SPECTRA-RSI: Counterfactual Spectral Sketching and Anytime-Valid Gating for Modular Recursive Self-Improvement

Andrew Kiruluta
UC Berkeley, School of Information

August 21, 2026

Abstract

Recursive self-improvement (RSI) is often described as a data-generation or optimization problem, yet every improvement cycle also creates a harder statistical question: which capabilities changed, which module caused the change, and whether the candidate can be accepted without accumulating hidden regressions. Contemporary efficient-evaluation methods reduce benchmark cost by selecting informative items or predicting static full-benchmark scores, while recent compressed-sensing work localizes capabilities in a fixed network through component knockouts. Neither setting directly addresses attribution of a *candidate update* under repeated, adaptively monitored model changes.

We introduce SPECTRA-RSI, a measurement and control layer for modular RSI based on *counterfactual spectral sketching*. Small signed interventions on low-rank experts are used to learn an interventional capability-response dictionary whose atoms are causal local fingerprints of expert updates. After a candidate is frozen, paired randomized sketches compare the base and candidate on common prompts and decoding randomness, and a router-informed structured estimator recovers the candidate's capability delta and expert support. The dictionary is transported across iterations and refreshed when residual confidence sequences detect basis drift. Discovery is separated from acceptance: bootstrap or debiased intervals support diagnosis, whereas protected anchor streams are evaluated by anytime-valid e-processes, allowing continuous monitoring and data-dependent stopping while controlling the probability of falsely accepting a materially regressive update.

The resulting novelty is the closed-loop coupling of (i) interventional expert fingerprints, (ii) realizable paired sketches of behavioral *deltas*, (iii) drift-aware structured recovery, and (iv) an optional-stopping-valid modular gate. We establish non-identifiability without structural assumptions, unbiasedness and variance reduction for paired sketches, a stable recovery bound with explicit compressibility, basis-drift, and nonlinear-response terms, and a finite-sample false-acceptance guarantee for the gate. We provide an end-to-end algorithm and a falsifiable experimental protocol; numerical gains are hypotheses to be tested rather than claimed results. The framework reframes RSI as repeated causal system identification with a protected statistical reference channel. Implementation repo can be found here: https://github.com/andrew-jeremy/compressed-sense-spectal-recursive-self-improvement

1 Introduction

Modern language-model improvement loops increasingly rely on synthetic data, self-critique, model-based judging, preference optimization, tool-mediated verification, and repeated post-training. These methods can improve a model without requiring a proportionate increase in human labeling, but they also create a central measurement problem: after a broad update, the system must determine which capabilities improved, which regressed, which changes are artifacts of the evaluator, and which component should receive credit or blame. A single average benchmark score is insufficient because it

can conceal offsetting gains and losses, while dense evaluation over every task, domain, difficulty level, demographic slice, and safety condition can dominate the cost of each iteration.

This paper develops a measurement-centered formulation of RSI. Rather than asking only how a model should generate its next training set, we ask how an improvement loop can estimate a high-dimensional capability change from a limited evaluation budget. The key observation is that model updates are often lower-dimensional than the space of all measured behaviors. Low-rank adaptation methods exploit rank deficiency in parameter updates [10]; empirical curvature spectra are often highly anisotropic; and mixture-of-experts (MoE) systems already impose conditional modularity through sparse routing [9, 15]. These properties suggest that per-iteration behavioral changes may be compressible in an appropriately learned basis even when raw benchmark vectors are not sparse.

Compressed sensing provides a mathematical language for recovering sparse or structured-compressible signals from fewer linear measurements than ambient dimension [6, 7, 4]. A direct transfer to model evaluation, however, is not valid without addressing four obstacles. First, benchmark scores are not arbitrary linear sensors; they arise from finite collections of prompts, stochastic generations, and nonlinear metrics. Second, capability coordinates are not fixed physical quantities and may drift as the model changes. Third, the sensing basis and the model update can co-adapt, creating a Goodhart channel. Fourth, safe RSI requires uncertainty estimates and regression guarantees, not merely a point reconstruction.

Spectra-RSI addresses these obstacles by introducing a local capability-response model and by separating *discovery measurements* from *acceptance measurements*. Discovery uses randomized composite evaluations formed from task-level loss differences or bounded scoring statistics. These measurements are designed after the candidate update is frozen. A spectral dictionary groups latent change directions into bands linked to low-rank experts. Router load, gate entropy, and expert-gradient signatures define a prior over active bands, which enters a weighted group-lasso or model-based recovery problem. Acceptance then uses a small, untouched anchor suite that is never used to fit the reconstruction. This separation is essential: compressed measurements propose where change occurred, while anchors decide whether an update may enter the next model generation.

1.1 Contributions and claims of novelty

The paper makes five claims that are narrower, more testable, and more distinct from the current state of the art than a generic claim of "compressed evaluation."

C1. Counterfactual capability-response dictionaries. We define spectral atoms from signed expert interventions, so each atom estimates a local causal response of capability slices to a realizable update direction. This differs from static benchmark compression and from identifying components that support an already-trained model.

C2. Paired delta sketching. We construct randomized measurements from paired base/candidate outcomes on the same item and, where possible, the same decoding randomness. The sketches are unbiased for linear capability aggregates and can have substantially lower variance than two independent evaluations.

C3. Drift-aware structured attribution. We combine group-sparse recovery, router-derived support priors, exploration floors, and dictionary transport. A residual process detects when the local response basis or linear model has ceased to be credible and triggers denser evaluation or dictionary refresh.

C4. Anytime-valid modular acceptance. We separate exploratory sensing from protected anchor streams and use e-processes or confidence sequences for non-inferiority tests. The resulting gate remains valid under repeated peeking, adaptive sample sizes, and early stopping, and can accept, reject, quarantine, or partially roll back individual experts.

C5. A falsifiable RSI evaluation benchmark. We specify controlled expert interventions, dense

ground-truth deltas, modern efficient-evaluation baselines, causal-attribution metrics, optional-stopping stress tests, and long-horizon safety endpoints.

The paper does not claim that behavioral change is always sparse, that compressed sketches replace dense catastrophic-risk evaluation, or that the proposed loop constitutes unconstrained autonomous RSI. These are empirical assumptions and governance boundaries to be tested explicitly.

2 Related Work

Self-improvement as a closed loop. Self-training, self-rewarding models, constitutional feedback, verifier-guided search, and iterative preference optimization automate parts of data acquisition, evaluation, or optimization [2, 19, 11]. Recent system-level treatments describe self-improvement as a lifecycle spanning data generation, selection, model update, inference refinement, and autonomous evaluation [28]. SPECTRA-RSI addresses the autonomous-evaluation and update-control layer: it is agnostic to how a candidate is proposed, but it requires that the candidate be frozen before protected measurements are sampled.

