assumption of Theorem 5, which concerns unobserved neighborhood references. The geometry motivates the score; calibration controls the observable correctness event.

D.3 Giving margin alone its own allowance

To calibrate a test that uses only margin, set $A = 0$ in Equation (21) and remove the residual-window checks. Each input contributes its largest margin among all wrongly decoded states, or zero if every candidate is correct. Use the complete $B = 32, D = 64$ grid, including the first and terminal updates. On the same separate allowance inputs, apply Equation (22) at $\alpha = .01$. The fitted J_{margin} is 18.25 on Sudoku and 11.5 on Maze. Deployment accepts the first candidate with $m_t > J_{\text{margin}}$; fallback is unchanged. Proposition D1 applies to this separately calibrated score. Merely removing A while retaining an MMI allowance changes the score without recalibrating it, so that proposition does not automatically apply to the omitted-charge ablation.

E Executable policies and data provenance

This appendix specifies the inputs, dynamics, and policy arithmetic needed to reproduce the evaluation. Appendix D gives the floor and allowance estimators.

Inputs and random seeds. The source splits are `sudoku-extreme-1k-aug-1000/test` (Sudoku) and, for Maze, `maze-30x30-unique-1k/test`.

SHA256 ranks canonical int32 input arrays using split seed 20260915, task and input hash. The first 4,000 Sudoku inputs and all 1,000 Maze inputs form the pools; their first 20% are calibration. Floor, allowance and evaluation inputs have disjoint exact identities. This does not rule out symmetry-equivalent puzzles. Seeds 20260916–20260920 change initialization and path noise through streams keyed by input and restart. Runs use frozen EMA checkpoints and bfloat16; the source package implements the main comparisons over all 30 budgets and the three full-budget motion-charge ablations.

Across five seeds, fast and reserve MMI each make $0/15745$ wrong/accepted returns on Sudoku; Maze counts are $14/3819$ and $3/2868$. These count seed–input outcomes, not independent puzzles. Both Sudoku paired-delay cohorts contain 3,136–3,139 inputs per seed; Maze fast contains 777–781 and reserve 779–782. Calibration is frozen before evaluation replay.

The input maximum controls whether any wrong candidate passes, even if CTE stops before visiting it. This grid-level event occurs in $10/16000$ Sudoku seed–input grids for either profile, and $130/4000$ and $49/4000$ Maze grids for fast and reserve. These rates concern the particular fitted allowances; the rank guarantee averages over calibration as well as new-input randomness.

MMI, maximum-margin selection, and the motion ablations use the same recorded input grids.

Grid convergence score. For endpoint b, write $h_{b,t,p}$ for its logit vector at position p and update t. The fixed-grid selector minimizes

$$S_b^{\text{logit}} = \frac{1}{3P} \sum_{t=D-2}^{D} \sum_{p=1}^{P} \|h_{b,t,p} - h_{b,t-1,p}\|_2 ,$$

with $D \geq 4$ in the evaluated grids. Subtraction and norms retain the iteration dtype; the released evaluator averages over positions and transitions. Exact score ties use native CUDA `argsort` in the score's iteration dtype.

Instantiating the general update. Equations (2)–(3) specify the internal composition of F_θ. HRM uses distinct modules; TRM and EqR share the refinement block. The released HRM/TRM implementations reset to fixed high/low vectors stored as buffers, not learned parameters or fresh per-trajectory draws. Diversity from input augmentation or adaptation is separate from this recurrence (Wang et al., 2025; Jolicoeur-Martineau, 2025).

The evaluated EqR checkpoints compose three cycles per outer update, interpolate with $\lambda = .95$, and draw fresh Gaussian noise at each internal call. High/low resets use truncated-normal draws. Our interpolation weight is one minus the damping coefficient called λ in the EqR paper (Huang et al., 2026). Both the five-seed evaluation and the margin diagnostic use $\beta = .01$. Internal updates can amplify noise, so the complete outer disturbance need not be additive Gaussian (Appendix A.1). All NFE counts refer to complete outer updates.

The online maximum retains only the best candidate and score, using strict improvement to preserve earliest-visit ties. Test histories reset between restarts. Fallback includes transient candidates and consumes the full budget. Algorithm 1 specifies the complete policy.

Exact test arithmetic. The decoder margin is the minimum top-two logit gap over evaluated positions. Residuals are nonnegative high-state Frobenius step norms. Compute $e_j = \max(r_j - \rho, 0)$, form w ratios from $w + 1$ observations, and assign zero whenever the denominator vanishes, even if the numerator is positive. Reject nonfinite/negative raw measurements and nonfinite ratios. Reject a recent maximum at least one; do not clip it. Implement Equation (5) as the strict computed-surplus comparison $m_t - A_t > J$, matching calibration. The first eligible step is $w + 1$; the terminal step is eligible. For example, excess history $(0, 0, 0, 10)$ gives three zero ratio scores, so the rule tests $m > J$ despite a positive latest excess. That arithmetic does not prove contraction across the jump. History $(1, 2, 0, 0)$ rejects because the earlier ratio two remains in the maximum. No agreement of decodes is tested by MMI.

Algorithm 1 CTE with a chosen stopping test and maximum-margin fallback

Require: Input x; budgets B, D; fixed decoder; test (MMI or decode lock) and its parameters

$m_{\text{best}} \leftarrow -\infty$; $a_{\text{best}} \leftarrow \varnothing$
for $b = 1, \ldots, B$ **do**
  Draw initialization $S_{b,0}$; initialize this restart's test history
  for $t = 1, \ldots, D$ **do**
    $S_{b,t} \leftarrow F_\theta(S_{b,t-1}, x; \xi_{b,t})$; decode $a_{b,t}$ and compute $m_{b,t}$
    if $m_{b,t} > m_{\text{best}}$ **then**
      $m_{\text{best}} \leftarrow m_{b,t}$; $a_{\text{best}} \leftarrow a_{b,t}$
    end if
    if the chosen test is eligible and accepts the current candidate **then**
      return $a_{b,t}$
    end if
  end for
end for
return a_{best}

Decode locking accepts w equal consecutive complete answers, first eligible at update w. The main comparison uses three states; MMI instead requires four residual observations for its three-ratio window.

