TRAJECTORY-TO-GO: FORECASTING CHANGES IN LANGUAGE MODEL PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model's prediction can change little in one step and still be far from a later output. Trajectory-to-Go (T2G) forecasts this remaining change from a short record of output and hidden-state motion. Fourteen scalar inputs describe current update size, magnitude history, direction, and confidence; small fitted readouts predict disagreement with a specified later output. On Huginn-0125 and two Ouro scales, the two-head Risk-T2G readout reduces error-ranking loss (EAURC) by 7.7–22.3% over confidence predictors with matched objectives and parameter counts. Component comparisons locate added value in history and hidden-state direction on Huginn and Ouro-1.4B, while current-update features suffice on Ouro-2.6B. Frozen forecasts retain gains on C4 and LAMBADA. Across four ordinary transformers up to 7B, trajectory features complement learned state probes: their combination improves on either alone in all eight model–lens settings. Stopping experiments separate these forecasting gains from allocation and execution benefits, where learned confidence remains competitive. T2G makes remaining output change predictable from compact motion features; the benefit of acting on that forecast depends on the task.

1 INTRODUCTION

When a language model's intermediate prediction appears stable, how much will it still change? This question arises naturally in depth-recurrent models, which repeatedly apply shared transformer blocks and can expose an output after each iteration (Geiping et al., 2025; Zhu et al., 2025). It also arises when reading ordinary transformer layers through a lens or choosing an early exit (Schuster et al., 2022; Geva et al., 2022). In both settings, we can inspect a prediction before running the remaining computation. We call such an intermediate prediction a *reading*.

Confidence describes the current prediction; one-step changes describe the transition that just occurred. Learned gates and stopping rules use these signals to allocate computation (Graves, 2016; Banino et al., 2021; Popescu et al., 2026; Logan, 2026). Yet the question is prospective: will small changes continue in the same direction, cancel each other, or give way to a larger turn? We call a reading with small recent change but substantial later discrepancy *stable but unfinished*. Local stability alone does not resolve these possibilities.

We forecast disagreement with a chosen later output, measured by KL divergence. This *terminal-fidelity forecasting* uses a later iteration or the final layer as its reference, not necessarily a converged or correct answer. Completing the model supplies training targets without human labels. The fitted predictor ranks new readings for selective use and calibrated stopping.

Trajectory-to-Go (T2G; Figure 1) summarizes the last two updates rather than only their most recent magnitude. Magnitude history tracks how the size of an update changes; directional features describe persistence and turns. We measure both in output and hidden space, where motion can remain visible even when it barely changes the current prediction. We distinguish three sources of information: current update size, its recent history, and direction. Our main readout, RISK-T2G, combines fourteen inputs in log-error and mean-error heads, with one floor coefficient fixed across models. Matched confidence controls test the full representation; staged input ablations isolate history and direction. Single-head comparisons separate this feature question from the choice of forecasting objective. Two extensions use the same trajectory representation: T2G + state adds current-state features, while TASK-T2G predicts answer changes for task-specific stopping.

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.

Looped models provide the primary test and an executable stopping setting: Huginn-0125 (3.5B) and Ouro at 1.4B and 2.6B, including Ouro's native halt gate as a baseline. Ordinary transformers test the scope of the representation: we use the same features and fitting procedure, with new model- and layer-specific heads, on Pythia-410m, Qwen1.5-0.5B and 1.8B, and Qwen2.5-7B. Across these settings, we make three contributions:

- A compact representation of output and hidden motion predicts later error beyond matched confidence controls. Staged ablations distinguish the contributions of current magnitude, history, and direction.
- The forecasting gain extends beyond the fitting corpus, and the same features remain useful with new fits at different reference horizons and along ordinary transformer layers, where trajectory and current-state information are complementary.
- We connect forecasting to selective use through KL allocation and MMLU exits, separating prediction quality, realized risk, and measured runtime against both published criteria and learned controls.

2 Related work

Looped models and adaptive computation. ACT, PonderNet, and the Universal Transformer learn recurrent halting (Graves, 2016; Banino et al., 2021; Dehghani et al., 2019). Confidence scores detect prediction errors (Hendrycks & Gimpel, 2017) and guide early exits (Kaya et al., 2019); CALM calibrates sequence consistency (Schuster et al., 2022). Mixture-of-Recursions and related systems learn token-wise allocation (Bae et al., 2025; Song et al., 2026; Fu et al., 2026). Huginn and Ouro supply pretrained recurrent-depth backbones (Geiping et al., 2025; Zhu et al., 2025); continuous depth batching addresses serving (Schwethelm et al., 2026). Early-exit readouts of ordinary transformers rely on the logit lens and its fitted variants (nostalgebraist, 2020; Belrose et al., 2023). We forecast residual change on frozen backbones of both kinds.

Post-hoc stopping in looped models. The closest comparisons are trajectory readouts that rival learned Ouro gates (Popescu et al., 2026) and training-free successive-KL stopping (Logan, 2026). They establish that inexpensive confidence and stability signals are strong baselines. Our contribution is a terminal-error target and a signed two-update representation that tests what remains predictable after one-step stability is controlled. We therefore compare both their confidence and convergence criteria on matched trajectories, alongside nonlinear controls with the same two-head rule.

Selective prediction and convergence. Selective prediction ranks examples so that retained coverage trades against risk (El-Yaniv & Wiener, 2010). We apply this framework to intermediate token readings, where the loss is terminal KL rather than task error. Fixed-point diagnostics ask whether an update has become small (Movahedi et al., 2026); deep equilibrium models use Broyden's method (Bai et al., 2019), while Anderson acceleration is a related fixed-point method (Walker & Ni, 2011). T2G uses a scalar secant to represent observed motion and learns its relationship to a finite-horizon output, using later outputs as forecasting targets.

3 Trajectory-to-Go

3.1 Problem

For one token, let $\ell_t \in \mathbb{R}^V$ be the logits of the reading at compute step t: the output after recurrent iteration t of a looped model, or a lens reading of the residual stream after layer t of an ordinary transformer. We call this distribution a *reading*, and distinguish it from the scalar *forecast* of its error. Let $p_t = \mathrm{softmax}(\ell_t)$ and let $T > t$ be a fixed terminal horizon. We define the prospective error

$$e_t = D_{\mathrm{KL}}(p_T \,\|\, p_t). \tag{1}$$

By contrast, the standard convergence signal $s_t = D_{\mathrm{KL}}(p_{t-1} \,\|\, p_t)$ observes only the last transition. A risk score r_t is useful when sorting token–checkpoint readings by r_t retains low e_t first. A stopping policy examines candidate checkpoints in order and accepts the first reading with $r_t \le \theta$; if none is accepted, it runs to T. If π denotes the selected checkpoint, the policy loss is e_π, including zero loss for terminal fallbacks. Thus policy risk averages over all inputs, whereas a risk–coverage curve measures error among retained readings.

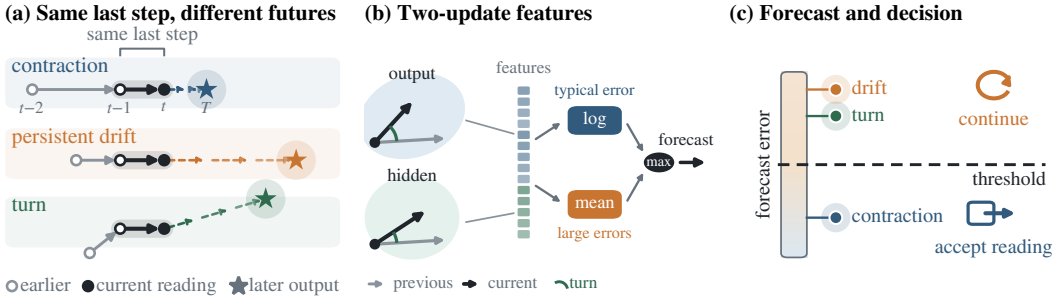


Figure 1: **Forecasting distance-to-go. (a)** The same last step, preceded by a larger (contraction), equal (drift), or rotated (turn) step; magnitudes separate the first two, direction the third. **(b)** Two-update features in output and hidden space feed a learned two-head forecaster. **(c)** A held-out threshold turns the forecast into an accept/continue decision.

3.2 Two-update output geometry

Write the two latest logit updates as $\Delta_t = \ell_t - \ell_{t-1}$ and $\Delta_{t-1} = \ell_{t-1} - \ell_{t-2}$. The categorical Fisher matrix at the current prediction is

$$F_t = \mathrm{diag}(p_t) - p_t p_t^\top, \qquad \langle a, b \rangle_{F_t} = a^\top F_t b. \tag{2}$$

This geometry is invariant to common logit shifts and locally approximates output KL. It can be evaluated without materializing F_t:

$$\langle a, b \rangle_{F_t} = \sum_j p_{t,j} a_j b_j - \Big(\sum_j p_{t,j} a_j \Big) \Big(\sum_j p_{t,j} b_j \Big). \tag{3}$$

We summarize trajectory magnitude by

$$u_t = \tfrac{1}{2} \langle \Delta_t, \Delta_t \rangle_{F_t}, \qquad u_{t-1} = \tfrac{1}{2} \langle \Delta_{t-1}, \Delta_{t-1} \rangle_{F_t}, \tag{4}$$

and let $m_t = \langle \Delta_t, \Delta_{t-1} \rangle_{F_t}$ be their cross-update moment. The Fisher secant coefficient, cosine alignment, and innovation energy reduce to one compact decomposition, where $\widetilde{\rho}_t = \mathrm{clip}(\rho_t, -4, 4)$ is used only in the innovation residual:

$$\rho_t = \frac{m_t}{2u_{t-1}}, \qquad c_t = \frac{m_t}{2\sqrt{u_t u_{t-1}}}, \qquad q_t = \tfrac{1}{2} \| \Delta_t - \widetilde{\rho}_t \Delta_{t-1} \|_{F_t}^2. \tag{5}$$

All five geometric features are coordinates of the same two-update Gram matrix. ρ_t measures signed persistence and c_t measures alignment. When clipping is inactive, the residual is Fisher-orthogonal to the preceding update and $q_t = u_t(1 - c_t^2)$; clipping stabilizes this residual near vanishing previous energy. These nonlinear coordinates expose magnitude and direction to a compact readout rather than treating them as independent observations.

Geometric motivation. Similar step energies can precede different errors under persistent or alternating motion. A scalar continuation example in Appendix B.1 illustrates this distinction; it motivates features rather than imposing a dynamics model. T2G learns their relationship to later error. Literal extrapolation underperforms update energy on all three looped models (Appendix C.5); Section 5.3 tests direction's incremental predictive value.

The output-space component has nine scalars. Update history contributes $\log s_t$, $\log s_{t-1}$, $\log u_t$, and $\log u_{t-1}$; direction contributes ρ_t, c_t, and $\log q_t$; and the current prediction contributes entropy $H(p_t)$ and top-two probability margin $M(p_t)$. We denote this vector by x_t. The five logged features use a 10^{-9} offset and base-10 logarithms; contraction, alignment, entropy, and margin remain unlogged.

