Contents

The Cosine–Causal Gap in Cross-Model Crosscoders

When Decoder Universality Overstates Causal Equivalence in Gemma-2-2B

Workshop draft for NeurIPS 2026 Mechanistic Interpretability Workshop. Apache-2.0. Reproducible on a single A100 in ∼5 hours of compute.

Abstract

Cross-model crosscoders are increasingly used to identify "shared" features between two models — typically a base and a fine-tuned variant — by training a single autoencoder over both residual streams and matching latents by decoder similarity. The standard universality test compares decoder cosine across the two model heads: features with high cosine and balanced norm are reported as "shared", and the analysis pipeline thereafter treats them as functionally equivalent representations. We argue that this conflates *representational* universality with *causal* universality, and we provide the first per-feature empirical measurement of the gap. We train a paper-grade BatchTopK crosscoder (73,728 latents, k=100, expansion 32) at layer 13 of Gemma-2-2B base/IT on 100M tokens of FineWeb-Edu and UltraChat-200k. For 80 high-firing shared features, we ablate each feature in both models on 256 probe inputs and measure the Pearson correlation between the two KL-divergence trajectories across probes — a per-feature causal-equivalence (CE) score. We find median decoder cosine 0.965 alongside median Pearson_CE 0.616, with **38.24% of shared features having cosine > 0.7 yet CE < 0.5**. Multiple outliers — including features with anti-aligned decoders (cosine ≈ -0.5) but identical

causal effect (CE $\approx +0.91$), and features with aligned decoders (cosine $\approx +0.95$) but opposite causal effect (CE ≈ -0.9) — show that decoder cosine is neither necessary nor sufficient for causal equivalence. We propose Pearson_CE as a mandatory complementary diagnostic for any crosscoder universality claim, release all artifacts under Apache-2.0, and discuss two future directions (optimal-ablation correction; per-layer cross-architecture replication).

Keywords: crosscoders, sparse autoencoders, mechanistic interpretability, universality, causal equivalence, Gemma-2

---

1. Introduction

Crosscoders (Lindsey et al., 2024; Anthropic, 2025) extend sparse autoencoders to operate jointly across multiple models' residual streams. Each latent in the crosscoder has one decoder vector per model, and the standard analysis pipeline classifies latents as "shared" (similar magnitude in both decoders), "base-only" or "chat-only" (large in one, small in the other), or "dead" (low activation), via the so-called **Δ_norm taxonomy**.

The claim attached to a "shared" latent is implicit but consistent across the literature: that the latent represents the *same feature* in both models, and ablating it should produce the *same downstream effect*. This claim is operational — crosscoder analyses use shared latents to pinpoint where models diverge by exclusion, and to validate that downstream interventions on a shared feature will transfer between checkpoints.

The standard test for this claim is decoder cosine similarity. When two decoders point in nearly the same direction (cosine ≈ 1), the latent is treated as universal; when they point in different directions, it is treated as model-specific. To our knowledge, **no published crosscoder paper measures whether ablating a "shared" latent in model A produces causally similar downstream effects to ablating it in model B**. The standard test is purely representational.

We measure that gap directly.

We train a paper-grade BatchTopK crosscoder on Gemma-2-2B base/IT at layer 13. For 80 prominent shared latents (firing $\geq$ 1% of validation tokens), we ablate each latent in both models on 256 probe inputs and compute the **Pearson correlation across probes between the two ablation-induced KL trajectories**. We call this per-feature score Pearson_CE.

The headline finding: median decoder cosine 0.965, median Pearson_CE 0.616. Of the 80 shared latents tested, **38.24% have cosine > 0.7 yet CE < 0.5** — representational match without causal match. Several features with anti-aligned decoders are causally equivalent. Several features with aligned decoders have *opposite* causal effects.

This paper contributes:

1. **Methodological**: Pearson_CE as the first per-feature, two-model causal pairing test for crosscoder latents. Compatible with any pre-trained crosscoder; $\sim$60 minutes of A100 compute for 80 features $\times$ 256 probes.
2. **Empirical**: Direct measurement of the cosine–causal gap on a paper-grade Gemma-2-2B crosscoder. Median cosine 0.965 vs median Pearson_CE 0.616 with tail behavior in both directions.
3. **Disclosure**: Our crosscoder uses BatchTopK + L1, which is known to over-constrain Δ_norm exclusivity (only 0.01% chat-only vs Minder et al. 2025's 4.3% with vanilla L1 on the same model pair). The cosine–causal gap result is robust to this choice — it is measured *within* the shared class, which dominates the population (53.7%) regardless of the exclusivity recipe.
4. **Reproducibility**: All artifacts public on HuggingFace under Apache-2.0, with a single notebook reproducer.

