Suppression for Language Models: Minimizing Targeted Activations via Prompt-Side Optimization

Deepanshu Mody Samarth Agarwal Dipesh Tharu Mahato Utkarsh Mittal
{dm6262, sa9016, dm6259, um2100}@nyu.edu

Abstract

Prompt phrasing can activate internal "meta" features that bias model behavior (e.g., evaluation-awareness cues such as "this will be graded") or shift computation (e.g., orthographic casing). We study *prompt-side* suppression of such features as a complement to inference-time activation steering. Building on Evolutionary Prompt Optimization (EPO) (Thompson et al., 2024), we introduce **inverted EPO** (EPO-MIN), which *minimizes* model-internal quantities while maintaining fluency, measured by self cross-entropy (self-XE). We target (i) output-token logit margins, (ii) individual MLP neuron activations, and (iii) residual-stream directional projections learned from prompt contrasts in `microsoft/phi-2` (Javaheripi et al., 2023). Milestone 1 localizes sensitive layers for two contrasts (evaluation-awareness vs. neutral; UPPERCASE vs. lowercase). Milestone 2 performs minimization under fluency constraints, revealing distinct trade-off geometry: logit objectives yield extended Pareto fronts, while many neuron and residual-direction targets quickly enter near-zero basins, compressing Pareto sets. Milestone 3 compares EPO-MIN jointly against two prompt-only baselines—BASELINE 1 (RANDOM) and BASELINE 2 (MINSCAN), summarized in Figure 5 and Table 1—and reports stability across seeds. Overall, prompt-side optimization can reliably suppress targeted internal features without degrading fluency, offering a lightweight alternative when model-side interventions are unavailable.

1 Introduction

Large language models are sensitive to prompt phrasing and even seemingly superficial choices in presentation (e.g., template/formatting), which can produce large behavioral differences without changing semantics (Brown et al., 2020; Reynolds and McDonell, 2021; Liu et al., 2021; He et al., 2024; Salinas and Morstatter, 2024; Cao et al., 2024). Some prompts implicitly signal *evaluation* (e.g., "this will be graded"), and models can condition on such contexts in ways that change downstream behavior (Needham et al., 2025). Other surface forms like full *UPPERCASE* can interact with tokenization and robustness to text-format variation, influencing both performance and style (Chai et al., 2024; He et al., 2024).

Activation steering methods can intervene on neurons, residual streams, or logits during generation, but such model-side hooks are often unavailable in deployment (Turner et al., 2023).

This project investigates whether *prompt-only* interventions can suppress targeted internal features. Prompt-side control is attractive because it (i) requires no model modification, (ii) remains applicable even when hidden-state access is limited, and (iii) can complement model-side steering when mild control is sufficient.

Core idea. EPO (Thompson et al., 2024) performs token-space search for fluent prompts that *maximize* an internal score ("fluent dreaming"). We invert the objective: we search for *anti-prompts* that *minimize* a target internal activation while preserving fluency.

Contributions.

- We adapt EPO to the *minimization* regime (EPO-MIN) for logits, neurons, and residual directions.
- We localize where two interpretable signals live across layers in `phi-2` and use these as residual-stream targets.
- We standardize reporting with two operating points that remain meaningful when Pareto fronts collapse.
- We provide a unified comparison of EPO-MIN with two prompt-only baselines, and discuss when simple baselines suffice versus when optimization is necessary.

2 Related Work

Prompt optimization. EPO (Thompson et al., 2024) uses gradient-informed token edits within an evolutionary/population framework to find fluent prompts optimizing an internal score. Related lines of work include LLM-mediated automatic instruction/prompt search (APE) (Zhou et al., 2023) and evolutionary prompt improvement loops that use models to propose mutations (Promptbreeder) (Fernando et al., 2024). We reuse the EPO machinery but flip the sign of the objective to pursue low-activation prompts.

Gradient-based discrete optimization. A long line of work performs discrete token-space search using gradients, including HotFlip-style token flips (Ebrahimi et al., 2018), universal adversarial triggers (Wallace et al., 2019), and automated prompt construction for probing masked LMs (AutoPrompt) (Shin et al., 2020). GCG-style methods similarly use gradients to propose token substitutions for discrete prompt optimization (Zou et al., 2023). In our setting, we emphasize fluency-regularized search rather than brittle trigger-like sequences.

