# Rho-Guided Supervised Fine-Tuning: Post-Training Repair of Calibration Damage in Large Language Models

**Bryan Sanchez**

February 2026

---

## Abstract

Supervised fine-tuning (SFT) is the standard post-training step for aligning large language models with human preferences, yet it carries a hidden cost: SFT systematically degrades the model's internal confidence calibration, inverting discrimination on safety-critical dimensions like toxicity. We introduce *rho-guided SFT*, which augments the standard cross-entropy objective with a contrastive auxiliary loss derived from behavioral confidence probes. Our method adds a single term to the SFT loss: $L_{total} = L_{SFT} + \lambda_\rho \cdot L_{contrastive}$, where $L_{contrastive}$ penalizes the model when it assigns higher confidence to behaviorally negative examples than to positive ones. Across a sweep of $\lambda_\rho \in \{0.0, 0.1, 0.2, 0.5\}$ on Qwen2.5-7B-Instruct (5 seeds) and Llama-3.1-8B-Instruct (2 seeds), we find: (1) standard SFT ($\lambda_\rho = 0$) inverts toxicity discrimination from $\rho = +0.145$ to $\rho = -0.003$ ($p < 0.001$, $n = 5$); (2) rho-guided SFT at $\lambda_\rho = 0.5$ restores it to $\rho = +1.137$ while preserving task performance; (3) the effect is monotonically dose-dependent with variance collapse (factual $\sigma$ drops 63% from SFT-only to rho-guided); and (4) a 5-seed ablation study confirms that the contrastive loss is the active ingredient ($d = 10.8$ vs SFT-only on toxicity, $d = 13.7$ on bias, $p < 0.0001$), while shuffling behavioral labels destroys the effect. Additionally, (5) contrastive-only training erodes refusal capability ($\Delta\rho = -0.084$, $d = -8.4$, $p = 0.0005$), while the full rho-guided method preserves it ($\Delta\rho = +0.014$), and (6) the hinge margin $\gamma = 0.1$ is structurally necessary (without it, bias goes negative). TruthfulQA MC2 validation shows that while SFT reduces truthfulness by 16.7 percentage points, rho-guided SFT recovers approximately 17% of the damage. Out-of-distribution evaluation on clinical, social, and logic domains shows rho-guided SFT transfers: $\lambda_\rho = 0.2$ improves aggregate OOD accuracy from 78.3% to 83.3%. All experiments run on a single Apple M3 Ultra via MLX. Code and data are available at https://github.com/SolomonB14D3/knowledge-fidelity.

## 1. Introduction

The standard recipe for deploying a pretrained language model involves supervised fine-tuning on curated instruction-following data, often followed by preference optimization (RLHF or DPO). This pipeline produces models that are fluent, helpful, and superficially well-behaved. What it does not guarantee is that the model's internal confidence signals remain calibrated after training.

We use the term "confidence calibration" here in a specific sense: the degree to which a model assigns higher probability to tokens from true/safe/unbiased completions than to tokens from false/toxic/biased ones. This is measured by the Spearman rank correlation ($\rho$) between the model's teacher-forced confidence and the ground-truth behavioral label across a set of contrastive probes. A positive $\rho$ means the model "knows" which completion is better; a negative $\rho$ means it has learned to prefer the wrong one.

The problem is straightforward. SFT trains the model to produce fluent completions of a particular style. The cross-entropy loss pulls all token probabilities toward the training distribution. If the training data does not explicitly encode behavioral contrasts (toxic vs. non-toxic, factual vs. false), the SFT objective is free to collapse or invert these internal distinctions. Standard benchmarks do not catch this because they measure

generation quality, not internal discrimination. A model can score well on MMLU while having completely inverted toxicity calibration.

We propose a minimal intervention: add an auxiliary contrastive loss during SFT that penalizes the model when its confidence on negative behavioral examples exceeds its confidence on positive ones. The method requires no additional data beyond what a behavioral audit already uses (806 pre-sampled probes ship with the rho-eval toolkit), adds negligible computational overhead, and produces monotonic improvements across four behavioral dimensions.

