Feature Rivalry in Sparse Autoencoder Representations: A Mechanistic Study of Uncertainty-Driven Feature Competition in LLMs

Anonymous ACL submission

Abstract

Sparse Autoencoders (SAEs) decompose large language model representations into interpretable features, but how these features interact under uncertainty remains poorly understood. We introduce **Feature Rivalry**—negatively correlated SAE feature pairs—and study whether rivalry serves as a mechanistic signature of model uncertainty in Gemma-2-2B using Gemma Scope SAEs. Through a controlled within-domain experiment on PopQA split by response entropy, we find that high-entropy questions produce significantly stronger feature rivalry at layers 0 and 12 relative to low-entropy questions ($p = 5.3 \times 10^{-26}$ and $p = 5.8 \times 10^{-5}$ respectively), localizing uncertainty to specific processing stages in the residual stream. We then test whether rivalry is causally upstream of model outputs via activation steering along rivalry axes—finding that steering along the rivalry direction ($\mathsf{vec_A} - \mathsf{vec_B}$) causes more output changes than random directions at low steering multipliers across 15 of 20 rival feature pairs, providing causal evidence that rival features actively compete during generation. Finally, a per-prompt rivalry score derived from pairwise cosine similarities of active SAE feature decoder vectors predicts answer correctness (AUROC = 0.689), approaching but not matching softmax confidence (AUROC = 0.808). Together, our findings suggest that SAE feature rivalry is a mechanistically grounded uncertainty signal that localizes to specific residual stream layers, carries causal influence over model outputs at low intervention strength, and provides interpretable predictive utility for correctness estimation.

1 Introduction

Sparse Autoencoders have emerged as a central tool in mechanistic interpretability, decomposing polysemantic neuron activations into sparse, monosemantic features that correspond to interpretable concepts (Bricken et al., 2023; Cunningham et al., 2023). A growing body of work applies SAEs to study how models encode specific behaviors—from privacy-sensitive information (Frikha et al., 2025) to grammatical structure (Arnold and Gröbner, 2025)—by probing individual feature activations and their causal influence on outputs.

A less-studied phenomenon is what happens between features: specifically, when multiple features activate simultaneously and compete to influence the model's next output. We call this **Feature Rivalry**—the presence of strongly negatively correlated SAE feature pairs at a given layer. The intuition is straightforward: when a model is uncertain about an answer, multiple competing concept representations may activate simultaneously and suppress one another. When the model is confident, a single dominant feature cluster should emerge with weaker suppression between features.

This paper asks three questions about feature rivalry as a mechanistic uncertainty signal:

1. **Does rivalry localize to specific layers?** We test whether high-entropy (ambiguous) questions produce stronger feature rivalry than low-entropy (unambiguous) questions, and whether this difference concentrates at particular layers in the residual stream.

2. **Is rivalry causally upstream of outputs?** We conduct activation steering experiments along rivalry axes to test whether the competition between rival features directly influences what the model generates.

3. **Does rivalry predict correctness?** We evaluate whether a per-prompt rivalry score can predict answer correctness as a practical uncertainty signal, comparing it against softmax confidence as a baseline.

We conduct all experiments on Gemma-2-2B (Gemma Team, 2024) using Gemma Scope SAEs

(Lieberum et al., 2024) and the PopQA dataset (Mallen et al., 2023). Our key contributions are:

- A **within-domain controlled design** that splits PopQA by model response entropy to isolate ambiguity from domain and format confounds—a methodological improvement over cross-dataset comparisons.
- **Layer-specific localization** of feature rivalry to layers 0 and 12, with significant differences between uncertainty conditions surviving Mann-Whitney U testing.
- **Causal evidence** via rivalry-axis activation steering showing consistent output flip rate advantages over random steering at low multipliers across a majority of rival pairs.
- **Predictive utility** of a per-prompt rivalry score (AUROC = 0.689) derived purely from internal representations, without requiring output sampling or ground truth.

