AGGREGATION ORDER IN ACTIVATION PATCHING: WHICH CIRCUIT IS SELECTED AND WHAT IT PRESERVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation-patching evidence supports a mechanistic claim only as far as the measurement pipeline behind it can distinguish that claim from alternatives. Here we study one step of that pipeline: how effects across prompts are aggregated. A circuit is often called faithful when ablating everything else preserves a task score aggregated over prompts, but opposite-signed prompt effects can be combined before or after taking their magnitude. We isolate this aggregation order: within each comparison, the model, candidate universe, prompt set, circuit size and exhaustive search are held fixed, and we measure which circuit each order selects and what that circuit preserves on held-out prompts. Across twelve independently drawn discovery-and-evaluation runs per model from one template bank, in the published indirect-object identification (IOI) universes of GPT-2 small and Pythia-160m, under mean ablation, averaging first selects circuits with $1.73\times$ and $3.65\times$ the full-ablation-normalized held-out KL of those selected by per-prompt-absolute scoring (raw gaps 0.049 and 0.129 nats; the two models' ablation scopes differ). The gap is already present on the discovery prompts. On held-out prompts, most of the excess lies in answer-set probability mass and, on normalized KL, average-first selections were not established to differ from random subsets of the same size. In Pythia, average-first circuits keep the clean model's top token on 75.6% of prompts versus 89.7% for per-prompt-absolute circuits (a prespecified descriptive endpoint), despite similar binary task accuracy. Under counterfactual patching the ratios fall to $1.11\times$ and $1.12\times$, and the registered operator interaction was confirmed on both models. A preregistered six-family study confirms the mean-ablation contrast, on the relative scale, in only two GPT-2 families, and a preregistered 8B layer-group extension missed its materiality gate. We propose neither a new discovery algorithm nor a universally preferable faithfulness score: changing only the aggregation order can change which circuit is selected and what it preserves, so faithfulness claims should state their aggregation and intervention protocol and evaluate directly the property they are intended to preserve.

1 INTRODUCTION

Circuit analyses report that a small set of components is *faithful*: ablating everything else preserves a task score, typically a logit difference averaged over prompts (Wang et al., 2023; Conmy et al., 2023; Mueller et al., 2025). Because faithfulness is used to choose among candidates, the scoring rule can change which circuit is reported.

Opposite-signed prompt effects can be combined before or after taking their magnitude. Average-first scoring (FIELD) averages the signed per-prompt metric changes, then takes the magnitude; per-prompt-absolute scoring (STRICT) averages their magnitudes. Existing methods use both, at different stages (§2).

The two criteria differ in exactly one respect: average-first scoring can cancel opposite-signed prompt effects (Scheurer et al., 2023; Kramár et al., 2024; Miller et al., 2024). Preserving the logit difference between the correct and distractor names fixes the ratio of their probabilities, not their total mass. Hence a failure mode: a candidate can look faithful after averaging despite substan-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

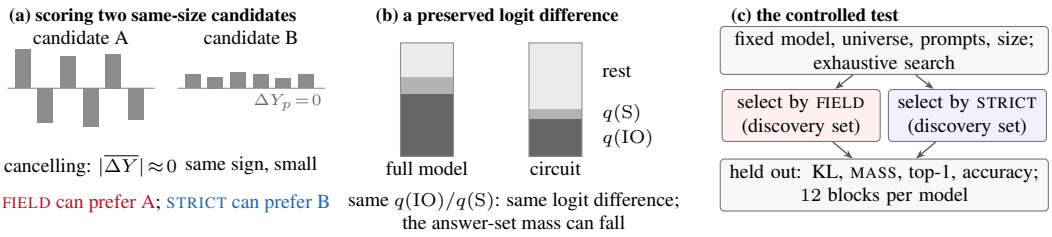

Figure 1: Schematic (illustration, not data). (a) Average-first scoring can prefer a candidate whose prompt effects cancel (Eq. 1); the selected circuit depends on the candidate set. (b) A preserved logit difference fixes $q(\mathrm{IO})/q(\mathrm{S})$, not the answer-set mass, so distribution and top-token changes can remain. (c) The comparison of §3 (MASS: answer-set probability mass).

tial per-prompt score changes and unconstrained shifts in answer-set mass, so the criteria can rank candidates differently (Figure 1a,b). Whether this ranking difference harms held-out distribution or prediction preservation is an empirical question.

We isolate this choice (Figure 1c). We fix the model, candidate universe, prompts and circuit size; exhaustive search removes search-quality variation. Each rule selects on discovery prompts; selections are scored on held-out templates. One discovery draw captures evaluation noise but not selection variability, so we repeat the whole procedure on 12 independent blocks per model under a frozen preregistration, with mean ablation and counterfactual patching (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024) on identical prompts.

Findings. (i) Under mean ablation in the published indirect-object identification (IOI) universes of GPT-2 small and Pythia-160m, average-first circuits have $1.73\times$ and $3.65\times$ the full-ablation-normalized held-out KL of per-prompt-absolute circuits, higher in all 12 blocks per model; operator scope differs between the models, so the magnitudes are not directly comparable. The gap is as large on discovery prompts as held out, so it is not explained by a discovery-to-held-out generalization gap. On normalized KL, average-first selections were not established to differ from random same-size subsets (§3). (ii) Most of the excess lies in answer-set mass. In Pythia it reaches predictions: the clean top token is kept on 75.6% of prompts against 89.7%, while accuracy differs little. (iii) The consequence is not uniform across protocols. The registered operator interaction was confirmed on both models, and counterfactual ratios are only $1.11\times$ and $1.12\times$ (the registered equivalence verdict has limited calibration; §3). A second preregistered study confirms the mean-ablation contrast in two of six model–task families, and a preregistered 8B extension missed its materiality gate (§4). (iv) An *exploratory* analysis of discovery-time acceptance rules suggests the distinction can persist at a stopping rule, with a size that depends on the candidate path and the rule; it vanishes on a shared path when both rules ignore per-prompt overshoot (§5).

Within the published universes, per-prompt-absolute scoring is the better-supported of the two for preserving the output distribution under mean ablation (both models) and clean-model predictions (Pythia). Average-first scoring shows no established compensating advantage; importance and discovery-KL selection preserve some endpoints better still. Nothing here shows that a published circuit is wrong or that either aggregation is generally correct. The selected circuits are retained subsets of declared universes (other heads intact), not outputs of an end-to-end pipeline. Our narrower conclusion: a faithfulness claim is incomplete unless it states the aggregation, the ablation operator, the candidate universe and the property it is meant to preserve (§6; claim status: Table 1).

2 SETTING: SELECTION CRITERIA AND WHAT THEY PRESERVE

A candidate circuit C lives in a fixed head universe U: every head in $U \setminus C$ is ablated, and heads outside U stay intact. Under *mean ablation*, GPT-2 replaces each ablated head's final-token output by its mean over the block's development prompts; Pythia uses role-aligned mean outputs at every position. Under *counterfactual patching*, the activations at the same sites come from a corrupted prompt. $\Delta Y_p(C)$ is the resulting change in the task metric on prompt p: the logit difference $z_{\mathrm{IO}} - z_{\mathrm{S}}$ between the correct indirect-object name (IO) and the distractor subject name (S). The two

claim	evidence	status	magnitude
Average-first selection raises held-out KL under mean ablation	repeated-discovery study, 12 blocks per model	preregistered; Holm-confirmed	$1.73\times$ / $3.65\times$ normalized; 0.049 / 0.129 nats
Operator interaction; counterfactual ratio within margin	same	preregistered; interaction Holm-confirmed; equivalence as registered (later audit: miscalibrated at the upper margin)	$1.11\times$ / $1.12\times$
Excess is answer mass; Pythia top-1 cost hidden by accuracy	same	prespecified secondary / descriptive	88% / 78% MASS; 75.6 vs 89.7%
Cost outside the published universes	scope study, 6 families $\times$ 2 constructions	preregistered; Holm over 22	$2/6$ families on the relative scale; material only in screened universes
Cost in an 8B model (layer-group universes)	Qwen3-8B, 3 cells	preregistered; composite gate missed	ratio 1.243 vs 1.25; $+0.785$ nats
Fresh-template transfer of frozen selections	frozen repeated-discovery selections, new 20-template bank	preregistered fresh-family evaluation; H1/H2 confirmatory, H3 descriptive after failed calibration audit	$1.49\times$ / $2.46\times$ normalized (confirmed / confirmed); Pythia top-1 7.9 points (confirmed); counterf. $1.06\times$ / $1.16\times$ (descriptive)
Accepted circuits at discovery thresholds can lose predictions	stored candidate sets (§5)	exploratory, post hoc; rule-dependent	Pythia 21.2 (own), 6.8 (shared) points fewer top-1 matches
Cost associated with negative Name Movers	120 universe restrictions	exploratory; nominated post hoc	log ratio 0.005 vs median 0.52
Fresh replication of nominated universe restriction	12 fresh GPT-2 IOI blocks, 120 restrictions	preregistered fresh-block replication of a post-hoc nominated restriction; two Holm-controlled claims	$C_G = 0.409$ (replicated), $C_A = 0.0188$ nats (replicated); FIELD $0.0581 \to 0.0401$, STRICT $0.0351 \to 0.0384$ nats

Table 1: Claims and their evidential status. Fresh-template row: H1–H3 per Appendix A.1. Restriction row: U^* is the nominated restriction and C_G, C_A are the median-minus-U^* log-ratio and raw-gap contrasts; point levels, median restriction $\to U^*$; only FIELD's change is resolved (Appendix A.2).

aggregation orders are

$$\text{FIELD}(C) = \left|\tfrac{1}{n}\textstyle\sum_p \Delta Y_p(C)\right|, \qquad \text{STRICT}(C) = \tfrac{1}{n}\textstyle\sum_p \left|\Delta Y_p(C)\right|.$$

FIELD is the form of the circuit-level faithfulness criteria of Wang et al. (2023, §4) and of the integrand of MIB's circuit–model distance (CMD) (Mueller et al., 2025, Eq. 1). The released EAP and EAP-IG code aggregates edge scores in the same order (Syed et al., 2024; Hanna et al., 2024; Appendix A); ACDC's default IOI objective is a per-prompt KL, which cannot cancel (Conmy et al., 2023). STRICT is the aggregation used by the AtP and AtP* estimators (Kramár et al., 2024, Eqs. 5, 11), although their target patching effect is defined on the mean task metric. The gap between the criteria is exact. With $P, N \geq 0$ the summed positive and negative magnitudes of the per-prompt effects,

$$\text{STRICT}(C) - \text{FIELD}(C) = \tfrac{2}{n}\min(P, N), \tag{1}$$

so the criteria agree exactly when positive and negative prompt effects do not coexist. The identity says when they differ, not which circuit each selects or what it preserves.

What a circuit can preserve. We score selected circuits on held-out prompts by four endpoints. The *task score* is each criterion's own held-out value. *Distribution preservation* is $\mathrm{KL}(P_{\text{full}} \,\|\, P_C)$ over the next-token distribution; normalized KL divides it by the full-ablation KL before averaging over circuit sizes. By the chain rule, the raw divergence splits exactly into a term for total answer-set probability (MASS) and two conditional terms: the split within the two-name answer set, and the rest of the vocabulary. *Prediction preservation* is agreement with the clean model's top vocabulary token. *Task accuracy* records whether the correct name outranks the distractor. Since $z_{\text{IO}} - z_{\text{S}} = \log q(\text{IO})/q(\text{S})$, with q the next-token probability, preserving the logit difference fixes the ratio of the two answer probabilities but not their total mass $q(\text{IO}) + q(\text{S})$. A circuit can thus preserve the task score while changing answer-set mass and the model's top prediction. We also score three reference procedures: an importance ranking (IMP: keep the heads with the largest single-ablation effect), exhaustive selection by discovery-prompt KL (DISC-KL), and random subsets.

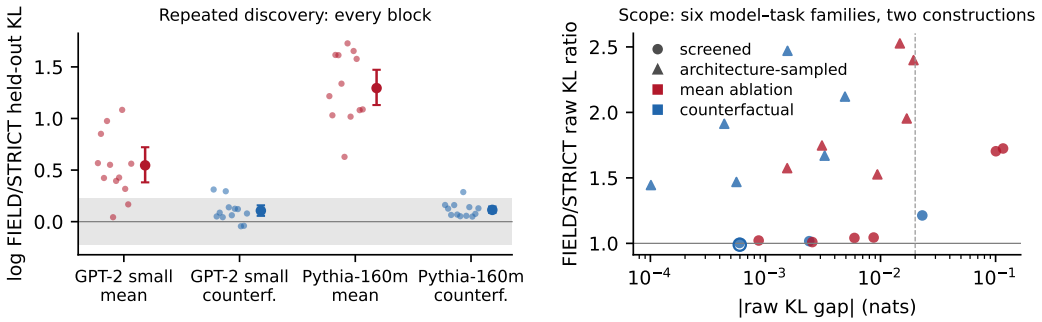

Figure 2: *Left:* every independent block of the repeated-discovery study: the log ratio of FIELD's to STRICT's normalized held-out KL, with the registered pooled estimate and its 95% (mean ablation) or 90% (counterfactual) block-bootstrap interval; grey band: the equivalence margin. *Right:* the scope study: FIELD/STRICT ratio of raw held-out KL against the raw gap in nats, for every model–task family, universe construction and operator with non-identical selections (open marker: STRICT worse; dashed: 0.02 nats). Most large ratios sit on small between-selector gaps, often in universes whose full ablation barely moves the output; a small gap between two selections does not by itself make the universe unimportant.