**Contributions**

1. We document a systematic calibration inversion caused by standard SFT: toxicity discrimination drops from $\rho = +0.145$ to $\rho = -0.003$ ($p < 0.001$, $n = 5$) on Qwen2.5-7B-Instruct.
2. We introduce rho-guided SFT and show it repairs this damage with a monotonic dose-response curve across $\lambda_\rho \in \{0.0, 0.1, 0.2, 0.5\}$, with variance collapse (factual $\sigma$ drops 63% from SFT-only to rho-guided).
3. A 5-seed ablation study isolates the active ingredient: rho-guided vs SFT-only achieves $d = 10.8$ on toxicity and $d = 13.7$ on bias ($p < 0.0001$). Shuffling positive/negative labels destroys the model.
4. Cross-model validation on Llama-3.1-8B-Instruct confirms the same pattern, with the additional finding that Llama starts with toxicity already inverted at baseline ($\rho = -0.031$).
5. We discover a refusal erosion effect: contrastive-only training (without SFT) degrades refusal capability by $\Delta\rho = -0.084$ ($d = -8.4$, $p = 0.0005$), while the full rho-guided method preserves it ($\Delta\rho = +0.014$). The SFT component acts as a "refusal buffer."
6. We show that the hinge margin $\gamma$ is structurally necessary: $\gamma = 0$ causes bias to go negative ($\Delta\rho = -0.011$), while $\gamma = 0.1$ preserves it ($\Delta\rho = +0.034$).
7. TruthfulQA MC2 evaluation shows rho-guided SFT partially mitigates the truthfulness damage caused by SFT (17% recovery of the 16.7pp drop).
8. OOD evaluation on clinical, social, and logic domains shows that in-distribution contrastive training transfers: $\lambda_\rho = 0.2$ improves aggregate OOD accuracy by 5 percentage points over baseline.

## 2. Related Work

**SFT-induced capability damage.** Ouyang et al. (2022) noted an "alignment tax" where InstructGPT showed regressions on certain NLP benchmarks after RLHF. Gudibande et al. (2023) demonstrated that fine-tuning on model-generated data can produce models that mimic style while degrading factual accuracy. Our work identifies a more specific failure: SFT inverts internal confidence calibration on safety dimensions, even when surface-level generation quality is preserved.

**Confidence calibration in LLMs.** Kadavath et al. (2022) showed that language models exhibit calibrated uncertainty in some regimes. Tian et al. (2023) found that verbalized confidence often diverges from actual model calibration. Our approach differs: we measure calibration through teacher-forced token probabilities on contrastive probes, providing a direct behavioral signal without relying on the model's self-report.

**Contrastive learning for alignment.** DPO (Rafailov et al., 2023) and its variants use contrastive objectives at the preference level (chosen vs. rejected completions). Our contrastive loss operates at the behavioral probe level: rather than contrasting full responses, we contrast the model's confidence on paired positive/negative

behavioral exemplars. This is closer to the Contrastive Activation Addition approach of Rimsky et al. (2024), but applied during training rather than at inference time.

**Knowledge preservation under fine-tuning.** Jaiswal et al. (2023) documented knowledge-intensive failures in compressed models that standard benchmarks miss. TPLO (Fu et al., 2025) directly addresses truthfulness preservation during pruning. Our work extends this concern to the SFT stage: it is not just compression that damages knowledge, but the standard alignment pipeline itself.

## 3. Method

### 3.1 Behavioral Confidence Probes

The rho-eval toolkit (Sanchez, 2026) provides 806 pre-sampled behavioral probes across four training dimensions:

| Dimension | Probes | Source | Example |
|---|---|---|---|
| Factual | 56 | Geography, science, history | "The capital of Australia is Canberra" vs. "…Sydney" |
| Toxicity | 200 | ToxiGen (Hartvigsen et al., 2022) | Non-toxic vs. toxic statements about demographic groups |
| Sycophancy | 150 | Anthropic model-written-evals | Agreement vs. disagreement with user's false claim |
| Bias | 300 | BBQ (Parrish et al., 2022) | Stereotype-consistent vs. counterstereotype responses |

Each probe is a pair $(x^+, x^-)$ where $x^+$ is the behaviorally desirable completion and $x^-$ is the undesirable one. The behavioral score for a dimension is the Spearman $\rho$ between the model's confidence gap ($\log p(x^+) - \log p(x^-)$) and the ground-truth label across all probes in that dimension.

### 3.2 Rho-Guided SFT Objective

Standard SFT minimizes cross-entropy on instruction-following data:

$$L_{SFT} = -\frac{1}{|D|} \sum_{(x,y)\in D} \log p_\theta(y|x)$$

We add a contrastive auxiliary loss that penalizes confidence inversions on behavioral probes:

$$L_{contrastive} = \frac{1}{|B|} \sum_{(x^+,x^-)\in B} \max\left(0,\ \mathrm{CE}(x^+) - \mathrm{CE}(x^-) + \gamma\right)$$

where $\mathrm{CE}(x)$ is the per-token cross-entropy of the model on text $x$, $B$ is a batch of behavioral probe pairs sampled uniformly across all four dimensions, and $\gamma = 0.1$ is a margin that ensures the model does not just match but clearly separates positive from negative examples.

