

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

Nozomu Fujisawa • Masaaki Kondo
Graduate School of Science and Technology, Keio University

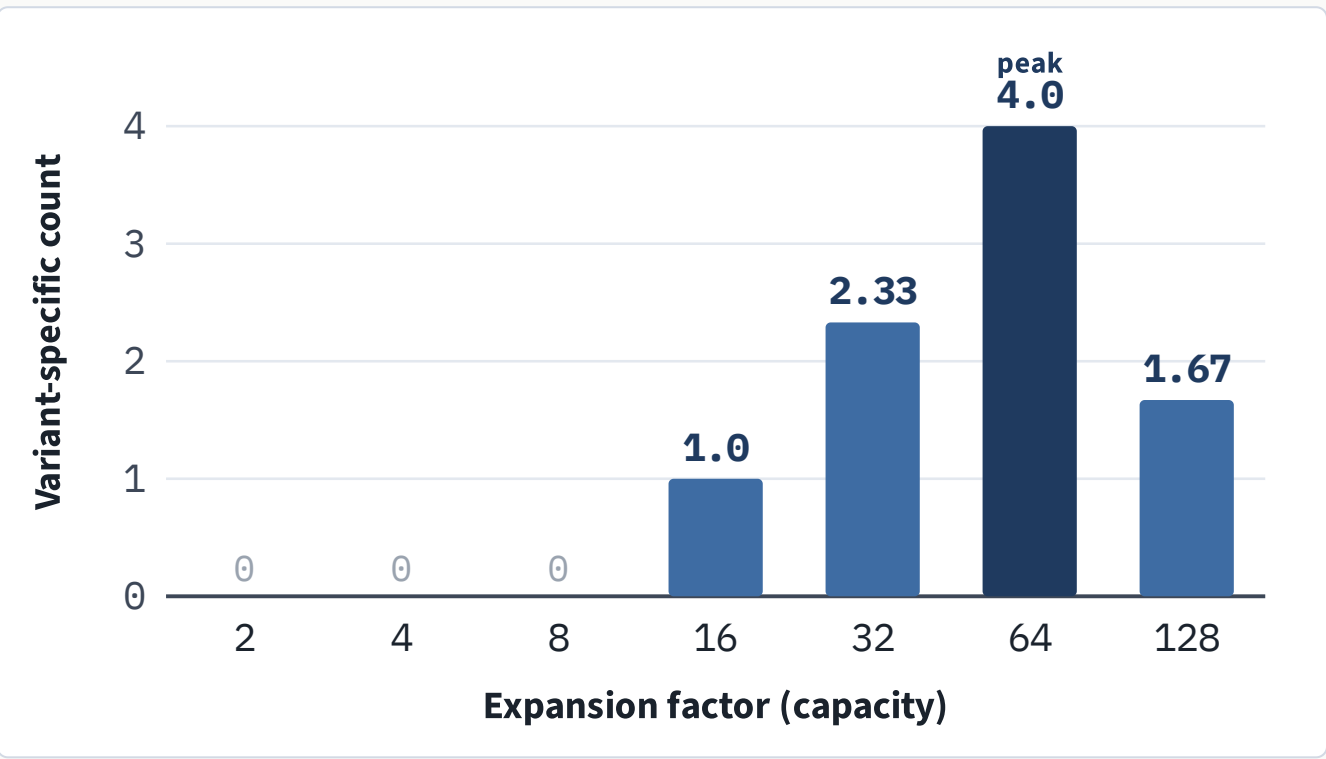
CLAIM

Seed-stable high- v features are **not sufficient evidence** of task-causal RL computation — two base-vs-base controls diagnose what the gate actually responds to.

STUDY SETUP Qwen3-4B → GRPO-LoRA math • Crosscoder layer 18 • gate $v > 0.7$ • Seeds 42 / 1337 / 2026

