Function-Vector Heads Are Two Populations: Writers and Cancellers in In-Context Learning

Anonymous ACL submission

Abstract

Function-vector (FV) heads (Todd et al., 2024) are typically identified by the *magnitude* of their causal contribution to in-context rule tasks, under the implicit assumption that the top set is a homogeneous functional class. **This assumption fails.** We replace magnitude-only ranking with a sign-preserving criterion (refined DLA + permutation FDR) and validate each candidate by path patching. The FV head population then splits into two opposing subpopulations: *writers* push the rule-correct logit up; *cancellers* push it down. A four-condition canonical verdict holds in $13/15$ cells across three model families and six Pythia scales, and a sign-shuffle rejects homogeneity in $5/6$ main cells. The structure is invisible to magnitude-only ranking: Todd's top-20 captures 64% of cancellers but only 4% of writers on the hierarchical task, and 59% of writers but only 8% of cancellers on the modular task. We rule out six artefact accounts on all 27 canceller (cell, head) pairs: induction overlap, sinks, generic importance, rank-1 copy-suppression, V-cascade, and rank-nearest non-FV controls. Zero-ablating cancellers yields $+0.13$ to $+0.29$ nats of logit gain in $6/6$ main cells with a directionally consistent $+2$ to $+7$ pp accuracy effect.

1 Introduction

Function-vector (FV) heads (Todd et al., 2024; Hendel et al., 2023; Variengien and Winsor, 2023; Olsson et al., 2022) [1] are widely studied in mechanistic interpretability as a sparse set of attention heads whose causal contribution to in-context-learning (ICL; Brown et al., 2020) rule tasks is large enough to support targeted steering, ablation, and circuit-level claims.

The standard pipeline ranks these heads by the *magnitude* of a causal-contribution metric and treats the top set as one homogeneous functional class. Implicit in this recipe is a sign assumption: each high-magnitude head pushes the model *toward* the demonstrated rule. Under this assumption, magnitude is identity: one number per head.

We show that this assumption fails empirically. Across the Pythia ladder (410M–12B, six sizes), two ICL rule tasks, and three spot-check architectures (Qwen2.5-$\{1.5, 7\}$B, GPT-2-medium), the high-magnitude population splits cleanly into two opposite roles: *writers* (refined-DLA > 0, push the rule-correct token's logit up) and *cancellers* (refined-DLA < 0, push it down). The split is large, reproducible across cells, and survives six confound rule-outs (attention sinks, induction overlap, copy-suppression, V-cascade, generic importance, rank-nearest non-FV controls). We lead with the population finding; the L11.H4 case study (§4.7) is one illustrative head.

Approach. Per (model, task) cell we identify the FV head set via refined DLA, validate it with path patching, and partition by sign of direct effect into $\mathcal{W}$ and $\mathcal{C}$. A four-condition canonical verdict ($W_{\text{shift}} < 0$, $C_{\text{shift}} > 0$, opposite signs, joint attenuated) arbitrates the population claim. Six artefact accounts are ruled out: per-source DLA bucketing for attention sinks, plus induction-overlap TOST, generic-importance ratio, rank-1 copy-suppression (McDougall et al., 2023), V-cascade (Wang et al., 2023), and rank-nearest non-FV controls. A single-head case study on the dominant canceller L11.H4 (Pythia-410M) uses four orthogonal interventions (V-shuffle, OV singular spectrum, V-composition with the dominant writer, and cross-template sign-flip) to test which standard mechanistic template, if any, it instantiates.

[1] Glossary. *FV head*: a function-vector head. *DLA* (direct logit attribution): a head's signed contribution to the correct−incorrect logit difference (Elhage et al., 2021; Geva et al., 2022). *OV*: the head's write-direction in the residual stream; *QK*: the head's attention-source pattern (Elhage et al., 2021). *Path patching* (Wang et al., 2023; Goldowsky-Dill et al., 2023) isolates a head's *direct* effect by patching its activation while holding downstream paths fixed.

Related work. The FV-head literature (Todd et al., 2024; Hendel et al., 2023; Variengien and Winsor, 2023; Subramani et al., 2022) identifies a sparse set of heads whose causal contribution to a rule task supports steering and ablation, and ranks them by the *magnitude* of that contribution. This pipeline implicitly treats the top set as one functional class, the assumption this paper overturns. Sign-opposed heads have appeared before, but only as *single-task, hand-built circuits*: IOI's negative name-movers (Wang et al., 2023) and GPT-2 L10.H7's copy-suppression (McDougall et al., 2023), each localised to one task by targeted circuit analysis. We lift the same sign-opposition to a population-level property holding across tasks, scales, and architectures (comparison in §5, Table 5).

Methodologically we use standard tooling: direct-logit attribution (Geva et al., 2022; Elhage et al., 2021), path patching (Wang et al., 2023; Goldowsky-Dill et al., 2023), causal mediation (Vig et al., 2020; Meng et al., 2022), and the OV/QK decomposition (Elhage et al., 2021). Sparse-autoencoder features (Bricken et al., 2023) and circuit-discovery catalogues (Hanna et al., 2023; Nanda et al., 2023; Conmy et al., 2023; Syed et al., 2024; Marks et al., 2025) are complementary; we instead re-examine the *sign* of effects within an existing head inventory.

Contributions. **(1) Population finding.** The FV head set splits into two opposing functional roles: *writers* push the rule-correct logit up; *cancellers* push it down. The split is causal ($6/6$ main lesions; $13/15$ with transfer) and resists six artefact accounts (induction overlap, BOS / format-prefix sinks, generic importance, rank-1 copy-suppression (McDougall et al., 2023), V-cascade (Wang et al., 2023), and a rank-nearest non-FV control) on all 27 canceller (cell, head) pairs. **(2) Single-head mechanism candidate.** L11.H4 (Pythia-410M, dominant canceller in both rule tasks) is a *distributed-bias content reader*: a single head whose role flips across templates as "suppress what was demonstrated" predicts, but whose mechanism fits neither rank-1 copy-suppression nor V-cascade. We do not extrapolate this template to the full canceller population. **(3) Calibration-free intervention.** Zero-ablating cancellers raises the correct-incorrect log-prob difference by $+0.13$ to $+0.29$ nats with CI excluding 0 in $6/6$ main cells, and shifts accuracy by $+2$ to $+7$ pp (CI-confirmed on hier-410M). This is *not* the Todd-style aggregation route, which we show is the wrong operation for the purpose (§5).

Model	arch.	L	H/layer	d_{model}	d_h
Pythia-410M	GPT-NeoX	24	16	1024	64
Pythia-1B	GPT-NeoX	16	8	2048	256
Pythia-1.4B	GPT-NeoX	24	16	2048	128
Pythia-2.8B	GPT-NeoX	32	32	2560	80
Pythia-6.9B	GPT-NeoX	32	32	4096	128
Pythia-12B	GPT-NeoX	36	40	5120	128
Qwen2.5-1.5B	Llama-style	28	12	1536	128
Qwen2.5-7B	Llama-style	28	28	3584	128
GPT-2-medium	Conv1D	24	16	1024	64

Table 1: Models studied. The first three rows form the main mechanism pipeline (Pythia 410M / 1B / 1.4B); the remaining six are extension cells used for behavioural-verdict transfer (§4.6).

2 Setup

Models. The main pipeline studies three Pythia checkpoints (Biderman et al., 2023) (410M, 1B, 1.4B), loaded in `fp32` with non-fused (eager) attention so per-head weights and post-softmax outputs remain exposed to forward hooks; see App. A.2 for the exact HuggingFace flag. Pythia provides a single architecture family across an order-of-magnitude scale range, isolating the scale axis from architectural confounds. §4.6 extends to the rest of the Pythia ladder (Pythia-2.8B/6.9B/12B; bf16 on the ≥ 6.9B sizes), and adds three single-task spot checks in different model families: Qwen2.5-1.5B/7B (Llama-style self-attention) and GPT-2-medium (Conv1D-style attention) on modular only, covering non-NeoX attention architectures.

Tasks. Two ICL rule-following tasks share a 4-shot surface form. Each prompt is a sequence of $K = 4$ demonstrations followed by a query, written as `f0,f1,Y ... f0q,f1q,`, where the model must emit the label for the query pair. Feature pairs $(f_0, f_1) \in \{0, \ldots, 7\}^2$ are sampled i.i.d. across demonstrations within a prompt; labels are $Y \in \{A, B\}$; a hidden rule $z \in \{0, 1, 2, 3\}$ is fixed within a prompt. A concrete *hierarchical* prompt with $z = 2$ (rule $f_0 + f_1 > 7 \Rightarrow$ label A) looks like:

`3,5,A 1,2,B 4,5,A 2,3,B 6,3,` $\rightarrow$ `A`

The two rule families differ in mechanism: *Hierarchical*: z selects one of $\{f_0 > 3,\ f_1 > 3,\ f_0 + f_1 > 7,\ f_0 = f_1\}$. *Modular*: z selects modulus $m \in \{2, 3, 5, 7\}$ (coprime, so no rule is a subset of another); the label is $\mathbf{1}[(f_0 + f_1) \bmod m = 0]$.

They share surface form but differ in rule type, so a head population that survives both is recruited by the rule, not the template.

