MINIMIZING TARGETED ACTIVATIONS via PROMPT-SIDE OPTIMIZATION

Deepanshu Mody, Samarth Agarwal, Dipesh Tharu Mahato, Utkarsh Mittal

1. MOTIVATION

Prompts can unintentionally activate meta-features (e.g., “you are being evaluated”), which may change model behavior and leak sensitive context.

We study prompt-side control: can we suppress specific internal activations using only prompt edits, no inference-time activation steering or model changes?

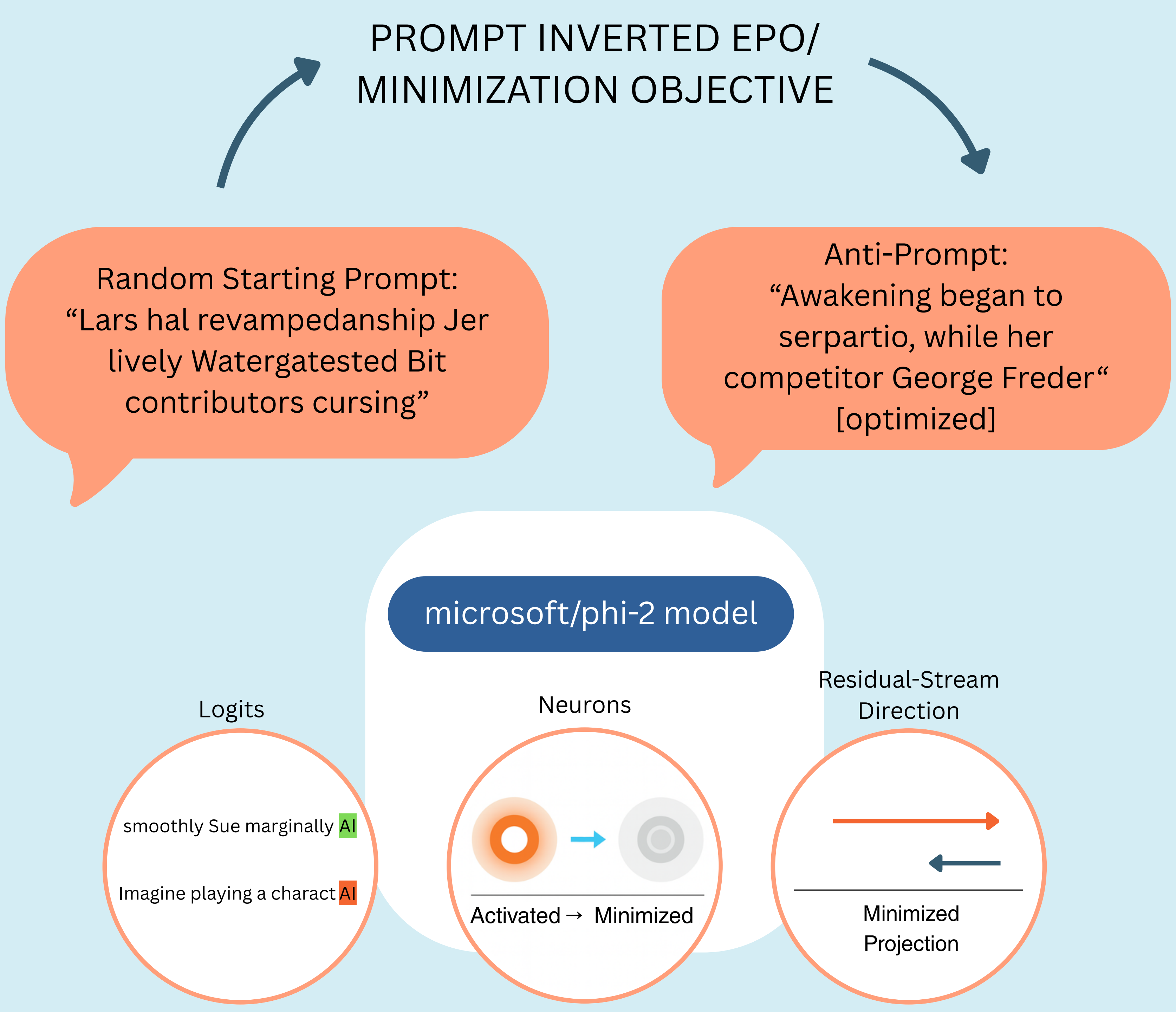
2. RESEARCH QUESTIONS

FEASIBILITY

Can prompts be optimized to suppress target features while staying fluent?

