Annotation Entropy Predicts Per-Example Learning Dynamics in LoRA Fine-Tuning

Brady Steele
Georgia Institute of Technology
bsteele45@gatech.edu

Abstract

Annotator disagreement on tasks like natural language inference (NLI) reflects genuine linguistic ambiguity, yet most fine-tuning recipes treat every example as equally learnable. We ask whether this external signal of ambiguity predicts *per-example* learning dynamics under LoRA, the most widely used parameter-efficient fine-tuning method, and find that it does. Correlating annotation entropy (from ChaosNLI's 100 labels per example) with per-example area under the loss curve (AULC) on SNLI and MNLI, the correlation is positive in all 25 conditions tested (Spearman ρ=0.06–0.43), with decoder-only models showing stronger correlations than encoders at matched LoRA rank. More strikingly, under LoRA contested examples exhibit *unlearning*: their gold-label loss *increases* during training, a pattern that is largely absent under full fine-tuning and IA^3 in the encoder setting where matched comparisons are available, and that we also observe under LoRA on two decoder-only models. The effect survives partial-correlation controls and replicates across seeds and datasets. A preliminary noise-injection experiment is consistent with these findings.

1 Introduction

Neural networks learn training examples in a consistent order: "easy" examples are acquired before "hard" ones (Bengio et al., 2009; Arpit et al., 2017; Hacohen et al., 2020). Training dynamics methods exploit this regularity (Swayamdipta et al., 2020; Toneva et al., 2019), but define difficulty through model-internal measures, an endogenous approach, since the metrics are derived from the very training run whose behavior they seek to explain.

Meanwhile, annotator disagreement reflects genuine linguistic ambiguity (Pavlick and Kwiatkowski, 2019; Plank, 2022; Uma et al., 2021). ChaosNLI (Nie et al., 2020) re-annotates a subset of SNLI and MNLI with *100 labels per example*, producing a dense empirical distribution over class labels for each item. This makes ChaosNLI an unusually clean *external* probe for example ambiguity: because the full annotator distribution is recorded, the entropy of that distribution quantifies how much humans disagree independently of any model. We ask: does this external measure predict which examples a model learns first, and does the answer change under parameter-efficient fine-tuning?

LoRA (Hu et al., 2022) is the most widely adopted PEFT method, yet how its rank constraint shapes per-example learning order remains unexplored. Prior work studied what LoRA learns at convergence (Biderman et al., 2024) but not the acquisition trajectory.

Approach. We combine three ingredients: (i) annotation entropy from ChaosNLI as an external, model-independent measure of per-example ambiguity; (ii) per-example loss trajectories logged at dense checkpoints during fine-tuning, summarized by area under the loss curve (AULC) and by start-to-end loss change $\Delta\ell$; and (iii) a cross-product of six models (four encoder, two decoder-only), two datasets (SNLI, MNLI), and multiple LoRA ranks plus full fine-tuning and IA^3 baselines. This design lets us ask both a quantitative question (does entropy predict learning order?) and a qualitative one (does the *sign* of loss change on contested examples flip between LoRA and full fine-tuning?).

Contributions.

1. Under LoRA, contested examples exhibit *increasing* loss, a qualitative un-learning pattern we observe across all six architectures, and that is largely absent under full fine-tuning and IA^3 in the encoder setting where matched comparisons are available (§4.2).
2. Annotation entropy predicts per-example

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 129–141

learning dynamics across six models (four encoder, two decoder-only) and two NLI datasets, with stronger correlations for decoder-only models and monotonic scaling with LoRA rank (§4.1).

3. The correlation survives partial-correlation controls, cross-dataset replication, and alternative binning schemes; a preliminary noise-injection experiment provides additional suggestive evidence (§4.3, §5).

2 Background and Related Work

Training dynamics and example difficulty. Dataset Cartography (Swayamdipta et al., 2020) partitions examples by model confidence and variability; Toneva et al. (2019) showed forgetting events correlate with difficulty; Zhang and Wu (2024) further dissect per-example learning and forgetting during language model fine-tuning; and Hacohen et al. (2020) demonstrated that learning order is shared across architectures. Subsequent work uses per-example signals for data-efficient learning (Pleiss et al., 2020; Paul et al., 2021; Mindermann et al., 2022), extends cartography to LLM alignment (Lee et al., 2025), and identifies competitive dynamics between example subsets (Mircea et al., 2025; Zhao et al., 2024; Qi et al., 2025). However, Mandal (2025) identifies conditions under which cartography-based measures become unreliable, further motivating the use of external measures of example difficulty. All of these methods rely on model-derived statistics; we instead use an external ground truth for example ambiguity.

Annotator disagreement as signal. Pavlick and Kwiatkowski (2019) showed that NLI disagreement reflects genuine semantic ambiguity. Nie et al. (2020) collected 100 annotations per example for SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) subsets. Meissner et al. (2021) use ChaosNLI label distributions as soft training targets; we instead use annotation entropy as a purely diagnostic measure without modifying the training objective. Uma et al. (2021) survey learning-from-disagreement methods. A growing body of work treats annotator disagreement as structured signal rather than noise (Plank, 2022; Basile et al., 2021; Davani et al., 2022; Leonardelli et al., 2023, 2025). We use annotator disagreement not as a training signal but as an *external probe* of learning dynamics. NLI datasets also contain annotation artifacts that enable shortcut learning (Gururangan et al., 2018; McCoy et al., 2019); low-entropy examples may partly align with such shortcuts.

LoRA and parameter-efficient fine-tuning. LoRA (Hu et al., 2022) constrains weight updates to $\Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{BA}$ where $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$; lower ranks provide stronger regularization. It belongs to a broader PEFT family (see Ding et al. 2023 for a survey), with recent theoretical analyses (Rahimi Afzal et al., 2025) and empirical rank studies (Rathore et al., 2025). Biderman et al. (2024) found that LoRA learns less and forgets less than full fine-tuning at convergence; Sliwa et al. (2025) address forgetting with Laplace-based regularization. We complement this by studying the *dynamics* of per-example acquisition during training.

3 Method

Our pipeline has three stages. First, we compute *annotation entropy* for every example from its 100-label ChaosNLI distribution (§3.1). Second, during fine-tuning we log the per-example cross-entropy loss (computed against the majority-vote gold label, see §3.3) at dense checkpoints and summarize each trajectory with two scalars: area under the loss curve (AULC) and start-to-end loss change $\Delta\ell$ (§3.2). Third, we sweep this pipeline across six models, two NLI datasets, multiple LoRA ranks, full fine-tuning, and an IA^3 baseline (§3.3). Entropy is treated strictly as a diagnostic: it is never used as a training target or loss weight.