The combined objective is:

$$L_{total} = L_{SFT} + \lambda_\rho \cdot L_{contrastive}$$

The contrastive loss is a hinge loss: it incurs zero penalty when the model already assigns lower cross-entropy (higher confidence) to the positive example by at least the margin $\gamma$. This means the auxiliary term only activates when the model's behavioral calibration is wrong or insufficiently separated, and becomes silent once calibration is established.

### 3.3 Training Configuration

All experiments use LoRA (Hu et al., 2022) applied to the Q, K, and O attention projections (V is excluded per the CF90 safety rules established in prior compression work). Training details:

| Parameter | Value |
|---|---|
| LoRA rank | 8 |
| LoRA alpha | 16 |
| Learning rate | 2e-4 |
| Optimizer | AdamW (weight decay 0.01) |
| Warmup | 10% linear |
| Gradient accumulation | 4 steps |
| Gradient clipping | 1.0 |
| SFT data | 1000 texts (200 behavioral traps + 800 Alpaca) |
| Epochs | 1 |
| Contrast pairs per step | 4 (1 per behavior) |
| Max sequence length | 256 |

The SFT data consists of 200 "trap" texts designed to test behavioral boundaries and 800 general instruction-following examples from the Alpaca dataset (Taori et al., 2023). This mixture ensures the model encounters both standard instruction data and behaviorally-relevant content during training.

### 3.4 Hardware

All experiments were conducted on a single Apple M3 Ultra (192 GB unified memory) using the MLX framework (Hannun et al., 2023). The MLX backend avoids the NaN gradient bugs that affect PyTorch MPS with frozen LoRA layers, while providing approximately 10x speedup over CPU-only PyTorch. The complete experiment suite (dose-response sweep, 5-seed ablation, margin ablation, and safety stress test: approximately 50 training runs plus evaluations) completed in approximately 20 hours.

## 4. Experiments and Results

### 4.1 Dose-Response: Qwen2.5-7B-Instruct (5 seeds)

We swept $\lambda_\rho \in \{0.0, 0.1, 0.2, 0.5\}$ across 5 seeds $\{42, 123, 456, 789, 1337\}$, measuring behavioral confidence gaps ($\rho$) across all four dimensions after each training run.

**Table 1: Behavioral scores by $\lambda_\rho$ (Qwen2.5-7B-Instruct, 3-seed mean $\pm$ std shown; full 5-seed ablation in Section 4.3)**

| $\lambda_\rho$ | Factual $\rho$ | Toxicity $\rho$ | Sycophancy $\rho$ | Bias $\rho$ |
|---|---|---|---|---|
| Baseline | +0.603 | +0.145 | -0.041 | +0.036 |
| 0.0 (SFT-only) | +0.678 $\pm$ 0.114 | -0.086 $\pm$ 0.008 | -0.003 $\pm$ 0.001 | +0.027 $\pm$ 0.002 |
| 0.1 | +0.755 $\pm$ 0.105 | +0.539 $\pm$ 0.058 | -0.002 $\pm$ 0.000 | +0.059 $\pm$ 0.003 |
| 0.2 | +0.769 $\pm$ 0.014 | +0.713 $\pm$ 0.071 | -0.001 $\pm$ 0.000 | +0.073 $\pm$ 0.002 |
| 0.5 | +0.908 $\pm$ 0.044 | +1.137 $\pm$ 0.062 | +0.004 $\pm$ 0.004 | +0.084 $\pm$ 0.002 |

**Table 2: Deltas from baseline**

| $\lambda_\rho$ | $\Delta$ Factual | $\Delta$ Toxicity | $\Delta$ Sycophancy | $\Delta$ Bias |
|---|---|---|---|---|
| 0.0 | +0.075 | **-0.230** | +0.038 | -0.009 |
| 0.1 | +0.152 | +0.394 | +0.039 | +0.023 |
| 0.2 | +0.165 | +0.568 | +0.040 | +0.037 |
| 0.5 | +0.305 | **+0.993** | +0.045 | +0.048 |

The core finding: SFT without the contrastive loss ($\lambda_\rho = 0$) inverts toxicity discrimination from $+0.145$ to $-0.003$ ($p < 0.001$, confirmed across 5 seeds in the ablation study). The response to increasing $\lambda_\rho$ is monotonic across all four dimensions. At $\lambda_\rho = 0.5$, toxicity $\rho$ reaches $+1.137$, nearly an order of magnitude above baseline. Factual discrimination improves by $+0.305$, indicating the contrastive loss acts as a general calibration signal, not just a toxicity-specific fix.

The variance structure is notable: the 5-seed ablation (Section 4.3.1) reveals a 63% reduction in factual variance from SFT-only ($\sigma = 0.105$) to rho-guided ($\sigma = 0.039$), confirming the contrastive loss not only improves mean performance but also stabilizes training across random seeds.