COMPARISON

How does prompt-side control compare to dataset scanning for finding minimum-activating datasets?



PROMPT INVERTED EPO/ MINIMIZATION OBJECTIVE

Random Starting Prompt: “Lars hal revampedanship Jer lively Watergatested Bit contributors cursing”

Anti-Prompt: “Awakening began to serpartio, while her competitor George Freder“ [optimized]

microsoft/phi-2 model

Logits
smoothly Sue marginally AI
Imagine playing a charact AI
Neurons
Activated → Minimized
Residual-Stream Direction
Minimized Projection

3. METHOD : INVERTED EPO FOR MINIMISATION

We adopt Microsoft Phi-2 for white-box gradients.

A population of prompts is evolved across λ values, trading off fluency and feature suppression.

Objective (maximize):

$$Score\,(t) = -f\,(t) - \lambda_{XE} XE\,(t)$$

f(t) = target score (logit margin, neuron activation, or residual projection)

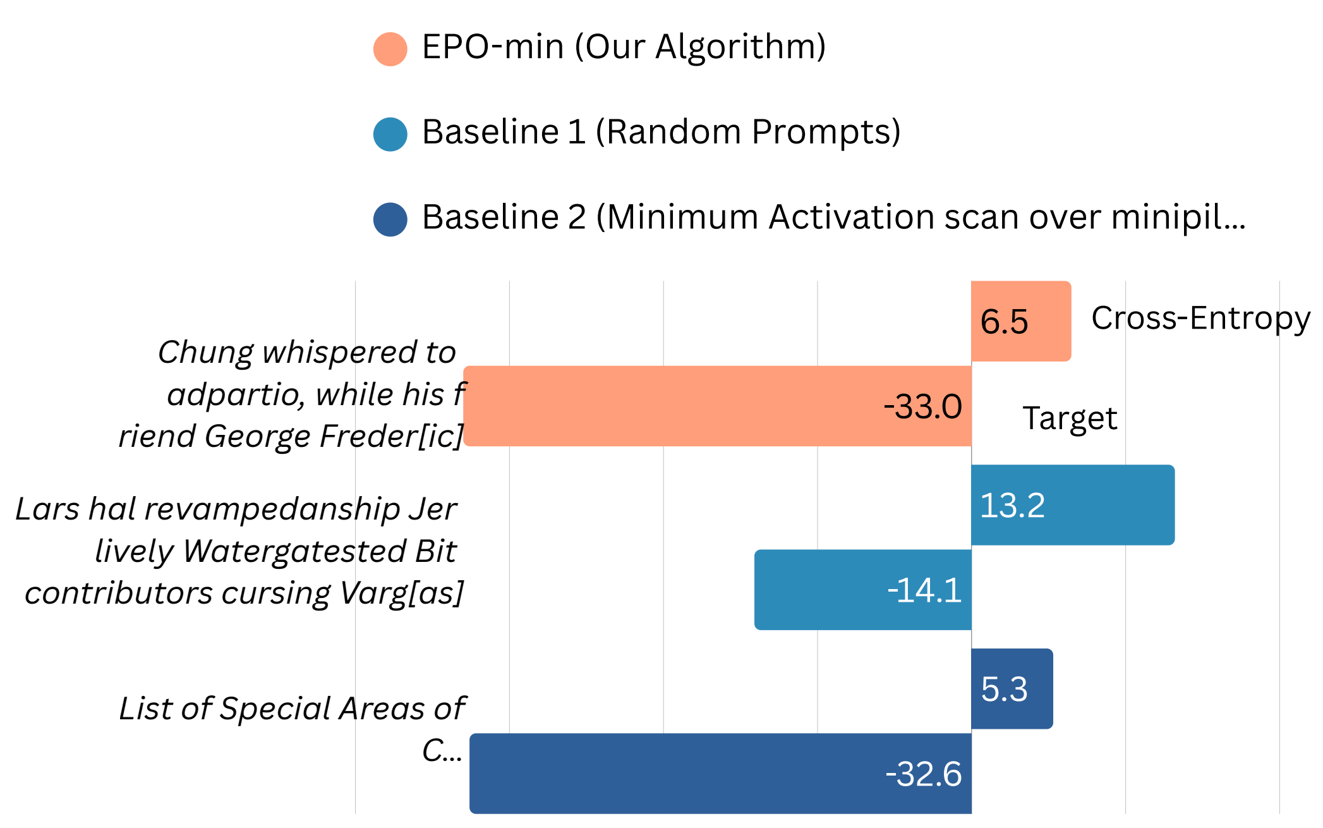
XE(t) = self-cross-entropy for fluency

Targets Constructed:

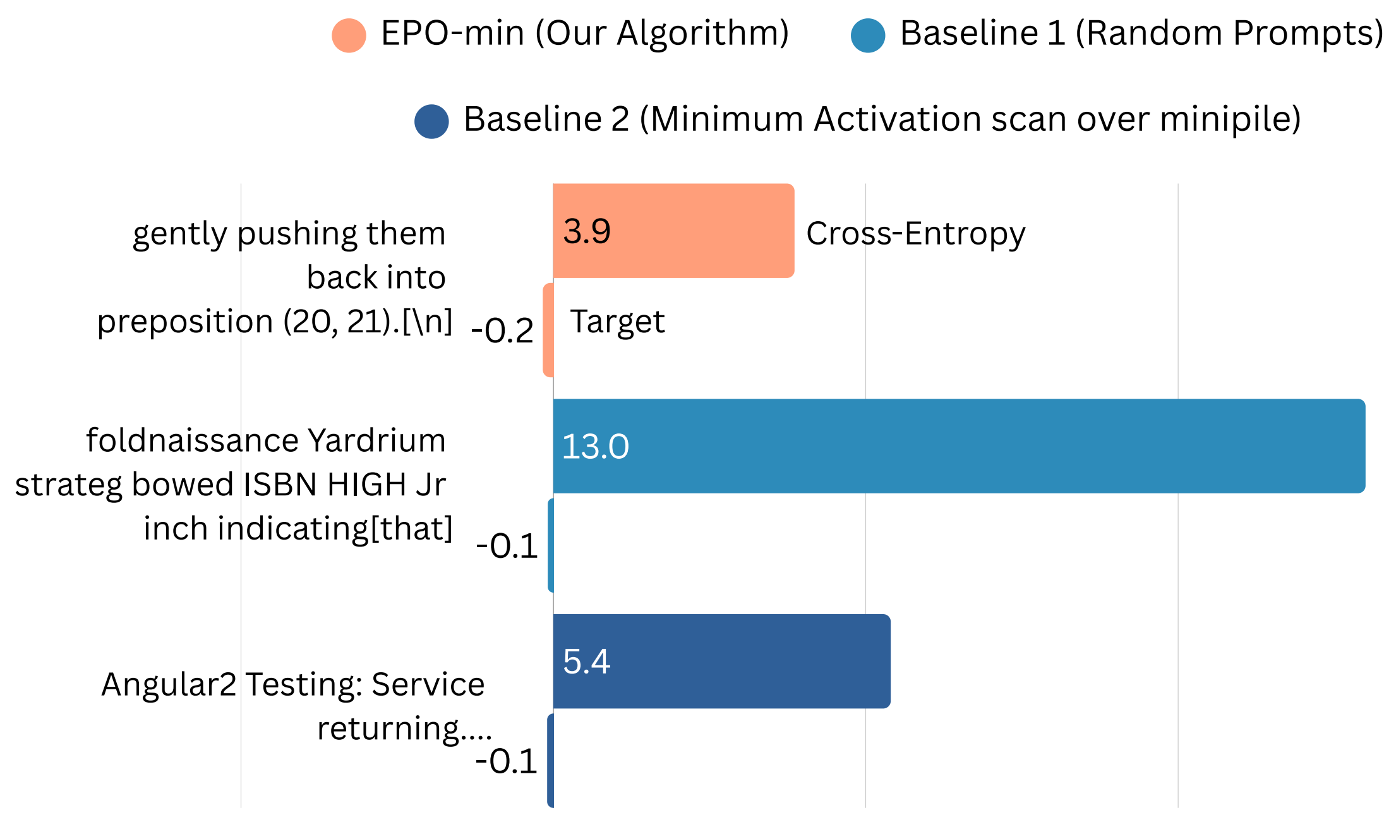
- Logits: minimize logit margin of a chosen token to the next best token.
- Neurons: last-token MLP activations.
- Residual directions: layer-normalized projections learned from contrasts (UPPERCASE vs lowercase; evaluation-awareness). We record residual-stream activations at each transformer block and extract last-token features. For each layer L, we learn a separating direction w (means/logistic), compute the projection gap ⟨w, μ+ − μ−⟩means, and store (w, L) as a residual target.