3 AVERAGE-FIRST SELECTION INCREASES HELD-OUT KL UNDER MEAN ABLATION

Design and registered predictions. Everything but the aggregation is fixed: the candidate universe (the ten published IOI heads of GPT-2 small; the fourteen-head published circuit of Pythia-160m, Tigges et al., 2024), the discovery prompts, and an *exhaustive* search over every subset of each size (2–8 and 3–11 heads). The repeated-discovery study repeats the whole procedure in 12 independent blocks per model (50 discovery and 200 held-out prompts each) from a fixed 21-template, 16-name IOI population (templates held out, names shared), under both operators on identical prompts (a single-draw precursor is in Appendix A). The block is the inferential unit, so intervals describe re-selection variability within this bank, not transfer to new templates (tested in §4, Appendix A.1). The primary contrast L is the log ratio of FIELD's to STRICT's mean normalized held-out KL. We registered three claims per model, Holm-corrected together (6 claims): under mean ablation $L > 0$; L is larger under mean ablation than under counterfactual patching; and under counterfactual patching e^L lies within $[0.8, 1.25]$. A block bootstrap and a delete-one-block jackknife must both support each claim. Hypotheses, code and a ledger of every previously exposed prompt were frozen before any prompt was run.

Result. Under mean ablation, FIELD-selected circuits have $1.73\times$ $[1.46, 2.06]$ (GPT-2) and $3.65\times$ $[3.10, 4.36]$ (Pythia) the normalized held-out KL of STRICT-selected circuits (Table 2), with raw gaps of 0.049 and 0.129 nats. The KL gap is as large on the discovery prompts (0.048 and 0.133 nats), so it is not a generalization gap; FIELD's own criterion, by contrast, shows no established held-out advantage (Table 3). The difference is positive in 12/12 and 12/12 blocks (Figure 2) and, in point estimate, at every retained size (Appendix A); the registered mean-ablation claim was confirmed on both models (Holm-adjusted Monte Carlo $p \leq 6 \times 10^{-4}$, at the resolution floor of 10,000 bootstrap replicates). The mean-versus-counterfactual contrast is 0.44 $[0.27, 0.62]$ and 1.18 $[1.03, 1.34]$ in L, and the registered interaction claim was likewise confirmed on both models. Under counterfactual patching the ratios are $1.11\times$ and $1.12\times$; FIELD has higher KL in 10/12 and 12/12 blocks, but both ratios fall inside the registered margin and the registered procedure declared equivalence. Later simulations for a follow-up study, in populations derived from these blocks, found the twelve-block equivalence procedure declaring equivalence in 9.4% and 9.2% of experiments at the upper margin rather than the nominal 5%, so we report the verdict as registered, not as a calibrated equivalence result (Appendix E). Two further facts limit the interpretation. Under mean ablation, FIELD's normalized KL was lower in point estimate than that of random same-size subsets, but the comparison did not establish a difference under our rule requiring both bootstrap and jackknife intervals

		mean ablation					counterfactual patching				
model	quantity	FIELD	STRICT	IMP	DISC-KL	random	FIELD	STRICT	IMP	DISC-KL	random
GPT-2	raw KL (nats)	0.117	0.068	0.063	0.047	0.136	0.148	0.132	0.474	0.124	0.579
	norm. KL	0.595	0.344	0.318	0.235	0.687	0.073	0.066	0.235	0.062	0.287
	accuracy (%)	99.2	99.6	99.8	99.2	98.6	89.2	90.1	60.9	88.4	52.5
	top-1 agreement (%)	92.6	92.7	91.6	93.6	91.6	83.2	84.1	57.8	83.2	50.1
Pythia	raw KL (nats)	0.178	0.049	0.035	0.033	0.202	0.082	0.073	0.105	0.072	0.793
	norm. KL	0.379	0.104	0.076	0.069	0.416	0.054	0.048	0.068	0.047	0.520
	accuracy (%)	96.6	97.4	97.4	97.5	85.2	86.9	87.7	78.5	87.1	27.8
	top-1 agreement (%)	75.6	89.7	90.2	90.9	73.1	85.1	85.5	81.1	85.8	46.4
GPT-2	FIELD/STRICT	$1.73\times$ [1.46, 2.06], Holm $p \le 6 \times 10^{-4\dagger}$					$1.11\times$ [1.06, 1.17], Holm $p = 0.002^{\ddagger}$				
Pythia	FIELD/STRICT	$3.65\times$ [3.10, 4.36], Holm $p \le 6 \times 10^{-4\dagger}$					$1.12\times$ [1.09, 1.16], Holm $p \le 6 \times 10^{-4\dagger\ddagger}$				

Table 2: Repeated-discovery study: 12 independent blocks per model; only the aggregation differs. Raw KL: held-out $\mathrm{KL}(P_{\mathrm{full}} \| P_C)$, size-averaged; norm. KL: the same relative to the block's full ablation, then averaged over blocks (lower preserves more); accuracy: IO ranked above S; top-1: the clean model's top vocabulary token preserved. Full ablation of the universe: raw KL 0.199 / 0.488 nats and accuracy 90% / 61% (GPT-2 / Pythia, mean ablation); 2.024 / 1.526 nats and 0% / 0% (counterfactual). †At the resolution floor of 10,000 bootstrap replicates. ‡Registered verdict: equivalent; a later calibration audit found the twelve-block equivalence test anti-conservative at the upper margin (Appendix E). Intervals: 95% block bootstrap (mean), 90% against $[0.8, 1.25]$ (counterfactual).

to exclude zero (Appendix A). The mean-ablation operators also differ in scope (the final token for GPT-2, every position for Pythia), so the difference between the two models' magnitudes cannot be attributed to the model alone.

Where the excess lies. In the prespecified secondary decomposition of the published-universe mean-ablation result, FIELD's excess normalized KL is predominantly MASS on both models (88% and 78% of the summed component estimates; unadjusted). Its direction differs post hoc between the models (Appendix A). FIELD's circuits preserve binary task accuracy nearly as well as STRICT's, but on Pythia they keep the clean model's top vocabulary token on only 75.6% of prompts versus 89.7% (random subsets: 73.1%; difference jackknife $[-17.1, -11.1]$ points; FIELD worse in 12/12 blocks; prespecified descriptive). The difference is present in point estimate at every retained size (-17.0 to -9.9 points), with intervals below zero at $8/9$ sizes (size average in Table 3). Post hoc, across the other repeated-discovery and scope-study cells, GPT-2 included, the size-averaged difference is at most 3.4 points; discovery-time operating points behave differently (§5).

Which predictions change (post hoc). Re-executing the selected circuits (bit-identical to the stored readouts; Appendix A) shows that the Pythia loss is confined to prompts whose clean top token is the correct name (55% of prompts; -30.3 points $[-34.4, -26.5]$), which FIELD's circuits more often displace with tokens outside the answer pair (21.9 such losses per 100 prompts versus 4.9 for STRICT). It persists at every clean top-two margin quartile and reverses where the clean top token lies outside the pair (Table 7). Comparable preservation under the FIELD criterion does not guarantee comparable predictions: among same-size Pythia pairs whose held-out FIELD errors differ by at most 0.05 logits (15/108, from 10 blocks; matched on that criterion only, so their STRICT errors still differ), the FIELD$-$STRICT difference in top-1 agreement is -10.3 points $[-14.5, -6.3]$ (block bootstrap; tolerance sweep in Appendix A). These subsets are selected using held-out outcomes, so they are descriptive rather than confirmatory estimates. As a secondary measure, FIELD's exact selected mask recurs across independent blocks in 5.4% and 0.7% of (block pair, size) cases, versus 32.9% and 25.6% for STRICT. Mask instability is not itself lower preservation (Appendix E).

4 Scope across tasks, universes and model scale

Other tasks and universe constructions (preregistered). The scope study confirmed the registered mean-ablation contrast, on the relative scale, in only $2/6$ model–task families (GPT-2 IOI and induction). It covers GPT-2 small and Pythia-160m on IOI, induction and subject–verb agreement,

model, operator	FIELD − STRICT on held-out prompts (all same-size pairs)				
	FIELD error (logits)	STRICT error (logits)	raw KL (nats)	top-1 (points)	accuracy (points)
GPT-2, mean abl.	-0.001 $[-0.072, +0.064]$	$+0.132$ $[+0.094, +0.171]$	$+0.049$ $[+0.033, +0.066]$	-0.1 $[-0.8, +0.7]$	-0.5 $[-0.8, -0.2]$
GPT-2, counterf.	-0.006 $[-0.054, +0.048]$	$+0.144$ $[+0.089, +0.201]$	$+0.015$ $[+0.007, +0.025]$	-0.9 $[-1.5, -0.2]$	-0.9 $[-1.8, -0.2]$
Pythia, mean abl.	-0.045 $[-0.119, +0.032]$	$+0.358$ $[+0.324, +0.389]$	$+0.129$ $[+0.108, +0.148]$	-14.1 $[-16.7, -11.5]$	-0.8 $[-1.2, -0.5]$
Pythia, counterf.	-0.010 $[-0.048, +0.030]$	$+0.108$ $[+0.084, +0.133]$	$+0.009$ $[+0.006, +0.013]$	-0.4 $[-0.7, -0.1]$	-0.9 $[-1.3, -0.4]$
Pythia, mean abl., matched on FIELD error (15 pairs)	-0.005 $[-0.022, +0.014]$	$+0.397$ $[+0.270, +0.533]$	$+0.073$ $[+0.040, +0.112]$	-10.3 $[-14.5, -6.3]$	-0.1 $[-1.8, +1.0]$

Table 3: What each selection preserves (repeated-discovery study, held-out prompts, size-averaged; 95% block-bootstrap intervals). Errors (in logits) and raw KL: positive means FIELD's circuit preserves less; top-1 agreement and accuracy: negative means FIELD's circuit preserves less. FIELD shows no established advantage on its own held-out criterion in any cell. Last row: the post hoc subset of same-size Pythia pairs matched on held-out FIELD error only (within 0.05 logits).

family	screened universe		architecture-sampled universe		registered mean-ablation claim	
	mean	counterfactual	mean	counterfactual	local p	verdict
GPT-2 IOI	1.70 (+0.101; 15)	1.21 (+0.023; 16)	2.53 (+0.015; 15)	2.47 (+0.002; 12)	$\leq 10^{-4}$	confirmed
GPT-2 induction	1.72 (+0.116; 16)	1.02 (+0.002; 11)	1.53 (+0.009; 16)	1.67 (+0.003; 14)	1.8×10^{-4}	confirmed
GPT-2 agreement	1.02 (+0.001; 5)	identical	1.57 (+0.002; 16)	1.45 ($+1 \times 10^{-4}$; 15)	0.010	not after Holm
Pythia IOI	1.04 (+0.006; 9)	0.99 (-0.001; 8)	1.95 (+0.017; 15)	1.47 (+0.001; 15)	0.195	not confirmed
Pythia induction	1.04 (+0.009; 12)	1.00 (+0.001; 4)	2.40 (+0.019; 15)	2.12 (+0.005; 14)	0.033	not after Holm
Pythia agreement	1.01 (+0.003; 6)	identical	1.75 (+0.003; 14)	1.91 ($+4 \times 10^{-4}$; 16)	0.032	not after Holm

Table 4: Scope study: FIELD/STRICT ratio of block-mean raw held-out KL (raw gap in nats; blocks of 16 with FIELD worse); "identical": same masks in every block. Local p is before Holm over 22 claims; all selectors' levels in Appendix B.

with sixteen fresh blocks per family. Each block holds an eight-head universe auto-screened from all 144 heads and an architecture-sampled one with one head per layer bin. The endpoint is raw held-out KL, with one Holm family of 22 registered claims (Table 4, Appendix B). In both confirmed families FIELD had higher KL on the registered relative scale under both constructions, but the architecture-sampled raw gaps were only 0.0147 and 0.0094 nats. These confirmations establish a construction-robust relative ordering, not a material absolute gap under both constructions. In the other four families, architecture-sampled universes gave large ratios on raw gaps below 0.02 nats; screened ones were near $1\times$. No family confirmed the registered cross-construction operator interaction or the counterfactual-equivalence conjunction ($0/6$ and $0/6$; conjunctions containing identical selections receive no verdict). Failure to confirm does not establish absence. Under counterfactual patching, architecture-sampled ratios of 1.45–$2.47\times$ rested on raw gaps of only 0.0001–0.0049 nats, against 0.0231 nats for screened GPT-2 IOI. Registered forecasters of the gap did not beat ordinary alternatives (Appendix B).

Model scale and fresh templates (preregistered). In a Qwen3-8B induction experiment over eight architectural head groups, the continuous mean-ablation estimates favoured STRICT. Average-first selection increased size-averaged held-out KL by 0.785 nats and task-accuracy loss by 10.76 percentage points, in the preregistered direction in 12/12 blocks. The FIELD/STRICT held-out KL ratio of 1.243 fell short of the preregistered 1.25 materiality requirement, so the composite claim was not confirmed and the subsequent cells are descriptive. In the agreement cell both criteria's circuits moved the output more than ablating every group (Appendix C). On a fresh 20-template IOI family with different first-mention and sentence structures, the frozen selections retained the mean-ablation preservation ordering in both models: FIELD had higher normalized held-out KL than STRICT in GPT-2 and Pythia ($1.49\times$ $[1.18, 1.85]$ and $2.46\times$ $[2.15, 2.79]$; Holm-adjusted $p = 0.012$

model	candidates	retained heads FIELD/ STRICT	top-1 agreement (%) FIELD/ STRICT	difference (points) FIELD−STRICT [95%]	raw KL (nats) FIELD/ STRICT	fallbacks FIELD/ STRICT
GPT-2	own exhaustive picks	$2.00^{\dagger}$ / 6.33	90.4 / 94.3	-3.9 $[-5.9, -1.6]$	0.236 / 0.036	0 / 0
GPT-2	shared importance path	6.00 / 7.08	94.0 / 95.6	-1.6 $[-2.8, -0.5]$	0.041 / 0.017	0 / 0
Pythia	own exhaustive picks	$3.00^{\dagger}$ / 5.92	67.2 / 88.5	-21.2 $[-29.5, -14.2]$	0.270 / 0.046	0 / 0
Pythia	shared importance path	3.58^{*} / 7.08	84.8 / 91.5	-6.8 $[-9.6, -4.6]$	0.063 / 0.029	0 / 0

Table 5: Discovery-only operating points at $\tau = 80\%$ (repeated-discovery study, mean ablation, 12 blocks; exploratory). Retained heads: mean accepted size; †: every block stopped at the smallest evaluated size (single heads also passed on discovery data in all Pythia and some GPT-2 blocks; their held-out outcomes were not evaluated); *: some blocks did. Fallbacks: blocks accepting the full universe. 90%, counterfactual patching and the 8B cells: Appendix E; 85%: Figure 3.

and $p = 5 \times 10^{-4}$). In Pythia, FIELD also lost more than the registered 2-percentage-point margin of clean-model top-token agreement (7.9 points $[6.2, 9.4]$, Holm $p = 5 \times 10^{-4}$).