E.1 Margins across the candidate population

To examine the margin ranking behind fallback (Section 4), this diagnostic uses a separate recording cohort of 1,024 Sudoku and 500 Maze inputs, each with $B = 32, D = 64$ and noise scale .01. Correct/wrong median minimum margins are 65.375/.25 on Sudoku and 9.961/.125 on Maze. Each hidden state contributes one weakest-position margin, labeled by whole-answer correctness. Separate normalization shows the correct and wrong score distributions; it does not estimate correctness probability given a margin. States from an input remain dependent.

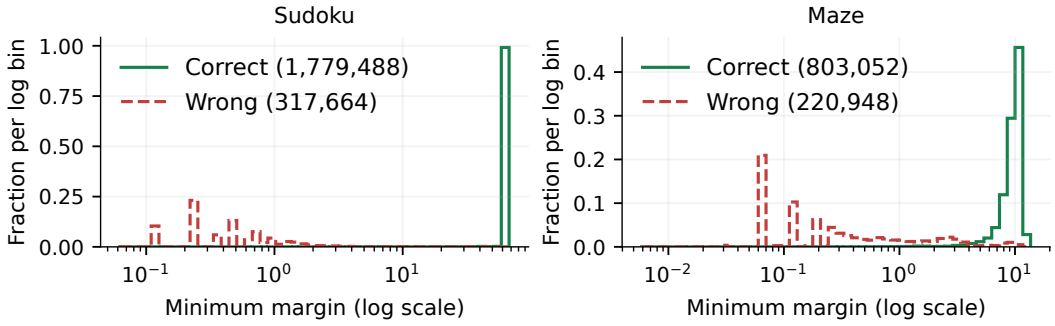

Figure 5: Why preserve a high-margin candidate? Correct whole answers (green) tend to have larger weakest-position margins than wrong ones (red), with overlap. Each class is normalized separately; the vertical axis is probability mass per log bin, not probability of correctness given margin. Zero margins, omitted from the log axis, comprise 18.29%/23.54% of wrong Sudoku/Maze states and less than .006% of correct states. Counts include dependent states.

F Relationship to prior work

Iterative reasoning and equilibrium geometry. Deep equilibrium models formulate implicit fixed points (Bai et al., 2019), while stability regularization and path independence address distinct properties of recurrent computation (Bai et al., 2021; Anil et al., 2022). Algorithmic recurrence and latent reasoning also benefit from additional computation (Schwarzschild et al., 2021; Bansal et al., 2022; Geiping et al., 2025; Hao et al., 2025). Our experiments concern EqR, with HRM and TRM providing architectural context (Huang et al., 2026; Wang et al., 2025; Jolicoeur-Martineau, 2025).

EqR's residual analysis uses contraction and an assumed correct attractor's margin; MMI uses the current margin to recognize the attractor's answer under persistent noise.

Halting and answer selection. ACT and PonderNet learn when to stop (Graves, 2016; Banino et al., 2021); confidence exits and prediction patience provide alternatives (Teerapittayanon et al., 2016; Xin et al., 2020; Zhou et al., 2020). CALM calibrates exits using probability gaps, representation saturation or learned confidence (Schuster et al., 2022). MMI instead compares an affine decoder margin against estimated remaining motion and an allowance. Maximum-confidence intermediate selection predates our work (Bansal et al., 2022); our fallback retains the largest weakest-position margin. Terminal voting, learned verifiers and process rewards use different selection signals (Wang et al., 2023; Cobbe et al., 2021; Lightman et al., 2024). Oracle analyses distinguish candidate generation from recognition (Brown et al., 2024; Schwartz et al., 2020); we measure this distinction within restarted trajectories and charge the actual depth and breadth consumed. This does not claim optimal computation allocation (Snell et al., 2025).

Robustness and supervision. Margin–Lipschitz certificates protect labels against bounded perturbations (Hein and Andriushchenko, 2017; Tsuzuku et al., 2018); inexact fixed-point analysis bounds persistent perturbations (Birken, 2015). Neither establishes that the protected label is correct. Our supervision analysis uses explicit coverage and selection conditions. Deep supervision and neural collapse motivate its local idealizations (Lee et al., 2015; Papyan et al., 2020; Zhu et al., 2021); implicit margin-bias results do not establish the same StableMax dynamics (Soudry et al., 2018; Lyu and Li, 2020; Prieto et al., 2025). Kernel coupling and gradient conflict motivate our transfer conditions (Jacot et al., 2018; Yu et al., 2020). Lazy-training theory studies when nonlinear training is approximated by its linearization (Chizat et al., 2019).

Calibration and the event being controlled. Selective prediction trades coverage against error, and confidence ranking need not calibrate correctness (Chow, 1970; Geifman and El-Yaniv, 2017; Guo et al., 2017). Split-conformal ranks yield marginal guarantees under exchangeability (Lei et al., 2018; Angelopoulos and Bates, 2021), distinct from calibration-conditional guarantees (Hulsman, 2022); conformal risk control treats more general losses (Angelopoulos et al., 2024). Both MMI allowances calibrate wrong early commitments using known answers on separate calibration inputs, at different risk levels. The guarantee excludes fallback and does not control precision conditional on acceptance. Grouping all states and restarts by input preserves their dependence in the calibration argument.