CERTIFICATION FRONTIERS FOR GAUSSIAN LORA: INDEPENDENT PRIORS, POSTERIOR RISK, AND PREDICTION-PRESERVING BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc Bayesian fine-tuning places Gaussians around trained low-rank adapters, yet a calibrated posterior does not by itself yield a useful generalization certificate. Such a posterior admits an informative PAC-Bayes certificate only when both the loss of its sampled predictors and its KL divergence from an admissible prior are small. In this research, we characterize this certification frontier for Gaussian LoRA posteriors and separate three interventions: changing the prior, changing the stochastic predictor, and changing only how its complexity is counted. First, an exact isotropic KL envelope eliminates the prior scale and yields a width threshold that excludes posterior widths before sampling, while a zero-KL floor identifies targets that no complexity reduction can reach at a measured risk bound. Second, we minimize KL in closed form over the full $GL(r)$ symmetry of the low-rank factors, leaving every sampled adapter product unchanged, and derive the noncentral objective required when the prior center is trained on an independent split. On a small-pool RoBERTa audit of 567 configurations, the recorded 64-draw summaries imply a certificate floor of 0.7298 even with zero KL and Chernoff accounting, so reducing complexity alone cannot certify these posteriors at the recorded budgets. For a stable posterior in a controlled Gaussian-factor task, Chernoff accounting certifies risk below 0.1 on 20 of 20 datasets with 1024 draws, whereas Hoeffding certifies none. On deliberately deformed synthetic rank-four factors, matrix balancing reduces KL by 29.3% on average beyond scalar balancing without changing any prediction. Numerical split-prior scenarios make the remaining risk and complexity budgets explicit.

1 INTRODUCTION

A Gaussian posterior over a trained low-rank adapter yields a useful PAC-Bayes certificate only when two quantities are small at once: the loss of the sampled predictors and the posterior's KL divergence from an admissible prior. Neither follows from LoRA's small factor count (Hu et al., 2022) or from the accuracy of the trained checkpoint, because the certificate concerns the loss averaged over sampled weights. We refer to the set of attainable pairs of risk and complexity as the *certification frontier* and study three questions: which posterior widths preserve prediction, which priors make those widths affordable, and which costs can be removed without altering any prediction.

These questions arise directly in post-hoc Bayesian fine-tuning. Laplace-LoRA fits a Gaussian around a trained adapter (Yang et al., 2024), and Training-Free Bayesianization searches for the largest admissible noise scale around one (Shi et al., 2025); BLoB instead learns the posterior during fine-tuning (Wang et al., 2024). All three target calibration, which does not by itself yield a PAC-Bayes certificate. Existing nonvacuous certificates rely on suitable priors, stochastic training, or compression (Dziugaite & Roy, 2017; Dziugaite et al., 2021; Dong, 2025; Lotfi et al., 2024a).

Our analysis separates what a posterior predicts from how its complexity is counted (Figure 1). We make three contributions:

- **Feasibility tests that need no prior scale.** An exact isotropic KL envelope yields a width threshold that excludes posterior widths before any sampling, and a complexity-free floor identifies targets that no KL correction can reach at the measured risk bound (Section 3).

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.

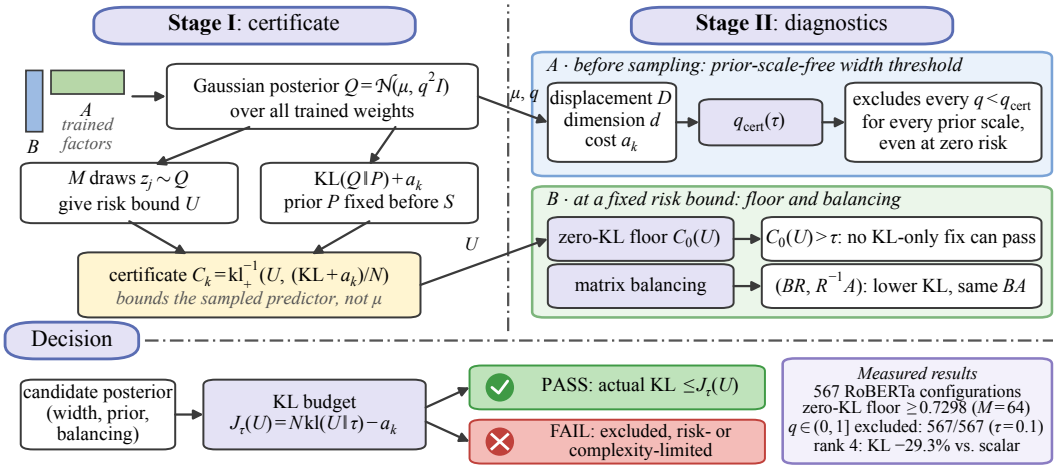


Figure 1: Overview. **Stage I** combines a Monte Carlo risk bound U for the Gaussian posterior Q with $\mathrm{KL}(Q\|P) + a_k$ into the certificate C_k. **Stage II** locates Q on the risk–complexity frontier: (A) a width threshold applied before sampling; (B) at fixed U, the zero-KL floor and matrix balancing, which leaves every sampled BA unchanged. **Decision**: pass when the actual KL fits $J_\tau(U)$.

- **Prediction-preserving matrix balancing.** We minimize the KL in closed form over the full $GL(r)$ factor symmetry while every sampled product BA stays unchanged, and we derive the noncentral objective required when the prior center is trained on an independent split (Sections 4 and 5).
- **Evidence on both sides of the frontier.** On 567 RoBERTa configurations, the recorded 64-draw summaries imply a certificate floor of 0.7298 even with zero KL and Chernoff accounting. Controlled tests show where tighter sampling helps, and matrix balancing gives a 29.3% mean KL reduction beyond scalar balancing on deformed rank-four factors (Section 6).

2 The Stochastic Predictor and Its Certificate

We certify a stochastic predictor that draws its trained weights z from a posterior Q. Let $S = ((x_i, y_i))_{i=1}^N \sim \mathcal{D}^N$, with $N \geq 2$ and a loss $\ell \in [0, 1]$. A configuration k fixes adapter placement, ranks, scaling, prior, and all other trained blocks. LoRA replaces a frozen matrix $W_0^{(\ell)}$ by $W_0^{(\ell)} + s_\ell B^{(\ell)} A^{(\ell)}$ with $\sum_\ell r_\ell(m_\ell + n_\ell)$ factor entries, but the certified vector $z \in \mathbb{R}^{d_k}$ must also include the classification head and every other parameter trained on S. Table 4 in Appendix A summarizes the notation.

For a posterior Q, define

$$R_{\mathcal{D}}(Q) = \mathbb{E}_{z\sim Q}\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(f_{k,z}(x), y), \qquad \widehat{R}_S(Q) = \mathbb{E}_{z\sim Q}\frac{1}{N}\sum_{i=1}^{N}\ell(f_{k,z}(x_i), y_i). \tag{1}$$

We refer to the posterior mean μ, which is the trained checkpoint, as the *center*. Its risk $\widehat{R}_S(\delta_\mu)$ can differ substantially from $\widehat{R}_S(Q)$; hence adding a Gaussian KL to a center score does not yield a certificate.

Theorem 1 (PAC-Bayes-kl; Maurer, 2004; Alquier, 2024)**.** *For a prior P fixed independently of S, with probability at least $1-\delta$ over S, simultaneously for all $Q \ll P$,*

$$\mathrm{kl}\Big(\widehat{R}_S(Q)\Big\|R_{\mathcal{D}}(Q)\Big) \leq \frac{\mathrm{KL}(Q\|P) + \log(2\sqrt{N}/\delta)}{N}. \tag{2}$$

Here kl is binary relative entropy and $\mathrm{kl}_+^{-1}(u, c)$ is the largest $v \in [u, 1]$ such that $\mathrm{kl}(u\|v) \leq c$. A prior may be trained on an independent sample S_0; conditioning on S_0 leaves the theorem applicable with denominator $|S|$, not $|S_0| + |S|$.

Proposition 2 (Configuration accounting). *Let $\mathcal{K}$ be countable and fixed independently of S, and let P_k and $w_k > 0$, $\sum_k w_k = 1$, be data-independent priors and weights. On the disjoint union, set $P = \sum_k w_k \delta_k \otimes P_k$. Then*

$$\mathrm{KL}(\delta_k \otimes Q_k \| P) = \mathrm{KL}(Q_k \| P_k) + \log(1/w_k). \tag{3}$$

Consequently, the PAC-Bayes event permits selection of k and Q_k using S.

Rank and prior choices are charged through w_k. A search over posterior widths incurs no additional penalty when P and the hypothesis space are fixed; the reuse of certification draws during that search is addressed by the Monte Carlo bound below.

The empirical risk $\widehat{R}_S(Q)$ is estimated by sampling. For each final frozen posterior, draw $z_1, \ldots, z_M \sim Q$ independently and let $Y_j = N^{-1} \sum_i \ell(f_{z_j}(x_i), y_i)$; the independent units are the M draws, not the MN predictions. For K_{MC} posteriors certified together, Hoeffding's inequality gives the simultaneous upper bound

$$U_H = \min\left\{1, \widetilde{r} + \sqrt{\frac{\log(K_{\mathrm{MC}}/\eta)}{2M}}\right\}, \qquad \widetilde{r} = \frac{1}{M}\sum_j Y_j. \tag{4}$$

Because the Y_j are bounded (though not Bernoulli), the tighter Chernoff inversion is also valid (Appendix C):

$$U_{\mathrm{kl}} = \mathrm{kl}_+^{-1}\left(\widetilde{r}, \frac{\log(K_{\mathrm{MC}}/\eta)}{M}\right). \tag{5}$$

With either upper bound U, the certificate is

$$C_k = \mathrm{kl}_+^{-1}\left(U_k, \frac{\mathrm{KL}(Q_k\|P_k) + \log(1/w_k) + \log(2\sqrt{N}/\delta_{\mathrm{PB}})}{N}\right). \tag{6}$$

It holds with probability at least $1 - \delta_{\mathrm{PB}} - \eta$ and bounds the risk of the stochastic predictor, not of its center. Minimizing C_k over k gives a selector; the accuracy of the center it selects is a separate empirical quantity.

3 Which isotropic certificates are feasible?

A useful posterior must preserve prediction while staying close to its prior. For isotropic Gaussians, the complexity side of this tradeoff can be optimized exactly over the prior scale, which yields tests that require no risk estimate.

Theorem 3 (Oracle isotropic envelope). *Fix centers $m, \mu \in \mathbb{R}^d$, write $D = \|\mu - m\|_2^2$, and let $Q_q = \mathcal{N}(\mu, q^2 I)$, $P_p = \mathcal{N}(m, p^2 I)$. Then*

$$\inf_{p>0} \mathrm{KL}(Q_q\|P_p) = \frac{d}{2}\log\left(1 + \frac{D}{dq^2}\right), \qquad p_*^2 = q^2 + D/d. \tag{7}$$

For $a_k = \log(1/w_k) + \log(2\sqrt{N}/\delta_{\mathrm{PB}})$ and $r(q) = \widehat{R}_S(Q_q)$, the ideal certificate using $r(q)$, and every certificate using an upper bound $U \geq r(q)$, can satisfy $C \leq \tau < 1$ only if $r(q) \leq \tau$ and

$$\frac{d}{2}\log\left(1 + \frac{D}{dq^2}\right) + a_k \leq N\,\mathrm{kl}(r(q)\|\tau). \tag{8}$$

Appendix C.3 gives the proof. The optimal scale p_* depends on S, so it is not an admissible prior. Minimizing over it can only strengthen an exclusion and cannot establish a certificate.

Corollary 4 (Prior-scale-free width threshold). *Suppose $D > 0$ and $\tau \in (0, 1)$. Define*

$$b_\tau = N\log\frac{1}{1-\tau} - a_k, \qquad q_{\mathrm{cert}}(\tau) = \begin{cases} \infty, & b_\tau \leq 0, \\ \sqrt{\dfrac{D}{d\,[\exp(2b_\tau/d) - 1]}}, & b_\tau > 0. \end{cases} \tag{9}$$

Any passing isotropic posterior has $q \geq q_{\mathrm{cert}}(\tau)$. Thus $q_{\mathrm{cert}} > q_{\max}$ excludes every $q \in (0, q_{\max}]$, for every isotropic prior scale, even if its empirical stochastic risk were zero.

The threshold grants zero empirical risk, since $\mathrm{kl}(r\|\tau) \leq \mathrm{kl}(0\|\tau)$ for $0 \leq r \leq \tau$, so no number of posterior draws can reverse an exclusion. A width above the threshold is only unexcluded and may still destroy prediction. Block-anisotropic posteriors fall outside this result and need their own KL.

3.1 THE BEST POSSIBLE ACCOUNTING-ONLY CORRECTION

We next ask whether any reduction in KL could change the decision, holding fixed the predictor, its draw-level losses, the sample size, and the confidence allocation.

Proposition 5 (Complexity-free certificate floor). *Let $U \in [0,1]$ be the fixed empirical-risk upper bound and $a_k = \log(1/w_k) + \log(2\sqrt{N}/\delta_{\mathrm{PB}})$. Every value of* (6) *with a nonnegative KL satisfies*

$$C_k \geq C_{0,k}(U) := \mathrm{kl}_+^{-1}\Big(U, \frac{a_k}{N}\Big) \geq U. \tag{10}$$

If $C_{0,k}(U) > \tau$, no correction that changes only the KL can make this certificate pass τ.

The bound holds because $\mathrm{kl}_+^{-1}(U, c)$ is nondecreasing in c. The floor is an arithmetic relaxation, not a certificate: setting the KL to zero does not make $Q = P$. Because the balancing of Section 4 preserves every draw-level loss, the same U, and hence the same floor, applies before and after balancing. The floor therefore separates a removable complexity cost from a binding risk or confidence cost.

4 CORRECTING KL WITHOUT CHANGING A STOCHASTIC PREDICTOR

Predictions depend only on the product BA, but the KL of a factor posterior also depends on how that product is split between the factors. Rescaling only the posterior mean would change the stochastic model, so we map every draw through $(B, A) \mapsto (BR, R^{-1}A)$ and optimize over the full group $GL(r)$ rather than one scalar per adapter.