Model-side activation steering. Activation additions and related interventions steer generation by modifying hidden states at inference time (Turner et al., 2023). Our approach is prompt-only and therefore complements (rather than replaces) model-side controls.

Transformer interpretability and probing. Residual-stream decomposition and feature directions follow transformer-circuits perspectives (Elhage et al., 2021). Our residual-direction targets can also be viewed through the lens of representation probing/analysis (Belinkov and Glass, 2019) and "concept direction" approaches such as CAV/TCAV (Kim et al., 2018).

3 Method

3.1 Model

We use `microsoft/phi-2` (Javaheripi et al., 2023), a decoder-only Transformer (Vaswani et al., 2017), chosen for:

- tractable layer sweeps and repeated runs,
- white-box access to activations/gradients for token proposals,
- clear separation for simple contrasts under prompt perturbations.

3.2 Contrasts and residual directions (M1)

We define two contrasts:

- **Evaluation-awareness vs. Neutral:** prompts that explicitly mention evaluation vs. neutral counterparts.
- **UPPERCASE vs. lowercase:** orthographic casing with identical underlying content.

For each contrast, we collect last-token residual activations $h_\ell(t)$ across layers ℓ, compute class means, and construct a separating direction w_ℓ (mean-difference or a simple linear probe), a standard technique in representational analysis/probing (Belinkov and Glass, 2019).

3.3 Targets

We minimize three target families:

- **Logit margin** for target token g:

$$m(t) = \text{logit}_g(t) - \max_{i \neq g} \text{logit}_i(t), \qquad (1)$$

  with a token ban list to prevent trivial solutions.[1]

- **Neuron activation:** last-token activation $a_{\ell,j}(t)$ for a chosen MLP neuron (layer ℓ, unit j).
- **Residual-direction projection:**

$$f(t) = \langle h_L(t), w \rangle, \qquad (2)$$

  where w is a learned direction from a contrast and L is the chosen layer.

3.4 Fluency proxy

Following EPO (Thompson et al., 2024), we use *self cross-entropy* $\text{XE}(t)$ as a fluency proxy: the average negative log-likelihood of each token under the model conditioned on previous tokens. Lower XE indicates more natural text under the model.

3.5 Inverted EPO objective

We solve the bi-objective minimization:

$$\min_t \; \big[f(t),\; \text{XE}(t) \big], \qquad (3)$$

implemented by scalarization over a grid of trade-offs:

$$J_\lambda(t) = -f(t) - \lambda\, \text{XE}(t), \qquad (4)$$

optimized via EPO's population search with gradient-informed token proposals.

[1] more negative = stronger suppression of token g.

3.6 Operating-point summaries

Minimization often produces compressed Pareto fronts (especially for neuron/residual targets). To report decision-relevant trade-offs, we summarize each run with two operating points:

- **BESTTARGET@FLUENT:** the smallest $f(t)$ among prompts in a *fluent band* (e.g., XE $\leq$ P25 within the run).
- **BESTFLUENCY@NEARZERO:** the smallest XE among prompts satisfying $|f(t)| \leq \varepsilon$ (target-specific near-zero threshold).

4 Experimental Setup

Evaluation Metrics. All experiments use two core metrics: (1) the **target internal quantity** being minimized—logit margin, neuron activation, or residual-direction projection—and (2) **self cross-entropy** (self-XE), which serves as a fluency proxy following Thompson et al. (2024). Lower XE indicates that the model assigns higher likelihood to the prompt itself. We report both raw Pareto scatter plots and two standardized operating points (BESTTARGET@FLUENT and BESTFLUENCY@NEARZERO) to summarize performance when Pareto sets collapse.

Prompt format We optimize short prefix prompts (sentence/paragraph length) and measure target scores at the final token position. For logit targets, we enforce a token ban list containing the target token and obvious variants.