---

2. Related Work

Lindsey et al. (2024) introduced cross-model crosscoders as a mechanism for cross-checkpoint feature identification, with the **Δ_norm taxonomy** as the primary classification tool. Their analyses use shared features for cross-model comparisons but do not measure cross-model causal equivalence per-feature.

Anthropic (Jan 2025) updated the methodology with refined shared/exclusive analyses on larger crosscoders, but the universality verification remains representational (decoder cosine + balanced norm). No per-feature cross-model intervention pairing.

Minder et al. (2025) introduced **Latent Scaling** diagnostics and a KL-bridging analysis as a population-level indicator of cross-model behavior. This is upstream of our test in two ways: it measures cumulative KL between models with and without a feature population active, not per-feature correlation. We view their tools as complementary; our Pearson_CE test complements their per-feature universality estimates with a direct two-model causal pairing.

Bhatt et al. (2026) introduced **Dedicated Feature Crosscoders (DFC)** which improve the geometry of shared-feature decoders. Their analysis is also purely representational (cosine geometry) and does not measure cross-model causal equivalence per-feature.

Universality at the SAE level (Templeton et al., 2024; Marks et al., 2024) addresses universality across different SAEs trained on the same model, not across models on the same crosscoder.

We position this paper as the missing per-feature causal pairing for the crosscoder universality claim.

---

3. Methods

3.1 Crosscoder training

We train a BatchTopK crosscoder over residual-stream activations at layer 13 of `google/gemma-2-2b` (base) and `google/gemma-2-2b-it` (chat instruction-tuned), following the recipe in Anthropic's Jan-2025 update with refinements from Minder et al. (2025).

Hyperparameter	Value
Latents (`d_sae`)	73,728 (32× expansion over `d_model = 2304`)
Sparsity	BatchTopK with k=100, k_warmup_init=1000, k_warmup_steps=5000
L1	$\lambda = 4.1 \times 10^{-2}$ (Minder Section K)
Decoder init norm	1.0
Tokens	100M (FineWeb-Edu 50% + UltraChat-200k 50%, BOS dropped)
Sequence length	512, batch sizes (fwd 4 / cc 4096)
Optimizer	Adam, LR = 1×10^{-4}, 1000 warmup steps, 20% LR decay frac
Per-model norm scaling	Pre-fit (norm A = 0.268, norm B = 0.236)
Compute	A100 40GB, 4h28min

Validation metrics: Variance Explained $\text{VE}_A = 0.877$, $\text{VE}_B = 0.867$, mean L0 = 100.5, dead fraction 42.9%.

3.2 Δ_norm taxonomy

Following Lindsey et al. (2024), we classify each latent by the relative norm of its two decoder vectors:

Class	Count	Proportion
shared	39,711	53.85%
dead	31,625	42.89%
unclassified	2,385	3.23%
base_only	4	0.005%
chat_only	3	0.004%

The under-representation of the exclusive classes (base/chat-only) under BatchTopK is a known artifact (Minder et al. 2025 obtained ~4.3% chat-only with vanilla L1 on the same model pair). Our Pearson_CE measurement is performed within the *shared* class, which dominates regardless of the exclusivity recipe, so this artifact does not affect our central finding.

3.3 Pearson causal-equivalence (Pearson_CE)

For a candidate shared feature f with decoder vectors d_f^A, d_f^B for models A and B, and 256 probe inputs $\{x_i\}_{i=1}^{256}$ (FineWeb-Edu samples, last token, max_length=128):

1. For each probe x_i, run a baseline forward pass in each model and capture the next-token logit distribution $p_A(x_i), p_B(x_i)$.
2. Run an ablated forward pass where the feature f is zeroed during the crosscoder reconstruction at layer 13 in each model. Capture $p_A^{-f}(x_i), p_B^{-f}(x_i)$.
3. Compute the per-probe KL divergence: $\mathrm{KL}_A(i) = D_{\mathrm{KL}}(p_A^{-f}(x_i) \| p_A(x_i))$, similarly $\mathrm{KL}_B(i)$.
4. Compute the Pearson correlation across probes: $\text{Pearson\_CE}(f) = \rho(\{\mathrm{KL}_A(i)\}_{i=1}^{256}, \{\mathrm{KL}_B(i)\}_{i=1}^{256})$.

