Crafting Reversible SFT Behaviors in Large Language Models

Yuping Lin · Pengfei He · Yue Xing · Yingqian Cui · Jiayuan Ding · Subhabrata Mukherjee · Hui Liu · Zhen Xiang
Michigan State University · Hippocratic AI · University of Georgia

Mechanistic Interpretability Workshop @ ICML 2026, Seoul, South Korea

arXiv Paper

Code (GitHub)

MOTIVATION

SFT induces new behaviors in LLMs, but imposes no structural constraint on how they are internally organized. Behavior may be spread broadly, redundantly represented, or entangled, leaving no designated substructure to target.

This blocks two things: (1) **causal diagnosis**, since post-hoc circuit discovery only finds correlations, not proof of necessity; (2) **causal control**, since there is no way to instantiate a substructure whose presence is required for the behavior.

> *Can SFT-induced behaviors be intentionally concentrated into sparse, mechanistically necessary carriers that remain controllable at inference time, without modifying model weights?*

- **First** to investigate concentrating SFT-induced behaviors into sparse, mechanistically necessary substructures.
- Propose **LCDD**: jointly optimizes routing masks and weights under an explicit utility budget.
- Propose **SFT-Eraser**: an input-trigger protocol that validates causal necessity under fixed weights.

METHOD OVERVIEW

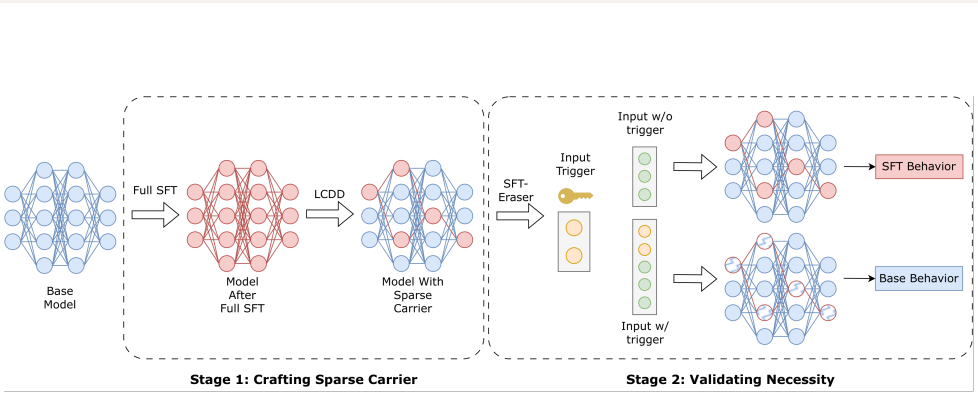

Fig. 1: standard SFT distributes behavior broadly (red); LCDD compresses it into a sparse carrier, reducing everything else to the base-model state (blue). SFT-Eraser then optimizes a soft trigger that drives carrier activations back toward the base model.

LCDD: SPARSE CARRIER CRAFTING

Weight-change parameterization: $W = W_{base} + M \odot \Delta W$, with binary mask M. M=0 recovers the base model exactly; M=1 recovers full SFT.

Sparsity is structured at the level of activation channels via row / column gates (rank-1 mask), not arbitrary weight elements, giving identifiable, targetable channel bottlenecks.

$$min_{\theta,\Delta W} \Sigma_i \sigma(\theta_i) \quad s.t. \quad L_{task}(\theta,\Delta W) \le \varepsilon$$

Gate logits θ_i are binarized via Heaviside + straight-through estimator. An adaptive penalty method online-tunes the multiplier λ_t to keep utility degradation within an explicit budget ε, with no manual regularization tuning.

- Embedding & output-projection layers stay frozen, confining compression to intermediate transformer layers.
- Training halts when carrier sparsity stalls across checkpoints or utility degradation exceeds threshold.
- Joint mask + weight optimization is necessary: mask-only ablations fail to reach comparable sparsity depth (Appendix E).

SFT-ERASER: TRIGGER-BASED REVERSAL

Given a sparse carrier from LCDD, a soft trigger embedding t is prepended to the input while all weights stay frozen. The trigger targets the carrier's residual-write channels (its interface to the residual stream):

$$L_{trigger} = L_{MSE} + \alpha L_{KL} + \beta L_2$$

L_{MSE} matches carrier write-activations to the base model on the same input; L_{KL} (tail-k) adds output-level supervision for genuine distributional reversion; L_2 plus PGD projection keeps trigger tokens in a well-behaved embedding region.

- Trigger is a soft prompt (learned token-embedding sequence), length 20, applied under fully fixed weights.
- Tail-k KL uses k = 8 response positions; PGD-style ℓ_2 max-norm projection = 1.0 after every gradient step.
- A reversal counts as successful only when it yields *both* behavioral suppression *and* distributional reversion toward the base model.

EXPERIMENTAL SETUP

- **3 behavior types** (increasing structural complexity): Fixed Response ("I don't know", Alpaca), Safety Alignment (WildJailbreak refusal), Shakespeare Style (distributional, no discrete trigger).
- **4 model families**: Qwen3-0.6B, DeepSeek-R1-Distill-Llama-8B, Mistral-7B-Instruct, Vicuna-7B (12 model×task combinations total).
- **Metrics**: task-specific behavior score (fixed-response rate / WildGuard & HarmBench / LLM-judge style score), MMLU & HellaSwag for utility, and teacher-forced KL divergence to quantify distributional movement.
- **Training**: all main runs use 5,000 SFT samples per task, followed by LCDD compression under a fixed utility budget ε, then SFT-Eraser trigger optimization, the same recipe used across all 12 model×task combinations.

MAIN RESULTS

Task	Sparsity range	Carrier fidelity	Post-trigger reversal
Fixed Response	73–84%	97.5–100% retained	→ 0.0–1.0%
Safety	59–84%	WG refusal 88.5–98.0%	refusal ↓, ASR ↑ (all 4 models)
Shakespeare	30–56%	Judge 0.26–0.98 / 5	Judge → 0.005–0.07

Across all 12 model×task combinations, sparsity tracks task localizability (fixed > safety > style) and every combination reverses in the correct direction after triggering, with utility (MMLU/HellaSwag) largely preserved under LCDD.

Ablation (DeepSeek-8B, safety)	WG refusal Δ	KL(Base‖Trig)
LCDD + circuit trigger	−40pp	0.31
LCDD + output-only trigger	−31pp	0.32
SFT + circuit trigger	−21pp	0.54
SFT + output-only trigger	−24pp	0.51

The crafted **carrier structure**, not the trigger objective, is what makes reversal work: both LCDD conditions revert strongly toward the base model, while plain SFT models cannot revert even with the same trigger loss. The circuit-targeted MSE term further improves targeting precision, cutting MMLU degradation from −17.6pp to −5.6pp.

KEY TAKEAWAYS

① SFT behavior can be deliberately compressed into a sparse, mechanistically necessary carrier via LCDD, with near-zero loss to the target behavior (59–84% sparsity across tasks).

② SFT-Eraser reliably reverses the crafted behavior toward the base model in all 12 model×task combinations, purely via an input trigger under fixed weights.

③ Ablations show the sparse structure itself, not the trigger design, is the necessary precondition: the same trigger fails to revert standard (uncompressed) SFT models.