WHEN DO STATE-BASED DIAGNOSTICS TRANSFER IN SOFT-PROMPT TUNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how adaptation changes language-model behavior is important for reliable evaluation and deployment. In soft-prompt tuning, we ask whether an internal diagnostic's advantage over answer loss transfers across prompt groups and output rules, and whether its ranking gains justify measurement costs. We propose a control-response evaluation framework to address these questions in frozen language models. First, in two Ministral–ECQA prompt pairs, removing continuation reverses the ranking of state and loss diagnostics under strict parsing, while preserving diagnostic scores and initially selected answers. This shows that detecting parseability changes can drive an apparent diagnostic advantage. Second, crossed prompt–question comparisons in Ministral–ECQA show larger changes in relative diagnostic performance across prompt groups than across question sets. The group contrast includes pair selection. Predictions fixed before confirmation reproduce this shift on fresh questions for the same prompts and improve on predicting no group difference. Pair-specific gains over a common-mean forecast remain uncertain, as do gains on unseen GSM8K prompts. Third, at the measured ECQA reference budgets, ranking gains do not offset checks lost to scoring, and direct checks outperform screening with the three original diagnostics on average. The framework provides a reproducible procedure for interpreting diagnostic advantages, testing their transfer, and deciding whether internal measurement is worth its evaluation cost.

1 INTRODUCTION

Researchers commonly use internal-state measurements to understand and monitor language-model behavior. Hidden-state probes can predict statement truthfulness (Azaria and Mitchell, 2023), while activation-based analyses inform interventions on model computation. Their interpretation depends on the evaluation setting: activation-patching conclusions vary with metrics and perturbations (Zhang and Nanda, 2024), and localization need not identify effective editing sites (Hase et al., 2023). A diagnostic that ranks behavioral changes well on existing prompts may not retain its advantage over answer loss when prompts, questions, or output rules change. When these factors change together, aggregate comparisons cannot distinguish their contributions.

Soft-prompt tuning adapts a frozen model using a trainable prefix (Lester et al., 2021; Li and Liang, 2021). We treat the prefix as a static control input and distinguish hidden-state response, continuous task readout, and parsed behavior. Fixed diagnostics rank questions whose correct-and-parseable status changes between two prompts.

We use three linked tests. First, removing continuation while preserving diagnostic scores and initial answers reverses state-versus-loss rankings in two Ministral–ECQA pairs under strict parsing, showing how parseability changes can favor a diagnostic. Second, crossed prompt–question panels compare old pairs selected for similar loss with new pairs fixed before training. Forecasts fixed before confirmation reproduce the group difference in relative diagnostic performance on fresh questions for the same Ministral prompts, with uncertain gains over a common mean. GSM8K tests unseen training runs. Third, direct checks outperform the three original screens on average at the measured ECQA reference budgets after charging scoring and checking costs. These tests provide a control-response framework for interpreting diagnostic advantages, testing transfer, and choosing between internal measurement and direct checks.

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.

2 STUDY DESIGN AND CONTROL-RESPONSE DIAGNOSTICS

Behavioral target and unit of analysis. We study fixed diagnostic comparisons in supervised offline evaluation. Predictive accuracy and measurement value are distinct targets. For question i and prompt u_s, let p_{is} indicate whether the generated answer is parseable. Define risk $b_{is} = 0$ for a correct parsed answer and $b_{is} = 1$ for an incorrect or unparseable answer. For two prompts, the target is

$$z_{i,st} = |b_{is} - b_{it}|. \tag{1}$$

A *risk change* therefore includes either direction of change in scored correctness, but excludes two different answers that are both wrong. Larger diagnostic scores predict $z = 1$. We evaluate this ranking with AUC, comparing changed and unchanged questions within a prompt pair and giving tied scores half credit. AUC is undefined if either class is absent. Model summaries weight pairs equally unless stated otherwise.

Prompt optimization and comparison groups. For question x_i with correct choice symbol y_i, ECQA loss $L_i(u)$ is the teacher-forced negative log probability of y_i under prompt u. For GSM8K, $L_i(u)$ is the mean token negative log likelihood of the complete reference solution, including its final-answer field. With the backbone frozen, we optimize an eight-vector prompt $u \in \mathbb{R}^{8\times d}$, where d is the prompt-vector (embedding) dimension, by minimizing

$$F(u) = |\mathcal{D}_{\text{opt}}|^{-1} \sum_{i\in\mathcal{D}_{\text{opt}}} L_i(u). \tag{2}$$

AdamW trains prompts on 128 questions. 64 disjoint questions select a checkpoint, including initialization (Appendix A). ECQA discovery (Aggarwal et al., 2021) uses Granite-4.1-8B and Ministral-3-8B-Instruct-2512-BF16 (IBM Granite Team, 2026; Mistral AI, 2025). Of sixteen runs per model, we select the first three pairs in seed order whose mean loss difference on 300 calibration/development questions has a 90% paired interval within $\pm 0.10 s_d$. s_d is the sample standard deviation of itemwise loss differences (Appendix A). Two Ministral pairs share a prompt and remain fixed cases with joint question resampling. Independent validation adds 32 prompts per model, including Qwen3.5-9B (Qwen Team, 2026), paired by adjacent seeds before training. Every prompt belongs to one pair and all pairs are evaluated.

Control response and task readout. Let q be the hidden-state dimension at a scored answer position and $h_i : \mathbb{R}^{8d} \to \mathbb{R}^q$ the frozen model's response to the vectorized prompt. Write $\Delta h_i = h_i(u_t) - h_i(u_s)$ and let $r_i : \mathbb{R}^q \to \mathbb{R}$ be a differentiable scalar task readout, mapping the relevant hidden state to a continuous task score. At u_s, the response Jacobian with respect to the vectorized prompt and the readout gradient satisfy

$$K_{i,s} = \left.\frac{\partial h_i}{\partial \operatorname{vec}(u)}\right|_{u_s}, \quad g_{i,s} = \nabla_h r_i(h_i(u_s)), \quad \nabla_{\operatorname{vec}(u)}(r_i \circ h_i)(u_s) = K_{i,s}^\top g_{i,s}. \tag{3}$$

This standard adjoint relation connects state measurement to prompt sensitivity. Form a response matrix R whose columns are calibration-perturbation state changes across items and scored positions. Its singular value decomposition defines an orthonormal rank-r empirical response basis:

$$R = V\Sigma W^\top, \qquad U = V_{[:,1:r]}, \qquad M = RR^\top/N,$$

where N is the number of response columns. Equivalently, U contains the leading eigenvectors of the uncentered second moment M. The projected term $g_{i,s}^\top UU^\top \Delta h_i$ measures alignment of these responses with the readout. The local readout approximation and its smoothness conditions are given in Appendix G. Discrete ranking also depends on score margins and behavioral labels (Appendix F.1).

Three fixed diagnostic scores. We estimate the response basis from 32 soft-prompt perturbations. Five-fold reconstruction assessment selects the smallest rank within one standard error of the best candidate in $\{1, 2, 4, 8, 16, 32\}$. All three ECQA models select rank 32. High residual energy indicates limited coverage by the low-rank response basis. Appendix A reports rank sensitivity. The answer-loss score is the absolute difference in teacher-forced loss:

$$s_{\text{loss}}(i, s, t) = |L_i(u_t) - L_i(u_s)|. \tag{4}$$

For ECQA, the *state/readout score* uses the differentiable correct-choice readout

$$r_i(h) = \ell_{i,y_i}(h) - \log \sum_{a \neq y_i} \exp \ell_{i,a}(h), \tag{5}$$

where $\ell_{i,a}(h)$ is the logit for choice symbol $a \in \{A, \ldots, E\}$ at the answer position for question x_i. The sum covers the other four choices, so the readout is the correct choice's log odds against those alternatives.

Our cross-fitted prompt-effect decomposition (CF-PED) partitions the readout response into projected and complementary terms. Its ridge diagnostic predicts $z_{i,st}$ from left-endpoint approximations to those terms, together with state magnitudes, prompt displacement, and calibration tail energy. For one ECQA scored position, with $\Delta u = u_t - u_s$, the six features and score are

$$x = \big(g_s^\top U U^\top \Delta h, \quad g_s^\top (I - UU^\top)\Delta h, \quad \|h_s\|_2, \quad \|h_t\|_2, \quad \|\Delta u\|_F, \quad e_{\text{tail}}(r)\big),$$
$$e_{\text{tail}}(r) = \text{tr}(M) - \text{tr}(U^\top M U), \qquad s_{\text{state}} = \beta_0 + \beta^\top \big[(x - \mu) \oslash \tilde{\sigma}\big].$$

Here $\oslash$ denotes elementwise division. Feature means μ and population standard deviations $\tilde{\sigma}$ are estimated from training rows, with zero scales replaced by one. Ridge regression fits an intercept and coefficients by penalized squared loss. Predictions are not clipped. Using risk-change labels from 200 calibration and 100 development questions, question-grouped five-fold validation selects the strongest penalty in $\{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ within one standard error of the best MAE, followed by fitting on all 300 questions.

Regularization selection by within-pair AUC and removal of pair-constant features do not consistently improve independent-pair ranking (Appendix A, Table 5).

The ECQA *state-norm score* separately fits $\|h_s\|_2$, $\|h_t\|_2$, and $\|\Delta u\|_F$ by the same procedure. It excludes gradient terms and fixed tail energy, while retaining supervised calibration. The discovery random-subspace control refits the predictor for each of 32 rank-matched random bases and takes the median prediction per record. Ministral and Granite reuse their original diagnostics. Qwen fits diagnostics using 16 separate calibration prompts with a non-thinking template. Bases, normalizers, coefficients, and directions remain fixed during validation.

Generation and uncertainty. The main condition uses native greedy decoding, strict single-letter parsing, and at most eight tokens. For independent and crossed panels, one trajectory of at most 32 tokens per endpoint supplies prefixes of 1, 2, 4, 8, 16, and 32. Strict parsing removes Unicode whitespace and accepts one uppercase A–E letter. Initial-letter parsing requires an opening A–E followed by end of text, whitespace, ASCII punctuation except underscore, or supported Chinese punctuation. Terminal-field parsing uses an explicit final answer field, otherwise falling back to the initial-letter rule. Malformed or conflicting fields are unparseable. Parsers do not consult gold answers, and all conditions reuse the same fixed diagnostics. Exact parsing rules appear in Appendix D.

Primary analyses share 20,000 crossed bootstrap draws over questions and independent prompt pairs. Bonferroni families contain 18 independent-ECQA comparisons, 24 crossed/fresh-question comparisons, and the original 22 GSM8K and readout comparisons, including those outside the reduced scope. Fixed discovery pairs receive joint question resampling only. Undefined AUCs remain undefined under the coverage rules in Appendix D.

3 Why the output rule changes diagnostic rankings

Diagnostic rankings depend on which parseability and correctness changes enter the scored behavioral target. Let $p_s, p_t \in \{0, 1\}$ indicate whether outputs under prompts s and t are parseable. Among behavioral-change positives ($z = 1$), define two disjoint channels:

$$Z_P = \{z = 1,\ p_s \neq p_t\}, \qquad Z_C = \{z = 1,\ p_s = p_t = 1\}. \tag{6}$$

A positive belongs to the parseability channel Z_P when exactly one output is recognized by the parser as a valid answer. It belongs to the correctness channel Z_C when both outputs are parseable but differ in correctness. Under our risk definition, these channels exhaust the positive examples. Both channel analyses use the same negative pool $Z_0 = \{z = 0\}$, including parseability changes that leave both outputs incorrect. No correctness value is assigned to an unparseable output.

Table 1: Independent ECQA equal-pair AUCs. All 16 pairs per model are valid. Additional probe and logit results are point estimates.

Model	State/ readout	Supervised norm	Raw norm difference	Linear probe	Loss	Gold margin	JS
Ministral	0.460	0.353	0.550	0.464	0.834	0.690	0.850
Granite	0.549	0.542	0.541	0.580	0.731	0.756	0.761
Qwen	0.622	0.642	0.481	0.781	0.895	0.658	0.885

In the discovery analysis, AUC pools the three fixed loss-matched prompt pairs and therefore includes comparisons across pairs. Later transfer analyses compute AUC within each pair before averaging across pairs. Channel AUC compares positives from Z_P or Z_C with the common negative pool Z_0, giving tied scores half credit. With $\pi = \Pr(Z_P \mid z = 1)$, the partition of positives and shared negative distribution give

$$\mathrm{AUC}(s) = \pi\,\mathrm{AUC}_P(s) + (1-\pi)\,\mathrm{AUC}_C(s). \tag{7}$$

Appendix F derives the finite-sample identity using the same positive–negative comparisons.

For two scores, define $\delta_k = \mathrm{AUC}_k(s_1) - \mathrm{AUC}_k(s_2)$ for $k \in \{P, C\}$. Their overall AUC difference is

$$\Delta(\pi) = \pi\delta_P + (1-\pi)\delta_C. \tag{8}$$

With $\delta_P > 0 > \delta_C$, the ordering changes at $\pi_* = -\delta_C/(\delta_P - \delta_C)$ when conditional score distributions and the negative pool stay fixed. New items and prompts can change both. Their transfer therefore requires measurement.

3.1 A SHARED RISK TARGET CONTAINS DIFFERENT CHANGES

Among 600 held-out records, Granite has 87 risk changes: 9 involve parseability and 78 occur between parseable answers. Ministral has 246: 187 and 59, respectively.

Loss favors correctness changes. Its correctness and parseability AUCs are 0.911 $[0.875, 0.944]$ and 0.324 $[0.189, 0.486]$ for Granite, and 0.869 $[0.827, 0.907]$ and 0.328 $[0.276, 0.380]$ for Ministral. These intervals describe the fixed-pair discovery panels (Appendix Figure 4).

Ministral's state/readout score instead favors parseability, with AUC 0.787 $[0.752, 0.821]$ versus 0.386 $[0.291, 0.475]$ for correctness. Norm and matched-random controls show the same parseability preference, with AUCs 0.785 and 0.791.

State and output baselines. A ridge probe predicts the same z from ordered endpoint states $[h_s; h_t]$, using original calibration questions and independent ECQA test pairs. Its AUC point estimates exceed CF-PED's in all three models. Loss remains higher (Table 1). Appendix A specifies splits and tuning.

Saved-logit baselines are gold-margin change $|r_i(h_t) - r_i(h_s)|$ (Eq. (5)) and Jensen–Shannon (JS) divergence between conditional A–E distributions. JS exceeds loss in Ministral and Granite but not Qwen. Only gold margin requires the gold answer to compute its score. Appendix A reports all five logit comparisons.

3.2 COMPOSITION ACCOUNTS FOR THE AGGREGATE REVERSAL

For Ministral, $\delta_P = 0.459$, $\delta_C = -0.483$, and $\pi = 0.760$ give an aggregate state/readout-minus-loss gap of 0.233 (Eq. (8), AUCs 0.691 versus 0.458), with crossover $\pi_* = 0.512$. Granite has only $\pi = 0.103$ parseability positives, and loss leads: 0.851 versus 0.596. Thus diagnostic rankings depend on which changes form the positive class. Appendix B gives channel and fixed-pair results. The data supplement retains all baseline and sparse-channel estimates.

3.3 FIXED-PAIR TRANSFER AND A CONTROLLED CHANGE OF READOUT

Within Ministral, two pairs have 96/107 and 90/101 risk changes involving parseability and favor state scores. A third has 1/38 and favors loss. A 200-item replication excludes all 892 previously used items

by source coordinate and normalized text. Its primary-minus-loss gaps are $+0.244$, $+0.354$, -0.449. Both state-score families repeat all three signs under six-comparison-adjusted, jointly item-resampled intervals (Appendix B).

The constrained-output control masks non-A–E tokens and stops after one answer token. All 1,000 native full-vocabulary first-token winners are already legal and preserved. On this panel, the intervention removes continuation without changing the initial choice. For example, `B. congratulated` becomes `B`, changing strict parseability. The 840 originally parseable outputs retain their labels. 124 of 160 newly parseable outputs expose a correct first letter. The two parse-dominated pairs' primary-minus-loss gaps change by -0.672 $[-0.893, -0.457]$ and -0.845 $[-1.069, -0.616]$, reversing their rankings. Both norm contrasts also pass the four-comparison correction. Appendices C and F.2 give the absolute AUCs and account for changes in both comparison sets. This supports an output-interface effect under the strict parser, without identifying a causal mechanism in the hidden states.

4 Transfer across prompt groups

For Ministral–ECQA, group adjustments recur on fresh questions. Gains over a common-mean forecast remain uncertain.

4.1 Independent training across three models

The three models supply 48 disjoint pairs and 38,400 endpoints on 400 shared new items. Under strict parsing, let G_L be the equal-pair mean state-minus-loss AUC gap at cap L, with $D = G_1 - G_8$. All 16 pairs per model and 20,000 shared item–pair draws are estimable (18 adjusted intervals, Table 13). Loss leads both state scores at both caps. Table 1 gives eight-token AUCs.

Stopping changes Ministral's norm gap by -0.065 $[-0.167, -0.002]$ while preserving loss's lead. Granite's state and loss AUCs both improve, with uncertain relative change. Qwen has 12,800 legal initial letters and parseable outputs throughout. Its 690 risk changes concern correctness. Gaps remain -0.273 and -0.253, favoring loss throughout. Appendix D and the supplement retain calibration, pair results, and coverage.

4.2 Separating prompt-group and question-set differences

We cross three loss-matched old pairs and 16 new pairs fixed by adjacent seeds before training on 200 old and 400 new questions. 200 further items confirm forecasts for the same prompts (Appendix E).