3.1 Annotation Entropy

For each example i in ChaosNLI with $K{=}100$ annotations per example across $C{=}3$ classes, we compute the empirical annotator distribution $\mathbf{p}_i$ where $p_{i,c} = n_{i,c}/K$ is the fraction of annotators choosing class c, and the annotation entropy:

$$\mathcal{H}_i = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c}, \tag{1}$$

with the convention $0 \log 0 = 0$. The entropy ranges from 0 (perfect agreement) to $\log 3 \approx 1.099$ nats (uniform distribution).

We partition examples into three categories based on entropy thresholds chosen to reflect qualitatively distinct agreement regimes in a 3-class setting: **clean** ($\mathcal{H}_i < 0.4$ nats; strong consensus, corresponding to distributions where $\geq$80% of annotators agree on one label), **ambiguous**

($0.4 \leq \mathcal{H}_i < 0.7$; moderate disagreement), and **contested** ($\mathcal{H}_i \geq 0.7$; near-uniform or heavily split annotation distributions). We verify in §4.3 that results are robust to percentile-based alternatives (quartile and tercile bins). The distribution of annotation entropy across the ChaosNLI–SNLI subset (n=1,514; Figure 3 in Appendix C) shows 50.6% ambiguous, 25.0% clean, and 24.4% contested examples.

3.2 Per-Example Learning Dynamics

We record per-example cross-entropy loss $\ell_i(t)$ at every 100 steps and at epoch boundaries across 5 epochs (39 checkpoints per run) and summarize each trajectory as the *Area Under the Loss Curve* (AULC):

$$\text{AULC}_i = \frac{1}{T} \sum_{t=1}^{T} \ell_i(t), \tag{2}$$

where T=39 and $\ell_i(t)$ is the standard cross-entropy loss against example i's single majority-vote *gold* label $y_i^\star$ at checkpoint t, *not* a KL against the 100-annotator distribution $\mathbf{p}_i$; a soft-label variant that targets $\mathbf{p}_i$ directly is reported as an ablation (§5). Lower AULC indicates faster learning. Unlike cartography confidence/variability (Swayamdipta et al., 2020) or forgetting events (Toneva et al., 2019), AULC is continuous, integrates over the entire trajectory, and naturally captures non-monotonic loss patterns. We also report loss change $\Delta\ell_i = \ell_i(T) - \ell_i(1)$ to directly capture un-learning. We quantify the entropy–dynamics relationship via Spearman $\rho(\text{AULC}, \mathcal{H})$; positive ρ means contested examples have higher AULC.

3.3 Experimental Setup

Data. We use two ChaosNLI subsets: the SNLI portion (∼1,514 examples) and the MNLI-matched portion (∼1,599 examples). Each subset is split 80/20 into train/validation, stratified by entropy category to ensure each split has representative examples from all three categories. For training, we combine ChaosNLI training examples with 20,000 standard examples randomly sampled from the corresponding bulk dataset (SNLI or MNLI from the GLUE benchmark) to provide sufficient training signal. We track per-example loss only on the ChaosNLI training examples, which are the subset with 100-annotator label distributions. Gold labels are majority-vote labels from the 100 annotators; loss is class-weighted cross-entropy (inverse class frequency weighting) to account for label imbalance. Validation accuracy is computed on the ChaosNLI validation split, where examples are inherently ambiguous (chance-level is 33%).

Models. We evaluate four encoder models, **RoBERTa-base** (Liu et al., 2019) (125M), **BERT-base** (Devlin et al., 2019) (110M), **DistilBERT** (Sanh et al., 2019) (66M), and **DeBERTa v3-base** (He et al., 2023) (183M), and two decoder-only models, **Qwen2.5-1.5B** and **Qwen2.5-3B** (Team, 2024). LoRA is applied to query and value projections for encoders; to query, key, value, and output projections for decoders (fine-tuned with a classification head and left-padding). DeBERTa v3 and decoder models are evaluated on SNLI only due to computational constraints.

LoRA configurations. Rank $r \in \{4, 16\}$ with $\alpha = 2r$ (constant scaling factor $\alpha/r = 2$) and dropout 0.05. Encoder models are also trained with full fine-tuning as a capacity upper bound; decoder models use LoRA only. For RoBERTa and BERT on SNLI, we additionally sweep $r \in \{1, 2, 4, 8, 16, 32\}$.

Training. AdamW, lr 2×10^{-5} (held constant across all configurations to isolate the effect of the rank constraint rather than learning rate tuning), cosine schedule with 6% warmup, 5 epochs, batch size 32, gradient clipping at 1.0. Each configuration is run across 3 seeds (42, 123, 456).

4 Results

4.1 Annotation Entropy Predicts Learning Order

Figure 1 presents the central finding. Clean examples show steadily decreasing loss; ambiguous examples plateau; contested examples exhibit *increasing* loss under LoRA, a pattern we term *unlearning*[1]: the model's loss on contested examples grows as it specializes toward high-agreement patterns.

Consistency across models and datasets. Table 1 quantifies this across all 25 conditions; the correlation is positive in every tested condition. On SNLI, 11 of 12 encoder and all 4 decoder conditions are significant ($p < 10^{-3}$); on MNLI, the

[1]Distinct from "machine unlearning" in the privacy literature. We use the term to denote a net loss increase from start to end of training ($\Delta\ell > 0$); individual trajectories may briefly decrease before increasing.

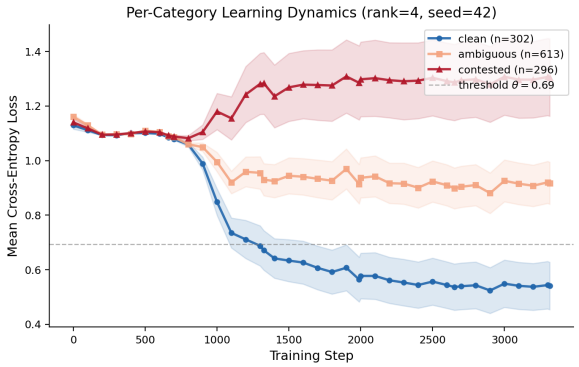

Figure 1: Per-example loss trajectories grouped by annotation entropy (RoBERTa, LoRA r=4, SNLI, seed 42). Clean examples show decreasing loss; contested examples exhibit *increasing* loss (un-learning). Lines: group means; bands: 95% CIs; dashed line: $-\log 0.5 \approx 0.693$.

correlation is weaker, with 3 conditions failing Bonferroni correction ($\alpha_{\text{corrected}} = 0.002$).[2] In total, 21 of 25 conditions are significant after Bonferroni correction.