Theorem 6 (Exact matrix-gauge envelope). *Let $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ have a nonsingular joint Gaussian posterior Q, and let $P = \mathcal{N}(0, p^2 I_{(m+n)r})$. Set*

$$S_B = \mathbb{E}_Q[B^\top B], \qquad S_A = \mathbb{E}_Q[AA^\top], \qquad \Delta = n - m.$$

Both moment matrices are positive definite. For $R \in GL(r)$, let $Q_R = (T_R)_{\#}Q$ with $T_R(B, A) = (BR, R^{-1}A)$ and write $\Gamma = RR^\top$. Then

$$\mathrm{KL}(Q_R \| P) = C_Q + \frac{\mathrm{tr}(S_B\Gamma) + \mathrm{tr}(S_A\Gamma^{-1})}{2p^2} + \frac{\Delta}{2}\log\det\Gamma, \tag{11}$$

where C_Q is independent of R. The unique minimizing $\Gamma \succ 0$ is

$$C = S_B^{1/2} S_A S_B^{1/2}, \qquad V_* = \frac{(p^4\Delta^2 I + 4C)^{1/2} - p^2\Delta I}{2},$$
$$\Gamma_* = S_B^{-1/2} V_* S_B^{-1/2}. \tag{12}$$

Every $R = \Gamma_*^{1/2} O$, with O orthogonal, is optimal and induces exactly the same distribution of BA as Q.

The change of variables contributes $(n - m)\log|\det R|$ to KL; omitting this term gives the wrong optimizer for rectangular factors. Stationarity in Γ yields the matrix equation

$$\Gamma S_B \Gamma + p^2\Delta\Gamma - S_A = 0. \tag{13}$$

After congruence by $S_B^{1/2}$, it becomes $V^2 + p^2\Delta V = C$. The positive root in each eigendirection gives (12). Coercivity and uniqueness of this positive solution establish global optimality, without assuming Euclidean convexity of the objective. Appendix D gives the full proof and a stable implementation.

Balancing leaves posterior risks, center predictions, and every sampled product unchanged; only the KL decreases. Because Theorem 1 holds simultaneously over posteriors, optimizing R costs no extra confidence. It does not, however, repair an unstable posterior. It contains scalar balancing as $\Gamma = c^2 I$ and applies layer by layer under a factorized prior.

4.1 INDEPENDENT NONZERO PRIOR CENTERS

A prior centered at an independent initialization, or at factors learned only on an independent sample S_0, breaks the rotational symmetry used in Theorem 6. Conditioning on all prior-side data and randomness, write $M_B = \mathbb{E}_Q B$, $M_A = \mathbb{E}_Q A$, and let $P_0 = \mathcal{N}((B_0, A_0), p^2 I)$ be the fixed prior. A direct change of variables gives

$$\mathrm{KL}(Q_R \| P_0) = C_Q + (n-m)\log|\det R| + \frac{\mathbb{E}_Q \|BR - B_0\|_F^2 + \mathbb{E}_Q \left\|R^{-1}A - A_0\right\|_F^2}{2p^2}, \tag{14}$$

where $C_Q = d\log(2\pi p^2)/2 - h(Q)$ and $d = (m+n)r$. The cross terms depend on R itself, not only on $RR^\top$, so the closed form of Theorem 6 is in general not the noncentral optimizer. In a split-trained prior both B_0 and A_0 may be nonzero, and resetting B_0 to zero after prior training would change the prior.

Proposition 7 (Exact elimination of the orthogonal factor)**.** *Set $R = \Gamma^{1/2}O$, $\Gamma \succ 0$, $O^\top O = I$, and*

$$F(\Gamma) = \Gamma^{1/2} M_B^\top B_0 + \Gamma^{-1/2} M_A A_0^\top. \tag{15}$$

With $C_{Q,0} = C_Q + (\|B_0\|_F^2 + \|A_0\|_F^2)/(2p^2)$,

$$\min_{O^\top O = I} \mathrm{KL}(Q_{\Gamma^{1/2}O} \| P_0) = C_{Q,0} + \frac{\mathrm{tr}(S_B\Gamma) + \mathrm{tr}(S_A\Gamma^{-1})}{2p^2} - \frac{\|F(\Gamma)\|_*}{p^2} + \frac{n-m}{2}\log\det\Gamma. \tag{16}$$

Here $\|\cdot\|_*$ is the nuclear norm. If $F(\Gamma) = L\Lambda V^\top$ is a full singular value decomposition, $O = LV^\top$ attains the minimum. A global minimum over $\Gamma \succ 0$ exists, but neither uniqueness nor a closed form is asserted.

The proof expands (14) and maximizes $\mathrm{tr}(O^\top F)$ over orthogonal O (Appendix E). In practice we compare numerical candidates with $R = I$ through the exact KL (14); any attained reduction is valid without a global optimum. A better prior center reduces the displacement, but it does not by itself make prediction-preserving widths affordable.

5 USING THE FRONTIER WITH AN INDEPENDENT PRIOR

An independent prior split changes which displacement is charged. Let S_0 be independent of S, fit a prior center μ_0 using only S_0, and fix the base model, tokenizer, head, adapter layout, and prior covariance before accessing S. Continue learning only the adapter factors on S. A head fitted on S_0 and then frozen is part of the hypothesis; a head fitted on S must be charged even if it is frozen afterwards.

For independent factor blocks with posterior widths q_b and fixed prior widths p_b, evaluate the full Gaussian KL in Equation (19). The displacement is $D = \|\mu - \mu_0\|^2$, measured in the declared factor coordinates. The identity $\mathrm{KL} = D/(2p^2)$ applies only when every posterior variance matches the common prior variance. Narrower widths add covariance-mismatch cost.

5.1 SCREENING BEFORE SAMPLING, DECIDING AFTER IT

The workflow has three steps. First, once the displacement is known, compute $q_{\mathrm{cert}}(\tau)$ and drop the excluded widths; this needs no posterior draws and applies to the raw isotropic family. Second, select candidate widths on pilot draws and freeze them. Third, compute U on an independent final stream and compare the actual KL, after any balancing, with the budget

$$J_\tau(U) = N\,\mathrm{kl}(U\|\tau) - a_k, \qquad U \le \tau. \tag{17}$$

A candidate passes exactly when its actual KL is at most $J_\tau(U)$. If $U > \tau$ or $J_\tau(U) < 0$, risk or confidence already blocks the certificate. If only the balanced KL fits within the budget, balancing

Table 1: Small-pool audit and accounting floors. The last column is the zero-KL floor with Chernoff accounting at the favorable $N = 1024$, computed conservatively from rounded summaries; it is a lower bound, not an issued certificate.

Draws	Configurations	Mean KL	Reported certificate	Accounting floor
8	567	12.49	1	0.9320
64	216	12.67	0.736–0.928	0.7298

changes the decision without changing any prediction. Appendix C.6 inverts the budget into sample-size and draw requirements.

These outcomes are decisions rather than learning-curve forecasts. A prospective study should therefore fix its rules before inspecting the final draws and report excluded, unexcluded-but-failing, and passing cases separately.

5.2 A MATCHED DESCRIPTION-LENGTH REFERENCE

A compression comparison concerns the decoded predictor, not the Gaussian stochastic predictor. Encode the balanced factor displacement relative to the same S_0-trained side information using a complete prefix-free code of length L bits, including quantization scales, tensor layout, and every reconstruction choice. Evaluating its actual bounded empirical error $\widehat{r}_{\mathrm{dec}}$ gives the standard description-prior comparison

$$C_{\mathrm{code}} = \mathrm{kl}_{+}^{-1}\left(\widehat{r}_{\mathrm{dec}}, \frac{L\log 2 + \log(2\sqrt{N}/\delta_{\mathrm{code}})}{N}\right). \tag{18}$$

It follows by giving the codeword prior mass at least 2^{-L} in Theorem 1. This fixed-code reference follows compression bounds (Lotfi et al., 2024a) but does not reproduce SubLoRA's training or loss, and decoded accuracy should be reported beside the code length. Appendix H specifies the code and its confidence allocation.

6 EXPERIMENTS

We report three measured studies: a small-pool RoBERTa audit with a conservative reanalysis of its summaries (Section 6.1), a controlled test of Monte Carlo slack (Section 6.2), and a controlled test of matrix balancing (Section 6.3). Appendix H works through numerical split-prior scenarios.

6.1 WHY THE SMALL-POOL POSTERIOR CANNOT BE REPAIRED BY ACCOUNTING

The RoBERTa-base audit uses seven GLUE tasks, requested pools of 256, 512, or 1024 examples, and three seeds. Each of its 63 problems contains nine rank–decay configurations, giving 567 fitted checkpoints. Query/value adapters and the classification head are all trained on the pool, so all are charged. The prior is origin-centered with $p = 1/\sqrt{\lambda}$ for LoRA weight decay λ. Because the historical seven-width search was not fully recorded, we treat this audit as a descriptive study rather than a reproducible validation. Appendix F gives full settings and selector results.

With eight posterior draws, the Hoeffding half-width of 0.6065 saturates every risk upper bound at one, although the complete KL is only 2.34–45.91 nats. A matched rerun with 64 draws on 24 problems (216 configurations) gives bounds of 0.736–0.928, but ranking by them selects centers with accuracy 0.7388, below the 0.7655 of training-risk selection. Across all 63 problems, the centers chosen by the center-risk proxy, training risk, and retrained validation reach 0.7053, 0.7066, and 0.7088, while their stochastic predictors reach only 0.511–0.524 (Appendix Figure 5). Accuracy and certification therefore concern different predictors: the deterministic center and its stochastic posterior.

Even with zero KL and the tighter Chernoff bound, the recorded summaries imply certificates of at least 0.9320 and 0.7298 (Table 1; Figure 2a), so no balancing can rescue these posteriors at the recorded budgets. The floor does not rule out narrower posteriors, other prior centers, or larger

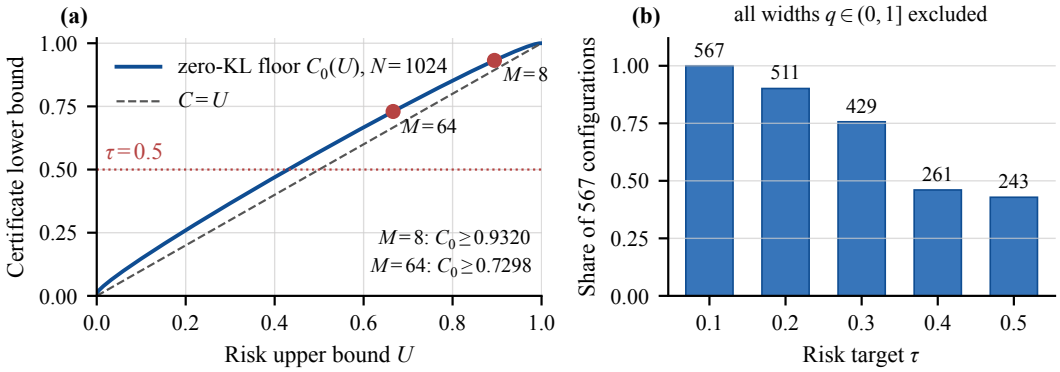


Figure 2: Accounting cannot repair the small-pool posteriors. (a) Zero-KL certificate floor $C_0(U)$ at the most favorable $N = 1024$; the markers are the Chernoff risk floors implied by the recorded 8- and 64-draw summaries, and both floors lie above $\tau = 0.5$. (b) Share of the 567 configurations for which the raw isotropic threshold excludes every width $q \in (0, 1]$, before any posterior sampling; bar labels are counts.

Table 2: Controlled Gaussian-factor classification: mean ± sample SD of the risk certificate across 20 independent datasets. The same posterior draws are used for both inequalities. The prior, trained center, KL, and confidence allocation are held fixed within each row.

Posterior width q	Draws M	Hoeffding certificate	Chernoff certificate
0.1	8	0.777 ± 0.005	0.721 ± 0.008
0.1	64	0.326 ± 0.003	0.215 ± 0.007
0.1	256	0.189 ± 0.003	0.104 ± 0.006
0.1	1024	0.117 ± 0.003	0.064 ± 0.005
4	8	0.999 ± 0.005	0.984 ± 0.021
4	64	0.804 ± 0.038	0.792 ± 0.033
4	256	0.684 ± 0.016	0.683 ± 0.016
4	1024	0.627 ± 0.009	0.627 ± 0.009

samples; Appendix E.4 gives the arithmetic. Monte Carlo draw budgets lead to the same conclusion. Inverting the Hoeffding budget on the 216 recorded 64-draw configurations, no draw budget suffices for any configuration at $\tau = 0.5$; at $\tau = 0.7$ none suffices for 28 of them and the median requirement is 343 draws; only at $\tau = 0.9$ do 64 draws suffice, for 201 of 216 (Appendix Figure 6).

The width screen reaches a similar conclusion before any sampling. The raw origin-centered isotropic threshold excludes every $q \in (0, 1]$ on $243, 261, 429, 511, 567$ configurations at targets $0.5, 0.4, 0.3, 0.2, 0.1$, respectively (Figure 2b). A nine-width risk sweep cannot sharpen this screen: its simultaneous half-width of 0.913 makes every risk lower bound zero.

6.2 Controlled separation of posterior risk and sampling uncertainty

To isolate the Monte Carlo mechanism, we evaluate a rank-one Gaussian-factor classifier on a two-dimensional Gaussian classification problem. The center is fitted from $N = 2048$ examples, the prior is fixed, and the exact empirical stochastic risk is available analytically. Across 20 independent datasets, a stable posterior with width $q = 0.1$ has exact empirical risk 0.0141 ± 0.0025, while $q = 4$ gives 0.4756 ± 0.0002 (mean ± sample standard deviation). All nine widths, four fixed nested draw budgets, and two inequalities are covered by $K_{\mathrm{MC}} = 72$.

At $q = 0.1$ and $M = 1024$, Chernoff certifies risk below 0.1 on $20/20$ datasets, compared with $0/20$ for Hoeffding. The unstable posterior remains far above that target under either inequality (Table 2; Figure 3). Saturation at eight draws therefore reflects insufficient sampling rather than an impossibility for the posterior family. This positive control validates the implementation on a tractable problem; it does not show that more draws alone would certify the language-model posteriors.