3.3 Hidden geometry and the Risk-T2G forecast

Output geometry is aligned with distributional change, while hidden states also record motion before it is weighted by the current prediction. Let h_t be the observed state: the pre-head state in the looped-model study, or the residual stream in the layer-wise study. Set $\delta_t = h_t - h_{t-1}$. The width-normalized

Euclidean product $\langle a, b\rangle_h = d^{-1}a^\top b$ gives a corresponding decomposition of hidden motion:

$$v_t = \tfrac{1}{2}\langle \delta_t, \delta_t\rangle_h, \quad v_{t-1} = \tfrac{1}{2}\langle \delta_{t-1}, \delta_{t-1}\rangle_h, \quad \gamma_t = \frac{\langle \delta_t, \delta_{t-1}\rangle_h}{\langle \delta_{t-1}, \delta_{t-1}\rangle_h},$$
$$c_t^h = \frac{\langle \delta_t, \delta_{t-1}\rangle_h}{\sqrt{\langle \delta_t, \delta_t\rangle_h \langle \delta_{t-1}, \delta_{t-1}\rangle_h}}, \quad \eta_t = \tfrac{1}{2d}\|\delta_t - \widetilde{\gamma}_t\delta_{t-1}\|_2^2, \quad \widetilde{\gamma}_t = \mathrm{clip}(\gamma_t, -4, 4). \tag{6}$$

The two views differ in what motion they emphasize. When readings share a linear head W applied to h_t, $\Delta_t = W\delta_t$, so output energy is $u_t = \frac{1}{2}\delta_t^\top W^\top F_t W\delta_t$, whereas hidden energy weights directions uniformly by I/d. Motion weakly weighted by the current output distribution can therefore remain visible to the hidden features. Layer-specific lenses need not share this linear relation; we compute their output updates from the lens logits and their hidden updates from the residual stream separately. We concatenate $h_t^{\mathrm{traj}} = [\log v_t, \log v_{t-1}, \gamma_t, c_t^h, \log \eta_t]$ with x_t to obtain the 14-input vector $z_t = [x_t, h_t^{\mathrm{traj}}]$.

Forecast objectives. Log-space regression emphasizes multiplicative accuracy and compresses large errors; a mean-error objective gives those errors greater influence. At each checkpoint, a ridge head fits $\log_{10}(e_t + 10^{-8})$ with penalty 10; a Gamma-deviance head with a log link targets mean error (McCullagh & Nelder, 1989). Their combination defines RISK-T2G, the main fourteen-input forecaster of terminal KL error:

$$r_t^{\log} = 10^{\beta_t^\top \phi_2(z_t) + b_t}, \qquad r_t^{\mathrm{mean}} = \bar{y}_t \exp\Big(\theta_t^\top \widetilde{\phi}_2(z_t) + a_t\Big),$$
$$r_t = \max\{r_t^{\log}, \kappa r_t^{\mathrm{mean}}\}, \qquad \kappa = 0.3. \tag{7}$$

The shared map has 119 linear and quadratic terms after input standardization; the mean head additionally standardizes expanded columns and normalizes its target by the fit mean $\bar{y}_t$, using penalty $10/n_t$ for n_t fit readings. The floor changes scores only where the mean forecast exceeds the log forecast by more than $1/\kappa$, preserving log ordering among unaffected readings. The coefficient is selected jointly across the three looped models by prompt-level development cross-validation and fixed before fresh confirmation. At inference, the current and two preceding states and readings supply every input; no terminal output is needed.

The objectives have a direct population interpretation. For $y = e_t + 10^{-8}$, unrestricted, unregularized readouts target the conditional geometric mean $10^{\mathbb{E}[\log_{10} y|z]}$ and arithmetic mean $\mathbb{E}[y \mid z]$, respectively. A group with many small errors and a few large ones can therefore have a low log forecast but a substantially larger mean forecast. The floor is an empirical compromise between these targets, not the minimizer of either evaluation metric. Mean-only is the natural comparison for expected retained KL; the log head describes relative error resolution. Both are evaluated alongside their combination.

Component ablations. Output-only and hidden-only readouts apply the same two-head rule to each geometry alone, with 54 and 20 terms per head. Log-only and mean-only readouts remove one head while retaining all fourteen inputs (Table 5). Corpus transfer freezes the fitted heads and standardizers. New models, layer-wise readings, and reference horizons use the same features and objectives with new fits. Appendix A.1 gives the objectives and calibration rule; base-model weights remain frozen.

Cost. Given three consecutive readouts, RISK-T2G stores two previous logit vectors and two hidden-state updates, computes Fisher and $O(d)$ hidden reductions, and shares one 119-term expansion between its two heads. It adds no vocabulary projection when these readouts are already available; collecting a missing previous readout requires its projection.

3.4 EXTENSIONS

State augmentation: T2G + state. To combine recent motion with the current representation, this extension appends PCA-compressed state coordinates to the quadratic trajectory features and refits the same two-head rule. The added inputs are fitted per model and checkpoint (Appendix C.3).

Answer-change forecasting: TASK-T2G. For multiple-choice stopping, this extension replaces terminal KL with whether the current and terminal argmax differ. A linear logistic head uses the same trajectory construction restricted to four answer tokens, five hidden-motion features, and Ouro's native gate survival: 15 inputs in total. Correctness labels are used only for evaluation. It projects directly onto the four answer tokens at passes 1–3; measured latency includes collection, gate evaluation, scoring, and execution. Appendix B.2 defines the endpoint.

4 EXPERIMENTAL DESIGN

Models and checkpoints. We evaluate the public Huginn-0125 (3.5B) checkpoint at $t \in \{4, 8, 16, 24\}$ with $T = 32$, and Ouro-1.4B and Ouro-2.6B at $t \in \{3, 4, 6\}$ with $T = 8$. The two Ouro sizes compare checkpoints within one family, not a controlled scaling law. Their KL reference extends beyond the trained four-pass depth; MMLU evaluates the native four-pass endpoint. Looped models run in BF16, while geometric features and KL values are accumulated in FP32. For depth trajectories we use Qwen1.5-0.5B, Qwen1.5-1.8B (Qwen Team, 2024), and Pythia-410m (Biderman et al., 2023), each with 24 layers, at layers $t \in \{12, 15, 18, 21\}$ with $T = 24$. Qwen2.5-7B (Yang et al., 2024) uses layers $\{14, 17, 21, 24\}$ with $T = 28$, the same fractional depths. All ordinary transformers run in FP32. The reading is either the logit lens, which applies the final norm and unembedding to the residual stream, or a moment lens, a per-layer ridge map to the final post-norm state. This is a linear state-mapping readout related to Yom Din et al. (2024), fitted on the fit prompts and then frozen. Output geometry uses the lens readings and hidden geometry the residual stream at $t, t-1, t-2$.

Data separation. Examples are 96-token sequences, with positions 16–95 eligible. The primary WikiText-2 (Merity et al., 2017) study uses 128 fit prompts, 128 calibration prompts, and 256 fresh confirmation prompts after readout development: 81,920 Huginn readings and 61,440 per Ouro scale. All primary readouts are evaluated on these same prompts; gate and top-1 studies use separate 256-prompt confirmations. Fitting, calibration, and confirmation data are disjoint from each other; confirmation text also excludes earlier development data. Original readouts and policies are fixed before confirmation; additional controls fit the original data without using confirmation labels.

For distribution transfer, all three models use 128 C4 validation sequences and 128 LAMBADA test sequences (Raffel et al., 2020; Paperno et al., 2016). The full readout and matched confidence control retain their WikiText-fitted coefficients, standardizers, checkpoint grid, and policy thresholds on both transfer corpora. The reference-horizon and depth-trajectory studies use further unused WikiText-2 text: 128 fit and 256 confirmation prompts each, with the comparisons specified before confirmation (Appendix D.2). The 7B replication reuses the depth-study passages with a new model-specific fit; it is not a new-text confirmation.

The task-facing study uses five-shot MMLU (Hendrycks et al., 2021) at Ouro's trained four-pass horizon. We fit on 1,024 validation questions, calibrate on the remaining 507, and confirm on 1,024 test questions disjoint from method development. Thresholds maximize pass-3 exits under a 2% calibration target for unconditional pass-3/pass-4 answer change. Timing uses a fixed 128-question subset of that confirmation. The readout construction and primary comparisons are frozen before confirmation at 1.4B and then replicated unchanged at 2.6B with model-specific fitting and calibration.

Baselines. Table 1 separates prior stopping criteria from constructed controls. We evaluate published entropy, maximum-probability, logit-margin, predictive-KL, logit-displacement, and hidden-displacement/cosine readouts (Popescu et al., 2026), plus Ouro's native gate. Predictive KL uses $D_{\mathrm{KL}}(p_t \| p_{t-1})$ (Logan, 2026); Huginn's released criterion and our original s_t use $D_{\mathrm{KL}}(p_{t-1} \| p_t)$. These scalar rules are applied per token on our shared checkpoint grid and terminal-fidelity endpoint. Appendix C.4 specifies the adaptations and further Huginn rules.

Constructed confidence controls use entropy, probability margin, and s_t: a 100-tree GBDT (Friedman, 2001), a mean-KL head, and a two-head control matching Risk's objectives, coefficient, and 119 terms per head. We also fit a two-head probe on standardized, PCA-compressed current states, choosing its dimension and penalty by four-fold prompt cross-validation. The T2G + state extension uses these same coordinates (Section 3.4). Output/hidden and log/mean ablations isolate geometry and objective, respectively. For MMLU, we compare fixed depth, answer margin, native gate, and linear, quadratic, and boosted confidence; a matched linear control also adds the gate. Appendices A.1 and F give additional capacity controls and component removals.

Metrics and inference. The primary KL-ranking metric is excess area under the risk–coverage curve (EAURC; lower is better), which subtracts the oracle ordering from the area under retained terminal KL. Spearman correlation is a secondary description of global ordering. For the binary top-1 endpoint, we report area under the retained-disagreement curve (AURC). The locally stable subset uses the calibration 25th percentile of successive KL; the descriptive matched-band figure uses its

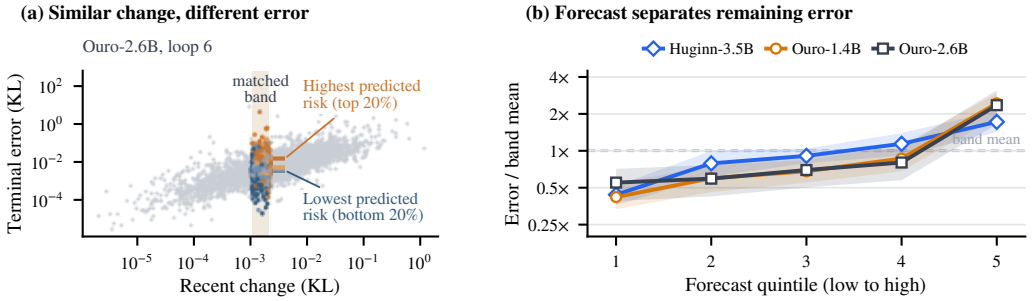


Figure 2: **Stable but unfinished. (a)** Ouro-2.6B readings at $t = 6$; the strip is the matched successive-KL band ($\pm.15$ dex), colored by output-only log-head forecast quintile with quintile means marked. **(b)** Within the band, terminal error rises across forecast quintiles for Huginn at step 24 and both Ouro models at loop 6 (2,403/3,222/3,261 readings); shading: 95% prompt-bootstrap intervals.

confirmation cohort's quartile. Primary uncertainty uses 2,000 paired bootstrap resamples of whole prompts, holding readouts fixed; shared texts do not count as extra cross-model evidence.