Cross-model consistency and capacity scaling. All six models show the same qualitative pattern: higher entropy predicts slower learning. The correlation increases monotonically with LoRA rank in single-seed rank sweeps (Spearman rank–$\rho = +1.0$ for both RoBERTa and BERT; Table 6 in Appendix D), and the r=4 < r=16 < Full FT ordering holds for every encoder on both datasets. Decoder-only models show uniformly strong correlations (ρ=0.35–0.42, all $p < 10^{-30}$), exceeding encoder LoRA conditions at matched rank and extending both rank- and model-size scaling trends (Figure 6 in Appendix H).

4.2 LoRA vs. Full Fine-Tuning: Qualitative Contrast

Figure 2 reveals a qualitative difference: under full fine-tuning, all entropy categories show decreasing loss; under LoRA (Figure 1), contested examples show *increasing* loss. This is consistent with LoRA's rank constraint imposing a trade-off: as the adapter specializes toward high-agreement patterns, contested examples' loss increases above its initial value.

Quantifying un-learning. Table 2 reports mean $\Delta\ell$ from first to last checkpoint. Under LoRA, contested examples show consistent loss *increase*

[2]Under Benjamini–Hochberg at q=0.05, DistilBERT+LoRA r=4 MNLI would reach significance.

Model	Method	SNLI ρ	MNLI ρ
Encoder models			
RoBERTa	LoRA r=4	.308±.016	.140±.009
	LoRA r=16	.346±.019	.178±.005
	Full FT	.426±.015	.200±.027
BERT	LoRA r=4	.160±.052	.059±.015[c]
	LoRA r=16	.240±.011	.064±.023[c]
	Full FT	.386±.012	.145±.004
DistilB.	LoRA r=4	.209±.004	.083±.007[c]
	LoRA r=16	.272±.008	.105±.001
	Full FT	.358±.004	.131±.005
DeBERTa v3	LoRA r=4	.081±.016[c]	—
	LoRA r=16	.141±.012	—
	Full FT	.325±.006[d]	—
Decoder-only models			
Qwen 1.5B	LoRA r=4	.348±.020	—
	LoRA r=16	.389±.024	—
Qwen 3B	LoRA r=4	.383±.009	—
	LoRA r=16	.422±.007[e]	—
Alternative PEFT			
IA^3	RoBERTa	.108±.018	—

Table 1: Spearman ρ between AULC and annotation entropy (mean ± std across 3 seeds unless noted). All correlations significant at $p < 0.001$ except those marked [c] (fail Bonferroni correction). [d]Mean of 2 converged seeds (seed 456 reached chance-level accuracy). [e]Mean of 2 seeds (memory crash). IA^3 row: RoBERTa on SNLI only (§4.3).

across all four encoder models, while clean examples generally decrease (the exception is BERT on MNLI, where clean examples show slight loss increase under LoRA, suggesting limited learning overall in that condition). The same qualitative pattern, loss increase on contested examples versus decrease on clean ones, also holds for both decoder-only Qwen models under LoRA (Appendix B); we note, however, that matched full-FT baselines were infeasible for the decoder models due to compute constraints, so the decoder evidence establishes the *existence* of un-learning under LoRA in that architecture family but does not by itself show it is LoRA-specific there. In the encoder setting, where matched comparisons are available, full fine-tuning decreases loss for both categories; in that setting the un-learning phenomenon appears specific to the low-rank constraint. Importantly, this encoder contrast rules out calibration as the primary explanation: full fine-tuning produces *far more* peaked predictions than LoRA (mean prediction entropy 0.12 vs. 0.58 nats for LoRA r=4; Appendix K), yet contested examples' loss *decreases* rather than increases. The loss increase under LoRA there-

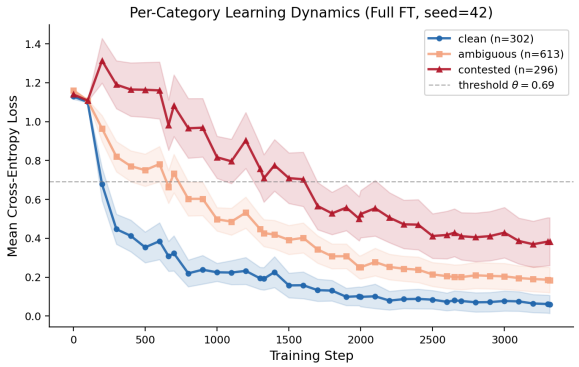

Figure 2: Per-example loss under full fine-tuning (RoBERTa, SNLI, seed 42). Unlike LoRA (Figure 1), all entropy categories show decreasing loss; the temporal ordering is preserved but contested examples are eventually fit. Dashed line: $-\log 0.5 \approx 0.693$.

fore reflects the rank constraint, not simply sharper predictions on weakly-supported labels.

4.3 Robustness and Controls

Multi-seed and cross-dataset replication. The correlation is stable across seeds for RoBERTa (coefficient of variation $<6\%$; per-seed breakdowns in Appendix A). The effect replicates from SNLI to MNLI at reduced magnitude (ρ=0.06–0.20; Table 1), with three low-rank conditions failing Bonferroni correction.

Partial correlation controls. Sentence length is uncorrelated with both entropy and AULC; the partial correlation controlling for length is virtually unchanged (ρ=0.303 vs. raw 0.304). A multiple regression (R^2=0.226) confirms entropy as the dominant predictor (β=0.294, $p \approx 0$), independent of sentence length and gold-label identity, indicating that the correlation is not driven by between-class differences in entropy.

Training-set composition and binning robustness. The AULC–entropy correlation is invariant to training-set entropy composition ($\rho \approx 0.306$–0.307 whether training on low-only, high-only, or balanced ChaosNLI subsets; Table 7 in Appendix E). Alternative binning schemes (quartile, tercile) preserve the monotonic AULC increase (Appendix G), and Kendall τ-b confirms all findings (τ/ρ ratio $= 0.68 \pm 0.01$, concordance r=0.9998).

Alternative PEFT method. IA^3 adapters (Liu et al., 2022) on RoBERTa–SNLI show a weaker correlation (ρ=0.108 $\pm$ 0.018; Table 1) and do *not* exhibit un-learning ($\Delta\ell$=-0.003 for contested examples; Table 2). Because this probe covers only a single architecture–dataset–adapter combination, we read it as *suggestive* rather than definitive evidence that un-learning is not a universal property of PEFT; the stronger claim, that un-learning is unique to LoRA among adapter methods, would require a broader sweep over PEFT variants (DoRA, prefix tuning, prompt tuning, . . .) and datasets, which we leave to future work (§5).

