PREDICTIVE UNCERTAINTY GEOMETRY IN LOW-RANK LANGUAGE MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian low-rank adaptation makes uncertainty estimation tractable, but the quality of predictive uncertainty depends on the geometry and coordinates of the posterior covariance, not only on the adapter dimension. For nonlinear models whose likelihood depends on an informed subspace, we derive a novel trace law for mean-field directional distortion: factorization can inflate uncertainty along informed predictive directions while collapsing it in an uninformed complement, whereas structured Gaussian approximations preserve the relevant geometry. We further show that covariance perturbations can substantially change posterior uncertainty while leaving integrated predictive probabilities and proper scores nearly unchanged. Controlled restricted-adapter experiments in Qwen3-4B and Phi-4-mini support these geometric mechanisms using HMC references. On Qwen, only 9 of 64 generalized covariance directions have MF/HMC variance ratio above two, yet 300 frozen predictive sensitivities place 78.4% of their HMC-whitened energy in those directions. Along a fitted covariance path, measured log-odds variance responses track first-order covariance-response predictions at Spearman 0.999, while matched-model predictive scores change negligibly; no selective-answering threshold is certified at target error rates through 20%. All ten matched-seed coordinate comparisons across Qwen and Phi favor the categorical-Fisher eigenbasis.

1 INTRODUCTION

Low-rank adaptation (LoRA) (Hu et al., 2022) has become one of the most widely adopted methods for adapting large language models to specialized tasks, and Bayesian extensions of LoRA now promise predictive uncertainty without placing a distribution over the entire pretrained model (Yang et al., 2024; Wang et al., 2024b; Samplawski et al., 2025; Shi et al., 2025). These methods share a common geometric question: *which posterior directions must be retained so that predictive uncertainty is faithful to the target posterior?*

The answer turns out to depend on geometry rather than on the LoRA rank alone. A rank-one weight update can contain thousands of stochastic coordinates, and diagonal Gaussian variational inference is coordinate-dependent: the exact Bayesian model is invariant under an orthogonal reparameterization, but the diagonal family is generally transformed into a non-diagonal one. Coordinate choice therefore becomes part of the inference method even when the physical model has not changed.

The possibility that mean-field VI underestimates marginal parameter variances yet overestimates predictive variance is well established (Margossian & Saul, 2023; Odgers et al., 2026; Margossian et al., 2025). Our goal is to go further and characterize the magnitude, directional structure, and interaction of this distortion with the predictive observables of interest. To the best of our knowledge, the results presented here are the first to give an exact nonlinear trace law for mean-field distortion in the LoRA setting: we establish an informed-subspace limit in which the exact posterior retains its prior complement but mean field contracts the wrong linear combinations, and an explicit trace identity quantifies the resulting inflation. We then distinguish posterior covariance fidelity from ordinary predictive scores: Price's theorem gives the generic first-order probability response, while our conditional blindness result identifies directions visible to uncertainty but invisible at first order to a finite collection of probability means.

Our main contributions can be summarized as follows.

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.

1. In Section 3, we establish a nonlinear directional-distortion theorem (Theorem 3.2) that simultaneously characterizes informed inflation and supported-null collapse, gives an explicit trace identity (Corollary 3.3), and identifies an ideal informed-curvature basis without leading-order informed distortion (Proposition 3.7).
2. In Section 4, we show that covariance perturbations can substantially change posterior uncertainty while integrated predictive probabilities and proper scores respond only at second order or higher (Theorem 4.2).
3. In Section 5, we provide empirical evidence corroborating these mechanisms in controlled 64-dimensional adapter posteriors on Qwen3-4B and Phi-4-mini, using Hamiltonian Monte Carlo (HMC) as a reference. Our experiments reveal that predictive sensitivities concentrate on the inflated part of the distortion spectrum (the nine of 64 directions with variance ratio above two carry 78.4% of the predictive energy), and all five matched-seed runs on each backbone favor the categorical-Fisher eigenbasis over identity coordinates.

All theoretical results are proved in Appendix A.

1.1 Related literature

Bayesian LoRA methods include Laplace-LoRA (Yang et al., 2024), BLoB (Wang et al., 2024b), stochastic variational subspace inference (Samplawski et al., 2025), training-free Bayesianization (Shi et al., 2025), orthogonalized Bayesian fine-tuning (Xiang et al., 2026), and the Bayesian Adaptation Gym benchmark (Samplawski et al., 2026). The shrinkage-delinkage trade-off and uncertainty impossibility results for factorized Gaussian approximations are well understood (Margossian & Saul, 2023; Margossian et al., 2025), and predictive over-dispersion under MFVI has been documented, with a posterior temperature proposed to correct it (Odgers et al., 2026) and linear-response corrections available post hoc (Giordano et al., 2018); Remark 3.6 gives the error no common rescaling repairs. Likelihood invariances can create variational gaps (Kurle et al., 2022), and recent symmetry-based work gives additional recovery guarantees (Zellinger & Vergari, 2026). Prior-complement and goal-oriented low-rank covariance approximations have substantial precedent in Bayesian inverse problems (Spantini et al., 2015; 2017). Our contribution lies in the nonlinear informed-subspace spectrum and its explicit trace law, the conditional covariance-response separation that explains why standard scores miss covariance errors, and the controlled same-physical-model coordinate experiment across two transformer backbones.

1.2 Notations

We write $[n] = \{1, \dots, n\}$, $\|\cdot\|$ for the Euclidean norm, and $\|\cdot\|_F$ for the Frobenius norm. For symmetric positive definite A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues. We use $\mathbb{E}$, Var, Cov for expectation, variance, and covariance under the distribution indicated by a subscript. The symbol σ denotes the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$.

2 Model, operators, and observables

We now introduce the model. At an adapted projection of a frozen transformer, we write

$$W = W_0 + \alpha BA, \qquad A \text{ fixed}, \qquad \phi = \phi_0 + Lz, \quad z \sim N(0, I_p). \tag{1}$$

Here ϕ stacks the stochastic entries of B, and L fixes their physical subspace and prior scale. Fixing A removes the two-factor ambiguity in BA from a matched-posterior comparison. The adapter matrix rank, stochastic dimension p, informed dimension m, and represented covariance rank k are all distinct quantities. We adopt the convention that uncertainty is conditional on the frozen backbone and any frozen adapter factors.

For categorical observations $D = \{(x_i, y_i)\}_{i=1}^n$, let $g_z(x)$ denote identifiable logit contrasts. At a specified anchor $\bar{z}$, set $J_i = \nabla_z g_z(x_i)|_{\bar{z}}$ and let F_i be the categorical Fisher matrix in contrast coordinates. For a scalar prediction functional h_x and an unlabeled reference distribution ν, define

$$H_D(\bar{z}) = \sum_{i=1}^n J_i^\top F_i J_i, \qquad a(x; \bar{z}) = \nabla h_x(\bar{z}), \qquad G_\nu(\bar{z}) = \mathbb{E}_\nu[a(X; \bar{z}) a(X; \bar{z})^\top]. \tag{2}$$

Both H_D and G_ν are positive semidefinite. Information is summed, not averaged, and prior-whitened coordinates make $I + H_D$ the precision of the chosen local Gaussian reference. Note that a direction can have low $v^\top H_D v$ and high $v^\top G_\nu v$; this is a geometric property of the two operators. For class probabilities $P_y(z) = p(y \mid x, z)$, the integrated predictive vector is $\bar{p} = \mathbb{E}_q P(z)$ and epistemic probability variance is $\mathrm{Var}_q\, P_y(z)$. Log-odds variance uses a different h_x. We carefully distinguish the nonlinear restricted-adapter posterior, its numerical HMC approximation, and the local-Gaussian diagnostic $(I+H)^{-1}$; all operators are evaluated at a common fixed anchor (details in Appendix B.1).

For the Gaussian reference $\pi_G = N(m, C)$ with $C = (I + H)^{-1}$, diagonal-Gaussian reverse-KL projection gives $q_{\mathrm{MF}} = N(m, D)$ where $D_{jj} = (I + H)^{-1}_{jj}$. Define $R_P = \Delta_P^{-1/2} P \Delta_P^{-1/2}$ with $P = I + H$ and $\Delta_P = \mathrm{diag}(P)$. The extrema of the generalized variance ratio $a^\top D a/(a^\top C a)$ are $\lambda_{\min}(R_P)$ and $\lambda_{\max}(R_P)$; if P is non-diagonal, $\lambda_{\min}(R_P) < 1 < \lambda_{\max}(R_P)$, even though $D_{jj} \le C_{jj}$ for every coordinate. This is a known correlation-removal mechanism (Margossian & Saul, 2023; Odgers et al., 2026); it motivates the nonlinear result below; the formal statement (Proposition A.1) and proof are in Appendix A.1.

3 Nonlinear distortion in the informed subspace

We now formalize the distortion mechanisms. Consider a fixed statistical model

$$\pi_n(z) \;\propto\; \exp\{-\|z\|^2/2 - n\ell(U^\top z)\}, \qquad U \in \mathbb{R}^{p\times m}, \quad U^\top U = I_m. \tag{3}$$

The likelihood depends only on $t = U^\top z$. Its null complement is retained exactly by Bayes. We study a fixed likelihood profile ℓ as its information multiplier n increases. For repeated classification designs this is a sample-size sequence; it is not an assertion about arbitrary random, changing designs.

Assumption 3.1. *Keep p, m, U, and ℓ fixed. Assume:*

(i) $\ell : \mathbb{R}^m \to \mathbb{R}$ *is convex and C^4, with globally bounded second, third, and fourth derivatives.*

(ii) ℓ *has a unique minimizer t_0 with $J = \nabla^2 \ell(t_0) \succ 0$.*

(iii) For some $c > 0$,

$$\ell(t) - \ell(t_0) \;\ge\; c \min\{\|t - t_0\|^2, \|t - t_0\|\} \quad \textit{for all } t. \tag{4}$$

These conditions provide local quadratic curvature and integrable tails without requiring global strong convexity. They are satisfied by the fixed logistic classification designs described in Appendix A.5, and they appear in a similar form in Laplace-type analyses (Wang & Blei, 2019; Katsevich & Rigollet, 2024). The growth condition (4) can be relaxed to log-Sobolev conditions at the cost of additional technical machinery; the present conditions suffice for the logistic designs in Appendix A.5.

Set $u_j = U^\top e_j$ and define

$$\begin{aligned} &h_j = u_j^\top J u_j, \quad A = \{j : h_j > 0\}, \quad s = |A|, \quad K = U^\top D_h U, \\ &D_h = \mathrm{diag}(d_{h,j}), \qquad d_{h,j} = \begin{cases} h_j^{-1}, & j \in A, \\ 0, & j \notin A. \end{cases} \end{aligned} \tag{5}$$

Since $J \succ 0$, inactive coordinates are precisely the zero rows of U.

Informally, the likelihood observes only $t = U^\top z$, yet a dense informed subspace forces a diagonal approximation to shrink coordinate variances even along combinations the likelihood never observes, while the structured family contracts within U without spending that contraction on $U^\perp$; Theorem 3.2 and Corollary 3.3 make both errors exact.