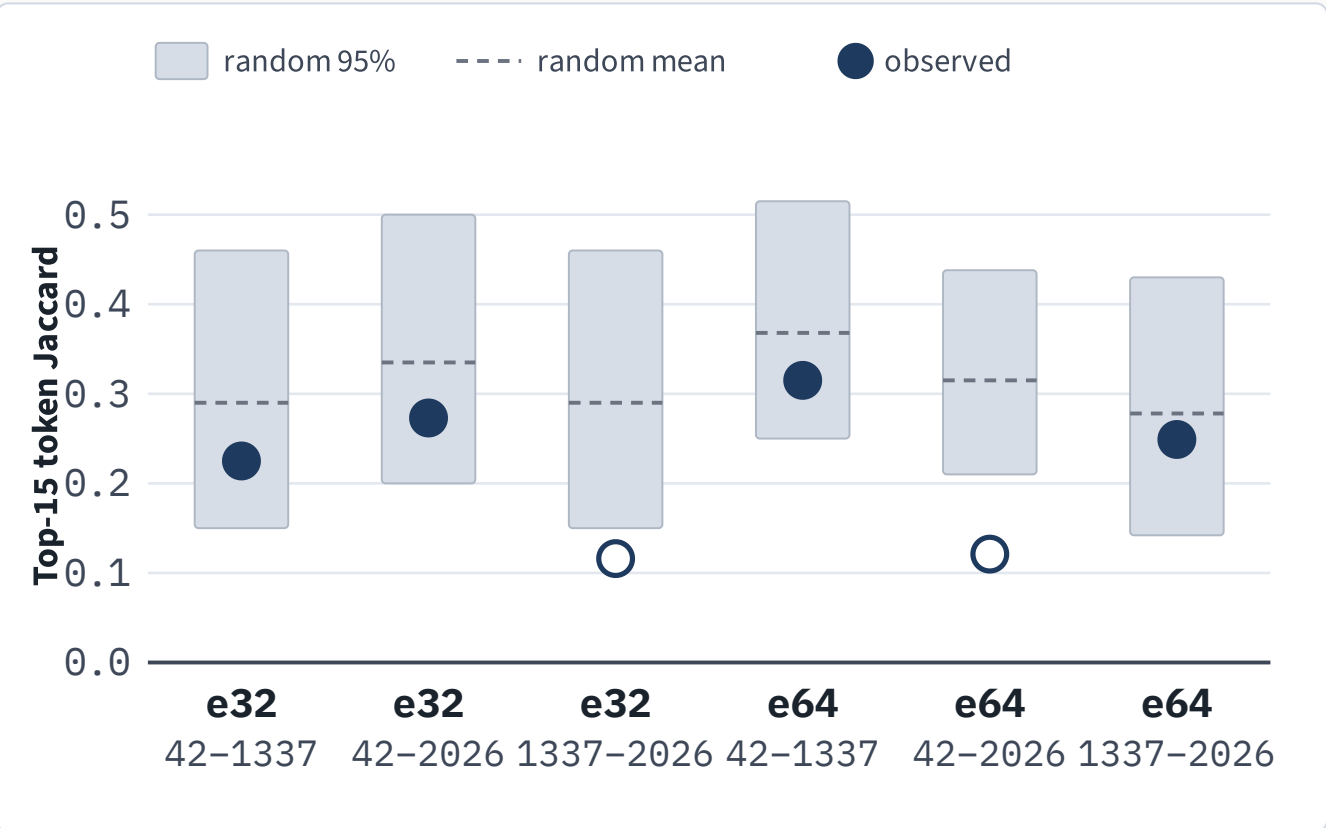
EVIDENCE • STABILITY

1 What high- v seems to show



Emergence is a capacity effect. High- v count peaks at expansion 64 ($\bar{n} = 4.0$); raising the token budget alone keeps it at 0. Mean over three seeds.

Non-lexical (decoder cosine): 22/22 cross-seed feature matches exceed the random 97.5th percentile — the feature **direction** is seed-stable.



Lexical (top tokens): 4/6 seed pairs sit inside the random 95% interval, 0 significantly above (2 below) — top tokens drift per seed.

Whether the features look “stable” **depends on the metric** — for the very same feature set.

EVIDENCE • CAUSALITY

2 Is it task-causal? No detectable effect



No task-specific drop beyond a matched perturbation.
 $n = 100$ • paired • MATH-500, against a magnitude-matched random control — a null result, not exact equivalence. Removing variant-specific features does **not detectably lower accuracy beyond the matched control**.