Efficient and adaptive language-model evaluation. TinyBenchmarks, metabench, anchor-point methods, active evaluation acquisition, amortized model-based evaluation, and Fluid Benchmarking reduce evaluation cost through item-response models, representative subsets, learned acquisition policies, difficulty prediction, or adaptive item selection [14, 13, 16, 24, 27, 22]. Feature-selection and regression methods show that static full-score prediction can often be formulated as supervised subset selection [5]. These methods primarily estimate a model's static score or rank from fewer items. Our target is different: a signed, high-dimensional *change vector* between a base and a candidate, together with attribution to update modules and a decision about whether that update may enter the next iteration. The distinction matters because two candidates with the same aggregate score can have opposite localized regressions.

Capability localization and sparse model structure. Compressed sensing has recently been used to localize capabilities to sparse sets of attention heads by applying randomized knockouts and regressing task degradation on ablated components [3]. This provides important evidence that some capabilities are compressibly associated with network components. The present work reverses the causal direction and changes the estimand: instead of asking which static components support a capability, we ask which proposed update directions caused an observed vector of capability changes. Signed local interventions build the response dictionary before recovery; the recovered coefficients describe candidate-update effects rather than static component importance.

Low-rank and mixture-of-expert adaptation. LoRA exposes low-dimensional update coordinates, while sparse MoE models and mixtures of low-rank experts provide conditional modularity [10, 15, 9, 20, 30, 26]. Current MoE-LoRA work focuses on routing, specialization, and parameter efficiency. SPECTRA-RSI uses those modules as intervention and rollback units and treats routing statistics only as fallible prior information for an external behavioral inverse problem.

Sequential validity, uncertainty, and repeated evaluation. Bootstrap and conformal methods are useful for descriptive uncertainty, but repeated benchmark inspection and adaptive stopping can invalidate fixed-sample error bars. Confidence sequences and e-processes provide evidence measures that remain valid at arbitrary stopping times [23, 25]. Adaptive learn-then-test demonstrates how e-processes can reduce testing rounds while preserving finite-sample risk control in model selection [29]. Small-sample LLM evaluations also require care because Gaussian or central-limit approximations

can underestimate uncertainty [21]. We therefore reserve protected anchor data for anytime-valid acceptance and use reconstructed intervals only for diagnosis and allocation.

Positioning. The closest adjacent methods optimize one of four objectives: predict a static benchmark, localize a static component, adapt a modular model, or sequentially certify a candidate. The proposed contribution is a single closed loop in which interventional response learning, paired sketches, structured update attribution, drift detection, and anytime-valid modular acceptance share a common estimand and audit trail.

3 Problem Formulation

Let f_{θ_t} denote the deployed model at iteration t and let $f_{\theta_t+\delta\theta_t}$ be a frozen candidate produced by any training, editing, self-training, or search procedure. Let $q_j(z,\omega;f) \in [a_j, b_j]$ be a bounded score for item z in capability slice $j \in \{1,\dots,n\}$ under generation randomness ω. Larger values are defined to be better. The population capability delta is

$$\Delta c_{t,j} = \mathbb{E}_{z\sim\mathcal{D}_{t,j},\omega}[q_j(z,\omega;f_{\theta_t+\delta\theta_t}) - q_j(z,\omega;f_{\theta_t})]\,. \tag{1}$$

The vector $\Delta c_t \in \mathbb{R}^n$ is the primary estimand. Dense evaluation estimates every coordinate; the proposed method seeks an accurate and safety-relevant reconstruction under a smaller evaluated-item or token budget.

3.1 Paired randomized sketches

For sampled item $z_{ij\ell}$ and common random seed $\omega_{ij\ell}$, define the paired difference

$$d_{ij\ell} = q_j(z_{ij\ell},\omega_{ij\ell};f_{\theta_t+\delta\theta_t}) - q_j(z_{ij\ell},\omega_{ij\ell};f_{\theta_t}). \tag{2}$$

A row $a_i \in \mathbb{R}^n$ of a sensing matrix A_t yields the realizable composite statistic

$$y_i = \sum_{j=1}^{n} a_{ij}\widehat{\Delta c}_{ij}, \qquad \widehat{\Delta c}_{ij} = \frac{1}{N_{ij}}\sum_{\ell=1}^{N_{ij}} d_{ij\ell}. \tag{3}$$

This is a weighted sum of task-level outcomes computed after model execution, not a synthetic linear combination of prompts. Evaluation cost is

$$B_{\text{total}} = B_{\text{probe}} + B_{\text{sense}} + B_{\text{anchor}} + B_{\text{audit}}, \tag{4}$$

with a separate accounting for base/candidate tokens, judge calls, and wall-clock time.

Proposition 1 (Unbiased paired sketch and variance advantage)**.** *If items are sampled from $\mathcal{D}_{t,j}$ and the paired score differences have finite variance, then $\mathbb{E}[y_i] = a_i^\top \Delta c_t$. For a single slice, using common randomness gives*

$$\operatorname{Var}(q_j^{\text{cand}} - q_j^{\text{base}}) = \operatorname{Var}(q_j^{\text{cand}}) + \operatorname{Var}(q_j^{\text{base}}) - 2\operatorname{Cov}(q_j^{\text{cand}}, q_j^{\text{base}}). \tag{5}$$

Thus paired evaluation weakly improves over independent randomness whenever the covariance is nonnegative, and the gain can be large when base and candidate outputs remain strongly coupled.

Proof. Linearity of expectation gives $\mathbb{E}[\widehat{\Delta c}_{ij}] = \Delta c_{t,j}$ and hence $\mathbb{E}[y_i] = a_i^\top \Delta c_t$. The variance identity follows from $\operatorname{Var}(X-Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) - 2\operatorname{Cov}(X,Y)$. Independent evaluation sets the covariance term to zero. □

3.2 Structural estimand and non-identifiability

Let $\Psi_t \in \mathbb{R}^{n\times p}$ be a capability-response dictionary and suppose

$$\Delta c_t = \Psi_t x_t + r_t, \tag{6}$$

where x_t is group-compressible and r_t is dictionary approximation error. Let ρ_t denote local nonlinear-response error, defined formally in Section 4. The sensing model is

$$y_t = A_t \Psi_t x_t + A_t(r_t + \rho_t) + \epsilon_t, \tag{7}$$

Proposition 2 (No-free-lunch for compressed delta evaluation)**.** *If $m < n$ and no structural restriction is imposed on Δc_t, then for any sensing matrix $A_t \in \mathbb{R}^{m\times n}$ there exist distinct capability deltas Δc and $\Delta c'$ with $A_t\Delta c = A_t\Delta c'$. Consequently, localized improvement and regression are not identifiable from compressed aggregate measurements without a dictionary, sparsity model, distributional prior, or additional interventions.*

Proof. Because $m < n$, the null space of A_t contains a nonzero vector v. For any Δc, set $\Delta c' = \Delta c + v$. Then $\Delta c' \neq \Delta c$ but $A_t\Delta c' = A_t\Delta c$. Choosing the signs and scale of v can alter localized gains and regressions without changing the observed sketch. □

This proposition clarifies the role of the spectral model: it is not merely a computational convenience but an identifying assumption. Failure of compressibility, conditioning, or local linearity must trigger a larger budget rather than a forced sparse explanation.

