THE LOTTERY CIRCUIT HYPOTHESIS: SPARSE MECHANISMS FROM RANDOMNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretable AI, and mechanistic interpretability in particular, aims to uncover the internal computations of language models, with circuit discovery remaining an important approach. We propose the *Lottery Circuit Hypothesis*: sparse, task-functional circuits can be constructed by zero-based masking of random-weight transformers, rather than merely extracted from computations established through pretraining. This reveals a possibility in circuit discovery that bears on what we call the *provenance assumption*: that a sparse subgraph reproducing a target task behaviour thereby localises computation already encoded in the pretrained weights. Motivated by the *strong lottery ticket hypothesis*, we perform circuit discovery beyond pretrained models and find that DiscoGP can construct sparse, high-performing weight-and-edge circuits in randomised transformers. We further demonstrate *concealment*: task-agnostic, small-magnitude fills obscure the trace of zero-based weight pruning while preserving task performance. This extends the perturbation perspective of *neural thickets* from neighbourhoods of pretrained models to neighbourhoods of masking-constructed subnetworks within random models. We give a theoretical account of what we call *margin synthesis* by linking this constructive capacity to mask granularity: in a simplified additive model, each edge contributes a sum of n independent, zero-mean Gaussian terms, normalised by $\sqrt{n}$, to the task margin. Weight pruning can retain favourable terms within an aggregate and attain $\Theta_p(\sqrt{n})$, whereas edge pruning selects among aggregates of magnitude $O_p(1)$. Together, these results call for a rethinking of circuit discovery: a discovered circuit may be a mask-induced construction, and interpreting it as a map of pretrained computation requires additional justification and evidence.[1]

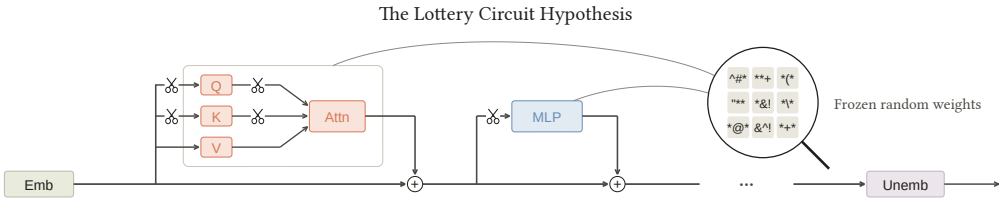


Figure 1: **A circuit in a random transformer.** Scissors mark pruned edges. Attention and MLP weights remain fixed at their original random values; embedding and unembedding parameters remain outside the randomised domain.

1 INTRODUCTION

Circuit discovery seeks to explain a language model's behaviour by identifying the internal computations that produce it. A discovered circuit is typically evaluated by how well it reproduces the target behaviour, with ablation or patching used to test the roles of its components (Elhage et al., 2021; Wang et al., 2023; Conmy et al., 2023; Bhaskar et al., 2024). Such evidence can show that

[1] Anonymous code: `https://anonymous.4open.science/r/circuits_from_random`.

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.

a selected computation is sufficient for the task. Whether it identifies a computation established by pretraining is a further question.

We call the assumption that a task-functional circuit thereby localises computation established by pretraining the *provenance assumption*. There is reason to examine it: task examples guide circuit search, and the learned mask determines which computations interact. A procedure that selects individual weights can change the contribution of a component even when every retained numerical value stays fixed. In our view, understanding what the search can construct is necessary for interpreting what it finds.

The lottery ticket literature shows why this matters. The original hypothesis concerns subnetworks that can be trained in isolation; supermasks and the strong lottery ticket hypothesis show that selection alone can obtain useful functions from frozen random networks (Frankle & Carbin, 2019; Ramanujan et al., 2020; Malach et al., 2020). Circuit discovery can likewise select individual parameters and connections between transformer components. Can these procedures construct task-functional circuits after the eligible pretrained values have been replaced by random weights?

We investigate this possibility through the *Lottery Circuit Hypothesis*: sparse, task-functional circuits can be constructed by zero-based masking of random-weight transformers. We randomise eligible attention and MLP parameters while keeping the remaining pretrained parameters fixed. Using DiscoGP's weight and edge masks (Yu et al., 2025), we find that selecting weights before edges yields high task accuracy with sparse connectivity where direct edge-only discovery is less successful. Thus, for these circuits, the eligible pretrained values are not required to obtain the target behaviour.

Weight pruning leaves a visible trace: the removed parameters become zeros. We test whether this trace can be concealed while preserving the circuit's behaviour. Taking the pruned vector as a reference, we draw on work on useful nearby models and weight perturbations (Gan & Isola, 2026; Neyshabur et al., 2018) to introduce task-agnostic *distributional completion*. It keeps selected values fixed and refills pruned coordinates to approach a fitted Gaussian distribution. We call the obscuring of the pruning trace *concealment*. Useful behaviour survives completion under the original edge support, although the effect on accuracy varies across model configurations. A shuffle control, which restores each tensor's original multiset of values, helps assess the effect of fresh refill.

Our theory examines how one mask can make several task margins positive, and when these margins survive refill. The models are idealised: they give sufficient conditions for margin synthesis and stability, and an optimal allocation of fill values for a fixed histogram target. These results help explain how task-guided selection can construct useful computation without changing retained weights. We make three contributions:

1. We study random-weight circuit construction across GPT-2 Small, GPT-2 Medium, Pythia-410M and Qwen2.5-0.5B on IOI, BLiMP and Docstring, comparing mask granularity and variation across initialisations (§3–4).
2. We demonstrate task-agnostic completion of pruned weights, evaluating task performance and Gaussian fit, with tensor-local shuffling as a control (§4).
3. We give sufficient conditions for shared-mask margin synthesis and stability under refill, and characterise optimal residual allocation in idealised models (§5).

2 Related work and preliminaries

Circuit discovery, DiscoGP, and the provenance assumption Circuit discovery begins with a choice of components. Architectural decompositions use units such as attention heads and MLPs; feature-based decompositions identify computations at a different granularity. Patching, attribution and mask optimisation then select circuits from these components (Elhage et al., 2021; Marks et al., 2025; Conmy et al., 2023; Syed et al., 2024; Bhaskar et al., 2024). We use an architectural graph $G = (V, E)$, where V denotes computational components and E directed edges carrying their outputs. In the unrolled residual stream, inputs are sums of earlier outputs. Zero-ablation removes one such contribution, while patching replaces it with a reference; either intervention is propagated through the downstream nonlinear computation.

DiscoGP also masks individual parameters, allowing selection at the level of scalar weights (Yu et al., 2025). Let $\mathcal{I}_\theta$ be the finite index set of eligible scalar attention and MLP parameters and $\theta = (\theta_j)_{j\in\mathcal{I}_\theta}$. Each index specifies a tensor and a coordinate; $|\mathcal{I}_\theta|$ counts coordinates, including distinct coordinates with equal values. Embeddings, positional embeddings, normalisation layers and the LM head remain untouched. A circuit is

$$C = (m_\theta, m_E), \qquad m_\theta \in \{0,1\}^{|\mathcal{I}_\theta|}, \quad m_E \in \{0,1\}^{|E|}. \tag{1}$$

Weight masking applies $m_\theta \odot \theta$, where $\odot$ denotes elementwise multiplication; edge masking zero-ablates disabled connections. Appendix B.2 describes the treatment of shared biases. Write $\mathbf{1}_\theta$ and $\mathbf{1}_E$ for the all-on masks. Discovery has the form $(\text{model}, \text{task}) \mapsto C$: numerical values remain frozen while task examples guide mask learning. Appendix A gives the discovery formulations; Appendix B specifies our implementation.

A circuit that performs the target task demonstrates the sufficiency of its selected computation under the evaluation procedure. Interpreting it as a mechanism established by pretraining invokes the provenance assumption. Work on weight-sparse models has also distinguished task performance from preservation of the original mechanism (Gao et al., 2026; Drori, 2026). We test whether discovery still succeeds when the eligible pretrained values are randomised. Parameter randomisation is an established sanity check for attribution methods (Adebayo et al., 2018); here it takes place before circuit search.

Variation among circuits Circuit non-uniqueness is already well documented. Useful circuits can differ with mask initialisation, overlap penalties, token position or prompt (Bhaskar et al., 2024; Chen et al., 2026; Haklay et al., 2025; Franco et al., 2026). Across models, interpretive equivalence asks whether distinct networks admit the same mechanistic account (Sun & Toneva, 2026). Subspace patching can isolate directions that carry an intervention without carrying the mechanism (Makelov et al., 2024), and mechanistic accounts can face broader identifiability problems (Méloux et al., 2025). Our question concerns the origin of a circuit's function: whether task-guided search can construct it after eligible pretrained weights have been randomised.

Strong lottery tickets: masks over random weights The lottery ticket hypothesis concerns subnetworks trainable in isolation (Frankle & Carbin, 2019). Its strong form asks whether pruning frozen random weights can yield useful functions. Supermasks and approximation results demonstrate that it can (Zhou et al., 2019; Ramanujan et al., 2020; Malach et al., 2020; Pensia et al., 2020; Otsuka et al., 2026). Random transformer blocks can also solve algorithmic tasks when their input and output embeddings are trained (Zhong & Andreas, 2024). In our experiments, parameters outside the randomised domain stay fixed, and task supervision acts through weight and edge selection. The theory studies how such selection can create positive margins on prescribed finite sets of examples.

Perturbations around a pruned weight vector The existence of useful models near a reference parameter vector (Gan & Isola, 2026) motivates examining the neighbourhood of a discovered circuit. We take its retained values and exact zeros as the reference, then refill the zeros using weight statistics and fresh randomness. Weight-noise analyses relate parameter perturbations to changes in prediction margins (Neyshabur et al., 2018). They suggest using small-magnitude fills to preserve the pruned model's behaviour without task feedback during completion.

3 Constructing lottery circuits in random transformers

We begin by asking whether edge-only discovery can find an accurate sparse circuit in randomised weights. We then allow scalar selection before edge search and test whether the resulting circuit survives refill of its pruned coordinates. Figure 2 shows this sequence; §4 evaluates each stage.

Edge circuits in random weights Let μ_{reinit} and σ_{reinit} be the mean and standard deviation of pretrained values pooled over $\mathcal{I}_\theta$, and let a_{reinit} and b_{reinit} be their minimum and maximum. Independently for each eligible coordinate, draw and clip

$$\mathrm{Z}_j \sim \mathcal{N}(\mu_{\text{reinit}}, \sigma^2_{\text{reinit}}), \qquad \theta_j^{\text{reinit}} = \min\{b_{\text{reinit}}, \max\{a_{\text{reinit}}, \mathrm{Z}_j\}\}. \tag{2}$$

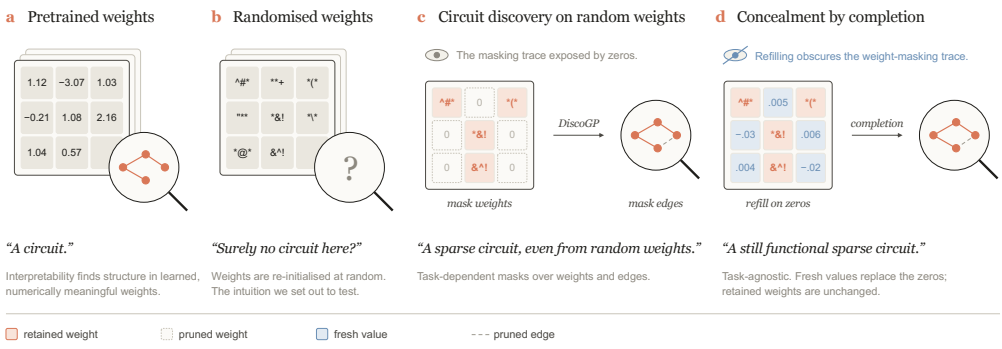


Figure 2: **Constructing and completing a lottery circuit.** Randomisation replaces eligible pretrained values; task-guided weight and edge masks construct a circuit. Completion retains selected values, fills pruned coordinates and removes the scalar mask.

We use this distribution for all eligible parameters, without layerwise rescaling. Numerical values then remain frozen during discovery. For target task $\mathcal{T}$, let $A_{\mathcal{T}}(\theta; C)$ denote circuit accuracy and $\rho_E(m_E) = \|m_E\|_0/|E|$ its retained edge fraction; $\|m_E\|_0$ counts enabled edges. Given desired accuracy α and maximum edge fraction ε, both in $[0, 1]$, we seek

$$C_E = (\mathbf{1}_\theta, m_E), \qquad A_{\mathcal{T}}(\theta^{\text{reinit}}; C_E) \geq \alpha, \qquad \rho_E(m_E) \leq \varepsilon. \tag{3}$$

Performance, sparsity and completeness follow DiscoGP's criteria (Yu et al., 2025); Appendix B.1 supplies the measurement definitions. We compare edge-only discovery with modes that also prune weights in §4.

Weight selection as a route to sparse edges An edge mask keeps or removes an aggregate contribution. A weight mask can change that contribution by zeroing individual parameters before they are combined. We first learn m_θ with every edge available, then freeze the pruned weights and learn m_E. Phase θ selects scalar contributions; Phase E selects routes through the resulting model. Completion subsequently refills the same pruned coordinates.

In the construction-and-completion experiments, all four models use $\lambda_\theta < 0$ to reward weight retention and $\lambda_E > 0$ to penalise edge density. Retaining more weights leaves fewer coordinates to refill, helping the completed array stay close to the original draw distribution; the edge penalty encourages sparse routing. Appendix B.3 gives the model-specific settings. We also compare weight-only, edge-only, reverse-order and joint pruning. The sequential protocol lets us evaluate the same edge mask before and after completion.

From zero pruning to distributional completion After weight-and-edge selection, the circuit still depends on an explicit scalar mask. We first test refill by shuffling the original removed values among pruned coordinates within each tensor, keeping retained values fixed. This restores each tensor's original empirical multiset without task feedback. It provides a control for completion with fresh, smaller-magnitude values.

Write the pruned index set as $\mathcal{P} = \{j \in \mathcal{I}_\theta : (m_\theta)_j = 0\}$. For fresh fill values $\boldsymbol{z} = (z_j)_{j\in\mathcal{P}}$, define

$$\theta_j^{\text{comp}}(\boldsymbol{z}) = \begin{cases} \theta_j^{\text{reinit}}, & j \notin \mathcal{P}, \\ z_j, & j \in \mathcal{P}. \end{cases} \tag{4}$$

To choose the fills, we use retained values, pruned locations and fresh randomness. A symmetric two-component reference distribution has standard deviation $f\sigma_{\text{reinit}}$, where $f > 0$; the parameter $a \in [0, 1]$ controls the separation of its component means. We fit a Gaussian mean and scale to the retained values together with this reference, then allocate fills to bins in proportion to their positive histogram deficits, with integer rounding. Fresh within-bin values are assigned to pruned coordinates in the rank order of the reference draws. Retained values stay fixed throughout, and there is no task-gradient or repair step. Appendix B.2 gives the full procedure.

We then remove the scalar mask and evaluate the completed weights under the original edge support:

$$(\mathbf{1}_\theta, \mathbf{1}_E) \xrightarrow{\text{Phase } \theta} (m_\theta, \mathbf{1}_E) \xrightarrow{\text{Phase } E} C = (m_\theta, m_E) \xrightarrow{\text{completion}} (\mathbf{1}_\theta, m_E) \text{ on } \theta^{\text{comp}}. \quad (5)$$

Keeping the original edge mask tests whether the circuit survives refill. We also enable all edges to measure performance without edge pruning, and run fresh edge-only discovery on the completed weights. All eligible values are randomly generated, but the learned weight mask determines which are retained and which are replaced. Pooled and layerwise Gaussian-fit diagnostics measure the distributional trace left by this selection.

4 Empirical evaluation and discussion

We evaluate GPT-2 Small, GPT-2 Medium, Pythia-410M and Qwen2.5-0.5B on IOI, a token-pair adaptation of BLiMP, and Docstring (Wang et al., 2023; Warstadt et al., 2020; Heimersheim & Janiak, 2023). The models share examples with valid target-token pairs that every pretrained model answers correctly. We compare pruning modes, follow weight-then-edge construction through completion, and run fresh edge-only searches on the completed weights. Appendices B and B.3 give the data, evaluation criteria, optimisation settings and replication.

Weight selection improves accuracy and edge sparsity Edge-only discovery reaches 99.0–100.0% accuracy with pretrained weights, but only 51.2–92.6% after randomisation under matched search settings (Table 1). The four modes that also prune weights reach 93.8–100.0% across the randomised models. The largest gains occur on IOI and Docstring; BLiMP is more accessible to edge-only discovery.

Table 1: **Pruning modes across models and tasks.** Pretrained / Randomised accuracy (%) from one search per weight setting and pruning mode, with shared search settings (Appendix B.3).

Model	Task	Weights only	Edges only	$\theta \to E$	$E \to \theta$	Joint
GPT-2 Small	IOI	100.0 / 99.8	100.0 / 60.1	100.0 / 100.0	100.0 / 99.6	100.0 / 100.0
	BLiMP	99.6 / 98.0	99.3 / 80.0	99.4 / 97.7	98.9 / 98.5	99.3 / 96.9
	Docstring	100.0 / 99.5	100.0 / 59.7	100.0 / 99.6	100.0 / 99.8	100.0 / 98.8
GPT-2 Medium	IOI	99.9 / 99.8	100.0 / 69.6	100.0 / 99.8	100.0 / 99.8	100.0 / 99.9
	BLiMP	99.5 / 97.6	99.1 / 91.3	99.0 / 97.5	99.4 / 98.0	98.9 / 97.6
	Docstring	100.0 / 99.1	100.0 / 62.1	100.0 / 99.3	100.0 / 99.5	99.9 / 98.3
Pythia-410M	IOI	99.8 / 99.8	100.0 / 89.0	100.0 / 99.9	99.9 / 99.8	100.0 / 99.9
	BLiMP	99.5 / 98.6	99.0 / 92.6	99.6 / 98.4	98.8 / 98.6	98.6 / 98.2
	Docstring	98.3 / 99.8	99.9 / 83.3	99.9 / 99.8	99.3 / 100.0	99.8 / 99.8
Qwen2.5-0.5B	IOI	99.7 / 99.8	100.0 / 51.2	100.0 / 99.9	100.0 / 99.9	100.0 / 100.0
	BLiMP	99.4 / 93.8	99.1 / 69.0	99.6 / 93.9	99.5 / 97.3	99.4 / 93.9
	Docstring	100.0 / 99.4	99.9 / 57.4	100.0 / 99.5	100.0 / 99.7	100.0 / 99.2

Weight selection also permits fewer edges. On Medium IOI, weight-then-edge discovery uses 56 edges at 99.8% accuracy, compared with 31,577 edges at 69.6% for edge-only discovery. Small IOI and Docstring, Medium Docstring and Pythia IOI show the same pattern (Appendix Figure 5). Qwen IOI reaches 99.9% with 26 edges, compared with 51.2% using 24,547 edges under edge-only discovery. Under weight-then-edge discovery, randomised circuits are also sparser than pretrained circuits in nine of the twelve pairs (Table 2). The ranking among pruning modes varies across settings. We use weight-then-edge pruning in the following experiments so that routing is learned on fixed pruned weights and can be tested again after refill.

From random initialisation to completed weights Across ten weight initialisations per model–task pair, selection raises mean accuracy from 48.7–53.1% to 92.6–99.9%, with 0.03–2.93% of edges retained (Table 3). GPT-2 and Pythia retain approximately half the eligible parameters. Qwen, whose search starts from a nearly dense weight mask, retains 93.2–97.9% and leaves far fewer coordinates to refill. These gains come entirely from selection: every retained numerical value stays fixed during discovery.

Table 2: **Accuracy and edge density with pretrained and randomised weights.** Weight-then-edge circuits from the pruning-mode comparison, reported as accuracy (%) / retained edge density (%). In nine of the twelve pairs the randomised circuit is the sparser of the two; the exceptions are Pythia on IOI and BLiMP and Qwen2.5-0.5B on BLiMP.

	IOI		BLiMP		Docstring	
Model	Pretrained	Randomised	Pretrained	Randomised	Pretrained	Randomised
GPT-2 Small	100.0 / 0.34	100.0 / 0.10	99.4 / 1.69	97.7 / 0.89	100.0 / 0.83	99.6 / 0.42
GPT-2 Medium	100.0 / 0.74	99.8 / 0.02	99.0 / 1.66	97.5 / 0.24	100.0 / 0.94	99.3 / 0.18
Pythia-410M	100.0 / 0.29	99.9 / 0.36	99.6 / 0.54	98.4 / 0.96	99.9 / 0.19	99.8 / 0.14
Qwen2.5-0.5B	100.0 / 0.16	99.9 / 0.01	99.6 / 0.42	93.9 / 0.58	100.0 / 0.12	99.5 / 0.04

Table 3: **Construction and completion across models and tasks.** Means over ten weight initialisations per model–task pair; all values are percentages. Paired completion retains the selected edge mask (Original edges) or enables every edge (All edges). Settings are given in Appendix B.3.

		Accuracy before completion		Retained density		Completed accuracy	
Model	Task	Dense	Selected circuit	Weights	Edges	Original edges	All edges
GPT-2 Small	IOI	49.6	98.6	50.2	1.24	96.0	95.3
	BLiMP	48.7	96.7	50.4	1.23	95.4	95.2
	Docstring	50.1	96.7	50.1	2.93	86.2	95.2
GPT-2 Medium	IOI	50.0	97.7	50.1	1.55	95.3	95.0
	BLiMP	53.1	96.5	50.4	0.66	95.3	93.8
	Docstring	49.8	96.2	50.1	1.66	82.8	92.6
Pythia-410M	IOI	49.9	92.6	50.1	2.47	86.3	88.2
	BLiMP	52.5	92.8	50.1	1.08	87.4	96.0
	Docstring	50.5	96.2	50.1	2.30	92.2	92.5
Qwen2.5-0.5B	IOI	51.1	99.9	97.9	0.25	99.2	98.1
	BLiMP	49.9	93.3	96.4	0.65	92.1	93.8
	Docstring	49.8	99.8	93.2	0.03	99.7	99.6

After task-agnostic completion, both GPT-2 models retain at least 95.3% accuracy on IOI and BLiMP with the original edge mask. Docstring is more sensitive to that mask: enabling all edges recovers 9–10 percentage points on both models. Pythia retains 86.3–92.2% accuracy on the original supports. Qwen retains 99.2%, 92.1% and 99.7% on IOI, BLiMP and Docstring, respectively. The zeros can therefore be replaced while preserving useful task behaviour, although the original routing does not always work as well on the completed weights.

Concealment and the refill tradeoff. A closer Gaussian fit can come at a cost to accuracy (Figure 3). For GPT-2 and Pythia, the highest-accuracy candidates in the completion sweep achieve 92.1–98.8% with best-fit histogram total variation (TV) of 0.1119–0.1671. The lowest-discrepancy candidates reach TV 0.0006, but accuracy falls to 57.1–92.4%. Qwen shows little such tradeoff in this sweep: its lowest-discrepancy candidates reach TV 0.0006–0.0009 at 100.0%, 92.8% and 99.3% accuracy, and accuracy varies by at most 0.8 percentage points within each task. Completion removes the pruning-induced point mass at zero. The remaining differences from a fitted Gaussian are measured separately in pooled and layerwise diagnostics (Appendix C.2).

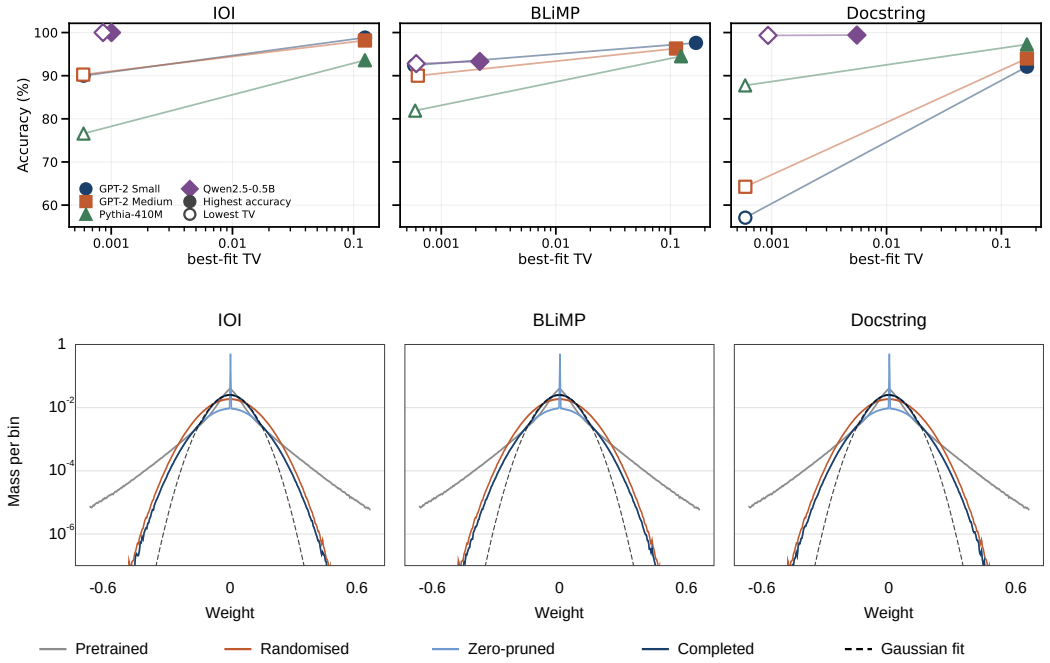


Figure 3: **Functional preservation and distributional concealment.** Top: extrema among 32 completion candidates per fixed model–task source (Table 12); lines pair the two extrema. Bottom: GPT-2 Small's four-stage pooled histograms and a fitted Gaussian at baseline completion $(f, a) = (0.5, 0.66)$. Bin mass uses a logarithmic scale; Appendix Figures 9–11 show GPT-2 and Pythia separately.

The shuffle control shows that the choice of refill matters (Figure 4). Restoring the original tensor multisets by shuffling removed values gives lower accuracy than fresh completion by 11.8–27.4 percentage points across the nine GPT-2 and Pythia pairs. Qwen differs by at most 1.0 percentage points between the two procedures. Thus the advantage of fresh completion is concentrated in GPT-2 and Pythia; Qwen tolerates both forms of refill under its configuration.

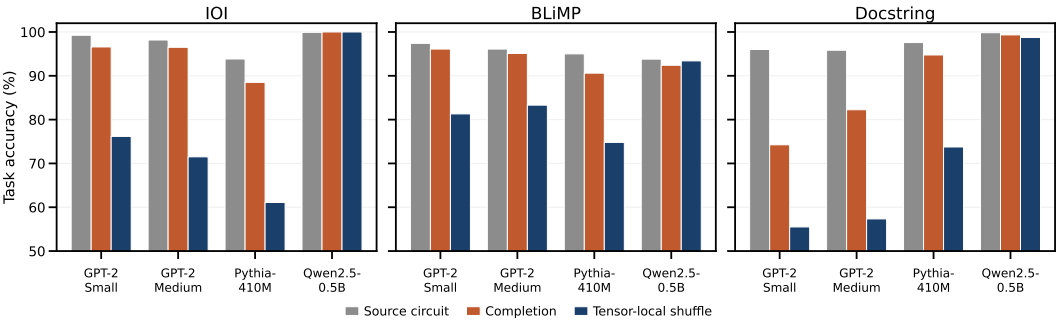


Figure 4: **Fresh completion and exact multiset restoration.** Accuracy on one fixed source per model–task pair, reusing its selected edge mask. Completion draws fresh values; shuffling reassigns removed values within their original tensors. Both preserve selected values. Their performance is similar on Qwen and differs substantially on GPT-2 and Pythia.

Fresh edge-only discovery after completion We next ask whether edge-only discovery can find accurate sparse circuits in the completed models, with all scalar parameters active. For each model–task pair, we fix one completed weight configuration and restart edge search from all edges with

ten random seeds. Table 4 reports the results, and Appendix Figure 13 compares them with the original-support baselines on the same weights.

Table 4: **Fresh edge-only discovery on completed weights.** Means over ten edge-search seeds on one fixed completed source per model–task pair. Accuracies are percentages; retained edges are counts. Raw complement disables the discovered edges on the same completed weights, without further optimisation.