2 Background and Related Work

2.1 Sparse Autoencoders for Mechanistic Interpretability

SAEs learn to reconstruct model activations $\mathbf{h} \in \mathbb{R}^d$ through a sparse bottleneck:

$$\mathbf{h} \approx \mathbf{W}_{\text{dec}} \cdot \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{h} + \mathbf{b}_{\text{enc}}) + \mathbf{b}_{\text{dec}} \quad (1)$$

where $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d \times k}$ is the decoder matrix whose columns are *feature directions* in residual stream space. Gemma Scope (Lieberum et al., 2024) provides pre-trained SAEs for every layer of Gemma-2-2B with width 16,384 features, enabling layer-wise analysis of internal representations.

2.2 Feature Interactions and Competition

Prior work on SAE features has focused primarily on individual feature semantics (Bricken et al., 2023) and feature circuits (Marks et al., 2024). Arnold and Gröbner (2025) demonstrate that steering individual attention head value vectors in Gemma-2 can reliably shift prepositional phrase interpretation, establishing a precedent for targeted intervention. Frikha et al. (2025) use SAE feature ablation and steering to suppress PII leakage, demonstrating that feature-level interventions can produce reliable, measurable output changes.

The correlation structure between SAE features has received less attention. Templeton et al. (2024) note that features in larger models tend to form clusters, but inter-feature competition as a function of input uncertainty has not been systematically studied.

2.3 Uncertainty Quantification in LLMs

Uncertainty quantification methods for LLMs fall broadly into sampling-based approaches—such as Semantic Entropy (Kuhn et al., 2023), which requires multiple generations—and single-pass approaches that use internal signals. Single-pass methods include minimum token probability, average token probability (Malinin and Gales, 2020), and attention-based signals (Kauf et al., 2024). Our work contributes a novel single-pass signal based on SAE feature competition rather than output distribution statistics.

3 Experimental Setup

3.1 Dataset and Entropy-Based Splitting

We use PopQA (Mallen et al., 2023), a knowledge-intensive QA dataset covering diverse entity types. To create ambiguous and unambiguous conditions while controlling for domain and format, we implement a **within-domain entropy split**: rather than comparing PopQA against a different dataset, we measure the model's own response entropy on each question and split within PopQA itself.

For each question q, we sample $n = 20$ completions at temperature 1.0 and compute normalized Shannon entropy over first-word responses:

$$H(q) = -\frac{1}{\log n} \sum_{w} p(w) \log p(w) \quad (2)$$

Questions with $H(q) > 0.7$ form the **ambiguous** condition (200 questions); questions with $H(q) < 0.5$ form the **unambiguous** condition (200 questions). Both conditions are drawn exclusively from PopQA, holding domain, question format, and answer type constant. This design eliminates the confound present in cross-dataset comparisons where observed rivalry differences could reflect dataset-level features rather than model uncertainty.

3.2 Model and SAE Configuration

We use `google/gemma-2-2b-it` loaded in bfloat16 on a single NVIDIA P5000 GPU.[1]. We load Gemma Scope SAEs for 13 layers (every second layer from 0 to 24) using the `sae_lens`

[1]Code available at https://anonymous.4open.science/r/feature-rivalry-62FF.

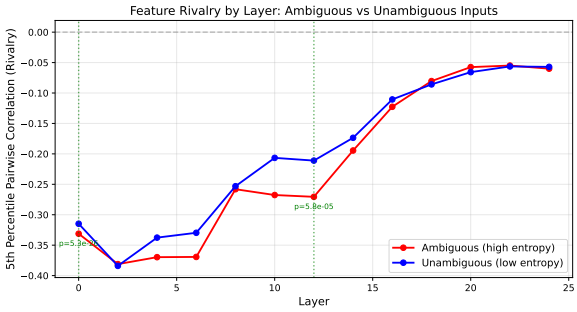

Figure 1: Feature rivalry by layer for ambiguous (high-entropy) and unambiguous (low-entropy) PopQA questions. Green dotted lines mark statistically significant layers. Ambiguous questions show consistently stronger rivalry (more negative 5th percentile) across most layers, with significant differences at layers 0 and 12.