Filters: pre-select features that fire on $\geq$ 1% of validation tokens (gives 80 candidates from 39,711 shared); post-filter features with `n_probes_fired $\geq$ 3` out of 256 (gives 68 effective).

Pearson_CE = +1 means ablation of f in A and in B has identically correlated downstream effects across probes. Pearson_CE = 0 means no correlation. Pearson_CE = -1 means the ablation effects are anti-correlated.

3.4 Control features (unclassified class)

We apply the same protocol to 28 features from the **unclassified** class (non-trivial activation but failing the Δ_norm shared/exclusive thresholds). This class serves as a within-crosscoder control for what Pearson_CE looks like when the standard taxonomy does not call them "shared".

3.5 Compute and infrastructure

A100 40GB on Colab Pro+. Crosscoder training: 4h28min. Causal validation pass for 80 shared + 28 unclassified features $\times$ 256 probes $\times$ 2 models: ~60 minutes. Total reproduction: ~5.5 hours from cold start.

---

4. Results

4.1 Aggregate cosine–causal gap

For the 68 shared features (80 candidates, 12 dropped by `n_probes_fired < 3`):

Metric	Shared (n=68)	Unclassified control (n=28)
Median decoder cosine	**0.965**	(varies, lower)

Metric	Shared (n=68)	Unclassified control (n=28)
Median Pearson_CE	**0.616**	0.358
Mean Pearson_CE	0.480	0.263
Min Pearson_CE	-0.667	-0.898
Max Pearson_CE	0.9997	0.999
% with cosine > 0.7 AND CE < 0.5	**38.24%**	—

Decoders for the typical shared feature are nearly co-linear (median cosine 0.965 implies ∼15° between them). However, the median causal-equivalence across the same features is 0.616 — substantial but far from the perfect correlation a "universal feature" claim would predict.

4.2 The 38.24% population: cosine ≫ CE

Of 68 shared features:

- **26 features (38.24%)** have decoder cosine > 0.7 AND Pearson_CE < 0.5.
- **5 features (7.4%)** have cosine > 0.9 AND CE < 0.3 — high-cosine features whose causal effects are nearly uncorrelated across the two models.
- **2 features** have cosine > 0.95 AND CE < 0 — high-cosine features whose causal effects across models are *anti-correlated*.

The standard universality test calls these "shared". The causal-equivalence test rejects the universality claim for nearly 40% of them.

4.3 The other tail: cosine ≪ CE

- **3 shared features** have cosine < 0.0 (anti-aligned decoders) but Pearson_CE > 0.7. The most striking: cosine ≈ -0.50 with CE $\approx +0.91$. The decoder geometry says these features point opposite directions, but ablating either produces near-identical downstream effects in both models.

This is the converse failure mode: decoder cosine *understates* causal universality. Some features are causally equivalent across A and B in spite of having anti-aligned decoder geometry.

4.4 Control features

Of 28 unclassified-class features:

- Median Pearson_CE 0.358 (lower than shared's 0.616, as expected).
- **Maximum Pearson_CE in the unclassified set: 0.999** — at least one feature classified as "not shared" by Δ_norm has near-perfect cross-model causal equivalence.
- Several unclassified features at cosine ≈ 0.95 with CE ≈ -0.9 — aligned decoders with opposite causal effect.

The control population confirms that the cosine–causal gap is a property of the decoder geometry, not an artifact of the shared/non-shared boundary.

4.5 Summary table — four outlier classes

Class	Cosine	Pearson_CE	Interpretation
Standard shared	$\approx +1$	$\approx +1$	decoders + causal both align
High cosine, low CE (38%)	> 0.7	< 0.5	decoders align, causal does not

Class	Cosine	Pearson_CE	Interpretation
Anti-aligned, equivalent	≈ -0.5	$\approx +0.9$	decoders oppose, causal aligns
Aligned, opposite causal	> 0.95	< 0	decoders align, causal opposes

All four classes exist in our crosscoder. The first matches the standard universality narrative; the other three are failure modes of cosine-based universality testing.

