Correctness-Optimized Residual Activation Lens (CORAL): Transferrable and Calibration-Aware Inference-Time Steering

Miranda Muqing Miao, Young-Min Cho, Lyle Ungar

ICML 2026, Seoul, South Korea

University of Pennsylvania

Department of Computer and Information Science

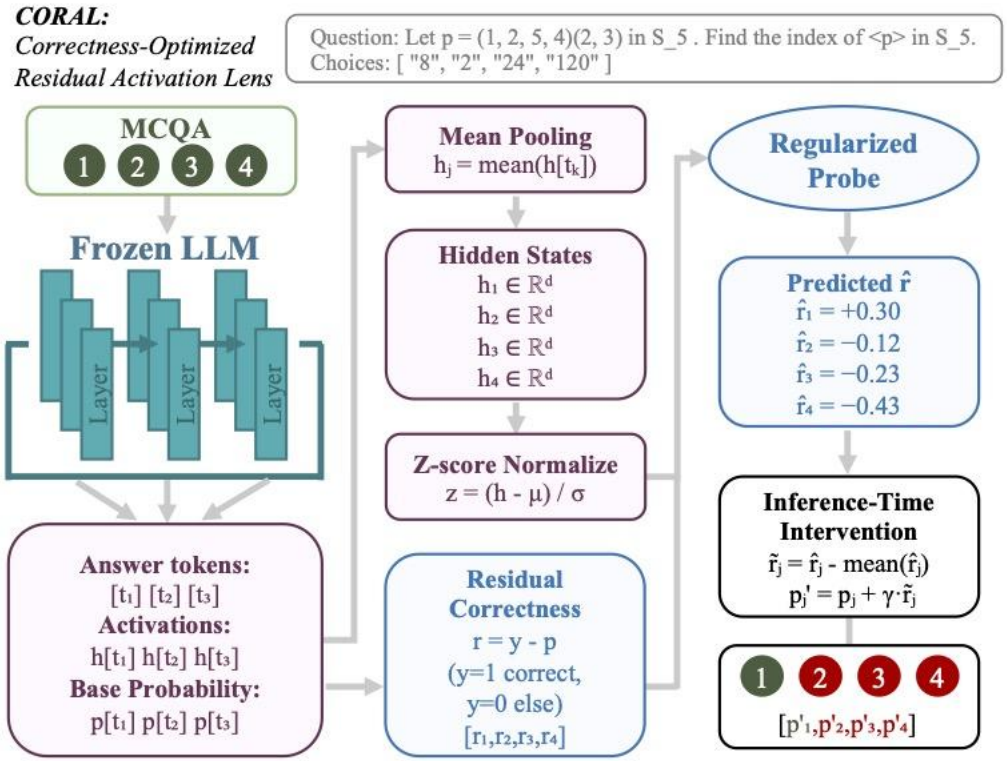
Penn Engineering

LLMs are Miscalibrated, and Alignment Makes it Worse

- **The Calibration Gap:** RLHF and DPO improve helpfulness but significantly worsen model calibration.

- **Proxy Optimization:** Existing steering methods target proxies like confidence or truthfulness rather than direct correctness.

CORAL: Direct Residual Correctness Steering



- **1. Extract**

  Mean-pool hidden states, z-score normalize.

- **2. Predict**

  MLP probe predicts residual correctness $r = y - p$.

- **3. Steer**

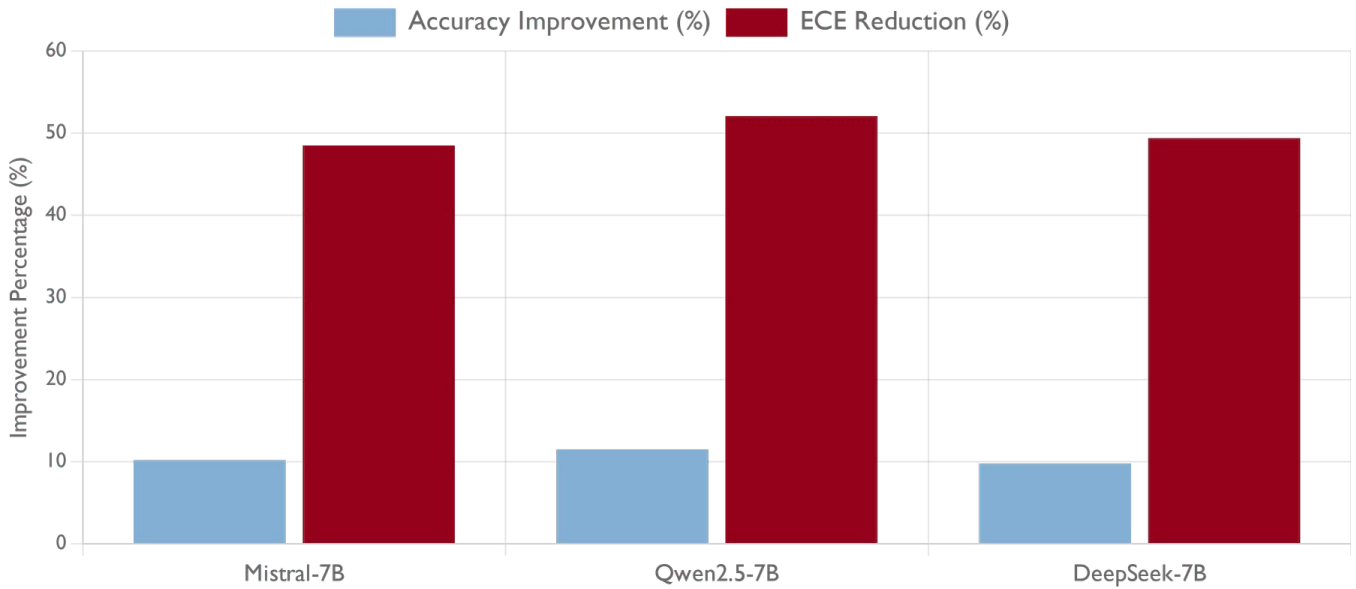
  Apply additive correction to base probabilities.