Model	Task	Fresh accuracy	Retained edges	Raw complement accuracy
GPT-2 Small	IOI	98.4	946	49.1
	BLiMP	96.6	697	64.4
	Docstring	97.6	1,380	53.3
GPT-2 Medium	IOI	97.6	3,960	49.6
	BLiMP	95.3	2,388	57.9
	Docstring	95.6	5,028	55.1
Pythia-410M	IOI	93.0	6,375	48.1
	BLiMP	96.8	3,505	57.6
	Docstring	97.0	6,144	55.0
Qwen2.5-0.5B	IOI	99.9	2,363	51.3
	BLiMP	92.9	3,019	66.2
	Docstring	99.5	3,108	53.6

Fresh discovery reaches 92.9–99.9% accuracy across all twelve model–task pairs with 697–6,375 retained edges. It exceeds both the original-support and all-edges baselines in each GPT-2 and Pythia pair. The improvement is particularly clear on Docstring: accuracy rises from 74.2% to 97.6% on Small and from 82.2% to 95.6% on Medium. Qwen remains close to its already accurate baselines. Disabling the newly discovered edges reduces accuracy to 48.1–66.2% in the complementary graph. Completion therefore leaves weights from which edge search can recover task-functional circuits, even where the original routing loses accuracy.

5 From random weights to task margins

We use *margin synthesis* to describe the creation of positive task margins by zeroing selected random weights while leaving retained values unchanged. The analysis draws on finite-task circuit models (Chen et al., 2026), attention lottery tickets (Otsuka et al., 2026) and weight perturbation bounds (Neyshabur et al., 2018). We first show how selection can increase margins, then characterise Gaussian histogram completion and give sufficient conditions for refill to preserve the margins. The models are idealised; Appendices D–E give their assumptions, statements and proofs.

One mask can increase several task margins. We seek one binary weight mask that makes all K prescribed comparison margins positive. With all edges enabled, consider n scalar weights $\sigma \mathrm{X}_j/\sqrt{n}$, where X_j are independent standard Gaussians and $\sigma > 0$. The additive margin is $\widetilde{r}_i = b_i + \sigma n^{-1/2} \sum_j h_{ij} (m_\theta)_j \mathrm{X}_j$: b_i is a fixed offset and h_{ij} is the fixed effect of weight j on comparison i. Assume signs $q_j \in \{-1, 1\}$, chosen from these readouts before drawing the weights, give positive average alignment $d_i = n^{-1} \sum_j h_{ij} q_j > 0$ for every comparison. Define the root-mean-square sensitivity $v_i = (n^{-1} \sum_j h_{ij}^2)^{1/2}$.

Retain a weight exactly when $q_j \mathrm{X}_j > 0$, and set it to zero otherwise. This gives $(m_\theta)_j \mathrm{X}_j = q_j [q_j \mathrm{X}_j]_+$, where $[t]_+ = \max\{t, 0\}$. Zeroing the opposite sign creates a nonzero mean along the common orientation:

$$\mathbb{E}\widetilde{r}_i = b_i + \frac{\sigma\sqrt{n}\, d_i}{\sqrt{2\pi}}, \qquad \mathrm{Var}(\widetilde{r}_i) = \sigma^2 v_i^2 \left(\frac{1}{2} - \frac{1}{2\pi}\right). \tag{6}$$

Before selection, the mean is b_i. The mask creates positive drift by removing weights with the opposite sign. With d_i bounded away from zero and bounded v_i, the mean gain grows as $\sqrt{n}$ while fluctuations stay bounded. At sufficient width, all prescribed margins are positive with high probability (Theorem 2). This is an oracle existence construction: it uses the sign of each realised weight.

In the centred disjoint-block model, each edge sums n independent, zero-mean Gaussian scalar contributions, normalised by $\sqrt{n}$. At fixed Gaussian scale, an intact edge is $O_p(1)$, bounded in probability as n grows; zeroing negative terms within one fixed edge instead gives $\Theta_p(\sqrt{n})$, growing at order $\sqrt{n}$ (Proposition 3). Edge-only selection remains $O_p(1)$ at fixed $|E|$ and fixed retained-edge budget. This gives an idealised explanation for the edge-density advantage in §4: scalar masks act before contributions cancel.

Retained values limit the Gaussian fit. Partition the real line into N_{bin} bins, including tails. Let ρ_b be retained bin counts divided by $|\mathcal{I}_\theta|$, $p_b(\eta)$ the bin probabilities of a Gaussian with mean and standard deviation $\eta = (\mu, \sigma)$, and $\alpha = |\mathcal{P}|/|\mathcal{I}_\theta| > 0$ the fill fraction. For fill masses x_b,

$$\min_{x_b \geq 0,\, \sum_b x_b = \alpha} \frac{1}{2} \sum_b |\rho_b + x_b - p_b(\eta)| = \sum_b [\rho_b - p_b(\eta)]_+. \tag{7}$$

Completion cannot remove retained mass that already exceeds the target in a bin. Allocating fills in proportion to the remaining deficits attains the resulting optimum; integer rounding adds at most $N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$ (Proposition 4). The identity holds for any Gaussian mean and scale, including those fitted after task-dependent selection.

Small fills can preserve the constructed margins. Let $F_i(\boldsymbol{z})$ be the margin after filling $\mathcal{P}$ at fixed m_E, with pruned baseline $r_i^{\mathrm{pruned}} = F_i(\mathbf{0})$. For a magnitude cap $u > 0$, suppose F_i is continuously differentiable near the fill cube and define $L_i(u) = \sup_{\|\boldsymbol{z}\|_\infty \leq u} \|\nabla F_i(\boldsymbol{z})\|_1$, its maximum aggregate fill sensitivity. Then

$$\|\boldsymbol{z}\|_\infty \leq u \quad \Longrightarrow \quad |F_i(\boldsymbol{z}) - r_i^{\mathrm{pruned}}| \leq u L_i(u). \tag{8}$$

Every positive margin exceeding this bound survives, including when all pruned coordinates are refilled (Theorem 5). The appendix also gives exact moments under random assignment, the Gaussian-fit cost of a fill cap, and a spectral bound from the cited perturbation lemma. Finally, a linear-feature construction shows that Gaussian weight marginals after completion can coexist with positive drift from selection (Proposition 24).

6 Conclusion

Task-guided masking can construct accurate, edge-sparse circuits after eligible attention and MLP parameters have been randomised. This is the phenomenon we call the Lottery Circuit Hypothesis. Task-agnostic completion can then obscure the trace of weight pruning while preserving useful behaviour, and the completed weights support accurate circuits under fresh edge-only discovery in all four models. Sensitivity to refill differs across configurations, especially between Qwen and the other models.

The idealised analysis shows how scalar selection can create positive task margins without changing retained weights. It also identifies what limits completion: retained mass constrains the Gaussian fit, while fill magnitude and sensitivity constrain margin preservation.

In our view, these findings call for examining what a discovery procedure can construct before attributing its circuits to pretraining. A task-functional mask provides evidence of sufficiency under the evaluation procedure; its provenance requires further evidence. For procedures that optimise task accuracy and can mask individual weights, random-weight controls directly test whether success requires the eligible pretrained values.

Reproducibility statement

Equation (5) specifies the pruning and completion order. Appendix B describes the parameter domain, common-example selection, shared data splits, evaluation criteria and completion algorithm. Appendix B.3 gives optimisation settings and replication; Appendix C presents extended results and diagnostics. Appendices D–E contain the theoretical assumptions, statements and proofs.

Ethics statement

This work uses public pretrained models and task data and involves no human subjects. Distributional completion can obscure the numerical trace of pruning. Pooled weight audits may therefore miss evidence that task-guided selection contributed to the model's behaviour.

AI use statement

Generative AI assisted with implementing prototype methods, developing the theoretical analysis (including mathematical claims, critical proof ingredients and proof writing), and interpreting results. The research question and central idea originated with the authors, who determined the theoretical direction and selected and assembled the framework's components. The authors take responsibility for the final content of this work.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/20fdaf67581e6d7157376d1ed584040a-Abstract-Conference.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://academic.oup.com/book/26549.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All circuits lead to Rome: Rethinking functional anisotropy in circuit and sheaf discovery for LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66420.

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*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Jacob Drori. Weight-sparse circuits may be interpretable yet unfaithful. LessWrong, February 2026. URL https://www.lesswrong.com/posts/sHpZZnRDLg7ccX9aF/weight-sparse-circuits-may-be-interpretable-yet-unfaithful.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gabriel Franco, Lucas M. Tassis, Azalea Rohr, and Mark Crovella. Finding interpretable prompt-specific circuits in language models. arXiv preprint arXiv:2602.13483, 2026. URL https://arxiv.org/abs/2602.13483.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65901.

Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. Weight-sparse transformers have interpretable circuits. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63567.

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, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.141. URL https://aclanthology.org/2025.acl-long.141/.

Stefan Heimersheim and Jett Janiak. A circuit for Python docstrings in a 4-layer attention-only transformer. AI Alignment Forum, February 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6682–6691. PMLR, 2020. URL https://proceedings.mlr.press/v119/malach20a.html.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Maxime Méloux, Silviu Maxwell, François Portet, and Cyril Labbé. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *arXiv preprint arXiv:2502.20914*, 2025. URL https://arxiv.org/abs/2502.20914.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Hikari Otsuka, Daiki Chijiwa, Yasuyuki Okoshi, Daichi Fujiki, Susumu Takeuchi, and Masato Motomura. The strong lottery ticket hypothesis for multi-head attention mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24639–24647, 2026. doi: 10.1609/aaai.v40i29.39648. URL https://ojs.aaai.org/index.php/AAAI/article/view/39648. Extended version: https://arxiv.org/abs/2511.04217.

Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2599–2610. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1b742ae215adf18b75449c6e272fd92d-Abstract.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11893–11902, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Ramanujan_Whats_Hidden_in_a_Randomly_Weighted_Neural_Network_CVPR_2020_paper.html.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/412e7a0143cb33d3a7a9b6abe8d90470-Abstract-Conference.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Kevin Ro 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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

Lei Yu, Jingcheng Niu, Zining Zhu, Xi Chen, and Gerald Penn. Sheaf discovery with joint computation graph pruning and flexible granularity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8822–8837. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.446. URL https://aclanthology.org/2025.emnlp-main.446/.

Ziqian Zhong and Jacob Andreas. Algorithmic capabilities of random transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104357–104382, 2024. doi: 10.52202/079017-3315. URL https://papers.nips.cc/paper_files/paper/2024/hash/bccdd196d798a51a4961989984a9ed4a-Abstract-Conference.html.

Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/1113d7a76ffceca1bb350bfe145467c6-Abstract.html.

APPENDIX CONTENTS

A Circuit-discovery frameworks **14**

B Experimental details **15**
B.1 Functional performance, sparsity and completeness 15
B.2 Residual allocation and the completed distribution 17
B.3 Optimisation settings and replication 18

C Additional empirical results **19**
C.1 Parameter and edge density 19
C.2 Matched completion and shuffle controls 22
C.3 Completion and fresh edge discovery 24

D Theoretical models and statements **32**
D.1 From attention lottery tickets to task margins 32
D.2 A common mask can synthesise margins 33
D.3 Weight selection before sparse edge selection 33
D.4 Gaussian fit as constrained completion 34
D.5 Refill stability from the pruned baseline 35
D.6 When fit and function are compatible 36

E Theory: full statements and proofs **36**
E.1 Attention products, Gaussian pruning and margin transfer 36
E.2 Shared-mask synthesis with overlapping task readouts 38
E.3 Weight and edge signatures with controlled error 39
E.4 From components and parameters to task signatures 39
E.5 Binary commitment and rounding 40
E.6 Sequential synthesis 41
E.7 Random scalar reachability and width 41
E.8 Edge-only ceilings and cancellation–selection 43
E.9 Search-space capacity 44
E.10 Deterministic stability from the pruned baseline 44
E.11 Exact inactivity and local completion stability 45
E.12 Deficit allocation: the exact histogram calculation 45
E.13 Rank assignment is a task-agnostic permutation 46
E.14 Independent noise: a sharper functional bound 47
E.15 Magnitude-constrained allocation and margin preservation 47
E.16 Layerwise refill and the spectral perturbation lemma 48
E.17 Constraints on Gaussian fit and task performance 49
E.18 A constructive example with all pruned coordinates affecting the task 50
E.19 Scope of the guarantees 51

F Notation glossary **52**

A CIRCUIT-DISCOVERY FRAMEWORKS

Decomposition and intervention. In the residual-stream formulation of Elhage et al. (2021), components communicate through sums of earlier outputs. Unrolling the stream makes these routes explicit while preserving the attention and MLP computations, including their nonlinearities. Architectural circuits select components or routes from this graph. Feature circuits begin with a learned decomposition: Marks et al. (2025) write an activation as $a = \sum_k f_k(a)v_k + b + \epsilon(a)$, where v_k are sparse-autoencoder features, $f_k(a)$ their sparse coefficients, b the decoder bias and $\epsilon(a)$ the reconstruction error. Their graph includes feature activations and error terms at token positions. They select nodes and edges by thresholding approximate indirect effects estimated through attribution patching or integrated gradients. These choices determine what a circuit contains; the ablation procedure and evaluation metric determine how it is tested.

Faithfulness under patching. For clean and corrupted inputs $(x, \widetilde{x})$, let $p_G(\cdot \mid x)$ be the full-model distribution. In the formulation of Bhaskar et al. (2024), an edge mask m_E, with $m_\theta = \mathbf{1}_\theta$, produces activations

$$a_v^C = F_v\left(\sum_{u\in \mathrm{pa}(v)} [(m_E)_{uv} a_u^C + (1-(m_E)_{uv})\widetilde{a}_u]\right), \qquad \widetilde{a}_u = a_u^G(\widetilde{x}), \tag{9}$$

where a_v^C is node v's output, F_v its computation, and $\mathrm{pa}(v)$ its parents; the embedding is included among the source nodes. Under interchange ablation, retained contributions are recomputed in the masked graph, while removed contributions are replaced with cached reference activations from the full model. For retained edges $H \subseteq E$, let p_H denote the patched output distribution. With target sparsity $s_0 \in [0,1]$ and Kullback–Leibler divergence D_{KL}, their objective is

$$\min_{H\subseteq E} \mathbb{E}_{(x,\widetilde{x})} D_{\mathrm{KL}}(p_G(\cdot \mid x) \,\|\, p_H(\cdot \mid x, \widetilde{x})) \quad \text{subject to} \quad 1 - \frac{|H|}{|E|} \geq s_0. \tag{10}$$

Edge Pruning optimises relaxed hard-concrete masks with frozen model weights and a sparsity Lagrangian, then discretises them. Its implementation also multiplies edge gates by source-node gates to encourage sparsity.

ACDC (Conmy et al., 2023) instead searches greedily from outputs toward inputs. Writing $D(H) = \mathbb{E}D_{\mathrm{KL}}(p_G \| p_H)$ under the specified patching intervention, its removal test is $D(H \setminus \{e\}) - D(H) < \tau$, where $\tau \geq 0$ is the pruning tolerance. The original study primarily uses interchange ablation and also evaluates zero ablation. EAP (Syed et al., 2024) approximates the effect of patching an individual edge. For a scalar output metric L and edge activations a_e^{clean} and a_e^{corr} from clean and corrupted full-model runs, its score is

$$\Delta_e L = (a_e^{\mathrm{corr}} - a_e^{\mathrm{clean}})^\top \nabla_{a_e} L(x)\big|_{a_e = a_e^{\mathrm{clean}}}. \tag{11}$$

It averages absolute attribution scores over examples and keeps the top-k edges. EAP uses task-specific metrics because KL divergence from the unchanged model has zero gradient at the clean forward pass.

DiscoGP's objective. DiscoGP jointly selects weight and edge masks, whose combination is called a *sheaf* (Yu et al., 2025). In our notation this is $C = (m_\theta, m_E)$. For family $\nu \in \{\theta, E\}$ and eligible coordinate j, the Gumbel-sigmoid relaxation and straight-through estimator are

$$\begin{aligned} \widetilde{m}_{\nu,j} &= \sigma\left(\frac{l_{\nu,j} - \log(\log U_{\nu,j,1} / \log U_{\nu,j,2})}{T}\right), \\ (m_\nu)_j &= [\mathbf{1}\{\widetilde{m}_{\nu,j} > 1/2\} - \widetilde{m}_{\nu,j}]_{\mathrm{detach}} + \widetilde{m}_{\nu,j}, \end{aligned} \tag{12}$$

where $\sigma(t) = (1 + e^{-t})^{-1}$, $l_{\nu,j}$ are learned logits, $T > 0$ is the temperature, and the $U_{\nu,j,r}$ are independent $\mathrm{Uniform}(0, 1)$ draws. The detach operator blocks gradients: forward masks are binary and backward derivatives pass through $\widetilde{m}_{\nu,j}$. For a label set $\{y_k\}_{k=1}^{Q}$, $Q \geq 1$, let $\widehat{y}_i = \arg\max_{y \in \{y_k\}_{k=1}^{Q}} p_G(y \mid x_i)$ and $\widetilde{C} = (\mathbf{1}_\theta - m_\theta, \mathbf{1}_E - m_E)$. The objective is defined by

$$\begin{aligned}
\mathcal{L}_{\text{fidelity}} &= -\sum_i \log p_C(\widehat{y}_i \mid x_i), \\
\mathcal{L}_{\text{complete}} &= -\sum_i \frac{1}{Q} \sum_{k=1}^{Q} \log p_{\widetilde{C}}(y_k \mid x_i), \\
\mathcal{L}_{\text{sparse}} &= \frac{1}{|\mathcal{I}_\theta|} \sum_{j \in \mathcal{I}_\theta} \sigma(l_{\theta,j}) + \frac{1}{|E|} \sum_{e \in E} \sigma(l_{E,e}), \\
\mathcal{L}_{\text{GP}} &= \mathcal{L}_{\text{fidelity}} + \lambda_c \mathcal{L}_{\text{complete}} + \lambda_s \mathcal{L}_{\text{sparse}}.
\end{aligned} \tag{13}$$

Here p_C is the label distribution of the masked model under zero ablation. Fidelity is the negative log-likelihood of the full model's predicted labels; completeness uses a uniform target for the complement, and sparsity penalises separately normalised mask densities. The coefficients λ_c, λ_s weight the latter two terms.

Our experimental configuration. Our parameter vector θ contains the eligible attention and MLP coordinates defined in Appendix B.2. Their numerical values stay frozen during discovery. Weight masks zero pruned coordinates, and edge masks use zero ablation ($\widetilde{a}_u = 0$ in Eq. 9). We optimise the designated task-token comparison with completeness and overlap losses disabled; Appendix B.3 gives the settings. Sequential search fixes the first mask before learning the second. The edge-only control uses the same DiscoGP edge masks, zero ablation and task objective. After discovery, we measure performance under the complement defined in Appendix B.1. The randomisation experiment then asks whether task success requires the eligible pretrained values.

Related evidence from weight-sparse models. Gao et al. (2026) train transformers with sparse numerical weights, then use mean ablation to prune nodes and isolate task-performing circuits. Follow-up experiments reported by Drori (2026) find that successful pruned circuits can alter component roles and attention patterns, and behave differently from their source models on shifted inputs. This motivates testing how a selected computation relates to its source. We examine that relation by learning task-guided masks on frozen random attention and MLP parameters, with the remaining components fixed.

Task definitions and adaptations. IOI was introduced by Wang et al. (2023): the completion names the indirect object, rather than the repeated subject, and the original study measures their logit difference and the indirect-object probability. Docstring was introduced by Heimersheim & Janiak (2023), who study argument-name completion in a four-layer attention-only transformer using reStructuredText-style prompts. The original score compares the correct argument's logit with the largest logit among competing argument names. Our common-correct filtering, model set, train/evaluation splits and designated two-token comparisons are the adaptations specified in Appendix B.

B Experimental details

We use the GPT-2 (Radford et al., 2019), Pythia (Biderman et al., 2023) and Qwen2.5 (Qwen et al., 2025) families. This appendix specifies how we select examples, learn masks and complete the pruned weights in Figure 2, together with the evaluation criteria and parameter conventions.

B.1 Functional performance, sparsity and completeness

Functional criterion. Let K be the number of held-out evaluation examples. For each $i \in \{1, \ldots, K\}$, let (y_i^+, y_i^-) be the ordered target token pair at the last evaluated position, with the

task requiring y_i^+ to receive a higher logit than y_i^-. With $\ell_{i,y}(C)$ denoting the circuit's logit for token y, define

$$r_i(C) = \ell_{i,y_i^+}(C) - \ell_{i,y_i^-}(C), \qquad \mathrm{Acc}(C) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}\{r_i(C) > 0\}. \tag{14}$$

Here $r_i(C)$ is the target-token-pair logit margin, and $\mathbf{1}\{\cdot\}$ equals one when its condition holds and zero otherwise. Positive margins count as correct; ties count as failures. Our BLiMP adaptation evaluates these token pairs. Evaluation examples are disjoint from mask-training data, and the task loss uses the same designated comparison. The 50% reference represents uniform random choice between the two answers. Model and task conventions are given in §4.

Sparsity and completeness. We report retained counts and densities for scalar parameters, nodes and edges, dividing each count by its eligible total. Sparsity is one minus density. To assess completeness, we evaluate the complement of the discovered mask. For fresh edge discovery, this means inverting the edge mask on the same completed weights, without retraining or repair. Completeness losses are disabled during search.

Relation to the idealised analysis. The experiments use clipped Gaussian reinitialisation; the synthesis analysis uses an unclipped Gaussian, with the qualification stated in Remark 2. The theoretical results concern prescribed finite collections of comparisons. Generalisation to held-out examples is evaluated empirically. The end of this appendix explains how shared parameters affect the refill perturbation model.

Pruning modes and the sequential decomposition. Weight-only, edge-only, weight-then-edge, edge-then-weight and joint search optimise the same circuit object $C = (m_\theta, m_E)$. In Eq. (5), Phase θ learns scalar masks through a Gumbel-sigmoid relaxation with every prunable edge active. We fix the resulting binary mask before Phase E, which learns edge masks without changing the scalar mask or numerical parameters. Both stages are evaluated on the same examples. Completion replaces the pruned entries with fresh values, giving parameters θ^{comp} evaluated with an all-on scalar mask and the retained edge mask. The pruning-mode comparison optimises each alternative independently (§4). All searches use the full nonlinear model; the theory in §5 uses idealised additive models.

Mask optimisation. Appendix B.3 gives the optimisation settings for construction and the pruning-mode comparison. The latter uses matched settings for independent searches on each pretrained or randomised model and pruning mode. Numerical parameters remain frozen, and completeness and overlap losses are disabled. We evaluate complement performance after discovery.

Common examples across models. After deduplicating prompt text, we keep examples whose designated answers are distinct, valid single-token continuations under every model's tokenizer. We further require a strictly positive target-token-pair margin in all four pretrained models. The resulting text examples and disjoint training and evaluation splits are shared across models, with model-specific tokenisation. Evaluation is therefore conditional on pretrained success.

The IOI pool is balanced between ABBA and BABA prompts. BLiMP uses pairs with identical prefixes and differing final lexical targets, after removing terminal punctuation; Docstring uses reStructuredText-style argument completion. Training/evaluation sizes are 4800/1200 for IOI, 4000/1000 for BLiMP and 4800/1200 for Docstring, shared across all four models.

Sources of variation. For construction and paired completion, we vary the weight initialisation while fixing the data split and mask-search randomisation. For fresh edge discovery, we vary edge-search randomisation on one frozen completed model. Appendix B.3 gives the replication and pairing for each experiment.

Task dependence of the source and task-agnostic refill. The source mask is learned from task examples. After discovery, completion uses the retained values, pruned locations and fresh randomness, with no further task feedback. This is the sense in which refill is task-agnostic; the source state on which the analysis conditions remains task-dependent.

Completion reference and target. Let $f > 0$ be the reference scale relative to σ_{reinit}, let $a \in [0, 1]$ control the separation of the reference components, and write $\nu = f\sigma_{\text{reinit}}$ for the reference standard deviation. For each pruned coordinate $j \in \mathcal{P}$, the raw reference value is $\xi_j = \nu(a\mathrm{B}_j + \sqrt{1 - a^2}\,\mathrm{G}_j)$, where B_j takes -1 and 1 with equal probability and $\mathrm{G}_j \sim \mathcal{N}(0, 1)$; all signs and Gaussian draws are independent. The two component centres are $-\nu a$ and νa. When the sample variance is positive, centring and rescaling give the reference values mean zero and standard deviation ν. Nominal moments combine the retained values with this zero-mean reference. They determine an initial Gaussian fit and an equal-width histogram with tail categories. The target is fitted numerically by adjusting its mean and scale to reduce retained excess, with the tails merged during fitting and separated for allocation. For each numerically fitted target, positive residual mass determines largest-remainder counts, and fresh values drawn within each bin form the fill multiset.

Assignment and numerical checks. We sort the fill values and assign them to pruned coordinates in the reference's rank order. Assignment uses coordinate locations, reference randomness and distributional parameters, with no task score, gradient or repair step. We check bin counts after conversion to the stored precision. Appendix E states the probabilistic assumptions for the assignment analysis.

B.2 Residual allocation and the completed distribution

The reference distribution has standard deviation $f\sigma_{\text{reinit}}$, with $f > 0$, and two symmetric components whose relative separation is controlled by $a \in [0, 1]$ (Appendix B). The allocator fits a Gaussian with mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma} > 0$ to the retained values and reference fill using binned total variation, half the sum of absolute differences between corresponding bin probabilities. For the resulting partition $(B_b)_{b=1}^{N_{\text{bin}}}$ of the real line into N_{bin} histogram bins, let k_b be the retained count in bin B_b and $q_b = \Pr_{G\sim\mathcal{N}(\widehat{\mu},\widehat{\sigma}^2)}(G \in B_b)$ be that bin's fitted Gaussian probability. Here G denotes a scalar draw from the fitted Gaussian. With $[t]_+ = \max\{t, 0\}$ denoting the positive part, the continuous fill allocation is

$$d_b = [|\mathcal{I}_\theta| q_b - k_b]_+, \qquad x_b^\star = |\mathcal{P}| \frac{d_b}{\sum_{c=1}^{N_{\text{bin}}} d_c}, \qquad (|\mathcal{P}| > 0). \tag{15}$$

Here d_b is the nonnegative count deficit in bin B_b, and $x_b^\star$ the fill count before rounding. Fills are allocated in proportion to these deficits. Including both tails ensures $\sum_b q_b = 1$, so the denominator is positive whenever $|\mathcal{P}| > 0$. We use largest-remainder rounding for integer counts, draw fresh values within each bin and assign them in the reference's rank order. None of these steps uses task feedback or changes retained values.

The resulting scalar distribution is the mixture

$$\widehat{P}_{\text{comp}} = \frac{|\mathcal{I}_\theta| - |\mathcal{P}|}{|\mathcal{I}_\theta|} \widehat{P}_{\text{ret}} + \frac{|\mathcal{P}|}{|\mathcal{I}_\theta|} \widehat{P}_{\text{fill}}, \tag{16}$$

where $\widehat{P}_{\text{ret}}$ and $\widehat{P}_{\text{fill}}$ are the empirical distributions of retained and fill values; an empty component contributes zero mass. Gaussian fit is evaluated for this complete mixture. We also measure realised fill magnitudes, since their effect on task margins is distinct from the histogram objective. Section 5 analyses both the constraint imposed by retained mass and the conditions for preserving predictions.

Distributional diagnostics. Target total variation (TV) compares the completed histogram with the allocator's fitted Gaussian. A separate Gaussian refit provides a numerical best-fit diagnostic; its fitting partition merges the tails. Each sweep candidate can have its own target and nominal bins, so these discrepancies quantify each candidate's agreement with its own fitted Gaussian. Supplementary diagnostics include quantile-bin TV (QTV), approximate Kolmogorov–Smirnov (KS) discrepancy, moments, skewness and excess kurtosis. QTV is half the ℓ_1 discrepancy from uniform bin mass on Gaussian quantile bins, using the completed array's mean and variance. Approximate KS is the largest empirical–Gaussian cumulative-distribution discrepancy on a finite quantile grid. QTV uses 100 quantile bins; approximate KS uses a 1000-point quantile grid. Allocation uses 200 equal-width interior bins spanning seven nominal standard deviations on either side of the nominal mean, with two tail bins. We measure realised fill magnitudes after allocation, including tail draws. The pooled and layerwise diagnostics describe marginal distributions on the specified bins and quantile grid; they do not measure independence between coordinates.

Fixed context and parameter domains. The eligible domain includes attention and MLP matrices and their internal biases, including gated-MLP matrices where present. These coordinates jointly determine the Gaussian mean, standard deviation and clipping bounds, and all participate in randomisation, pruning, completion and distributional diagnostics. Numerical values are frozen during mask optimisation. Embeddings, LM heads, positional embeddings where present and normalisation operations remain outside this domain. Normalisation factors folded into loaded linear parameters are randomised with those parameters. Counts and densities refer to the loaded circuit representation, including expanded grouped-query key/value projections where applicable.

Shared attention-output biases. An attention-output bias is one physical parameter used by several head contributions. The notation $m_\theta \odot \theta$ omits the head-specific gates, which may retain different uses of the same value. For completion, we retain a physical coordinate if any head gate keeps it, refill it only if every gate removes it, and count it once in parameter statistics. Removing scalar masks can therefore restore a head's use of a retained value without changing that value. A refill-only perturbation bound must account separately for any such restored shared-bias contribution; its routing is otherwise fixed.

