FSR: ASSESSING RELIABILITY IN LANGUAGE MODEL INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

An intermediate prediction can be confident yet differ from a language model's final output, and a large estimated intervention effect can have the wrong sign. We introduce *Fisher–Schur Reliability* (FSR), which assesses these discrepancies using reference errors projected onto the quantity being measured. For token distributions, an anisotropic Fisher statistic describes residual variation along likely token contrasts. Across five models up to 27B parameters, adding this statistic reduces error-ranking loss (EAURC) by 4.1–27.3% over a matched three-input control and by 3.1% in macro average over an eleven-input predictor. The compact predictor's gain also holds on a disjoint cohort with its fit frozen. For intervention reporting, FSR normalizes fixed lens or attribution-patching estimates by their residual error scales. On 512 fresh Qwen prompts at about 10% coverage, conditioning the scale on each reading reduces wrong-direction reports from 12.31% to 11.50% for lenses and from 1.45% to $.07\%$ for attribution patching. Response magnitude supplies most of the latter gain. FSR separates effect estimation from error assessment, providing measurement-specific scores for ranking discrepancies and selecting intervention directions to report.

1 INTRODUCTION

A researcher using an intermediate readout to interpret a large language model faces a basic question: would the full model support the same conclusion? A confident token distribution may disagree with the final output, and an estimated intervention effect may have the wrong sign. Affine lenses make such intermediate readings possible: the logit lens applies the unembedding directly (nostalgebraist, 2020), while the Tuned Lens learns a translator (Belrose et al., 2023). These tools make computation readable, but the clarity of a reading is not its fidelity to a specified endpoint. We study that fidelity, without treating every difference between an intermediate and final prediction as a mistaken description of the intermediate computation.

Prediction confidence describes how concentrated a readout's distribution is, while input atypicality describes how far a state lies from the fitting data. Neither captures which token contrasts the readout tends to get wrong. An error between likely tokens matters more to a distributional reading than the same error between unlikely ones. Similarly, two estimated intervention effects may be equally large but differ in reliability because their token contrasts and the sizes of the underlying changes imply different error scales.

Fisher–Schur Reliability (FSR) assesses these error patterns for the question being asked. We compare a fixed readout with the full model on reference prompts and fit the mean and covariance of its errors.

For a full distribution, an error matters more when it changes the relative probabilities of likely tokens. Starting with an affine adaptation of a linear readout (Yom Din et al., 2024), we use the softmax Fisher metric to weight these error directions. Subtracting the contribution of direction-independent spread gives the anisotropic statistic A_{FS}: it measures whether the current prediction emphasizes contrasts with unusually large or small readout errors. For example, two equally confident readings can place mass on different token pairs whose residual differences have very different variance. FSR distinguishes these cases using the identities of the likely tokens. We test the additional information beyond confidence, atypicality, and frequency, including a richer eleven-input predictor.

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.

For an intervention, FSR projects the residual moments onto the requested contrast and compares its error scale with the estimated effect. For lens differences, this accounts for shared errors that cancel between endpoints. The same construction applies to attribution patching (AtP), using residuals of its first-order response. AtP gives more accurate estimates in our matched WikiText comparison; FSR further improves which directions it reports. Global moments distinguish contrasts; a small conditional model captures variation across readings using estimated response and change magnitudes. An independent 512-prompt evaluation confirms the additional gain for both estimators. A second-family study evaluates Gemma-2-2B using explicitly defined pre-softcap logit contrasts.

The fitted moments support multiple contrasts for a fixed intervention and full-distribution assessment. Global scoring needs no additional backward pass; per-reading scoring also uses estimator-dependent response-vector features. FSR assesses reliability separately from how the estimate was obtained. For distributional readings, the target is agreement with the final model, not factual correctness or a stopping policy. When that final output is already available, the discrepancy can be measured directly. For interventions, the exact target instead requires the changed computation; our experiments ask which estimates to report at a given coverage.

Contributions. (1) We derive an anisotropic Fisher statistic for distributional reliability and evaluate it exactly with one unembedding matrix–vector product. (2) We establish its value across five models through matched ablations, richer controls, and frozen-score validation. (3) We use projected residual scales to improve direction reporting for both lens and attribution-patching estimates, and confirm the value of per-reading scales on fresh prompts, with additional model and intervention checks.

2 Related Work

Affine readings. Logit, Tuned, Future, Entropy, and Jacobian lenses expose intermediate computation (nostalgebraist, 2020; Belrose et al., 2023; Pal et al., 2023; Ali et al., 2025; Gurnee et al., 2026). Our moment lens is a ridge-regularized affine adaptation of Yom Din et al. (2024); Patchscopes and DecoderLens offer alternative readings (Ghandeharioun et al., 2024; Langedijk et al., 2024). Recent work characterizes the J-lens as a causal-transfer operator (Yan et al., 2026) and separates behavioral visibility from intervention efficacy (Pettyjohn et al., 2026). Our main question is how to assess the error of a given affine reading. We use the moment lens for the distributional experiments and test transfer to a KL-trained affine lens. Early-exit methods such as CALM (Schuster et al., 2022) turn intermediate predictions into stopping policies. FSR studies error ranking for a fixed readout, without changing decoding or evaluating adaptive execution.

Confidence and residual geometry. Leverage measures extrapolation (Hoaglin & Welsch, 1978; Gramatica, 2007). Representation density and distance also inform OOD detection and language-model uncertainty (Lee et al., 2018; Podolskiy et al., 2021; Jiang et al., 2018; Liu et al., 2020; Postels et al., 2020; Yuksekgonul et al., 2023; Vazhentsev et al., 2025). Uncertainty-aware probes and trajectory methods target labels, correctness, or abstention (Wang et al., 2023; Eusebi et al., 2026), whereas layerwise confidence need not distinguish correct predictions (Kim et al., 2025). FSR targets agreement with the frozen model's endpoint. It combines covariance propagation (Nilsen et al., 2022) with local Fisher geometry and selective risk–coverage evaluation (Geifman & El-Yaniv, 2017; Noskov et al., 2024); the anisotropic term separates excess residual spread from its isotropic ambiguity component. Its predictive value is evaluated by matched ablations, rather than assumed from the decomposition.

Intervention measurements. Attribution patching and AtP* approximate intervention effects using local derivatives at the current input (Kramár et al., 2024). Zhang & Wang (2026) also propose a directly related reliability score: the magnitude of a Hessian correction divided by the first-order effect, followed by selective correction. FSR instead fits residual scales on reference interventions and leaves the estimator fixed. The distinction is fitted error statistics versus query-specific curvature, not reliability assessment versus effect estimation; a matched comparison of their reporting quality and derivative cost remains open. Intervention-centered evaluation and MIB study control and circuit localization (Bhalla et al., 2024; Mueller et al., 2025). Certified Interventional Fidelity treats distribution-averaged causal estimands (Asiaee, 2026). FSR addresses individual affine effect estimates and the decision to report them. This matters because circuit estimates and patching conclusions depend on sampling, metrics, subspaces, and linearization (Méloux et al., 2025; Zhang

& Nanda, 2024; Miller et al., 2024; Makelov et al., 2024). We focus on reporting decisions for individual estimates (Orgad et al., 2026).

3 MEASUREMENT-ALIGNED RELIABILITY

Setup. Fix a decoder-only transformer and a source layer ℓ. For a position t in a prompt, let $x = h_\ell(t) \in \mathbb{R}^d$ be the residual state after layer ℓ and $y = h_{\text{final}}(t) \in \mathbb{R}^d$ the post-final-norm state whose unembedding gives pre-output-transform logits $z = W_U y$. Write $g(z)$ for the model's output transform: identity for Qwen/Pythia and coordinatewise softcapping for Gemma. Layer labels Lk denote h_k, after exactly k blocks. Collect n pairs (x_i, y_i) from forward passes over a fitting corpus (positions $t \geq 16$; Sec. 4.1).

3.1 AFFINE READINGS AND RESIDUAL GEOMETRY

Moment-based reading. The moment lens estimates the joint first and second moments of (x, y) and uses a ridge-regularized affine predictor:

$$\widehat{y}_\ell(x) = \widehat{\mu}_y + M_\lambda \delta, \qquad M_\lambda = \widehat{\Sigma}_{yx}\widehat{\Sigma}_\lambda^{-1}, \qquad p_{\text{lens}}(\cdot \mid x) = \text{softmax}(g(W_U \widehat{y}_\ell(x))), \tag{1}$$

where $\delta = x - \widehat{\mu}_x$. The distributional experiments use one fixed ridge covariance

$$\widehat{\Sigma}_\lambda = \widehat{\Sigma}_{xx} + \lambda I, \qquad \lambda = 10^{-2}, \tag{2}$$

for both the reading and its geometry, without layer-specific tuning. One Cholesky factor $\widehat{\Sigma}_\lambda = LL^\top$ serves both. Gaussian assumptions enter only the population analysis, not the fit.

The same joint fit supplies the ridge-consistent Schur covariance

$$S_\lambda = \widehat{\Sigma}_{yy} - \widehat{\Sigma}_{yx}\widehat{\Sigma}_\lambda^{-1}\widehat{\Sigma}_{xy}. \tag{3}$$

The fit therefore gives three quantities: source precision $P_\lambda = \widehat{\Sigma}_\lambda^{-1}$, the reading map M_λ, and global spread S_λ. The empirical ridge-residual covariance is $S_\lambda - \lambda M_\lambda M_\lambda^\top$; the Schur surrogate instead matches the regularized joint model (Appendix F.2). This is a modeling choice, not an empirical claim that the Schur surrogate outperforms the fitted-map residual covariance.

Measurement risk. Let $\mathbf{r}$ stack the residuals needed for a measurement, with fitted mean $\mathbf{m}$ and covariance $\mathbf{S}$. If J maps these residuals into measurement space and $G \succeq 0$ is its loss metric, their quadratic error is

$$\mathcal{R}(J, G; \mathbf{m}, \mathbf{S}) = \mathbf{m}^\top J^\top G J \mathbf{m} + \text{tr}\big(GJ\mathbf{S}J^\top\big). \tag{4}$$

The two terms are the projected squared bias and variance. For intervention effects, the projection follows a token contrast; for distributions, the loss metric is the local Fisher matrix. We develop distributional reliability before intervention reporting, following the construction in Figure 1.

3.2 RELIABILITY OF DISTRIBUTIONAL READINGS

Source-state atypicality. The Cholesky factor also gives

$$\tau(x) = \delta^\top \widehat{\Sigma}_\lambda^{-1}\delta = \|L^{-1}\delta\|_2^2, \tag{5}$$

the squared Mahalanobis distance from the fit. High-τ readings extrapolate the affine map beyond typical fitting states; Section E relates this radius to readout error.

3.2.1 ANISOTROPIC COVARIANCE ALIGNMENT

Radial atypicality and entropy describe extrapolation and ambiguity, but not *which vocabulary directions* carry residual uncertainty. Readings with identical τ and entropy can differ in their uncertainty along the token contrasts that matter to the current distribution.