For stopping, each score threshold is chosen on the policy-calibration split to maximize mean loop iterations avoided while the estimated risk plus 1.645 standard errors remains below its target. Thresholds and readouts are then frozen for confirmation. Deployment latency uses batch size one, five warmup prompts per worker, three repeats, deterministic method-order rotation, CUDA-synchronized boundaries, and a per-prompt median. Allocation comparisons use common calibration targets and report realized held-out risk; avoided iterations exclude scoring cost, whereas runtime measurements include it. Appendix A gives the full data-separation protocol; hyperparameters and target grids appear in Appendix A.1.

5 Results

We test three questions: what motion predicts beyond confidence, which components supply that information, and when the resulting scores improve decisions. Transfer and layer-wise studies probe the reach of the forecasts.

5.1 What remains after apparent stability

A targeted stress test asks whether history helps after the latest update already looks small. This component diagnostic uses only output features and a log-error head. Within the lowest successive-KL quartile, adding output geometry to confidence and successive KL reduces EAURC from .737 to .389 on Huginn and from 26.651 to 22.962 on Ouro-1.4B (all $\times 10^{-3}$; Table 7). Within a narrow successive-KL band, terminal KL still differs four- to six-fold between the lowest and highest forecast quintiles at the latest checkpoints (Figure 2): the one-step signal leaves residual variation that the forecast can rank. These output-only controls motivate the full output/hidden comparison; they do not isolate history from current geometry.

5.2 Forecasting beyond matched controls

Table 1 compares published stopping criteria and matched readouts. Predictive KL gives the lowest prior-criterion EAURC on Huginn, and hidden cosine on both Ouro scales. Risk-T2G lowers their EAURC by 18.3%, 39.6%, and 17.9%, respectively, with positive paired intervals, and improves correlation over every prior row. It also lowers EAURC over boosted confidence by 8.9%, 43.3%, and 30.6%. Against confidence with the same two-head rule and capacity, reductions are 8.6%, 22.3%, and 7.7%; all six paired intervals exclude zero. It also improves global correlation over both controls at every scale and both point metrics over GBDT at all ten checkpoints. The mean-only ablation also has lower point EAURC than its capacity-matched mean-confidence control at all three scales.

Table 1: Prior stopping criteria and matched controls on identical WikiText-2 readings, all scored on our tokenwise fidelity endpoint. E is terminal-KL EAURC $\times 10^3$; ρ is global rank correlation.

	Huginn-0125		Ouro-1.4B		Ouro-2.6B	
Score	$E\downarrow$	$\rho\uparrow$	$E\downarrow$	$\rho\uparrow$	$E\downarrow$	$\rho\uparrow$
Published stopping criteria						
Entropy (Popescu et al., 2026)	67.691	.4142	87.738	.4151	47.243	.4473
Max probability (Popescu et al., 2026)	77.367	.3683	90.593	.3689	48.879	.4052
Logit margin (Popescu et al., 2026)	101.806	.2696	101.257	.2750	57.804	.3142
Predictive KL (Logan, 2026)	10.793	.9244	48.874	.7071	19.703	.8182
Logit L2 (Popescu et al., 2026)	16.660	.8454	37.859	.5437	16.894	.6533
Hidden L2 (Popescu et al., 2026)	15.814	.8496	38.490	.5556	19.010	.6564
Hidden cosine (Popescu et al., 2026)	15.823	.8500	31.835	.5333	15.122	.6329
Native gate (Zhu et al., 2025)	–	–	36.552	.6598	21.579	.7043
Constructed readout controls						
GBDT confidence	9.681	.9405	33.899	.7883	17.893	.8300
Mean confidence (119 terms)	9.562	.9371	20.973	.6808	11.665	.7525
Risk confidence (119/head)	9.652	.9399	24.773	.7628	13.451	.8117
Learned state probe	12.873	.8936	22.350	.7370	16.028	.7531
T2G: main method and objective ablation						
RISK-T2G (ours)	**8.822**	**.9500**	19.237	**.8083**	12.416	**.8357**
T2G, mean-only (ablation)	8.858	.9475	**15.895**	.7549	**11.234**	.7928

Table 2: Current geometry, history, and direction. EAURC $\times 10^3$ (↓), with the same two-head fitting rule. Positive incremental reductions favor added inputs; brackets give paired 95% prompt intervals. Band: small last-step KL at the last checkpoint.

Inputs / incremental reduction	Huginn-0125	Ouro-1.4B	Ouro-2.6B
One-update geometry	9.345	21.193	12.016
+ magnitude history	9.139	20.352	12.362
+ direction: RISK-T2G	8.822	19.237	12.416
History reduction	$0.206\ [0.154, 0.258]$	$0.841\ [0.443, 1.276]$	$-0.346\ [-0.752, -0.027]$
Direction reduction	$0.317\ [0.238, 0.407]$	$1.115\ [0.522, 1.827]$	$-0.054\ [-0.725, 0.410]$
Direction reduction in band	$0.119\ [0.055, 0.217]$	$0.633\ [0.478, 0.797]$	$0.017\ [-0.027, 0.060]$

The learned state probe has 16–46% higher EAURC than RISK-T2G, with paired intervals excluding zero in all three models. The state-augmented extension improves further: on 512 fresh prompts per model, T2G + state lowers EAURC by 8.7–13.0% relative to RISK-T2G, with positive paired intervals in all three models (Appendix C.3).

5.3 WHAT GEOMETRY AND THE FITTING OBJECTIVE CONTRIBUTE

With the two-head rule held fixed, joint geometry improves both point metrics over either view alone (Table 5, Appendix C). Its EAURC gains over output-only are 3.5%, 12.7%, and 8.5%, all with positive paired intervals. Energy features supply the largest component increments (Appendix C.1).

Table 2 starts from current successive KL, output and hidden energies, entropy, and margin, then adds magnitude history and direction. History improves Huginn and Ouro-1.4B, but worsens Ouro-2.6B EAURC by $.346 \times 10^{-3}$, with a paired interval excluding zero. Direction lowers overall EAURC by 3.5% on Huginn and 5.5% on Ouro-1.4B; the relative reductions within the matched band are larger, at 32.0% and 18.4%. Ouro-2.6B shows no clear benefit overall or in the band. Hidden direction supplies the clearest incremental signal. Across ordinary-transformer readings, direction lowers EAURC by 7.0–9.2%, with positive paired intervals in the six sub-2B model–lens cells. Appendix C.2 gives the retrospective design and space-specific ablations.

The objective comparison is a separate tradeoff. Mean-only gives the lowest Ouro EAURC, improving over Risk by 17.4% and 9.5%; log-only gives the strongest global ordering. Risk reduces Ouro

Table 3: Frozen two-head transfer. E is EAURC $\times 10^3$; KL is realized policy risk at calibration target .05. C: matched Risk confidence; R: RISK-T2G. Δ saved: gross iteration avoidance in points (95% intervals), at different realized risks.

Model	Corpus	E_{C}	E_{R}	KL_{C}	KL_{R}	Δ saved $\uparrow$
Huginn-0125	C4	12.111	10.961	.0514	.0549	$+1.16$ $[.86, 1.46]$
Huginn-0125	LAMBADA	7.010	6.550	.0407	.0432	$+1.04$ $[.83, 1.25]$
Ouro-1.4B	C4	10.471	8.010	.0316	.0340	$+2.78$ $[2.53, 3.03]$
Ouro-1.4B	LAMBADA	9.213	8.327	.0308	.0340	$+2.70$ $[2.44, 2.96]$
Ouro-2.6B	C4	6.496	5.955	.0342	.0342	$+1.10$ $[.89, 1.30]$
Ouro-2.6B	LAMBADA	5.086	4.585	.0312	.0333	$+1.58$ $[1.43, 1.74]$

EAURC over log-KL by 25.2% and 16.0% while lowering ρ by .0124 and .0113. The fixed two-head rule preserves more global ordering, but does not outperform mean-only on the primary Ouro metric. We therefore test policy value directly rather than infer it from correlation.

5.4 FROZEN FORECASTS TRANSFER ACROSS CORPORA

Does the learned geometry forecast residual change beyond its fitting corpus? Table 3 applies the frozen WikiText-fitted RISK-T2G readouts to C4 and LAMBADA. Against confidence with the same two-head rule and 119 terms per head, EAURC falls by 6.6–23.5% across all six cells, with five positive paired intervals; Ouro-1.4B on LAMBADA overlaps zero. Global correlation also improves over this control in every cell. Reductions against GBDT confidence are 6.9–43.5%, all with positive paired intervals. The predictive benefit thus extends beyond WikiText without changing features, coefficients, or readout weights.

Frozen thresholds yield 1.04–2.78 additional allocation points over matched confidence, but realized KL is also higher in five of six cells. This is a coverage–risk tradeoff, not an efficiency gain at matched risk. Risk avoids 43.9–65.1% of iterations. On Huginn/C4, Risk and confidence exceed the .05 target, reaching .0549 and .0514. Appendix D.1 gives the stricter budget and successive-KL comparison.
Changing the reference horizon. With heads fitted for each reference horizon, RISK-T2G improves over matched confidence in all nine model–horizon cells, including Huginn steps 48/64 and Ouro loops 12/16. The advantage covers both Huginn's nearly stable later outputs and Ouro's continued drift beyond its trained depth. Appendix D.2 compares horizon-specific fits with native-horizon transfer.

5.5 FORECASTING ALONG ORDINARY TRANSFORMER LAYERS

A transformer layer need not repeat the preceding layer's computation for its recent readings to contain useful history. Table 4 tests the same fourteen features and two-head fitting procedure on four ordinary transformers from 410M to 7B, with heads fitted per model, lens, and checkpoint. RISK-T2G lowers EAURC over matched confidence by 10.8–21.1%, with every paired interval excluding zero. At 7B, the reductions over matched confidence are 10.8% and 13.2% for the logit and moment lenses, respectively.