Why Residual Correctness?

- **The Formulation:** Residual $r_j = 1 - p_j$ if correct, and $-p_j$ if incorrect.
  - Directly captures the gap between ideal and predicted probabilities.
- **Directional Insights:** Residuals provide interpretable signals for steering.
  - Positive residuals: Underconfidence on correct answers.
  - Negative residuals: Overconfidence on incorrect answers.
- **Mathematical Foundation:** Directly targets the Brier score.
  - Brier score decomposes as $\sum r_j^2$, making residual prediction a principled way to minimize error.

Main Results: Significant Gains in Accuracy and Calibration

CORAL consistently improves performance across all 7B-parameter models tested.



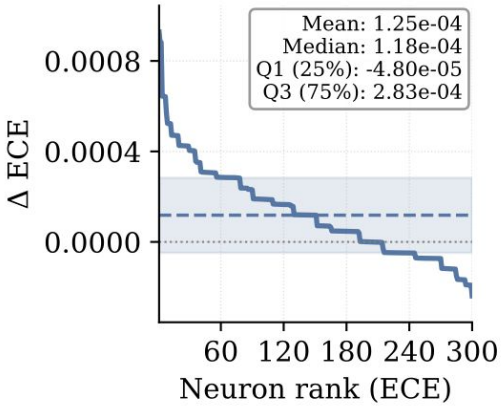
**Average improvements across MMLU, RACE, and CommonsenseQA benchmarks.*

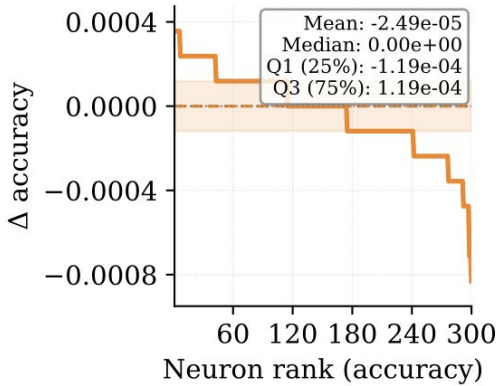
Transfer Results: Zero-Shot Generalization

Benchmark	Model	Baseline Acc	CORAL Acc	ECE Change
ARC-Challenge	Mistral-7B	63.1%	**73.5%**	-65%
HellaSwag	DeepSeek-7B	75.4%	**82.0%**	-20%
OpenBookQA	Qwen2.5-7B	49.8%	**65.4%**	-59%
Math-MC	Mistral-7B	27.4%	**45.7%**	-84%
Average	All Models	—	**+14.0%**	-49.0%

** Probe trained on CommonsenseQA + RACE generalizes to science, math, and reasoning without retraining.*

Correctness is Distributed, Not Sparse





- **SAE Ablation Analysis**

  Individual SAE features show negligible causal impact (mean 0.01 percentage points) on calibration.

- **Distributed Nature**

  Correctness is a collective property of the activation space, requiring supervised aggregation.

- **Probe Advantage**

  Regularized probes capture distributed signals that sparse feature selection methods miss.

Layer Analysis and Practical Efficiency

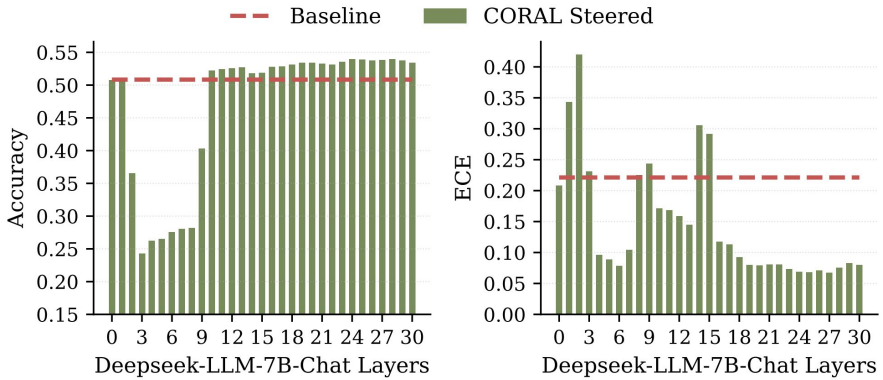


Figure 2: Layer-wise accuracy and ECE for DeepSeek-7B

- **Optimal Steering Layers**

  Middle layers (17-21) yield the strongest improvements in both accuracy and calibration.

- **Training Efficiency**

  Requires <5 hours on a single RTX 2080 Ti using only cached activations.

- **Inference Overhead**

  Minimal +0.3% latency (~20ms per question) with no backpropagation through the base model.

Summary and Key Contributions

- **Lightweight & Practical Steering**
  - Achieves +10% accuracy and -50% ECE with minimal compute overhead.
  - Inference latency increases by only 0.3% (~20ms per question).
- **Robust Zero-Shot Transferability**
  - Generalizes to science, math, and reasoning benchmarks without retraining.
  - Captures a general correctness subspace rather than task-specific patterns.
- **Distributed Signal Discovery**
  - SAE analysis confirms correctness is a collective property of the activation space.
  - Code and probe weights are publicly available for the community.