FINDING SUBSPACE CIRCUITS WITH SUBSPACE ATTRIBUTION PATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Edge attribution patching (EAP) has been widely adopted as a tool for finding *circuits*—small, task-specific computational subgraphs—within language models. A growing body of evidence suggests that the model components within a circuit communicate via low-rank subspaces of the residual stream. However, while circuit localization is well-studied, the study of the subspaces used by a circuit has received little attention. This work proposes *subspace attribution patching* (SAP), a finer-grained formulation of EAP, which finds circuit edges jointly with the causally-relevant subspace for each edge. We rigorously evaluate a range of methods for constructing subspace bases for SAP, relying on spectral decompositions of model weights and activations. Intriguingly, we demonstrate that several plausible SAP approaches can perform well while not being faithful to the model's computations, falling prey to the Interpretability Illusion of Makelov et al. (2024). We conclude by discussing a method to diagnose this illusion. Our findings enable researchers 1) to jointly localize circuit edges and low-rank information subspaces for a given task, and 2) to detect illusory results, paving the way for more reliable research in this area.

1 INTRODUCTION

As the capabilities and scale of language models (LMs) advance, it becomes increasingly important to understand their inner workings. To do this, researchers have turned to **circuit localization**, which attempts to find **circuits**—small computational subgraphs of model components—which allow an LM to perform a certain task such as indirect object identification (IOI Wang et al., 2022), greater-than operations (Hanna et al., 2023), arithmetic (Stolfo et al., 2023), *inter alia* (Olah et al., 2020; Elhage et al., 2021; Conmy et al., 2023; Marks et al., 2025; Huang & Hahn, 2026). To do so, circuit localization literature treats an LM as a graph $\mathcal{M}$. The units (or **nodes**) of the graph are typically coarse structural model components, such as attention heads or MLPs. In Transformers, all inter-component interaction is mediated by the **residual stream**, which is the additive activation space written to by earlier (upstream) nodes and read from by later (downstream) nodes. As such, the **edge** $u \to v$ between two nodes u, v constitutes part of a circuit for a specific task and is typically taken to mean that "the residual stream output of u, as read/processed by v, matters for the task". Analyzing the functional role of u and v, as well as the contents of the communication between them, remains difficult and involves significant manual analysis (e.g., Wang et al., 2022).

Elhage et al. (2021) theorize that model components involved in a particular task communicate across low-rank subspaces of the residual stream. There has since been an increased interest not only in localizing a circuit to a subset of model components, but also in attempting to locate and interpret the subspace which carries the task-relevant information between them (see Mueller et al. (2026) for a review). Most closely to the current work, Geiger et al. (2024) attempt to find orthogonal projections of the residual stream which align high-level human-specified variables in a causal model with non-basis-aligned subspaces of the residual stream. Marks et al. (2025) decompose the residual stream outputs of every component using sparse autoencoders (SAEs), and localize circuits with respect to the SAE features which activate at every token position. Dunefsky et al. (2024); Paulo et al. (2025); Lindsey et al. (2024) take a similar approach with Transcoders and Cross-Layer Transcoders (CLTs), which approximate the input-output mapping of a layer, rather than simply decomposing the input or output as with an SAE. However, Engels et al. (2025) argue that some task-relevant subspaces are irreducibly multidimensional (such as a circular representation of the

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.

days of the week), which cannot be represented by an SAE or related approach. Additionally, Lange et al. (2026) illustrate that CLTs-based circuit localization is frequently unfaithful to the underlying computations. Huang & Hahn (2026) partition the residual stream into uncorrelated subspaces, with an iterative algorithm that merges irreducibly high-correlation directions into one subspace. They propose that such subspaces can be akin to "variables" in the model and can form good fundamental units for finding circuits, but leave this to future work.