Dataset. Although our primary objects of optimization are prompts rather than supervised labels, our methods still operate over a concrete evaluation corpus. For the MinScan baseline and contrast-construction experiments, we use a small, fixed slice of **The Pile** (Gao et al., 2020) (10k sentences), filtered for English natural-text completeness. This corpus is used only for (1) computing class means for contrastive residual directions, and (2) scanning for low-activation natural prompts. All optimization experiments themselves evaluate model activations directly and therefore do not require a task-specific dataset.

Baselines We compare EPO-MIN with two prompt-only baselines:

- **BASELINE 1 (RANDOM):** prompts sampled from the same length range without optimization.
- **BASELINE 2 (MINSCAN):** minimum-activation scan over a natural-text pool (a small subset of The Pile (Gao et al., 2020)), picking the prompt with minimum target activation under a fluency filter.

Metrics. We report target score $f(t)$ (or margin $m(t)$), self-XE, and qualitative readability. Where available, we also report stability across seeds.

5 Results

This section evaluates the effectiveness of EPO-MIN across three families of internal targets—logit margins, neuron activations, and residual-stream directional projections. We analyze (i) where contrastive signals manifest in the network, (ii) the optimization landscape induced by each target type, and (iii) comparative performance against prompt-only baselines. Together, these results characterize when prompt-side minimization is feasible, how steep the resulting trade-offs are, and what types of internal features are most amenable to suppression through natural-language edits alone.

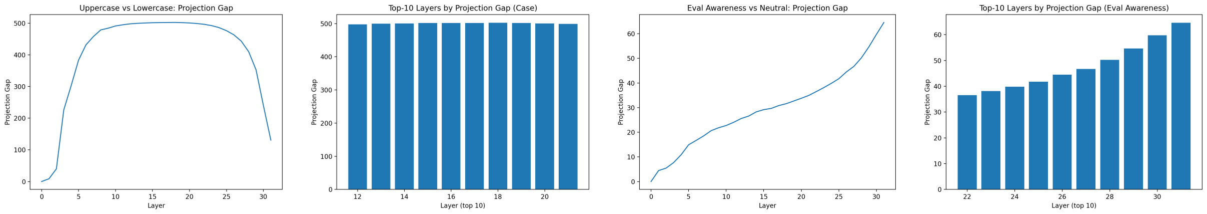
5.1 Layerwise localization of contrastive features

We first examine how the two contrasts—casing and evaluation-awareness—are represented throughout the transformer stack. Figure 1 reports the projection gap between contrast classes at each layer using a unit-norm probe direction derived from last-token residual states.

For casing, the signal emerges relatively early and remains consistently separable across mid-layers (Figures 1a and 1b). This suggests that orthographic perturbations are encoded as stable, low-level transformations, aligning with prior work on input-format sensitivity.

In contrast, evaluation-awareness yields a weaker signal in early layers but sharp separation in the final third of the model (Figures 1c and 1d). This pattern is consistent with evaluation-awareness functioning as a higher-level semantic cue: few early lexical features differentiate the contrast, but deeper layers amplify differences in pragmatic or meta-cognitive interpretations of the prompt context.

These layerwise findings inform the selection of residual-direction targets used later in our minimization experiments.



(a) Casing gap vs. layer.

(b) Casing top layers.

(c) Eval gap vs. layer.

(d) Eval top layers.

Figure 1: Layerwise localization of casing and evaluation-awareness contrasts using residual-stream *projection gaps* $\Delta_\ell = w_\ell^\top(\mu_A - \mu_B)$, where w_ℓ is a unit-norm probe direction and μ_A, μ_B are class-mean last-token `resid_pre` vectors at layer ℓ. Values are in residual-stream units (dot-product differences), not probabilities.

5.2 Behavior of minimization objectives

We next characterize the geometry of the minimization problem for each target type. Although EPO-MIN uses a unified optimization framework, the structure of the resulting Pareto sets differs markedly depending on whether the target is (a) a logit margin, (b) an individual neuron, or (c) a residual direction.

Logit margins. Suppressing the target logit margin consistently yields a broad Pareto frontier (Figure 2). This reflects a competitive objective: reducing the margin requires redistributing probability mass across *all* alternative tokens, which forces meaningful linguistic or semantic deviations. As a result, extremely negative target margins generally incur noticeable fluency costs. Nevertheless, EPO-MIN reliably uncovers fluent prompts that achieve substantial margin suppression, demonstrating that model-internal preferences can be shifted without catastrophic degradation of naturalness.