Notation. Let (L, H) index a head, with $h_{(L,H)}(x) \in \mathbb{R}^{d_h}$ its output at the final query-token position before $W_O^{(L)}$. W_U is the unembedding matrix; γ is the final-LayerNorm gain; $r(x)$ is the pre-LN residual. We aggregate correct/incorrect-label probabilities over leading-space variants (e.g. {`A', `A'} for label A) and define the unembed contrast $u(x) = W_U[y_+] - W_U[y_-]$, the readout $\Delta\ell(x) = \log p(y_+ \mid x) - \log p(y_- \mid x)$, and the per-pair *swing* $\Delta\ell(x_c) - \Delta\ell(x_r)$ between correct- and rule-flipped prompts.

Three prompt sets are sampled per (model, task) cell, with seeds fixed in advance: 192 balanced *discovery* prompts (used for refined-DLA, seed 42), 200 paired *path-patching* prompts (seed 43), and 500 paired *evaluation* prompts (seed 44) used for group lesions, cross-task transfer, and specificity. Logits are aggregated over the leading-space label variants above. Full prompt examples and the format-prefix string are in App. A.2.

3 Methodology

The pipeline runs independently per (model, task) cell: *discovery* screens candidate FV heads (§3.1–3.2), group lesions *arbitrate* the cancellation signature against two nulls (§3.3), and a *mechanism* and *robustness* battery (§3.4–3.5) tests that the writer/canceller labels are intrinsic rather than artefacts of the discovery data. All thresholds and the 2×3 Pythia grid are fixed in advance; statistical conventions (paired-prompt percentile bootstraps with $B = 10{,}000$, BH-FDR (Benjamini and Hochberg, 1995) at $q = 0.10$) and a full per-test inventory are in App. A.1.

3.1 Refined Direct Logit Attribution

We attribute $\Delta\ell$ to head (L, H) by projecting its post-W_O contribution onto the unembed difference under a frozen-LN linearisation (Geva et al., 2022; Wang et al., 2023):

$$\widehat{\mathrm{DLA}}_{(L,H)}(x) = u(x)^\top \left(\gamma \odot \frac{c_{(L,H)}(x) - \overline{c_{(L,H)}}(x)}{\sigma(r(x)) + \varepsilon} \right), \tag{1}$$

where $c_{(L,H)}$ is the head's post-W_O contribution to the residual stream, $u(x) = W_U[y_+] - W_U[y_-]$ is the correct-minus-incorrect unembed direction, $\overline{c_{(L,H)}}$ is the discovery-batch mean (the LN linearisation reference), and $\sigma(r(x))$, γ are the standard LN scale and gain.[2] Per-head signed means are computed on the discovery batch (seed 42, $N = 192$); a 20-seed label-permutation null with BH-FDR at $q = 0.10$ ($L \geq 1$) defines the FDR-passing set. The final FV-candidate set is its union with the top-K heads by mean $|\widehat{\mathrm{DLA}}|$, ensuring a tractable shortlist when the FDR-passing set is sparse (exact K rule in App. A.2).

3.2 Path Patching with Ratio-of-Means

We form 200 paired prompts (x_c, x_r) sharing query (f_{0q}, f_{1q}) but differing in rule (Wang et al., 2023; Goldowsky-Dill et al., 2023). After filtering pairs with $|\Delta\ell^{\text{full}}| < 0.5$, the ratio-of-means

$$\mathrm{direct}_{(L,H)} = \frac{\mathbb{E}\left[\Delta\ell_i^{\text{direct}}\right]}{\mathbb{E}\left[|\Delta\ell_i^{\text{full}}|\right]} \tag{2}$$

(and analogously *total*, *indirect*) is bootstrapped ($B = 10{,}000$) for 95% CIs and a two-sided p_{direct}. PP p-values are BH-corrected at $q = 0.10$. The $\pm 5\%$ gate is fixed in advance; sensitivity to gate width ($\pm 3\%, \pm 7\%, \pm 10\%$) is reported in App. A.3. We also impose the two-hop ordering $L_W < L_C$ for any (W, C) pair entered into edge-level analysis.

Operational definition of writers and cancellers. $\mathcal{F}$ is the sign-preserving FV candidate set from refined-DLA screening (§3.1): unlike Todd et al. (2024)'s top-K by signed AIE,[3] our criterion admits heads with both signs of direct effect. A head $(L, H) \in \mathcal{F}$ is a *writer* if its direct PP effect is $\geq +5\%$ of the swing, a *canceller* if $\leq -5\%$, and *weak* otherwise; only writers $\mathcal{W}$ and cancellers $\mathcal{C}$ are entered into the group lesion (weak heads are listed in $|\mathcal{F}|$ for context but excluded from ablations). Per-cell counts appear in Table 3.

Asymmetric overlap with Todd's magnitude-only set. Todd et al. (2024)'s mean-ablation top-20 captures only one of the two populations per task family (Table 2; full 6-cell breakdown in App. C.2). Magnitude-only ranking systematically surfaces whichever sub-population locally dominates and obscures the other; only sign-aware screening recovers both.

[2] $\varepsilon = 10^{-5}$ is the LN numerical stabiliser. The additive identity $\sum_{L,H} \widehat{\mathrm{DLA}} + (\text{MLP}, \text{embed}, \text{biases}) = \Delta\ell$ is audited end-to-end (rel. tol. 5%).

[3] Cancellers have positive but smaller AIE than writers and are systematically missed by the AIE-ranked top-K; we use "magnitude-only" as shorthand for this selection rule.

Table 2: **Todd top-20 overlap, pooled across 6 main cells.** Magnitude-only ranking surfaces opposite populations on the two task families.

task	W in Todd-20	C in Todd-20
hierarchical	1/27 (4%)	**9/14 (64%)**
modular	**16/27 (59%)**	1/13 (8%)

3.3 Group Canceller Lesion: Primary Causal Test

For each ablation strategy $s \in \{\text{zero}, \text{mean}\}$ on $n_{\text{eval}} = 500$ paired prompts we ablate (a) writers $\mathcal{W}$, (b) cancellers $\mathcal{C}$, and (c) their union, bootstrapping 95% CI per group ($B = 10{,}000$).

The canonical cancellation-circuit verdict requires four conditions:

(i) $|\Delta\ell_{\mathcal{W}}| \geq 0.10$ with CI excluding 0,
(ii) $|\Delta\ell_{\mathcal{C}}| \geq 0.10$ with CI excluding 0,
(iii) $\Delta\ell_{\mathcal{W}}$ and $\Delta\ell_{\mathcal{C}}$ have opposite signs,
(iv) $|\Delta\ell_{\text{both}}| < |\Delta\ell_{\mathcal{W}}| + |\Delta\ell_{\mathcal{C}}|$ (joint attenuation).

Sub-canonical patterns (writer-dominant, canceller-dominant, opposite-signs-without-attenuation, and co-writer) are recorded so non-canonical outcomes remain interpretable.

Sign-shuffle null. Conditions (i)–(iii) are mechanically forced by sign-partitioning; only joint attenuation (iv) is diagnostic. We test it with a sign-shuffle null: for $n_{\text{perm}} = 10{,}000$ draws we randomly partition $\mathcal{W} \cup \mathcal{C}$ into two subsets matching $|\mathcal{W}|, |\mathcal{C}|$ and record the contrast $\Delta\ell^{\text{permW}} - \Delta\ell^{\text{permC}}$ under zero-strategy ablation. The empirical two-sided p is the fraction of permutations with $|\text{contrast}|$ above the observed, floored at $1/n_{\text{perm}} = 10^{-4}$; FDR control uses the empirical p (Gaussian back-stop in Table 8). A complementary $|\text{DLA}|$-matched head-randomised control confirms that the structural attenuation is FV-specific in magnitude (App. E.1).

3.4 QK Source Decomposition

Each head's last-position attention mass is partitioned into five buckets (BOS, format prefix, demonstration input, demonstration label, query input) and averaged over $n_{\text{prompts}} = 32$ eval prompts to produce a per-head bucket distribution. We report the per-bucket mean attention mass within $\mathcal{W}$ and $\mathcal{C}$ and their per-bucket difference (canceller minus writer) per cell. We ask whether the canceller-minus-writer sign is consistent across cells on the demonstration-label and format-prefix buckets.

3.5 Auxiliary tests

Six auxiliary tests support §4.4 (full specs in App. A.1, Table 6). *Split-half*: 5 random 50/50 prompt splits; a cell passes if $\geq \lceil n/2 \rceil$ splits reach canonical on the held-out half. *Cross-task transfer*: W/C labels from cell A applied to cell B's eval prompts under a relaxed (i)–(iii) variant (sign persistence, not the strict attenuation gate). *Within-layer OV null*: per (W, C) pair, ≥ 100 random same-layer non-FV pairs define a one-tailed cosine null; a cell passes if $\geq 30\%$ of pairs are significantly anti-aligned at $p \leq 0.05$. *TOST equivalence*: margin $\Delta = \sqrt{\text{Var}_{\text{hg}}}$ from the hypergeometric under random allocation, used for the induction-overlap rule-out. *Specificity*: zero-ablate each canceller on rule and random-token sequences; rule-specific iff $\mathbb{E}[\Delta\text{NLL}^{\text{rule}}]/\mathbb{E}[\Delta\text{NLL}^{\text{rand}}] \geq 5$. *Induction-head overlap*: top-10 heads by prefix-matching (Olsson et al., 2022) intersected with $\mathcal{W} \cup \mathcal{C}$, compared against the hypergeometric expectation.