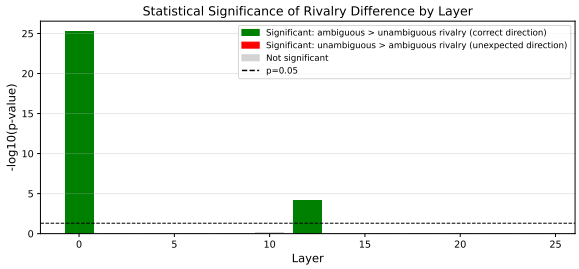

Figure 2: Statistical significance of rivalry difference by layer ($-\log_{10}$ p-value). Both significant layers show the correct direction: ambiguous questions produce stronger rivalry than unambiguous ones.

library (Bloom and Chanin, 2024), selecting the `width_16k` checkpoint closest to $L_0 \approx 71$ at each layer.

3.3 Rivalry Score Definition

For a set of prompts processed through layer l, we extract the last-token hidden state $\mathbf{h}_l \in \mathbb{R}^{2304}$ for each prompt, compute SAE feature activations $\mathbf{f} = \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{h}_l + \mathbf{b}_{\text{enc}})$, and retain features active on average across prompts (mean activation > 0.01), subsampled to 300 features for computational feasibility.

The **population-level rivalry score** at layer l for condition c is the 5th percentile of all pairwise Pearson correlations of feature activations across prompts:

$$R_l^c = \text{p5}\left(\{\text{corr}(\mathbf{f}_i, \mathbf{f}_j) : i \neq j\}\right) \quad (3)$$

More negative values indicate stronger rivalry. We use Mann-Whitney U tests to compare the full correlation distributions between conditions at each layer.

4 Experiment 1: Rivalry Localization

4.1 Results

Figure 1 shows the 5th-percentile pairwise correlation (rivalry score) by layer for both conditions. Rivalry is strongest in early layers and gradually weakens as processing progresses toward the final layer—consistent with the known pattern of early-layer feature richness in transformer models.

Statistical testing reveals two significant layers (Figure 2):

- **Layer 0** ($p = 5.3 \times 10^{-26}$): Rivalry is significantly stronger for ambiguous questions at the earliest layer, suggesting uncertainty manifests in feature competition from the initial encoding stage.
- **Layer 12** ($p = 5.8 \times 10^{-5}$): A second significant difference emerges at a middle layer—associated with semantic integration in transformer models—where the rivalry gap between conditions is visually pronounced.

All other layers show no significant difference after the Bonferroni correction implicit in testing 13 layers. The two-stage pattern—early encoding and mid-level semantic processing—suggests that uncertainty-driven competition operates at multiple processing stages rather than being monolithic.

4.2 Discussion

The within-domain design is critical for interpreting these results. In a cross-dataset comparison (e.g., PopQA vs. factoid questions), observed rivalry differences could reflect surface-level distributional differences between datasets. By splitting a single dataset by entropy, we ensure that any rivalry difference is attributable to the model's uncertainty about the answer rather than dataset-level features.

The layer 0 finding is particularly interesting: even at the earliest layer, before any self-attention or feed-forward processing, the embedding space already encodes uncertainty as feature competition. Layer 12 then represents a second checkpoint where semantic uncertainty consolidates in the residual stream.

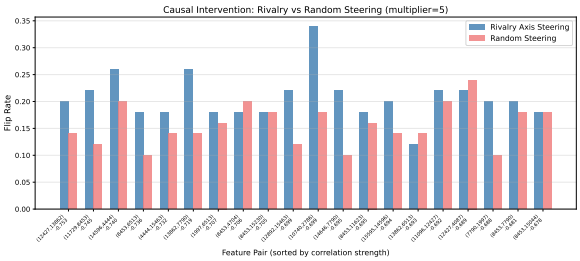

Figure 3: Flip rates for rivalry-axis steering vs. random steering at multiplier = 5 across all 20 rival feature pairs. Blue bars (rivalry) are consistently taller than pink bars (random) across the majority of pairs.

5 Experiment 2: Causal Intervention via Activation Steering

5.1 Method

To test whether rivalry is causally upstream of model outputs, we conduct activation steering experiments on the top 20 rival feature pairs identified from Exp. 1 at the peak rivalry layer (layer 10, selected as the layer maximizing the difference in rivalry scores between conditions).