On the same 200 old questions, eight adjacent historical pairs yield Ministral norm/loss AUCs 0.410/0.776, versus 0.535/0.523 for three selected pairs. Pairing and endpoint coverage change without new training. Group contrasts therefore include selection (Table 15).

For a state-score family, let $G_{a,d}$ be its equal-pair mean AUC minus the loss-score AUC for prompt cohort a and item split d, with O and N denoting old and new. Strict parsing at eight tokens is the primary condition. The cohort, item, and interaction contributions are

$$\begin{aligned} P &= \tfrac{1}{2}[(G_{N,O} - G_{O,O}) + (G_{N,N} - G_{O,N})], \\ Q &= \tfrac{1}{2}[(G_{O,N} - G_{O,O}) + (G_{N,N} - G_{N,O})], \\ H &= G_{N,N} - G_{N,O} - G_{O,N} + G_{O,O}. \end{aligned} \tag{9}$$

The diagonal difference satisfies $P+Q = G_{N,N} - G_{O,O}$ exactly. Old pairs remain fixed equal-weight references. The new cohort receives pair resampling as well as joint item resampling across pairs and scores.

The Ministral change is larger across cohorts than across items. For state/readout, Ministral's cohort contribution is $P = -0.402$ $[-0.544, -0.266]$, compared with $Q = 0.007$ $[-0.093, 0.108]$ and $H = -0.022$ $[-0.202, 0.149]$. Norm yields $P = -0.464$ $[-0.591, -0.335]$, $Q = -0.030$ $[-0.149, 0.086]$, and $H = -0.025$ $[-0.214, 0.166]$. For both families, the smallest cohort magnitude allowed by these simultaneous intervals exceeds the largest allowed item magnitude. The gap change is predominantly associated with the prompt cohort, including its selection procedure. Granite's P values are -0.006 and -0.009, with intervals spanning zero (Figure 1).

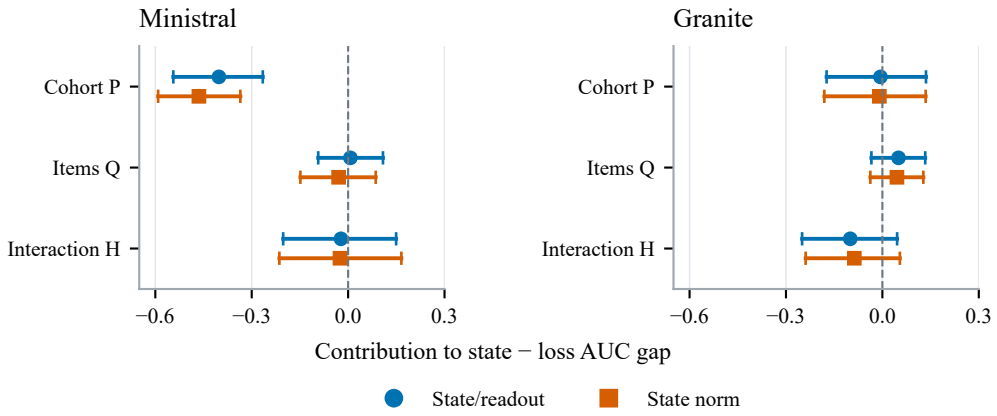

Figure 1: The crossed design locates the Ministral gap change in the prompt cohort contrast. Points and bars are P, Q, H estimates and intervals adjusted jointly with the confirmation tests for 24 comparisons. Each model compares 16 new pairs with three fixed old references on both item splits. State/readout and norm scores are unchanged. The cohort contrast includes training history and pair selection.

4.3 Predicting the group difference on fresh questions

For each new prompt pair j, the source panels determine

$$\widehat{\Delta G}_j = \tfrac{1}{2} \sum_{d \in \{O,N\}} (G_{j,d} - G_{O,d}), \qquad \Delta G_{j,F} = G_{j,F} - G_{O,F}. \tag{10}$$

Confirmation uses fresh items F with the same prompts. Pair forecasts, their cohort mean, and the analysis are fixed before generating outputs. The zero baseline predicts no group difference. The uniform baseline assigns every pair the source-mean forecast. Both are evaluated against fresh-item AUC gaps.

Ministral's fresh mean differences are -0.365 $[-0.550, -0.190]$ and -0.433 $[-0.611, -0.257]$, preserving the locked source directions. Prediction MAEs are 0.063 and 0.047, versus 0.365 and 0.433 for zero, and 0.123 and 0.103 for the uniform forecast. Improvements over zero are 0.302 $[0.014, 0.453]$ and 0.385 $[0.082, 0.516]$. Improvements over the uniform forecast are 0.060 $[-0.055, 0.139]$ and 0.056 $[-0.054, 0.123]$. These results support a group adjustment on fresh questions for the same Ministral prompts. The uniform-baseline comparison asks whether pair-specific adjustments add predictive value. Its intervals allow both improvement and deterioration. It therefore leaves the value of pair-specific forecasting unresolved.

Granite prediction MAEs are 0.085 and 0.085, and its mean-difference and both error-improvement intervals span zero. All four model/score families and every pair appear in Appendix E.1 and Figure 2.

Uncertainty and output conditions. The 24-comparison family jointly covers crossed contributions and fresh-question prediction contrasts. All 16 new pairs, three old references, and 20,000 shared crossed draws are valid (Owen, 2007) (Appendices E and E.1).

4.4 Forecasting unseen prompt pairs on GSM8K

GSM8K (Cobbe et al., 2021) tests unseen Ministral training runs and longer solutions. Its 80 prompts comprise 16 calibration, 32 forecast-development, and 32 confirmation prompts. Each latter cohort has 16 disjoint pairs, half adjacent-seed and half selection-loss pairs. Rank-32 state/readout diagnostics retain both gradient-dependent terms. Only the norm control zeroes them. Both use teacher-forced reference-solution states, and the readout is negative mean solution loss. Behavioral evaluation instead uses question-only greedy generation, at most 512 tokens, and strict final-number scoring.

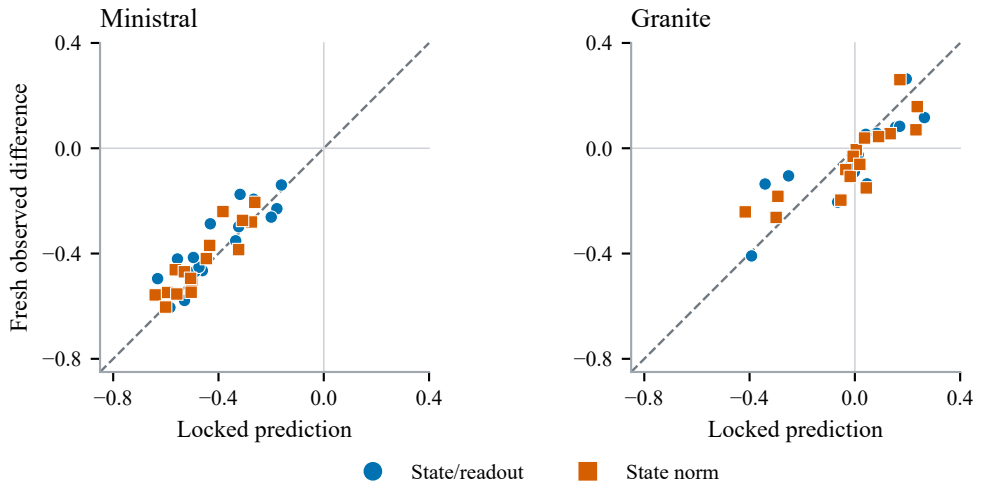

Figure 2: Predicted and observed reference-group differences on fresh questions for the same 16 prompt pairs, with both state-score families. The diagonal shows equality. Adjusted MAE-improvement intervals over the uniform source-mean forecast include zero for both families in both models (Table 17).

For each pair, the forecast target is the signed AUC gap $G_j = \mathrm{AUC}(s_{\mathrm{state}}, z_j) - \mathrm{AUC}(s_{\mathrm{loss}}, z_j)$. The item-level diagnostic predicts z. The separate pair-level forecaster predicts G_j. Cheap forecasting features are the absolute selection-loss difference, prompt RMS displacement, and pairing type. Full ridge forecasts add four control-response summaries from 32 fixed calibration probes: response RMS divided by prompt-displacement RMS, signed cosine alignment between g_s and ΔH, the fixed-subspace share of response energy, and the relative linear residual

$$\frac{|\Delta r - \langle g_s, \Delta H \rangle_F|}{\max(|\Delta r|, 10^{-8})}, \qquad \Delta r = L_i(u_s) - L_i(u_t).$$

Here H contains all teacher-forced reference-answer positions, $\Delta H = H_t - H_s$, and $g_s = \nabla_{H_s}[-L_i(u_s)]$. The four probe features are averaged equally across the 32 items and use no generated-behavior features.

Table 2: GSM8K MAE for unseen prompt pairs (lower is better). Full uses control features. Cheap omits them. Mean is the fixed development mean.

State diagnostic	Full	Cheap	Mean	Zero
State/readout	0.097	0.105	0.095	0.087
State norm	0.098	0.100	0.096	0.091

All six MAE-improvement intervals in the retained 22-comparison family include zero. Mean predicted gaps are $-0.056/-0.076$, versus observed means $-0.004/-0.014$. Appendix I retains all predictions, intervals, and 256-token results.

4.5 Separating readout labels from diagnostic calibration

On 400 fresh ECQA questions, 75 existing prompts yield 30,000 endpoints. At fixed states, reference and entropy readouts replace the gold-margin gradient using the unprompted model's fixed A–E choice or conditional A–E entropy. They retain supervised calibration but require no gold answer at measurement. Evaluation still uses gold-defined risk labels.

Table 3 compares replacement-minus-gold AUC changes between 16 new and three old pairs. All four intervals include zero under the retained 22-comparison correction. All pairs and 20,000 draws are valid. Appendix K gives exclusion, calibration, and state-context details.

Table 3: Prompt-group interaction of each readout replacement, relative to the gold readout. Positive values favor the replacement more in the new group. Intervals retain the 22-comparison correction.

Model	Readout	Interaction	Lower	Upper
Ministral	Reference	+0.005	-0.082	+0.090
Ministral	Entropy	-0.059	-0.180	+0.043
Granite	Reference	+0.016	-0.012	+0.051
Granite	Entropy	+0.008	-0.029	+0.049

5 Does diagnostic ranking improve evaluation efficiency?

A diagnostic must concentrate enough risk changes to pay for the checks lost while obtaining its scores. Under budget B, let $K_s(B)$ count checks that complete generation and parsing at both prompt endpoints, and let $Y_s(B)$ count discovered risk changes. Define $p_s(B) = Y_s(B)/K_s(B)$ when $K_s(B) > 0$. Subscript 0 denotes direct inspection in fixed hash order. Suppressing B when both hit rates are defined, the realized discovery difference satisfies the standard accounting identity

$$Y_s - Y_0 = \underbrace{K_s(p_s - p_0)}_{\text{hit-rate gain}} - \underbrace{(K_0 - K_s)p_0}_{\text{coverage loss}}. \tag{11}$$

For zero checks, we compare raw discovery counts. Cost replay evaluates the three original diagnostics and selector on 54 existing pairs and 400 common ECQA questions under strict eight-token scoring. All measurement and checking costs are charged (Bischl et al., 2016) (Appendix J).

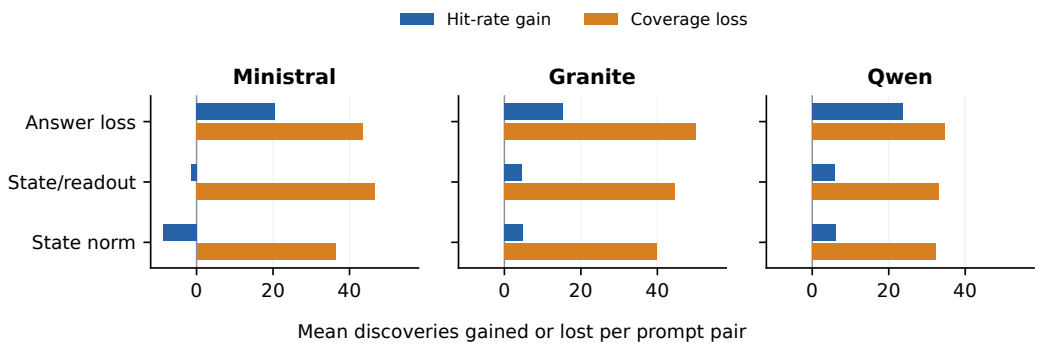

Figure 3: Measured hit-rate gains and coverage losses, averaged across pairs using Eq. 11. Gains must exceed losses for a diagnostic to pay off. Every displayed mean has larger coverage loss, even before selector probes. Appendix J.1 gives the condition matrix and bounded Granite timing study.

At the reference budgets (121.29, 83.67, and 377.01 seconds per pair), loss finds 37.16, 32.95, and 32.25 changes versus direct inspection's 60.21, 67.68, and 43.12. In Ministral, loss's higher hit rate adds 20.51 discoveries to its smaller queue, but reduced coverage costs 43.56. Neither state score beats direct inspection in any of the 54 pairs. Free per-pair choice of the best method among all diagnostics and direct inspection gains at most 1.55% in model mean across seven budgets. The charged selector finds 27.26, 28.42, and 25.25 changes and makes constant choices within each model. On three existing Granite pairs, strict-parser-preserving early stopping reduces measured eight-token generation and parsing time by 31.2% (Appendix J.1).

Batching changes relative scoring costs and rankings. 51/1,152 Granite cross-batch endpoint comparisons differ in tokens. Appendix K.3 consolidates primitive times and cost ratios, separately from full-strategy budget replay.

Monitoring methods trade predictive information against data and compute (Tillman and Mossing, 2025). Compare ranking gains with lost checks, charging probes even after fallback. Our short-answer replay favors direct checks. Appendix J.1 gives cost conditions and the runnable anonymous case.

6 INTERPRETATION, RELATED WORK, AND SCOPE

Control-response measurements organize comparisons. Behavioral transfer and evaluation value remain empirical.

6.1 FROM A MEASURED RESPONSE TO A PROMPT UPDATE

Prompted neurofeedback can support monitoring and control of activation directions (Li et al., 2025). Here the selected state objective backpropagates the projected readout gradient $K^\top UU^\top g$ to update a soft prompt. We compare it with randomly selected auxiliary objective families and a fixed route objective, using eight streams per model, shared task-loss gradients, equal update counts, and matched final prompt displacements.

For the initial and updated prompts, total risk reduction decomposes as

$$\mathbb{E}[b_0 - b_1] = \mathbb{E}[(b_0 - b_1)\mathbf{1}\{p_0 \neq p_1\}] + \mathbb{E}[(b_0 - b_1)\mathbf{1}\{p_0 = p_1\}]. \tag{12}$$

Ministral's selected update reduces risk by 39.00 percentage points: 37.81 points come from changed parseability and 1.19 from the remaining questions. Granite gains 3.88 points, comprising -0.25 and $+4.13$ points. The gain over random objective selection is uncertain: 1.13 $[-5.75, 8.00]$ points for Ministral and 1.31 $[-1.25, 4.88]$ for Granite. Thus the response-guided objective produces measurable recovery. Appendix G.1 reports all arms and intervals.

6.2 RELATION TO PRIOR WORK

State monitors assess truthfulness (Azaria and Mitchell, 2023), response consistency (Chen et al., 2024), and confidence from multiple layers (Beigi et al., 2024). IRIS uses verification-context activations and uncertainty pseudolabels instead of gold calibration (Srey et al., 2025). Other probes predict format failures before generation (Ashok and May, 2025) or link toxicity to representations across models and prompts (Waldis et al., 2026).

Format sensitivity (Sclar et al., 2024; Lu et al., 2022) includes the separation of format following and answer knowledge through soft-prompt tuning (Erkan et al., 2026). We test transfer of a fixed diagnostic's AUC advantage over loss using crossed panels with zero and common-mean forecast baselines. Appendix K.3 contrasts diagnostic targets and measurement requirements.

6.3 SCOPE AND PRACTICAL ASSESSMENT

The findings concern supervised, reference-answer evaluation with greedy generation. MATH has 107/114 changes between parseable outputs but loss/primary AUCs of 0.524/0.494 (Appendix H). Long-answer confirmation uses one model, and costs describe the measured host and pipelines. These conditions require task-specific reassessment.

7 CONCLUSION

We propose a control-response evaluation framework linking internal measurements, output controls, transfer tests, and costs. In two Ministral–ECQA pairs, removing continuation reverses state-versus-loss rankings under strict parsing without changing diagnostic scores or initial answers. Crossed Ministral–ECQA comparisons show larger prompt-group than question-set differences. The group contrast includes pair selection. Fixed forecasts reproduce this contrast on fresh questions for the same prompts. Gains over a common mean and on unseen GSM8K prompts remain uncertain. At the measured ECQA reference budgets, direct checks find more changes on average than screening with the three original diagnostics.

For researchers, the framework separates detecting behavioral changes from predicting diagnostic advantages across groups. For practitioners, it specifies three checks: define the target behavior, validate comparisons on intended prompts and output rules, and weigh ranking gains against checks lost to measurement.