4 Results

The FV head set has size $|\mathcal{F}| \in [12, 23]$ across the 6 main cells. Since conditions (i)–(iii) of the canonical verdict are mechanically forced by sign-partitioning, the load-bearing evidence is the canceller subgroup's causal effect (mean 5.2× larger than rank-nearest non-FV controls), the sign-shuffle null on joint attenuation, the structural separation of writers and cancellers, and the task-asymmetric overlap with Todd's top-K, developed in §4.1–4.5.

4.1 Cancellers suppress the correct-label logit

Ablating the canceller subgroup shifts the readout $+0.13$ to $+0.29$ nats *toward* the correct label in 6/6 cells; ablating writers shifts -0.25 to -0.99 nats *away* from it (Table 3; the 6 main cells in Figure 1). The four-condition canonical verdict (§3.3) holds in 11/12 outcomes across 6 cells × 2 ablation strategies (zero, mean); the single miss is (mod-1.4B, mean), where the canceller magnitude ($+0.09$) sits just below the 0.10 threshold. Table 4 reports the per-cell verdicts under six independent tests, so partial failures (cross-task hier→mod-1B, mean-strategy mod-1.4B, OV gate hier-1B + mod-1.4B) are surfaced rather than buried.

Magnitudes and budgets. Joint ablation attenuates the writer effect by mean 48% (range 21–100%). Writers outnumber cancellers 2:1 with

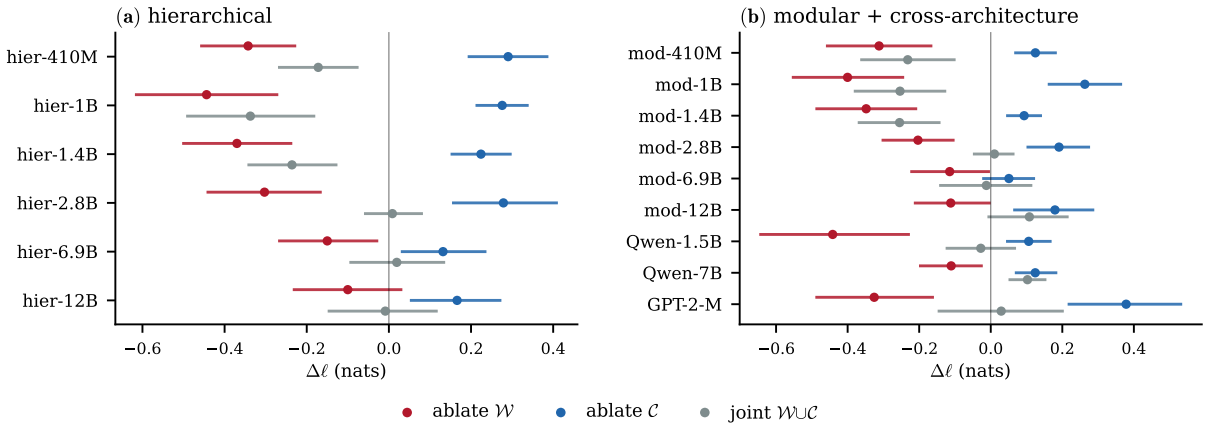

Figure 1: **Function-vector heads are two populations.** Mean-ablation logit shifts (BCa 95% CI) for writers, cancellers, and their joint, per cell. Writer ablation shifts the readout away from the correct label, canceller ablation toward it, and joint ablation attenuates. The four-condition canonical verdict holds in $13/15$ cells. *(a)* Hierarchical (6 Pythia cells, 410M–12B); *(b)* Modular + cross-architecture (9 cells: Pythia mod 410M–12B, Qwen2.5-$\{1.5, 7\}$B, GPT-2-medium).

$\sim 1.8\times$ larger per-head magnitudes; per-head canceller shift is uniform (CV 0.19 vs. 0.56). Joint $\mathcal{W} \cup \mathcal{C}$ ablation on mod-410M yields $+8.4$ pp accuracy (CI $[+2.3, +14.5]$; the only joint effect with CI excluding 0). Per-cell budgets in App. B.3.

Geometric independence. Writer and canceller mean OV directions are nearly orthogonal in the residual stream (perpendicular fraction mean 0.96, 95% CI $[0.91, 0.99]$; App. D.4). Cancellers therefore write to a near-orthogonal subspace while still exerting a negative direct effect on u, which lets the two populations co-exist at the same readout and explains why joint ablation is attenuated rather than fully additive.

The verdict survives a sign-shuffle null. Because the W/C sign definition mechanically forces sign-balance, the appropriate null is 10,000 random sign assignments *within* $\mathcal{W} \cup \mathcal{C}$ (zero-strategy ablation). The observed $|W - C|$ contrast exceeds the null 95th-percentile in $5/6$ cells, with empirical BH-$q \leq 0.014$ on those five; mod-1.4B is the single boundary case at $p_{\text{emp}} = 0.104$ ($p_{\text{Gauss}} = 0.088$, Table 8). We discuss this boundary case in Limitations.

A per-cell verdict matrix summarises six independent tests (zero 4c $6/6$, sign-shuffle $5/6$, split-half $6/6$, cross-task $5/6$); full breakdown in App. B.2. Per-cell magnitudes are reported in Table 3 (below, following §4.4).

4.2 Writers and cancellers read different prompt positions

QK-source decomposition rules out the hypothesis that cancellers are sign-flipped writer copies. Writers and cancellers read structurally different prompt positions, with the same direction of difference in $6/6$ cells on the two load-bearing buckets (Figure 2; per-cell numerics in App. C.3, Table 12). Writers concentrate 34–52% of attention mass on demonstration labels and 28–44% on BOS; cancellers shift 0.09–0.17 of mass off the label position and gain 0.08–0.18 on format-prefix tokens (15–29% of canceller mass vs. 4–17% for writers): cancellers are structurally distinct heads, not sign-flipped writer copies.

4.3 Mechanism is content-driven, not a sink artefact

The format-prefix pattern raises a sink confound: cancellers might be format sinks whose effect is an artefact of where attention lands rather than what they read. Per-source DLA bucket attribution rules this out. We split each canceller's direct effect by source bucket and label it *content* when its negative DLA is dominated by demonstration-content buckets, *sink* otherwise. 20 of 27 (74%) cancellers are content-driven, and they carry the main magnitude (-0.10 to -0.23 nats); the 7 sink heads cluster near zero (-0.02 to -0.09 nats) and reappear cross-task at fixed scale, consistent with format readers rather than rule-specific suppressors. Per-cell numerics in App. C.4, Table 13; an independent V-shuffle protocol on the strongest canceller per cell confirms content-driving in $6/6$. L11.H4 is fully content-

driven ($\sim 100\%$ demonstration-label sourced in both 410M cells); its single-head mechanism is analysed in §4.7.

4.4 Robustness and specificity

Stability. Split-half ($5 \times 50/50$ prompt splits) is verdict-stable on all 5 seeds in $6/6$ cells. Cross-task transfer (W/C labels from a source cell on a destination cell's eval prompts) is canonical in $5/6$ ordered pairs; $\mathcal{W} \cap \mathcal{W}$ Jaccard 0.45–0.70 and $\mathcal{C} \cap \mathcal{C}$ 0.43–0.60 (partial overlap, consistent with the task-conditional populations of §4.5); zero role flips at any scale (App. B.1).

Specificity. $19/27\,(70\%)$ cancellers are rule-specific (rule-NLL drop $\geq 5\times$ random-NLL drop; per-cell median ratio 4.8–62×, up to 151× for L11.H4 on mod-410M), trending up with scale (410M: $5/9$, 1B: $6/10$, 1.4B: $8/8$; Cochran–Armitage $p = 0.049$). No above-chance overlap with the induction top-10.

Null comparisons. Within-layer OV anti-alignment holds in $4/6$ cells (all $5/5$ same-layer pairs significant on mod-410M; App. B.4). Edge-level: real (W, C) pairs are significantly stronger than random layer-respecting pairs in $6/6$ cells (Mann–Whitney BH-$q \leq 1.4 \times 10^{-11}$; App. B.4).

Layer geometry. Cancellers cluster in early-to-mid layers and $23/27$ have an upstream writer in their own cell, ruling out late-stage suppression (Figure 3). Maximum writer layer reaches $L =$ 21–23 in 24-layer Pythia; canceller layers stop at $L = 19$.

Named-circuit rule-outs. Three named circuits are tested as universal explanations of cancellation and ruled out (full numerics in App. C.7). *Rank-1 copy-suppression* (McDougall et al., 2023): $0/27$ cancellers exceed the top-1 OV Frobenius share of 0.5 (observed max 7.5%). *V-cascade* (our shorthand for V-composition in the S-Inhibition pathway of Wang et al., 2023): ruled out in $4/6$ cells ($2/6$ active counterfactual on the dominant upstream writer, $2/6$ vacuous with no upstream writer at $L < L_{\mathcal{C}}$); a weak partial residual cascade remains in the other $2/6$. *V-shuffle*: permuting the head's value vectors across source positions collapses cancellation by 86–113% in $6/6$ cells, marking the mechanism as content-driven rather than position-driven. With the W vs. C OV near-orthogonality reported above (App. D.4, Table 21), no single named template explains the canceller population.

