Variant-specific crosscoder features are seed-stable but not detectably task-causal in a GRPO-LoRA math setting

Nozomu Fujisawa [1] **Masaaki Kondo** [1]

Abstract

We test whether variant-specific features identified by joint-norm pairwise crosscoders, trained on activation pairs from a base LLM and its RL fine-tune, correspond to task-causal mechanisms underlying the observed fine-tuned behavior. In a Qwen3-4B vs. GRPO-LoRA math setting, high-ν features pass a non-lexical cross-seed reproducibility check but fail task-causal specificity under $n = 100$ paired ablation against a magnitude-matched random control. Two complementary base-vs-base controls diagnose what the gate is responding to: under *paired-identity* ($A_t = B_t$ exactly) it produces zero high-ν features in all 12 trained crosscoders, while under *disjoint halves* it produces 50–200$\times$ as many features as base/GRPO with similar non-lexical seed-stability. The gate therefore responds to systematic between-side distributional difference, and a large high-ν population can arise from unpaired-pair reconstruction asymmetry; in the paired base/GRPO setting, the remaining high-ν population is much smaller, is consistent with model-pair distributional drift, and is not detectably task-causal under our ablations. A high-ν gate, even combined with non-lexical seed stability, is insufficient evidence for task-causal mechanisms underlying the observed fine-tuned behavior.

1. Introduction

Reinforcement-learning fine-tuning of large language models reliably improves task performance, yet the mechanistic basis of those improvements is poorly understood. Sparse autoencoders (Cunningham et al., 2024; Bricken et al., 2023; Templeton et al., 2024) and their pairwise extension — joint-norm *crosscoders* (Lindsey et al., 2024) — decompose pairs of activation streams (a base model and its fine-tune) into a shared dictionary of features whose decoder weights are split between the two sides. The latent-scaling statistic $\nu \in [0, 1]$ summarizes that split: $\nu \to 1$ means a feature expressed only by the variant. The implicit assumption underlying the methodology is that *features with ν above a high threshold (e.g. > 0.7) correspond to computations introduced or sharpened by the fine-tune*.

This paper tests that assumption. Working with Qwen3-4B (Qwen Team, 2025) and its GRPO-LoRA (Shao et al., 2024; Hu et al., 2022) math fine-tune at layer 18, we ask: (1) do variant-specific features emerge as a function of training resources? (2) Are detected features stable across crosscoder training seeds? (3) Are they causally relevant to the task (Conmy et al., 2023)?

Our answers (Figures 1–4): (1) yes for the capacity axis above an emergence threshold, no for the token axis; (2) under lexical top-token Jaccard they match a size-matched random baseline, but under non-lexical metrics (decoder cosine, held-out activation correlation) every feature pairs above all 1,000 random trials; (3) no detectable accuracy effect, even against a magnitude-matched random ablation. Two complementary base-vs-base controls diagnose the gate: a *paired-identity* control ($A_t = B_t$ exactly) produces zero high-ν features in all 12 trained crosscoders, while a *disjoint-halves* control on independent base activations produces 50–200$\times$ as many high-ν features as base/GRPO with similar non-lexical seed-stability. The high-ν gate therefore responds to systematic between-side distributional difference but a large high-ν population can arise from unpaired-pair reconstruction asymmetry; in the paired base/GRPO setting it is much smaller and not detectably task-causal. We claim only that this setting falsifies a high-ν gate as sufficient evidence; the diagnostics generalize to other base/RL pairings.

2. Method

Models. Our object of study is the GRPO-LoRA math fine-tune of Qwen3-4B (Qwen Team, 2025), trained with the GRPO objective of Shao et al. (2024) and the parameter-efficient adapters of Hu et al. (2022); the pre-RL Qwen3-4B

[1]Keio University, Japan. Correspondence to: Nozomu Fujisawa <teruyo.sui4@gmail.com>, Masaaki Kondo <kondo@acsl.ics.keio.ac.jp>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. *Copyright 2026 by the author(s).*

serves only as the *paired side* of the crosscoder. Under the 512-token greedy decoding budget used throughout our ablation pipeline, the GRPO-LoRA fine-tune emits a final `\boxed{}` answer on every problem in the audited $n = 100$ subset and reaches **0.57** audited accuracy, while the pre-RL model rarely emits a final answer under identical decoding settings (verified by inspection: $0/100$ pre-RL outputs emitted a parseable `\boxed{}` before the cutoff) — confirming that GRPO fine-tuning was not a no-op here, but leaving any behavioral gap entangled with answer-trajectory length and format compliance. We therefore make no fair-comparison claim about mathematical reasoning ability between the two models, and we ask only whether high-ν crosscoder features ablation accounts for any portion of the GRPO-LoRA model's own behavior. We extract paired hidden states at layer 18, the same indices on both sides; the *capacity-axis* experiments use 2.3M paired states, while the *token-axis* sweep varies the number of paired states from 0.3M to 5M.