3.3 Evaluation objective

The system is judged at fixed evaluated-token or item cost by: reconstruction error of Δc_t; recall of materially negative coordinates; attribution accuracy for perturbed experts; interval or confidence-sequence validity; and the probability of falsely accepting a candidate with at least one protected regression. Claims such as "bits per evaluation" are avoided because a scalar score can contain variable information and can itself require many model calls.

4 Counterfactual Spectral Response Geometry

4.1 Local behavioral Jacobian

Let

$$c_j(\theta) = \mathbb{E}[q_j(z, \omega; f_\theta)], \qquad c(\theta) = (c_1(\theta), \ldots, c_n(\theta))^\top. \tag{8}$$

For a sufficiently small candidate update,

$$\Delta c_t = J_t \delta\theta_t + \rho_t, \qquad J_t = \nabla_\theta c(\theta_t), \tag{9}$$

where ρ_t is a nonlinear remainder. Directly forming J_t is infeasible and, for non-differentiable scores, undefined. We instead estimate its action on realizable modular update directions.

4.2 Signed expert probes

Let $v_{t,e,r}$ be a normalized update direction for rank component r of expert e. Before evaluating the candidate, apply small positive and negative probes of magnitude h to the base model and define the central-difference fingerprint

$$h_{t,e,r} = \frac{c(\theta_t + h v_{t,e,r}) - c(\theta_t - h v_{t,e,r})}{2h}. \tag{10}$$

Stacking fingerprints gives the interventional response matrix

$$H_t = [h_{t,1,1}, \ldots, h_{t,E,R_E}] \in \mathbb{R}^{n \times d}. \tag{11}$$

These columns are causal local responses to controlled update directions. Historical deltas, task-gradient products, and Fisher information may regularize H_t, but they do not replace the intervention when causal attribution is required.

Within expert e, compute a rank-r_e factorization $H_{t,e} \approx U_{t,e}\Sigma_{t,e}V_{t,e}^\top$. The global dictionary is

$$\Psi_t = [U_{t,1}, \ldots, U_{t,E}], \tag{12}$$

with group G_e containing the coordinates of expert e. If the candidate update has expert coefficients β_t, then

$$\Delta c_t = H_t\beta_t + \rho_t + \zeta_t = \Psi_t x_t + r_t + \rho_t, \tag{13}$$

where ζ_t captures probe estimation error and r_t captures low-rank truncation or unmodeled directions.

Assumption 1 (Local response regularity)**.** *There exists $L_t > 0$ such that for all accepted probes and candidate sub-updates in a trust region $\|\delta\theta\|_2 \le \eta_t$,*

$$\|\rho_t\|_2 \le \frac{L_t}{2}\|\delta\theta_t\|_2^2. \tag{14}$$

A pilot set tests this assumption by comparing probe-superposition predictions with observed candidate deltas. Large residuals trigger candidate decomposition, selected interaction probes, or dense evaluation.

4.3 Dictionary transport and drift

Because capabilities, tasks, routers, and experts change over time, Ψ_t cannot be treated as fixed. Let Q_t solve the orthogonal Procrustes alignment

$$Q_t = \operatorname*{arg\,min}_{Q^\top Q = I} \|\Psi_t - \Psi_{t-1}Q\|_F. \tag{15}$$

Aligned coordinates permit longitudinal support tracking, while principal-angle and residual statistics quantify basis change. Define a protected pilot residual

$$e_{t,\ell} = a_{t,\ell}^\top \widehat{\Delta c}_{t,\ell} - a_{t,\ell}^\top \Psi_t \hat{x}_t. \tag{16}$$

A confidence sequence or e-process for excess residual risk monitors whether the deployed dictionary remains calibrated. A threshold crossing initiates signed re-probing, dictionary expansion, or a dense fallback. Thus drift is treated as a sequentially detected model violation rather than a fixed refresh interval.

Remark 1 (Relation to static component localization)**.** *A knockout study estimates the importance of existing components for a static capability. Equation (10) estimates the derivative-like behavioral effect of an admissible update direction. The former answers "where is a capability implemented?"; the latter answers "which update caused this capability change, and can that update be rolled back?"*

5 Spectral Mixture of Low-Rank Experts

We partition the p dictionary coordinates into E groups $\mathcal{G} = \{G_1, \dots, G_E\}$. Group boundaries may be contiguous in singular-value order, clustered by task-loading similarity, or learned through graph partitioning of the response covariance. Expert e owns low-rank update coordinates $u^{(e)}$ whose behavioral image lies primarily in $\mathrm{span}(\Psi_{t,G_e})$.

For input z, the router emits probabilities $\pi_e(z)$. Over the candidate-generation corpus, define the occupancy statistics

$$\bar{\pi}_e = \mathbb{E}_z[\pi_e(z)], \qquad H_e = -\mathbb{E}_z[\pi_e(z) \log \pi_e(z)], \tag{17}$$

and an update-energy statistic g_e from expert gradient norms or adapter-factor changes. These values induce a prior activity score

$$p_e = \sigma(\alpha_0 + \alpha_1 \log(\bar{\pi}_e + \varepsilon) + \alpha_2 g_e - \alpha_3 H_e), \tag{18}$$

where coefficients are calibrated only on previous, fully evaluated iterations. Router statistics are not treated as evidence of improvement. We clip p_e to $[p_{\min}, p_{\max}]$ and impose an exploration floor so that every group retains nonzero measurement probability. Disagreement between routed support and behavioral support is itself logged as a diagnostic of router gaming or expert entanglement.