Exploratory: the contrast is associated with the negative Name Movers. After inspecting the twelve GPT-2 blocks, we varied which heads could be ablated (Appendix A). Making heads 10.7 and 11.10 (negative Name Movers) and 9.0 (a backup Name Mover) ineligible, while leaving them intact to recompute, gave the smallest size-averaged FIELD/STRICT log-KL ratio of all 120 seven-of-ten restrictions: 0.005 $[-0.13, 0.15]$, against a median of 0.52 over the others (breakdowns: Appendix A). FIELD remained worse by point estimate in $120/120$ restrictions. Because the trio was nominated after inspection, this shows an association with universe composition in these blocks (nominal; not adjusted for selection among the 120), not a mechanism. A retrospective association between opposing single-head effects and the selection cost across all settings (Appendix B) motivates a prospective composition test. On 12 fresh GPT-2 IOI blocks, the previously nominated seven-head restriction excluding 10.7, 11.10 and 9.0 again reduced the FIELD–STRICT preservation contrast relative to the median of the other 119 restrictions, on both the log-ratio and raw-gap estimands ($C_G = 0.409$ $[0.207, 0.605]$, $C_A = 0.0188$ $[0.0098, 0.0274]$ nats; Holm-adjusted p-values 0.005 and 0.005). This replicates the restriction-specific sensitivity, not a causal effect of head sign.

Endpoint trade-offs (post hoc). No selector is best on every endpoint (Table 2). On common endpoints (Appendix B; nominal 95% intervals, unadjusted), FIELD has higher held-out STRICT error in 21/28 study cells, while no cell establishes that FIELD is worse on its own raw held-out criterion. In 4/4 model–operator cells of the published IOI universes, FIELD has no established advantage on its own held-out criterion and is worse on raw and normalized KL (under counterfactual patching, within the equivalence margin), the STRICT criterion and accuracy. Screened scope universes differ: FIELD is better on its own held-out criterion in 9 of 12 screened settings (9 of 10 with non-identical selections). The trade-off is objective-dependent, not a universal ordering: neither aggregation is a generally preferred objective.

5 Exploratory operating-point analysis: passing a discovery-time threshold need not preserve predictions

Acceptance rules. Sections 3–4 compare the two criteria at matched circuit sizes. A stopping rule instead accepts the smallest evaluated size that passes a faithfulness threshold, so the criteria may stop at different sizes. With the discovery normalizer $D = \mathbb{E}_{\text{disc}}|\Delta Y_p(\text{ALL})|$ (every universe head ablated), the errors are $J_F = |\mathbb{E}_{\text{disc}}\Delta Y_p|/D$ and $J_S = \mathbb{E}_{\text{disc}}|\Delta Y_p|/D$, with recovery $f_X = 1 - J_X$. Under this common normalizer $J_F \le J_S$ for every circuit, so the average-first rule is never harder to pass at the same threshold. A rule at threshold τ accepts the smallest evaluated size whose own discovery-selected circuit has $J_X \le 1 - \tau$; if none does, it falls back to the full universe. Thus τ bounds an aggregated task-score error, not the share of predictions or task accuracy preserved. We applied these discovery-only rules to the stored candidate sets at $\tau = 80, 85$ and 90%, retaining every block and fallback. Candidates were each criterion's own exhaustive selections and a shared importance-ranked path, on which the rules differ only in where they stop. The analysis is exploratory: we fixed these rules after analysing the registered studies and inspecting an earlier pass that used each criterion's own normalizer (Appendix E).

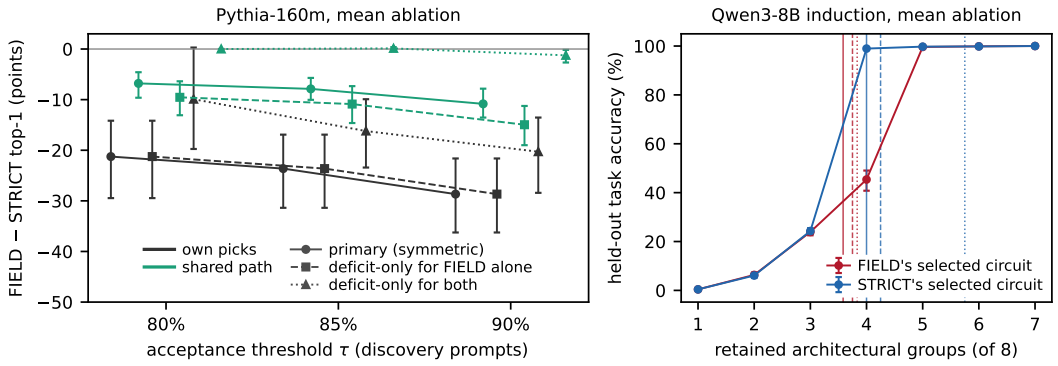

Figure 3: Operating points (exploratory). *Left:* Pythia-160m, mean ablation: FIELD−STRICT agreement with the clean top prediction at each rule's accepted circuits, for each criterion's own exhaustive picks and the shared importance path, under three acceptance rules: symmetric (primary), deficit-only for both (ignoring per-prompt overshoot), and deficit-only for FIELD alone (added after the others had been read; Appendix E); 95% block-bootstrap intervals. *Right:* Qwen3-8B induction, mean ablation: held-out task accuracy of each criterion's discovery-selected circuit by retained group count; vertical lines mark the mean accepted size at 80% (solid), 85% (dashed) and 90% (dotted).

Result. Under mean ablation in Pythia, circuits accepted by the average-first rule at 80% agreed with the clean model's top prediction less often than those accepted by the per-prompt-absolute rule. For each criterion's own picks the difference was -21.2 percentage points. On the shared path, where only the acceptance rule changes, it was a smaller -6.8 points (intervals: Table 5). For its own picks, FIELD stopped at the smallest size with held-out outcomes (three heads) at every threshold. In separate discovery-only checks at 80%, single-head candidates also passed in all twelve blocks, but their held-out outcomes were not evaluated. FIELD's measured operating point thus stayed at this grid floor, while stricter thresholds moved STRICT to larger circuits; the observed difference grew to -28.7 points at 90%. At the smallest evaluated size at which both discovery rules passed, the Pythia difference remained -15.9 points $[-23.5, -9.2]$. For GPT-2 the operating-point difference (-3.9; Table 5) attenuated to -0.8 $[-2.6, 1.0]$ at this common size, consistent with size choice accounting for much of the observed difference. Passing either rule is not a calibrated held-out guarantee (held-out pass rates in Appendix E).

The consequence depends on what the rule penalizes. Deficit-only rules ignore per-prompt overshoot, as recovery thresholds do: $J_{F,+} = (\sigma\, \mathbb{E}_{\text{disc}} \Delta Y_p)_+ / D$ and $J_{S,+} = \mathbb{E}_{\text{disc}} (\sigma \Delta Y_p)_+ / D$, with σ the sign of the full-ablation effect (Figure 3). With both rules deficit-only, the Pythia own-pick difference at 80% shrank in magnitude to -9.9 points $[-19.8, 0.3]$. On the Pythia shared path at 80% both rules accepted the same grid-floor circuit, so their difference is exactly zero; that circuit kept the clean top prediction on 82.0% of prompts. Paths were not re-optimized.

Other settings. In the scope study, FIELD−STRICT top-1 differences at the 80% operating point were negative in 12 of the 12 mean-ablation settings (point estimates), with 10 nominal block-bootstrap intervals excluding zero. Most differences attenuated at a common accepted size. STRICT fell back to the full universe in 22 of 192 setting–block units at 80% and in 100 at 90%. All were retained, so fallback increasingly shapes comparisons at stricter thresholds. The settings share model–task blocks and are not independent replications. In the 8B induction cell the own-pick difference was large but concentrated at one size. This post hoc result does not change that experiment's failed composite gate (Appendix E).

6 Implications for Reporting Circuits

Report four choices with every circuit: the aggregation, the ablation operator, the candidate universe and the endpoint. Each can change the result. The right criterion depends on what the circuit should preserve. For an aggregate task score, state the aggregation and test whether the conclusion depends

on cross-prompt cancellation, e.g. by reporting both selections when the aggregation orders disagree. For per-prompt task-score fidelity, score prompts individually so signed effects cannot cancel. For the output distribution or the model's predictions, select or at least validate on that endpoint. Neither task-score aggregation optimizes it; importance and discovery-KL selection preserve some of these endpoints better than both (Table 2; Appendix C). Always evaluate the selected circuit on held-out prompts and report absolute levels alongside ratios.

7 Related work

Aggregation and cancellation. Scheurer et al. (2023) identify cancellation under averaged loss, and Kramár et al. (2024) define a cancellation ratio. Miller et al. (2024, §§4–5) show that faithfulness depends on ablation semantics and on averaging before versus after recovery normalization, with substantial prompt-level variation. Haklay et al. (2025) show that aggregation changes top-K circuit selection but do not test the selected circuits on held-out outcomes. Méloux et al. (2026, §§3–5) treat circuit discovery as statistical estimation and quantify sampling and methodological variability via prediction disagreement and KL. Some of their EAP-IG configurations differ in aggregation, but jointly with other changes.

Multiplicity and evaluation of circuits. Several circuits can meet the same criteria (Méloux et al., 2025), structurally different circuits can implement interchangeable computation (Bayat Makou et al., 2026), and circuit-level conclusions can change across defensible analytic choices (Mahale, 2026). Evaluation work decomposes patching estimands (Vaidyanathan et al., 2026), warns against subspace-patching illusions (Makelov et al., 2024), certifies interventional fidelity (Asiaee, 2026), tests circuit hypotheses (Shi et al., 2024), analyses discovery variance (Wu et al., 2026) and formalizes intervention semantics (Geiger et al., 2025).

What is different here. The closest work tests aggregation's effect on selection without held-out preservation, varies it jointly with other choices, or studies broader variability among recovered circuits. We change only the aggregation order, $|\mathbb{E}\,\Delta Y(C)|$ versus $\mathbb{E}|\Delta Y(C)|$, and hold the candidate universe, prompts, circuit size and exhaustive search fixed. We repeat selection over independent draws, score held-out distribution and prediction preservation, and test operator and universe scope. We do not claim to be the first to observe circuit instability or behavioural disagreement between circuits.

8 Limitations and conclusion

The aggregation results are conditional on the tested prompt populations, candidate universes and operators. The mean-ablation contrast is strongest in the published IOI universes; on the registered relative scale it is confirmed in only two of six scope families. KL's aggregation structure is closer to STRICT than to FIELD. In the published universes, much of the excess divergence lies in answer-set mass that neither task-score criterion directly constrains. The Pythia prediction effect is specific to that universe, and the Qwen3-8B extension did not pass its registered materiality gate. A later calibration audit found the twelve-block procedure anti-conservative at the upper margin, which limits the registered counterfactual-equivalence verdict. The selections are partial ablations against an intact background. We test exhaustive selection inside fixed head universes, which may favour cancelling candidates more than ranked search (§5). We do not test attribution-based discovery over whole-model graphs, where aggregation can enter both edge scoring and circuit-size choice; whether the same consequences appear there is open. The effect is established at the tested discovery size (50 prompts per block), where average-first optima were far less stable (Appendix E). Whether the gap persists with larger discovery samples has not been tested prospectively. As circuit discovery becomes more automated, faithfulness claims should name the aggregation and intervention protocol and directly evaluate the property they intend to preserve.

Reproducibility statement

Every estimate, interval and test result in the paper is emitted by a claim ledger, or for tables by the generator `gen_tables2.py`, from hashed artifacts (design constants, head names and a few

descriptive counts and bounds are typed in the text); both ledgers verify inside a fresh unpacking of the distributed supplement (the first, which covers the addendum's design analyses, re-reads raw captures that only the full supplement contains). The repeated-discovery and scope studies, the 8B extension, the fresh-template transfer study and the restriction replication were preregistered and frozen before official evaluation; the block-level analyses of the first two were independently reimplemented from the sealed records, and the last two were independently verified before inclusion. The supplement contains the preregistrations and amendments, the complete 50-study registry, analysis code, the reported block and lattice artifacts, artifact hashes, the bibliography audit and the build files; large raw discovery banks are omitted but pinned by recorded digests. The registry distinguishes preregistered, exploratory, failed, corrected, stopped and invalid studies. A separate technical addendum contains the correction and provenance history of earlier intervention-design and search work; no unavailable result from ongoing follow-ups supports this submission.