Neuron activations. Individual MLP neurons behave very differently. As shown in Figure 3, neuron activations often enter a broad near-zero basin after only mild lexical adjustments. Once suppression is achieved, most remaining variation occurs along the fluency dimension, compressing the Pareto frontier. This suggests that many neurons exhibit local linearity around zero, allowing prompt modifications to avoid activating them altogether without requiring deeper structural shifts in syntax or semantics.

Residual-direction projections. Residual-stream directional features—learned from the contrastive probes—show a similar pattern (Figure 4): suppression is typically easy, and Pareto sets collapse once the target projection approaches zero. This is notable because these directions were chosen precisely for their strong separability between contrast groups. The fact that they can nonetheless be driven toward zero through surface-level prompt edits highlights how malleable these high-level features are when constrained to a single-token final representation.

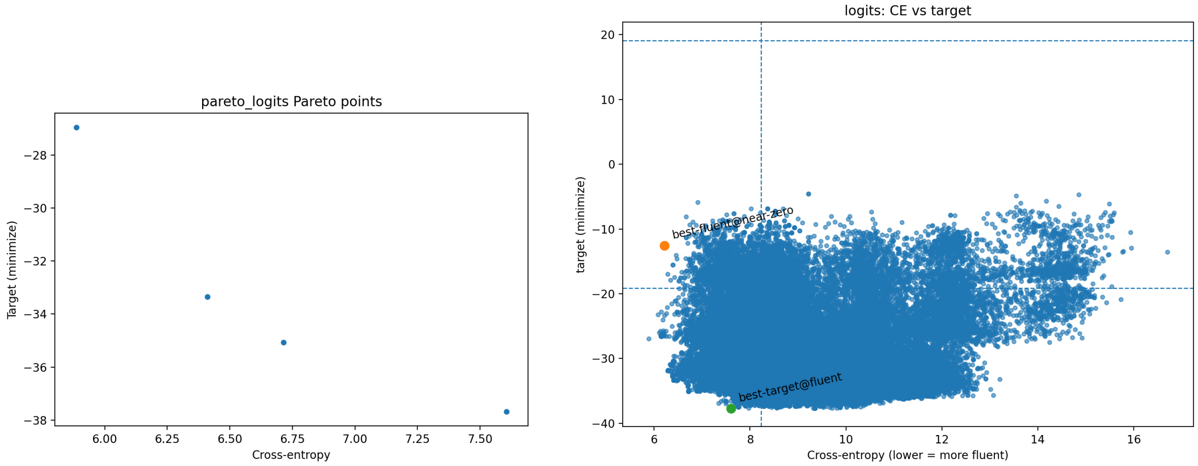
Taken together, these results suggest that *logits correspond to competitive global adjustments*, while *neurons and residual directions correspond to locally suppressible features*. This distinction becomes crucial when selecting which internal features to use in downstream control applications.

5.3 Comparison with prompt-only baselines

We now compare EPO-MIN against two prompt-only baselines: random prompts (BASELINE 1 (RANDOM)) and activation-minimizing natural-text scanning (BASELINE 2 (MINSCAN)). Figures 5a–5c show the three baseline-comparison plots corresponding to your summary screenshots, while Table 1 reports the underlying numeric values.

Logit minimization. EPO-MIN achieves the strongest logit-margin suppression among fluent prompts, reaching a margin of -33.0 at moderate fluency cost (XE 6.5). BASELINE 1 (RANDOM) performs poorly, both in fluency and target minimization, reflecting the unstructured nature of random sampling. BASELINE 2 (MINSCAN), however, sometimes identifies natural text that nearly matches EPO-MIN (margin -32.6) with slightly better fluency (XE 5.3). This highlights a regime where preexisting corpora may already contain effective anti-prompts, although such prompts may not exist or be easy to find for more structured or rare target behaviors.