Theorem 3.2 (Nonlinear informed-subspace limit and structured repair)**.** *Under Assumption 3.1, the exact posterior retains an independent standard Gaussian law on $U^\perp$ for every n. Its informed moments satisfy*

$$\mathbb{E}_{\pi_n} t = t_0 + O(n^{-1}), \qquad n\,\mathrm{Cov}_{\pi_n}(t) = J^{-1} + O(n^{-1}). \tag{6}$$

The unique diagonal-Gaussian reverse-KL optimum q_n has mean $U\beta_n$ and coordinate variances $s_{n,j}$. Its inactive variances equal one, and

$$\beta_n = t_0 + O(n^{-1}), \quad n s_{n,j} = h_j^{-1} + O(n^{-1}) \;\; (j \in A), \quad n\,\mathrm{Cov}_{q_n}(t) = K + O(n^{-1}). \tag{7}$$

Check	Theorem limit	$n = 10^5$ calculation
Scalar MF probability variance, $s = 4$	4	3.997–3.998
Scalar structured probability variance	1	1.0000
Supported-null MF linear variance	0	6.3×10^{-4}
$m = 3, s = 9$ informed spectrum	$(1.67105, 3.15216, 4.17679)$	$(1.67095, 3.15155, 4.17581)$
$m = 3, s = 9$ spectral trace	9	8.99831

Table 1: Deterministic verification of the trace law under Assumption 3.1. The checks use deterministic nonlinear examples rather than the transformer posterior.

The reverse-KL optimum over $N(Ub, I - UU^\top + USU^\top)$, $S \succ 0$, also exists uniquely and obeys

$$b_n = t_0 + O(n^{-1}), \qquad nS_n = J^{-1} + O(n^{-1}). \tag{8}$$

It preserves the exact null-complement law at every n; we write PC for this prior-complement optimum.

The proof, given in Appendix A.3, first establishes concentration of the variational optimizer before taking limits in its stationarity equations.

Corollary 3.3 (Distortion spectrum and two simultaneous errors)**.** *Under Theorem 3.2, let $R = J^{1/2} K J^{1/2}$. Then*

$$\operatorname{tr}(R) = s, \qquad \lambda_{\max}(R) \geq s/m. \tag{9}$$

If $s > m$, at least one informed direction has limiting variance ratio greater than one. Simultaneously, every unit $v \in U^\perp$ supported on A satisfies

$$\operatorname{Var}_{\pi_n}(v^\top z) = 1, \qquad \operatorname{Var}_{q_n}(v^\top z) = \frac{v^\top D_h v}{n} + O(n^{-2}). \tag{10}$$

For any C^3 function $f : \mathbb{R}^m \to \mathbb{R}$ with bounded derivatives of orders one through three and $a = \nabla f(t_0) \neq 0$,

$$\operatorname{Var}_{\pi_n} f(t) = \frac{a^\top J^{-1} a}{n} + O(n^{-2}), \qquad \operatorname{Var}_{q_n} f(t) = \frac{a^\top K a}{n} + O(n^{-2}). \tag{11}$$

The structured optimum has the same leading variance as the exact posterior.

The proof is given in Appendix A.4. This corollary provides the core quantitative mechanism: the trace identity $\operatorname{tr}(R) = s$ fixes the total informed distortion, and whenever $s > m$ the distortion simultaneously inflates informed directions and collapses supported null ones.

Because the trace identity is one of the main theoretical claims, we verify it in deterministic nonlinear examples satisfying the theorem's assumptions, rather than treating the transformer covariance spectrum as if it satisfied the same asymptotic model. Table 1 summarizes the checks; the complete numerical ledger and convergence plots are in Appendix B.4.

Remark 3.4 (Logistic class and the support factor). The logistic profile $\ell(t) = \sum_\ell \omega_\ell \{\log(1 + \exp(x_\ell^\top t)) - \rho_\ell x_\ell^\top t\}$ satisfies Assumption 3.1 (see Appendix A.5).

Corollary 3.5 (Logistic support factor)**.** *In the scalar case $m = 1$, let $U = u$ be a unit vector with $s \geq 2$ nonzero coordinates and $c_\rho = \rho(1 - \rho)$. The exact informed variance is $(nc_\rho)^{-1} + O(n^{-2})$, while the mean-field variance is $s/(nc_\rho) + O(n^{-2})$. For every fixed $b \in \mathbb{R}$,*

$$\frac{\operatorname{Var}_{q_n}\{\sigma(b + u^\top z)\}}{\operatorname{Var}_{\pi_n}\{\sigma(b + u^\top z)\}} = s + O(n^{-1}), \qquad \frac{\operatorname{Var}_{\mathrm{PC},n}\{\sigma(b + u^\top z)\}}{\operatorname{Var}_{\pi_n}\{\sigma(b + u^\top z)\}} = 1 + O(n^{-1}). \tag{12}$$

The support factor s is independent of the success fraction ρ. For the balanced case $\rho = 1/2$, symmetry gives a zero mean and the stationarity equation $s_{n,j}^{-1} = 1 + n u_j^2 \mathbb{E}\sigma'(\sqrt{v_n} Z)$ with $v_n = \sum_j u_j^2 s_{n,j}$ and $n v_n \to 4s$, which makes the support-factor convergence transparent. The structured optimum recovers the exact variance; see Appendix A.5 for the proof and a finite-data Gaussian proxy that explains when tiny active coordinates delay the support-factor regime.

Remark 3.6 (No common multiplier repairs both errors). Rescaling the mean-field covariance by any $\gamma_n > 0$ that corrects the informed linear variance forces $\gamma_n \to 1/s$, and the supported-null variance still tends to zero. This is a posterior covariance rescaling statement; temperature scaling of output logits is a different calibration operation. A formal statement is in Appendix A.5.

3.1 COORDINATES AND EQUIVARIANCE

Proposition 3.7 (Equivariance and the mean-field restriction). *For an orthogonal O, replace the physical map L in (1) by LO and use $z' = O^\top z$. The prior, physical parameter ϕ, likelihood, and exact predictive distribution are unchanged. A prior-complement Gaussian with subspace U maps to the same family with subspace $O^\top U$, with invariant reverse KL. A diagonal covariance maps to $O^\top DO$, which is generally non-diagonal; the mean-field family is therefore not invariant under arbitrary such rotations.*

The proof is in Appendix A.6. In the theorem's idealized limit, the coordinate implication is sharper than non-invariance alone. Let V diagonalize J and choose an orthogonal O whose first m coordinate axes are the informed-curvature eigenvectors UV. Then $O^\top U = [V^\top; 0]$: exactly m coordinate rows are active, their h_j values are the eigenvalues of J, and therefore $s = m$, $K = J^{-1}$, and $R = I_m$ i.e. the entire informed distortion spectrum collapses to the identity. Thus mean field has no leading-order informed distortion in this ideal eigenbasis. A finite-sample nonlinear posterior need not factorize after such a rotation, but the calculation motivates a local curvature or Fisher eigenbasis as a targeted coordinate control: the ideal informed-curvature basis removes leading-order informed distortion in the theorem, and our experiments in Section 5.5 test this finite-sample analogue on two transformer backbones.

4 WHAT PREDICTIVE EVALUATIONS CAN DETECT

We now turn to a complementary question: *how much can predictive probabilities and proper scores reveal about posterior covariance errors?* We start with a generic covariance-response identity, then give a condition under which the probability response is weak even though the uncertainty response is strong. The proof of the following lemma is in Appendix A.7.

Lemma 4.1 (Gaussian covariance response). *Fix a mean μ and a symmetric covariance direction Δ. Let $q_t = N(\mu, \Sigma_0 + t\Delta)$ on an interval where the covariance is positive definite. If $f \in C^4$ with at most polynomial-growth derivatives through order four, then*

$$\frac{d}{dt}\mathbb{E}_{q_t} f = \frac{1}{2}\operatorname{tr}\{\Delta\, \mathbb{E}_{q_t} \nabla^2 f\}. \tag{13}$$

The remainder of the first-order approximation is $O(t^2)$ with an explicit bound given in Appendix A.7.

This is Price's theorem from Voigtlaender (2021), stated with the remainder order needed for Theorem 4.2; the identity shows that covariance error generally produces a *linear* probability error.

At q_0, collect prompt/class probability functions $f_1, \dots, f_L$ and set $H_\ell^{\mathrm{prob}} = \frac{1}{2}\mathbb{E}_{q_0}\nabla^2 f_\ell$, $\mathcal{H} = \operatorname{span}\{H_1^{\mathrm{prob}}, \dots, H_L^{\mathrm{prob}}\}$. For an uncertainty observable h, define

$$B_h = \tfrac{1}{2}\mathbb{E}_{q_0}\nabla^2(h^2) - (\mathbb{E}_{q_0} h)\mathbb{E}_{q_0}\nabla^2 h. \tag{14}$$

For linear $h(z) = a^\top z$, this reduces to $B_h = aa^\top$. The proof of the next theorem is in Appendix A.7.

Theorem 4.2 (Conditional first-order blindness). *Assume $f_1, \dots, f_L, h, h^2$ satisfy the polynomial-growth conditions of Lemma 4.1. Let $B_\perp = B_h - \operatorname{Proj}_{\mathcal{H}} B_h$ in the Frobenius inner product. If $B_\perp \neq 0$, choose $\Delta = B_\perp / \|B_\perp\|_F$. Then $\Sigma_0 + t\Delta \succ 0$ for $|t| < \lambda_{\min}(\Sigma_0)$, and as $t \to 0$,*

$$\mathbb{E}_{q_t} f_\ell - \mathbb{E}_{q_0} f_\ell = O(t^2) \quad (\ell = 1, \dots, L), \tag{15}$$

$$\operatorname{Var}_{q_t} h - \operatorname{Var}_{q_0} h = t\|B_\perp\|_F + O(t^2). \tag{16}$$

For linear h, the variance response is exactly linear.

The theorem supplies a testable covariance geometry for a specified finite prompt collection. The constructed direction need not equal the difference between two fitted covariances, but it identifies directions along which standard evaluation metrics are first-order uninformative.

Remark 4.3 (Proper scores respond at second order). For multiclass Brier loss, $R_{\mathrm{Br},\tau}(p+\delta) - R_{\mathrm{Br},\tau}(p) = 2(p-\tau)^\top\delta + \|\delta\|_2^2$; when $\tau = p$, the regret is exactly quadratic in δ. For log loss with $p_y \geq \alpha > 0$, the regret under the reference distribution is $\frac{1}{2}\sum_y \delta_y^2/p_y + O_\alpha(\|\delta\|_2^3)$. Along the

direction in Theorem 4.2, the probability-mean perturbation is $\delta = O(t^2)$, so proper-score regret under the reference distribution is $O(t^4)$. Under a different outcome distribution or for a realized label, the score change is generally only $O(t^2)$. The formal expansions are in Appendix A.8. Crucially, a small empirical score difference leaves the posterior covariance accuracy unresolved, while a large covariance error may still fail to produce a measurable decision improvement.

Example 4.4 (Equal means, different uncertainties). For $Z \sim N(0,1)$ and $P_v = \sigma(\sqrt{v}Z)$, symmetry gives $\mathbb{E}P_v = 1/2$ for every $v \geq 0$. NLL and Brier are therefore identical for every realized binary label at every v. Yet both $\mathrm{Var}(P_v)$ and $\mathrm{MI}(v) = \log 2 - \mathbb{E}\mathsf{h}(P_v)$ increase strictly with v, with $\mathrm{Var}(P_v) = v/16 + O(v^2)$. More generally, if $P_y \geq \alpha > 0$ almost surely and $\bar{p} = \mathbb{E}P$, the entropy expansion $H(\bar{p}) - \mathbb{E}H(P) = \frac{1}{2}\sum_y \mathrm{Var}(P_y)/\bar{p}_y + O_\alpha(\mathbb{E}\|P - \bar{p}\|_2^3)$ shows that mutual-information acquisition (Houlsby et al., 2011) is directly sensitive to variance errors that integrated single-label scoring cannot detect.