5.1 Weighted group-sparse recovery

Let $w_e = (p_e + \epsilon_w)^{-\gamma}$ with $\gamma \in [0, 1]$. We estimate

$$\hat{x}_t = \arg\min_x \frac{1}{2} \|y_t - A_t \Psi_t x\|_2^2 + \lambda_1 \|x\|_1 + \lambda_g \sum_{e=1}^{E} w_e \|x_{G_e}\|_2 . \tag{19}$$

The elementwise term captures within-band sparsity; the group term favors a small number of active experts. A debiased least-squares fit on the selected support is used for effect-size estimation.

6 Measurement Design

6.1 Composite evaluations

Rows of A_t are constructed from signed, variance-normalized task weights. A practical design is

$$a_{ij} = \frac{s_{ij} \xi_{ij} / \hat{\sigma}_j}{\sqrt{\sum_{\ell=1}^{n} s_{i\ell} \xi_{i\ell}^2 / \hat{\sigma}_\ell^2}}, \tag{20}$$

where $s_{ij} \sim \mathrm{Bernoulli}(d/n)$ controls row sparsity, $\xi_{ij} \in \{-1, +1\}$ is Rademacher, and $\hat{\sigma}_j^2$ estimates task-score variance. Sparse rows simplify implementation and prevent a measurement from spreading its item budget too thinly across all tasks.

6.2 Two-stage adaptive design

Stage I uses a coarse design across all groups to identify likely active bands. Stage II concentrates measurements within nominated groups but reserves an exploration fraction $\varepsilon_{\text{explore}}$ for non-nominated groups. To avoid adaptive overfitting, the candidate model is frozen before measurement matrices are sampled, and matrices are not exposed to the candidate-generation policy.

6.3 Item allocation

A measurement row specifies weights over capability slices, while the finite item budget determines how accurately the corresponding aggregate is estimated. Neyman allocation assigns more samples to high-variance slices:

$$N_{ij} \propto |a_{ij}|\hat{\sigma}_j. \tag{21}$$

Repeated seeds or paired decoding randomness can reduce variance when comparing base and candidate models.

7 Recovery Guarantees

Let $\mathcal{M}_{s,k}$ denote vectors supported on at most s groups with at most k nonzero coordinates in total. Let $\sigma_{\mathcal{M}}(x)_1$ be the best structured approximation error in ℓ_1.

Assumption 2 (Model-restricted isometry)**.** *The effective sensing matrix $M_t = A_t\Psi_t$ satisfies, for every v in the difference set $\mathcal{M}_{2s,2k}$,*

$$(1-\delta)\,\|v\|_2^2 \le \|M_t v\|_2^2 \le (1+\delta)\,\|v\|_2^2 \tag{22}$$

for a sufficiently small δ.

Theorem 1 (Stable local recovery)**.** *Suppose Eq. (7) holds, M_t satisfies the model-restricted isometry assumption, and the total perturbation obeys*

$$\|\epsilon_t + A_t r_t + A_t \rho_t\|_2 \le \varepsilon_t. \tag{23}$$

Then a model-based basis-pursuit or appropriately tuned group-sparse estimator admits constants $C_1, C_2 > 0$ such that

$$\|\hat{x}_t - x_t\|_2 \le C_1 \frac{\sigma_{\mathcal{M}}(x_t)_1}{\sqrt{k}} + C_2\varepsilon_t. \tag{24}$$

Consequently,

$$\|\Psi_t\hat{x}_t - \Delta c_t\|_2 \le \|\Psi_t\|_{2\to 2}\left(C_1\frac{\sigma_{\mathcal{M}}(x_t)_1}{\sqrt{k}} + C_2\varepsilon_t\right) + \|r_t\|_2\,. \tag{25}$$

Proof sketch. Under the model-restricted isometry condition, standard model-based compressed-sensing arguments establish robust instance optimality for the structured feasible set: the estimation error is bounded by the best (s,k)-structured approximation error plus the norm of the effective perturbation. Applying the operator norm of Ψ_t and then adding the approximation residual r_t yields Eq. (25). The constants depend on the restricted-isometry level and on the precise recovery program. □

Interpretation. The bound makes clear that sparse recovery cannot compensate for a poor dictionary or a large nonlinear update. Performance degrades through four distinct channels: noise in finite evaluation, non-sparse latent change, dictionary approximation error, and local-linearization error. Each term is measurable or upper-bounded empirically.

Proposition 3 (Effect of basis drift)**.** *Let the true response dictionary be $\Psi_t^\star$ and the deployed dictionary be Ψ_t, with $\|\Psi_t - \Psi_t^\star\|_{2\to 2} \le \delta_{\Psi,t}$. Then basis mismatch contributes at most*

$$\|A_t(\Psi_t - \Psi_t^\star)x_t\|_2 \le \|A_t\|_{2\to 2}\,\delta_{\Psi,t}\,\|x_t\|_2 \tag{26}$$

to the perturbation term in Theorem 1.

Proof. Apply submultiplicativity of the spectral norm:

$$\|A_t(\Psi_t - \Psi_t^\star)x_t\|_2 \le \|A_t\|_{2\to 2}\,\|\Psi_t - \Psi_t^\star\|_{2\to 2}\,\|x_t\|_2\,.$$

□

This motivates dictionary-refresh triggers based on predictive residuals rather than a fixed schedule.

8 Uncertainty, Anytime-Valid Acceptance, and Drift Control

8.1 Diagnostic uncertainty

For discovery, we use a block bootstrap over items, prompts, judges, and decoding seeds, refitting the structured estimator in each replicate. A debiased least-squares fit on the selected support estimates effect sizes, while selection frequencies communicate support instability. Split conformal calibration may be used for held-out task residuals when exchangeability is credible. These intervals guide diagnosis and data allocation, but they are not the sole basis for accepting a repeatedly monitored candidate.

8.2 Protected non-inferiority hypotheses

Let $\mathcal{H}_t$ be a protected anchor suite disjoint from dictionary construction, sensing, hyperparameter selection, and proposal generation. For anchor slice j, define a material-regression margin $\tau_j > 0$ and the null

$$H_{0,j} : \Delta c_{t,j} \le -\tau_j, \qquad H_{1,j} : \Delta c_{t,j} > -\tau_j. \tag{27}$$

For sequential paired observations on anchor j, construct a nonnegative e-process $E_{j,s}$ satisfying

$$\sup_{P \in H_{0,j}} \mathbb{E}_P[E_{j,\tau}] \le 1 \tag{28}$$

for every stopping time τ. Equivalently, an anytime-valid lower confidence sequence $L_{j,s}$ may be used. The candidate establishes non-inferiority on anchor j when $E_{j,s} \ge 1/\alpha_j$ or $L_{j,s} > -\tau_j$.