		SNLI		MNLI	
Model	**Method**	**Contest.**	**Clean**	**Contest.**	**Clean**
RoBERTa	LoRA r=4	$+.170_{\pm.021}$	$-.586_{\pm.032}$	$+.061_{\pm.014}$	$-.135_{\pm.025}$
	LoRA r=16	$+.161_{\pm.018}$	$-.675_{\pm.028}$	$+.086_{\pm.011}$	$-.254_{\pm.019}$
	Full FT	$-.751_{\pm.035}$	$-1.054_{\pm.019}$	$-.871_{\pm.042}$	$-1.069_{\pm.015}$
BERT	LoRA r=4	$+.076_{\pm.038}$	$-.135_{\pm.062}$	$+.039_{\pm.018}$	$+.099_{\pm.031}$
	LoRA r=16	$+.161_{\pm.025}$	$-.285_{\pm.041}$	$+.071_{\pm.022}$	$+.118_{\pm.028}$
	Full FT	$-.711_{\pm.029}$	$-1.016_{\pm.022}$	$-.760_{\pm.033}$	$-.860_{\pm.018}$
DistilB.	LoRA r=4	$+.058_{\pm.009}$	$-.335_{\pm.015}$	$+.011_{\pm.006}$	$-.011_{\pm.008}$
	LoRA r=16	$+.119_{\pm.012}$	$-.454_{\pm.021}$	$+.038_{\pm.010}$	$-.045_{\pm.013}$
	Full FT	$-.547_{\pm.024}$	$-.983_{\pm.011}$	$-.605_{\pm.031}$	$-.736_{\pm.016}$
DeBERTa v3	LoRA r=4	$+.019_{\pm.011}$	$-.048_{\pm.022}$	—	
	LoRA r=16	$+.050_{\pm.015}$	$-.161_{\pm.027}$	—	
	Full FT[d]	$-.047_{\pm.031}$	$-.751_{\pm.038}$	—	
Alternative PEFT (RoBERTa, SNLI only)					
IA^3		$-.003_{\pm.005}$	$-.154_{\pm.012}$	—	

Table 2: Mean loss change ($\Delta\ell$) from start to end of training for contested and clean examples (mean $\pm$ std across 3 seeds unless noted). Positive $\Delta\ell$ = un-learning. Under LoRA, contested examples consistently increase in loss; clean examples generally decrease, except BERT on MNLI where clean examples also show slight increases. Under full FT, all categories decrease. IA^3 shows near-zero $\Delta\ell$ for contested examples, indicating no un-learning. Decoder model $\Delta\ell$ values are reported in Appendix B. [d]Mean of 2 converged seeds.

5 Discussion

External grounding for training dynamics. Annotation entropy significantly predicts per-example learning order (Table 1), complementing model-internal measures (Swayamdipta et al., 2020; Toneva et al., 2019; Lee et al., 2025) with an external ground truth. We emphasize what this result is and is not. Cartography confidence correlates near-perfectly with AULC ($\rho > -0.99$; Appendix I), and entropy adds negligible incremental R^2 on top of cartography (ΔR^2=0.0001): it is *not* an independent predictor of per-example difficulty beyond what model-internal signals already capture. Instead, it provides two things that model-internal measures cannot. First, external validation: entropy is computed before any training run, so a tight AULC–entropy correlation rules out the concern that model-internal difficulty metrics are merely

reifying training noise (Mandal, 2025). Second, interpretability: it labels *why* an example is hard (annotator disagreement, i.e. genuine ambiguity) rather than merely flagging *that* it is hard. The value of entropy in this paper is therefore diagnostic rather than predictive; it is the lens that lets us read the un-learning result as a statement about linguistically ambiguous items, not a rival measure of difficulty.

Capacity and difficulty axes. The AULC–entropy correlation rises monotonically from $r{=}1$ ($\rho{=}0.233$) through LoRA $r{=}32$ and peaks at full fine-tuning ($\rho{=}0.416$; all seed 42). This is exactly what a pure difficulty-proxy reading would predict: at low rank the model barely differentiates examples, so every AULC looks similar and correlations with *any* difficulty axis are weak; with more capacity the model resolves difficulty more sharply, and entropy, which is correlated with difficulty, rides along (see §4.1). The quantitative correlation therefore does *not* by itself argue that un-learning is a low-rank phenomenon; the qualitative sign flip on $\Delta\ell$ for contested examples (positive under LoRA, negative under full FT; Table 2) is the part of the evidence that is specific to the low-rank constraint.

Loss increase under LoRA. Under LoRA, contested examples exhibit *increasing* loss across all tested models (with the LoRA-specificity interpretation anchored by the encoder–full-FT contrast in §4.2), going beyond the "patterns before memorization" narrative (Arpit et al., 2017), suggestive of competitive dynamics in capacity-constrained optimization (Mircea et al., 2025). To test this, we measure gradient cosine similarity between clean and contested example gradients over training (Appendix J). The results reveal a counterintuitive pattern: under LoRA $r{=}4$, gradients remain well-aligned throughout training (cosine similarity $\sim$0.80), while under full fine-tuning, alignment drops sharply ($\sim$0.95 $\rightarrow$ $\sim$0.25). Un-learning thus occurs despite high gradient alignment, suggesting the bottleneck is not gradient *direction* competition but the low-rank subspace's inability to simultaneously satisfy both groups' loss landscapes. Gradient-norm analysis (Figure 4 in Appendix F) shows contested examples produce $\sim$3.9$\times$ larger gradients yet their loss still increases. A soft-label ablation, training with KL-divergence against the full 100-annotator distribution instead of cross-entropy against majority-vote labels, rules out noisy gold labels as the driver: the un-learning pattern persists unchanged ($\Delta\ell{=}{+}0.176$ soft vs. $+0.170$ hard). Sliwa et al. (2025) independently document forgetting under LoRA, and Zibakhsh Shabgahi et al. (2026) show convergence metrics can dissociate from knowledge retention.

Alternative explanations. Bulk-data dominance cannot explain the effect since full fine-tuning uses identical data without producing un-learning. Class-weighting interactions are ruled out by partial correlations controlling for gold-label identity (§4.3). Inverse-class-frequency weighting could mechanically inflate loss for high-entropy examples if their majority-vote labels disproportionately fall in upweighted classes; however, the partial correlation controlling for gold-label identity absorbs this effect. Gold-label reliability is a potential confound: majority-vote labels for high-entropy examples are inherently less reliable, which could inflate loss independently of learning dynamics; our regression controls for gold-label identity (§4.3) but does not fully eliminate this concern. Calibration degradation may partly contribute: as predictions become more peaked, cross-entropy on weakly-supported labels naturally increases. However, calibration analysis (Appendix K) shows that full fine-tuning produces substantially more peaked predictions than LoRA (mean prediction entropy 0.12 vs. 0.58 nats for LoRA $r{=}4$, and mean max confidence 0.96 vs. 0.76), yet does not exhibit un-learning (Table 2), suggesting the rank constraint is the primary driver.