For a token distribution, Equation 4 uses $J = W_U$ and $G = F(p)/2$. Let $z = W_U\widehat{y}(x)$, $p = \text{softmax}(z)$, and $F(p) = \text{diag}(p) - pp^\top$ be the softmax Fisher information. Set $C = W_U S_\lambda W_U^\top$.

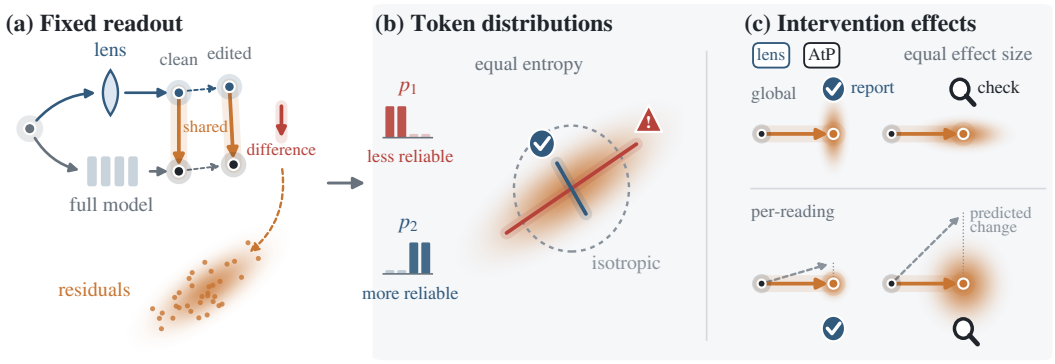

Figure 1: **From residual geometry to reliable reporting.** **(a)** Fit residual moments on reference prompts with the lens fixed; shared errors cancel between clean and edited readings. **(b)** Equal-entropy readings stress different token pairs; A_{FS} compares their residual spread with an isotropic reference. **(c)** For lens or AtP estimates, error scales vary by contrast and, per reading, with the predicted change.

For a zero-mean local logit residual with covariance C, a second-order expansion of the KL between the model's and the lens's token distributions gives the spread contribution

$$G(p) \;=\; \tfrac{1}{2}\,\mathrm{tr}[F(p)\,C]\,. \tag{6}$$

This projects the global covariance S_λ onto the token contrasts weighted by the current distribution. It also includes output ambiguity. Because $F(p)\mathbf{1} = 0$, only centered-logit residuals matter. The following decomposition identifies the part that cannot be reduced to average spread.

Proposition 1 (Canonical decomposition of local Fisher risk)**.** *Fix $p \in \Delta^{V-1}$. Let a logit residual r have mean b and covariance C, and set*

$$P_V = I - \mathbf{1}\mathbf{1}^\top/V, \quad C_\perp = P_V C P_V, \quad \alpha = \frac{\mathrm{tr}(C_\perp)}{V-1}, \quad D = C_\perp - \alpha P_V.$$

Then, for an ordinary softmax,

$$\frac{1}{2}\mathbb{E}[r^\top F(p) r] = \underbrace{\frac{1}{2}(P_V b)^\top F(p)(P_V b)}_{\textit{mean bias}} + \underbrace{\frac{\alpha}{2}\big(1 - \|p\|_2^2\big)}_{\textit{isotropic spread}} + \underbrace{\tfrac{1}{2}\,\mathrm{tr}[F(p)D]}_{A_{\mathrm{FS}}(p)\textit{: anisotropic alignment}}. \tag{7}$$

Moreover, $C_\perp = \alpha P_V + D$ is the unique decomposition into an isotropic operator on the centered-logit subspace and a centered trace-free remainder; its two terms are Frobenius-orthogonal. The decomposition is invariant to adding any (possibly residual-dependent) common-logit offset to r.

The proof is in Appendix E. The intercept centers residuals over the fitting split, but does not estimate the conditional bias $\mathbb{E}[r \mid x]$; our primary surrogate uses the covariance terms only. Because D is trace-free, it is generally indefinite and A_{FS} can be positive or negative. Only the sum of the isotropic and anisotropic spread terms is necessarily nonnegative. The isotropic term is proportional to Gini impurity $1 - \|p\|_2^2$, which overlaps with, but is not identical to, readout entropy. The last term, A_{FS}, asks whether likely token contrasts have larger or smaller residual spread than this isotropic reference.

For example, a distribution placing equal mass on tokens u and v has $A_{\mathrm{FS}} = [\mathrm{Var}(r_u - r_v) - 2\alpha]/8$. Two such readings have the same entropy and Gini impurity, yet can have different A_{FS} because their token contrasts have different residual variances, a distinction that confidence alone cannot capture.

Exact evaluation. The diagonal-minus-rank-one form $F(p) = \mathrm{diag}(p) - pp^\top$ gives $\mathrm{tr}[F(p)M] = \sum_v p_v M_{vv} - p^\top M p$ for symmetric M. We can therefore evaluate A_{FS} without forming the $V \times V$ matrix C:

$$A_{\mathrm{FS}}(p) = \tfrac{1}{2}\Big(\textstyle\sum_v p_v C_{vv} - p^\top C p\Big) - \tfrac{\alpha}{2}\big(1 - \|p\|_2^2\big). \tag{8}$$

We cache the V diagonal entries C_{vv}; $p^\top C p$ requires one probability-weighted unembedding matvec and an $O(d^2)$ quadratic form. Beyond the fitted $d \times d$ covariance factor, this adds $O(V)$ storage. A coordinatewise logit softcap is handled by its diagonal Jacobian without changing this complexity.

Appendix F.1 gives the factorization and numerical details. Here, exactness refers to evaluating the quadratic statistic; its connection to finite-residual KL remains local.

Lexical frequency. We complement the geometric signals with a token prior and use it for nuisance control in the covariance ablation:

$$\ell_{\mathrm{freq}}(p) = \sum_{v=1}^{V} p_v \left[-\log \frac{n_v + 1}{N + V} \right], \tag{9}$$

where counts (n_v) use fitting prompts only and $N = \sum_v n_v$.

Compact distributional score. We call the four-input predictor Compact FSR: it combines $\tau, H, \ell_{\mathrm{freq}}, A_{\mathrm{FS}}$, with interactions that depend on the reading:

$$\boxed{s_{\mathrm{FSR}}(x) = g_\theta\Big(\widehat{F}_\tau(\tau), \widehat{F}_H(H), \widehat{F}_L(\ell_{\mathrm{freq}}), \widehat{F}_A(A_{\mathrm{FS}})\Big).} \tag{10}$$

Here $H = H[p]$ and $\widehat{F}$ denotes empirical calibration percentiles. The depth-3 boosted regressor g_θ (Friedman, 2001) fits $\log(1 + e(x))$ with squared loss, where $e(x) = D_{\mathrm{KL}}(p_{\mathrm{model}} \| p_{\mathrm{lens}})$. We fix 200 iterations, learning rate .05, and seed 0 across every model and layer; the lens and its moments remain frozen. The percentile maps and predictor are fitted separately per cell, then frozen before test scoring. Low scores are retained first. Additive FSR combines the same four inputs with nested scalar weights rather than a boosted regressor; Additive FSR w/o freq. is its three-input ablation (Appendix B.2). Rich FSR adds A_{FS} to an eleven-input Rich GBM control. All variants share the same lens and residual moments.

3.3 Reliability of intervention effects

An intervention reading compares matched clean and changed states. Write $\Delta h = h_1 - h_0$ and $\Delta y = y_1 - y_0$. Our linear-contrast target is $\delta(q) = q^\top W_U \Delta y$. For $q = e_u - e_v$ and an identity output transform, it is the exact change in token log-odds. Gemma's intervention experiments use this *pre-softcap* target, not its output log-odds. For a general output transform, that latter target is instead

$$\delta_{\mathrm{out}}(q) = q^\top [g(W_U y_1) - g(W_U y_0)], \qquad \widehat{\delta}_{\mathrm{AtP,out}}(q) = q^\top Dg(W_U y_0) W_U J_c \Delta c. \tag{11}$$

Here J_c differentiates the final normalized state with respect to the intervened activation. The fixed projection below is exact for δ; it is not an exact finite-error model for nonlinear δ_{out}.

Estimate the residual scale. A fixed state lens estimates the response by differencing its endpoints, $\widehat{\Delta y} = \widehat{y}_1 - \widehat{y}_0$. Its error is $r_\Delta = \Delta y - \widehat{\Delta y} = r_1 - r_0$, with $r_i = y_i - \widehat{y}_i$. Stacking $\mathbf{r} = (r_0, r_1)$ and setting $T = [-I\ I]$ shows why its covariance must retain the cross-terms:

$$m_\Delta = \mathbb{E}[T\mathbf{r}], \qquad S_\Delta = \mathrm{Cov}(T\mathbf{r}) = S_1 + S_0 - S_{10} - S_{01}. \tag{12}$$

Estimate these moments on matched intervention pairs. Projecting onto the requested contrast gives the linear-effect error scale:

$$U_\Delta(q) = q^\top W_U S_\Delta W_U^\top q + (q^\top W_U m_\Delta)^2. \tag{13}$$

This is Equation 4 with $J = q^\top W_U$ acting on the response residual and $G = 1$: an exact second moment under the fitted residual model. Direct response fits supply these moments from their own residuals; endpoint differences instead account for shared clean and changed errors.

From error scale to a reporting decision. Let $\widehat{\delta}(q) = q^\top W_U \widehat{\Delta y}$ be the estimated effect. Reversing its direction requires an error at least as large as $|\widehat{\delta}(q)|$, motivating the normalized score

$$\rho_\Delta(q) = \frac{\sqrt{U_\Delta(q)}}{|\widehat{\delta}(q)| + 10^{-6}}. \tag{14}$$

Lower values are reported first, with a threshold fixed by a calibration quantile. Held-out error is the fraction of reports that reverse the exact direction. Normalization lets a large effect remain reportable despite substantial absolute error. Each estimator stays fixed when comparing scores.

Table 1: **Reliability comparison.** EAURC divided by readout-entropy EAURC, averaged over two layers per model (lower is better); Macro averages all ten cells. Bold marks group minima. Full values and adaptation details appear in Appendix B.1.

Method	0.5B	1.8B	27B	Pythia	Gemma	Macro
Reliability baselines and FSR						
MSP (2017)	1.139	1.126	1.229	1.119	1.146	1.152
Entropy-Lens (2025)	1.991	1.594	1.965	1.030	.966	1.509
Readout entropy	1.000	1.000	1.000	1.000	1.000	1.000
Token MD (2018)	1.134	1.170	2.656	.891	.914	1.353
kNN density (2025)	1.189	1.478	3.055	1.357	1.224	1.661
SATMD+MSP (2025)	.723	.769	1.105	.655	.624	.775
Trajectory probe (2026)	.799	.805	.810	.592	.527	.706
Confidence GBM	.734	.732	.761	.766	.664	.732
Compact FSR (ours)	**.556**	**.566**	**.756**	**.485**	**.492**	**.571**
FSR feature in a richer predictor (11 / 12 inputs)						
Rich GBM	.516	.532	.566	.427	.406	.490
Rich FSR (ours)	**.505**	**.509**	**.565**	**.407**	**.386**	**.474**

Attribution patching. Let c be the intervened activation and J_c the clean-state Jacobian of the final normalized state. AtP uses $\widehat{\Delta y} = J_c \Delta c$; its residual $r_\Delta = \Delta y - J_c \Delta c$ supplies estimator-specific moments for Equation 14. The effect estimator remains unchanged when comparing reporting rules.