Current-state information complements this history, rather than uniformly dominating it. The state probe is stronger under every logit lens, but RISK-T2G is stronger under the moment lens in Qwen1.5-1.8B and Qwen2.5-7B. The T2G + state extension gives the lowest EAURC among these readouts in all eight settings. At 7B it improves on RISK-T2G by 30.0% and 10.9%. Appendix C.3 gives fitting details and intervals.

5.6 STOPPING WITH THE EXTENSIONS

KL policies and allocation. On the primary cohort, a common empirical KL cap separates objective from feature gains (Table 12). Mean-only and T2G + state improve over mean confidence on Huginn at .02 and Ouro-1.4B at both caps, but not on Ouro-2.6B. Thus better ranking than matched two-head confidence does not imply a better policy than a mean-confidence readout. Against published criteria, T2G + state raises the allocation frontier on 512 fresh prompts per model by .34–3.91 percentage

Table 4: Layer-wise forecasting. EAURC $\times 10^3$ (↓), 256 prompts. C: matched confidence; R: RISK-T2G; S: state probe. C−R: paired 95% intervals. State comparisons are descriptive.

Reading	Model	C	R	S	T2G + state	C−R [95%]
Logit lens	Qwen1.5-0.5B	974.2	798.8	459.2	445.4	175.4 [159.3, 191.4]
Logit lens	Pythia-410m	411.2	344.3	267.0	249.9	66.8 [61.1, 73.3]
Logit lens	Qwen1.5-1.8B	1069.4	922.4	614.1	580.0	147.0 [131.0, 163.1]
Logit lens	Qwen2.5-7B	1627.7	1452.0	1097.0	1016.4	175.7 [155.0, 196.1]
Moment lens	Qwen1.5-0.5B	249.4	214.4	203.1	167.8	35.0 [31.3, 38.6]
Moment lens	Pythia-410m	229.5	181.1	165.8	141.4	48.5 [45.2, 51.9]
Moment lens	Qwen1.5-1.8B	300.3	247.3	270.3	202.2	53.0 [48.8, 57.6]
Moment lens	Qwen2.5-7B	690.3	599.1	740.0	533.6	91.2 [77.7, 105.7]

points over the calibration-selected prior at KL caps .02 and .05 nats. Intervals are positive in five of six settings; Huginn at .02 is inconclusive (Table 13). Frontier thresholds are evaluation-selected. Separately, all six policies with frozen calibration thresholds have observed mean KL within budget, and five have positive allocation intervals. Iteration avoidance excludes scoring costs.

Answer preservation and execution cost. TASK-T2G accepts the pass-3 MMLU answer or runs pass 4, with its classifier and threshold frozen before evaluation. It exits on 74.61% and 80.76% of questions, changing 1.86% and 1.66% of pass-4 answers. Including all scoring, batch-1 B200 latency falls by 13.02% [10.82, 14.96] and 18.85% [17.13, 20.39] relative to fixed pass 4. Confidence plus gate reaches similar coverage and is faster at 1.4B and close at 2.6B: these are savings over full execution, not a demonstrated latency advantage over the learned control. Table 14 reports policies, accuracy, and paired intervals.

6 DISCUSSION

What trajectory history predicts. Current magnitude, magnitude history, and direction answer different questions about progress. History and mainly hidden-state direction improve Huginn and Ouro-1.4B, especially among stable readings. On Ouro-2.6B, the one-update control is sufficient. The cause remains open: the ablations measure added predictive information, not a mechanism for the difference. Across models, output and hidden motion provide complementary views.

From a forecast to a policy. Forecast targets should match the decision. Mean error is relevant to retained KL; answer-change risk is relevant to preserving a task decision. KL can change while the leading answer stays fixed, explaining why confidence and gates remain competitive for MMLU exits. Allocation frontiers describe attainable tradeoffs; calibration and scoring cost determine what a deployed policy actually achieves.

Limitations. The ordinary-transformer study ranks intermediate readings, not deployable exits, and does not evaluate a KL-trained tuned lens. Forecasts are fitted per model, reading, checkpoint, and horizon. Prompt-bootstrap intervals condition on those fits; they do not include refitting variability. The 96-token studies and batch-1 MMLU timing do not establish long-context or autoregressive serving gains, where early exits leave missing deeper-layer KV-cache states. Corpus shift also requires checking risk calibration.

7 CONCLUSION

Trajectory-to-Go turns recent output and hidden motion into a compact forecast of disagreement with a later output. It improves error ranking over matched confidence controls, with model-dependent contributions from history and direction. Frozen forecasts transfer across corpora, and state information complements trajectory features along ordinary transformer layers up to 7B. Stopping studies show how these forecasts can be used while distinguishing ranking gains from additional savings over learned policies. The central result is that recent stability leaves predictable remaining change, and that the useful evidence extends beyond confidence alone.

AI USE STATEMENT

In developing and evaluating trajectory forecasts, the authors used generative AI models for research ideation and refinement, mathematical derivations, experimental design and implementation, and result analysis. These tools also assisted with literature searches, manuscript drafting and revision, and figure and reference preparation. The authors set the research questions, hypotheses, and methodology, directed the manuscript's development, and reviewed the claims, derivations, experiments, code, and reported results. The authors approved the final text and take responsibility for all content, including the material developed with AI assistance.

REPRODUCIBILITY STATEMENT

Forecast fitting, stopping-policy calibration, and evaluation use distinct data roles. Section 4 defines the comparisons and metrics; Appendix A specifies source separation, fitted readouts, frozen thresholds, and runtime measurement. Appendix B derives the features and endpoint-specific forecasts. Forecasting and generalization results appear in Appendices C and D, with stopping policies and execution results in Appendices E and F. The supplementary code includes prompt hashes, saved fits and thresholds, evaluation scripts, results, and table checks. Model weights and benchmark data come from their original providers.

ETHICS STATEMENT

Stopping early saves computation but can discard changes that matter to a task. T2G evaluates this tradeoff through output fidelity and answer preservation, with stopping thresholds calibrated on held-out prompts. These targets concern agreement with later computation, not whether an answer is true or safe. Policies should be validated for the intended task and data distribution, particularly when either differs from calibration. Model and benchmark reuse remains subject to the original licenses and access conditions.

REFERENCES

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, and Se-Young Yun. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. In *Proceedings of the Eighth ICML Workshop on Automated Machine Learning*, 2021.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Tianyu Fu, Yichen You, Zekai Chen, Guohao Dai, Huazhong Yang, and Yu Wang. Think-at-Hard: Selective latent iterations to improve reasoning language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62678.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022. doi: 10.18653/v1/2022.emnlp-main.3.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3301–3310, 2019.

Joe Logan. Per-token fixed-point convergence in depth-recurrent transformers. *arXiv preprint arXiv:2607.14427*, 2026.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989. ISBN 9780412317606.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Sajad Movahedi et al. Fixed-point reasoners: Stable and adaptive deep looped transformers. *arXiv preprint arXiv:2606.18206*, 2026.

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

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534, 2016.

Andrei Cristian Popescu, Haitz Sáez de Ocáriz Borde, and Pietro Liò. Adaptive depth in looped transformers: Diagnosing learned halting gates and trajectory readouts. *arXiv preprint arXiv:2607.20519*, 2026.

Qwen Team. Introducing Qwen1.5. Qwen Blog, 2024. URL https://qwenlm.github.io/blog/qwen1.5/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kristian Schwethelm, Daniel Rueckert, and Georgios Kaissis. Depth-adaptive inference of looped language models via continuous depth batching. *arXiv preprint arXiv:2608.09444*, 2026.

Shixiang Song, He Li, Zitong Wang, Boyi Zeng, Feichen Song, Yixuan Wang, Zhiqin John Xu, Ziwei He, and Zhouhan Lin. AdaPonderLM: Gated pondering language models with token-wise adaptive depth. *arXiv preprint arXiv:2603.01914*, 2026.

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011. doi: 10.1137/10078356X.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9615–9625. ELRA and ICCL, 2024.

Rui-Jie Zhu et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

A DATA, FITTING, CALIBRATION, AND RUNTIME MEASUREMENT

Language-model cohorts. Examples contain 96 tokens, with positions 16–95 eligible for evaluation. The primary study uses 128 fitting prompts, 128 policy-calibration prompts, and 256 confirmation prompts per model, yielding 81,920 Huginn readings and 61,440 per Ouro scale. Huginn fitting text comes from WikiText-2 test; Ouro fitting text, and all primary calibration and confirmation text, come from WikiText-2 train. These corpus split names identify the text sources; forecast-fitting and evaluation roles are defined by separate prompt sets. These roles are source-disjoint from each other, and confirmation excludes earlier development and evaluation data. The three confirmation cohorts contain 277 distinct source texts across tokenizers; uncertainty is computed within each model rather than treating shared texts as independent replicates. State augmentation and KL allocation are also evaluated on a separate 512-prompt cohort per model from an unused WikiText-2 train offset. Source-text and token-hash checks exclude overlap with all previous study roles. Readout settings, calibration thresholds, and comparator selection are fixed before collection. The deterministic state-augmented fit reproduces its stored predictions on both earlier evaluation and calibration rows before the new cohort is scored. Appendix C.3 reports ranking; Appendix E reports allocation and calibrated policies.

Fitting and evaluation. Features, polynomial degree, checkpoints, horizons, original readouts, and policy rules are fixed before primary confirmation. Published criteria require no fitting. The two-head confidence and separate-space controls are retrospective comparisons: they retain the original coefficient and fitting data without tuning on confirmation labels. Earlier evaluations used for readout development are excluded from the reported confirmations. Endpoint-specific heads are selected before their respective evaluations.

Corpus transfer. Each corpus has three disjoint 128-sequence cohorts for output-only ranking, output-only policy evaluation, and joint-readout evaluation (Table 3). C4 uses consecutive eligible validation documents of at least 600 characters; LAMBADA packs consecutive test examples until reaching 800 characters. Sequences are tokenized and truncated to 96 tokens. The full-method cohort uses source chunks 256–383. All transferred readouts, standardizers, and thresholds use only WikiText fitting and calibration data and remain unchanged. Source-text and token-hash checks exclude overlap with prior evaluation roles.

MMLU cohorts. The task study uses disjoint validation sets of 1,024 and 507 questions for fitting and calibration, followed by 1,024 test questions; timing uses the first 128 confirmation questions. Three earlier 1,024-question test cohorts used for development are excluded. A second evaluation applies unchanged 2% thresholds to 1,021 questions after removing exact duplicates and overlaps from a separate 1,024-question cohort. The primary fitting, calibration, and confirmation sets have no exact prompt overlap.