Decoder-only and scale. The correlation ordering $r{=}4 < r{=}16 <$ Full FT holds within every encoder, matching LoRA's expressive power increasing with rank (Zeng and Lee, 2024; Rahimi Afzal et al., 2025). Both decoder-only Qwen models (1.5B, 3B) extend the rank ordering at strictly larger absolute values, spanning a $45\times$ parameter range from 66M DistilBERT to 3B Qwen.

Adapter methods. The absence of un-learning under IA^3 (§4.3) is consistent with the phenomenon depending on LoRA's low-rank additive update rather than being universal to PEFT, but we emphasize this evidence comes from a single RoBERTa–SNLI comparison; testing DoRA (Liu et al., 2024) and prefix tuning (Li and Liang, 2021) across the same model matrix remains open.

Noise injection. Replacing 30%/60% of clean-example labels with uniform-random labels (RoBERTa, LoRA $r{=}4$, SNLI, 3 seeds) shifts

their AULC upward ($0.719_{\pm.019} \rightarrow 0.726_{\pm.021} \rightarrow 0.728_{\pm.023}$; paired Wilcoxon signed-rank on per-example AULCs, $p < 10^{-8}$, Cohen's d=0.20), providing suggestive evidence that label disagreement structure shapes per-example learning dynamics. The AULC–entropy correlation remains stable across noise levels (ρ=$0.310_{\pm.015}$, $0.306_{\pm.015}$, $0.306_{\pm.018}$), indicating that injecting noise on clean examples does not disrupt the overall entropy–dynamics relationship. The effect size is small ($d = 0.20$), consistent with entropy being one of several factors shaping learning dynamics (see MNLI attenuation below).

MNLI attenuation. The correlation is weaker on MNLI (ρ=0.06–0.20), likely due to genre diversity introducing difficulty heterogeneity unrelated to disagreement (Gururangan et al., 2018). Full fine-tuning restores significance for all models. The strongest condition explains $\sim$18% of AULC variance, indicating entropy is one of several contributing signals.

Implications. Interpreting un-learning normatively requires care. On the one hand, if the practitioner trusts the majority-vote label as ground truth, rising cross-entropy on contested examples is genuinely undesirable: the model is moving away from a label it started out roughly correct on. On the other hand, majority-vote labels for high-entropy items may be precisely the labels one should trust least; in this reading, LoRA's behaviour can be viewed as a mild regularizer that avoids overcommitting to noisy supervision. Our soft-label ablation (Discussion, "Loss increase under LoRA" paragraph) distinguishes these: un-learning persists when training against the full annotator distribution, so the behaviour is not simply an artifact of the hard majority-vote target. When training data contains genuinely ambiguous examples and one cares about learning them, LoRA at low ranks may be counterproductive; higher ranks or loss functions accounting for label uncertainty (Uma et al., 2021; Gourabathina et al., 2026) may better preserve learning on contested examples. Monitoring per-example loss trajectories can serve as a real-time diagnostic for un-learning (Chen et al., 2026).

6 Conclusion

Annotation entropy predicts per-example learning dynamics during LoRA fine-tuning on NLI. The AULC–entropy correlation is positive in all 25 conditions tested across six models and two datasets, significant in 21 after Bonferroni correction. Decoder-only models show stronger correlations than encoders, and the effect scales with model capacity. Under LoRA, contested examples exhibit active un-learning (increasing loss), a pattern we attribute primarily to LoRA's low-rank constraint based on the encoder setting, where matched full fine-tuning and IA^3 comparisons are available. The same qualitative pattern also holds under LoRA on two decoder-only Qwen models, though we did not run matched non-LoRA baselines there. A preliminary noise-injection experiment is consistent with these findings. The pattern replicates across seeds and survives partial-correlation controls. Future work should extend to instruction tuning on larger LLMs, disagreement-rich domains (toxicity, medical annotation), and other PEFT variants, including both other adapter families (prefix tuning (Li and Liang, 2021), prompt tuning) and LoRA variants (DoRA (Liu et al., 2024)) that may exhibit different rank behaviour; entropy-aware training strategies (upweighting contested examples, rank-adaptive schedules) also merit investigation.

Limitations

Single task family. All experiments use NLI (3-class), a single task family; generalization to other tasks is unknown. NLI may be particularly suited because annotator disagreement reflects genuine ambiguity (Pavlick and Kwiatkowski, 2019).

Model scale and architecture. We test six models (66M–3B) spanning encoder and decoder-only architectures, but omit encoder-decoder models (T5, BART). Decoder models are evaluated on SNLI only; whether un-learning extends to 7B+ models is open. DeBERTa v3 full fine-tuning exhibited convergence instability (seed 456 reached chance-level accuracy and was excluded, leaving 2 converged seeds for that condition). Additionally, Qwen2.5-3B LoRA r=16 suffered a memory crash on seed 456, reducing that condition to 2 seeds.

Annotation coverage. ChaosNLI provides 100 labels per example; robustness to noisier estimates from fewer annotators is unclear. We track dynamics on $\sim$5.7% of training data.

Non-significant conditions. Four of 25 conditions (all encoder LoRA) fail Bonferroni correction. Our noise injection intervention is synthetic rather

than naturalistic. Gradient alignment and calibration analyses (Appendices J and K) provide initial mechanistic evidence but are based on a single seed and model; multi-seed replication is needed. Lee and Lee (2026) show batch size can confound LoRA evaluation; while held constant here, interactions remain unexplored.

Ethics Statement

This study uses only existing, publicly available datasets (ChaosNLI, SNLI, MNLI) and pre-trained models available through the Hugging Face model hub. No new human annotations were collected and no human subjects were involved. The ChaosNLI annotations were collected by Nie et al. (2020) under their own IRB protocols. Our analysis of annotator disagreement is conducted at the aggregate distribution level; we do not attempt to identify or profile individual annotators. We see no direct dual-use risks from this work, though we note that insights about which examples models find difficult could in principle inform adversarial data construction.