4.5 Cross-template transfer to vocabulary ICL

We transfer Pythia-410M W/C labels to two vocabulary-ICL templates (antonym, country-capital): canonical in $3/4$ pairs (App. C.6). The $1/4$ "failure" (hier→antonym) flips cancellers to super-writers ($\Delta\ell = -1.56$ nats), the exact pattern copy-suppression (McDougall et al., 2023) predicts when the demonstrated label *is* the answer. The W/C role is therefore task-conditional rather than intrinsic.

4.6 Scale and cross-architecture spot checks

The verdict extends to the rest of the Pythia ladder (2.8B/6.9B/12B) and three cross-architecture spot checks (Qwen2.5-{1.5, 7}B, GPT-2-medium; modular only, App. B.5, Table 9). All 9 extension cells pass the signed-direction condition; 6.9B mod and 12B hier shift to PARTIAL (one CI grazes 0). Combined with the main cells: $13/15$ canonical, $2/15$ partial, $0/15$ sign-flipped.

4.7 Mechanistic case study of L 11.H 4

The per-source decomposition of §4.3 identifies L11.H4 as the dominant canceller in both Pythia-410M rule cells, with $\sim 100\%$ of its negative DLA sourced from demonstration-label tokens. We narrow its mechanism with four interventions (numerics in App. C.8).

Content reader (V-shuffle + OV spectrum). Permuting the head's value vectors across source positions collapses cancellation by $\sim 82\%$ (hier) / $\sim 53\%$ (mod): the head reads source content rather than a fixed direction. The OV singular spectrum of $W_O W_V$ is consistent (top-1 Frobenius share 2.8%, top-10 25%, no rank-1 plateau), ruling out rank-1 copy-suppression (McDougall et al., 2023).

V-composition. Ablating the dominant upstream writer L10.H9 leaves L11.H4's DLA unchanged, ruling out S-inhibition V-cascade (Wang et al., 2023).

Cross-template sign-flip. L11.H4's solo ablation flips sign across templates (Figure 4): canceller on hier and mod, writer on antonym, null on country-capital. The role is task-conditional, consistent with "suppress what was demonstrated" rather than unconditional logit suppression.

L11.H4 is therefore a *distributed-bias content reader*, distinct from both standard templates. We

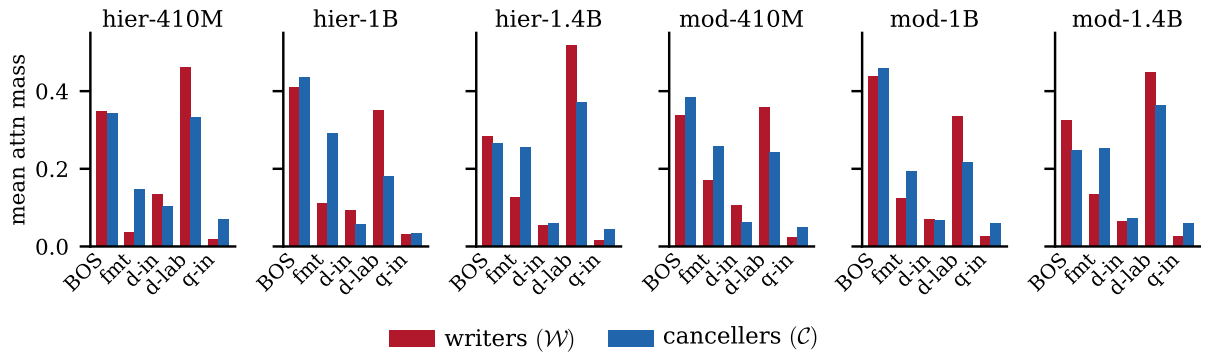

Figure 2: **Cancellers re-route attention from labels to format.** Mean per-head attention mass at the readout token for writers (red) vs cancellers (blue), per bucket, one panel per cell. Writers concentrate on demonstration labels; cancellers shift mass to format-prefix tokens; the direction of difference is consistent in $6/6$ cells (per-cell mass in App. C.3, Table 12).

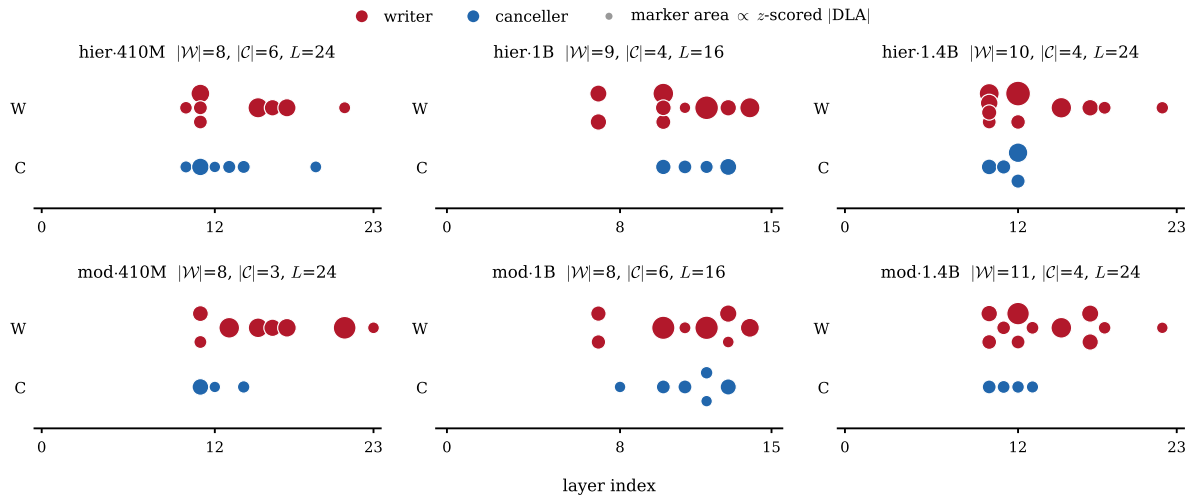

Figure 3: **Layer distribution of writers and cancellers.** Writers (red) and cancellers (blue) by layer, one panel per cell; marker area $\propto$ z-scored $|\mathrm{DLA}|$. Cancellers cluster in early-to-mid layers and $23/27$ have an upstream writer (a writer at a lower layer in the same cell), ruling out late-stage suppression.

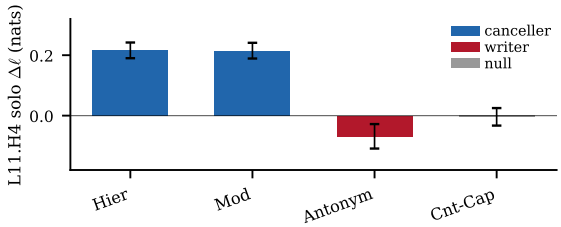

Figure 4: **L11.H4 case study: cross-template sign-flip.** Solo-ablation $\Delta\ell$ on the four templates (paired bootstrap 95% CI). Blue = canceller, red = writer, grey = null. The head's address transfers across all four templates but its role is task-conditional.

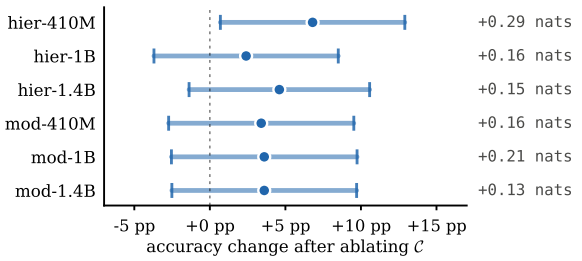

Figure 5: **Canceller ablation: actionable consequence.** Per-cell accuracy delta after zero-ablating $\mathcal{C}$, with paired-bootstrap 95% CI; logit-shift magnitude annotated on the right. Sign is positive in $6/6$ main cells; CI excludes 0 on hier-410M. App. D.3 Table 20.

do not extrapolate this single-head characterisation to the full canceller population.

5 Discussion

The FV literature treats high-magnitude heads as one functional class; we split them into opposing sub-circuits at the same readout. Five independent lines support this: sign-shuffle rejects homogeneity ($5/6$ cells), QK reads differ ($6/6$), OV directions are near-orthogonal, $23/27$ cancellers sit downstream of a writer, and L11.H4 flips

Table 3: **Group lesion (zero strategy).** Writer ablation shifts the readout -0.25 to -0.99 nats away from the correct label; canceller ablation $+0.13$ to $+0.29$ nats toward it; joint attenuates the writer effect by mean 48% (median 41%). Four-condition canonical (§3.3) holds $6/6$ zero-strategy, $11/12$ mean-strategy (mod-1.4B mean canceller below 0.10; App. B.2). $|\mathcal{F}|$ is the FV candidate set; only $\mathcal{W} \cup \mathcal{C}$ are ablated (weak heads with $|\text{direct}\%| < 5\%$ excluded).

Task	Model	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{W}\rvert$	$\lvert\mathcal{C}\rvert$	$\Delta\ell_{\mathcal{W}}$	$\Delta\ell_{\mathcal{C}}$	$\Delta\ell_{\text{both}}$	verdict
hier	410M	22	8	6	-0.25	$+0.29$	$+0.00$	canonical
hier	1B	23	9	4	-0.56	$+0.16$	-0.43	canonical
hier	1.4B	19	10	4	-0.32	$+0.15$	-0.13	canonical
mod	410M	12	8	3	-0.39	$+0.16$	-0.14	canonical
mod	1B	16	8	6	-0.99	$+0.21$	-0.78	canonical
mod	1.4B	19	11	4	-0.49	$+0.13$	-0.38	canonical

Table 4: **Per-cell verdict matrix across six independent tests.** ✓ pass, **x** fail, ∘ boundary. The cancellation signature holds in $6/6$ cells under the strongest test (zero-strategy 4-condition); secondary tests rule out role-flipping, induction overlap, and writer/canceller equivalence on 4–6 of 6 cells. FWER analysis in App. B.2.