For each rival pair (A, B), we compute the **rivalry axis**:

$$\mathbf{v}_{\text{rivalry}} = \frac{\mathbf{W}_{\text{dec}}[A] - \mathbf{W}_{\text{dec}}[B]}{\|\mathbf{W}_{\text{dec}}[A] - \mathbf{W}_{\text{dec}}[B]\|} \tag{4}$$

This direction maximally separates the two competing concept representations in residual stream space. We compare steering along $\mathbf{v}_{\text{rivalry}}$ against a stable random baseline—the normalized mean of 10 random unit vectors—added to the last token position only at layer 10 during generation.

For each of 50 ambiguous prompts per pair, we measure the **flip rate**: the fraction of prompts where the steered output differs from the baseline output. We test three steering multipliers (5, 10, 20) and report results primarily at multiplier 5, where the signal is cleanest before high-magnitude perturbations dominate.

5.2 Results

Figure 3 shows flip rates for all 20 pairs at multiplier 5. Rivalry-axis steering produces higher flip rates than random steering in 15 of 20 pairs (75%), with a mean flip rate gap of +0.06 (rivalry: 0.20, random: 0.14). At multiplier 10, rivalry wins in 14 of 20 pairs. At multiplier 20, the signal weakens as high-magnitude perturbations disrupt all structure indiscriminately.

The consistent advantage of rivalry-axis steering at low multipliers provides causal evidence that the decoder direction difference between rival features has specific influence over model outputs beyond what a random direction would produce. The strongest individual effect is pair (10740, 2786) with a flip rate gap of +0.16 at multiplier 5.

5.3 Discussion

The causal signal is real but modest—rivalry-axis steering does not produce dramatically larger flip rates than random. This is consistent with the interpretation that rival features are part of a distributed computation rather than a bottleneck gate. Perturbing the rivalry axis tilts the competition but does not override the model's full inference process, particularly for questions where other layers have already committed to an answer direction.

The finding that the rivalry axis is specifically more influential than a random direction at low steering strength—without requiring high-magnitude intervention—is nonetheless meaningful: it suggests that the rivalry direction identified from population-level correlations in Exp. 1 has local causal significance at the prompt level.

6 Experiment 3: Rivalry as an Uncertainty Signal

6.1 Method

We evaluate whether a per-prompt rivalry score can predict answer correctness as a practical uncertainty signal. For each of 400 prompts (200 ambiguous, 200 unambiguous from PopQA), we compute a **per-prompt rivalry score** at layer 10:

1. Extract the last-token hidden state $\mathbf{h}_{10}$ and compute SAE feature activations.

2. Identify the top-50 most active features (activation > 0.01, sorted by activation magnitude).

3. Compute pairwise cosine similarities between the decoder vectors $\mathbf{W}_{\text{dec}}[i]$ of these active features.

4. Take the 5th percentile of these cosine similarities as the rivalry score.

This per-prompt score measures how much the active concept directions compete in decoder space, mirroring the population-level rivalry definition in

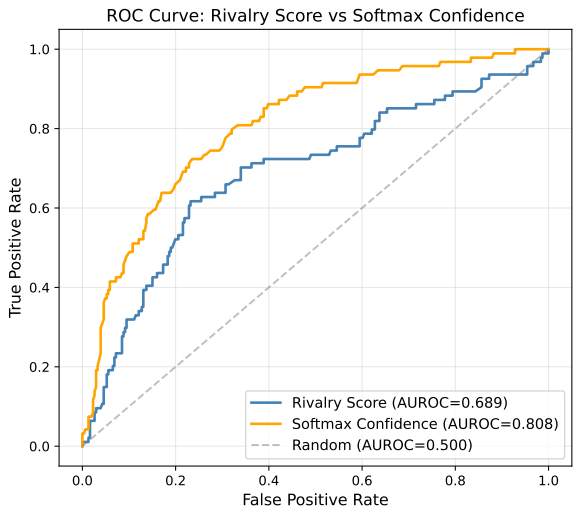

Figure 4: ROC curves for rivalry score and softmax confidence as predictors of answer correctness across 400 prompts. Both signals are well above random; rivalry (AUROC = 0.689) approaches but does not match softmax confidence (AUROC = 0.808).