5 Controlled experiments on language models

We now provide empirical evidence corroborating the mechanisms established in Sections 3 and 4, using restricted Bayesian adapters in two transformer families. The experimental design holds the physical Bayesian model fixed while varying the covariance approximation or the parameter coordinates, and uses HMC as a numerical reference for the 64-dimensional restricted posterior. The emphasis throughout is on controlled investigation of covariance geometry in the LoRA posterior rather than on proposing a new method. The LoRA structure is central to the design: by fixing the physical adapter and varying only the covariance or coordinates, we isolate the geometric mechanisms that Sections 3 and 4 predict.

5.1 Setup

We use Qwen3-4B-Instruct-2507 (Yang et al., 2025) and Phi-4-mini-instruct (Abouelenin et al., 2025) with MMLU-Pro (Wang et al., 2024a). The source benchmark split is repurposed into disjoint adaptation, reference, and evaluation sets, so reported scores are not canonical MMLU-Pro benchmark numbers. Unless stated otherwise, the headline Qwen comparisons use $n = 2000$ adaptation prompts and 300 frozen evaluation functions. A one-sided rank-one adapter modifies the final block's `mlp.down_proj`. For this placement, we cache the upstream activations and evaluate the remaining nonlinear block computation, normalization, and answer-token head exactly for each posterior draw. The full Qwen stochastic adapter has $p = 2560$ coordinates; the HMC comparison fixes a $p = 64$-dimensional orthonormal physical adapter subspace and hence a named restricted Bayesian model. Predictive NLL, Brier score, and accuracy are computed after softmax normalization over the cached multiple-choice answer logits; the same normalization applies throughout. The same-model score audit (Section 5.4) also uses the original frozen 1200 calibration prompts and 1500 held-out test prompts, evaluated under the same fixed-mean restricted posterior approximations.

The Qwen HMC reference uses four NUTS chains with 250 warmup and 500 retained draws per chain, target acceptance 0.8, and maximum tree depth 6; the worst rank-normalized $\widehat{R}$ is 1.009, the minimum centered-square ESS is above 800, and all chains report zero divergences. The Phi reference uses four fresh chains with 500 warmup and 1500 retained draws each; worst $\widehat{R} = 1.004$, minimum bulk ESS 10703, zero divergences. Restricted HMC serves as a numerical reference for these 64-dimensional posteriors; extending to the full adapter remains an important open problem. Full diagnostics are in Appendix B.2.

5.2 Distortion spectrum and predictive energy concentration

The generalized MF/HMC variance ratios range from 0.38 to 4.75 across 64 eigendirections, with 29 above one and 35 below one; the discrepancy is not a scalar temperature error. In native coordinates, 40 of 64 marginal ratios are below one (median 0.903), yet this coordinatewise view is misleading. Along all 300 frozen log-odds sensitivity directions, the MF/HMC linear variance ratio exceeds two (median 2.58, range 2.00–3.92); no predictive direction attains the eigendirection maximum 4.75.

Only 9 of 64 generalized eigendirections have variance ratio above two, but the 300 predictive sensitivity directions place a mean 78.4% (median 78.8%) of their HMC-whitened energy in those

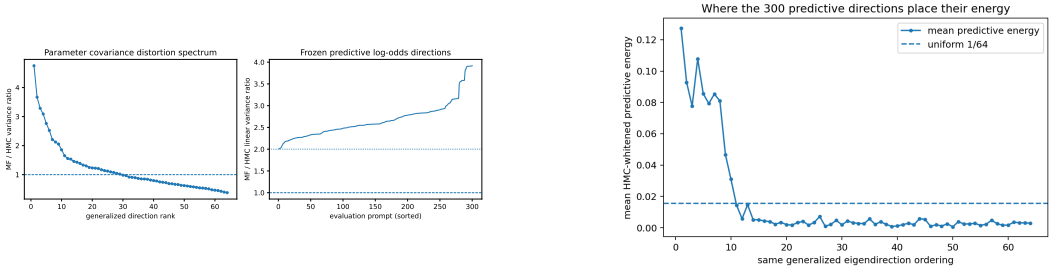


Figure 1: Qwen covariance geometry at $n = 2000$. Left: generalized MF/HMC variance ratios and their action on 300 frozen log-odds sensitivity directions. Right: mean HMC-whitened predictive energy by eigendirection, ordered by variance ratio.

nine directions; the top eight inflated eigendirections alone contain a median 75.9% of predictive energy. The 29 directions with ratio above one receive 90.5% of predictive energy on average. Promptwise variance inflation is positively associated with energy in the $\lambda > 2$ subspace (Spearman 0.586). Thus the deployment observables concentrate on a small part of the posterior spectrum where MF is most distorted, even though coordinatewise the distortion appears mild. Projecting both covariances onto the leading eigenspace of the common-anchor categorical Fisher gives mean restricted ratios of 2.62 (rank 9) and 2.45 (rank 10), bracketing the median predictive-direction ratio of 2.58. This numerical agreement serves as a finite-sample consistency check; the empirical Fisher eigenspace has nonzero leverage on all 64 coordinates, so the identification with the theorem's active support s remains approximate. The detailed threshold sensitivity analysis and a dense eigenvalue sweep are in Appendix B.5.

5.3 The covariance path produces the predicted response

Let $\Sigma(t) = (1-t)\Sigma_{\mathrm{MF}} + t\Sigma_{\mathrm{PC}}$ with a common fixed mean, where Σ_{PC} is the Prediction-PC16 covariance: the prior-complement family of Theorem 3.2 with subspace spanned by the sixteen common-anchor Fisher eigenvectors v_i of largest $\lambda_i(1+\lambda_i)^{-1}v_i^\top G v_i$, where G is the Gram matrix of the frozen log-odds sensitivities, refit by reverse KL at that mean. We evaluate the exact cached nonlinear predictor along this path using common random numbers. The fitted MF-to-PC path tests the generic Price/covariance-response identity (Lemma 4.1); the orthogonality premise of Theorem 4.2 is tested separately in Appendix B.10. For log-odds variance, the measured derivative at $t = 1/2$ agrees closely with the linear covariance-response calculation: Spearman 0.999, Pearson 0.999, through-origin slope 1.014. This constitutes a direct empirical contact point with the covariance-response identity. The ordinary predictive mean is far less responsive along this fitted path: from MF to PC covariance, the median absolute relative focal-probability change is below 10^{-4} percent, whereas the median relative log-odds-variance change is -38.5% with 88% of prompts decreasing. Using posterior-mean probabilities, mean NLL and Brier move by less than 10^{-3} (5.37672 to 5.37713 and 0.95848 to 0.95851; Table 2). Even on the 30 least-saturated prompts (baseline focal probabilities 0.380–0.616), the median absolute relative probability change is only 0.011%, versus a 37.8% change in log-odds variance (Appendix B.7). The mean/covariance swap in Figure 2 isolates the effect further: with either fitted mean, replacing MF covariance by PC covariance changes the median log-odds variance from 0.00456 to 0.00247, while changing the mean at fixed covariance barely moves it. This confirms that the uncertainty change is driven by the covariance source, not by the posterior mean.

An independent control-variate Price-response audit (Appendix B.8) gives Pearson 0.758 between predicted and measured focal-probability derivatives; the comparison is noise-limited, with 285 of 300 differences within two Monte Carlo standard errors. A separate frozen construction of the $B_\perp$ direction from Theorem 4.2 suppresses cross-fitted first-order probability responses relative to a generic control (ratio 0.133); the estimated response system is extremely ill-conditioned, so we treat this audit as mixed mechanistic evidence and report the full details in Appendix B.10.

5.4 Scores, selective answering, and the variational objective

The covariance-path results demonstrate a separation between uncertainty fidelity and conventional evaluation. We now verify that this separation persists in a controlled same-model score comparison.

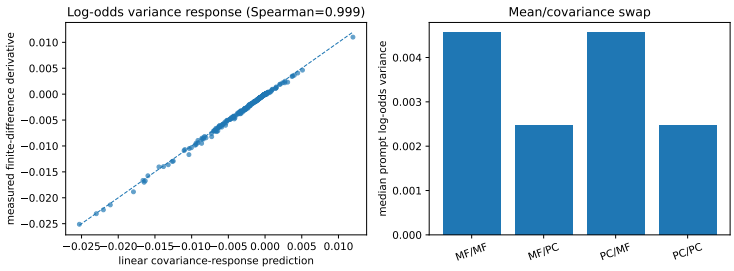


Figure 2: Left: measured log-odds variance derivative versus first-order covariance-response prediction (Spearman 0.999). Right: mean/covariance swap; the uncertainty change follows the covariance source.

Panel	Covariance	Accuracy	NLL	Brier	ECE	Predictive MI
Frozen fidelity 300	MF	0.460	5.37672	0.95848	0.44809	0.000323
	Prediction-PC16	0.460	5.37713	0.95851	0.44814	0.000227
Held-out 1500	MF	0.446	5.34755	0.98338	0.46169	0.000302
	Prediction-PC16	0.446	5.34808	0.98342	0.46173	0.000202

Table 2: Same restricted $p = 64$ Bayesian model, same posterior mean, two covariance approximations. ECE uses 15 equal-width bins; NLL and Brier use posterior-mean answer-choice-normalized probabilities. NLL exceeds the uniform ten-choice baseline $\log 10 \approx 2.30$; errors are frequently confident (see ECE).

Without refitting the posterior or the backbone, we evaluate the fixed-mean $p = 64$ MF and Prediction-PC16 covariance approximations on both the 300 frozen fidelity prompts and the original 1500 held-out prompts (Table 2). On both panels, accuracy is identical (0.460 and 0.446, respectively), and NLL, Brier, and ECE differences are below 10^{-3}. Thus the substantial HMC-referenced covariance-fidelity difference is not accompanied by a meaningful change in these particular predictive scores. Using a Clopper–Pearson/Bonferroni calibrated selective-answering protocol on the original 1200-calibration/1500-test split, neither approximation admits any certified threshold at target error rates $\alpha \in \{0.05, 0.10, 0.20\}$ for any of three uncertainty scores; details are in Appendix B.6.

Prediction-PC16 reduces the median HMC-referenced log-variance error from 0.840 to 0.348 (a 2.4-fold improvement), while its estimated ELBO advantage over diagonal MF is only $+1.73$ (Monte Carlo $z = 1.03$), which remains unresolved at the current precision. This illustrates that the optimized divergence and the evaluated uncertainty functional can separate: large uncertainty-fidelity improvements can coexist with small or unresolved objective differences, consistent with known factorized-VI limitations (Margossian et al., 2025). A separate generic rank-16-plus-diagonal control was not successfully optimized: its three achieved ELBOs are 12.9–17.9 nats below diagonal MF, and its worse covariance fidelity appears to reflect an optimization failure rather than a family limitation (Appendix B.9). The representation costs also differ: diagonal covariance stores $O(p)$ numbers, while rank-k plus diagonal stores $O(pk)$.