Cell	Zero 4c	Mean 4c	Sign-shuf q	Split-half	Cross-task 4c	OV gate
hier-410M	✓	✓	✓	✓	✓	✓
hier-1B	✓	✓	✓	✓	∘ W-dom	**x**
hier-1.4B	✓	✓	✓	✓	✓	✓
mod-410M	✓	✓	✓	✓	✓	✓
mod-1B	✓	✓	✓	✓	✓	✓
mod-1.4B	✓	∘ W-dom	∘ (q=0.104)	✓	✓	**x**
pass	**6/6**	5/6	5/6	6/6	5/6	4/6

Table 5: **Sign-opposed attention in prior work vs. this paper.** We lift the phenomenon from single-task hand-built circuits to a structural FV-population property.

	Scope	Mechanism	Detection
IOI neg. name-mover (Wang et al., 2023)	single task (IOI)	hand-built circuit	task-specific analysis
GPT-2 L10.H7 (McDougall et al., 2023)	single head, copy task	rank-1 copy-suppression	task-specific analysis
W/C population (this paper)	population, $13/15$ cells, 3 architectures	distributed; no single template fits	sign-aware DLA + path patching

canceller→writer on antonym (the prediction of "suppress what was demonstrated").

Connection to FV identification. Todd's top-K is task-conditionally biased toward the locally-dominant sub-population (hier C 64% vs. W 4%; mod W 59% vs. C 8%), so any aggregation built on this ranking inherits a mixture of the two opposing roles (Table 5; App. C.5).

Ablation, transplant, and additive steering are different operations. The W/C identification has different downstream consequences depending on the intervention. We separate three operations the literature often blurs.

Ablation (zero out canceller outputs during the forward pass) raises the correct logit by $+0.13$–0.29 nats and shifts accuracy directionally $+2$–7 pp in $6/6$ main cells (Figure 5; App. D.3).

Transplant (replace head outputs with their task-means) goes the *opposite* direction: writer-only underperforms FV-mean in accuracy in $6/6$ cells (App. D.2), since FV-mean incidentally regularises cancellers' per-prompt outputs while writer-only leaves cancellers active.

Additive steering (inject a mean vector at the readout token) gives a larger absolute held-out logit shift for $v_{\mathcal{W}}$ than for v_{FV} in $6/6$ main cells (App. D.1), but α is tuned on logit shift; under norm-matched injection with held-out α the partition effect on accuracy is not established (see Limitations). Population-level identification thus says which heads to intervene on; it does not imply that an aggregated steering vector is the right operation, since cancellers' per-prompt behaviour makes writer-only lose to FV-mean transplant under any α-free criterion (App. D.2).

6 Conclusion

The W/C split is a structural property of FV head sets, not a prescription: it is invisible to magnitude-only ranking and is inherited by any downstream FV-based aggregation.

Limitations

Task scope. The primary analysis uses two synthetic ICL rule-following families (hierarchical, modular) with shared 4-shot surface form. Cross-template transfer extends to two vocabulary-ICL templates (antonym, country-capital). All tasks are short, closed-label, and in-context-learnable from ≤ 5 demonstrations; we do not test long-form generation, open-ended NL classification at scale, or instruction-following.

Model coverage. Six Pythia main cells (410M, 1B, 1.4B $\times$ hier, mod) carry the full statistical pipeline. Extension cells (Pythia-2.8B/6.9B/12B; Qwen2.5-$\{1.5, 7\}$B; GPT-2-medium) use a lighter protocol with $n_{\text{eval}} = 200$. We do not test Meta's Llama series or instruction-tuned variants (Qwen2.5 shares the Llama-style attention architecture but is a separate model family). One main cell (mod-1.4B) is the boundary case on sign-shuffle ($p_{\text{emp}} = 0.104$); two extension cells (6.9B mod, 12B hier) shift to PARTIAL (one CI grazes 0).

Steering claims are limited. We make no positive claim that $v_{\mathcal{W}}$ improves practical steering. The only α-free comparison (transplant, App. D.2) shows writer-only is at best tied with FV-mean on accuracy, with the gap predicted by Todd's top-K canceller fraction. Held-out α on the 6 main cells (App. D.1) is on logit shift only; accuracy under held-out α with norm-matched injection and path-patching W/C across all 15 cells is left to future work.

Mechanistic case study has narrow generality. L11.H4 (Pythia-410M) is the dominant canceller in both rule cells, but its "distributed-bias content reader" characterisation is established only at this one head. We do not claim the case-study mechanism generalises to the full 27 canceller (cell, head) pairs; we do show population-wide that rank-1 copy-suppression holds for $0/27$ pairs, V-cascade is ruled out in $4/6$ cells ($2/6$ active, $2/6$ vacuous), and per-source DLA is content-dominated for $20/27$ (74%) cancellers.

Statistical scope. Cross-cell aggregation uses Holm-Bonferroni FWER at $\alpha = 0.05$ over the 6 Pythia main cells (5/6 reject). The verdict matrix (six independent tests, App. B.2) reports per-test BH-FDR within each test family but not family-wise across the matrix as a whole, so the matrix is presented as a per-test pass-rate summary rather than a single composite test.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing

model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2023. Copy suppression: Comprehensively understanding an attention head. In *Workshop on Attributing Model Behavior at Scale (NeurIPS)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *BlackBoxNLP Workshop at EMNLP*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*.

Alexandre Variengien and Eric Winsor. 2023. Look before you leap: A universal emergent decomposition of retrieval tasks in language models. *arXiv preprint arXiv:2312.10091*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.

A Methods and reproducibility

A.1 Statistical inventory

Every test the paper relies on. Bootstraps are paired-prompt percentile bootstraps with $B = 10{,}000$ unless noted; BH-FDR at $q = 0.10$ on raw p-values; display p-values floored at 1×10^{-6}.

A.2 Reproducibility

The anonymized implementation and reproducibility package are available at `https://anonymous.4open.science/r/function-vectors-two-populations-A3C6/`.

Models. Pythia checkpoints `EleutherAI/pythia-410m-deduped`, `pythia-1b-deduped`, `pythia-1.4b-deduped`; final training step, `fp32`, `attn_implementation=eager`.

Prompts. 4-shot template `f0,f1,Y...f0q,f1q,` with $f_0, f_1 \sim \mathrm{Unif}\{0,\ldots,7\}$, $z \sim \mathrm{Unif}\{0,1,2,3\}$; format-prefix is the literal `", "`. Per cell: 192 DLA prompts (seed 42), 200 PP prompts (seed 43), 500 eval prompts (seed 44). Logit aggregation: $\log \sum p$ over leading-space variants {`' A'`, `' A'`} vs. {`' B'`, `' B'`}.

FV head set. Sign-preserving refined-DLA passes at $q = 0.10$ unioned with top-K by $|\Delta\ell|$, $K = \mathrm{clip}(2n_{\mathrm{FDR}}, 8, 50)$, intersected with PP gate $|\mathrm{direct}\%| \geq 5\%$ to yield $\mathcal{F}$.

Path patching. Direct-path receiver-edge ablation of Wang et al. (2023); filter $|\Delta\ell^{\mathrm{full}}| < 0.5$.

QK bucketing. Five deterministic bucket spans: BOS = $\{0\}$; format prefix (whitespace/punctuation between demos); demonstration input ((f_0, f_1) spans); demonstration label (Y tokens); query input ((f_{0q}, f_{1q})).

Edge-level null. 50 random head pairs (L_a, L_b) with $L_a < L_b$ from the full head population per cell.

Table 6: **Statistical inventory.**

Test	Null	Statistic	p-floor	Multcomp
Refined-DLA	20 label permutations	per-head signed mean	0.05	BH q=0.10
Path patching	B=10,000 paired bootstrap	ratio-of-means / two-sided	$1/B$	BH q=0.10
Group lesion	B=10,000 paired bootstrap	4-condition CI gate	$1/B$	—
Sign-shuffle null	n_{perm}=10,000 in-pool sign	$\lvert W-C\rvert$ contrast	10^{-4}	BH q=0.10
Within-layer OV null	≥ 100 same-layer non-(W∪C) pairs	cosine $\leq$ obs	0.01	—
Edge-null ACDC	50 random layer-respecting pairs	MWU on $\lvert\Delta\ell\rvert$	1×10^{-6}	BH q=0.10
Specificity ratio	B=5,000 canceller bootstrap	rule/random NLL ratio	—	—
Induction TOST	hypergeometric E, margin $\sqrt{\text{Var}_{\text{hg}}}$	two one-sided test	—	—
Cross-task verdict	sign+magnitude on dst eval (relaxed gate, §3.5)	4-condition variant	—	—
Split-half	n_{split}=5 random 50/50 splits	verdict-stable in $\geq n/2$ splits	—	—
Per-source DLA	per-token-source bucket attribution	|sink| vs. |content| DLA share	—	—
Vocab transfer	B=2,000 paired bootstrap, n=100	$\Delta\ell^{\text{correct}}$ shift / sign agreement	$1/B$	—
Cochran–Armitage	rule-specific rate by ordinal scale	trend z, two-sided	—	—

Statistics. Bootstrap $B = 10{,}000$; sign-shuffle $n_{\text{perm}} = 10{,}000$; refined-DLA 20-seed permutation null; specificity $B = 5{,}000$. Thresholds ($\pm5\%$ direct effect, 0.10 nat floor, $5\times$ rule/random ratio) fixed pre-analysis.