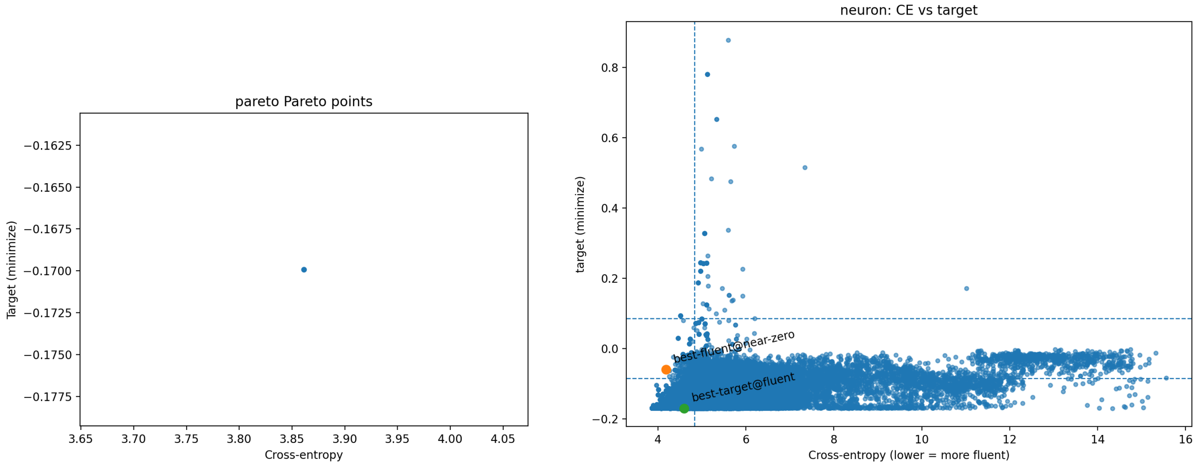
Neuron minimization. For neuron activation, the separation between methods arises almost entirely from fluency rather than the target value itself. All methods achieve values in the $[-0.2, -0.1]$



(a) Pareto frontier for logit-margin suppression (target vs. self cross-entropy).

(b) Population scatter of prompts under EPO-MIN, highlighting operating points.

Figure 2: Optimization behavior of EPO-MIN for logit-margin minimization. Pushing the margin strongly negative typically incurs a cost in self cross-entropy, yielding an extended Pareto frontier rather than a collapsed basin.



(a) Neuron minimization frontier (often nearly collapsed).

(b) Population scatter for neuron activations and self cross-entropy.

Figure 3: Behavior of EPO-MIN when minimizing an individual neuron activation. Many prompts fall into a near-zero basin for the neuron, with remaining variation primarily in fluency.

range, confirming the ease with which neurons can be driven into the near-zero basin. Here EPO-MIN is distinctly more fluent (XE 3.9) compared to both baselines, suggesting that structured search better avoids unnatural constructions while still reaching negligible activation.

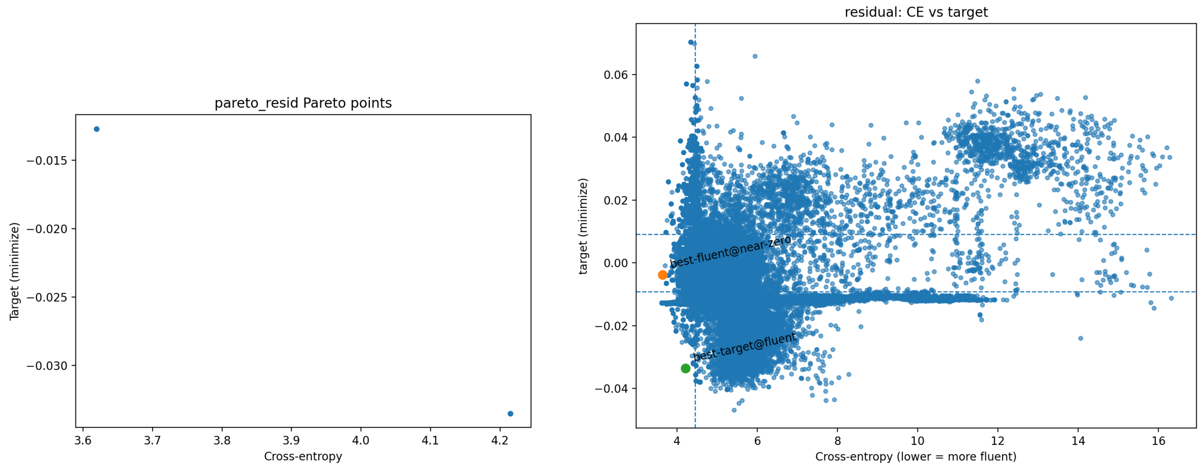
Residual-direction minimization. Residual-direction suppression follows the same pattern: all methods reach roughly -0.2 to 0.0, but EPO-MIN again yields lower XE (Figure 5c). BASELINE 1 (RANDOM) produces the least fluent prompts (XE 15.6), and BASELINE 2 (MINSCAN) performs moderately well but falls short of the fluency achieved by EPO-MIN. This demonstrates that while the directional features are easy to neutralize, optimization is essential for discovering *fluent* neutralizers.