Proposition 4 (Anytime-valid false-acceptance control)**.** *Assume each protected anchor stream is sequestered until the candidate and sensing policy are frozen, and let $E_{j,s}$ be valid under $H_{0,j}$. If a candidate is accepted only after every critical anchor rejects its material-regression null at level α_j, then under arbitrary data-dependent monitoring and stopping,*

$$\mathbb{P}(\textit{accept while at least one critical } H_{0,j} \textit{ is true}) \le \sum_{j \in \mathcal{J}_{\mathrm{crit}}} \alpha_j. \tag{29}$$

Sharper closed-testing or e-value multiple-testing procedures may replace the Bonferroni allocation.

Proof. For any true null $H_{0,j}$, Ville's inequality gives $\mathbb{P}(\sup_s E_{j,s} \ge 1/\alpha_j) \le \alpha_j$, regardless of the stopping rule. False acceptance requires at least one true material-regression null to cross its rejection boundary. Taking a union bound over critical anchors proves Eq. (29). □

This guarantee addresses a failure mode that fixed-sample bootstrap intervals do not: developers may inspect results repeatedly, stop when a candidate looks favorable, or allocate more samples to ambiguous anchors. The e-process gate remains valid under those choices, provided the protected data and null model are respected.

8.3 Modular gate

For expert e, let $\widehat{\mathcal{J}}_e$ be capability slices with substantial recovered loading on group G_e. The system may accept expert e only if:

1. its protected critical anchors establish non-inferiority;
2. at least one declared benefit target satisfies a pre-registered improvement criterion or the update serves a non-performance objective;
3. global composed-model anchors pass, guarding against expert interactions;
4. integrity, contamination, evaluator-independence, and security checks pass; and
5. residual monitoring does not reject the local response model.

Experts may therefore be accepted, rolled back, quarantined for denser diagnosis, or split when an apparently single module has heterogeneous effects. Partial acceptance is permitted only when the composed accepted subset is re-evaluated on global anchors.

8.4 Acceptance is not discovery

The sensing branch may adaptively search for changed coordinates, but anchor outcomes are never used to fit Ψ_t, tune sparsity penalties, nominate experts, or generate new training examples. This firewall limits Goodhart pressure: discovery proposes an explanation, while protected sequential tests decide whether the candidate may alter the deployed model.

9 The SPECTRA-RSI Loop

10 Architecture

Figure 1 summarizes the separation among proposal generation, sensing, and independent verification.

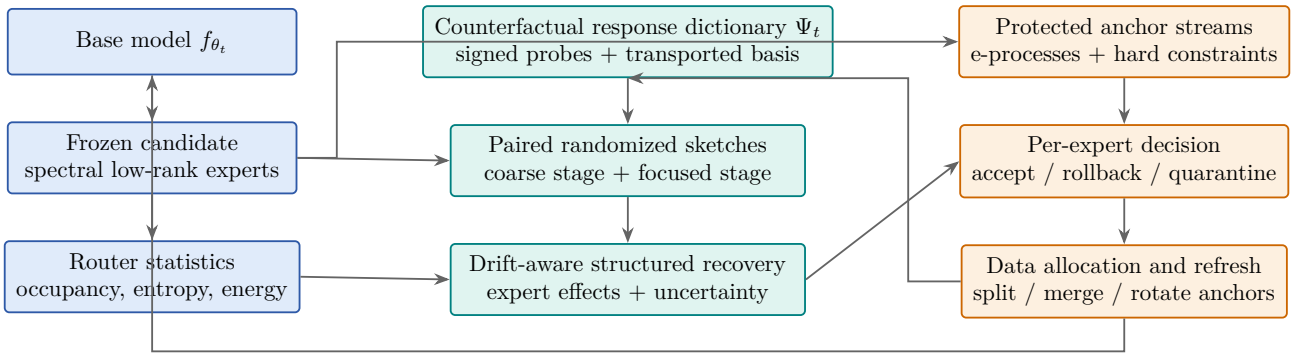

Figure 1: The proposed loop deliberately separates adaptive discovery from acceptance. The sensing branch estimates a structured capability delta; the anchor branch decides whether the candidate is safe to incorporate.

11 Novelty Relative to the State of the Art

Table 1 distinguishes the estimand, intervention, and statistical decision made by adjacent approaches. The manuscript's novelty does not rest on compressed sensing, LoRA, MoE routing, or sequential testing individually. It rests on using them to solve one new closed-loop problem: *counterfactual attribution and risk-controlled acceptance of a modular model update under a drifting behavioral response basis.*

Require: base model f_{θ_t}, candidate proposal, expert directions $\{v_{t,e,r}\}$, sensing budget B_t, protected anchors $\mathcal{H}_t$, risk budget α

Freeze the candidate, training-data lineage, evaluators, and proposal policy
Align the previous response basis and test residual-drift confidence sequences
if refresh is required **then**
    apply small signed expert probes and estimate H_t and Ψ_t
end if
Estimate clipped router priors and reserve a nonzero exploration probability for every group
Run a local-linearity pilot using paired common-random-number evaluations
if pilot residual exceeds the trust-region threshold **then**
    split the candidate into smaller sub-updates or invoke dense evaluation; **return**
end if
Collect a coarse randomized paired sketch over all capability groups
Recover preliminary group support and quantify instability by bootstrap refits
Choose a focused sketch by expected error reduction, retaining exploration mass
Solve weighted sparse-group recovery and debias the selected support
Map recovered effects to experts and nominate accept/rollback/quarantine actions
Start protected anchor e-processes for each nominated expert and the composed candidate
while no accept/reject boundary is crossed and budget remains **do**
    acquire the next most informative protected anchor item without revealing it to the proposal policy
    update e-processes, confidence sequences, and residual-drift monitor
end while
if all critical non-inferiority tests and global integrity conditions pass **then**
    accept the verified expert subset and re-evaluate the composed accepted model
    append signed probes and accepted deltas to the auditable response history
else
    rollback failed experts and quarantine unresolved effects for dense diagnosis
end if
rotate protected anchors, version all thresholds, and emit an immutable audit record