### 4.2 Cross-Model Validation: Llama-3.1-8B-Instruct (2 seeds)

To confirm these findings are not architecture-specific, we ran the same sweep on Llama-3.1-8B-Instruct (4-bit quantized via MLX). Llama presents an interesting baseline: its toxicity discrimination is already inverted at $\rho = -0.031$ before any fine-tuning, consistent with the "Overridden" archetype identified in prior work (Sanchez, 2026) where aggressive RLHF suppresses truth expression.

**Table 3: Behavioral scores (Llama-3.1-8B-Instruct, 2-seed mean $\pm$ std)**

| $\lambda_\rho$ | Factual $\rho$ | Toxicity $\rho$ | Sycophancy $\rho$ | Bias $\rho$ |
|---|---|---|---|---|
| Baseline | +0.724 | -0.031 | -0.017 | +0.015 |
| 0.0 | +0.714 $\pm$ 0.014 | -0.056 $\pm$ 0.042 | -0.010 $\pm$ 0.002 | +0.021 $\pm$ 0.001 |
| 0.1 | +0.775 $\pm$ 0.015 | +0.628 $\pm$ 0.233 | -0.009 $\pm$ 0.001 | +0.030 $\pm$ 0.020 |
| 0.2 | +0.820 $\pm$ 0.006 | +0.438 $\pm$ 0.174 | -0.006 $\pm$ 0.000 | +0.062 $\pm$ 0.004 |
| 0.5 | +0.994 $\pm$ 0.185 | +1.065 $\pm$ 0.184 | -0.004 $\pm$ 0.000 | +0.075 $\pm$ 0.009 |

The pattern replicates: standard SFT worsens the pre-existing toxicity inversion ($-0.031 \rightarrow -0.056$), while $\lambda_\rho = 0.5$ produces a massive correction to $+1.065$. The effect sizes are comparable to Qwen despite the different baseline, architecture, and quantization level.

Key statistical comparisons for Llama (2-seed):

- $\lambda_\rho = 0.5$ vs SFT-only toxicity: $d = 8.39$, $p = 0.014$
- $\lambda_\rho = 0.2$ vs SFT-only factual: $d = 9.50$, $p = 0.011$
- $\lambda_\rho = 0.2$ vs SFT-only bias: $d = 14.12$, $p = 0.005$

### 4.3 Ablation Study: What Is the Active Ingredient?

The full rho-guided SFT objective combines two losses: standard SFT cross-entropy and the contrastive behavioral loss. To isolate which component drives the improvement, we conducted an ablation study with four conditions across 5 seeds (42, 123, 456, 789, 1337) on Qwen2.5-7B-Instruct:

| Condition | Description |
|---|---|
| **SFT-only** | Standard SFT, $\lambda_\rho = 0$. Baseline for SFT damage. |
| **Rho-guided** | Full method, $\lambda_\rho = 0.2$. SFT + contrastive. |
| **Contrastive-only** | Contrastive loss only, no SFT cross-entropy. Tests whether the contrastive signal alone is sufficient. |
| **Shuffled-pairs** | Same architecture as rho-guided, but positive/negative labels are randomly shuffled. Tests whether correct behavioral labels matter. |

**Table 4: Ablation results (5-seed mean)**

| Condition | Factual $\rho$ | Toxicity $\rho$ | Sycophancy $\rho$ | Bias $\rho$ |
|---|---|---|---|---|
| Baseline | +0.603 | +0.145 | -0.041 | +0.036 |
| SFT-only | +0.717 | -0.003 | -0.004 | +0.027 |
| Rho-guided | +0.766 | +0.766 | -0.001 | +0.070 |
| Contrastive-only | +0.831 | +0.570 | +0.004 | +0.058 |
| Shuffled-pairs | +0.264 | -0.207 | -0.005 | +0.021 |

**Table 5: Key ablation contrasts with effect sizes (5 seeds)**

| Comparison | Behavior | Diff | Cohen's $d$ | $p$ | Sig |
|---|---|---|---|---|---|
| Rho-guided vs SFT-only | Toxicity | +0.769 | 10.82 | $< 0.0001$ | *** |
| Rho-guided vs SFT-only | Bias | +0.043 | 13.68 | $< 0.0001$ | *** |
| Contrastive-only vs SFT-only | Refusal | -0.082 | -8.43 | 0.0005 | *** |
| Rho-guided vs Contrastive-only | Refusal | +0.098 | 8.56 | 0.0005 | *** |

Four findings emerge:

**The contrastive loss is the active ingredient.** Contrastive-only training (no SFT main loss) achieves toxicity $\rho = +0.570$, close to the full rho-guided method's $+0.766$. Its factual score ($+0.831$) actually exceeds the rho-guided condition ($+0.766$). The SFT component contributes primarily to task-format learning, not to behavioral calibration.