Scope of the fixed context. The experiment tests whether task-functional circuits require the pretrained attention and MLP values in the declared domain. Components outside that domain remain fixed and provide part of the computation. Testing their contribution would require extending the randomisation domain.

Implementation checks. We use the same bias-inclusive parameter domain for discovery and completion. Implementation checks verify that retained values stay unchanged, completion performs no repair, and assignment preserves reference-coordinate ranks. Edge totals are 231,877 for GPT-2 Medium, 231,493 for Pythia-410M and 179,749 for Qwen2.5-0.5B. Pythia's parallel attention/MLP block has no same-layer attention-to-MLP edges, accounting for its smaller total. We also check that sequential runs reproduce the first-stage density of the corresponding single-stage control and that the second stage leaves the first mask fixed.

B.3 Optimisation settings and replication

The search objective includes retained weight and edge densities with coefficients λ_θ and λ_E. Negative λ_θ rewards retention; positive coefficients penalise density. Table 5 lists settings for construction, the pruning-mode comparison and fresh edge discovery. Completeness and overlap regularisation are disabled in all three.

Table 5: **Mask-search settings.** Paired entries specify Phase θ / Phase E. A dash denotes a mask family held fixed.

Experiment	Epochs	Learning rates	Density coefficients
Construction (GPT-2 and Pythia)	$120/40$	$0.1/0.07$	$-0.1/1$
Construction (Qwen)	$160/60$	$0.01/0.03$	$-0.1/0.1$
Pruning-mode comparison	$160/60$	$0.01/0.03$	$0.5/0.1$
Fresh edge discovery	$-/40$	$-/0.07$	$-/1$

Qwen construction settings. Qwen construction uses the epoch counts and learning rates of the pruning-mode comparison in Table 5, with $\lambda_\theta = -0.1$ as in the other models. Under the GPT-2 and Pythia settings, Qwen's IOI accuracy stays near chance: 50.0% at $\lambda_\theta = -0.1$, 50.5% at 0 and 52.3% at +0.5, with retained edge density near 25%. The successful Qwen configuration changes several settings together. Weight-mask logits start with mean 3.0, the Gumbel-sigmoid temperature is 1.0, and optimiser weight decay is zero. The density coefficient begins its linear schedule after 32 epochs and reaches its base value over the next 96. GPT-2 and Pythia instead use logit mean 0.01, temperature 0.01 and weight decay 0.01. Both configurations use logit standard deviation 0.01, and all reported randomised runs include eligible biases.

The weight-stage density schedule. The value λ_θ is the base coefficient. GPT-2 Small, GPT-2 Medium and Pythia-410M use a three-phase schedule: ramp linearly to $1000\times$ the base during the

first 96 epochs, return to the base over the next 12, then hold. Thus $\lambda_\theta = -0.1$ gives a peak retention reward near -100. Qwen2.5-0.5B holds the coefficient at 0 for the first 32 epochs, ramps to the base over the next 96, then holds it. Its peak magnitude is three orders of magnitude below the default schedule. Initial logits of 3.0 give a retained density near 95% rather than 50%, so Qwen starts nearly dense and uses task loss alone for the first fifth of its weight stage. These runs establish that the combined configuration succeeds; they do not isolate the effects of its individual changes.

Joint search updates both mask families for $160 + 60 = 220$ epochs, using their respective learning rates and density schedules rescaled over that total.

Replication and pairing. For each model–task pair, we run construction from ten weight initialisations, fixing the data split and mask-search randomisation. Each source receives one completion with the same refill settings and random seed. Construction and paired-completion tables report arithmetic means over these ten sources. The pruning-mode comparison has one search per combination of model, task, weight origin and pruning mode.

Completion and fixed-source comparisons. All paired completions use the baseline $(f, a) = (0.5, 0.66)$. The shuffle control, completion sweep and four-stage distribution plots share one prespecified source per model–task pair. For the sweep, we cross $f \in \{0.1, 0.2, \ldots, 0.8\}$ with $a \in \{0.2, 0.4, 0.6, 0.8\}$ while fixing source weights, masks, refill randomisation and the evaluation split. The highest-accuracy and lowest-discrepancy candidates describe sensitivity to (f, a); paired-completion results continue to use the fixed baseline.

Fresh edge discovery. We freeze each baseline completed source and run ten fresh edge searches. Only edge-search randomisation varies; data and training order stay fixed. We report mean accuracy, retained edge count, edge-complement accuracy and overlap with the original edge support, all evaluated on the same completed weights.

C Additional empirical results

C.1 Parameter and edge density

Table 6 gives the full pruning-mode grid; Table 7 supplements the construction results with node density. The selected GPT-2 and Pythia circuits retain 34.8–83.4% of nodes and 0.66–2.93% of edges: few edges can connect a much larger fraction of the components. Qwen retains 11.9%, 41.0% and 3.5% of nodes on IOI, BLiMP and Docstring, with corresponding edge densities of 0.25%, 0.65% and 0.03%. Table 8 gives retained edge counts for the pruning-mode comparison in Table 1.

Table 6: **Full pruning-mode comparison.** Accuracy (%) from one search per model, task, origin and mode under shared search settings (Appendix B.3).

Model	Task	Origin	Weights only	Edges only	$\theta \to E$	$E \to \theta$	Joint
GPT-2 Small	IOI	Pretrained	100.0	100.0	100.0	100.0	100.0
		Randomised	99.8	60.1	100.0	99.6	100.0
	BLiMP	Pretrained	99.6	99.3	99.4	98.9	99.3
		Randomised	98.0	80.0	97.7	98.5	96.9
	Docstring	Pretrained	100.0	100.0	100.0	100.0	100.0
		Randomised	99.5	59.7	99.6	99.8	98.8
GPT-2 Medium	IOI	Pretrained	99.9	100.0	100.0	100.0	100.0
		Randomised	99.8	69.6	99.8	99.8	99.9
	BLiMP	Pretrained	99.5	99.1	99.0	99.4	98.9
		Randomised	97.6	91.3	97.5	98.0	97.6
	Docstring	Pretrained	100.0	100.0	100.0	100.0	99.9
		Randomised	99.1	62.1	99.3	99.5	98.3
Pythia-410M	IOI	Pretrained	99.8	100.0	100.0	99.9	100.0
		Randomised	99.8	89.0	99.9	99.8	99.9
	BLiMP	Pretrained	99.5	99.0	99.6	98.8	98.6
		Randomised	98.6	92.6	98.4	98.6	98.2
	Docstring	Pretrained	98.3	99.9	99.9	99.3	99.8
		Randomised	99.8	83.3	99.8	100.0	99.8
Qwen2.5-0.5B	IOI	Pretrained	99.7	100.0	100.0	100.0	100.0
		Randomised	99.8	51.2	99.9	99.9	100.0
	BLiMP	Pretrained	99.4	99.1	99.6	99.5	99.4
		Randomised	93.8	69.0	93.9	97.3	93.9
	Docstring	Pretrained	100.0	99.9	100.0	100.0	100.0
		Randomised	99.4	57.4	99.5	99.7	99.2

Table 7: **Construction accuracy and sparsity.** Means over ten weight initialisations per model–task pair. Accuracy and retained weight, node and edge densities are percentages.

Model	Task	Dense	Circuit	Weights	Nodes	Edges
GPT-2 Small	IOI	49.6	98.6	50.2	34.8	1.24
	BLiMP	48.7	96.7	50.4	37.3	1.23
	Docstring	50.1	96.7	50.1	68.1	2.93
GPT-2 Medium	IOI	50.0	97.7	50.1	77.9	1.55
	BLiMP	53.1	96.5	50.4	47.7	0.66
	Docstring	49.8	96.2	50.1	78.3	1.66
Pythia-410M	IOI	49.9	92.6	50.1	83.4	2.47
	BLiMP	52.5	92.8	50.1	71.3	1.08
	Docstring	50.5	96.2	50.1	79.7	2.30
Qwen2.5-0.5B	IOI	51.1	99.9	97.9	11.9	0.25
	BLiMP	49.9	93.3	96.4	41.0	0.65
	Docstring	49.8	99.8	93.2	3.5	0.03

Table 8: **Edge counts under randomised weights.** The last column is the ratio of edge-only to weight-then-edge retained counts. Eligible totals are 231,877 for GPT-2 Medium, 231,493 for Pythia-410M and 179,749 for Qwen2.5-0.5B.

Model	Task	Edges only		$\theta \to E$		Count ratio
		Count	Accuracy	Count	Accuracy	
GPT-2 Small	IOI	3,584	60.1	31	100.0	115.6×
GPT-2 Small	Docstring	3,787	59.7	137	99.6	27.6×
GPT-2 Medium	IOI	31,577	69.6	56	99.8	563.9×
GPT-2 Medium	Docstring	31,065	62.1	411	99.3	75.6×
Pythia-410M	IOI	46,608	89.0	842	99.9	55.4×
Qwen2.5-0.5B	IOI	24,547	51.2	26	99.9	944.1×
Qwen2.5-0.5B	Docstring	37,620	57.4	66	99.5	570.0×

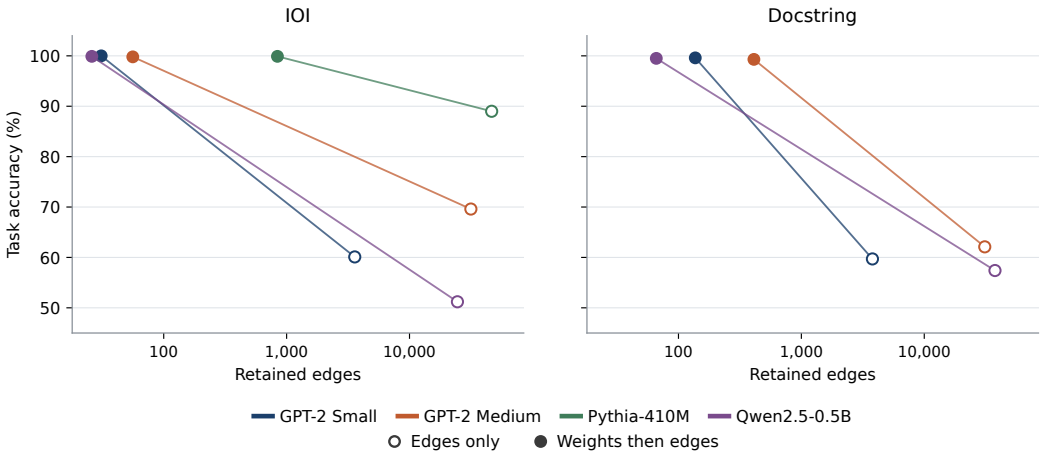


Figure 5: **Scalar selection enables more accurate circuits with fewer edges.** Circuits found with randomised weights in the pruning-mode comparison. Lines connect edge-only (open) to weight-then-edge (filled) discovery for the same model and task. The horizontal axis shows retained edge counts on a logarithmic scale. Table 8 supplies the numerical comparisons.

C.2 MATCHED COMPLETION AND SHUFFLE CONTROLS

Table 9: **Completion versus tensor-local shuffling.** One prespecified source per model–task pair; accuracy (%) on the same examples. Source, completion and shuffle use the original edge support; the last column enables all edges after shuffling.

Model	Task	Source circuit	Completion	Shuffle	Shuffle, all edges
GPT-2 Small	IOI	99.2	96.6	76.2	83.6
	BLiMP	97.4	96.1	81.3	87.7
	Docstring	96.0	74.2	55.5	62.4
GPT-2 Medium	IOI	98.2	96.5	71.5	84.8
	BLiMP	96.1	95.1	83.3	81.9
	Docstring	95.8	82.2	57.3	62.7
Pythia-410M	IOI	93.8	88.5	61.1	56.8
	BLiMP	95.0	90.6	74.8	89.7
	Docstring	97.6	94.8	73.8	72.5
Qwen2.5-0.5B	IOI	99.9	100.0	100.0	99.9
	BLiMP	93.8	92.4	93.4	94.3
	Docstring	99.8	99.3	98.8	99.2

Table 9 compares refill procedures on the same source circuit for each task. Shuffling moves original removed values only among pruned coordinates within their tensor, keeping selected values fixed and restoring the original tensor multisets exactly. Fresh completion changes both the fill distribution and the assignment, so their separate effects cannot be inferred from this comparison. GPT-2 and Pythia perform better under fresh completion. Qwen retains high accuracy under both procedures; its configuration also retains more weights and uses sparser routing.

Pooled distribution diagnostics. Across ten paired completions, GPT-2 and Pythia have best-fit histogram TV of 0.0254–0.0258, QTV of 0.0470–0.0475 and excess kurtosis of 0.853–0.860 (Table 11). Qwen's discrepancies are smaller: best-fit TV is 0.0017–0.0050, QTV 0.0023–0.0069 and excess kurtosis 0.030–0.093. Qwen retains 93.2–97.9% of eligible weights, leaving only 2–7% to refill, so its pooled distribution remains largely that of the original random draws. Target TV compares with the allocator's fitted Gaussian; best-fit TV uses a separate numerical refit.

Layer and parameter-family diagnostics. For one fixed source per model–task pair at $(f, a) = (0.5, 0.66)$, Table 10 reports refill fractions by parameter family and Gaussian-fit diagnostics by layer. The former counts refilled physical coordinates; the latter measures the combined retained and filled values. Shared attention-output biases follow the convention in Appendix B.2.

GPT-2 Small, GPT-2 Medium and Pythia have layerwise best-fit TV between 0.0250 and 0.0278, with a within-pair range of at most 0.0025 across layers. Their excess kurtosis of 0.845–0.867 shows a consistent departure from Gaussianity despite the small histogram TV. Qwen's layerwise TV is 0.0023–0.0068 and its excess kurtosis 0.026–0.082, an order of magnitude smaller. With 93.2–97.9% of eligible weights retained, most values in Qwen's completed layer histograms were never removed.

GPT-2 and Pythia refill 38.71–40.28% of bias coordinates and 49.41–49.95% of matrix coordinates, retaining a larger fraction of biases. Qwen refills less of every family: bias fill fractions are 0.0275–0.0921, compared with 0.0225–0.0831 for matrices. The bias–matrix ordering varies by task; biases are refilled most on BLiMP and least on IOI. These are physical-coordinate counts, including the shared-bias convention. Figure 6 summarises the fractions and layer ranges, and Figure 14 shows each layer.

Fresh edge discovery. We start fresh searches with all edges enabled and completed weights frozen, then measure accuracy, edge-complement accuracy, retained edge count and overlap with the original support. Intersection over union (IoU) is the number of shared edges divided by the size of their nonempty union.

Table 10: **Completion diagnostics by parameter family and layer.** One fixed source per model–task pair, completed with $(f, a) = (0.5, 0.66)$. A fill fraction is the proportion of coordinates in that parameter family replaced during completion. The final two columns give the minimum and maximum across transformer layers of the completed-weight histogram TV to a best-fit Gaussian and excess kurtosis, respectively.

		Fill fraction			Range across layers	
Model	Task	Attention matrices	MLP matrices	Biases	Best-fit TV	Excess kurtosis
GPT-2 Small	IOI	0.4979	0.4985	0.3900	0.0255–0.0262	0.855–0.863
	BLiMP	0.4960	0.4963	0.3871	0.0250–0.0272	0.845–0.864
	Docstring	0.4988	0.4988	0.3886	0.0252–0.0258	0.847–0.861
GPT-2 Medium	IOI	0.4993	0.4993	0.3887	0.0252–0.0259	0.848–0.858
	BLiMP	0.4941	0.4961	0.3874	0.0253–0.0278	0.850–0.867
	Docstring	0.4992	0.4992	0.3892	0.0251–0.0261	0.848–0.857
Pythia-410M	IOI	0.4991	0.4990	0.4018	0.0252–0.0263	0.849–0.860
	BLiMP	0.4991	0.4987	0.3985	0.0254–0.0264	0.851–0.862
	Docstring	0.4995	0.4994	0.4028	0.0252–0.0258	0.848–0.857
Qwen2.5-0.5B	IOI	0.0601	0.0365	0.0275	0.0039–0.0051	0.046–0.066
	BLiMP	0.0483	0.0225	0.0921	0.0023–0.0050	0.026–0.056
	Docstring	0.0831	0.0508	0.0540	0.0053–0.0068	0.063–0.082

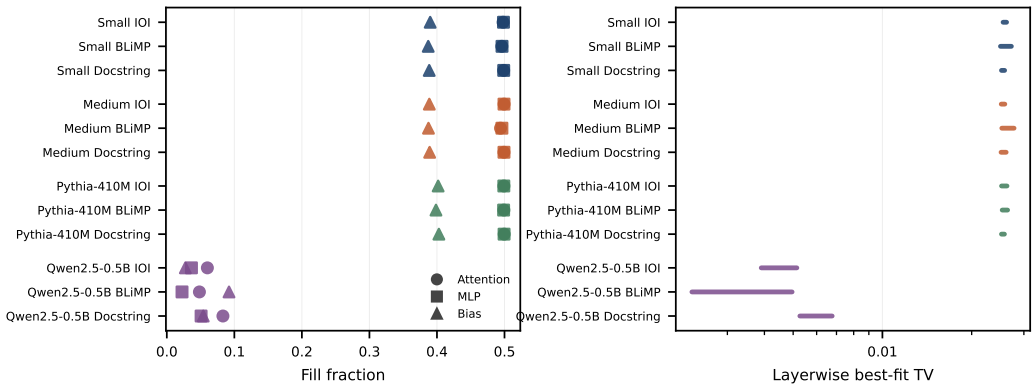


Figure 6: **Completion across parameter families and layers.** One fixed source per model–task pair at $(f, a) = (0.5, 0.66)$. Left: measured fill fractions by parameter family; the remaining coordinates retain their selected values. Right: the minimum and maximum completed-weight histogram TV to a best-fit Gaussian across layers. Segments span the observed layer minima and maxima. Table 10 also reports excess kurtosis.

C.3 COMPLETION AND FRESH EDGE DISCOVERY

Construction variability. Figure 7 follows each weight initialisation from the dense randomised model through selection and completion. Connecting each completion to its source shows how the effect of refill varies across initialisations.

Distributional completion. Table 11 gives Gaussian-fit diagnostics for completed weights. Figures 9–12 compare the four stages across all four models. The point mass introduced by zero pruning disappears after completion, while the central profile remains concentrated. Figure 6 separates layers and parameter families.

Table 11: **Completion diagnostics across the full grid.** Means over ten paired completions per model–task pair. Target TV uses the allocator Gaussian; best-fit TV uses a separately fitted Gaussian. QTV uses 100 Gaussian quantile bins. Kurtosis is excess kurtosis.

Model	Task	Target TV	Best-fit TV	QTV	Kurtosis
GPT-2 Small	IOI	0.0476	0.0257	0.0475	0.860
	BLiMP	0.0474	0.0257	0.0473	0.857
	Docstring	0.0473	0.0255	0.0472	0.856
GPT-2 Medium	IOI	0.0471	0.0254	0.0471	0.854
	BLiMP	0.0476	0.0258	0.0475	0.859
	Docstring	0.0471	0.0254	0.0470	0.853
Pythia-410M	IOI	0.0472	0.0254	0.0471	0.855
	BLiMP	0.0474	0.0256	0.0473	0.858
	Docstring	0.0471	0.0254	0.0470	0.853
Qwen2.5-0.5B	IOI	0.0023	0.0017	0.0023	0.030
	BLiMP	0.0040	0.0029	0.0040	0.052
	Docstring	0.0069	0.0050	0.0069	0.093

Scale and separation sweep. Figure 8 compares original-support accuracy and Gaussian discrepancy across the 8×4 scale–separation grid for each source. Table 12 summarises the highest-accuracy and lowest-discrepancy settings in each grid.

Table 12: **Completion sweep extrema.** One fixed source per model–task pair, with 32 candidates. Accuracy (%) uses the original edge support; TV is best-fit equal-width histogram total variation. Extrema use the shared evaluation split. Values rounding to zero at four decimal places are reported as < 0.00005.

		Highest accuracy			Lowest discrepancy		
Model	Task	(f, a)	Accuracy	TV	(f, a)	Accuracy	TV
GPT-2 Small	IOI	$(0.2, 0.2)$	98.8	0.1239	$(0.8, 0.8)$	90.0	0.0006
	BLiMP	$(0.1, 0.8)$	97.6	0.1671	$(0.8, 0.8)$	92.4	0.0006
	Docstring	$(0.1, 0.8)$	92.1	0.1669	$(0.8, 0.8)$	57.1	0.0006
GPT-2 Medium	IOI	$(0.2, 0.4)$	98.2	0.1238	$(0.8, 0.6)$	90.2	0.0006
	BLiMP	$(0.2, 0.8)$	96.3	0.1119	$(0.8, 0.8)$	90.0	0.0006
	Docstring	$(0.1, 0.8)$	94.0	0.1668	$(0.8, 0.6)$	64.2	0.0006
Pythia-410M	IOI	$(0.2, 0.2)$	93.6	0.1238	$(0.8, 0.8)$	76.6	0.0006
	BLiMP	$(0.2, 0.2)$	94.5	0.1238	$(0.8, 0.6)$	81.9	0.0006
	Docstring	$(0.1, 0.8)$	97.3	0.1668	$(0.8, 0.6)$	87.8	0.0006
Qwen2.5-0.5B	IOI	$(0.1, 0.4)$	100.0	0.0010	$(0.8, 0.4)$	100.0	0.0009
	BLiMP	$(0.6, 0.2)$	93.3	0.0022	$(0.8, 0.4)$	92.8	0.0006
	Docstring	$(0.3, 0.4)$	99.4	0.0055	$(0.8, 0.6)$	99.3	0.0009

Fresh edge-only discovery. Freshly discovered supports have mean IoU of 0.124–0.247 with the original supports across the nine GPT-2 and Pythia pairs. Qwen's means are 0.011, 0.111 and 0.014 for IOI, BLiMP and Docstring. Table 4 reports accuracy and edge-complement performance; Figure 13 plots fresh-search accuracy against edge density and marks the original-support baselines.

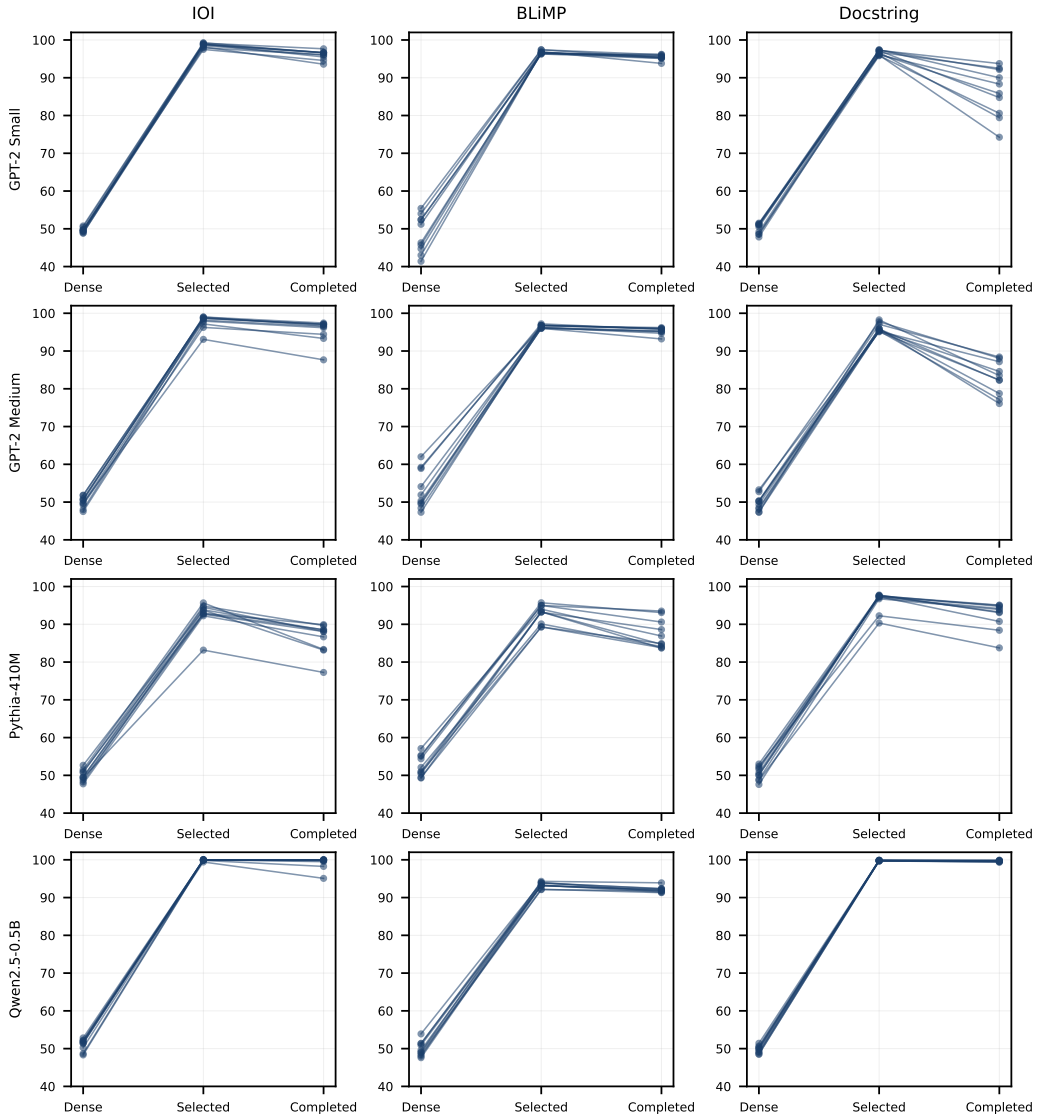


Figure 7: **Variation across weight initialisations.** Held-out accuracy (%) for dense randomised, selected and completed models, with the selected edge mask retained during completion. Each trajectory is one of ten paired weight initialisations. Rows identify models; columns identify tasks.

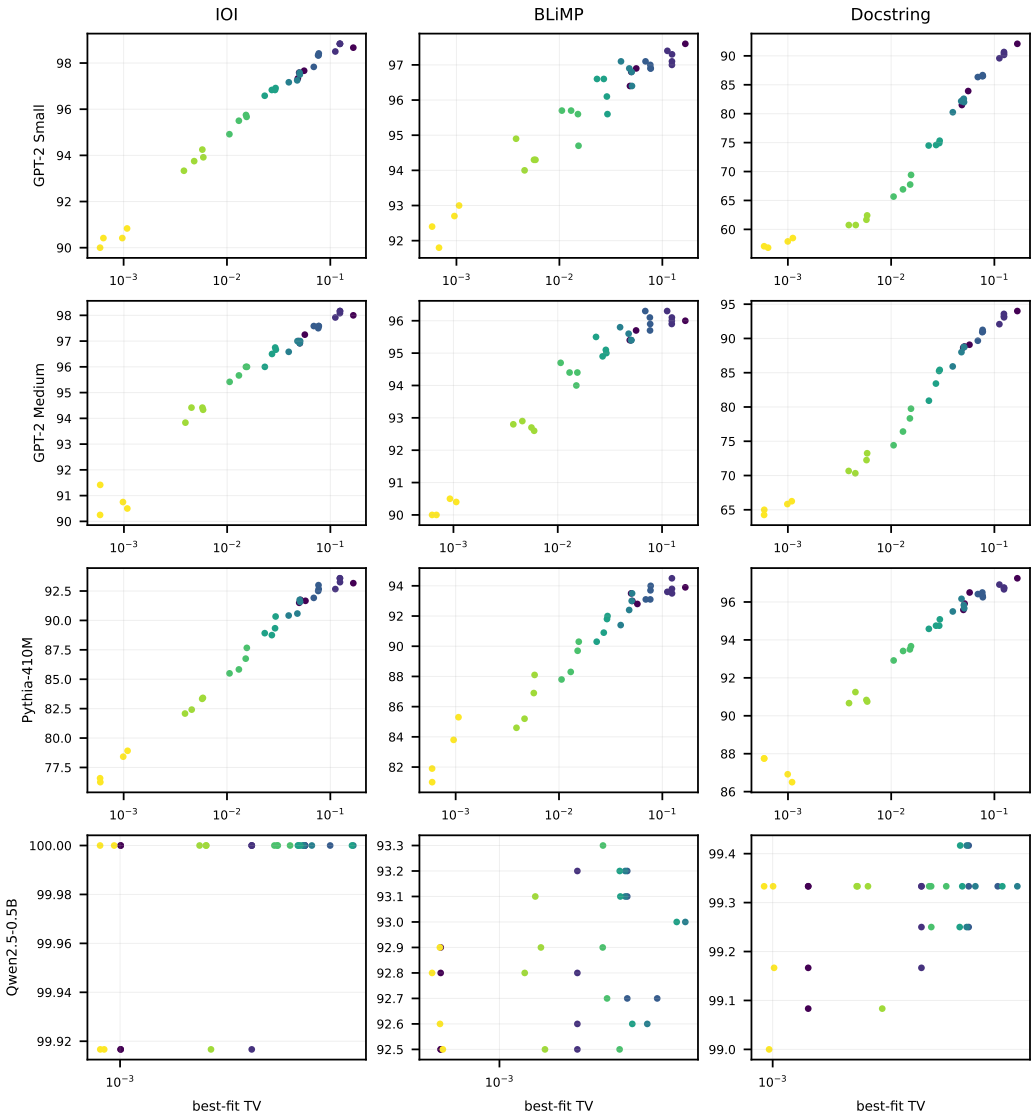


Figure 8: **Completion sensitivity across the full sweep.** Each point is one of 32 scale–separation candidates on a fixed source circuit. Horizontal axes show best-fit histogram TV on a logarithmic scale; vertical axes show original-support accuracy (%). Rows identify models and columns identify tasks; axis ranges differ across panels.