---

5. Discussion

5.1 Decoder cosine is neither necessary nor sufficient for causal equivalence

The standard test treats decoder cosine as a proxy for causal universality. Our results show this proxy fails in both directions:

- **High cosine, low CE** (38% of our shared features): the proxy gives a false positive. The decoders point the same way, but ablation produces different downstream effects.
- **Anti-aligned, high CE**: the proxy gives a false negative. The decoders point opposite directions, but ablation has identical effect.

For any operational claim that depends on cross-model functional equivalence of a feature — for instance, "we identified the persona-vector direction in the chat model by matching to the base" — neither cosine direction is load-bearing. Pearson_CE is.

5.2 Why per-feature pairing matters

Population-level diagnostics (KL bridging, latent scaling) confirm that shared features as a *set* contribute to cross-model output similarity. They do not adjudicate which specific features are universal. Cross-model probing, steering, or feature surgery requires per-feature evidence. The standard practice — extract a feature, identify its high-cosine cross-model match, operate on the matched pair — passes the population-level test but, as we show, fails 38% of the time at the per-feature level.

This failure mode mirrors a broader pattern in alignment evaluation. Anthropic Alignment (2026) document that training on the evaluation distribution can reduce a measured metric while leaving held-out automated auditing metrics unchanged. Decoder cosine is the in-distribution proxy here — measured on the same activations the crosscoder was trained to reconstruct. Pearson_CE is the held-out automated audit — measured under ablation on a probe set the crosscoder never saw and on a downstream output (KL between models) the crosscoder is not trained to align. The 38% gap is the size of the overfitting. OpenAI Alignment (2026), auditing accidental CoT-text grading in RL training, document the same pattern in a different substrate: a training-time signal (CoT-aware reward) reports stability while held-out auditors (CoT-blind detectors) reveal degraded recall on specific distributions. Our cosine-CE gap is the crosscoder-universality instance of the same in-distribution-vs-held-out-audit pattern these two alignment-team findings document at the training level.

5.3 Compatible with existing pipelines

Pearson_CE is a *complementary* diagnostic, not a replacement. It can be applied to any pre-trained crosscoder without retraining. Its compute cost is dominated by the $2 \times N \times 256$ forward passes ($\sim$60 minutes on an A100 for 80 features). The single change we recommend for crosscoder universality analyses: **before publishing a "shared feature X transfers between models" claim, report Pearson_CE for that feature with at least 256 probes**.

5.4 Connection to Anthropic Persona Vectors and feature-level steering

Persona-vector steering (Anthropic, 2025) and related interpretability applications routinely identify a feature in one model and apply steering-equivalent operations on the matched feature in another. If 38% of "matched" features have substantially different causal effects across models, the steering effect generalizes correspondingly less. We do not re-run any persona-vector experiment in this paper, but we note that any cross-model steering claim should be cosine + CE-validated.

5.5 Why this measurement was tractable

Two infrastructure pieces made the measurement tractable in 60 minutes of GPU compute: (1) a paper-grade crosscoder with VE > 0.85 in both models (many available open-source crosscoders sit at VE 0.5–0.7, where causal ablation noise overwhelms signal); (2) a sufficiently large probe set (256 inputs, FineWeb-Edu) to estimate Pearson at the per-feature level. Smaller probe sets (∼64) yield noisier per-feature CE estimates and would need bootstrapping — we leave this to follow-up.

---

6. Limitations

- **Pre-filter bias**: We test 80 of 39,711 shared features (0.2%) — those firing on $\geq$ 1% of validation tokens. The vast tail of low-firing shared features remains untested. The cosine–causal gap may be larger or smaller in the tail.
- **Effective n=68 of 80 candidates**: 12 features did not pass the `n_probes_fired $\geq$ 3` threshold even at the prominent class. Pearson at n=3 is high-variance; we filter to keep estimates stable.
- **Zero ablation is OOD** (Li & Janson, 2024). The KL trajectories from zero ablation may differ from optimal-ablation trajectories. A natural follow-up is to repeat the measurement with optimal-ablation (the Anthropic Apr-2024 method) and check whether the cosine–causal gap shrinks.
- **Single layer, single model pair**. We measure at L13 of Gemma-2-2B base/IT only. Generalization across layers and across model architectures (Llama, Qwen, Gemma-2-9B) is future work. We have a Qwen-3.5-1.7B crosscoder configured but not yet trained at paper-grade scale.
- **BatchTopK collapsed exclusivity taxonomy**: our chat-only/base-only classes are essentially empty (0.01%). The vanilla-L1 recipe (Minder et al. 2025) recovers richer exclusive classes. Re-running our Pearson_CE pipeline on a vanilla-L1 crosscoder would test whether the cosine–causal gap is consistent across crosscoder recipes.
- **Single architecture choice (BatchTopK + JumpReLU inference)**. SAE architecture choices propagate non-trivially to feature semantics; the finding may be sensitive to this.

