PRISM: A GEOMETRIC RISK BOUND FOR DECOMPOSING DRIFT INTO SCALE, SHAPE, AND HEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

A single base LLM now comes with dozens of post-training variants, quantized, LoRA-adapted, or distilled, and each has to be checked before release. Existing evaluations provide only a partial picture: benchmark scores and likelihood screens say *that* a variant has degraded, similarity scores such as CKA and SVCCA say *how* its features moved, and nothing connects the two. We connect them with one structural fact and one design choice: the prediction head is linear, so feature geometry reaches the loss, and we compare the two feature sets through an orthogonal map, which leaves the geometry being measured unchanged. From these we derive **PRISM**, a closed-form upper bound on the cross-entropy risk gap between a target model and a proxy variant, and prove that it splits *exactly* into three measurable axes: scale, shape, and head. Each axis names a failure mode and where to intervene: low-bit quantization distorts shape and, at the lowest bit-widths, also inflates activation scale; quantizing the output embedding inflates the head term, which dominates the bound at high bit-widths. Because the shape term is differentiable, the same geometry becomes a regularizer that curbs catastrophic forgetting more than experience replay. Across two model families and five benchmarks, PRISM ranks quantized and fine-tuned variants from a single forward pass at mean Spearman above 0.8, and eight reference sequences already recover the ranking that the risk gap itself needs over a hundred to reach.

1 INTRODUCTION

As large language models (LLMs) move from pre-training to deployment, a new bottleneck emerges: a single base model now produces many post-training variants, quantized (GPTQ (Frantar et al., 2023), GGUF, BitsAndBytes (Dettmers et al., 2022)), LoRA-adapted (Hu et al., 2022), or distilled (Hinton et al., 2015), that must be evaluated before release (Grattafiori et al., 2024; Qwen Team, 2025). Existing evaluation largely relies on aggregate accuracy or perplexity, which can reveal that a variant has degraded but not *why*, leaving developers to debug low-bit quantization failures, catastrophic forgetting, or prediction-head corruption by costly trial-and-error. What is missing is a diagnostic that not only predicts degradation, but also identifies which component of the model has drifted from the base checkpoint.

A natural alternative is to compare internal representations directly. Prior representational similarity methods such as SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), and generalized shape metrics summarize two feature matrices with a single alignment score, which is descriptive but not diagnostic. Yet the prediction head of an LLM is linear, so any change in feature geometry reaches the loss through a fixed linear map. Can the loss gap between a model and its variant therefore be bounded, and attributed to its source, from feature geometry alone? Three obstacles have stood in the way: (i) no prior similarity has been tied to downstream cross-entropy risk on the deployed prediction head (Ding et al., 2021; Davari et al., 2023; Klabunde et al., 2025); (ii) feature drift does bound the loss through a Lipschitz argument, but the standard constant grows with vocabulary size and is 50 to 60 times larger than ours on 8B models (Appendix A.3); and (iii) a usable diagnostic must further decompose the bound into geometrically interpretable axes that map to distinct failure mechanisms.

Our key insight is that one structural property of modern LLMs, paired with one choice of comparison map, resolves all three obstacles. The linear `lm_head` over a non-linear backbone yields a sharper Lipschitz constant, set by pairwise distances between token embeddings rather than by the spectral

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.

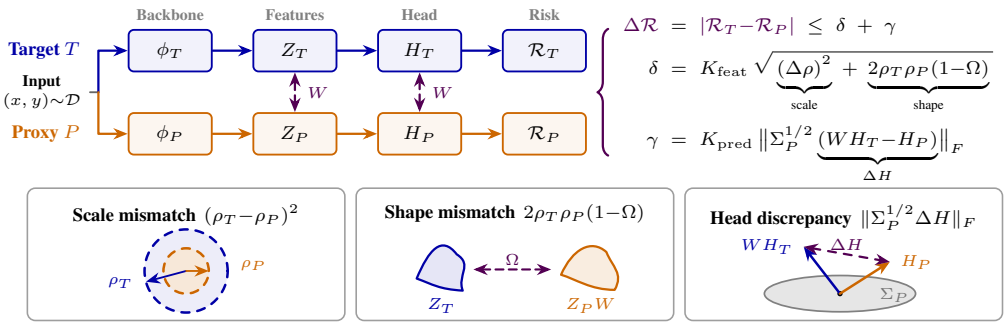

Figure 1: **PRISM** (**P**roxy **R**isk **I**nference via **S**tructural **M**apping) **decomposition of the risk-gap bound.** *Top:* target T and proxy P run parallel pipelines on a shared input (backbone $\rightarrow$ features Z $\rightarrow$ head $H \rightarrow$ risk $\mathcal{R}$), related by an orthogonal alignment $W \in \mathcal{O}(d)$; Thm. 1 bounds their risk gap by a feature alignment error δ plus a head discrepancy γ. *Bottom:* the three axes the bound separates. Scale compares the feature radii ρ_T, ρ_P; shape compares the arrangement of the two feature clouds at fixed radius, and Prop. 1 splits δ exactly into these two; head compares the aligned heads through their residual $\Delta H = W H_T - H_P$, measured only where the data has support. Each axis localizes a distinct empirical failure mode (Sec. 4.4).

norm of the head, and therefore independent of vocabulary size (Sec. 3.2). Comparing the two feature sets through an orthogonal map then leaves their internal distances and angles untouched, so the alignment residual splits *exactly* into a scale part and a shape part read at the original feature scale. Together with a covariance-weighted head-side term, these give **PRISM** (**P**roxy **R**isk **I**nference via **S**tructural **M**apping; Fig. 1), a closed-form upper bound on the cross-entropy risk gap between a *target* model (the base checkpoint) and a *proxy* model (a post-training variant).

PRISM exposes three diagnostic axes: *scale mismatch* (activation-magnitude drift), *shape mismatch* (feature-geometry distortion), and a covariance-weighted *head discrepancy* (prediction-head divergence). Unlike a scalar similarity score, these axes correspond to distinct empirical failure modes. Low-bit quantization mainly distorts shape. LoRA fine-tuning drifts mostly in shape as well, but its scale axis adds a task-dependent signal that shape alone misses: on Llama, TruthfulQA fine-tuning inflates activation magnitude while BBQ shrinks it on ARC and MMLU. Quantizing the output projection inflates the head term.

Beyond post-hoc diagnosis, PRISM also provides a training signal. Under frozen-head LoRA fine-tuning the head term vanishes, leaving a differentiable shape term that is a clean regularization target for backbone drift. Penalizing feature-geometry distortion during fine-tuning cuts catastrophic forgetting further than experience replay does on the same reference data (Sec. 4.5).

Our contributions are threefold.

1. **Theory: a closed-form CE risk bound with three diagnostic axes.** The cross-entropy risk gap between a target model and its proxy admits a closed-form upper bound that splits into a scale term and a shape term, obtained from an exact Procrustes residual decomposition, plus a covariance-weighted head-discrepancy term (Theorem 1; Fig. 1).

2. **Framework: a unified diagnostic that doubles as a training objective.** Computed from features and head weights alone, the bound applies across both post-training quantization (PTQ) and frozen-head LoRA; its differentiable shape term further turns it into a training-time regularizer against catastrophic forgetting (Sec. 4.5).

3. **Empirical: rank-consistent, axis-localizing, and robust on little data.** On Llama and Qwen over five benchmarks, the bound ranks PTQ variants at mean Spearman 0.875 and LoRA checkpoints at 0.831, from a single forward pass, with no decoding or accuracy scoring (*predictiveness*); it ranks as well as CKA and SVCCA on a full reference set, and with only eight reference sequences it still reproduces the full-slice ordering, where CKA, SVCCA, and the directly measured risk gap all degrade (*efficiency*); its axes separate failure modes, namely shape distortion at Q2/Q3 PTQ, scale separability under cross-task LoRA drift, and head divergence at Qwen3 Q6_K/Q8_0 (*decomposability*); and the shape regularizer outperforms experience replay in aggregate at mitigating downstream forgetting (*actionability*).

2 RELATED WORK

Representational similarity and Procrustes shape metrics. Representational similarity statistics such as SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), and generalized shape metrics (Williams et al., 2021) quantify geometric resemblance between activation matrices, but tying them to downstream behavior remains open (Ding et al., 2021; Davari et al., 2023; Klabunde et al., 2025). The closest attempt relates shape distance to decodability (Harvey et al., 2024), yet it bounds the distance between *optimal* linear readouts fit to a distribution of decoding tasks, not the loss of the head a model actually ships with. PRISM bounds cross-entropy risk on the deployed head itself, and returns three separately measurable components in place of one scalar (Sec. 3).

Post-training quantization and its evaluation. Post-training quantization spans calibration-based reconstruction (GPTQ (Frantar et al., 2023)) and data-free round-to-nearest schemes (GGUF k-quants, BitsAndBytes (Dettmers et al., 2022)), with low-bit quality often limited by activation outliers (Xiao et al., 2023). The diagnostics available at quantization time are layer-wise: reconstruction error (Frantar et al., 2023) or weight quantization error, computed per block and blind to how that error accumulates through the remaining non-linear layers. PRISM measures the accumulated drift end-to-end on the assembled model, in one forward pass (Sec. 4.2).

Catastrophic forgetting under parameter-efficient fine-tuning. LoRA fine-tuning (Hu et al., 2022) can degrade pre-training knowledge even with a frozen `lm_head`. Remedies constrain either the weights (EWC (Kirkpatrick et al., 2017), L2-SP (Li et al., 2018)) or the data (experience replay); neither constrains the feature geometry that the frozen head actually reads. PRISM adds a third option: the shape term of its bound is differentiable in the current features, so the same quantity that certifies forgetting post hoc can be minimized during training (Sec. 3.4).

Efficient evaluation and performance forecasting. Scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) forecast aggregate loss but not variant-specific degradation, and efficient-evaluation subsets (Polo et al., 2024; Perlitz et al., 2024) preserve variant orderings without explaining them. Both still require the benchmark to be run and scored. PRISM bounds the target-vs-proxy gap from a feature comparison alone, with no decoding and no accuracy scoring (Sec. 3).

3 A GEOMETRIC BOUND ON THE CROSS-ENTROPY RISK GAP

This section derives **PRISM**, a closed-form upper bound on the cross-entropy risk gap $|\mathcal{R}_T - \mathcal{R}_P|$ that decomposes into three diagnostic axes: scale, shape, and head.

3.1 SETUP

Let T be a *target model* (e.g., a reference checkpoint) and P a *proxy model* (e.g., a quantized or fine-tuned variant), sharing hidden dimension d. Each model $M \in \{T, P\}$ factors into a Transformer backbone $\phi_M : \mathcal{X} \to \mathbb{R}^d$ followed by a linear prediction head $H_M \in \mathbb{R}^{d\times V}$ (the `lm_head`, vocabulary size V). The risk of M under data distribution $\mathcal{D}$ is the cross-entropy expectation

$$\mathcal{R}_M \;=\; \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell\big(\phi_M(x)H_M,\, y\big)\big], \qquad \ell(v,y) = -v_y + \log\textstyle\sum_{j=1}^{V} e^{v_j},\; v \in \mathbb{R}^V. \tag{1}$$

All geometry is read off a finite *diagnostic reference set* $\mathcal{D}_{\text{diag}} = \{(x_i, y_i)\}_{i=1}^{N}$: per benchmark, one held-out subset drawn once and reused by every variant of a given base, on which both the geometry below and the risk gap it bounds are computed, so every variant is scored on identical inputs. We write $\widehat{\mathcal{R}}_M = \frac{1}{N}\sum_i \ell(\phi_M(x_i)H_M, y_i)$ for the empirical risk on $\mathcal{D}_{\text{diag}}$, and $|\Delta\mathcal{R}| := |\widehat{\mathcal{R}}_T - \widehat{\mathcal{R}}_P|$ for the empirical risk gap. Stacking the ϕ_M outputs on $\mathcal{D}_{\text{diag}}$ as rows, we form $Z_T, Z_P \in \mathbb{R}^{n\times d}$, where n counts scored tokens, and define the RMS feature scale $\rho_M = \|Z_M\|_F/\sqrt{n}$ and the empirical (uncentered) covariance $\Sigma_P = Z_P^\top Z_P/n$. We restrict the alignment between the two backbones to $W \in \mathcal{O}(d)$ because an orthogonal map leaves distances and angles within Z_P unchanged, which is what lets Proposition 1 split the residual into scale and shape at the original feature scale; a general linear map would absorb both into the alignment. This is a design choice, not an isometry assumption: the fit of W affects only how tight the feature term is, and the measured Ω reports that fit (Sec. 4.4). Prior reports of near-isometry across related representations (Park et al., 2024; Moschella et al., 2023) suggest it is often good.

3.2 THE UNIFIED RISK BOUND

To bound $|\mathcal{R}_T - \mathcal{R}_P|$ we introduce a hybrid risk $\mathcal{R}_{P\to T} := \mathbb{E}[\ell(\phi_P(x)\, W \cdot H_T, y)]$ that evaluates the target's head on the proxy's aligned features. The triangle inequality $|\mathcal{R}_T - \mathcal{R}_P| \le |\mathcal{R}_T - \mathcal{R}_{P\to T}| + |\mathcal{R}_{P\to T} - \mathcal{R}_P|$ splits the gap into two components, each upper-bounded by a closed-form geometric quantity: a *feature alignment error* δ and a *head discrepancy* γ (full derivations in Appendix A).

Feature error δ. The cross-entropy loss is Lipschitz in features with constant $K_{\text{feat}} = \max_{j,k} \|h_{T,j} - h_{T,k}\|_2$, where $h_{T,j}$ is the j-th column of H_T (Appendix A.3; tighter than the spectral bound $\sqrt{2}\|H_T\|_2$, which grows with vocabulary size). The alignment residual $\|Z_T - Z_P W\|_F^2/n$ admits an exact identity for every orthogonal W:

Proposition 1 (Exact Scale–Shape Decomposition)**.** *For any $W \in \mathcal{O}(d)$,*

$$\frac{1}{n}\|Z_T - Z_P W\|_F^2 \;=\; \underbrace{(\rho_T - \rho_P)^2}_{\text{Scale Mismatch } (\Delta\rho)^2} \;+\; \underbrace{2\rho_T\rho_P\big(1 - \Omega_W\big)}_{\text{Shape Mismatch}}, \tag{2}$$

where Ω_W is the trace Procrustes similarity*:*

$$\Omega_W(Z_T, Z_P) \;:=\; \frac{\operatorname{Tr}(Z_T^\top Z_P\, W)}{\|Z_T\|_F\, \|Z_P\|_F} \;\in [-1,\, 1]. \tag{3}$$

The Lipschitz bound then gives $|\mathcal{R}_T - \mathcal{R}_{P\to T}| \le K_{\text{feat}}\sqrt{\frac{1}{n}\|Z_T - Z_P W\|_F^2} =: \delta$, expanded via Eq. (2) in Theorem 1 below.

Head error γ. Let $\Delta H := W H_T - H_P$; we define $\gamma := K_{\text{pred}}\, \|\Sigma_P^{1/2} \Delta H\|_F$ with $K_{\text{pred}} \le \sqrt{2}$ (Appendix A.5), so that $|\mathcal{R}_{P\to T} - \mathcal{R}_P| \le \gamma$. The covariance weighting $\Sigma_P^{1/2}$ ensures that only head misalignment in the *active subspace* (directions where the data has support) contributes to γ.

Theorem 1 (Unified Risk Bound)**.** *For any $W \in \mathcal{O}(d)$,*

$$|\mathcal{R}_T - \mathcal{R}_P| \;\le\; \mathcal{B} \;:=\; \underbrace{K_{\text{feat}}\sqrt{(\rho_T - \rho_P)^2 + 2\rho_T\rho_P\big(1 - \Omega_W\big)}}_{\delta\text{: feature alignment error}} \;+\; \underbrace{K_{\text{pred}}\, \|\Sigma_P^{1/2}(W H_T - H_P)\|_F}_{\gamma\text{: head discrepancy}}. \tag{4}$$

We refer to $\mathcal{B} = \delta + \gamma$ as the PRISM bound*.*

Choice of alignment. Any $W \in \mathcal{O}(d)$ gives a valid bound; we evaluate two. The trace form $\Omega := \Omega_{W=I}$ is the main-text default and the Procrustes-optimal nuclear form $\Omega_N := \Omega_{W=W_N}$, with bound $\mathcal{B}_N$, is in Appendix B. W_N minimizes the alignment residual but requires a per-step SVD and leaves γ nonzero even when $H_T \approx H_P$; $W{=}I$ sends γ to exactly zero whenever $H_T{=}H_P$ (frozen-head LoRA and FP16-head PTQ, though not GGUF k-quant) and is directly differentiable, which enables the regularizer of Sec. 3.4. Both rank variants comparably well (Sec. 4.2). PTQ perturbs weights in place and LoRA adapts a frozen base, so the inherited coordinates, $W{=}I$, are the natural convention.

Empirical form. Every step of the proof is valid with each expectation replaced by a mean over $\mathcal{D}_{\text{diag}}$, so Theorem 1 holds verbatim as $|\Delta\mathcal{R}| \le \widehat{\mathcal{B}}$ on the hatted quantities, which is the form the experiments test. The population gap follows under two explicit assumptions (i.i.d. sampling and an almost-surely bounded per-sample gap), at the cost of an $O(N^{-1/2})$ concentration term, with N counting reference sequences rather than tokens (Corollary 1).

Autoregressive form. The bound applies unchanged to autoregressive next-token prediction over a whole target sequence, scored with teacher forcing. For a sequence (c, y) with context c and target continuation $y = (y_1, \ldots, y_{|y|})$, the autoregressive risk is

$$\mathcal{R}_M^{\text{AR}} \;=\; \mathbb{E}_{(c,y)\sim\mathcal{D}}\left[\frac{1}{|y|}\sum_{\tau=1}^{|y|} \ell\big(\phi_M(c, y_{<\tau}) \cdot H_M,\; y_\tau\big)\right], \tag{5}$$

and teacher forcing extracts its $|y|$ token-level features in one forward pass, stacked into $Z_M^{\text{AR}} \in \mathbb{R}^{|y|\times d}$, to which Theorem 1 applies directly (full corollary in Appendix C).

3.3 THREE DIAGNOSTIC AXES

Theorem 1 splits the bound along three *independently measurable* axes (Fig. 1), each attached to a distinct failure mode of the proxy. **Scale** $\Delta\rho$ isolates divergence in activation magnitude, and its sign is informative: the most aggressive bit-width inflates ρ_P in most cells we measure, whereas fine-tuning moves it in either direction (Sec. 4.4). **Shape** $1-\Omega_W$ is the normalized feature mismatch left once scale is accounted for, measured under the chosen alignment: at $W{=}I$ it includes any change of orientation in the inherited coordinates, and at $W{=}W_N$ it is what survives the best orthogonal fit. **Head** $\|\Sigma_P^{1/2}(WH_T - H_P)\|_F$ measures how differently the two prediction heads read the same features.

3.4 FROM DIAGNOSTIC TO TRAINING: SHAPE REGULARIZATION

Standard LoRA leaves the prediction head frozen, so at $W{=}I$ we have $\gamma{=}0$ and Theorem 1 reduces to backbone drift alone between the base θ_0 and a checkpoint θ_t: $|\mathcal{R}_0 - \mathcal{R}_t| \le \delta$ with $\gamma{=}0$ and $\Omega = \Omega(Z_0, Z_t)$. Since $\Omega(Z_0, Z_t)$ falls as fine-tuning proceeds and $1-\Omega$ is differentiable in Z_t, the shape term is not only a diagnostic but a training signal: penalizing it on a fixed reference set holds the backbone geometry near the base, which is what the bound says forgetting depends on. We therefore add a *shape regularizer* to the objective,

$$L_{\text{total}} \;=\; \underbrace{L_{\text{CE}}\big(\theta_t;\, \mathcal{D}_{\text{FT}}\big)}_{\text{task loss}} \;+\; \lambda\,\big(1 - \Omega\big(Z_0^{\text{ref}},\, Z_t^{\text{ref}}\big)\big), \tag{6}$$

where Z_0^{ref} and Z_t^{ref} are the base and current feature matrices on the *regularizer reference set* $\mathcal{D}_{\text{ref}}$: 32 sequences drawn once from the held-out split of the fine-tuning task, disjoint from $\mathcal{D}_{\text{FT}}$ and from every benchmark on which forgetting is measured, and held fixed for the whole run. Z_0^{ref} is computed once from the frozen base; Z_t^{ref} is recomputed at every optimizer step. The result is insensitive to which sequences are drawn (Appendix I).

4 EXPERIMENTS

We organize the experiments around four claims. **Predictiveness**: the bound tracks $|\Delta\mathcal{R}|$ in rank order across both PTQ and LoRA variants (Sec. 4.2). **Efficiency**: one forward pass and a handful of reference sequences suffice, and the bound adds what the risk gap from the same pass cannot (Sec. 4.3). **Decomposability**: its three axes localize three distinct failure modes (Sec. 4.4). **Actionability**: the differentiable shape term doubles as a training-time regularizer that suppresses forgetting (Sec. 4.5).

4.1 EXPERIMENTAL SETUP

Models. The main analysis spans two model families: Llama-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B-Base (Qwen Team, 2025) (written Qwen3-8B below); PTQ replications on Ministral-3-8B (Liu et al., 2026), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), and three instruction-tuned counterparts (seven 8B families total) are in Appendix E.3.