A scale for each reading. For a fixed contrast, $U_\Delta(q)$ is constant within an estimator–layer cell. To capture variation between readings, we fit a conditional squared-error scale $\sigma_i^2 = \exp(\phi_i^\top \beta)$ and report by $\sigma_i/(|\widehat{\delta}_i| + \epsilon)$. The inputs are log projected risk, log response and change norms, and clean confidence where available. Up to five coefficients are fitted on reference residuals, with standardized inputs and a regularized squared-error likelihood. Calibration prompts fix the reporting threshold; neither fit uses evaluation outcomes. AtP's response norm requires $J_c \Delta c$, beyond a scalar contrast gradient. Appendix C.3 gives the objective and implementation.

4 Experiments

We first evaluate distributional error ranking, ablate the score's inputs, and test frozen predictors on a separate cohort. We then assess direction reporting for fixed lens and attribution-patching estimates, separating the effects of residual geometry and reading-specific scale features.

4.1 Distributional reliability

For full token distributions, we test how well FSR ranks the error $e(x) = D_{\mathrm{KL}}(p_{\mathrm{model}} \| p_{\mathrm{lens}})$ for Qwen1.5-0.5B/1.8B, Qwen3.6-27B, Pythia-410m, and Gemma-2-2b (Qwen Team, 2024; 2026; Biderman et al., 2023; Gemma Team, 2024). Completed-block depths $\lfloor .75 n_{\mathrm{blocks}} \rfloor$ and $\lfloor .875 n_{\mathrm{blocks}} \rfloor$, where n_{blocks} is the decoder block count, give ten model–layer cells. WikiText-2 contributes 448 fitting prompts, 64 disjoint calibration prompts, and 256 official-test prompts; positions before token 16 are excluded. A further 128 disjoint validation prompts test the frozen predictors. EAURC subtracts oracle AURC from the risk–coverage area; lower is better. Learned-score intervals use 2,000 paired prompt bootstraps conditional on the fitted predictors. Macro results weight models equally, then their layers. Appendix A.1 gives the full protocol.

On the same lens and split, we adapt MSP (Hendrycks & Gimpel, 2017), Entropy-Lens (Ali et al., 2025), readout entropy, token Mahalanobis distance (Token MD, exactly the radial τ component), and kNN density (Lee et al., 2018; Vazhentsev et al., 2025). We also adapt supervised multi-layer density (SATMD+MSP) (Vazhentsev et al., 2025) and a sparse geometric trajectory probe (Eusebi et al., 2026) to lens error, using only the available prefix. Both use the same calibration labels; Appendix B.1 specifies the adaptations. A matched confidence GBM uses maximum lens probability, entropy, and frequency with the same settings as Compact FSR. Rich GBM uses eleven non-covariance diagnostics; Rich FSR adds A_{FS} as the twelfth feature.

Table 2: **Readout and feature ablation.** EAURC (↓). The "No" columns remove one input from Compact FSR. Brackets give paired pointwise 95% intervals for the covariance contribution.

Model	Layer	Additive FSR	No A_{FS}	No freq.	Compact FSR	A_{FS} gap [95% CI]
0.5B	L18	.216	.208	.192	**.189**	$+.019\,[+.016, +.022]$
0.5B	L21	.117	.106	.124	**.097**	$+.009\,[+.008, +.010]$
1.8B	L18	.235	.235	.220	**.206**	$+.029\,[+.024, +.033]$
1.8B	L21	.122	.117	.128	**.100**	$+.017\,[+.015, +.019]$
27B	L48	.651	.621	.626	**.596**	$+.026\,[+.020, +.032]$
27B	L56	.366	.335	.341	**.307**	$+.028\,[+.023, +.033]$
Pythia	L18	.167	.163	.158	**.143**	$+.020\,[+.017, +.022]$
Pythia	L21	.109	.104	.104	**.086**	$+.018\,[+.016, +.020]$
Gemma	L19	.592	.605	.484	**.440**	$+.165\,[+.147, +.183]$
Gemma	L22	.320	.265	.245	**.230**	$+.036\,[+.029, +.042]$

Compact FSR has lower layer-averaged EAURC than both published multi-layer adaptations for each model. Its normalized macro EAURC is .571, versus .706 for trajectory and .775 for SATMD+MSP. Paired improvements are positive in all ten cells against SATMD+MSP and nine against trajectory; trajectory has a lower point value at Gemma L22, with an interval overlapping zero. The matched confidence GBM is stronger at 27B L48 but weaker in the other nine cells.

Rich FSR. Rich GBM already outperforms Compact FSR when averaged over layers. Adding A_{FS} reduces normalized macro EAURC from .490 to .474 (3.1% from unrounded values), with positive point gaps in all ten cells and nine positive intervals. For absolute scale, at 0.5B L18 the Compact FSR, Rich GBM, and Rich FSR EAURCs are .189, .1745, and .1717; at 27B L48 they are .596, .4154, and .4138. These comparisons favor Rich FSR when the additional diagnostics are available.

4.2 What the covariance signal contributes

The ablations in Table 2 ask two questions: does learning how to combine the signals help, and does each signal contribute? Additive FSR uses the same four inputs as Compact FSR; the no-A_{FS} and no-frequency controls retain Compact FSR's boosted architecture.

Compact FSR lowers EAURC by 8.6–28.3% over Additive FSR, with positive paired intervals in all ten cells. Adding A_{FS} to the matched three-input GBM improves EAURC by 4.1–27.3%, again with ten positive intervals. Removing frequency also worsens all ten cells. Both choices matter: A_{FS} contributes beyond $\{\tau, H, \ell_{\mathrm{freq}}\}$, and the learned combination improves on additive weighting.

What the decomposition establishes. Subtracting the isotropic term gives A_{FS} an interpretation as excess spread, not an automatic ranking advantage. In a fixed ten-feature comparison it outperforms the undecomposed Fisher spread $G(p)$ in six of ten cells. Since their difference is proportional to Gini impurity, the present evidence establishes useful covariance information, not the necessity or uniform superiority of isotropic subtraction.

Additive FSR. The covariance gain also appears without a learned combination. Under full-pipeline inference, all ten frequency-controlled gaps are positive (1.5–12.5%), and nine simultaneous intervals exclude zero. Appendices A.1, B.2, and D test the covariance signal across additional depths, ridge values, and readouts.

4.3 Frozen validation and scope

We next evaluate Compact FSR on 128 separate WikiText-2 prompts, without refitting moments, percentile maps, or predictor weights. It improves over Additive FSR by 8.2–27.8% and over the no-A_{FS} control by 4.2–25.8%, with all twenty paired intervals positive. Table 3 shows the complete held-out comparison; Appendix B.3 specifies cohort construction. These validation prompts are text-disjoint from all fitting, calibration, and primary evaluation prompts.

With architecture and other inputs fixed, residual geometry improves rankings beyond the prompts used to develop the score.

Table 3: **Frozen-score validation.** EAURC on 128 disjoint prompts (↓), using the variants in Table 2. Brackets give paired pointwise 95% intervals for adding A_{FS}.

Model	Layer	Additive FSR	No A_{FS}	No freq.	Compact FSR	A_{FS} gap [95% CI]
0.5B	L18	.210	.198	.186	**.182**	$+.016\,[+.012, +.019]$
0.5B	L21	.115	.103	.124	**.095**	$+.008\,[+.006, +.010]$
1.8B	L18	.216	.212	.200	**.189**	$+.024\,[+.018, +.029]$
1.8B	L21	.114	.107	.122	**.092**	$+.015\,[+.013, +.018]$
27B	L48	.586	.562	.567	**.538**	$+.024\,[+.016, +.032]$
27B	L56	.326	.291	.310	**.269**	$+.022\,[+.016, +.028]$
Pythia	L18	.166	.168	.163	**.148**	$+.019\,[+.014, +.024]$
Pythia	L21	.104	.101	.103	**.087**	$+.014\,[+.012, +.017]$
Gemma	L19	.575	.579	.482	**.430**	$+.149\,[+.122, +.176]$
Gemma	L22	.305	.251	.230	**.221**	$+.031\,[+.022, +.040]$

Table 4: **Reliability on fresh evaluations.** Wrong-direction reports (%) at the 10% calibration quantile; Gemma uses pre-softcap contrasts. Gap: per-reading minus global FSR, in percentage points with paired 95% prompt intervals.

Cohort	Estimator	Magnitude	Global FSR	Per-reading	Gap [95%]
Qwen, fresh	Lens	13.78	12.31	11.50	$-.81\,[-1.09, -.52]$
	AtP	1.76	1.45	.07	$-1.37\,[-1.52, -1.24]$
Gemma-2-2B	Lens	25.09	22.90	21.35	$-1.56\,[-2.27, -.84]$
	AtP	6.85	5.76	3.91	$-1.84\,[-2.21, -1.49]$

Other readouts and depths. Compact FSR uses the moment lens at two late depths. With a KL-trained affine lens, the covariance term in Additive FSR helps in all six Qwen/Pythia cells; Gemma is inconclusive. With the moment lens, the gain is also positive in ten early/mid-depth cells ($.25/.50$; Appendices D.1 and D.2).

4.4 Reliable direction reporting

We assess lens and AtP estimates of attention- and MLP-mean ablations at L18/L21 on WikiText-2. Each query compares the clean lens's top two tokens using the linear contrast in Section 3.3: output log-odds for Qwen and pre-softcap logits for Gemma. Both exact and estimated effects use this same target within each model. All reporting rules select from identical estimates. Residual scales use 128 fitting prompts and thresholds use 64 calibration prompts. We compare magnitude, global FSR, and per-reading FSR at the 10% calibration quantile; loss is the wrong-sign frequency among reports. Rates average cells equally, with 2,000 paired prompt-bootstrap resamples.

Reliability beyond estimator accuracy. The matched Qwen comparison gives AtP lower all-reading MAE than the lens (.172 versus .622 nats), while residual scaling improves reporting for both (Appendix C.2). Table 4 tests the complete construction on fresh Qwen prompts and a new reporting model, Gemma-2-2B. The former freezes the existing coefficients and thresholds; the latter uses the same features and fitting procedure with new model-specific fits.

Fresh prompts and a second family. On 512 previously unused paragraphs, per-reading FSR improves reporting for both Qwen estimators. Global/per-reading coverage is $10.50/10.59\%$ for the lens and $10.16/10.04\%$ for AtP. The AtP gain persists at 25% and 50% calibration quantiles (Appendix C.3). On Gemma's 256 test prompts, global FSR improves over magnitude by 2.19 $[1.51, 2.84]$ points for the lens and 1.09 $[.77, 1.42]$ for AtP; conditioning further improves both. Global/per-reading coverage is $11.46/11.08\%$ and $10.15/9.88\%$, respectively. The same fits and thresholds replicate both gains on 683 reserved prompts.