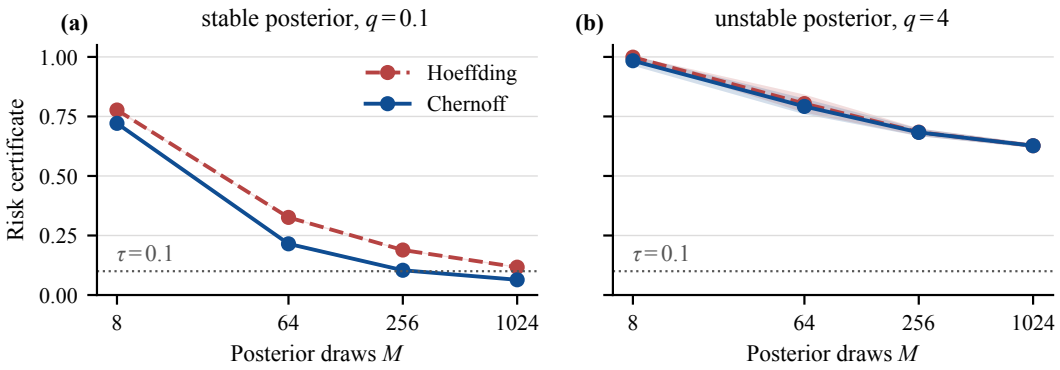


Figure 3: Posterior behavior and Monte Carlo slack have different effects. (a) The stable posterior ($q = 0.1$) benefits from more draws and from the Chernoff inequality, which crosses $\tau = 0.1$ at $M = 1024$. (b) The unstable posterior ($q = 4$) stays far above the target under both inequalities. Lines are means and bands ± 1 sample SD over 20 datasets (Table 2); both inequalities use the same posterior draws and simultaneous confidence family.

Table 3: Gaussian factor experiment, $m = 12, n = 8$. KL is in nats; entries are mean $\pm$ sample SD over 20 seeds. The last column is the mean per-case percentage reduction relative to scalar balancing, not a ratio of pooled means. All three columns describe the same stochastic product.

Rank	Raw KL	Scalar optimum	Matrix optimum	Extra reduction (%)
1	61.75 $\pm$ 16.60	39.04 $\pm$ 1.26	39.04 $\pm$ 1.26	0.0 $\pm$ 0.0
2	120.40 $\pm$ 22.29	117.59 $\pm$ 19.87	68.35 $\pm$ 1.31	40.6 $\pm$ 8.4
4	190.76 $\pm$ 18.60	186.95 $\pm$ 16.82	131.21 $\pm$ 1.55	29.3 $\pm$ 5.6
8	332.50 $\pm$ 15.37	329.86 $\pm$ 15.22	256.67 $\pm$ 2.33	22.1 $\pm$ 3.1

6.3 Controlled prediction-preserving matrix balancing

This experiment uses synthetic anisotropic factor posteriors. We generate Gaussian factor posteriors for three rectangular or square shapes, four ranks, and 20 seeds, yielding 240 cases. For each posterior, we compare its raw KL, the exact scalar optimum, and Theorem 6; all use the same fixed prior and stochastic product distribution. At $m = 12, n = 8, r = 4$, full balancing reduces KL from the scalar optimum 186.95 ± 16.82 to 131.21 ± 1.55 nats (Table 3; Figure 4). Sampled predictions are unchanged by construction, so the entire reduction occurs in the complexity term.

Across all shapes and ranks, the largest relative stationarity residual is 1.06×10^{-13}, the largest relative disagreement with independent numerical minimization is 9.91×10^{-16}, and paired sampled products differ by at most 2.71×10^{-15}. The scalar and matrix optima coincide at rank one, as required. These measurements validate the closed form. Because the deformation is deliberate, these percentages show that balancing can reduce KL substantially; they do not estimate the reduction for trained LoRA factors. Its effect on a given posterior depends on whether that posterior's zero-KL floor lies below the target, which is not the case in the small-pool audit (Section 6.1).

7 Related work

Post-hoc uncertainty and risk certification have different objectives. Laplace-LoRA, Training-Free Bayesianization, and BLoB build low-rank Gaussian posteriors for calibration (Yang et al., 2024; Shi et al., 2025; Wang et al., 2024); we ask whether such a posterior also affords a bounded-loss certificate, without substituting calibration quality for empirical Gibbs risk.

Nonvacuous neural-network bounds optimize posterior risk and complexity together (Dziugaite & Roy, 2017; Pitas, 2020). Prior-data allocation and curvature affect that tradeoff (Dziugaite et al., 2021; Yang et al., 2022); localized priors, PAC-tuning, and randomized LoRA modify training or the certificate construction (Dong, 2025; Liu et al., 2023; Lei et al., 2024). Our independent split-trained

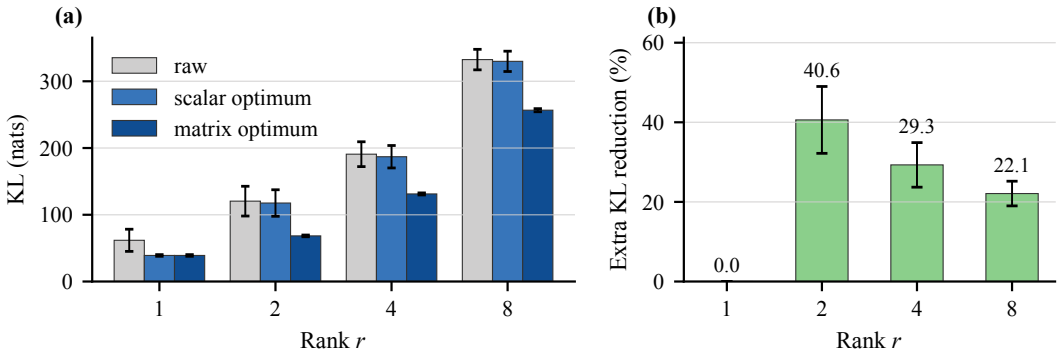


Figure 4: Prediction-preserving balancing on deformed Gaussian factors ($m = 12$, $n = 8$; 20 seeds, error bars ± 1 sample SD). (a) Raw, scalar-balanced, and matrix-balanced KL describe the same stochastic product. (b) Mean per-case KL reduction of the matrix optimum beyond scalar balancing; the two optima coincide at rank one.

prior is not the same-sample localized objective. Compression-based language-model bounds (Lotfi et al., 2024a) provide a description-length reference but certify a different decoded predictor and loss construction. The matrix calculation specializes the established role of rescaling invariance (Rouchouse et al., 2025); it does not introduce invariance as a new principle. Appendix B reviews further work on PAC-Bayes certificates, Bayesian LoRA, low-rank adaptation theory, parameter symmetries, and derandomization.

8 Limitations

Our language-model evidence is a small-pool audit (RoBERTa-base, seven GLUE tasks, pools of at most 1024 examples, an origin-centered prior) whose historical width search was not fully recorded. Its zero-KL floors use the recorded 8- and 64-draw summaries at the favorable $N = 1024$; more draws, narrower posteriors, or other prior centers can move them. The width threshold covers only raw isotropic posteriors, the balancing gains come from deliberately deformed synthetic factors, and the noncentral objective has no global-optimality guarantee. The split-prior study and the description-length comparison are specified with numerical scenarios but not yet measured on large tasks.

9 Discussion and Conclusion

We study when a Gaussian distribution around a trained low-rank adapter supports a useful PAC-Bayes certificate, and answer with three tools that locate a posterior on the risk–complexity frontier (Figure 1): a prior-scale-free width threshold that excludes isotropic widths before sampling, a zero-KL floor that decides whether any KL reduction can change the outcome at a measured risk bound, and closed-form matrix balancing over $GL(r)$ that leaves every sampled product unchanged, with a noncentral objective for independently trained priors. On the small-pool audit, the recorded summaries alone show that the certificate is limited by risk and confidence rather than complexity: even at zero KL it stays at or above 0.7298, so improvement requires a lower risk bound or a larger certification sample rather than further accounting. The controlled studies show that the other levers can be effective when their conditions hold: with 1024 draws, Chernoff accounting certifies a stable posterior below 0.1 on all 20 datasets, and balancing reduces KL by a further 29.3% on synthetic rank-four factors, a closed-form instance of the rescaling invariance studied by Rouchouse et al. (2025). In practice, a certification pipeline can compute the threshold and floor before sampling, balance only when the floor lies below the target, and certify the stochastic predictor itself. Whether these tools yield nonvacuous certificates for trained adapters remains to be measured on large tasks with independently trained priors.

REPRODUCIBILITY STATEMENT

The appendix contains complete proofs, the measured tables, and the specified inputs of every scenario. The accompanying scenario file records assumed losses and complexities separately from their formula-derived certificates, thresholds, margins, and workload counts.

AI USE STATEMENT

Generative AI assisted with literature discovery, mathematical exposition, code preparation, numerical analysis, and language editing. The authors remain responsible for the claims, experimental interpretation, and final submission.

ETHICS STATEMENT

The study uses established language-understanding benchmarks and synthetic numerical examples. It collects no new human-subject data. A bound for a specified bounded loss is not a general safety guarantee.

REFERENCES

Md Sakir Ahmed, Kumaresh Sarmah, and Hemen Dutta. Fisher-geometric sharpness and the implicit bias of SGD toward flat minima. *arXiv preprint arXiv:2606.20469*, 2026.

Nicola Aladrah and Fabio Anselmi. PAC–Bayes bounds on quotient parameter spaces: Geometry-induced implicit-bias priors. *arXiv preprint arXiv:2607.18422*, 2026.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

J. Arunan. Tight sample complexity for low-rank adaptation: Matching bounds and rank selection. *arXiv preprint arXiv:2607.27680*, 2026.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. doi: 10.1007/978-3-319-00227-9.

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. In *ICLR Workshop on Quantify Uncertainty and Hallucination in Foundation Models*, 2025.

Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mathieu Bazinet, Valentina Zantedeschi, and Pascal Germain. Bound to disagree: Generalization bounds via certifiable surrogates. In *Conference on Uncertainty in Artificial Intelligence*, 2026.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Felix Biggs and Benjamin Guedj. On margins and derandomisation in PAC-Bayes. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3709–3731, 2022.

Sergey G Bobkov and Michel Ledoux. Weighted poincaré-type inequalities for cauchy and other convex measures. *Annals of Probability*, 2009.

Herm Jan Brascamp and Elliott H. Lieb. On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976. doi: 10.1016/0022-1236(76)90004-5.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007.

Patrick Cattiaux and Arnaud Guillin. Semi log-concave markov diffusions. *Séminaire de Probabilités*, 2014.

Bai Cong, Nico Daheim, Yuesong Shen, Rio Yokota, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Improving LoRA with variational learning. *arXiv preprint arXiv:2506.14280*, 2025.

Dario Coscia, Sindy Löwe, and Max Welling. BaLoRA: Bayesian low-rank adaptation of large scale models. *arXiv preprint arXiv:2605.08110*, 2026.

Wen Dong. Certifying deep network risks and individual predictions with PAC-Bayes loss via localized priors. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes bounds. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 604–612, 2021.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, 2021.

Diego García-Pérez, Emilio Parrado-Hernández, and John Shawe-Taylor. Some theoretical improvements on the tightness of PAC-Bayes risk certificates for neural networks. *arXiv preprint arXiv:2510.07935*, 2025.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998. doi: 10.1214/ss/1028905934.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806, 2024.

Vasant G. Honavar, Satish Kumar Keshri, Neil Ashtekar, and Zehao Liu. PAC-Bayes beyond parameter space: Behavioral equivalence, Z-information, and exact complexity decomposition. *arXiv preprint arXiv:2608.11465*, 2026. doi: 10.48550/arXiv.2608.11465.

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.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21306–21328, 2024.

Taehoon Kim, Henry Gouk, Minyoung Kim, and Timothy Hospedales. Model merging is secretly certifiable: Non-vacuous generalisation bounds for low-shot learning. In *Conference on Uncertainty in Artificial Intelligence*, 2026.

Anastasis Kratsios, Tin Sum Cheng, Aurelien Lucchi, and Haitz Sáez de Ocáriz Borde. Sharp generalization bounds for foundation models with asymmetric randomized low-rank adapters. *arXiv preprint arXiv:2506.14530*, 2025.

Ismail Labiad, Mathurin Videau, Matthieu Kowalski, Marc Schoenauer, Alessandro Leite, Julia Kempe, and Olivier Teytaud. Tuning without peeking: Provable generalization bounds and robust LLM post-training. *arXiv preprint arXiv:2507.01752*, 2025.

Tom A. Lamb, Philip Torr, and Tim G. J. Rudner. Tight PAC-Bayes generalisation guarantees for large language model safety monitoring. In *ICML Workshop on Foundations of Generalization in Modern Machine Learning (FoGen)*, 2026.

John Langford and Rich Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Benjamin Leblanc, Mathieu Bazinet, Nathaniel D'Amours, Alexandre Drouin, and Pascal Germain. Generalization bounds via meta-learned model representations: PAC-Bayes and sample compression hypernetworks. In *International Conference on Machine Learning*, 2025.

Geneviève Lefebvre, Russell Steele, and Alain C. Vandal. A path sampling identity for computing the Kullback–Leibler and J divergences. *Computational Statistics & Data Analysis*, 54(7):1719–1731, 2010. doi: 10.1016/j.csda.2010.01.018.

Zijian Lei, Dong Qian, and William K. Cheung. Fast randomized low-rank adaptation of pre-trained language models with PAC regularization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5236–5249, 2024. doi: 10.18653/v1/2024.findings-acl.310.

Alexandre Lemire Paquin, Brahim Chaib-Draa, and Philippe Giguère. Smoothness-based derandomization of PAC-Bayes bounds. *arXiv preprint arXiv:2606.19105*, 2026. doi: 10.48550/arXiv.2606.19105.

Zhuoyun Li, Boxuan Wang, Jinwei Hu, Xiaowei Huang, and Yi Dong. FragileFlow: Spectral control of correct-but-fragile predictions for foundation model robustness. *arXiv preprint arXiv:2605.08896*, 2026. doi: 10.48550/arXiv.2605.08896.

Moule Lin, Shuhao Guan, Andrea Patane, David Gregg, and Goetz Botterweck. Bayesian-LoRA: Probabilistic low-rank adaptation of large language models. *arXiv preprint arXiv:2601.21003*, 2026.

Guangliang Liu, Zhiyu Xue, Xitong Zhang, Kristen Johnson, and Rongrong Wang. PAC-tuning: Fine-tuning pre-trained language models with PAC-driven perturbed gradient descent. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12178–12189, 2023. doi: 10.18653/v1/2023.emnlp-main.748.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuhang Liu, Tao Li, Zhehao Huang, Zuopeng Yang, and Xiaolin Huang. Bi-LoRA: Efficient sharpness-aware minimization for fine-tuning large-scale models. In *International Conference on Learning Representations*, 2026.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818, 2024a.