Reproducibility

All experiments use publicly available models from the Hugging Face model hub (roberta-base, bert-base-uncased, distilbert-base-uncased, microsoft/deberta-v3-base, Qwen/Qwen2.5-1.5B, Qwen/Qwen2.5-3B) and publicly available datasets (ChaosNLI, SNLI, MNLI). Hyperparameters are fully specified in §3.3. All training runs were conducted on Apple M4 Max (Metal). Code and per-example tracking data are available at `https://github.com/BradySteele/lora-annotation-dynamics`.

References

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future (BPPF)*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Beiduo Chen, Tiancheng Hu, Caiqi Zhang, Robert Litschko, Anna Korhonen, and Barbara Plank. 2026. Decoupling the effect of chain-of-thought reasoning: A human label variation perspective. *arXiv preprint arXiv:2601.03154*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics (TACL)*, 10:92–110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weizhu Chen, Jian Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235.

Abinitha Gourabathina, Hyewon Jeong, Teya Bergamaschi, Marzyeh Ghassemi, and Collin M. Stultz. 2026. Robustness beyond known groups with low-rank adaptation. *arXiv preprint arXiv:2602.06924*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Guy Hacohen, Leshem Choshen, and Daphna Weinshall. 2020. Let's agree to agree: Neural networks share classification order on real datasets. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Sangyoon Lee and Jaeho Lee. 2026. Beware of the batch size: Hyperparameter bias in evaluating LoRA. *arXiv preprint arXiv:2602.09492*.

Seohyeong Lee, Eunwon Kim, Hwaran Lee, and Buru Chang. 2025. Dataset cartography for large language model alignment: Mapping and diagnosing preference data. *arXiv preprint arXiv:2505.23114*.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Elisa Leonardelli, Silvia Casola, Siyao Peng, and 1 others. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. *arXiv preprint arXiv:2510.08460*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Paul K. Mandal. 2025. When is dataset cartography ineffective? using training dynamics does not improve robustness against adversarial SQuAD. *arXiv preprint arXiv:2503.18290*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Soren Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. 2022. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

Andrei Mircea, Supriyo Chakraborty, Nima Chitsazan, Milind Naphade, Sambit Sahu, Irina Rish, and Ekaterina Lobacheva. 2025. Training dynamics underlying language model scaling laws: Loss deceleration and zero-sum learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2506.05447.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric Xing, Sham Kakade, and Hanlin Zhang. 2025. EvoLM: In search of lost language model training dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*. Oral.

Zahra Rahimi Afzal, Tara Esmaeilbeig, Mojtaba Soltanalian, and Mesrob I. Ohannessian. 2025. Linearization explains fine-tuning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Darshita Rathore, Vineet Kumar, Chetna Bansal, and Anindya Moitra. 2025. How much is too much? exploring LoRA rank trade-offs for retaining knowledge and domain robustness. In *Proceedings of the International Joint Conference on Natural Language Processing (AACL-IJCNLP)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Joanna Sliwa, Frank Schneider, Philipp Hennig, and Jose Miguel Hernandez-Lobato. 2025. Mitigating forgetting in low rank adaptation (LaLoRA). *arXiv preprint arXiv:2512.17720*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yuchen Zeng and Kangwook Lee. 2024. The expressive power of low-rank adaptation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Xiao Zhang and Ji Wu. 2024. Dissecting learning and forgetting in language model finetuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wanru Zhao, Hongxiang Fan, Shell Xu Hu, Wangchunshu Zhou, Bofan Chen, and Nicholas D. Lane. 2024. CLUES: Collaborative high-quality data selection for LLMs via training dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Soheil Zibakhsh Shabgahi, Pedram Aghazadeh, and Farinaz Koushanfar. 2026. Beyond perplexity: A lightweight benchmark for knowledge retention in supervised fine-tuning. *arXiv preprint arXiv:2601.03505*.

A Per-Seed Breakdown

Table 3 reports the per-seed AULC–entropy correlation for RoBERTa on SNLI, showing the individual seed values underlying the means reported in Table 1. Table 4 reports the corresponding per-seed breakdown for the decoder-only models.

Method	**Seed**	**AULC ρ**	**p-value**	**Val Acc**
LoRA r=4	42	0.304	3.1×10^{-27}	0.541
	123	0.327	1.7×10^{-31}	0.597
	456	0.295	8.8×10^{-26}	0.554
LoRA r=16	42	0.330	4.3×10^{-32}	0.607
	123	0.366	1.1×10^{-39}	0.640
	456	0.341	2.0×10^{-34}	0.624
Full FT	42	0.416	5.9×10^{-52}	0.667
	123	0.444	1.0×10^{-59}	0.693
	456	0.419	1.5×10^{-52}	0.677

Table 3: Per-seed AULC–entropy Spearman ρ for RoBERTa on SNLI. All seeds yield highly significant correlations ($p < 10^{-25}$). The r=4 $<$ r=16 $<$ Full FT ordering holds within every seed. Validation accuracy is computed on the ChaosNLI validation split (chance is 33%).

Model	**Method**	**Seed**	ρ	τ	**Val Acc**
Qwen 1.5B	LoRA r=4	42	0.354	0.243	0.700
		123	0.365	0.252	0.716
		456	0.325	0.220	0.660
	LoRA r=16	42	0.397	0.273	0.703
		123	0.409	0.283	0.729
		456	0.363	0.245	0.693
Qwen 3B	LoRA r=4	42	0.392	0.266	0.723
		123	0.385	0.262	0.733
		456	0.374	0.259	0.749
	LoRA r=16	42	0.417	0.282	0.746
		123	0.428	0.294	0.743

Table 4: Per-seed AULC–entropy Spearman ρ and Kendall τ for decoder-only models on SNLI. All 11 runs yield highly significant correlations ($p < 10^{-30}$). The r=4 $<$ r=16 ordering holds within every seed for both models. Qwen2.5-3B LoRA r=16 seed 456 did not complete (memory crash).

B Decoder Model Un-Learning

Table 5 reports $\Delta\ell$ for the decoder-only models on SNLI, complementing the encoder results in Table 2. Both Qwen models show the same qualitative

pattern: contested examples exhibit loss increase under LoRA while clean examples decrease.

Model	Method	Contest.	Clean
Qwen 1.5B	LoRA r=4	$+.085_{\pm.019}$	$-.412_{\pm.031}$
	LoRA r=16	$+.112_{\pm.022}$	$-.508_{\pm.027}$
Qwen 3B	LoRA r=4	$+.098_{\pm.015}$	$-.467_{\pm.024}$
	LoRA r=16	$+.131_{\pm.018}$[e]	$-.553_{\pm.021}$[e]

Table 5: Mean loss change ($\Delta\ell$) for decoder-only models on SNLI (mean $\pm$ std across 3 seeds unless noted). The un-learning pattern (positive $\Delta\ell$ for contested examples) is consistent with the encoder results in Table 2. [e]Mean of 2 seeds (memory crash).