5.5 Coordinate dependence across backbones

The central evidence is the five matched-seed comparisons on each backbone. Holding the physical model fixed, Qwen Fisher-minus-identity median absolute log-variance errors are $-0.585, -0.683, -0.653, -0.768, -0.565$; all five runs and all frozen evaluation prompts favor Fisher coordinates. With the independently certified 6000-draw HMC reference, the corresponding Phi differences are $-0.633, -0.481, -0.612, -0.407, -0.598$; all five runs favor Fisher, with seedwise prompt fractions from 88.3% to 100% (Figure 3; per-seed Phi endpoints in Appendix B.11).

The experimental logic here is worth emphasizing. By Proposition 3.7, replacing L by LO and coordinates by $O^\top z$ leaves the standard-normal prior, physical adapter, likelihood, predictions, and exact posterior unchanged. Qwen scoring equality is checked numerically and HMC draws transform exactly. The ideal-basis calculation from Section 3.1 predicts that aligning and diagonalizing informed curvature sends s to m and R to I_m; the empirical categorical-Fisher basis is a finite-sample approximation to this ideal eigenbasis. For a concrete Qwen endpoint, identity-coordinate MF error is 0.815 versus 0.146 in Fisher coordinates; ten Haar-random controls have errors 0.778–0.915, confirming that the improvement is specific to the Fisher rotation rather than a generic effect of

Seed	Identity	Fisher	Fisher − identity
20271001	0.849	0.216	-0.633
20271002	0.892	0.411	-0.481
20271003	0.783	0.171	-0.612
20271004	0.861	0.454	-0.407
20271005	0.825	0.227	-0.598

Table 3: Phi-4-mini matched-seed coordinate replication (certified long-chain HMC). Errors are median absolute log-variance ratios to HMC.

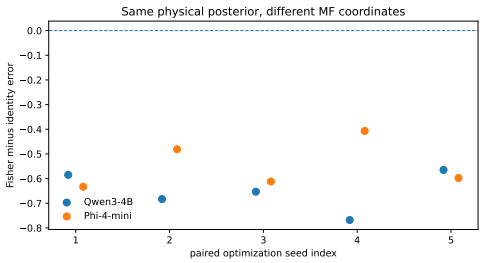


Figure 3: Fixed-physical coordinate experiment. Negative values indicate lower MF log-odds variance error in the Fisher eigenbasis. All five paired runs favor Fisher on both backbones.

any coordinate change. An exactly transformed Prediction-PC16 covariance has error 0.275 in both coordinates, serving as an implementation-level equivariance check. Table 3 gives the certified Phi per-seed endpoints. The Fisher basis serves as a diagnostic establishing that coordinate choice matters for LoRA uncertainty; extending it to full-size adapters is an important direction for future work.

6 Concluding Remarks

We have shown that covariance geometry and parameter coordinates are integral parts of the inference method in Bayesian low-rank adapters. Our results yield a precise characterization of how diagonal variational inference misallocates posterior variance in the LoRA setting, and why conventional predictive metrics can miss substantial covariance errors. Since LoRA and its variants have become the dominant paradigm for adapting large language models, we expect the geometric mechanisms identified here to be broadly relevant for any Bayesian LoRA method that employs a diagonal or block-diagonal posterior. The trace law (Corollary 3.3) provides an exact identity $\mathrm{tr}(R) = s$ showing that when $s > m$, mean-field factorization forces some informed inflation while collapsing supported null directions; the structured prior-complement approximation preserves the correct geometry. The support factor s in the logistic case makes the distortion magnitude transparent and gives a concrete explanation of why diagonal LoRA posteriors can simultaneously overestimate and underestimate uncertainty. The conditional blindness result (Theorem 4.2) explains why proper scores respond only at higher order to covariance perturbations that produce first-order uncertainty changes. The controlled experiments on Qwen3-4B and Phi-4-mini are consistent with these mechanisms: only 9 of 64 eigendirections carry 78.4% of predictive sensitivity energy, the covariance-response predictions track measured responses at Spearman 0.999, and all ten matched-seed coordinate experiments favor the Fisher eigenbasis.

Several extensions are crucial for future work. The theory assumes fixed dimension and a convex, smooth informed-subspace likelihood, and extending it to actual transformer posteriors remains open. The HMC reference covers a 64-dimensional restricted adapter; both backbone studies use the same multiple-choice task family and analogous final-block placement; and the independent probability-derivative and finite-path $B_\perp$ audits are noise-limited (Appendices B.8 and B.10). Despite these scope constraints, the geometric mechanisms identified here—the trace law, the covariance-response separation, and the coordinate dependence—are generic features of diagonal variational inference applied to LoRA posteriors. It is crucial to extend these results to multi-layer adapters, to develop scalable informed-curvature eigenbases, and to investigate whether the documented covariance-geometry improvements translate to downstream active-learning or out-of-distribution detection gains.

Reproducibility statement

The accompanying repository contains the complete sources for paper-asset generation scripts, frozen same-model empirical results and figures, corrected executable notebooks, and provenance checksums. The score notebooks reuse the 8544-prompt original Qwen last-block cache and the already fitted restricted posteriors without any retraining. The certified Phi result uses longer fresh HMC chains. Large deterministic model caches are documented but not duplicated; the notebooks consume or reconstruct them at user-specified workspace paths.

AI use statement

AI-assisted tools were used for code development and debugging, result organization, literature checking, feedback on research methodology and experimental design, manuscript preparation (including restructuring, compression, and copyediting), and checking of mathematical arguments and proofs. The authors reviewed the resulting text, computations, theorem statements, proofs, and experimental provenance, and take responsibility for the paper's contents.

References

A. Abouelenin et al. Phi-4-Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Ryan Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness, and variational bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Anya Katsevich and Philippe Rigollet. On the approximation accuracy of gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409, 2024. doi: 10.1214/24-AOS2393.

Richard Kurle, Ralf Herbrich, Tim Januschowski, Yuyang Wang, and Jan Gasthaus. On the detrimental effect of invariances in the likelihood for variational inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4531–4542, 2022.

Charles C. Margossian and Lawrence K. Saul. The shrinkage-delinkage trade-off: An analysis of factorized gaussian approximations for variational inference. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1358–1367, 2023.

Charles C. Margossian, Loucas Pillaud-Vivien, and Lawrence K. Saul. Variational inference for uncertainty quantification: An analysis of trade-offs. *Journal of Machine Learning Research*, 26 (202):1–41, 2025.

James Odgers, Bernhard Riegler, Siddharth Swaroop, and Vincent Fortuin. Gaussian mean field variational inference can overestimate predictive variance. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Colin Samplawski, Adam D. Cobb, M. Acharya, R. Kaur, and S. Jha. Scalable bayesian low-rank adaptation of large language models via stochastic variational subspace inference. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3587–3604, 2025.

Colin Samplawski, R. Kaur, M. Acharya, A. Roy, and Adam D. Cobb. Bayesian adaptation gym: A benchmark for the bayesian low-rank adaptation of multi-modal language models. In *Proceedings of the Forty-Second Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5919–5966, 2026.

H. Shi, Y. Wang, L. Han, H. Zhang, and H. Wang. Training-free bayesianization for low-rank adapters of large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46633–46670, 2025.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

Alessio Spantini, Tiangang Cui, Karen Willcox, Luis Tenorio, and Youssef Marzouk. Goal-oriented optimal approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 39(5):S167–S196, 2017.

Felix Voigtlaender. A general version of price's theorem. *Journal of Theoretical Probability*, 34: 1474–1485, 2021.

Y. Wang, X. Ma, G. Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024a.

Y. Wang, H. Shi, L. Han, D. Metaxas, and H. Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 67758–67794, 2024b.

Yixin Wang and David M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

H. Xiang, B. Li, and Q. Lu. Scalable variational bayesian fine-tuning of llms via orthogonalized low-rank adapters. *arXiv preprint arXiv:2604.03388*, 2026.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

L. Zellinger and A. Vergari. Even more guarantees for variational inference in the presence of symmetries. *arXiv preprint arXiv:2604.21407*, 2026.

A Proofs and scope of the theoretical results

All asymptotic results below keep the dimension, informed subspace, and likelihood profile fixed. Gaussian reference identities and nonlinear likelihood results have distinct scopes. The proofs establish the mathematical statements under their premises; numerical illustrations and the LLM experiments do not establish those premises for the full transformer.

A.1 Gaussian directional distortion and a rotated null direction

Proposition A.1 (Gaussian directional distortion)**.** *Let $P = I + H$, $\Delta_P = \mathrm{diag}(P)$, and $R_P = \Delta_P^{-1/2} P \Delta_P^{-1/2}$. Then*

$$\min_{a \neq 0} \frac{a^\top D a}{a^\top C a} = \lambda_{\min}(R_P), \qquad \max_{a \neq 0} \frac{a^\top D a}{a^\top C a} = \lambda_{\max}(R_P).$$

If P is non-diagonal, $\lambda_{\min}(R_P) < 1 < \lambda_{\max}(R_P)$, although $D_{jj} \leq C_{jj}$ for every coordinate.

Let $\Delta = \mathrm{diag}(P)$. For $q = N(\mu, D)$ with diagonal $D \succ 0$ and $\pi_G = N(m, P^{-1})$,

$$2\,\mathrm{KL}(q \| \pi_G) = \mathrm{tr}(PD) + (\mu - m)^\top P(\mu - m) - p - \log\det P - \log\det D.$$

The unique optimum has $\mu = m$ and $D_{jj} = P_{jj}^{-1}$. For any coordinate, a Schur complement gives

$$C_{jj} = \{P_{jj} - P_{j,-j} P_{-j,-j}^{-1} P_{-j,j}\}^{-1} \geq P_{jj}^{-1} = D_{jj}.$$

The extrema of the generalized Rayleigh quotient $a^\top Da/(a^\top Ca)$ are the eigenvalues of $C^{-1/2}DC^{-1/2} = P^{1/2}\Delta^{-1}P^{1/2}$. Conjugating by $P^{-1/2}$ and then by $\Delta^{1/2}$ shows this matrix has the same spectrum as $\Delta^{-1}P$ and hence as $R = \Delta^{-1/2}P\Delta^{-1/2}$. Since R has unit diagonal, its eigenvalues sum to p. If P is non-diagonal, $R \neq I$, because R carries the off-diagonal entries $P_{ij}/\sqrt{P_{ii}P_{jj}}$; a real symmetric matrix with every eigenvalue one would equal I. At least one eigenvalue must therefore be below one and another above one. This proves Proposition A.1.

For the classical null-direction example, take $H_n = nH_0$ with $H_0 \succeq 0$ singular and $h_* = \min_j (H_0)_{jj} > 0$. If $v \in \ker H_0$ is unit, then

$$v^\top (I + nH_0)^{-1} v = 1, \qquad v^\top D_n v \leq (1 + nh_*)^{-1} \longrightarrow 0.$$

The positive diagonal condition expresses misalignment between the null space and factorized coordinates. It fails for a coordinate that is entirely unused, whose mean-field variance can remain one. In the controlled Gaussian check ($p = 20$, information rank 5), the rotated mean-field null variance decreases from 0.378 at $n = 10$ to 0.0068 at $n = 1000$. An informed direction has exact variance 9.99×10^{-4} but mean-field variance 5.13×10^{-3} at $n = 1000$. These are reference-experiment values used only as finite numerical checks of the Gaussian calculation.