Joint-norm pairwise top-k crosscoder. Following Lindsey et al. (2024) and the top-k SAE objective of Gao et al. (2024), a crosscoder with n_f features and sparsity budget k encodes $(\mathbf{h}^{\text{base}}, \mathbf{h}^{\text{var}})$ to a sparse code $\mathbf{z}$ via a top-k ReLU on a single linear encoder of the concatenation, and decodes back via two separate decoders $W_{\text{dec}}^{\text{base}}, W_{\text{dec}}^{\text{var}}$. Each feature has a per-side norm $w_i^{\text{base}} = \|W_{\text{dec}}^{\text{base}}[:, i]\|$ and w_i^{var} similarly. The latent-scaling statistic is

$$\nu_i = \frac{w_i^{\text{var}}}{w_i^{\text{base}} + w_i^{\text{var}}}. \tag{1}$$

We follow the convention $\nu_i > 0.7 \Rightarrow$ *variant-specific*, $\nu_i < 0.3 \Rightarrow$ *base-exclusive*.

Two-axis scaling. Three seeds for each axis. *Token axis*: capacity fixed at $(\text{expansion} = 8, k = 32)$, budget over $\{0.3, 0.75, 1.5, 2.3, 5.0\}$M. *Capacity axis*: budget fixed at 2.3M, $(\text{expansion}, k)$ over seven points spanning $(2, 8)$ to $(128, 512)$.

Cross-seed inspection. For every crosscoder at expansion-32 and expansion-64 across all three seeds, we record top-15 activating tokens per detected variant-specific feature on a fixed inspection corpus of 200 problems.

Causal ablation. On expansion-64 seed 42 (5 variant-specific features) we generate MATH-500 (Lightman et al., 2023; Hendrycks et al., 2021) answers under three interventions on the residual stream at layer 18: CLEAN; ABLATE-VS (subtract the decoder contribution of the 5 features); ABLATE-RANDOM (same on 5 control features sampled from $\nu \in [0.4, 0.6]$). At each step we run base and variant in parallel on the same input ids, encode through the crosscoder, and subtract $\mathbf{z}[\text{ablate}] \cdot W_{\text{dec}}^{\text{var}}[:, \text{ablate}]^\top$ from $\mathbf{h}^{\text{var}}$.

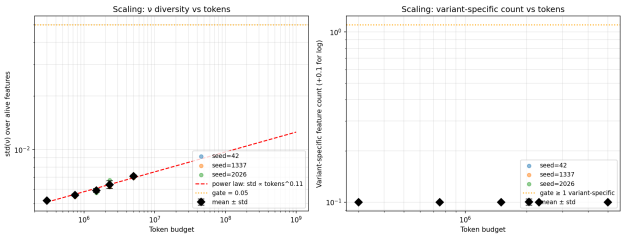

Figure 1. Token-axis scaling at expansion $= 8, k = 32$. *Left*: $\text{std}(\nu)$ across alive features grows sublinearly with budget (slope 0.11 on log–log). *Right*: variant-specific count is identically zero across all five budgets and three seeds.

The random control band has small variant-side decoder norms; the resulting asymmetry is discussed in Section 4.

Diagnostics. **D1 — ablation effectiveness**: per-problem Levenshtein ratio between generated tails of the three conditions. **D2a — lexical cross-seed stability**: pairwise Jaccard of top-15 token sets between seeds, compared against a *size-matched* per-pair random baseline (10,000 random pairs of subsets at sizes $|A_{s_i}|, |A_{s_j}|$ from the per-capacity token universe), with empirical p-values $P(J_{\text{rand}} \leq J_{\text{obs}})$. **D2b — non-lexical cross-seed stability**: for every detected variant-specific feature i in seed s we find its best match j^* among the variant-specific features of seed s' under three metrics (decoder cosine $\cos(W_{\text{dec}}^{\text{var}}[:, i]^{(s)}, W_{\text{dec}}^{\text{var}}[:, j^*]^{(s')})$; activation correlation on a held-out 200K-token slice; top-50 activating-position overlap), and compare against a *best-of-m_{vs}* random baseline: 1,000 trials, each drawing $m_{\text{vs}} = |A_{s'}|$ random non-variant-specific features from seed s' and recording their max similarity to feature i. The reported percentile is the fraction of these best-of-m_{vs} trials below the observed best within-VS match. Random non-VS features are drawn from a fixed-seed pool of 1,000 features per crosscoder (Section E).

Pre-registration and code. We maintained a written analysis protocol (hypotheses, gates, seed grid, random-control band, verdict thresholds) and a chronological amendment log. Both, along with all training/evaluation code, raw and aggregated results, and runnable supplementary notebooks (Appendix E), are released in a public repository: `https://github.com/teruyo14/variant-specific-crosscoder-features`. The amendments include the post-hoc switch to per-pair size-matched D2 reported here.