Quantization protocols. The target T is the base model and the proxy P a quantized variant from one of three PTQ families across bit-widths 2–8: **GGUF** (round-to-nearest, `Q2_K`–`Q8_0`), **GPTQ** (Frantar et al., 2023) (4-bit, second-order reconstruction), and **BitsAndBytes** (Dettmers et al., 2022) (INT8, NF4, FP4). Only GGUF quantizes `lm_head` as well, so γ vanishes for GPTQ and BitsAndBytes.

Fine-tuning tasks. LoRA fine-tunes on **TruthfulQA** (Lin et al., 2022) (factual grounding) and **BBQ** (Parrish et al., 2022) (Bias Benchmark for QA, social-context reasoning). The target $T{=}\theta_0$ is the frozen base and the proxy $P{=}\theta_t$ a LoRA checkpoint; downstream forgetting is measured on benchmarks disjoint from both fine-tuning tasks, so that $|\Delta\mathcal{R}|$ isolates forgetting.

Benchmarks and scoring. Five benchmarks: **MMLU** (Hendrycks et al., 2021), **ARC** (Clark et al., 2018) (multiple-choice knowledge), **TriviaQA** (Joshi et al., 2017), **SQuAD** (Rajpurkar et al., 2016) (short-horizon QA), and **GSM8K** (Cobbe et al., 2021) (multi-step reasoning). $\mathcal{D}_{\text{diag}}$ holds 512 held-out examples per benchmark for PTQ and 256 for LoRA forgetting. All risks are computed

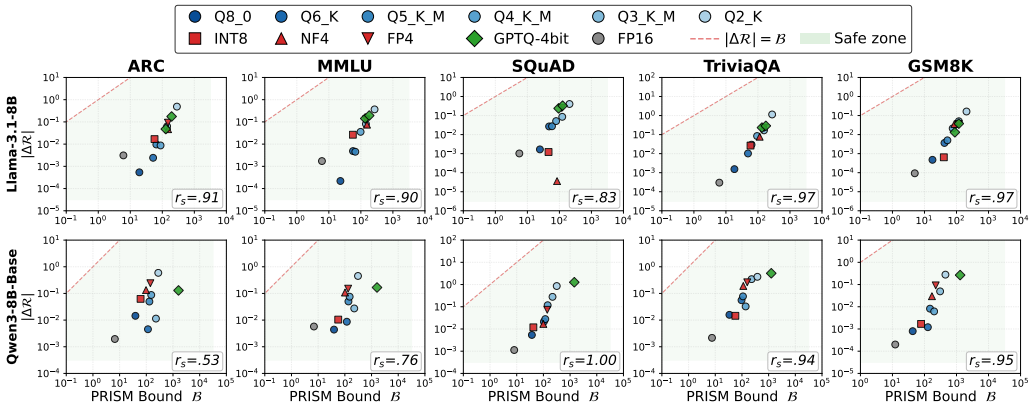

Figure 2: **The PRISM bound $\mathcal{B}$ tracks the empirical risk gap across two model families and five benchmarks.** Each subplot scatters the PRISM bound $\mathcal{B}$ (x-axis, log) against the empirical cross-entropy risk gap $|\Delta\mathcal{R}|$ (y-axis, log). Each point is one quantization variant; colors denote PTQ family (GGUF / GPTQ / BitsAndBytes). Rows: Llama-3.1-8B, Qwen3-8B. Columns: ARC, MMLU, SQuAD, TriviaQA, GSM8K. Spearman r_s is annotated per panel (mean r_s=0.875 across the 2×5 grid); the red dashed $y{=}x$ line marks bound tightness, and the green shaded region below it is where the measured gap respects the bound.

Table 1: Geometric decomposition for **Llama-3.1-8B** under identity alignment ($W{=}I$) on MMLU. The first column reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	138.95	138.97	0.9999	6.0918	0	6.0918	0.0017
($\mathbf{r_s{=}0.90}$)	GGUF	Q8_0	138.95	139.17	0.9989	17.0010	5.9609	22.9620	0.0002
		Q6_K	138.95	139.82	0.9946	37.9361	18.7745	56.7106	0.0049
		Q5_K_M	138.95	139.93	0.9913	48.2559	18.7916	67.0475	0.0045
		Q4_K_M	138.95	140.35	0.9765	79.1674	18.8939	98.0613	0.0356
		Q3_K_M	138.95	140.32	0.9414	124.9549	18.8431	143.7980	0.0808
		Q2_K	138.95	143.86	0.7751	248.1663	19.5549	267.7212	0.3658
	BnB	INT8	138.95	139.09	0.9881	56.1660	0	56.1660	0.0265
		NF4	138.95	144.15	0.9125	155.3570	0	155.3570	0.0750
		FP4	138.95	138.10	0.9197	145.0905	0	145.0905	0.1306
	GPTQ	GPTQ-4bit	138.95	140.36	0.9299	136.6923	0	136.6923	0.1422

teacher-forced: prompt c and targets y are scored in a single forward pass over the gold span, giving a deterministic per-sample CE loss.

Hyperparameters. LoRA (rank 32) fine-tunes with frozen head (AdamW, lr 10^{-5}, batch size 16, bf16). The shape regularizer sweeps $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$ and the replay-CE baseline sweeps $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, each range matched to its penalty's natural scale ($1{-}\Omega \sim 0.1$ vs. CE reference loss ~ 1 nat). Checkpoints every 25 steps, analysis at step 300; all experiments use a single NVIDIA RTX 5090 (32 GB).

4.2 Predictiveness: The Bound Tracks the Risk Gap

Quantization. The PRISM bound $\mathcal{B}$ tracks $|\Delta\mathcal{R}|$ with strong rank correlation across Llama and Qwen (mean Spearman $r_s = 0.875 \pm 0.0425$ SEM over the 2×5 grid; Fig. 2). Bit-width drives the bound monotonically, Q8/Q6 < Q5/Q4 < Q3/Q2 in both families, with Q2 in the upper-right corner where shape distortion dominates (Table 1). Every evaluated variant lies below the $y{=}x$ line, as Theorem 1 requires. Per-benchmark tables, the feature-only (δ) scatter, and the Ministral and DeepSeek replications are in Appendix E.

LoRA forgetting. The same holds for fine-tuning drift (Fig. 3): across LoRA checkpoints on TruthfulQA or BBQ, the bound computed on *downstream* benchmarks tracks $|\Delta\mathcal{R}|$ at mean Spearman $r_s = 0.831 \pm 0.0722$ over the 2×5 cells, comparable to the PTQ grid, and at 0.958 on the TruthfulQA row alone. The Qwen3-8B replication (Appendix H) reproduces this in 8 of 10 cells; the two

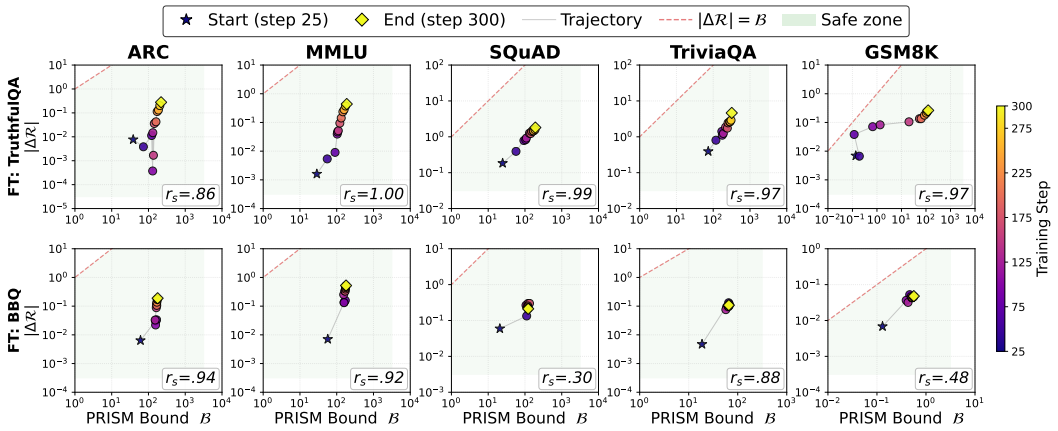

Figure 3: **Llama-3.1-8B: the PRISM bound tracks catastrophic forgetting across LoRA fine-tuning steps.** Each subplot scatters the bound $\mathcal{B}$ (x-axis, log) against the empirical forgetting $|\Delta\mathcal{R}|$ (y-axis, log) on a downstream benchmark, with one point per LoRA checkpoint colored by training step. Rows: fine-tuning task (TruthfulQA, BBQ). Columns: downstream benchmark (ARC, MMLU, SQuAD, TriviaQA, GSM8K). Under LoRA's frozen `lm_head`, $\gamma{=}0$ at $W{=}I$, so $\mathcal{B}$ reduces to backbone scale ($\Delta\rho$) and shape ($1{-}\Omega$) drift. Spearman r_s is annotated per panel.

exceptions, BBQ fine-tuning evaluated on MMLU and TriviaQA, leave nothing to rank: TriviaQA's forgetting stays near 10^{-3} nats per token and MMLU's plateaus by step 75, so later checkpoints differ by noise.

The bound is loose, and why ranking survives it. $\mathcal{B}$ exceeds $|\Delta\mathcal{R}|$ by a large factor (Table 1), most for the variants that drift least. Two steps of the proof account for this. The Lipschitz step charges every token the worst-case sensitivity K_{feat}, although the realized gradient $\|H_T(\hat{p} - e_y)\|_2$ is near zero on confidently and correctly predicted tokens. Moving the absolute value inside the expectation then bounds the absolute per-token change, whereas signed changes partly cancel in the mean: when perturbations have a negligible mean first-order contribution, the mean loss change is second order, which helps explain why the gap shrinks faster than the residual as drift decreases. These sources of looseness neither establish nor rule out ranking fidelity; that is an empirical question. Within a cell K_{feat} and ρ_T are fixed, so the feature term is K_{feat} times the RMS alignment residual $\|Z_T - Z_P W\|_F/\sqrt{n}$, and the question is whether $|\Delta\mathcal{R}|$ increases with this residual, which the feature-only scatter (Appendix E.2) confirms cell by cell, and Fig. 2 with the head term included. We therefore use $\mathcal{B}$ to rank variants and diagnose their drift, rather than as a tight estimate of the risk gap.

PRISM ranks on par with CKA and SVCCA, and adds what they cannot. Against $1{-}$CKA (Kornblith et al., 2019) and $1{-}$SVCCA (Raghu et al., 2017) on the same ten cells, every PRISM quantity, down to the shape term $1{-}\Omega_N$ alone (0.898), lands in the same band: mean r_s of 0.875–0.906, with $\mathcal{B}_N$ best at 0.906 and both baselines at 0.899 (Table 16, Appendix F). The per-metric SEM is 0.02–0.05 and no pairwise difference survives multiplicity correction, so ranking quality does not separate these metrics. What separates PRISM is what comes with the score: an exact split into scale, shape and head (Sec. 4.4), a shape term that can be trained on (Sec. 4.5), and stability on reference sets too small for these statistics or for the risk gap itself (Sec. 4.3).

4.3 Efficiency: one forward pass and eight reference sequences

One pass, no decoding. PRISM needs one teacher-forced forward pass per benchmark, with no decoding and no grading. On 256 matched GSM8K prompts that pass takes 8.9 s where greedy decoding takes 556.7 s, 62.6× more, and the gap is structural: scoring is one pass parallel over the sequence, generation is one pass per token (Appendix G). The saving is against generative benchmarks; likelihood-scored ones such as MMLU already cost one pass.

The risk gap comes from the same pass, so what does the bound add? The pass that yields Z_P also yields $\widehat{\mathcal{R}}_P$, so on a full reference set $|\Delta\mathcal{R}|$ is free, and Sec. 4.2 is a check that the bound tracks what it bounds, not a cheaper substitute for measuring it. What the pass does not give without

PRISM is the attribution of the gap to scale, shape and head (Sec. 4.4), a differentiable surrogate for it (Sec. 4.5), and, as we show next, an ordering that survives a small reference set.

Eight sequences fix the ordering; the risk gap itself needs over a hundred. In an independent round on Llama-3.1-8B (Appendix F.2) we shrink $\mathcal{D}_{\text{diag}}$ from 512 to 8 sequences per benchmark and ask how well each quantity computed on the small slice recovers the ordering that $|\Delta\mathcal{R}|$ on the full 512 gives. The bound does not move: $\mathcal{B}_N$ scores $r_s = 0.932$ at both sizes, and the $W{=}I$ feature term δ scores 0.930 and 0.925. The risk gap measured on the same 8 sequences reaches only 0.795 and needs 128 to match what the bound gives with 8; $1-$CKA drops to 0.903; $1-$SVCCA is degenerate in this regime, since canonical correlations need more samples than dimensions, and Table 17 reports it for completeness. Plausibly, the bound reads $d{=}4096$ feature coordinates per scored token where the gap reads one loss value. Qwen3-8B replicates the pattern (0.827 against 0.755 at $n{=}8$).

4.4 Decomposability: three axes, three failure modes

Ranking says *whether* a variant degraded; the decomposition says *which* axis moved, read directly from the per-variant numbers.

Shape distortion (low-bit PTQ). Low-bit PTQ leaves ρ almost untouched and moves Ω instead. At Llama-Q2_K on MMLU (Table 1), $\Delta\rho \approx 4.9$ puts the scale term near 24, yet Ω drops to 0.78 and the shape term $2\rho_T\rho_P(1-\Omega)$ reaches $\approx 9{,}000$; Ministral-Q2_K on SQuAD shows the same (~ 3.5 vs $\sim 9{,}000$), and at Q2 and Q3 shape dominates scale across all four families tested, often by orders of magnitude (Appendix E.3). LoRA forgetting shows the same pattern: at Llama TruthfulQA-FT on TriviaQA (no-reg; Table 18), Ω drops to 0.76 and shape ($\approx 10{,}000$) outweighs scale (≈ 35) by $\sim 280\times$, which motivates the shape regularizer of Sec. 4.5.

Scale-axis separability. The scale axis matters where Ω holds but ρ does not, or where its sign separates variants that shape alone would not. In PTQ, Qwen3-8B Q2_K on GSM8K moves ρ_P from 273 to 318 ($|\Delta\rho|{=}45$ against below 7.5 for every other Qwen3 variant; Table 10). In LoRA forgetting on Llama the sign depends on the source task: TruthfulQA-FT drives $\rho_P > \rho_T$ on all five benchmarks while BBQ-FT drives $\rho_P < \rho_T$ on ARC and MMLU; on Qwen3-8B both tasks shrink ρ on nine of ten cells, so the sign is model-dependent as well. Magnitude also matters: on Qwen3-8B fine-tuned on BBQ, ρ falls by 11.0 on SQuAD while Ω holds at 0.9985, putting the scale term at roughly half the shape term, against $\sim 1/280$ for Llama TruthfulQA above (Tables 18–21).

Head divergence (GGUF k-quant tiers that quantize `lm_head`). At Qwen3-8B Q6_K on SQuAD (Table 10) the backbone is nearly intact, $(\Delta\rho)^2 \approx 0.18$ and $\Omega{=}0.9997$, yet the bound stays high: the feature term is only $\delta{=}26.83$ while the quantized output embedding contributes $\gamma{=}77.42$, three quarters of the bound (Q8_0: $\delta{=}12.49$, $\gamma{=}24.48$). BnB INT8 on the same model keeps $\gamma{\equiv}0$ by construction and its whole bound is $\delta{=}42.06$, $2.5\times$ below Q6_K's $\mathcal{B}{=}104.25$; the difference is entirely which protocol quantizes `lm_head`.

From diagnosis to remediation. Each axis suggests a different fix. A scale mismatch points to activation outliers, which per-channel scaling (Xiao et al., 2023) or keeping outlier channels in higher precision (Dettmers et al., 2022) addresses. A dominant head term says to quantize `lm_head` more gently, or to leave it unquantized, which sets γ to zero by construction. Shape distortion calls for more bits, or mixed precision, on the layers that introduce it at quantization time, and for the shape regularizer of Sec. 4.5 at fine-tuning time. Only the last is tested in this paper (Sec. 5).

4.5 Actionability: from diagnostic to training objective

We compare the shape regularizer of Eq. 6 against replay-CE on Llama-3.1-8B fine-tuned on TruthfulQA, with no regularization as the anchor (Fig. 4, Table 2). Both use the same reference set and cadence, so replay separates shape preservation from ordinary data re-fitting, and each is reported at the best λ in its sweep (1.0 for the shape regularizer, 0.01 for replay). The shape regularizer cuts mean downstream $|\Delta\mathcal{R}|$ from 0.843 to 0.681 (-19%) where replay reaches only 0.764 (-9%), and it lifts mean Ω from 0.906 to 0.931 against replay's 0.915. This is consistent with the objectives: replay re-fits reference data, whereas the shape regularizer directly targets the shape term. The effect is regime-dependent. On the low-drift settings, Llama BBQ and both Qwen3-8B tasks, where baseline Ω is already at or above 0.93, the regularizer moves mean $|\Delta\mathcal{R}|$ by less than 10% in either direction;

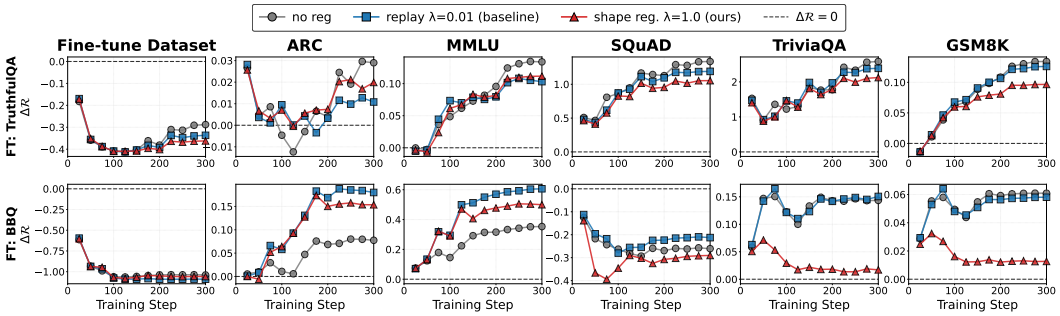

Figure 4: **Shape regularization vs. replay-CE on Llama-3.1-8B.** Signed CE change $\Delta\mathcal{R}$ of each LoRA checkpoint relative to the frozen base; first column: the fine-tuning task (negative = learning), other columns: downstream benchmarks (positive = forgetting). Three configurations: *no reg* (anchor), the *replay* baseline, and *our shape regularizer* (Eq. 6); the latter two share a 32-sample reference set; each is shown at the best λ in its sweep (1.0 and 0.01). Per-benchmark values for TruthfulQA: Table 2; BBQ analysis: Appendix J; Qwen3-8B replication: Appendix H.

Table 2: Regularization comparison for **Llama-3.1-8B** fine-tuned on **TruthfulQA** (identity alignment $W{=}I$; metrics at step 300). Rows: Ω (higher = more shape preserved) and $|\Delta\mathcal{R}|$ (lower = less forgetting). **Bold** / underline: 1st / 2nd-best per benchmark group. Columns per benchmark: no regularization, replay (λ=0.01), and our shape regularizer (λ=1.0); across-benchmark means are given in the text. In this frozen-`lm_head` LoRA setting $\gamma \equiv 0$ and $\rho_T \approx \rho_P$, so δ and $\mathcal{B}$ track Ω; the full decomposition is in Appendix Tables 18–21.

	ARC			MMLU			SQuAD			TriviaQA			GSM8K		
	no reg	**replay**	**shape reg**	**no reg**	**replay**	**shape reg**	**no reg**	**replay**	**shape reg**	**no reg**	**replay**	**shape reg**	**no reg**	**replay**	**shape reg**
$\Omega\uparrow$	0.916	0.918	**0.932**	0.943	0.945	**0.954**	0.923	0.935	**0.948**	0.759	0.781	**0.821**	0.991	0.996	**1.000**
$\lvert\Delta\mathcal{R}\rvert\downarrow$	0.029	**0.011**	0.020	0.134	**0.103**	0.112	1.337	1.191	**1.054**	2.583	2.388	**2.124**	0.134	0.126	**0.097**

on Llama BBQ it ends slightly above no regularization: it cuts forgetting on TriviaQA and GSM8K but adds some on ARC and MMLU, where shape drift had not turned into forgetting (Appendix J).

Replay is the natural baseline because it sees the same reference data as the shape regularizer, but not the strictest test, so Appendix I also compares EWC (Kirkpatrick et al., 2017), L2-SP (Li et al., 2018) and layer freezing, each at the sweep configuration whose target-task loss matches the shape-regularized run's. In this separate round the shape regularizer forgets least among all methods that learn the task to a comparable level: mean downstream $|\Delta\mathcal{R}|$ of 0.680 against 0.751 for EWC and 0.763 for L2-SP. It also has the lowest target-task loss of every method in that table, so the reduction does not come at the expense of plasticity.

5 Discussion

Screening, diagnosis, intervention. PRISM is not a substitute for evaluation. Benchmarks measure what a variant can do, and the teacher-forced risk gap measures its change in cross-entropy relative to the base; PRISM supports decisions about what to evaluate next. From one forward pass, it helps practitioners prioritize variants for benchmarking, identify what changed in their representations and heads, and constrain shape drift during training. Its closed-form bound and exact decomposition into scale, shape, and head connect screening, diagnosis, and intervention (Secs. 4.2–4.5).