Sanae Lotfi, Yilun Kuang, Brandon Amos, Micah Goldblum, Marc Finzi, and Andrew Gordon Wilson. Unlocking tokens as data points for generalization bounds on larger language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

David Madras, Joshua Safyan, and Qiuyi Zhang. Prompts generalize with low data: Non-vacuous generalization bounds for optimizing prompts with more informative priors. In *ICML Workshop on Exploration in AI Today (EXAIT)*, 2025.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23610–23641, 2023.

Patryk Marszałek, Klaudia Bałazy, Jacek Tabor, and Tomasz Kuśmierczyk. Minimal ranks, maximum confidence: Parameter-efficient uncertainty quantification for LoRA. *arXiv preprint arXiv:2502.12122*, 2025.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999. doi: 10.1145/307400.307435.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Emre Onal, Klemens Flöge, Emma Caldwell, Arsen Sheverdin, and Vincent Fortuin. Gaussian stochastic weight averaging for Bayesian low-rank adaptation of large language models. *arXiv preprint arXiv:2405.03425*, 2024.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 2000.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

Antoine Picard-Weibel, Eugenio Clerico, Roman Moscoviz, and Benjamin Guedj. How good is PAC-Bayes at explaining generalisation? *arXiv preprint arXiv:2503.08231*, 2025.

Konstantinos Pitas. Dissecting non-vacuous generalization bounds based on the mean-field approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7739–7749. PMLR, 2020.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 30th Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703, 2017.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 2023.

Damien Rouchouse, Antoine Gonon, Rémi Gribonval, and Benjamin Guedj. Non-vacuous generalization bounds: Can rescaling invariances help? *arXiv preprint arXiv:2509.26149*, 2025.

Colin Samplawski, Adam D. Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. Scalable Bayesian low-rank adaptation of large language models via stochastic variational subspace inference. In *Conference on Uncertainty in Artificial Intelligence*, 2025.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Haizhou Shi, Yibin Wang, Ligong Han, Huan Zhang, and Hao Wang. Training-free Bayesianization for low-rank adapters of large language models. In *Advances in Neural Information Processing Systems*, 2025.

Zhengkai Sun, Dibyakanti Kumar, Alejandro F. Frangi, Anirbit Mukherjee, and Mingfei Sun. Convergent stochastic training of multi-headed attention and understanding LoRA. *arXiv preprint arXiv:2605.07959*, 2026.

Khoat Than and Dat Phan. Non-vacuous generalization bounds for deep neural networks without any modification to the trained models. *arXiv preprint arXiv:2503.07325*, 2025.

Mame Diarra Toure and David A. Stephens. When does a low-rank bayesian neural network certify its deterministic center? In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2026a.

Mame Diarra Toure and David A. Stephens. Singular bayesian neural networks. *arXiv preprint arXiv:2602.00387*, 2026b.

Anna Van Elst and Debarghya Ghoshdastidar. Tight PAC-Bayesian risk certificates for contrastive learning. *SIAM Journal on Mathematics of Data Science*, 7(4):1904–1927, 2025.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Leveraging PAC-Bayes theory and Gibbs distributions for generalization bounds with complexity measures. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris N. Metaxas, and Hao Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 67758–67794, 2024.

Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérief-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A frequentist approach to sequential prior updates with no information loss. *arXiv preprint arXiv:2405.14681*, 2024.

Haotian Xiang, Bingcong Li, and Qin 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.

Rubing Yang, Jialin Mao, and Pratik Chaudhari. Does the data induce capacity control in deep learning? In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25166–25197. PMLR, 2022.

Xinping Yi, Gaojie Jin, Xiaowei Huang, and Shi Jin. Towards a unified PAC-Bayesian framework for norm-based generalization bounds. *arXiv preprint arXiv:2601.08100*, 2026. doi: 10.48550/arXiv.2601.08100.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2019.

Yuxuan Zhu, Rohan Alur, and Daniel Kang. Non-vacuous generalization bounds for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2607.14506*, 2026.

A NOTATION

Table 4 collects the symbols used in the main text, grouped by topic. Symbols used only inside a single proof are defined where they appear.

Table 4: Notation.

Symbol	Meaning
Data, predictor, and certificate	
S, N, $\mathcal{D}$	certification sample, its size, and the data distribution
$\ell \in [0,1]$	bounded loss
$k \in \mathcal{K}$, w_k	configuration and its prior weight, $\sum_k w_k = 1$
$W_0^{(\ell)}$, $B^{(\ell)}$, $A^{(\ell)}$, s_ℓ	frozen matrix, LoRA factors, and scaling of adapted layer ℓ
r, m, n	adapter rank and factor dimensions, $B \in \mathbb{R}^{m\times r}$, $A \in \mathbb{R}^{r\times n}$
$z \in \mathbb{R}^{d_k}$	all trained coordinates (adapters, head, and other trained blocks)
P, Q (P_k, Q_k)	prior and posterior over z (for configuration k)
μ	posterior mean, the trained checkpoint (the *center*)
$R_{\mathcal{D}}(Q)$, $\widehat{R}_S(Q)$	population and empirical risk of the stochastic predictor
kl, $\mathrm{kl}_+^{-1}(u,c)$	binary KL and its upper inversion
M, z_j, Y_j, $\widetilde{r}$	posterior draws, a draw, its empirical loss, and their mean
K_{MC}	number of posteriors covered by one Monte Carlo family
U_H, U_{kl}, U	Hoeffding, Chernoff, and generic upper bounds on $\widehat{R}_S(Q)$
δ_{PB}, η	PAC-Bayes and Monte Carlo failure probabilities
C_k	certificate of configuration k
Isotropic feasibility	
q, p, p_*	posterior width, prior width, and KL-optimal prior width
m, D, d	prior center, displacement $D = \|\mu - m\|_2^2$, and dimension
a_k	confidence cost $\log(1/w_k) + \log(2\sqrt{N}/\delta_{\mathrm{PB}})$
τ	target risk
b_τ, $q_{\mathrm{cert}}(\tau)$	available complexity at zero risk and the width threshold
$C_{0,k}(U)$	zero-KL (complexity-free) certificate floor
Matrix balancing	
$R \in GL(r)$, T_R, Q_R	gauge matrix, map $(B,A) \mapsto (BR, R^{-1}A)$, and pushforward of Q
$\Gamma = RR^\top$, Γ_*	positive-definite gauge variable and its optimum
S_B, S_A	second moments $\mathbb{E}_Q[B^\top B]$ and $\mathbb{E}_Q[AA^\top]$
Δ	dimension gap $n - m$
C, V_*	$S_B^{1/2} S_A S_B^{1/2}$ and the root defining Γ_*
C_Q	R-independent constant of the KL, with entropy $h(Q)$
S_0, μ_0	independent prior sample and prior center fitted on it
B_0, A_0, P_0	prior factor centers and the noncentral prior
M_B, M_A, $F(\Gamma)$, $C_{Q,0}$	posterior factor means, orientation matrix, and noncentral constant
Workflow and experiments	
$J_\tau(U)$	KL budget $N\,\mathrm{kl}(U\|\tau) - a_k$
L, $\widehat{r}_{\mathrm{dec}}$, δ_{code}, C_{code}	code length in bits, decoded error, its failure probability, and certificate
λ	LoRA weight decay; the audit prior has $p = 1/\sqrt{\lambda}$

B EXTENDED RELATED WORK

PAC-Bayes foundations. The certificate of Theorem 1 belongs to the PAC-Bayesian line, in which the empirical term averages the loss over the posterior, complexity is measured against a data-independent prior, and the guarantee holds simultaneously over posteriors (McAllester, 1999; Langford & Caruana, 2001; Seeger, 2002; Maurer, 2004; Catoni, 2007; Alquier, 2024). Data-dependent priors require a separate argument, such as differential privacy or an independent data split (Dziugaite & Roy, 2018; Dziugaite et al., 2021).

Tighter and nonvacuous certificates. Following the first nonvacuous bounds for stochastic networks (Dziugaite & Roy, 2017), later work tightens the risk–complexity pair through training

objectives (Pérez-Ortiz et al., 2021), compression (Zhou et al., 2019; Lotfi et al., 2022), recursive or meta-learned priors (Wu et al., 2024; Leblanc et al., 2025), sharper kl inversions (García-Pérez et al., 2025), tail-robust and complexity-measure bounds (Rodríguez-Gálvez et al., 2023; Viallard et al., 2024), certifiable surrogates (Bazinet et al., 2026), and task-specific certificates (Van Elst & Ghoshdastidar, 2025). Bounds for unmodified trained networks (Than & Phan, 2025) and an assessment of how well PAC-Bayes explains generalization (Picard-Weibel et al., 2025) are closest in spirit to our post-hoc setting.

Certificates for language models and post-training. Compression bounds for language models trained from scratch (Lotfi et al., 2024a) extend to larger models by treating tokens as data points (Lotfi et al., 2024b), and quantized adapters yield certificates for language-model safety monitors (Lamb et al., 2026). Black-box post-training, model merging, prompt optimization, and reinforcement learning with verifiable rewards also admit nonvacuous bounds (Labiad et al., 2025; Kim et al., 2026; Madras et al., 2025; Zhu et al., 2026). These methods train or encode for the bound, whereas we evaluate a Gaussian attached after ordinary fine-tuning.

Bayesian and uncertainty-aware LoRA. Beyond the methods discussed in Section 7, variational, subspace, weight-averaging, ensemble, and orthogonalized constructions target uncertainty and calibration (Lin et al., 2026; Cong et al., 2025; Marszałek et al., 2025; Xiang et al., 2026; Onal et al., 2024; Samplawski et al., 2025; Coscia et al., 2026; Balabanov & Linander, 2025). Calibration gains do not by themselves imply PAC-Bayes validity, although noise-stable posteriors may improve the empirical-risk side of the frontier.

Theory of low-rank adaptation. Kernel, neural-tangent, expressivity, convergence, and sample-complexity analyses ask whether low-rank adaptation can represent, reach, or learn a solution (Malladi et al., 2023; Jang et al., 2024; Zeng & Lee, 2024; Sun et al., 2026; Arunan, 2026), and randomized asymmetric adapters with one frozen factor admit sharp generalization bounds (Kratsios et al., 2025). Studies of learning rates, forgetting, and sharpness-aware training show that rank alone does not determine adaptation behavior (Hayou et al., 2024; Biderman et al., 2024; Liu et al., 2026). None of these analyses certifies a declared stochastic posterior.

Symmetry, curvature, and posterior geometry. Rescaling invariances and continuous symmetries make weight-space KL depend on the parameterization, and recent work moves the analysis to quotient or behavioral spaces (Rouchouse et al., 2025; Aladrah & Anselmi, 2026; Honavar et al., 2026). Low-rank Bayesian networks induce singular posteriors on the rank-r product manifold (Toure & Stephens, 2026b), and curvature-, norm-, and sharpness-based analyses relate posterior or loss geometry to generalization (Pitas, 2020; Yang et al., 2022; Yi et al., 2026; Ahmed et al., 2026). Our matrix balancing resolves the $GL(r)$ symmetry of Gaussian factor posteriors in closed form under a fixed isotropic prior.

From stochastic to deterministic predictors. Transferring a stochastic bound to a deterministic checkpoint requires margin, smoothness, or stability arguments (Bartlett et al., 2017; Neyshabur et al., 2018; Biggs & Guedj, 2022; Lemire Paquin et al., 2026; Li et al., 2026); for low-rank Bayesian networks, balanced factors resolve the gauge ambiguity that such a transfer faces (Toure & Stephens, 2026a). We certify and deploy the stochastic predictor and therefore require no such transfer.

Gibbs posteriors and Langevin dynamics. An alternative to the Gaussian route certifies the Gibbs posterior targeted by regularized stochastic training. Its KL admits a path-sampling identity (Gelman & Meng, 1998; Lefebvre et al., 2010) that functional inequalities can bound (Brascamp & Lieb, 1976; Otto & Villani, 2000; Bobkov & Ledoux, 2009; Bakry et al., 2014; Cattiaux & Guillin, 2014), and Langevin convergence results relate such posteriors to stochastic training (Raginsky et al., 2017; Vempala & Wibisono, 2019; Erdogdu & Hosseinzadeh, 2021). We do not pursue this route, because a numerically useful certificate would require these constants to be controlled along the entire temperature path.

C CERTIFICATE PROOFS AND BOUNDARY CASES

C.1 GAUSSIAN KL AND CONFIGURATION SELECTION

For independent Gaussian blocks $Q_b = \mathcal{N}(\mu_b, q_b^2 I_{d_b})$ and $P_b = \mathcal{N}(m_b, p_b^2 I_{d_b})$, direct evaluation of the log-density ratio gives

$$\mathrm{KL}(Q\|P) = \frac{1}{2}\sum_b \left[\frac{\|\mu_b - m_b\|_2^2}{p_b^2} + d_b\left(\frac{q_b^2}{p_b^2} - 1 - \log\frac{q_b^2}{p_b^2}\right)\right]. \tag{19}$$

Every block contributing to the trained predictor must appear. The theorem does not require its optimizer's regularizer to match the prior, so a head trained without weight decay can still be charged to a fixed Gaussian prior. It does not, however, acquire a Gibbs or maximum-a-posteriori interpretation from that accounting choice.

On component k of the disjoint union, the density ratio is

$$\frac{d(\delta_k \otimes Q_k)}{dP}(k, z) = \frac{1}{w_k}\frac{dQ_k}{dP_k}(z).$$

Its posterior expectation gives (3). Applying Theorem 1 on the union proves simultaneous validity after posterior and configuration selection. A data-dependent prior requires separate accounting; posterior-side optimization does not turn a data-dependent prior into an independent one.

C.2 MONTE CARLO UPPER BOUNDS

Conditional on S and a frozen Q_k, the variables $Y_{k,j}$ are independent across posterior draws, bounded in $[0, 1]$, and have mean $\widehat{R}_S(Q_k)$. Hoeffding's inequality gives

$$\Pr\{\widehat{R}_S(Q_k) > \widetilde{r}_k + e \mid S, Q_k\} \le e^{-2Me^2}.$$

Taking $e = \sqrt{\log(K_{\mathrm{MC}}/\eta)/(2M)}$ and union-bounding over the fixed family proves (4). Dependence between different candidate streams is allowed; independence within a posterior's draw sequence is required.