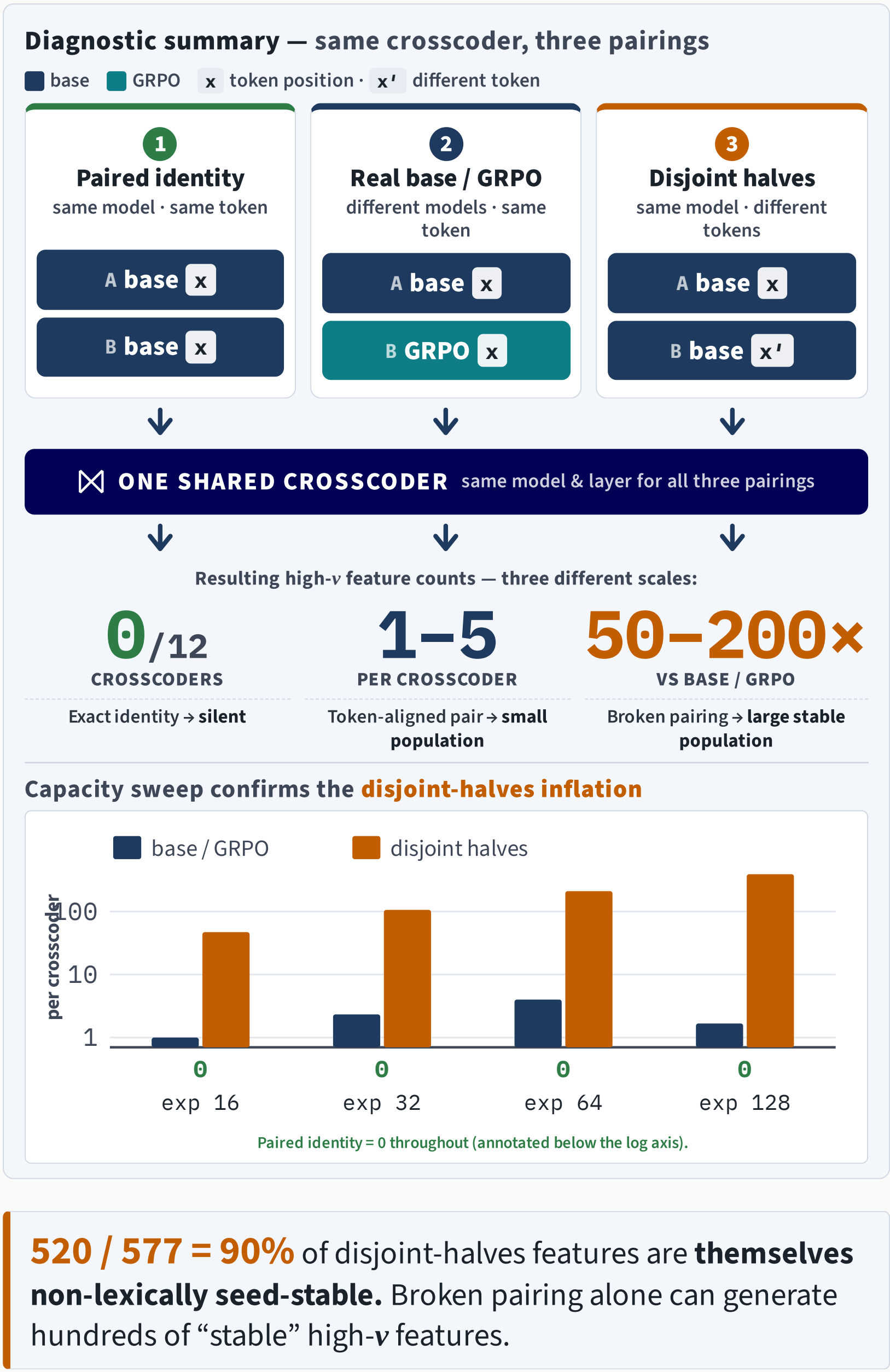
PERTURBATION VALIDITY CHECK		
Both interventions changed 17/100 outputs.		
INTERVENTION	OUTPUT CHANGED VS. CLEAN	CORRECTNESS-DISCORDANT
Ablate-vs	17 / 100	3 / 100
Mag-matched	17 / 100	3 / 100

Both interventions altered 17/100 outputs; their task-accuracy effects remained indistinguishable.

The magnitude-matched random control **sets the bar** — and variant-specific ablation does not clear it.

DIAGNOSIS • MAIN CONTRIBUTION

3 What is the gate responding to?



Candidate reading: mispairing can route consistent input-distribution directions to one side — making non-causal high- v features seed-stable.

Non-lexical seed-stability is itself insufficient: a high- v gate alone is **not sufficient evidence of task-causal RL computation**.

DIAGNOSTIC CHECK	RESULT	READING
Lexical top-token Jaccard	4/6 within random • 0 above	Top tokens drift across seeds
Non-lexical decoder cosine	22/22 feature matches above random	Feature direction is seed-stable
$n=100$ magnitude-matched ablation	$\Delta = 0.00$ [-0.03, +0.03]	No detectable task-causality
Paired-identity b2b ($A_t = B_t$)	0 / 12 crosscoders	Gate is silent on identical input
Disjoint-halves b2b (mispaird)	50–200× • 520/577 stable	Mispairing produces seed-stable high- v features
Top-token category audit	7 / 265 math-words	Consistent with surface / formatting signatures
Protocol: pair lexical + non-lexical seed checks • use magnitude-matched controls (a fixed v -band control was byte-identical to clean) • pair small- n with larger- n ablation (+0.05 at $n=20 \rightarrow -0.01$ at $n=100$) • run both base-vs-base controls.		