A second intervention. Resample ablation replaces each sublayer output with a donor prompt's output. With new intervention-specific fits and the same scoring procedure, global FSR improves both estimators across the three reporting models. Per-reading scaling lowers AtP's wrong-direction rate from 3.90% to 1.53%, replicating 3.99% to 1.56% on reserved prompts at 10% nominal coverage. The added lens gain does not replicate (Appendix C.4).

What supplies the gain. Global residual scaling accounts for most of the original lens improvement, with pairing adding $.23/.33$ points. In the per-reading model, projected risk remains important for lenses, while response-vector magnitude supplies most of AtP's gain, consistent with size-dependent linearization error. Matched-feature controls on the 512-prompt cohort support this distinction; a direct sign classifier also slightly improves lens reporting (Appendix C.3). Appendices C.1 and C.3 give pairing, effect-floor, and feature ablations.

4.5 Using a reliability score

Reporting and verification. Calibration quantiles specify a reporting fraction. Table 4 measures error among the selected readings, which remains substantial for Gemma lenses despite the improvement. A required error tolerance instead needs risk-targeted calibration and validation on the intended population. Appendix C.3 reports realized coverage and higher-coverage operating points.

Computational scope. Distributional discrepancy is directly available after a completed forward pass; verifying an intervention requires its changed downstream computation. An AtP gradient scores many candidate changes, and global FSR adds no derivative. Per-reading AtP also uses a response-vector derivative (Table 5). Our evidence establishes reporting quality; end-to-end savings under a fixed verification budget remain to be tested.

5 Discussion

Measurement-specific reliability. The two endpoints show why error assessment should follow the question, rather than the hidden state alone. A distributional reading is sensitive to errors between likely tokens; an intervention report is sensitive to errors along its requested contrast. The Fisher and contrast projections express these distinctions using the same residual-moment construction. Matched ablations establish an incremental covariance signal beyond confidence, frequency, and richer diagnostics. The isotropic subtraction defines excess spread; comparisons to undecomposed spread separately test its predictive usefulness.

Estimation and selection. FSR keeps each estimator fixed and changes which readings it reports. The controls separate two useful signals: contrast-specific residual geometry for lenses and response-dependent error scale for AtP. Their contributions depend on the estimator's error structure.

Practical implications. Report the contrast, signed estimate, error scale, and observed selective error together. New contrasts reuse the fitted moments but need a new projection and calibrated threshold. The same reference errors thus support several questions. Comparing fitted scales with query-specific curvature on large candidate sets is a natural next test of budgeted reporting quality.

Limitations. We measure agreement with model outputs, not factual correctness or complete causal mechanisms. Intervention residual fits are specific to the intervention. Direction reporting covers two model families and mean and resample ablations; Gemma's intervention target is pre-softcap. Semantic-contrast evidence remains limited to the CounterFact check in the appendix. The conditional-KL interpretation is limited by global covariance, marginal centering, a local Taylor approximation, and an absolute ridge parameter. Transfer is mixed on code and Gemma. Each model needs fitting and calibration; deployment speedups remain untested.

6 Conclusion

Interpreting an affine reading requires judging its error for the question being asked. FSR does this by projecting fitted residual moments onto that measurement. For token distributions, anisotropic Fisher alignment improves error ranking beyond confidence and richer diagnostics. For interventions, residual normalization improves reporting for both lens and AtP estimates; per-reading scales further improve selection on fresh Qwen prompts, with additional evidence for Gemma's linear contrasts. For a fixed intervention, changing the token contrast changes the projection, not the fitted readout or residual moments. This lets a researcher assess several questions with the same fit and compare reporting rules at an explicit coverage and measured error rate. A diagnostic readout should report both an estimate and its measurement-specific error scale.

AI USE STATEMENT

During the development of FSR, generative AI models assisted with research ideation and refinement, mathematical derivations and proofs, experimental design, implementation, and result analysis. They also supported literature searches, manuscript drafting and revision, and figure and reference preparation. The authors set the research questions and methodology, made the final research decisions, and reviewed the claims, proofs, experimental design, code, results, and manuscript, including AI-assisted material. They approved the final text and take responsibility for its content.

REPRODUCIBILITY STATEMENT

Section 3 defines FSR's residual moments, measurement projections, and reliability scores. Appendix A specifies the data splits, distributional fitting, calibration, and uncertainty estimates; Appendix C gives the settings for direction reporting and per-reading error scales. Assumptions and proofs appear in Appendix E, and the exact Fisher-risk evaluator and covariance identities in Appendix F. The anonymous interactive demo illustrates token-level reliability and presents saved results for direction reporting in a self-contained browser page. Model weights and benchmark data are obtained from their original providers.

ETHICS STATEMENT

FSR assesses agreement with a model's output for a specified measurement. A small predicted error does not make that output true, nor does selective reporting establish a causal mechanism or a system's safety. The scores guide which readings merit interpretation or exact checking; their calibration must match the intended measurement and data. Model and dataset use remains subject to the providers' licenses and access conditions.

REFERENCES

Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: Uncovering decision strategies in LLMs. *arXiv preprint arXiv:2502.16570*, 2025.

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 210–227. PMLR, 2026.

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

Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Aliai Eusebi, Alexander Herzog, Xiaoyu Liang, Marie Vasek, Enrico Mariconti, and Lorenzo Cavallaro. Reading calibrated uncertainty from language model trajectories. *arXiv preprint arXiv:2605.22864*, 2026.

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

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning (ICML)*, 2024.

Paola Gramatica. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, 26(5):694–701, 2007. doi: 10.1002/qsar.200610151.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. URL `https://transformer-circuits.pub/2026/workspace/index.html`.

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

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978. doi: 10.1080/00031305.1978.10479237.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sunwoo Kim, Haneul Yoo, and Alice Oh. On the effect of uncertainty on layer-wise inference dynamics. In *Actionable Interpretability Workshop at ICML*, 2025. arXiv:2507.06722.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. DecoderLens: Layerwise interpretation of encoder-decoder transformers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4764–4780, 2024. doi: 10.18653/v1/2024.findings-naacl.296.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. *arXiv preprint arXiv:2510.00845*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *First Conference on Language Modeling*, 2024.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *International Conference on Machine Learning (ICML)*, 2025.

Geir K. Nilsen, Antonella Z. Munthe-Kaas, Hans J. Skaug, and Morten Brun. Epistemic uncertainty quantification in deep learning classification by the Delta method. *Neural Networks*, 145:164–176, 2022. doi: 10.1016/j.neunet.2021.10.014.

Fedor Noskov, Alexander Fishkov, and Maxim Panov. Selective nonparametric regression via testing. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pp. 1023–1038, 2024.

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

Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron Wallace, Sarah Wiegreffe, Eric Wong, Ian Tenney, and Mor Geva. Interpretability can be actionable. *arXiv preprint arXiv:2605.11161*, 2026. Accepted at ICML 2026.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 548–560, 2023. doi: 10.18653/v1/2023.conll-1.37.

Jordan Pettyjohn, Mansi Sakarvadia, Nathaniel Hudson, Daniel McKenzie, Kyle Chard, and Ian Foster. Interpreting language model hidden states at scale. *arXiv preprint arXiv:2608.10260*, 2026.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 13675–13682, 2021. doi: 10.1609/aaai.v35i15.17612.

Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural network's activations. *arXiv preprint arXiv:2012.03082*, 2020.

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

Qwen Team. Qwen3.6-27B model card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.6-27B.

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

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2246–2262. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.113.

Zi Wang, Alexander Ku, Jason Baldridge, Thomas L. Griffiths, and Been Kim. Gaussian process probes (GPP) for uncertainty-aware probing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shi-Qi Yan, Kai-Xuan Ding, Chao-Hong Tan, Qian Chen, Wen Wang, Xiangang Li, and Zhen-Hua Ling. Short horizons and sparse concepts: A mathematical view of the readout in the J-lens. *arXiv preprint arXiv:2608.25347*, 2026.

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

Mert Yuksekgonul, Linjun Zhang, James Zou, and Carlos Guestrin. Beyond confidence: Reliable models should also consider atypicality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction. *arXiv preprint arXiv:2606.09899*, 2026.

Table 5: **FSR scoring roles.** Costs are additional to the fixed estimator after fitting. Reference labels are endpoint KL or exact intervention residuals; calibration sets the reporting threshold.

Score	Target and inputs	Supervision	Query computation
Compact / Rich	Distributional KL; 4 / 12 diagnostics including A_{FS}	Reference states; calibration KL for the ranker	Fisher statistic and a small boosted predictor
Global scale	Effect sign; projected residual moments and effect magnitude	Exact residuals from reference interventions	Contrast quadratic form; no additional derivative
Per-reading scale	Effect sign; global risk, response/change norms, confidence	Same residuals; conditional second-moment fit	Small scale model; AtP also needs vector $J_c\Delta c$

A Experimental setup and calibration

The appendix provides experimental definitions and supporting evidence. Distributional comparisons and intervention reporting appear in Appendices B and C; Appendices D–F cover generalization, proofs, and exact computation.

A.1 Distributional fitting and calibration

Models and readout target. We evaluate Qwen1.5-0.5B, Qwen1.5-1.8B, Qwen3.6-27B, Pythia-410m, and Gemma-2-2b. Qwen1.5, Pythia, and Gemma run in fp32; Qwen3.6-27B runs in bf16 through its text-only causal-LM path. Its 64-block hybrid Gated-DeltaNet/attention stack is treated as a decoder sequence. Every affine readout maps a same-position state h_ℓ to the final post-norm state and then through the model's unembedding. The per-position error is $D_{\mathrm{KL}}(p_{\mathrm{model}} \| p_{\mathrm{lens}})$. Probed layers follow the single relative-depth rule in Section 4.1.

Data and splits. WikiText-2 prompts contain at least 600 characters, are truncated to 128 tokens, and contribute positions $t \geq 16$. The lens fit uses 448 prompts and calibration uses 64 disjoint prompts. The frequency-controlled primary evaluation uses 256 prompts selected with a fixed seed from the official WikiText-2 test partition; the official test split is not used by any fitting or calibration step. Component and endpoint studies use separate held-out samples. The Python study draws 448/64/256 functions from the official CodeSearchNet train/validation/test partitions, with one function per repository and pairwise repository-disjoint splits.

Risk–coverage metric. For m readings ordered from lowest to highest predicted error by π, the retained risk at coverage k/m is $R_k = k^{-1}\sum_{j=1}^{k} e_{\pi_j}$. We use $\mathrm{AURC} = m^{-1}\sum_{k=1}^{m} R_k$ and subtract the area obtained by sorting the same exact errors to obtain EAURC. Ties use a fixed observation order. Bootstrap resampling preserves within-prompt dependence by drawing entire prompts rather than individual positions.

Additive full-pipeline uncertainty. The point estimate uses the complete 448/64/256 prompt pools. Each of 1,000 bootstrap replicates independently draws prompts with replacement within the fit, calibration, and test pools; the same three draw-index arrays are shared across all models and layers. Every replicate refits lens moments, the fit-only frequency prior, percentile maps, scalar weights, and test EAURCs. The primary A_{FS}-after-frequency contrast uses a standardized max-absolute-deviation simultaneous 95% interval over all ten cells. The model-equal macro first averages the two layers within each model and then the five models.