AI USE STATEMENT

The authors conceived and directed this research and made the final decision on every hypothesis, design and claim. Generative AI tools assisted with the following tasks with required disclosure: proposing and refining hypotheses and preregistered study designs; designing experiments and giving feedback on methodology, including adversarial review of every protocol before it was frozen; helping develop the theoretical and conceptual framework; formulating mathematical claims and assisting with proofs (every algebraic statement was re-derived symbolically, and the proofs are in Supplement S16); implementing methods and analysis code; drafting prompt templates for the synthetic task sets, which are then generated programmatically; screening the fresh-template transfer set, in which two independent model-based reviewers audited every clean template and its twin for semantic validity, templates that either reviewer failed were replaced before any execution, and their naturalness judgments are reported (Appendix A.1); and interpreting results, including adversarial review of every claim against the frozen artifacts. That review caught the support confound of Supplement S2.3.1, a lattice-ordering defect in a development analysis, and provenance errors, all recorded in a correction log. We have not used generative AI tools to generate or alter any experimental measurement; translation is not applicable to this work.

Additionally, we used generative AI tools to create and edit code, draft and edit the text, design figures, search and summarize the literature, and format references.

We have reviewed all AI-assisted work. Two safeguards constrain it: every estimate, interval and test result is emitted mechanically from hashed artifacts by the claim ledgers and the table generator rather than written by hand, and every cited work was checked against its primary source (the audit is in the supplement). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Uncertainty in Artificial Intelligence (UAI)*, volume 337 of *PMLR*, pp. 210–227, 2026.

Alireza Bayat Makou, Jingcheng Niu, Subhabrata Dutta, and Iryna Gurevych. Many circuits, one mechanism: Input variation and evaluation granularity in circuit discovery. *Transactions on Machine Learning Research*, 2026. arXiv:2606.06267.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. Position-aware automatic circuit discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2792–2817, 2025.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling (COLM)*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255, 2024.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Ajay Pravin Mahale. Explanation multiplicity: Circuit-level interpretability evidence does not survive defensible analytic variation. *arXiv preprint arXiv:2608.13754*, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations (ICLR)*, 2025.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2510.00845.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling (COLM)*, 2024.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, 2025.

Jérémy Scheurer et al. Practical pitfalls of causal scrubbing. LessWrong / Alignment Forum, `https://www.lesswrong.com/posts/DFarDnQjMnjsKvW8s/practical-pitfalls-of-causal-scrubbing`, 2023.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 94539–94567. Curran Associates, Inc., 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *BlackboxNLP*, 2024. arXiv:2310.10348.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

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

Frank Zhengqing Wu, Francesco Tonin, and Volkan Cevher. Demystifying variance in circuit discovery of LLMs. *arXiv preprint arXiv:2606.16920*, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

A THE REPEATED-DISCOVERY STUDY AND THE EARLIER CONDITIONAL EXPERIMENT IN FULL

Selectors. For a candidate circuit C inside a fixed head universe U, every head of $U \setminus C$ is ablated (mean ablation: final-token dev means of each head on GPT-2, role-aligned means at every position on Pythia; the corrupted-run donor under counterfactual patching), all other heads are left intact, and the per-prompt change $\Delta Y_p(C)$ in the IO−S logit difference is recorded. FIELD selects, per size, $\arg\min_C |\frac{1}{n}\sum_p \Delta Y_p(C)|$ and STRICT selects $\arg\min_C \frac{1}{n}\sum_p |\Delta Y_p(C)|$, both by exhaustive enumeration over all subsets of that size. In P9, P10 and P19B two reference selectors are reported: an importance ranking (IMP: retain the heads with the largest single-ablation effect) and random same-size subsets; P30, P32 and P28 add exhaustive selection on discovery KL. The evaluation endpoint is $\mathrm{KL}(P_{\text{full}} \,\|\, P_C)$ on 200 evaluation prompts from held-out templates (in P19B and P30 the names are drawn from the same 16-name pool as discovery; P9 and P10 also held out the names), normalized by the KL of the full ablation and averaged over sizes.

Where each form is used. Wang et al. (2023) define faithfulness as $|F(M) - F(C)|$ with F a prompt average (§4; completeness and minimality likewise): FIELD. MIB's circuit metric (Mueller et al., 2025) is $f = (m(\mathcal{C}) - m(\varnothing))/(m(\mathcal{N}) - m(\varnothing))$ (Eq. 1) with m averaged over examples before the ratio in the released evaluation code, so the CMD integrand $|1 - f|$ is FIELD up to a C-independent normalizer, and CPR is a signed ratio of averages. The released EAP and EAP-IG code sums signed per-prompt attributions before taking magnitudes (FIELD-form edge scores), whereas the EAP paper describes a mean of absolute attributions. AtP* (Kramár et al., 2024) scores nodes by $|\mathbb{E}[\mathcal{I}(n)]|$ (Eq. 1, FIELD) and estimates with $\mathbb{E}[|\hat{\mathcal{I}}(n)|]$ (Eqs. 5, 11, STRICT); its cancellation ratio is $(\text{STRICT} - \text{FIELD})/\text{STRICT}$. Scheurer et al. (2023) show that averaged loss cancels and recommend inspecting per-sample losses. Exact code locations are in the supplement's bibliography audit.

The cancellation identity. With $P = \sum_{\Delta Y_p > 0} \Delta Y_p$ and $N = \sum_{\Delta Y_p < 0} (-\Delta Y_p)$, $\text{FIELD} = |P - N|/n$ and $\text{STRICT} = (P + N)/n$, so $\text{STRICT} - \text{FIELD} = (P + N - |P - N|)/n = 2\min(P, N)/n$; the two agree exactly when one of P, N is zero. This is elementary and is the quantity that the cancellation ratio of Kramár et al. (2024) normalizes; we use it only to state precisely what the aggregation discards.

Three runs. P9 (GPT-2 small, ten published IOI heads, the original five discovery templates) and P10 (Pythia-160m, the fourteen-head circuit of Tigges et al. (2024), role-aligned donors) were preregistered and run once each; their primary-family FIELD−STRICT differences in normalized KL AUC are 0.201 $[0.123, 0.293]$ and 0.347 $[0.302, 0.425]$, and both were reproduced by blinded reimplementations from the preregistration text. P9 also registered three other execution families on GPT-2: with mean ablation but downstream heads clamped to their clean values (no recomputation) the difference was -0.019 $[-0.082, 0.017]$, with zero ablation 0.111 $[0.019, 0.229]$ (recomputing) and 0.023 $[-0.004, 0.056]$ (clamped): in P9 the contrast appeared only when downstream heads recompute, which fits its association with the negative Name Movers (§4). P19B, the first prospective replication, used 15 new templates and 16 new names drawn before any model run (the evaluation *templates* are held out from discovery; the names are not — one pool of 16 serves dev, discovery and evaluation, as preregistered), 50 discovery and 200 evaluation prompts, identical prompts under both operators, a frozen preregistration with one permitted analysis run, and dual intervals (a $10,000$-replicate template-cluster bootstrap and a template-level t interval) both required for every verdict. Its verdicts are literal: the mean-ablation hypothesis is CONFIRMED on both models; the counterfactual hypothesis is "EQUIVALENT (small positive)" on both, meaning both 90% intervals lie inside $\pm \log 1.25$ with lower bounds above zero (GPT-2 bootstrap $[0.080, 0.166]$, t $[0.094, 0.182]$; Pythia $[0.067, 0.148]$, $[0.014, 0.196]$); the interaction is CONFIRMED ($p = 0.002$ and $p < 10^{-4}$).

Table 6: P30, every block: the log ratio L of FIELD's to STRICT's normalized held-out KL per block and operator, and the size-averaged exact-mask agreement of each selector between independent blocks. Generated from the sealed record.

block	0	1	2	3	4	5	6	7	8	9	10	11
GPT-2 L, mean	0.57	0.85	0.42	0.98	0.55	0.04	0.39	0.43	1.08	0.32	0.17	0.56
GPT-2 L, cf	0.31	0.05	0.09	0.04	0.29	0.14	0.06	0.13	0.12	-0.04	-0.04	0.08
Pythia L, mean	1.22	1.03	1.62	1.61	1.34	0.63	1.73	1.02	1.65	1.58	1.08	1.09
Pythia L, cf	0.16	0.13	0.06	0.16	0.07	0.05	0.29	0.06	0.14	0.05	0.08	0.13

size-averaged exact-mask agreement between independent blocks (%)
GPT-2, mean: field 5.4, strict 32.9, imp 48.3, disc-kl 44.4
GPT-2, cf: field 24.0, strict 43.9, imp 74.2, disc-kl 37.9
Pythia, mean: field 0.7, strict 25.6, imp 43.6, disc-kl 49.8
Pythia, cf: field 37.0, strict 43.4, imp 65.0, disc-kl 49.3

The first P19B attempt failed its own state-zero validity gate on a float32 batch-size artifact and was restarted under a stricter gate; it is reported as a failed attempt.

P30: the procedure repeated. P19B's intervals condition on one discovery draw and the circuits it selected. P30 (a new study; it changes none of P19B's estimands, data or verdicts) draws 12 independent blocks per model from the 21 templates that passed the P19 screen and the 16 P19B names: per block 5 discovery and 10 evaluation templates without replacement, 8 donor and 10 discovery prompts per discovery template, 20 evaluation prompts per evaluation template, excluding every previously exposed (template, name, name) triple recorded in a frozen exposure ledger. The same blocks serve both models and operators. Per block the full discovery lattice is collected, FIELD, STRICT, IMP and DISC-KL select exhaustively per size, and held-out KL is normalized by that block's full-ablation KL. L is the log ratio of the mean normalized risks over blocks; inference uses a 10,000-replicate block bootstrap and a delete-one-block jackknife t interval, claim p values are the larger of the two, and Holm runs over the 6 claims. The jackknife 95% intervals for the mean-ablation ratio are $[1.41, 2.11]$ (GPT-2) and $[2.99, 4.45]$ (Pythia); for the counterfactual ratio the 90% jackknife intervals are $[1.05, 1.18]$ and $[1.08, 1.16]$. Table 6 gives every block. Per-size ratios under mean ablation range over 1.18–3.98× (GPT-2, $7/7$ sizes positive) and 1.61–23.14× (Pythia, $9/9$); these are point estimates, not simultaneous tests. P19B's single-draw L lies above $4/12$ of the GPT-2 blocks and $2/12$ of the Pythia blocks, which places it within the block distribution; it is not a second replication statistic. The chain-rule components sum to the KL to numerical precision in every block (identity and bridge checks in the sealed record). One exact tie occurred in FIELD's discovery objective (Pythia, block 11, mean ablation, size 10); the registered lowest-index rule broke it and the tied alternative was never evaluated. Post hoc, giving that alternative any normalized AUC from 0 to 2 moves Pythia's L only within $[1.293, 1.341]$ (observed 1.295). Under counterfactual patching the raw gap is 0.015 and 0.009 nats, below the preregistered 0.02-nat floor for a component label.

Direction of the answer-mass excess (post hoc). The MASS term is the binary KL between the clean and the intervened total probability of the two names, so it does not say whether the circuit lost or gained answer mass. Splitting it by prompt according to that direction (a decomposition suggested by an external review of the sealed results, not preregistered), FIELD's normalized excess under mean ablation comes from prompts that *gain* answer mass on GPT-2 (0.244 against -0.024 from prompts that lose it; gain-side excess positive in $12/12$ blocks) and from prompts that *lose* it on Pythia (0.225 against -0.011; loss-side excess positive in $12/12$ blocks). "Answer mass collapses" is therefore not a general description, and gaining answer mass on the two names combined is not automatically better task behaviour; what the two models share is that the logit-difference criteria leave this quantity unconstrained.

Universe composition (exploratory, post hoc). On the twelve GPT-2 P30 blocks we evaluated all 120 seven-of-ten restrictions of the published universe: the three listed heads of a restriction are ineligible for ablation (they stay in the model and recompute under P30's semantics), FIELD and STRICT select exhaustively at sizes 2–6 from the stored discovery lattices, and held-out states not already evaluated were executed post hoc under P30's frozen semantics (re-execution of stored states reproduced the stored readouts bit-exactly). The statistic is the log ratio of block-mean raw held-out KL, with block resampling shared across all restrictions. Making $\{10.7, 11.10, 9.0\}$ ineligible gives

0.005 $[-0.13, 0.15]$ under mean ablation (rank 1/120, below the median restriction in 12/12 blocks); under counterfactual patching -0.002 (rank 4/120), where it reflects largely identical selections. All eight restrictions that make both negative Name Movers ineligible fall in the lower tail, and making both ineligible changes the log ratio by -0.38 $[-0.57, -0.20]$ on average against -0.03 $[-0.11, 0.06]$ for either alone. The pooled ratio is a STRICT-risk-weighted average of size-specific ratios, and sizes 5 and 6 still show log ratios of 0.33 and 0.26 on small absolute risks, so the pooled difference is near zero rather than absent. The trio was nominated after P30 was inspected; rank 1 is a stability summary, not a p value, and the restriction also changes the admissible interventions and the clean background. GPT-2's top-1 and accuracy differences are below two points in every restriction, so this analysis does not bear on the Pythia prediction cost.