Scope and limitations. Three limitations qualify these claims. First, the bound provides an upper limit on the risk gap, but our empirical validation supports its use for ranking rather than tight absolute estimation. Second, all features are extracted under teacher forcing, which needs no decoding and is the default we recommend. The theorem also covers trajectories the proxy generates itself, and on GSM8K rollouts the ranking is essentially unchanged ($r_s = 0.947$ against 0.958 teacher-forced; Appendix C). Third, only the shape axis has an implemented mitigation, the regularizer of Sec. 4.5; the scale and head remediations that Sec. 4.4 points to follow from the diagnosis but remain untested.

Future work. We describe future directions in detail in Appendix K, including **beyond LoRA forgetting** (analyze and regularize SFT/distillation drift), **diagnostic applications** (OOD detection, hyperparameter transfer, drift monitoring), and **beyond LLMs** (ViT, contrastive image encoders).

AI USE STATEMENT

We used generative AI tools to implement methods and to clean and reformat data: specifically, to write the scripts that run the experiments, evaluate the models, and aggregate results, and to prepare benchmark subsets and result tables. The PRISM bound, its decomposition, and the training objective were designed and implemented by the authors. We did not use generative AI tools to generate synthetic data, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or give feedback on the methodology or experiments, or interpret results; translation and qualitative data analysis do not apply to this work. We also used generative AI tools to edit the language of author-written text and to produce plotting code. All AI-assisted work was reviewed: generated code was tested for correctness by the authors, and edited text was checked against the underlying results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a diagnostic for comparing post-training variants of publicly released language models. It releases no new model or dataset, involves no human subjects or personal data, and uses every checkpoint, quantized variant, and benchmark under its published license (Sec. 4.1, Appendix D). BBQ and TruthfulQA serve only as fine-tuning tasks for measuring representational drift; we make no claims about social groups or about the truthfulness of any model. PRISM is a diagnostic applied to existing models and introduces no misuse pathway beyond those of the underlying LLMs; its limitations (Sec. 5) mean it should not be read as a safety certificate.

REPRODUCIBILITY STATEMENT

Complete proofs of Proposition 1 and Theorem 1, including the constants K_{feat} and K_{pred}, are in Appendix A. Section 4.1 specifies the models, quantization tiers, fine-tuning tasks, benchmarks, hyperparameters, and reference-set sizes; Appendix D lists the exact public checkpoint of every quantized variant, and Appendices E and H tabulate the per-variant decompositions behind every figure. Teacher-forced scoring on fixed held-out subsets is deterministic, so all reported bounds and risk gaps are recomputable. All experiments ran on a single NVIDIA RTX 5090 with PyTorch 2.10 and `transformers` 5.2. Code for the bound, the quantization sweep, and LoRA training with the shape regularizer and the replay baseline is provided as anonymized supplementary material with pinned dependencies and will be released on acceptance.

REFERENCES

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332. Curran Associates, Inc., 2022.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

John C Gower and Garmt B Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sarah E Harvey, David Lipshutz, and Alex H Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 (ICLR)*, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2825–2834, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.

Alexander H Liu et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2519–2536, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.139.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning (ICML)*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264.

Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott W Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *International Conference on Machine Learning (ICML)*, pp. 38087–38099. PMLR, 2023.

APPENDIX

Table of Contents

A Proof of the unified risk bound (Theorem 1) 13
- A.1 Relation to classical Procrustes shape metrics . . . 14
- A.2 Step 1: risk decomposition via triangle inequality . . . 14
- A.3 Step 2: bounding the feature alignment error δ . . . 14
- A.4 Step 3: exact geometric decomposition . . . 15
- A.5 Step 4: bounding the head discrepancy γ . . . 17
- A.6 Step 5: assembling the unified bound . . . 17
- A.7 Step 6: empirical form and the population corollary . . . 18

B Choosing the orthogonal alignment 18
- B.1 The bound as a function of W . . . 18
- B.2 The trace specialization ($W = I$) . . . 19
- B.3 Joint optimization over W . . . 19

C Formal extension to autoregressive generation 19

D Model and quantization details 20
- D.1 Target models . . . 20
- D.2 Quantization backends and bit-widths . . . 20
- D.3 Backend coverage and GPTQ checkpoints . . . 20
- D.4 GGUF repositories . . . 21

E Quantization replication, feature-only scatter, and per-model tables 21
- E.1 Replication on Ministral and DeepSeek . . . 21
- E.2 Feature-only scatter . . . 21
- E.3 Per-model decomposition tables . . . 22

F Similarity baselines, component ablation, and how small the reference set can be 30
- F.1 Full-size comparison . . . 30
- F.2 How the comparison changes with the size of the reference set . . . 32

G Teacher forcing versus decoding: the cost of one screening pass 33

H Qwen3-8B forgetting and shape-regularization replications 33
- H.1 Forgetting: bound tracks cross-task drift on Qwen3-8B . . . 33
- H.2 Shape regularization: Qwen3-8B replication . . . 34
- H.3 Additional decomposition tables . . . 35

I Additional regularization baselines: EWC, L2-SP and layer freezing 35

J When shape regularization helps: task dependence 37

K Future work 38

A PROOF OF THE UNIFIED RISK BOUND (THEOREM 1)

We provide the complete proof of Theorem 1, which the main text states for general $W \in \mathcal{O}(d)$ (trace family Ω_W, Eq. 3). The proof proceeds in six steps: risk decomposition via the triangle inequality, Lipschitz analysis of cross-entropy in the features, the alignment-residual decomposition (whose Procrustes-optimal specialization $W = W_N$ yields the nuclear form Ω_N via SVD), the covariance-adjusted head bound, the final assembly, and the empirical form with its population corollary. The identity-alignment specialization $W = I$ (used throughout Sec. 4 and the regularizer of Sec. 3.4) and the $W = W_N$ specialization (used in the Sec. 4.2 ablation) follow as corollaries of the same proof.

A.1 Relation to classical Procrustes shape metrics

At $W = W_N$, the left-hand side of Eq. (2) is the squared Procrustes size-and-shape distance of the generalized shape metrics framework (Williams et al., 2021; Harvey et al., 2024). Our contribution is not that distance but its exact split into a scale term and a shape term, which that literature treats as one descriptive dissimilarity, and its lifting to a bound on cross-entropy risk (Theorem 1).

A.2 Step 1: risk decomposition via triangle inequality

Setup. Recall that the risk of model M is $\mathcal{R}_M = \mathbb{E}_{(x,y)}[\ell(\phi_M(x) \cdot H_M, y)]$, where ℓ is the cross-entropy loss (Eq. 1). For any orthogonal alignment $W \in \mathcal{O}(d)$, introduce the hybrid risk:

$$\mathcal{R}_{P\to T} := \mathbb{E}_{(x,y)}\Big[\ell\big(\phi_P(x)\, W \cdot H_T,\ y\big)\Big]. \tag{7}$$

This evaluates the target's head H_T on proxy features aligned by W, serving as a bridge between $\mathcal{R}_T$ and $\mathcal{R}_P$. The proof below holds for any $W \in \mathcal{O}(d)$; the two specializations relevant to the paper are $W = I$ (identity alignment, trace form Ω, used throughout the main text) and $W = W_N := \arg\min_W \|Z_T - Z_P W\|_F^2$ (Procrustes-optimal, nuclear form Ω_N, Sec. 4.2 ablation).

Decomposition. By the triangle inequality on absolute values:

$$|\mathcal{R}_T - \mathcal{R}_P| \ \le\ \underbrace{\big|\mathcal{R}_T - \mathcal{R}_{P\to T}\big|}_{\text{Feature Alignment Error } \delta} \ +\ \underbrace{\big|\mathcal{R}_{P\to T} - \mathcal{R}_P\big|}_{\text{Head Discrepancy } \gamma}. \tag{8}$$

δ measures how features differ through the same head H_T; γ measures how heads differ on the same aligned features.

A.3 Step 2: bounding the feature alignment error δ

Lipschitz property of cross-entropy with respect to features. We bound δ by showing that the cross-entropy loss is Lipschitz continuous in the feature vector z. Define the end-to-end composition $g_T(z, y) := \ell(z \cdot H_T, y)$, treating the classification head H_T and the cross-entropy loss as a single map from features to scalar loss. We bound g_T via direct gradient analysis rather than the Lipschitz composition rule (loss-Lipschitz $\sqrt{2}$ times head operator norm $\|H_T\|_2$): the composition rule is agnostic to the softmax output and gives the spectral constant $\sqrt{2}\|H_T\|_2$ (Remark below), whereas the gradient $\nabla_z g_T = H_T(\hat{p} - e_y)$ exposes the predicted distribution $\hat{p}$ and lets us derive a constant set by pairwise token-embedding distances instead. Our goal is to find the tightest constant K_{feat} such that:

$$|g_T(z_1, y) - g_T(z_2, y)| \le K_{\text{feat}} \|z_1 - z_2\|_2, \quad \forall z_1, z_2, y. \tag{9}$$

Computing the gradient. The logit vector is $v = z \cdot H_T \in \mathbb{R}^V$, and the cross-entropy loss is $\ell(v, y) = -v_y + \log \sum_{j=1}^V e^{v_j}$. The gradient with respect to the logits is:

$$\frac{\partial \ell}{\partial v} = \hat{p} - e_y, \tag{10}$$

where $\hat{p} = \mathrm{softmax}(v) \in \mathbb{R}^V$ and e_y is the one-hot vector at index y. Applying the chain rule to obtain the gradient with respect to features:

$$\nabla_z \ell = H_T(\hat{p} - e_y) = \sum_{j=1}^{V} (\hat{p}_j - \delta_{jy})\, h_{T,j}, \tag{11}$$

where $h_{T,j}$ denotes the j-th column of H_T (the embedding vector for token j).

Simplex polarization. We now derive a tight bound on $\|\nabla_z \ell\|_2$ by exploiting the probability simplex constraint $\sum_j \hat{p}_j = 1$. Expanding the gradient for true class y:

$$\nabla_z \ell = (\hat{p}_y - 1)\, h_{T,y} + \sum_{j \ne y} \hat{p}_j\, h_{T,j}. \tag{12}$$

Since $1 - \hat{p}_y = \sum_{j \ne y} \hat{p}_j$, we can rewrite:

$$\nabla_z \ell = -\Big(\sum_{j \ne y} \hat{p}_j\Big) h_{T,y} + \sum_{j \ne y} \hat{p}_j\, h_{T,j} = \sum_{j \ne y} \hat{p}_j\, (h_{T,j} - h_{T,y}). \tag{13}$$

This reveals that the gradient is a *non-negative linear combination* of pairwise differences between incorrect-class embeddings and the correct-class embedding, with weights $\hat{p}_j$ summing to $1 - \hat{p}_y \leq 1$. Taking the norm:

$$\|\nabla_z \ell\|_2 \leq \sum_{j \neq y} \hat{p}_j \|h_{T,j} - h_{T,y}\|_2 \leq \max_{j \neq y} \|h_{T,j} - h_{T,y}\|_2 \cdot \underbrace{\sum_{j \neq y} \hat{p}_j}_{\leq 1}. \tag{14}$$

The bound depends on the true class y only through $\max_{j \neq y} \|h_{T,j} - h_{T,y}\|_2$. Taking the supremum over $y \in \{1, \ldots, V\}$ extends the maximum to all ordered pairs (j, y) with $j \neq y$ (and the $j = y$ case contributes 0), giving the full pairwise diameter:

$$K_{\text{feat}} = \max_{j,k} \|h_{T,j} - h_{T,k}\|_2. \tag{15}$$

Remark. This bound depends on the *relative distances* between token embeddings, not their absolute magnitudes. A uniform shift $H_T \to H_T + c\mathbf{1}^\top$ does not change K_{feat}. A naive Cauchy–Schwarz bound gives $K_{\text{feat}}^{\text{naive}} = \sqrt{2}\|H_T\|_2$, which is substantially looser. Since the decomposition of Eq. (8) routes through $\mathcal{R}_{P \to T}$ (which uses H_T), K_{feat} is determined by H_T alone; it takes a single value across all proxy variants of a fixed target model and enters as a constant scaling on the feature term, not as a proxy-dependent quantity that could perturb variant ranking.

From gradient bound to Lipschitz inequality. Since $g_T(\cdot, y)$ is continuously differentiable, the gradient bound $\|\nabla_z g_T(z, y)\|_2 \leq K_{\text{feat}}$ implies the Lipschitz inequality stated above by the mean value theorem: for any z_1, z_2,

$$g_T(z_2, y) - g_T(z_1, y) = \int_0^1 \nabla_z g_T\big(z_1 + t(z_2 - z_1),\, y\big)^\top (z_2 - z_1)\, dt,$$

and Cauchy–Schwarz gives $|g_T(z_2, y) - g_T(z_1, y)| \leq K_{\text{feat}}\, \|z_2 - z_1\|_2$.

Applying the Lipschitz bound. Using the Lipschitz property with $z_1 = \phi_T(x)$ and $z_2 = \phi_P(x)W$:

$$\begin{aligned} \delta = \big|\mathcal{R}_T - \mathcal{R}_{P \to T}\big| &= \Big|\mathbb{E}\big[g_T(\phi_T(x), y) - g_T(\phi_P(x)W, y)\big]\Big| \\ &\leq \mathbb{E}\Big[\big|g_T(\phi_T(x), y) - g_T(\phi_P(x)W, y)\big|\Big] \\ &\leq K_{\text{feat}} \cdot \mathbb{E}_x\big[\|\phi_T(x) - \phi_P(x)W\|_2\big]. \end{aligned} \tag{16}$$

Applying Jensen's inequality ($\mathbb{E}[\|X\|] \leq \sqrt{\mathbb{E}[\|X\|^2]}$) to connect this to the alignment residual:

$$\delta \leq K_{\text{feat}} \cdot \sqrt{\mathbb{E}_x\big[\|\phi_T(x) - \phi_P(x)W\|_2^2\big]} = K_{\text{feat}} \cdot \sqrt{\frac{1}{n}\|Z_T - Z_P W\|_F^2}. \tag{17}$$

Empirical values. Table 3 reports K_{feat} for all 8B models studied here, ranging from 0.93 (Mistral family) to 3.46 (Qwen3-8B-Base); $K_{\text{pred}} = \sqrt{2}$ holds universally. These values are bounded by pairwise token-embedding distances and remain independent of V (Llama V=128,256, Qwen3 V=151,936), in contrast to the naive bound $\sqrt{2}\|H_T\|_2$, whose spectral norm grows with vocabulary size: it equals 133.8 on Llama-3.1-8B and 220.9 on Qwen3-8B-Base, 51 and 64 times K_{feat} respectively.

A.4 Step 3: exact geometric decomposition

We derive the exact identity linking the alignment residual to scale and shape mismatches. The identity holds for every $W \in \mathcal{O}(d)$; we then specialize to the two cases relevant to this paper.

General identity. For any $W \in \mathcal{O}(d)$, expand the squared Frobenius norm:

$$\begin{aligned} \|Z_T - Z_P W\|_F^2 &= \operatorname{Tr}\big[(Z_T - Z_P W)^\top (Z_T - Z_P W)\big] \\ &= \|Z_T\|_F^2 + \|Z_P W\|_F^2 - 2\operatorname{Tr}(Z_T^\top Z_P W). \end{aligned} \tag{18}$$

Since W is orthogonal, $\|Z_P W\|_F^2 = \operatorname{Tr}(W^\top Z_P^\top Z_P W) = \operatorname{Tr}(Z_P^\top Z_P) = \|Z_P\|_F^2$ (cyclic trace + $W^\top W = I_d$). Therefore:

$$\|Z_T - Z_P W\|_F^2 = \|Z_T\|_F^2 + \|Z_P\|_F^2 - 2\operatorname{Tr}(Z_T^\top Z_P W). \tag{19}$$

Table 3: Empirical Lipschitz constants of the PRISM bound (Theorem 1) for each evaluated model. $K_{\text{pred}} = \sqrt{2}$ holds universally (Appendix A.5, simplex polarization), independent of model. K_{feat} is empirical per model, depending on the spread of token embeddings in H_T; for the 8B-scale models studied here, K_{feat} ranges from ≈ 0.93 (Mistral family) to ≈ 3.46 (Qwen3-8B-Base). Since K_{feat} is a constant per model, it scales the bound's magnitude but does not affect the within-model rank correlation that PRISM is used for.

Model	K_{feat}	K_{pred}
Llama-3.1-8B	2.61	$\sqrt{2}$
Ministral-3-8B	0.98	$\sqrt{2}$
Qwen3-8B	3.46	$\sqrt{2}$
DeepSeek-R1-8B	2.60	$\sqrt{2}$
Llama-3.1-8B-Instruct	2.60	$\sqrt{2}$
Ministral-3-8B-Instruct	0.93	$\sqrt{2}$
Qwen3-8B-Instruct	3.41	$\sqrt{2}$

Recall $\rho_M = \|Z_M\|_F/\sqrt{n}$ (so $\|Z_M\|_F^2/n = \rho_M^2$) and $\Omega_W := \text{Tr}(Z_T^\top Z_P W)/(\|Z_T\|_F\|Z_P\|_F)$ (main-text Eq. 3). Dividing Eq. (19) by n:

$$\frac{1}{n}\|Z_T - Z_P W\|_F^2 = \rho_T^2 + \rho_P^2 - 2\rho_T\rho_P\,\Omega_W. \tag{20}$$

Adding and subtracting $2\rho_T\rho_P$ completes the square:

$$\frac{1}{n}\|Z_T - Z_P W\|_F^2 = \underbrace{(\rho_T - \rho_P)^2}_{\text{Scale Mismatch}} + \underbrace{2\rho_T\rho_P(1 - \Omega_W)}_{\text{Shape Mismatch}}. \tag{21}$$

This is Proposition 1 in its general form. □

Specialization 1: $W = I$ **(identity alignment, main text).** Setting $W = I$ gives the trace form $\Omega \equiv \Omega_{W=I} = \text{Tr}(Z_T^\top Z_P)/(\|Z_T\|_F\|Z_P\|_F)$, and Eq. (21) becomes

$$\frac{1}{n}\|Z_T - Z_P\|_F^2 = (\rho_T - \rho_P)^2 + 2\rho_T\rho_P(1 - \Omega), \tag{22}$$

which is the form used throughout the main-text experiments and the regularizer of Sec. 3.4.

Specialization 2: $W = W_N$ **(Procrustes-optimal, ablation).** The alignment that minimizes the residual is the classical orthogonal Procrustes problem (Schönemann, 1966; Gower & Dijksterhuis, 2004):

$$W_N := \arg\min_{W \in \mathcal{O}(d)} \|Z_T - Z_P W\|_F^2 = \arg\max_{W \in \mathcal{O}(d)} \text{Tr}(Z_T^\top Z_P W), \tag{23}$$

where the second equality follows from Eq. (19) since the first two terms are W-independent. Its closed-form SVD solution is well known; we restate it here for completeness. Let $Z_T^\top Z_P = U\Sigma V^\top$ be the SVD with $U, V \in \mathcal{O}(d)$ and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$, $\sigma_i \geq 0$. By cyclic trace:

$$\text{Tr}(Z_T^\top Z_P W) = \text{Tr}(\Sigma\, V^\top W U) = \text{Tr}(\Sigma R), \quad R := V^\top W U \in \mathcal{O}(d). \tag{24}$$

Since $\sigma_i \geq 0$ and $|R_{ii}| \leq 1$ for any orthogonal R, $\text{Tr}(\Sigma R) = \sum_i \sigma_i R_{ii}$ is maximized at $R = I_d$, giving:

$$W_N = VU^\top, \qquad \max_W \text{Tr}(Z_T^\top Z_P W) = \sum_i \sigma_i = \|Z_T^\top Z_P\|_*. \tag{25}$$

Therefore $\Omega_{W_N} = \Omega_N := \|Z_T^\top Z_P\|_*/(\|Z_T\|_F\|Z_P\|_F)$ (nuclear form, Eq. 38), and Eq. (21) at $W = W_N$ becomes

$$\min_{W \in \mathcal{O}(d)} \frac{1}{n}\|Z_T - Z_P W\|_F^2 = (\rho_T - \rho_P)^2 + 2\rho_T\rho_P(1 - \Omega_N), \tag{26}$$

the form underlying the Sec. 4.2 ablation.

Remark 1 (W_N is the feature-side optimum, not the joint optimum)**.** *The Procrustes alignment W_N minimizes only the feature alignment residual $\delta(W)$, not the full bound $\delta(W) + \gamma(W)$, since the head term also depends on W via $WH_T - H_P$. Appendix B.3 discusses the joint optimization.*

Assembling the feature bound. Combining Eq. (17) with the general decomposition Eq. (21):

$$\delta \le K_{\text{feat}} \sqrt{(\rho_T - \rho_P)^2 + 2\rho_T \rho_P (1 - \Omega_W)}. \tag{27}$$

At $W = I$, replace Ω_W with Ω; at $W = W_N$, with Ω_N.

A.5 Step 4: bounding the head discrepancy γ

Lipschitz property with respect to logits. The head discrepancy measures how much predictions differ when two different "effective heads" (WH_T vs. H_P) operate on the same proxy features. Let $z = \phi_P(x)$ denote the (unaligned) proxy feature. Then:

$$\gamma = \left|\mathcal{R}_{P\to T} - \mathcal{R}_P\right| = \left|\mathbb{E}\big[\ell(z \cdot WH_T, y) - \ell(z \cdot H_P, y)\big]\right|. \tag{28}$$

To bound this, we use the Lipschitz property of cross-entropy with respect to *logits*. The gradient of ℓ with respect to v is $\nabla_v \ell = \hat{p} - e_y$, with norm:

$$\|\nabla_v \ell\|_2 = \|\hat{p} - e_y\|_2. \tag{29}$$

Since $\hat{p}$ lies on the probability simplex, $\|\hat{p}\|_2 \le \|\hat{p}\|_1 = 1$, hence

$$\|\hat{p} - e_y\|_2^2 \;=\; \|\hat{p}\|_2^2 - 2\hat{p}_y + 1 \;\le\; 1 - 2\hat{p}_y + 1 \;=\; 2(1 - \hat{p}_y) \;\le\; 2.$$

The supremum $\sqrt{2}$ is approached as the model's predicted distribution concentrates on an incorrect class ($\hat{p}_y \to 0$). Thus $K_{\text{pred}} \le \sqrt{2}$, yielding the Lipschitz inequality $|\ell(v_1, y) - \ell(v_2, y)| \le K_{\text{pred}} \|v_1 - v_2\|_2$ on logits.