Algorithm 1: One counterfactual, risk-controlled SPECTRA-RSI iteration

Approach	Primary estimand	Intervention / acquisition	Structural prior	Decision unit	Repeated-monitoring protection
IRT / tiny benchmark [14, 13, 22]	static ability or rank	adaptively selected items	item difficulty / low-dimensional ability	whole model	usually fixed-sample interval
Active or amortized evaluation [24, 27]	static full score	learned item acquisition or generation	cross-item predictability	whole model	task-dependent stopping
mRMR + regression [5]	static full score	selected predictive items	supervised feature relevance	whole model	held-out prediction error
CS capability localization [3]	static component importance	randomized component knockouts	sparse critical heads	attention head	diagnostic only
MoE-LoRA adaptation [20, 30, 26]	task performance	routed low-rank updates	expert specialization	adapter / expert	conventional validation
Adaptive learn-then-test [29]	candidate risk validity	sequential tests	risk hypotheses	hyperparameter / policy	e-process validity
SPECTRA-RSI	**base-to-candidate capability delta and causal expert effects**	**signed expert probes + paired randomized sketches**	**group-compressible interventional response with drift**	**expert subset and composed candidate**	**protected e-process gate under arbitrary stopping**

Table 1: State-of-the-art positioning. Existing methods solve important subproblems, but none of the listed approaches jointly learns causal update fingerprints, compressively estimates a candidate delta, transports the basis over iterations, and performs anytime-valid per-module acceptance.

11.1 Novel technical object: the counterfactual response spectrum

The key object is not a Hessian spectrum or a singular-value decomposition by itself. It is the spectrum of the *interventional map* from admissible modular update directions to behavioral deltas. Its singular vectors define cross-capability response modes; its grouped right-side coordinates retain the identity of experts that can be rolled back. This creates a bridge between behavioral evaluation and controllable parameter modules.

11.2 Novel statistical architecture: sketches for discovery, e-processes for acceptance

Current efficient benchmarks generally use the same evaluation outcomes to estimate performance and support a downstream choice. Here the high-adaptivity discovery channel is deliberately denied authority to accept the candidate. Protected anchor streams support non-inferiority hypotheses with optional-stopping validity. The separation is a structural safeguard, not merely a held-out split, because it assigns different statistical roles and access controls to the two channels.

11.3 Novel dynamic element: basis drift as a monitored failure event

Static benchmark compressors and static localization methods ordinarily assume a stable correlation or component structure. In an RSI loop, the model, router, task distribution, and update basis evolve. SPECTRA-RSI aligns dictionaries across iterations, measures subspace motion, and treats excess predictive residual as a sequentially monitored model violation that expands the evaluation budget.

11.4 Scope of the claim

A literature search cannot prove absolute novelty, and future or contemporaneous work may overlap. The defensible claim is that, among the adjacent methods identified above, the complete counterfactual-

sketching and anytime-valid modular-control loop is not provided. Empirical publication will still require showing that its extra structure improves attribution or safety at equal total evaluation cost.

12 Experimental Protocol

No empirical gains are asserted. The following protocol is designed to falsify the structural, statistical, and systems claims.

12.1 Research questions

RQ1: Delta reconstruction.
At equal evaluated-token cost, do paired structured sketches recover dense base-to-candidate capability deltas more accurately than modern static-score compressors adapted by differencing?

RQ2: Causal attribution.
Under known single- and multi-expert interventions, does the counterfactual response dictionary identify the perturbed expert groups and the sign of their effects?

RQ3: Optional-stopping safety.
Across repeated peeking, early stopping, and adaptive anchor allocation, does the e-process gate control false acceptance at its declared level while fixed-sample intervals fail under adversarial stopping?

RQ4: Basis drift.
How quickly does residual monitoring detect changes caused by task-distribution shift, router retraining, expert replacement, or nonlinear update size?

RQ5: Evaluation economics.
Does the method reduce total items, tokens, judge calls, and wall-clock time at matched regression recall and false-acceptance rate after accounting for probe cost?

RQ6: Long-horizon control.
Over 20–100 update cycles, does modular rollback reduce cumulative worst-slice regression relative to whole-candidate acceptance?

12.2 Models and update interventions

Use at least two open-weight model families and two scales. Instantiate standard LoRA, sparse MoE-LoRA, and a full-parameter or model-editing condition where feasible. Construct: (i) single-expert beneficial updates; (ii) single-expert regressions; (iii) mixtures with canceling aggregate effects; (iv) interacting expert pairs; (v) broad non-compressible updates; and (vi) out-of-trust-region updates. Dense evaluation supplies ground-truth deltas and expert support labels for the research benchmark.

12.3 Capability space and split discipline

Form n capability slices by crossing task family, difficulty, language, response format, perturbation type, demographic or safety category, and evaluator. Suggested families include mathematical reasoning, code with executable tests, factuality, instruction following, tool use, robustness, refusal, and over-refusal. Partition examples into probe-calibration, sensing, bootstrap-calibration, protected-anchor, and final-audit pools. The proposal generator must never access protected-anchor items or measurement seeds.

12.4 Baselines

Compare against:

1. dense evaluation with fewer items per slice;
2. uniform, variance-stratified, and sequential confidence-interval sampling;
3. tinyBenchmarks, metabench, anchor points, and Fluid Benchmarking adapted to base/candidate differencing [14, 13, 16, 22];
4. active evaluation acquisition and reliable amortized model-based evaluation [24, 27];
5. mRMR feature selection with kernel-ridge prediction [5];
6. low-rank matrix completion over the task-by-iteration delta matrix;
7. unstructured random-projection compressed sensing;
8. group-sparse recovery without signed probes, router priors, or exploration;
9. router-only attribution and exhaustive per-expert ablation;
10. fixed-sample bootstrap or asymptotic intervals versus the anytime-valid gate; and
11. an oracle dictionary learned from dense deltas, providing an upper bound on recoverability.

12.5 Primary metrics

$$\text{normalized delta error} = \frac{\left\|\widehat{\Delta c} - \Delta c\right\|_2}{\|\Delta c\|_2 + \epsilon}, \tag{30}$$

$$\text{expert support F1} = \mathrm{F1}(\widehat{\mathcal{S}}_{\text{expert}}, \mathcal{S}_{\text{expert}}), \tag{31}$$

$$\text{weighted regression recall} = \frac{\sum_j \omega_j \mathbf{1}[j \text{ detected}]\mathbf{1}[\Delta c_j < -\tau_j]}{\sum_j \omega_j \mathbf{1}[\Delta c_j < -\tau_j]}, \tag{32}$$

$$\text{false-acceptance rate} = \mathbb{P}(\text{accepted} \mid \exists j \in \mathcal{J}_{\text{crit}} : \Delta c_j \le -\tau_j), \tag{33}$$

$$\text{drift detection delay} = \mathbb{E}[(T_{\text{alarm}} - T_{\text{shift}})_+]. \tag{34}$$

Also report confidence-sequence coverage over time, false rollback, support stability, principal-angle error, probe overhead, evaluated items, generated tokens, judge tokens, solver time, wall-clock time, energy where measurable, and cumulative worst-slice regression.