Comparisons use matched readings within each cohort; absolute values across different cohorts need not agree. The matched-band diagnostic in Figure 2 uses the dedicated output-only confirmations, separate from the primary joint-readout cohort. We use one late checkpoint per model: $t = 24$ for Huginn and $t = 6$ for both Ouro scales. Readings are restricted to successive KL within ± 0.15 dex of that checkpoint's confirmation-set 25th percentile, then grouped into quintiles of the frozen output-only log-KL forecast. Terminal KL is normalized by the selected band's mean; intervals use 1,000 prompt-bootstrap resamples. The lowest/highest quintile means are $0.00043/0.00170$ nats for Huginn, $0.00730/0.04240$ for Ouro-1.4B, and $0.00354/0.01516$ for Ouro-2.6B. These are distributional discrepancies, not task effects.

A.1 READOUTS AND CALIBRATION

The quadratic readout standardizes its inputs using fit statistics at each checkpoint. The nine-feature output core and five hidden-geometry scalars are fixed across architectures. Both heads of RISK-T2G use the concatenated views; output-only and hidden-only controls use each alone. For both geometries, the secant coefficient is clipped to $[-4, 4]$ only when computing residual innovation. Core, hidden-only, and joint maps contain 54, 20, and 119 nonconstant terms. Let $y_i = e_i + 10^{-8}$ and $\phi_i = \phi_2(z_i)$. The joint log head minimizes $\sum_i (\log_{10} y_i - \beta^\top \phi_i - b)^2 + 10\|\beta\|_2^2$. For the mean head, each polynomial column is additionally standardized using fit statistics, giving $\widetilde{\phi}_i$. With $\bar{y} = n^{-1} \sum_i y_i$, $u_i = y_i/\bar{y}$, and $\widehat{u}_i = \exp(\theta^\top \widetilde{\phi}_i + a)$, the penalized Gamma objective has the

following form, up to additive constants:

$$\mathcal{L}_\Gamma = \frac{1}{n}\sum_i \left(\frac{u_i}{\widehat{u}_i} + \log \widehat{u}_i\right) + \frac{10}{2n}\|\theta\|_2^2, \qquad r_i^{\text{mean}} = \bar{y}\,\widehat{u}_i. \tag{8}$$

At the population level, the unpenalized pointwise minimizer is the conditional mean of u_i; fitting this objective does not assume an exact Gamma law for terminal errors. We use L-BFGS with at most 2,000 iterations and tolerance 10^{-6}. Log predictions are clipped to $[-12, 4]$ before base-10 exponentiation, and mean predictions to $[10^{-12}, 10^4]$. Equation 7 combines the heads with a mean-error floor. The coefficient was selected from $\{.01, .03, .1, .3, 1\}$ by four-fold prompt-level cross-validation on the original 128 fitting prompts of each of the three looped models, considering EAURC and Spearman jointly. The shared value is held fixed for every model and checkpoint; it is not reselected on the confirmation cohorts or the ordinary-transformer studies. Input and polynomial standardizers use fit prompts only; their affine transforms can be absorbed into the heads at evaluation. Mean-confidence controls use degree-two, -five, and -seven expansions of successive KL, entropy, and margin, giving 9, 55, and 119 terms under the same Gamma objective and regularization. Risk-confidence controls add a log-KL head with the same expansion and ridge penalty, then apply the same .3 floor. The 119-term-per-head control has 238 coefficients plus two intercepts, matching the full readout. Successive KL uses the core's base-10 log transform with a 10^{-9} offset. The boosted log-KL control uses 100 iterations, learning rate .05, at most 15 leaves, minimum leaf size 20, and ℓ_2 regularization 1.

TASK-T2G uses the two successive answer-space KL values, two Fisher update energies, contraction, alignment, innovation, entropy, and answer margin, all in their raw (unlogged) form. It adds the five hidden-trajectory scalars, with energies and innovation logged as in the joint representation, and native-gate continuation. Its standardized linear logistic readout uses $C = .1$, balanced classes, and L-BFGS. The self-supervised label is pass-3/pass-4 answer change.

Policy calibration searches 1,001 empirical score quantiles plus the reject-all option. It selects the threshold with greatest mean loop avoidance among those whose prompt-mean terminal KL satisfies $\bar{e} + 1.645\,\mathrm{SE}(\bar{e}) \le \epsilon$. Confirmation intervals resample whole prompts, preserving all positions and checkpoints within each cluster. The mean-readout confirmation uses KL targets $\{.02, .05\}$, with .05 primary; all fits and thresholds are saved before confirmation arrays are opened. The MMLU policy maximizes pass-3 acceptance subject to $\bar{y} + 1.645\sqrt{\bar{y}(1-\bar{y})/n} \le .02$, where y is the unconditional answer-change indicator after applying the exit decision. These normal-approximation criteria guide calibration; they do not give a finite-sample risk guarantee after threshold selection. Policy comparisons use a common calibration target rather than matching realized test risk.

A.2 COMPUTE

Experiments use NVIDIA B200 GPUs and NGC PyTorch 25.08. Deployment timing uses batch size one, five warmup prompts per worker, synchronized CUDA boundaries, rotating method order, three repeats, and per-prompt medians; it includes projections, feature construction, classifier inference, and control flow. The matched MMLU benchmark uses cached on-device coefficients for every learned readout on the fixed 128-question timing subset. Reported time saved is the mean paired relative reduction $n^{-1}\sum_i (1 - L_i^\pi / L_i^T)$, where each L_i is a per-prompt median; it is not the ratio of aggregate mean latencies. This includes the online forecasting cost but excludes the offline collection and fitting costs.

B GEOMETRIC DERIVATION AND ENDPOINT-SPECIFIC READOUTS

B.1 LOCAL GEOMETRIC DERIVATION

Let $B(\ell) = \log \sum_j \exp(\ell_j)$ and $b = \ell_T - \ell_t$. The exponential family identity gives

$$D_{\mathrm{KL}}(p_T \| p_t) = b^\top \nabla B(\ell_t + b) - B(\ell_t + b) + B(\ell_t) = \tfrac{1}{2} b^\top F_t b + O(\|b\|^3). \tag{9}$$

If future updates obey the local scalar model $\Delta_{t+k} = \rho^k \Delta_t$, then $b = (\sum_{k=1}^{T-t} \rho^k)\Delta_t$; substitution yields

$$e_t \approx u_t \left(\sum_{k=1}^{T-t} \rho^k \right)^2 . \tag{10}$$

This is a local approximation: it requires small total displacement for the Taylor expansion and persistent update direction for the scalar model. For $\rho \neq 1$, the sum is $\rho(1-\rho^{T-t})/(1-\rho)$; for $\rho = 1$, it is $T-t$.

As an illustration, take $u_t = u$, $u_{t-1} = 4u$, and zero innovation. Aligned and reversed updates have $\rho = 1/2$ and $-1/2$ despite identical step energies. With four steps remaining, Equation 10 gives $225u/256$ versus $25u/256$, a ninefold difference under the assumed continuation.

For nonzero previous energy, the least-squares coefficient minimizing $\|\Delta_t - a\Delta_{t-1}\|_{F_t}^2$ is $a = m_t/(2u_{t-1})$. Its residual is Fisher-orthogonal to Δ_{t-1}, and its half squared norm is $u_t - m_t^2/(4u_{t-1}) = u_t(1-c_t^2)$. The implementation floors small denominators and clips the coefficient for residual computation; the exact orthogonality identity applies when these stabilizers are inactive. Hidden geometry follows the same argument with F_t replaced by I/d.

B.2 Endpoint-specific readouts

For a discrete decision, the target becomes disagreement with the terminal answer. Let $\mathcal{A}$ be the candidate vocabulary or answer set, and let $\xi_t^{\mathcal{A}}$ be its trajectory representation. A logistic readout fits

$$y_t = \mathbf{1}\left[\arg\max_{j \in \mathcal{A}} \ell_{t,j} \neq \arg\max_{j \in \mathcal{A}} \ell_{T,j} \right], \qquad \widehat{\Pr}(y_t = 1) = \sigma\big(f_t(\xi_t^{\mathcal{A}})\big) . \tag{11}$$

For multiple-choice execution, TASK-T2G forms the output geometry on the softmax restricted to the four answer tokens, then adds hidden motion and an available native gate.

When the base model supplies a learned halt gate, its signal can be added rather than replaced. Ouro exposes a scalar exit-gate logit a_j after each recurrent pass; under its released halting rule, the probability mass remaining after checkpoint t is

$$g_t = \prod_{j=1}^{t} (1 - \sigma(a_j)) . \tag{12}$$

We use g_t directly as the *native-gate* baseline, or append it to a trajectory readout. TASK-T2G uses a 15-input linear logistic map for pass-3/pass-4 answer change. Representation and map are selected on separate development data for each endpoint; Appendix A.1 specifies their transforms. Correct answers are used only to evaluate task accuracy, never to fit or calibrate a policy.

For the single-exit task policy, let $I_t = \mathbf{1}[r_t \leq \theta]$. Its exit coverage is $C = \mathbb{E}[I_t]$ and answer-change risk is $R = \mathbb{E}[I_t y_t]$. This risk has a direct task interpretation:

$$|\mathrm{Accuracy}(\pi) - \mathrm{Accuracy}(T)| \leq R, \tag{13}$$

because only changed answers can alter correctness. The bound holds for the realized evaluation set as well as in expectation; a calibration target for R is evaluated on held-out questions.

C Forecasting comparisons and ablations

Table 5 reports the geometry and objective ablations discussed in Section 5.3; Figure 3 shows the retained-error profiles of the Table 1 cohort.

The paired intervals distinguish improvements from numerical ties. For example, mean confidence has lower point EAURC than Risk on Ouro-2.6B, but their paired interval includes zero. Changing the fitting objective and adding trajectory information therefore remain separate comparisons.

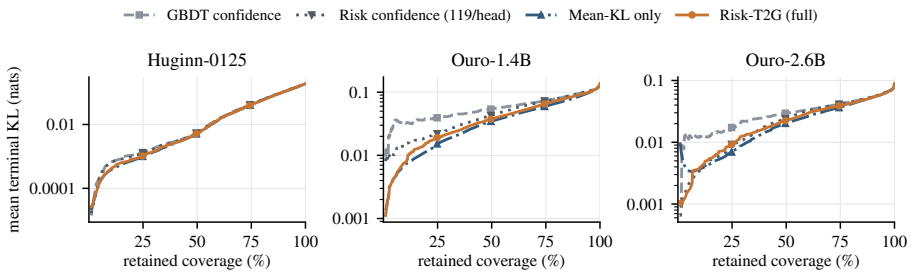


Figure 3: Selective terminal fidelity on the Table 1 cohort. Lower retained KL is better (log scale). Risk confidence uses the same two-head rule and parameter count as RISK-T2G; mean-only uses the same trajectory inputs.

Table 5: Geometry and objective ablations on the Table 1 cohort. Each block holds the other design choice fixed; bold marks block-wise best. E is EAURC $\times 10^3$; ρ measures global rank correlation.