Examples of same-size selections (post hoc). Among the 108 same-size Pythia mean-ablation pairs, 7 satisfy a pre-stated eligibility rule (the two selections differ; held-out FIELD errors and mean logit differences both within 0.05 logits). The pair with the median top-1 gap (block 1, four of fourteen heads) does not reproduce the population pattern: both selections overshoot the task score, and most of FIELD's prediction changes move *into* the answer pair. The pair with the largest gap (block 2, eleven heads; FIELD ablates a mixed-sign head triple whose single-head effects do not add to its joint effect) shows the population pattern: matched logit difference and accuracy, top-1 agreement 73.5% against 96.5%, with the correct name displaced by high-frequency tokens. We therefore rest the main text on the population-level analyses rather than on a single example; the full table of the seven pairs and both examples are in the supplement (`probes/tradeoff_example`). The seven pairs of this stricter rule are a different subset from the $15/108$ pairs of the main text, which are matched on held-out FIELD error only. The FIELD$-$STRICT top-1 difference ranges from -14.7 to -9.8 points across 10 matching tolerances (fixed post hoc) and is -15.6 points in the 60 of all 108 pairs where FIELD is not worse on its own held-out criterion; the delete-one-block jackknife interval for the 0.05-logit rule is $[-15.9, -4.7]$.

Prediction cost by clean top token and margin (post hoc). Table 7 re-executes every selected Pythia mean-ablation circuit (bit-identical to the stored readouts) and splits the top-1 difference by the clean model's top token and by its top-two logit margin, with strata fixed before the re-executed outputs were read. The marginal pattern by margin is a composition effect: prompts with small margins mostly have a clean top token outside the answer pair, where FIELD's circuits keep it *more* often. Where the clean top token is the correct name, the loss is largest at small margins but persists at the largest, so it is not only a reordering of near-ties. STRICT's circuits more often move the top token onto the correct name (5.2 such changes per 100 against 2.2).

clean top token	Q1 [0.00, 0.42]	Q2 [0.42, 0.88]	Q3 [0.88, 1.52]	Q4 [1.52, 3.46]	all
correct name (IO)	-45.1 $[-54.9, -35.1]$ (176)	-47.1 $[-52.4, -41.7]$ (239)	-32.5 $[-37.9, -27.3]$ (373)	-16.4 $[-19.4, -13.9]$ (538)	-30.3 $[-34.4, -26.5]$ (1326)
outside the pair	$+10.2$ $[+6.9, +13.6]$ (422)	$+4.7$ $[+2.0, +6.8]$ (361)	$+1.9$ $[+0.3, +3.7]$ (227)	$+0.0$ $[-2.6, +2.8]$ (62)	$+6.0$ $[+4.1, +7.7]$ (1072)
all	-6.1 $[-9.2, -3.6]$ (599)	-16.0 $[-19.8, -12.7]$ (601)	-19.5 $[-24.2, -15.2]$ (600)	-14.7 $[-17.5, -12.3]$ (600)	-14.1 $[-16.7, -11.5]$ (2400)

Table 7: Pythia, mean ablation: FIELD$-$STRICT top-1 agreement (points; 95% block-cluster bootstrap; (block, prompt) count) by clean top token and by quartile of the clean top-two logit margin (range in logits). Two prompts whose clean top token is the distractor are in "all" only.

Threshold sensitivity in P19B (mean ablation). P19B also registered a nested threshold protocol, reported here and not tabulated: heads enter in the order of their single-ablation importance on discovery prompts, and each criterion keeps the shortest prefix whose discovery faithfulness reaches the threshold, so the two criteria can differ only in where they stop. At the 80% threshold FIELD stops after one head on both models. At 85% it admits STRICT's 7-head circuit on GPT-2 while FIELD still stops after one head on Pythia; at 90% both criteria select the same circuit on GPT-2, and on Pythia FIELD reaches eight heads against STRICT's ten. The effect is a property of the threshold as well as the operator (under counterfactual patching neither criterion stops at one head), so we report all three thresholds.

Size-averaged, not absent. The counterfactual equivalence was declared (as registered) on the size-averaged endpoint, not in the circuits themselves: the sign coherence $|\overline{\Delta Y}|/\overline{|\Delta Y|}$ of the complements FIELD selects is near zero at most sizes under *both* operators, and generally higher for

STRICT's (the preregistered mediator in the sealed record), so cancellation is present under counterfactual patching too and costs less held-out KL there. In the earlier single draw, cancellation still produced a FIELD/STRICT ratio above 2 on GPT-2 at size 6, and in P19B the FIELD−RANDOM contrast under mean ablation is inconclusive on GPT-2 and favours FIELD on Pythia (in P30 it is inconclusive on both models under the both-intervals rule: log ratio of normalized KL -0.145 and -0.094, bootstrap $[-0.283, -0.012]$ and $[-0.246, 0.050]$, jackknife $[-0.305, 0.015]$ and $[-0.271, 0.084]$), so the earlier P10 finding that FIELD selects worse than random (0.068 $[0.035, 0.097]$) did not replicate and is not claimed.

Universe sensitivity (exploratory). Resampling 200 random ten-of-fourteen sub-universes of the Pythia circuit keeps the sign of the FIELD−STRICT contrast in 98.5% of draws (median ratio $2.07\times$) while moving its magnitude over a $2.4\times$ range between the 10th and 90th percentiles. We first read that spread as a property of the universe. A later exploratory check contradicts that reading: the same 200 sub-universes evaluated on a second, independent role-mean prompt set (P10 against P19B) give per-universe log ratios that correlate only 0.28 across the two sets, and FIELD's selected masks agree between the two sets in 4% of (universe, size) cases against 31% for STRICT. Most of the spread is therefore selection variability tied to the discovery draw, concentrated in FIELD's picks. It is not general: in P28's architectural-group universes the two criteria's picks were comparably reproducible across independent blocks (66% against 70% in cell A under role-mean replacement). Headline magnitudes are therefore conditional on the discovery draw as well as on the (universe, operator) pair. P30 later measured pick reproducibility prospectively, as a registered secondary analysis (Section 3, Table 6); the exploratory comparison above motivated that analysis and is not additional evidence. The two vary different things, so the universe-conditioning caveat stands.

Mechanism (exploratory, on the P9 lattices). In the complements of minimal FIELD-faithful circuits, most of the individual-effect magnitude cancels at first order by opposite signs, and on both models nearly every minimal FIELD-faithful circuit fails hereditary faithfulness per prompt: some sub-ablation of its complement breaks tolerance. (This mechanism analysis produced no sealed artifact, so we give it qualitatively rather than as ledger numbers, and we make no claim about alternative explanations such as self-repair or head-class composition.) STRICT is not advertised as an objective: under mean ablation IMP beats exhaustive STRICT on both models (normalized KL AUC 0.314 against 0.475 on GPT-2, 0.115 against 0.133 on Pythia, fresh prompts), an objective mismatch between a logit-difference criterion and a KL endpoint; under counterfactual patching it is worse than both aggregations (0.261 and 0.066). Exhaustive selection by discovery KL comes within a factor 1.047 of the test-informed oracle's held-out KL in three of the four model×operator cells and 1.202 in the fourth, so selecting on discovery prompts is not in itself what makes FIELD's circuits preserve less. Nor is a failure to generalize: in the repeated-discovery study FIELD's excess raw KL on the discovery prompts themselves is 0.048 $[0.035, 0.063]$ and 0.133 $[0.099, 0.178]$ nats, as large as held out, so the KL gap is not a generalization gap; FIELD's own criterion, by contrast, loses its in-sample advantage held out.

A.1 FRESH-TEMPLATE TRANSFER OF THE FROZEN SELECTIONS (P34)

Design. P34 keeps every selection of the repeated-discovery study fixed (all selectors, sizes and operators in its 12 GPT-2 and 12 Pythia blocks) and evaluates it on a new bank of 20 IOI templates in 5 frames, written blind to the tested models. No new template introduces the two names as one coordinated noun phrase, as every earlier template in the audited inventory does (the original bank included); the frames vary first-mention packaging, name order and sentence boundaries and keep the transfer-dative query. Two independent model-based reviewers audited every clean template and its twin (three were judged mildly more natural in the counterfactual direction, which limits a purely syntactic reading of transfer). Block b pairs block b's sealed selections and original donors with 200 new prompts, and normalized endpoints divide by the full ablation on the new prompts. No discovery is repeated: the study measures transfer of the selected circuits to this fixed target bank of new templates (non-coordinated first mentions, retaining the transfer-dative query), not to a template population. One Holm family of 5 slots, never shrunk: H1 per model ($L > 0$, mean ablation), H2 in Pythia (STRICT−FIELD top-1 above 2 points) and H3 per model (counterfactual ratio in $[0.8, 1.25]$); each p is the larger of the block-bootstrap and jackknife p. The first official launch was refused before any study output was written because Python 3.13's multiprocessing im-

port aliases the frozen entry module as `__mp_main__`, while the frozen provenance guard allowed only the name `__main__`. A pre-data implementation amendment permitted only the identity alias `sys.modules['__mp_main__'] is sys.modules['__main__']`; distinct modules under that name remain refused. No scientific design, input, seed, intervention, estimator, or analysis rule changed. Two clean-room reimplementations, a provenance audit and a preregistration audit verified the official analysis before inclusion (supplement, `VERIFICATION34.md`); they found one implementation detail: one-sided bootstrap tail probabilities of discrete statistics resolve resampled differences of exactly zero by floating-point summation order rather than counting them in both tails, which moves no tail probability by more than 0.0041 and changes no verdict or printed value.

Registered claims, gates and competence. Table 9 gives all five slots. The literal verdicts were "H1 CONFIRMED" for GPT-2's H1, "H1 CONFIRMED" for Pythia's H1 and "H2 CONFIRMED" for Pythia's H2 (secondary label: none). The H1/H2 validity gate was passed for GPT-2 and passed for Pythia; the H3 gate, which needs every frame to pass on its own ($5/5$ and $5/5$ did), was passed for GPT-2 and passed for Pythia (Table 10). Per-template clean correctness was at least 75% (GPT-2) and 74% (Pythia; Table 13); no block or template is ever excluded.

Raw and normalized risks, top-1 and accuracy. Table 11 gives every selector's levels. Under mean ablation the FIELD/STRICT raw-KL ratio was $1.49\times$ $[1.19, 1.85]$ (GPT-2) and $2.46\times$ $[2.16, 2.78]$ (Pythia; raw, distinct from the normalized H1 ratios); FIELD$-$STRICT differences were -2.8 and -7.9 points in top-1 agreement and -4.0 and -0.9 in accuracy; the difference on FIELD's own held-out criterion was "unresolved" for GPT-2 and "unresolved" for Pythia (an unresolved difference is not evidence of a match).

Frames (descriptive, unadjusted). The registered frame contrasts (Table 8) give positive point estimates of L in every frame of both models under mean ablation, with the smallest values in GPT-2's frame F2 (0.26) and Pythia's frame F8 (0.51). The Pythia top-1 result is not uniform: in frame F8 the STRICT$-$FIELD top-1 gap reverses to -4.8 points ($[-7.4, -2.4]$; jackknife $[-7.8, -1.9]$), so the pooled H2 result is carried by the other four frames.

frame	GPT-2 L	Pythia L	GPT-2 top-1 gap	Pythia top-1 gap
F1	0.57 $[0.35, 0.78]$	0.91 $[0.77, 1.04]$	2.2 $[0.1, 4.3]$	9.3 $[6.1, 12.8]$
F2	0.26 $[0.10, 0.42]$	0.94 $[0.75, 1.14]$	3.9 $[-0.1, 7.7]$	11.5 $[8.0, 14.5]$
F5	0.43 $[0.10, 0.73]$	1.05 $[0.90, 1.19]$	4.2 $[-1.3, 8.9]$	14.5 $[12.3, 16.6]$
F7	0.43 $[0.02, 0.77]$	1.00 $[0.86, 1.14]$	1.8 $[-1.9, 5.1]$	8.8 $[7.0, 10.5]$
F8	0.37 $[0.24, 0.51]$	0.51 $[0.40, 0.63]$	2.1 $[-0.7, 5.0]$	-4.8 $[-7.4, -2.4]$

Table 8: P34 per-frame contrasts under mean ablation (descriptive, unadjusted; 95% block-bootstrap intervals). L: log ratio of FIELD's to STRICT's normalized held-out KL; top-1 gap: STRICT$-$FIELD clean-top-token agreement, points (positive: FIELD preserves fewer predictions).

Paired comparison with the original blocks (descriptive). For context, pairing each frozen selection with its original P30 block gives $L_{\text{new}} - L_{\text{old}} = -0.15$ $[-0.32, 0.01]$ in GPT-2 and -0.39 $[-0.48, -0.33]$ in Pythia. This is a descriptive change in effect magnitude, not a registered equivalence or retention test. The template family, the target prompts, the normalizer and the syntax all changed together, so no change is attributed to any one of them (Table 12).

Counterfactual patching (descriptive). Counterfactual estimates are descriptive only because the preregistered twelve-block equivalence procedure failed its pre-run calibration audit. In that audit the procedure declared equivalence in 9.4% (GPT-2) and 9.2% (Pythia) of simulated experiments at the upper margin, against a nominal 5% (Appendix E), so both H3 slots enter Holm at $p = 1$. The counterfactual FIELD/STRICT ratio was $1.06\times$ in GPT-2 (90% bootstrap $[0.95, 1.23]$, jackknife $[0.92, 1.23]$) and $1.16\times$ in Pythia ($[1.13, 1.20]$, $[1.13, 1.20]$); no equivalence or non-equivalence is claimed.