12.6 Optional-stopping stress test

For each candidate, simulate stopping rules that inspect the running evidence and stop at the most favorable time, after a fixed deadline, or after an adaptive budget increase. Measure empirical type-I error of the e-process gate, fixed-horizon t or normal intervals, percentile bootstrap intervals, and naive repeated testing. This experiment directly tests the claimed benefit of anytime-valid acceptance.

12.7 Ablations and phase diagram

Remove signed counterfactual probes, pairing/common randomness, group structure, router priors, exploration floors, dictionary transport, drift monitoring, e-process acceptance, sealed anchors, and per-expert rollback. Sweep active-group count, update norm, interaction strength, score noise, judge correlation, task shift, and probe magnitude. Present a phase diagram identifying regions in which structured sensing is cheaper than dense evaluation and regions in which the method correctly falls back.

12.8 Statistical analysis and reporting

Pre-register normalized delta error at fixed total token budget and protected-regression recall at fixed false-acceptance rate as primary endpoints. Use paired comparisons across identical candidate interventions and task pools. Report exact finite-sample or anytime-valid intervals where available;

avoid relying on central-limit approximations for small slices. Correct secondary comparisons and publish all rejected candidates, stopping decisions, and dense ground truth.

13 Worked Budget Analysis

The original draft used $n = 10{,}000$, $k = 100$, $E = 64$, and $s = 8$ to suggest a measurement count near 530. This number should be interpreted only as an order-of-magnitude design target. Constants depend on coherence, signal-to-noise ratio, dynamic range, group structure, and the number of items required to estimate each measurement. A more defensible planning equation is

$$B_{\text{total}} = B_{\text{probe}} + B_{\text{sense}} + B_{\text{anchor}} + B_{\text{audit}}, \tag{35}$$

where sensing itself obeys $B_{\text{sense}} = \sum_{i,j:a_{ij}eq0} N_{ij}$. The empirical question is whether reuse of the interventional dictionary and paired sketches reduces B_{total} at a target attribution, regression-detection, and false-acceptance accuracy.

Method	Aggregate measurements	Items per measurement	Total item cost
Dense per-slice	n	N_0	nN_0
Uniform task subsampling	m	N_0	mN_0
Sparse composite sensing	m	adaptive N_{ij}	$\sum_{ij} N_{ij}$
Two-stage structured sensing	$m_1 + m_2$	adaptive, group-focused	$B_1 + B_2$

Table 2: Evaluation efficiency must be compared in evaluated items or tokens, not only in the number of aggregate scores.

14 Failure Modes and Safeguards

Non-compressible capability change. Broad updates may affect many coordinates. Rising tail energy or unstable support indicates that sparse sensing is inappropriate. The system should increase budget, reduce update size, or fall back to dense stratified evaluation.

Nonlinear interactions. A combination of individually safe expert updates may be unsafe. The gate must evaluate both individual experts and the composed candidate. Interaction terms can be added through a sparse quadratic extension, but this increases sample complexity.

Dictionary capture and benchmark gaming. A model trained repeatedly against reconstructed coordinates may exploit blind spots. Measurement matrices must be sampled after candidate freezing; anchor tasks must be sealed; and task distributions must be periodically refreshed by external or procedurally verified generators.

Router manipulation. Router occupancy is a prior, not an observation of improvement. Exploration measurements, prior clipping, and comparison between router and behavioral support prevent an expert from hiding behind or inflating routing statistics.

Evaluator dependence. When scores come from a judge model, correlated errors can violate noise assumptions. Use heterogeneous judges, rule-based checks, executable tests, and human audits for high-impact slices. Judge identity should be a random effect in uncertainty analysis.

Safety externalities. No compressed evaluation scheme should replace dense checks for catastrophic or legally constrained outcomes. Hard safety anchors, adversarial testing, access controls, and human authorization remain mandatory for consequential deployment.

15 Governance and Auditability

Each iteration should emit an immutable audit record containing the base and candidate hashes, training-data lineage, dictionary version, router statistics, measurement seed, task pools, recovered support, uncertainty intervals, anchor outcomes, expert-level decisions, and rollback state. The sensing policy and acceptance thresholds should be versioned separately from the model. This makes it possible to reconstruct why an update was accepted and to detect systematic drift in the evaluator itself.

The framework is best viewed as a control system with a protected reference channel. The adaptive sensing branch maximizes diagnostic efficiency; the anchor branch maintains a stable safety reference. Governance should prohibit the proposal generator from reading sealed-anchor contents or measurement seeds before candidate freezing.

16 Discussion

Spectra-RSI treats iterative model improvement as a changing-input, changing-basis system-identification problem. The framework is most plausible when updates are modular, candidate steps are small enough for local superposition, capability changes are group-compressible, and protected outcomes can be scored with bounded or otherwise sequentially testable statistics. These assumptions are not guaranteed by scale. They are operational contracts that the pilot residual, drift monitor, and dense fallback are designed to audit.

The counterfactual dictionary adds cost because expert probes require model executions. That cost is justified only when the dictionary can be reused across several candidates or when the value of causal rollback exceeds the probe overhead. A static one-off benchmark may be better served by IRT, active item acquisition, or regression-based subset selection. The proposed method targets repeated model-development loops in which attribution, safe partial acceptance, and longitudinal auditability matter.

Anytime-valid testing controls a declared statistical error, not the completeness of the safety specification. An omitted capability slice, biased judge, contaminated anchor, or invalid null model can still produce a formally valid but substantively inadequate decision. High-consequence capabilities therefore require external red teams, procedurally generated fresh tasks, executable checks, human authorization, and dense evaluation where failure is unacceptable.

The framework also applies outside autonomous RSI. It can support continuous integration for models, adapter marketplaces, federated or multi-tenant fine-tuning, knowledge-edit auditing, and safety regression triage. In these uses, the model need not generate its own training data; only modular candidate updates and repeated evaluation are required.

17 Conclusion

This paper proposed Spectra-RSI, a counterfactual and statistically protected measurement layer for modular recursive self-improvement. Signed expert probes estimate a local map from admissible low-rank updates to capability changes. Paired randomized sketches recover a structured candidate delta, router information supplies a clipped support prior, and dictionary transport plus residual monitoring expose basis drift. Crucially, exploratory reconstruction does not authorize deployment: protected

anchor e-processes provide anytime-valid non-inferiority evidence for expert-level and composed-model acceptance under adaptive sample sizes and stopping.