A.2 A moment-controlled Laplace expansion on the informed subspace

Lemma A.2 (Fixed-profile moment expansion)**.** *Under Assumption 3.1, let T_n have density proportional to $\exp\{-\|t\|^2/2 - n\ell(t)\}$. Then*

$$\mathbb{E}T_n = t_0 + O(n^{-1}), \qquad \mathrm{Cov}(T_n) = n^{-1}J^{-1} + O(n^{-2}).$$

For $e_n = T_n - t_0$, each entry of $\mathbb{E}e_n^{\otimes 3}$ is $O(n^{-2})$, and $\mathbb{E}\|e_n\|^r = O(n^{-r/2})$ for every fixed positive integer r.

Proof. Write $\phi(t) = \ell(t) - \ell(t_0)$ and $y = \sqrt{n}(t - t_0)$. After removing a constant, the unnormalized density in y is

$$g_n(y) = \exp\{-\tfrac{1}{2}y^\top Jy - n^{-1/2}b(y) - r_n(y)\}, \quad b(y) = t_0^\top y + \tfrac{1}{6}D^3\ell(t_0)[y,y,y],$$

where Taylor's theorem and the bounded fourth derivative give

$$|r_n(y)| \leq n^{-1}\{\tfrac{1}{2}\|y\|^2 + (L_4/24)\|y\|^4\}.$$

The function b is odd. We show, for each fixed integer $r \geq 0$, that

$$\int (1 + \|y\|^r) \left| g_n(y) - e^{-y^\top Jy/2}\{1 - n^{-1/2}b(y)\} \right| dy = O(n^{-1}). \tag{17}$$

Choose a sufficiently small fixed $d > 0$. On $\|y\| \leq d\sqrt{n}$, the cubic and quartic terms in $n^{-1/2}|b(y)| + |r_n(y)|$ are bounded by a constant times $(d + d^2)\|y\|^2$. Young's inequality bounds $n^{-1/2}|t_0^\top y|$ by any prescribed positive multiple of $\|y\|^2$ plus C/n. For small enough d and large enough n, the Gaussian factor times $\exp\{n^{-1/2}|b| + |r_n|\}$ is bounded by $C\exp(-c_1\|y\|^2)$ with $c_1 > 0$. The elementary inequalities

$$|e^{-(a+r)} - e^{-a}| \leq |r|e^{|a|+|r|}, \qquad |e^{-a} - 1 + a| \leq \tfrac{1}{2}a^2 e^{|a|}$$

then bound the weighted integral of the error on this ball by C_r/n.

For the complement, condition (4) implies $\phi(t_0 + e) \geq c\min(d, 1)\|e\|$ whenever $\|e\| \geq d$. The prior contribution is $-t_0^\top e - \|e\|^2/2$. For sufficiently large n, it follows that $g_n(y) \leq \exp(-c_2\sqrt{n}\|y\|)$ on $\|y\| > d\sqrt{n}$ for some $c_2 > 0$. Every fixed polynomially weighted integral of this tail is exponentially small in n, as is the Gaussian approximation's tail. This proves (17).

The normalizer equals the positive Gaussian normalizer plus $O(n^{-1})$, since the first correction is odd. The same parity cancellation for $yy^\top$ gives $\mathbb{E}YY^\top = J^{-1} + O(n^{-1})$. Applying the expansion to y and to each entry of $y^{\otimes 3}$ gives $\mathbb{E}Y = O(n^{-1/2})$ and $\mathbb{E}Y^{\otimes 3} = O(n^{-1/2})$. All fixed absolute moments of Y are bounded. Rescaling by $n^{-1/2}$ and subtracting the mean product establishes the claims. □

Only the m informed coordinates undergo this expansion. The orthogonal complement retains its prior distribution, so the lemma is not a full-dimensional Bernstein–von Mises assertion.

A.3 Existence, concentration, and covariance of the mean-field optimizer

Proof of the exact and mean-field parts of Theorem 3.2. Orthogonally decompose $z = Ut + w$ with $w \in U^{\perp}$. The standard Gaussian prior makes t and w independent. Since the likelihood depends only on t, the exact posterior factorizes and leaves w unchanged. Lemma A.2 gives (6).

For a diagonal Gaussian $q = N(\mu, \mathrm{diag}(d_j^2))$, write reverse KL up to a constant as

$$F_n(\mu, d) = \tfrac{1}{2}\|\mu\|^2 + \tfrac{1}{2}\sum_{j=1}^{p}(d_j^2 - 2\log d_j - 1) + n\mathbb{E}\phi\{U^{\top}\mu + U^{\top}\,\mathrm{diag}(d)Z\}, \quad d_j > 0, \tag{18}$$

where $Z \sim N(0, I_p)$. The first two terms are strictly convex in (μ, d). The last is convex because, for each Z, its argument is affine and ϕ is convex. The objective is finite by bounded Hessian and Gaussian moments. Its prior term is nonnegative and coercive, including at $d_j \downarrow 0$, while $\phi \geq 0$. Thus a unique interior minimizer exists.

For any fixed $U^{\top}\mu = \beta$, the minimum-norm mean is $U\beta$. An inactive row of U never enters the likelihood, so its optimum has $d_j = 1$. Set $s_j = d_j^2$, $S = U^{\top}\,\mathrm{diag}(s_j)U$, and $T \sim N(\beta, S)$. Differentiation in β and Gaussian integration by parts in d_j give

$$\beta + n\mathbb{E}\nabla\ell(T) = 0, \qquad s_j^{-1} = 1 + nu_j^{\top}\mathbb{E}\nabla^2\ell(T)u_j. \tag{19}$$

The positive semidefinite Hessian implies $s_j \leq 1$.

We next establish concentration rather than presupposing it. The trial mean $\mu = Ut_0$, active variances $s_j = 1/n$, and inactive variances one have $S = I_m/n$. Bounded Hessian gives $\mathbb{E}\phi(T) \leq L_2 m/(2n)$ at this trial. Its prior KL is at most $(s/2)\log n + C$. At the optimizer, nonnegativity of the prior KL therefore gives

$$\mathbb{E}\phi(T_n) \leq \{(s/2)\log n + C\}/n. \tag{20}$$

The growth condition forces $T_n \to t_0$ in probability. For Gaussian $T_n \sim N(\beta_n, S_n)$ this entails $\beta_n \to t_0$ and $S_n \to 0$: convergence of characteristic-function moduli gives $a^{\top}S_n a \to 0$ for every fixed $a \in \mathbb{R}^m$, and the remaining Gaussian mean must converge to the same point mass. Hence boundedness and continuity of the Hessian imply $\mathbb{E}\nabla^2\ell(T_n) \to J$.

For every active j, (19) now yields $ns_{n,j} \to h_j^{-1}$ and consequently $S_n = O(n^{-1})$. A centered Taylor expansion of the gradient, using bounded third derivatives, gives

$$\mathbb{E}\nabla\ell(T_n) = \nabla\ell(\beta_n) + O(\mathrm{tr}\, S_n).$$

The mean stationarity equation and bounded β_n imply $\nabla\ell(\beta_n) = O(n^{-1})$. Positive definiteness of J and $\beta_n \to t_0$ then imply $\beta_n - t_0 = O(n^{-1})$.

Finally, expand the Hessian at β_n. Its linear centered term averages to zero and its quadratic remainder is bounded by the fourth-derivative bound. Therefore

$$\mathbb{E}\nabla^2\ell(T_n) = \nabla^2\ell(\beta_n) + O(\mathrm{tr}\, S_n) = J + O(n^{-1}).$$

Substitution into (19) gives $ns_{n,j} = h_j^{-1} + O(n^{-1})$ for active j. Inactive rows contribute nothing to S_n, so $nS_n = K + O(n^{-1})$. □

A.4 The structured optimum and smooth predictive functions

Proof of the structured part of Theorem 3.2. The complement factors out exactly for $q = N(Ub, I - UU^{\top} + USU^{\top})$. Parameterize $S = LL^{\top}$ with lower-triangular L and positive diagonal. Its nonconstant objective is

$$\tfrac{1}{2}\|b\|^2 + \tfrac{1}{2}\|L\|_F^2 - \sum_j \log L_{jj} - m/2 + n\mathbb{E}\phi(b + LZ), \qquad Z \sim N(0, I_m).$$

It is strictly convex in (b, L) on a convex domain, with the same boundary coercivity as above. It therefore has a unique interior optimum. The trial $b = t_0$, $S = I_m/n$ has objective at most $(m/2)\log n + C$, proving concentration by the preceding argument. Differentiation with respect to b and to the symmetric covariance S gives

$$b + n\mathbb{E}\nabla\ell(T) = 0, \qquad S^{-1} = I_m + n\mathbb{E}\nabla^2\ell(T).$$

Concentration first gives $nS \to J^{-1}$. The two centered expansions used above then give $b - t_0 = O(n^{-1})$ and $nS = J^{-1} + O(n^{-1})$. The complement is unchanged by construction for every n. □

Proof of Corollary 3.3. The informed linear ratios converge to $a^\top Ka/(a^\top J^{-1}a)$, whose extremal values are the eigenvalues of $R = J^{1/2}KJ^{1/2}$. Its trace is

$$\operatorname{tr}(R) = \operatorname{tr}(D_h U J U^\top) = \sum_{j\in A} h_j^{-1} h_j = s.$$

The $s \times m$ active-row restriction of U has rank m. Its transpose has nullity $s - m$, establishing existence of supported null directions when $s > m$. Their exact laws are standard Gaussian, their mean-field means vanish, and their mean-field variances follow from (7).

For smooth f, let $e = t - t_0$ and Taylor expand

$$f(t) = f(t_0) + a^\top e + \tfrac{1}{2} e^\top \nabla^2 f(t_0) e + r(e), \qquad |r(e)| \le C\|e\|^3.$$

For the exact posterior, Lemma A.2 gives covariance $J^{-1}/n + O(n^{-2})$, raw third moments $O(n^{-2})$, and fourth and sixth absolute moments $O(n^{-2})$ and $O(n^{-3})$. For either Gaussian variational optimum, its mean displacement $O(n^{-1})$ and covariance $O(n^{-1})$ give the same moment orders. The covariance of the linear and quadratic terms is $O(n^{-2})$ by the third-moment bound; the linear–remainder covariance is $O(n^{-2})$ by the fourth-moment bound and the corresponding expectation products. The variance of the quadratic term is $O(n^{-2})$, and all remaining terms are no larger by Cauchy–Schwarz. This gives (11) and the structured counterpart. The nonzero gradient ensures a positive leading exact variance, permitting the stated ratio limits. □

A.5 Logistic assumptions, support factor, and finite-data proxy

Let $x_\ell \in \mathbb{R}^m$ span $\mathbb{R}^m$, let $\omega_\ell > 0$ sum to one, and fix success fractions $\rho_\ell \in [\eta, 1-\eta]$ for $\eta > 0$. The logistic profile

$$\ell(t) = \sum_{\ell=1}^{L} \omega_\ell \{\log(1 + \exp(x_\ell^\top t)) - \rho_\ell x_\ell^\top t\} \tag{21}$$

satisfies Assumption 3.1 (Lemma A.3 below). For the scalar case $m = 1$, $U = u$ unit, and $\ell_\rho(t) = \log(1 + e^t) - \rho t$, we have $t_0 = \log\{\rho/(1-\rho)\}$ and $J = c_\rho = \rho(1-\rho)$. For the balanced case $\rho = 1/2$, symmetry gives a zero mean and the stationarity equation

$$s_{n,j}^{-1} = 1 + n u_j^2 \mathbb{E}\sigma'(\sqrt{v_n} Z), \qquad v_n = \sum_j u_j^2 s_{n,j}, \qquad n v_n \longrightarrow 4s. \tag{22}$$