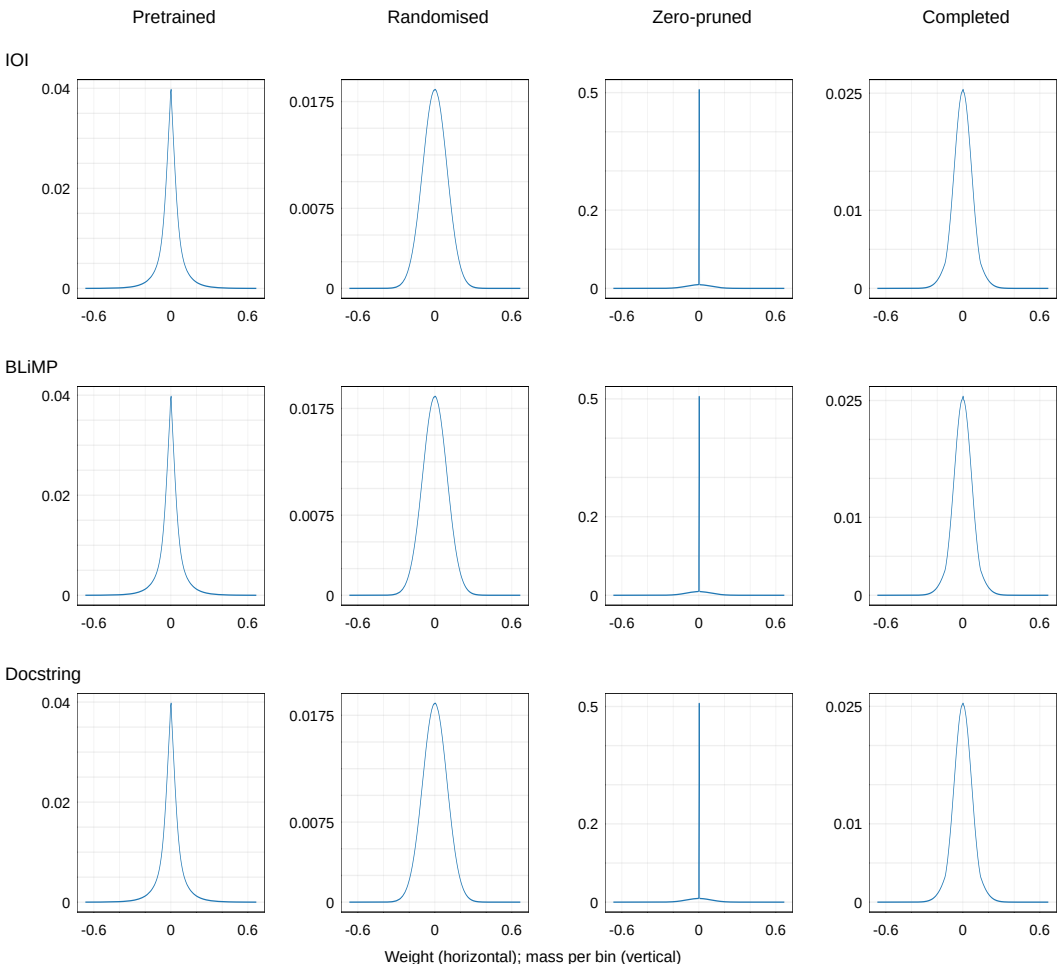


Figure 9: **Four-stage pooled distributions: GPT-2 Small.** Rows follow IOI, BLiMP and Docstring; columns follow Pretrained, Randomised, zero-pruned and completed weights. Each row uses one fixed source and common bin boundaries across stages. Vertical axes show mass per bin on stage-specific linear scales; completion uses $(f, a) = (0.5, 0.66)$.

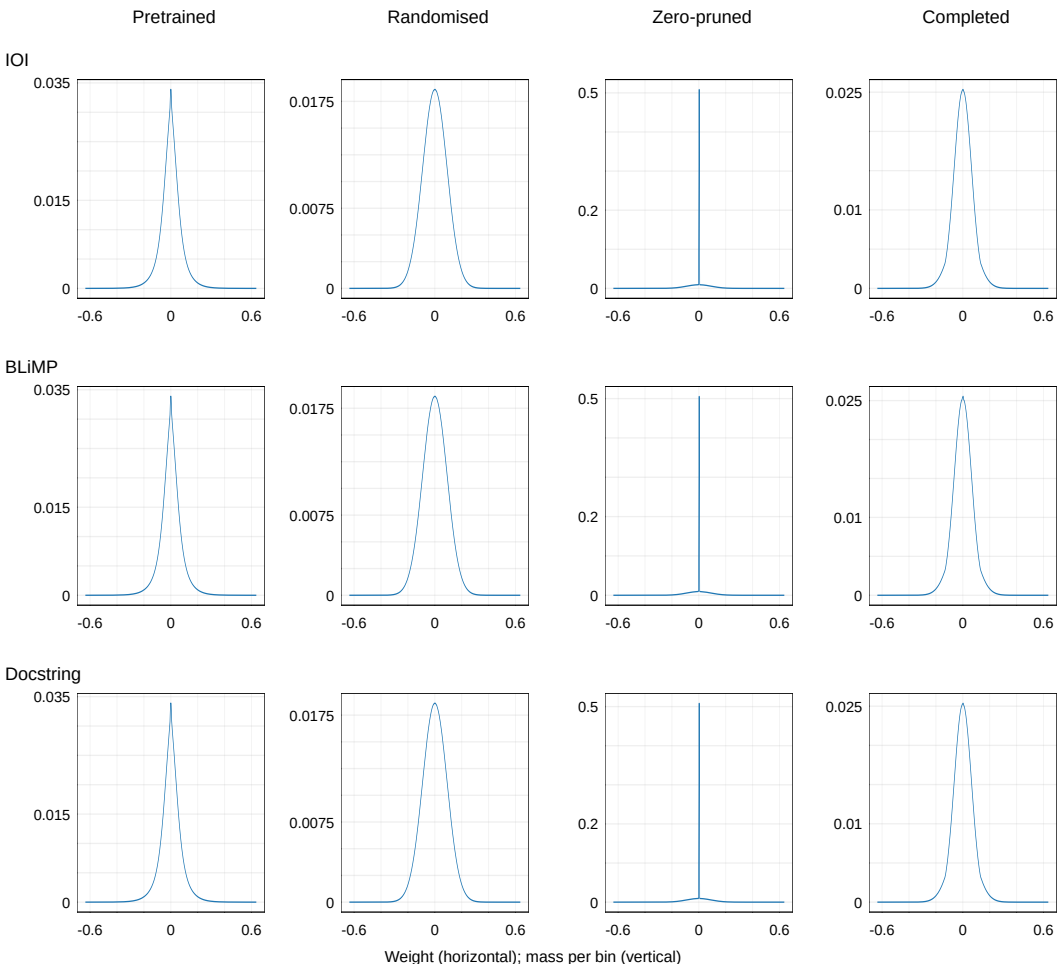


Figure 10: **Four-stage pooled distributions: GPT-2 Medium.** Rows follow IOI, BLiMP and Docstring; columns follow Pretrained, Randomised, zero-pruned and completed weights. Each row uses one fixed source and common bin boundaries across stages. Vertical axes show mass per bin on stage-specific linear scales; completion uses $(f, a) = (0.5, 0.66)$.

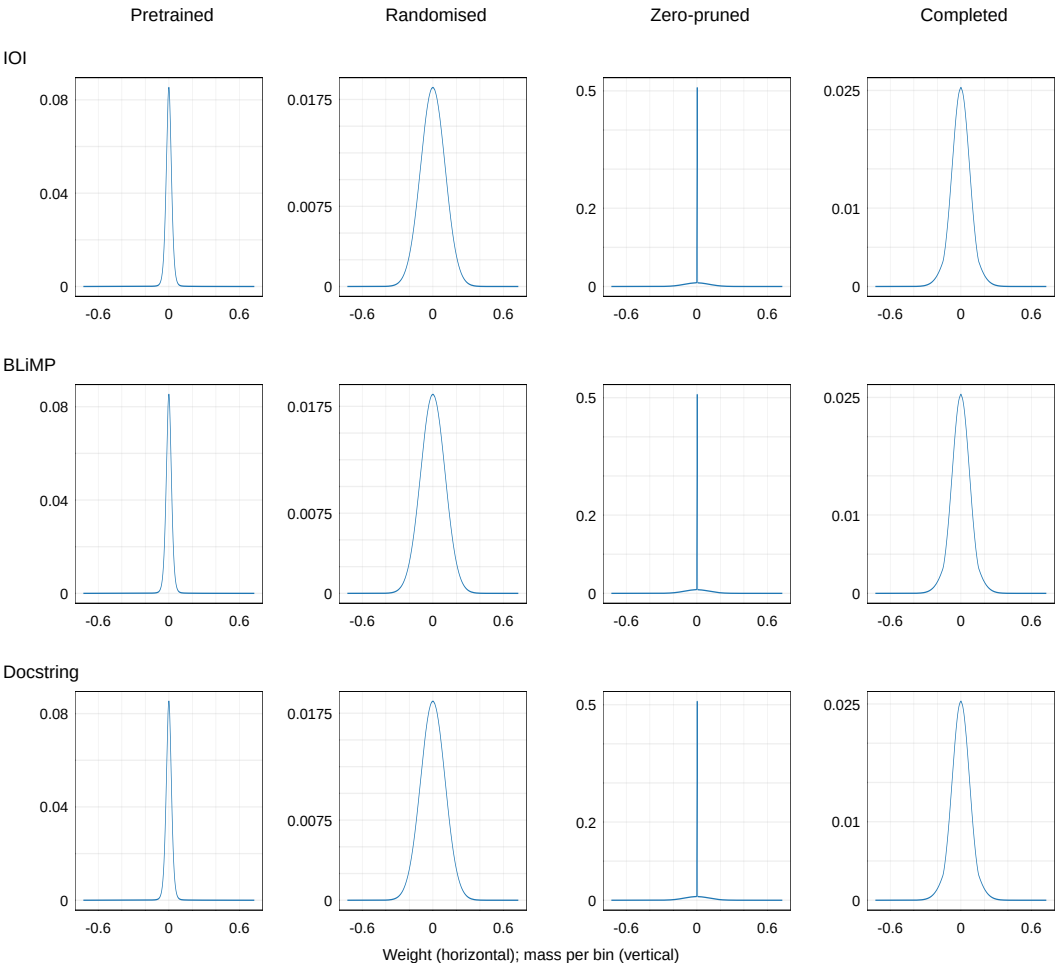


Figure 11: **Four-stage pooled distributions: Pythia-410M.** Rows follow IOI, BLiMP and Docstring; columns follow Pretrained, Randomised, zero-pruned and completed weights. Each row uses one fixed source and common bin boundaries across stages. Vertical axes show mass per bin on stage-specific linear scales; completion uses $(f, a) = (0.5, 0.66)$.

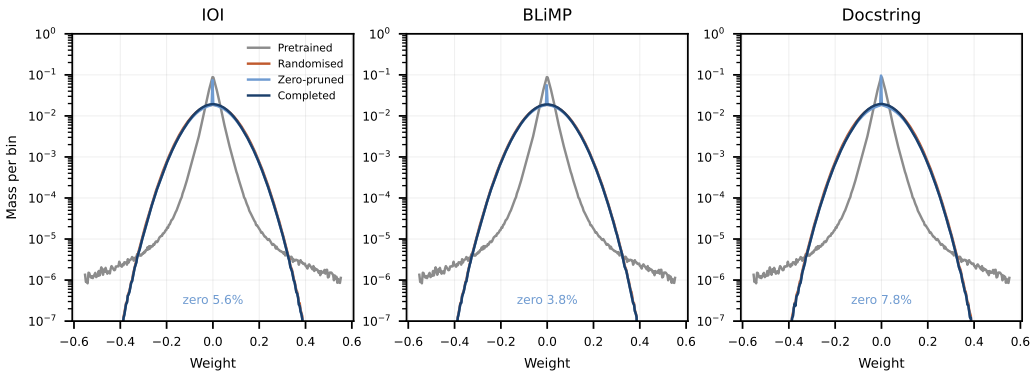


Figure 12: **Four-stage pooled distributions: Qwen2.5-0.5B.** Pooled histograms for pretrained, randomised, zero-pruned and completed weights on common bin boundaries, using one fixed source per task and completion $(f, a) = (0.5, 0.66)$. Panels follow IOI, BLiMP and Docstring. Vertical axes show bin mass on a logarithmic scale.

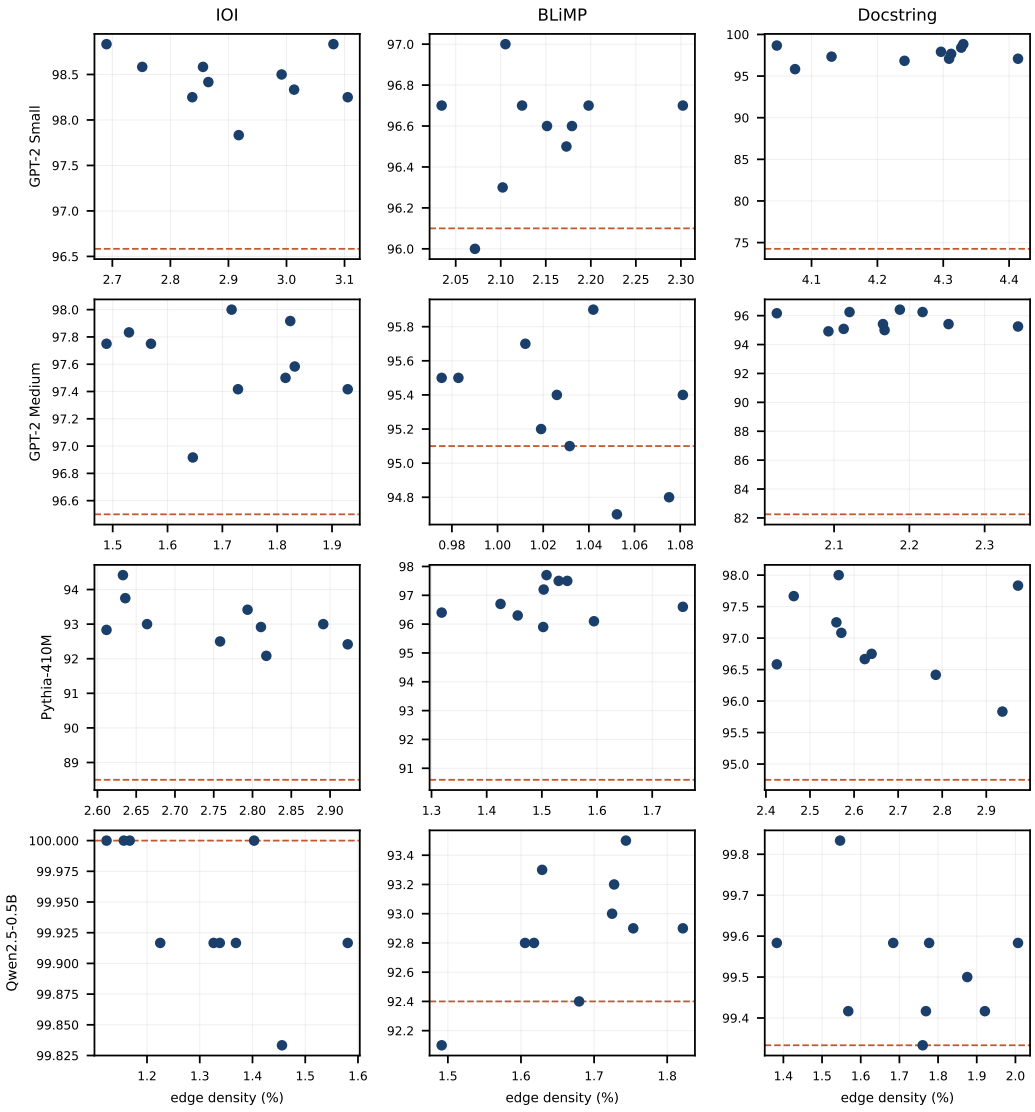


Figure 13: **Variation across fresh edge searches.** Held-out accuracy (vertical axis, %) against retained edge density (horizontal axis, %). Each point is one of ten edge-search seeds on a fixed completed source. Dashed lines show original-support accuracy; rows identify models and columns identify tasks. Table 4 gives complement performance.

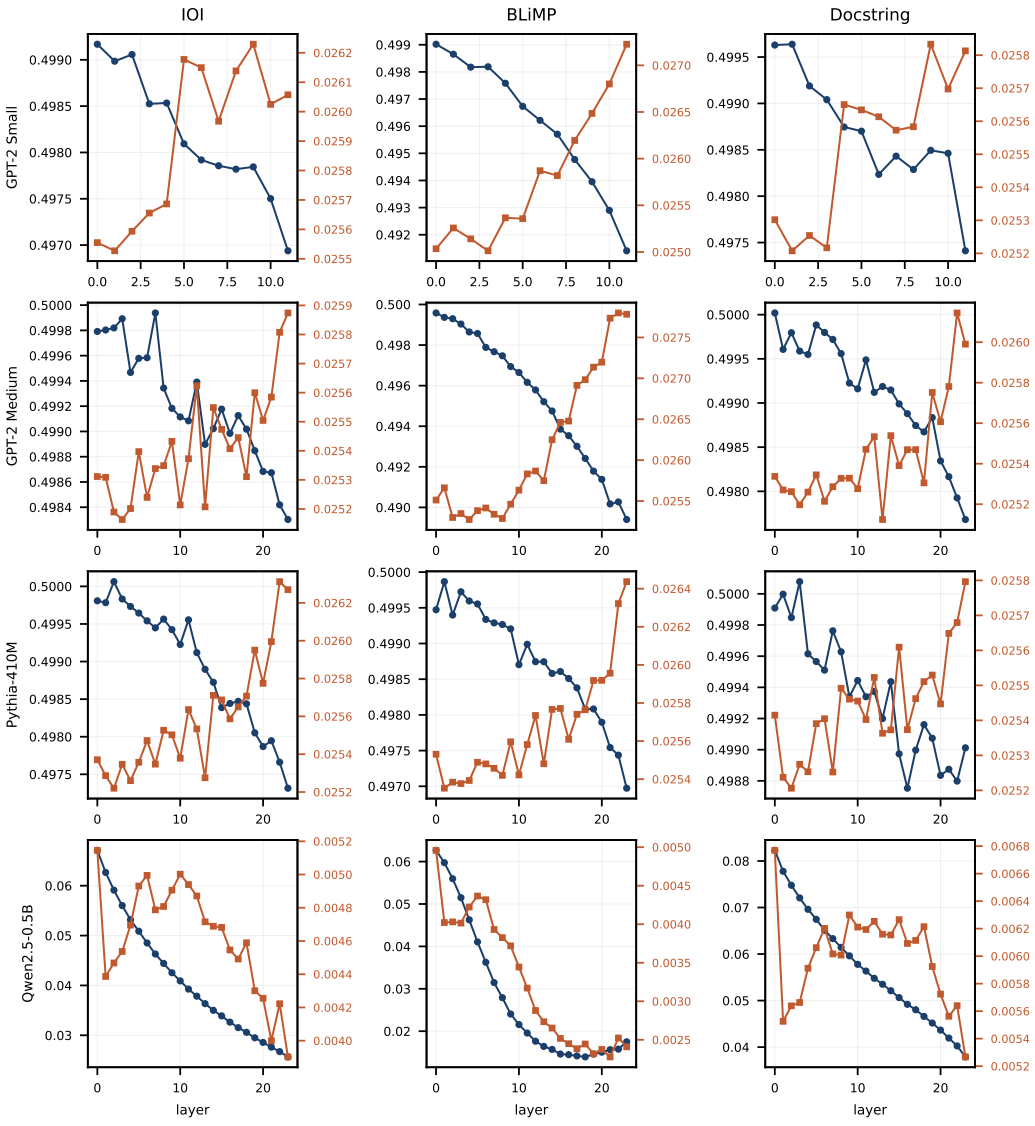


Figure 14: **Layerwise completion profiles.** Blue circles show the filled-coordinate fraction (left axis); orange squares show completed-weight best-fit histogram TV (right axis), indexed by layer. Rows identify models and columns identify tasks. Each panel uses one fixed source; the two axes have separate scales. Figure 6 and Table 10 summarise these diagnostics.

D THEORETICAL MODELS AND STATEMENTS

We study three aspects of the construction in simplified models: how zeroing random weights creates task margins, how refill changes the weight distribution, and when the margins survive. Finite-task readouts (Chen et al., 2026) let us express the contribution of each selected term. Attention lottery-ticket theory (Otsuka et al., 2026) provides a connection to pruning attention projections, while weight-perturbation bounds (Neyshabur et al., 2018) motivate the stability analysis. Each result states the assumptions it requires.

For vectors, $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the sum of absolute entries, Euclidean norm and largest absolute entry. For matrices, $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ the Frobenius norm. Logarithms are natural, and $\delta \in (0,1)$ denotes a failure-probability bound. Fix K prescribed comparisons between target token pairs, with desired positive margins τ_i, collected in $\boldsymbol{\tau} = (\tau_i)_{i=1}^K$. One circuit must satisfy all comparisons. The target region $\mathcal{M}_\tau = \{\boldsymbol{r} \in \mathbb{R}^K : r_i \geq \tau_i \ \forall i\}$ collects margin vectors that meet every threshold. Write $\boldsymbol{r} = (r_i)_{i=1}^K$ for the exact margin vector and $\widetilde{\boldsymbol{r}}$ for its surrogate. A uniform error bound $\kappa \geq 0$ controls their discrepancy through $\|\boldsymbol{r} - \widetilde{\boldsymbol{r}}\|_\infty \leq \kappa$ over the masks under consideration. Proofs and model details are in Appendix E; Appendix F collects the notation.

D.1 FROM ATTENTION LOTTERY TICKETS TO TASK MARGINS

Attention-specific lottery-ticket theory studies how pruning random query, key, value and output projections can approximate a target attention module (Otsuka et al., 2026). Let H be the number of attention heads and $h \in \{1, \ldots, H\}$ index a head. Its query, key, value and output projection matrices are $\mathbf{W}_{Q,h}$, $\mathbf{W}_{K,h}$, $\mathbf{W}_{V,h}$ and $\mathbf{W}_{O,h}$. For each projection Proj $\in \{Q, K, V, O\}$, write $\mathbf{W}^{m_\theta}_{\text{Proj},h} = (m_\theta)_{\text{Proj},h} \odot \mathbf{W}_{\text{Proj},h}$, with the mask restricted to the corresponding matrix. With query/key width n_K, the effective products are

$$\mathbf{A}_h = \frac{\mathbf{W}^{m_\theta}_{Q,h}(\mathbf{W}^{m_\theta}_{K,h})^\top}{\sqrt{n_K}}, \qquad \mathbf{P}_h = \mathbf{W}^{m_\theta}_{V,h}\mathbf{W}^{m_\theta}_{O,h}. \tag{17}$$

The query–key product $\mathbf{A}_h$ acts on the input feature space; the value–output product $\mathbf{P}_h$ maps that space to the common output feature space. Both are determined by the masks on their projection factors. Let $\boldsymbol{x}_t$ be the input token vector at position t, written as a column, and let $\mathbf{X}$ stack the rows $\boldsymbol{x}_t^\top$. Fix $B > 0$ as a uniform upper bound on the token-vector Euclidean norms: $\|\boldsymbol{x}_t\|_2 \leq B$ for every position and input under consideration. A head computes $\mathrm{softmax}_{\mathcal{A}_t}(\boldsymbol{x}_t^\top \mathbf{A}_h \mathbf{X}^\top)\mathbf{X}\mathbf{P}_h$, where $\mathcal{A}_t$ is the fixed nonempty set of positions available to token t. The restricted softmax normalises scores over that set and assigns zero elsewhere. The multi-head output sums these weighted value averages. Each task comparison applies a fixed linear readout at the queried token position. Its coefficients are the difference between the two target-token rows of a fixed logit projection.

Proposition 1 (attention approximation transfers task margins)**.** *Let $(\mathbf{A}^\star_h, \mathbf{P}^\star_h)$ be target products and let $\epsilon_{A,h}, \epsilon_{P,h} \geq 0$ bound the spectral approximation errors of the pruned products: $\|\mathbf{A}_h - \mathbf{A}^\star_h\|_2 \leq \epsilon_{A,h}$ and $\|\mathbf{P}_h - \mathbf{P}^\star_h\|_2 \leq \epsilon_{P,h}$. For comparison i, let L_i be the Euclidean norm of its pairwise readout and $r^\star_i$ its margin under the target products. With the same readout and fixed offset, the pruned margin satisfies*

$$r_i((m_\theta, \mathbf{1}_E)) \geq r^\star_i - L_i \sum_{h=1}^{H} \left(B\epsilon_{P,h} + B^3 \|\mathbf{P}^\star_h\|_2 \epsilon_{A,h} \right). \tag{18}$$

Thus one common set of projection masks achieves all desired margins whenever the right-hand side is at least τ_i for every comparison. The bound is uniform in sequence length.

Pruning can approximate attention products (Otsuka et al., 2026). Appendix E.1 gives an independent construction for centred Gaussian projection factors and explicit sufficient widths. Proposition 1 then transfers the approximation error to task margins. We next examine directly how selection can increase those margins.

D.2 A COMMON MASK CAN SYNTHESISE MARGINS

Consider a frozen random linear-feature model with n individually maskable scalar weight coordinates, indexed by $j \in \{1, \dots, n\}$:

$$\widetilde{r}_i((m_\theta, \mathbf{1}_E)) = b_i + \frac{\sigma}{\sqrt{n}} \sum_{j=1}^{n} h_{ij}(m_\theta)_j \mathrm{X}_j, \qquad \mathrm{X}_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad \sigma > 0. \tag{19}$$

The scalar weights are $\sigma \mathrm{X}_j/\sqrt{n}$. The readout matrix $\mathbf{H} = (h_{ij}) \in \mathbb{R}^{K \times n}$ and baseline offsets $\boldsymbol{b} = (b_i)_{i=1}^K$ are fixed before the draw. Entry h_{ij} gives the effect of scalar weight j on comparison i; a column records that weight's effects across all comparisons. This adapts the finite-task edge-readout abstraction of Chen et al. (2026) to scalar-weight signatures. Selection acts on the random coefficients along these fixed task directions.

The same mask must raise every comparison margin. We assume that the fixed readouts admit a common sign orientation with positive average alignment, while allowing an individual weight to help some comparisons and oppose others. Keeping Gaussian draws whose signs agree with that orientation makes their means follow the chosen signs. The resulting margin drift grows as $\sqrt{n}$; the $n^{-1/2}$ normalisation keeps the variance of independent fluctuations bounded.

Theorem 2 (shared-mask Gaussian synthesis)**.** *Fix an orientation-sign vector $\boldsymbol{q} \in \{-1, 1\}^n$ before drawing the coefficients. For comparison i, define its oriented average readout $d_i = n^{-1} \sum_j h_{ij} q_j$ and root-mean-square readout $v_i = (n^{-1} \sum_j h_{ij}^2)^{1/2}$. The mask $(m_\theta)_j = \mathbf{1}\{q_j \mathrm{X}_j > 0\}$ keeps each retained weight unchanged and sets every other weight exactly to zero. Its surrogate margins satisfy*

$$\mathbb{E}\widetilde{r}_i = b_i + \frac{\sigma\sqrt{n}}{\sqrt{2\pi}} d_i, \qquad \mathrm{Var}(\widetilde{r}_i) = \sigma^2 v_i^2 \left(\frac{1}{2} - \frac{1}{2\pi}\right). \tag{20}$$

For $0 < \delta < 1$, with probability at least $1 - \delta$, this same mask satisfies

$$r_i((m_\theta, \mathbf{1}_E)) \geq b_i + \frac{\sigma\sqrt{n}}{\sqrt{2\pi}} d_i - \sigma v_i \sqrt{2\log(K/\delta)} - \kappa \quad \textit{for every } i, \tag{21}$$

whenever the uniform surrogate error is at most κ. The retained-coordinate count satisfies $\|m_\theta\|_0 \sim \mathrm{Binomial}(n, 1/2)$.