For the Chernoff form, convexity on $[0, 1]$ gives $\mathbb{E}e^{tY} \le 1 - r + re^t$ when $\mathbb{E}Y = r$. For $v < r$, exponential Markov with $t < 0$ and minimization over t yield

$$\Pr\{\widetilde{r} \le v\} \le \exp[-M\,\mathrm{kl}(v\|r)].$$

Inverting at level η/K_{MC} proves (5). This argument does not assume that a draw-level average is Bernoulli. Intersecting the Monte Carlo event with the PAC-Bayes event gives total failure probability at most $\delta_{\mathrm{PB}} + \eta$. When several budgets or inequality choices are inspected, their union count must also be included. The controlled experiment pays for both; the original fixed-budget experiment uses its separately stated count.

C.3 PROOF OF THE ISOTROPIC ENVELOPE

For $x = p^2 > 0$, the Gaussian expression is

$$K(x, q) = \frac{1}{2}\left[\frac{D + dq^2}{x} - d + d\log\frac{x}{q^2}\right].$$

Its derivative is $\frac{1}{2}[-(D + dq^2)/x^2 + d/x]$, so the only stationary point is $x_* = q^2 + D/d$. The objective diverges as $x \downarrow 0$ or $x \uparrow \infty$, making this the global minimum. Substitution gives (7).

For fixed $r \le \tau$, binary KL is increasing in its second argument above r. Hence $\mathrm{kl}_+^{-1}(r, c) \le \tau$ is equivalent to $c \le \mathrm{kl}(r\|\tau)$. The Monte Carlo upper bound is no smaller than r on its validity event, and replacing that upper bound by the true empirical posterior risk only relaxes feasibility. Replacing the KL by its oracle minimum relaxes it further, yielding (8). An oracle scale depending on S cannot support a positive claim under the independent-prior theorem.

C.4 Proof of the width threshold

For $0 < r \le \tau$,

$$\frac{\partial}{\partial r}\,\mathrm{kl}(r\|\tau) = \log\frac{r(1-\tau)}{\tau(1-r)} \le 0.$$

Continuity at $r = 0$ yields $\mathrm{kl}(r\|\tau) \le \log(1/(1-\tau))$. Thus any passing width satisfies $\frac{d}{2}\log(1 + D/(dq^2)) \le b_\tau$. For $D > 0$, its left side is strictly positive at every finite width. If $b_\tau \le 0$, none is feasible; otherwise exponentiation gives (9). Monotonicity in q makes exclusion over $(0, q_{\max}]$ exact. If $D = 0$, the envelope is zero and no positive width threshold follows: the necessary condition is $a_k \le N\log(1/(1-\tau))$. The implementation uses `expm1` for the denominator when its argument is small.

C.5 Data and posterior-draw budgets

Fix $U = u < \tau$ and $\kappa = \mathrm{KL}(Q\|P) + \log(1/w_k)$. Writing $b = \mathrm{kl}(u\|\tau) > 0$, inversion gives the exact arithmetic condition

$$C \le \tau \quad\Longleftrightarrow\quad Nb - \tfrac{1}{2}\log N \ge \kappa + \log(2/\delta_{\mathrm{PB}}). \tag{20}$$

The upper root yields a sample-size threshold at fixed (u, κ). It is not a learning-curve prediction because both quantities may change with more training data. No positive-cost certificate passes when $u \ge \tau$.

At fixed N and κ, let $u_\tau \in [0, \tau)$ solve $\mathrm{kl}(u_\tau\|\tau) = (\kappa + \log(2\sqrt{N}/\delta_{\mathrm{PB}}))/N$. A solution exists only if that cost is at most $\mathrm{kl}(0\|\tau)$. Conditional on an observed mean $\widehat{r} < u_\tau$, the Hoeffding implementation passes exactly when

$$M \ge \frac{\log(K_{\mathrm{MC}}/\eta)}{2(u_\tau - \widehat{r})^2}. \tag{21}$$

These are two different resource limitations. Insufficient posterior draws enlarge U; an excessive KL cannot be repaired by Monte Carlo precision. The budgets must be fixed beforehand or covered by a simultaneous or sequentially valid construction.

C.6 Budget inversion proofs

For $u < \tau$, rearranging the certificate condition gives (20). The function $f(N) = Nb - \frac{1}{2}\log N$ decreases to its unique minimum at $1/(2b)$ and increases thereafter. Its upper crossing gives the eventual passing threshold; checking integers directly handles a crossing outside the admissible range $N \ge 2$. This arithmetic calculation conditions on fixed u and κ rather than assuming that learning leaves them unchanged.

The map $u \mapsto \mathrm{kl}(u\|\tau)$ is continuous and decreasing on $[0, \tau]$, from $\mathrm{kl}(0\|\tau)$ to zero. This gives the stated existence and uniqueness of u_τ for positive complexity cost. Substituting U_H and solving $U_H \le u_\tau$ proves (21). At the boundary $u_\tau = 0$, no finite Hoeffding budget passes with a nonnegative observed mean. Repeatedly inspecting the bound and stopping on success is not justified by a fixed-M inequality.

D Matrix and scalar gauge proofs

D.1 Density transformation and global optimum

Write $d = (m + n)r$. Since T_R acts independently on each of the m rows of B and n columns of A, its absolute Jacobian determinant is $|\det R|^{m-n}$. Differential entropy therefore transforms as

$$h(Q_R) = h(Q) + (m - n)\log|\det R|.$$

The Gaussian prior's negative log density is $\frac{d}{2}\log(2\pi p^2) + (\|B\|_F^2 + \|A\|_F^2)/(2p^2)$. Taking expectation under the pushed-forward posterior, and writing $\Gamma = RR^\top$, gives

$$\mathbb{E}\,\|BR\|_F^2 = \mathrm{tr}(S_B\Gamma), \qquad \mathbb{E}\,\left\|R^{-1}A\right\|_F^2 = \mathrm{tr}(S_A\Gamma^{-1}).$$

Combining these expressions proves (11), with $C_Q = \frac{d}{2}\log(2\pi p^2) - h(Q)$. Cross-covariances between B and A do not alter these two second-moment terms or the Jacobian calculation. For the independent isotropic blocks used in the experiment,

$$C_Q = -\frac{d}{2} + \frac{d}{2}\log p^2 - \frac{1}{2}\left(mr\log s_B^2 + nr\log s_A^2\right).$$

Nonsingularity makes S_B, S_A positive definite: for any nonzero v, $\mathbb{E}\,\|Bv\|^2$ and $\mathbb{E}\,\left\|v^\top A\right\|^2$ are strictly positive.

Differentiating over symmetric positive-definite Γ gives

$$\nabla_\Gamma \mathrm{KL} = \frac{1}{2p^2}(S_B - \Gamma^{-1}S_A\Gamma^{-1}) + \frac{\Delta}{2}\Gamma^{-1}.$$

Multiplying stationarity on both sides by Γ gives (13). Set $V = S_B^{1/2}\Gamma S_B^{1/2}$. Then $V^2 + p^2\Delta V = C$. In particular, V commutes with C because C is a polynomial in V. On each common eigendirection, $v > 0$ solves $v^2 + p^2\Delta v - c = 0$, which has exactly one positive root for $c > 0$. Functional calculus gives (12), including repeated eigenvalues.

For completeness, a global optimum exists. If $s_B, s_A > 0$ are lower spectral bounds for the two moment matrices and λ_i are eigenvalues of Γ, the nonconstant objective is bounded below by

$$\sum_{i=1}^{r}\left[\frac{s_B\lambda_i + s_A/\lambda_i}{2p^2} + \frac{\Delta}{2}\log\lambda_i\right].$$

Each summand is bounded below and diverges as its argument tends to zero or infinity. Sublevel sets are therefore compact inside the positive-definite cone. Every minimizer is stationary, and the stationary solution is unique. Any factor R with $RR^\top = \Gamma_*$ attains it, explaining the remaining orthogonal freedom.

Finally, $(BR)(R^{-1}A) = BA$ for every draw, not merely in expectation. A fixed base matrix and LoRA scaling preserve this equality. The full network prediction is consequently unchanged when every transformed adapter is used with its paired factors. Replacing the transformed Gaussian by a new isotropic distribution would discard this guarantee.

D.2 Scalar special case

Proposition 8 (Scalar gauge envelope)**.** *Let $U_B = \mathbb{E}\,\|B\|_F^2$, $U_A = \mathbb{E}\,\|A\|_F^2$, $d_B = mr$, and $d_A = rn$. For $(B, A) \mapsto (cB, A/c)$ with $c > 0$, the unique optimum is*

$$c_*^2 = \frac{-p^2(d_A - d_B) + \sqrt{p^4(d_A - d_B)^2 + 4U_AU_B}}{2U_B}. \tag{22}$$

Substitute $\Gamma = c^2 I$ into (11). Its nonconstant terms are $(c^2U_B + c^{-2}U_A)/(2p^2) + (d_A - d_B)\log c$. Differentiation and $x = c^2$ give $U_Bx^2 + p^2(d_A - d_B)x - U_A = 0$. The only positive root is (22). The matrix family contains this scalar family, so its optimum cannot be worse. At $r = 1$ they coincide.

The matrix operations are on $r \times r$ moments. Forming moments has costs depending on the factor representation; the spectral step costs $O(r^3)$. This is an arithmetic complexity statement, not a measured language-model latency gain.

E Noncentral gauge geometry and fixed-predictor diagnostics

E.1 The fixed noncentral Gaussian identity

Let $P_0 = \mathcal{N}((B_0, A_0), p^2 I_d)$, where (B_0, A_0) and p are fixed independently of the certification sample. Its negative log density at (B', A') is

$$\frac{d}{2}\log(2\pi p^2) + \frac{\|B' - B_0\|_F^2 + \|A' - A_0\|_F^2}{2p^2}.$$

Algorithm 1 Behavior-preserving matrix balancing

Require: posterior moments $S_B, S_A \succ 0$, factor dimensions m, n, fixed prior width p

$C \leftarrow S_B^{1/2} S_A S_B^{1/2}$; diagonalize $C = V \operatorname{diag}(c_i) V^\top$
$d_0 \leftarrow p^2(n-m)$
for each eigenvalue c_i **do**
  if $d_0 \geq 0$ **then**
    $v_i \leftarrow 2c_i / (\sqrt{d_0^2 + 4c_i} + d_0)$
  else
    $v_i \leftarrow (\sqrt{d_0^2 + 4c_i} - d_0)/2$
  end if
end for
$\Gamma_* \leftarrow S_B^{-1/2} V \operatorname{diag}(v_i) V^\top S_B^{-1/2}$; $R \leftarrow \Gamma_*^{1/2}$
transform each complete draw to $(BR, R^{-1}A)$ and evaluate KL by (11)

The posterior entropy after the linear change of variables is $h(Q_R) = h(Q) + (m-n)\log|\det R|$. Substitution of $B' = BR$, $A' = R^{-1}A$ proves (14). This calculation allows correlation between the two Gaussian factors and transforms the full covariance, not just its marginal variances.

For the polar parameterization $R = \Gamma^{1/2} O$, $R^{-1} = O^\top \Gamma^{-1/2}$. The noncentral cross terms become

$$\langle M_B R, B_0 \rangle_F = \operatorname{tr}\Big(O^\top \Gamma^{1/2} M_B^\top B_0 \Big),$$
$$\langle R^{-1} M_A, A_0 \rangle_F = \operatorname{tr}\Big(O^\top \Gamma^{-1/2} M_A A_0^\top \Big).$$

The remaining quadratic terms are $\operatorname{tr}(S_B \Gamma)$ and $\operatorname{tr}(S_A \Gamma^{-1})$. Thus the orientation enters only through $-2\operatorname{tr}(O^\top F(\Gamma))$. If $F = L\Lambda V^\top$, orthogonality gives

$$\operatorname{tr}(O^\top F) = \operatorname{tr}\big(V^\top O^\top L \Lambda\big) \leq \sum_i \Lambda_{ii} = \|F\|_*.$$

Taking $O = LV^\top$ attains equality, including when F is rank deficient. This proves Proposition 7.

To prove existence over Γ, define the centered second moments $G_B = \mathbb{E}[(B - M_B)^\top (B - M_B)] \succ 0$ and $G_A = \mathbb{E}[(A - M_A)(A - M_A)^\top] \succ 0$. Positive definiteness follows from nonsingularity of the joint Gaussian. Variance decomposition gives

$$\mathbb{E}\,\|BR - B_0\|_F^2 \geq \operatorname{tr}(G_B \Gamma), \qquad \mathbb{E}\,\big\|R^{-1}A - A_0\big\|_F^2 \geq \operatorname{tr}(G_A \Gamma^{-1}).$$

For lower eigenvalue bounds $g_B, g_A > 0$, the objective is therefore bounded below by a constant plus

$$\sum_i \left[\frac{g_B \lambda_i + g_A/\lambda_i}{2p^2} + \frac{n-m}{2} \log \lambda_i \right],$$

where λ_i are the eigenvalues of Γ. Each summand diverges at zero and infinity. The reduced objective is continuous, so it attains a minimum on a compact sublevel set. This is an existence proof, not a convexity or uniqueness argument. At $B_0 = A_0 = 0$, $F(\Gamma) = 0$ and the zero-centered theorem is recovered exactly.

E.2 Numerical use without assuming the zero-centered optimizer

The identity transformation is always a valid candidate. For a fixed Γ, the polar step minimizes over the complete orthogonal group, including both determinant signs. One may optimize the remaining objective over a positive-definite parameterization, compare all attained values with the identity, and retain the smallest actual Gaussian KL. Rank deficiency of $F(\Gamma)$ can make the nuclear-norm term nondifferentiable; numerical stationarity is not a global-optimality certificate. The exact zero-centered optimizer is a possible initialization, not a substitute for the noncentral objective.

For a direct matrix-variable check, let $L_B = M_B^\top B_0$, $L_A = M_A A_0^\top$, and $\Delta = n - m$. At nonsingular R, differentiation of (14) gives

$$\nabla_R \operatorname{KL}(Q_R \| P_0) = p^{-2}\big[S_B R - L_B - R^{-\top} R^{-1} S_A R^{-\top} + R^{-\top} L_A^\top R^{-\top}\big] + \Delta R^{-\top}. \tag{23}$$

This gradient evaluates the original smooth objective on $GL(r)$, whereas the reduced objective has already optimized out the orthogonal factor. For each retained R, paired sampling must use $(BR, R^{-1}A)$ exactly. Replacing this pushforward by a freshly isotropic Gaussian changes the stochastic predictor.