The finite-data Gaussian proxy is defined by $\kappa = n c_\rho$, $a_{\min} = \min_{j\in A} u_j^2$, and

$$R_G(\kappa, u) = (1+\kappa) \sum_{j\in A} \frac{u_j^2}{1 + \kappa u_j^2}. \tag{23}$$

This is the exact informed variance ratio for the Gaussian reference with information $\kappa u u^\top$, and

$$0 \le s - R_G = \sum_{j\in A} \frac{1 - u_j^2}{1 + \kappa u_j^2} \le \frac{s-1}{1 + \kappa a_{\min}}. \tag{24}$$

Lemma A.3 (Fixed logistic designs satisfy the nonlinear assumptions)**.** *The profile in (21), with the stated spanning and interior-fraction conditions, satisfies all assumptions of Theorem 3.2.*

Proof. The second through fourth derivatives of scalar softplus are bounded; the finite fixed features preserve these bounds. For every finite t, the Hessian is

$$\sum_\ell \omega_\ell \sigma(x_\ell^\top t)\{1 - \sigma(x_\ell^\top t)\} x_\ell x_\ell^\top \succ 0$$

by spanning. Also $\log(1 + e^a) - \rho a \ge \eta |a|$ for $\rho \in [\eta, 1-\eta]$. The function $\sum_\ell \omega_\ell |x_\ell^\top t|$ is a norm equivalent to the Euclidean norm, so $\ell(t) \ge c_0 \|t\|$. Coercivity and strict convexity yield a unique finite minimizer. Positive curvature gives a quadratic lower bound near it; linear growth gives a linear lower bound sufficiently far away after subtracting $\ell(t_0)$. On the compact annulus between these regions, the positive continuous ratio to $\min(\|t - t_0\|^2, \|t - t_0\|)$ has a positive minimum. This proves (4). □

Proof of Corollary 3.5. For $U = u$, $J = c_\rho$, $h_j = c_\rho u_j^2$, and $K = s/c_\rho$. Apply Theorem 3.2 and Corollary 3.3 to $f(t) = \sigma(b + t)$, whose bounded derivatives and strictly positive derivative at finite t_0 satisfy the hypotheses. A supported null projection has zero variational mean and variance $O(n^{-1})$. The logistic map is $1/4$-Lipschitz, so its variance is at most one sixteenth of this linear variance, for example by using two independent copies. The exact null law and its structured counterpart remain $N(0, 1)$, giving positive constant probability variance. For balance, the objective is invariant under $\mu \mapsto -\mu$; uniqueness gives $\mu = 0$ and (19) reduces to (22). □

Proposition A.4 (Finite-data Gaussian proxy)**.** *Put $\kappa = nc_\rho$, $a_{\min} = \min_{j\in A} u_j^2$. Then $R_G(\kappa, u)$ in (23) is the exact informed variance ratio for the Gaussian reference with information $\kappa uu^\top$, and the bound (24) holds. For each fixed ρ, u, b, both the nonlinear informed linear-variance ratio and the nonlinear probability-variance ratio differ from R_G by at most C/n for $n \geq n_0$, with finite model-dependent C, n_0.*

Corollary A.5 (A common covariance multiplier cannot repair both)**.** *Under Corollary 3.5, rescale the mean-field covariance by any $\gamma_n > 0$. If the informed linear variance becomes asymptotically correct, $\gamma_n \to 1/s$, and the supported-null variance still tends to zero.*

Proof of Proposition A.4 and Corollary A.5. For precision $I + \kappa uu^\top$, the exact informed variance is $(1 + \kappa)^{-1}$ and diagonal reverse KL has variance $\sum_j u_j^2/(1 + \kappa u_j^2)$ in direction u, giving (23). For each active j,

$$1 - \frac{(1 + \kappa)u_j^2}{1 + \kappa u_j^2} = \frac{1 - u_j^2}{1 + \kappa u_j^2}.$$

Summing and using $\sum_j u_j^2 = 1$ proves (24). Each nonlinear ratio is $s + O(n^{-1})$ by the established moment expansions, while (24) gives $R_G = s + O(n^{-1})$ for fixed $c_\rho, a_{\min} > 0$. The triangle inequality supplies finite C, n_0 for the displayed nonlinear-to-proxy bound. This argument does not supply sharp numerical constants or uniformity as $a_{\min}$ or c_ρ tends to zero.

For covariance rescaling, the informed variance ratio before scaling tends to s. Making it tend to one forces $\gamma_n \to 1/s$. Its supported-null variance is consequently still $O(n^{-1})$, whereas the exact variance is one. □

A.6 Coordinate equivariance

Proof of Proposition 3.7. The transformation $z = Oz'$ has unit absolute Jacobian and preserves the standard Gaussian density. It leaves $\phi_0 + Lz = \phi_0 + LOz'$ and hence every likelihood value and physical prediction unchanged. Transforming both densities in a KL integral leaves KL unchanged. Moreover,

$$O^\top(I - UU^\top + USU^\top)O = I - (O^\top U)(O^\top U)^\top + (O^\top U)S(O^\top U)^\top.$$

The transformed mean is $O^\top\mu$, so the structured families correspond bijectively. Diagonal matrices do not have this property under general orthogonal conjugation; only signed coordinate permutations preserve the entire diagonal family. For a Gaussian reference, choosing an orthogonal eigenbasis of its precision makes the target diagonal. For a nonlinear target this diagonalizes only the selected local curvature and does not generally remove higher-order dependence. □

A.7 Probability and variance response along Gaussian covariance paths

For reference, the explicit remainder bound is: if $\sup_z \|D^4 f(z)\|_F \leq M_4$, then

$$\left|\mathbb{E}_{q_t} f - \mathbb{E}_{q_0} f - \frac{t}{2}\operatorname{tr}\{\Delta\, \mathbb{E}_{q_0}\nabla^2 f\}\right| \leq \frac{t^2}{8} M_4 \|\Delta\|_F^2. \tag{25}$$

Proof of Lemma 4.1. Let φ_t be the density of $N(\mu, \Sigma_0 + t\Delta)$. Differentiating the Gaussian density gives the heat-equation identity

$$\partial_t \varphi_t(z) = \tfrac{1}{2}\sum_{ij} \Delta_{ij}\partial_{z_i}\partial_{z_j}\varphi_t(z).$$

On a compact interval of positive definite covariance matrices, Gaussian tails dominate the polynomial derivatives in the assumptions. Differentiation under the integral and two integrations by parts are therefore justified, yielding (13). Applying the same identity once more gives

$$\frac{d^2}{dt^2}\mathbb{E}_{q_t} f = \tfrac{1}{4}\sum_{ijkl} \Delta_{ij}\Delta_{kl}\mathbb{E}_{q_t}\partial_{ijkl} f.$$

The Frobenius norm of the tensor with entries $\Delta_{ij}\Delta_{kl}$ is $\|\Delta\|_F^2$. Cauchy–Schwarz gives a second derivative bound $M_4\|\Delta\|_F^2/4$, and the integral Taylor remainder gives (25). With polynomial growth alone the second derivative is still bounded on a compact interior path, giving a local $O(t^2)$ remainder. □

Proof of Theorem 4.2. Apply Lemma 4.1 to h and h^2 and differentiate $\mathrm{Var}_{q_t} h = \mathbb{E}_{q_t} h^2 - (\mathbb{E}_{q_t} h)^2$. At zero its derivative is $\mathrm{tr}(\Delta B_h)$, with a locally bounded second derivative. Orthogonal projection gives

$$\langle \Delta, H_\ell^{\mathrm{prob}} \rangle_F = 0, \qquad \langle \Delta, B_h \rangle_F = \|B_\perp\|_F.$$

These identities give the two response orders. Since $\|\Delta\|_{\mathrm{op}} \le \|\Delta\|_F = 1$, the minimum eigenvalue of the perturbed covariance is at least $\lambda_{\min}(\Sigma_0) - |t| > 0$. If $h = a^\top z$, its variance is exactly $a^\top(\Sigma_0 + t\Delta)a$, so the second response is exactly linear. □

For a nonconstant mean path μ_t, differentiation also contributes $\dot{\mu}_t^\top \mathbb{E}_{q_t}\nabla f$ to a probability response. Thus the fixed-mean premise cannot be dropped when comparing trained approximations with different means.

A.8 Proper scores, nonidentification, and acquisition quantities

For reference, the proper-score expansions used in Remark 4.3 are as follows. For multiclass Brier loss,

$$R_{\mathrm{Br},\tau}(q) - R_{\mathrm{Br},\tau}(p) = 2(p-\tau)^\top \delta + \|\delta\|_2^2. \tag{26}$$

For log loss with $p_y \ge \alpha > 0$ and $\|\delta\|_\infty \le \alpha/2$,

$$R_{\log,p}(q) - R_{\log,p}(p) = \mathrm{KL}(p\|q) = \frac{1}{2}\sum_y \frac{\delta_y^2}{p_y} + O_\alpha(\|\delta\|_2^3). \tag{27}$$

For $Z \sim N(0,1)$ and $P_v = \sigma(\sqrt{v}Z)$, the mutual information is

$$\mathrm{MI}(v) = \log 2 - \mathbb{E}\mathsf{h}(P_v), \qquad \mathsf{h}(p) = -p\log p - (1-p)\log(1-p). \tag{28}$$

More generally, if $P_y \ge \alpha > 0$ almost surely and $\bar{p} = \mathbb{E}P$,

$$H(\bar{p}) - \mathbb{E}H(P) = \frac{1}{2}\sum_y \frac{\mathrm{Var}(P_y)}{\bar{p}_y} + O_\alpha\big(\mathbb{E}\|P-\bar{p}\|_2^3\big). \tag{29}$$

Brier risk. For an outcome indicator e_Y with $\mathbb{E}e_Y = \tau$,

$$\mathbb{E}_\tau\|q - e_Y\|^2 = \|q\|^2 - 2\tau^\top q + 1.$$

Subtract the same expression at p and substitute $q = p + \delta$ to obtain (26). This is exact for any two probability vectors.

Log risk. For $p_y \ge \alpha$ and $|\delta_y| \le p_y/2$, Taylor expansion of $-\log(1+\delta_y/p_y)$ has remainder bounded by a constant times $|\delta_y/p_y|^3$. Under $\tau = p$, the linear terms sum to $-\sum_y \delta_y = 0$, giving (27). More generally,

$$R_{\log,\tau}(p+\delta) - R_{\log,\tau}(p) = -\sum_y \frac{\tau_y}{p_y}\delta_y + \frac{1}{2}\sum_y \frac{\tau_y}{p_y^2}\delta_y^2 + O_\alpha(\|\delta\|_2^3).$$

The linear term need not cancel. A single realized label corresponds to a point-mass τ and likewise has a linear term in general. Substituting $\delta = O(t^2)$ proves the response orders stated after Theorem 4.2; no analogous smooth expansion was asserted for ECE, accuracy, or ranking metrics.