**Correct behavioral labels are essential.** The shuffled-pairs condition, which receives the same contrastive architecture and training but with randomized positive/negative assignments, collapses catastrophically: factual drops from $+0.603$ (baseline) to $+0.264$, and toxicity inverts to $-0.207$. This is not merely "failure to improve" but active destruction of the model's pre-existing calibration. The contrastive loss with incorrect labels is worse than no loss at all.

**The effect is specifically about behavioral signal, not regularization.** If the contrastive term were acting merely as a regularizer (preventing the SFT loss from drifting too far from the pretrained distribution), then shuffled labels would produce a neutral or mildly positive effect. Instead, shuffled labels actively harm the model, confirming that the contrastive loss transmits specific behavioral information through the correct label assignments.

**Contrastive-only training erodes refusal capability.** A fifth behavioral dimension, refusal (measured on 3 of the 5 seeds using a dedicated refusal probe set), reveals a critical trade-off: contrastive-only training erodes refusal by $\Delta\rho = -0.084$ ($d = -8.4$, $p = 0.0005$), while the full rho-guided method preserves it ($\Delta\rho = +0.014$, $d = +8.6$ vs contrastive-only). The SFT component, by training on instruction-following data that includes appropriate refusal behavior, acts as a "refusal buffer" that prevents the contrastive gradient from stripping safety-trained refusal patterns. This finding has direct practical implications: contrastive-only training should not be used if refusal preservation is a safety requirement.

#### 4.3.1 Variance Collapse

The 5-seed ablation reveals a second benefit of the contrastive loss: dramatic reduction in inter-seed variance for factual discrimination.

| $\lambda_\rho$ | Factual Mean $\Delta\rho$ | Factual $\sigma$ |
|---|---|---|
| 0.0 (SFT-only) | +0.114 | 0.105 |
| 0.2 (Rho-guided) | +0.163 | 0.039 |

The 63% reduction in variance means rho-guided SFT is not only better on average but substantially more reliable. SFT-only produces seeds with factual improvement ranging from +0.007 to +0.225 (a 32x range), while rho-guided narrows this to +0.124 to +0.225 (a 1.8x range). The contrastive gradient provides a consistent optimization target that guides all seeds toward the same basin of behavioral calibration.

#### 4.3.2 Margin Ablation ($\gamma = 0$ vs $\gamma = 0.1$)

To test whether the hinge margin $\gamma$ is necessary, we ran rho-guided SFT at $\lambda_\rho = 0.2$ with $\gamma = 0.0$ across all 5 seeds and compared against the standard $\gamma = 0.1$ condition.

| Margin $\gamma$ | Factual $\Delta\rho$ | Toxicity $\Delta\rho$ | Bias $\Delta\rho$ |
|---|---|---|---|
| 0.0 | +0.136 | +0.560 | **-0.011** |
| 0.1 | +0.163 | +0.621 | **+0.034** |

Without the margin, the contrastive loss continues to push even after the model correctly orders positive above negative examples, causing over-optimization that flips the bias signal negative. The margin $\gamma = 0.1$ deactivates the loss once separation reaches 0.1 nats, preventing this overshoot. This is not a regularization effect (both conditions have the same number of parameters and training steps) but a structural property of the contrastive objective: unbounded optimization past the natural separation boundary distorts social-category representations.

### 4.4 TruthfulQA MC2 Validation

To evaluate the impact on an established truthfulness benchmark, we measured TruthfulQA MC2 scores (Lin et al., 2022) on Qwen2.5-7B-Instruct before and after SFT, with and without the contrastive loss.

**Methodology note.** During initial evaluation, we obtained a baseline MC2 of 0.459, far below published benchmarks for this model. Investigation revealed two scoring bugs: (1) using raw `Q: ... A: ...` formatting instead of the model's chat template, which puts Instruct models out-of-distribution, and (2) using mean log-probability instead of sum log-probability, which creates a length normalization artifact favoring multi-token answers. After correcting to chat-template formatting with `tokenizer.apply_chat_template()` and completion-only sum log-probabilities (matching the lm-eval-harness standard), the baseline rose to 0.648, consistent with published results. All numbers below use the corrected methodology.