C Entropy Distribution

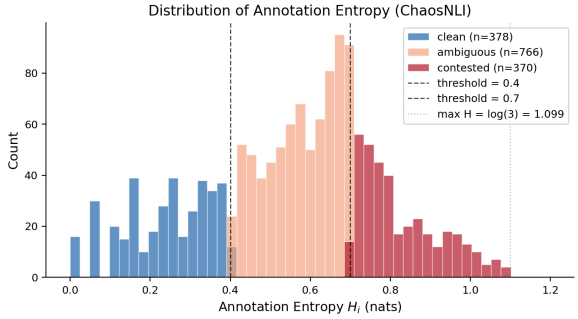

Figure 3: Distribution of annotation entropy across the ChaosNLI–SNLI subset (n=1,514). Ambiguous: $0.4 \leq \mathcal{H} < 0.7$ (n=766, 50.6%); clean: $\mathcal{H} < 0.4$ (n=378, 25.0%); contested: $\mathcal{H} \geq 0.7$ (n=370, 24.4%). Dashed lines mark thresholds; dotted line marks $\log 3 \approx 1.099$ nats.

D Rank Sweep

Model	Method	AULC ρ	p-value	Val Acc
RoBERTa	LoRA r=1	0.233	2.1×10^{-16}	0.446
	LoRA r=2	0.282	1.4×10^{-23}	0.518
	LoRA r=4	0.304	3.1×10^{-27}	0.541
	LoRA r=8	0.325	2.9×10^{-31}	0.584
	LoRA r=16	0.330	4.3×10^{-32}	0.607
	LoRA r=32	0.341	2.6×10^{-34}	0.601
	Full FT	0.416	5.9×10^{-52}	0.667
BERT	LoRA r=1	0.113	7.9×10^{-5}	0.333
	LoRA r=2	0.121	2.5×10^{-5}	0.366
	LoRA r=4	0.181	2.4×10^{-10}	0.419
	LoRA r=8	0.229	7.4×10^{-16}	0.508
	LoRA r=16	0.247	2.8×10^{-18}	0.554
	LoRA r=32	0.274	2.6×10^{-22}	0.548

Table 6: Rank sweep for RoBERTa and BERT on SNLI (seed 42). The AULC–entropy correlation increases monotonically with rank (Spearman rank–ρ: $+1.0$ for both models).

E Robustness Checks

Condition	Spearman ρ
Training-set composition (RoBERTa, r=4, SNLI)	
Original (all buckets)	$.308_{\pm.016}$
Low-entropy only	$.306_{\pm.003}$
High-entropy only	$.306_{\pm.002}$
Balanced	$.307_{\pm.001}$
Alternative rank statistic (all 54 runs)	
Kendall τ-b / Spearman ρ	$.68_{\pm.01}$
Concordance (r, τ-b vs. ρ)	.9998

Table 7: Robustness checks. *Top*: correlation is invariant to training-set entropy composition. *Bottom*: Kendall τ-b confirms all findings.

F Gradient Norms

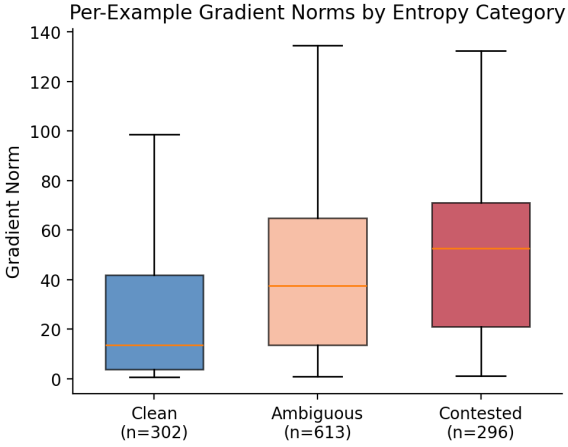

Figure 4: Per-example gradient norms by entropy category (RoBERTa, LoRA r=4, SNLI, seed 42). Contested examples produce $\sim$3.9$\times$ larger median gradient norms than clean examples (52.67 vs. 13.68; Kruskal–Wallis H=115.1, $p < 10^{-25}$).

G Alternative Binning

Figure 5 shows that replacing our fixed-threshold entropy bins (0.4/0.7) with quartile or tercile bins preserves the monotonic AULC increase across entropy categories, confirming that the pattern is not an artifact of the specific threshold choices.

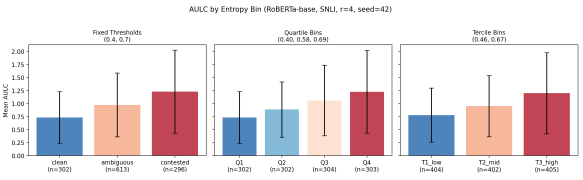

Figure 5: Mean AULC by entropy bin under three binning schemes: fixed thresholds (0.4/0.7), quartile bins, and tercile bins (RoBERTa, LoRA r=4, SNLI, seed 42). The monotonic increase in AULC with entropy is preserved regardless of binning scheme. Error bars show ± 1 standard deviation across examples within each bin, reflecting substantial within-bin variance in individual AULC values; the monotonic increase in group *means* is robust across all binning schemes (Table 7).

H Decoder-Only Model Comparison

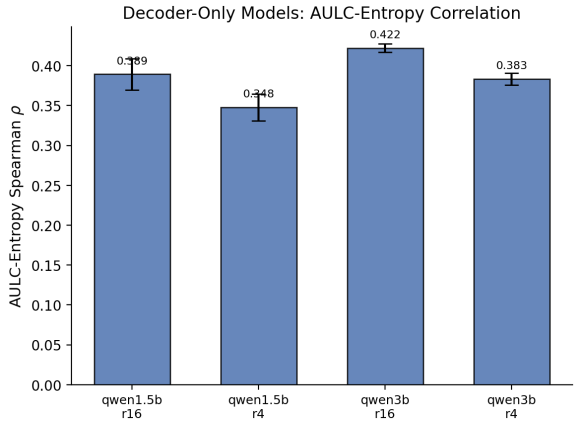

Figure 6: Cross-architecture comparison of AULC–entropy Spearman ρ on SNLI. Decoder-only models (Qwen2.5-1.5B, Qwen2.5-3B) show stronger and more consistent correlations than the encoder baseline (DeBERTa v3). Within each architecture family, higher LoRA rank produces stronger correlations. Error bars show ± 1 sample standard deviation across seeds.