Exact equal-mean example. Symmetry and $\sigma(-a) = 1 - \sigma(a)$ imply $\mathbb{E}\sigma(\sqrt{v}Z) = 1/2$. All integrated Bernoulli predictive distributions are identical, so every realized-label log loss and Brier score is identical. For $Z \neq 0$, $|\sigma(\sqrt{v}Z) - 1/2|$ is strictly increasing in v, making its squared expectation, $\mathrm{Var}(P_v)$, strictly increasing. Binary entropy is strictly decreasing as $|p - 1/2|$ increases, so $\mathsf{h}(\sigma(\sqrt{v}Z))$ strictly decreases almost surely when v increases; its expectation does also. Hence (28) is strictly increasing.

Near zero the scalar functions have expansions

$$\{\sigma(x) - 1/2\}^2 = x^2/16 + O(x^4), \qquad \mathsf{h}(\sigma(x)) = \log 2 - x^2/8 + O(x^4).$$

Their fourth derivatives are globally bounded (the entropy composition is smooth with exponentially decaying derivatives times polynomials). Taylor's theorem and $\mathbb{E}Z^4 = 3$ therefore justify integration of the remainders, proving the stated small-v expansions. The Gaussian tails pose no probability-floor assumption in this example.

General entropy expansion. For $P_y \geq \alpha$ almost surely, the segment between P and $\bar{p} = \mathbb{E}P$ stays in that interior region. The entropy Hessian at $\bar{p}$ is $-\mathrm{diag}(1/\bar{p}_y)$, and its third derivatives are bounded there by α^{-2}. Taylor expansion of $H(P)$, followed by expectation, removes the linear term and gives (29), with an absolute remainder bounded by $(6\alpha^2)^{-1}\mathbb{E}\|P - \bar{p}\|_2^3$.

One-step acquisition approximation. For a finite candidate set, let A be a reference acquisition objective and $\widehat{A}$ an approximation satisfying $\sup_x |\widehat{A}(x) - A(x)| \leq \epsilon$. If $x^\star$ maximizes A and $\widehat{x}$ maximizes $\widehat{A}$, then

$$A(x^\star) - A(\widehat{x}) \leq A(x^\star) - \widehat{A}(x^\star) + \widehat{A}(\widehat{x}) - A(\widehat{x}) \leq 2\epsilon.$$

The omitted middle term $\widehat{A}(x^\star) - \widehat{A}(\widehat{x})$ is nonpositive. This is objective regret for one selection, not a guarantee of downstream error reduction or multiround learning. No such uniform acquisition error is established for the current LLM fits.

B Experimental details and additional audits

B.1 Operator and anchor convention

A model Fisher averages score outer products over model-sampled labels; an observed-label score outer product does not generally equal that expectation or the observed likelihood Hessian. At a common fixed anchor, summed model Fisher information is additive across disjoint batches. Recomputing at each fitted MAP changes this relation. These distinctions matter empirically: throughout the controlled experiments we name the nonlinear HMC reference, the local information operator, the anchor, and the predictive functional separately.

B.2 HMC reference diagnostics

The Qwen HMC reference uses four NUTS chains with 250 warmup and 500 retained draws per chain, target acceptance 0.8, and maximum tree depth 6. It was audited on the same 300 predictive functions used in the main text: the worst rank-normalized/folded $\widehat{R}$ is 1.0094, the minimum centered-square ESS is above 800, and all chains report zero divergences. Phi is independently cached and validated; the final certification uses four fresh chains with 500 warmup and 1500 retained draws each. Its locked gate passes with parameter max $\widehat{R} = 1.0044$, minimum parameter bulk ESS 10703, predictive max $\widehat{R} \leq 1.0031$, and zero divergences. Restricted HMC is a numerical reference for these 64-dimensional posteriors, not for the full adapter. The Phi certification details appear in Appendix B.11.

B.3 Post-review experiment ledger

The final audit sequence separates theorem verification from transformer diagnostics. The deterministic nonlinear verification checks the support-factor, structured-recovery, supported-null, and multidirectional trace limits under the assumptions of Theorem 3.2. The existing-results audit then

uses saved Qwen HMC and variational fits to measure finite-sample trace diagnostics, predictive energy in the generalized covariance eigenbasis, and achieved variational objectives. The direct $B_\perp$ audit constructs the covariance direction in Theorem 4.2 on a frozen prompt panel, estimates the probability-response span with independent Gaussian-Stein draw banks, and evaluates an exact cached nonlinear covariance path. None of these audits refits the transformer backbone.

B.4 Deterministic nonlinear verification

For the scalar logistic construction with four active coordinates, the MF probability-variance ratio at $n = 10^5$ is 3.997–3.998 across the tested unbalanced success fractions, while the structured ratio is 1.0000 and the supported-null MF linear variance is at most 6.3×10^{-4}. In the $m = 3$, $s = 9$ construction, the limiting generalized eigenvalues are $(1.67105, 3.15216, 4.17679)$ with trace 9; at $n = 10^5$ they are $(1.67095, 3.15155, 4.17581)$ with trace 8.99831.

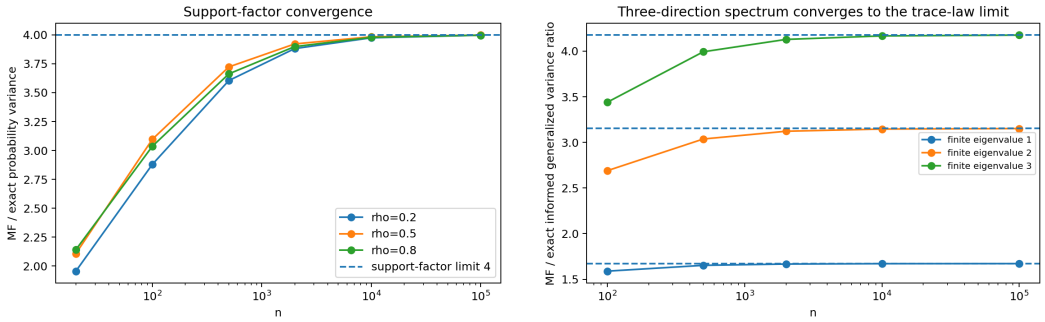


Figure 4: Deterministic nonlinear verification. Left: the MF predictive-variance ratio approaches the active-coordinate support factor while the structured approximation approaches one. Right: the multidirectional generalized spectrum approaches the three predicted limits whose sum is nine.

B.5 Finite-sample trace diagnostic and predictive eigenspace alignment

The full 64-dimensional generalized MF/HMC traces are 68.49, 66.81, and 76.26 for $n = 100, 500, 2000$. In the theorem's asymptotic model, the corresponding full-space trace tends p, not s, because the informed contribution s is supplemented by $p - s$ inactive-coordinate ratios equal to one; the full Qwen trace is therefore not a support diagnostic. For the restricted audit, we project both covariances onto the leading eigenspace of the same $n = 2000$-MAP categorical Fisher at every sample size. At $n = 2000$, relative Fisher thresholds 10^{-2} and 10^{-3} give rank 9 and trace 23.5507, while threshold 10^{-4} gives rank 10 and trace 24.4625. The matched mean generalized ratios are 2.6167 and 2.4462, respectively, which bracket the median predictive-direction inflation 2.58. This numerical agreement is a finite-sample consistency check, not an identification of the theorem's s: all 64 coordinate rows have nonzero leverage under the numerical tolerance and the inverse-participation effective supports are 53.0–54.2. Moreover, the restricted trace is a mean over Fisher directions whereas 2.58 is a median over predictive directions. The independent eigenspace audit strengthens the connection: 9 of 64 generalized directions have ratio above two but receive 78.4% of predictive energy on average (78.8% median), and the top eight directions receive 75.9% median energy.

Relative Fisher threshold	Rank	Restricted trace	Trace/rank
$1e-01$	9	23.5507	2.6167
$1e-02$	9	23.5507	2.6167
$1e-03$	9	23.5507	2.6167
$1e-04$	10	24.4625	2.4462

Table 4: Original common-anchor $n = 2000$ Fisher-restricted trace across preselected relative rank thresholds. Rank 9 is stable from 10^{-1} through 10^{-3}; at 10^{-4} an additional weak Fisher direction enters. These traces are finite-sample diagnostics, not estimates of theorem active support.

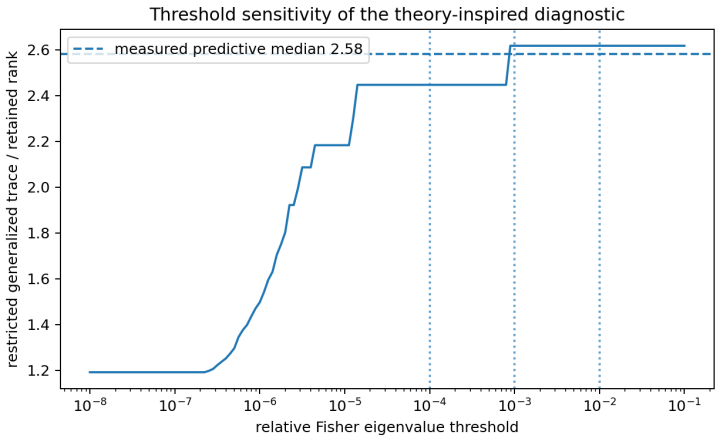


Figure 5: Dense relative-Fisher threshold sweep for the restricted-trace/rank diagnostic. The rank-9 plateau is broad; dropping the threshold sufficiently far admits further low-information directions and changes the mean generalized ratio.

B.6 Frozen same-model score and selective-risk audit

The compact R0/R1 cache lacked the original calibration prompts, so the initial audit could compute descriptive scores but could not validly redo the calibration protocol. The final audit imports the frozen complete `qwen_lastdown_core` cache (8544 prompt records) and checks every one of the original 1200 calibration and 1500 test IDs against its manifest before evaluation. It applies the already fitted restricted $p = 64$ fixed-mean MF and Prediction-PC16 endpoints; neither the physical model, fitted posterior, nor HMC reference is altered. The original scoring profile is `plain_answer_newline` for `Qwen/Qwen3-4B-Instruct-2507`. The frozen covariance-fidelity panel contains 300 prompts; the held-out 1500-prompt score panel is reported separately. The original split uses a repurposed MMLU-Pro source split rather than a standard benchmark evaluation, and these panels should not be treated as independent populations.

For the fidelity and held-out score calculations, each posterior is represented by 4096 fixed-seed Gaussian draws; the calibration threshold scores use 1024 draws. Predictive probability vectors are posterior-mean answer-choice probabilities. The multiclass Brier loss is evaluated on the same choice-normalized categorical vector; ECE uses 15 equal-width confidence bins with accuracy minus confidence in absolute value. The selective procedure constructs 50 frozen calibration-quantile thresholds separately for each score (mutual information, predictive entropy, or $1-$confidence); each candidate's number of calibration errors receives a one-sided Clopper–Pearson bound at confidence level $1 - 0.05/50$, and the largest accepted threshold meeting its target error bound is evaluated on the original 1500 held-out prompts. If no threshold passes, the certified answer rate is zero, not a missing result. In the final audit no threshold passes for either approximation at $\alpha = 0.05, 0.10, 0.20$ and any of the three scores. The smallest calibrated upper risks are 0.2249 (MI) and 0.2155 (entropy and $1-$confidence). The calibration-point answer-rate denominator is the number of frozen calibration prompts (1200), and the held-out answer-rate denominator is 1500; the earlier v2 output mistakenly used the number of dictionary fields in its displayed calibration rate. The v3 selective table corrects that display error. It does not affect which thresholds pass or the all-zero certification result.