Deriving the covariance projection. Define the head misalignment matrix $\Delta H := WH_T - H_P \in \mathbb{R}^{d\times V}$. The logit difference for a single sample is $z \cdot \Delta H = \phi_P(x)(WH_T - H_P)$, and:

$$\begin{aligned} \gamma &\le \mathbb{E}\big[|\ell(z \cdot WH_T, y) - \ell(z \cdot H_P, y)|\big] \\ &\le K_{\text{pred}} \cdot \mathbb{E}_z\big[\|z \cdot \Delta H\|_2\big]. \end{aligned} \tag{30}$$

Applying Jensen's inequality on the concave square root:

$$\big(\mathbb{E}[\|z \cdot \Delta H\|_2]\big)^2 \le \mathbb{E}\big[\|z \cdot \Delta H\|_2^2\big]. \tag{31}$$

Expanding the squared norm using the trace operator:

$$\begin{aligned} \mathbb{E}\big[\|z \cdot \Delta H\|_2^2\big] &= \mathbb{E}\Big[\operatorname{Tr}\big(\Delta H^\top z^\top z\, \Delta H\big)\Big] \\ &= \operatorname{Tr}\Big(\Delta H^\top \underbrace{\mathbb{E}[z^\top z]}_{\Sigma_P} \Delta H\Big) \\ &= \|\Sigma_P^{1/2} \Delta H\|_F^2. \end{aligned} \tag{32}$$

Taking the square root:

$$\gamma \le K_{\text{pred}} \cdot \|\Sigma_P^{1/2}(WH_T - H_P)\|_F. \tag{33}$$

Interpretation. The covariance weighting $\Sigma_P^{1/2}$ projects the head error onto the *active subspace* of the data. If the heads disagree only in directions orthogonal to the feature distribution (the null space of Σ_P), this disagreement has zero cost, since it does not affect any prediction. This is sharper than the loose spectral bound $\gamma \le K_{\text{pred}} \|WH_T - H_P\|_2 \cdot \mathbb{E}[\|z\|_2]$, which conservatively assumes worst-case alignment between features and head error. The covariance weighting $\Sigma_P^{1/2}$ instead aligns the bound with the data-active directions: head misalignment in directions the features rarely traverse contributes little, and disagreement on $\operatorname{null}(\Sigma_P)$ contributes nothing.

A.6 Step 5: assembling the unified bound

For any $W \in \mathcal{O}(d)$, combining the risk decomposition (Eq. 8), the feature bound (Eq. 27), and the head bound (Eq. 33):

$$|\mathcal{R}_T - \mathcal{R}_P| \;\le\; \delta + \gamma \;\le\; K_{\text{feat}} \sqrt{(\rho_T - \rho_P)^2 + 2\rho_T \rho_P (1 - \Omega_W)} + K_{\text{pred}} \cdot \|\Sigma_P^{1/2}(WH_T - H_P)\|_F. \tag{34}$$

This is Theorem 1 in its general form. The two specializations used in the paper:

- $W = I$ **(identity alignment, main text):** $\Omega_W = \Omega$ (trace form) and $WH_T - H_P = H_T - H_P$, giving the bound used throughout Sec. 4 and the regularizer of Sec. 3.4.

- $W = W_N$ **(Procrustes-optimal, ablation):** $\Omega_W = \Omega_N$ (nuclear form, achieves the tightest feature-side specialization), giving the bound underlying the Sec. 4.2 ablation.

This completes the proof of Theorem 1. □

A.7 Step 6: empirical form and the population corollary

The proof above is written with expectations under $\mathcal{D}$, but nothing in it requires them: the triangle inequality through $\mathcal{R}_{P\to T}$ and the simplex-polarization Lipschitz step (Appendix A.3) are pointwise, Jensen applies to any average, and Prop. 1 is already an exact finite-sample identity. Replacing each expectation by the mean over a reference sample $S = \{(x_i, y_i)\}_{i=1}^N$ therefore gives, with $\widehat{\mathcal{R}}_M = \frac{1}{N}\sum_i \ell(\phi_M(x_i)H_M, y_i)$ and $\widehat{\rho}_M, \widehat{\Omega}_W, \widehat{\Sigma}_P$ computed from the Z_T, Z_P extracted on S,

$$\left|\widehat{\mathcal{R}}_T - \widehat{\mathcal{R}}_P\right| \;\le\; \widehat{\mathcal{B}} \;:=\; K_{\text{feat}}\sqrt{(\widehat{\rho}_T - \widehat{\rho}_P)^2 + 2\widehat{\rho}_T\widehat{\rho}_P(1-\widehat{\Omega}_W)} \;+\; K_{\text{pred}}\left\|\widehat{\Sigma}_P^{1/2}(WH_T - H_P)\right\|_F. \tag{35}$$

This is the form the experiments test: every quantity reported in this paper is hatted and measured on $\mathcal{D}_{\text{diag}}$, and the $360/360$ satisfied rows of Sec. 4.2 verify Eq. (35), not a population statement. Reaching the population gap is a separate and standard step, whose two extra assumptions we state rather than absorb.

Corollary 1 (Population risk gap)**.** *Assume S is drawn i.i.d. from $\mathcal{D}$ and that the per-sample loss gap $D_i := \ell(\phi_T(x_i)H_T, y_i) - \ell(\phi_P(x_i)H_P, y_i)$ satisfies $|D_i| \le b$ almost surely. Since $\mathcal{R}_T - \mathcal{R}_P = \mathbb{E}[D_1]$ and $\widehat{\mathcal{R}}_T - \widehat{\mathcal{R}}_P = \frac{1}{N}\sum_i D_i$, McDiarmid's inequality for that mean (bounded differences $2b/N$), combined with Eq. (35), gives for any $\delta_c \in (0,1)$, with probability at least $1-\delta_c$ over the draw of S,*

$$|\mathcal{R}_T - \mathcal{R}_P| \;\le\; \widehat{\mathcal{B}} \;+\; b\sqrt{\frac{2\ln(2/\delta_c)}{N}}. \tag{36}$$

Splitting D_i at the hybrid loss as in Step 1 bounds the constant explicitly: if $\sup_x \|\phi_T(x) - \phi_P(x)W\|_2 \le D_Z$ and $\sup_x \|\phi_P(x)\|_2 \le C_z$, then $b \le K_{\text{feat}}D_Z + \sqrt{2}\,C_z\,\|WH_T - H_P\|_{\text{op}}$.

The deviation term is a generic $N^{-1/2}$ concentration penalty, the price of any finite-sample estimate of a population gap rather than a property of PRISM, and for sequence data the per-sample loss is the token-average within a sequence, so N counts reference sequences rather than tokens. We state the empirical form as the theorem because it is what ranking on a fixed reference set needs, and because b, unlike $\widehat{\mathcal{B}}$, is a worst-case constant that is not directly observable.

B Choosing the orthogonal alignment

Theorem 1 is stated for every $W \in \mathcal{O}(d)$, so the question this appendix answers is not whether the bound generalizes but which member of the family to report. We write out the bound as a function of W, characterize the two specializations the paper uses, the $W = I$ trace form and the Procrustes-optimal $W = W_N$ nuclear form (the latter in the Sec. 4.2 ablation), and treat the joint optimization over the feature and head terms. The two specializations encode different priorities: W_N minimizes the feature alignment residual but generally inflates the head term, while $W = I$ makes the head term vanish exactly under a frozen `lm_head` and is SVD-free differentiable, properties that the LLM settings studied here (PTQ with FP16 head and frozen-head LoRA) require.

B.1 The bound as a function of W

The exact decomposition derived in Appendix A.4 (Eq. 21) holds for every $W \in \mathcal{O}(d)$ via the alignment-dependent similarity Ω_W of Eq. 3, yielding a family of risk bounds parameterized by W:

$$|\mathcal{R}_T - \mathcal{R}_P| \;\le\; \underbrace{K_{\text{feat}}\sqrt{(\rho_T - \rho_P)^2 + 2\rho_T\rho_P(1-\Omega_W)}}_{\delta(W)} + \underbrace{K_{\text{pred}}\|\Sigma_P^{1/2}(WH_T - H_P)\|_F}_{\gamma(W)}. \tag{37}$$

The Procrustes solution W_N minimizes $\delta(W)$ alone, yielding the nuclear form

$$\Omega_{W_N} = \Omega_N := \frac{\|Z_T^\top Z_P\|_*}{\|Z_T\|_F\|Z_P\|_F} = \max_{W\in\mathcal{O}(d)} \frac{\operatorname{Tr}(Z_T^\top Z_P\, W)}{\|Z_T\|_F\|Z_P\|_F}, \tag{38}$$

the maximum of the trace family over orthogonal alignments (Appendix A.4). However, W_N enters the head term $\gamma(W_N)$, which can be inflated when the rotation is applied to a nearly preserved head ($H_T \approx H_P$). This creates a fundamental trade-off: W_N gives the tightest feature bound but may worsen the head bound, while $W = I$ preserves head coherence at the cost of a slightly looser feature bound.

B.2 The trace specialization ($W = I$)

Setting $W = I$ recovers the main-text trace similarity $\Omega = \Omega_{W=I}$ (Eq. 3). Two structural properties characterize this specialization.

Ordering of similarities. Since the Procrustes solution maximizes $\mathrm{Tr}(Z_T^\top Z_P W)$ over $\mathcal{O}(d)$:

$$\mathrm{Tr}(Z_T^\top Z_P) \;\le\; \|Z_T^\top Z_P\|_*, \tag{39}$$

which gives $\Omega \le \Omega_N$. Equality holds if and only if $W_N = I$, i.e., $Z_T^\top Z_P$ is symmetric positive semidefinite.

When does $\Omega = \Omega_N$? Let $Z_T^\top Z_P = U\Sigma V^\top$ be the SVD. Then $\Omega = \Omega_N$ iff $VU^\top = I$, i.e., $U = V$, which holds when $Z_T^\top Z_P$ is symmetric positive semidefinite (SPSD). In post-training quantization with mild noise, $Z_T^\top Z_P \approx Z_T^\top Z_T$ is approximately SPSD, so $\Omega \approx \Omega_N$. As quantization becomes more aggressive, asymmetric perturbations cause Ω to fall below Ω_N, and this gap reflects genuine asymmetric distortion that a rotation-invariant metric would mask.

Relation to CKA. The Frobenius-form similarity $\Omega_F := \|Z_T^\top Z_P\|_F / (\|Z_T\|_F \|Z_P\|_F)$ satisfies $\Omega_F \le \Omega_N$, since $\|\sigma\|_2 \le \|\sigma\|_1$ on the singular values of $Z_T^\top Z_P$, but it does not arise from the alignment residual at any W. Linear CKA (Kornblith et al., 2019), $\|Z_T^\top Z_P\|_F^2 / (\|Z_T^\top Z_T\|_F \, \|Z_P^\top Z_P\|_F)$, differs from Ω_F^2 only in its denominator and satisfies $\mathrm{CKA} \ge \Omega_F^2$ because $\|A\|_F \le \mathrm{Tr}(A)$ for positive semidefinite A. Neither quantity is derived from an alignment, so neither can be substituted into the residual of Proposition 1; Ω_N is the member of that family that exactly solves the orthogonal Procrustes problem.

Head-side simplification. At $W = I$, the head term $K_{\mathrm{pred}} \, \|\Sigma_P^{1/2}(H_T - H_P)\|_F$ carries no rotational contamination and vanishes whenever the prediction head is preserved (FP16 head under PTQ, frozen head under LoRA), recovering the regimes used throughout Sec. 4 and the regularizer of Sec. 3.4.

B.3 Joint optimization over W

Minimizing $\delta(W) + \gamma(W)$ jointly over $\mathcal{O}(d)$ has no closed form, since the head term is a weighted orthogonal Procrustes objective (Gower & Dijksterhuis, 2004), and a numerical solution would fold scale, shape and head into the alignment itself, defeating the purpose of reading them separately. The question does not arise in the regimes studied here: with $H_T = H_P$ (GPTQ, BitsAndBytes, frozen-head LoRA) the head term vanishes exactly at $W = I$ and is generically positive elsewhere, so $W = I$ is the natural choice for analysis and for the regularizer of Sec. 3.4, while W_N tightens the feature term at the price of a nonzero head term; jointly rotated features and divergent heads (e.g., full-parameter SFT) are left to future work.

C Formal extension to autoregressive generation

The bound of Theorem 1 applies to any matrix pair $(Z_T, Z_P) \in \mathbb{R}^{n \times d}$ of feature vectors, whatever the origin of the rows. We spell out here the autoregressive instantiation informally summarized in Sec. 3.2.

From tokens to matrices. Fix a single sequence $(c, y_1, \ldots, y_{|y|})$. Each position τ supplies a (context, target-token) pair instantiating Eq. (1) with feature $\phi_M(c, y_{<\tau}) \in \mathbb{R}^d$ and loss $\ell(\phi_M(c, y_{<\tau})H_M, \, y_\tau)$. Stacking the $|y|$ feature vectors into a $|y| \times d$ block recovers exactly the per-row setting that Theorem 1 controls.

Corollary 2 (Autoregressive Risk Bound)**.** *For a single sequence $(c, y_1, \ldots, y_{|y|})$, let $Z_T^{\mathrm{AR}}, Z_P^{\mathrm{AR}} \in \mathbb{R}^{|y| \times d}$ be the feature matrices stacked as above, with RMS scales $\rho_T^{\mathrm{AR}}, \rho_P^{\mathrm{AR}}$, Procrustes similarity*

Ω^{AR}*, and covariance* Σ_P^{AR}*. Then*

$$|\mathcal{R}_T^{\mathrm{AR}} - \mathcal{R}_P^{\mathrm{AR}}| \le K_{\mathrm{feat}} \sqrt{(\rho_T^{\mathrm{AR}} - \rho_P^{\mathrm{AR}})^2 + 2\,\rho_T^{\mathrm{AR}}\,\rho_P^{\mathrm{AR}}(1 - \Omega^{\mathrm{AR}})} + K_{\mathrm{pred}} \left\| (\Sigma_P^{\mathrm{AR}})^{1/2} (H_T - H_P) \right\|_F. \tag{40}$$

Proof. Each row of Z_M^{AR} is a feature vector of dimension d, and the Lipschitz analysis of Appendix A.3 together with the Procrustes decomposition of Appendix A.4 depend only on the matrix shape, not on the provenance of its rows. Applying Theorem 1 to $(Z_T^{\mathrm{AR}}, Z_P^{\mathrm{AR}})$ yields the stated inequality. □

This unification enables one bound to cover point-wise classification (MMLU, ARC, where $|y|{=}1$), short-horizon QA (SQuAD, TriviaQA, where y is a 1–5 token span), and multi-step reasoning (GSM8K, where y is a chain-of-thought solution).

Proxy-generated trajectories. Nothing in Corollary 2 requires y to be a reference answer: any sequence that both models score is a valid y, including one the proxy generated itself, in which case its own early errors shape its later context. We test this on GSM8K, the only one of our benchmarks whose answers are long enough for trajectories to diverge. For each of the 12 Llama-3.1-8B PTQ variants, the proxy greedily generates a continuation for 100 prompts (at most 128 new tokens; 72–81 on average), and both models are then scored on that trajectory with the generated tokens as targets. Because the proxy chose these tokens greedily, its own loss on them is small, and the gap mostly measures how far the target's predictions depart from the proxy's along its trajectory: a divergence between the two models' generative distributions rather than a task loss, valid as a bound for each variant on its own trajectory. Over five disjoint prompt subsets, the bound computed on these rollouts ranks the rollout gaps at $r_s = 0.947 \pm 0.015$, against 0.958 ± 0.011 for the teacher-forced bound and gap on the same prompts; the teacher-forced bound alone already orders the rollout gaps at 0.958 ± 0.011, and the two bounds agree on the variant ordering at 0.959 ± 0.010. Free-running use is therefore covered, but it requires decoding, which is what teacher forcing avoids, so teacher forcing remains our recommended default.

D Model and quantization details

D.1 Target models

Table 4 lists all target (full-precision) models evaluated in the quantization experiments. All targets are loaded in BF16 as the reference precision; an FP16 proxy is additionally included for each model to measure the BF16→FP16 precision-drift baseline.

Table 4: Target models used in quantization experiments. All models are loaded in BF16.

#	Family	HuggingFace ID	Variant	Params	d	$\lvert V\rvert$
1	Qwen 3	`Qwen/Qwen3-8B-Base`	Base	8B	4096	151K
2	Llama 3.1	`meta-llama/Meta-Llama-3.1-8B`	Base	8B	4096	128K
3	Ministral 3	`mistralai/Ministral-3-8B-Base-2512`	Base	8B	4096	131K
4	Qwen 3	`Qwen/Qwen3-8B`	Instruct	8B	4096	151K
5	Llama 3.1	`meta-llama/Meta-Llama-3.1-8B-Instruct`	Instruct	8B	4096	128K
6	Ministral 3	`mistralai/Ministral-3-8B-Instruct-2512`	Instruct	8B	4096	131K
7	DeepSeek R1	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	Distilled	8B	4096	128K

D.2 Quantization backends and bit-widths

Every target model is evaluated against up to three quantization backends. Table 5 summarises the shared GGUF and BitsAndBytes (BnB) configurations applied uniformly across all models.

D.3 Backend coverage and GPTQ checkpoints

Table 6 lists, per model, the backends evaluated and the pre-quantized GPTQ checkpoint loaded where one was available. GGUF (six bit-widths) and BitsAndBytes (three configurations) are run on every model; GPTQ is 4-bit (the setting reported in Sec. 4.1) except for Qwen3-8B-Instruct, where

Table 5: Shared quantization tiers applied to all target models. GGUF quants use the `llama.cpp` k-quant scheme; BnB quants use on-the-fly quantization from the full-precision checkpoint.

Tier	Backend	Tag	Description
FP16 reference	dtype	`dtype:float16`	BF16→FP16 precision drift baseline
8-bit	GGUF	`Q8_0`	Round-to-nearest, 8-bit
	BnB	`bnb:int8`	LLM.int8() absmax
6-bit	GGUF	`Q6_K`	k-quant, 6-bit (super-blocks)
5-bit	GGUF	`Q5_K_M`	k-quant, 5-bit (medium)
4-bit	GGUF	`Q4_K_M`	k-quant, 4-bit (medium)
	BnB	`bnb:nf4`	NormalFloat4 (QLoRA format)
	BnB	`bnb:fp4`	FP4 with FP16 compute
3-bit	GGUF	`Q3_K_M`	k-quant, 3-bit (medium)
2-bit	GGUF	`Q2_K`	k-quant, 2-bit

the only public checkpoint we could load uses 8-bit. Cells marked "—" were outside the compute budget of this study.

Table 6: Quantization backend coverage per model, with the GPTQ checkpoint used. ✓ = included; — = not evaluated.

#	Model	BnB	GGUF	GPTQ repository
1	Qwen3-8B-Base	✓	✓	—
2	Llama-3.1-8B	✓	✓	`ModelCloud/Meta-Llama-3.1-8B-gptq-4bit`
3	Ministral-3-Base	✓	✓	—
4	Qwen3-8B	✓	✓	`JunHowie/Qwen3-8B-GPTQ-Int8`
5	Llama-3.1-8B-Inst.	✓	✓	`ModelCloud/Meta-Llama-3.1-8B-Instruct-gptq-4bit`
6	Ministral-3-Inst.	✓	✓	—
7	DeepSeek-R1-Distill	✓	✓	`jakiAJK/DeepSeek-R1-Distill-Llama-8B_GPTQ-int4`

D.4 GGUF REPOSITORIES

Table 7 lists the GGUF repositories and filename conventions for each model.

Table 7: GGUF repositories and filename templates. Each repository provides quants `Q8_0`, `Q6_K`, `Q5_K_M`, `Q4_K_M`, `Q3_K_M`, and `Q2_K` unless otherwise noted.

#	Model	GGUF Repository	Filename Template
1	Qwen3-8B-Base	`mradermacher/Qwen3-8B-Base-GGUF`	`Qwen3-8B-Base.{Q}.gguf`
2	Llama-3.1-8B	`QuantFactory/Meta-Llama-3.1-8B-GGUF`	`Meta-Llama-3.1-8B.{Q}.gguf`
3	Ministral-3-Base	`mradermacher/Ministral-3-8B-Base-2512-GGUF`	`Ministral-3-8B-Base-2512.{Q}.gguf`
4	Qwen3-8B	`Qwen/Qwen3-8B-GGUF`	`Qwen3-8B-{Q}.gguf`
5	Llama-3.1-8B-Inst.	`bartowski/Meta-Llama-3.1-8B-Instruct-GGUF`	`...-Instruct-{Q}.gguf`
6	Ministral-3-Inst.	`bartowski/mistralai_Ministral-3-8B-Instruct-2512-GGUF`	`...-2512-{Q}.gguf`
7	DeepSeek-R1-Distill	`bartowski/DeepSeek-R1-Distill-Llama-8B-GGUF`	`...-Llama-8B-{Q}.gguf`

E QUANTIZATION REPLICATION, FEATURE-ONLY SCATTER, AND PER-MODEL TABLES

E.1 REPLICATION ON MINISTRAL AND DEEPSEEK

Figure 5 replicates the main quantization scatter (Fig. 2) on the two additional 8B families, Ministral-3-8B and DeepSeek-R1-Distill-Llama-8B. The bound tracks $|\Delta\mathcal{R}|$ on these two families with patterns consistent with the Llama–Qwen main analysis, supporting the cross-family generality of the PRISM framework. Per-variant numerical decompositions for all four families are in Appendix E.3.