Suppose a common orientation gives a positive lower bound $d_i \geq d_* > 0$ and a root-mean-square upper bound $v_i \leq v_*$ for every comparison, both uniform in n. The drift then grows as $\sqrt{n}$ while fluctuations remain bounded. For fixed K, fixed $\sigma > 0$ and bounded target margins, offsets and approximation error, the following bound makes all prescribed margins attainable with any fixed success probability $1 - \delta$ once n is sufficiently large. Let $T = \max_i [\tau_i - b_i + \kappa]_+$ be the largest margin deficit after accounting for the baseline and approximation error. A sufficient width for all target margins is

$$n \geq \frac{2\pi}{\sigma^2 d_*^2} \left(T + \sigma v_* \sqrt{2\log(K/\delta)}\right)^2. \tag{22}$$

The sign mask retains half the coordinates in expectation and, under the common-orientation assumption, creates positive drift for every comparison. This is a pruning-only existence result for finite-task margins. Appendix E.7 gives a complementary subset-sum construction: under its directional assumptions, binary masks approximate a growing box of target readouts with error $O(\sqrt{k \log(kn/\delta)/n})$ for fixed coefficient scales and k prescribed directions. Both connect margin synthesis to the strong lottery ticket hypothesis.

D.3 WEIGHT SELECTION BEFORE SPARSE EDGE SELECTION

To compare granularities on one random draw, use the nested additive model

$$\widetilde{\boldsymbol{r}}((m_\theta, m_E)) = \boldsymbol{o} + \sum_{e \in E} (m_E)_e \Big(\bar{\boldsymbol{a}}_e + \sum_{j \in \mathcal{I}_\theta} (m_\theta)_j \boldsymbol{a}_{ej} \Big). \tag{23}$$

Here $\boldsymbol{o}, \bar{\boldsymbol{a}}_e, \boldsymbol{a}_{ej} \in \mathbb{R}^K$ are, respectively, a fixed offset, an edge-specific baseline and the contribution of scalar coordinate j through edge e, each recording effects on all task comparisons. For fixed m_θ, this is a sum of edge signatures; scalar pruning changes those signatures before edge selection.

For the one-margin comparison below, specialise to zero offsets and disjoint blocks of n scalar parameters per edge. Label these coordinates by $(e, j) \in E \times \{1, \dots, n\}$, so the weight-mask entry is $(m_\theta)_{(e,j)}$. Thus n is the number of individually maskable scalar weight coordinates in each idealised edge block. Let $\sigma_\theta > 0$ set their weight scale. The surrogate is

$$\widetilde{r}((m_\theta, m_E)) = \frac{\sigma_\theta}{\sqrt{n}} \sum_{e \in E} (m_E)_e \sum_{j=1}^{n} (m_\theta)_{(e,j)} \mathrm{X}_{ej}, \qquad \mathrm{X}_{ej} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{24}$$

Each X_{ej} is a standardised Gaussian coefficient; its scaled scalar contribution is $\sigma_\theta \mathrm{X}_{ej}/\sqrt{n}$. The disjoint blocks isolate the difference between selecting scalar contributions and selecting their aggregate along an edge.

Proposition 3 (centred cancellation–selection separation)**.** *In the model of Eq. (24), fix $0 < \delta < 1$. Let $s_E \in \{1, \dots, |E|\}$ be the maximum number of retained edges and assume $n \geq 16\pi \log(4/\delta)$. With probability at least $1 - \delta$,*

$$\sup_{\|m_E\|_0 \leq s_E} \widetilde{r}((\mathbf{1}_\theta, m_E)) \leq \sigma_\theta s_E \sqrt{2 \log(4|E|/\delta)}, \qquad \widetilde{r}((m_\theta, m_E^{\{1\}})) \geq \sigma_\theta \sqrt{n}/6 \tag{25}$$

where edge 1 is fixed in advance, $m_E^{\{1\}}$ retains only that edge, and $(m_\theta)_{(e,j)} = \mathbf{1}\{e = 1\}\mathbf{1}\{\mathrm{X}_{1j} > 0\}$; hence $n > 72 s_E^2 \log(4|E|/\delta)$ gives a strict surrogate margin-capacity separation, and exact separation follows once the gap exceeds the two applicable remainders.

The mechanism is

$$\underbrace{n^{-1/2} \textstyle\sum_j \mathrm{X}_{ej}}_{O_p(1)\text{: cancellation}} \quad \text{versus} \quad \underbrace{n^{-1/2} \textstyle\sum_j [\mathrm{X}_{ej}]_+}_{\Theta_p(\sqrt{n})\text{: selection}}. \tag{26}$$

The identity $\mathbf{1}\{\mathrm{X}_{ej} > 0\}\mathrm{X}_{ej} = [\mathrm{X}_{ej}]_+$ expresses weight zeroing, with every retained value unchanged. The order notation means that the centred sum is bounded in probability, while the positive-part sum grows at order $\sqrt{n}$; Proposition 3 gives explicit bounds. An edge gate sees the sum after cancellation; a scalar gate can zero an unfavourable summand before aggregation. The selected mask works with all edges enabled because all other blocks are zero, so the same construction applies when weights are selected before edges. The comparison measures margin capacity in the stated centred, disjoint-block model. General edge-only ceilings and conditional sequential certificates are given in Appendices E.8 and E.6.

D.4 Gaussian fit as constrained completion

For completion, the parameter array is indexed by $\mathcal{I}_\theta$; parameters outside this array remain fixed within each comparison. Fix a histogram partition of $\mathbb{R}$ into N_{bin} interval bins $I_1, \dots, I_{N_{\text{bin}}}$, each with nonempty interior and with both tails included. For $|\mathcal{P}| \geq 1$, write $\alpha = |\mathcal{P}|/|\mathcal{I}_\theta|$ for the removed fraction. The retained and fill masses in bin b are normalised coordinate counts:

$$\rho_b = \frac{|\{j \in \mathcal{I}_\theta \setminus \mathcal{P} : \theta_j^{\text{reinit}} \in I_b\}|}{|\mathcal{I}_\theta|}, \qquad \phi_b(\boldsymbol{z}) = \frac{|\{j \in \mathcal{P} : z_j \in I_b\}|}{|\mathcal{I}_\theta|}. \tag{27}$$

Thus $\sum_b \rho_b = 1 - \alpha$ and $\sum_b \phi_b(\boldsymbol{z}) = \alpha$. For a fixed nonempty Gaussian family $\Lambda \subseteq \mathbb{R} \times (0, \infty)$, let $p_b(\eta) = \Pr_{G \sim \mathcal{N}(\mu, \sigma^2)}(G \in I_b)$ for $\eta = (\mu, \sigma) \in \Lambda$. The best Gaussian fit of the completed histogram and the least retained excess over this family are, respectively,

$$\begin{aligned} \mathrm{GFit}(\boldsymbol{z}) &= \inf_{\eta \in \Lambda} \frac{1}{2} \sum_b |\rho_b + \phi_b(\boldsymbol{z}) - p_b(\eta)|, \\ R_{\text{Ghist}} &= \inf_{\eta \in \Lambda} \sum_b [\rho_b - p_b(\eta)]_+. \end{aligned} \tag{28}$$

The objective selects the Gaussian mean and scale within Λ for the complete mixture of retained values and fills. It measures total variation between bin masses, with the partition and Gaussian family held fixed within each statement.

Proposition 4 (immutable retained mass and deficit allocation)**.** *For unrestricted fresh fills, $R_{\text{Ghist}} \leq \inf_z \mathrm{GFit}(\boldsymbol{z}) \leq R_{\text{Ghist}} + N_{\text{bin}}/(2|\mathcal{I}_\theta|)$. At any target η, let $E_\eta = \sum_b [\rho_b - p_b(\eta)]_+$ be its excess retained mass and $D_\eta = \alpha + E_\eta$ its total positive deficit. Allocating continuous fill mass $x_b = \alpha [p_b(\eta) - \rho_b]_+ / D_\eta$ to bin b attains total variation E_η. Largest-remainder integer allocation changes this discrepancy by at most $N_{\text{bin}}/(2|\mathcal{I}_\theta|)$.*

Retained values cannot be moved, so any excess over the target remains after completion. Filling the deficits proportionally attains the best continuous fit at that target. This proves optimality of the experimental allocator before rounding, for each fitted target separately. At the numerically selected target η, the completed discrepancy is at most $E_\eta + N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$; the extra term accounts for integer fill counts.

D.5 Refill stability from the pruned baseline

Let $F_i(\boldsymbol{z})$ be the *exact* margin of $\theta^{\mathrm{comp}}(\boldsymbol{z})$ with $C_{\mathrm{comp}} = (\mathbf{1}_\theta, m_E)$ after refilling the $|\mathcal{P}|$ holes. The pruned index set $\mathcal{P}$ and the edge mask learned on those pruned weights are held fixed, while the completed array is evaluated with $\mathbf{1}_\theta$. The baseline $r_i^{\mathrm{pruned}} = F_i(\mathbf{0}) = r_i((m_\theta, m_E))$ is the margin before refill. The argument $\mathbf{0}$ leaves every pruned coordinate at zero, so $F_i(\mathbf{0})$ evaluates the pruned model. This equality assumes that removing the scalar masks does not restore additional uses of retained shared parameters; Appendix B.2 describes the shared-bias qualification for the experimental implementation. Here $u > 0$ denotes a cap on each fill's absolute magnitude. Suppose F_i is continuously differentiable on a neighbourhood of the fill cube $[-u, u]^{|\mathcal{P}|}$ and define $L_i(u) = \sup_{\|\boldsymbol{z}\|_\infty \le u} \|\nabla F_i(\boldsymbol{z})\|_1$. This is the worst-case aggregate sensitivity of margin i over that cube: the gradient is taken with respect to the pruned-coordinate fill values.

Theorem 5 (small-magnitude refill, arbitrary assignment)**.** *For every fill with $\|\boldsymbol{z}\|_\infty \le u$,*

$$|F_i(\boldsymbol{z}) - r_i^{\mathrm{pruned}}| \le uL_i(u), \qquad \mathrm{Acc}_{\mathrm{fill}}(\mathbf{0}) - \mathrm{Acc}_{\mathrm{fill}}(\boldsymbol{z}) \le \frac{1}{K}\sum_i \mathbf{1}\{0 < r_i^{\mathrm{pruned}} \le uL_i(u)\}. \tag{29}$$

Here $\mathrm{Acc}_{\mathrm{fill}}(\boldsymbol{z}) = K^{-1}\sum_i \mathbf{1}\{F_i(\boldsymbol{z}) > 0\}$, with m_E fixed. For desired positive post-refill margins τ_i^{comp}, the conditions $r_i^{\mathrm{pruned}} \ge \tau_i^{\mathrm{comp}} + uL_i(u)$ ensure $F_i(\boldsymbol{z}) \ge \tau_i^{\mathrm{comp}}$ for every comparison.

This bound allows dependent, histogram-constrained fills at every pruned coordinate. It compares the available margin with the largest change allowed by the fill magnitude and sensitivity. These are sufficient conditions for preserving task performance; the allocator itself uses only weight statistics.

A layerwise certificate for refill. Consider a bias-free feedforward ReLU model of depth L, with fixed edge routing and a linear logit output layer. For layer $\ell \in \{1, \dots, L\}$, let $\mathbf{W}_\ell$ be its weight matrix and $\mathbf{W}_\ell^{\mathrm{pruned}} = (m_\theta)_\ell \odot \mathbf{W}_\ell$ its pruned baseline. Let $\mathbf{Z} = (\mathbf{Z}_\ell)_{\ell=1}^L$ collect same-shaped fill arrays, so $\mathbf{U}_\ell = ((\mathbf{1}_\theta)_\ell - (m_\theta)_\ell) \odot \mathbf{Z}_\ell$ is the perturbation supported on that layer's pruned entries. Assume $\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2 > 0$. In this model, $B > 0$ uniformly bounds the Euclidean norms of the evaluated input vectors $\boldsymbol{x}_i$.

Corollary 6 (spectral certificate for dependent refill)**.** *If $\|\mathbf{U}_\ell\|_2 \le \|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2/L$ for every layer, then for every input $\|\boldsymbol{x}_i\|_2 \le B$ the pairwise logit margin obeys*

$$|F_i(\mathbf{Z}) - r_i^{\mathrm{pruned}}| \le D_{\mathrm{spec}}(\mathbf{Z}) := \sqrt{2}\, eB \left(\prod_{\ell=1}^L \|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2\right) \sum_{\ell=1}^L \frac{\|\mathbf{U}_\ell\|_2}{\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2}. \tag{30}$$

Consequently $\mathrm{Acc}_{\mathrm{fill}}(\mathbf{0}) - \mathrm{Acc}_{\mathrm{fill}}(\mathbf{Z}) \le K^{-1}\sum_i \mathbf{1}\{0 < r_i^{\mathrm{pruned}} \le D_{\mathrm{spec}}(\mathbf{Z})\}$. This deterministic statement permits dependent, non-centred, histogram-constrained fills.

This applies Lemma 2 of Neyshabur et al. (2018) at the *pruned* weights. The factor $\sqrt{2}$ is the Euclidean norm of the pairwise target logit readout, and $e = \exp(1)$. The certificate applies to arbitrary realised fills satisfying the spectral condition. If layer ℓ has u_ℓ pruned entries, each filled with magnitude at most u, then $\|\mathbf{U}_\ell\|_2 \le \|\mathbf{U}_\ell\|_F \le u\sqrt{u_\ell}$. Appendix E.16 gives the proof, the general product bound, conditions for both Gaussian fit and margin preservation, and extensions to affine layers.

Exact moments under task-agnostic random assignment. A complementary calculation separates the fill's values from their placement. For an additive refill margin $F_i(\mathbf{z}) = r_i^{\mathrm{pruned}} + \boldsymbol{g}_i^\top \mathbf{z}$, let $\boldsymbol{g}_i \in \mathbb{R}^{\mathcal{P}}$ be the fixed vector of per-coordinate margin sensitivities. Choose a fixed ordering of $\mathcal{P}$ and condition on a fill multiset $a_1, \dots, a_{|\mathcal{P}|}$, $|\mathcal{P}| \ge 2$: the sampled values before assignment to coordinates, with repeated values counted separately. Let $\mathbf{z}$ be the vector obtained by a uniformly random

permutation of this multiset. The multiset mean and population variance are $\bar{a} = |\mathcal{P}|^{-1} \sum_j a_j$ and $v^2 = |\mathcal{P}|^{-1} \sum_j (a_j - \bar{a})^2$. Writing $\mathbf{1} \in \mathbb{R}^{|\mathcal{P}|}$ for the all-ones vector, the completed-margin mean μ_i and variance V_i over assignment randomness are

$$\begin{aligned} \mu_i &:= \mathbb{E}F_i = r_i^{\text{pruned}} + \bar{a}\mathbf{1}^\top \boldsymbol{g}_i, \\ V_i &:= \mathrm{Var}(F_i) = \frac{v^2}{|\mathcal{P}| - 1} \left(|\mathcal{P}| \|\boldsymbol{g}_i\|_2^2 - (\mathbf{1}^\top \boldsymbol{g}_i)^2 \right). \end{aligned} \tag{31}$$

If every $\mu_i > 0$, a one-sided variance bound and a union bound give

$$\Pr\{F_i(\mathbf{z}) > 0 \text{ for all } i\} \geq \left[1 - \sum_{i=1}^{K} \frac{V_i}{V_i + \mu_i^2} \right]_+ . \tag{32}$$

The multiset's histogram is invariant under assignment. Its distributional fit and this functional bound therefore hold together, without protecting any coordinate from refill. Appendix E.13 proves these formulas and gives the exchangeability and independence conditions under which reference-rank assignment implements the uniform conditional permutation. Appendix E.11 adds nonlinear error control.

D.6 When fit and function are compatible

A magnitude cap limits the perturbation to task margins, but can prevent fills from reaching bins needed for a close Gaussian fit. Theorem 19 gives the exact additional distributional cost of that cap, with every pruned coordinate available for refill.

Proposition 24 gives a random linear-feature model in which Gaussian fit and positive margins coexist. Here σ is the original Gaussian weight scale, α the expected removed fraction and ν the refill standard deviation. For $\alpha\sigma \leq \nu \leq \sigma$, independent task-agnostic refill restores an exactly Gaussian *population* marginal. Both pruned and completed margins have mean $c\sqrt{n}$, where $c > 0$ is the selection-induced drift coefficient, and variance at most σ^2. At half removal, the refill scale can be half the original Gaussian scale. Every refilled coordinate affects the task: the completed weights retain an alignment with the readouts even though their marginal distribution is Gaussian. The empirical study measures finite-bin fit and task performance.

E Theory: full statements and proofs

All task guarantees in this appendix concern a fixed finite set of K last-token comparisons. For completion, we condition on the realised initialisation, selected masks and retained values. The edge mask may depend on the weight mask and initialisation; fresh completion randomness is independent of mask training. We follow the main-text notation and also use $\|\cdot\|_{\text{op}}$ for the spectral norm.

E.1 Attention products, Gaussian pruning and margin transfer

Architectural connection. The effective-product viewpoint in Eq. (17) follows Otsuka et al. (2026). In their extended version, Lemma 4 constructs binary masks for a product of two uniform random matrices; Theorem 3 combines that construction for query–key and value–output products with attention stability. Theorem 6 treats a specified transformer without normalisation. We use the product viewpoint, prove the task-margin transfer below, and give an independent Gaussian-factor construction.

Lemma 7 (length-independent masked-softmax stability)**.** *For a fixed nonempty allowed-position set $\mathcal{A}$, let $\boldsymbol{p}(\boldsymbol{s})$ be the probability vector obtained by applying softmax to scores at positions in $\mathcal{A}$ and setting all other entries to zero. For arbitrary score vectors $\boldsymbol{s}, \boldsymbol{t}$ of the same length,*

$$\|\boldsymbol{p}(\boldsymbol{s}) - \boldsymbol{p}(\boldsymbol{t})\|_1 \leq \|(\boldsymbol{s} - \boldsymbol{t})_{\mathcal{A}}\|_\infty. \tag{33}$$

For value vectors $\boldsymbol{v}_j$ in a common Euclidean space with a uniform norm bound $R \geq 0$, if $\|\boldsymbol{v}_j\|_2 \leq R$, then $\| \sum_j (p_j(\boldsymbol{s}) - p_j(\boldsymbol{t}))\boldsymbol{v}_j \|_2 \leq R \|(\boldsymbol{s} - \boldsymbol{t})_{\mathcal{A}}\|_\infty$.

Proof. Restrict to $\mathcal{A}$. The Jacobian is $\mathbf{J} = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$. For any vector $\boldsymbol{a}$,

$$\|\mathbf{J}\boldsymbol{a}\|_1 = \sum_j p_j |a_j - \sum_k p_k a_k| \le \sqrt{\mathrm{Var}_{j\sim\boldsymbol{p}}(a_j)} \le \|\boldsymbol{a}\|_\infty.$$

Integrate along the segment from $\boldsymbol{s}$ to $\boldsymbol{t}$. The weighted-output claim follows by the triangle inequality. The bound holds for arbitrary score errors and is independent of the number of allowed positions. □

Proof of Proposition 1. Use the same tokens and allowed-position sets in both models. For head h, the score error at any allowed position is at most $B^2\epsilon_{A,h}$. Write $\boldsymbol{p}_h, \boldsymbol{p}_h^\star$ for the two attention rows. Decompose the output difference as

$$\boldsymbol{p}_h\mathbf{X}\mathbf{P}_h - \boldsymbol{p}_h^\star\mathbf{X}\mathbf{P}_h^\star = \boldsymbol{p}_h\mathbf{X}(\mathbf{P}_h - \mathbf{P}_h^\star) + (\boldsymbol{p}_h - \boldsymbol{p}_h^\star)\mathbf{X}\mathbf{P}_h^\star.$$

The first term has norm at most $B\epsilon_{P,h}$ because $\boldsymbol{p}_h$ is a probability vector. The second is at most $B\|\mathbf{P}_h^\star\|_2 B^2\epsilon_{A,h}$ by Lemma 7. Sum over heads and apply the fixed pairwise readout. The offset cancels. One set of projection masks controls every input satisfying the token-norm bound. □

An elementary Gaussian product construction. With frozen centred Gaussian factors, positive and negative products can span a target range. The next result gives sufficient conditions for this to occur and uses scalar rounding to turn the relaxed coefficients into binary pruning masks.

Proposition 8 (binary approximation by Gaussian factors)**.** *Let $\mathbf{U} \in \mathbb{R}^{a\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times b}$ have mutually independent entries $\mathrm{U}_{ij} \sim \mathcal{N}(0, \sigma_U^2)$ and $\mathrm{V}_{jk} \sim \mathcal{N}(0, \sigma_V^2)$, with $\sigma_U, \sigma_V > 0$. The positive integers a, b are the target matrix dimensions, $n = bn_{\mathrm{grp}}$ is the intermediate width, and the positive integer n_{grp} is the number of intermediate coordinates assigned to each target column. Fix a product normalisation $\lambda > 0$, a band threshold $c > 0$ and $0 < \delta < 1$. Write $\omega = \sigma_U\sigma_V$ for the product scale and $\beta_c = \mathbb{P}(c \le \mathrm{GH} \le 2c) > 0$ for the probability of the positive product band, where G, H are independent standard Gaussian scalars. If*

$$n_{\mathrm{grp}} \ge \frac{8}{\beta_c}\log\frac{4ab}{\delta}, \tag{34}$$

then with probability at least $1-\delta$, every target $\mathbf{T} \in \mathbb{R}^{a\times b}$ satisfying $\|\mathbf{T}\|_{\max} \le \lambda\omega c\beta_c n_{\mathrm{grp}}/2$ admits binary masks $(m_\theta)_U, (m_\theta)_V$ such that

$$\|\lambda(\mathbf{U}\odot(m_\theta)_U)(\mathbf{V}\odot(m_\theta)_V) - \mathbf{T}\|_2 \le \lambda\omega c\sqrt{ab}. \tag{35}$$

Here $\|\mathbf{T}\|_{\max}$ is its largest absolute entry. The masks change no retained numerical value.

Proof. Partition the n hidden coordinates into b predetermined groups J_k of size n_{grp}. Set $(m_\theta)_{\mathbf{V},jk} = 1$ for $j \in J_k$ and zero otherwise. For output row i and input column k, the remaining product is a sum of n_{grp} independent terms $\mathrm{U}_{ij}\mathrm{V}_{jk}/\omega$, each distributed as GH. Its positive-band and negative-band counts, for $[c, 2c]$ and $[-2c, -c]$, are each binomial with success probability β_c. Chernoff gives probability at most $e^{-n_{\mathrm{grp}}\beta_c/8}$ that a specified count is below $n_{\mathrm{grp}}\beta_c/2$. A union bound over $2ab$ counts is at most $\delta/2$ under Eq. (34); independence between different matrix entries is unnecessary.

On the resulting event, the relaxed subset sum of just these band-limited terms contains $[-\lambda\omega c\beta_c n_{\mathrm{grp}}/2, \lambda\omega c\beta_c n_{\mathrm{grp}}/2]$. For each desired T_{ik}, scalar nullspace reduction (Lemma E.1 with one row) leaves at most one fractional coefficient without changing that sum. Round this last coefficient to the nearest binary value. Every band coefficient has magnitude at most $2\lambda\omega c$, so the error is at most $\lambda\omega c$. Implement these choices through $(m_\theta)_{\mathbf{U},ij}$ for $j \in J_k$; different target entries use distinct entries of $(m_\theta)_U$, although they may share a retained value of $\mathbf{V}$. All out-of-band terms are zeroed. The Frobenius norm of the entrywise error is at most $\lambda\omega c\sqrt{ab}$ and bounds its spectral norm. The count event is independent of the target, so the result holds for all targets simultaneously. □

Explicit sufficient widths for both attention products. For the normalised product $\lambda = n^{-1/2}$, choose $c = 1$. To approximate every target with maximum entry magnitude at most $R > 0$ within a prescribed spectral error $\epsilon > 0$, it suffices that n be a multiple of b satisfying

$$n \geq \max\left\{\frac{8b}{\beta_1}\log\frac{4ab}{\delta},\quad \frac{4b^2R^2}{\omega^2\beta_1^2},\quad \frac{\omega^2 ab}{\epsilon^2}\right\}. \tag{36}$$

For the unnormalised product $\lambda = 1$, choose $c = \epsilon/(\omega\sqrt{ab})$. It suffices that $n = bn_{\mathrm{grp}}$ with

$$n_{\mathrm{grp}} \geq \max\left\{\frac{8}{\beta_c}\log\frac{4ab}{\delta},\quad \frac{2R}{\omega c\beta_c}\right\}. \tag{37}$$

Both follow by rearranging the conditions on the counts, target range and error; $\beta_c > 0$ for every $c > 0$, so these widths are finite. Apply the first to $\mathbf{QK}^\top/\sqrt{n_K}$ and the second to $\mathbf{VO}$, with per-product failure probabilities $\delta/(2H)$. A union bound supplies one set of all projection masks with probability at least $1-\delta$. Combine with Eq. (18) to realise every target margin with sufficient buffer. These widths therefore suffice to approximate the target attention module with one common set of masks and the prescribed error tolerance.

A direct refill consequence for a fixed-input attention module. Take the target in Proposition 1 to be the zero-pruned attention module, with pruned projections $\mathbf{Q}_h^{\mathrm{pruned}}, \mathbf{K}_h^{\mathrm{pruned}}, \mathbf{V}_h^{\mathrm{pruned}}, \mathbf{O}_h^{\mathrm{pruned}}$. Let $\Delta\mathbf{R}_h$ be its refill perturbation, for $R \in \{Q, K, V, O\}$. Expanding the products gives the sufficient errors

$$\begin{aligned}\epsilon_{A,h} &= \frac{\|\mathbf{Q}_h^{\mathrm{pruned}}\|_2\|\Delta\mathbf{K}_h\|_2 + \|\Delta\mathbf{Q}_h\|_2\|\mathbf{K}_h^{\mathrm{pruned}}\|_2 + \|\Delta\mathbf{Q}_h\|_2\|\Delta\mathbf{K}_h\|_2}{\sqrt{n_K}},\\ \epsilon_{P,h} &= \|\mathbf{V}_h^{\mathrm{pruned}}\|_2\|\Delta\mathbf{O}_h\|_2 + \|\Delta\mathbf{V}_h\|_2\|\mathbf{O}_h^{\mathrm{pruned}}\|_2 + \|\Delta\mathbf{V}_h\|_2\|\Delta\mathbf{O}_h\|_2.\end{aligned} \tag{38}$$

With $\mathbf{P}_h^{\mathrm{pruned}} = \mathbf{V}_h^{\mathrm{pruned}}\mathbf{O}_h^{\mathrm{pruned}}$, substitution in Eq. (18) bounds the change of each margin for arbitrary, possibly dependent fill values. This is a perturbation bound for the module with unchanged inputs and readout; it includes the cross-products of simultaneous query/key and value/output changes. The feedforward bound in Corollary 6 and this attention bound give conditions for margin stability in each module. Applying them through multiple blocks requires propagation of the corresponding intermediate-input bounds.

Attention assumptions and interpretation. The Gaussian construction uses centred independent factors with fixed variances, while retaining the query–key normalisation $n_K^{-1/2}$. The deterministic transfer applies more broadly to any factors satisfying the product-error bounds, with bounded module inputs and a fixed linear pairwise readout. The sets $\mathcal{A}_t$ specify allowed attention positions; the circuit mask m_E selects residual-graph edges. Thus scalar projection masks establish an attention approximation, and the subsequent edge analysis studies routing under its own margin conditions. The product construction and the Gaussian-drift model give complementary explanations of mask-based margin synthesis.