Reproducibility Statement

The supplied evaluation artifact includes the manuscript, figures, numerical tables, saved item-level scores and risk labels, fixed forecasts, and runnable CPU statistical replay. The entry `bootstrap_replay/run.py`, relative to the artifact root, recomputes 20,000 shared resamples for the three-model ECQA validation, crossed comparisons, fresh-question confirmation, GSM8K 512-token forecasts, and ECQA readout interactions. It checks 52 primary intervals against the retained results, preserving the original seeds, shared-draw structure, coverage rules, and 18/24/22 comparison families. Appendix L maps claims to files. Other reported intervals are supplied results unless explicitly covered by this entry.

The crossed and independent-training studies contain 60,000 and 38,400 endpoints, respectively, with 25,600 reused records: their union is 72,800, with Qwen calibration additional. Cost measurements reuse existing prompts and questions and contribute no independent training runs or test panels. GSM8K adds 80 prompts and 17,600 main native endpoint trajectories, with timing and state-only probes counted separately. Historical raw-data replay checks are recorded in `data/gsm8k/independent_replay.json`. The supplied GSM8K case reproduces all 16 pairs from saved scores and forecasts. The readout supplement contributes 30,000 prompted endpoints and 800 unprompted references on 400 further ECQA questions. Historical raw-data replay checked fixed scores, parsing, statistics and 3,456 batching endpoint executions. The supplied ECQA case recomputes pairwise AUCs and batching costs from saved records. Both successful and failed-attempt costs are retained. The cost records retain control probes, operation costs, queues, folds, the development decision, and the balanced Granite timing study. Appendices A, D and H specify model, parser and numerical conditions. Raw hidden-state tensors, token trajectories, checkpoints, model weights, and full training datasets are not included in this ZIP. The separately retained raw archives and their historical receipts must not be confused with the supplied statistical replay. New checksums verify the anonymous artifact's integrity, not the historical timing of prediction locks.

Ethics Statement

This study uses public benchmarks and open-weight models with deterministic answer scoring. It contains no human-subject experiment or private user data. The analysis clarifies the behavioral targets and conditions of diagnostic comparisons.

AI Use Statement

Generative-AI tools assisted research design, implementation and testing, record inspection, literature search, figures, and translation. Dataset labels and deterministic parsers supply the scores. The authors are responsible for verifying the experiments, analyses, citations, and final text.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.238. URL https://aclanthology.org/2021.acl-long.238/.

Dhananjay Ashok and Jonathan May. Language models can predict their own behavior. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2102. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5a9c1af5f76da0bd37903b6f23e96c74-Abstract-Conference.html.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. InternalInspector I^2: Robust confidence estimation in LLMs through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12847–12865, 2024. doi: 10.18653/v1/2024.findings-emnlp.751. URL https://aclanthology.org/2024.findings-emnlp.751/.

Bernd Bischl, Pascal Kerschke, Lars Kotthoff, Marius Lindauer, Yuri Malitsky, Alexandre Fréchette, Holger Hoos, Frank Hutter, Kevin Leyton-Brown, Kevin Tierney, and Joaquin Vanschoren. ASlib: A benchmark library for algorithm selection. *Artificial Intelligence*, 237:41–58, 2016. doi: 10.1016/j.artint.2016.04.003. URL https://arxiv.org/abs/1506.02465.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

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. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Selen Erkan, Bastian Boll, Kristian Kersting, Björn Deiseroth, and Letitia Parcalabescu. Soft-prompt tuning for fair and efficient LLM benchmark evaluation, 2026. URL https://arxiv.org/abs/2606.12117.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2301.04213.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 34, 2021. URL https://arxiv.org/abs/2103.03874.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection, 2025. URL https://arxiv.org/abs/2509.11536v2.

Pyuyi Chufeng Huang and Zikang Song. Horizon-uniform sensitivity and decay of terminal reward perturbations in discrete-time pontryagin systems, 2026. URL https://arxiv.org/abs/2606.17762v3. Version 3.

IBM Granite Team. Granite 4.1 8b model card. Hugging Face model repository, 2026. URL https://huggingface.co/ibm-granite/granite-4.1-8b. Retrieved 2026-08-26.

Bum Jun Kim, Shohei Taniguchi, Makoto Kawano, Yusuke Iwasawa, and Yutaka Matsuo. Residual Koopman spectral profiling for predicting and preventing transformer training instability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 3036–3060. PMLR, 2026. URL https://proceedings.mlr.press/v337/kim26d.html.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Ji-An Li, Huadong Xiong, Robert Wilson, Marcelo G. Mattar, and Marcus K. Benna. Language models are capable of metacognitive monitoring and control of their internal activations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2009. URL https://papers.neurips.cc/paper_files/paper/2025/hash/56a225639da77e8f7c0409f6d5ba996b-Abstract-Conference.html.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Mistral AI. Ministral 3 8b instruct 2512 bf16 model card. Hugging Face model repository, 2025. URL https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512-BF16. Retrieved 2026-08-26.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122. URL https://arxiv.org/abs/0712.1111.

Qwen Team. Qwen3.5-9B model card. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Non-thinking mode setting; retrieved 2026-09-10.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. Unsupervised hallucination detection by inspecting reasoning processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22117–22129, 2025. doi: 10.18653/v1/2025.emnlp-main.1124. URL https://aclanthology.org/2025.emnlp-main.1124/.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Henk Tillman and Dan Mossing. Investigating task-specific prompts and sparse autoencoders for activation monitoring, 2025. URL https://arxiv.org/abs/2504.20271.

Andreas Waldis, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych. Aligned probing: Relating toxic behavior and model internals. *Transactions of the Association for Computational Linguistics*, 14:271–291, 2026. doi: 10.1162/tacl.a.613. URL https://aclanthology.org/2026.tacl-1.14/.

Yujing Wang, Yuanbang Liang, Yukun Lai, Hainan Zhang, and Hanqi Yan. GRADE: Probing knowledge gaps in LLMs through gradient subspace dynamics, 2026. URL https://arxiv.org/abs/2604.02830v1.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors, 2024. URL https://arxiv.org/abs/2406.01026v2.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

The appendix retains the fixed protocols, all adjusted primary comparisons, pairwise confirmation results, and the assumptions needed to interpret them. The one-page index in Appendix L links these claims to the complete data supplement.

A PROMPT TRAINING, PAIR SELECTION, AND FIXED PREDICTORS

Table 4: Training and evaluation settings used for the ECQA cases.

Component	Setting
Soft prompt	Eight vectors initialized from distinct non-special-token embeddings
Optimization	AdamW, 100 updates, learning rate 0.1, weight decay 0, effective batch 8, gradient clipping norm 1
Checkpoints	Every ten updates, including initialization, select by 64-item loss
Training runs	Seeds 2101–2116 for each model
Item roles	128 optimization, 64 checkpoint selection, 200 calibration, 100 development, 200 discovery holdout
Native generation	Greedy, up to eight new tokens, model-bundled tokenizer and chat template
Pair criterion	Mean loss difference has a 90% paired interval inside $\pm 0.10 s_d$ on the 300 calibration/development items
Ridge selection	Five item-grouped folds, penalties $10^{-4}, 10^{-3}, \ldots, 10^{4}$, strongest penalty within one standard error of best validation MAE
Independent replication	200 additional items, same model, three pairs, and fitted predictors

Here s_d in the pair criterion is the sample standard deviation of the 300 itemwise loss differences. Granite has three loss-equivalent pairs among sixteen runs. Ministral has four. The selected discovery pairs are the first three eligible pairs in seed order. Non-finite losses or degenerate paired variance make a pair ineligible. The pair comparisons in Table 9 include every selected discovery pair, without selection by endpoint-channel results.

Model and numerical conditions. The models are Ministral-3-8B-Instruct-2512-BF16, Granite-4.1-8B and Qwen3.5-9B, using their bundled tokenizers and chat templates. Backbone weights use BF16 without quantization. Prompts and optimizer states use FP32. New measurements use an NVIDIA A800 80GB PCIe GPU, PyTorch 2.6.0 and Transformers 5.15.1. Snapshot hashes and complete environment records are supplied with the data. All costs refer to this observed host and pipeline, with resource waiting recorded separately.

State features and ridge fitting. The calibration operator is the uncentered second moment $M = N^{-1} \sum_j \Delta h_j \Delta h_j^\top$ of perturbation states in the final hidden space. With the response columns collected in R, this is $M = RR^\top/N$, and U consists of its leading eigenvectors. Calibration uses 32 random prompt draws. The candidate ranks are $1, 2, 4, 8, 16, 32$. Five-fold assessment and a smallest-rank one-standard-error rule select rank 32 in both ECQA cases. Mean held-out residual energy fractions at rank 32 are 0.5545 for Granite and 0.8264 for Ministral. These measure reconstruction of calibration responses. High residual energy indicates variation outside the selected subspace. The primary diagnostic still includes the gradient-weighted complementary response. The estimated basis is tied to its model, positions, and calibration distribution.

For one scored ECQA position, the primary six-feature vector is

$$x = \big(g_s^\top U U^\top \Delta h, \quad g_s^\top (I - UU^\top)\Delta h, \quad \|h_s\|_2, \quad \|h_t\|_2, \quad \|u_t - u_s\|_F, \quad e_{\mathrm{tail}}(r)\big). \tag{13}$$

Here $e_{\mathrm{tail}}(r) = \mathrm{tr}(M) - \mathrm{tr}(U^\top M U)$. The first two terms are left-endpoint readout approximations, rather than path integrals. For several scored positions, these terms sum positionwise and the state norms are divided by the square root of the number of positions. The predictor standardizes features using training means and population standard deviations, replacing a zero scale with one. It fits ridge regression with an unpenalized intercept equal to mean training risk-change rate. Predictions are not clipped before ranking. Five-fold selection groups each item across all prompt pairs, with fold index equal to its analysis ordinal modulo five. The final fit uses all 300 original calibration and

development items. The independent replication reuses that fit without updating its coefficients or normalizers.

The ECQA norm control is fitted on exactly three features, $\|h_s\|_2$, $\|h_t\|_2$, and $\|u_t - u_s\|_F$, using the same supervised fitting procedure. It excludes the gradient-dependent terms and the fixed calibration tail energy. The GSM8K norm control instead sets the first two coordinates of the six-feature vector to zero before both fitting and scoring. Its primary state/readout diagnostic retains both gradient-dependent coordinates. Both GSM8K variants use teacher-forced reference-solution states (Appendix I). The matched-random control replaces U with each of 32 rank-matched random bases, refits for each basis, and takes the median prediction per record. These controls show how much ranking information is already available from generic state magnitude and subspace choices.

Fitting target and constant features In a subsequent sensitivity analysis, we selected regularization by equal-pair AUC within each calibration fold and refitted after removing prompt displacement and fixed tail energy. These changes did not consistently improve independent-pair ranking across the four panels. Subtracting the two constant contributions with coefficients held fixed left every defined within-pair AUC unchanged. Refitting the reduced model can change the remaining coefficients.

Table 5: Rank-32 feature and regularization-selection sensitivity. Entries are equal-pair mean AUCs. Reduced removes prompt displacement and fixed tail energy. Confirmation forecasts and original intervals are unchanged.

Panel	Full MAE	Full AUC	Reduced MAE	Reduced AUC
ECQA / Ministral	0.460	0.447	0.443	0.449
ECQA / Granite	0.549	0.547	0.548	0.547
ECQA / Qwen	0.622	0.622	0.622	0.622
GSM8K / Ministral	0.484	0.484	0.485	0.485

Ordinary activation probe. For each ECQA model, the probe input is the ordered concatenation $v_{i,st} = [h_{is}; h_{it}] \in \mathbb{R}^{8192}$ of two 4096-dimensional endpoint states. It predicts the same binary risk-change target using standardized ridge regression with an unpenalized intercept and squared error. It uses neither a response basis nor readout gradients. The 300 calibration/development questions give 900 rows for each of Ministral and Granite (three original pairs) and 2,400 for Qwen (eight calibration pairs). A question and all its pairs stay in one fold, assigned by its ordinal modulo five. Means and population standard deviations use the training fold only, with zero scales replaced by one. For each $\alpha \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$, we average AUC equally over pairs within each fold, then over the five folds. The largest α within one standard error of the best mean is selected. The standard error is the sample standard deviation of fold means divided by $\sqrt{5}$. All three models select $\alpha = 10^4$. A final fit uses all calibration rows and evaluates 16 independent pairs on 400 new questions (6,400 rows per model). Larger scores predict $z = 1$. Test labels do not select parameters or score direction.

Rank and logit sensitivity. Ranks 8, 16, 32, and 64 form the sensitivity analysis. Each fit uses only original calibration rows and the original MAE rule. Table 6 shows changes in absolute AUC across ranks. The original rank-32 scores and confirmation forecasts remain unchanged.

Table 6: State/readout AUC by response rank, with the unchanged loss baseline. Values are equal-pair means over 16 valid pairs.

Panel	Rank 8	Rank 16	Rank 32	Rank 64	Loss
ECQA / Ministral	0.371	0.493	0.460	0.407	0.834
ECQA / Granite	0.542	0.567	0.549	0.549	0.731
ECQA / Qwen	0.629	0.624	0.622	0.623	0.895
GSM8K / Ministral	0.485	0.487	0.484	0.478	0.488

For five-choice logits ℓ_s, ℓ_t, let p_s, p_t be their conditional softmax distributions. The gold-margin score is $|r_i(h_t) - r_i(h_s)|$ with r_i defined by the correct choice's log odds against all four alternatives

(Eq. (5)). The remaining scores are absolute changes in maximum probability, the largest-minus-second-largest logit gap, and entropy, plus $\mathrm{JS}(p_s, p_t) = [\mathrm{KL}(p_s\|m) + \mathrm{KL}(p_t\|m)]/2$, where $m = (p_s + p_t)/2$. The comparison uses saved logits. Only gold margin uses the gold answer to compute a score. All AUC evaluations still use gold-defined risk labels. Entropy change is distinct from the entropy-gradient readout replacement. Table 7 retains all five comparisons.

Table 7: All saved-logit baselines on the independent ECQA panels. Equal-pair AUCs. All 16 pairs per model are valid.

Model	Gold margin	Max. prob.	Top-two gap	Entropy	JS
Ministral	0.690	0.661	0.430	0.553	0.850
Granite	0.756	0.685	0.582	0.644	0.761
Qwen	0.658	0.723	0.420	0.602	0.885

A.1 Selected checkpoints and prompt displacements

Selection-set losses refer to the chosen checkpoint on 64 questions. The original pair-screening statistic instead used 300 calibration and development questions. The two quantities need not agree. The distance $\|u_t - u_s\|_F$ compares selected prompt tensors, not the length of an optimization path. Table 8 summarizes the new training runs. Full selected steps, losses, and distances for all old and new pairs remain in the data supplement. The old Ministral pairs share seed 2102. All new pairs have distinct endpoints. No pair was reassigned using these summaries.

Table 8: Selected checkpoints and attained controls for the independent validation prompts. Ranges cover all runs or pairs, as appropriate. A step-zero selection follows a completed 100-update run. It is not a failed or omitted training run. Full values are in training_pairs.csv.

Model	Prompts	Step zero	Selection loss	$\lvert\Delta L_{\mathrm{sel}}\rvert$	$\|\Delta u\|_F$
Ministral	32	3	[0.795, 1.430]	[0.001, 0.599]	[98.45, 164.88]
Granite	32	0	[0.687, 1.621]	[0.010, 0.927]	[126.86, 204.70]
Qwen	32	4	[0.574, 0.884]	[0.010, 0.193]	[69.96, 169.35]

B Discovery and fixed-pair replication

Table 9: Every selected discovery prompt pair. Each row has 200 items. "Parse" counts risk changes involving parseability, a subset of the preceding flip count. The last three columns are AUCs.

Case	Pair	Flips	Parse	State/readout	Norm	Loss
ECQA Granite	2101–2114	30	0	0.645	0.649	0.921
	2108–2112	23	2	0.633	0.643	0.912
	2111–2113	34	7	0.613	0.605	0.737
ECQA Ministral	2102–2111	107	96	0.602	0.642	0.324
	2102–2112	101	90	0.626	0.614	0.350
	2103–2105	38	1	0.402	0.350	0.896
ECQA Ministral, new	2102–2111	121	111	0.530	0.596	0.286
	2102–2112	107	96	0.648	0.632	0.295
	2103–2105	35	2	0.440	0.340	0.889

Replication construction. Eligible ECQA items satisfy the tokenizer intersection and deterministic canonical-answer checks used for the original study. A stratified source order uses seed 20260828. Starting beyond the original 500-item analysis pool, the selection takes the next 200 eligible items after excluding every previously used coordinate and normalized question/options text. The exclusion

Table 10: Independent-item contrasts and sensitivity. Simultaneous intervals account for the six primary comparisons. All leave-one-item-out ranges preserve the expected sign.

Pair	Score minus loss	ΔAUC	Simultaneous interval	Leave-one-item-out range
2102–2111	State/readout	0.244	[0.081, 0.406]	[0.235, 0.258]
2102–2111	State norm	0.310	[0.136, 0.479]	[0.302, 0.327]
2102–2112	State/readout	0.354	[0.179, 0.517]	[0.347, 0.367]
2102–2112	State norm	0.337	[0.160, 0.504]	[0.330, 0.351]
2103–2105	State/readout	-0.449	[-0.629, -0.260]	[-0.470, -0.433]
2103–2105	State norm	-0.549	[-0.688, -0.386]	[-0.577, -0.538]

set contains 692 original items and 200 intervention-holdout items. Text normalization applies Unicode NFKC and collapses whitespace after removing the common answer-format instruction. No new model outputs enter item selection, score fitting, or the expected sign specification.