Comparators and selection. The compact comparison uses MSP, Entropy-Lens, readout entropy, target-layer Mahalanobis distance, and kNN density. The kNN reference fixes $k = 10$ and a seeded 16,384-state subsample of fitting states. The learned reference is a depth-3 gradient-boosted model over ten deployable diagnostics spanning state geometry, lens confidence, cross-lens disagreement, position, and density. Token frequency is the eleventh non-covariance feature and A_{FS} the twelfth (Appendix B.1). The cross-readout study trains its KL-trained affine lens only on fitting prompts, with early stopping on a held-out subset of those prompts. Empirical percentile maps and w^*, η^*, γ_A^*

are fitted on calibration prompts. The primary ten-cell, cross-readout, code, and nonlinear comparator studies evaluate A_{FS} exactly by Equation 8.

Covariance stability. Refitting on disjoint halves of the fitting prompts gives split-half Spearman .993–.996 for the cached $\mathrm{diag}(C)$ and .854–.992 for A_{FS} rankings; the lower values occur for Gemma and 27B. A label-free isotropic shrinkage check selects the unshrunk covariance in eight of ten cells; we retain the unshrunk S_λ throughout.

Ridge and calibration sensitivity. The covariance increment remains positive in all ten cells at ridge values $10^{-3}, 10^{-2}, 10^{-1}$, with positive conditional prompt-bootstrap intervals. Reducing calibration from 64 to 32 or 16 prompts raises model-cell-equal EAURC by median .00077 or .00216, respectively. The tested ranges support local robustness to the fitting and calibration choices.

Compute. Reliability experiments ran on NVIDIA B200 GPUs. The 1,000-replicate full-pipeline analysis used 8.52 B200-hours with peak memory of 59.6 GiB.

B Distributional reliability and ablations

B.1 Baselines, components, and learned comparators

Single-state controls. MSP ranks the maximum lens probability; Entropy-Lens uses the entropy obtained by applying the final normalization and unembedding directly to the source state. Readout entropy instead uses the fitted moment lens. Token MD is the squared Mahalanobis radius τ; kNN density is the mean Euclidean distance to ten reference states. Low confidence and large distance indicate greater predicted error.

Rich GBM uses eleven inputs: τ, moment-lens entropy, its top-two logit gap and maximum log probability, raw-logit-lens entropy, the KL from the raw logit lens to the moment lens, distance to the fitting-state mean, source-state norm, token position, kNN distance, and ℓ_{freq}. Each input is mapped to calibration percentiles. Rich GBM and Rich FSR use the same boosted regressor settings as Compact FSR; only Rich FSR receives A_{FS}. Thus the richer comparison isolates covariance information without changing the predictor or its fitting examples.

Published multi-layer comparators. SATMD+MSP (Vazhentsev et al., 2025) combines token Mahalanobis distances from $h_1, \ldots, h_k$, ten principal components, and maximum lens probability in a linear predictor. Its reference population is the lower-KL half of the 448-prompt fitting split, with ridge 0.01; PCA and regression use the 64 calibration prompts. Our adaptation of the trajectory probe (Eusebi et al., 2026) uses ten geometric features per layer along cumulative MLP outputs, with maximum-probability interactions. We truncate the path at k and use its available endpoint, so neither score observes the suffix. We adapt both response-level correctness predictors to per-reading $\log(1 + \mathrm{KL})$; the trajectory probe uses elastic-net regression, selecting its 64 penalty configurations by four-fold prompt-level cross-validation within calibration. Both are frozen before test scoring. These methods retain prefix history, whereas FSR scores a single state. Table 6 reports all ten cells in the multi-layer group; paired intervals condition on the fitted predictors.

Table 6 gives raw cellwise EAURCs underlying the compact and richer-input comparisons in Section 4.1. For the eleven-to-twelve-feature covariance increment, paired intervals are positive in nine cells and overlap zero at 27B L56.

Factorization check. In a fixed ten-feature GBM ablation, A_{FS} has lower point EAURC than undecomposed Fisher spread $G(p)$ in six of ten cells. The two statistics differ by an isotropic Gini term, which overlaps with the ambiguity information in entropy but is not determined by it. The decomposition expresses A_{FS} as signed excess residual spread along likely token contrasts.

Covariance beyond the four-input score. We also test A_{FS} in the sparse trajectory predictor, using cells selected after inspecting the comparison. On 27B L48/L56 and Gemma L19/L22, adding its percentile feature while keeping the calibration-selected penalty and mixing ratio fixed lowers EAURC by 1.0–5.9%; all four conditional pointwise intervals are positive.

Table 6: **Distributional reliability comparators.** Raw EAURC (↓) on identical model–layer cells. Bold marks column-wise minima in the first two groups only. Additive FSR includes frequency; Rich FSR adds A_{FS} to Rich GBM. Multi-layer methods use prefix history.

	Qwen1.5 0.5B		Qwen1.5 1.8B		Qwen3.6 27B		Pythia 410M		Gemma-2 2B	
Method	L18	L21	L18	L21	L48	L56	L18	L21	L19	L22
Reference scores and additive FSR										
MSP	.3202	.2590	.3481	.2471	.8701	.5674	.3116	.2116	1.0047	.5449
Entropy-Lens	.6925	.3325	.5686	.2911	1.4776	.8464	.2893	.1929	.8757	.4438
Readout entropy	.2663	.2405	.2965	.2293	.6776	.4837	.2707	.1947	.8690	.4799
Token MD	.4097	.1755	.4163	.2147	2.2637	.9536	.2555	.1633	.8326	.4172
kNN	.4191	.1934	.4779	.3082	2.2032	1.3830	.3294	.2914	1.1135	.5594
Additive FSR (ours)	**.2160**	**.1169**	**.2348**	**.1221**	**.6514**	**.3663**	**.1667**	**.1087**	**.5919**	**.3199**
Rich learned predictors										
Rich GBM	.1745	.0906	.2036	.0867	.4154	.2515	.1339	.0699	.3629	.1892
Rich FSR (ours)	**.1717**	**.0877**	**.1940**	**.0836**	**.4138**	**.2509**	**.1267**	**.0675**	**.3374**	**.1845**
Adapted multi-layer predictors										
SATMD+MSP (2025)	.2248	.1449	.2456	.1628	.8247	.4807	.1707	.1323	.5769	.2801
Trajectory probe (2026)	.2635	.1463	.2788	.1534	.6409	.3257	.1788	.1018	.5160	.2208

B.2 Additive component ablation

Radial–entropy base. We first combine two signals available from the reading itself:

$$s_w(x) = w\,\widehat{F}_\tau(\tau(x)) + (1-w)\,\widehat{F}_H(H[p_{\mathrm{lens}}(\cdot \mid x)])\,, \qquad w \in \{0, .05, \ldots, 1\}, \tag{15}$$

where $\widehat{F}_\tau$ and $\widehat{F}_H$ are empirical percentile maps on a prompt-disjoint calibration set and w minimizes calibration AURC. It uses only the geometry score and the lens reading's own entropy: no observed next token or final-model output is required. Equation 15 is the *base rule*: it combines radial atypicality with readout entropy.

Additive reference. A transparent nested rule adds the anisotropic term and nests the radial–entropy base at $\gamma_A = 0$:

$$\boxed{s_{\mathrm{add}}(x) = (1-\gamma_A)\,s_w(x) + \gamma_A\,\widehat{F}_A\big(A_{\mathrm{FS}}(p(\cdot \mid x))\big), \qquad \gamma_A \in \{0, .05, \ldots, 1\},} \tag{16}$$

where $\widehat{F}_A$ is the calibration percentile map of A_{FS} and γ_A minimizes calibration AURC. We denote this two-weight reference Additive FSR w/o freq.; its nested structure separates radial, confidence, and covariance information.

Frequency enters through $s_{w,\eta} = (1-\eta)s_w + \eta\widehat{F}_L(\ell_{\mathrm{freq}})$. Replacing s_w by $s_{w,\eta}$ in Equation 16 and refitting the last weight gives Additive FSR, with $\eta \in \{0, .05, \ldots, 1\}$.

The additive reference selects weights sequentially (w, then η, then γ_A) on the calibration pool. Table 7 isolates the covariance increment without changing predictor capacity. Unlike the learned-score intervals, these simultaneous intervals include uncertainty from refitting lens moments, frequency counts, maps, and scalar weights.

B.3 Frozen learned-score validation

Compact FSR uses $\{\tau, H, \ell_{\mathrm{freq}}, A_{\mathrm{FS}}\}$; removing A_{FS} or frequency changes only the input set. Every predictor uses histogram gradient boosting with squared loss on $\log(1 + \mathrm{KL})$, depth 3, 200 iterations, learning rate .05, seed 0, and no early stopping. Feature percentiles use only the 64 calibration prompts. All ten cells use the same settings, fixed before the additional-cohort evaluation.

We select the first 128 official WikiText-2 validation entries with more than 600 characters and verify text-level disjointness from the 448/64/256 primary pools. All learned heads, percentile references, and additive weights are reused unchanged. Table 3 compares the compact, additive, and feature-ablation scores without additional tuning.

Table 7: **Additive component ablation** (EAURC; lower is better). Additive FSR w/o freq. adds A_{FS} to the radial–entropy base; Additive FSR adds it after frequency control. Brackets give simultaneous 95% intervals for this increment; bold marks the best additive rule per cell.

Model	Layer	τ	H	Base	Additive FSR w/o freq.	Base + freq.	Additive FSR	A_{FS} gap [95% CI]
0.5B	L18	.4097	.2663	.2193	**.2160**	.2193	**.2160**	$+.0033\,[+.0003, +.0063]$
0.5B	L21	.1755	.2405	.1477	.1391	.1251	**.1169**	$+.0082\,[+.0048, +.0116]$
1.8B	L18	.4163	.2965	.2453	**.2348**	.2446	**.2348**	$+.0099\,[+.0037, +.0160]$
1.8B	L21	.2147	.2293	.1597	.1417	.1342	**.1221**	$+.0121\,[+.0085, +.0157]$
27B	L48	2.2637	.6776	.6731	**.6514**	.6731	**.6514**	$+.0217\,[+.0075, +.0359]$
27B	L56	.9536	.4837	.4213	.3873	.4116	**.3663**	$+.0453\,[+.0233, +.0673]$
Pythia-410m	L18	.2555	.2707	.1827	.1703	.1802	**.1667**	$+.0135\,[+.0070, +.0200]$
Pythia-410m	L21	.1633	.1947	.1296	.1168	.1240	**.1087**	$+.0153\,[+.0104, +.0201]$
Gemma-2-2b	L19	.8326	.8690	.7214	.6753	.6766	**.5919**	$+.0847\,[+.0438, +.1257]$
Gemma-2-2b	L22	.4172	.4799	.3524	.3494	.3316	**.3199**	$+.0117\,[-.0020, +.0253]$

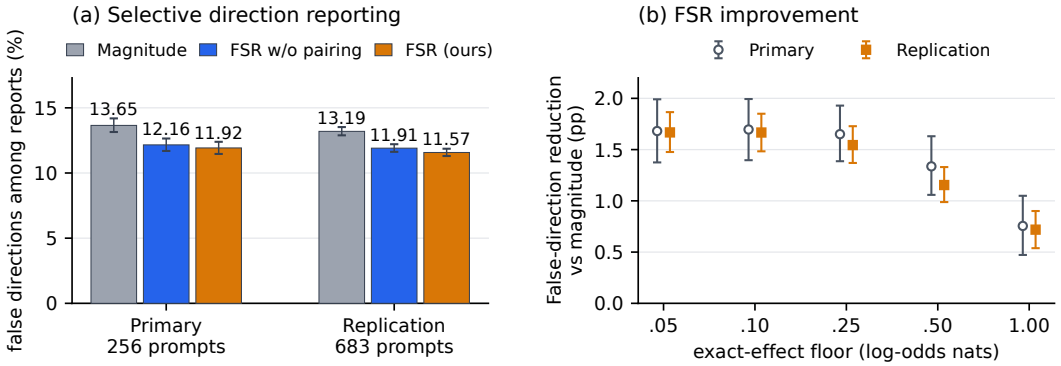

Figure 2: **Selecting intervention directions to report. (a)** FSR, its no-pairing ablation, and magnitude at a frozen 10% calibration quantile. **(b)** FSR's false-direction reduction over magnitude after exact-effect filtering. Both average eight cells; whiskers are 95% prompt-bootstrap intervals. Replication reuses the same fits and thresholds.