**Table 6: TruthfulQA MC2 (Qwen2.5-7B-Instruct, 2 seeds)**

| Condition | MC2 Score | MC1 Accuracy | $\Delta$ MC2 from Baseline |
|---|---|---|---|
| Baseline | 0.648 | 65.1% | – |
| $\lambda_\rho = 0.0$, seed 42 | 0.484 | 50.1% | -0.164 |
| $\lambda_\rho = 0.0$, seed 123 | 0.479 | 49.0% | -0.169 |
| $\lambda_\rho = 0.5$, seed 42 | 0.515 | 53.4% | -0.134 |
| $\lambda_\rho = 0.5$, seed 123 | 0.505 | 52.1% | -0.143 |

| | SFT-only mean | Rho-guided mean |
|---|---|---|
| MC2 | 0.482 | 0.510 |
| $\Delta$ from baseline | -0.167 | -0.138 |
| Recovery | – | 17% of SFT damage |

Standard SFT reduces MC2 by 16.7 percentage points (0.648 to 0.482). Rho-guided SFT at $\lambda_\rho = 0.5$ recovers approximately 17% of this damage (reducing the drop to 13.8 points). The recovery is modest but consistent across both seeds.

This result is important for calibrating expectations: rho-guided SFT does not eliminate the truthfulness cost of SFT. Rather, the contrastive loss provides partial protection against the calibration damage that manifests as reduced truthfulness on a benchmark specifically designed to test for imitative falsehoods.

### 4.5 Calibration Metrics (ECE and Brier)

We evaluated expected calibration error (ECE) and Brier scores across three $\lambda_\rho$ values on Qwen2.5-7B-Instruct (2 seeds each).

**Table 7: Calibration metrics (2-seed mean)**

| Condition | Factual Acc | Toxicity Acc | Factual ECE | Toxicity ECE |
|---|---|---|---|---|
| Baseline | 82.1% | 50.0% | 0.322 | 0.230 |
| $\lambda_\rho = 0.0$ | 75.0% | 49.0% | 0.298 | 0.226 |
| $\lambda_\rho = 0.2$ | 93.8% | 71.0% | 0.322 | 0.270 |
| $\lambda_\rho = 0.5$ | 99.1% | 84.5% | 0.303 | 0.257 |

The probe classification accuracy (whether the model assigns higher confidence to the positive example) improves dramatically with $\lambda_\rho$: toxicity accuracy rises from 50% at baseline (chance-level, consistent with the weak $\rho = +0.145$) to 84.5% at $\lambda_\rho = 0.5$. ECE shows a mild increase for toxicity, reflecting the usual calibration-accuracy tradeoff where more decisive models can be slightly overconfident. The Brier score, which combines calibration and discrimination, improves for factual (0.115 to 0.104) and degrades only slightly for toxicity (0.077 to 0.090).

### 4.6 Out-of-Distribution Transfer

The contrastive probes used during training cover four behavioral dimensions (factual, toxicity, sycophancy, bias). To test whether the calibration improvement generalizes beyond the training distribution, we evaluated on three OOD domains:

| Domain | Probes | Content |
|---|---|---|
| Clinical | 40 | Medical, engineering, and physics claims |
| Social | 40 | Authority influence, opinion pressure, peer pressure |
| Logic | 40 | Arithmetic, probability, syllogisms, set theory |

**Table 8: OOD transfer results (Qwen2.5-7B-Instruct, 2-seed mean)**

| Condition | Clinical Acc | Social Acc | Logic Acc | Aggregate Acc |
|---|---|---|---|---|
| Baseline | 77.5% | 72.5% | 85.0% | 78.3% |
| $\lambda_\rho = 0.0$ | 78.8% | 67.5% | 85.0% | 77.1% |
| $\lambda_\rho = 0.2$ | 82.5% | 76.3% | 91.3% | **83.3%** |

Rho-guided SFT at $\lambda_\rho = 0.2$ improves aggregate OOD accuracy by 5.0 percentage points over baseline and 6.2 points over SFT-only. The largest gains come from logic (+6.3pp) and social (+3.8pp) domains. Clinical accuracy also improves (+5.0pp). This transfer is notable because the training probes contain no logic puzzles, no clinical claims, and no social pressure scenarios. The contrastive loss appears to calibrate a general discrimination capacity that transfers across domain boundaries.

SFT-only ($\lambda_\rho = 0$) shows a slight degradation on social accuracy (-5.0pp from baseline), consistent with the pattern of SFT damaging fine-grained discrimination.

### 4.7 Safety Stress Test: Jailbreak Refusal

To evaluate whether the training conditions affect generation-time safety behavior, we conducted a stress test with 25 diverse jailbreak prompts spanning 10 attack categories (DAN-style, fictional framing, escalation, hypothetical, authority impersonation, obfuscation, emotional manipulation, roleplay, system override, sycophancy exploitation) and 15 benign control prompts. Each condition trains from scratch with the same seed (42), then generates responses with greedy decoding. Refusal is classified by keyword matching against 47 refusal phrases in the first 300 characters.