Computational budget and software. The released reproduction package is a single-notebook pipeline with modular Python stages for prompt generation, refined DLA, path patching, group lesion, QK-source attribution, rule-outs, transfer, scale extension, case study, and aggregation. It uses `torch`, `transformer_lens`, `transformers`, `numpy`, `scipy`, and `matplotlib`. A full regeneration of all released results takes approximately 18 hours on one NVIDIA A100 GPU; discovery and path patching dominate the runtime, with Pythia-6.9B and Pythia-12B accounting for most of the extension-cell cost. The released `results/` directory and `extracted_numbers.json` provide the seeded outputs used for all paper-cited numbers, so rerunning the notebook is not required to inspect the reported results.

Artifacts, licenses, and intended use. We use and release scientific artifacts: deterministic prompt generators, analysis code, figure scripts, per-cell JSON results, auxiliary result payloads, and an aggregation file containing every paper-cited number. The released code is MIT-licensed. Pretrained model artifacts retain their original terms: Pythia and Qwen2.5 under Apache 2.0, GPT-2 under MIT. The intended use of the released package is research reproduction and inspection of the mechanistic claims in this paper; it is not intended as a deployment or safety-control system.

Data, human subjects, and risk profile. No human subjects or annotators are used. The rule-learning prompts are synthetic strings generated from integer feature pairs and deterministic label rules; the vocabulary-transfer checks use fixed common-word pairs such as antonyms and country–capital pairs. We did not collect private, personal, or newly authored human text data. The main foreseeable risk is technical over-interpretation or misuse of head-level interventions as a general model-control method; accordingly, the paper limits its claims to short closed-label ICL settings and explicitly separates population identification from practical steering.

A.3 Threshold sensitivity

The primary $11/12$ canonical result is robust to $\pm50\%$ perturbation of any single threshold ($\pm5\%$ direct-effect gate, 0.10-nat magnitude floor, $5\times$ rule/random ratio); the single boundary case (mod-1.4B mean, canceller $+0.09$) flips to canonical at floors ≤ 0.09.

B Discovery and validation

B.1 Refined DLA and path patching: per-cell tables

Full per-head refined-DLA and PP tables released as JSONs; cell sizes in Table 3. End-to-end sum-to-$\Delta\ell$ reconstruction passes (rel. tol. 5%) in $9/9$ tests for $6/6$ cells.

B.2 Group lesion: zero *and* mean strategies

Table 7 mirrors Table 3 under mean ablation: $5/6$ canonical, mod-1.4B writer-dominant (canceller $+0.09$ just below the 0.10 threshold).

Table 7: **Group lesion (mean strategy).** $11/12$ canonical across both strategies; the only miss is mod-1.4B mean.

Task	Model	$\Delta\ell_{\mathcal{W}}$	$\Delta\ell_{\mathcal{C}}$	$\Delta\ell_{\text{both}}$	verdict
hier	410M	-0.34	$+0.29$	-0.17	canonical
hier	1B	-0.44	$+0.28$	-0.34	canonical
hier	1.4B	-0.37	$+0.22$	-0.24	canonical
mod	410M	-0.31	$+0.13$	-0.23	canonical
mod	1B	-0.40	$+0.26$	-0.25	canonical
mod	1.4B	-0.35	$+0.09$	-0.26	writer-dominant

FV-set sign-shuffle null per-cell. Table 8: observed $|W-C|$ contrast vs. in-pool sign-shuffle null with Gaussian back-stop and BH-q. $5/6$ cells reach $q_{\text{emp}} \leq 0.014$; mod-1.4B is the boundary ($p_{\text{emp}} = 0.104$, $p_{\text{Gauss}} = 0.088$).

Per-cell verdict matrix: FWER. Body Table 4 reports the per-cell verdict matrix. Primary FWER: Holm–Bonferroni on sign-shuffle p-values $\{0.001, 0.001, 0.001, 0.014, 0.014, 0.104\}$ at family-wise $\alpha = 0.05$ rejects $5/6$ cells; mod-1.4B fails the final threshold. BH-FDR $q = 0.10$ yields the same conclusion. Other matrix entries are conjunctive threshold tests at fixed gates and reported descriptively.

B.3 Asymmetric budgets and per-task breakdown

Per-head $|\text{direct}\%|$: $\mathcal{W} = 20.3\%$ ($n = 54$), $\mathcal{C} = 11.2\%$ ($n = 27$); per-head canceller shift is more uniform than writer (CV 0.19 vs. 0.56; canceller budget ~ 0.04 nats/head, task-independent). Modular elicits $\sim 1.7\times$ larger writer magnitudes than hierarchical ($|\Delta\bar{\ell}_{\mathcal{W}}| = 0.62$ vs. 0.38); cancellers not correspondingly amplified (0.17 vs. 0.20), so joint attenuation is lower on modular (36% vs. 61%).

B.4 ACDC edge ablation: pair-by-pair

Real-vs-null per cell: Mann–Whitney BH-FDR pass $6/6$ at $q \leq 1.4 \times 10^{-11}$, rank-biserial $r \in [0.60, 0.97]$. Fisher canonical-rate BH-FDR pass: $1/6$ (mod-410M, $q = 0.0016$; others raw $p \in [0.10, 0.21]$). Magnitude dominance is uniform; canonical-edge enrichment is task- and size-specific.

B.5 Scale and cross-architecture spot checks

C Mechanism rule-outs

C.1 Specificity controls: per-cell tables

Table 10: per-cell rule-specific / indeterminate / generally-important breakdown for 27 cancellers and the top-10 induction overlap. Rule-specific rate sharpens with scale; no above-chance induction overlap. TOST equivalence margin is $\sqrt{\text{Var}_{\text{hg}}} \approx 0.5$–$0.7$ heads, so the test has limited power for 1–2 head intersections; the induction-distinction is also supported by QK-source attribution (§4.2).

Highest-specificity cancellers. L11.H4 ($115\times$ hier-410M, $151\times$ mod-410M), L13.H1 ($109\times$ hier-410M), L11.H2 ($82\times$ mod-1.4B).

C.2 FV-overlap controls

Intersection of $\mathcal{F}$ with Todd et al. (2024)'s mean-ablation top-K at $K \in \{5, 10, 15, 20\}$ in Table 11. Total Jaccard 0.06–0.25 at top-20. The W vs. C asymmetry is task-dependent: hierarchical concentrates $\mathcal{C}$ in Todd's top-20 ($9/14$ vs. writer $1/27$); modular reverses ($16/27$ writers vs. $1/13$ cancellers).

C.3 QK source distribution: per-cell attention mass

Per-cell mean attention mass by group and source bucket in Table 12. In $6/6$ cells, demo-label mass drops from W to C (mean -0.13) and format-prefix rises (mean $+0.13$); other buckets shift by ≤ 0.05.

C.4 Per-source DLA decomposition: cancellers vs. sinks

Per-canceller DLA attributed to source buckets via $\sum_{\text{pos}} \alpha_{\text{pos}} W_O W_V r_{\text{pos}}$ projected onto $u(x)$. Per-cell summary in Table 13; full JSON released.

Named-head detail. L11.H4: $\sim 100\%$ demo-label sourced in both cells.

C.5 Comparison to GPT-2 copy-suppression

Head-level detail behind body Table 5. Three load-bearing differences distinguish L11.H4 (this work) from GPT-2 L10.H7 (McDougall et al., 2023). *QK signature:* L10.H7 attends to copies of the query token; L11.H4's attention is invariant across rule tasks but its DLA flips sign across templates (§4.7). *OV map:* L10.H7 has a rank-1 promote/suppress structure; L11.H4's $W_O W_V$

Table 8: **FV-set sign-shuffle null with Gaussian back-stop ($n_{\mathrm{perm}} = 10{,}000$).** Empirical and Gaussian p-values; BH-corrected q-values across the 6-cell family at $q = 0.10$. Empirical p floored at $1/n_{\mathrm{perm}} = 10^{-4}$.

Cell	observed	null p_{50}	null p_{95}	z	emp p	Gauss p	q_{emp}	q_{Gauss}
hier-410M	-0.435	-0.01	$+0.24$	-2.79	0.0024	5.2×10^{-3}	0.007	0.016
hier-1B	-0.607	-0.10	$+0.24$	-2.47	0.0047	1.3×10^{-2}	0.007	0.027
hier-1.4B	-0.275	-0.04	$+0.13$	-2.32	0.0035	2.0×10^{-2}	0.007	0.031
mod-410M	-0.361	-0.04	$+0.17$	-2.08	0.0116	3.7×10^{-2}	0.014	0.045
mod-1B	-0.683	-0.03	$+0.35$	-2.87	0.0010	4.1×10^{-3}	0.006	0.016
mod-1.4B	-0.223	$+0.01$	$+0.22$	-1.71	0.1040	8.8×10^{-2}	0.104	0.088

Table 9: **Verdict across architectures and scale.** Group-lesion W / C / joint logit shifts (mean strategy, $n_{\mathrm{eval}} = 200$). $W < 0$, $C > 0$ in all 9 extension cells; combined with the 6 main cells: $13/15$ canonical, $2/15$ partial (signed direction holds, one CI grazes 0).