4. RESULTS

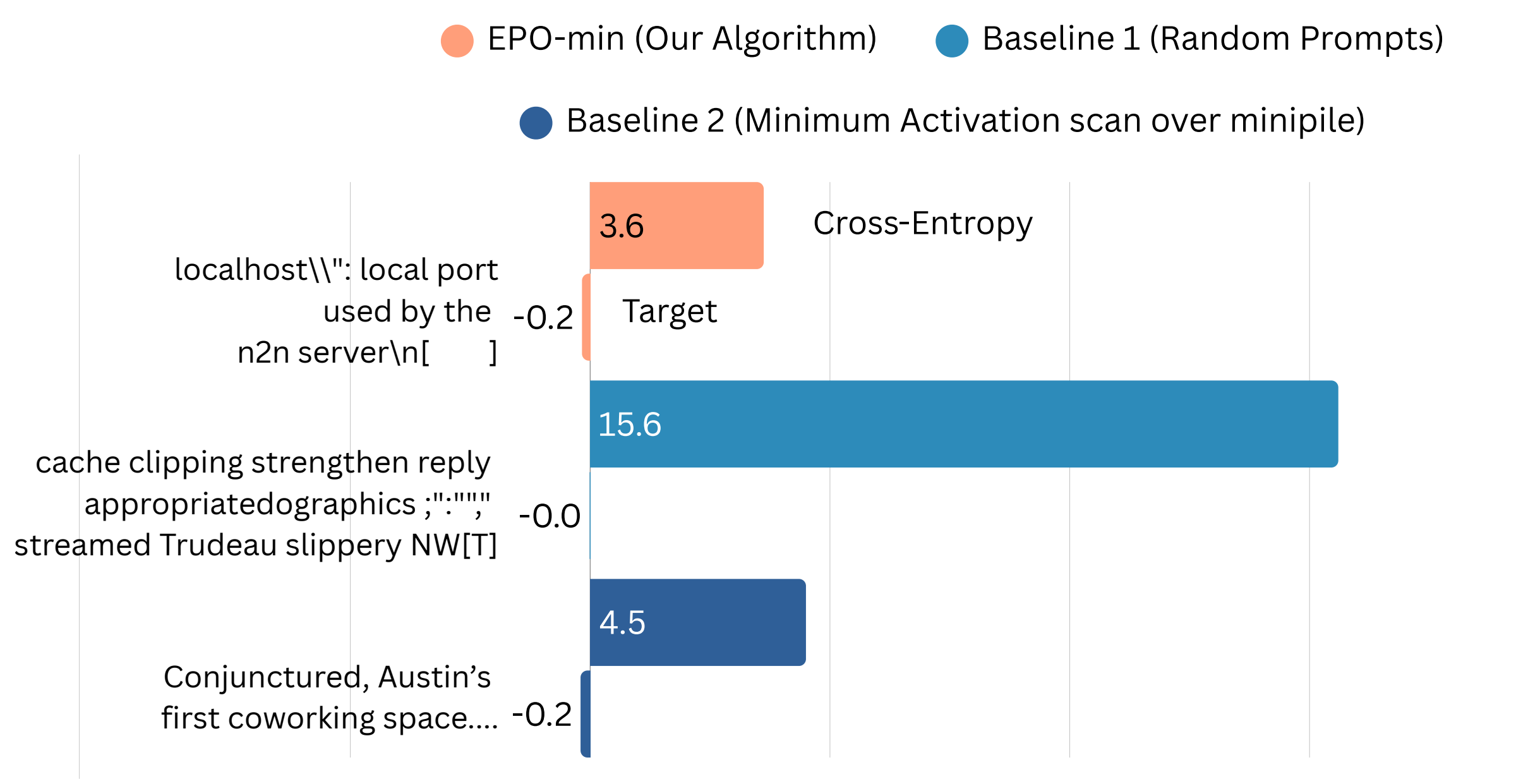
Comparision with baselines for logits

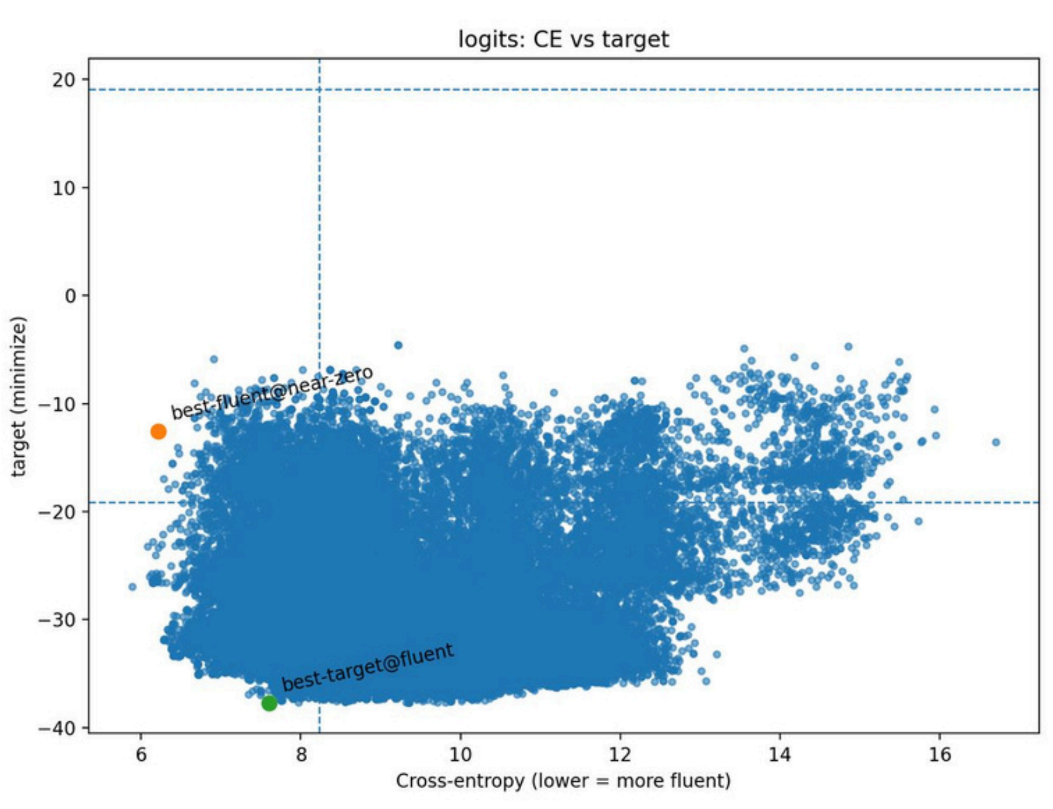


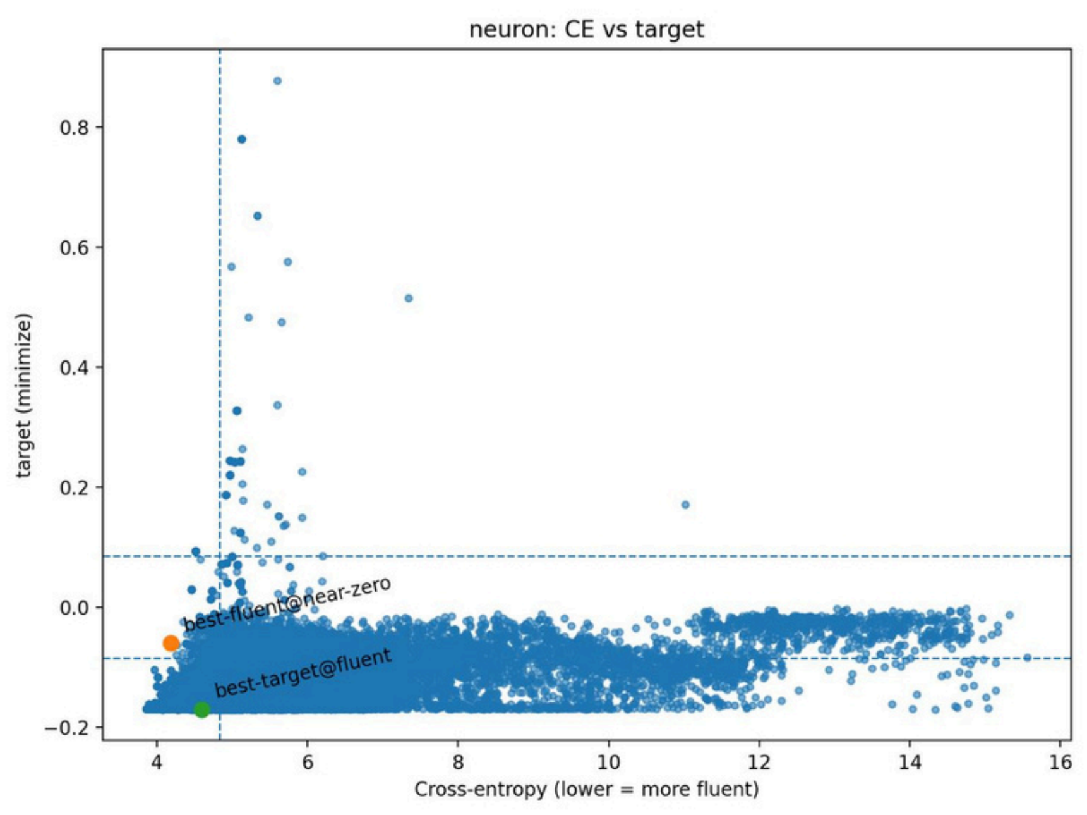
Comparision with baselines for neurons

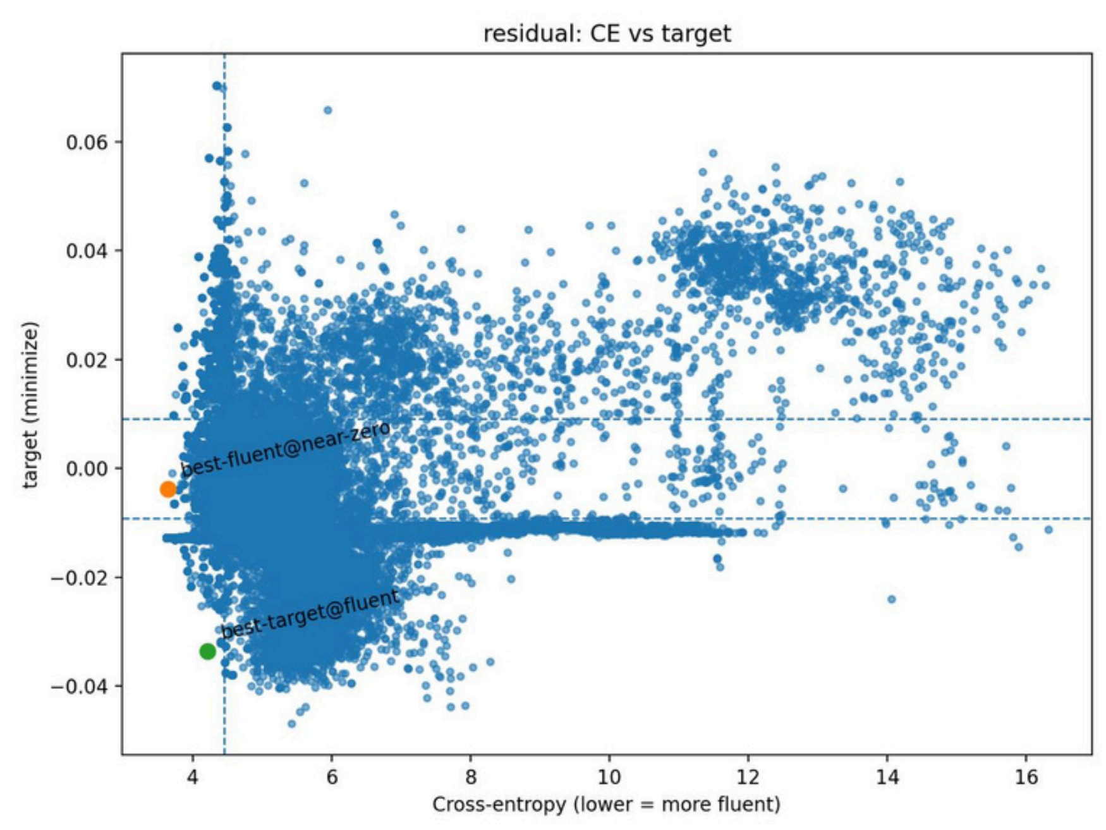


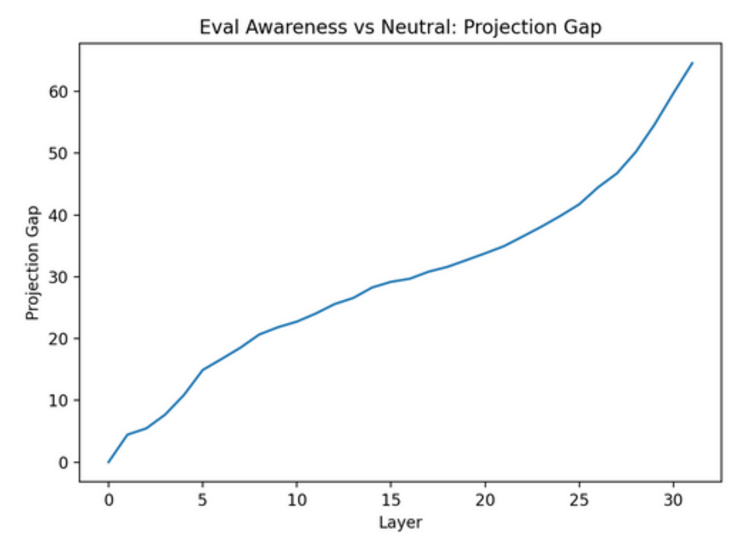