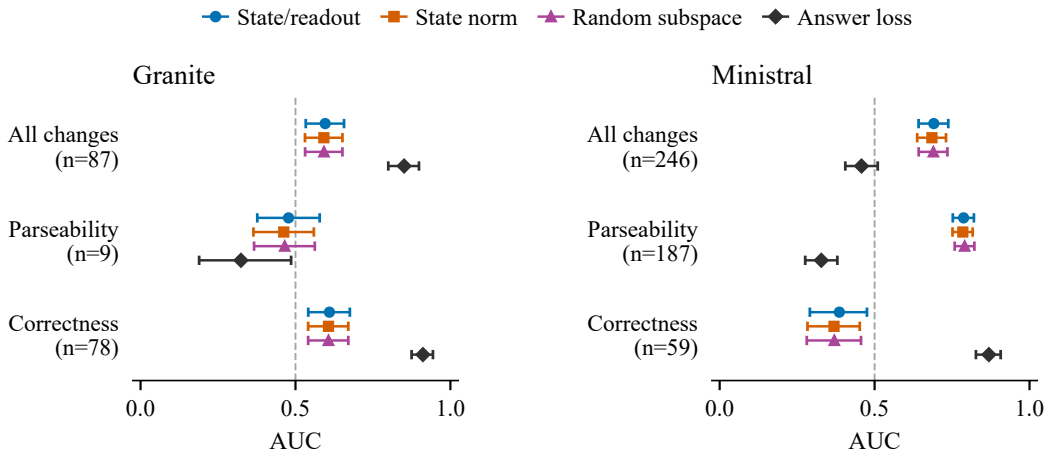

Figure 4: Endpoint-conditioned diagnostic ranking on the discovery data. Each channel uses the same unchanged-risk negative pool within its model. Points are AUCs and bars are item-clustered 95% intervals. Counts identify the positive examples available for each comparison. Granite's parseability channel has only nine. The state/readout and norm scores are close, while their contrasts with answer-loss differences depend on the endpoint channel.

C Output-interface intervention

This secondary probe was specified after the independent-item study completed and before any constrained-output measurement. It reuses all 200 replication items and all five prompt endpoints: 2102, 2103, 2105, 2111, and 2112. The source model, prompts, item manifests, and fitted predictors remain identical. Five original-item reference replays again produce zero maximum state and loss error and identical behavior.

For each item, supervised encoding of the five labels A–E identifies five distinct single-token candidates. At the native BF16 output head, the intervention preserves their logits and sets every other logit to the minimum finite value of the same dtype. Greedy generation then emits exactly one token. Candidate selection is symmetric in the five labels. The gold answer is used only to score correctness. Each measurement stores all five native candidate logits, their token IDs, the chosen token, and the resulting output. All 1,000 outputs are strictly parseable. The unmasked vocabulary argmax is already legal in every measurement. Its token equals the constrained argmax in all 1,000 cases, and the controlled letter matches the leading letter of the original native output. The mask therefore does not change the first answer choice on these items. Stopping after that choice removes the continuation. Of 840 originally parseable endpoints, 583 remain correct and 257 incorrect. The remaining 160 become parseable, with 124 correct and 36 incorrect initial letters.

The loss, state/readout, and state-norm scores are copied unchanged from the native study. New risk-change labels are computed from the constrained outputs. For the two parse-dominated pairs and the two state-score families, the measured contrast is

$$D_f = \left(\mathrm{AUC}_f - \mathrm{AUC}_{\mathrm{loss}}\right)_{\mathrm{constrained}} - \left(\mathrm{AUC}_f - \mathrm{AUC}_{\mathrm{loss}}\right)_{\mathrm{native}}. \tag{14}$$

The same item bootstrap multiplicities apply to both conditions, all pairs, and all scores. We use 10,000 draws with seed 20260907 and percentile endpoints $0.05/8$ and $1 - 0.05/8$ for the four specified contrasts. All draws have a defined AUC. The third pair's changes and all marginal AUC intervals are descriptive. This analysis has a separate four-comparison adjustment from the six-comparison replication family.

Table 11: All prompt pairs under native and constrained generation. Each condition contains the same 200 items and uses unchanged diagnostic scores.

Pair	Condition	Flips	State/readout AUC	Norm AUC	Loss AUC
2102–2111	Native	121	0.530	0.596	0.286
2102–2111	Constrained	25	0.451	0.371	0.879
2102–2112	Native	107	0.648	0.632	0.295
2102–2112	Constrained	33	0.329	0.310	0.821
2103–2105	Native	35	0.440	0.340	0.889
2103–2105	Constrained	33	0.423	0.330	0.915

Table 12: The four specified changes in state-minus-loss AUC gap. Intervals have nominal family confidence 95% across these four contrasts.

Pair	Score	Change in ΔAUC	Simultaneous interval
2102–2111	State/readout	-0.672	[-0.893, -0.457]
2102–2111	State norm	-0.818	[-1.018, -0.616]
2102–2112	State/readout	-0.845	[-1.069, -0.616]
2102–2112	State norm	-0.848	[-1.060, -0.629]

The intervention changes continuation, parseability, and the positive/negative sets used for ranking while preserving initial answer choice in this sample. It identifies an effect of the output rule under this parser. It does not identify which hidden-state features causally produce the continuation, or an effect for arbitrary free-form answer evaluators.

D Independent training: design, verification, and all pairs

Independent training and held-out items. We train 32 new validation soft prompts for each of Ministral-3-8B, Granite-4.1-8B, and Qwen3.5-9B, using the same eight-vector prompt, fixed backbone, optimizer, and 128 optimization and 64 checkpoint-selection items as the discovery study. The seeds are 7201–7232, 7101–7132, and 7301–7332, respectively. Adjacent seeds form 16 pairs per model, with each prompt appearing in exactly one pair. The pairing is specified before training, and all pairs enter evaluation. Similarity of selected losses is a descriptive property of these pairs.

All three models use the same 400 new ECQA items. Selection follows the original deterministic source ordering and excludes all 1,092 previously used items by source coordinate and normalized question/options hash. The original calibration-derived subspaces, feature normalizers, ridge coefficients, and score directions remain fixed for Ministral and Granite. Qwen fits its diagnostic using 16 separate prompts (6901–6916, eight adjacent pairs) and the original 200 calibration/100 development items. Its 32 perturbation draws use scale 0.0025, 168 core calibration items, and candidate ranks 1, 2, 4, 8, 16, and 32. 32 reserved calibration items remain outside the subspace core. Item-grouped five-fold rank selection chooses rank 32. Ridge regularization follows the same original fitting procedure. The Qwen template sets `enable_thinking=False` for direct answers. Its 32 validation prompts and 400 new items enter neither subspace nor ridge fitting. The 112 new training runs comprise 96 validation prompts and 16 calibration prompts. The new prompts and items therefore test transfer of the existing diagnostic functions across training runs and items jointly.

Output conditions. One native full-vocabulary greedy sequence of at most 32 tokens is retained for every prompt–item endpoint. Its 1, 2, 4, 8, 16, and 32 token prefixes define six budgets, preserving the original end-of-sequence rule. Three deterministic parsers score every prefix: the strict single-letter parser, an initial-letter parser requiring a clear boundary, and a parser that prioritizes a terminal explicit answer field and otherwise uses the initial-letter rule. Conflicting or malformed explicit fields are unparseable. The diagnostic scores are identical across all 18 conditions. Appendix D specifies the exact boundary and field rules.

The main stopping comparison uses the native first token, without an A–E mask. Its interpretation as preserving an initial answer letter is assessed from the token's actual legality. A descriptive control uses the fixed pair–item subset on which both endpoint first tokens are legal. We also reconstruct a five-choice control from saved logits, with a gold-blind argmax. These controls retain their distinct item and pair domains.

Paired estimands and uncertainty. For state-score family f and strict-parser generation cap L, let $G_{f,L} = \mathrm{AUC}(s_f) - \mathrm{AUC}(s_{\mathrm{loss}})$. The primary change is $D_f = G_{f,1} - G_{f,8}$. We compute each prompt-pair AUC with half credit for ties, then average equally over the original intersection of pairs with defined AUC at both budgets. A crossed bootstrap resamples the 400 item identities and 16 pair identities independently, sharing each draw across scores and output conditions (Owen, 2007). Item draws are also shared across models.

We use 20,000 draws and retain the 18-comparison Bonferroni family of the three-model evaluation. The interval quantiles are $0.05/36$ and $1 - 0.05/36$. A model-level comparison supports inference when at least 12 original pairs are jointly estimable and at least 95% of draws are valid for all three estimands. An included pair with undefined AUC makes that draw undefined. We retain these records and report their coverage. A supported shift requires the D_f interval to exclude zero. A supported reversal additionally requires the two G intervals to lie on opposite sides of zero. All pair estimates and leave-one-pair-out checks are reported. The full parser–budget matrix uses exploratory pointwise 95% intervals.

Exact parsing rules. Strict parsing removes every Unicode whitespace character and accepts exactly one uppercase ASCII A–E letter. Initial-letter parsing strips leading whitespace and accepts A–E only at the end of the text or before whitespace, ASCII punctuation other than underscore, or the listed common Chinese punctuation characters. Thus `A.` is accepted, whereas `Apple`, `A2`, and `B_B` are rejected. The machine-readable parser source supplies the exact Unicode delimiter set.

Terminal parsing recognizes English `Answer:` and `Final answer:` fields case-insensitively, and the corresponding Chinese answer field. Both colon widths are accepted. Each field value must be an uppercase A–E followed by at most one optional ASCII or Chinese full stop. Any malformed field or conflicting valid fields is unparseable. The final nonempty line must consist entirely of a valid field. Earlier fields may repeat the same answer. If there is no explicit field, the parser uses the initial-letter rule. A valid terminal field takes priority over an ordinary initial answer. None of these decisions consults the gold label. Empty outputs and early EOS retain their actual decoded prefixes.

The leading parser changes parseability between budgets one and two on 426 Granite endpoints: 26 from seed 7101 and 400 from seed 7123. Their first letters remain unchanged, but continuation joins them to additional letters, as in `A` becoming `AIN`. Ministral has no such transition. Qwen likewise has none. The supplement retains every changed coordinate and prefix. Thus initial-token legality is a token-level property. A longer prefix also needs the parser's boundary condition.

Controls and coverage. The main endpoint panel retains all native first tokens. Its 38,399 legal first tokens support the statement that taking their one-token prefix preserves the initial answer letter. The remaining Ministral endpoint supports no such interpretation. The jointly legal pair–item control retains 6,399 Ministral and 6,400 coordinates each for Granite and Qwen, using the same subset at both budgets. A separate five-choice control computes a gold-blind argmax over the saved A–E logits. Both are descriptive controls, with their own estimable-pair counts. The strict-8/strict-1 primary analysis retains all 6,400 pair–item coordinates per model.

Coverage and retained measurements. All 112 training runs complete the specified schedule. Three Ministral runs select initialization (7204, 7216, 7228). They remain in their original pairs.

Source-coordinate and normalized-text checks exclude all 1,092 previously used items. Saved states reconstruct both fixed scores, and tokens reconstruct every parser and prefix. Qwen's separate calibration and both ridge fits also replay. These checks reproduce retained measurements. They do not regenerate model outputs. All primary cells retain 16 pairs and 20,000 valid resamples.

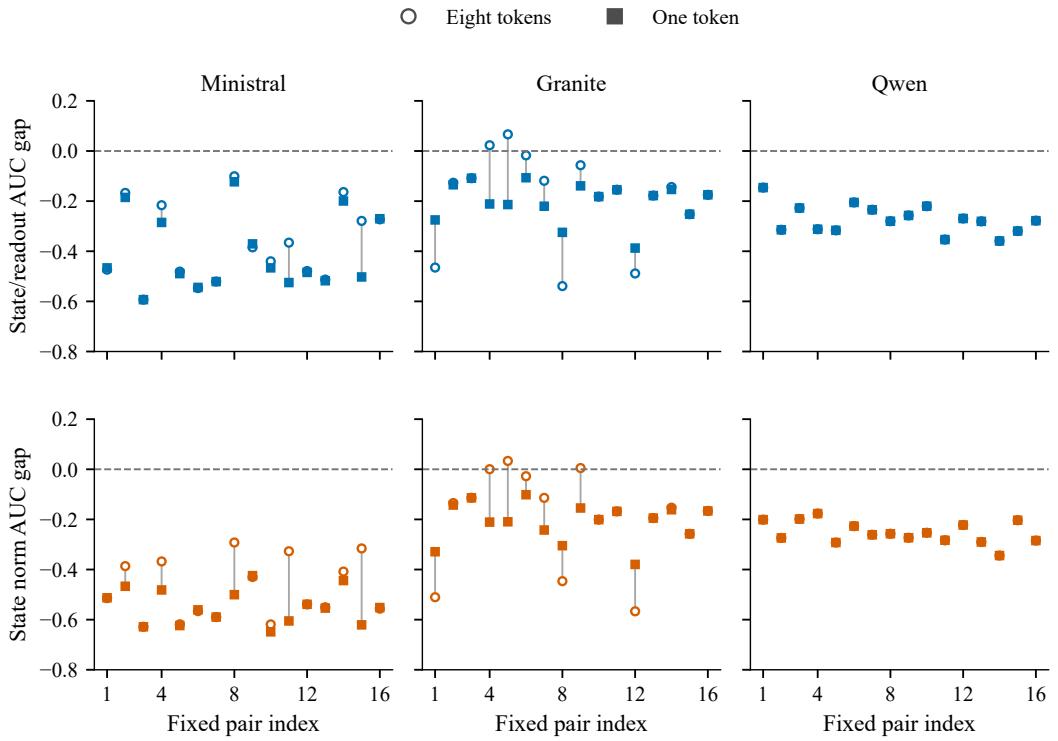

Figure 5: Every independent training pair at the two primary strict-parser budgets. Points are pair AUC gaps, without individual-pair significance claims. Lines join the same pair's two output conditions. Each model's 16 pairs contain 32 distinct prompts. The model means and adjusted intervals are in Table 13. Complete numerical estimates, positive counts, selected-loss differences, and leave-one-pair-out results are supplied in the accompanying CSV files.

Table 13: All 18 primary intervals. Every row uses 16 common pairs and 20,000 valid crossed-bootstrap draws. Intervals use one 18-comparison Bonferroni family.

Model	Score	G_8 [interval]	G_1 [interval]	D [interval]
Ministral	State/readout	-0.3748 $[-0.4944, -0.2467]$	-0.4089 $[-0.5252, -0.2865]$	-0.0341 $[-0.1020, 0.0018]$
Ministral	State norm	-0.4818 $[-0.5861, -0.3704]$	-0.5470 $[-0.6271, -0.4608]$	-0.0652 $[-0.1668, -0.0018]$
Granite	State/readout	-0.1823 $[-0.3296, -0.0568]$	-0.2010 $[-0.2817, -0.1328]$	-0.0187 $[-0.1271, 0.0842]$
Granite	State norm	-0.1885 $[-0.3423, -0.0626]$	-0.2086 $[-0.2881, -0.1409]$	-0.0201 $[-0.1260, 0.0799]$
Qwen	State/readout	-0.2733 $[-0.3575, -0.1861]$	-0.2733 $[-0.3575, -0.1861]$	0.0000 $[0.0000, 0.0000]$
Qwen	State norm	-0.2526 $[-0.3340, -0.1746]$	-0.2526 $[-0.3340, -0.1746]$	0.0000 $[0.0000, 0.0000]$

Table 14: Leave-one-pair-out ranges for the three-model primary estimates. These are descriptive sensitivity ranges. The complete 288 estimates remain in the CSV supplement.

Model	Score	G_8	G_1	D
Ministral	State/readout	[-0.393, -0.360]	[-0.428, -0.397]	[-0.037, -0.021]
Ministral	State norm	[-0.494, -0.472]	[-0.555, -0.540]	[-0.070, -0.049]
Granite	State/readout	[-0.199, -0.158]	[-0.207, -0.189]	[-0.034, -0.001]
Granite	State norm	[-0.203, -0.163]	[-0.216, -0.197]	[-0.034, -0.005]
Qwen	State/readout	[-0.282, -0.268]	[-0.282, -0.268]	[0.000, 0.000]
Qwen	State norm	[-0.258, -0.246]	[-0.258, -0.246]	[0.000, 0.000]

Secondary channel estimability. All overall parser–budget cells have 16 estimable pairs and 20,000 valid draws. Channel-specific estimates can use fewer pairs and valid draws. For example, the strict-8 parse channel has 14 estimable Ministral pairs and 61.985% valid draws, versus 10 Granite pairs and 90.49% valid draws. These are descriptive estimates. At strict-1, Granite has no parse-channel positive and Ministral has only one. Those estimates remain undefined or sparse in the supplied tables, rather than being filled with chance AUC. Pointwise secondary intervals are accompanied by all coverage counts and are not additional primary replication criteria.

D.1 Parser and generation-budget conditions

Strict parse failures fall from 285 to one in Ministral and from 2,569 to zero in Granite when replacing the eight-token output by its one-token prefix. Scores remain fixed. All strict risk labels agree at 8, 16 and 32 tokens in these independent panels. This equality does not imply equality for other parsers. Granite's terminal-parser failure counts are 426, 428 and 429 at those caps. All 12,800 Qwen trajectories end at two tokens and have identical risk and channel labels across all 18 conditions. Its parseability-channel AUC is undefined because that channel has no positives. The complete matrix preserves every estimate and coverage count in the data supplement. Observed-label channel reweighting is exploratory and is kept there separately from the locked predictions.