For any invertible measurable T applied to both distributions, the change-of-variables Jacobians cancel in their density ratio:

$$\mathrm{KL}(T_{\#}Q \| T_{\#}P_0) = \mathrm{KL}(Q \| P_0).$$

The improvement studied here instead keeps P_0 fixed and changes the parameter distribution within a prediction-equivalent orbit. A zero initial B_0 does not restore the zero-centered theorem when $A_0 \neq 0$: the surviving term $\Gamma^{-1/2} M_A A_0^\top$ still determines a preferred orientation.

E.3 A VALID INITIALIZATION-CENTERED COMPARISON

A matched comparison must save the actual initialization of every trained block, including the classification head, before learning from S. For a posterior with block widths q_b, the two candidate prior centers use the same posterior and therefore exactly the same predictions. Their Gaussian KLs differ only in the squared displacements when prior covariances are identical. A posterior selected for prediction stability is a separate choice: changing q_b changes both its loss and covariance cost, so the two choices cannot be combined into an attributed gain without the matched comparisons.

If both origin and initialization centers are selected together with the nine rank–regularization configurations, a uniform declared mixture has 18 components, not nine. The width search remains posterior-side when the prior is fixed. Final certification draws must be independent of that search, or a simultaneous bound must cover all inspected candidates. Hoeffding and Chernoff comparisons use the same draw-level losses; selecting between them requires covering both unless the inequality was fixed independently of those draws. All these comparisons can use Theorem 1, but none supplies an initialization-centered empirical result in the absence of the corresponding checkpoints and draws.

E.4 REPRODUCING THE ACCOUNTING FLOORS FROM RECORDED SUMMARIES

At $K_{\mathrm{MC}} = 9$ and $\eta = .025$, let $c = \log 360$. For $M = 8$, clipping $\widetilde{r} + \sqrt{c/16}$ to one implies $\widetilde{r} \geq 0.3934676414$. Hence

$$U_{\mathrm{kl}} \geq \mathrm{kl}_+^{-1}(1 - \sqrt{c/16}, c/8) = 0.8941321961\ldots.$$

For $M = 64$, the conservative lower endpoint of the reported three-decimal interval is 0.6715. Since it is below one, clipping is inactive at that minimum. Therefore every recorded mean satisfies

$$\widetilde{r} \geq 0.6715 - \sqrt{c/128} = 0.4570584281\ldots,$$

which yields $U_{\mathrm{kl}} \geq 0.6662649085\ldots$. These inequalities use the stated rounded summaries, not unobserved per-configuration means.

The function $[\log 9 + \log(2\sqrt{N}/.025)]/N$ decreases for $N \geq 2$. Thus $N \leq 1024$ permits the common favorable confidence cost

$$c_{\mathrm{PB},0} = \frac{\log 9 + \log(64/.025)}{1024}.$$

Applying $\mathrm{kl}_+^{-1}(\cdot, c_{\mathrm{PB},0})$ gives lower bounds $0.9320526640\ldots$ and $0.7298471427\ldots$, respectively. Table 9 rounds downward. Both exceed 0.5, even though KL is optimistically set to zero. Enlarging the confidence family increases these floors. Increasing M with newly observed losses, changing posterior widths, or changing the empirical-risk estimator is outside this fixed-summary calculation.

F LANGUAGE-MODEL EXPERIMENTAL DETAILS

The center-risk proxy is explicitly

$$\widehat{R}_S(\delta_{\mu_k}) + \sqrt{\frac{\mathrm{KL}_{\mathrm{LoRA}} + \log(2\sqrt{N}/.05)}{2(N-1)}}.$$

Table 5: Paired selector experiment. Requested training-pool sizes are capped by the available split. Confidence guarantees hold separately for each selection problem.

Component	Setting
Backbone	RoBERTa-base and its tokenizer; fitted sequence-classification head
Training pools	Requested sizes $256, 512, 1024$; three seeds $0, 1, 2$; seven tasks
Sequence processing	Maximum 128 tokens; dynamic padding
Optimization	AdamW, base learning rate 5×10^{-5}, head multiplier 10; three epochs with a floor of 400 optimizer steps; batch size 16
Adapters	Query/value matrices; ranks $4, 16, 64$; scaling $\alpha = 2r$; dropout zero; no adapter bias
Regularization	LoRA decay $10^{-3}, 10^{-2}, 10^{-1}$; other trained blocks have zero decay
Prior	Origin-centered isotropic Gaussian with $p = 1/\sqrt{\lambda}$; nine components with weight $1/9$
Posterior	Gaussian on all trained coordinates; one width selected from seven candidates before final independent draws
Confidence	$\delta_{\mathrm{PB}} = .025$, $\eta = .025$, $K_{\mathrm{MC}} = 9$ for a frozen final posterior per configuration
Evaluation	Public development examples, capped at 800; MRPC 408, RTE 277, WNLI 71
Hardware	One NVIDIA A10G
Inference	Paired task means; all 2^7 sign flips; 10,000 task-bootstrap draws

Table 6: Secondary paired comparisons. Intervals are descriptive 95% task-cluster bootstrap intervals. Cross-validation rows use only their matched 21-problem grid. Native metrics re-score frozen accuracy-selected configurations; selection was not rerun with a different metric.

Selector on matched 21 problems	Accuracy	Effect relative to validation
Five-fold cross-validation	.7180	$+.0062$ $[-.0059, .0209]$
Center-risk proxy	.7166	$+.0047$ $[-.0046, .0132]$
Proxy minus cross-validation	–	$-.0014$ $[-.0225, .0142]$

Task	Metric	Proxy	Training risk	Validation
CoLA	Matthews correlation	.416	.412	.379
MRPC	F1	.847	.847	.846
QQP	F1	.654	.653	.664

It omits the fitted head, hierarchical penalty, and Monte Carlo upper bound. The literal certificate uses (6) with the actual sample size. The historical width search has seven candidates, but their numerical values and whether they were absolute or prior-relative are not recorded in the available experimental specification. This prevents exact replication of that historical search. The explicitly listed nine-width frontier and controlled-study grids below are different experiments and do not reconstruct it. The certified dimension is obtained from all trained blocks rather than an adapter-only name filter.

Validation and five-fold cross-validation select configurations without access to evaluation labels. The retrained validation baseline uses the entire training pool only after selection. The unretrained validation model reaches 0.695 mean accuracy, compared with 0.705 for the proxy, but this comparison gives the proxy more fitting data. It is not the primary comparison. A test-label oracle has center accuracy 0.7339 and stochastic accuracy 0.5145; it is an upper reference, not an admissible selector. Complexity-only selection coincides with the smallest-rank rule in this audit.

F.1 Secondary selector comparison

RoBERTa-base (Liu et al., 2019) is fine-tuned on seven GLUE tasks (Wang et al., 2019): CoLA, MRPC, QNLI, QQP, RTE, SST-2, and WNLI. A selection problem combines one task, a requested training pool in $\{256, 512, 1024\}$, and a seed in $\{0, 1, 2\}$. Each problem contains the nine combinations of rank $\{4, 16, 64\}$ and LoRA weight decay $\{10^{-3}, 10^{-2}, 10^{-1}\}$, for 63 problems and 567

Table 7: Selection on 63 problems. Accuracy is evaluated at the selected deterministic center and under its posterior. Effects are relative to validation with retraining, with descriptive 95% task-cluster bootstrap intervals. Wins count tasks, not individual configurations.

Selector	Center	Stochastic	Effect [95% interval]	Wins
Center-risk Gaussian-KL proxy	.7053	.5107	$-.0035$ $[-.0209, .0070]$	5/7
Training risk	.7066	.5186	$-.0023$ $[-.0206, .0086]$	5/7
Validation, retrained	.7088	.5152	0	–
BIC	.6883	.5235	$-.0205$ $[-.0373, -.0002]$	1/7
Smallest rank	.6877	.5229	$-.0211$ $[-.0375, -.0017]$	1/7

Table 8: The original eight-draw certificate and a matched 64-draw rerun. The latter covers 216 configurations in 24 problems. Selected accuracies in the last row are compared only within this common subset; stochastic evaluation predictions were not recorded for the rerun.

Quantity	Center-risk proxy	Certificate, $M = 8$	Certificate, $M = 64$
Charged blocks	LoRA only	All trained	All trained
Mean KL (nats)	4.29	12.49	12.67
Risk input	Center	Upper bound 1	Upper bound .672–.761
Score or certificate	Mean .1935	Constant 1	.736–.928
Selected accuracy, matched 24	.7634	.7257	.7388

configurations. Query and value adapters and the classification head are trained. Evaluation uses the public development split, capped at 800 examples; actual smaller sets contain 408 MRPC, 277 RTE, and 71 WNLI examples. WNLI has only 635 training examples, so requested and actual pool sizes are distinguished.

The center-risk proxy adds a Gaussian-KL square-root penalty to deterministic training error, with only LoRA factors charged. Its prior is $\mathcal{N}(0, \lambda^{-1}I)$ and its risk input is not posterior-averaged. The valid certificate instead charges every trained coordinate, uses a uniform nine-configuration prior, and estimates stochastic risk with independent final draws. The total per-problem failure probability is 0.05, split equally between PAC-Bayes and Monte Carlo. Posterior widths are selected before final certification; Table 5 lists the protocol, and the limitations of its recorded settings follow the table.

Validation selects on a 25% training-pool holdout and then retrains the chosen configuration on the complete pool. This is the appropriate comparison with full-pool-trained checkpoints, not the weaker unretrained validation model. Effects are paired within problems and reduced to seven task means. Exact sign flips use task-level effects and require joint sign symmetry under the null; task-cluster bootstrap intervals are descriptive because the benchmark tasks are not sampled from a defined population of tasks.

The proxy reaches 0.7053 mean accuracy, below training risk (0.7066) and retrained validation (0.7088). Its task-level sign-flip p-value against validation is 0.9844; this does not establish equivalence. The more important difference is between predictors: the selected centers are substantially more accurate than their associated stochastic models (Table 7). A detailed MRPC cell has center and posterior training errors 0.0703 and 0.6040, respectively. On evaluation examples their errors are 0.1912 and 0.6379, with 54.2% disagreement. The sampled models collapse to constant-class predictions. A small Gaussian KL therefore coexists with an unusable stochastic predictor. Figure 5 shows the same gap for every selector, together with the distributions of proxy scores and literal certificates.

F.2 Historical draw-budget comparison

At $M = 8$ and $K_{\mathrm{MC}} = 9$, the Hoeffding half-width is 0.6065. Added to the observed posterior errors, it clips every empirical-risk upper bound to one. All 567 certificates consequently equal one mathematically; the stored inversion clips to $1 - 10^{-12}$. Every problem has a nine-way tie, and enumeration selects the smallest rank in 63/63 problems. The complete KL is only 2.34–45.91 nats, with mean $\mathrm{KL}\,/N = 0.0289$. Monte Carlo uncertainty, rather than complexity, causes this saturation.

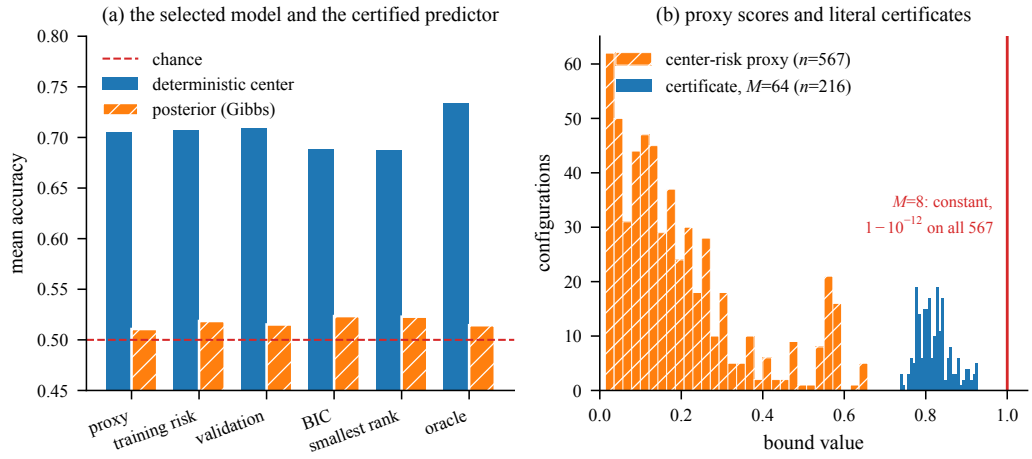


Figure 5: Selected models versus the certified predictor on the 63 small-pool problems. (a) Mean evaluation accuracy of each selector's chosen configuration at its deterministic center and under its Gaussian posterior; the oracle uses test labels and is not an admissible selector. (b) Center-risk proxy scores on all 567 configurations and literal 64-draw certificates on the matched 216; every 8-draw certificate equals one.

Table 9: Arithmetic lower bounds for the recorded LoRA grids. Each entry is a conservative lower bound, not a rerun mean or a certificate issued after setting KL to zero. The calculation grants the favorable $N = 1024$, uses $w_k = 1/9$ and $\delta_{\mathrm{PB}} = \eta = .025$, and rounds downward.

Draws	Configurations	Chernoff risk UCB floor	Zero-KL certificate floor
8	567	0.8941	0.9320
64	216	0.6662	0.7298

With $M = 64$, the half-width falls to 0.2144. All 216 certificates in the matched rerun are distinct, with range 0.736–0.928. Their selected center accuracy is 0.7388, below the proxy (0.7634), training risk (0.7655), and retrained validation (0.7582) on those same 24 problems. Removing a tie does not make the certificate a better selector. Conversely, near-one values alone do not prove poor ordering: selected-model accuracy must be evaluated explicitly.

F.3 What tighter accounting can change on the recorded LoRA grid

The small KL in Table 8 makes the complexity-free floor directly relevant. At $M = 8$, Hoeffding saturation implies $\widetilde{r} \geq 1 - \sqrt{\log(360)/16}$. Even the more favorable single-inequality Chernoff charge $K_{\mathrm{MC}} = 9$ then gives $U_{\mathrm{kl}} \geq 0.8941$. For the 64-draw grid, the smallest reported Hoeffding upper bound is 0.672. Taking its three-decimal rounding conservatively as 0.6715 gives

$$\widetilde{r} \geq 0.6715 - \sqrt{\frac{\log 360}{128}}, \qquad U_{\mathrm{kl}} \geq 0.6662. \tag{24}$$

Because every training pool has $N \leq 1024$, Equation (10) strengthens these to the conservative certificate floors in Table 9. Thus even a hypothetical zero-KL representation, combined with Chernoff instead of Hoeffding, cannot meet $\tau = 0.5$ at either recorded budget. This conclusion holds before estimating whether real trained factors benefit from matrix balancing.