	Huginn-0125		Ouro-1.4B		Ouro-2.6B	
Variant	$E\downarrow$	$\rho\uparrow$	$E\downarrow$	$\rho\uparrow$	$E\downarrow$	$\rho\uparrow$
Geometry: same two-head risk-floor readout						
Output only	9.143	.9467	22.026	.7961	13.576	.8283
Hidden only	15.414	.8600	23.913	.6704	13.716	.6887
Joint: RISK-T2G	**8.822**	**.9500**	**19.237**	**.8083**	**12.416**	**.8357**
Objective: same joint output/hidden geometry						
Joint log-KL	8.828	**.9501**	25.712	**.8207**	14.775	**.8470**
T2G, mean-only	8.858	.9475	**15.895**	.7549	**11.234**	.7928
RISK-T2G (both heads)	**8.822**	.9500	19.237	.8083	12.416	.8357

Table 6: Paired terminal-KL EAURC differences on held-out confirmations. Entries are point estimates [95% CI] scaled by 10^3; positive differences favor the second score.

First − second	Huginn-0125	Ouro-1.4B	Ouro-2.6B
Terminal-KL forecasting			
Predictive KL − Risk	1.972 [1.734, 2.229]	29.637 [23.440, 36.955]	7.287 [5.611, 9.167]
Hidden cosine − Risk	7.001 [6.437, 7.604]	12.598 [11.007, 14.120]	2.707 [1.310, 3.931]
Native gate − Risk	–	17.315 [14.400, 20.284]	9.164 [7.727, 10.606]
GBDT − Risk	.859 [.716, 1.020]	14.662 [10.500, 19.525]	5.477 [3.917, 7.309]
Risk conf. (119) − Risk	.830 [.700, .989]	5.536 [3.958, 7.729]	1.036 [.213, 1.856]
Mean conf. (119) − Risk	.741 [.608, .894]	1.736 [.128, 3.492]	−.751 [−1.927, .268]
Output only − Risk	.321 [.234, .424]	2.788 [1.705, 4.107]	1.160 [.432, 2.024]
Hidden only − Risk	6.592 [6.086, 7.132]	4.676 [3.580, 5.725]	1.300 [.050, 2.406]
Joint (log) − Risk	.006 [.001, .011]	6.475 [4.298, 8.973]	2.359 [1.462, 3.434]
Joint (mean) − Risk	.036 [−.029, .096]	−3.343 [−4.509, −2.270]	−1.182 [−2.083, −.049]

C.1 OUTPUT AND HIDDEN FEATURE ABLATIONS

Table 7 tests the output representation separately from the risk-floor readout. Fisher energy supplies the largest overall increment; direction improves all four cells, with positive paired direction-over-energy intervals on both models. Table 5 then tests the contribution of each complete geometry under the same two-head objective.

C.2 DIRECTION BEYOND MAGNITUDE HISTORY

We refit the published two-head procedure on the original fitting role with five input sets: all 14 inputs; eight inputs without direction; 11 without output direction; 11 without hidden direction; and five one-update inputs (current successive KL, output and hidden energies, entropy, and margin). The

Table 7: Output-geometry ablation with quadratic log-KL heads on dedicated confirmations. EAURC $\times 10^3$; "Stable" denotes the lowest successive-KL quartile. Each row adds the indicated features.

	Huginn-0125		Ouro-1.4B	
Features	All	Stable	All	Stable
Confidence and successive KL	10.358	.737	28.589	26.651
+ previous KL	10.094	.514	28.451	27.083
+ Fisher energies	9.704	.418	25.848	23.827
+ direction (ρ, c, q): core T2G	**9.540**	**.389**	**24.995**	**22.962**

Table 8: Input ablations: EAURC differences $\times 10^3$ with paired 95% prompt intervals. Direction columns remove only the named group; history contrasts one-update with nondirectional features.

Model / reading	Output direction	Hidden direction	Magnitude history
Huginn-0125	$0.027\ [-0.004, 0.059]$	$0.178\ [0.122, 0.241]$	$0.206\ [0.154, 0.258]$
Ouro-1.4B	$-0.109\ [-0.371, 0.163]$	$0.633\ [0.234, 1.097]$	$0.841\ [0.443, 1.276]$
Ouro-2.6B	$-0.086\ [-0.300, 0.097]$	$0.132\ [-0.340, 0.568]$	$-0.346\ [-0.752, -0.027]$
Qwen1.5-0.5B / logit	$-3.935\ [-10.815, 1.658]$	$62.677\ [55.303, 70.070]$	$65.665\ [54.842, 76.922]$
Qwen1.5-0.5B / moment	$5.666\ [4.375, 6.816]$	$12.721\ [11.252, 14.258]$	$16.885\ [15.083, 18.642]$
Pythia-410m / logit	$1.982\ [0.308, 3.760]$	$23.111\ [19.719, 26.558]$	$21.861\ [18.551, 25.084]$
Pythia-410m / moment	$3.543\ [2.758, 4.319]$	$9.233\ [8.241, 10.229]$	$22.094\ [20.481, 23.763]$
Qwen1.5-1.8B / logit	$0.960\ [-7.755, 10.774]$	$62.843\ [53.517, 72.367]$	$41.766\ [32.171, 51.162]$
Qwen1.5-1.8B / moment	$6.804\ [5.797, 7.902]$	$15.743\ [13.793, 17.700]$	$12.429\ [10.969, 13.857]$
All direction removed: depth-study EAURC difference			
Model	Logit lens	Moment lens	
Qwen1.5-0.5B	$70.113\ [61.226, 79.447]$	$20.208\ [18.220, 22.212]$	
Pythia-410m	$26.225\ [22.513, 30.106]$	$14.364\ [13.029, 15.679]$	
Qwen1.5-1.8B	$69.027\ [57.653, 80.180]$	$25.037\ [22.604, 27.572]$	

six directional inputs are the secant, alignment, and innovation coordinates in each space. All variants retain degree-two expansion, the original penalties, and $\kappa = .3$. Their term counts per head are 119, 44, 77, 77, and 20: this tests input contribution, not matched capacity.

This is a retrospective analysis of previously examined confirmation cohorts. Intervals use 2,000 paired whole-prompt bootstrap draws. For the primary cohort, at step 24 of Huginn and loop 6 of Ouro, the matched band lies within .15 dex of the confirmation-set 25th percentile of log successive KL. This follows Figure 2's band definition, but uses the primary cohort and the joint two-head forecast. Band membership is held fixed during bootstrap resampling; intervals are conditional on that selected band. No confirmation labels are used to fit the readouts.

The one-update comparison shows a benefit from magnitude history on Huginn, Ouro-1.4B, and all six sub-2B depth settings, but the ordering reverses on Ouro-2.6B. Huginn is larger at 3.5B, so these checkpoints do not establish a monotonic model-size trend or identify the cause of the Ouro difference. Removing hidden direction increases EAURC on the first two looped models and all tested sub-2B depth settings. Output direction adds a smaller increment under moment lenses. Each comparison measures a feature group's incremental predictive value after refitting with the remaining, correlated inputs.

C.3 Current-state probes and trajectory information

Matched target and fitting roles. The state probe predicts terminal KL from the current pre-head residual state, using the same log and mean heads and coefficient .3 as RISK-T2G. It is a learned-state baseline motivated by early-exit classifiers (Schuster et al., 2022), not a reproduction of CALM's sequence-level policy. We standardize states and fit PCA within each training fold. Four-fold prompt cross-validation on the 128 fitting prompts selects dimension and penalty; no evaluation labels enter this choice. T2G + state appends the selected standardized PCA coordinates to the 119 quadratic trajectory terms, then selects its penalty on the same folds. Each checkpoint has its own fit.

Table 9: Current state and trajectory features on looped models. EAURC $\times 10^3$ (↓). R: RISK-T2G; S: state probe; R+S: T2G + state. Gaps are R−(R+S), with paired 95% prompt intervals.

(a) Main and second cohorts: 256 prompts per model

Model	Cohort	R	S	R+S	Reduction	Gap [95%]
Huginn-0125	Main	8.822	12.873	8.054	8.7%	.768 [.599, .945]
	Second	8.287	12.865	7.750	6.5%	.538 [.389, .686]
Ouro-1.4B	Main	19.237	22.350	16.322	15.2%	2.915 [2.005, 3.818]
	Second	16.861	18.898	14.518	13.9%	2.343 [1.380, 3.273]
Ouro-2.6B	Main	12.416	16.028	11.427	8.0%	.989 [.143, 1.770]
	Second	10.450	14.345	10.149	2.9%	.301 [−.522, 1.105]

(b) Fresh policy cohort: 512 prompts per model

Model	Risk-T2G	T2G + state	Reduction (%)	Risk − joint [95% CI]
Huginn	8.583	7.838	8.7	.745 [.638, .849]
Ouro-1.4B	18.636	16.205	13.0	2.431 [1.761, 3.090]
Ouro-2.6B	11.645	10.620	8.8	1.025 [.424, 1.588]

For looped models, dimensions are selected from $\{64, 128, 256, 512\}$ and penalty multipliers from $\{1, 10, 100\}$. Huginn selects 512 coordinates; both Ouro scales select 256. For ordinary layers, the dimension grid is $\{32, 64, 128, 256\}$ and the penalty grid is $\{.1, 1, 10\}$; all eight cells select 256 coordinates and multiplier 10. The comparison is target- and data-matched, but not parameter-count matched to the fourteen-input T2G readout.

Looped models. On the main cohort, the state probe's excess EAURC over RISK-T2G is 4.051 $[3.594, 4.552]$, 3.113 $[1.765, 4.488]$, and 3.613 $[2.495, 4.857]$ in model order (EAURC $\times 10^3$). Table 9 evaluates their combination on the main cohort and two additional cohorts, without changing the fitting data or configuration. On the 512 unused prompts per model used for policy evaluation, augmentation reduces EAURC by 8.7%, 13.0%, and 8.8%, respectively, with positive paired intervals throughout. Ranking is a pre-specified secondary outcome of that study, distinct from the policy comparison in Appendix E. The separate 256-prompt cohort had been evaluated in an earlier readout study but was not used to select state-based predictors. There, Huginn and Ouro-1.4B improve, while Ouro-2.6B's 2.9% gain has an interval crossing zero. We retain the fourteen-input RISK-T2G as the primary readout; the extension remains optional.

Ordinary transformer layers. These probe comparisons reuse the depth-study evaluation rows, with hyperparameters chosen only on the fitting role. The linear state probe has lower EAURC than RISK-T2G in six of eight cells; the exceptions are Qwen1.5-1.8B and Qwen2.5-7B with the moment lens. State augmentation improves on RISK-T2G in all eight cells. On the six sub-2B cells, descriptive re-scoring against the state probe gives reductions of 3.0–6.4% under the logit lens and 14.7–25.2% under the moment lens, with all paired intervals positive. These comparisons support complementarity, not a common ordering of state and trajectory predictors across architectures. At 7B, RISK-T2G lowers EAURC by 19.0% relative to the moment-lens state probe, a gap of 140.9 $[119.5, 162.4]$; the ordering reverses under the logit lens. State augmentation improves on RISK-T2G by 435.6 under the logit lens and 65.5 $[58.9, 72.0]$ under the moment lens. The selected joint penalty is 10 for the logit lens and 1 for the moment lens. The corresponding mean terminal KLs are 9.76 and 2.72 nats.