E Crossed prompt–question comparison and prediction

Panels, prompts, and output conditions. For each model, the old prompt cohort contains the three fixed references used in the discovery analysis, and the new prompt cohort contains sixteen adjacent seed pairs formed before training. The old references are Ministral 2102–2111, 2102–2112, and 2103–2105, and Granite 2101–2114, 2108–2112, and 2111–2113. The new pairs are Ministral 7201–7202 through 7231–7232 and Granite 7101–7102 through 7131–7132. The prompt-group contrast therefore includes both training-run differences and the historical selection rule for the old references. It is not interpreted as an isolated effect of training randomness.

The item splits are old source items (200), new source items (400), and a fresh confirmation split (200). Fresh items are selected deterministically from the same ECQA snapshot after excluding $1{,}492$ previously used items by item identity, source coordinate, and normalized question/options hash. Token lengths are checked against both model tokenizers. Every endpoint is generated by native full-vocabulary greedy decoding for at most 32 tokens in BF16, with no answer mask. The eighteen output conditions are the Cartesian product of three parsers, strict, leading-letter, and terminal-field, with budgets $1, 2, 4, 8, 16, 32$. The primary condition is strict parsing at budget eight.

Scores and labels. For a model m, score family $f \in \{\mathrm{primary}, \mathrm{norm}\}$, prompt pair j, item i, and output condition c, let $Y_{cij} = |b^R_{cij} - b^L_{cij}|$ be the risk-change label between the right and left prompt endpoints. Risk is one for an unparseable or incorrect answer and zero for a correct parsed answer. Larger diagnostic scores always predict a risk change. No signs are refit or flipped. For condition c, define

$$G^{a,d}_{m,f} = \frac{1}{|J^{a,d}_c|} \sum_{j \in J^{a,d}_c} \left\{ \mathrm{AUC}(s^{a,d}_{m,f,j}, Y^{a,d}_{cj}) - \mathrm{AUC}(s^{a,d}_{m,\mathrm{loss},j}, Y^{a,d}_{cj}) \right\},$$

where $a \in \{\mathrm{O}, \mathrm{N}\}$ indexes old or new prompt cohorts and $d \in \{\mathrm{O}, \mathrm{N}, \mathrm{F}\}$ indexes old, new, or fresh item splits. The old cohort uses the three fixed references with equal weight. The source

analysis uses the same new-pair set in both item splits: the intersection of pairs with defined AUCs on old and new source items, for each output condition. Confirmation intersects this locked set with fresh-item estimability. AUCs are tie-aware positive-negative ranking fractions. Undefined AUCs remain undefined.

Source crossed decomposition. The source analysis uses only the old and new item splits. For the primary strict/8 condition, and also descriptively for the other seventeen conditions, write $G_{\mathrm{OO}}, G_{\mathrm{ON}}, G_{\mathrm{NO}}, G_{\mathrm{NN}}$ for the four prompt-cohort by item-split gaps. The crossed estimands are

$$P = \frac{(G_{\mathrm{NO}} - G_{\mathrm{OO}}) + (G_{\mathrm{NN}} - G_{\mathrm{ON}})}{2}, \tag{15}$$

$$Q = \frac{(G_{\mathrm{ON}} - G_{\mathrm{OO}}) + (G_{\mathrm{NN}} - G_{\mathrm{NO}})}{2}, \tag{16}$$

$$H = G_{\mathrm{NN}} - G_{\mathrm{NO}} - G_{\mathrm{ON}} + G_{\mathrm{OO}}. \tag{17}$$

The implementation checks the identity

$$P + Q = G_{\mathrm{NN}} - G_{\mathrm{OO}}$$

to numerical tolerance in every draw. P averages the prompt-group contrast over source item splits, Q averages the source item-split contrast over prompt cohorts, and H is the interaction contrast. The magnitudes of P and Q are not used by themselves to claim that a measured change mainly comes from prompts or items. Such a comparison requires sufficient separation of their simultaneous uncertainty intervals. The prompt-group interpretation continues to include training history and pair selection.

Locked source predictions. Before generating any fresh confirmation output, the source stage locks a per-pair prediction and a constant baseline. For each new prompt pair j,

$$\widehat{\Delta G}_{m,f,j} = \frac{1}{2} \sum_{d \in \{\mathrm{O},\mathrm{N}\}} \left(G^{\mathrm{N},d}_{m,f,j} - \overline{G}^{\mathrm{O},d}_{m,f} \right),$$

where $\overline{G}^{\mathrm{O},d}_{m,f}$ is the equal-reference mean over the three old prompt pairs on item split d. The uniform-mean baseline is the mean of $\widehat{\Delta G}_{m,f,j}$ over source-valid new pairs. The zero baseline predicts no prompt-group difference. The prediction artifact records all parser/budget cells, but strict/8 is the confirmatory cell. The prediction file, source analysis, derived source-panel manifest, and provenance are hash-bound in a prospective lock whose fresh-record count must be zero. The confirmation stage verifies this lock and does not re-estimate predictions, baselines, coefficients, score directions, or source-valid pair sets.

Fresh confirmation estimands. On the fresh item split, the observed target for new pair j is

$$\Delta G_{m,f,j} = G^{\mathrm{N,F}}_{m,f,j} - \overline{G}^{\mathrm{O,F}}_{m,f}.$$

The confirmation reports the mean difference $\overline{\Delta G}$, the mean bias $\overline{\widehat{\Delta G} - \Delta G}$, the prediction MAE $\mathrm{mean}\,|\widehat{\Delta G} - \Delta G|$, the zero-baseline MAE $\mathrm{mean}\,|\Delta G|$, and the uniform-mean baseline MAE $\mathrm{mean}\,|\overline{\widehat{\Delta G}} - \Delta G|$. The two MAE improvements are baseline MAE minus prediction MAE. These quantities answer different questions: a nonzero $\overline{\Delta G}$ tests a mean prompt-group difference on fresh items. Mean-direction transfer additionally requires that its interval exclude zero and match the sign of the locked source mean prediction. An MAE improvement over zero tests whether the locked pairwise predictions beat a zero-difference baseline. An MAE improvement over the uniform mean tests whether per-pair predictions add value beyond applying one source mean to every pair. A mean difference by itself is not counted as predictive success.

Resampling, intervals, and missing AUCs. All primary intervals use $20{,}000$ bootstrap draws with seed 20260908 and two-sided Bonferroni percentile endpoints for a family of twenty-four comparisons: two models, two state-score families, and six primary estimands ($P, Q, H, \overline{\Delta G}$, improvement over zero, and improvement over the uniform mean). The source component contributes twelve

intervals and the fresh confirmation contributes twelve. The bootstrap uses independent item streams for old, new, and fresh item splits and one shared multinomial stream for the sixteen new prompt pairs. The same new-pair weights are mapped to each model's sorted pair identities and shared across scores, conditions, and splits. Old references are always the three fixed equal-weight cases. They receive item resampling but no old-pair bootstrap.

An AUC is undefined when a pair-condition has no positive or no negative examples. Undefined AUCs are never filled with chance values. A primary estimate is inferential only when the full production layout, draw count, and seed are used. All old references are defined. At least twelve common new pairs are estimable. For confirmation, at least twelve source-common pairs are also available, and every primary estimand has at least 95% valid bootstrap draws. Otherwise the corresponding cell is retained as ND or reported only as descriptive. The artifacts also retain leave-one-new-pair-out sensitivity and leave-one-old-reference-out checks, without changing the primary masks, bootstrap weights, or frozen predictions.

Control-response summaries. For $\Delta h = h_R - h_L$, $\Delta u = u_R - u_L$ and left readout gradient g_L, the retained quantities are control distance $\|\Delta u\|$, response gain $\|\Delta h\|/\|\Delta u\|$, alignment $g_L^\top \Delta h/(\|g_L\|\|\Delta h\|)$, subspace share $\|U^\top \Delta h\|^2/\|\Delta h\|^2$, and readout residual $R = \Delta r - g_L^\top \Delta h$ (also stored as $R/|\Delta r|$). Zero denominators remain undefined with a reason. Selected-checkpoint losses and local projected and complementary responses accompany these descriptive summaries. They do not fit a new score.

The measured gain describes each finite prompt pair. The local derivative relevant for projected state repair is $K^\top U U^\top g$. Appendix G gives its implementation and relates it to the regenerated-output intervention.

E.1 Results and sensitivity

Expanded historical pairing On the same 200 old questions, expanding the historical prompt set from three selected pairs to eight adjacent-seed pairs changes Ministral's mean norm/loss AUCs from 0.535/0.523 to 0.410/0.776. The sixteen new adjacent pairs give 0.372/0.811. The corresponding Granite values are 0.632/0.857, 0.551/0.825, and 0.554/0.744.

Table 15: Historical-pairing comparison on the same 200 old questions. Each cell reports norm/loss AUCs.

Model	Selected old 3 pairs	Adjacent old 8 pairs	Adjacent new 16 pairs
Ministral	0.535 / 0.523	0.410 / 0.776	0.372 / 0.811
Granite	0.632 / 0.857	0.551 / 0.825	0.554 / 0.744

All primary comparisons retain 16 new pairs, all three fixed references, and all 20,000 resamples. Tables 16 and 17 retain all 24 adjusted intervals and the complete error comparison. Individual predictions are plotted in Figure 2. All primary pair values appear in Table 18. Absolute AUC summaries below show the performance levels behind the relative differences. The complete deletion checks are supplied as data. Ministral's prompt-group contribution remains negative when deleting one new pair or one old reference. These descriptive checks keep the predictions and baseline fixed and do not repeat primary inference.

The two aggregation methods answer different questions. For example, Ministral's old-prompt/old-question state/readout gap is 0.020 when AUCs are averaged over pairs, whereas its pooled discovery gap is 0.233. The latter also compares scores from different pairs. The crossed analysis consistently uses the former, as defined in Eq. 9.

Output-condition dependence. Primary inference uses Table 16. The full descriptive parser–budget matrix is supplied as data, with intervals and coverage. In particular, Ministral's state/readout group contribution is $+0.119$ at one token and -0.402 at eight. Thus the primary conclusion applies to strict eight-token scoring.

Table 16: All 24 primary comparisons. Error improvements are baseline MAE minus forecast MAE. Positive values favor the forecast.

Model	Score	Estimand	Estimate [adjusted interval]
Ministral	State/readout	P	-0.402 [-0.544, -0.266]
Ministral	State/readout	Q	0.007 [-0.093, 0.108]
Ministral	State/readout	H	-0.022 [-0.202, 0.149]
Ministral	State/readout	$\overline{\Delta G}$	-0.365 [-0.550, -0.190]
Ministral	State/readout	$E_0 - E_{\mathrm{pred}}$	0.302 [0.014, 0.453]
Ministral	State/readout	$E_{\mathrm{mean}} - E_{\mathrm{pred}}$	0.060 [-0.055, 0.139]
Ministral	State norm	P	-0.464 [-0.591, -0.335]
Ministral	State norm	Q	-0.030 [-0.149, 0.086]
Ministral	State norm	H	-0.025 [-0.214, 0.166]
Ministral	State norm	$\overline{\Delta G}$	-0.433 [-0.611, -0.257]
Ministral	State norm	$E_0 - E_{\mathrm{pred}}$	0.385 [0.082, 0.516]
Ministral	State norm	$E_{\mathrm{mean}} - E_{\mathrm{pred}}$	0.056 [-0.054, 0.123]
Granite	State/readout	P	-0.006 [-0.174, 0.136]
Granite	State/readout	Q	0.050 [-0.034, 0.133]
Granite	State/readout	H	-0.100 [-0.250, 0.046]
Granite	State/readout	$\overline{\Delta G}$	-0.037 [-0.201, 0.111]
Granite	State/readout	$E_0 - E_{\mathrm{pred}}$	0.034 [-0.093, 0.149]
Granite	State/readout	$E_{\mathrm{mean}} - E_{\mathrm{pred}}$	0.032 [-0.092, 0.147]
Granite	State norm	P	-0.009 [-0.181, 0.135]
Granite	State norm	Q	0.045 [-0.037, 0.127]
Granite	State norm	H	-0.087 [-0.239, 0.054]
Granite	State norm	$\overline{\Delta G}$	-0.043 [-0.199, 0.104]
Granite	State norm	$E_0 - E_{\mathrm{pred}}$	0.036 [-0.074, 0.143]
Granite	State norm	$E_{\mathrm{mean}} - E_{\mathrm{pred}}$	0.034 [-0.075, 0.139]

Table 17: Confirmation error and signed bias (forecast minus observation). MAEs are descriptive point estimates. Adjusted error-improvement intervals are in Table 16.

Model	Score	E_{pred}	E_0	E_{mean}	Bias [95% interval]
Ministral	State/readout	0.063	0.365	0.123	-0.037 [-0.137, 0.061]
Ministral	State norm	0.047	0.433	0.103	-0.032 [-0.134, 0.071]
Granite	State/readout	0.085	0.119	0.117	0.031 [-0.054, 0.109]
Granite	State norm	0.085	0.122	0.120	0.034 [-0.047, 0.110]

Table 18: Every ECQA fresh-question prediction under strict eight-token scoring. Targets are changes in the state-minus-loss AUC gap relative to the three old reference pairs. Each model retains all 16 pairs. The prompts themselves were already observed in the source phase. Full errors and both baselines accompany the data.

Model	Pair	$\Delta G_{\text{readout}}$	$\widehat{\Delta G}_{\text{readout}}$	ΔG_{norm}	$\widehat{\Delta G}_{\text{norm}}$
Ministral	7201–7202	-0.465	-0.461	-0.516	-0.501
	7203–7204	-0.230	-0.178	-0.385	-0.323
	7205–7206	-0.495	-0.631	-0.557	-0.640
	7207–7208	-0.175	-0.318	-0.241	-0.383
	7209–7210	-0.470	-0.490	-0.548	-0.593
	7211–7212	-0.605	-0.585	-0.603	-0.601
	7213–7214	-0.452	-0.474	-0.547	-0.504
	7215–7216	-0.140	-0.161	-0.280	-0.275
	7217–7218	-0.287	-0.431	-0.369	-0.434
	7219–7220	-0.415	-0.495	-0.462	-0.564
	7221–7222	-0.351	-0.335	-0.274	-0.309
	7223–7224	-0.578	-0.529	-0.554	-0.558
	7225–7226	-0.421	-0.556	-0.469	-0.530
	7227–7228	-0.262	-0.200	-0.419	-0.446
	7229–7230	-0.194	-0.267	-0.206	-0.262
	7231–7232	-0.298	-0.323	-0.494	-0.506
Granite	7101–7102	-0.105	-0.252	-0.183	-0.292
	7103–7104	-0.135	0.046	-0.151	0.043
	7105–7106	-0.015	0.002	-0.005	0.005
	7107–7108	0.080	0.155	0.055	0.134
	7109–7110	0.117	0.264	0.070	0.232
	7111–7112	0.264	0.195	0.261	0.171
	7113–7114	0.057	0.083	0.044	0.090
	7115–7116	-0.409	-0.393	-0.263	-0.299
	7117–7118	0.083	0.170	0.158	0.237
	7119–7120	-0.060	-0.013	-0.081	-0.034
	7121–7122	-0.025	0.013	-0.031	-0.006
	7123–7124	-0.136	-0.341	-0.242	-0.416
	7125–7126	-0.090	-0.002	-0.107	-0.018
	7127–7128	0.053	0.041	0.039	0.037
	7129–7130	-0.205	-0.065	-0.197	-0.053
	7131–7132	-0.065	-0.001	-0.061	0.019

Table 19: Absolute AUCs in the crossed comparison, averaged equally over prompt pairs. Old and fresh question sets contain 200 questions each. The new source set contains 400. These are descriptive means. All individual-pair AUCs and source intervals are retained in the data supplement.

Model	Prompts	Questions	Pairs	State/readout	State norm	Loss
Ministral	Old	Old	3	0.543	0.535	0.523
Ministral	Old	New	3	0.535	0.492	0.496
Ministral	Old	Fresh	3	0.492	0.461	0.485
Ministral	New	Old	16	0.441	0.372	0.811
Ministral	New	New	16	0.460	0.353	0.834
Ministral	New	Fresh	16	0.443	0.344	0.800
Granite	Old	Old	3	0.630	0.632	0.857
Granite	Old	New	3	0.746	0.737	0.872
Granite	Old	Fresh	3	0.707	0.713	0.822
Granite	New	Old	16	0.561	0.554	0.744
Granite	New	New	16	0.549	0.542	0.731
Granite	New	Fresh	16	0.555	0.555	0.708

Response interpretation. In particular, the gain is a finite-pair ratio, not a Jacobian operator norm. Mean signed residuals are not absolute approximation errors. These measurements describe the observed response geometry without identifying a cause of the group difference.

F AUC DECOMPOSITION AND STABILITY CONDITIONS

Let n_0, n_P, n_C be the numbers of risk-stable, parseability-channel-positive, and correctness-channel-positive records. For score s, empirical AUC is

$$\widehat{\mathrm{AUC}}(s) = \frac{1}{(n_P + n_C)n_0} \sum_{i \in Z_P \cup Z_C} \sum_{j \in Z_0} \kappa(s_i, s_j) \tag{18}$$

$$= \frac{n_P}{n_P + n_C} \widehat{\mathrm{AUC}}_P(s) + \frac{n_C}{n_P + n_C} \widehat{\mathrm{AUC}}_C(s). \tag{19}$$

This is a partition of the same positive–negative comparisons. It requires the same negative pool for both channel AUCs. Removing risk-stable parseability switches from one channel's negative pool would define a different conditional ranking problem and invalidate this identity. The population version follows by conditioning a positive draw on its channel. Subtraction gives Eq. (8). Its zero is $-\delta_C/(\delta_P - \delta_C)$ when $\delta_P > 0 > \delta_C$.