3. Results

3.1. Token-axis scaling (Figure 1)

The diversity of ν grows weakly with budget ($\text{std}(\nu) \propto \text{tokens}^{0.11}$) but does not cross the variant-specific gate at any of the 15 (budget, seed) cells: training longer at fixed capacity does not surface variant-specific features in this

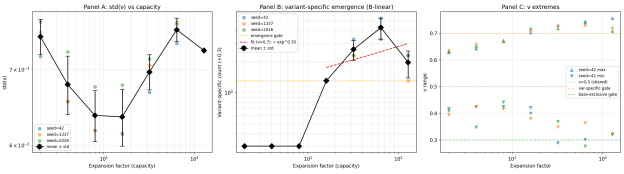

Figure 2. Capacity-axis scaling at fixed 2.3M-token budget. Variant-specific count is zero below expansion $= 16$, grows to a peak of $\bar{n} = 4.0$ at expansion $= 64$, then drops to $\bar{n} = 1.7$ at expansion $= 128$ where per-feature samples become sparse.

setting.

3.2. Capacity-axis scaling (Figure 2)

Holding the token budget fixed and varying capacity, the variant-specific count is zero up to expansion $= 8$, rises non-monotonically to a peak of $\bar{n} = 4.0$ at expansion $= 64$ (per-seed counts $3, 4, 5$), then drops to $\bar{n} = 1.7$ at expansion $= 128$. A descriptive log–log fit on the post-threshold range gives slope ≈ 0.30, but the trend is non-monotonic and we therefore avoid power-law claims. Read on its own this admits a positive interpretation ("features emerge once the dictionary is large enough"), which the next subsection refutes.

3.3. Inspection, ablation, and diagnostics (Figs 3, 4)

Inspection. Of the 265 token records collected for variant-specific features across the six crosscoders, only 7 fall in the problem-content math-words category (*math_words_problem* in Figure 3; words such as *coordinates*, *integers*); the remainder are dominated by LaTeX delimiters and bare letter/variable tokens. The detected features look more consistent with surface-form or formatting signatures than with semantically specific math-reasoning features in this setting. The ν values cluster tightly just above the 0.7 gate, without the heavy tail one would expect from features expressed only in the variant.

Ablation ($n = 20$). Accuracy is 0.500 in CLEAN, 0.550 in ABLATE-VS, and 0.500 in ABLATE-RANDOM under the parser used in the main pipeline. A supplementary symbolic-equivalence audit (Section D) recovers one parser false negative per condition (a $\left(3, \frac{\pi}{2}\right)$ vs. $(3, \frac{\pi}{2})$ case), shifting all three numbers by an identical $+0.05$ to $(0.55, 0.60, 0.55)$ and leaving the ranking unchanged. The point estimate for ABLATE-VS is on the wrong side of CLEAN for a hypothesis that the features support correctness; binomial 95% intervals at $n = 20$ overlap heavily ($\approx [0.30, 0.70]$ each). We read this as evidence *against* a strong positive interpretation rather than a strong claim of causal irrelevance, given the underpowered design (Section 4).

Ablation ($n = 100$, paired, magnitude-matched control). A follow-up evaluation (Section E) extends the ablation to $n = 100$ MATH-500 problems and adds a magnitude-matched random control (5 features matched to the variant-specific set on $(\|W_{\text{dec}}^{\text{var}}[:, i]\| \cdot \mathbb{E}[z_i],\ \Pr[z_i > 0])$). Paired audited deltas are $\Delta_{\text{vs}} = -0.01$ (95% bootstrap CI $[-0.04, +0.02]$), $\Delta_{\text{mm}} = -0.01$ ($[-0.05, +0.02]$), and a paired between-condition $\Delta = 0.00$ ($[-0.03, +0.03]$); all intervals contain zero, so we read this as a null result, not as exact equivalence (full numerics in Section E, E.2). The ablation targets the 5 variant-specific features in the expansion-64 seed-42 crosscoder; we do not claim no high-ν feature in any crosscoder is causal, but rather that the most direct high-ν set in this crosscoder fails to show task-causal specificity under this paired evaluation.

D1. Ablation is not a no-op on the $n = 20$ subset: 75% of ABLATE-VS tails differ from CLEAN, and the magnitude-matched control perturbs 100%. The original ABLATE-RANDOM condition is byte-identical to CLEAN on all 20 problems and is therefore not used for specificity claims. On the larger $n = 100$ set the non-identical-output rate is 0.17 for both ablations (Section E, E.2); we read $\Delta_{\text{vs}\to\text{mm}} = 0$ as a paired task-level null at this perturbation strength, not as evidence that every prompt was strongly perturbed.

D2a (lexical). Pairwise Jaccard of top-15 token sets between seeds, against 10,000 size-matched random pairs per (capacity, seed-pair) cell. Of six cells, four lie inside the random 95% interval, two lie significantly below ($p < 0.01$), zero lie above. Pooled observed Jaccard means 0.21 (expansion-32) and 0.23 (expansion-64) versus random ≈ 0.31. *Under this lexical diagnostic the features show no more cross-seed agreement than random subsets of the relevant vocabulary.*

D2b (non-lexical). The same six comparisons under decoder cosine, held-out activation correlation, and top-50 activating-position overlap give the opposite verdict. For each of the two non-lexical metrics, all 22 within-VS best-match comparisons (one per (capacity, seed-pair, feature)) exceed the best-of-m_{vs} random-baseline 97.5 percentile; mean within-VS similarities are ~ 0.64 and ~ 0.70 respectively, against best-of-m_{vs} random-max means ≈ 0. Top-position overlap is more permissive but still significantly above random for $14/22$ features. *The features are geometrically and activation-space reproducible across seeds even though their top tokens are lexically indistinguishable from random* (per-cap and per-pair breakdowns in Section E). Table 1 summarizes all six checks.