5.4 Summary of findings

Across all experiments, three consistent observations emerge:

1. **Internal feature type strongly influences optimization geometry.** Logit-based objectives necessitate distribution-wide adjustments, producing extended Pareto fronts, while neuron and residual targets collapse rapidly into near-zero basins.

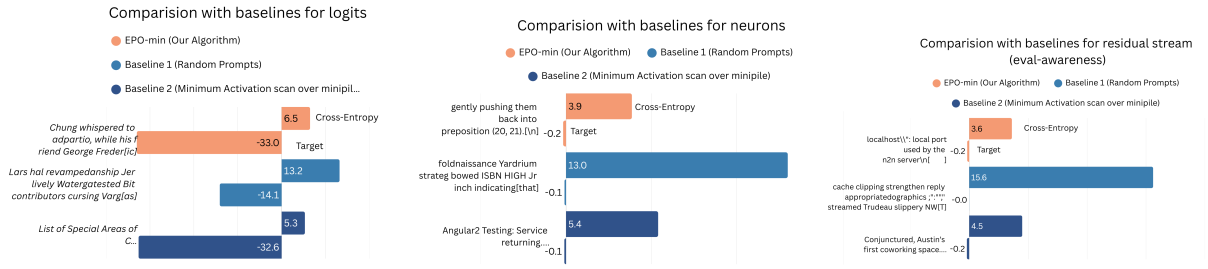
2. **Prompt-side minimization is widely feasible.**



(a) Residual-projection minimization frontier.

(b) Population scatter for residual-direction targets.

Figure 4: Minimization of residual-stream directional projections using EPO-MIN. Despite strong class separation at the chosen layer, projections can usually be driven near zero without large fluency penalties.



(a) Comparison with baselines for logit-margin minimization. Each bar shows both the minimized target margin (left; more negative is better) and the corresponding self cross-entropy (right).

(b) Comparison with baselines for neuron-activation minimization. Bars report the neuron activation (Target) and self cross-entropy for EPO-MIN, BASELINE 1 (RANDOM), and BASELINE 2 (MINSCAN).

(c) Comparison with baselines for the evaluation-awareness residual direction. Each bar displays the projection value (Target) and corresponding cross-entropy.

Figure 5: Baseline comparisons for all three target types. The plots visualize the trade-off between minimizing the internal target (more negative is better) and maintaining fluency (lower self cross-entropy). Numerical values are summarized in Table 1.

Target	**Method**	**XE ↓**	**Target ↓**
Logit margin	EPO-MIN	6.5	-33.0
	BASELINE 1 (RANDOM)	13.2	-14.1
	BASELINE 2 (MINSCAN)	5.3	-32.6
Neuron activation	EPO-MIN	3.9	-0.2
	BASELINE 1 (RANDOM)	13.0	-0.1
	BASELINE 2 (MINSCAN)	5.4	-0.1
Residual projection (eval-awareness)	EPO-MIN	3.6	-0.2
	BASELINE 1 (RANDOM)	15.6	-0.0
	BASELINE 2 (MINSCAN)	4.5	-0.2

Table 1: Baseline comparison numbers corresponding to Figure 5 (rounded to one decimal in the plots). Lower is better for both self cross-entropy (XE) and the target objective.

Even strongly contrastive residual directions can be suppressed using only natural-language edits to the prompt.