The mixture explains a fixed collection of conditional distributions. It does not imply that changing prompts or tasks changes only π. Both conditional positive distributions and the negative distribution can change in a new setting. We evaluate the fixed prompt pairs on new items and the independent training cohorts to measure these distinct forms of transfer.

Clustered resampling. Each discovery or replication case contains 200 item identities shared across three prompt pairs. A bootstrap draw resamples those identities with replacement and applies each item's multiplicity to all its pair records and all score families. We compute weighted AUC with half credit for ties. Discovery intervals use the 0.025 and 0.975 quantiles of 10,000 draws with seed 20260906. Replication intervals for the six primary contrasts use quantiles $0.05/12$ and $1 - 0.05/12$, giving nominal family confidence 0.95 by Bonferroni adjustment. These are bootstrap intervals, not an exact finite-sample coverage guarantee. All six primary contrasts have 10,000 valid draws. A leave-one-item-out analysis removes each identity from all three pairs together.

For a channel with no positives, its AUC is undefined. For sparse channels, a bootstrap draw with no positive or no negative weight is omitted and the valid count is reported. For example, Granite's discovery parseability channel uses 9 positive records and 9,998 valid bootstrap draws. This does not affect its aggregate AUC or the six primary replication comparisons. Full draw counts and intervals are supplied in the machine-readable summary.

F.1 FINITE-PANEL STABILITY OF AUC-GAP COMPARISONS

The following deterministic bound connects control-sensitive scores to their behavioral comparison target on a fixed panel. It separates score drift from changes in which items form positive–negative comparisons.

Proposition 1 (Fixed-panel AUC-gap stability)**.** *Fix item indices $i = 1, \ldots, n$ and two prompt/readout conditions $a \in \{0, 1\}$. Let $y_{a,i} \in \{0, 1\}$, with $n_{a,+} = \sum_i y_{a,i}$ and $n_{a,-} = n - n_{a,+}$ both nonzero. For score family $k \in \{f, \ell\}$, let $s^k_{a,i} \in \mathbb{R}$ be fixed score values, and define the tie-aware comparison kernel*

$$\kappa(x, z) = \mathbf{1}\{x > z\} + \tfrac{1}{2}\mathbf{1}\{x = z\}.$$

Set

$$w^a_{ij} = \frac{\mathbf{1}\{y_{a,i} = 1,\ y_{a,j} = 0\}}{n_{a,+} n_{a,-}}, \qquad c^a_{ij} = \kappa(s^f_{a,i}, s^f_{a,j}) - \kappa(s^\ell_{a,i}, s^\ell_{a,j}),$$

and $G_a = \sum_{i,j} w^a_{ij} c^a_{ij}$. Suppose that known numbers $\epsilon^k_i \geq 0$ satisfy $|s^k_{1,i} - s^k_{0,i}| \leq \epsilon^k_i$ for every i, k. Define

$$M^k_{ij} = \mathbf{1}\left\{\epsilon^k_i + \epsilon^k_j > 0, \quad |s^k_{0,i} - s^k_{0,j}| \leq \epsilon^k_i + \epsilon^k_j\right\}.$$

Then

$$|G_1 - G_0| \leq \min\left\{2,\ \sum_{i,j} |w^1_{ij} - w^0_{ij}| + \sum_{i,j} w^0_{ij}\left(M^f_{ij} + M^\ell_{ij}\right)\right\}. \tag{20}$$

Proof. Since each G_a is the difference of two AUCs, $G_a \in [-1, 1]$, giving $|G_1 - G_0| \le 2$. For the refined bound, add and subtract $\sum_{i,j} w^0_{ij} c^1_{ij}$:

$$\begin{aligned} |G_1 - G_0| &\le \left| \sum_{i,j} (w^1_{ij} - w^0_{ij}) c^1_{ij} \right| + \left| \sum_{i,j} w^0_{ij} (c^1_{ij} - c^0_{ij}) \right| \\ &\le \sum_{i,j} |w^1_{ij} - w^0_{ij}| + \sum_{i,j} w^0_{ij} |c^1_{ij} - c^0_{ij}|, \end{aligned}$$

because $c^1_{ij} \in [-1, 1]$. It remains to bound the kernel changes. Fix $k \in \{f, \ell\}$ and write $d^k_a = s^k_{a,i} - s^k_{a,j}$ and $E^k_{ij} = \epsilon^k_i + \epsilon^k_j$. If $E^k_{ij} = 0$, then $d^k_1 = d^k_0$, including the tie case, so the kernel is unchanged. If $E^k_{ij} > 0$ and $|d^k_0| > E^k_{ij}$, then d^k_1 has the same strict sign as d^k_0 because $|d^k_1 - d^k_0| \le E^k_{ij}$. The kernel is again unchanged. In all remaining cases, the absolute change of a tie-aware comparison is at most one, and those cases are exactly covered by M^k_{ij}. Therefore

$$|\kappa(s^k_{1,i}, s^k_{1,j}) - \kappa(s^k_{0,i}, s^k_{0,j})| \le M^k_{ij}.$$

Using this bound for $k = f$ and $k = \ell$ gives $|c^1_{ij} - c^0_{ij}| \le M^f_{ij} + M^\ell_{ij}$, completing the proof. □

Two immediate cases are useful for interpreting the experiments. If the score values are unchanged, all $\epsilon^k_i = 0$ and the score-drift term vanishes: only the change in AUC comparison weights can move the gap. If the labels are also unchanged, then $w^1 = w^0$ and $G_1 = G_0$. This is the finite-panel reason that a model/readout condition with unchanged empirical targets has no measured AUC-gap change. Conversely, a smooth response bound alone does not control ranking when many old pairwise score differences lie within the drift margin, or when the readout changes enough positives and negatives to make $\sum_{ij} |w^1_{ij} - w^0_{ij}|$ large.

For a fixed ridge diagnostic, the score-drift envelope can be obtained from feature-drift envelopes. If

$$s_{a,i} = \theta_0 + \sum_{m=1}^{p} \theta_m \frac{\phi_{a,i,m} - \mu_m}{\tilde{\sigma}_m}, \qquad \tilde{\sigma}_m > 0,$$

with coefficients, means, signs, and scales fixed before the comparison, and if $|\phi_{1,i,m} - \phi_{0,i,m}| \le \eta_{i,m}$, then

$$|s_{1,i} - s_{0,i}| \le \sum_{m=1}^{p} \frac{|\theta_m|}{\tilde{\sigma}_m} \eta_{i,m}. \tag{21}$$

This is only a path from feature sensitivity to the finite-panel certificate. Pontryagin sensitivity results such as the horizon-uniform Green certificate, local reconstruction, and a posteriori bounds of Huang and Song (Huang and Song, 2026, v3 HTML, Thm. 3.2, Thm. 4.1, Prop. 5.3) could supply an input or response envelope only after their stationarity, hyperbolicity, transversality, and feature-Lipschitz hypotheses are verified for the system at hand. The present transformer measurements do not certify those hypotheses, and the output budget is a parser/generation condition rather than the control horizon in that theory.

Retained-record check under fixed scores. All three diagnostic values are identical across the 18 parser–budget conditions at every retained pair–item coordinate. Thus the score-drift term is exactly zero. For strict eight versus one token, all 96 pair–score comparisons satisfy the comparison-weight bound. Direct tie-aware positive–negative comparisons reproduce the model mean changes within 10^{-12} of the reported estimates. Table 20 reports descriptive means over the 16 pairs. The complete per-pair check is included in the reproduction material. In Qwen, every risk, parseability, correctness, and channel label is also unchanged across all 18 conditions at all 6,400 pair–item coordinates. The resulting zero bound explains its exact empirical invariance. These are deterministic checks on observed labels, with no additional confirmatory test or forecast of an unobserved score-drift envelope.

Table 20: Fixed-score comparison-weight bound, strict eight versus one token. The bound is checked within each pair before averaging. Both state-score families use the same behavioral labels and hence the same bound.

Model	Mean $\|w^1 - w^0\|_1$	Mean $\|D_j\|$, state/readout	Mean $\|D_j\|$, norm
Ministral	0.238	0.037	0.067
Granite	0.671	0.082	0.084
Qwen	0.000	0.000	0.000

F.2 Changes in both AUC comparison sets

Changing an output condition changes the pair-level risk-change label, even though all diagnostic scores are fixed. Let a_i and b_i be these labels under strict budgets eight and one, and group items by $g_i = (a_i, b_i) \in \{00, 01, 10, 11\}$. For score family f, define the mean kernel contrast between groups g, h as

$$H^f_{gh} = \frac{1}{n_g n_h} \sum_{i \in g, j \in h} \{\kappa(s_{f,i}, s_{f,j}) - \kappa(s_{\mathrm{loss},i}, s_{\mathrm{loss},j})\}, \tag{22}$$

when both groups are nonempty. Let $w_q(g,h) = n_g n_h/(N_{q,+}N_{q,-})$ when g is positive and h negative under condition q, and zero otherwise. Whenever both AUCs are defined, partitioning their comparisons gives

$$D_f = \sum_{g,h} \{w_1(g,h) - w_8(g,h)\} H^f_{gh}. \tag{23}$$

An empty comparison block has zero weight. The 11–00 block retains both labels but changes its normalization. Other blocks involve at least one relabeled pair–item record. The released table supplies every block, and their sum reproduces each reported paired change. This is exact descriptive accounting for target relabeling. It assigns no causal mediation effect to the groups.

The full block decomposition is retained as a numerical table in the data supplement. No block is selected or omitted from the sum.

Directional and answer-change checks Directional checks distinguish improvement from deterioration rather than treating every risk change alike. Each direction is compared with unchanged risk, excluding the opposite direction. We also evaluate parsed-answer changes only on records where both endpoints parse, and token-sequence changes separately. AUCs without both classes remain undefined, and each table reports its valid-pair coverage. These alternative targets are descriptive checks of the specified behavioral readout.

G Control response and prompt-update interventions

Write $h_i : \mathbb{R}^{8d} \to \mathbb{R}^q$ for the frozen model's scored-position response to the vectorized prompt. At u_s, where h_i and r_i are differentiable, the chain rule gives $D(r_i \circ h_i)(u_s)[v] = g_{i,s}^\top K_{i,s} v$ for every prompt perturbation v. Taking the Euclidean adjoint yields Eq. (3). This is a local sensitivity identity. It does not establish controllability or identify a dynamical system from terminal states alone.

Suppose r_i is twice continuously differentiable in a neighborhood of the state segment and $\|\nabla_h^2 r_i(h_s + t\Delta h_i)\|_{\mathrm{op}} \le M_i$ for $0 \le t \le 1$. Then

$$r_i(h_t) - r_i(h_s) = g_{i,s}^\top \Delta h_i + R_i, \qquad |R_i| \le \tfrac{1}{2} M_i \|\Delta h_i\|^2. \tag{24}$$

The integral remainder is

$$R_i = \int_0^1 (1-t) \Delta h_i^\top \nabla_h^2 r_i(h_s + t\Delta h_i) \Delta h_i \, dt.$$

The chord evaluates a specified smooth readout. Actual prompt updates follow the frozen model's response map and are evaluated by regenerated outputs.

Table 21: Alternative behavioral targets for the original state/readout score. AUCs average defined pairs. Positive and included-row counts describe the full specified target panel. Undefined pair AUCs are not replaced by 0.5.

Panel	Target	AUC	Defined pairs	Positive/ included rows
ECQA / Ministral	Deterioration	0.358	16/16	609/5896
ECQA / Ministral	Improvement	0.561	16/16	504/5791
ECQA / Ministral	Parsed-answer change	0.469	16/16	1181/6117
ECQA / Ministral	Risk change	0.460	16/16	1113/6400
ECQA / Ministral	Token change	0.457	16/16	1462/6400
ECQA / Granite	Deterioration	0.607	13/16	821/4761
ECQA / Granite	Improvement	0.611	15/16	1639/5579
ECQA / Granite	Parsed-answer change	0.753	13/16	756/3912
ECQA / Granite	Risk change	0.549	16/16	2460/6400
ECQA / Granite	Token change	0.672	13/16	3244/6400
ECQA / Qwen	Deterioration	0.504	16/16	378/6088
ECQA / Qwen	Improvement	0.761	16/16	312/6022
ECQA / Qwen	Parsed-answer change	0.616	16/16	850/6400
ECQA / Qwen	Risk change	0.622	16/16	690/6400
ECQA / Qwen	Token change	0.616	16/16	850/6400
GSM8K / Ministral	Deterioration	0.481	16/16	218/2970
GSM8K / Ministral	Improvement	0.488	16/16	230/2982
GSM8K / Ministral	Parsed-answer change	0.489	16/16	633/3177
GSM8K / Ministral	Risk change	0.484	16/16	448/3200
GSM8K / Ministral	Token change	0.536	16/16	3024/3200

For an orthonormal empirical response basis U, $g_{i,s}^\top \Delta h_i = (U^\top g_{i,s})^\top (U^\top \Delta h_i) + g_{i,s}^\top (I - UU^\top)\Delta h_i$. CF-PED integrates these complementary terms along the readout chord. The primary diagnostic instead uses their left-endpoint approximations as fixed regression features. The norm control omits both gradient terms. The empirical basis and a numerical readout reconstruction therefore play different roles from an independently verified behavior prediction.

For the selected state arm, let $P = UU^\top$. At the current prompt u_s, the implementation computes $g_{i,s}$, holds $Pg_{i,s}$ fixed during backpropagation, and differentiates the local scalar surrogate

$$\phi_{i,s}(u) = \langle Pg_{i,s}, h_i(u) \rangle, \qquad \nabla_u \phi_{i,s}(u_s) = K_{i,s}^\top P g_{i,s}.$$

The auxiliary minimization gradient is the negative of this direction. The surrogate is updated with the current readout gradient at each step. This is the implemented adjoint connection between a projected state readout and a prompt update. Our continuation experiment evaluates the task-loss plus auxiliary objectives with equal update budgets and matched final control displacements. Its regenerated-output comparison measures intervention effects, while the selected-minus-random-family contrast measures the extra benefit of auxiliary-objective selection.

Integrated readout accounting. For $h(\alpha) = h_s + \alpha \Delta h$, CF-PED partitions the scalar change as

$$\tau = A + O + \varepsilon, \qquad A = \int_0^1 \nabla r(h(\alpha))^\top UU^\top \Delta h \, d\alpha, \tag{25}$$

where O replaces $UU^\top$ by $I - UU^\top$ and ε is the numerical residual. This applies integrated-gradient completeness (Sundararajan et al., 2017) to the chosen state path. The prediction features use the local terms rather than this integral.

G.1 Update objectives and observed behavioral components

The continuation starts from the worse-calibration-risk member of the first eligible pair: Granite 2101 relative to 2114, and Ministral 2102 relative to 2111. In both cases the selected auxiliary family is state repair, using the projected direction $K^\top Pg$. The random-family control balances control, state, route, and parse repair across eight streams, while the wrong-link control uses the direct pre-answer route objective. Stream identities are 4101–4108. All arms share the task-loss gradient and set the

first-step weighted auxiliary-gradient norm to 0.10 times its norm. They train for 25 Adam updates with learning rate 0.03 and batch size eight on 200 continuation examples.

After optimization, the final displacement of every arm is normalized to the median norm of the eight selected-arm displacements within its model: 31.0723 for Granite and 31.9965 for Ministral. This matches final displacement, not optimization-path length. Evaluation uses the same 200 fresh holdout items across arms and streams. These holdout items are disjoint from the original analysis and the later independent replication.

For each bootstrap draw, item multiplicities and stream multiplicities are sampled independently and applied jointly to all arms. The endpoint component analysis uses 10,000 draws and seed 20260906. The two terms of Eq. (12) add to total recovery in every record and every draw. A component can be negative. Its fraction of total recovery can exceed one when the other component offsets it. We therefore report signed changes as the primary decomposition.

Table 22: Net risk reduction and endpoint components in percentage points, with item-by-stream 95% bootstrap intervals.

Model	Arm	Parse switch	Same-parse correctness	Total
Granite	Selected	-0.250 [-0.812, 0.000]	4.125 [-0.250, 8.625]	3.875 [-0.500, 8.438]
	Random family	-1.688 [-4.250, 0.000]	4.250 [-0.125, 8.688]	2.562 [-2.502, 7.625]
	Wrong-link	-5.312 [-9.062, -2.250]	2.250 [-2.252, 6.750]	-3.062 [-8.875, 2.564]
Ministral	Selected	37.812 [26.562, 47.812]	1.188 [-0.562, 3.312]	39.000 [27.312, 49.438]
	Random family	37.250 [25.688, 47.812]	0.625 [-1.250, 2.625]	37.875 [26.125, 48.938]
	Wrong-link	36.000 [28.813, 43.125]	-0.125 [-2.250, 2.000]	35.875 [28.312, 43.252]

The selected arm's contribution from changed parseability divided by net recovery is 0.970 $[0.919, 1.016]$ for Ministral. Its contribution from unchanged parseability is 1.188 percentage points, with an interval spanning zero. The corresponding ratios are 0.983 and 1.003 for random-family and wrong-link arms. These ratios summarize paired outcome changes, without assigning latent correctness to unparseable answers.

The previously computed paired direction contrasts use the original item-by-stream resamples. Granite's selected-minus-wrong-link gain is 6.938 points $[2.000, 12.063]$, and Ministral's is 3.125 points $[-9.250, 13.250]$. Together with the selected-minus-random-family contrasts in Section 6.1, they show why total recovery and objective selection answer different empirical questions.

