RLAE Technical Experimentation Report

- **Date:** January 16, 2026
- **Author:** Pardhu Varma
- **Experimental Subject:** RLAE (Runtime Low-Rank Adaptive Environments)
- **Experimental Scope:** Sprints 1 through 6
- **Experimental Objective:** Comparative Analysis of Reversibility in Parameter-Efficient Adapters versus Weight Mutation.
- **Research Paper:** On the Structural Limitations of Weight-Based Neural Adaptation and the Role of Reversible Behavioral Learning.

---

1. Abstract

This report summarizes the experimental findings of the RLAE (Runtime Low-Rank Adaptive Environments) research. The objective was to examine the structural recoverability of Large Language Models (LLMs) following behavioral adaptation. Using a standardized *Identity Stress* protocol, we compared two adaptation paradigms: **Behavioral Adapters (RLAE)** and **Weight Mutation (traditional fine-tuning)**. The results reveal a consistent structural asymmetry. Behavioral adapters enable full post-adaptation recovery, exhibiting near-zero Kullback–Leibler (KL) divergence after reset, whereas weight mutation introduces persistent, intensity-dependent deviations in the model's output distribution. These residual deviations are referred to here as *identity scars*.

---

2. Methodology: The Identity Stress Protocol

All experiments followed a controlled state-transition lifecycle $S_0 \rightarrow S_{\text{adapt}} \rightarrow S_{\text{reset}}$ to measure behavioral information loss and recoverability.

1. **Baseline (S_0):** Deterministic state of the frozen base model (Seed 1337).

2. **Adaptation (S_{adapt}):** Application of behavioral modification via either adapter injection or direct weight mutation.

3. **Reset (S_{reset}):** Attempted restoration of the baseline state via adapter unloading or reset training (KL minimization).

4. **Verification:** Divergence between S_0 and S_{reset} was measured using:

    - **Kullback–Leibler (KL) Divergence:** $D_{KL}(P \parallel Q) = \sum_x P\,(x) \log \frac{P(x)}{Q(x)}$
    - **Recoverability Factor (RF):** A normalized metric where $\text{RF} = 100$ denotes exact reproduction of baseline logits.

3. Phase I: Preliminary Staging (Sprint 1)

Objective: Validation of experimental controls and determinism.

Before executing the M-series experiments, staged runs were conducted to validate the stability and determinism of the evaluation pipeline.

- **Findings:** Initial baseline measurements exhibited small entropy variations due to non-deterministic CUDA behavior.
- **Resolution:** Enforcing strict seeding (torch.manual_seed(1337)) and deterministic execution eliminated these fluctuations.
  - *Reference artifact (exp1_results.json):* Prompt p1 baseline entropy = 0.1727.
- **Conclusion:** A stable experimental floor was established, ensuring that subsequent divergence measurements ($\Delta > 10^{-6}$) reflected adaptation effects rather than stochastic noise.

4. Phase II: The M-Series Experiments (Sprints 2–6)

The core investigation was organized into four experimental modules (M1–M4).

M1 & M2: RLAE Validation (Sprint 2)

- **Objective:** Empirical validation of the reversibility properties of RLAE.
- **Protocol:** Behavioral adapters were applied and subsequently unloaded at varying elimination factors (ϵ).
- **Quantitative Results (exp2_rlae_results.json):**
  - $\epsilon = 0.0$: $D_{KL} \approx 0.0599$ (baseline deviation present)
  - $\epsilon = 0.2$: $D_{KL} \approx 0.0468$
  - $\epsilon \geq 0.6$: $D_{KL} = 0.0000$ (full state recovery)
- **Observation:** For elimination factors $\epsilon \geq 0.6$, the model returned to a state statistically indistinguishable from the baseline, demonstrating that RLAE supports temporary behavioral modification without permanent structural alteration.

M2.5: Structural Variance & Adversarial Robustness (Sprint 3)

- **Objective:** Evaluation of model stability under structural perturbations (SVAR).