E.2 Shared-mask synthesis with overlapping task readouts

Proof of Theorem 2. Put $\mathrm{Y}_j = q_j\mathrm{X}_j$. Since $\boldsymbol{q}$ is fixed before the Gaussian draw, the Y_j are independent standard Gaussians and $(m_\theta)_j\mathrm{X}_j = q_j[\mathrm{Y}_j]_+$. Thus

$$\widetilde{r}_i((m_\theta, \mathbf{1}_E)) - b_i = \frac{\sigma}{\sqrt{n}}\sum_j h_{ij}q_j[\mathrm{Y}_j]_+.$$

The moments $\mathbb{E}([\mathrm{Y}]_+) = 1/\sqrt{2\pi}$ and $\mathrm{Var}([\mathrm{Y}]_+) = 1/2 - 1/(2\pi)$ give Eq. (20). The full covariance matrix is $\sigma^2(1/2 - 1/(2\pi))\mathbf{HH}^\top/n$, retaining the dependence between task margins. As a function of the random vector $\mathbf{Y} = (\mathrm{Y}_j)_{j=1}^n$, margin i is σv_i-Lipschitz, because the positive-part function is coordinatewise 1-Lipschitz and Cauchy–Schwarz bounds the weighted sum. Gaussian concentration (Boucheron et al., 2013, Section 5.4) therefore gives

$$\mathbb{P}\{\widetilde{r}_i < \mathbb{E}\widetilde{r}_i - t\} \leq \exp\{-t^2/(2\sigma^2 v_i^2)\}, \qquad t > 0,$$

when $v_i > 0$; a row with $v_i = 0$ is constant. Use $t = \sigma v_i \sqrt{2\log(K/\delta)}$, take a union bound over the K comparisons, and subtract the uniform approximation-error bound κ. All comparisons hold under one common mask. Finally, $(m_\theta)_j = \mathbf{1}\{\mathrm{Y}_j > 0\}$ are independent fair Bernoulli variables. Hoeffding also gives $\mathbb{P}\{|\|m_\theta\|_0/n - 1/2| > t\} \le 2e^{-2nt^2}$. Rearranging the lower bound under $d_i \ge d_*$ and $v_i \le v_*$ proves Eq. (22). □

The directional assumption. The signs $\boldsymbol{q}$ can depend on the task readouts, but must be fixed before the Gaussian draw. It suffices that every row of $\mathbf{H}\,\mathrm{diag}(\boldsymbol{q})$ have positive average and bounded mean square; supports may overlap fully across examples. This assumption excludes incompatible comparisons. For example, if $b_1 = b_2 = 0$ and $h_{2j} = -h_{1j}$ for every j, then $\widetilde{r}_2 = -\widetilde{r}_1$ under every mask, so both margins cannot be positive. The $\sqrt{n}$ asymptotic conclusion also requires the directional constants and surrogate errors to remain uniformly controlled as n grows.

Relation to pruning-only approximation. Theorem 2 constructs a one-sided task margin. The reachability result below approximates prescribed coefficients through subset sums of frozen random weights, as in strong-lottery-ticket analyses (Malach et al., 2020; Pensia et al., 2020). Our targets are finite task-margin vectors, with the readout geometry, normalisation and uniform approximation error made explicit.

E.3 Weight and edge signatures with controlled error

Conventions. The generic signature matrix $\mathbf{C} = [\boldsymbol{c}_1 \cdots \boldsymbol{c}_M] \in \mathbb{R}^{K\times M}$ has M candidate columns, and $\boldsymbol{x} \succeq \boldsymbol{y}$ means $x_i \ge y_i$ for every i. Masks are binary unless explicitly relaxed to $[0,1]$. For a support $S \subseteq E$, $(m_E^S)_e = \mathbf{1}\{e \in S\}$.

Assumption 1 (stagewise additive signatures)**.** Let $\boldsymbol{b}_\theta \in \mathbb{R}^K$ be a margin offset, $\mathbf{A} \in \mathbb{R}^{K\times|\mathcal{I}_\theta|}$ the matrix of scalar-coordinate signatures, and $\kappa_\theta \ge 0$ a uniform approximation-error bound. On the specified relaxed and binary masks, the full-edge margins satisfy

$$\boldsymbol{r}((m_\theta, \mathbf{1}_E)) = \boldsymbol{b}_\theta + \mathbf{A}m_\theta + \boldsymbol{e}_\theta(m_\theta), \qquad \|\boldsymbol{e}_\theta(m_\theta)\|_\infty \le \kappa_\theta. \tag{39}$$

After a particular binary m_θ is fixed, let $\boldsymbol{b}_E(m_\theta) \in \mathbb{R}^K$ be the edge-stage offset, $\mathbf{G}(m_\theta) = [\boldsymbol{g}_1(m_\theta)\cdots\boldsymbol{g}_{|E|}(m_\theta)] \in \mathbb{R}^{K\times|E|}$ the edge-signature matrix and $\kappa_E(m_\theta) \ge 0$ its uniform approximation-error bound, with

$$\boldsymbol{r}((m_\theta, m_E)) = \boldsymbol{b}_E(m_\theta) + \mathbf{G}(m_\theta)m_E + \boldsymbol{e}_{E,m_\theta}(m_E), \qquad \|\boldsymbol{e}_{E,m_\theta}(m_E)\|_\infty \le \kappa_E(m_\theta). \tag{40}$$

The vectors $\boldsymbol{e}_\theta(m_\theta), \boldsymbol{e}_{E,m_\theta}(m_E) \in \mathbb{R}^K$ collect approximation errors across task comparisons. Each remainder bound is uniform over the relaxed witnesses and binary candidates used in the result; an edge-only impossibility result requires the bound to hold for every mask within the budget.

Column $\boldsymbol{a}_j$ of $\mathbf{A}$ is a scalar coordinate's *task-margin signature*, and $\boldsymbol{g}_e(m_\theta)$ is a routed edge's signature after scalar selection. Each signature is a vector of effects on the K task margins (Appendix E.4 maps them to parameter and gate derivatives). The edge representation adapts the stacked-readout abstraction (Chen et al., 2026) to last-token margins. The nested model (23) gives $\boldsymbol{b}_\theta = \boldsymbol{o} + \sum_e \bar{\boldsymbol{a}}_e$, $\boldsymbol{a}_j = \sum_e \boldsymbol{a}_{ej}$ and $\boldsymbol{g}_e(m_\theta) = \bar{\boldsymbol{a}}_e + \sum_j (m_\theta)_j \boldsymbol{a}_{ej}$, providing one concrete instance of the stagewise representation.

E.4 From components and parameters to task signatures

Exact zero-ablation computation. Choose an acyclic, explicitly unrolled computation graph. For an input example $\boldsymbol{x}$, write u_v and h_v for the input and output of component v, F_v for its computation, and $\mathcal{I}_{\theta,v}$ for its maskable parameter indices. A route $e = (u, v)$ carries the source output h_u through a fixed map T_e; u_v^{fixed} collects fixed input terms. Each input port receives a sum of the schematic form

$$\begin{aligned} u_v(m_\theta, m_E; \boldsymbol{x}) &= u_v^{\mathrm{fixed}}(\boldsymbol{x}) + \sum_{e=(u,v)} (m_E)_e T_e h_u(m_\theta, m_E; \boldsymbol{x}), \\ h_v(m_\theta, m_E; \boldsymbol{x}) &= F_v\big(u_v(m_\theta, m_E; \boldsymbol{x});\ (m_\theta)_{\mathcal{I}_{\theta,v}} \odot \theta^{\mathrm{reinit}}_{\mathcal{I}_{\theta,v}}\big), \end{aligned} \tag{41}$$

one such sum per port. Fixed routes are nonmaskable edges; the graph expresses the implemented placement of normalisation, projections, attention and residual additions rather than moving gates

across nonlinear operations; shared scalar parameters use the same mask coordinate at every occurrence. Setting $(m_E)_e = 0$ deletes exactly that routed term and recomputes everything downstream. Equation (41) gives the exact computation under the declared zero-ablation intervention, including its nonlinear operations.

Weight signatures. At a relaxed full-edge reference $\bar{m}_\theta$,

$$A_{ij} = \frac{\partial r_i((m_\theta, \mathbf{1}_E))}{\partial (m_\theta)_j}\Big|_{\bar{m}_\theta} = \theta_j^{\mathrm{reinit}} \frac{\partial r_i}{\partial \theta_j}\Big|_{\theta = \bar{m}_\theta \odot \theta^{\mathrm{reinit}}}, \qquad \boldsymbol{b}_\theta = \boldsymbol{r}((\bar{m}_\theta, \mathbf{1}_E)) - \mathbf{A}\bar{m}_\theta, \tag{42}$$

the derivative summing all uses of a shared parameter. A scalar coordinate is mapped to a K-vector by the task readout, and the coefficients multiplying $\theta_j^{\mathrm{reinit}}$ may depend on the realised weights. Assumption 2 isolates a tractable regime in which these readout directions are fixed before the coefficient draw.

Edge signatures. Freeze m_θ and choose an edge reference $\bar{m}_E$. Then

$$\begin{aligned} G_{ie}(m_\theta) &= \frac{\partial r_i((m_\theta, m_E))}{\partial (m_E)_e}\Big|_{\bar{m}_E} \\ &= \Big\langle \frac{\partial r_i}{\partial u_v},\, T_e h_u \Big\rangle_{(m_\theta, \bar{m}_E)}, \qquad e = (u, v), \\ \boldsymbol{b}_E(m_\theta) &= \boldsymbol{r}((m_\theta, \bar{m}_E)) - \mathbf{G}(m_\theta)\bar{m}_E. \end{aligned} \tag{43}$$

These expressions include all tensor indices and downstream paths. Stacking these effects across examples gives the margin signatures; the intercept centres the approximation at the reference.

Bounding the remainder. Taylor's formula gives, for instance, $|e_{\theta,i}(m_\theta)| \le \frac{1}{2} H_{\theta,i} \|m_\theta - \bar{m}_\theta\|_2^2$ when the mask-space Hessian is bounded by a nonnegative curvature bound $H_{\theta,i}$ in spectral norm on the segment from $\bar{m}_\theta$ to m_θ; a uniform κ_θ needs that control on every relevant segment. Binary rounding can move far in mask space even when $\mathbf{A}(m_\theta - \boldsymbol{\alpha})$ is small. Alternatively, Assumption 1 can be imposed directly on a candidate family containing the relaxed witnesses and binary masks used in the theorem. The nested model (23) makes an additional approximation to (41): although residual addition is exact, fixed signatures need not be, and products of gates along paths remain even in a linear network with frozen activations.

E.5 Binary commitment and rounding

For a relaxed coefficient vector $\boldsymbol{\alpha} \in [0,1]^M$, let $F(\boldsymbol{\alpha}) = \{j : 0 < \alpha_j < 1\}$ be its fractional support. For $\mathbf{C} \in \mathbb{R}^{K \times M}$ let $B_2 = \max_j \|\boldsymbol{c}_j\|_2$, $B_\infty = \max_{i,j} |(\mathbf{C})_{ij}|$ and $S(\mathbf{C}) = \max_i \max_{|J| \le \operatorname{rank} \mathbf{C}} \sum_{j \in J} (\mathbf{C})_{ij}^2$. These are the largest column norm, largest entry magnitude and largest row energy over at most $\operatorname{rank} \mathbf{C}$ columns, respectively. Here $\operatorname{rank} \mathbf{C}$ is the matrix rank and $J \subseteq \{1, \dots, M\}$. Define the unrestricted and exact-budget rounding radii by

$$\varepsilon_\star(\mathbf{C}) = \min\Big\{ \tfrac{1}{2} B_2 \sqrt{\operatorname{rank} \mathbf{C}},\ \sqrt{\tfrac{1}{2} S(\mathbf{C}) \log(4K)} \Big\}, \qquad \varepsilon^{\mathrm{bud}}(\mathbf{C}) = \tfrac{1}{2} B_\infty \min\{\operatorname{rank} \mathbf{C} + 1, M\}. \tag{44}$$

Lemma 9 (rounding, with a separate budgeted bound)**.** *Every $\boldsymbol{\alpha} \in [0,1]^M$ admits $\boldsymbol{v} \in \{0,1\}^M$ with $\|\mathbf{C}(\boldsymbol{v} - \boldsymbol{\alpha})\|_\infty \le \varepsilon_\star(\mathbf{C})$. If $\sum_j \alpha_j = k_{\mathrm{mask}} \in \mathbb{Z}$, there is a possibly different $\boldsymbol{v}$ with $\sum_j v_j = k_{\mathrm{mask}}$ and error at most $\varepsilon^{\mathrm{bud}}(\mathbf{C})$.*

The lemma supplies a binary witness through fractional reduction and rounding; the second bound additionally preserves the prescribed integer retention count k_{mask}.

Fractional reduction.

Lemma E.1 (fractional reduction)**.** *For $\mathbf{C} \in \mathbb{R}^{K \times M}$ and $\boldsymbol{\alpha} \in [0,1]^M$ there exists $\widetilde{\boldsymbol{\alpha}} \in [0,1]^M$ with $\mathbf{C}\widetilde{\boldsymbol{\alpha}} = \mathbf{C}\boldsymbol{\alpha}$ and at most $\operatorname{rank} \mathbf{C}$ fractional coordinates. Appending a row of ones additionally preserves $\sum_j \alpha_j$ and leaves at most $\min\{\operatorname{rank} \mathbf{C} + 1, M\}$ fractional coordinates.*

Proof. If $|F(\boldsymbol{\alpha})| > \operatorname{rank}\mathbf{C}$, choose a nonzero $\boldsymbol{u}$ supported on $F(\boldsymbol{\alpha})$ with $\mathbf{C}\boldsymbol{u} = 0$. The set of t with $\boldsymbol{\alpha} + t\boldsymbol{u} \in [0,1]^M$ is a bounded closed interval with 0 in its interior; moving to an endpoint preserves $\mathbf{C}\boldsymbol{\alpha}$, leaves integral coordinates unchanged and makes at least one fractional coordinate integral. Iterate. For $[\mathbf{C}; \mathbf{1}^\top]$ the same operation also preserves the coordinate sum, and the rank is at most $\min\{\operatorname{rank}\mathbf{C} + 1, M\}$. □

Lemma 9 (restated). *Every $\boldsymbol{\alpha} \in [0,1]^M$ admits $\boldsymbol{v} \in \{0,1\}^M$ with $\|\mathbf{C}(\boldsymbol{v} - \boldsymbol{\alpha})\|_\infty \le \varepsilon_\star(\mathbf{C})$. If $\sum_j \alpha_j = k_{\mathrm{mask}} \in \mathbb{Z}$, there is a possibly different $\boldsymbol{v}$ with $\sum_j v_j = k_{\mathrm{mask}}$ and error at most $\varepsilon^{\mathrm{bud}}(\mathbf{C})$.*

Proof. Apply Lemma E.1; the resulting fractional set satisfies $|F(\widetilde{\boldsymbol{\alpha}})| \le \operatorname{rank}\mathbf{C}$. Let $\mathrm{v}_j \sim \mathrm{Bernoulli}(\widetilde{\alpha}_j)$ independently for $j \in F(\widetilde{\boldsymbol{\alpha}})$ and fix the integral coordinates. Then $\mathbb{E}\|\mathbf{C}(\mathbf{v} - \widetilde{\boldsymbol{\alpha}})\|_2^2 = \sum_{j \in F(\widetilde{\boldsymbol{\alpha}})} \widetilde{\alpha}_j(1 - \widetilde{\alpha}_j)\|\boldsymbol{c}_j\|_2^2 \le \frac{1}{4}\operatorname{rank}\mathbf{C}B_2^2$, so some realisation has Euclidean, hence coordinate-wise, error at most $B_2\sqrt{\operatorname{rank}\mathbf{C}}/2$. For the second radius each row is a sum of independent centred variables with ranges $|(\mathbf{C})_{ij}|$, and Hoeffding gives $\mathbb{P}\{|[\mathbf{C}(\mathbf{v} - \widetilde{\boldsymbol{\alpha}})]_i| \ge t\} \le 2\exp(-2t^2/S(\mathbf{C}))$; at $t = \sqrt{S(\mathbf{C})\log(4K)/2}$ the union bound over rows is at most $1/2$, so a realisation satisfies all rows (a zero-energy row is identically zero). Each radius has its own witness, so the smaller is attainable. For the exact budget reduce with $[\mathbf{C}; \mathbf{1}^\top]$; let w be the sum of the remaining fractional entries, an integer. If $|F(\widetilde{\boldsymbol{\alpha}})| = 0$ no rounding is needed; otherwise round the w largest fractions to one and the rest to zero. Their sum U satisfies $U \ge w^2/|F(\widetilde{\boldsymbol{\alpha}})|$, so $\sum_j |v_j - \widetilde{\alpha}_j| = 2(w - U) \le 2w(1 - w/|F(\widetilde{\boldsymbol{\alpha}})|) \le |F(\widetilde{\boldsymbol{\alpha}})|/2$, and the row error is at most $B_\infty|F(\widetilde{\boldsymbol{\alpha}})|/2 \le \varepsilon^{\mathrm{bud}}(\mathbf{C})$ with the sum preserved. $\mathbf{C}\widetilde{\boldsymbol{\alpha}} = \mathbf{C}\boldsymbol{\alpha}$ transfers every bound to $\boldsymbol{\alpha}$. □

E.6 Sequential synthesis

Theorem 10 (weight-stage synthesis)**.** *Under (39), let $\widetilde{m}_\theta \in [0,1]^{|\mathcal{I}_\theta|}$. If $\boldsymbol{b}_\theta + \mathbf{A}\widetilde{m}_\theta \succeq \boldsymbol{\tau} + (\varepsilon_\star(\mathbf{A}) + \kappa_\theta)\mathbf{1}$ then some binary m_θ satisfies $\boldsymbol{r}((m_\theta, \mathbf{1}_E)) \succeq \boldsymbol{\tau}$; in terms of the exact relaxed network the buffer is $\varepsilon_\star(\mathbf{A}) + 2\kappa_\theta$. If $\sum_j(\widetilde{m}_\theta)_j$ is an integer, replacing $\varepsilon_\star(\mathbf{A})$ by $\varepsilon^{\mathrm{bud}}(\mathbf{A})$ gives a mask with exactly that many retained weights.*

Theorem 11 (edge-stage synthesis on the frozen pruned weights)**.** *Fix the binary m_θ entering the edge stage and $\widetilde{m}_E \in [0,1]^{|E|}$. If $\boldsymbol{b}_E(m_\theta) + \mathbf{G}(m_\theta)\widetilde{m}_E \succeq \boldsymbol{\tau} + (\varepsilon_\star(\mathbf{G}(m_\theta)) + \kappa_E(m_\theta))\mathbf{1}$ then some binary m_E satisfies $\boldsymbol{r}((m_\theta, m_E)) \succeq \boldsymbol{\tau}$; if $\sum_e(\widetilde{m}_E)_e = s_E \in \mathbb{Z}$ the conclusion holds with exactly s_E edges after replacing $\varepsilon_\star$ by $\varepsilon^{\mathrm{bud}}$. A premise stated using the exact relaxed margin requires the corresponding rounding radius plus $2\kappa_E(m_\theta)$.*

The edge-stage premise must hold on the *same* binary m_θ produced by the weight stage. It is enough for both stages to reach the required margin region; they need not produce identical margin vectors.

Proof of Theorem 10. Take the binary m_θ of Lemma 9 for $\mathbf{C} = \mathbf{A}$. Then $r_i((m_\theta, \mathbf{1}_E)) \ge b_{\theta,i} + (\mathbf{A}m_\theta)_i - \kappa_\theta \ge b_{\theta,i} + (\mathbf{A}\widetilde{m}_\theta)_i - \varepsilon_\star(\mathbf{A}) - \kappa_\theta \ge \tau_i$. If the premise uses exact relaxed margins, subtract κ_θ to enter the surrogate and again to return to the exact binary model, which is the buffer $\varepsilon_\star(\mathbf{A}) + 2\kappa_\theta$; the budgeted version substitutes the budgeted rounding bound. □

Proof of Theorem 11. Condition on the frozen m_θ and apply Lemma 9 to $\mathbf{C} = \mathbf{G}(m_\theta)$, $\boldsymbol{\alpha} = \widetilde{m}_E$: $r_i((m_\theta, m_E)) \ge b_{E,i}(m_\theta) + (\mathbf{G}(m_\theta)\widetilde{m}_E)_i - \varepsilon_\star(\mathbf{G}(m_\theta)) - \kappa_E(m_\theta) \ge \tau_i$. The appended-sum reduction enforces exactly the prescribed edge count s_E through $\varepsilon^{\mathrm{bud}}(\mathbf{G}(m_\theta))$. □

The proofs establish the order $\exists m_\theta$ from the weight stage, followed by $[\exists \widetilde{m}_E$ on that same $m_\theta] \Rightarrow \exists m_E$. Thus the sequential guarantee requires a relaxed edge witness after the weight stage.

E.7 Random scalar reachability and width

The following random-feature regime supplies a relaxed task-margin witness and an explicit width condition for rounding it to a binary mask.

Assumption 2 (prespecified scalar directions)**.** Let $k \in \{1, \dots, K\}$ count orthonormal task directions $\boldsymbol{v}_\ell \in \mathbb{R}^K$, and let $n \geq 1$ count scalar coordinates per direction. These directions and the resulting kn coordinates are specified before inspecting their coefficients, with full-edge signatures $\boldsymbol{a}_{\ell j} = n^{-1/2}(\mu_\theta + \sigma_\theta \mathrm{X}_{\ell j})\boldsymbol{v}_\ell$, $\mathrm{X}_{\ell j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, where $\mu_\theta \in \mathbb{R}$ and $\sigma_\theta > 0$ are the coefficient mean and standard deviation before the $n^{-1/2}$ normalisation. Here $\mathbf{A}$ contains these columns; all remaining gates are fixed and their contributions are absorbed into $\boldsymbol{b}_\theta$.

For a coefficient-magnitude threshold $c_0 > 0$, let $\mathrm{X} \sim \mathcal{N}(0,1)$ and let p_0 be the smaller of the probabilities that $\mu_\theta + \sigma_\theta \mathrm{X}$ exceeds c_0 or falls below $-c_0$, and set $R_G = \frac{1}{2} c_0 p_0 \sqrt{n}$, $\varepsilon_G = \frac{\sqrt{k}}{2\sqrt{n}}(|\mu_\theta| + \sigma_\theta \sqrt{2\log(4kn/\delta)})$. Here R_G is the reachable radius along each task direction and ε_G is the Euclidean error allowed when converting a relaxed witness into a binary mask.

Proposition 12 (random scalar margin reachability)**.** *If $n \geq (8/p_0)\log(4k/\delta)$ then with probability at least $1-\delta$ every $\boldsymbol{w}^\star = \sum_\ell w_\ell^\star \boldsymbol{v}_\ell$ with $|w_\ell^\star| \leq R_G$ admits a binary weight mask with $\|\mathbf{A} m_\theta - \boldsymbol{w}^\star\|_2 \leq \varepsilon_G$; if $\boldsymbol{b}_\theta + \boldsymbol{w}^\star \succeq \boldsymbol{\tau} + (\varepsilon_G + \kappa_\theta)\mathbf{1}$ that mask synthesises the target.*

For a fixed coefficient distribution, the reachable radius grows as $\sqrt{n}$ while relative binary error is $\widetilde{O}(\sqrt{k}/n)$, where $\widetilde{O}$ suppresses logarithmic factors: more maskable contributions enlarge the reachable region and reduce the relative rounding error. Solving both requirements for n gives a width bound of the form used in strong-lottery-ticket results.

Corollary 13 (over-parameterisation suffices)**.** *Under Assumption 2, fix $\delta \in (0,1)$, an available margin buffer $\gamma > 0$ and a target-coefficient bound $W \geq 0$. Suppose $\max_\ell |w_\ell^\star| \leq W$ and $\boldsymbol{b}_\theta + \boldsymbol{w}^\star \succeq \boldsymbol{\tau} + (\gamma + \kappa_\theta)\mathbf{1}$. If*

$$n \;\geq\; \max\left\{ \frac{8}{p_0}\log\frac{4k}{\delta},\;\; \frac{4W^2}{c_0^2 p_0^2},\;\; \frac{k\mu_\theta^2}{\gamma^2} + \frac{2k\sigma_\theta^2}{\gamma^2}\log\frac{8k^2\sigma_\theta^2}{\gamma^2\delta} \right\}, \tag{45}$$

then with probability at least $1-\delta$ some binary weight mask has $\boldsymbol{r}((m_\theta, \mathbf{1}_E)) \in \mathcal{M}_\tau$, with all target margins achieved on the same event.

The three requirements control two-sided coefficient availability, target reach and binary resolution, with task dimension, coefficient distribution and approximation error entering explicitly.

Positive drift from literal zeroing. For the centred special case $k = K$, $\boldsymbol{v}_\ell$ the coordinate basis and $\mu_\theta = 0$, one shared mask sets $(m_\theta)_{ij} = \mathbf{1}\{\mathrm{X}_{ij} > 0\}$. Each scalar is either unchanged or exactly zero; no sign flip or weight update occurs. Its surrogate margins satisfy

$$\widetilde{r}_i - b_{\theta,i} = \frac{\sigma_\theta}{\sqrt{n}} \sum_{j=1}^{n} [\mathrm{X}_{ij}]_+, \qquad \mathbb{E}(\widetilde{r}_i - b_{\theta,i}) = \frac{\sigma_\theta \sqrt{n}}{\sqrt{2\pi}}, \quad \mathrm{Var}(\widetilde{r}_i - b_{\theta,i}) = \sigma_\theta^2 \left(\frac{1}{2} - \frac{1}{2\pi}\right). \tag{46}$$

With probability at least $1-\delta$, that same mask gives, for all i, $r_i \geq b_{\theta,i} + \sigma_\theta\sqrt{n/(2\pi)} - \sigma_\theta\sqrt{2\log(K/\delta)} - \kappa_\theta$. Thus drift grows with available contributions while fluctuations stay bounded under the specified normalisation. In this special case, disjoint task directions provide an explicit shared-mask construction across all comparisons.

Proofs of reachability and width.

Proof of Proposition 12. For direction ℓ let N_ℓ^+ count coefficients $\mu_\theta + \sigma_\theta \mathrm{X}_{\ell j} \geq c_0$ and N_ℓ^- those at most $-c_0$; each is binomial with success probability at least p_0, and the multiplicative Chernoff bound gives $\mathbb{P}\{N_\ell^\pm < np_0/2\} \leq e^{-np_0/8}$, so the union over $2k$ counts is at most $\delta/2$ under the stated condition on n. On the complement, with $c_{\ell j} = (\mu_\theta + \sigma_\theta \mathrm{X}_{\ell j})/\sqrt{n}$, the relaxed scalar sum ranges over $[-\sum_j [-c_{\ell j}]_+,\ \sum_j [c_{\ell j}]_+] \supseteq [-R_G, R_G]$, and independent control in the k orthogonal directions yields the product box. Separately $\mathbb{P}\{\max_{\ell,j} |\mathrm{X}_{\ell j}| > \sqrt{2\log(4kn/\delta)}\} \leq \delta/2$. On the intersection every column has norm at most $(|\mu_\theta| + \sigma_\theta\sqrt{2\log(4kn/\delta)})/\sqrt{n}$, and the matrix $\mathbf{A}$ has rank at most k, so the Euclidean rounding radius is at most ε_G for every point of the box on this one event; orthogonality converts coefficient error to margin error, and the exact statement follows by subtracting $\varepsilon_G + \kappa_\theta$. □

Corollary 13 (restated). *Under Assumption 2, fix $\delta \in (0,1)$ and $\gamma > 0$, and suppose $\max_\ell |w^\star_\ell| \le W$ and $\boldsymbol{b}_\theta + \boldsymbol{w}^\star \succeq \boldsymbol{\tau} + (\gamma + \kappa_\theta)\mathbf{1}$. If n satisfies (45), then with probability at least $1-\delta$ there is a binary weight mask m_θ with $\boldsymbol{r}((m_\theta, \mathbf{1}_E)) \in \mathcal{M}_\tau$.*

Proof. Three requirements hold at once. The first term of (45) is the hypothesis of Proposition 12. The second makes the box contain the target: $R_G \ge W$ reads $n \ge 4W^2/(c_0^2 p_0^2)$. The third makes rounding cost at most γ: $\varepsilon_G \le \gamma$ reads $n \ge k(|\mu_\theta| + \sigma_\theta b_\delta)^2/(4\gamma^2)$ with the Gaussian tail threshold $b_\delta = \sqrt{2\log(4kn/\delta)}$, implied via $(x+y)^2 \le 2x^2 + 2y^2$ by $n \ge C_\mu + C_\sigma \log(4kn/\delta)$, where $C_\mu = k\mu_\theta^2/(2\gamma^2)$ and $C_\sigma = k\sigma_\theta^2/\gamma^2$ collect the mean and variance contributions to rounding cost. The elementary inequality $\log t \le t/u + \log u - 1$ at $t = 4kn/\delta$, $u = 8kC_\sigma/\delta$ gives $C_\sigma \log(4kn/\delta) \le n/2 + C_\sigma \log(8kC_\sigma/\delta)$, so the implicit condition holds as soon as $n \ge 2C_\mu + 2C_\sigma \log(8kC_\sigma/\delta)$, the third term. On the good event Proposition 12 supplies m_θ with $\|\mathbf{A}m_\theta - \boldsymbol{w}^\star\|_2 \le \gamma$, whence $r_i((m_\theta, \mathbf{1}_E)) \ge b_{\theta,i} + w^\star_i - \gamma - \kappa_\theta \ge \tau_i$. □

Proof of the shared-mask drift calculation. In Eq. (46), the mask $(m_\theta)_{ij} = \mathbf{1}\{\mathrm{X}_{ij} > 0\}$ is one binary vector over all Kn coordinates. For $\mathrm{X} \sim \mathcal{N}(0,1)$, $\mathbb{E}([\mathrm{X}]_+) = 1/\sqrt{2\pi}$ and $\mathbb{E}([\mathrm{X}]_+^2) = 1/2$, proving the moments by independence. The map $\mathbf{X}_{i,\cdot} \mapsto \sigma_\theta n^{-1/2} \sum_j [\mathrm{X}_{ij}]_+$ is σ_θ-Lipschitz. Gaussian concentration bounds its lower deviation by $\exp[-t^2/(2\sigma_\theta^2)]$. Set $t = \sigma_\theta \sqrt{2\log(K/\delta)}$, union-bound over the K rows, and subtract the uniform exact-model remainder κ_θ. This mask is shared across all examples.