4. Discussion and limitations

The high-ν features (i) cluster tightly above $\nu = 0.7$, (ii) concentrate on LaTeX delimiters and bare letter/variable tokens, (iii) carry a reproducible identity across seeds under decoder cosine and held-out activation correlation (D2b: 22/22) even though their top tokens are statistically indistinguishable from random (D2a), and (iv) yield no detectable

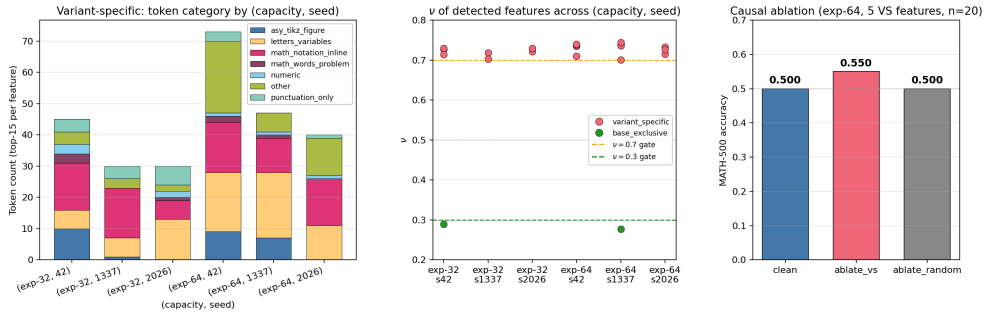

Figure 3. Left: top-15 token categories of detected variant-specific features across expansion-32 and expansion-64, three seeds each. Math notation and bare letters/variables dominate; *math_words_problem* is rare. *Center*: ν values cluster just above the gate. *Right*: under the main parser, ablation does not produce a detectable accuracy drop ($0.50 \rightarrow 0.55$, $n = 20$); a supplementary symbolic audit (Appendix D) shifts all conditions by $+0.05$.

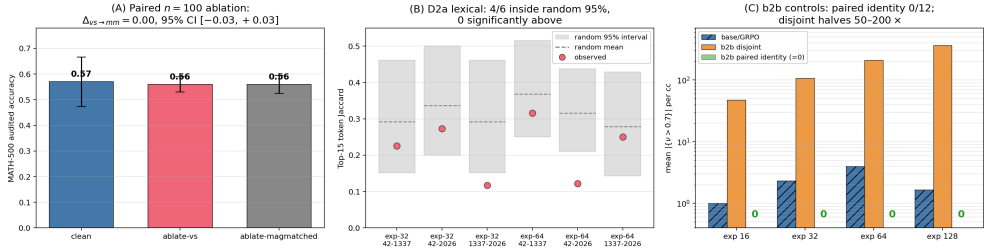

Figure 4. Diagnostics. *(A)* Paired $n = 100$ ablation, audited accuracy: $\Delta_{\text{vs}\rightarrow\text{mm}} = 0.00$ (95% bootstrap CI $[-0.03, +0.03]$); no detectable task-causal specificity. *(B)* Pairwise top-15 token Jaccard versus a size-matched per-pair random baseline (10,000 trials per pair): observed values are statistically indistinguishable from random ($4/6$ inside the 95% interval, 0 significantly above, 2 significantly below at $p < 0.01$). *(C)* High-ν population per crosscoder across capacity sweep (mean over 3 seeds, log y-axis): paired-identity b2b control produces zero high-ν features at every capacity, while disjoint-halves produces 50–200$\times$ as many as base/GRPO. The $n = 20$ Levenshtein perturbation diagnostic is in Appendix E.

Check	Result	Reading
$n = 20$ ablation	no detectable drop	underpowered
$n = 100$ paired, mag-matched	$\Delta = 0$, CI $[-0.03, +0.03]$	no task-causal specificity
D2a lexical Jaccard	$4/6$ inside random 95%	top tokens unstable
D2b decoder cos / act. corr	$22/22 >$ best-of-m_{vs} $p_{97.5}$	non-lexical reproducible directions exist
b2b paired identity (E.6)	$0/12$ crosscoders produce any high-ν feature	gate is silent under exact identity
b2b disjoint halves, count (E.4)	50–200$\times$ vs. base/GRPO	unpaired-pair asymmetry inflates count
b2b disjoint halves, geometric (E.5)	$520/577$, $567/577 >$ best-of-m_{vs} $p_{97.5}$	inflated count is also seed-stable