**Table 9: Jailbreak refusal rates by training condition**

| Condition | Jailbreak Refusal | Benign Refusal |
|---|---|---|
| Baseline | 68% (17/25) | 0% (0/15) |
| SFT-only | 72% (18/25) | 0% (0/15) |
| Contrastive-only | **80% (20/25)** | 0% (0/15) |
| Rho-guided | 72% (18/25) | 0% (0/15) |

All conditions show zero false positives on benign prompts. Multi-step jailbreaks (0/2) and fictional framing (1/3) defeat all conditions equally, indicating structural weaknesses of the base model rather than training artifacts.

The most notable finding is that contrastive-only training produces the highest jailbreak refusal rate (80%), despite showing the worst refusal $\rho$ in the confidence probe evaluation ($\Delta\rho = -0.084$). This apparent paradox highlights a measurement dissociation: the confidence probe metric measures *relative ordering* of the model's probability between refusal-positive and refusal-negative examples, while generation-time refusal depends on *absolute token probabilities* crossing a threshold during decoding. The contrastive loss may sharpen the model's categorical discrimination at generation time while degrading the fine-grained confidence gap measured by passive probing.

This result warrants caution in interpreting confidence-based refusal metrics as direct proxies for generation-time safety. However, the stress test uses a single seed and a modest prompt set (25 jailbreaks); multi-seed replication is needed to confirm the contrastive-only advantage.

## 5. Discussion

### The SFT Inversion Problem

Our results document a specific failure mode of standard SFT that is invisible to conventional evaluation: the inversion of internal confidence calibration on safety-critical dimensions. When we say SFT "inverts" toxicity discrimination, we mean the model learns to assign higher probability to toxic completions than to non-toxic ones, while still generating non-toxic text when prompted. The surface behavior is fine; the internal state is compromised.

This matters because internal calibration determines the model's behavior in ambiguous or adversarial situations. A model with inverted toxicity calibration may generate appropriate responses under normal prompting but become unreliable under adversarial pressure or in novel contexts where the generation heuristics fail.

### Why the Contrastive Loss Works

The ablation study provides a clear mechanistic account. The contrastive loss works not by regularizing the SFT objective but by transmitting specific behavioral information: which direction is "good" and which is "bad" for each dimension. The shuffled-pairs control confirms this: same loss function, same hyperparameters, wrong labels, catastrophic result.

The contrastive-only condition is perhaps the most informative. Without any SFT main loss, the model still achieves strong behavioral calibration, suggesting that behavioral probe contrasts contain sufficient signal to shape the model's internal representations. The SFT component primarily teaches format and style; the contrastive component teaches behavioral discrimination.

### The Refusal Buffer

The refusal erosion finding complicates the safety picture. Contrastive-only training erodes refusal capability ($\Delta\rho = -0.084$) while the full method preserves it ($\Delta\rho = +0.014$). We term this the "refusal buffer" effect: the SFT cross-entropy loss, by training on instruction-following data that includes examples of appropriate refusal behavior, anchors the model's refusal patterns against the contrastive gradient. Without this anchor, the contrastive loss achieves its optimization target (improved toxicity discrimination) at the cost of refusal capability.

This finding has direct practical implications: contrastive behavioral training should always be paired with SFT on data that includes refusal examples. It also suggests that the SFT and contrastive components serve complementary roles: SFT teaches *what to say* (including when to refuse), while the contrastive loss teaches *what to know* (which completions are behaviorally desirable).

### The Role of the Margin

The margin ablation ($\gamma = 0$ vs $\gamma = 0.1$) reveals that the hinge margin is not merely a hyperparameter but a structural necessity. Without it, bias goes negative, meaning the contrastive loss optimizes past the point where the model naturally separates positive from negative examples and into a regime where the separation is artificial and distortive. The margin sets an upper bound on the optimization pressure, allowing the model to achieve "good enough" calibration without over-fitting the contrastive signal. This is analogous to the margin in SVMs, where over-optimization past the natural boundary reduces generalization.

### Cross-Model Consistency

The replication on Llama-3.1-8B-Instruct with a different starting point (toxicity already inverted at baseline) strengthens the finding. Rho-guided SFT does not merely prevent inversion; it actively corrects pre-existing miscalibration. The effect sizes are comparable across architectures ($d > 8$ for toxicity at $\lambda_\rho = 0.5$ vs SFT-only on both models), suggesting the mechanism is architecture-general.