E.2 FEATURE-ONLY SCATTER

Figure 6 reports the feature-alignment-only replication of the main quantization result (Fig. 2), with the x-axis replaced by the head-free term $\delta = K_{\text{feat}}\sqrt{(\rho_T-\rho_P)^2 + 2\rho_T\rho_P(1-\Omega)}$. The grid isolates the backbone contribution to the bound.

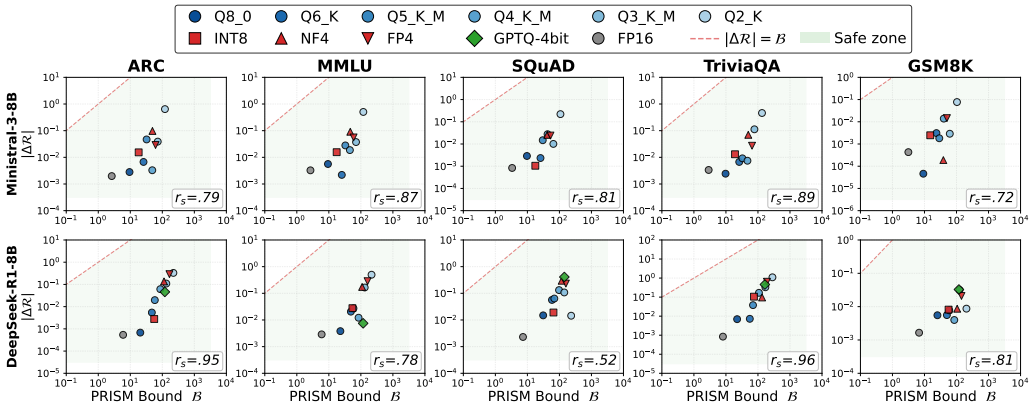

Figure 5: **PTQ replication on Ministral-3-8B and DeepSeek-R1-Distill-Llama-8B.** Counterpart of main-text Fig. 2 (Llama-3.1-8B and Qwen3-8B), with identical axes, PTQ family color coding, and benchmark set. Rows: Ministral-3-8B, DeepSeek-R1-Distill-Llama-8B. Columns: ARC, MMLU, SQuAD, TriviaQA, GSM8K. Per-subplot Spearman r_s is annotated in each panel.

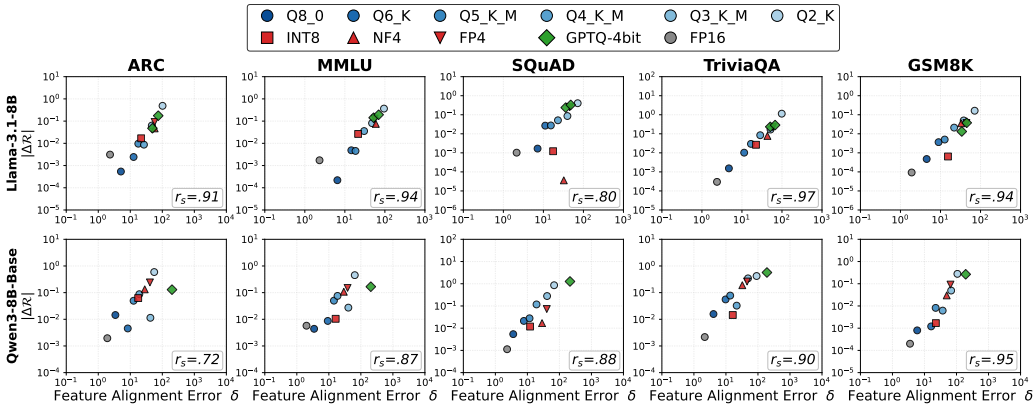

Figure 6: **Feature alignment error δ alone is already highly predictive of the risk gap.** Identical layout to Fig. 2 but with x-axis replaced by the feature-only term δ. Because most protocols retain the `lm_head` in FP16, the head-discrepancy term γ is zero for GPTQ and BitsAndBytes and smaller than δ for most GGUF configurations that quantize output embeddings; δ therefore captures most of the predictive signal of Fig. 2, so scale drift $(\Delta\rho)^2$ and shape mismatch $(1 - \Omega)$ account for most of the degradation signal under PTQ (Sec. 4.1).

E.3 Per-model decomposition tables

We report the full geometric decomposition (Eq. 2) for each evaluated model, covering all quantization variants and benchmarks. Each table column lists the base RMS scale ρ_T, proxy RMS scale ρ_P, identity-alignment similarity Ω, feature alignment error δ, head discrepancy γ, full PRISM bound $\mathcal{B}$, and empirical cross-entropy gap $|\Delta\mathcal{R}|$. Per-dataset Spearman rank correlations $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ appear in the left-hand column of each benchmark block.

The base-model tables cover the four 8B families: Llama-3.1-8B extended benchmarks in Table 9, Ministral-3-8B in Table 11, Qwen3-8B in Table 10 (also referenced in Sec. 4.4), and DeepSeek-R1-Distill-Llama-8B in Table 12. The instruction-tuned counterparts of Llama, Ministral, and Qwen are reported in Tables 13, 15, and 14 respectively, and reproduce the same patterns as their base counterparts. Across ARC/MMLU/SQuAD/TriviaQA the per-benchmark mean Spearman lies in $r_s \in [0.76, 0.89]$ (Table 8), and GSM8K reaches $r_s \approx 0.73$. GSM8K's long teacher-forced chain-of-thought answer spans dilute per-token loss, shrinking mean $|\Delta\mathcal{R}|$ to ≈ 0.019, four to nine times below the other benchmarks (0.07–0.16), so it is the benchmark with the least headroom between predictive signal and measurement noise, but the bound still ranks its variants. The single cell where that headroom is exhausted is Qwen3-8B-Instruct, whose mean $|\Delta\mathcal{R}|$ on GSM8K is only 0.0036:

Table 8: Per-benchmark mean empirical risk gap $|\Delta\mathcal{R}|$ (averaged over per-method aggregated quantization rows, matching the per-model body tables) and per-benchmark mean Spearman $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ (over raw variants, matching the figure's per-subplot r_s) across seven 8B model families. GSM8K's small $|\Delta\mathcal{R}|$—driven by long teacher-forced answer spans diluting per-token loss—leaves the least headroom between predictive signal and measurement noise; the effect shows up per family rather than in the benchmark mean.

	Empirical risk gap $\|\Delta\mathcal{R}\|$				
Model	**ARC**	**MMLU**	**SQuAD**	**TriviaQA**	**GSM8K**
Llama-3.1-8B	0.0766	0.0812	0.1030	0.1839	0.0311
Ministral-3-8B	0.0886	0.0761	0.0330	0.0711	0.0117
Qwen3-8B	0.1198	0.0893	0.1411	0.1412	0.0468
DeepSeek-R1-8B	0.0910	0.1112	0.1224	0.2680	0.0123
Llama-3.1-8B-Instruct	0.0709	0.1080	0.0744	0.2086	0.0204
Ministral-3-8B-Instruct	0.0274	0.0265	0.0372	0.1108	0.0052
Qwen3-8B-Instruct	0.0307	0.0275	0.0861	0.1559	0.0036
Mean $\|\Delta\mathcal{R}\|$	0.0721	0.0743	0.0853	0.1628	**0.0187**
Mean r_s	0.758	0.786	0.794	0.889	0.726

at that scale per-variant differences are dominated by measurement noise rather than the bound's predictive signal, and the per-cell r_s is correspondingly near zero.

Table 9: Geometric decomposition for **Llama-3.1-8B** under identity alignment ($W{=}I$) on extended benchmarks (excl. MMLU). Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	**Family**	**Method**	ρ_T	ρ_P	Ω	δ	γ	**PRISM $\mathcal{B}$**	$\|\Delta\mathcal{R}\|$
ARC	–	FP16	137.59	137.63	0.9999	6.0881	0	6.0882	0.0031
($\mathbf{r_s{=}0.91}$)	GGUF	Q8_0	137.59	137.68	0.9993	13.2980	5.8902	19.1882	0.0005
		Q6_K	137.59	138.04	0.9958	33.1354	18.4531	51.5885	0.0024
		Q5_K_M	137.59	138.57	0.9917	46.4894	18.5484	65.0377	0.0096
		Q4_K_M	137.59	137.70	0.9812	69.7101	18.4173	88.1274	0.0087
		Q3_K_M	137.59	138.67	0.9424	122.5642	18.5484	141.1126	0.0631
		Q2_K	137.59	144.67	0.7413	265.8843	19.6871	285.5715	0.4872
	BnB	INT8	137.59	137.98	0.9874	57.1181	0	57.1181	0.0168
		NF4	137.59	142.78	0.9159	150.8955	0	150.8955	0.0470
		FP4	137.59	135.86	0.9116	150.2889	0	150.2889	0.0926
	GPTQ	GPTQ-4bit	137.59	138.70	0.9369	128.3108	0	128.3108	0.0476
TriviaQA	–	FP16	140.94	140.94	0.9999	6.2233	0	6.2233	0.0003
($\mathbf{r_s{=}0.97}$)	GGUF	Q8_0	140.94	140.90	0.9994	12.3313	6.0651	18.3964	0.0015
		Q6_K	140.94	141.12	0.9968	29.7139	19.1269	48.8408	0.0102
		Q5_K_M	140.94	141.19	0.9929	43.8865	19.1393	63.0257	0.0296
		Q4_K_M	140.94	141.15	0.9794	74.7724	19.1382	93.9106	0.0842
		Q3_K_M	140.94	141.40	0.9329	135.2098	19.1887	154.3985	0.1675
		Q2_K	140.94	142.26	0.7574	257.8427	19.3354	277.1781	1.1405
	BnB	INT8	140.94	141.08	0.9874	58.6196	0	58.6196	0.0266
		NF4	140.94	142.03	0.9532	113.1651	0	113.1651	0.0773
		FP4	140.94	142.55	0.9124	155.1706	0	155.1706	0.2248
	GPTQ	GPTQ-4bit	140.94	141.30	0.9346	133.3713	0	133.3713	0.2326
SQuAD	–	FP16	138.27	138.28	0.9999	5.6049	0	5.6049	0.0010
($\mathbf{r_s{=}0.83}$)	GGUF	Q8_0	138.27	138.21	0.9987	18.6994	5.9026	24.6020	0.0017
		Q6_K	138.27	138.25	0.9968	29.0704	18.5785	47.6489	0.0268
		Q5_K_M	138.27	138.20	0.9938	40.1193	18.5692	58.6885	0.0274
		Q4_K_M	138.27	138.19	0.9861	60.3256	18.5695	78.8951	0.0512
		Q3_K_M	138.27	138.29	0.9586	104.0296	18.5864	122.6161	0.0857

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		Q2_K	138.27	138.56	0.8652	187.8263	18.6283	206.4546	0.4039
	BnB	INT8	138.27	138.17	0.9920	45.7751	0	45.7751	0.0012
		NF4	138.27	138.36	0.9727	84.5191	0	84.5191	0.0000
		FP4	138.27	139.22	0.9462	119.0043	0	119.0043	0.2498
	GPTQ	GPTQ-4bit	138.27	138.37	0.9660	94.3404	0	94.3404	0.2379
GSM8K	–	FP16	138.93	138.94	0.9999	4.9758	0	4.9758	0.0001
($\mathbf{r_s}$=**0.97**)	GGUF	Q8_0	138.93	138.93	0.9995	11.8111	5.9378	17.7489	0.0005
		Q6_K	138.93	138.92	0.9979	23.3889	18.7001	42.0890	0.0036
		Q5_K_M	138.93	138.95	0.9958	33.2591	18.7055	51.9645	0.0050
		Q4_K_M	138.93	138.93	0.9875	57.3563	18.6941	76.0504	0.0211
		Q3_K_M	138.93	139.33	0.9627	99.3114	18.7726	118.0840	0.0502
		Q2_K	138.93	139.17	0.8683	186.5094	18.7291	205.2385	0.1630
	BnB	INT8	138.93	138.92	0.9939	40.1892	0	40.1892	0.0006
		NF4	138.93	139.00	0.9721	85.8494	0	85.8494	0.0356
		FP4	138.93	139.53	0.9491	116.1543	0	116.1543	0.0367
	GPTQ	GPTQ-4bit	138.93	138.87	0.9700	88.9440	0	88.9440	0.0132

Table 10: Geometric decomposition for **Qwen3-8B-Base** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	332.95	333.07	1.0000	6.9340	0	6.9341	0.0057
($\mathbf{r_s}$=**0.76**)	GGUF	Q8_0	332.95	332.80	0.9999	11.9995	27.2373	39.2369	0.0044
		Q6_K	332.95	332.19	0.9996	32.0060	86.4603	118.4663	0.0087
		Q5_K_M	332.95	332.09	0.9991	49.8368	86.4112	136.2480	0.0500
		Q4_K_M	332.95	333.38	0.9984	64.6037	86.7488	151.3525	0.0754
		Q3_K_M	332.95	324.48	0.9927	140.6362	84.4380	225.0742	0.0274
		Q2_K	332.95	336.96	0.9815	223.1666	87.5885	310.7551	0.4521
	BnB	INT8	332.95	332.30	0.9988	56.4563	0	56.4563	0.0104
		NF4	332.95	329.78	0.9961	101.3736	0	101.3736	0.1080
		FP4	332.95	332.85	0.9933	133.0910	0	133.0910	0.1509
ARC	–	FP16	332.77	332.85	1.0000	6.6202	0	6.6203	0.0020
($\mathbf{r_s}$=**0.53**)	GGUF	Q8_0	332.77	332.35	0.9999	11.8269	27.1907	39.0176	0.0145
		Q6_K	332.77	332.43	0.9997	28.7997	86.4728	115.2725	0.0045
		Q5_K_M	332.77	333.35	0.9993	43.8386	86.6632	130.5018	0.0496
		Q4_K_M	332.77	333.88	0.9984	66.2247	86.7784	153.0031	0.0875
		Q3_K_M	332.77	320.15	0.9923	146.4970	83.2223	229.7194	0.0114
		Q2_K	332.77	338.63	0.9863	193.0062	87.9769	280.9831	0.5916
	BnB	INT8	332.77	329.24	0.9986	61.0306	0	61.0306	0.0626
		NF4	332.77	328.23	0.9964	97.7019	0	97.7019	0.1328
		FP4	332.77	333.53	0.9924	141.9107	0	141.9107	0.2417
TriviaQA	–	FP16	241.16	240.98	1.0000	7.5430	0	7.5430	0.0022
($\mathbf{r_s}$=**0.94**)	GGUF	Q8_0	241.16	241.37	0.9999	14.1202	19.8002	33.9205	0.0158
		Q6_K	241.16	241.94	0.9992	33.7306	62.7352	96.4659	0.0556
		Q5_K_M	241.16	240.12	0.9984	47.0098	62.2697	109.2795	0.0776
		Q4_K_M	241.16	241.89	0.9960	74.9351	62.7311	137.6662	0.0325
		Q3_K_M	241.16	235.24	0.9792	169.1604	61.0151	230.1755	0.3429
		Q2_K	241.16	253.25	0.9338	313.6651	65.6364	379.3015	0.4199
	BnB	INT8	241.16	239.40	0.9977	56.5553	0	56.5553	0.0145
		NF4	241.16	241.25	0.9909	112.3114	0	112.3114	0.1900
		FP4	241.16	241.16	0.9821	158.0163	0	158.0163	0.2615
SQuAD	–	FP16	298.58	298.42	1.0000	8.0509	0	8.0510	0.0011

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
($\mathbf{r_s{=}1.00}$)	GGUF	Q8_0	298.58	298.71	0.9999	12.4941	24.4758	36.9699	0.0054
		Q6_K	298.58	299.00	0.9997	26.8333	77.4197	104.2530	0.0213
		Q5_K_M	298.58	298.32	0.9992	40.3879	77.2515	117.6394	0.0278
		Q4_K_M	298.58	300.42	0.9979	66.6773	77.7849	144.4622	0.1173
		Q3_K_M	298.58	298.07	0.9905	142.5569	77.1934	219.7503	0.2796
		Q2_K	298.58	308.09	0.9751	236.3322	79.7480	316.0802	0.8567
	BnB	INT8	298.58	297.45	0.9992	42.0565	0	42.0565	0.0119
		NF4	298.58	293.30	0.9955	99.3043	0	99.3043	0.0169
		FP4	298.58	305.28	0.9913	139.4570	0	139.4570	0.0730
GSM8K	–	FP16	273.35	273.43	0.9999	12.2277	0	12.2277	0.0002
($\mathbf{r_s{=}0.95}$)	GGUF	Q8_0	273.35	273.04	0.9998	20.4881	22.3962	42.8843	0.0008
		Q6_K	273.35	276.88	0.9984	55.9547	71.5294	127.4840	0.0012
		Q5_K_M	273.35	277.03	0.9968	77.2788	71.5690	148.8478	0.0082
		Q4_K_M	273.35	280.38	0.9914	127.9424	72.4245	200.3668	0.0062
		Q3_K_M	273.35	280.79	0.9702	235.4422	72.5002	307.9424	0.0493
		Q2_K	273.35	318.23	0.9462	368.8037	82.2336	451.0373	0.2787
	BnB	INT8	273.35	274.18	0.9966	78.3422	0	78.3422	0.0017
		NF4	273.35	280.04	0.9843	171.0420	0	171.0420	0.0299
		FP4	273.35	277.91	0.9724	224.4546	0	224.4546	0.0917

Table 11: Geometric decomposition for **Ministral-3-8B-Base** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	191.05	191.05	0.9999	2.6652	0	2.6652	0.0033
($\mathbf{r_s{=}0.87}$)	GGUF	Q8_0	191.05	191.03	0.9997	4.6604	4.7230	9.3834	0.0056
		Q6_K	191.05	191.12	0.9984	10.6224	14.7564	25.3788	0.0022
		Q5_K_M	191.05	191.55	0.9954	18.0046	14.7573	32.7619	0.0280
		Q4_K_M	191.05	191.11	0.9867	30.6699	14.7318	45.4017	0.0187
		Q3_K_M	191.05	190.43	0.9556	55.9842	14.6636	70.6477	0.0374
		Q2_K	191.05	193.99	0.8484	104.4459	14.8564	119.3023	0.5035
	BnB	INT8	191.05	191.21	0.9957	17.5506	0	17.5506	0.0156
		NF4	191.05	190.91	0.9691	46.7289	0	46.7289	0.0910
		FP4	191.05	191.57	0.9504	59.3479	0	59.3479	0.0558
ARC	–	FP16	191.12	191.11	0.9999	2.6333	0	2.6334	0.0020
($\mathbf{r_s{=}0.79}$)	GGUF	Q8_0	191.12	191.08	0.9997	4.7790	4.6769	9.4559	0.0028
		Q6_K	191.12	191.06	0.9982	11.2339	14.6347	25.8685	0.0067
		Q5_K_M	191.12	191.53	0.9956	17.6667	14.6523	32.3190	0.0469
		Q4_K_M	191.12	190.52	0.9837	33.9737	14.5890	48.5628	0.0033
		Q3_K_M	191.12	189.88	0.9532	57.4450	14.5253	71.9702	0.0388
		Q2_K	191.12	194.15	0.8437	106.1286	14.7282	120.8569	0.6435
	BnB	INT8	191.12	191.45	0.9953	18.2446	0	18.2446	0.0154
		NF4	191.12	190.50	0.9658	49.1696	0	49.1696	0.0973
		FP4	191.12	191.92	0.9469	61.4591	0	61.4591	0.0291
TriviaQA	–	FP16	194.06	194.06	0.9999	2.8871	0	2.8872	0.0034
($\mathbf{r_s{=}0.89}$)	GGUF	Q8_0	194.06	194.07	0.9996	5.1726	4.5582	9.7308	0.0024
		Q6_K	194.06	194.09	0.9981	11.8017	14.2389	26.0405	0.0067
		Q5_K_M	194.06	194.14	0.9953	18.4400	14.2344	32.6744	0.0092
		Q4_K_M	194.06	194.23	0.9854	32.6894	14.2345	46.9239	0.0075
		Q3_K_M	194.06	193.98	0.9452	63.2679	14.2302	77.4981	0.1125
		Q2_K	194.06	195.16	0.8057	119.4964	14.2513	133.7477	0.4595
	BnB	INT8	194.06	194.15	0.9950	19.2048	0	19.2048	0.0130
		NF4	194.06	194.10	0.9660	49.8162	0	49.8162	0.0697