Table 1. Summary of diagnostics and controls (numerics and protocols in Section E). The paired-identity b2b control (E.6) confirms the gate does not gratuitously emit asymmetry; the disjoint-halves b2b controls (E.4–E.5) show that a large high-ν population can arise from unpaired-pair reconstruction asymmetry; the $n = 100$ paired ablation finds no detectable task-causal specificity for the smaller population that appears in the base/GRPO setting.

task-causal effect at $n = 20$ or $n = 100$ under magnitude-matched ablation. Two base-vs-base controls diagnose the gate. *Paired-identity* ($A_t = B_t$, E.6): all 12 trained crosscoders produce zero high-ν features (ν std collapses to 0.004); the gate does not gratuitously emit asymmetry. *Disjoint halves* (E.4–E.5): the gate produces 50–200$\times$ as many features as base/GRPO with similar non-lexical seed-stability ($520/577 = 90\%$ above the best-of-m_{vs} random 97.5 percentile). The gate responds to systematic between-side difference (silent under exact identity), and a large high-ν population can arise from unpaired-pair reconstruction asymmetry. In the base/GRPO setting, where token pairing is preserved, the high-ν population is much smaller (1–5 features per crosscoder); it is consistent with model-pair distributional drift but is not detectably the task-causal RL computation. The methodological lesson: non-lexical seed stability is itself insufficient—the disjoint base-vs-base control produces hundreds of seed-stable high-ν features with no model-pair difference.

Four concrete remediations follow. *(1)* Pair a ν-gate with both lexical and non-lexical cross-seed checks, which gave opposite verdicts here. *(2)* Use magnitude-matched random controls (decoder-norm $\times$ activation-magnitude) rather than a fixed ν-band: our nominal $\nu \in [0.4, 0.6]$ control was byte-identical to CLEAN on all 20 problems. *(3)* Pair small-n ablations with paired analysis at larger n: $n = 20$'s $+0.05$ delta tightens to -0.01 (95% CI $[-0.04, +0.02]$) at $n = 100$. *(4)* Run *both* a paired-identity and a disjoint-halves b2b control: opposite verdicts here (0 vs. 50–200$\times$) rule out a single "no-difference" label.

Limitations. One base model, layer, fine-tune, domain; $n = 100$ CI $[-0.03, +0.03]$ rules out specific effects but not ± 0.02 equivalence, and greedy decoding amplifies low-margin trajectories. The high-ν gate alone is insufficient

here; the diagnostics make the difference falsifiable. This negative result poses no new deployment risks.

Acknowledgements

This work was supported by JST-Research and Development Program for Next-generation Edge AI Semiconductors Grant Number JPMJES2512.

References

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step, 2023. Source of the MATH-500 evaluation subset.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Cunningham, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html. Anthropic technical update.

Qwen Team. Qwen3 technical report, 2025. Model card for Qwen3 family.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

A. Pre-registration and amendments

The pre-registered protocol (in `prereg/prereg_v1.0.0.yaml`) specified the central hypothesis to be tested (*variant-specific features detected by the crosscoder reflect RL-induced computation*), the gates ($\nu_{\text{vs}} = 0.7$, $\nu_{\text{be}} = 0.3$), the seed grid ($\{42, 1337, 2026\}$), the two scaling axes, the inspection corpus size (200), the causal ablation protocol with random control, and the verdict thresholds for both diagnostics. The amendment log (`prereg/amendments.yaml`) records five revisions: a power-law fit on the token axis (quantitative reporting only); two extension capacity points to distinguish among three pre-specified shape verdicts; a specification of the random control band that the original protocol did not pin down (the asymmetry this introduces is disclosed in Section 4); the addition of the two diagnostics; and a final claim revision to reflect the negative result. We do not claim the protocol was timestamp-certified "before any analysis"; we claim only that a written protocol and amendment log accompany the work.

B. Per-condition ablation samples

The supplementary repository contains `results/ablation_clean.json` and `results/ablation_ablate_vs.json` with per-problem generated tails and extracted answers for all 20 problems. Of the 15 problems where ABLATE-VS differs from CLEAN, the parsed answer string is identical in 14 cases and differs in 1.

C. Random-baseline distribution

The random baseline used in Figure 4B samples 10,000 random pairs of *size-matched* subsets from the per-capacity union of variant-specific top tokens; for each observed seed pair (s_i, s_j) the random subsets have sizes $|A_{s_i}|, |A_{s_j}|$ matching the observed signature sizes. Of the six cells, four lie inside the random 95% interval, two lie significantly below ($p < 0.01$, namely expansion-32 seeds 1337-2026 with $J = 0.118$ and expansion-64 seeds 42-2026 with $J = 0.122$), and zero lie significantly above. The empirical p-values, per-pair $2.5/97.5$ percentiles, and per-pair sample sizes are tabulated in `results/diagnostics_cross_seed_random_baseline.csv`.