Building on the above, this work proposes **subspace attribution patching** (SAP), a circuit localization method which finds not just the edges involved in a circuit, but also the subspace which carries the task-relevant information for every circuit edge. We present a range of methods for constructing the basis directions for SAP, and rigorously evaluate the faithfulness of the circuits they yield on a range of tasks using the framework proposed by the Mechanistic Interpretability Benchmark (Mueller et al., 2025). Intriguingly, we find that the SAP basis with the highest circuit compactness and faithfulness actually finds circuits that are completely unfaithful to the underlying model's computation. This reinforces the findings of Makelov et al. (2024), who call this phenomenon the Interpretability Illusion and hypothesize that it may be prevalent in mechanistic interpretability work. To this end, we present a normalized conditional variance metric which can diagnose illusory SAP findings, and demonstrate that the unfaithful basis scores worst on the metric. In further support of Elhage et al. (2021), we close with an exhaustive qualitative analysis of the subspaces used by GPT-2 on the IOI dataset, demonstrating that the subspaces found by SAP are frequently low-rank and highly interpretable.

2 PRELIMINARIES

Notation. Let index n denote a particular token position in a particular input sequence. Let $\boldsymbol{m}_u^{(n)} = \boldsymbol{z}_u^{(n)} \boldsymbol{W}_{u,out}^\top \in \mathbb{R}^{1\times d_m}$ be the output (or *message*) written by writer u to the residual stream, where $\boldsymbol{z}_u^{(n)} \in \mathbb{R}^{1\times d_u}$ is its hidden activation u and $\boldsymbol{W}_{u,out} \in \mathbb{R}^{d_m\times d_u}$ is its output weight matrix. Let $\boldsymbol{h}_v^{(n)} = \mathrm{LN}(\boldsymbol{r}_v^{(n)})\boldsymbol{W}_{v,in} \in \mathbb{R}^{1\times d_v}$ be the hidden activation of reader v, $\boldsymbol{r}_v^{(n)} \in \mathbb{R}^{1\times d_m}$ being the value of the residual stream at its input, and $\boldsymbol{W}_{v,in} \in \mathbb{R}^{d_m\times d_u}$ its weight matrix. For all data-derived values a (e.g., $\boldsymbol{m}_u^{(n)}, \boldsymbol{r}_v^{(n)}$, etc.), let a' denote the same value on the same n on the counterfactual (corrupt) prompt.

Datasets and models To compare with prior results from the Mechanistic Interpretability Benchmark (Mueller et al., 2025), we test our methods on the Indirect Object Identification dataset (IOI; Wang et al., 2022). We also rely on datasets from Hanna et al. (2024): Subject-Verb Agreement (SVA Finlayson et al., 2021; Newman et al., 2021), Greater-Than (Hanna et al., 2023), Gender-Bias (Vig et al., 2020), Capital-Country and Hypernymy. Please see Appendix A for more details. We conduct our experiments with GPT2-Small (Radford et al., 2019), as it is the most widely-studied model for mechanistic interpretability, with known circuits for several of the tasks above.

2.1 FINDING CIRCUITS

The gold standard approach for circuit localization is **activation patching**, also known as causal mediation analysis (Vig et al., 2020). Given a minimal pair of a *clean* and a *corrupt* prompt (x, x') which elicit the desired change in model behaviour[1], activation patching caches model activations $(\boldsymbol{m}_u, \boldsymbol{m}_u{}') \in \mathbb{R}^{1\times d_m}$ at every component u across both prompts. It then iterates through every component u, and attempts to change the model prediction when processing x by substituting (*patching*) $\boldsymbol{m}_u{}'$ for $\boldsymbol{m}_u$ as the activation of u, leaving the rest of the model operation unadulterated. The extent to which the intervention scores highly on a test metric of choice L, such as logit difference under the model $L(\mathcal{M}) = y' - y$ between the correct answer y given x and the incorrect answer y' given x', reflects the causal importance of component u in the circuit for the task:

$$\Delta L_u := L(\mathcal{M}) - L(\mathcal{M}|\,\mathrm{do}(\boldsymbol{m}_u \leftarrow \boldsymbol{m}_u{}')) \qquad \text{(activation patching)}$$

This approach can be easily extended to **edge activation patching** (EActP). Given that the residual stream is linearly additive, $\boldsymbol{r}_v$ already includes the contribution of $\boldsymbol{m}_u$. Therefore, EActP estimates