A.2 Fresh replication of the nominated universe restriction (P35)

Design. The restriction analysis of Appendix A made three of the ten GPT-2 IOI heads ineligible for ablation (they stay in the model and recompute) in each of the 120 possible ways and,

model	slot	estimand	estimate	95% bootstrap	95% jackknife	p	Holm p	verdict
GPT-2	H1	FIELD/STRICT normalized KL, mean abl.	1.49	$[1.18, 1.85]$	$[1.14, 1.96]$	0.004	0.012	confirmed
GPT-2	H3	FIELD/STRICT normalized KL, counterf.	1.06	$[0.93, 1.26]$	$[0.90, 1.27]$	–	1.000	descriptive
Pythia	H1	FIELD/STRICT normalized KL, mean abl.	2.46	$[2.15, 2.79]$	$[2.11, 2.87]$	$\leq 10^{-4}$	5.0×10^{-4}	confirmed
Pythia	H2	STRICT −FIELD top-1 (points; margin 2)	7.9	$[6.2, 9.4]$	$[6.0, 9.7]$	$\leq 10^{-4}$	5.0×10^{-4}	confirmed
Pythia	H3	FIELD/STRICT normalized KL, counterf.	1.16	$[1.12, 1.20]$	$[1.12, 1.21]$	–	1.000	descriptive
H3 boundary contrasts (descriptive; decided nothing): $\mathrm{E}[X - 0.8Y]$ and $\mathrm{E}[1.25Y - X]$, X, Y = FIELD's, STRICT's normalized KL								
GPT-2	H3	$\mathrm{E}[X - 0.8Y]$	0.062	$[0.034, 0.088]$	$[0.030, 0.094]$			
GPT-2	H3	$\mathrm{E}[1.25Y - X]$	0.043	$[-0.002, 0.089]$	$[-0.010, 0.097]$			
Pythia	H3	$\mathrm{E}[X - 0.8Y]$	0.012	$[0.011, 0.013]$	$[0.010, 0.013]$			
Pythia	H3	$\mathrm{E}[1.25Y - X]$	0.003	$[0.002, 0.004]$	$[0.001, 0.004]$			

Table 9: P34, the five registered slots (one Holm family; p is the larger of the bootstrap and jackknife p; H3 enters at $p = 1$ after its failed calibration audit) and the H3 boundary contrasts. Ratios are FIELD/STRICT; intervals are unadjusted.

	GPT-2				Pythia			
prompts	correct (%)	median Δ	flip (%)	gate	correct (%)	median Δ	flip (%)	gate
all (H1/H2 gate)	96.2	2.45	95.7	pass	96.2	3.29	88.7	pass
F1	96.5	2.40	97.9	pass	90.0	1.79	99.8	pass
F2	96.7	1.66	91.2	pass	100.0	4.65	70.8	pass
F5	99.6	2.43	94.4	pass	100.0	3.33	79.4	pass
F7	96.0	2.67	95.8	pass	99.0	3.73	94.0	pass
F8	92.3	3.26	99.2	pass	92.3	2.87	99.4	pass
full abl. $\lvert\Delta Y\rvert$, mean / cf. (≥ 0.5)		2.96 / 5.63				3.19 / 5.99		
H3 gate (all frames, flips, floors)		pass				pass		

Table 10: P34 validity gates on the new prompts, pooled over blocks: clean correctness, median clean logit difference and twin flip, overall (the H1/H2 gate) and per frame (each must pass for H3).

after inspecting the repeated-discovery blocks, nominated the restriction U^* that makes 10.7 and 11.10 (negative Name Movers) and 9.0 ineligible. P35 repeats the protocol on 12 fresh blocks (same model, universe, intervention and prompt generation), sealing every restricted pick before evaluation. The blocks are new prompt draws from the 20 templates of the original bank that still had enough unexposed instances, not new templates (contrast Appendix A.1); as in the repeated-discovery study, instances may recur across blocks. With $R_{X,U,b}$ selector X's raw held-out KL averaged over retained sizes 2–6, $G_U = \log(\sum_b R_{F,U,b} / \sum_b R_{S,U,b})$ and $A_U = \mathrm{mean}_b(R_{F,U,b} - R_{S,U,b})$, the two registered claims are $C_G = \mathrm{median}_{U \neq U^*} G_U - G_{U^*} > 0$ and $C_A = \mathrm{median}_{U \neq U^*} A_U - A_{U^*} > 0$ (mean ablation; Holm over the two; every statistic and the median recomputed in each of $10,000$ bootstrap replicates and each jackknife deletion; p the larger of the two). A confirmed claim replicates the restriction-specific effect, not a prospective discovery of the trio; the rank of U^* is no criterion. Two clean-room reimplementations, a provenance audit and a preregistration audit verified the official analysis before inclusion (supplement, `VERIFICATION35.md`); every figure, table and number below reproduces, and bootstrap tail probabilities of the discrete top-1 contrast, which this appendix does not print, depend on floating-point ties.

The two primary contrasts. $C_G = 0.409$ (bootstrap $[0.207, 0.605]$, jackknife $[0.149, 0.668]$, Holm $p = 0.005$): REPLICATED. $C_A = 0.0188$ nats ($[0.0098, 0.0274]$, $[0.0070, 0.0306]$, Holm $p = 0.005$): REPLICATED (Table 14; every restriction in Figure 4).

Absolute risks and which risk changed. Under U^*, FIELD's size-averaged held-out KL was 0.0401 nats against a median of 0.0581 over the other restrictions and STRICT's 0.0384 against 0.0351; on the log scale (negative: lower under U^*) $\Delta_F = -0.37$ $[-0.52, -0.19]$ and $\Delta_S = 0.09$ $[-0.02, 0.22]$ (Table 15). A smaller contrast from STRICT's risk rising is not an improvement of FIELD. The analysis's point reading against the single median-G restriction is "FIELD's risk is lower and STRICT's is higher (the ratio is smaller): the smaller ratio comes from FIELD improving and STRICT worsening"; in the registered decomposition only FIELD's decrease has intervals excluding zero, so STRICT worsening is not established. The separation is smaller than in the repeated-discovery blocks in which U^* was nominated ($G_{U^*} = 0.005$ against a median of 0.52 there; 0.043 against 0.452 here), as expected for a restriction chosen as the extreme of 120.

		mean ablation						counterfactual					
model	endpoint	FIELD	STRICT	IMP	DISC-KL	random	full	FIELD	STRICT	IMP	DISC-KL	random	full
GPT-2	normalized KL	0.345	0.231	0.264	0.212	0.367	–	0.250	0.235	0.419	0.218	0.276	–
	raw KL (nats)	0.185	0.124	0.141	0.114	0.196	0.537	0.285	0.266	0.479	0.249	0.315	1.143
	top-1 (%)	73.8	76.7	73.0	76.2	73.2	–	64.5	64.0	48.5	64.8	60.7	–
	accuracy (%)	80.9	84.9	78.2	82.5	79.9	41.5	64.8	62.7	42.4	63.1	57.7	1.6
Pythia	normalized KL	0.367	0.149	0.134	0.114	0.446	–	0.038	0.032	0.035	0.030	0.414	–
	raw KL (nats)	0.139	0.057	0.051	0.043	0.169	0.380	0.024	0.021	0.023	0.020	0.269	0.649
	top-1 (%)	77.1	85.0	85.1	86.5	72.9	–	89.7	90.9	90.7	91.5	66.3	–
	accuracy (%)	91.7	92.5	93.0	92.8	77.7	53.7	91.4	92.7	89.4	92.3	54.6	13.1

Table 11: P34, every selector's size-averaged levels on the new family (block means); "full" ablates the whole universe.

model	operator	L_{new}	L_{old}	$L_{\text{new}} - L_{\text{old}}$	95% bootstrap	95% jackknife	raw-KL Δ	old-denominator Δ
GPT-2	mean abl.	0.40	0.55	-0.15	$[-0.32, 0.01]$	$[-0.35, 0.06]$	-0.14	-0.13
GPT-2	counterf.	0.06	0.11	-0.04	$[-0.22, 0.16]$	$[-0.26, 0.17]$	-0.04	-0.02
Pythia	mean abl.	0.90	1.29	-0.39	$[-0.48, -0.33]$	$[-0.48, -0.31]$	-0.40	-0.38
Pythia	counterf.	0.15	0.11	+0.04	$[0.00, 0.07]$	$[0.00, 0.08]$	+0.04	+0.04

Table 12: P34 against the same blocks' original values (descriptive, paired by block). Δ: new minus original, for the raw-KL log ratio and for L with the original full-ablation denominators.

	GPT-2		Pythia			GPT-2		Pythia	
template	correct	flip	correct	flip	template	correct	flip	correct	flip
F1-1	100	100	99	100	F5-3	100	83	100°	52
F1-2	94	92	90	99	F5-4	100	97	100	100
F1-3	100	100	97	100	F7-1	100	100	100	84
F1-4	92	100	74°	100	F7-2	88	90	99	97
F2-1	98	97	100°	41	F7-3	97	93	97	97
F2-2	99	81	100	95	F7-4	100	100	100	98
F2-3	98	90	100	98	F8-1	75°	100	78°	100
F2-4	91	98	100°	50	F8-2	100	100	98	100
F5-1	98	100	100°	79	F8-3	94	98	92	100
F5-2	100	98	100	87	F8-4	100	99	100	98

Table 13: P34 clean-model competence per template (%, pooled over blocks). °: below the earlier $0.8/0.8$ template rule (reported, never used).

Rank and secondary selectors. U^* had midrank $1/120$ by G (bootstrap $[1, 19]$; first in 28.0% of replicates) and $2/120$ by A ($[1, 17]$; first in 30.7%); rank is descriptive, and no replication is inferred from it. The prespecified secondary references, with no criterion attached, had level contrasts 0.21 $[0.16, 0.26]$ (IMP) and 0.29 $[0.24, 0.32]$ (DISC-KL). The STRICT$-$FIELD top-1 difference was -0.5 points under U^* and 0.2 at the comparator median (median-minus-U^* contrast 0.7 $[-0.7, 2.1]$ points; for IO-over-S accuracy 0.3 $[0.1, 0.4]$); restrictions making both negative Name Movers ineligible averaged $G_U = 0.19$ and $A_U = 0.0057$ nats, one of them 0.45 and 0.0190, neither 0.53 and 0.0286. FIELD and STRICT picked the same circuit in 37% of U^*'s (block, size) cases. In the unrestricted ten-head universe of these blocks, the repeated-discovery study's comparison gave $L = 0.55$ (sizes 2–8) and raw-KL log ratios of 0.50 (sizes 2–6) and 0.54 (sizes 2–8). Under counterfactual patching (secondary, nominal) $C_G = 0.159$ and $C_A = 0.0088$ nats, but there both of U^*'s risks were higher than the median restriction's ($\Delta_F = 0.88$, $\Delta_S = 1.11$; identical picks in 98% of cases): the smaller ratio comes from STRICT worsening, not a repair of FIELD.

B THE SCOPE STUDY: TASKS AND UNIVERSE CONSTRUCTIONS

Design. Six model–task families (GPT-2 small and Pythia-160m on IOI, induction and subject–verb agreement), sixteen fresh blocks each, with prompts generated from new grammars and checked against an exposure ledger of every earlier study. In each block two eight-head universes are built from separate screening prompts: a *screened* universe of the eight heads with the best aver-

operator	contrast	U^*	median of others	C	95% bootstrap	95% jackknife	p	Holm p	verdict
mean abl.	C_G (log ratio)	0.043	0.452	0.409	$[0.207, 0.605]$	$[0.149, 0.668]$	0.003	0.005	replicated
	C_A (nats)	0.0017	0.0205	0.0188	$[0.0098, 0.0274]$	$[0.0070, 0.0306]$	0.003	0.005	replicated
counterf.	C_G (log ratio)	0.004	0.163	0.159	$[0.119, 0.190]$	$[0.113, 0.205]$	$\leq 10^{-4}$	–	secondary
	C_A (nats)	0.0006	0.0093	0.0088	$[0.0063, 0.0102]$	$[0.0059, 0.0117]$	$\leq 10^{-4}$	–	secondary

Table 14: P35, the two registered contrasts (C = median over the other restrictions minus U^*; Holm over the two, mean ablation) and the same statistics under counterfactual patching (secondary, nominal). Intervals are unadjusted.

selector	role	$R(U^*)$	median R	Δ	95% bootstrap	95% jackknife
FIELD	decomposition	0.0401	0.0581	−0.37	$[-0.52, -0.19]$	$[-0.58, -0.17]$
STRICT	decomposition	0.0384	0.0351	+0.09	$[-0.02, 0.22]$	$[-0.08, 0.26]$
IMP	secondary	–	–	+0.21	$[0.16, 0.26]$	$[0.10, 0.32]$
DISC-KL	secondary	–	–	+0.29	$[0.24, 0.32]$	$[0.26, 0.32]$

retained size	FIELD U^*	FIELD median	STRICT U^*	STRICT median	census U^*	census median
2	0.0664	0.1067	0.0658	0.0757	0.0845	0.1295
3	0.0531	0.0876	0.0531	0.0533	0.0656	0.1030
4	0.0417	0.0553	0.0374	0.0301	0.0482	0.0782
5	0.0266	0.0261	0.0290	0.0105	0.0315	0.0536
6	0.0126	0.0035	0.0067	0.0006	0.0155	0.0283

Table 15: P35 raw held-out KL (nats; block means over retained sizes 2–6) under U^* and the median restriction, the level contrast $\Delta = \log R(U^*) - \mathrm{median}_U \log R(U)$ (FIELD, STRICT: the registered decomposition; IMP, DISC-KL: secondary references), and the size profiles with the exact random-subset census.