C.4 PUBLISHED STOPPING CRITERIA ON MATCHED TRAJECTORIES

Table 1 applies Popescu et al. (2026)'s criteria per token: $H(p_t)$, negative maximum probability and logit margin, $D_{\mathrm{KL}}(p_t \| p_{t-1})$, $\|\ell_t - \ell_{t-1}\|_2$, $\|h_t - h_{t-1}\|_2$, and $1 - \cos(h_t, h_{t-1})$. Predictive KL also follows Logan (2026); native-gate survival is Equation 12. We use exposed pre-head states, our checkpoint grid and terminal-fidelity endpoint, without fitting transformations or reproducing the original task-accuracy and generation protocols.

For Huginn's additional released rules, relative latent change is $\|s_t - s_{t-1}\|_2 / \|s_t\|_2$: s_t is Huginn's core state before the coda or Ouro's recurrent readout. Sequence pooling becomes tokenwise

evaluation. Argmax patience counts unchanged transitions from iteration 1, resetting to zero on a change; it observes all iterations but uses the shared candidate exits. Its AURC averages over tied orders. Successive KL retains the opposite orientation to predictive KL.

All criteria use the original 128 calibration and 256 confirmation prompts per model, with identical rows and terminal labels. Paired EAURC intervals favor Risk over every tested raw criterion; Table 6 reports the strongest prior comparisons.

C.5 Closed-form extrapolation

Equation 10 can be evaluated without fitting: $u_t\big(\sum_{k=1}^{T-t}\widehat{\rho}_t^{\,k}\big)^2$ with $\widehat{\rho}_t = \mathrm{clip}(\rho_t, -.99, .99)$, or with the energy ratio $\sqrt{u_t/u_{t-1}}$ in place of ρ_t. On the Table 1 cohort, secant EAURCs are 53.717, 60.411, and 37.776; decay gives 35.434, 59.853, and 26.708, versus 10.342, 46.891, and 19.049 for u_t alone (model order; $\times 10^3$). Both rank terminal error worse than raw update energy u_t, with paired intervals excluding zero. The comparison supports learning how observed geometry relates to terminal error instead of imposing a fixed geometric tail. Update energy alone is the strongest tested zero-parameter criterion on Huginn.

D Generalization across corpora, horizons, and layers

D.1 Cross-corpus transfer

Full-method transfer. On the six cells of Table 3, paired EAURC reductions over matched Risk confidence, scaled by 10^3, are 1.150 $[.702, 1.608]$ and $.461$ $[.361, .559]$ for Huginn, 2.461 $[1.183, 4.507]$ and $.887$ $[-.322, 1.778]$ for Ouro-1.4B, and $.541$ $[.182, .907]$ and $.501$ $[.121, 1.067]$ for Ouro-2.6B (C4, then LAMBADA). At the stricter $.02$ calibration target, realized Risk KL ranges from $.0142$ to $.0195$, below the target in all six cells. At the primary $.05$ target, allocation differences against successive KL are $+.74$ and $-.47$ points for Huginn, $+6.72$ and $+7.09$ for Ouro-1.4B, and $+.26$ and $+.22$ for Ouro-2.6B; each paired interval excludes zero. Realized risks differ across these frozen policies, so allocation gains at a shared calibration target need not imply gains at matched test risk.

Output-only transfer ablation. On separate source-disjoint cohorts, output-only log readouts also transfer from WikiText-2 to C4 and LAMBADA. Relative to GBDT confidence, core T2G reduces EAURC by 5.4–6.4% on Huginn and 11.7–16.6% on Ouro, with positive paired intervals throughout.

D.2 Reference horizons and transformer layers

The horizon and sub-2B depth studies use source-disjoint WikiText-2 train text, with 128 fitting prompts and 256 confirmation prompts per study. Model–lens comparisons are specified before confirmation, and intervals use 2,000 paired whole-prompt bootstrap resamples. Heads are fitted separately for each model, checkpoint, and target; the fourteen features, two objectives, and mixing coefficient remain fixed. Qwen2.5-7B uses the same depth-study passages with its own 128-prompt fit and 256-prompt evaluation. This evaluates a larger model on shared source text.

Reference horizon. Table 10 refits RISK-T2G and the matched two-head confidence control to farther references. Huginn runs to 64 recurrences; Ouro runs to 16 loops. Huginn changes little beyond step 32, so its farther references produce similar error distributions. Ouro, trained for four loops, drifts when run past loop 8: mean KL from loop 6 to loops 8, 12, and 16 is .037, .217, and .629 on 1.4B, and top-1 disagreement rises from 9% to 30%. The farther Ouro targets therefore measure instability under continued recurrence rather than further convergence. The advantage over matched confidence persists in all nine cells. The relative gain is approximately flat on Huginn and not monotone on Ouro. The native-fit readout transfers to the farther Huginn references within $\pm.02$ EAURC$\times 10^3$ of refitting; on Ouro, the drift targets instead require horizon-specific fitting.

Depth trajectories. Section 5.5 uses layers 12, 15, 18, and 21 of Qwen1.5-0.5B, Qwen1.5-1.8B, and Pythia-410m, and layers 14, 17, 21, and 24 of Qwen2.5-7B, all in FP32. The logit lens applies the model's final norm and unembedding to the pre-norm residual stream. The moment lens is a per-layer affine map $M_l h_l + b_l$ from the pre-norm residual stream to the final post-norm state, fitted by ridge regression on the fit prompts with relative ridge $10^{-2}\,\mathrm{tr}(\Sigma_{xx})/d$ and frozen before confirmation; the reading is the unembedding of that map. Under the moment lens, mean final-layer KL falls to 1.2–1.6

Table 10: Reference-horizon confirmation. E is EAURC $\times 10^3$ for matched two-head confidence (C), RISK-T2G refitted to the reference (R), and RISK-T2G fitted to the native reference and applied unchanged (R, native). Gap is C−R with paired 95% intervals.

Model	T'	Mean KL	E_{C}	E_{R}	$E_{\mathrm{R,\,native}}$	Reduction	Gap [95%]
Huginn-0125	32	0.1863	9.377	8.590	8.590	8.4%	0.787 [0.651, 0.920]
Huginn-0125	48	0.1877	9.964	9.124	9.117	8.4%	0.840 [0.701, 0.979]
Huginn-0125	64	0.1882	10.264	9.392	9.377	8.5%	0.872 [0.723, 1.016]
Ouro-1.4B	8	0.1331	21.352	18.546	18.546	13.1%	2.806 [1.405, 3.952]
Ouro-1.4B	12	0.3529	108.119	90.103	163.677	16.7%	18.016 [12.674, 23.767]
Ouro-1.4B	16	0.7863	284.613	246.824	548.418	13.3%	37.789 [27.345, 48.029]
Ouro-2.6B	8	0.0865	11.312	9.792	9.792	13.4%	1.520 [0.927, 2.179]
Ouro-2.6B	12	0.2471	78.521	73.308	134.789	6.6%	5.214 [0.285, 10.537]
Ouro-2.6B	16	0.6343	256.344	235.647	517.108	8.1%	20.697 [8.674, 33.053]

Table 11: Frozen token policies at calibration KL budget .05, on the Table 1 cohort. KL is realized unconditional risk; saved is gross iterations avoided (%). Gain compares mean-KL only with each policy in percentage points, with paired 95% intervals.

Model	Policy	KL ↓	Saved ↑	Mean − policy
Huginn-0125	Mean conf. (119)	.0447	63.08	.55 [.36, .73]
	RISK-T2G (ours)	.0451	63.26	.37 [.26, .48]
	T2G, mean-only	.0463	63.63	–
Ouro-1.4B	Mean conf. (119)	.0477	36.99	.81 [.57, 1.05]
	RISK-T2G (ours)	.0444	36.25	1.55 [1.32, 1.80]
	T2G, mean-only	.0453	37.80	–
Ouro-2.6B	Mean conf. (119)	.0497	47.92	.73 [.57, .89]
	RISK-T2G (ours)	.0496	47.00	1.65 [1.46, 1.84]
	T2G, mean-only	.0521	48.65	–

nats from 2.8–10.0 under the logit lens in the three sub-2B models; at 7B the corresponding means are 2.72 and 9.76. The relative advantage of RISK-T2G persists. Per-layer relative reductions under the moment lens are 6/7/14/34% (Qwen1.5-0.5B), 12/16/30/22% (Pythia-410m), and 10/8/11/41% (Qwen1.5-1.8B) at layers 12/15/18/21. Zero-parameter criteria are weak here: under the logit lens, successive KL between adjacent layers has Spearman correlation .00–.46 with final-layer KL across the three models.

E KL ALLOCATION AND STOPPING POLICIES

E.1 FROZEN KL POLICIES AND FITTING OBJECTIVES

All policies are fitted to the same trajectories and independently calibrated by the same rule before confirmation. Their realized risks differ, so the positive paired allocation gaps do not establish dominance at matched test risk. For mean-KL only, realized-KL intervals are $[.0434, .0494]$, $[.0430, .0476]$, and $[.0489, .0554]$; the 2.6B point exceeds .05. Risk's corresponding 2.6B interval, $[.0461, .0532]$, spans the same target.

E.2 LEARNED CONTROLS AT A COMMON RISK CAP

Table 12 compares all learned predictors on the same 256 primary prompts, using their saved scores without refitting. Every method uses first-accepted-checkpoint stopping and the same 1,001-quantile threshold grid. For each KL cap, we maximize gross iteration avoidance subject to the empirical risk constraint; this is the frontier defined below, not a frozen deployable policy. The comparison is retrospective. Paired intervals use 2,000 whole-prompt resamples and reoptimize the grid threshold within each resample, holding fitted predictors fixed.

Table 12: Learned policy controls on the primary 256-prompt cohort. Evaluation-selected frontiers under common KL caps, excluding scoring cost. Paired intervals condition on existing fits; this comparison is retrospective.