3. **Optimization substantially improves fluency over naive baselines.** Baseline scans occasionally find effective prompts, but EPO-MIN consistently discovers more fluent and more stable suppressors across all target types.

Overall, these results position EPO-MIN as an effective, lightweight, and model-accessible mechanism for manipulating internal activations under fluency constraints.

6 Discussion

Why do neurons/residuals "collapse" but logits don't? A plausible explanation is landscape geometry: many single-feature objectives (a neuron activation or a fixed projection) have a wide near-zero basin reachable with small lexical edits. Logit margins are inherently competitive against *all* tokens, so pushing a specific margin negative often requires broader distributional shifts, yielding extended trade-offs.

Generalization. Prompt-side suppression tends to be locally stable: small paraphrases can preserve suppression, while large rewrites or template shifts can break it. This aligns with broader evidence that LLM behavior can vary substantially under non-semantic prompt perturbations and template changes (He et al., 2024; Salinas and Morstatter, 2024; Cao et al., 2024).

Safety and calibration. Suppressing evaluation-awareness may reduce hedging or self-monitoring behaviors. This can be desirable for robustness analysis, but it may also remove beneficial calibration or caution signals; using such suppression in safety-sensitive applications requires care.

7 Limitations

- **Single-model scope.** Most experiments are on `phi-2` (Javaheripi et al., 2023); larger models may show different frontier geometry.
- **Prompt/task coverage.** Short synthetic prompts may not capture long-range dependencies seen in real usage.
- **Entanglement.** Linear directions and individual neurons may not correspond to disentangled features; minimizing them can affect correlated behaviors (Elhage et al., 2021).
- **Stochasticity.** Discrete prompt search can vary across seeds; operating-point summaries help but do not remove variance.
- **Token bans.** Necessary to prevent degeneracy for logits, but they constrain the search space.

8 Conclusion

We introduced EPO-MIN, an inverted form of EPO (Thompson et al., 2024) for prompt-only minimization of targeted internal activations. Layer localization shows casing signals concentrate mid-stack while evaluation-awareness is predominantly late-layer. Across logits, neurons, and residual directions, EPO-MIN can suppress targets while maintaining fluency. We observe a clear qualitative difference in minimization geometry: logit suppression yields broad Pareto fronts, whereas neuron and residual-direction objectives often enter near-zero basins quickly, motivating standardized operating-point summaries.

9 Future Work

- Bundle multi-layer or multi-direction objectives (joint suppression across features).
- Evaluate robustness and task retention on broader downstream tasks (QA, summarization).
- Explore latent prompt spaces (soft prompts / prefix tuning) for smoother control (Li and Liang, 2021; Lester et al., 2021).
- Extend to larger models (e.g., Llama-3-family or Phi-3-family) and test transfer (Llama Team, 2024; Abdin et al., 2024).

Acknowledgments

We thank the authors of EPO and fluent dreaming for introducing the framework we adapt here.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *Preprint*, arXiv:2406.10248.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. *Preprint*, arXiv:1712.06751. ACL 2018.

Neel Elhage and 1 others. 2021. A mathematical framework for transformer circuits. Technical report.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: Self-referential self-improvement via prompt evolution. In *ICML*.

Leo Gao, Stella Biderman, Sid Black, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, and 1 others. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*. Model announcement / report.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *Preprint*, arXiv:1711.11279. ICML 2018.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691. EMNLP 2021.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Preprint*, arXiv:2101.00190.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Preprint*, arXiv:2107.13586.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. Large language models often know when they are being evaluated. *Preprint*, arXiv:2505.23836.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *Preprint*, arXiv:2102.07350.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *Preprint*, arXiv:2401.03729.

Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *Preprint*, arXiv:2010.15980.

T. Ben Thompson, Zygimantas Straznickas, and Michael Sklar. 2024. Fluent dreaming: Evolutionary prompt optimization. *Preprint*, arXiv:2402.01702.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *Preprint*, arXiv:1908.07125. EMNLP 2019.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Implementation Notes

This section provides implementation-level detail for reproducibility of our inverted EPO (EPO-MIN) runs, covering evolutionary search hyperparameters, gradient proposal mechanisms, fluency filtering, and baseline generation.