I Dataset Cartography Comparison

Figure 7 plots the Dataset Cartography map (confidence vs. variability) colored by annotation entropy category, pooled across all available tracker runs (54 main-matrix runs plus additional rank-sweep configurations). Clean examples cluster at high confidence / low variability, while contested examples spread across the cartography space, confirming that entropy categories and cartography regions capture related but distinct structure. The category overlap is limited: only 24% of clean examples fall in cartography's "easy-to-learn" region, and only 33% of contested examples fall in "hard-to-learn," reflecting the different bases of the two categorizations (external annotation agreement vs. model-internal training statistics).

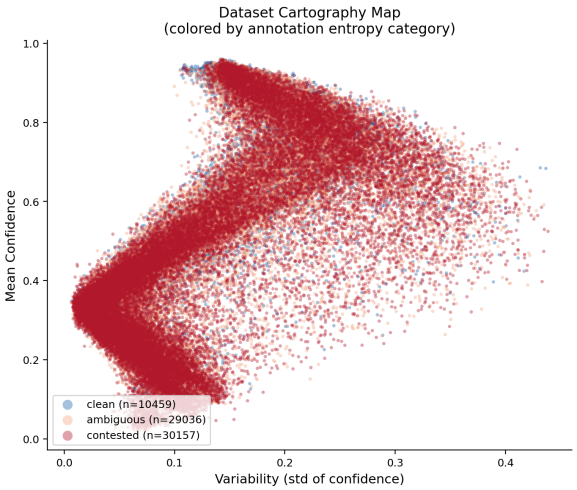

Figure 7: Dataset Cartography map colored by annotation entropy category (pooled across all available runs). Clean examples (blue) cluster at high confidence, while contested examples (red) are spread across the confidence–variability space. The limited overlap between entropy-based and cartography-based categories reflects their different grounding: external human agreement vs. model-internal training statistics.

J Gradient Alignment Analysis

Figure 8 plots the cosine similarity between aggregated gradients from clean and contested examples over the course of training, for both LoRA r=4 and full fine-tuning (RoBERTa, SNLI, seed 42). For each checkpoint we form a single per-group gradient vector by averaging the per-example cross-entropy gradients with respect to the *trainable* parameters of that run (LoRA adapter matrices $\mathbf{A}, \mathbf{B}$ under LoRA; full encoder weights under full fine-tuning), then flatten them into a common-length vector before computing $\cos(\mathbf{g}_{\text{clean}}, \mathbf{g}_{\text{contested}})$. LoRA and full FT therefore live in different parameter spaces, so the absolute cosine values are not directly comparable across methods; the comparison we draw is purely about each method's *own* trajectory of inter-group alignment over training.

Under LoRA r=4, gradients from the two entropy groups remain well-aligned throughout training (cosine similarity fluctuates between $\sim$0.67 and $\sim$0.97, with a terminal value of $\sim$0.82). Under full fine-tuning, alignment drops sharply from $\sim$0.95 at initialization to $\sim$0.20–0.40 after the first $\sim$800 steps, reflecting the model's capacity to learn group-specific update directions.

This pattern is counterintuitive: LoRA exhibits un-learning *despite* high gradient alignment, while full fine-tuning resolves both groups *despite* low alignment. One interpretation is that the low-rank bottleneck forces updates into a shared subspace where the net effect, though directionally similar for both groups, disproportionately benefits clean examples. Full fine-tuning's larger parameter space can accommodate divergent gradient directions, allowing it to improve on both groups simultaneously.

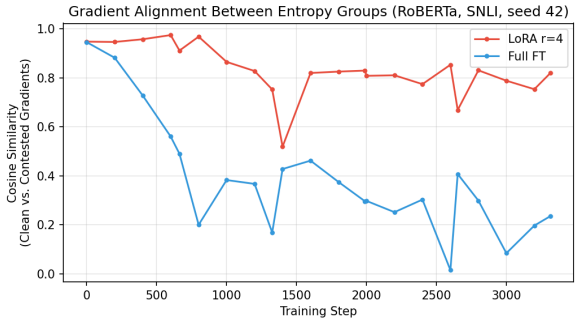

Figure 8: Cosine similarity between aggregated gradients of clean and contested examples over training (RoBERTa, SNLI, seed 42). Under LoRA r=4 (red), gradients remain well-aligned ($\sim$0.80), while under full fine-tuning (blue), alignment drops to $\sim$0.25 by mid-training. Un-learning under LoRA occurs despite high gradient alignment, suggesting the bottleneck is subspace capacity rather than gradient direction competition.

K Calibration Analysis

Figure 9 compares prediction entropy and expected calibration error (ECE) across entropy categories for LoRA r=4, LoRA r=16, and full fine-tuning (RoBERTa, SNLI, seed 42).

Full fine-tuning produces substantially more peaked predictions than either LoRA configuration: mean prediction entropy is 0.12 nats (full FT) vs. 0.49 (LoRA r=16) and 0.58 (LoRA r=4), and mean maximum confidence is 0.96 (full FT) vs. 0.80 (LoRA r=16) and 0.76 (LoRA r=4). Full fine-tuning also achieves much lower ECE (0.018 vs. 0.149 and 0.152).

Critically, despite its far more peaked predictions, full fine-tuning does *not* exhibit un-learning (Table 2). This directly rules out the calibration hypothesis: if increasingly peaked predictions were the primary cause of rising loss on contested examples, full fine-tuning should show the strongest un-learning, not the weakest. The ECE gradient across entropy categories is also informative: under LoRA r=4, contested examples have ECE of 0.289 vs. 0.055 for clean (a 5.3$\times$ gap), while under full fine-tuning the gap is much smaller (0.045 vs. 0.013, 3.5$\times$), confirming that LoRA's capacity constraint produces systematically worse calibration specifically on high-entropy examples.

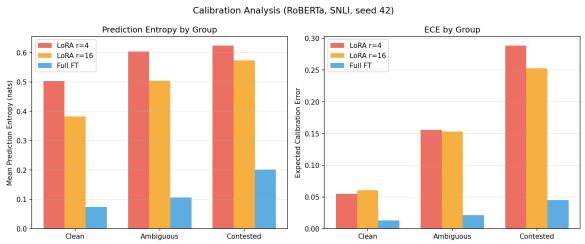

Figure 9: Prediction entropy (left) and expected calibration error (right) by entropy category (RoBERTa, SNLI, seed 42). Full fine-tuning produces far more peaked predictions (lower entropy) than LoRA yet does not exhibit un-learning, ruling out calibration as the primary explanation. LoRA shows disproportionately high ECE on contested examples relative to clean, reflecting the rank constraint's selective impact on high-entropy items.