---

7. Conclusion

We measured per-feature causal equivalence for 80 high-firing shared features in a paper-grade Gemma-2-2B base/IT BatchTopK crosscoder, the first such measurement we are aware of in the crosscoder literature. We found a substantial cosine–causal gap: 38.24% of "shared" features (cosine > 0.7) fail the causal-equivalence test (Pearson_CE < 0.5). In both extreme tails, decoder cosine misclassifies the causal relationship: anti-aligned decoders can be causally equivalent, and aligned decoders can be causally opposite.

Pearson_CE is a cheap (∼60 minutes A100 per crosscoder) diagnostic that crosscoder analyses should report alongside cosine. We open-source the crosscoder, the validation pipeline, and the result tables under Apache-2.0 and invite replication on other crosscoders — particularly across additional layers, additional model pairs, and vanilla-L1 recipes.

Reproducibility Statement

All artifacts are public under Apache-2.0:

Component	Location
Crosscoder model	HuggingFace `caiovicentino1/gemma2-2b-crosscoder-model-diff-papergrade`
Training notebook	`OpenInterpretability/notebooks/17b_crosscoder_model_diff_`
Causal validation data (96 features × cosine + Pearson_CE)	`data/gemma_causal_validation.csv` (this repo)
Crosscoder config	`data/gemma_cfg.json`
openinterp SDK (`pip install openinterp`)	provides `safe_load_qwen36_lora` and related utilities

Reproduction time on a single A100 40GB: - Crosscoder training: 4h28min (or download from HF, ∼3 minutes) - Causal validation pass (80 shared + 28 unclassified features × 256 probes): ∼60 minutes - Total cold start: ∼5.5 hours; with cached crosscoder: ∼70 minutes.

Acknowledgments

We thank the Gemma team (Google DeepMind) for the open-weights base/IT pair, the Anthropic interpretability team for the crosscoder methodology and ongoing public-research updates, and the Minder et al. (2025) team for the training-recipe diagnostics that anchored our hyperparameter choices.

Compute was provided by Google Colab Pro+ (single A100 40GB). All training and validation runs are reproducible on consumer-tier cloud GPU within ∼6 hours of wall-clock time.

References

Anthropic. (2024). *Crosscoders*. Transformer Circuits. https://transformer-circuits.pub/2024/crosscoders/index.html

Anthropic. (2025). *Crosscoder diffing update*. Transformer Circuits. https://transformer-circuits.pub/2025/crosscoder-diffing-update/index.html

Anthropic. (2025). *Persona vectors: Identifying and modulating personality traits in language models*. Anthropic Research Blog.

Anthropic Alignment Team. (2026). *Teaching Claude Why: Principle-based training generalizes better than behavioral imitation*. Anthropic Alignment Research. https://alignment.anthropic.com/2026/teaching-claude-why/

OpenAI Alignment Team. (2026). *Accidental Chain-of-Thought Grading: Audit and Monitorability Analysis*. OpenAI Alignment Research. https://alignment.openai.com/accidental-cot-grading/

Bhatt, M., et al. (2026). Dedicated Feature Crosscoders. *arXiv preprint arXiv:2602.11729*.

Cobbe, K., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (GSM8K).

Lindsey, J., Cunningham, H., et al. (2024). *Crosscoders for cross-checkpoint model diffing*. Anthropic. https://transformer-circuits.pub/2024/crosscoders/index.html

Li, J., & Janson, L. (2024). On the use of zero-ablation in interpretability. *arXiv preprint*.

Marks, S., et al. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

Minder, J., Dumas, L., Juang, A., Chughtai, B., & Nanda, N. (2025). Latent Scaling and KL bridging for crosscoder diagnostics. *NeurIPS 2025*. arXiv:2504.02922.

Templeton, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Research*.

Submitted to NeurIPS Mechanistic Interpretability Workshop 2026 (preprint).

Code, data, and reproducer notebook: github.com/OpenInterpretability — Apache-2.0.