All additive and no-A_{FS} contrasts have positive paired intervals. The frequency interval overlaps zero only at Gemma L22. Intervals use 2,000 prompt bootstraps conditional on the fitted predictors.

C Intervention Reliability and Reporting Controls

C.1 Intervention Reliability and Decision Scope

Section 4.4 defines the query, moment construction, split sizes, and main comparison. The resulting second moment is $U_\Delta(q)$ in Equation 13. The independence ablation replaces S_Δ by $S_0 + S_1$ and retains the same mean-difference term. Maximum endpoint Additive FSR is the endpoint control. All comparisons evaluate identical held-out intervention pairs.

Effect-risk ranking. In Table 8, paired risk improves over endpoint risk in all eight cells and over independence in seven; the 0.5B L18 MLP interval includes zero. Magnitude alone is stronger for this loss; its blend with paired risk improves by 8.9–19.1% in all eight cells. The blend combines calibration percentiles of $|\widehat{\delta}|$ and U_Δ, selecting $\gamma \in \{0, .05, \ldots, 1\}$ by calibration EAURC and retaining low scores. For direction reporting, large magnitude is reassuring: its control ranks $1/(|\widehat{\delta}| + 10^{-6})$, while paired geometry uses Equation 14.

Selective direction reporting. For directional reporting, the scale-free score ρ_Δ uses calibration quantiles without fitting another risk model. Directional metrics exclude effects with estimated or exact absolute magnitude at most 10^{-8}, for which numerical sign is undefined. The false-direction rate is computed among reports within each cell, then averaged equally over the eight cells; paired bootstrap draws resample the same prompts across all cells. Its reduction over magnitude alone

Table 8: **Paired-risk ablation.** Absolute log-odds error EAURC (↓). Endpoint, independent, and paired scores isolate covariance design; the last two columns test its addition to magnitude.

Model	Layer	Ablation	Endpoint	Indep.	U_Δ	Magnitude	Mag. $+U_\Delta$
0.5B	L18	Attn.	.307	.237	.215	.139	**.119**
0.5B	L18	MLP	.335	.261	.261	.242	**.204**
0.5B	L21	Attn.	.369	.224	.138	.078	**.071**
0.5B	L21	MLP	.301	.264	.254	.182	**.147**
1.8B	L18	Attn.	.318	.190	.178	.130	**.105**
1.8B	L18	MLP	.405	.296	.285	.239	**.200**
1.8B	L21	Attn.	.276	.169	.153	.102	**.086**
1.8B	L21	MLP	.510	.334	.315	.192	**.167**

replicates on the 683 official-test prompts excluded from the original test set. A diagnostic applies fixed exact-effect floors of .05, .10, .25, .50, and 1.00 log-odds nats after selection; the paired-versus-magnitude interval remains positive at every floor on both test sets. At .25 nats, magnitude/paired coverage is $16.50/16.52\%$ on the primary cohort and $16.62/16.64\%$ on replication.

With the original lens estimates fixed, the global paired score reduces wrong-direction reports from 13.65% to 11.92% on the primary cohort (gap 1.73 points, $[1.41, 2.06]$) and from 13.19% to 11.57% on replication (1.62 points, $[1.43, 1.82]$). The independence control reaches 12.16% on the primary cohort; pairing adds $.23/.33$ points across the two cohorts. At the .25-nat effect floor, the paired advantage over magnitude remains $1.65/1.54$ points. These values underlie Figure 2.

Appendix C.3 extends the per-reading comparison to CounterFact object contrasts.

C.2 Matched WikiText attribution-patching reports

The matched direction-reporting task uses both estimators: Qwen1.5-0.5B and 1.8B, L18/L21 attention and MLP mean ablations at all positions, the clean lens's top-two contrast, and the same primary and reserved prompt splits. Residual moments use the first 128 of the lens's 448 fitting prompts. Separate, disjoint sets contain 64 calibration prompts, 256 primary test prompts, and 683 replication prompts. Both cohorts use the same fits and calibration thresholds. Mean ablation replaces a sublayer's output at every position with its reference mean, estimated by averaging positions within each residual-fit prompt and then averaging prompts. AtP uses the final-state directional derivative $J_c\Delta c$, computed by automatic differentiation with eager attention, and its residual $\Delta y - J_c\Delta c$ supplies the mean and covariance. The magnitude rule ranks $1/(|\widehat{\delta}| + 10^{-6})$; FSR uses Equation 14. Intervals use 2,000 paired prompt-bootstrap draws, shared across cells, with rates averaged equally over the eight cells.

C.3 Conditioning the error scale on the reading

Fit and information available at query time. The per-reading score retains the effect estimator and learns only its projected squared-error scale. On WikiText, ϕ contains an intercept, $\log U_\Delta(q)$, $\log \|\widehat{\Delta y}\|$, the log change norm, and the clean lens's top-two log-probability gap. The change norm is $\|\Delta h\|$ for a lens and the intervention-position $\|\Delta c\|$ for AtP. Inputs are standardized on the 128 residual-fit prompts. With residual $e_i = \delta_i - \widehat{\delta}_i$, the five coefficients minimize a zero-mean Gaussian working loss by Newton updates with backtracking:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_i \left[e_i^2 e^{-\phi_i^\top \beta} + \phi_i^\top \beta \right] + 10^{-3} \|\beta_{\text{nonint}}\|_2^2, \qquad \rho_i^{\text{cond}} = \frac{\exp(\phi_i^\top \widehat{\beta}/2)}{|\widehat{\delta}_i| + \epsilon}. \tag{17}$$

This estimates a conditional second moment, including squared bias, rather than assuming the residual is centered. No exact intervened output is used as a scoring feature. The lens uses its estimated response and changed source state. AtP uses the full directional derivative $J_c\Delta c$, implemented by the double-backward identity with eager attention. Its norm is not recoverable from one scalar contrast gradient; this extension therefore has extra derivative cost in that setting.

Table 9: **Error-scale controls.** Wrong-direction rate and coverage (%) at the 10% calibration quantile. Retrospective controls on the 512-prompt Qwen cohort, with fitting and calibration data unchanged.

	Lens		AtP	
Score	Wrong ↓	Coverage	Wrong ↓	Coverage
Magnitude	13.780	10.42	1.764	9.64
Global scale	12.307	10.50	1.447	10.16
Response norm	13.564	10.36	.063	9.56
Response + change norms	13.894	10.43	.060	9.71
Norms + projected risk	11.545	10.66	.087	10.09
Full per-reading scale	11.498	10.59	.074	10.04
Direct sign classifier	10.966	10.74	.085	10.05

Fresh-cohort confirmation. The Qwen scale coefficients, standardization, and thresholds at 10%, 25%, and 50% were frozen before collecting 512 WikiText-2 validation paragraphs. Eligible paragraphs are sampled without text overlap with the 1,451 fitting, calibration, and earlier evaluation prompts. The primary endpoint is the per-reading-minus-global wrong-direction gap at the 10% calibration quantile, separately for each estimator. Both paired 95% intervals exclude zero (Table 4). Table 10 reports higher coverage levels, with the same $\epsilon = 10^{-6}$ and numerical sign-validity filter.

Generalization to Gemma. Gemma-2-2B uses the same L18/L21 attention and MLP ablations, 448 lens-fit prompts, 128 residual-fit prompts, 64 calibration prompts, and feature and fitting rules. Its 256 test and 683 reserved prompts had not been used for reporting on this model; model-specific fits and thresholds are fixed before evaluation. Rates average its four cells. The global-versus-magnitude and per-reading-versus-global intervals exclude zero for both estimators on both cohorts. On reserved prompts, global FSR reduces lens and AtP error by 2.02 $[1.66, 2.37]$ and $.95$ $[.73, 1.13]$ points; conditioning adds 1.86 $[1.43, 2.27]$ and 1.81 $[1.59, 2.03]$ points. This tests the same construction on a new family, not transfer of Qwen's fitted coefficients. The endpoint is the final normalized state followed by linear unembedding, before Gemma's $g(z) = 30\tanh(z/30)$ output softcap. The model's internal computation, including attention softcapping, is unchanged. Both the exact contrast and AtP derivative stop before the output softcap. These results therefore concern pre-softcap contrast changes; their signs need not equal changes in the model's output log-odds. The distributional experiment instead includes the output transform and its Fisher pullback.

Feature contributions. The feature-removal analysis on the earlier Qwen cohorts is retrospective for the conditional score. Descriptive feature removal shows that the response-vector norm nearly recovers the full AtP gain by itself, while projected risk is important for the lens. The two estimators therefore benefit from different components of the same scale model.

Matched-feature controls. Table 9 evaluates these explanations on the 512-prompt Qwen cohort, which was already viewed before designing these controls. All coefficients use only the original 128 residual-fit prompts, and the 64 calibration prompts set each 10% threshold. Norm-only models retain Equation 17's loss, ridge, standardization, and magnitude denominator while changing only its inputs. The direct sign classifier uses the full four scale features plus $\log|\widehat{\delta}|$, fitting logistic loss to wrong-sign labels with ridge 10^{-3} on non-intercept coefficients. There is no evaluation-set tuning. Intervals use 2,000 paired prompt draws, conditional on these fits and thresholds, with equal weight across eight cells.

For lenses, adding projected risk to the two norms lowers wrong-direction reports by 2.35 $[2.09, 2.60]$ points. Thus norm variation alone does not explain the lens gain. For AtP, response norm alone matches the full model within uncertainty; the large reduction is attributable to conditional scale estimation, not a unique benefit of covariance geometry. The direct classifier improves on the full lens score by $.53$ $[.23, .82]$ points, while their AtP interval overlaps zero. It uses an explicitly sign-specific target rather than the reusable squared-error target. All comparisons use nominally equal coverage, with the realized values shown rather than assumed identical.