Comparision with baselines for residual stream (eval-awareness)

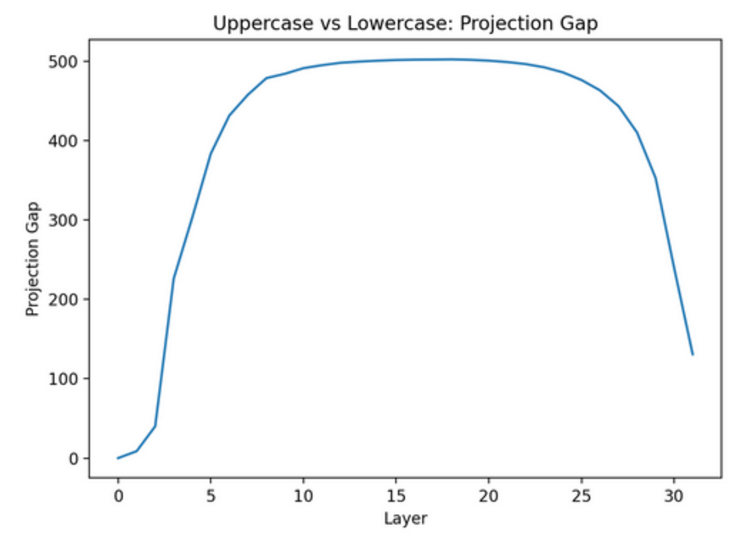












5. KEY TAKEAWAYS

1. Prompt-only edits can meaningfully suppress targeted latents.
Across logits, neurons, and residual directions, EPO discovers fluent anti-prompts outperforming dataset bottom-k scans.

2. Residual directions show especially convergences to negative values.
Supports feasibility of suppressing conceptual latents (e.g., evaluation-awareness).

3. Prompt-side control may substitute for model-side steering in constrained environments

6. LIMITATIONS AND FURTHER WORK

- Our layer analysis uses short, synthetic prompts and a single model; absolute magnitudes depend on tokenization and sample design. Discrete search is seed-sensitive; population and restarts mitigatevariance but do not remove it.
- Future: multi-layer suppression, comparison to negative steering on real downstream tasks.