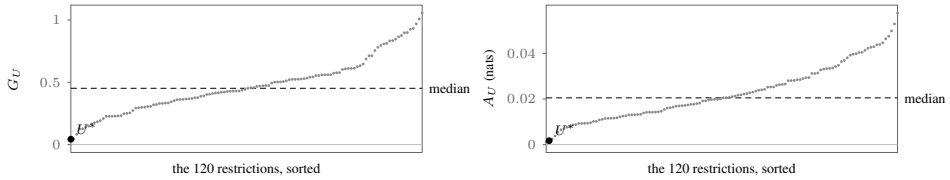

Figure 4: P35, every restriction under mean ablation: G_U (left) and A_U (right), sorted; U^* is marked and the dashed line is the median over the others (C is the gap between that line and U^*). Rank is descriptive.

age rank over four single-head screens (mean $|\Delta Y|$ and single-head KL, under both operators), and an *architecture-sampled* universe of one head drawn uniformly from each of eight layer bins. FIELD and STRICT select exhaustively at sizes 2–6 on selection prompts; the endpoint is raw held-out KL averaged over sizes, and the estimand is $\ell = \log(\bar{R}_F/\bar{R}_S)$, the log ratio of block-mean risks (not normalized by full ablation, unlike P30's; the two estimands differ after averaging over blocks). The 22 confirmatory claims form one Holm family: per family, the mean-ablation contrast, the operator interaction and counterfactual equivalence within $\pm \log 1.25$, each required under *both* constructions; and four forecast claims (two sealed forecasters of L, two of the accuracy-loss difference) against every registered ordinary alternative. Forecasts for all 384 units were sealed and committed before any evaluation prompt was run.

Outcome. Holm-adjusted p for the mean-ablation claim: GPT-2 IOI 0.002, induction 0.004 (both CONFIRMED), agreement 0.205; Pythia IOI 1.000, induction 0.606, agreement 0.606 (local p in Table 4). No interaction or equivalence claim was confirmed; 71 of the 384 units (47 counterfactual, 24 mean) had identical FIELD and STRICT selections; they enter the estimates as $L = 0$, and only settings identical in every block receive no verdict. Read locally, before multiplicity correction, the counterfactual equivalence test declared equivalence in $3/4$ screened universes with non-identical selections, but not in screened GPT-2 IOI ($1.21\times$) or in any architecture-sampled universe. The per-family mean-ablation, interaction and equivalence claims were promoted to confirmatory status before any official outcome existed, after the development analysis showed the forecaster unlikely to beat direct validation. Architecture-sampled universes barely move the task outputs (Table 16: full-ablation KL small except for Pythia induction under mean ablation), which is why large ratios

Table 16: P32, every selector: size-averaged held-out KL (nats; "full" ablates the whole universe) and task-accuracy loss (points), pooled over the sixteen blocks. Generated from the sealed record.

			held-out KL (nats)						accuracy loss (points)				
family	universe	operator	FIELD	STRICT	IMP	DISC-KL	random	full	FIELD	STRICT	IMP	DISC-KL	random
GPT-2 IOI	screened	mean	0.244	0.143	0.157	0.098	0.202	0.569	5.8	6.0	14.4	9.4	15.2
GPT-2 IOI	screened	cf	0.131	0.108	0.255	0.102	0.331	1.168	11.7	10.7	28.8	12.2	43.6
GPT-2 IOI	arch.	mean	0.024	0.010	0.010	0.008	0.038	0.071	−0.2	−0.1	−0.1	0.2	0.7
GPT-2 IOI	arch.	cf	0.003	0.001	0.001	0.001	0.027	0.049	0.1	0.0	0.0	0.0	2.9
GPT-2 induction	screened	mean	0.276	0.160	0.351	0.149	0.411	1.207	2.8	1.9	7.1	1.9	6.1
GPT-2 induction	screened	cf	0.146	0.143	0.263	0.134	0.639	2.485	1.9	1.9	5.6	1.8	14.2
GPT-2 induction	arch.	mean	0.027	0.018	0.017	0.016	0.046	0.100	0.1	0.0	0.0	0.0	0.3
GPT-2 induction	arch.	cf	0.008	0.005	0.004	0.004	0.030	0.063	0.0	0.1	0.0	0.0	0.5
GPT-2 agreement	screened	mean	0.041	0.040	0.042	0.035	0.096	0.280	3.4	3.4	3.4	3.7	17.3
GPT-2 agreement	screened	cf	0.095	0.095	0.095	0.094	0.369	0.998	19.4	19.4	19.4	19.2	56.8
GPT-2 agreement	arch.	mean	0.004	0.003	0.003	0.002	0.006	0.012	0.0	0.0	0.0	0.0	0.1
GPT-2 agreement	arch.	cf	0.000	0.000	0.000	0.000	0.002	0.006	0.0	0.0	0.0	0.0	0.5
Pythia IOI	screened	mean	0.150	0.144	0.093	0.070	0.146	0.272	3.2	3.2	5.9	9.3	16.4
Pythia IOI	screened	cf	0.059	0.059	0.057	0.053	0.348	0.907	9.6	9.6	9.3	10.2	48.2
Pythia IOI	arch.	mean	0.035	0.018	0.017	0.015	0.047	0.093	0.0	0.0	−0.1	0.4	1.0
Pythia IOI	arch.	cf	0.002	0.001	0.001	0.001	0.025	0.058	−0.1	0.0	0.0	0.0	3.7
Pythia induction	screened	mean	0.204	0.196	0.247	0.193	0.595	1.961	3.6	4.5	7.9	4.8	19.2
Pythia induction	screened	cf	0.223	0.223	0.225	0.222	0.908	3.019	9.6	9.5	9.5	9.5	31.4
Pythia induction	arch.	mean	0.033	0.014	0.014	0.013	0.243	0.472	−0.2	−0.1	−0.2	−0.1	6.0
Pythia induction	arch.	cf	0.009	0.004	0.005	0.004	0.022	0.045	0.4	0.1	0.1	0.1	0.9
Pythia agreement	screened	mean	0.264	0.262	0.268	0.118	0.194	0.346	2.0	2.0	2.0	3.0	9.6
Pythia agreement	screened	cf	0.032	0.032	0.032	0.032	0.363	0.776	3.6	3.6	3.6	3.7	59.2
Pythia agreement	arch.	mean	0.007	0.004	0.004	0.004	0.027	0.059	0.0	0.0	0.0	0.0	0.3
Pythia agreement	arch.	cf	0.001	0.000	0.000	0.000	0.030	0.064	0.0	0.0	0.0	0.0	5.4

there sit on raw gaps of only 0.0015–0.0194 nats under mean ablation and 0.0001–0.0049 nats under counterfactual patching. The secondary decomposition labels are OFFSETTING_COMPONENTS under mean ablation and GAP_NEGLIGIBLE_ADDITIVE_ONLY under counterfactual patching (pooled over families), so P30's answer-mass account does not carry over to the pooled panel. Table 17 gives every forecaster's error: neither sealed forecaster established its registered advantage; the lowest point error among the alternatives belongs to C6_discovery_raw, which is a statement about point errors, not a demonstrated superiority. Direct validation on held-out prompts is the natural baseline a forecaster must beat.

A retrospective moderator. Across the 4 P30 cells and 24 P32 settings, the share of universe heads whose single-head mean effect has the sign opposite to the full ablation's correlates with the log FIELD/STRICT raw-KL ratio (Spearman 0.81; leaving out one model–task family, 0.78–0.86). It is the largest of 18 candidate moderators computed post hoc, and its association with the raw gap is weaker (0.35, against 0.82 for the discovery lattice's own cancellation). This association was examined after the outcomes were available; the 28 settings are not independent mechanistic regimes (model, operator, universe construction and effect scale vary together), a leave-one-family-out recomputation is not a prospective forecast, and it does not revise P32's registered forecast result. It motivated the universe-composition analysis of Appendix A and a prospective composition test.

P30 and P32 on common endpoints (post hoc). The two studies use different primary estimands (P30 normalizes held-out KL by each block's full-ablation KL; P32 uses raw KL), so we scored both on the same endpoints with block-bootstrap intervals: raw and normalized KL, each criterion's own held-out error, full-vocabulary top-1 agreement and task accuracy. Normalizing changes no resolved direction (0/28 cells differ). On the held-out STRICT criterion FIELD's circuits are worse in 21/28 cells; on the raw held-out FIELD criterion they are never worse (0/28) and better in 9/28, all in screened P32 universes. In the published universes FIELD shows no advantage on its own criterion while losing on the others; the objective-dependent trade-off appears only in those screened settings. The top-1 cost of FIELD on Pythia under mean ablation is -14.1 points $[-16.7, -11.5]$ (FIELD worse in 12/12 blocks); nothing of that size appears in any screened P32 universe, so it belongs to the published Pythia universe. These comparisons are descriptive, nominal and unadjusted.

Table 17: P32 forecast contest: point mean absolute error over all 384 sealed units for the two sealed forecasters of each target (P_DIAG, P_END for L; A_DIAG, A_END for the accuracy-loss difference) and every registered alternative, listed by their registered identifiers. Generated from the sealed record.

predictor of L	MAE
C6_discovery_raw	0.093
C4_direct_validation_raw	0.104
P_END	0.111
C_STACK	0.131
C5_direct_validation_cal	0.132
C7_discovery_cal	0.140
C_VAL_META	0.149
C9_cv_discovery_raw	0.150
C10_cv_discovery_cal	0.178
P_DIAG	0.222
C3_D3	0.244
C2_operator_scale	0.261
C3_D2	0.267
C3_D1	0.277
C1_operator	0.283
C0_constant	0.292
C8_P20_coherence	0.328
predictor of the accuracy-loss difference	**MAE (points)**
A4_direct_validation_raw	0.68
A5_direct_validation_cal	1.15
A_END	1.38
A0_constant	1.76
A_VAL_META	1.80
A2_operator_scale	2.35
A_DIAG	2.71

C THE 8B EXTENSION BEYOND CURATED IOI CIRCUITS

Design. Qwen/Qwen3-8B at a pinned snapshot (36 layers, 32 query heads, 1,152 output heads). Three cells: **A** induction/copying over eight architectural groups whose boundaries are integer floors of $j \cdot L/8$, so every attention layer belongs to exactly one group; **B** subject–verb agreement over the same universe; **C** induction again with twelve heads screened automatically from all 1,152 heads, screened independently inside each repetition so that screening variability enters the intervals. Each cell runs twelve independent discovery/evaluation blocks under both role-mean and counterfactual replacement. The universes of A and B are fixed by architecture, not by any task response, which removes the task-informed head curation that §3 inherits from the published IOI circuit; C's universe is chosen by an automatic task screen rather than by hand, so it tests the claim on individual heads. Only scalar readouts, identifiers, manifests, hashes and receipts are persisted.

What was frozen, and when. Hypotheses, gates, selectors, estimands, sample sizes and the analysis were fixed in a preregistration committed before the model was loaded, under a guard that pins every implementation file by hash and refuses to collect if any of them, the runtime, the input manifests or the library versions differ from those validated on the device. The GPU validation itself is a separate mode whose receipt must match the live source hashes; collection refuses without it.

The registered decision rule. A distributional pass for the confirmatory cell requires *all* of: pooled FIELD/STRICT KL ratio ≥ 1.25; the 95% decision interval for L entirely above zero; pooled raw size-averaged KL gap ≥ 0.02 nats; and the predicted direction in at least 9 of 12 blocks. A behaviourally consequential pass additionally requires a size-averaged task-accuracy gap ≥ 2 percentage points with its own 95% interval above zero. The sole first-stage confirmatory hypothesis is cell A under role-mean replacement, and the registered sequence is A, then B, then C, with formal confirmatory claims advancing only while the preceding cell passes its distributional gate. Predeclared alternatives include a small/equivalent effect (the 90% ratio interval inside $[1/1.1, 1.1]$ and the behavioural interval inside ± 1 point), a reversal, KL-only harm, and an unstable procedure at fewer than $9/12$ blocks.

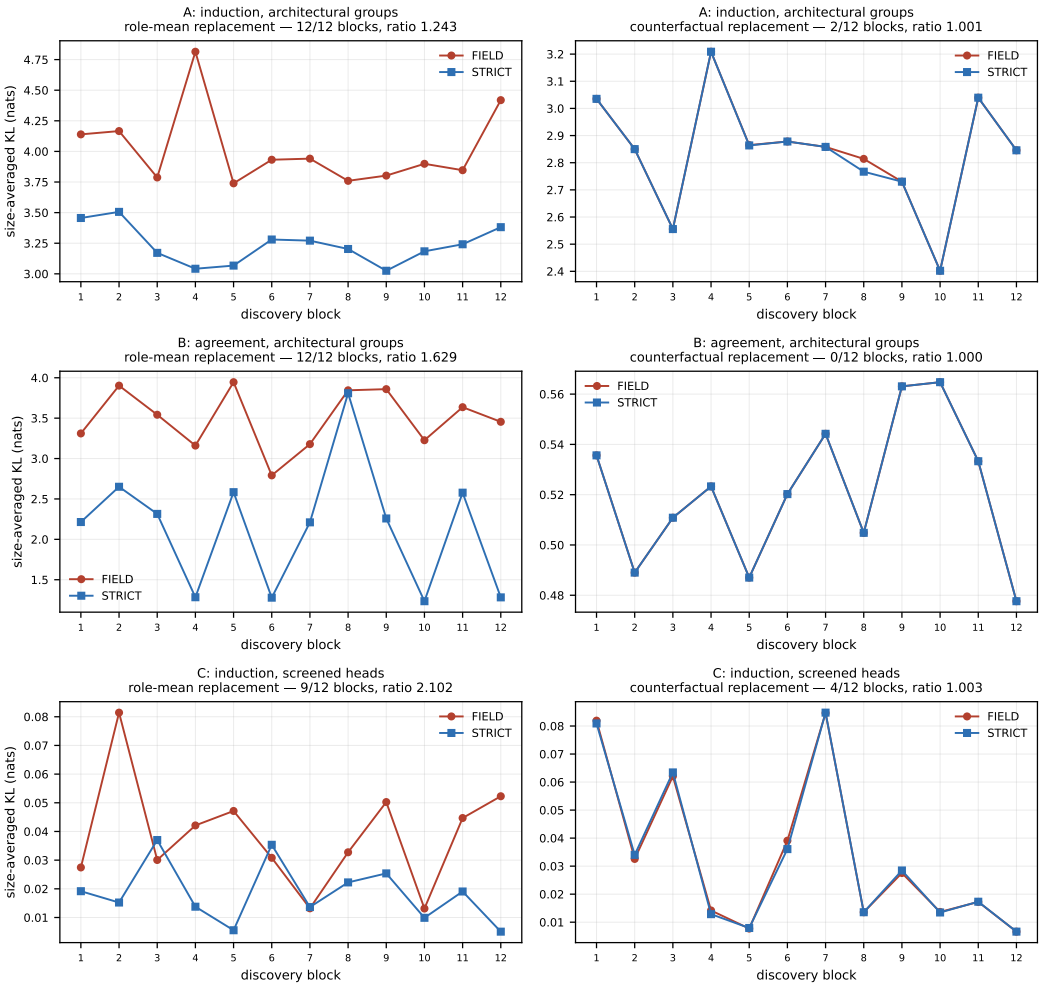

Figure 5: P28, every registered block. Each panel plots the two criteria's size-averaged held-out KL in each of the twelve independent discovery blocks; no quantity here is pooled other than as the preregistration defines, and no gate is redrawn. Under role-mean replacement FIELD's circuit has the higher held-out KL in $12/12$, $12/12$ and $9/12$ blocks (cells A, B, C); under counterfactual replacement the curves nearly coincide, which is the predeclared small-effect outcome in A and C (in B the selected masks are identical). The panel titles carry the registered block count and pooled ratio, including the ratio that fell below the 1.25 materiality gate.