Relative to efficient static benchmarking and static compressed-sensing localization, the new object is the causal capability delta of a candidate update and the new decision is whether a reversible subset of that update can safely enter the next model generation. Theoretical guarantees identify the required assumptions, while the proposed experiments are designed to reveal when those assumptions fail. A successful empirical result would not demonstrate unconstrained self-improvement; it would establish a more efficient, attributable, and auditable control channel for repeated model change.

References

[1] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[2] Y. Bai et al. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073, 2022.

[3] A. Bair, Y. E. Xu, M. Sun, and J. Z. Kolter. Compressed sensing for capability localization in large language models. arXiv:2603.03335, 2026.

[4] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[5] S. Bowyer, A. Locatelli, and K. Cao. Efficient benchmarking is just feature selection and multiple regression. arXiv:2605.25773, 2026.

[6] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[7] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[8] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[9] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[10] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[11] A. Huang et al. Self-improvement in language models: The sharpening mechanism. arXiv:2412.01951, 2024.

[12] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models. *IEEE Transactions on Signal Processing*, 59(5):1985–2001, 2011.

[13] A. Kipnis, K. Voudouris, L. M. Schulze Buschoff, and E. Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, 2025.

[14] F. Maia Polo, L. Weber, L. Choshen, Y. Sun, G. Joshi, and M. Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning*, 2024.

[15] N. Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[16] R. Vivek, K. Ethayarajh, D. Yang, and D. Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2024.

[17] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[18] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

[19] W. Yuan et al. Self-rewarding language models. arXiv:2401.10020, 2024.

[20] Y. Zhu et al. SiRA: Sparse mixture of low-rank adaptation. arXiv:2311.09179, 2023.

[21] S. Bowyer, L. Aitchison, and D. R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[22] V. Hofmann, D. Heineman, I. Magnusson, K. Lo, J. Dodge, M. Sap, P. W. Koh, C. Wang, H. Hajishirzi, and N. A. Smith. Fluid language model benchmarking. In *Conference on Language Modeling*, 2025.

[23] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[24] Y. Li, J. Ma, M. Ballesteros, Y. Benajiba, and G. Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 35581–35602, 2025.

[25] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[26] B. Shi, W. Chen, S. Zhao, J. Shen, S. Guo, S. Wang, and H. Wan. SAMoRA: Semantic-aware mixture of LoRA experts for task-adaptive learning. arXiv:2604.19048, 2026.

[27] S. T. Truong, Y. Tu, P. Liang, B. Li, and S. Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 60238–60265, 2025.

[28] H. Yang, M. Xerri, S. Park, H. Zhang, Y. Feng, S. A. Kogilathota, and J. Zhou. Self-improvement of large language models: A technical overview and future outlook. arXiv:2603.25681, 2026.

[29] M. Zecchin, S. Park, and O. Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 74018–74036, 2025.

[30] D. Zhang, K. Zhang, S. Chu, L. Wu, X. Li, and S. Wei. MoRE: A mixture of low-rank experts for adaptive multi-task learning. arXiv:2505.22694, 2025.

A Dictionary Construction Options

Historical delta PCA. Stack dense or high-confidence capability deltas from previous iterations into $D = [\Delta c_1, \ldots, \Delta c_T]$ and use a robust or shrinkage SVD. This is simple but can reproduce historical blind spots.

Task-gradient response dictionary. For differentiable scores, estimate task gradients or Fisher-vector products and project them through candidate expert update bases. This directly approximates $G_t = J_t B_t$ but can be computationally expensive.

Interventional dictionary. Apply small random perturbations to experts, measure resulting task deltas, and learn a low-rank response model. This provides causal grounding but consumes calibration budget.

Hybrid dictionary. Concatenate historical, gradient, and interventional atoms, then orthogonalize or learn an overcomplete dictionary with a sparsity penalty. Dictionary atoms should be labeled by provenance for auditability.

B Possible Quadratic Extension

When expert interactions are important, augment the local model with selected pairwise terms:

$$\Delta c_t \approx \Psi_t x_t + \sum_{(e,f)\in\mathcal{P}} \Psi^{(2)}_{ef} z_{ef}. \tag{36}$$

A hierarchical sparsity constraint requires an interaction z_{ef} to be active only when one or both parent groups are active. The extension should be used only when held-out residual reduction justifies its increased sample complexity.

C Construction of a Bounded E-Process for Paired Anchors

For illustration, suppose paired anchor differences $D_s \in [-b, b]$ are conditionally independent given the past and the null is $\mathbb{E}[D_s \mid \mathcal{F}_{s-1}] \le -\tau$. Let $X_s = D_s + \tau$, so the null implies nonpositive conditional mean. A betting process of the form

$$E_s(\lambda) = \prod_{r=1}^{s} \exp(\lambda X_r - \psi_r(\lambda)), \tag{37}$$

with a valid conditional cumulant upper bound ψ_r is a nonnegative supermartingale under the null. Mixing over $\lambda \ge 0$ yields an e-process that adapts to unknown effect size. In implementation, the exact construction should match score bounds, dependence, and stratification; the paper's guarantee assumes validity of the chosen process rather than a universal parametric form.

D Causal Fingerprint Design

Probe directions should be admissible update operations rather than arbitrary dense parameter perturbations. For LoRA expert e with factors $A_e B_e$, candidate directions may include singular directions of the current adapter update, rank-one signed perturbations, or a small orthogonal design in adapter-coordinate space. Probe magnitude h is selected by balancing central-difference bias,

finite-evaluation variance, and operational safety. A three-level pilot $\{-h, 0, +h\}$ permits a curvature diagnostic; large asymmetry indicates that linear superposition is unreliable.

E Audit Record Schema

Each iteration should store: base and candidate hashes; expert and router versions; training-data lineage; probe directions and magnitudes; task-pool versions; sensing matrices and seeds released after evaluation; paired item identifiers; solver settings; recovered support and uncertainty; anchor nulls, margins, alpha allocation, e-process trajectories, and stopping reasons; drift alarms; expert-level actions; composed-model verification; and rollback state. The record should allow an independent auditor to replay both the statistical decision and the model composition.

F Minimum Reproducibility Checklist

1. Publish all task-slice definitions and data-splitting rules.
2. Report sensing and anchor item counts separately.
3. Release measurement-matrix generators and random seeds after evaluation.
4. Report all candidate updates, including rejected ones.
5. Provide dense ground-truth deltas for the research benchmark.
6. Include calibration plots, worst-slice metrics, and false-acceptance rates.
7. Document dictionary refresh, split, and merge thresholds.
8. Separate exploratory results from pre-registered primary endpoints.