The floors compare accounting on the same realized draw losses; they are not lower confidence bounds on population risk. More draws or a different posterior can change them. Inspecting both inequalities with a simultaneous guarantee requires an additional Monte Carlo allocation, which can only increase these optimistic floors. In particular, an initialization-centered prior cannot rescue these fixed losses through a smaller KL alone, although initialization-centered, narrower posteriors remain a different and unmeasured comparison.

Figure 6 inverts the Hoeffding budget (21) on the 216 recorded 64-draw configurations, holding each configuration's complete KL and observed mean fixed. At $\tau = 0.5$ no finite number of draws suffices for any configuration, because the observed mean is not below the allowed risk u_τ or no such risk exists. At $\tau = 0.6$ this holds for 145 configurations and at $\tau = 0.7$ for 28, with median requirements of 6137 and 343 draws among the rest. Only at $\tau = 0.8$ and 0.9 do 64 draws suffice for 71 and 201 configurations. This is fixed-summary arithmetic, not a sequentially valid stopping rule.

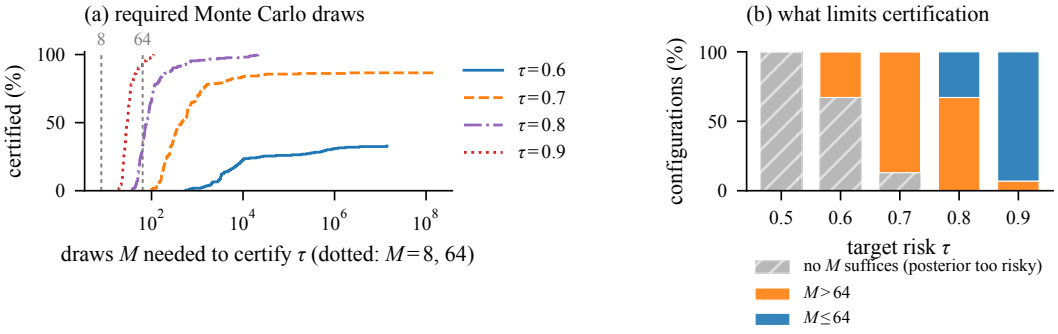


Figure 6: Monte Carlo draws needed on the 216 recorded 64-draw configurations, with each configuration's KL and observed mean held fixed and the Hoeffding bound at $K_{\mathrm{MC}} = 9$. (a) Cumulative share certifiable at target τ as a function of the required draws M; curves stop below 100% when no finite M suffices. (b) Binding constraint per target: no finite M suffices (observed mean not below u_τ, or no u_τ exists), more than 64 draws, or at most 64 draws.

F.4 Family exclusion and a behavior-preserving KL correction

For each of the 567 configurations, we evaluate Corollary 4 with the origin-centered prior, $q_{\max} = 1$, and the complete dimension. The exclusion counts at risk targets $\tau = 0.5, 0.4, 0.3, 0.2, 0.1$ are, respectively, $243, 261, 429, 511, 567$. These exclusions cover all widths in $(0, 1]$, not just sampled grid points. At the empirical training-pool majority-error reference, $297/567$ configurations are excluded; that reference is not itself a certified population risk. A nine-width risk sweep has a simultaneous half-width of 0.913, giving zero lower bounds throughout and no sharper exclusions.

Figure 7 shows the full distribution of $q_{\mathrm{cert}}(\tau)$ behind these counts. The median threshold moves above $q = 1$ as the target tightens, so the exclusions at small τ are not driven by a few extreme configurations.

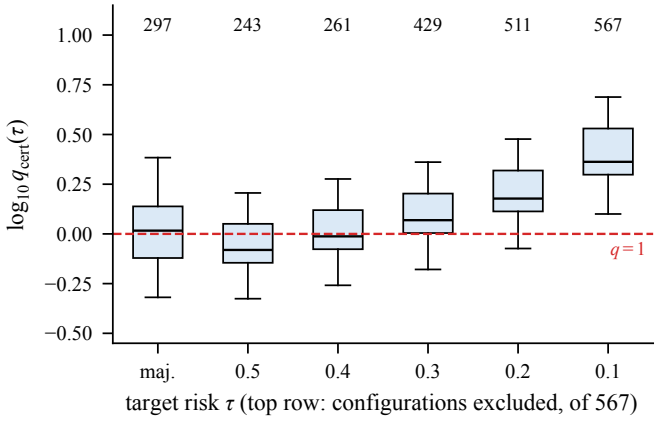


Figure 7: Prior-scale-free width threshold $q_{\mathrm{cert}}(\tau)$ over the 567 configurations (origin-centered prior, complete dimension). Boxes show quartiles and whiskers 1.5 interquartile ranges; values above the dashed line exclude every $q \in (0, 1]$. The top row gives the excluded count; "maj." uses each configuration's training-pool majority error as its target.

F.5 Task dependence and the WNLI sensitivity analysis

The exact two-sided task sign-flip p-values are 0.9844 for the proxy and training-risk selector, and 0.1094 for BIC and the smallest rank. Equal absolute statistics are included in the tail. These tests rely on joint sign symmetry and do not transform related benchmark tasks into an independent random sample of NLP problems.

Excluding WNLI post hoc reverses the proxy's mean effect to $+0.0048$ with interval $[0.0016, 0.0080]$ and $p = 0.0625$. Training risk then has effect $+0.0065$, interval $[0.0030, 0.0095]$, and the same p-value. WNLI has only 71 evaluation examples and every evaluated selector is below its majority baseline. The exclusion is a sensitivity analysis, not a prospective claim of superiority on a newly defined task population.

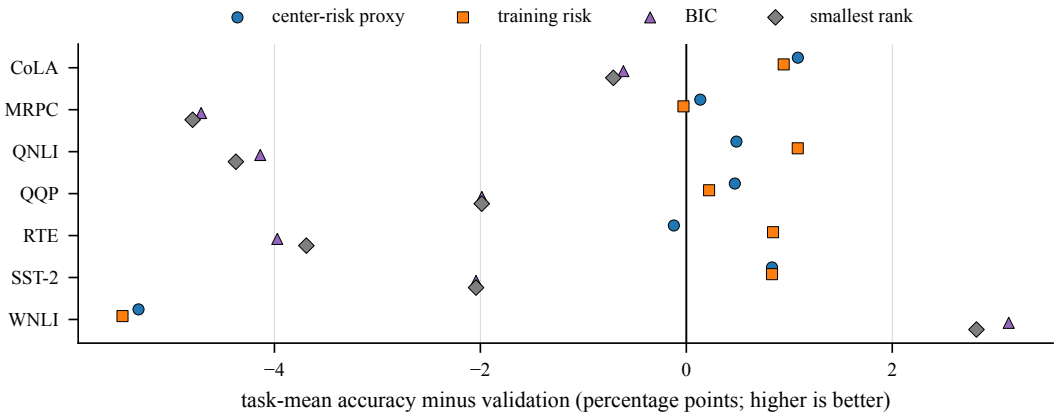


Figure 8: Task-level selector effects relative to retrained validation, in percentage points of task-mean accuracy over the nine problems of each task (higher is better). WNLI, with 71 evaluation examples, drives the negative mean effect of the proxy and training-risk selectors.

F.6 Posterior and confidence diagnostics

On the MRPC cell with $N = 512$, rank 16, and decay 10^{-2}, posterior draws attain the two constant-class evaluation risks 0.3162 and 0.6838. The $M = 16$ posterior-risk upper bound is 0.9435; replacing the center-risk input by this upper bound moves the inverse-KL expression from 0.1251 to 0.9757. The latter is mathematically below one but does not establish a useful classifier.

The $M = 64$ rerun uses CoLA, MRPC, RTE, and SST-2. A more conservative variant uses $K_{\mathrm{MC}} = 63$ and also charges $\log(63/9)$ extra nats in the mixture penalty. It yields 0.772–0.946 and the same selected configuration as $K_{\mathrm{MC}} = 9$ on 23 of 24 problems. This variant is not needed when width selection precedes a fresh final stream, but shows the consequence of a larger declared family.

The width screen uses $D = \|\mu\|^2$ and $a_k = \log 9 + \log(2\sqrt{N}/.025)$. A consistency check between the analytic endpoint at $q = 1$ and nine logarithmic widths in $[10^{-4}, 1]$ yields no disagreements. The mean measured posterior risk at $q = 1$ is 0.50, but the simultaneous half-width 0.913 makes all risk lower bounds zero. These risk measurements therefore do not contribute to the exclusion counts. Initialization-centered priors and widths above one remain outside the reported exclusion.

G Controlled numerical studies

G.1 Gaussian-factor classification

For each of 20 seeds, sample $N = 2048$ independent $x_i \sim \mathcal{N}(0, I_2)$ and labels $y_i = \mathrm{sign}(x_{i1})$. Define $\bar{a} = N^{-1} \sum_i y_i x_i$ and $\mu_A = 2\bar{a}/\|\bar{a}\|$. The posterior has independent factors $B \sim \mathcal{N}(1, q^2)$ and $A \sim \mathcal{N}(\mu_A, q^2 I_2)$, and predicts $\mathrm{sign}(BAx)$. The fixed prior is $\mathcal{N}(0, I_3)$. Its exact KL is

$$\mathrm{KL}(Q_q \| P) = \tfrac{1}{2}\left[1 + \|\mu_A\|^2 + 3q^2 - 3 - 3\log q^2\right].$$

The center is fitted on S, while the prior is independent of it. A uniform configuration label costs $\log 9$ conservatively; all widths share the same prior, so this penalty is not mathematically necessary for posterior-width selection.

The widths are $\{.05, .1, .2, .4, .8, 1.2, 2, 4, 8\}$. Each receives 1024 independent draws, with fixed nested prefixes of sizes $8, 64, 256, 1024$. We allocate $\eta = .025$ across nine widths, four budgets, and two inequalities, so both reported Monte Carlo layers use $\log(72/.025)$. The PAC-Bayes allocation is $\delta_{\mathrm{PB}} = .025$. All certificate arithmetic uses (6) and the same sampled predictions within a comparison.

The empirical stochastic risk has an analytic reference. Set $b = \Phi(-1/q)$ and $a_i = \Phi(-y_i \mu_A^\top x_i / (q \, \|x_i\|))$. Independence of B and A gives

$$\widehat{R}_S(Q_q) = b + (1 - 2b) \frac{1}{N} \sum_i a_i. \tag{25}$$

This formula checks Monte Carlo coverage without treating a longer Monte Carlo stream as ground truth. Across the 720 width–budget–seed rows, neither upper-bound implementation undercovers this reference. This is an implementation check, not empirical proof of its nominal coverage. The sample SD in Table 2 includes dataset and posterior-sampling variation through the independent seeds; it is not a standard error.

G.2 Matrix-gauge construction and independent optimization

We use $(m, n) \in \{(12, 8), (8, 12), (12, 12)\}$, ranks $r \in \{1, 2, 4, 8\}$, and seeds 0–19. A random orthogonal matrix O rotates a deformation with eigenvalues $\exp(\mathrm{linspace}(-2, 2, r))$; for rank one the eigenvalue is e^2. Let this deformation be G. Independent Gaussian arrays generate $M_B = (G_B/\sqrt{m})G$ and $M_A = G^{-1}(G_A/\sqrt{n})$. The posterior uses independent isotropic factor noises $s_B = .15, s_A = .10$ and the prior width is $p = 1$. Thus

$$S_B = M_B^\top M_B + m s_B^2 I, \qquad S_A = M_A M_A^\top + n s_A^2 I.$$

The deformation deliberately makes scalar and matrix balancing distinguishable; it is not a claim about the distribution of trained language-model factors.

For every case, we compute raw, scalar, and matrix-optimal KL. An independent BFGS solve parameterizes $\Gamma = LL^\top$ with exponential diagonal entries in L, starts from the identity, and minimizes the same objective with an analytic gradient, tolerance 10^{-8}, and at most 1000 iterations. We report the objective difference, not a blanket optimizer-convergence assertion. The maximum relative difference, normalized by $1 + |\,\mathrm{KL}_*\,|$, is 9.91×10^{-16}. The matrix stationarity residual is normalized by $1 + \|S_A\|_F$.

For each case, 128 independent factor draws are paired before and after the exact pushforward. Product discrepancy is $\|B'A' - BA\|_F / (1 + \|BA\|_F)$. Its maximum is 2.71×10^{-15}. The inclusion ordering $\mathrm{KL}_{\mathrm{matrix}} \le \mathrm{KL}_{\mathrm{scalar}} \le \mathrm{KL}_{\mathrm{raw}}$ is checked in every row. Seed-level outputs and complete summaries for all three matrix shapes accompany the code.

H Split-prior numerical evaluation scenarios

Displacements, stochastic losses, matrix-KL reductions, and decoded errors are assumed inputs; the raw Gaussian KLs, confidence bounds, width thresholds, certification margins, and evaluation counts are calculated from them.

H.1 Independent prior and posterior specification

The split-prior design uses RoBERTa-base with rank-four query and value adapters, a prior-data partition of 33,000 units, and nominal certification-sample sizes $N \in \{1000, 4000, 16000, 33000\}$ for SST-2 and $N \in \{1000, 4000, 16000, 70000\}$ for QNLI. These are numerical design sizes, not verified independent-unit counts. The classification head is fitted only on the prior partition and is frozen before certification data are accessed. Query and value factors across the twelve width-768 layers contain 147,456 certified coordinates. Source-level independence and available sample counts remain conditions of the application; row disjointness alone does not establish independence.

Table 10: Split-prior sample-size frontier. $\widetilde{r}$ and D are assumed, K_R is an assumed attained matrix KL, and U_{kl} and C_R are calculated. KL is in nats; risks are proportions. The same posterior losses apply before and after a valid pushforward.