Covariance	MI AURC	Entropy AURC	$1-$confidence AURC
MF	0.374710	0.355167	0.355532
Prediction-PC16	0.374347	0.355181	0.355538

Table 5: Held-out 1500-prompt area under the risk–coverage curve (AURC), calculated under the same restricted posterior as the covariance-fidelity experiment. Small method differences do not establish equivalence or improved risk control. No candidate threshold was certified at any of the three prespecified error targets.

B.7 Least-saturated robustness for the fitted covariance path

The weak probability response in Section 5 could be confounded by softmax saturation. Without rerunning the model, we therefore formed a post-hoc robustness subset from the saved path output: the 30 prompts with largest $\min\{\bar{p}_{\mathrm{MF}}, 1-\bar{p}_{\mathrm{MF}}\}$ at the MF endpoint. Their baseline focal probabilities range from 0.3803 to 0.6160. Between the MF and Prediction-PC covariance endpoints, the median absolute relative focal-probability change on this subset is 0.01095%, whereas the median absolute relative log-odds-variance change is 37.78%; 27 of 30 prompts decrease in log-odds variance. The subset is explicitly post hoc and is used only to show that near-pinned probabilities are not the sole source of the response separation.

B.8 Independent probability-response audit and Monte Carlo limits

The original fitted MF-to-PC path uses a common fixed restricted posterior mean μ, with midpoint covariance $\Sigma_0 = (\Sigma_{\mathrm{MF}} + \Sigma_{\mathrm{PC}})/2$ and derivative direction $\Delta = \Sigma_{\mathrm{PC}} - \Sigma_{\mathrm{MF}}$. To test the probability version of Lemma 4.1, the final closure experiment predicts each prompt's focal-probability derivative from an independent Gaussian-Stein sample. With $\Sigma_0 = LL^\top$, $Z = \mu + L\varepsilon$, $\varepsilon \sim N(0, I)$ and $A = L^{-1}\Delta L^{-\top}$, the estimator is based on

$$\frac{d}{dt}\mathbb{E}_{N(\mu,\Sigma_0+t\Delta)} f_j(Z)\Big|_{t=0} = \frac{1}{2}\mathbb{E}\big[\{f_j(\mu+L\varepsilon) - f_j(\mu)\}\{\varepsilon^\top A\varepsilon - \mathrm{tr}(A)\}\big]. \tag{30}$$

Subtracting the deterministic control variate $f_j(\mu)$ is *exact*, because $\mathbb{E}(\varepsilon^\top A\varepsilon - \mathrm{tr}\, A) = 0$; it avoids assigning spurious derivative variation to a nearly constant (saturated) probability. The final calculation uses 8192 antithetically paired Gaussian evaluations (4096 independent pairs), estimates the Monte Carlo standard error from pair means, and compares the result with an independent 4096-draw common-random-number central finite difference with covariance steps $\Sigma_0 \pm 0.025\Delta$. Independent draw banks keep the comparison from inheriting matched Monte Carlo noise.

Across all 300 prompts, predicted versus finite-difference derivatives have Pearson correlation 0.758, Spearman 0.524, and through-origin regression slope 0.968. The median absolute predicted derivative is 3.28×10^{-7}, the median absolute measured derivative is 5.37×10^{-7}, and their median absolute discrepancy is 4.02×10^{-7}. Their median reported Monte Carlo standard errors are 1.36×10^{-7} and 5.84×10^{-7}, respectively. At the level of the reported independent Monte Carlo errors, 285 of 300 promptwise discrepancies are within two combined standard errors and all 300 within three (the largest standardized discrepancy is 2.49). These compatibility diagnostics depend on the estimated standard errors and are not formal proof of the analytic formula, which is already exact under Lemma 4.1.

For the 30 least-saturated focal probabilities in this independent derivative audit, baseline focal probabilities range from 0.381 to 0.615, Pearson correlation is 0.346, and the median absolute measured derivative 1.23×10^{-4} is below its median Monte Carlo standard error 1.56×10^{-4}. The through-origin slope there is 0.493. Therefore this smaller subgroup does not establish accurate prompt-by-prompt probability-derivative recovery. It complements, but is distinct from, the previously reported post-hoc *endpoint* robustness check in the preceding subsection; both suggest weak probability response relative to log-odds-variance response, without proving the exact conditional-blindness orthogonality premise.

B.9 Generic low-rank objective audit

The corrected generic rank-16+diagonal fits use the standard latent prior $N(0, I)$. All three generic fits have both worse HMC log-variance fidelity and lower achieved ELBO than diagonal MF, so the experiment cannot distinguish family inadequacy from optimization failure. Prediction-PC16, by contrast, reduces median HMC log-variance error from 0.840 to 0.348 while its ELBO difference from diagonal MF is only $+1.73$ with Monte Carlo $z = 1.03$ (standard error 1.7). We therefore use this row as evidence that the covariance observable can differ sharply even when the achieved-objective difference is unresolved at current precision, not as evidence that the two ELBOs are equal. The generic fits are retained as an *optimization-failure control*, not as a comparison of covariance families. Fitting them longer or initializing them differently could change their outcomes; the available runs do not isolate representation quality from optimization. Earlier generic-low-rank runs incorrectly reused

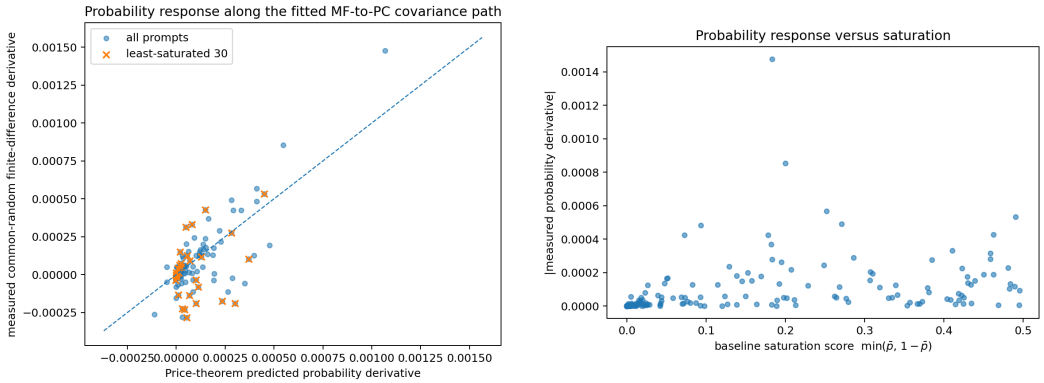


Figure 6: Independent focal-probability derivative audit along the frozen MF-to-PC covariance path. Left: control-variate Price/Stein predictions and independent finite differences; the least-saturated 30 are marked separately. Right: the response estimates against the initial probability-saturation score. Large relative uncertainty changes should not be conflated with numerically precise prompt-level probability derivatives.

Approximation	ELBO − MF	z vs. MF	HMC log-var error
Diagonal MF	0.00	–	0.840
Prediction-PC16	+1.73	1.03	0.348
Generic LR16, seed 1	−12.89	−4.05	1.957
Generic LR16, seed 2	−13.16	−5.64	1.874
Generic LR16, seed 3	−17.87	−7.69	2.030

Table 6: Achieved full-data ELBO audit for the saved Qwen approximations. The generic low-rank fits are worse than diagonal MF on both the variational objective and HMC covariance fidelity; accordingly, no family-level superiority claim is made.

the physical adapter prior scale inside the latent KL and are excluded from all manuscript results; their existence is documented in `provenance/EXCLUDED_SUPERSEDED_RUNS.md`.

B.10 Direct constructed $B_\perp$ audit

The frozen direct test uses 24 evenly spaced Qwen evaluation prompts, the midpoint covariance of fixed-mean MF and Prediction-PC16, and a linear uncertainty functional chosen before path outcomes as the leading principal direction of the 300 saved log-odds sensitivities. The probability-response span is estimated by a Gaussian Stein identity with 8192 antithetic construction draws and independently re-estimated with another 8192 draws. The constructed direction has $\|B_\perp\|_F = 0.584$ and therefore an exact linear uncertainty slope 0.584. Its independent RMS probability Price slope is 3.21×10^{-4}, versus 2.42×10^{-3} for the frozen generic symmetric control, a ratio 0.133. Along the exact cached nonlinear path, the largest tested displacement changes the selected linear uncertainty variance by about $\pm 12.4\%$, while the median absolute focal-probability change on the construction panel is only a few 10^{-9} and the maximum is below 3.7×10^{-5}.

The preregistered finite-path control comparison nevertheless fails: at the smallest tested path scale, the RMS probability slope is 4.94×10^{-5} for the constructed direction and 4.19×10^{-6} for the random control, so the former is roughly 11.8 times larger. This comparison has a different outcome from the independent *first-order* response-bank comparison (ratio 0.133), and neither can be substituted for the other. A numerical issue is visible in the construction: the estimated probability-response system has a largest singular value about 2.48 and smallest singular value about 3.6×10^{-12}, a condition number of order 6.9×10^{11}. Orthogonal projection onto nearly singular estimated response directions can be sensitive to Monte Carlo perturbations; moreover, the finite-path slopes may be close to the Monte Carlo floor. These are plausible contributors to the discrepancy, *not* demonstrated explanations of the failed preregistered comparison. Post-hoc singular-value truncation would define a different constructed direction and cannot retrospectively turn the original test into a success. Accordingly, we preserve both results, refrain from claiming empirical verification of the orthogonality premise, and treat the direct construction as mixed mechanistic evidence.

Quantity	Value
$\|B_\perp\|_F = a^\top \Delta a$	0.584
Construction RMS probability Price slope	7.9×10^{-18}
Independent cross-fit RMS slope, $B_\perp$	3.21×10^{-4}
Independent cross-fit RMS slope, generic control	2.42×10^{-3}
Cross-fit ratio ($B_\perp$/control)	0.133
Finite-path smallest-scale RMS slope, $B_\perp$	4.94×10^{-5}
Finite-path smallest-scale RMS slope, control	4.19×10^{-6}
Finite-path ratio ($B_\perp$/control)	11.8

Table 7: Direct constructed-direction audit. The independent first-order response is suppressed relative to the generic control, but the preregistered finite-path control comparison does not favor the constructed direction. The construction-sample Price slope is zero to floating-point precision by construction, since the direction is projected orthogonal to the response span estimated from those same draws; the independent cross-fit rows quantify the out-of-sample response.

B.11 PHI REFERENCE CERTIFICATION

The preliminary Phi reference narrowly missed the locked parameter $\widehat{R} \leq 1.01$ threshold. The certification therefore ran four fresh NUTS chains with 500 warmup and 1500 retained draws per chain, leaving target acceptance, tree-depth limit, and every gate threshold unchanged. The certified gate has max parameter rank-$\widehat{R} = 1.00442$, minimum parameter bulk ESS 10703, max log-odds rank-$\widehat{R} = 1.00191$, max probability rank-$\widehat{R} = 1.00314$, and zero divergences. The ten identity/Fisher VI fits were not retrained. The seedwise Fisher-minus-identity effect changed by at most 0.047 relative to the preliminary HMC reference.