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		FP4	194.06	194.44	0.9417	65.3210	0	65.3210	0.0272
SQuAD	–	FP16	186.71	186.71	0.9998	3.3405	0	3.3406	0.0008
($\mathbf{r_s}$=**0.81**)	GGUF	Q8_0	186.71	186.72	0.9996	5.1520	4.5230	9.6750	0.0029
		Q6_K	186.71	186.71	0.9980	11.5662	14.0954	25.6616	0.0023
		Q5_K_M	186.71	186.67	0.9962	16.0711	14.0900	30.1610	0.0148
		Q4_K_M	186.71	186.88	0.9880	28.5358	14.0865	42.6223	0.0276
		Q3_K_M	186.71	186.93	0.9621	50.6392	14.0996	64.7388	0.0101
		Q2_K	186.71	188.59	0.8719	93.5709	14.1253	107.6963	0.2209
	BnB	INT8	186.71	186.77	0.9954	17.6942	0	17.6942	0.0010
		NF4	186.71	186.54	0.9743	41.6677	0	41.6677	0.0256
		FP4	186.71	187.28	0.9592	52.6147	0	52.6147	0.0240
GSM8K	–	FP16	186.46	186.45	0.9999	3.1380	0	3.1380	0.0004
($\mathbf{r_s}$=**0.72**)	GGUF	Q8_0	186.46	186.46	0.9997	4.8023	4.4029	9.2052	0.0000
		Q6_K	186.46	186.35	0.9986	9.7288	13.8810	23.6098	0.0031
		Q5_K_M	186.46	186.41	0.9966	15.1044	13.8844	28.9888	0.0018
		Q4_K_M	186.46	186.61	0.9898	26.2032	13.8893	40.0925	0.0139
		Q3_K_M	186.46	186.80	0.9662	47.8137	13.8987	61.7125	0.0029
		Q2_K	186.46	188.47	0.8824	89.5578	13.9425	103.5003	0.0772
	BnB	INT8	186.46	186.51	0.9966	15.1307	0	15.1307	0.0025
		NF4	186.46	186.04	0.9776	38.8596	0	38.8596	0.0002
		FP4	186.46	186.75	0.9627	50.1973	0	50.1973	0.0146

Table 12: Geometric decomposition for **DeepSeek-R1-Distill-Llama-8B** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	130.42	130.31	0.9998	5.9618	0	5.9618	0.0029
($\mathbf{r_s}$=**0.78**)	GGUF	Q8_0	130.42	130.04	0.9987	17.3076	5.5805	22.8882	0.0038
		Q6_K	130.42	130.50	0.9957	31.3821	17.3731	48.7552	0.0205
		Q5_K_M	130.42	130.31	0.9925	41.3732	17.3657	58.7389	0.0273
		Q4_K_M	130.42	130.72	0.9803	67.3439	17.4300	84.7740	0.0121
		Q3_K_M	130.42	133.79	0.9453	113.8700	17.8649	131.7349	0.1651
		Q2_K	130.42	134.50	0.8356	197.6929	17.8740	215.5669	0.4977
	BnB	INT8	130.42	131.29	0.9872	54.5359	0	54.5359	0.0277
		NF4	130.42	133.20	0.9491	109.5658	0	109.5658	0.1732
		FP4	130.42	135.18	0.8904	162.0623	0	162.0623	0.2850
	GPTQ	GPTQ-4bit	130.42	132.76	0.9408	117.8484	0	117.8484	0.0075
ARC	–	FP16	129.12	129.02	0.9998	5.9643	0	5.9644	0.0005
($\mathbf{r_s}$=**0.95**)	GGUF	Q8_0	129.12	128.88	0.9990	15.0077	5.5150	20.5226	0.0007
		Q6_K	129.12	129.06	0.9960	29.9552	17.1525	47.1076	0.0054
		Q5_K_M	129.12	128.94	0.9925	41.1496	17.1540	58.3036	0.0197
		Q4_K_M	129.12	130.17	0.9790	69.1055	17.3378	86.4434	0.0612
		Q3_K_M	129.12	133.16	0.9432	115.3884	17.7527	133.1411	0.1094
		Q2_K	129.12	133.96	0.8232	203.6563	17.7987	221.4549	0.3300
	BnB	INT8	129.12	129.36	0.9862	55.8249	0	55.8249	0.0028
		NF4	129.12	132.24	0.9448	113.1453	0	113.1453	0.1359
		FP4	129.12	135.31	0.8830	166.9927	0	166.9927	0.2899
	GPTQ	GPTQ-4bit	129.12	132.03	0.9378	119.9134	0	119.9134	0.0460
TriviaQA	–	FP16	143.70	143.69	0.9998	8.0572	0	8.0572	0.0008
($\mathbf{r_s}$=**0.96**)	GGUF	Q8_0	143.70	143.70	0.9991	16.1854	6.1872	22.3726	0.0069
		Q6_K	143.70	143.67	0.9953	36.1650	19.4953	55.6603	0.0073
		Q5_K_M	143.70	143.78	0.9908	50.5854	19.5117	70.0971	0.0379

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		Q4_K_M	143.70	143.72	0.9723	87.8958	19.5003	107.3961	0.1682
		Q3_K_M	143.70	144.11	0.9186	150.9558	19.5505	170.5063	0.3294
		Q2_K	143.70	144.13	0.7514	263.7436	19.5567	283.3003	1.0967
	BnB	INT8	143.70	143.85	0.9799	74.9842	0	74.9842	0.1057
		NF4	143.70	143.85	0.9350	134.6969	0	134.6969	0.0959
		FP4	143.70	144.30	0.8695	191.1835	0	191.1835	0.6411
	GPTQ	GPTQ-4bit	143.70	144.15	0.9032	164.5857	0	164.5857	0.4587
SQuAD	–	FP16	134.00	134.00	0.9998	7.2905	0	7.2905	0.0023
($\mathbf{r_s}$=**0.52**)	GGUF	Q8_0	134.00	133.94	0.9973	25.3724	5.7383	31.1106	0.0149
		Q6_K	134.00	133.72	0.9932	40.4765	18.0566	58.5331	0.0564
		Q5_K_M	134.00	133.80	0.9891	51.3145	18.0684	69.3829	0.0636
		Q4_K_M	134.00	133.67	0.9740	79.3096	18.0431	97.3527	0.1320
		Q3_K_M	134.00	134.80	0.9373	123.7284	18.1983	141.9266	0.1077
		Q2_K	134.00	135.88	0.8131	214.4921	18.3499	232.8420	0.0144
	BnB	INT8	134.00	134.27	0.9825	65.2211	0	65.2211	0.0190
		NF4	134.00	134.35	0.9427	118.0269	0	118.0269	0.2955
		FP4	134.00	135.40	0.8968	159.0941	0	159.0941	0.2321
	GPTQ	GPTQ-4bit	134.00	137.68	0.9188	142.5730	0	142.5730	0.4087
GSM8K	–	FP16	124.85	124.79	0.9998	6.7882	0	6.7882	0.0017
($\mathbf{r_s}$=**0.81**)	GGUF	Q8_0	124.85	124.79	0.9981	20.0711	5.2866	25.3577	0.0055
		Q6_K	124.85	124.63	0.9947	33.4395	16.6108	50.0503	0.0056
		Q5_K_M	124.85	124.52	0.9913	42.7436	16.5981	59.3417	0.0077
		Q4_K_M	124.85	124.59	0.9776	68.6009	16.6073	85.2082	0.0040
		Q3_K_M	124.85	125.49	0.9452	107.6943	16.7478	124.4420	0.0316
		Q2_K	124.85	126.71	0.8359	187.3607	16.9196	204.2803	0.0087
	BnB	INT8	124.85	125.14	0.9849	56.5404	0	56.5404	0.0081
		NF4	124.85	124.98	0.9478	104.9387	0	104.9387	0.0086
		FP4	124.85	126.58	0.9044	142.9376	0	142.9376	0.0210
	GPTQ	GPTQ-4bit	124.85	128.00	0.9347	118.9811	0	118.9811	0.0327

Table 13: Geometric decomposition for **Llama-3.1-8B-Instruct** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	144.62	144.62	0.9997	8.6575	0	8.6575	0.0006
($\mathbf{r_s}$=**0.71**)	GGUF	Q8_0	144.62	144.52	0.9955	35.4732	6.1943	41.6675	0.0225
		Q6_K	144.62	144.62	0.9924	46.2823	19.4739	65.7563	0.0147
		Q5_K_M	144.62	144.61	0.9852	64.7792	19.4811	84.2603	0.0134
		Q4_K_M	144.62	143.10	0.9402	129.4149	19.1528	148.5677	0.1047
		Q3_K_M	144.62	141.89	0.8549	200.7990	18.9551	219.7541	0.0306
		Q2_K	144.62	144.23	0.7332	274.3229	19.4327	293.7556	0.6372
	BnB	INT8	144.62	144.43	0.9772	80.2094	0	80.2094	0.0423
		NF4	144.62	145.40	0.9156	154.9492	0	154.9492	0.1038
		FP4	144.62	144.76	0.8619	197.6982	0	197.6982	0.0551
	GPTQ	GPTQ-4bit	144.62	143.76	0.8709	190.5542	0	190.5542	0.0867
ARC	–	FP16	145.61	145.62	0.9997	9.2035	0	9.2035	0.0007
($\mathbf{r_s}$=**0.52**)	GGUF	Q8_0	145.61	145.62	0.9971	28.6787	6.2640	34.9426	0.0076
		Q6_K	145.61	145.65	0.9937	42.4023	19.6986	62.1009	0.0160
		Q5_K_M	145.61	145.68	0.9882	58.1170	19.7166	77.8336	0.0122
		Q4_K_M	145.61	145.13	0.9583	109.1316	19.5783	128.7099	0.0093
		Q3_K_M	145.61	143.70	0.8825	182.3922	19.3790	201.7712	0.0044
		Q2_K	145.61	145.43	0.7465	269.4514	19.6428	289.0941	0.4684
	BnB	INT8	145.61	145.48	0.9796	76.4152	0	76.4152	0.0206

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		NF4	145.61	146.23	0.9208	151.0218	0	151.0218	0.0418
		FP4	145.61	146.17	0.8654	196.8382	0	196.8382	0.0813
	GPTQ	GPTQ-4bit	145.61	145.41	0.8791	186.0583	0	186.0583	0.0111
TriviaQA	–	FP16	143.92	143.94	0.9998	6.8100	0	6.8100	0.0010
($\mathbf{r_s}$=**0.97**)	GGUF	Q8_0	143.92	143.96	0.9994	13.4786	6.1746	19.6533	0.0021
		Q6_K	143.92	144.01	0.9968	29.9684	19.3978	49.3661	0.0210
		Q5_K_M	143.92	143.92	0.9925	45.8194	19.3799	65.1993	0.0654
		Q4_K_M	143.92	144.01	0.9780	78.4840	19.4013	97.8853	0.0653
		Q3_K_M	143.92	144.32	0.9287	141.4994	19.4554	160.9548	0.3116
		Q2_K	143.92	145.04	0.7730	253.1576	19.5960	272.7536	1.0636
	BnB	INT8	143.92	144.33	0.9842	66.5723	0	66.5723	0.0496
		NF4	143.92	144.84	0.9449	124.6002	0	124.6002	0.1588
		FP4	143.92	144.93	0.9044	164.2644	0	164.2644	0.3070
	GPTQ	GPTQ-4bit	143.92	144.14	0.9330	137.0685	0	137.0685	0.1271
SQuAD	–	FP16	138.77	138.77	0.9999	5.5351	0	5.5352	0.0026
($\mathbf{r_s}$=**0.91**)	GGUF	Q8_0	138.77	138.79	0.9987	18.5955	5.8932	24.4886	0.0044
		Q6_K	138.77	138.84	0.9971	27.4213	18.4556	45.8770	0.0162
		Q5_K_M	138.77	138.79	0.9949	36.4015	18.4439	54.8454	0.0138
		Q4_K_M	138.77	138.92	0.9868	58.5698	18.4742	77.0440	0.0170
		Q3_K_M	138.77	138.96	0.9667	93.1453	18.4843	111.6297	0.0719
		Q2_K	138.77	138.33	0.8987	162.1759	18.3676	180.5435	0.3164
	BnB	INT8	138.77	138.76	0.9916	46.6886	0	46.6886	0.0158
		NF4	138.77	138.86	0.9688	90.1510	0	90.1510	0.0507
		FP4	138.77	139.39	0.9454	119.4940	0	119.4940	0.1932
	GPTQ	GPTQ-4bit	138.77	138.92	0.9661	94.0248	0	94.0248	0.0639
GSM8K	–	FP16	139.11	139.10	0.9999	6.1426	0	6.1426	0.0004
($\mathbf{r_s}$=**0.90**)	GGUF	Q8_0	139.11	139.14	0.9991	15.0734	5.9007	20.9741	0.0007
		Q6_K	139.11	139.08	0.9972	26.8665	18.4980	45.3645	0.0019
		Q5_K_M	139.11	139.15	0.9936	40.8336	18.5109	59.3445	0.0087
		Q4_K_M	139.11	138.82	0.9816	69.2330	18.4532	87.6862	0.0028
		Q3_K_M	139.11	139.25	0.9484	116.2355	18.5245	134.7600	0.0215
		Q2_K	139.11	138.59	0.8532	195.6384	18.4046	214.0431	0.1025
	BnB	INT8	139.11	139.13	0.9894	52.7757	0	52.7757	0.0046
		NF4	139.11	139.03	0.9542	109.3880	0	109.3880	0.0274
		FP4	139.11	139.50	0.9179	146.7755	0	146.7755	0.0217
	GPTQ	GPTQ-4bit	139.11	139.54	0.9438	121.4749	0	121.4749	0.0308

Table 14: Geometric decomposition for **Qwen3-8B-Instruct** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	159.13	159.08	0.9999	8.2992	0	8.2992	0.0044
($\mathbf{r_s}$=**0.72**)	GGUF	Q8_0	159.13	159.36	0.9996	14.7119	12.7219	27.4338	0.0013
		Q6_K	159.13	160.11	0.9978	36.5955	41.0117	77.6072	0.0219
		Q5_K_M	159.13	159.95	0.9951	53.8170	40.9531	94.7701	0.0075
		Q4_K_M	159.13	157.24	0.9887	81.2492	40.2340	121.4832	0.0812
	BnB	INT8	159.13	161.48	0.9926	66.9423	0	66.9423	0.0584
		NF4	159.13	161.58	0.9772	117.0377	0	117.0377	0.0248
		FP4	159.13	161.02	0.9462	179.2144	0	179.2144	0.0456
	GPTQ	GPTQ-8bit	159.13	158.72	0.9996	14.4978	0	14.4978	0.0021
ARC	–	FP16	164.96	164.96	0.9999	8.0669	0	8.0669	0.0002
($\mathbf{r_s}$=**0.80**)	GGUF	Q8_0	164.96	165.18	0.9997	13.6871	13.1770	26.8642	0.0018

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		Q6_K	164.96	165.27	0.9982	33.3705	42.3272	75.6977	0.0289
		Q5_K_M	164.96	165.56	0.9961	49.6842	42.3762	92.0603	0.0044
		Q4_K_M	164.96	163.62	0.9907	76.7102	41.8486	118.5588	0.0895
	BnB	INT8	164.96	167.51	0.9933	66.3043	0	66.3043	0.0258
		NF4	164.96	165.86	0.9775	119.6165	0	119.6165	0.0211
		FP4	164.96	166.75	0.9530	173.4687	0	173.4687	0.1012
	GPTQ	GPTQ-8bit	164.96	164.40	0.9997	13.9945	0	13.9945	0.0032
TriviaQA	–	FP16	151.03	151.03	0.9998	9.1749	0	9.1750	0.0102
($\mathbf{r_s}$=**0.53**)	GGUF	Q8_0	151.03	151.15	0.9995	16.2266	12.1886	28.4151	0.0260
		Q6_K	151.03	150.68	0.9972	38.7761	38.4032	77.1793	0.0351
		Q5_K_M	151.03	151.43	0.9931	60.5157	38.5897	99.1053	0.0280
		Q4_K_M	151.03	149.73	0.9816	98.4393	38.1577	136.5970	0.0152
	BnB	INT8	151.03	151.55	0.9903	71.9123	0	71.9123	0.2055
		NF4	151.03	153.18	0.9567	152.7903	0	152.7903	0.5815
		FP4	151.03	151.52	0.9335	188.1664	0	188.1664	0.4596
	GPTQ	GPTQ-8bit	151.03	151.00	0.9995	16.1822	0	16.1822	0.0421
SQuAD	–	FP16	150.58	150.60	0.9998	10.5434	0	10.5434	0.0106
($\mathbf{r_s}$=**0.85**)	GGUF	Q8_0	150.58	150.57	0.9995	15.9146	12.0746	27.9892	0.0051
		Q6_K	150.58	150.07	0.9974	36.6765	38.3071	74.9836	0.0136
		Q5_K_M	150.58	151.31	0.9940	56.6100	38.6214	95.2314	0.1466
		Q4_K_M	150.58	150.35	0.9861	85.6449	38.3745	124.0193	0.1370
	BnB	INT8	150.58	151.00	0.9914	67.6118	0	67.6118	0.0658
		NF4	150.58	155.78	0.9640	141.2028	0	141.2028	0.2207
		FP4	150.58	154.60	0.9486	167.3859	0	167.3859	0.1603
	GPTQ	GPTQ-8bit	150.58	150.71	0.9995	15.9466	0	15.9466	0.0152
GSM8K	–	FP16	136.73	136.77	0.9996	12.3928	0	12.3928	0.0032
($\mathbf{r_s}$=$-$**0.02**)	GGUF	Q8_0	136.73	136.79	0.9991	19.3309	10.9439	30.2747	0.0058
		Q6_K	136.73	136.24	0.9960	41.5145	34.7398	76.2542	0.0006
		Q5_K_M	136.73	137.14	0.9907	63.5659	34.9687	98.5346	0.0044
		Q4_K_M	136.73	133.89	0.9771	99.1792	34.1249	133.3041	0.0007
	BnB	INT8	136.73	137.14	0.9862	77.6898	0	77.6898	0.0004
		NF4	136.73	138.06	0.9576	136.5375	0	136.5375	0.0139
		FP4	136.73	144.83	0.9345	175.8865	0	175.8865	0.0008
	GPTQ	GPTQ-8bit	136.73	136.70	0.9993	17.6924	0	17.6924	0.0026

Table 15: Geometric decomposition for **Ministral-3-8B-Instruct** under identity alignment ($W{=}I$) on all benchmarks. Each benchmark section reports Spearman's $r_s(\mathcal{B}, |\Delta\mathcal{R}|)$ across all quantization variants. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ when the head is preserved ($H_T{=}H_P$).

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
MMLU	–	FP16	189.49	189.49	1.0000	0.0000	0	0.0000	0.0000
($\mathbf{r_s}$=**0.76**)	GGUF	Q8_0	189.49	189.43	0.9969	13.8312	4.4501	18.2813	0.0037
		Q6_K	189.49	189.43	0.9960	15.7838	13.9847	29.7685	0.0170
		Q5_K_M	189.49	189.77	0.9936	19.8389	13.9930	33.8319	0.0267
		Q4_K_M	189.49	189.46	0.9836	31.8793	13.9821	45.8614	0.0119
		Q3_K_M	189.49	187.05	0.9626	47.8645	13.8643	61.7289	0.0236
		Q2_K	189.49	187.52	0.8843	84.1785	13.9004	98.0789	0.1117
	BnB	INT8	189.49	189.33	0.9953	16.9969	0	16.9969	0.0089
		NF4	189.49	189.62	0.9625	48.1688	0	48.1688	0.0049
		FP4	189.49	189.14	0.9441	58.7566	0	58.7566	0.0570
ARC	–	FP16	189.39	189.39	1.0000	0.0000	0	0.0000	0.0000
($\mathbf{r_s}$=**0.81**)	GGUF	Q8_0	189.39	189.29	0.9970	13.6593	4.4225	18.0818	0.0013
		Q6_K	189.39	189.20	0.9960	15.7277	13.8684	29.5961	0.0116

continued on next page

(continued from previous page)

Dataset	Family	Method	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
		Q5_K_M	189.39	189.43	0.9936	19.8754	13.8794	33.7548	0.0194
		Q4_K_M	189.39	189.60	0.9840	31.4817	13.8730	45.3547	0.0059
		Q3_K_M	189.39	187.28	0.9605	49.1690	13.7588	62.9279	0.0097
		Q2_K	189.39	187.77	0.8819	85.0600	13.7394	98.7994	0.1136
	BnB	INT8	189.39	189.25	0.9953	17.1177	0	17.1177	0.0038
		NF4	189.39	189.80	0.9607	49.3239	0	49.3239	0.0363
		FP4	189.39	188.99	0.9400	60.8374	0	60.8374	0.0722
TriviaQA	–	FP16	193.86	193.86	1.0000	0.0000	0	0.0000	0.0000
($\mathbf{r_s}$=**0.95**)	GGUF	Q8_0	193.86	193.78	0.9968	14.4688	4.3140	18.7829	0.0083
		Q6_K	193.86	193.70	0.9956	16.9176	13.5500	30.4676	0.0032
		Q5_K_M	193.86	193.82	0.9932	21.0329	13.5489	34.5819	0.0264
		Q4_K_M	193.86	193.86	0.9835	32.6625	13.5470	46.2095	0.0389
		Q3_K_M	193.86	193.84	0.9510	56.3049	13.5562	69.8611	0.1490
		Q2_K	193.86	194.74	0.8363	103.1718	13.5514	116.7232	0.5791
	BnB	INT8	193.86	193.95	0.9945	18.9217	0	18.9217	0.0204
		NF4	193.86	194.07	0.9634	48.7213	0	48.7213	0.0883
		FP4	193.86	194.53	0.9354	64.7848	0	64.7848	0.1946
SQuAD	–	FP16	184.45	184.45	1.0000	0.0000	0	0.0000	0.0000
($\mathbf{r_s}$=**0.65**)	GGUF	Q8_0	184.45	184.40	0.9966	14.0550	4.2532	18.3082	0.0261
		Q6_K	184.45	184.22	0.9961	15.1235	13.3434	28.4669	0.0046
		Q5_K_M	184.45	184.37	0.9938	18.9821	13.3481	32.3302	0.0275
		Q4_K_M	184.45	184.43	0.9862	28.4493	13.3458	41.7952	0.0092
		Q3_K_M	184.45	183.86	0.9646	45.5030	13.3320	58.8351	0.0225
		Q2_K	184.45	184.78	0.8958	78.2128	13.3350	91.5478	0.1201
	BnB	INT8	184.45	184.51	0.9959	15.5770	0	15.5770	0.0169
		NF4	184.45	184.62	0.9709	41.3127	0	41.3127	0.0662
		FP4	184.45	185.79	0.9518	53.3291	0	53.3291	0.0789
GSM8K	–	FP16	182.81	182.81	1.0000	0.0000	0	0.0000	0.0000
($\mathbf{r_s}$=**0.76**)	GGUF	Q8_0	182.81	182.74	0.9973	12.4551	4.1489	16.6040	0.0025
		Q6_K	182.81	182.62	0.9965	14.2621	13.0572	27.3192	0.0015
		Q5_K_M	182.81	182.60	0.9950	17.0314	13.0591	30.0905	0.0032
		Q4_K_M	182.81	182.70	0.9886	25.5626	13.0618	38.6244	0.0005
		Q3_K_M	182.81	182.25	0.9709	40.8767	13.0354	53.9121	0.0100
		Q2_K	182.81	183.87	0.9081	72.9412	13.0756	86.0168	0.0266
	BnB	INT8	182.81	182.84	0.9969	13.4258	0	13.4258	0.0002
		NF4	182.81	182.68	0.9773	36.1520	0	36.1520	0.0001
		FP4	182.81	184.08	0.9599	48.2423	0	48.2423	0.0076

F SIMILARITY BASELINES, COMPONENT ABLATION, AND HOW SMALL THE REFERENCE SET CAN BE

F.1 FULL-SIZE COMPARISON

Table 16 scores two standard representational-similarity metrics against the nested PRISM quantities on exactly the same footing: the same base models, an identical variant set for every column within each cell, feature matrices Z_T, Z_P extracted on $\mathcal{D}_{\text{diag}}$ under the identical protocol, and the same $|\Delta\mathcal{R}|$ targets. The two extraction runs are independent but agree closely: the Ω values recomputed here match the main pipeline's to three decimals (median absolute difference 2.7×10^{-5}), and the $|\Delta\mathcal{R}|$ targets are bit-identical, so the blocks of the table are paired cell by cell. Every entry averages Spearman correlations between a degradation score and $|\Delta\mathcal{R}|$ computed within one (family, benchmark) cell, so higher is better and all columns of a row are computed over an identical variant set. Linear CKA (Kornblith et al., 2019) and SVCCA (Raghu et al., 2017) are similarities, so we score $1-\text{CKA}$ and $1-\text{SVCCA}$; we also ran the Procrustes distance $\min_{W\in\mathcal{O}(d)} \|Z_T - Z_P W\|_F/\sqrt{n}$, which is already a distance, and it lands at 0.898 over the ten cells, indistinguishable from the other two.