D. Symbolic answer audit

The MATH parser used in the main ablation pipeline normalizes whitespace and a few delimiters but does not perform symbolic equivalence. We re-evaluate all 20 problems on each of the three conditions (CLEAN, ABLATE-VS, ABLATE-RANDOM) using a sequence of three checks: a strict whitespace-stripped match, a canonical-LaTeX match (stripping `\left/\right`, normalizing fraction and spacing macros), and a SymPy-based symbolic equivalence check that decomposes tuples and tests `simplify(a-b)==0` per component. The audited accuracy and the parser accuracy differ by exactly one false negative per condition (problem 0: $\left(3, \frac{\pi}{2}\right)$ vs. $(3, \frac{\pi}{2})$), shifting every cell by $+0.05$:

Condition	Parser	Audited
CLEAN	0.500	0.550
ABLATE-VS	0.550	0.600
ABLATE-RANDOM	0.500	0.550

The qualitative ranking is unchanged. The full per-problem table is in `results/audit_per_problem.csv`; the audit code is in `notebooks/05_manual_answer_audit.ipynb`.

E. Supplementary ablation experiments and follow-up protocols

This appendix contains six executed supplementary experiments (E.1–E.6). All six are released as complete, runnable end-to-end scripts and ship with their per-problem outputs and aggregated CSV/JSON in the supplementary repository.

E.1 Magnitude-matched random ablation control ($n = 20$). (`notebooks/06_magmatched_control.ipynb`.) Replaces the main paper's $\nu \in [0.4, 0.6]$ random control with five features matched to the variant-specific set in joint (decoder-norm $\times$ mean activation, activation frequency) space. Per-feature contribution magnitude $c_i = \|W_{\text{dec}}^{\text{var}}[:, i]\| \cdot \mathbb{E}[z_i]$ and frequency $f_i = \Pr[z_i > 0]$ are computed on the inspection corpus, candidates restricted to alive features with $\nu < 0.7$,

and a greedy 1-1 matching (top-3 nearest neighbours per variant-specific feature, random tie-break, seed=42) is performed in $(\log c, \log f)$ space. The five matched indices have contribution ratios 0.73–1.05 and matched activation frequencies (`ablation/magmatched_indices.json`). On the original 20-problem subset the magnitude-matched ablation perturbs all 20 outputs (mean Levenshtein 0.40 vs. CLEAN), versus 15/20 for ABLATE-VS (mean Levenshtein 0.60); audited accuracy is 0.60 for the magnitude-matched control, identical to ABLATE-VS (0.60) and within binomial noise of CLEAN (0.55). The original $\nu \in [0.4, 0.6]$ ABLATE-RANDOM is byte-identical to CLEAN on all 20 problems, confirming the no-op behavior assumed in the main text.

E.2 $n = 100$ ablation with paired analysis. (`notebooks/07_n100_ablation.ipynb`.) Extends the evaluation to 100 MATH-500 problems (the original 20 plus 80 additional, drawn by fixed-seed random sampling, seed=2026) on three conditions: CLEAN, ABLATE-VS, and ABLATE-MAGMATCHED. Audited accuracies are $0.57, 0.56, 0.56$ respectively. Paired analysis (bootstrap 95% CI on the paired audited delta, 10 000 resamples; McNemar test on discordant pairs; exact sign test): $\Delta_{\text{vs}} = -0.01$ (CI $[-0.04, +0.02]$, McNemar $p = 1.0$, sign $p = 1.0$, 3 discordant pairs out of 100); $\Delta_{\text{mm}} = -0.01$ (CI $[-0.05, +0.02]$, McNemar $p = 1.0$, sign $p = 1.0$, 3 discordant pairs); $\Delta_{\text{vs}\to\text{mm}} = 0.00$ (CI $[-0.03, +0.03]$, McNemar $p = 0.48$, 2 discordant pairs). All intervals contain zero; none of the paired comparisons reaches significance at $\alpha = 0.05$. We read this as a null result — no detectable task-causal difference between variant-specific and magnitude-matched random ablation under the resources used here — not as positive evidence that the two ablations are exactly equivalent. The non-identical-output rate on the full $n = 100$ set is 0.17 for both conditions (mean Levenshtein 0.91 for ABLATE-VS, 0.91 for ABLATE-MAGMATCHED), substantially lower than on the 20-problem subset; under greedy decoding only steps with small top-1/top-2 logit margins propagate hidden-state perturbations to the output, so the original 20 problems appear to have been over-represented in low-margin trajectories. The full per-problem table is in `ablation/n100/paired_table.csv` and the aggregated statistics in `ablation/n100/paired_summary.json`.