- **Protocol:** Injection of noise, dropout, and adversarial tensors during inference time to measure deviation resilience.
- **Quantitative Results (exp3_svar_results.json):**
-

PERTURBATION TYPE	INTENSITY	D_{KL} DEVIATION	ANALYSIS
Layer Dropout	0.25	≈ 0.001	Highly Robust
Weight Decay	0.1	≈ 0.002	Negligible
Gaussian Noise	0.01	≈ 0.015	Minor Drift
Adversarial	0.05	≈ 0.265	**Significant Divergence**

- **Observation:** The model exhibits high structural rigidity against random noise and dropout, maintaining low divergence (< 0.02). However, adversarial perturbations induce significant deviations (≈ 0.265), defining the upper bound of the model's stability envelope.

M2.6: Operational Stress Testing (Sprint 4)

- **Objective:** Verification of inference stability under sustained operational load.
- **Protocol:** High-frequency inference loops (ITER_0) to detect memory leaks or logical drift over time.
- **Findings (exp4_stress_results.json):**
    - **Memory Stability:** Consistent usage (~5.9 GB) with no accumulative leakage.
    - **Output Coherence:** Text integrity maintained verifying operational reliability.

M3: Mutation Intensity Sweep (Sprint 5)

- **Objective:** Assessment of irreversibility under weight mutation as a function of mutation intensity (α).
- **Protocol:** Direct weight perturbations were applied at increasing intensities, followed by reset attempts.
- **Quantitative Results (exp5_m3_sweepresults.json):**

INTENSITY (α)	POST-RESET D_{KL}	QUALITATIVE OUTPUT	STATUS
0.001 (Low)	**0.462**	Coherent, altered style	**Scarred**
0.010 (Medium)	**10.928**	Severely degraded	**Degraded**
0.050 (High)	**18.933**	Corrupted output	**Corrupted**

- **Observation:** Weight mutation resulted in monotonic and irreversible divergence. Even the smallest tested intensity ($\alpha = 0.001$) produced measurable residual divergence after reset, confirming structural hysteresis. At $\alpha \geq 0.01$, the model entered a regime of catastrophic degradation, indicating that direct weight modification fundamentally compromises reversibility.

M4: Scale Invariance (Sprint 6)

- **Objective:** Validation of structural behavior across different model scales.
- **Artifact:** exp_m4_multimodelrun
- **Aggregate Results:**

MODEL SIZE	METHOD	RECOVERABILITY FACTOR (RF)	ANALYSIS
1.5B (Small)	RLAE	**100.0%**	Full recovery
	Mutation	30.0%	Partial deviation
3B (Medium)	RLAE	**100.0%**	Full recovery
	Mutation	40.0%	Partial deviation
7B (Large)	RLAE	**100.0%**	Full recovery
	Mutation	**10.0%**	Significant degradation

- **Observation:** Susceptibility to identity scarring increased with model scale. Larger parameter spaces exhibited lower recoverability under weight mutation, suggesting increased vulnerability to irreversible representational drift.

---

5. Discussion

The experimental results support a structural distinction between the two adaptation paradigms:

1. **RLAE (Isomorphic Adaptation):** Acts as an additive modification. Unloading the adapter restores the original parameter state W_0 exactly.
2. **Weight Mutation (Anisotropic Adaptation):** Acts as a transformative modification ($W_0 \rightarrow W'$). The inverse path $W' \rightarrow W_0$ is ill-posed; information lost during gradient updates cannot be reliably reconstructed through optimization alone.

---

6. Conclusion

The M-series experiments demonstrate that **weight mutation introduces irreversible behavioral changes** whose severity increases with both training intensity and model scale. In contrast, **RLAE enables near-zero divergence recovery** by structurally isolating adaptive behavior from core parameters.

For systems requiring auditability, controlled rollback, or long-term behavioral governance, reversible behavioral adaptation provides a structurally robust alternative to direct weight modification.