Table 16: **Component-wise ablation and similarity baselines: every metric ranks variants about equally well.** Per-cell Spearman r_s between each degradation score and $|\Delta\mathcal{R}|$, averaged over the five benchmarks of each family; higher is better, and all columns are computed on identical features, identical variants and identical $|\Delta\mathcal{R}|$ targets. Top block: representational-similarity baselines. Middle/bottom blocks: the nested PRISM quantities (shape $\rightarrow$ + scale $\rightarrow$ + head) under the identity alignment $W{=}I$ and the Procrustes-optimal $W{=}W_N$. **Bold**: best in column (ties bolded). The last column is the SEM of the mean over the ten cells; it exceeds every between-metric gap in the table, so $\mathcal{B}_N$'s lead over the baselines is not resolved at this sample size (Appendix F). The main text adopts $W{=}I$ for SVD-free differentiability and the $H_T{=}H_P$ head-term simplification under a frozen `lm_head`.

Metric		Llama (5 benchmarks)	Qwen3 (5 benchmarks)	Mean r_s (10 cells)	$\pm$ SEM (10 cells)
$1-$CKA	(linear CKA; Kornblith et al., 2019)	0.924	**0.873**	0.899	0.025
$1-$SVCCA	(SVCCA; Raghu et al., 2017)	0.929	0.869	0.899	0.024
Ω	(shape similarity; baseline)	0.915	0.865	0.890	0.024
δ	(+ scale; feature alignment error)	0.910	0.865	0.888	0.024
$\mathcal{B}$	(+ head; full bound, $W{=}I$)	0.916	0.835	0.875	0.045
Ω_N	(shape similarity; baseline)	**0.933**	0.864	0.898	0.027
δ_N	(+ scale; feature alignment error)	**0.933**	0.864	0.898	0.027
$\mathcal{B}_N$	(+ head; full bound, $W{=}W_N$)	0.923	**0.889**	**0.906**	0.021

The baselines rank variants about as well as PRISM does (0.899, 0.899 and 0.898 over the ten cells, against 0.875 for $\mathcal{B}$ and 0.906 for $\mathcal{B}_N$), and none of the eighteen paired comparisons (three baselines $\times$ six PRISM quantities, $n{=}10$ paired cells) survives a multiplicity correction. The full bound is the clearest case: $|t| \leq 1.2$ and $p \geq 0.26$ against all three baselines under both alignments. The single nominally significant comparison is $1-$CKA over Ω, a gap of 0.009 at uncorrected $p{=}0.043$ (Wilcoxon $p{=}0.062$), which is what one expects to see once among eighteen tests. This is expected rather than disappointing: within one cell the variants differ mainly in how far the feature manifold has been distorted, and any statistic that responds monotonically to that distortion will produce nearly the same order.

What the nested ablation does and does not measure. Under $W{=}I$, the shape-only similarity Ω already reaches 0.890, adding scale to form δ leaves it at 0.888, and adding the head term to form $\mathcal{B}$ costs a further 0.013. Reading that as "the extra axes are redundant" would mistake what the ablation measures. It asks a single question, which nested quantity best orders variants *inside* one cell, and the answer is governed by what PTQ happens to do to these models, not by whether an axis is informative. Scale is the clearest case: across the 360 rows of the grid the shape term exceeds the scale term by a median factor of $\sim$7,700 and dominates in all but 5 rows, so $\Delta\rho$ has almost no variance to contribute to a pooled PTQ order. That is a finding about post-training quantization, not a defect of the axis: the same axis is the discriminating channel exactly where it does move (Qwen3-8B-Base Q2_K on GSM8K, Sec. 4.4) and under LoRA drift (Sec. 4.2), which is the point of keeping it separate rather than folded into a scalar.

The head term's effect on ranking depends on the alignment. Under $W{=}I$ the head term costs 0.013; under $W{=}W_N$, where the alignment absorbs the $H_T \rightarrow H_P$ rotation, the same term *adds* 0.008 ($\delta_N{=}0.898 \rightarrow \mathcal{B}_N{=}0.906$). So γ carries genuine signal, and what a pooled $W{=}I$ order penalizes is not head geometry but a protocol offset: $\gamma{>}0$ on the 200 of 360 rows whose GGUF k-quant tier quantizes `lm_head` and $\gamma{\equiv}0$ on GPTQ/BnB, nearly constant across bit-widths within a family. Interleaving two such regimes in one rank order is the questionable operation. Qwen3-8B-Base ARC, the cell with the largest gap against a baseline, shows it cleanly: pooled over both protocols $\mathcal{B}$ ranks poorly, yet within the FP16-head variants alone it ranks almost perfectly and within the GGUF variants alone clearly better than pooled. The split is not a general repair, since in some cells the GGUF subgroup is itself noisy, but it reveals protocol-dependent ranking behavior hidden by the pooled correlation, which a scalar similarity score cannot do. $\mathcal{B}$ remains our default: it is Theorem 1's certified upper bound and the only one of these metrics that retains that guarantee when $\gamma{>}0$. Ω and δ remain useful ranking scores, but δ alone does not retain the risk-bound guarantee when the head term is nonzero. The $W{=}I$ trace specialization trades $\sim$0.03 Spearman for SVD-free differentiability, frozen-`lm_head` head-term simplification, and consistency with the Sec. 3.4 regularizer, so the choice between the two alignments is primarily design-driven.

F.2 How the comparison changes with the size of the reference set

Appendix F.1 compares the scores at the paper's operating point, 512 sequences. The practically interesting question is the opposite one: how small can $\mathcal{D}_{\text{diag}}$ be before a score stops ordering variants correctly? This matters because it sets how much reference data, and how many forward passes, a screen needs.

Protocol. An independent round on Llama-3.1-8B over its 12 variants, with three seeds $(43/44/45)$ not used for the main quantization grid. Reference slices of $n \in \{8, 32, 128, 512\}$ sequences are nested by construction (shuffle, then take the first n), so a small slice is a subset of the large one and no distribution shift is introduced. Each entry is the Spearman correlation between the score computed on the n-sequence slice and the risk gap measured on the *full* 512-sequence slice (the decision-relevant question: can a cheap slice recover the ordering the full protocol would give?), reported as mean $\pm$ sd over the three seeds.

Table 17: **Eight reference sequences already fix the ordering; the risk gap measured on the same slice does not.** Llama-3.1-8B, 12 variants, seeds $43/44/45$; every column is r_s against $|\Delta\mathcal{R}|$ measured on the full 512-sequence slice, mean $\pm$ sd over seeds. The first column is $|\Delta\mathcal{R}|$ itself measured on the small slice; at $n{=}512$ it is the target and equals 1 by definition.

Reference size	$\lvert\Delta\mathcal{R}\rvert$ on the slice	$1-$CKA	$1-$SVCCA	**PRISM** δ $(W{=}I)$	**PRISM** $\mathcal{B}_N$
8 sequences	0.795 ± 0.059	0.903 ± 0.015	0.083 ± 0.036	0.930 ± 0.012	$\mathbf{0.932 \pm 0.016}$
32	0.825 ± 0.071	0.932 ± 0.010	0.821 ± 0.100	0.927 ± 0.012	0.931 ± 0.009
128	0.934 ± 0.033	0.934 ± 0.013	0.894 ± 0.001	0.927 ± 0.011	0.933 ± 0.011
512 (paper protocol)	1 (target)	0.931 ± 0.008	0.941 ± 0.014	0.925 ± 0.011	0.932 ± 0.011

The bound does not degrade; the risk gap and CKA do, and SVCCA leaves its domain of validity. $\mathcal{B}_N$ is flat at $r_s \approx 0.932$ from 8 sequences to 512, and the $W{=}I$ feature term δ, computed from the same stored features, is as flat (0.930 at $n{=}8$, 0.925 at $n{=}512$). The weakest small-sample statistic in the table is the risk gap measured on the slice itself: 0.795 at $n{=}8$ and 0.825 at $n{=}32$, reaching the bound's level only at $n{=}128$. This is the direct answer to why one would compute a bound when the same forward pass yields the gap (Sec. 4.3): on a small slice the bound orders the variants and the gap does not. The feature arm behaves the same way (δ_N: 0.924 ± 0.007 at $n{=}8$, 0.944 ± 0.012 at $n{=}512$). $1-$CKA loses 0.03 at $n{=}8$ but degrades gracefully. $1-$SVCCA reads 0.083, which we do not count as a comparison: canonical correlations need more samples than dimensions, and with fewer than d rows both centered feature matrices span the same subspace, so every canonical correlation is 1 by construction ($1-$SVCCA is numerically zero on ARC, MMLU, SQuAD and TriviaQA at $n{=}8$). The statistic recovers as the slice grows, and GSM8K, whose eight sequences already supply hundreds of tokens, is unaffected at every size.

Sequences are not tokens. The metrics live on answer-region tokens, and the answer is a single token for multiple-choice tasks, so a fixed number of *sequences* buys very different numbers of *rows* per benchmark. At 512 sequences on Llama we measure 511 answer tokens on MMLU, 527 on ARC, 1,524 on TriviaQA, 2,330 on SQuAD and 52,184 on GSM8K (Qwen3-8B-Base: 511, 527, 1,556, 2,515, 63,056). This is exactly why SVCCA works at every size on GSM8K and nowhere else at $n{=}8$: eight GSM8K sequences already supply $\sim$815 rows, while eight MMLU sequences supply eight. The requirement is therefore best stated in answer tokens rather than in sequences, and stating it that way also sharpens the eight-sequence result: on the multiple-choice benchmarks eight sequences means on the order of eight rows against $d{=}4096$, and the bound still recovers the full-slice ordering from them, which is the regime where a canonical-correlation statistic cannot function at all.

Qwen3-8B replication. The same round on Qwen3-8B-Base (11 variants, same seeds and sizes) gives the same picture at a lower ceiling: $\mathcal{B}_N$ is flat at 0.827 ± 0.035 ($n{=}8$) against 0.827 ± 0.031 ($n{=}512$); the risk gap on the slice reads 0.755 ± 0.109 at $n{=}8$, 0.878 at $n{=}32$ and 0.928 at $n{=}128$; $1-$CKA reads 0.819 ± 0.024 and $1-$SVCCA 0.427 ± 0.017 at $n{=}8$.

Provenance. This is a separate round from Table 16: different seeds, different 512-sequence draws, and, for the full bound, the $W{=}W_N$ arm only; the $W{=}I$ head term needs the quantized heads, which this round did not store, so the $W{=}I$ column reports the feature term δ. Its $n{=}512$ row therefore reproduces rather than repeats the Llama column of Table 16 ($1-$CKA 0.931 vs. 0.924, $1-$SVCCA 0.941 vs. 0.929, $\mathcal{B}_N$ 0.932 vs. 0.923), which also serves as a seed-robustness check on that table.

G TEACHER FORCING VERSUS DECODING: THE COST OF ONE SCREENING PASS

This appendix substantiates the efficiency claim of Sec. 4.2 with a controlled measurement. All timings are wall-clock seconds on one RTX 5090 in the same unoptimized HuggingFace runner, with no inference server such as vLLM on either side.

Teacher forcing versus decoding, measured directly. GSM8K generation dominates any benchmark run that includes it, so we measure it directly rather than assuming a generation budget. On 256 GSM8K prompts at batch size 4, with the same model and the same prompts on both sides:

	wall clock	**tokens**
Teacher-forced scoring $\times 1$ (the PRISM extraction call)	8.9 s	26,341 scored
Greedy decode $\times 1$ to natural EOS (ceiling 1024)	556.7 s	53,457 generated
ratio	$\geq 62.6\times$	

Three points on how to read this number. *(i) It is not a per-token comparison, and should not be.* The decode side emits roughly twice as many tokens as the teacher-forced side scores, because the model's natural chain of thought averages 208.8 tokens per prompt against 102 for the gold answer. That is not a confound to be normalized away: obtaining a GSM8K score *requires* generating the chain of thought, and its length is not known in advance. The comparison is per screening decision, which is the unit a practitioner actually pays for. *(ii) It is a lower bound.* $34/256$ prompts (13.3%) were still generating at the 1024-token ceiling, base models that never emit EOS; the remaining 222 averaged only 84 tokens. Truncation removes work from the decode side, so raising the ceiling would only increase the ratio. We report $\geq 62.6\times$ rather than re-running with a larger ceiling. *(iii) The gap is structural.* Teacher forcing scores a whole sequence in one forward pass; decoding needs one forward pass per generated token, about 209 in sequence here. Batching and a faster serving stack shrink both sides by a constant but leave this ratio. Greedy decoding is deterministic, and a repeat of the measurement agreed to within 0.1 s.

H QWEN3-8B FORGETTING AND SHAPE-REGULARIZATION REPLICATIONS

This appendix reports the Qwen3-8B replications of Sec. 4.2 and Sec. 4.5 along with the remaining decomposition tables (Llama BBQ and both Qwen fine-tuning tasks). Qwen3-8B differs from Llama-3.1-8B in depth (36 vs. 32 layers), vocabulary size (151,936 vs. 128,256), and pre-training corpus (multilingual, ∼36T tokens vs. English-dominant, ∼15T tokens); reproducing the main-text patterns under these shifts provides cross-family evidence for the claims of Secs. 4.2–4.5.

H.1 FORGETTING: BOUND TRACKS CROSS-TASK DRIFT ON QWEN3-8B

Figure 7 reproduces the two structural observations of the Llama main-text result (Fig. 3) on Qwen3-8B. First, *the PRISM bound rises in lockstep with empirical forgetting across most* $(task, benchmark)$ *combinations*: as LoRA fine-tuning proceeds on TruthfulQA or BBQ, $\Omega(Z_0, Z_t)$ drifts away from 1 and ρ_t from ρ_0, and this backbone drift tracks the empirical cross-entropy drift on most held-out benchmarks, including benchmarks (e.g., GSM8K) whose distributions differ substantially from the fine-tuning data. Second, *different source tasks induce qualitatively different drift geometries*, matching Llama: TruthfulQA fine-tuning drives predominantly shape drift while BBQ moves both scale and shape, a distinction the scale–shape split of Theorem 1 makes visible whereas a unified distance would collapse it. The replication holds in 8 of 10 cells; the two exceptions are analyzed next.

Per-subplot variability under low-forgetting regimes. A small subset of subplots in Fig. 7, most notably Qwen3-8B fine-tuned on BBQ evaluated on MMLU and TriviaQA, shows weak or slightly negative per-subplot r_s. In both cells the checkpoints have little to rank. On TriviaQA, $|\Delta\mathcal{R}|$ at $\lambda{=}0$ stays at order 10^{-3} nats per token ($\approx$0.0035 at step 300, against $\approx$0.288 on ARC under the same fine-tuning run; Table 21), two orders of magnitude below the scale at which the bound is informative across the rest of the grid. On MMLU, forgetting rises to $\approx$0.13 by step 75 and then

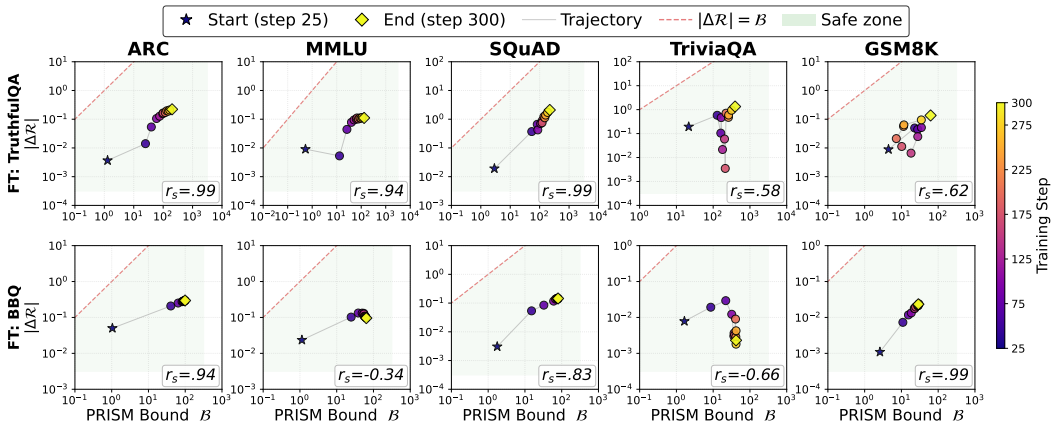

Figure 7: **Qwen3-8B: the PRISM bound tracks catastrophic forgetting across LoRA fine-tuning steps.** Each subplot scatters the PRISM bound $\mathcal{B}$ (x-axis, log) against the empirical forgetting $|\Delta\mathcal{R}|$ (y-axis, log) on a held-out benchmark, with one point per LoRA checkpoint colored by training step. Rows: fine-tuning task (TruthfulQA, BBQ). Columns: held-out evaluation benchmark (ARC, MMLU, SQuAD, TriviaQA, GSM8K). Because LoRA keeps the `lm_head` frozen, the head-discrepancy term γ vanishes at $W{=}I$ and the bound is governed entirely by backbone scale ($\Delta\rho$) and shape ($1-\Omega$) drift, as derived in Sec. 3.4. Spearman r_s per subplot is annotated in each panel.

stays within 0.11–0.13, so later checkpoints differ only by noise. In both cases the per-checkpoint r_s reflects checkpoint-trajectory noise rather than a failure of the bound, and is a property of the (model, fine-tuning task, benchmark) triple rather than of the diagnostic. The same noise-dominated effect appears in Llama-3.1-8B BBQ→SQuAD ($r_s \approx 0.30$, Fig. 3): $|\Delta\mathcal{R}|$ saturates early in fine-tuning, so late checkpoints sit at roughly the same level and the per-checkpoint Spearman is driven by checkpoint noise rather than continued drift.