E.3 Non-lexical cross-seed stability. (`notebooks/08_cross_seed_geometric_stability.ipynb`.) Strengthens Diagnostic 2 with three non-lexical similarity metrics between variant-specific features across seeds within expansion-32 and expansion-64: (i) decoder cosine similarity $\cos(W_{\text{dec}}^{\text{var}}[:, i]^{(s)}, W_{\text{dec}}^{\text{var}}[:, j]^{(s')})$; (ii) Pearson correlation between $z_i(x)$ from crosscoder s and $z_j(x)$ from crosscoder s' on a fixed held-out 200K-token slice; (iii) Jaccard overlap of the top-50 activating token positions of each feature on the same slice. For each variant-specific feature i in seed s we record the best within-VS match j^* in seed s' and the percentile of the resulting similarity against a *best-of-m_{vs}* random baseline (1,000 trials, each drawing $m_{\text{vs}} = |A_{s'}|$ random non-variant-specific features from a fixed pool of 1,000, seed=2026, recording the max similarity per trial). This makes observed (best-of-m_{vs} within-VS) and baseline (best-of-m_{vs} random) directly comparable. Aggregating over all 22 (cap, seed-pair, feature) cells:

	decoder cos	act. corr	top-pos overlap
mean observed	0.64	0.70	0.36
mean best-of-m_{vs} rand	-0.0003	$+0.0002$	$\sim 2\times10^{-5}$
n above best-of-m_{vs} $p_{97.5}$	22/22	22/22	14/22
mean NN percentile	1.00	1.00	0.64

Per-cap means are $(0.61, 0.64, 0.34)$ for expansion-32 and $(0.66, 0.73, 0.37)$ for expansion-64 for the three metrics respectively. Within each metric the per-feature similarity has a substantial spread ($\sigma \approx 0.26$–0.38): a fraction of features pair near-perfectly across seeds while others pair more weakly. The qualitative conclusion is robust: under both decoder cosine and held-out activation correlation *every* detected variant-specific feature outpairs the random non-variant-specific best-of-m_{vs} 97.5 percentile, and the population mean is two to three orders of magnitude above the random mean. Top-position overlap shows the same direction more weakly. Per-pair breakdowns are tabulated in `results/diagnostics_cross_seed_geometric.csv` and the verdict in `results/diagnostics_cross_seed_geometric_summary.json`.

E.4 Base-vs-base disjoint-halves control crosscoder. (`notebooks/09_base_vs_base_control.ipynb`.) Trained a joint-norm pairwise crosscoder on $(\mathbf{h}^A, \mathbf{h}^B)$ pairs where A and B are disjoint halves of the same base model's layer-18 activations (1.15M tokens per side, drawn by fixed-seed permutation), with the same architecture, hyperparameters, and $\nu > 0.7$ gate as the main paper. Sweep over expansions $\{16, 32, 64, 128\}$ and seeds $\{42, 1337, 2026\}$ (12 crosscoders). This setting eliminates model-pair difference but breaks token-level pair alignment: the crosscoder sees pairs $(\mathbf{h}^{\text{base}}(x_t), \mathbf{h}^{\text{base}}(x_{t'}))$ for

unrelated tokens t, t'. The bulk of the ν distribution sits tightly at 0.500 ± 0.02 as expected when both sides come from the same base, but the extreme right tail grows steeply with capacity:

expansion	disjoint b2b mean $\lvert\{\nu > 0.7\}\rvert$	base/GRPO mean $\lvert\{\nu > 0.7\}\rvert$	ratio
16	47.7	1.00	47.7×
32	106.7	2.33	45.8×
64	210.7	4.00	52.7×
128	364.3	1.67	218×

The high-ν tail is populated 50–200× more strongly than in the base/GRPO sweep. Read in isolation, this looks like a crosscoder-architecture artifact, but the paired-identity control in E.6 shows the gate produces *zero* high-ν features when $A_t = B_t$ exactly. The disjoint-halves inflation therefore reflects *unpaired-pair reconstruction asymmetry* — per-token noise that the crosscoder routes preferentially to one side — rather than a generic crosscoder behavior. In the base/GRPO setting, where token pairing is preserved, the absolute high-ν count is small (1–5 per crosscoder) and is consistent with model-pair distributional drift; we do not directly identify it as such here. The `base_vs_base/results.csv` and `summary.json` contain the per-(cap, seed) counts.

E.5 Geometric stability on b2b crosscoders. (`notebooks/10_b2b_geometric_stability.ipynb`.) Re-runs E.3's three non-lexical similarity diagnostics on the base-vs-base crosscoders trained in E.4. Because the b2b crosscoders contain 43–379 variant-specific features per crosscoder (vs. 2–5 in the base/GRPO sweep), we sample up to 50 vs features per crosscoder ($n = 577$ feature-pair comparisons total across expansions $16, 32, 64, 128$); other settings (best-of-m_{vs} random baseline with random pool size 1000, held-out 200K-token slice, percentile threshold 97.5) match E.3 exactly.

	decoder cos	act. corr	top-pos overlap
mean observed	0.45	0.43	0.11
mean best-of-m_{vs} rand	0.073	0.0013	$\sim 3 \cdot 10^{-4}$
n above best-of-m_{vs} $p_{97.5}$	$520/577$	$567/577$	$287/577$
fraction	90.1%	98.3%	49.7%