Task	N	D	$\widetilde{r}$	U_{kl}	K_I	K_R	C_R
SST-2	1,000	0.42	0.059	0.0973	525.00	490.35	0.5719
SST-2	4,000	0.46	0.066	0.1059	575.00	547.40	0.3321
SST-2	16,000	0.50	0.071	0.1121	625.00	601.25	0.2182
SST-2	33,000	0.54	0.076	0.1181	675.00	652.73	0.1928
QNLI	1,000	0.62	0.081	0.1241	775.00	713.77	0.7003
QNLI	4,000	0.70	0.089	0.1336	875.00	821.62	0.4314
QNLI	16,000	0.82	0.098	0.1442	1025.00	976.82	0.2939
QNLI	70,000	0.90	0.107	0.1547	1125.00	1081.12	0.2255

The training specification uses seeds $0, 1, 2$, three prior-training epochs, maximum sequence length 128, batch size 16, LoRA scale $\alpha = 2r$, zero adapter dropout, and AdamW. The prior-stage factor and head learning rates are 5×10^{-5} and 5×10^{-4}, with factor weight decay 10^{-2}. Each continued checkpoint starts from its saved prior center and takes 400 factor-only steps at learning rate 2×10^{-6} with zero weight decay. Both prior factors can be nonzero after this training. The numerical tables instantiate single assumed profiles, not averages over executed seed replicates.

Prior widths belong to $\{.005, .01, .02, .04\}$ with mixture weights $1/4$. The posterior-width grid is $\{.005, .01, .015, .02, .025, .03, .04\}$. A pilot searches the Cartesian product of the factor-group widths using 32 draws on at most 1024 certification examples. The final family retains the seven equal-width profiles, twelve profiles varying one group around .02, and at most four additional pilot choices, for at most 23 distinct profiles. All final profiles are frozen before independent final draws.

The confidence allocation for each task–seed study is .02 for Gaussian PAC-Bayes, .02 for Monte Carlo, .005 for compression, and .005 for the fixed-label reference. Division over the four sample sizes gives $\delta_{\mathrm{PB}} = \eta = .005$ and $\delta_{\mathrm{code}} = \delta_{\mathrm{base}} = .00125$ per size. The Monte Carlo family covers 23 profiles, two fixed budgets, and both inequalities, giving $K_{\mathrm{MC}} = 92$. Balancing adds no draw-level multiplicity when it preserves the complete paired stochastic predictor. The principal tables use $M = 1024$. Guarantees across additional task–seed studies require their own allocation.

H.2 Sample-size frontier

Table 10 instantiates the equal-width case $p = q_A = q_B = .02$. Its raw KL is exactly $D/(2p^2)$; no covariance-mismatch term is omitted. The Chernoff upper bound is computed from the assumed draw mean with the full final-family charge. The matrix-KL entries are stipulated attainable values, not values derived from the scalar displacement alone. Their use in the certificate is conditional on a valid noncentral pushforward attaining them.

Under these assumed inputs, the SST-2 certificate decreases from 0.5719 at $N = 1000$ to 0.1928 at $N = 33000$. The QNLI scenario remains above the target $\tau = .2$ even at $N = 70000$, where its value is 0.2255. This contrast is a conditional calculation: more data can reduce the bound while a larger stochastic loss still prevents a fixed target from being met. Figure 9 plots Tables 10 and 11.

H.3 Posterior-risk inputs and prior-width sensitivity

For each assumed posterior width, Table 11 chooses the lowest raw Gaussian KL among the four declared prior widths, with their mixture cost retained. This is not substitution of the oracle prior scale. The posterior-risk entries are stipulated draw means. The nonmonotone complexity at intermediate widths is a consequence of the coarse admissible prior grid: a small displacement does not remove covariance mismatch. Matrix-KL reductions outside the equal-width point are assumed to be smaller than the principal reduction, rather than extrapolated from the synthetic balancing experiment.

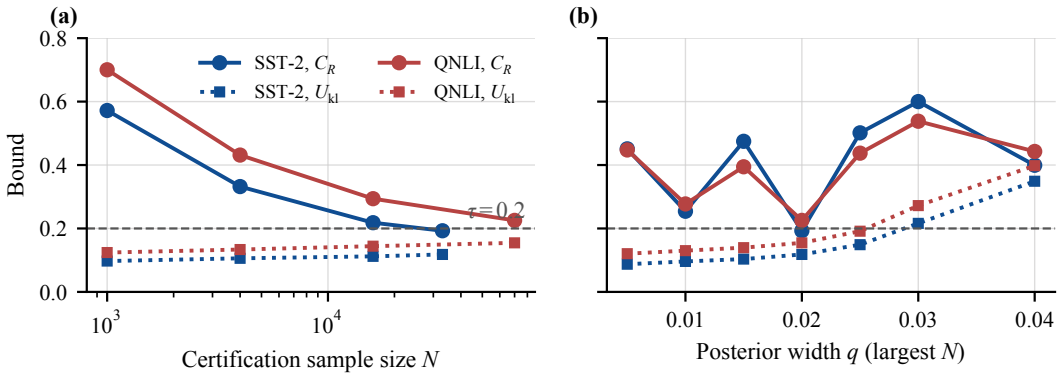


Figure 9: Split-prior scenarios (assumed inputs; values are formula-derived, not measured). (a) Balanced certificate C_R and Chernoff risk bound U_{kl} versus certification sample size at $p = q = .02$. (b) The same quantities versus posterior width at the largest N; the nonmonotone certificate follows the coarse prior-width grid. The dashed line is $\tau = 0.2$.

Table 11: Posterior-width curves. Nominal N is 33,000 for SST-2 and 70,000 for QNLI. Raw KL includes both displacement and covariance mismatch. All displayed certificate entries are recalculated from the row inputs.

Task	q	p	$\widetilde{r}$	U_{kl}	K_I	K_R	C_R
SST-2	0.005	0.005	0.051	0.0873	10800.0	10568.3	0.4510
SST-2	0.010	0.010	0.058	0.0961	2700.0	2642.1	0.2543
SST-2	0.015	0.020	0.064	0.1035	10839.4	10606.9	0.4748
SST-2	0.020	0.020	0.076	0.1181	675.0	652.7	0.1928
SST-2	0.025	0.020	0.102	0.1489	9243.1	9044.9	0.5015
SST-2	0.030	0.040	0.161	0.2161	10333.2	10111.6	0.6000
SST-2	0.040	0.040	0.284	0.3489	168.8	165.1	0.3993
QNLI	0.005	0.005	0.078	0.1205	18000.0	17543.7	0.4475
QNLI	0.010	0.010	0.086	0.1301	4500.0	4385.9	0.2773
QNLI	0.015	0.020	0.094	0.1395	11289.4	11003.3	0.3945
QNLI	0.020	0.020	0.107	0.1547	1125.0	1081.1	0.2255
QNLI	0.025	0.020	0.139	0.1914	9693.1	9447.4	0.4374
QNLI	0.030	0.040	0.212	0.2721	10445.7	10180.9	0.5381
QNLI	0.040	0.040	0.331	0.3980	281.2	274.1	0.4428

H.4 Noncentral balancing and the binding constraint

A scalar or matrix transformation is compared with the identity using Equation (14), with the prior center held fixed. Numerical optimization begins at the identity and at the zero-centered optimizer, includes both orientation branches through the polar step, and retains the smallest exact KL among valid candidates. A lower attained value is sufficient for a conditional improvement; numerical stationarity is not global optimality. Every draw uses the paired factors $(BR, R^{-1}A)$, rather than newly isotropic perturbations of the rescaled mean.

Table 12 separates the assumed KL reduction from the exactly calculable change in the certificate. Neither the percentage reduction nor its implication for trained factors is inferred from the deliberately deformed synthetic experiment. A full stochastic-predictor comparison additionally checks paired product and logit discrepancies at the chosen evaluation precision.

A positive width-screen result means only that the raw isotropic candidate has not been excluded. The balanced anisotropic distribution is evaluated with its own KL. In the larger QNLI scenario, $C_0 = 0.1618$ leaves room below the target, but the assumed remaining matrix complexity exceeds the allowable budget. Thus its failure is complexity-limited at the specified draw-level risk bound, not proof that every posterior is unusable.

Table 12: Noncentral-gauge comparison. Identity, scalar, and matrix KLs refer to one assumed stochastic predictor. The last columns are formula-derived certificates using the same U_{kl}. Reduction is percent of the identity KL.

Task	K_I	K_{scalar}	K_R	Reduction (%)	C_I	C_R
SST-2	675.00	667.87	652.73	3.30	0.1942	0.1928
QNLI	1125.00	1110.96	1081.12	3.90	0.2270	0.2255

Table 13: Frontier decisions at $\tau = .2$. q_{cert} applies to the raw isotropic family, C_0 is the complexity-free floor, J is the available KL budget, and $J - K_R$ is the balanced margin. Pass is 1 exactly when the margin is nonnegative and $U_{\mathrm{kl}} \leq \tau$.

Task	N	q_{cert}	C_0	J	$J - K_R$	Pass
SST-2	1,000	0.03143	0.1467	28.1	-462.3	0
SST-2	4,000	0.01611	0.1308	116.7	-430.7	0
SST-2	16,000	0.00828	0.1248	430.9	-170.3	0
SST-2	33,000	0.00591	0.1272	770.9	118.1	1
QNLI	1,000	0.03818	0.1779	9.3	-704.4	0
QNLI	4,000	0.01987	0.1609	49.1	-772.5	0
QNLI	16,000	0.01061	0.1583	156.0	-820.8	0
QNLI	70,000	0.00509	0.1618	464.1	-617.1	0

H.5 Fixed-reference and decoded-predictor comparisons

The reference label is selected from the prior-data partition and then fixed. The descriptive empirical reference is distinct from its one-sided population-risk lower bound. In Table 15, integer reference errors are set to the nearest count corresponding to rates .47 and .49, and the lower endpoints are computed by one-sided binomial inversion at $\delta_{\mathrm{base}} = .00125$. A Gaussian upper bound below this lower endpoint would establish the population comparison on the joint event; being below the empirical reference alone would not.

The description-length comparator encodes the transformed center relative to the same prior-trained side information. Quantization depths are 4, 8, and 16 bits. Tensor layouts, dyadic scales, all quantized integers, and a 64-bit payload-length header are included in the assumed complete code lengths. A fixed lossless codec and decoder define a prefix-free grammar for payloads shorter than 2^{64} bytes. Table 16 uses Equation (18) on the decoded predictor rather than reusing the Gaussian posterior risk.

H.6 Fully enumerated small-pool replacement grid

The historical seven-point width search is not reconstructed by the following numerical scenario. A replacement design records the prior-relative grid $q/p \in \{1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1\}$ before any pilot stream. Table 17 instantiates one rank-sixteen, $N = 512$ case with 1,181,954 charged factor-and-head coordinates, $D = 2300$, and $p = 10$. These assumed quantities are not retroactively assigned to a measured historical checkpoint. Final confidence uses $\delta_{\mathrm{PB}} = \eta = .025$, a nine-component prior, and a Monte Carlo family of 36 frozen posterior–budget–inequality combinations; the independent pilot chooses one width per configuration.

H.7 Evaluation workload

Table 18 gives arithmetic workload counts rather than measured latency. Each posterior draw is evaluated on the entire certification sample, so a worst-case 23-profile final family with 1024 draws requires $23 \cdot 1024 \cdot N$ example–posterior evaluations. Cached logits, model batching, hardware, and precision determine wall-clock cost and are not inferred from these counts.

Table 14: Pilot-to-final decision comparison. The pilot projection substitutes its assumed mean into the final-budget calculation and is not a pilot confidence certificate. Forecast and pass use the target .2.

Task	N	Pilot mean	Projected C	Forecast	Final C	Pass
SST-2	1,000	0.056	0.5671	0	0.5719	0
SST-2	4,000	0.068	0.3356	0	0.3321	0
SST-2	16,000	0.068	0.2135	0	0.2182	0
SST-2	33,000	0.078	0.1957	1	0.1928	1
QNLI	1,000	0.078	0.6966	0	0.7003	0
QNLI	4,000	0.091	0.4345	0	0.4314	0
QNLI	16,000	0.095	0.2894	0	0.2939	0
QNLI	70,000	0.109	0.2282	0	0.2255	0

Table 15: Fixed-label reference comparison. The reference error counts are assumed and their lower confidence bounds are calculated. The final indicator is 1 only when the Gaussian upper bound is below the reference lower bound.

Task	N	Reference error	Reference lower	C_R	Below lower
SST-2	1,000	0.4700	0.4221	0.5719	0
SST-2	4,000	0.4700	0.4461	0.3321	1
SST-2	16,000	0.4700	0.4581	0.2182	1
SST-2	33,000	0.4700	0.4617	0.1928	1
QNLI	1,000	0.4900	0.4419	0.7003	0
QNLI	4,000	0.4900	0.4660	0.4314	1
QNLI	16,000	0.4900	0.4780	0.2939	1
QNLI	70,000	0.4900	0.4843	0.2255	1

Table 16: Description-length comparator. L includes the complete assumed code in bits. C_{code} is computed from the assumed decoded error and actual formula, not copied from the Gaussian column. Values rounding to 1.000000 are uninformative at displayed precision.

Task	N	Bit depth	L	Decoded error	C_{code}
SST-2	33,000	4	226,144	0.088	0.996054
SST-2	33,000	8	607,264	0.079	0.999999
SST-2	33,000	16	1,605,248	0.078	1.000000
QNLI	70,000	4	251,808	0.123	0.961728
QNLI	70,000	8	638,624	0.113	0.999462
QNLI	70,000	16	1,687,168	0.111	1.000000

Table 17: Small-pool replacement sweep. Every candidate width is explicitly numeric. The assumed draw means produce the displayed certificates at the indicated budgets; all-trained-coordinate Gaussian KL is computed exactly.

q/p	q	$\widetilde{r}$	KL	C_{256}	C_{1024}
$1/64$	0.15625	0.463	4324787.3	1.0000	1.0000
$1/32$	0.31250	0.474	3505952.0	1.0000	1.0000
$1/16$	0.62500	0.486	2688415.3	1.0000	1.0000
$1/8$	1.25000	0.495	1876072.8	1.0000	1.0000
$1/4$	2.50000	0.501	1084506.7	1.0000	1.0000
$1/2$	5.00000	0.504	376046.8	1.0000	1.0000
1	10.00000	0.507	11.5	0.7542	0.7025

Table 18: Calculated workload. The pilot column counts the full 49-profile pilot at 32 draws over 1024 examples. The final column counts complete-data posterior evaluations, not independent PAC-Bayes observations.

Task	N	Profiles	Draws	Final evaluations	Pilot evaluations
SST-2	33,000	23	1024	777,216,000	1,605,632
QNLI	70,000	23	1024	1,648,640,000	1,605,632