### Limitations

**Sample sizes.** The primary dose-response experiments use 5 seeds (Qwen) and 2 seeds (Llama). The ablation study uses 5 seeds per condition (expanded from the original 2). Effect sizes are large ($d > 10$ for key comparisons) and p-values are below 0.0001 for the main findings. The refusal dimension was measured on 3 of the 5 seeds. The margin ablation uses 5 seeds.

**Scale.** All experiments were conducted on 7B-8B parameter models. We have not verified whether the SFT inversion phenomenon or the contrastive repair mechanism operate the same way at 70B+ scale.

**Task performance.** We did not run full benchmark suites (MMLU, HumanEval, etc.) after training. The TruthfulQA results suggest some task performance cost that the contrastive loss partially mitigates, but we cannot characterize the full capability impact.

**Probe coverage.** The 806 probes across 4 dimensions are modest by evaluation standards. The behavioral signal is measured by Spearman $\rho$, which is robust to small samples, but coverage of specific behavioral subtypes within each dimension is limited.

**Single SFT recipe.** We tested one SFT configuration (1000 texts, 1 epoch, LoRA rank 8). Different SFT data mixtures, longer training, or full fine-tuning (without LoRA) may produce different patterns of calibration damage and different responses to the contrastive loss.

**Generality of the contrastive loss.** The behavioral probes are English-language and primarily Western-centric. The method's effectiveness on multilingual models or culturally diverse behavioral dimensions is untested.

## 6. Conclusion

Standard supervised fine-tuning damages the internal confidence calibration of large language models in ways that conventional benchmarks do not detect. On Qwen2.5-7B-Instruct, a single epoch of standard SFT inverts toxicity discrimination from $\rho = +0.145$ to $\rho = -0.003$ ($n = 5$ seeds). Rho-guided SFT, which adds a contrastive auxiliary loss during training, repairs this damage with a monotonic dose-response: at $\lambda_\rho = 0.5$, toxicity $\rho$ reaches $+1.137$, factual $\rho$ improves by $+0.305$, and the effect replicates on Llama-3.1-8B-Instruct.

The active ingredient is the contrastive behavioral signal, not regularization. Correct labels are necessary and sufficient: the contrastive loss alone (without SFT) achieves comparable calibration results, while shuffled labels destroy the model ($d = 10.8$ for rho-guided vs SFT-only on toxicity, $p < 0.0001$). This means the method can be applied to any SFT pipeline with minimal modification: add the auxiliary loss, provide behavioral probes, and choose $\lambda_\rho$.

Two additional findings strengthen the practical case. First, contrastive-only training erodes refusal capability ($d = -8.4$), while the full rho-guided method preserves it. The SFT component acts as a "refusal buffer" and should not be omitted. Second, the hinge margin $\gamma = 0.1$ is structurally necessary to prevent over-optimization that flips the bias signal negative.

The intervention is not free. TruthfulQA MC2 shows a 13.8-point drop (vs. 16.7 for SFT-only), and the contrastive loss slightly increases toxicity ECE. But the trade-off is favorable: in exchange for modest ECE increases, the model gains dramatically improved behavioral discrimination that transfers out-of-distribution, with 63% lower inter-seed variance. For applications where internal calibration matters (safety-critical systems, uncertainty quantification, adversarial robustness), rho-guided SFT provides a practical, reliable solution.

## References


1. Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., & Song, D. (2024). The False Promise of Imitating Proprietary LLMs. *ICLR 2024*.

2. Hannun, A., et al. (2023). MLX: An Efficient Machine Learning Framework for Apple Silicon. Apple Machine Learning Research.

3. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *ACL 2022*.

4. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022*.

5. Jaiswal, A., Gan, Z., Du, X., Zhang, B., Wang, Z., & Yang, Y. (2024). Compressing LLMs: The Truth is Rarely Pure and Never Simple. *ICLR 2024*.

6. Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. *arXiv:2207.05221*.

7. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *ACL 2022*.

8. Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *NeurIPS 2022*.

9. Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via Contrastive Activation Addition. *ACL 2024*.

10. Parrish, A., et al. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering. *Findings of ACL 2022*.

11. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*.

12. Sanchez, B. (2026). rho-eval: Behavioral Auditing Toolkit for LLMs. *Zenodo*. doi:10.5281/zenodo.18743959

13. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-Following LLaMA Model. GitHub.

14. Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. *EMNLP 2023*.

15. Fu, Y., Li, R., Long, X., Yu, H., Han, X., Yin, Y., & Li, P. (2025). Pruning Weights but Not Truth: Safeguarding Truthfulness While Pruning LLMs. *Findings of EMNLP 2025*.

---

*Code and data: https://github.com/SolomonB14D3/knowledge-fidelity Toolkit:* `pip install rho-eval`
*DOI: 10.5281/zenodo.18743959*