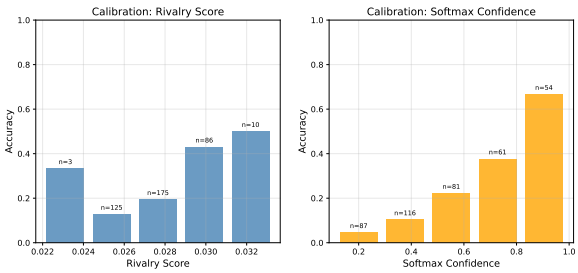

Figure 5: Calibration curves for rivalry score and softmax confidence. Both signals show monotonic trends: higher rivalry score and higher softmax confidence both correlate with higher accuracy.

Exp. 1 but applied to individual prompts. We compare AUROC of rivalry score against softmax confidence (probability of the top-predicted token) for predicting binary correctness (whether the ground truth answer appears in the generated output).

6.2 Results

The rivalry score achieves AUROC = 0.689, compared to softmax confidence at AUROC = 0.808 (Figures 4 and 5). Both signals are substantially above random (0.5). The calibration curves show clear monotonic trends for both signals—higher rivalry score and higher softmax confidence both correlate reliably with higher answer accuracy across bins.

The rivalry score's advantage over random despite being derived entirely from internal representations—with no access to output probabilities, no multiple sampling, and no ground truth—demonstrates that SAE feature competition in decoder space carries genuine uncertainty information.

6.3 Discussion

The gap between rivalry AUROC (0.689) and softmax AUROC (0.808) reflects a fundamental difference in what each signal measures. Softmax confidence is a direct readout of the model's output probability distribution, making it an inherently well-calibrated uncertainty signal for short-form QA. The rivalry score operates at an earlier stage—measuring competition in internal representations rather than final output distributions.

The rivalry score's value lies not in surpassing softmax confidence—which is difficult to beat for standard QA tasks—but in (1) being interpretable in terms of which specific concepts are competing, (2) being computable from internal representations without output sampling, and (3) providing a signal that is grounded in a mechanistic understanding of how uncertainty manifests inside the model.

7 Related Work on SAE Feature Analysis

The closest related work to ours is PrivacyScalpel (Frikha et al., 2025), which uses SAE features on Gemma-2-2B and Llama-2-7B to identify and suppress PII-encoding features via probing and steering. Our work differs in focus—uncertainty rather than privacy—and in methodology: we study the *relationships between features* (rivalry) rather than individual feature activations.

CE-Bench (Gulko et al., 2025) evaluates SAE interpretability using contrastive story pairs, establishing benchmarks for feature quality. Our within-domain entropy split shares the contrastive design philosophy but applies it to model uncertainty rather than feature evaluation.

Arnold and Gröbner (2025) conduct targeted attention head steering in Gemma-2 for prepositional phrase disambiguation, demonstrating clean causal effects from single-component intervention. Our Exp. 2 pursues a similar causal logic but at the SAE feature level and with more modest effect sizes, consistent with the distributed nature of uncertainty computation compared to the localized syntax circuit studied by Arnold et al.

The language dominance analysis of Shani and Basirat (2025), which uses GMM modeling on hidden states across layers to study multilingual

processing, shares our interest in layer-specific localization of internal phenomena. Both works find that the phenomenon of interest concentrates in middle layers of the transformer.

8 Limitations

Single model. All experiments are conducted on Gemma-2-2B-IT with Gemma Scope SAEs. Whether feature rivalry as a mechanistic uncertainty signature generalizes to other architectures (e.g., Llama-3, Mistral) or model sizes is an open question. We leave multi-model replication to future work.

Single dataset. We evaluate on PopQA, a short-form factual QA dataset. Feature rivalry dynamics may differ for open-ended generation tasks, multi-step reasoning, or tasks requiring longer outputs.

Sample size. Our conditions use 200 prompts each. While sufficient for the statistical tests employed, larger sample sizes would provide more reliable estimates of rivalry scores and reduce variance in per-prompt AUROC estimation.