Family / Arch	Size	Task	W_{shift}	C_{shift}	joint	verdict
Pythia (NeoX)	2.8B	hier	-0.30	$+0.28$	$+0.01$	canonical
Pythia	2.8B	mod	-0.20	$+0.19$	$+0.01$	canonical
Pythia	6.9B	hier	-0.15	$+0.13$	$+0.02$	canonical
Pythia	6.9B	mod	-0.11	$+0.05$	-0.01	partial
Pythia	12B	hier	-0.10	$+0.17$	-0.01	partial
Pythia	12B	mod	-0.11	$+0.18$	$+0.11$	canonical
Qwen2.5 (Llama-style)	1.5B	mod	-0.44	$+0.11$	-0.03	canonical
Qwen2.5	7B	mod	-0.11	$+0.12$	$+0.10$	canonical
GPT-2-medium (Conv1D)	355M	mod	-0.33	$+0.38$	$+0.03$	canonical

Table 10: **Per-cell specificity and induction-overlap controls.** *rule-spec*: cancellers with rule/random NLL ratio ≥ 5. *ind. ovlp.*: $|\mathcal{F} \cap \text{top-10}_{\text{ind}}|/|\mathcal{F}|$ (intersection of FV head set with top-10 induction heads).

Cell	$\lvert\mathcal{C}\rvert$	rule-spec	indet	gen-imp	ind. ovlp.
hier-410M	6	4	0	2	$1/22$
hier-1B	4	3	1	0	$3/23$
hier-1.4B	4	4	0	0	$0/19$
mod-410M	3	1	1	1	$0/12$
mod-1B	6	3	2	1	$1/16$
mod-1.4B	4	4	0	0	$0/19$
total	27	**19** (70%)	4	4	—

Table 11: **Full 6-cell FV-overlap.** W/C set from the PP-validated set used throughout the paper; FV top-K from Todd's mean-ablation identification on the same task. The W- vs C-asymmetry column shows W-in-FV / C-in-FV at $K = 20$; `(C-heavy)` marks cells where cancellers concentrate in Todd's top-20, `(W-heavy)` marks the opposite.

Cell	$\lvert\mathcal{W}\rvert$	$\lvert\mathcal{C}\rvert$	Jaccard	W in FV-20	C in FV-20	pattern
hier-410M	8	6	0.06	0	**2**	C-heavy
hier-1B	9	4	0.18	1	**4**	C-heavy
hier-1.4B	10	4	0.10	0	**3**	C-heavy
mod-410M	8	3	0.19	**5**	0	W-heavy
mod-1B	8	6	0.17	**5**	0	W-heavy
mod-1.4B	11	4	0.25	**6**	1	W-heavy
hier total	27	14	—	**1** (4%)	**9** (64%)	—
mod total	27	13	—	**16** (59%)	**1** (8%)	—

top-1 Frobenius share is 2.8% with no rank-1 plateau (App. C.8). *Negation:* L10.H7's behaviour under negation is not reported; L11.H4 flips canceller→writer on antonym, the expected behaviour of a "suppress what was demonstrated" mechanism. We do not yet test whether L11.H4 exhibits L10.H7's full copy-suppression signature on copy-task contexts; this is left as future work.

Table 12: **Per-cell mean attention mass at the final query token, by W/C group and source bucket.** Each row is one (group, cell). n is the number of heads in the group. demo-lab/format-prefix differences are the load-bearing buckets of Figure 2.

Cell	Grp	n	BOS	format	d-in	d-lab	q-in
hier-410M	W	8	0.35	0.04	0.14	0.46	0.02
	C	6	0.34	0.15	0.11	0.33	0.07
hier-1B	W	9	0.41	0.11	0.09	0.35	0.03
	C	4	0.44	0.29	0.06	0.18	0.03
hier-1.4B	W	10	0.28	0.13	0.05	0.52	0.02
	C	4	0.27	0.26	0.06	0.37	0.05
mod-410M	W	8	0.34	0.17	0.11	0.36	0.02
	C	3	0.38	0.26	0.06	0.24	0.05
mod-1B	W	8	0.44	0.12	0.07	0.34	0.03
	C	6	0.46	0.20	0.07	0.22	0.06
mod-1.4B	W	11	0.32	0.13	0.06	0.45	0.03
	C	4	0.25	0.25	0.07	0.36	0.06

Table 13: **Per-cell content vs. sink decomposition of cancellers.** A canceller is content-driven if $|\text{BOS} + \text{fmt}| < |\text{demo}_{\text{in}} + \text{demo}_{\text{lab}} + \text{query}_{\text{in}}|$ in DLA contribution, sink-driven otherwise. $\sum \Delta\ell$ sums total per-head DLA over each sub-population. Content cancellers carry the primary magnitude in $5/6$ cells; the hier-1B exception splits $2/4$ heads as sink-classified. Stable cross-task sink heads: L14.H3 (both 410M), L10.H2 (both 1B), L11.H2 (both 1.4B).

Cell	content $\lvert\mathcal{C}\rvert$	sink $\lvert\mathcal{C}\rvert$	content $\sum \Delta\ell$	sink $\sum \Delta\ell$
hier-410M	5	1 (L14.H3)	-0.38	-0.03
hier-1B	2	2 (L10.H2, L12.H5)	-0.15	-0.13
hier-1.4B	3	1 (L11.H2)	-0.25	-0.04
mod-410M	2	1 (L14.H3)	-0.26	-0.02
mod-1B	5	1 (L10.H2)	-0.21	-0.02
mod-1.4B	3	1 (L11.H2)	-0.16	$+0.00$

C.6 Cross-template transfer to vocabulary ICL

Two vocab-ICL tasks (antonym, 60 pairs; country-capital, 50 pairs; $k = 4$, format `word:answer`). Zero-strategy ablation, $n_{\text{eval}} = 100$, $B = 2{,}000$. Sub-population transfer in Table 14.

L11.H4 solo on vocab tasks. The *address* transfers but the *role* is task-conditional: on antonym L11.H4 acts as a writer ($\Delta\ell = -0.069$, 95% CI $[-0.109, -0.028]$); on country-capital it has no significant effect ($\Delta\ell = -0.005$, CI $[-0.033, +0.025]$).

C.7 Multi-cell mechanism rule-outs (rank-1 / V-cascade / V-shuffle)

Three named-mechanism rule-outs extended to the dominant canceller per cell: rank-1 (top-1 OV Frobenius share, all cancellers); V-cascade (zero-ablate top-3 anti-aligned upstream writers, $n_{\text{boot}} = 2000$ BCa); V-shuffle (relative collapse ≥ 0.5 = content-driven).

Rank-1 ruled out population-wide ($0/27$ cancellers exceed 7.5% top-1 OV vs. 50% threshold). V-shuffle content-driven in $6/6$ (rel-collapse 0.86–1.13). V-cascade ruled out $4/6$; hier-1B and mod-1B show a small residual cascade (Δ DLA$\in [+0.02, +0.03]$) from a near-orthogonal upstream writer (cosine$[-0.04, -0.01]$).

L11.H4 detailed CIs. Table 16 gives the paired-prompt bootstrap CIs for L11.H4 V-shuffle and V-composition (the case-study head also reported in §4.7). V-shuffle CIs exclude 0 in both cells (content-driven); V-composition CIs include 0 (parallel, not cascaded).

C.8 OV singular spectrum of L11.H4

D W/C aggregation: clean comparisons

We report two methodologically clean comparisons of v_{FV} vs. $v_{\mathcal{W}}$. The (in-sample α, per-vector tuning, top-K heuristic W/C on extension cells) variant from an earlier pilot is excluded as confounded; results below use the path-patching W/C partition and either held-out α or α-free transplant.

Table 14: **Cross-template W/C transfer: $3/4$ canonical pairs.** Mean shift in correct-token logit; positive = ablation helps correct; negative = ablation hurts. Canonical pattern requires (i) writers negative, (ii) cancellers positive, (iii) joint attenuates writers.

src cell	target task	writer $\Delta\ell$	canc $\Delta\ell$	joint $\Delta\ell$	verdict
hier-410M	antonym	-0.48	-1.56	-2.00	writer-dominant (fail)
hier-410M	country-capital	-0.52	$+0.14$	-0.37	canonical
mod-410M	antonym	-0.33	$+0.03$	-0.24	canonical (C-CI just over 0)
mod-410M	country-capital	-0.59	$+0.23$	-0.36	canonical

Table 15: **Multi-cell mechanism rule-outs.** $|\mathcal{C}|$: cancellers in the cell. $\sigma_1^{\max}/\|\mathrm{OV}\|_F^2$: max top-1 OV Frobenius share across all cancellers in the cell. $n_{\text{rank-1}}$: number of cancellers with top-1 OV ≥ 0.5. $\Delta\ell$ V-shuffle: V-shuffle relative collapse on the top canceller. V-cascade verdict on the top canceller from ablating the top-3 anti-aligned upstream writers (cos <0): RULED OUT means CIs include 0 for all 3; PARTIAL means ≥ 1 writer's CI excludes 0; VACUOUS means no upstream writer at $L < L_{\text{canceller}}$.