CounterFact evaluation. We use true-versus-counterfactual single-token object contrasts (Meng et al., 2022) on Qwen1.5-1.8B and Pythia-410M. Attention/MLP mean replacement acts at the final prompt position at L18/L21, without filtering the model's initial correctness. Subject-disjoint roles contain 448 lens-fit, 256 residual-fit, 512 development, 512 calibration, and 2,048 evaluation facts.

Table 10: **Coverage and replication.** Wrong-direction rate / realized coverage (%) at fixed calibration quantiles. Gaps compare per-reading with global FSR, in percentage points with paired 95% intervals.

Cohort	Estimator	Quantile	Global FSR	Per-reading	Gap [95%]
Qwen, fresh	Lens	25%	16.95 / 25.56	16.19 / 25.56	$-.76$ $[-.92, -.60]$
	Lens	50%	22.19 / 50.66	21.67 / 50.34	$-.52$ $[-.62, -.43]$
	AtP	25%	1.62 / 25.27	.17 / 25.27	-1.45 $[-1.56, -1.35]$
	AtP	50%	2.21 / 50.47	.71 / 50.20	-1.50 $[-1.58, -1.42]$
Gemma, reserved	Lens	10%	23.36 / 11.05	21.50 / 10.57	-1.86 $[-2.27, -1.43]$
	AtP	10%	6.06 / 10.24	4.25 / 9.86	-1.81 $[-2.03, -1.59]$

Table 11: **Per-reading scales on CounterFact.** Wrong-direction rate / realized coverage (%) at the 10% calibration quantile. Gaps are per-reading minus global risk in percentage points, with paired 95% subject intervals.

Model	Estimator	Global scale	Per-reading scale	Gap [95%]
Qwen1.5-1.8B	Lens	20.40 / 11.36	17.45 / 10.56	-2.95 $[-5.26, -.71]$
Pythia-410m	Lens	6.48 / 10.08	4.96 / 9.89	-1.52 $[-2.83, -.22]$
Qwen1.5-1.8B	AtP	.39 / 10.11	.00 / 9.95	$-.39$ $[-.88, .00]$
Pythia-410m	AtP	.00 / 9.11	.00 / 8.79	.00 $[.00, .00]$

Table 12: **Resample-ablation reporting.** Wrong-direction rate / coverage (%); twelve-cell means at 10% nominal coverage, with pre-softcap Gemma targets. Gaps compare with magnitude and global FSR (points; paired recipient-prompt 95% intervals).

Cohort	Estimator	Magnitude	Global	Per-reading	Global gap	Cond. gap
Test	Lens	15.24 / 10.76	14.36 / 10.80	14.08 / 10.87	$-.88$ $[-1.15, -.62]$	$-.28$ $[-.52, -.07]$
Reserved	Lens	15.44 / 10.57	14.26 / 10.56	14.14 / 10.58	-1.18 $[-1.35, -1.01]$	$-.13$ $[-.26, .01]$
Test	AtP	4.77 / 9.90	3.90 / 10.26	1.53 / 9.81	$-.87$ $[-1.06, -.70]$	-2.37 $[-2.57, -2.17]$
Reserved	AtP	4.71 / 9.77	3.99 / 10.10	1.56 / 9.88	$-.72$ $[-.82, -.62]$	-2.43 $[-2.56, -2.30]$

The estimators are an affine state lens and the clean-state AtP derivative; the comparison changes only their error scales. The CounterFact comparison refits the same scale model using each cell's residual-fit subjects and sets thresholds on separate calibration subjects. This version uses the three log features and an intercept, without clean confidence, and $\epsilon = 10^{-8}$. Features and fitting rules were fixed before re-analysis of the existing evaluation cohort, so this tests cross-task applicability with new fits rather than independent replication. At the pre-specified 10% calibration quantile, lenses improve in both models (Table 11). The AtP comparison is inconclusive: its interval reaches zero in Qwen and both scores make no errors in Pythia.

C.4 Reporting under resample ablation

We apply the same scoring construction to resample ablation of attention and MLP outputs at L18/L21 in Qwen1.5-0.5B, Qwen1.5-1.8B, and Gemma-2-2B. Each prompt receives the corresponding sublayer output of the next prompt in its role's cyclic order. Positions beyond the donor's length stay clean. The lens uses the same 448 fitting prompts; residual moments and per-reading coefficients are refitted on 128 resample-ablation prompts, and 64 separate calibration prompts set new thresholds. Thus this tests a second intervention with new fits, not transfer of mean-ablation coefficients. The 256 test and 683 reserved texts are reused from the earlier reporting study; their resample-ablation outcomes are new.

The primary comparisons use the 10% calibration quantile, separately for each estimator, with equal weight across twelve model–layer–module cells. Qwen cells use output log-odds and Gemma cells use pre-softcap contrasts, as defined in Section 3.3; the pooled rate combines within-cell sign errors, not common-scale effect magnitudes. Intervals resample recipient prompts jointly across cells, keeping the cyclic donor assignments fixed. Global scaling improves both estimators on both cohorts (Table 12). Per-reading scaling further improves AtP, while its smaller lens gain does not

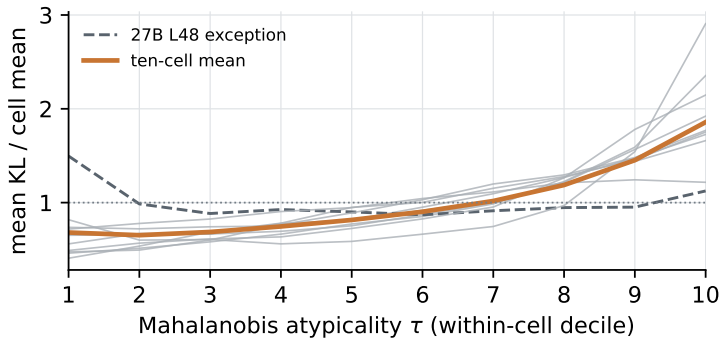

Figure 3: **Empirical Mahalanobis-tail diagnostic.** Curves show decile mean KL divided by the cell-wide mean KL. Gray curves are individual cells; orange is their equal-weight mean, and dashed 27B L48 is the exception.

Table 13: **Early/mid-depth stress test** (EAURC; lower is better). Depth is the fraction of completed blocks. Gaps and 95% intervals compare the frequency-controlled base with Additive FSR under conditional prompt bootstrap.

Model	Depth (layer)	Base	Additive FSR	Gap [95% CI]	Rel.
0.5B	.25 (L6)	.7311	.7153	$+.0158\,[+.0117, +.0201]$	2.2%
0.5B	.50 (L12)	.5205	.5011	$+.0194\,[+.0132, +.0251]$	3.7%
1.8B	.25 (L6)	.8743	.8315	$+.0428\,[+.0317, +.0532]$	4.9%
1.8B	.50 (L12)	.6076	.5884	$+.0192\,[+.0141, +.0240]$	3.2%
27B	.25 (L16)	.9882	.9698	$+.0183\,[+.0120, +.0245]$	1.9%
27B	.50 (L32)	.9754	.9686	$+.0067\,[+.0042, +.0093]$	0.7%
Pythia	.25 (L6)	.8230	.6602	$+.1627\,[+.1388, +.1849]$	19.8%
Pythia	.50 (L12)	.5372	.4578	$+.0794\,[+.0655, +.0928]$	14.8%
Gemma	.23 (L6)	1.8543	1.4621	$+.3922\,[+.3477, +.4328]$	21.2%
Gemma	.50 (L13)	1.7637	1.3470	$+.4167\,[+.3805, +.4522]$	23.6%

replicate on reserved prompts. At 25% and 50% nominal coverage, global scaling still improves both estimators and conditioning still improves AtP; conditioning slightly worsens the lens point estimates. A descriptive circular-block bootstrap with blocks of two, four, or eight adjacent recipient pairs checks donor-sharing dependence; all interval conclusions at 10% nominal coverage are unchanged. These results extend the global reporting rule to resample ablation and locate the consistent conditional gain in AtP estimates.

D Generalization across domains, depths, and readouts

D.1 Additional single-state evaluations

The domain, depth, and cross-readout checks below use the additive score to isolate the covariance increment; they do not refit the learned readout.

On repository-disjoint CodeSearchNet Python, only 1.8B L18 retains A_{FS}: its EAURC improvement is $+.00387$ (95% CI $[+.00201, +.00546]$). The other three cells assign zero weight to covariance.

Radial tail diagnostic. Figure 3 bins held-out readings by τ within each cell. Nine cells have positive rank association and higher normalized mean KL in the top than bottom decile, with positive prompt-bootstrap intervals; 27B L48 is a clear exception. It also has the worst τ-only EAURC in Table 7, whereas L56 has a positive tail trend. The radial pattern motivating Theorem 1 therefore holds in most cells, but is not uniform across depths.

Table 14: **Transfer to a KL-trained affine lens.** Non-cov. is radial–entropy–frequency; Additive FSR adds centered A_{FS}. A positive gap favors Additive FSR; partial association conditions on $\{\tau, H, \ell_{\mathrm{freq}}\}$.

Model	Layer	non-cov.	Additive FSR	γ_A	gap [95% CI]	partial(A_{FS})
0.5B	L18	.1398	**.1361**	.05	$+.00371\,[+.00291, +.00451]$	$+.177$
0.5B	L21	.0643	**.0599**	.10	$+.00442\,[+.00375, +.00508]$	$+.237$
1.8B	L18	.1681	**.1580**	.10	$+.01018\,[+.00816, +.01197]$	$+.230$
1.8B	L21	.0724	**.0634**	.15	$+.00897\,[+.00774, +.01009]$	$+.284$
Pythia	L18	.1104	**.0997**	.15	$+.01062\,[+.00902, +.01228]$	$+.230$
Pythia	L21	.0607	**.0473**	.20	$+.01340\,[+.01207, +.01488]$	$+.363$
Gemma	L19	.3680	.3671	.05	$+.00089\,[-.00088, +.00261]$	$+.062$
Gemma	L22	**.1899**	.1899	.00	$.00000\,[.00000, .00000]$	$-.046$

D.2 Cross-readout transfer

For an empirical-covariance readout, A_{FS} must use $\mathrm{Cov}(r)$ rather than the uncentered second moment $\mathbb{E}[rr^\top]$. We estimate the fit-split residual mean and covariance jointly for a separately fit, Tuned-Lens-style KL-trained affine map $Ah_k + b$ into the post-final-norm target space (Belrose et al., 2023). The moment readout already supplies centered residual geometry through its affine intercept and Schur covariance. Gemma's position-dependent softcap Jacobian is included in the Fisher pullback, retaining the same $O(V)$ cache and exact closed form.

Table 14 tests centered A_{FS} after the radial–entropy–frequency rule. Six of eight paired intervals are positive, all in Qwen/Pythia; Gemma L19 is inconclusive and L22 selects zero weight. Both Gemma moment-lens cells have positive point gaps, showing that the covariance contribution depends on both the model and its readout fit.

E Theoretical results and proofs

Geometry and readout error.