H Numerical accounting and the MATH extension

Readout reconstruction. Adaptive integration uses orders 8, 16, 32, and 64 for the state-path readout in Eq. (25). We reassess the retained final residuals with a mixed absolute–relative accuracy criterion,

$$|\varepsilon| \leq \max(10^{-6}, 0.01|\tau|). \tag{26}$$

All 96 primary paths and all 96 single-position paths in each of the three model–task cases satisfy this criterion. The two MATH paths requiring 64 nodes have small readout changes. Their absolute residuals are below 10^{-6}. The complete residual arrays and node counts are retained in the data supplement. This check evaluates stored numerical residuals. It does not rerun quadrature or change the predictors used in this paper. The residual arrays and their reconstruction script are included.

Longer-answer task. The MATH extension uses the MATH dataset (Hendrycks et al., 2021), with deterministic answer extraction and a 512-token generation limit. Of sixteen runs per model, Granite has no pair satisfying the loss-equivalence criterion. Ministral has pairs 5103–5106, 5109–5112, and 5113–5114. Its completed calibration has 32 random prompt draws and its held-out panel has 600 item–pair records.

Among those records, 114 risk changes comprise 7 with changed parseability and 107 between parseable answers. A further 40 parseability switches leave overall risk unchanged. The loss score's correctness-channel AUC is 0.527 $[0.462, 0.591]$, compared with 0.494 for the primary score. Both scores have parseability-channel AUC 0.492 on only seven positives. The full per-pair MATH records are retained in the data supplement. These measurements show that the ECQA ordering does not

establish useful discrimination on this longer-answer task. They do not isolate an architectural or numerical explanation for the difference.

I GSM8K TRANSFER TO UNSEEN PROMPT PAIRS

The diagnostic and forecasting stages use separate prompts and items. The following protocol precedes the complete confirmation results.

Data and prompt optimization. The GSM8K extension uses Ministral-3-8B-Instruct-2512-BF16 and 80 new eight-vector soft prompts, with model weights fixed. Items are ordered by the SHA256 of their NFKC, case-folded, whitespace-normalized questions. Source coordinates and normalized question hashes are disjoint across roles. From the training split, 128 items support optimization, 64 support checkpoint selection, 200 plus 100 support diagnostic calibration, and 200 support pair-forecast development. Confirmation uses 200 items from the official test split. A separate timing pool supplies no fitting labels. Reference solutions retain all calculation steps after removal of calculator annotations. The final delimiter becomes `Final Answer:`. The teacher-forced loss $L_i(u)$ averages negative log likelihood over the complete reference solution, including its final answer field. AdamW takes 100 updates at learning rate 0.1, batch size eight, and clipping norm one. The minimum selection-loss checkpoint is chosen from steps $0, 10, \ldots, 100$. Backbone computation uses BF16, with FP32 gold-token readout.

Separate prompt cohorts and pairing. The first 16 seeds form eight adjacent-seed calibration pairs. The next 32 and final 32 seeds form development and confirmation cohorts, respectively. Within each latter cohort, four consecutive eight-seed blocks are paired independently. Odd blocks pair adjacent seeds. Even blocks sort by selected checkpoint loss, breaking ties by seed, and pair neighboring entries. Each cohort therefore has 16 endpoint-disjoint pairs, eight of each type. The lower seed is always the left endpoint. Calibration, development, and confirmation prompts share no endpoints. Diagnostic fitting uses only calibration pairs. The separate pair-level forecasters use development pairs. The three native-generation panels contain 4,800 calibration endpoints, 6,400 development endpoints, and 6,400 confirmation endpoints: 17,600 in total. The independent unseen-prompt evaluation is the last panel alone. The state-only calibration probes and response perturbations are accounted for separately and do not enlarge its behavioral sample.

Response subspace and fixed item-level scores. Of the 200 calibration items, 32 fixed probes are excluded from the 168-item subspace core. For every core item, 32 Gaussian perturbations cycle through the eight left calibration prompts. Each direction has Frobenius norm $0.0025m\sqrt{8}$, where m is the median FP32 embedding-row norm. Response differences include every teacher-forced solution-token position. FP32 differences are accumulated in FP64 into an uncentered second moment, with equal weight per token position and perturbation. Five item folds (canonical index modulo five) select from ranks $\{1, 2, 4, 8, 16, 32\}$ using the smallest rank within one standard error of the best mean held-out relative reconstruction residual. The final basis U uses the complete core moment. Tail energy is the sum of discarded eigenvalues. The selected rank is 32. Its mean held-out residual energy fraction is 0.8998 (standard error 0.0043), quantifying response energy outside the fitted subspace.

For an item with n_i scored positions, let $H_s, H_t \in \mathbb{R}^{n_i \times d}$, $\Delta H = H_t - H_s$, and $g_s = \nabla_{H_s}[-L_i(u_s)]$. The primary diagnostic uses

$$\left(a_{\|},\ a - a_{\|},\ \frac{\|H_s\|_F}{\sqrt{n_i}},\ \frac{\|H_t\|_F}{\sqrt{n_i}},\ \|u_t - u_s\|_F,\ e_{\text{tail}}\right),$$
$$a = \langle g_s, \Delta H\rangle_F, \quad a_{\|} = \langle g_s U, \Delta H U\rangle_F.$$

The norm diagnostic zeroes the first two coordinates. Each standardized ridge fit uses all 300×8 calibration item–pair rows and the binary risk-change target defined below. Five-fold item-grouped MAE selection chooses the largest regularization within one standard error of the best candidate, from $\{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$. Its objective is summed squared error plus an ℓ_2 penalty. Both diagnostics are then fixed. The comparator score is $s_{\text{loss}} = |L_i(u_t) - L_i(u_s)|$.

Forecasting a new pair's diagnostic gain. For family $f \in \{\text{primary}, \text{norm}\}$, the development target is

$$G_{j,f} = \text{AUC}(s_f, z_j) - \text{AUC}(s_{\text{loss}}, z_j), \tag{27}$$

computed separately for each of the 16 development pairs on 200 items. The cheap forecaster uses the absolute selection-loss gap, prompt RMS displacement, and pairing type. The full forecaster adds response RMS gain relative to prompt RMS displacement, signed cosine alignment of g_s with ΔH, subspace response-energy fraction, and the relative readout residual $|\Delta r - \langle g_s, \Delta H\rangle_F| / \max(|\Delta r|, 10^{-8})$, where $\Delta r = L_i(u_s) - L_i(u_t)$. Probe-derived features are equal-item means over the same 32 calibration probes, using all solution positions and no generated behavior.

Separate full and cheap ridge forecasts for each family use nested four-fold development validation, with four inner folds, training-fold standardization, and MAE selection over $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Pair folds are stratified by pairing type and keep any shared endpoints together. Exact MAE ties select stronger regularization. A final four-fold selection on all development pairs sets the fit minimizing mean squared error plus $\alpha\|\beta\|_2^2$. Baselines also predict zero or the same mean of defined development gains for every pair. Invalid features or unavailable ridge fits fall back to that fixed mean, with a flag and incurred costs retained. Forecast rules are fixed before confirmation-prompt training. The 16 confirmation predictions are fixed from their selection and probe features before any confirmation behavior is measured. Confirmation labels enter no fitting or standardization step.

Native behavior and estimability. Native greedy generation uses the chat-formatted problem prefix, without reference-solution tokens, and stops at EOS or 512 new tokens. Strict parsing requires a terminal `Final Answer:` numeric field, rejects malformed or conflicting fields, and compares finite decimals or fractions exactly. Risk b_{is} is zero for a correct parsed answer and one otherwise. $z_{i,st} = |b_{is} - b_{it}|$. Thus the target is a correctness-risk change, including parser failures, rather than any change in output text. The primary analysis uses 512-token behavior. The first 256 tokens of the same trace give paired descriptive support with unchanged scores and forecasts. Parser status and stopping reasons accompany both budgets.

Confirmation uncertainty. For each family and each baseline $b \in \{\text{zero}, \text{mean}, \text{cheap}\}$, the contrast is the equal-pair mean of $|\widehat{G}_{j,b} - G_{j,f}| - |\widehat{G}_{j,\text{full}} - G_{j,f}|$. These six MAE-improvement contrasts use 20,000 bootstrap draws, jointly resampling items and prompt pairs within pairing type. Every family and baseline shares the same item counts, pair counts, and valid-draw mask. All forecasts remain fixed. All contrasts use the common set of pairs with defined gains and finite predictions. AUC gives ties half credit and remains undefined when either risk-change class is absent. Inference requires at least 12 common valid original pairs, at least six of each pairing type, and at least 95% valid draws. Otherwise estimates are descriptive and intervals are withheld. The retained Bonferroni family size is 22 at family level 0.05, giving percentile endpoints $0.05/(2 \cdot 22)$ and $1 - 0.05/(2 \cdot 22)$. Undefined AUCs are never filled. Each pair retains its gain, errors or undefined status, and delete-one-pair estimates. Mean observed gain, mean forecasts, and signed mean bias use the common valid set. Group summaries have no separate transfer-success test or prespecified equivalence margin.

I.1 All confirmation comparisons

All 16 confirmation pairs are common-valid: eight per pairing type. All 20,000 resamples meet the coverage conditions. Positive MAE improvement favors the full forecast. The six contrasts retain the original 22-comparison correction, including comparisons outside this reduced experimental scope. The family size is not reduced after observing results.

Parsing and reproduction. At 512 tokens, 6,377 of 6,400 confirmation outputs (99.64%) parse successfully. 6,369 (99.52%) do so at 256 tokens. The latter uses prefixes of the same native trajectories and unchanged forecasts, and is descriptive support. Full parser and stopping categories are retained in `parser_coverage.csv`. `support_256.json` retains all secondary comparisons. Independent CPU reconstruction checks saved scores, fixed forecasts, native token parsing and point statistics. The shared locked statistical implementation replays all 20,000 draws. The data index links

Table 23: GSM8K MAE improvements and simultaneous intervals.

Diagnostic	Baseline	Improvement	Lower	Upper
State/readout	Zero	-0.0100	-0.0544	0.0431
State/readout	Development mean	-0.0012	-0.0050	0.0051
State/readout	Cheap features	0.0079	-0.0292	0.0362
State norm	Zero	-0.0076	-0.0753	0.0648
State norm	Development mean	-0.0025	-0.0271	0.0261
State norm	Cheap features	0.0023	-0.0130	0.0172

Table 24: All GSM8K observed gaps and full forecasts at 512 tokens. Each pairing type has eight pairs. Other forecasts and every error are supplied in the accompanying CSV.

Pair	Pairing	G_{readout}	$\widehat{G}_{\mathrm{readout}}$	G_{norm}	$\widehat{G}_{\mathrm{norm}}$
1	Adjacent seed	-0.0954	-0.0548	-0.1027	-0.0723
2	Adjacent seed	0.0195	-0.0564	-0.0237	-0.0815
3	Adjacent seed	-0.1448	-0.0569	-0.1233	-0.0877
4	Adjacent seed	0.0233	-0.0625	0.0089	-0.1094
5	Selection loss	0.1047	-0.0536	0.0796	-0.0784
6	Selection loss	0.0706	-0.0516	0.0919	-0.0397
7	Selection loss	0.1253	-0.0460	0.1449	-0.0294
8	Selection loss	-0.1972	-0.0614	-0.2261	-0.1130
9	Adjacent seed	0.1023	-0.0612	0.0677	-0.1156
10	Adjacent seed	0.0286	-0.0610	0.0147	-0.0765
11	Adjacent seed	-0.0380	-0.0657	-0.1026	-0.1202
12	Adjacent seed	-0.0681	-0.0597	-0.0847	-0.0898
13	Selection loss	0.1628	-0.0529	0.1639	-0.0664
14	Selection loss	0.0216	-0.0542	0.0408	-0.0753
15	Selection loss	-0.1019	-0.0487	-0.0898	-0.0320
16	Selection loss	-0.0835	-0.0480	-0.0827	-0.0329

these receipts separately from document rebuilding. These checks authenticate the measurements. They do not turn intervals spanning zero into evidence of a forecast improvement.

J Evaluation cost and diagnostic selection

Question and evidence unit. The source study screens prompt-induced risk changes on labeled ECQA with greedy strict eight-token scoring (Qwen non-thinking). It uses 54 existing pairs and 107 unique prompts on 400 shared items: three old and sixteen new pairs for each of Ministral and Granite, plus sixteen Qwen pairs. Ministral's shared endpoint stays in one validation block. Repeated endpoint measurement adds no independent prompts. No new confirmation prompts were trained after this check.

Control-response inputs. Let $m = 32$ denote calibration probes, $\Delta u = u_t - u_s$, $\Delta h_i = h_i(u_t) - h_i(u_s)$, $g_i = \nabla r_i(h_i(u_s))$, $\Delta r_i = r_i(h_i(u_t)) - r_i(h_i(u_s))$, and $q = \dim(u)$. The six fixed features are

$$x_1 = |L_{\mathrm{sel}}(u_t) - L_{\mathrm{sel}}(u_s)|, \qquad x_2 = \|\Delta u\|/\sqrt{q}, \tag{28}$$

$$x_3 = m^{-1}\sum_i \|\Delta h_i\|/\|\Delta u\|, \qquad x_4 = m^{-1}\sum_i \frac{g_i^\top \Delta h_i}{\|g_i\|\|\Delta h_i\|}, \tag{29}$$

$$x_5 = m^{-1}\sum_i \frac{\|U^\top \Delta h_i\|^2}{\|\Delta h_i\|^2}, \qquad x_6 = m^{-1}\sum_i \frac{|\Delta r_i - g_i^\top \Delta h_i|}{|\Delta r_i| + |g_i^\top \Delta h_i| + 10^{-8}}. \tag{30}$$

The basis and scalar readout are unchanged from diagnostic calibration. These are finite control-response descriptors, not estimates of global controllability or certified operator gains. Probes are chosen deterministically from the original calibration set and use its known answers under teacher forcing. They generate no new behavior labels. Zero control, state, or gradient norms and nonfinite

Table 25: Development yields, equal weight per source pair. Uniform means the outer-training-fold choice, and may differ from the final all-source fit.

Model	Seconds	Selector	Loss	Uniform	Direct
Ministral	121.29	27.26	37.16	37.16	60.21
Granite	83.67	28.42	32.95	30.68	67.68
Qwen	377.01	25.25	32.25	32.25	43.12

features trigger the uniform source strategy, retaining already incurred costs. Every measured source pair has valid features.

Rule formation and endpoint isolation. A ridge model predicts the budget yield of each of the three fixed diagnostics. The rule selects the highest mean prediction across a group, breaking action ties in favor of loss. Endpoint-connected components are assigned to four outer folds. Each fold simulates an unseen group using one common action. Within its training partition, four-fold selection chooses regularization from $10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100$. Scaling, fitting, and the uniform-mean baseline use that training partition only. Invalid-feature fallback uses the same source-only uniform action. A fold is a development case made from existing prompts, not a newly trained independent group.

Cost and complete discoveries. Each diagnostic scores all 400 items before inspection. The budget is the median measured cost of a full loss scan followed by the first 80 loss-ranked paired checks on three preselected timing pairs per model. Every source pair receives that model's budget. All applicable probes, feature aggregation, selection, scans, transfer, sorting, generation and parsing are charged. Both endpoints must finish within budget for a discovery to count. Direct inspection skips scoring and uses fixed item-hash order. Model loading and prior training are common preparation costs. Research logging and reference-panel analysis are separate.

These are replays of measured operation costs on existing records, not new-group executions of complete policies. Caches receive no free cost. Full timing, CPU/GPU co-load, queue and accounting records are in the supplement. The seven budget points reuse primary-budget outer-fold choices without retuning. The measured wall times remain host-specific.

Choices and baseline overlap. The selector chooses loss in all four Ministral and Qwen folds and norm in all four Granite folds. The outer-fold uniform baseline chooses loss throughout Ministral and Qwen. Granite uses norm, norm, loss, loss. Thus the uniform baseline coincides with always-loss for Ministral and Qwen, but not for Granite. The final all-source uniform fit chooses loss in all three models. That fit is not used as a held-out reference.

Opportunity, regret, and overhead. For a held-out group, write $V^0(a)$ for mean discoveries with action a before selector overhead and $V^c(a)$ after overhead. For the selected action $\hat{a}$, the accounting identity

$$\max_a V^0(a) - V^c(\hat{a}) = \underbrace{\max_a V^0(a) - V^0(\hat{a})}_{\text{selection regret}} + \underbrace{V^0(\hat{a}) - V^c(\hat{a})}_{\text{overhead loss}} \tag{31}$$

separates choosing a diagnostic from paying to choose it. The maximum uses one common action for the group, not a different oracle for each pair. Here a ranges over the three diagnostics. Direct inspection is included in the broader per-pair upper bound in Section 5. It is not an action in this diagnostic-selection decomposition. This is standard selection accounting, not a new control theorem.

Development decision. Advancement required at least one model with nonconstant choices, group-oracle space above the strongest reference, and at least 5% outer-fold net improvement (one extra mean discovery if the reference were zero). No model satisfies these conditions. Selector deficits against the strongest reference are 54.72%, 58.01%, and 41.45% for Ministral, Granite, and Qwen. The new 384-prompt confirmation phase therefore remains unrun. Prepared seed and item manifests document its isolation without being counted as experimental observations. This development

result identifies a practical constraint on using response diagnostics for short-answer screening. The independent cohort-transfer conclusions come from the earlier panels.