Remark 1 (relation to strong-lottery-ticket width bounds). Corollary 13 gives a pruning-only approximation guarantee for task-margin targets in $\mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. Its sufficient width conditions express the strong-lottery-ticket mechanism (Malach et al., 2020; Pensia et al., 2020) in finite-task coordinates: enough independently maskable random contributions provide both target reach and binary resolution with high probability.

Remark 2 (clipping and independence). Clipping independent proof coefficients to an interval containing $\pm c_0$ preserves the two-sided counts and cannot increase the maximum absolute coefficient, so the reachability proof retains its bounds. An asymmetric clip can change the mean and invalidate the centred granularity premise. The synthesis results therefore state their assumptions on margin signatures explicitly, including the fixed-direction and independence conditions; Eq. (42) describes how these signatures arise from a network. The completion bounds permit arbitrary task-dependent retained values.

E.8 Edge-only ceilings and cancellation–selection

Let $s_E \in \{0, \ldots, |E|\}$ be the maximum allowed number of active edges. With all scalar weights retained, $m_\theta = \mathbf{1}_\theta$, write $\boldsymbol{g}_e^{\mathrm{full}} = \boldsymbol{g}_e(\mathbf{1}_\theta)$ and $U_i^E(s_E) = b_{E,i}(\mathbf{1}_\theta) + \mathrm{Top}^+_{s_E}(g^{\mathrm{full}}_{1,i}, \ldots, g^{\mathrm{full}}_{|E|,i})$, with $\mathrm{Top}^+_{s_E}$ the sum of the largest at most s_E positive entries.

Theorem 14 (edge-only obstruction)**.** *Every mask m_E with $\|m_E\|_0 \le s_E$ satisfies $r_i((\mathbf{1}_\theta, m_E)) \le U_i^E(s_E) + \kappa_E(\mathbf{1}_\theta)$. Hence $\sup_{\|m_E\|_0 \le s_E} \mathrm{Acc}((\mathbf{1}_\theta, m_E)) \le \overline{A}_E := \frac{1}{K}\sum_i \mathbf{1}\{U_i^E(s_E) + \kappa_E(\mathbf{1}_\theta) > 0\}$, and one upper bound below its τ_i shows that no such support reaches $\mathcal{M}_\tau$.*

At full weights, the best s_E edges determine this ceiling through their positive task contributions. Proposition 3 bounds those contributions in the centred independent-contribution model. The next certificate compares the ceiling with a weight-and-edge witness on the same initialisation and edge budget, corresponding to the pruning-mode comparison in §4.

Proof of Theorem 14. For any m_E with at most s_E active edges, $b_{E,i}(\mathbf{1}_\theta) + \sum_e (m_E)_e g^{\mathrm{full}}_{e,i} \le U_i^E(s_E)$; adding the uniform remainder gives the exact upper bound. A nonpositive bound fails that comparison under every allowed mask, and bounding the remaining indicators by one gives the accuracy ceiling; a bound below τ_i makes the target region infeasible. □

Corollary 15 (same-realisation, same-budget certificate)**.** *Exact-budget rounding of a relaxed edge gate with $\sum_e (\widetilde{m}_E)_e = s_E$ on the frozen m_θ yields a binary weight-and-edge circuit of accuracy at least $\underline{A}_{\theta\to E} = \frac{1}{K}\sum_i \mathbf{1}\{r_i((m_\theta, \widetilde{m}_E)) > \varepsilon^{\mathrm{bud}}(\mathbf{G}(m_\theta)) + 2\kappa_E(m_\theta)\}$; if $\underline{A}_{\theta\to E} > \overline{A}_E$ it strictly outperforms every edge-only support at that budget on the same initialisation and comparisons.*

Proof. Budgeted rounding gives one binary m_E with exactly s_E edges and $r_i((m_\theta, m_E)) \geq r_i((m_\theta, \widetilde{m}_E)) - \varepsilon^{\mathrm{bud}}(\mathbf{G}(m_\theta)) - 2\kappa_E(m_\theta)$ for all i simultaneously, so every positive indicator in $\underline{A}_{\theta \to E}$ remains positive under that one support; combine with Theorem 14 on the same initialisation and comparisons. □

Proof of Proposition 3. The coarse signature is $g_e^{\mathrm{full}} = \sigma_\theta n^{-1/2} \sum_j \mathrm{X}_{ej} \sim \mathcal{N}(0, \sigma_\theta^2)$, and $\mathbb{P}\{\max_e |g_e^{\mathrm{full}}| > \sigma_\theta \sqrt{2\log(4|E|/\delta)}\} \leq \delta/2$; on the complement every budgeted edge sum is at most s_E times that maximum. For the fine selector set $(m_\theta)_{(e,j)} = \mathbf{1}\{e = 1\}\mathbf{1}\{\mathrm{X}_{1j} > 0\}$, keeping positive coefficients in block 1 and zeroing every other coordinate; with $S_n = \sum_j [\mathrm{X}_{1j}]_+$ the margin is $\sigma_\theta S_n/\sqrt{n}$, $\mathbb{E}S_n = n/\sqrt{2\pi}$, and S_n is $\sqrt{n}$-Lipschitz in its Gaussian vector, so $\mathbb{P}\{S_n \leq \frac{1}{2}\mathbb{E}S_n\} \leq e^{-n/(16\pi)} \leq \delta/2$ and $\sigma_\theta S_n/\sqrt{n} \geq \sigma_\theta \sqrt{n}/(2\sqrt{2\pi}) \geq \sigma_\theta \sqrt{n}/6$. Comparing the two bounds gives $n > 72 s_E^2 \log(4|E|/\delta)$; for exact margins the surrogate gap must exceed the two applicable remainders. □

Remark 3 (what drives the gap). If each contribution has mean $\mu_\theta/\sqrt{n} \neq 0$, the unselected edge already has mean $\mu_\theta \sqrt{n}$ and the centred ceiling no longer applies. Matching the pooled pretrained weight mean and centring task-margin coefficients are distinct conditions because the readout multiplies each weight. Scaling centred coefficients changes margin magnitudes; classification additionally depends on their signs and offsets. The proposition gives a one-margin capacity comparison, and the common-direction hypothesis with Eq. (46) gives the simultaneous-margin construction.

E.9 Search-space capacity

Write $H_2(\rho) = -\rho \log_2 \rho - (1-\rho)\log_2(1-\rho)$ for binary entropy in bits. Here ρ is a scalar retention fraction.

Proposition 16 (mask counts). *For fixed* $0 < \rho < 1$, $\log_2 \binom{|\mathcal{I}_\theta|}{\lfloor \rho |\mathcal{I}_\theta| \rfloor} = |\mathcal{I}_\theta| H_2(\rho) + O(\log |\mathcal{I}_\theta|)$, *and for an integer* $1 \leq s_E \leq |E|/2$, $\log_2 \sum_{j \leq s_E} \binom{|E|}{j} \leq s_E \log_2(e|E|/s_E)$.

Proof. Stirling's formula gives the first assertion. For the second put $t = s_E/|E| \leq 1/2$; expanding $(t + (1-t))^{|E|} = 1$ and lower-bounding each weight $t^j(1-t)^{|E|-j}$, $j \leq s_E$, by $t^{s_E}(1-t)^{|E|-s_E}$ yields $\sum_{j \leq s_E} \binom{|E|}{j} \leq 2^{|E| H_2(t)}$, and $-(1-t)\log(1-t) \leq t$ gives $|E| H_2(t) \leq s_E \log_2(e|E|/s_E)$. □

At about half retention, a scalar mask has order $|\mathcal{I}_\theta|$ bits available for selection. This counts possible masks. The margin constructions establish what some of those masks can compute.

E.10 Deterministic stability from the pruned baseline

Proof of Theorem 5. The line segment $t\boldsymbol{z}$, $0 \leq t \leq 1$, lies in the cube. The fundamental theorem of calculus and duality of the ℓ_1 and ℓ_∞ norms give

$$|F_i(\boldsymbol{z}) - F_i(\mathbf{0})| = \left| \int_0^1 \langle \nabla F_i(t\boldsymbol{z}), \boldsymbol{z} \rangle \, dt \right| \leq L_i(u) \|\boldsymbol{z}\|_\infty \leq u L_i(u).$$

Every initially positive margin exceeding $u L_i(u)$ remains positive. Only the initially successful comparisons counted in Eq. (29) can be lost; new successes can only lower $\mathrm{Acc}_{\mathrm{fill}}(\mathbf{0}) - \mathrm{Acc}_{\mathrm{fill}}(\boldsymbol{z})$. The target-margin claim is immediate. □

This certificate holds for every fill assignment within the declared magnitude cap. The sensitivity $L_i(u)$ bounds the gradient throughout the cube traversed by the fill.

Let $\mathcal{Q} \subseteq \mathcal{P}$ be the structural dependency closure of the evaluated logits under m_E, formed with every hole treated as live; coordinates outside $\mathcal{Q}$ cannot affect the evaluated logits. The set $\mathcal{Q}$ is used only to analyse sensitivity; every coordinate in $\mathcal{P}$ remains eligible for refill.

Assumption 3 (local smoothness of exact completed margins)**.** Let $T > 0$ be a perturbation radius. Each exact completed margin $F_i(\boldsymbol{z}_{\mathcal{Q}})$ is twice continuously differentiable on a neighbourhood of the closed Euclidean ball of radius T about zero. Write $\boldsymbol{g}_i = \nabla F_i(\mathbf{0})$ for its vector of first-order refill sensitivities, $H_i \geq 0$ for a uniform Hessian spectral-norm bound $\|\nabla^2 F_i\|_{\mathrm{op}} \leq H_i$ on that neighbourhood, and $r_i^{\mathrm{pruned}} = F_i(\mathbf{0}) = r_i((m_\theta, m_E))$ for its pruned baseline margin.

The smoothness assumption applies to the exact completed network without requiring additivity. For the next probabilistic bound, we fix the fill multiset and vary only its uniformly random assignment. The histogram then stays fixed while the task margins can change.

Theorem 17 (task-agnostic random assignment of a fitted fill multiset)**.** *Condition on the discovered masks and on a fresh multiset $a_1, \ldots, a_{|\mathcal{P}|}$, $|\mathcal{P}| \geq 2$, assigned to the holes uniformly at random. Write $\bar{a} = |\mathcal{P}|^{-1} \sum_j a_j$, $v^2 = |\mathcal{P}|^{-1} \sum_j (a_j - \bar{a})^2$ and $M = \max_j |a_j|$ for the multiset mean, variance and maximum magnitude, respectively. Write $\mathbf{z}$ for the randomly assigned fill vector and extend $\boldsymbol{g}_i$ by zero outside $\mathcal{Q}$. If $\sqrt{|\mathcal{Q}|}M \leq T$ then with probability at least $1 - \delta$, simultaneously for all i,*

$$F_i(\mathbf{z}) \geq r_i^{\mathrm{pruned}} - \Delta_i, \quad \Delta_i = |\bar{a}\, \mathbf{1}^\top \boldsymbol{g}_i| + \sqrt{\tfrac{KV_i}{\delta}} + \tfrac{1}{2} H_i |\mathcal{Q}| M^2, \quad V_i = \tfrac{v^2}{|\mathcal{P}|-1} \big(|\mathcal{P}| \|\boldsymbol{g}_i\|_2^2 - (\mathbf{1}^\top \boldsymbol{g}_i)^2 \big). \tag{47}$$

Here V_i is the variance of the linear margin change and Δ_i combines its mean, random fluctuation and nonlinear remainder. The multiset's histogram is invariant under assignment, so a deficit-allocated multiset obeys the target-TV bound $E_\eta + N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$ and (47) at once, and $\mathrm{Acc}_{\mathrm{fill}}(\mathbf{z}) \geq K^{-1} \sum_i \mathbf{1}\{r_i^{\mathrm{pruned}} > \Delta_i\}$ on that event.

For independent fills $\mathrm{z}_j \sim \mathcal{N}(0, \nu^2)$, we obtain the bound

$$F_i(\mathbf{z}) \geq r_i^{\mathrm{pruned}} - \nu \|\boldsymbol{g}_i\|_2 \sqrt{2 \log(4K/\delta)} - \tfrac{1}{2} H_i \nu^2 \big(\sqrt{|\mathcal{Q}|} + \sqrt{2 \log(4/\delta)} \big)^2 \tag{48}$$

with probability at least $1 - \delta$, provided $\nu(\sqrt{|\mathcal{Q}|} + \sqrt{2 \log(4/\delta)}) \leq T$ (Corollary 18). The linear effect scales with sensitivity $\nu \|\boldsymbol{g}_i\|_2$, while curvature costs order $H_i \nu^2 |\mathcal{Q}|$. Small entries need not yield a small Euclidean perturbation when many coordinates are refilled. The bound quantifies the neighbourhood in which margins are preserved through sensitivity, curvature and the number of active pruned coordinates; the histogram results quantify the accompanying distributional fit.

E.11 Exact inactivity and local completion stability

We fix m_E and treat every pruned coordinate as live when constructing $\mathcal{Q}$. A path with zero value and zero gradient can otherwise be missed even though two or more fills could reactivate it. The closure includes weight sharing, normalisation, biases, fixed skip connections and every use of a parameter.

Lemma E.2 (structural inactivity and Taylor control)**.** *Under the declared dependency graph the completed margins depend only on $\boldsymbol{z}_{\mathcal{Q}}$, and under Assumption 3, for $\|\boldsymbol{z}_{\mathcal{Q}}\|_2 \leq T$, $|F_i(\boldsymbol{z}_{\mathcal{Q}}) - r_i^{\mathrm{pruned}} - \langle \boldsymbol{g}_i, \boldsymbol{z}_{\mathcal{Q}} \rangle| \leq \frac{1}{2} H_i \|\boldsymbol{z}_{\mathcal{Q}}\|_2^2$.*

Proof. A parameter outside the structural ancestor set of the evaluated logits appears in no active directed dependency leading to them; induction through the acyclic graph, including every use of shared parameters and fixed paths, shows that changing it leaves the logits unchanged for arbitrary values of the other holes. For the second claim apply Taylor's formula with integral remainder on the segment $t\boldsymbol{z}_{\mathcal{Q}}$, $0 \leq t \leq 1$, and bound the quadratic form by the operator norm. □

A closure based on every output coordinate and allowed position gives a conservative set of dependencies that can be computed without token-pair preference labels. Its size $|\mathcal{Q}|$ depends on parameter sharing and routing within the active support; the sensitivity $\|\boldsymbol{g}_i\|_2$ further distinguishes strong from weak dependencies.

E.12 Deficit allocation: the exact histogram calculation

Use the bins and normalised masses from §D.4. For a fixed target η, abbreviate $p_b = p_b(\eta)$. The total positive deficit satisfies $D_\eta = \sum_b [p_b - \rho_b]_+ = \alpha + E_\eta$.

Proof of Proposition 4. For feasible fill masses $x_b \geq 0$ summing to α put $f_b = \rho_b + x_b - p_b$; as $\sum_b f_b = 0$, $\frac{1}{2}\sum_b |f_b| = \sum_b [f_b]_+ \geq \sum_b [\rho_b - p_b]_+ = E_\eta$. Conversely allocate $x_b = \alpha[p_b - \rho_b]_+ / D_\eta$, whose multiplier α/D_η is at most one, so no deficit bin becomes an excess bin and the excess bins are unchanged: the total variation is exactly E_η. For actual values the desired counts $|\mathcal{I}_\theta| x_b$ sum to the integer $|\mathcal{P}|$; take floors and distribute the missing count to the largest remainders, so each count moves by less than one and the histogram by at most $N_{\mathrm{bin}}/|\mathcal{I}_\theta|$ in ℓ_1, hence $N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$ in total variation, realised by fresh values in the bin interiors (an atomless within-bin law almost surely reproduces no removed value). Taking the infimum over targets proves the two-sided bound; without attainment choose a target within ζ of the infimum and let $\zeta \downarrow 0$. □

For its chosen Gaussian target, the allocator attains the optimal unrestricted fill histogram up to integer-rounding error $N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$. Thus the target fitted from the reference-completed histogram has discrepancy at most $E_\eta + N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$. The separate optimum over the Gaussian family is R_{Ghist}.

E.13 Rank assignment is a task-agnostic permutation

Lemma E.3 (exchangeable ranks)**.** *Let $\xi \in \mathbb{R}^{|\mathcal{P}|}$ be exchangeable with pairwise distinct coordinates almost surely. Conditional on its order statistics, its rank permutation is uniform over the $|\mathcal{P}|!$ coordinate permutations and independent of those order statistics. If a fill multiset is constructed from permutation-invariant functions of ξ and independent randomness, assigning that multiset in ξ's rank order is conditionally a uniform permutation of the multiset.*

Proof. Exchangeability maps each strict ordering event to every other by a coordinate permutation without changing any order statistic, so their conditional probabilities are equal and sum to one; independent auxiliary randomness and functions of the order statistics cannot distinguish the permutations, and conditioning on the generated multiset preserves uniformity. □

Centring an exchangeable reference and multiplying it by a positive permutation-invariant scale preserves its ranks. If the multiset uses only symmetric reference statistics and independent auxiliary randomness, the lemma applies. Conditional uniformity therefore follows from exchangeability, permutation-invariant multiset construction and independent auxiliary draws. Finite-precision ties can be handled by a symmetric tie-breaking rule that preserves these conditions. The moment and success bounds below use this assignment model.

Exact additive moments and simultaneous success. Conditionally uniform assignment has $\mathbb{E}\mathrm{z}_j = \bar{a}$, $\mathrm{Var}(\mathrm{z}_j) = v^2$ and $\mathrm{Cov}(\mathrm{z}_j, \mathrm{z}_k) = -v^2/(|\mathcal{P}| - 1)$ for $j \neq k$. Expanding the mean and variance of $r_i^{\mathrm{pruned}} + \boldsymbol{g}_i^\top \mathbf{z}$ gives Eq. (31). If $\mu_i > 0$ and $V_i > 0$, Cantelli's inequality gives $\mathbb{P}(F_i \leq 0) \leq V_i/(V_i + \mu_i^2)$; for $V_i = 0$, F_i is the positive constant μ_i. A union bound proves Eq. (32). No independence between comparisons is used. If $|\mathcal{P}| = 1$ there is no random assignment; its single fill has a deterministic effect. For a desired post-refill margin $\tau_i^{\mathrm{comp}} > 0$, the same bound uses $\mu_i - \tau_i^{\mathrm{comp}} > 0$ in the denominator.

Proof of Theorem 17. Condition on the multiset. Every assigned coordinate has mean $\bar{a}$ and variance v^2, and sampling without replacement gives $\mathrm{Cov}(\mathrm{z}_j, \mathrm{z}_k) = -v^2/(|\mathcal{P}| - 1)$ for $j \neq k$. For the zero-extended $\boldsymbol{g}_i$, $\mathbb{E}(\boldsymbol{g}_i^\top \mathbf{z}) = \bar{a}\mathbf{1}^\top \boldsymbol{g}_i$ and $\mathrm{Var}(\boldsymbol{g}_i^\top \mathbf{z}) = v^2 \sum_j g_{ij}^2 - \frac{v^2}{|\mathcal{P}|-1} \sum_{j \neq k} g_{ij} g_{ik} = V_i$; Cauchy–Schwarz gives $V_i \geq 0$. If $V_i > 0$ Chebyshev bounds a deviation beyond $\sqrt{KV_i/\delta}$ by δ/K, and if $V_i = 0$ the statistic is deterministic. Take a union bound over the K comparisons. Deterministically $\|\mathbf{z}_{\mathcal{Q}}\|_2 \leq \sqrt{|\mathcal{Q}|}M \leq T$, so Lemma E.2 supplies the curvature term. The empirical histogram is invariant under permutation, so the fixed-target fit bound holds on every assignment. □

For $|\mathcal{P}| = 1$ the assignment is deterministic and $V_i = 0$; for $|\mathcal{Q}| = 0$ all task effects vanish exactly while the histogram can change. The negative covariance from assignment without replacement is retained explicitly in V_i.

E.14 INDEPENDENT NOISE: A SHARPER FUNCTIONAL BOUND

Corollary 18 (fresh Gaussian stability, allowing mean drift)**.** *Condition on the discovered circuit. Let the fills on $\mathcal{Q}$ be independent Gaussians with mean $\mu_F \in \mathbb{R}$ and standard deviation $\nu > 0$, and assume $|\mathcal{Q}| \geq 1$. Define a high-probability perturbation radius T_δ and a margin-loss bound D_i by $T_\delta = |\mu_F|\sqrt{|\mathcal{Q}|} + \nu(\sqrt{|\mathcal{Q}|} + \sqrt{2\log(4/\delta)})$ and $D_i = |\mu_F \mathbf{1}^\top \boldsymbol{g}_i| + \nu\|\boldsymbol{g}_i\|_2\sqrt{2\log(4K/\delta)} + \frac{1}{2}H_i T_\delta^2$. If $T_\delta \leq T$ then with probability at least $1-\delta$, $F_i(\mathbf{z}) \geq r_i^{\mathrm{pruned}} - D_i$ for all i; in particular $r_i^{\mathrm{pruned}} \geq \tau_i^{\mathrm{comp}} + D_i$ suffices for all completed target margins.*

Proof. For a standard Gaussian $\xi \in \mathbb{R}^{|\mathcal{Q}|}$ the norm is 1-Lipschitz with $\mathbb{E}\|\xi\|_2 \leq \sqrt{|\mathcal{Q}|}$, so $\mathbb{P}\{\|\xi\|_2 > \sqrt{|\mathcal{Q}|} + \sqrt{2\log(4/\delta)}\} \leq \delta/4$; each $\boldsymbol{g}_i^\top \xi$ is $\mathcal{N}(0, \|\boldsymbol{g}_i\|_2^2)$, and a two-sided tail bound with a union over i gives $\mathbb{P}\{\exists i : |\boldsymbol{g}_i^\top \xi| > \|\boldsymbol{g}_i\|_2\sqrt{2\log(4K/\delta)}\} \leq \delta/2$. On the intersection $\mathbf{z}_\mathcal{Q} = \mu_F \mathbf{1} + \nu\xi$ has norm at most T_δ and its linear effect is bounded below by minus the first two terms of D_i; Taylor control supplies the curvature term. Setting $\mu_F = 0$ gives (48). □

The random linear cost is proportional to $\nu\|\boldsymbol{g}_i\|_2$, and curvature costs order $H_i\nu^2|\mathcal{Q}|$. For any radius $0 \leq t \leq T$, the condition $r_i^{\mathrm{pruned}} \geq \tau_i^{\mathrm{comp}} + \|\boldsymbol{g}_i\|_2 t + H_i t^2/2$ for every i certifies margin preservation throughout that ball. Structural inactivity extends the guaranteed set by $\mathbb{R}^{|\mathcal{P}|-|\mathcal{Q}|}$ in the remaining coordinates. For independent centred sub-Gaussian fills with proxy variance v_F^2, use the norm bound $v_F\sqrt{2|\mathcal{Q}|\log(4|\mathcal{Q}|/\delta)}$ and the corresponding linear concentration bound. For histogram-constrained rank fills, the conditional assignment law above gives the corresponding moment bounds, while Theorem 5 applies to arbitrary assignments.

E.15 MAGNITUDE-CONSTRAINED ALLOCATION AND MARGIN PRESERVATION

Gaussian fit and margin preservation can hold with all pruned coordinates refilled. For a cap $u > 0$, let $J_u = \{b : \mathrm{int}(I_b) \cap (-u, u) \neq \varnothing\}$ and define

$$D_{\mathrm{out}}(\eta, u) = \sum_{b \notin J_u} [p_b(\eta) - \rho_b]_+, \qquad \mathcal{C}(u) = \inf_{\eta \in \Lambda} \max\{E_\eta, D_{\mathrm{out}}(\eta, u)\}. \tag{49}$$

Here J_u is the set of bins reachable by a fill of magnitude below u, $D_{\mathrm{out}}(\eta, u)$ is the target deficit in all other bins, and Λ is the specified family of Gaussian parameters. The cost $\mathcal{C}(u)$ balances retained excess against deficit outside the permitted fill range.

Theorem 19 (Gaussian fit with a fill-magnitude constraint)**.** *For continuous fill masses supported in J_u, the optimal TV at target η is $\max\{E_\eta, D_{\mathrm{out}}(\eta, u)\}$. With $|\mathcal{P}|$ actual fresh values,*

$$\mathcal{C}(u) \leq \inf_{\boldsymbol{z} \in (-u,u)^{|\mathcal{P}|}} \mathrm{GFit}(\boldsymbol{z}) \leq \mathcal{C}(u) + \frac{N_{\mathrm{bin}}}{2|\mathcal{I}_\theta|}. \tag{50}$$

If $r_i^{\mathrm{pruned}} \geq \tau_i^{\mathrm{comp}} + uL_i(u)$ for every i, then for every $\zeta > 0$ there exists a fresh fill, with no protected coordinate set, satisfying both $F_i(\boldsymbol{z}) \geq \tau_i^{\mathrm{comp}}$ and $\mathrm{GFit}(\boldsymbol{z}) \leq \mathcal{C}(u) + N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|) + \zeta$.

The cap limits perturbation size, while $\mathcal{C}(u)$ measures the Gaussian-fit cost of enforcing it. Allocation uses retained bin counts, the fixed cap and distributional randomness; the margin inequality supplies the condition for preserving the task. The empirical allocator samples unrestricted Gaussian tails, so its realised fill magnitudes must be used in the stability bounds.

Proof of Theorem 19. Fix η and put $D_{\mathrm{in}} = \sum_{b \in J_u}[p_b - \rho_b]_+$ and $D_{\mathrm{out}} = \sum_{b \notin J_u}[p_b - \rho_b]_+$. Since $D_{\mathrm{in}} + D_{\mathrm{out}} = \alpha + E_\eta$, the residual-allocation lower bound is E_η. Deficits outside the allowed bins cannot change, so TV is at least D_{out}.

If $\alpha \leq D_{\mathrm{in}}$, allocate α proportionally to the inside deficits. No excess is added and TV is E_η, with $E_\eta \geq D_{\mathrm{out}}$. If $\alpha > D_{\mathrm{in}}$, first fill every inside deficit, then distribute the remaining $\alpha - D_{\mathrm{in}}$ inside J_u. This is possible because J_u is nonempty. TV is $E_\eta + \alpha - D_{\mathrm{in}} = D_{\mathrm{out}}$. These cases prove the exact continuous optimum, including $D_{\mathrm{in}} = 0$.

The desired bin counts sum to $|\mathcal{P}|$. Largest-remainder rounding within J_u changes the histogram by at most $N_{\mathrm{bin}}/(2|\mathcal{I}_\theta|)$ in TV. Sample fresh values from atomless laws on $\mathrm{int}(I_b) \cap (-u, u)$ to realise

those counts; zeros and all finitely many deleted values are avoided almost surely. The target and within-bin rules can depend on retained counts and u, without task feedback. Taking the infimum over η proves Eq. (50). If no minimiser exists, choose a target within ζ of $\mathcal{C}(u)$. Every resulting fill obeys the common magnitude bound, so Theorem 5 gives all target margins simultaneously under the stated slack. □

Restricting allocation to J_u enforces the magnitude cap. Allowing all bins recovers the unrestricted proportional-residual allocation used in the experiment.

E.16 Layerwise refill and the spectral perturbation lemma

Use the feedforward representation of Corollary 6, writing $f_{\mathbf{W}^{\mathrm{pruned}}}$ and $f_{\mathbf{W}^{\mathrm{pruned}}+\mathbf{U}}$ for its baseline and refilled logit functions. The proof also permits any intervening activation that is 1-Lipschitz and vanishes at zero. Edge routing remains fixed throughout the comparison.

Lemma 20 (product perturbation bound)**.** *For $\|\boldsymbol{x}\|_2 \le B$ and arbitrary perturbations $\mathbf{U}_\ell$,*

$$\|f_{\mathbf{W}^{\mathrm{pruned}}+\mathbf{U}}(\boldsymbol{x}) - f_{\mathbf{W}^{\mathrm{pruned}}}(\boldsymbol{x})\|_2 \le B\left[\prod_{\ell=1}^{L}(\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2 + \|\mathbf{U}_\ell\|_2) - \prod_{\ell=1}^{L}\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2\right]. \quad (51)$$

The bound holds for arbitrary deterministic perturbations, including layers with zero baseline spectral norm.