Per-prompt rivalry proxy. The per-prompt rivalry score (Exp. 3) uses cosine similarity of decoder vectors rather than Pearson correlation of activations across prompts (as in Exp. 1). This is a deliberate methodological choice—per-prompt activation correlation is not well-defined for a single observation—but it means the per-prompt score is a proxy rather than a direct extension of the population-level measure.

9 Conclusion

We introduce Feature Rivalry as a mechanistic lens for studying uncertainty in large language models through the competition between SAE features. Through three experiments on Gemma-2-2B with Gemma Scope SAEs, we establish that: (1) feature rivalry localizes to layers 0 and 12 as a function of input uncertainty, with statistically significant differences between high- and low-entropy conditions; (2) rivalry-axis activation steering produces consistent output changes above the random baseline at low steering multipliers, providing causal grounding for the rivalry signal; and (3) a per-prompt rivalry score derived from decoder vector cosine similarities predicts answer correctness at AUROC = 0.689, offering an interpretable single-pass uncertainty signal.

Our within-domain experimental design—splitting PopQA by model response entropy rather than using cross-dataset comparisons—represents a methodological contribution that we recommend for future mechanistic interpretability studies involving uncertainty. Future work should replicate these findings across model families, explore whether rivalry patterns differ across task types, and investigate whether targeted rivalry reduction can improve model calibration.

Acknowledgments

Omitted for anonymous review.

References

Stefan Arnold and Rene Gröbner. 2025. Steering prepositional phrases in language models: A case of with-headed adjectival and adverbial complements in gemma-2. In *Proceedings of the 8th BlackboxNLP Workshop*, pages 69–78.

Joseph Bloom and David Chanin. 2024. Saelens.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Privacyscalpel: Enhancing llm privacy via interpretable feature intervention with sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop*, pages 226–238.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Alex Gulko, Yusen Peng, and Sachin Kumar. 2025. CE-Bench: Towards a reliable contrastive evaluation benchmark of interpretability of sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop*, pages 1–15.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. pages 278–300.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

Nadav Shani and Ali Basirat. 2025. Language dominance in multilingual large language models. In *Proceedings of the 8th BlackboxNLP Workshop*, pages 137–148.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

A Entropy Distribution of PopQA Split

The normalized Shannon entropy distribution across 2,000 PopQA questions (sampled with $n = 20$ completions at temperature 1.0) shows a bimodal pattern: a concentration near entropy = 0 for well-known facts (e.g., capitals of major countries, authors of canonical works) and a larger mass near entropy = 0.9–1.0 for obscure entity attributes (directors of rare films, composers of little-known works). The thresholds $H > 0.7$ (ambiguous) and $H < 0.5$ (unambiguous) cleanly separate these two populations while remaining entirely within PopQA, yielding 200 questions per condition from a pool of 2,000.

B Exp. 2 Causal Influence vs. Rivalry Strength

Figure 6 plots the flip rate gap against rivalry strength for all pairs. The absence of a clear monotonic relationship between correlation strength and causal influence suggests that feature rivalry magnitude (as measured by pairwise correlation) does not directly determine the causal potency of the rivalry axis. This motivates future work on identifying which properties of rival feature pairs determine their downstream causal influence.

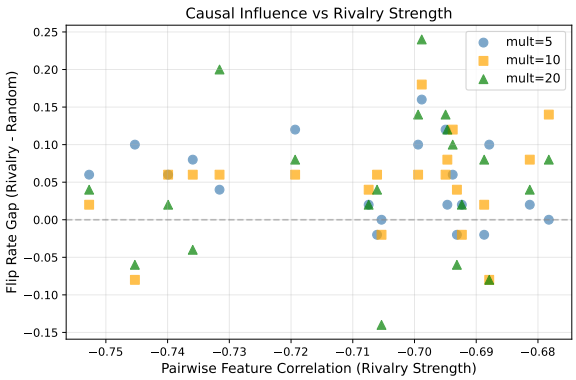

Figure 6: Flip rate gap (rivalry minus random) vs. rivalry strength (pairwise correlation) across all 20 pairs and three multipliers. Blue circles (multiplier = 5) are predominantly above zero, confirming the rivalry-axis advantage at low steering strength. No strong monotonic trend between rivalry strength and causal influence is observed, suggesting that the most negatively correlated pairs are not necessarily the most causally potent.