Outcome. Figure 5 shows every registered block. Cell A under role-mean replacement returned INCONCLUSIVE_OR_BELOW_MATERIAL_GATE: the ratio was 1.243 against the required 1.25, while the same cell's other registered endpoints separated clearly (95% interval for L $[0.162, 0.272]$; raw gap 0.785 nats; $12/12$ blocks in the predicted direction; task-accuracy gap 10.76 points, $[9.61, 11.91]$). Under counterfactual replacement the same cell returned EQUIVALENT_SMALL_EFFECT (ratio 1.001; $2/12$ blocks; behavioural gap 0.02 points, $[-0.03, 0.08]$), satisfying the predeclared small-effect clause.

Under role-mean replacement, cell B (subject–verb agreement over the same universe) returned PASS_BEHAVIORALLY_CONSEQUENTIAL — ratio 1.629, 95% interval for L $[0.302, 0.674]$, raw gap 1.346 nats, $12/12$ blocks, task-accuracy gap 8.44 points $[6.74, 10.14]$ — which clears every registered threshold. Under counterfactual replacement it returned DEGENERATE_NO_CONFIRMATORY_VERDICT: the two criteria selected the same mask at every size in every block, so all twelve paired differences are exactly zero, the bootstrap has no variance to re-

Table 18: P28, every preregistered selector. Size-averaged held-out KL (nats; "full abl." is the pooled KL of ablating the whole universe) and size-averaged task-accuracy loss (percentage points), pooled over the twelve blocks. Generated post hoc from the sealed per-block records (exploratory).

		full	held-out KL (nats)					accuracy loss (points)			
cell	operator	abl.	FIELD	STRICT	IMP	DISC-KL	random	FIELD	STRICT	IMP	DISC-KL
A induction, groups	role-mean	9.43	4.02	3.24	2.25	2.16	5.74	44.9	34.1	37.7	36.3
A induction, groups	counterfactual	15.36	2.84	2.84	2.91	2.52	9.38	37.6	37.5	37.8	36.8
B agreement, groups	role-mean	1.53	3.49	2.14	0.61	0.51	1.45	10.2	1.7	3.4	9.5
B agreement, groups	counterfactual	2.05	0.52	0.52	0.52	0.53	1.33	32.4	32.4	32.4	34.7
C induction, heads	role-mean	0.16	0.04	0.02	0.03	0.02	0.05	1.1	0.4	0.5	0.3
C induction, heads	counterfactual	0.25	0.03	0.03	0.04	0.02	0.08	0.6	0.6	0.6	0.3

sample, and the analysis declines a verdict rather than reporting a spurious equivalence interval. That exact coincidence is the operator dependence in its strongest form on this task.

Because the confirmatory cell did not pass its distributional gate, the registered sequence makes cells B and C descriptive; they are reported because the preregistration requires continuous estimates regardless of gates, and they are not combined with A into a single headline. We note explicitly that the strongest verdict in this experiment belongs to a cell whose result the preregistration does not permit us to advance.

Cell C, induction over twelve screened heads, returned KL_ONLY_HARM under role-mean replacement: ratio 2.102, 95% interval for L $[0.237, 1.249]$, raw gap 0.020 nats, $9/12$ blocks, task-accuracy gap 0.71 points $[0.18, 1.24]$, below the registered 2-point materiality threshold. The absolute KL levels in this cell are small (Table 18), so a large ratio corresponds to a raw gap of hundredths of a nat. Under counterfactual replacement it returned EQUIVALENT_SMALL_EFFECT (ratio 1.003; $4/12$ blocks; behavioural gap 0.03 points).

The preregistered competitors. The preregistration also named an importance ranking (IMP), exhaustive selection on discovery-prompt KL (DISC-KL) and random same-size subsets, all scored on the same held-out blocks. Table 18 gives every selector's absolute levels. Under role-mean replacement in cells A and B the two KL-oriented selectors preserve the output distribution better than either aggregation, and STRICT preserves task accuracy better than they do (in cell C, whose absolute levels are small, the orderings are mixed); in cell B FIELD's circuit has higher held-out KL than ablating every group in 12/12 blocks and STRICT's in 8/12, while STRICT has the lowest accuracy loss of any selector. Under counterfactual replacement FIELD and STRICT coincide or nearly so, but DISC-KL does not: it has $0.89\times$ and $0.56\times$ STRICT's held-out KL on induction in cells A and C. No selector is best on both endpoints in every cell.

D STUDY NAMES AND CODES

The main text names studies by what they do; the supplement, the study registry and the machine-readable outputs use programme codes. Table 19 maps one to the other.

name in the paper	code	where
repeated-discovery study	P30	§3, Appendix A
earlier conditional experiment	P19B (with P9, P10)	Appendix A
scope study	P32	§4, Appendix B
8B extension	P28	§4, Appendix C
exploratory composition analysis	(probe)	§4, Appendix A
fresh restriction replication	P35	§4, Appendix A.2
fresh-template transfer study	P34	§4, Appendix A.1
search study	E1, P26	Supplement S21

Table 19: Study names and programme codes.

E OPERATING POINTS AND SELECTION INSTABILITY IN FULL

Status and chronology. Both analyses use data that had already been analysed: the discovery lattices and held-out records of the repeated-discovery study (with its registered block-bootstrap indices), the scope study and the 8B extension. A first pass used each criterion's own normalizer, as in the earlier conditional experiment; after it had been read, the rules reported here were fixed (thresholds 80, 85, 90%; the common normalizer D; the symmetric rule as the lead specification; the full-universe fallback; the evaluated grid of registered sizes plus $K-1$) before their outputs were computed. The pairing "FIELD one-sided, STRICT symmetric" was added later, after the other rule outputs had been read. An independent reimplementation from the raw arrays, written without the analysis code, reproduced every value quoted in §5 (digests in the supplement). The choice of normalizer moves the repeated-discovery headlines by less than one point of top-1 agreement, but it is not immaterial study-wide: in one scope-study setting on the shared path it moves the difference by about six points; the per-cell maxima are in the supplement.

Higher thresholds. Table 20 repeats Table 5 at 90%. STRICT's accepted circuits grow with the threshold while FIELD's own picks stay at the grid floor, which is why the own-pick difference widens. On the shared path both rules walk the same nested circuits, so its differences arise only from where each rule stops.

model	candidates	retained heads FIELD/ STRICT	top-1 agreement (%) FIELD/ STRICT	difference (points) FIELD−STRICT [95%]	raw KL (nats) FIELD/ STRICT	fallbacks FIELD/ STRICT
GPT-2	own exhaustive picks	$2.00^{\dagger}$ / 8.17	90.4 / 97.5	-7.1 $[-9.4, -5.1]$	0.236 / 0.005	0 / 0
GPT-2	shared importance path	7.17 / 8.17	95.9 / 97.4	-1.5 $[-2.5, -0.5]$	0.017 / 0.006	0 / 0
Pythia	own exhaustive picks	$3.00^{\dagger}$ / 10.17	67.2 / 95.9	-28.7 $[-36.2, -21.6]$	0.270 / 0.009	0 / 0
Pythia	shared importance path	5.00^{*} / 10.67	86.1 / 97.0	-10.8 $[-13.5, -7.8]$	0.052 / 0.005	0 / 0

Table 20: As Table 5, at $\tau = 90\%$. Retained heads: mean accepted size; †: every block stopped at the smallest evaluated size; *: some blocks did.

Held-out pass rates are not like-for-like. Each criterion's accepted circuit is checked against its own rule recomputed on evaluation prompts. Under the common denominator FIELD's error is pointwise no larger than STRICT's ($|\mathbb{E}\Delta Y_p| \le \mathbb{E}|\Delta Y_p|$), so its threshold is never harder to satisfy, and the higher FIELD pass rate (11 of 12 against 2 in Pythia) shows that the rule does not detect the prediction cost, not that FIELD is more reliable.

Deficit-only rules. With σ the sign of the full ablation's discovery effect, the deficit-only errors $J_{F,+} = (\sigma\,\mathbb{E}\Delta Y_p)_+/D$ and $J_{S,+} = \mathbb{E}(\sigma\Delta Y_p)_+/D$ ignore deviations opposite to the full-ablation direction. The STRICT variant is a modified criterion: it counts only per-prompt deficits. It is applied to the stored candidate paths and does not re-optimize any size. On the Pythia shared path both rules then accept the same three-head circuit in every block; that circuit keeps the clean top prediction on 82.0% of prompts (the STRICT-accepted circuit under the primary rule: 91.5%).

The 8B cells and counterfactual patching. In the 8B extension the operating-point analysis uses architectural groups and the width constraints of that study. In the induction cell under mean ablation the own-pick difference at 80% is -61.0 accuracy points and is identical under all five rule variants; at the common size four it is -53.6 points, and at 90% the common-size difference is 0.0. Under counterfactual patching the two criteria accepted identical circuits at every threshold in that cell. The agreement and screened-induction cells show no comparable difference. In the repeated-discovery study under counterfactual patching, the two criteria accepted the same or nearly the same circuits at 80–85%, but not uniformly at 90%: GPT-2's own-pick prediction difference reached -4.9 points $[-6.5, -3.3]$.

Earlier related analyses. Two studies in the registry bear on this appendix. An unregistered practitioner-protocol comparison of averaged versus per-example recovery on archived lattices (P18) found the averaged criterion stopping at smaller or worse circuits in most mean-ablation comparisons but at identical minimal sizes under counterfactual patching; it was superseded before its preregistered replication ran. An exploratory mechanism analysis on reused data (P9-MECH) rejected

a winner's-curse account of the earlier single-draw result, but its synthesis was not completed; the observations below neither confirm nor overturn it.

Selection instability. We scale the discovery gap between the best and second-best candidate of each size by a noise scale (for FIELD the standard error of the signed discovery mean; a heuristic scale, not a test). Under mean ablation in Pythia this gap is 0.008 for FIELD against 1.06 for STRICT, 313 FIELD candidates lie within one such unit of the optimum against 4.5 for STRICT, and a bootstrap over discovery prompts re-selects FIELD's pick in 0.9% of resamples against 54%. Held-out values exist only for the uniformly random evaluation masks of the original study, so near-optimal sets are compared on those members alone: under a tolerance set by the discovery-to-held-out noise, 1877 FIELD and 232 STRICT members were evaluated, and 30 STRICT (block, size) cells have none. On the 78 cells where both criteria have members, weighted equally within blocks, the FIELD near-optimal members have 0.199 $[0.155, 0.243]$ higher held-out normalized KL than STRICT's and a top-1 agreement difference of -9.2 $[-11.5, -6.9]$ points. This tolerance uses held-out information, so the neighbourhood is descriptive. Under counterfactual patching these comparisons depend on the cell weighting and are not quoted. Within the recorded discovery sample of 50 prompts, nested template-balanced subsamples (the same for every selector; the full sample reproduces the original picks) left the GPT-2 held-out gap at the two fully evaluated sizes at 0.20 $[0.15, 0.25]$ with 10 prompts and 0.30 $[0.13, 0.50]$ with 50; this range says nothing about larger discovery samples. Under counterfactual patching FIELD's masks also recur less often across independent blocks (24.0% and 37.0% of (block pair, size) cases against 43.9% and 43.4% for STRICT) while the two selectors' held-out KL ratios lie inside the registered margin, so an unstable circuit identity does not have a uniform functional consequence. These observations are consistent with both criterion-dependent preservation differences and finite-sample selection sensitivity; they do not separate the two at the population level.

Calibration of the twelve-block equivalence test. In sensitivity simulations prepared for a follow-up study, using populations constructed from these block contrasts (5000 experiments per scenario, five-claim Holm family), the twelve-block equivalence procedure declared equivalence in 9.4% (GPT-2) and 9.2% (Pythia) of experiments whose true counterfactual ratio sat at the upper margin 1.25, against a nominal 5%; at the lower margin it was conservative. P30's own preregistered design simulation had put this rate near the nominal level; it resampled the earlier conditional experiment's lattices with their fixed donors and template roles and set the boundary on the mean log ratio, whereas the later audit resamples P30's sealed blocks, which carry the rebuilt donors, the template-role variation and the right-skewed FIELD risks, and sets the boundary on the ratio of expected risks. The counterfactual verdict of §3 is reported as registered, but the procedure should not be regarded as generally calibrated at its upper boundary.