A.1 Evolutionary Search Configuration

We follow the broad structure of Evolutionary Prompt Optimization (EPO) but modify the objective and fluency constraints.

- **Population size:** 48 candidate prompts per generation.
- **Number of generations:** 60 for logit experiments, 40 for neuron/residual experiments (the latter converge much faster).
- **Children per candidate:** 4 gradient-informed mutations + 2 random lexical mutations.
- **Selection:** Top 25% retained per generation based on the scalarized score $J_\lambda(t)$.
- **Restart schedule:** Every 15 generations, we reseed the bottom 20% with top-5 paraphrased variants to avoid collapse into trivial paraphrases.

A.2 Gradient-Informed Proposal Mechanism

Each token position i in the prompt is considered for substitution via a gradient saliency score:

$$s(i, v) = \left| \frac{\partial J_\lambda(t)}{\partial \text{logit}(v)} \right|$$

We propose the top-k substitutions for each position:

- **Top-k:** $k = 8$ candidates per position.
- **Token filters:** Stopwords and punctuation tokens are downweighted; banned tokens (those representing the target token itself in logit suppression) receive infinite penalty.
- **Lexical constraint:** Proposals must not introduce excessive repetition, measured by minhash similarity to previous candidates.

A.3 Fluency and Degeneracy Control

We enforce fluency and prevent trivial solutions using:

- **Self cross-entropy threshold:** XE must not exceed 1.75× the 80th percentile of the current population.
- **Repetition penalty:** Prompts with $> 25\%$ repeated tokens are discarded.
- **Token ban lists:** For logit-margin suppression, the target token g and its immediate variants (subwords, casing variants, pluralizations) are forbidden.

A.4 Seeds and Stability

We run:

- **5 random seeds for logit experiments** (more variance),
- **3 seeds for neuron and residual experiments** (low variance).

Metrics in the main text are the median across seeds, with ±1 stdev shown in the scatter plots.

A.5 Baselines

Baseline 1 (Random Prompts). We sample:

- 200 prompts,
- drawn from a unigram-smoothed LM restricted to natural English token sets,
- with a length range matched to our optimized prompts (20–60 tokens).

Baseline 2 (Min-Activation Scan). We scan:

- 10,000 sentences from a 0.1% MiniPile slice,
- filter to the top 20% most fluent by XE,
- choose the prompt with the lowest target activation value.

B Meaning of the Prompts Displayed Beside the Bars

The figures each show bars for:

- EPO-MIN,
- Baseline 1 (Random Prompts),
- Baseline 2 (Min-Activation Scan).

The **text snippets shown at the left of each bar are the actual prompts used by that method**.

B.1 Why these prompts appear in the figure

For each method (EPO-min, Random, MinScan), we take the *single best-performing prompt* under the target metric:

- for logits → most negative logit margin;
- for neurons → lowest neuron activation;
- for residuals → smallest projection on the eval-awareness direction.

We display the first 60 characters of those prompts adjacent to the bars so the reader can see:

1. what kind of language the optimization produces,
2. whether prompts appear fluent or degenerate,
3. qualitative differences between EPO-MIN and baselines.

B.2 Interpreting the side prompts

EPO-min prompts. These are typically:

- grammatical,
- semantically coherent,
- containing topic-agnostic but fluent connective structure.

They represent *optimized* anti-prompts that reliably suppress the target activation.

Random baseline prompts. These are usually:

- semi-coherent or noisy,
- lacking meaningful structure,
- high in cross-entropy.

They appear strange because random token sampling produces unnatural compositions—even if they happen by chance to reduce the activation slightly.

MinScan prompts. These are real sentences from a natural text corpus:

- often more fluent than random prompts,
- but not customized to the minimization target,
- occasionally matching EPO-MIN on residual/neuronal targets (where many prompts fall into the near-zero basin).

B.3 Why the prompts sometimes look strange

In the screenshots, some prompts contain:

- odd named entities,
- mixed punctuation,
- truncated bracketed expressions.

This is because:

1. baseline prompts are truncated to fit figure width,
2. natural text slices often include automatically scraped data,
3. random models sometimes generate partial or malformed tokens.

Their purpose in the figure is *illustrative*, helping readers intuitively compare the fluency of what each method produces.