Risk improvements and deteriorations both count as discoveries. Their separate counts depend on the recorded seed orientation, which is not an optimization trajectory. All pair yields, fold assignments, six control features, and signed counts are in the data supplement.

J.1 Measurement value and prediction error

Operational quantities and a testable decision. For an evaluation budget B in seconds, strategy s first incurs C_s for its diagnostic scan, data transfer, sorting, and any control probes and selection. It then inspects a fixed queue. A check is complete only after generating and parsing both prompt endpoints. Let $K_s(B)$ count complete checks, $Y_s(B)$ count absolute parsed-risk changes, and $p_s(B) = Y_s(B)/K_s(B)$ when $K_s > 0$. Improvements and deteriorations are retained separately and sum to Y_s. For direct hash-order inspection, use subscript 0. The identity

$$Y_s - Y_0 = K_s(p_s - p_0) - (K_0 - K_s)p_0 \tag{32}$$

separates the value of concentrating changes from the cost of reduced coverage. It holds for each realized pair and budget. We average these pairwise terms. Products of separately averaged precision and coverage need not obey it. If $K_s, K_0, p_0 > 0$, positive net yield is equivalent to $p_s/p_0 > K_0/K_s$. Zero checks or zero discoveries use the original counts. The coverage term is signed: a cheaper strategy may inspect more items.

This accounting equation becomes a prospective rule only when the quantities on its decision side are predicted before new behavior labels are available. A first decision may use training logs, control displacement, output settings, and source cost estimates to decide whether to buy control probes. A second decision may use the measured response features after paying their cost. Fallback to a source-fixed action retains sunk probe costs. These stages make the informational value of control measurement distinct from the value of cheap metadata. Feature acquisition and the virtual-best upper bound follow algorithm-selection methodology (Bischl et al., 2016).

Proposition 2 (Conditional robustness of a predicted net yield)**.** *For $a \in \{s, 0\}$ suppose $K_a, \widehat{K}_a \geq 0$, $0 \leq p_a, \widehat{p}_a \leq 1$, and nonnegative error bounds such that $|K_a - \widehat{K}_a| \leq \eta_a$, $|p_a - \widehat{p}_a| \leq \epsilon_a$ hold simultaneously. Define $\widehat{\Delta} = \widehat{K}_s\widehat{p}_s - \widehat{K}_0\widehat{p}_0$. Then*

$$|(Y_s - Y_0) - \widehat{\Delta}| \leq \eta_s + \widehat{K}_s\epsilon_s + \eta_0 + \widehat{K}_0\epsilon_0.$$

Consequently a predicted margin larger than this bound suffices for positive realized net yield under the stated simultaneous errors.

Proof. For each action, $|K_a p_a - \widehat{K}_a\widehat{p}_a| \leq |K_a - \widehat{K}_a|p_a + \widehat{K}_a|p_a - \widehat{p}_a| \leq \eta_a + \widehat{K}_a\epsilon_a$. Apply the triangle inequality to the two actions. □

This standard error calculation specifies what an independent usefulness claim must calibrate. The present source data establish an observed boundary. They supply no independently validated simultaneous prediction-error bound. Variable costs also make check counts discontinuous. For a fixed queue, upper bounds on setup and each paired check give a conservative complete prefix. Lower bounds give an optimistic prefix. A fixed mean token cost cannot replace this prefix calculation when early EOS and long continuations differ.

Connection to control theory. In the Pontryagin analysis of Huang and Song (2026), the control horizon T differs from the generation cap L and evaluation budget B used here. Its Green estimate (v3 Theorem 3.2) requires hyperbolicity, exponential dichotomy, and invertible scaled boundary matrices. Local reconstruction (Theorem 4.1) also requires smooth maps and a contraction region. The a posteriori result (HTML Proposition 5.3, PDF Proposition 5.2) uses full interior and boundary residuals together with inverse-Jacobian and derivative bounds. Terminal sensitivity (HTML Theorem 5.9, PDF Theorem 5.5) assumes the corresponding stationary branches and boundary conditions. Our finite-pair response gains and scalar readout residuals are empirical measurements, not verification of these dynamical assumptions. Appendix F.1 states the additional score-margin and label conditions needed for a ranking bound.

Source conditions and opportunity. All 54 source pairs retain their fixed diagnostic values across three parsers and six generation caps. The accompanying condition matrix reports all results at 80 paired checks, separating old and new prompt groups. The complete per-pair table also records hit rate, signed change counts, and exact label-equivalence identifiers. Strict risk labels agree at $L = 8, 16, 32$ in all three models. Qwen's new endpoints all end at two tokens. Only one of Ministral's 12,800 new endpoints exceeds eight tokens. Granite has 2,282 such endpoints, which motivates the bounded timing comparison.

Strict $L = 8$ endpoint costs were measured on the full 400-item panel. Budget outcomes replay this historical ledger. At its reference budget neither state score exceeds direct inspection in any of the 54 pairs. Over seven specified budgets, even subsequently choosing the best option for each pair from direct inspection and all three diagnostics at no selection cost improves the best model mean over direct inspection by at most 1.55%. This permissive observed upper bound already includes choices unavailable to a prospective selector. Other full condition cells report equal-check outcomes, with their generation costs marked unmeasured rather than extrapolated from token counts.

Bounded Granite timing and parser-preserving stopping. The three existing pairs use the first 32 common-panel items in fixed hash order. Native generation at $L = 8, 16, 32$ and a strict-parser early-stop implementation are each executed three times in balanced phase order. Each score is also recomputed through independent forwards in every repetition. Input preparation, transfer, native computation and operational stopping consume the measurement clock. Research-only trace rendering is outside it. Parser time is recorded separately. The repeated 32-item micro-panel uses $B = 6.69$ seconds, defined as $32/400$ of the original source reference budget. This scales the budget definition, not any measured token cost.

The early stop accepts only an irrevocable non-whitespace ASCII violation under the verified ByteLevel decoder with cleanup disabled. A later UTF-8 continuation cannot remove that ASCII witness. EOS and the native cap otherwise remain unchanged. All 1,728 early-mode endpoint traces preserve the full native strict answer or invalid status, and all native prefixes match the retained reference. Of these, 297 endpoints actually trigger irreversible early stopping. The early implementation is shared by every strict-parser strategy, including direct inspection. Leading and terminal parsers retain full native traces.

Across the repeated native and early conditions, generation and strict parsing require 469.46 and 216.95 seconds, respectively (53.8% aggregate reduction). Detailed timings, all repetitions, waiting and process costs remain in the cost supplement.

The full cost replay retains each measured paired completion time, and charges complete scans, sorting, and the original actual 32-probe ledger where applicable. All seven budgets, all three parsers for native generation, and strict early stopping appear in the delivered per-pair table. Across all seven budgets and three caps, the strict-parser free per-pair/repetition oracle adds zero discoveries over direct inspection in both modes. Over the native leading/terminal conditions, its largest relative gain over the strongest fixed mean is 1.61%, below the specified 5% entry margin. The micro-panel provides runtime sensitivity on existing pairs. It is not a new-group efficiency confirmation.

Reproducible decision accounting. The supplied evaluation artifact replays fixed scores, paired risk labels and complete operation costs as historical observations. A separate interface accepts forecasts locked before evaluation labels and checks their model, prompt-pair, decode-cap, and budget scope. Before-probe decisions use cheap information. After-probe decisions may use control features and retain their costs. A positive predicted margin identifies a candidate use condition. An efficiency claim still requires the planned new-group validation against cheap-information rules and direct inspection.

K Readout dependence and batched execution

K.1 A fixed-coefficient readout comparison on fresh questions

The supplement reuses 37 Ministral and 38 Granite prompts: three original fixed pairs and 16 disjoint validation pairs per model, with the original Ministral shared endpoint retained. The 400 new questions use the same data snapshot and deterministic ordering, with source-coordinate and

normalized-text exclusion against the 1,692 previously used questions. The native generation protocol is greedy, at most eight tokens, with the strict single-letter parser. Each model has 400 unprompted reference states. The resulting 14,800 Ministral and 15,200 Granite prompted endpoints have stored states, native tokens, losses and execution receipts.

For each question, the reference choice is the unprompted model's largest conditional A–E logit, breaking ties by the smallest token ID. This is a choice among the five answer symbols, not the unrestricted vocabulary winner. The reference readout uses Eq. (5) with that fixed choice. The entropy readout is $-\sum_{a\in\{A,\dots,E\}} p_a \log p_a$, where p is the conditional five-choice softmax. All three readouts use the same state and FP32 head rows. Their scalar values and gradients are reconstructed in FP64. The alternative gradients replace only the first two features in the original state/readout score. This fixed-coefficient comparison tests sensitivity to the measurement readout. Separate calibration of the alternatives would answer a different prediction question.

State context. Ministral uses the retained pre-answer prefix path. Granite's state is read before a constant dummy A in the complete answer template, keeping the finite-precision context consistent with the retained production path. The dummy does not depend on the gold answer. Native generation continues to use its original prefix. Before any fresh Granite behavior was measured, four old probes and all five answer substitutions passed the original 10^{-4} state/readout checks. The maximum state difference was zero, gradient error 8.04×10^{-9}, and scalar error 3.59×10^{-6}. Prompted states and their same-question reference states bind the full token context. This check authenticates the measurement context. It does not attribute earlier numerical differences to a particular kernel.

Statistical target. For readout $a \in \{\text{reference}, \text{entropy}\}$, the interaction is

$$I_a = \text{mean}_{j\in N}[\text{AUC}(s_a, z_j) - \text{AUC}(s_{\text{gold}}, z_j)] - \text{mean}_{j\in O}[\text{AUC}(s_a, z_j) - \text{AUC}(s_{\text{gold}}, z_j)].$$

The three old pairs remain fixed references. The 16 new pairs and common question weights are resampled jointly across readouts. All four comparisons retain 20,000 draws and the original 22-comparison family, including comparisons omitted from the reduced experimental scope. Table 3 gives every interval. All common pairs and draws are valid. Complete per-pair AUCs and deletion checks are supplied in `data/reviewer_supplement/readout/interactions.json`. This panel supplies new-question evidence for existing prompts.

Table 26: Equal-pair mean state/readout AUCs on the 400-question readout-substitution panel, with all calibration parameters fixed. These descriptive estimates supplement, rather than replace, the corrected interaction intervals.

Model	Group	Gold	Reference	Entropy
Ministral	Old (3 pairs)	0.517	0.522	0.510
Ministral	New (16 pairs)	0.459	0.469	0.393
Granite	Old (3 pairs)	0.725	0.713	0.722
Granite	New (16 pairs)	0.533	0.537	0.538

K.2 What a readout perturbation bound certifies

Let $\alpha = \theta_1/\tilde{\sigma}_1$ and $\beta = \theta_2/\tilde{\sigma}_2$ be the fixed standardized coefficients of the two gradient-dependent features. With δg_i the change in readout gradient, set

$$v_i = \alpha U U^\top \Delta h_i + \beta(I - UU^\top)\Delta h_i.$$

Holding the state, subspace and remaining features fixed gives the exact score change and its Cauchy–Schwarz bound,

$$s_{a,i} - s_{\text{gold},i} = \delta g_i^\top v_i, \qquad |s_{a,i} - s_{\text{gold},i}| \le E_i := \|\delta g_i\|\,\|v_i\|.$$

Thus a gold-score margin larger than $E_i + E_j$ preserves the ordering of items i, j. This is a standard sensitivity calculation and supplies the score envelope required by Appendix F.1.

All 30,400 stored score changes satisfy the bound. Among the 1,516,200 within-pair item orderings per model, the reference substitution certifies 700,264 for Ministral (46.19%) and 857,851 for

Granite (56.58%). Every certified ordering is preserved. Coverage has a specific source: all Ministral certificates and all but one Granite certificate have $E_i + E_j = 0$, so they compare items whose readout scores are unchanged. Entropy certifies no orderings. These counts distinguish exact invariance from a useful positive-error certificate: nearly all observed coverage comes from the former. Complete counts, including the positive-bound subset, accompany the data.

K.3 Batching changes cost and numerical agreement

The fixed Granite pairs 2101–2114, 7101–7102 and 7131–7132 use the same hash-ordered 32-question queue at batch sizes 1, 4 and 8, with three balanced repetitions. Both native generation and strict-parser-preserving early stopping are measured. This gives 54 generation conditions and 3,456 endpoint executions. The three diagnostic scans give 81 scoring conditions. Original batch-one token and scoring checks precede timing. Each timing includes input rendering, transfer and the operation itself. Full process costs separately include loading, warmup, serialization and all attempted work. In this implementation, template-consistency checks are included in the measured preprocessing cost.

Table 27: Granite mean primitive seconds on the fixed 32-item queue. Each cell averages three pairs and three repetitions. Generation checks both prompt endpoints. Each scoring column computes one diagnostic for the same 32 paired questions. The last column divides state/readout scoring by early-stop checking. These are operation costs.

Batch	Native	Early stop	Loss	State/readout	State norm	Ratio
1	12.236	9.305	5.626	5.610	5.086	0.603
4	4.682	3.545	3.044	2.837	2.440	0.800
8	3.067	2.420	2.418	2.297	1.901	0.949

All 1,728 early-stop/native endpoint comparisons preserve the native prefix and strict parsed answer. Across batch sizes, 51 of 1,152 native endpoint comparisons differ in their tokens. The 54 diagnostic-order comparisons all differ somewhere in the complete ranking. Top-eight overlap is seven or eight. These measurements establish output and ranking sensitivity within the observed pipeline. They do not identify a kernel-level cause, and token differences alone do not count risk changes. The saved batch scores and preflight receipts were checked. Batch hidden states were not retained for a second score reconstruction.

Relative operation costs Relative to paired checking with the same parser-preserving early stopping, state/readout scoring costs 0.603, 0.800, and 0.949 times as much at batch sizes 1, 4, and 8. Each timing mean uses nine conditions on the fixed 32-question queue. The reported ratios divide the corresponding timing means. These relative operation costs describe the measured pipeline. They are not a measured 400-question policy, a cross-hardware normalization, or a prospective efficiency result.

Verification scope. The full cost ledger, including failed work, is retained separately from the successful primitive times. CPU replay reconstructs fixed readout scores and native parsing. A separate pairwise AUC implementation agrees with all four 20,000-draw estimates and intervals within 10^{-12}.

Scope of diagnostic comparisons. The norm and random-subspace controls share our paired target. GRADE-pre and HARP instead predict answerability and answer-level hallucination, respectively (Wang et al., 2026; Hu et al., 2025). A matched comparison must align targets, supervision, state context, and measurement cost. No head-to-head benchmark of these systems is reported here. Symbol-binding research studies the association between option content and answer symbols (Xue et al., 2024). Our stopping intervention preserves the first symbol and measures its continuation's effect on parsed risk. RKSP distinguishes spectral prediction from intervention tests (Kim et al., 2026). Smooth Pontryagin sensitivity results (Huang and Song, 2026) motivate explicit assumptions, but do not themselves certify the discrete behavior studied here.

L CLAIM–EVIDENCE–DATA INDEX

Section links below navigate this PDF. Paths are relative to the extracted paper directory: **E** denotes `extended_materials/` and **D** denotes `data/`. Undefined values remain undefined. None is replaced by a chance AUC.

Claim and PDF evidence	Numerical records or fixed protocol
Training and calibration Appendices A, D, I	E/`protocol_evidence/MANIFEST.json`, E/`editorial_data/training_pairs.csv`.
Three-model ECQA transfer Appendix D	E/`table_data/independent_ecqa/`: `primary_intervals.csv`, `pair_conditions.csv`, `condition_matrix.csv`, `leave_one_pair_out.csv`.
Crossed effects and fresh-question prediction Appendix E.1	E/`editorial_data/`: `crossed_all_intervals.csv`, `crossed_all_pair_predictions.csv`, `crossed_all_sensitivity.csv`, E/`table_data/crossed_ecqa/prediction_lock.json`.
Unseen GSM8K prompts Appendix I.1	D/`gsm8k/pair_predictions.csv`, `gsm8k/mae_improvements.csv`, `gsm8k/independent_replay.json`.
Control response and stability Appendices F, G	E/`table_data/crossed_ecqa/all_pair_control_respo` `nses.csv`, D/`finite_panel_stability_audit.json`, `endpoint_observability.json`.
Readout dependence Appendix K	D/`reviewer_supplement/readout/`: `interactions.json`, `ranking_stability.json`, `fixed_scores.csv`.
Measurement value and batching Appendices J, J.1, K.3	E/`table_data/measurement_value/`: `source_cost_pairs.csv`, `retiming_cost_pairs.csv`, D/`reviewer_supplement/batch_primitives.csv`.
Detailed and exploratory support Appendices D.1, H	E/`SUPPLEMENT_INDEX.md` locates full parser matrices, MATH records, channel reweighting, omitted duplicate figures and archived detail text.
Shared statistical replay	`../bootstrap_replay/README.md`, `../bootstrap_replay/run.py`, `../bootstrap_replay/inputs/`. Saved scores, labels and fixed forecasts reproduce 52 primary intervals using 20,000 draws and the original 18/24/22 families.

E/`EVIDENCE_INDEX.json` lists the exact files. `build_compact_appendix.py` checks every path and rebuilds the pair tables. Full reproduction instructions are in **E/**`SUPPLEMENT_INDEX.md`. Raw tensors, tokens and checkpoints are excluded. **D/**`delivery_index.json` identifies separately retained archives and historical replay receipts, not files in this ZIP.