Proof. Let $\boldsymbol{h}_\ell, \boldsymbol{h}'_\ell$ be the baseline and perturbed activations, and $\Delta_\ell = \|\boldsymbol{h}'_\ell - \boldsymbol{h}_\ell\|_2$. The Lipschitz property and vanishing at zero imply $\|\boldsymbol{h}'_\ell\|_2 \le B\prod_{j\le\ell}(\|\mathbf{W}_j^{\mathrm{pruned}}\|_2 + \|\mathbf{U}_j\|_2)$ and

$$\Delta_\ell \le \|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2\Delta_{\ell-1} + \|\mathbf{U}_\ell\|_2\|\boldsymbol{h}'_{\ell-1}\|_2, \qquad \Delta_0 = 0.$$

Induction yields $\Delta_\ell \le B[\prod_{j\le\ell}(\|\mathbf{W}_j^{\mathrm{pruned}}\|_2+\|\mathbf{U}_j\|_2) - \prod_{j\le\ell}\|\mathbf{W}_j^{\mathrm{pruned}}\|_2]$. The final linear map satisfies the same inequality, proving the result. □

Proof of Corollary 6. Lemma 2 of Neyshabur et al. (2018) applies with baseline layers $\mathbf{W}_\ell^{\mathrm{pruned}}$ and perturbations $\mathbf{U}_\ell$. The preceding product bound also gives a direct proof: put $\eta_\ell = \|\mathbf{U}_\ell\|_2/\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2 \le 1/L$ and $S = \sum_\ell \eta_\ell \le 1$. Then

$$\prod_\ell(1+\eta_\ell) - 1 \le e^S - 1 \le eS,$$

so the output change is at most $eB(\prod_\ell \|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2)\sum_\ell \eta_\ell$. A target-token-pair logit margin is the inner product with $\boldsymbol{e}_{y_i^+} - \boldsymbol{e}_{y_i^-}$, where $\boldsymbol{e}_y$ is the unit vector for vocabulary coordinate y. Its Euclidean norm is $\sqrt{2}$. Cauchy–Schwarz gives Eq. (30). Only positive baseline margins no larger than the bound can become nonpositive; this proves the accuracy statement, with ties counted as failures. □

Converting fill size into a norm certificate. If layer ℓ has u_ℓ holes, at most r_ℓ holes per row and c_ℓ per column, then a common cap $|\mathrm{z}_j| \le u$ yields

$$\|\mathbf{U}_\ell\|_2 \le u\min\{\sqrt{u_\ell}, \sqrt{r_\ell c_\ell}\}. \quad (52)$$

The first inequality uses the Frobenius norm; the second uses $\|\mathbf{U}\|_2 \le \sqrt{\|\mathbf{U}\|_1\|\mathbf{U}\|_\infty}$, where the matrix norms on the right are the maximum absolute column sum and row sum, respectively. Both hold for every coordinate assignment. A realised multiset also gives direct Frobenius or spectral norm certificates; the refill scale determines such a certificate through the applicable magnitude bound. For the distinct case of conditionally independent centred Gaussian fills at scale ν, a union bound gives $\max_{j\le|\mathcal{P}|}|\mathrm{z}_j| \le \nu\sqrt{2\log(2|\mathcal{P}|/\delta)}$ with probability at least $1-\delta$, which can be substituted in Eq. (52). This Gaussian sampling case supplies an explicit probabilistic cap.

Corollary 21 (joint fit and spectral stability without protected coordinates)**.** *Use the bins, Gaussian family and cost $\mathcal{C}(u)$ of Theorem 19, and the feedforward model above. Define a layerwise perturbation bound $t_\ell(u)$ and the induced pairwise margin-loss bound $D(u)$ by*

$$t_\ell(u) = u\min\{\sqrt{u_\ell}, \sqrt{r_\ell c_\ell}\}, \qquad D(u) = \sqrt{2}B\left[\prod_\ell (\|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2 + t_\ell(u)) - \prod_\ell \|\mathbf{W}_\ell^{\mathrm{pruned}}\|_2\right].$$

If every baseline margin obeys $r_i^{\mathrm{pruned}} \geq \tau_i^{\mathrm{comp}} + D(u)$ with $\tau_i^{\mathrm{comp}} > 0$, then for every $\zeta > 0$ the capped allocation supplies a fresh fill with

$$F_i(\boldsymbol{z}) \geq \tau_i^{\mathrm{comp}} \quad \forall i, \qquad \mathrm{GFit}(\boldsymbol{z}) \leq \mathcal{C}(u) + \frac{N_{\mathrm{bin}}}{2|\mathcal{I}_\theta|} + \zeta.$$

The fill rule uses retained counts, a declared cap and fresh randomness; every hole is refilled.

Proof. The magnitude-constrained histogram construction in the proof of Theorem 19 gives the stated distributional bound with $|z_j| < u$. Equation (52) holds for every such fill. Apply Lemma 20 and the pairwise readout bound; no local differentiability assumption or conditional permutation law is required. □

We fix the cap before allocation and use the margins and norms to check the stability condition. The allocator chooses values from retained bin counts and fresh randomness. For the empirical procedure with unrestricted tail sampling, the spectral bound instead uses the realised perturbation norms.

Module representation and biases. The spectral argument applies to the stated feedforward representation and uses the Lipschitz property of ReLU, including at its nondifferentiable points. Attention has the separate product certificate in Eq. (38); composing modules requires bounds for the intervening transformations and parameter uses. For an affine ReLU network, let $\boldsymbol{b}_\ell$ and $\Delta\boldsymbol{b}_\ell$ be the baseline bias and its refill perturbation. Homogeneous coordinates incorporate them as

$$\widetilde{\mathbf{W}}_\ell = \begin{pmatrix} \mathbf{W}_\ell & \boldsymbol{b}_\ell \\ 0 & 1 \end{pmatrix}, \qquad \widetilde{\mathbf{U}}_\ell = \begin{pmatrix} \mathbf{U}_\ell & \Delta\boldsymbol{b}_\ell \\ 0 & 0 \end{pmatrix}.$$

Start with $(\boldsymbol{x}, 1)$, use ReLU on hidden augmented coordinates (the last stays at one), and project away the final dummy coordinate. The proof then uses augmented spectral norms and input radius $\sqrt{B^2+1}$. The augmented constant row represents the affine map. Each repeated parameter use enters the layer representation and its norm bound; Appendix B.2 gives the empirical shared-bias routing convention. The perturbation lemma therefore gives a deterministic bound on the changes in task margins.

E.17 Constraints on Gaussian fit and task performance

Proposition 22 (small-fill concentration obstruction)**.** *Let $t \geq 0$ be a fill-magnitude bound and $J(t)$ the union of bins meeting $[-t, t]$. Write $\rho(J(t))$ for the retained mass in this union and $p_\eta(J(t))$ for its target Gaussian probability. Let $\mathrm{TV}_\eta(\boldsymbol{z}) = \frac{1}{2}\sum_b |\rho_b + \phi_b(\boldsymbol{z}) - p_b(\eta)|$ be the binned total variation at that target, using the fill masses $\phi_b(\boldsymbol{z})$ in Eq. (27). If all fills lie in $[-t, t]$, then at any target η, $\mathrm{TV}_\eta(\boldsymbol{z}) \geq [\rho(J(t)) + \alpha - p_\eta(J(t))]_+$. If every allowed target has standard deviation at least $\sigma_{\min} > 0$ and $J(t)$ has finite total length $\lambda(t)$, then $\mathrm{GFit}(\boldsymbol{z}) \geq [\alpha - \lambda(t)/(\sigma_{\min}\sqrt{2\pi})]_+$.*

Proof. On bins, total variation is the supremum of absolute probability differences over unions of bins, and the completed mass on $J(t)$ is $\rho(J(t)) + \alpha$; take that union. A Gaussian density with $\sigma \geq \sigma_{\min}$ is at most $1/(\sigma_{\min}\sqrt{2\pi})$, so its mass on $J(t)$ is at most $\lambda(t)/(\sigma_{\min}\sqrt{2\pi})$; drop the nonnegative retained mass and take the infimum over targets. □

If interior bins have width at most $h > 0$ and $J(t) \subseteq [-t-h, t+h]$ avoids the tails, a fit $\mathrm{GFit}(\boldsymbol{z}) \leq \epsilon$ requires $t \geq \frac{1}{2}[\sigma_{\min}\sqrt{2\pi}(\alpha - \epsilon) - 2h]$.

Proposition 23 (retained spread forces a target scale)**.** *Let $a < b$ be two separating locations on the weight axis. Suppose two unions of bins lie in $(-\infty, a]$ and $[b, \infty)$, each containing retained mass at least β. Fix a fit tolerance ϵ with $0 < \epsilon < \beta < 1/2$. Every Gaussian whose binned total variation to a completion is at most ϵ satisfies $\sigma \geq (b-a)/(2\Phi^{-1}(1-\beta+\epsilon))$, where Φ is the standard normal CDF.*

Proof. A fit within ϵ implies Gaussian mass at least $\beta - \epsilon$ in each union, hence in each containing half-line. With $z = \Phi^{-1}(1-\beta+\epsilon) > 0$ the left tail gives $\mu \leq a + z\sigma$ and the right tail $\mu \geq b - z\sigma$, so $b - a \leq 2z\sigma$. □

Thus the fixed retained mass places a lower bound on the Gaussian scale, so the small-fill obstruction applies to every candidate with a sufficiently close fit.

Population versus empirical completion. Let H be a retained subprobability measure of mass $1-\alpha$ and G a nondegenerate Gaussian probability measure, with $0 < \alpha < 1$ the refill fraction. A probability law F for fresh fills satisfies $H + \alpha F = G$ exactly when $G - H$ is a nonnegative measure, in which case $F = (G-H)/\alpha$. When densities h and g exist, the condition is $h \leq g$ almost everywhere. At finite $|\mathcal{I}_\theta|$ the retained empirical measure has atoms a continuous Gaussian cannot dominate, so this applies to population or limiting laws while Proposition 4 applies to finite histograms; if retained empirical sub-CDFs converge uniformly to H, $|\mathcal{P}|/|\mathcal{I}_\theta| \to \alpha$ and fill empirical CDFs converge uniformly to F, the completed CDF converges uniformly to $H + \alpha F$, which is convergence to *some* Gaussian whenever a feasible limiting completion law exists.

E.18 A CONSTRUCTIVE EXAMPLE WITH ALL PRUNED COORDINATES AFFECTING THE TASK

The following one-task random linear-feature model realises Gaussian completion and margin preservation together, with every refilled coordinate contributing to the task.

Proposition 24 (Gaussian-preserving fresh refill with growing margins)**.** *Fix an original weight scale $\sigma > 0$, an expected pruning fraction $0 < \alpha < 1$ and a refill standard deviation ν with $\alpha\sigma \leq \nu \leq \sigma$. For each of n scalar coordinates, independently draw $\mathrm{X}_j \sim \mathcal{N}(0, \sigma^2)$, a feature sign $\mathrm{B}_j \in \{-1, +1\}$ uniformly and a uniform selection threshold $\mathrm{U}_j \sim \mathrm{Unif}[0,1]$. With $\varphi_\sigma, \varphi_\nu$ the centred Gaussian densities, define the retention probability averaged over feature signs, $p(x)$, and its allowable task-directed adjustment, $h(x)$, by*

$$\begin{aligned} p(x) &= 1 - \alpha\frac{\varphi_\nu(x)}{\varphi_\sigma(x)}, \qquad h(x) = \min\{p(x), 1-p(x)\}, \\ (m_\theta)_j &= \mathbf{1}\{\mathrm{U}_j \leq p(\mathrm{X}_j) + h(\mathrm{X}_j)\,\mathrm{sgn}(\mathrm{B}_j\mathrm{X}_j)\}. \end{aligned} \tag{53}$$

Here $\mathrm{sgn}(t)$ is -1, 0 or 1 according as t is negative, zero or positive. Refill each removed coordinate with $\mathrm{z}_j \sim \mathcal{N}(0, \nu^2)$, independent across j and of all previous variables, and put $\mathrm{Y}_j = (m_\theta)_j\mathrm{X}_j + (1-(m_\theta)_j)\mathrm{z}_j$. Then $\mathbb{P}((m_\theta)_j = 0) = \alpha$; each Y_j has exactly the marginal $\mathcal{N}(0,\sigma^2)$ and the completed pooled empirical CDF converges to it almost surely as $n \to \infty$; with $r^{\mathrm{full}} = n^{-1/2}\sum_j \mathrm{B}_j\mathrm{X}_j$, $r^{\mathrm{pruned}} = n^{-1/2}\sum_j \mathrm{B}_j(m_\theta)_j\mathrm{X}_j$ and $F(\mathbf{z}) = n^{-1/2}\sum_j \mathrm{B}_j\mathrm{Y}_j$ one has $r^{\mathrm{full}} \sim \mathcal{N}(0,\sigma^2)$, $\mathbb{E}r^{\mathrm{pruned}} = \mathbb{E}F(\mathbf{z}) = c\sqrt{n}$, where $c = \mathbb{E}[|\mathrm{X}_j|h(\mathrm{X}_j)] > 0$ is the per-coordinate drift coefficient. For every margin threshold $\tau < c\sqrt{n}$, $\mathbb{P}\{r^{\mathrm{pruned}} < \tau \text{ or } F(\mathbf{z}) < \tau\} \leq 2\sigma^2/(c\sqrt{n}-\tau)^2$. Every refilled coordinate is task-visible, so $|\mathcal{Q}| = |\mathcal{P}|$.

Proof. Since $\nu \leq \sigma$, $\varphi_\nu/\varphi_\sigma \leq \sigma/\nu \leq 1/\alpha$, so $0 \leq p \leq 1$ and $0 \leq p \pm h \leq 1$ and the mask is well defined. Averaging over the symmetric sign gives $\mathbb{P}((m_\theta)_j = 1 \mid \mathrm{X}_j = x) = p(x)$, so the retained subdensity is $p\varphi_\sigma = \varphi_\sigma - \alpha\varphi_\nu$ with mass $1-\alpha$, and adding independent fills of mass α gives density φ_σ. Independence across coordinate tuples makes the Y_j i.i.d. unconditionally, and Glivenko–Cantelli gives the empirical-CDF convergence. Conditional on $\mathrm{X}_j = x$, averaging $\mathrm{B}_j(m_\theta)_j\mathrm{X}_j$ over B_j and U_j gives $|x|h(x)$; h is positive on a set of positive Gaussian measure so $c > 0$, and fresh fills have zero conditional mean so $\mathbb{E}[\mathrm{B}_j\mathrm{Y}_j] = c$ too. $\mathbb{E}[(\mathrm{B}_j(m_\theta)_j\mathrm{X}_j)^2] \leq \sigma^2$ and $\mathbb{E}[(\mathrm{B}_j\mathrm{Y}_j)^2] = \sigma^2$, so $\mathrm{Var}r^{\mathrm{pruned}}, \mathrm{Var}F(\mathbf{z}) \leq \sigma^2$; Chebyshev and a union bound give the failure bound. Symmetric signs leave the full Gaussian law unchanged, and the refill derivatives $\mathrm{B}_j/\sqrt{n}$ are nonzero. □

These probabilities are taken over random features, original weights and fills. The mask depends on task-feature signs, while refill draws fresh Gaussian samples at scale ν. When $\alpha = 1/2$, setting $\nu = \sigma/2$ restores the pooled Gaussian law with a smaller refill standard deviation. Correlation between completed weights and feature signs preserves the positive task margin; the Gaussian marginal alone does not describe that relationship.

E.19 SCOPE OF THE GUARANTEES

The computation graph fixes how masks intervene. Under the stagewise additive assumptions, rounding gives sufficient conditions for binary masks to reach the target margins. The random-direction assumptions yield explicit Gaussian-weight constructions, and the centred block model separates scalar selection from edge selection on the same draw. These are existence results for finite collections of margins. The experiments test whether DiscoGP finds task-functional circuits in the evaluated transformers.

For completion, retained values may be task-dependent. Their excess mass limits histogram fit, while uniform sensitivity and spectral norms bound the effect of refill on margins. Conditional uniform assignment gives additional moment and success formulas. The capped construction combines a fit bound with deterministic margin preservation; the Gaussian-preserving example shows that both can hold even when every pruned coordinate affects the task. In these constructions, allocation uses distributional information and fresh randomness. Margins and sensitivities are used to certify the effect, not to choose fills.

The results apply to the stated models and finite comparisons. They explain how a task-guided mask can create positive margins and how task-agnostic refill can preserve them while removing the pooled point mass at zero.

F NOTATION GLOSSARY

Bold letters distinguish vectors and matrices from nonbold scalar entries; upright letters identify random variables (for example, X_j and the random fill vector $\mathbf{z}$, with realised value $\boldsymbol{z}$). The parameter array θ and circuit masks m_θ, m_E keep their established notation. The empirical construction and theory use the same circuit notation. Weight and edge masks are m_θ and m_E; subscripts on a mask specify a parameter array or coordinate. The circuit $C = (m_\theta, m_E)$ has margin $r_i((m_\theta, m_E))$. Unsubscripted $\mathbf{1}$ denotes an all-ones vector of the required size; $\mathbf{1}\{\cdot\}$ is an indicator, equal to one when its condition holds and zero otherwise. For vectors, $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are the sum of absolute entries, Euclidean norm and maximum absolute entry. For matrices, $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ is the Frobenius norm. The count $\|\cdot\|_0$ gives the number of nonzero entries, and $\odot$ denotes entrywise multiplication. Logarithms are natural; $\delta \in (0, 1)$ bounds a failure probability. Symbols reused by different simplified models have their local meanings specified below and at their point of use.

Table 13: Circuit, task and margin-synthesis notation.

Symbol	Meaning and scope
$\mathcal{T}, E$	Target task and eligible directed-edge set.
$\mathcal{I}_\theta, \lvert\mathcal{I}_\theta\rvert$	Finite index set of eligible scalar parameter coordinates and its cardinality. An index identifies a tensor and a location within it; distinct coordinates are counted separately even when their values coincide.
$\theta = (\theta_j)_{j\in\mathcal{I}_\theta}$	Scalar parameter values indexed by the specified domain.
$\mu_{\text{reinit}}, \sigma_{\text{reinit}}$	Mean and standard deviation of pooled eligible pretrained values, used for the pre-clipping Gaussian draw.
$a_{\text{reinit}}, b_{\text{reinit}}$	Minimum and maximum of that pool, used as lower and upper clipping bounds.
$\theta^{\text{reinit}}, \theta^{\text{comp}}(\boldsymbol{z})$	Frozen randomised parameters and completed parameters after filling pruned coordinates with $\boldsymbol{z}$.
$C = (m_\theta, m_E)$	Circuit: binary scalar-weight mask and binary edge mask, with entries $(m_\theta)_j$ and $(m_E)_e$.
$\mathbf{1}_\theta, \mathbf{1}_E$	All-on masks over eligible scalar weights and eligible edges.
$\widetilde{m}_\theta, \widetilde{m}_E$	Relaxed mask witnesses with entries in $[0, 1]$. A bar, as in $\bar{m}_\theta$, denotes a differentiation reference.
$(m_\theta)_\ell, (m_\theta)_{\mathrm{Proj},h}$	Restrictions to a layer or head projection. Subscripts U, V similarly restrict a mask to Gaussian factors.
m_E^S, s_E	Edge mask supported on $S \subseteq E$; maximum retained-edge count.
$A_\mathcal{T}(\theta; C), \mathrm{Acc}(C)$	Empirical task accuracy and its finite-comparison form, evaluated on the specified parameter state and examples; ties fail.
ρ_θ, ρ_E	Retained fractions of eligible weights and edges. Sparsity is one minus density.
α, ε	In Section 3: desired accuracy and maximum edge density. In completion theory, α is the removed fraction, or its expectation in the population example.
$K, i, r_i(C), \widetilde{r}_i(C)$	Number and index of prescribed token comparisons; exact target-token-pair logit margin and its surrogate.
$\boldsymbol{\tau}, \tau_i, \mathcal{M}_\tau, \kappa$	Target-margin vector and its entries, region satisfying all target margins, and uniform bound on exact–surrogate error.
$\mathbf{H} = (h_{ij}), \boldsymbol{b}, \boldsymbol{q}$	Fixed feature-readout matrix, baseline offsets and orientation signs for shared-mask synthesis. Entry h_{ij} describes the effect of weight j on comparison i.
$n, \mathrm{X}_j, \sigma, d_i, v_i$	Number of maskable scalars, standard Gaussian coefficient, weight scale, oriented average readout and root-mean-square readout in the feature model.
d_*, v_*, T	Uniform positive lower bound on d_i, upper bound on v_i, and largest target-margin deficit after baseline and error adjustments.
$\boldsymbol{o}, \bar{\boldsymbol{a}}_e, \boldsymbol{a}_{ej}$	Fixed offset vector, edge baseline and contribution of a scalar coordinate through an edge in the nested additive model, all in $\mathbb{R}^K$.
$(e, j), \mathrm{X}_{ej}, \sigma_\theta$	Edge and within-edge scalar-coordinate indices, standard Gaussian coefficient and weight scale in the disjoint-block model. Here n counts scalar coordinates per edge block.
$[x]_+$	Positive part $\max\{x, 0\}$. The identity $\mathbf{1}\{x > 0\}x = [x]_+$ describes literal zero-based selection.

Table 14: Attention, approximation and rounding notation.

Symbol	Meaning and scope
H, h, n_K	Number of heads, head index and query/key width in the attention model.
$\mathbf{X}, \boldsymbol{x}_t$	Token-input matrix and column vector at position t; row t of $\mathbf{X}$ is $\boldsymbol{x}_t^\top$.
B	Uniform upper bound on input Euclidean norms: $\|\boldsymbol{x}_t\|_2 \le B$ for attention tokens, or $\|\boldsymbol{x}_i\|_2 \le B$ for inputs to the feedforward certificate.
$\mathcal{A}_t$	Fixed nonempty set of positions available to token t. Restricted softmax normalises over these positions and is zero elsewhere.
$\mathbf{W}_{\mathrm{Proj},h}^{m_\theta}$	Masked query, key, value or output projection, $\mathrm{Proj} \in \{Q, K, V, O\}$; the mask has the same shape as its matrix.
$\mathbf{A}_h, \mathbf{P}_h$	Query–key product acting on the input feature space and value–output product mapping it to the common output feature space. The former includes $n_K^{-1/2}$ normalisation.
$\mathbf{A}_h^\star, \mathbf{P}_h^\star, r_i^\star$	Target effective products and resulting target margin under the same pairwise readout.
$\epsilon_{A,h}, \epsilon_{P,h}, L_i$	Spectral approximation tolerances and Euclidean norm of the fixed target-token-pair readout. The function $L_i(u)$ below measures refill sensitivity.
$\mathbf{U}, \mathbf{V}, a, b, n_{\mathrm{grp}}$	Gaussian factors of shapes $a \times n$ and $n \times b$, where $n = bn_{\mathrm{grp}}$. The group width n_{grp} counts hidden coordinates assigned to each output column.
$\sigma_U, \sigma_V, \omega, \lambda$	Factor standard deviations, their product $\omega = \sigma_U\sigma_V$, and product normaliser.
c, β_c, R, ϵ	Product-band threshold, probability $\beta_c = \Pr(c \le \mathrm{GH} \le 2c)$ for independent standard Gaussians G, H, target-entry magnitude bound and spectral approximation tolerance.
$\mathbf{T}$	Target matrix in Gaussian product approximation. Its local role differs from the scalar margin deficit used in Eq. (22).
$\mathbf{A}, \boldsymbol{b}_\theta, \boldsymbol{e}_\theta, \kappa_\theta$	Weight-stage signature matrix, intercept, remainder and uniform remainder bound.
$\mathbf{G}(m_\theta), \boldsymbol{b}_E, \boldsymbol{e}_{E,m_\theta}, \kappa_E$	Edge-stage signatures, intercept, remainder and bound on the same frozen weight mask. $\boldsymbol{g}_e^{\mathrm{full}} = \boldsymbol{g}_e(\mathbf{1}_\theta)$ denotes an edge signature with all scalar weights retained.
$\mathbf{C}, M, \boldsymbol{\alpha}, \boldsymbol{v}$	Generic signature matrix, column count, relaxed coefficient vector and binary rounded vector in the rounding lemmas.
$k_{\mathrm{mask}}, \varepsilon_\star, \varepsilon^{\mathrm{bud}}$	Integer budget for the sum of coefficients, unconstrained rounding radius and rounding radius with an exact budget.
$\mu_\theta, \sigma_\theta, k, c_0, p_0$	Gaussian coefficient mean and scale, number of target directions, signed threshold and lower tail probability in directional reachability.
$S, T, \mathrm{IoU}(S, T)$	Edge supports and their intersection over union, $\lvert S \cap T\rvert / \lvert S \cup T\rvert$, for nonempty union.

Table 15: Distributional completion notation.

Symbol	Meaning and scope
$\mathcal{P}, \lvert\mathcal{P}\rvert, \alpha$	Pruned-coordinate index set, its size, and removed fraction $\lvert\mathcal{P}\rvert/\lvert\mathcal{I}_\theta\rvert$ in completion theory.
$\boldsymbol{z}, \mathbf{z}, C_{\mathrm{comp}}$	Realised fill vector, random fill vector, and completed circuit $(\mathbf{1}_\theta, m_E)$.
f, a, ν, ξ_j	In the empirical reference law: scale relative to σ_{reinit}, mode-separation parameter, standard deviation $\nu = f\sigma_{\mathrm{reinit}}$, and raw reference value. Here a controls separation, not a fill cap.
$\mathrm{B}_j, \mathrm{G}_j$	Independent equiprobable signs and standard Gaussian coefficients in the empirical reference draw.
$B_b, I_b, N_{\mathrm{bin}}$	Empirical and theoretical histogram intervals and the number of bins, including tails.
$\widehat{\mu}, \widehat{\sigma}, k_b, q_b$	Fitted Gaussian mean and standard deviation, retained bin count and fitted target bin probability in the empirical allocator.
$d_b, x_b^\star$	Positive target-count deficit and continuous number of fills allocated to a bin before integer rounding.
$\widehat{P}_{\mathrm{ret}}, \widehat{P}_{\mathrm{fill}}, \widehat{P}_{\mathrm{comp}}$	Empirical distributions of retained values, fills and their completed mixture.
ρ_b, ϕ_b	Retained and fill bin masses: coordinate counts divided by the total eligible count $\lvert\mathcal{I}_\theta\rvert$.
$\Lambda, \eta = (\mu, \sigma), p_b(\eta)$	Allowed Gaussian mean–standard-deviation pairs, one candidate pair and its bin probabilities.
$\mathrm{GFit}, R_{\mathrm{Ghist}}$	Best binned total-variation fit of the completed array and smallest excess retained mass over the specified Gaussian family.
E_η, D_η, x_b	Excess retained mass, total positive deficit $\alpha + E_\eta$, and continuous fill mass allocated to a bin. Unlike $x_b^\star$, x_b is normalised by the total coordinate count.
$u, J_u, \mathcal{C}(u)$	In bounded-fill theory: cap on absolute fill magnitude, bins reachable within the cap, and optimal continuous capped-fit cost.

Table 16: Refill stability and random assignment notation.

Symbol	Meaning and scope
$F_i(\boldsymbol{z}), r_i^{\mathrm{pruned}}, \tau_i^{\mathrm{comp}}$	Exact completed margin, margin before refill and desired positive completed margin.
$\mathrm{Acc}_{\mathrm{fill}}(\boldsymbol{z})$	Fraction $K^{-1}\sum_i \mathbf{1}\{F_i(\boldsymbol{z}) > 0\}$ of successful comparisons on the completed weights at fixed edge routing.
$L_i(u)$	Worst-case aggregate margin sensitivity: the supremum of $\|\nabla F_i(\boldsymbol{z})\|_1$ over the fill cube $\|\boldsymbol{z}\|_\infty \le u$.
$\boldsymbol{g}_i, H_i$	Per-coordinate additive margin sensitivities (or gradient at zero in the smooth model) and a local Hessian operator-norm bound.
$\mathcal{Q}, \lvert\mathcal{Q}\rvert$	Set of pruned coordinates on which the evaluated logits structurally depend, and its size; every pruned coordinate remains available for refill.
$a_j, \bar{a}, v^2, M$	Fixed multiset values, their mean, population variance and maximum absolute value in the assignment bound. This local amplitude differs from the rounding column count.
μ_i, V_i, Δ_i	Completed-margin mean, variance and simultaneous margin-loss bound over random assignment.
$\mathbf{W}_\ell^{\mathrm{pruned}}, \mathbf{U}_\ell, \mathbf{Z}_\ell, L$	Pruned layer matrix, refill perturbation, fill array and network depth in the feedforward certificate.
$u_\ell, D_{\mathrm{spec}}$	Number of pruned entries in a layer and margin-change bound from baseline and perturbation spectral norms.
$\Delta\mathbf{R}_h$	Refill perturbation to head projection $R \in \{Q, K, V, O\}$ in the attention bound.
σ, α, ν, c	Original Gaussian standard deviation, expected removed fraction, refill standard deviation and positive drift coefficient in the constructive population example.
$r^{\mathrm{full}}, r^{\mathrm{pruned}}, F(\mathbf{z})$	Unpruned, pruned and completed one-task margins in that example.