(a) Iterations avoided at a common empirical KL cap (%)

	Huginn		Ouro-1.4B		Ouro-2.6B	
Score / KL cap	.02	.05	.02	.05	.02	.05
Selected prior	55.68	63.93	21.68	36.44	31.03	46.79
GBDT confidence	55.27	63.59	17.21	33.47	28.31	46.55
Risk confidence (119)	55.63	63.91	22.17	36.38	31.34	46.63
Mean confidence (119)	55.56	64.25	23.35	37.72	33.27	48.02
State probe	53.76	61.91	21.75	36.03	28.97	45.08
Risk-T2G	55.94	64.52	24.30	38.33	31.62	47.11
T2G, mean-only	56.28	64.48	25.99	39.60	32.40	47.91
T2G + state	56.11	64.39	25.59	39.62	32.72	47.90

(b) Gain over Mean confidence (119), percentage points [95% CI]

Model / KL cap	Risk-T2G	T2G, mean-only	T2G + state
Huginn / .02	.38 [.18, .81]	.72 [.31, 1.09]	.55 [.18, .89]
Huginn / .05	.27 [−.05, .60]	.23 [−.04, .59]	.14 [−.19, .60]
Ouro-1.4B / .02	.96 [−.31, 2.35]	2.64 [1.61, 3.55]	2.24 [.85, 3.33]
Ouro-1.4B / .05	.61 [−.47, 1.56]	1.88 [1.15, 2.52]	1.90 [1.05, 2.60]
Ouro-2.6B / .02	−1.64 [−3.58, −.24]	−.87 [−1.97, .40]	−.55 [−2.18, .79]
Ouro-2.6B / .05	−.92 [−1.90, .06]	−.11 [−1.01, .56]	−.12 [−1.06, .74]

Mean-only and state augmentation outperform mean confidence on Huginn at .02 and Ouro-1.4B at both caps. Their gains are inconclusive on Huginn at .05 and at both Ouro-2.6B caps. Risk-T2G is below mean confidence on Ouro-2.6B at .02 by 1.64 points $[.24, 3.58]$. The state-only probe has a lower point frontier than every T2G readout in all six settings. These controls support model-dependent use of trajectory information, not uniform policy superiority of the two-head rule.

E.3 Allocation frontiers and calibrated policies

We distinguish a score's attainable allocation from the performance of a threshold chosen on separate calibration data. Table 13 evaluates Risk-T2G and T2G + state on 512 unused prompts per model, with settings and calibrated policies fixed before collection. The published comparator is selected by calibration saving. At budgets .02 and .05, the selected criteria are predictive KL/predictive KL for Huginn, hidden L2/hidden cosine for Ouro-1.4B, and logit L2/predictive KL for Ouro-2.6B. The same fitted readouts and calibration choices apply at both cohort sizes; neither predictors nor policies are tuned to these new prompts.

Empirical allocation frontier. For each score, we sweep the reject-all option and 1,001 evaluation-score quantiles. Let $\widehat{R}(\theta)$ be mean terminal KL under first-accepted checkpoint stopping, including zero loss for terminal fallbacks, and let $\widehat{S}(\theta)$ be mean gross iteration avoidance. We report $S^*(\epsilon) = \max_{\theta:\widehat{R}(\theta)\le\epsilon} \widehat{S}(\theta)$: the best saving on this grid under a common empirical KL cap, not necessarily identical realized risks. Thresholds are chosen on the evaluation cohort, so this measures the score's allocation frontier rather than a deployable policy. Paired 95% intervals recompute the maximizing threshold within each of 2,000 whole-prompt bootstrap samples.

T2G + state's gains range from .34 to 3.91 points, with positive intervals for both Ouro models at both caps and for Huginn at .05. At Huginn's .02 cap the gain is .34 $[-.002, .68]$; Risk-T2G gains .35 $[-.03, .57]$. The pre-specified requirement of positive intervals in all six settings is therefore not met by either readout. Against the best published criterion chosen within each evaluation bootstrap, the same five comparisons have positive intervals. Descriptively, both T2G readouts match or exceed the best prior's point frontier at every tested cap from .005 to .10. State-only and learned-confidence policies appear in the primary-cohort comparison above, not in this fresh-cohort evaluation.

Table 13: Allocation on 512 fresh WikiText prompts per model. Prior: calibration-selected criterion; Risk: RISK-T2G; + state: T2G + state. Iterations avoided are percentages; gaps are percentage points with paired 95% intervals. KL is in nats.

(a) Evaluation-selected allocation frontier

Model / KL cap	Prior	Risk	+ state	Risk − prior	+ state − prior
Huginn / .02	55.28	55.63	55.63	.35 [−.03, .57]	.34 [−.002, .68]
Huginn / .05	63.91	64.33	64.68	.42 [.14, .59]	.77 [.45, .93]
Ouro-1.4B / .02	20.81	23.65	24.72	2.84 [1.56, 3.94]	3.91 [2.68, 5.14]
Ouro-1.4B / .05	36.42	38.57	39.45	2.14 [1.29, 2.86]	3.03 [2.32, 3.86]
Ouro-2.6B / .02	30.53	32.35	33.10	1.82 [.47, 2.94]	2.56 [1.26, 3.65]
Ouro-2.6B / .05	46.71	47.47	48.00	.76 [.43, 1.19]	1.29 [.85, 1.67]

(b) Frozen-threshold policies: realized KL / iterations avoided

Model / target	Prior	Risk	+ state	+ state − prior
Huginn / .02	.0184 / 54.61	.0187 / 55.12	.0172 / 54.61	.01 [−.10, .13]
Huginn / .05	.0462 / 63.18	.0451 / 63.26	.0445 / 63.43	.25 [.05, .45]
Ouro-1.4B / .02	.0172 / 19.06	.0157 / 20.67	.0198 / 24.64	5.58 [4.94, 6.21]
Ouro-1.4B / .05	.0473 / 35.32	.0433 / 35.81	.0421 / 36.34	1.03 [.45, 1.59]
Ouro-2.6B / .02	.0198 / 30.45	.0201 / 32.44	.0178 / 31.52	1.07 [.62, 1.51]
Ouro-2.6B / .05	.0485 / 46.17	.0472 / 46.38	.0476 / 47.04	.87 [.67, 1.07]

On the earlier 256-prompt Table 1 cohort, the corresponding T2G + state gains were .44–3.91 points with positive intervals in all six settings. Together, the cohorts support allocation gains beyond the selected prior, while the small incremental benefit on Huginn at .02 remains uncertain.

Separately calibrated policies. The T2G + state predictor is fitted on the original 128 fitting prompts using the configuration in Appendix C.3. Its thresholds use the separate 128-prompt calibration role and the rule in Appendix A.1; Risk and prior thresholds are unchanged. All six state-augmented policies have point-estimate KL within their target, and five of six saving intervals over the selected prior exclude zero. Four policies also have lower observed KL than the prior: Huginn at .05, Ouro-1.4B at .05, and Ouro-2.6B at both caps. At Huginn's .02 cap, the calibrated saving gap is .01 $[-.10, .13]$. Point-estimate compliance is not a risk guarantee; the risk intervals cross the cap for Ouro-1.4B at .02 and Ouro-2.6B at .05. Realized risks differ between policies, so the calibrated gaps are distinct from the frontier comparison. Iteration avoidance excludes state projection and scoring; this analysis does not measure execution latency.

F MMLU EXECUTION, MATCHED CONTROLS, AND REPLICATION

Table 14 reports the primary MMLU policies of Section 5.6 on 1,024 questions. Timing uses the matched 128-question subset and includes all scoring costs.

F.1 MATCHED TASK CONTROLS

The additional controls use the same 1,024 fit and 507 calibration questions as TASK-T2G. Linear confidence uses answer entropy, margin, and successive answer-space KL with the same standardized logistic map ($C = .1$, balanced classes); confidence plus gate adds only g_t. The quadratic control uses a degree-two expansion and $C = 1$ without class balancing, and GBDT uses the tree settings above. Removing hidden motion or the gate retains the Task-T2G readout class and penalty. All controls are fitted without confirmation labels and use the same 2% calibration rule. These are retrospective matched controls on saved cohorts; the original TASK-T2G fit and thresholds remain unchanged.

At the calibrated operating point, TASK-T2G exits more often than the linear, quadratic, and boosted confidence controls at both scales. Adding the native gate makes the simpler confidence control competitive. The primary Task-minus-gated-confidence exit gaps are $-.29$ points $[-1.46, .78]$ and

Table 14: Five-shot MMLU at a 2% calibration target (percentages). (a) Policies and timing; (b) Task-T2G's 95% intervals. Exit gains compare inverse margin; accuracy changes and time savings compare pass 4. Answer change is unconditional.

(a) Policies and execution cost

Model	Policy	Change	Exit	Accuracy	Time saved
Ouro-1.4B	Fixed pass 4	0	0	66.50	0
	Inverse margin	1.46	69.63	66.21	14.23
	Confidence + gate	1.95	74.90	66.02	14.33
	TASK-T2G (ours)	1.86	74.61	66.11	13.02
Ouro-2.6B	Fixed pass 4	0	0	75.00	0
	Inverse margin	1.17	77.44	75.10	18.65
	Confidence + gate	1.95	81.35	74.61	18.50
	TASK-T2G (ours)	1.66	80.76	74.80	18.85

(b) Task-T2G uncertainty estimates

Model	Answer change	Accuracy Δ	Exit gain	Time saved
Ouro-1.4B	$1.86\ [1.07, 2.73]$	$-.39\ [-1.17, .39]$	$4.98\ [3.42, 6.54]$	$13.02\ [10.82, 14.96]$
Ouro-2.6B	$1.66\ [.88, 2.44]$	$-.20\ [-.98, .59]$	$3.32\ [1.86, 4.88]$	$18.85\ [17.13, 20.39]$

Table 15: Matched MMLU controls on the primary 1,024-question confirmation. Exit and change are percentages; change is unconditional. AURC is scaled by 10^3 and measures answer-disagreement ranking over all coverages.

	Ouro-1.4B			Ouro-2.6B		
Policy	Exit	Change	AURC	Exit	Change	AURC
Inverse margin	69.63	1.46	22.04	77.44	1.17	10.13
Successive KL	23.24	1.46	91.93	39.06	1.27	42.97
Native gate	46.78	1.86	43.43	43.16	.98	30.07
Confidence (linear)	70.90	1.46	21.64	77.73	1.56	10.28
Confidence (quadratic)	70.41	1.56	23.44	77.54	1.17	17.86
Confidence (GBDT)	65.53	1.17	28.10	73.44	1.07	11.48
Confidence + gate	74.90	1.95	20.45	81.35	1.95	10.72
Task without hidden	75.20	1.95	20.41	79.49	1.76	10.62
Task without gate	73.54	1.76	20.59	78.91	1.56	10.48
TASK-T2G	74.61	1.86	20.04	80.76	1.66	10.90

$-.59$ points $[-1.95, .78]$. The hidden component has a model-dependent contribution, and higher exit coverage need not imply a lower full-curve AURC.

F.2 SECOND-COHORT EVALUATION

On the second, deduplicated cohort, the unchanged TASK-T2G thresholds exit on 76.79% and 81.29% of questions, with answer changes of 1.86% and 1.76%. The exit gains over margin are 4.21 points $[2.74, 5.68]$ and 1.18 points $[-.29, 2.64]$. Gated confidence exits on 76.49% and 83.45%, with changes of 1.76% and 2.45%, respectively. These paired checks characterize the coverage–risk tradeoff at different realized risks on held-out questions.