Measurement geometry specifies which residual directions matter; source-state geometry identifies where an affine approximation is strained. Under a quadratic conditional mean and Gaussian whitened states, the population-best affine predictor's conditional risk decomposes as $R_A(t) = V_A(t) + B_A(t)$, with $B'_A(t) \geq 0$ for Mahalanobis radius $t \geq d$. Thus curvature increases tail error even at constant residual spread (Theorem 1). For ordinary softmax, the exact KL integral gives $\mathbb{E}[\mathrm{KL} \mid \tau = t] \leq \frac{1}{4} R_{W_U^\top P_V W_U}(t)$ (Proposition 2). These results motivate radial screening; A_{FS} additionally resolves which token contrasts carry residual spread. For intervention effects, Equation 13 is the exact projected second moment under the fitted residual model. Appendix E gives the proofs.

E.1 Proposition 1

Since $F(p)\mathbf{1} = 0$, $F(p)P_V = P_V F(p) = F(p)$. The mean–covariance identity therefore gives

$$\mathbb{E}[r^\top F(p) r] = b^\top F(p) b + \mathrm{tr}\{F(p)C\} = (P_V b)^\top F(p)(P_V b) + \mathrm{tr}\{F(p)C_\perp\}.$$

Substituting $C_\perp = \alpha P_V + D$ and using $\mathrm{tr}\{F(p)P_V\} = \mathrm{tr}\, F(p) = 1 - \|p\|_2^2$ proves Equation 7. For uniqueness, suppose also that $C_\perp = aP_V + E$, where $P_V E P_V = E$ and $\mathrm{tr}(E) = 0$. Taking traces yields $a = \mathrm{tr}(C_\perp)/(V-1) = \alpha$, hence $E = D$. Finally, $\langle P_V, D\rangle_F = \mathrm{tr}(P_V D) = \mathrm{tr}(D) = 0$. If $r' = r + \xi\mathbf{1}$ for any square-integrable scalar random variable ξ, then $P_V r' = P_V r$ samplewise, proving the stated invariance. □

E.2 Population tail-risk theorem

Theorem 1 (Curvature–spread tail risk)**.** *Let $u = \Sigma_{xx}^{-1/2}(x - \mu_x) \sim \mathcal{N}(0, I_d)$ and $\tau = \|u\|^2$. Suppose $y \in \mathbb{R}^m$ satisfies $\mathbb{E}\|y\|_2^2 < \infty$ and has conditional covariance $\Omega(u)$ and conditional mean*

$$\mathbb{E}[y \mid u] = a + L_0 u + \tfrac{1}{2}\kappa(u), \qquad \kappa_j(u) = u^\top H_j u, \tag{18}$$

where each H_j is symmetric. For any fixed output metric $A \succeq 0$, let m_{aff} be the population L^2-best affine predictor and define

$$R_A(t) = \mathbb{E}[\|y - m_{\mathrm{aff}}(u)\|_A^2 \mid \tau = t], \qquad V_A(t) = \mathbb{E}[\mathrm{tr}\{A\Omega(u)\} \mid \tau = t].$$

Then

$$R_A(t) = V_A(t) + B_A(t), \qquad B_A'(t) \geq 0 \quad \textit{for } t \geq d, \tag{19}$$

where B_A is the exact A-weighted squared affine bias, given in Equation 20 below. The inequality is strict for $t > d$ whenever curvature is visible through A. Consequently, if $V_A(t)$ is nondecreasing, $R_A(t)$ is nondecreasing throughout the Mahalanobis tail. In the scalar radial case $\Omega(u) = s(\tau)\Omega_0$, the spread term reduces to $V_A(t) = s(t)\,\mathrm{tr}(A\Omega_0)$.

Proof. Write $h_j = \mathrm{tr}(H_j)$, collect these traces in $h \in \mathbb{R}^m$, and define the curvature Gram matrix $\Gamma \in \mathbb{R}^{m\times m}$ by $\Gamma_{jk} = \mathrm{tr}(H_jH_k)$. Gaussian odd moments vanish, so the quadratic component is orthogonal to every linear function of u, while $\mathbb{E}[\kappa(u)] = h$. The population-best affine predictor is therefore

$$m_{\mathrm{aff}}(u) = a + L_0u + \tfrac{1}{2}h.$$

Let $\varepsilon = y - \mathbb{E}[y \mid u]$ and $b(u) = \frac{1}{2}\{\kappa(u) - h\}$. Conditional mean zero of ε makes the cross term vanish, giving the exact risk decomposition

$$R_A(t) = \mathbb{E}[\varepsilon^\top A\varepsilon \mid \tau = t] + \mathbb{E}[b(u)^\top Ab(u) \mid \tau = t] = V_A(t) + B_A(t).$$

Conditional on $\|u\|^2 = t$, $u/\sqrt{t}$ is uniform on the unit sphere. Its second and fourth moments give

$$\mathbb{E}[\kappa_j(u) \mid t] = \frac{t}{d}h_j, \qquad \mathbb{E}[\kappa_j(u)\kappa_k(u) \mid t] = \frac{t^2}{d(d+2)}\{2\Gamma_{jk} + h_jh_k\}.$$

Consequently, with $\gamma = \mathrm{tr}(A\Gamma)$ and $\eta = h^\top Ah$,

$$B_A(t) = \frac{1}{4}\left[\frac{t^2}{d(d+2)}(2\gamma + \eta) - \frac{2t}{d}\eta + \eta\right]. \tag{20}$$

Its derivative can be written

$$B_A'(t) = \frac{d\gamma - \eta}{d(d+2)} + \frac{(t-d)(2\gamma+\eta)}{2d(d+2)}.$$

To see that the first numerator is nonnegative, factor $A = D^\top D$ and define $\widetilde{H}_r = \sum_j D_{rj}H_j$. Then

$$\gamma = \sum_r \mathrm{tr}(\widetilde{H}_r^2), \qquad \eta = \sum_r \{\mathrm{tr}(\widetilde{H}_r)\}^2,$$

and $d\,\mathrm{tr}(\widetilde{H}_r^2) \geq \{\mathrm{tr}(\widetilde{H}_r)\}^2$ term by term. Thus $B_A'(t) \geq 0$ for $t \geq d$, and it is strictly positive for $t > d$ if some $\widetilde{H}_r \neq 0$. If V_A is nondecreasing, their sum is nondecreasing. In the radial case $\Omega(u) = s(\tau)\Omega_0$, conditional expectation gives $V_A(t) = s(t)\,\mathrm{tr}(A\Omega_0)$. □

E.3 Softmax-KL identity and bound

Proposition 2 (Softmax-KL integral identity and bound)**.** *For logits $z^\star, z \in \mathbb{R}^V$, set $\Delta = z - z^\star$, $P_V = I - \mathbf{1}\mathbf{1}^\top/V$, and $F(z) = \mathrm{diag}(p(z)) - p(z)p(z)^\top$. Then*

$$\mathrm{KL}(p(z^\star)\,\|\,p(z)) = \int_0^1 (1-r)\Delta^\top F(z^\star + r\Delta)\Delta\, dr \leq \frac{1}{4}\|P_V\Delta\|_2^2. \tag{21}$$

For $z^\star = W_Uy$ and $z = W_Um_{\mathrm{aff}}(u)$, this implies

$$\mathbb{E}[\mathrm{KL}(p(z^\star)\|p(z)) \mid \tau = t] \leq \tfrac{1}{4}R_{W_U^\top P_V W_U}(t). \tag{22}$$

A coordinatewise 1-Lipschitz softcap gives the looser metric $A = W_U^\top W_U$.

Proof. Let $\phi(z) = \log \sum_v e^{z_v}$. Its gradient is $p(z)$ and its Hessian is $F(z)$. The softmax KL is the Bregman divergence of ϕ:

$$\mathrm{KL}(p(z^\star)\|p(z)) = \phi(z) - \phi(z^\star) - p(z^\star)^\top (z - z^\star).$$

Taylor's theorem with integral remainder gives the equality in Equation 21. Moreover, $F(z)\mathbf{1} = 0$. Each Gershgorin absolute row sum is $2p_v(1-p_v) \le \frac{1}{2}$, so the symmetric positive-semidefinite matrix $F(z)$ has $\|F(z)\|_2 \le \frac{1}{2}$. Therefore

$$\Delta^\top F(z)\Delta = (P_V\Delta)^\top F(z)(P_V\Delta) \le \tfrac{1}{2}\|P_V\Delta\|^2.$$

Integrating $(1-r)$ over $[0,1]$ proves the upper bound. Substituting $\Delta = W_U\{m_{\mathrm{aff}}(u) - y\}$ and taking the conditional expectation gives Equation 22. For a coordinatewise 1-Lipschitz map g, $\|P_V\{g(W_U m_{\mathrm{aff}}) - g(W_U y)\}\| \le \|W_U(m_{\mathrm{aff}} - y)\|$, which proves the softcap extension. □

F Exact computation and covariance identities

F.1 Exact evaluator and softcap pullback

Let $L_S L_S^\top = S_\lambda$ and let w_v be the v-th unembedding row. Equation 8 uses $C_{vv} = \|L_S^\top w_v\|^2$ and $p^\top C p = \|L_S^\top (W_U^\top p)\|^2$. The former is cached as V scalars beyond the stored $d \times d$ factor L_S; the latter requires one probability-weighted unembedding matvec and an $O(d^2)$ quadratic form. The isotropic reference is precomputed as

$$\alpha = \frac{\sum_v C_{vv} - \|L_S^\top (W_U^\top \mathbf{1})\|^2/V}{V-1}.$$

We evaluate the final subtraction in `float64`: A_{FS} can be small even when the two quantities being subtracted are $O(1)$.

For a coordinatewise logit softcap with diagonal Jacobian $J(x)$ evaluated at the pre-cap logits, write $R = C - \alpha P_V$. The transported statistic and its isotropic complement are

$$A_{\mathrm{FS}}^{\mathrm{cap}}(p) = \tfrac{1}{2}\,\mathrm{tr}\{F(p)J(x)RJ(x)\}, \qquad \tfrac{\alpha}{2}\,\mathrm{tr}\{F(p)J(x)P_V J(x)\}.$$

The diagonal-plus-rank-one structure preserves the $O(V)$ cache and one-matvec evaluation. It reduces to the canonical decomposition when $J = I$; for nonuniform J, centering after the softcap would define a different statistic.

F.2 Ridge-consistent Schur identity

The Schur covariance in Equation 3 is the conditional covariance of the ridge-adjusted joint Gaussian. To compare it with fitted-map residuals, expand $S_{\mathrm{residual}} = \mathrm{Cov}(y - M_\lambda x)$ and use $M_\lambda(\widehat{\Sigma}_{xx} + \lambda I) = \widehat{\Sigma}_{yx}$, obtaining

$$S_\lambda - S_{\mathrm{residual}} = \lambda M_\lambda M_\lambda^\top \succeq 0. \tag{23}$$

The empirical residual covariance and Schur covariance describe fitted-map residuals and regularized conditional spread, respectively. FSR chooses the latter to keep P_λ, M_λ, and S_λ consistent. Their positive-semidefinite ordering alone does not imply an ordering of held-out prediction accuracy.