H.2 Shape regularization: Qwen3-8B replication

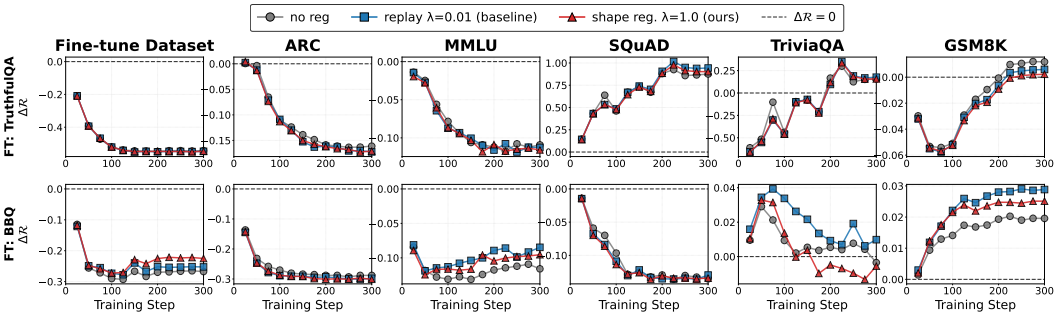

Figure 8: **Shape regularization vs. replay-CE on Qwen3-8B (replication of Fig. 4).** Same three configurations as Llama: *no reg*, *replay* (λ=0.01), and *shape regularizer* (λ=1.0). Qwen3-8B's baseline forgetting is already small (mean $|\Delta\mathcal{R}| < 0.27$ on both fine-tuning tasks vs. Llama's 0.84 on TruthfulQA), leaving little room for improvement; the regularizer-vs-replay gap narrows accordingly, and neither method changes forgetting beyond noise. Per-config decompositions: Tables 20 and 21.

Figure 8 reproduces the main-text Llama comparison (Fig. 4) on Qwen3-8B. Because Qwen3-8B's baseline forgetting is already small, both methods stay near the no-regularization level: the shape regularizer moves Ω by less than 10^{-3}, and neither method changes forgetting beyond noise. Per-config decompositions are in Tables 20 and 21 below.

Table 18: Regularization comparison for **Llama-3.1-8B** fine-tuned on **TruthfulQA** under identity alignment ($W{=}I$). Rows group by evaluation benchmark; each benchmark block lists metrics at step 300 for: no reg, replay $\lambda{=}0.01$ (baseline), shape regularizer $\lambda{=}1.0$ (ours). Bold / underline: 1st / 2nd-best smallest $|\Delta\mathcal{R}|$ and largest Ω per block. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ under frozen `lm_head`.

Dataset	Method (λ)	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
ARC	no reg	137.55	137.75	0.9158	147.6081	0	147.6081	0.0290
	replay $\lambda{=}0.01$ (baseline)	137.55	137.35	0.9185	145.0853	0	145.0853	**0.0108**
	shape reg. $\lambda{=}1.0$ (ours)	137.55	137.32	**0.9319**	132.5490	0	132.5490	0.0199
MMLU	no reg	138.91	139.50	0.9426	123.2722	0	123.2722	0.1339
	replay $\lambda{=}0.01$ (baseline)	138.91	138.94	0.9451	120.3033	0	120.3033	**0.1031**
	shape reg. $\lambda{=}1.0$ (ours)	138.91	139.09	**0.9542**	109.9868	0	109.9868	0.1116
SQuAD	no reg	138.20	140.07	0.9231	142.6639	0	142.6639	1.3372
	replay $\lambda{=}0.01$ (baseline)	138.20	139.58	0.9345	131.4408	0	131.4408	1.1911
	shape reg. $\lambda{=}1.0$ (ours)	138.20	139.04	**0.9478**	117.0684	0	117.0684	**1.0543**
TriviaQA	no reg	140.93	146.85	0.7593	261.3595	0	261.3595	2.5829
	replay $\lambda{=}0.01$ (baseline)	140.93	146.18	0.7810	248.6873	0	248.6873	2.3879
	shape reg. $\lambda{=}1.0$ (ours)	140.93	145.10	**0.8210**	223.8704	0	223.8704	**2.1242**
GSM8K	no reg	137.15	138.32	0.9905	49.6934	0	49.6934	0.1342
	replay $\lambda{=}0.01$ (baseline)	137.15	138.01	0.9957	33.3099	0	33.3099	0.1261
	shape reg. $\lambda{=}1.0$ (ours)	137.15	137.65	**1.0000**	1.3192	0	1.3192	**0.0966**

Table 19: Regularization comparison for **Llama-3.1-8B** fine-tuned on **BBQ** under identity alignment ($W{=}I$). Rows group by evaluation benchmark; each benchmark block lists metrics at step 300 for: no reg, replay $\lambda{=}0.01$ (baseline), shape regularizer $\lambda{=}1.0$ (ours). Bold / underline: 1st / 2nd-best smallest $|\Delta\mathcal{R}|$ and largest Ω per block. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ under frozen `lm_head`.

Dataset	Method (λ)	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
ARC	no reg	137.55	133.78	0.8804	173.7045	0	173.7045	**0.0775**
	replay $\lambda{=}0.01$ (baseline)	137.55	135.25	0.8802	174.6273	0	174.6273	0.1799
	shape reg. $\lambda{=}1.0$ (ours)	137.55	137.06	**0.9762**	78.3406	0	78.3406	0.1536
MMLU	no reg	138.91	135.04	0.8744	179.6944	0	179.6944	**0.3531**
	replay $\lambda{=}0.01$ (baseline)	138.91	136.63	0.8766	178.9974	0	178.9974	0.6061
	shape reg. $\lambda{=}1.0$ (ours)	138.91	138.66	**0.9758**	79.8590	0	79.8590	0.4995
SQuAD	no reg	138.20	139.04	0.9281	137.3737	0	137.3737	0.2608
	replay $\lambda{=}0.01$ (baseline)	138.20	139.16	0.9366	129.0819	0	129.0819	**0.2127**
	shape reg. $\lambda{=}1.0$ (ours)	138.20	138.34	**0.9694**	89.4630	0	89.4630	0.2903
TriviaQA	no reg	140.93	141.47	0.9780	77.4262	0	77.4262	0.1435
	replay $\lambda{=}0.01$ (baseline)	140.93	141.60	0.9784	76.6896	0	76.6896	0.1506
	shape reg. $\lambda{=}1.0$ (ours)	140.93	140.99	**0.9955**	34.9074	0	34.9074	**0.0171**
GSM8K	no reg	137.15	137.41	**1.0000**	0.6918	0	0.6918	0.0609
	replay $\lambda{=}0.01$ (baseline)	137.15	137.41	1.0000	0.6895	0	0.6895	0.0581
	shape reg. $\lambda{=}1.0$ (ours)	137.15	137.10	1.0000	0.1255	0	0.1255	**0.0126**

H.3 Additional decomposition tables

Tables 18–21 give the full per-(model, fine-tuning task) decomposition at step 300 for the three configurations (no reg, replay $\lambda{=}0.01$, shape regularizer $\lambda{=}1.0$); the main-text Table 2 compresses the Llama-TruthfulQA combination to Ω and $|\Delta\mathcal{R}|$ only.

I Additional regularization baselines: EWC, L2-SP and layer freezing

Sec. 4.5 compares the shape regularizer against replay-CE, which isolates shape preservation from data re-fitting because both methods see the same $\mathcal{D}_{\text{ref}}$. Replay is not, however, the strongest continual-learning baseline available. This appendix adds the two standard weight-space penalties,

Table 20: Regularization comparison for **Qwen3-8B-Base** fine-tuned on **TruthfulQA** under identity alignment ($W{=}I$). Rows group by evaluation benchmark; each benchmark block lists metrics at step 300 for: no reg, replay $\lambda{=}0.01$ (baseline), shape regularizer $\lambda{=}1.0$ (ours). Bold / underline: 1st / 2nd-best smallest $|\Delta\mathcal{R}|$ and largest Ω per block. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ under frozen `lm_head`.

Dataset	Method (λ)	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
ARC	no reg	332.98	327.46	0.9963	100.0550	0	100.0550	**0.1612**
	replay $\lambda{=}0.01$ (baseline)	332.98	327.21	0.9959	105.4953	0	105.4953	0.1725
	shape reg. $\lambda{=}1.0$ (ours)	332.98	327.29	**0.9965**	97.1908	0	97.1908	0.1708
MMLU	no reg	333.53	328.76	0.9984	67.4051	0	67.4051	**0.1092**
	replay $\lambda{=}0.01$ (baseline)	333.53	328.53	0.9981	71.7638	0	71.7638	0.1125
	shape reg. $\lambda{=}1.0$ (ours)	333.53	328.70	**0.9985**	64.5258	0	64.5258	0.1158
SQuAD	no reg	297.68	292.04	0.9930	122.0758	0	122.0758	**0.8757**
	replay $\lambda{=}0.01$ (baseline)	297.68	292.31	0.9926	125.2384	0	125.2384	0.9437
	shape reg. $\lambda{=}1.0$ (ours)	297.68	293.13	**0.9932**	120.3011	0	120.3011	0.9025
TriviaQA	no reg	240.42	224.70	0.9667	214.4331	0	214.4331	**0.1552**
	replay $\lambda{=}0.01$ (baseline)	240.42	223.97	0.9647	220.7355	0	220.7355	0.1767
	shape reg. $\lambda{=}1.0$ (ours)	240.42	225.12	**0.9674**	212.1221	0	212.1221	0.1575
GSM8K	no reg	269.06	271.54	**1.0000**	8.5518	0	8.5518	0.0118
	replay $\lambda{=}0.01$ (baseline)	269.06	272.36	1.0000	11.3864	0	11.3864	0.0060
	shape reg. $\lambda{=}1.0$ (ours)	269.06	272.50	1.0000	11.8835	0	11.8835	**0.0023**

Table 21: Regularization comparison for **Qwen3-8B-Base** fine-tuned on **BBQ** under identity alignment ($W{=}I$). Rows group by evaluation benchmark; each benchmark block lists metrics at step 300 for: no reg, replay $\lambda{=}0.01$ (baseline), shape regularizer $\lambda{=}1.0$ (ours). Bold / underline: 1st / 2nd-best smallest $|\Delta\mathcal{R}|$ and largest Ω per block. Shading: $\Omega{<}0.80$ / $\Omega{<}0.95$ on $(\Omega, \delta, \mathcal{B}, |\Delta\mathcal{R}|)$; $\gamma{=}0$ under frozen `lm_head`.

Dataset	Method (λ)	ρ_T	ρ_P	Ω	δ	γ	PRISM $\mathcal{B}$	$\lvert\Delta\mathcal{R}\rvert$
ARC	no reg	332.98	325.46	0.9976	82.4401	0	82.4401	**0.2881**
	replay $\lambda{=}0.01$ (baseline)	332.98	324.38	0.9967	96.6146	0	96.6146	0.2975
	shape reg. $\lambda{=}1.0$ (ours)	332.98	327.55	**0.9987**	60.7069	0	60.7069	0.3009
MMLU	no reg	333.53	328.73	0.9991	52.3490	0	52.3490	0.1159
	replay $\lambda{=}0.01$ (baseline)	333.53	328.17	0.9987	61.9216	0	61.9216	**0.0847**
	shape reg. $\lambda{=}1.0$ (ours)	333.53	330.08	**0.9992**	46.8071	0	46.8071	0.0954
SQuAD	no reg	297.68	286.65	0.9985	67.1415	0	67.1415	0.1353
	replay $\lambda{=}0.01$ (baseline)	297.68	285.88	0.9982	72.9559	0	72.9559	**0.1296**
	shape reg. $\lambda{=}1.0$ (ours)	297.68	287.67	**0.9991**	55.9438	0	55.9438	0.1342
TriviaQA	no reg	240.42	235.59	0.9992	36.1152	0	36.1152	**0.0035**
	replay $\lambda{=}0.01$ (baseline)	240.42	235.16	0.9989	42.8307	0	42.8307	0.0098
	shape reg. $\lambda{=}1.0$ (ours)	240.42	236.13	**0.9994**	33.0855	0	33.0855	0.0055
GSM8K	no reg	269.06	261.82	**1.0000**	25.0486	0	25.0486	**0.0196**
	replay $\lambda{=}0.01$ (baseline)	269.06	260.60	1.0000	29.2693	0	29.2693	0.0288
	shape reg. $\lambda{=}1.0$ (ours)	269.06	262.96	1.0000	21.1073	0	21.1073	0.0251

EWC (Kirkpatrick et al., 2017) with a diagonal Fisher and L2-SP (Li et al., 2018), and one architectural control, layer freezing (LoRA adapters on the top-K layers only, $K \in \{4, 8, 16\}$).

Protocol, and why plasticity has to be matched. Every method runs under the paper's exact fine-tuning protocol (Sec. 4.1): LoRA rank 32, lr 10^{-5}, 300 steps, identical $\mathcal{D}_{\text{ref}}$ and identical checkpoint cadence, differing only in the penalty. EWC's Fisher is accumulated over the LoRA-trainable parameters on $\mathcal{D}_{\text{ref}}$ at the adapter's initialization, where the model equals the base model and is therefore at the optimum of the "old task"; the Fisher is normalized to unit mean so that EWC and L2-SP share one λ scale. A forgetting number is meaningless without the target-task loss beside it, since any method can drive downstream drift to zero by not learning, so each method is reported at the sweep configuration whose target-task loss comes closest to the shape-regularized run's, and the target loss is printed in the table. Three seeds $(42/43/44)$; downstream forgetting is the unweighted mean $|\Delta\mathcal{R}|$ over the five held-out benchmarks at step 300.

Table 22: **Regularizer baselines at matched plasticity** (Llama-3.1-8B, TruthfulQA, three seeds, mean ± sd). Downstream forgetting is the mean $|\Delta\mathcal{R}|$ over the five held-out benchmarks; lower is better. Target-task loss is the answer-only CE on the fine-tuning task, and is what makes the comparison honest: layer freezing reaches a lower gap only by learning the task less.

Method (TruthfulQA, 3 seeds)	**Downstream forgetting ↓**	**Target-task loss**
no regularization	0.815 ± 0.042	0.950 ± 0.008
replay-CE, λ=0.01	0.771 ± 0.027	0.899 ± 0.007
L2-SP, λ=0.01	0.763 ± 0.019	0.901 ± 0.003
EWC (diagonal Fisher), λ=0.1	0.751 ± 0.015	0.905 ± 0.010
shape regularizer (ours), $\boldsymbol{\lambda}$=1.0	$\mathbf{0.680 \pm 0.016}$	$\mathbf{0.872 \pm 0.005}$
layer freezing, top 16	0.404 ± 0.017	0.924 ± 0.001

Result. Among the methods that learn the task to a comparable level, the shape regularizer gives the lowest downstream forgetting, it wins in each of the three seeds individually by 0.050 to 0.157, and it does so while carrying the *lowest* target-task loss in the table (0.872 against 0.899–0.950), so the reduction does not come at the expense of plasticity. The ordering among the baselines is itself informative: the two weight-space penalties (L2-SP 0.763, EWC 0.751) improve on replay only slightly, which is what the decomposition predicts. Under frozen-`lm_head` LoRA the bound reduces to ρ and Ω on the backbone, and constraining weights constrains features only indirectly, whereas the shape regularizer contracts the shape term itself.

Layer freezing is the exception that proves the point of matching plasticity. Its raw number, 0.404, is by far the best in the table, but at a target-task loss of 0.924 against the shape-regularized run's 0.872: it reaches retention by learning the fine-tuning task less. Reported without the second column it would look like the strongest method here; reported with it, it is a different point on the plasticity–stability trade-off rather than a better one.

Consistency with the main text. These runs are a separate three-seed round; the paper's Table 2 reports seed 42 alone. The two agree: seed-42 no-reg / replay / shape regularizer read 0.843 / 0.764 / 0.681 against the three-seed 0.815 ± 0.042 / 0.771 ± 0.027 / 0.680 ± 0.016, i.e. every main-text value falls inside its three-seed band, and the shape-regularizer figure the paper's headline rests on reproduces to 0.001.

Sensitivity to the reference draw. The main-text run uses one draw of 32 reference sequences. Holding everything else fixed (Llama-3.1-8B, TruthfulQA, λ=1.0, seed 42, step 300), we redraw $\mathcal{D}_{\text{ref}}$ as four pairwise-disjoint windows of one shuffle of the task's held-out split, at $n \in \{8, 16, 32\}$. Mean downstream $|\Delta\mathcal{R}|$ over the four draws is 0.690 ± 0.019 at n=8, 0.685 ± 0.011 at n=16 and 0.676 ± 0.008 at n=32; all twelve runs fall in 0.666–0.716, i.e. 15–21% below the no-regularization 0.843, and the spread shrinks with n as sampling variation should. Re-running the paper's own draw gives 0.683 against the reported 0.681, so the draw effect is small but real, and the specific 32 sequences are not load-bearing.

Scope. Methods for sequential continual instruction tuning target a stream of tasks rather than the single fine-tuning run studied here, so a fair comparison would require their benchmarks and protocols; that setting is listed among the future directions of Appendix K.

J When shape regularization helps: task dependence

TruthfulQA-FT, where shape drift dominates downstream forgetting, illustrates the regularizer's intended use case: the shape regularizer cuts mean $|\Delta\mathcal{R}|$ from 0.843 (no reg) to 0.681 (-19%), outperforming experience replay (0.764, -9%). BBQ-FT serves as a robustness check in the small-$|\Delta\mathcal{R}|$ regime: averaging the per-benchmark $|\Delta\mathcal{R}|$ values from Table 19 gives means of 0.179 (no reg), 0.195 (shape reg.), and 0.241 (replay), substantially smaller than TruthfulQA's 0.843, reflecting BBQ's modest baseline shape drift (mean baseline $\Omega = 0.932$ vs 0.906). The regularizer is therefore ahead of replay but, in aggregate, 9% above no regularization; the rest of this appendix explains why, and why the same regularizer is neutral on Qwen3-8B.

Geometric mechanism remains intact. On BBQ, the contrast between the shape regularizer and experience replay is most visible in the geometric signal: the shape regularizer lifts Ω from 0.93 to

0.98 (a 73% reduction in $1-\Omega$) while experience replay leaves Ω flat ($0.93 \rightarrow 0.93$). The regularizer's targeted quantity moves as designed.

Why BBQ is worse in aggregate. The shape regularizer's largest $|\Delta\mathcal{R}|$ reductions appear where forgetting is substantive: on BBQ-FT (Table 19), TriviaQA $|\Delta\mathcal{R}|$ drops from 0.143 to 0.017 (-88%) and GSM8K from 0.061 to 0.013 (-79%); on TruthfulQA-FT (Table 2), SQuAD from 1.337 to 1.054 (-21%) and TriviaQA from 2.583 to 2.124 (-18%). The BBQ aggregate is dominated by ARC and MMLU, where the no-reg run shows substantial shape drift (Ω drops to 0.88 and 0.87) but small forgetting ($|\Delta\mathcal{R}|$ of 0.077 and 0.353): on these cells shape drift does not translate into $|\Delta\mathcal{R}|$, so restoring shape has little forgetting to remove, and both regularizers add some (the shape regularizer less than replay). A regularizer that targets $1-\Omega$ helps on the cells where $1-\Omega$ and $|\Delta\mathcal{R}|$ move together, which the decomposition of Sec. 4.4 identifies before training.

Qwen3-8B replication. The Qwen3-8B replication (Tables 20–21, Fig. 8) extends the regime dependence from *tasks* (TruthfulQA-FT vs. BBQ-FT on Llama) to *models*. Qwen3-8B is markedly more robust to LoRA forgetting on these two fine-tuning sources than Llama-3.1-8B: across-benchmark mean $|\Delta\mathcal{R}|$ at $\lambda{=}0$ is 0.263 on Qwen TruthfulQA-FT vs. Llama's 0.843 ($3.2\times$ smaller), and 0.112 on Qwen BBQ-FT vs. Llama's 0.179. Baseline shape preservation is correspondingly closer to its ceiling on Qwen (Ω at $\lambda{=}0$: 0.991 on Qwen TruthfulQA-FT, 0.999 on Qwen BBQ-FT vs. Llama's 0.906 and 0.932), leaving the shape regularizer with essentially no drift to repair: $1-\Omega$ collapses from Llama's 0.094/0.068 to Qwen's 0.009/0.001, an order of magnitude smaller. The differences across no-reg, replay, and the shape regularizer correspondingly shrink to the evaluation-noise floor: across-benchmark mean $|\Delta\mathcal{R}|$ on Qwen TruthfulQA-FT is 0.263 / 0.282 / 0.270 (no-reg / replay / shape reg.), and on Qwen BBQ-FT is 0.112 / 0.110 / 0.112, with method-to-method spread under 0.02 on both. We report the Qwen3-8B regularization numbers in full for transparency and reproducibility, but emphasize that the shape regularizer's mechanism, contracting $1-\Omega$ at the source of LoRA backbone drift, can demonstrate a measurable benefit only when $1-\Omega$ has substantive headroom to contract. On Qwen3-8B there is no shape drift to repair, so these numbers do not contradict the Llama TruthfulQA-FT result; they confirm the regime dependence the bound itself predicts.

K FUTURE WORK

- **Beyond LoRA forgetting.** PRISM's diagnostic and regularization roles (Sec. 4.5) extend naturally to broader fine-tuning (full SFT, distillation, continual learning), where backbone drift is substantially larger. The shape regularizer constrains drift; Ω and $\Delta\rho$ together monitor it per step in ways validation loss does not.
- **Diagnostic applications.** The single-forward-pass computation of $\mathcal{B}$ enables three deployment uses: *per-sample OOD detection* via scale and shape residuals on individual inputs; *cross-scale hyperparameter transfer* between a small proxy and its scaled target to predict tuning robustness; and *production drift monitoring* on live traffic.
- **Beyond LLMs.** Vision Transformers and contrastive image encoders share LLMs' backbone-then-linear-head structure, with final-layer LayerNorm (and strict L2 normalization in CLIP/SigLIP) enforcing tight feature scales, making the PRISM scale axis naturally small. Natural targets include ViT quantization, vision-encoder distillation, and cross-modal alignment (consistent with the Platonic Representation Hypothesis (Huh et al., 2024)).