The cross-seed geometric reproducibility seen in E.3 largely holds in the disjoint-halves base-vs-base control: 90.1% ($520/577$) of detected variant-specific features pair above the best-of-m_{vs} random 97.5-percentile under decoder cosine, and 98.3% ($567/577$) under activation correlation. The fraction-above-random declines with capacity — decoder-cosine fractions are $96.9\%, 92.7\%, 88.7\%, 83.3\%$ for expansions $16, 32, 64, 128$ respectively, with mean observed similarities $0.78, 0.53, 0.32, 0.24$ over the same sweep — consistent with larger dictionaries spreading similar modes across more asymmetric units. The unpaired-pair asymmetry features identified in E.4 are therefore themselves seed-stable in the geometric sense: which directions the crosscoder allocates to one side is determined by consistent features of the input distribution, not seed-specific noise. This does not generalise to base/GRPO directly because the paired-identity control (E.6) shows the gate produces zero features under exact identity; we cannot rerun the geometric diagnostic on zero features. The base/GRPO setting's $22/22$ reproducibility (E.3) could reflect either real RL drift signal that is seed-stable, or the same kind of input-distribution-driven reproducibility we see in the disjoint b2b features. Distinguishing these is left to follow-up work.

E.6 Paired-identity base-vs-base control. (`notebooks/11_b2b_paired_identity_control.ipynb`.) Trains the same crosscoders as E.4 but with $\mathbf{h}_t^A = \mathbf{h}_t^B$ for every token t: the two sides of every training pair are the same base activation. The $\nu > 0.7$ gate, expansion sweep $(16, 32, 64, 128)$, seed grid ($\{42, 1337, 2026\}$), and reconstruction objective match E.4 exactly; only the pair construction differs.

expansion	paired-identity $\lvert\{\nu > 0.7\}\rvert$	ν mean	ν std
16	0/3 seeds	0.500	0.0045
32	0/3 seeds	0.500	0.0044
64	0/3 seeds	0.500	0.0042
128	0/3 seeds	0.500	0.0044

All 12 trained crosscoders produce *zero* variant-specific features. The ν standard deviation collapses by an order of magnitude versus the disjoint-halves control ($\sigma \approx 0.004$ vs. 0.02); the bulk of the distribution is essentially a δ-function at $\nu = 0.5$. This rules out the simplest reading of E.4–E.5 as "the crosscoder gratuitously emits asymmetry under the joint-norm objective regardless of input." The crosscoder needs *some* between-side difference to populate the high-ν tail; under exact identity it does not. The disjoint-halves inflation of E.4 (50–200$\times$ the base/GRPO rate) is therefore an unpaired-pair reconstruction artifact: per-token noise that the crosscoder routes to one side. In the base/GRPO setting (token alignment preserved) the high-ν count is much smaller (1–5 per crosscoder) and is consistent with model-pair distributional drift — a candidate signal that, under the ablation diagnostics in E.2, is not detectably task-causal for MATH-500 accuracy.

F. Two distinct populations: a candidate reading

The supplementary controls (E.4–E.6) reveal that "high-ν features" is not a single population. The paired-identity control (E.6) confirms that with $A_t = B_t$ exactly, the gate produces zero features: there is no inherent crosscoder bias toward asymmetry. The disjoint-halves control (E.4) produces 50–200$\times$ as many features as base/GRPO, with similar non-lexical seed-stability (E.5). The base/GRPO setting itself produces a small number (1–5 per crosscoder), with the same lexical/non-lexical pattern. Here we sketch a candidate reading consistent with all of these observations; we do not claim it is established.

Two populations, two mechanisms. We distinguish:

1. *Unpaired-pair asymmetry* (the disjoint-halves population). When A and B come from independent token slices, a single training pair contains the per-token reconstruction targets for two unrelated activations. The joint-norm objective, given enough dictionary capacity, will allocate some directions preferentially to one side because the per-token noise on the two sides is independent. This produces a large, capacity-scaling high-ν population that is seed-stable in the geometric sense: which directions get routed to which side is determined by consistent properties of the token distribution, not by initialisation.

2. *Candidate model-pair drift* (the base/GRPO population). When $A_t = \mathbf{h}^{\text{base}}(x_t)$ and $B_t = \mathbf{h}^{\text{GRPO}}(x_t)$ with token alignment preserved, the per-token noise present in the disjoint case is suppressed. What remains is consistent with the systematic distributional drift between base and GRPO models, which for a LoRA fine-tune is small in absolute terms. This produces a much smaller high-ν population, qualitatively distinct in size from the unpaired case but qualitatively similar in seed-stability and token concentration (LaTeX delimiters, letter/variable tokens).

The capacity-axis non-monotonicity in the base/GRPO sweep (Figure 2; rises to $\bar{n} = 4.0$ at expansion 64, drops to 1.7 at 128) is harder to explain on either mechanism alone. It may reflect that as dictionary capacity grows, the small drift signal is divided across more units, with some falling below the $\nu = 0.7$ gate. Characterising this directly, and explaining why the base/GRPO seed-stability happens to mimic the disjoint-b2b pattern even though the absolute counts differ by 50–200$\times$, is the natural next step. We leave this to follow-up work; the diagnostics we describe here are what would falsify either reading.