Cell	top canc.	$\|\mathcal{C}\|$	$\sigma_1^{\max}$	$n_{\text{rank-1}}$	$\Delta\ell$ V-shuf rel	V-cascade
hier-410M	L11.H4	6	0.070	$0/6$	0.86	RULED OUT
hier-1B	L13.H1	4	0.075	$0/4$	1.00	PARTIAL (L7.H2 cos$=-0.04$)
hier-1.4B	L12.H7	4	0.030	$0/4$	0.98	RULED OUT
mod-410M	L11.H4	3	0.070	$0/3$	1.13	VACUOUS (no upstream W)
mod-1B	L13.H1	6	0.075	$0/6$	1.05	PARTIAL (L7.H0 cos$=-0.01$)
mod-1.4B	L10.H7	4	0.030	$0/4$	0.95	VACUOUS (no upstream W)
cells		**27**	≤ 0.075	$\mathbf{0/27}$	≥ 0.86 in $6/6$	$4/6$ ruled out, $2/6$ partial

Table 16: **V-shuffle and V-composition diffs on L11.H4 DLA.** V-shuffle: random permutation of per-head V across source positions ($n_{\text{eval}} = 200$, 3 seeds, paired-prompt bootstrap). V-composition: zero-ablate L10.H9 (dominant upstream writer).

Intervention	Cell	baseline DLA	post DLA	diff	95% CI
V-shuffle	hier-410M	-0.216	-0.040	$+0.176$	$[+0.080, +0.279]$
	mod-410M	-0.215	-0.103	$+0.113$	$[+0.015, +0.206]$
V-composition	hier-410M	-0.216	-0.209	$+0.007$	$[-0.001, +0.015]$
	mod-410M	-0.215	-0.211	$+0.005$	$[-0.002, +0.011]$

Table 17: **OV singular spectrum.** L11.H4's $W_O W_V \in \mathbb{R}^{1024\times 1024}$, $\|OV\|_F = 7.64$. The top-1 left-singular mode anti-aligns with task-mean $u(x)$ at $\cos = -0.13$, but its Frobenius share is only 2.8% (top-5: 13%; top-10: 25%): the map is dense, with no rank-1 plateau (McDougall et al., 2023).

mode k	σ_k	$\cos(u_k, u)$	signed contrib. $\sigma_k \cos$
0	1.27	-0.13	-0.16
1	1.26	$+0.03$	$+0.04$
2	1.24	$+0.01$	$+0.01$
3	1.23	$+0.07$	$+0.09$
4	1.21	-0.00	-0.00

D.1 Held-out α on 6 main cells

$192/500$ calibration–evaluation split; per-cell α selected on calibration over the grid $\{0.25, 0.5, 1, 2, 4, 8, 16\}$ to maximise $\overline{\Delta\ell}$, evaluated on disjoint hold-out. All three vectors injected at layer $\lfloor n_{\text{layers}}/2 \rfloor$ on the readout token.

D.2 Transplant accuracy: writer-only vs FV-mean

Each head's activation slot is *replaced* with its task-mean ($n_{\text{eval}} = 200$). The asymmetry $\Delta\mathrm{acc}(\mathcal{W}) - \Delta\mathrm{acc}(\mathcal{F})$ is ≤ 0 in $6/6$ cells, the opposite direction from the additive-injection logit-shift result. Replacement neutralises cancellers' per-prompt outputs; sign-aware drop does not.

Table 18: Held-out logit shift at calibrated α.

Task	Model	$\Delta\ell(v_{\text{FV}})$	$\Delta\ell(v_{\mathcal{W}})$	$\Delta\ell(v_{\text{PCA}})$
hier	410M	-0.002	-0.003	-0.007
hier	1B	-0.002	-0.007	-0.003
hier	1.4B	-0.004	-0.022	$+0.003$
mod	410M	$+0.095$	$+0.159$	-0.002
mod	1B	-0.010	$+0.012$	-0.002
mod	1.4B	-0.005	$+0.097$	-0.032

Table 19: Transplant accuracy ($n_{\text{eval}} = 200$).

Cell	baseline	$\Delta\text{acc}(\mathcal{F})$	$\Delta\text{acc}(\mathcal{W})$	asymmetry
hier-410M	0.525	$+0.025$	-0.040	-0.065
hier-1B	0.595	$+0.000$	-0.100	-0.100
hier-1.4B	0.600	$+0.005$	-0.110	-0.115
mod-410M	0.480	$+0.090$	$+0.085$	-0.005
mod-1B	0.515	$+0.065$	$+0.060$	-0.005
mod-1.4B	0.520	$+0.085$	$+0.065$	-0.020

D.3 Per-cell ablation accuracy effects

Group-lesion accuracy effects on the 6 main cells. Sign of the shift matches the W/C direct-effect sign in $6/6$ cells; Δacc(W) CI excludes 0 in $3/6$, Δacc(C) CI excludes 0 in $1/6$.

D.4 Cancellers are writer-orthogonal in OV space

Decomposition $\bar{\mathbf{a}}_{\mathcal{C}} = \alpha_{\parallel}\bar{\mathbf{a}}_{\mathcal{W}} + \bar{\mathbf{a}}_{\mathcal{C}}^{\perp}$; perpendicular fraction 0.84–1.00 (mean 0.96, 95% CI $[0.91, 0.99]$); cos mean -0.18, 95% CI $[-0.36, -0.03]$.

E Robustness controls

E.1 Head-randomised control: the tautology rule-out

Tests whether the canonical 4-condition verdict is a tautology of sign-partitioning a high-magnitude population. Per cell: $n_{\text{eval}} = 100$, top-$K = 14$ FV by |DLA| vs. 10 random seeds of |DLA|-matched non-FV partitions, zero-strategy group lesions on both. Table 22 reports per-cell aggregates.

FV ablations are $5.5\times$ (W) and $5.2\times$ (C) larger than |DLA|-matched random, attenuation gap $2.5\times$ (0.047 vs. 0.019 nats). The sign part is mechanically forced; the loose attenuation flag fires for 77% of random seeds, so magnitude (not sign) distinguishes FV from non-FV. Matching is rank-closest (non-FV |DLA| averages 0.04 vs. FV 0.20 nats); the mod-1.4B FV/random gap < 1 is consistent with its weak canceller pool ($|\mathcal{C}| = 4$).

Table 20: Δacc with Wald 95% CI.

Cell	baseline	Δacc(W)	Δacc(C)	Δacc(W$\cup$C)
hier-410M	0.55	$-0.05\,[-0.11, +0.01]$	$+0.07\,[+0.01, +0.13]$	$+0.03\,[-0.03, +0.09]$
hier-1B	0.58	$-0.16\,[-0.22, -0.10]$	$+0.02\,[-0.04, +0.09]$	$-0.12\,[-0.18, -0.06]$
hier-1.4B	0.61	$-0.07\,[-0.13, -0.01]$	$+0.05\,[-0.01, +0.11]$	$0.00\,[-0.07, +0.06]$
mod-410M	0.56	$-0.01\,[-0.07, +0.05]$	$+0.03\,[-0.03, +0.10]$	$+0.08\,[+0.02, +0.15]$
mod-1B	0.55	$-0.23\,[-0.29, -0.17]$	$+0.04\,[-0.03, +0.10]$	$-0.16\,[-0.22, -0.10]$
mod-1.4B	0.56	$-0.01\,[-0.07, +0.06]$	$+0.04\,[-0.03, +0.10]$	$-0.06\,[-0.12, +0.00]$

Table 21: Writer/canceller OV geometry on the 6 main cells.

cell	$\cos(\mathcal{W}, \mathcal{C})$	perp. frac.	$\alpha_{\parallel}$
hier-410M	$+0.05$	1.00	$+0.08$
hier-1B	-0.32	0.95	-1.02
hier-1.4B	$+0.03$	1.00	$+0.04$
mod-410M	-0.54	0.84	-1.91
mod-1B	-0.24	0.97	-0.22
mod-1.4B	-0.09	1.00	-0.26

Table 22: **Head-randomised control across** 6 **Pythia cells.** For each cell we report the FV-set's writer/canceller/joint ablation logit shifts and the mean over 10 random |DLA|-matched non-FV partitions. Sign convention: writer = DLA >0 (pushes correct $\uparrow$), canceller = DLA <0 (pushes correct $\downarrow$); ablating writers shifts $-$, ablating cancellers shifts $+$. Magnitudes $|W|$, $|C|$, gap $= |W| + |C| - |both|$. Last column: fraction of random matched seeds whose joint ablation passes the loose binary attenuation criterion $|both| < |W| + |C| - 0.005$.

	FV-set			Random matched (mean over 10 seeds)				
Cell	$\lvert W\rvert$	$\lvert C\rvert$	gap	$\lvert W\rvert$	$\lvert C\rvert$	gap	atten frac	FV/rand gap
hier-410M	0.184	0.068	0.034	0.037	0.034	0.008	8/10	4.3×
hier-1B	0.161	0.206	0.101	0.039	0.019	0.019	6/10	5.3×
hier-1.4B	0.097	0.097	0.082	0.046	0.031	0.042	9/10	1.9×
mod-410M	0.166	0.349	0.013	0.019	0.007	0.005	6/10	2.6×
mod-1B	0.148	0.311	0.045	0.015	0.043	0.017	8/10	2.6×
mod-1.4B	0.102	0.172	0.010	0.060	0.031	0.023	9/10	0.4×
mean	**0.143**	**0.200**	**0.047**	**0.036**	**0.027**	**0.019**	**46/60 (77%)**	**2.5×**