[1]For example, in the IOI task, x might be "When Mary and John went to the store, John gave a drink to", and x' may be "When [...], **Mary** gave a drink to". This pair of prompts, differing in one token, is meant to change the next-token prediction from 'Mary' (for x) to 'John' (for x').

the causal importance of specific edges $u \to v$ by patching the residual stream input $\boldsymbol{r}_v$ of the reader node v with the difference of messages $\boldsymbol{m}_u{}' - \boldsymbol{m}_u$. This approximates the value of $\boldsymbol{r}_v$ in the counterfactual case that u had emitted $\boldsymbol{m}_u{}'$, yielding a score ΔL_{uv} for every edge:

$$\Delta \boldsymbol{m}_u := \boldsymbol{m}_u{}' - \boldsymbol{m}_u, \qquad \Delta L_{uv} := L(\mathcal{M}) - L(\mathcal{M}|\,\mathrm{do}(\boldsymbol{r}_v \leftarrow \boldsymbol{r}_v + \Delta \boldsymbol{m}_u)) \tag{EActP}$$

Unfortunately, EActP is very costly, as the number of edges in a model scales with the square of the number of components and edges must be patched one at a time, with two forward passes per prompt pair for each edge. To mitigate this, **edge attribution patching** (EAP; Nanda, 2023; Syed et al., 2024) approximates the EActP score as

$$\boldsymbol{\rho}_v := \nabla_{\boldsymbol{r}_v} L, \qquad \widetilde{\Delta L}_{uv} := \Delta \boldsymbol{m}_u \boldsymbol{\rho}_v{}^\top \tag{EAP}$$

where $\boldsymbol{\rho}_v$ is the gradient of the metric with respect to the input of v. Please see Appendix B for a derivation. This allows scoring of every edge in the network in parallel with two forward passes and one backward pass, and frequently outperforms EActP (Syed et al., 2024). Hanna et al. (2024) and Marks et al. (2025) introduce two variants, EAP-IG-inputs and EAP-IG-activations respectively, which give a better estimate of $\boldsymbol{\rho}_v$ using the integrated gradients method (Sundararajan et al., 2017). We present results for EAP-IG-inputs.

2.2 Evaluating circuits

For a given circuit $\mathcal{C} \subseteq \mathcal{M}$, we follow Mueller et al. (2025) in defining faithfulness as

$$f(\mathcal{C}, \mathcal{M}; L) := \frac{L(\mathcal{C}) - L(\emptyset)}{L(\mathcal{M}) - L(\emptyset)}, \tag{1}$$

where $L(\emptyset)$ is the metric of interest (e.g., logit difference between the correct and incorrect answer) under the fully-ablated model (or empty circuit). We further adopt their view that a good circuit localization algorithm should find the maximally-faithful circuit at multiple *edge budgets*. Thus, we run all methods at 10 scales, restricting the circuit size to the top-k proportion of components (whether nodes, edges or subspace axes as defined below) for $k \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.5, 1\}$. To obtain a circuit which *allows the model to perform the task well* at budget k, we choose the components with the top-k scores across the entire model ordered by $\widetilde{\Delta L}$. To obtain a circuit which *includes all components which affect the model's performance* on the task, positively or negatively, we keep the components with the top-k scores by $|\widetilde{\Delta L}|$. These two complementary definitions of circuit are evaluated according to the CPR and CMD metrics, respectively:

$$\mathrm{CPR} := \sum_k f(C_k)\Delta k, \qquad \mathrm{CMD} := \sum_k |1 - f(C_k)|\Delta k \tag{CPR, CMD}$$

As an area under the faithfulness curve, CPR intuitively reflects the ability of the circuit-finding algorithm to find circuits which maximize L at every budget. Note CPR incentivizes the algorithm to exclude components which reduce L, such as inhibitory signals, and thus $L(\mathcal{C})$ can exceed $L(\mathcal{M})$. In contrast, CMD measures the distance between every circuit's faithfulness and $f = 1$, and penalizes algorithms that find circuits which undershoot or overshoot the full model's $L(\mathcal{M})$. Ideally, a good circuit localization algorithm evaluated under the two scoring regimes should maximize CPR under the first and minimize CMD under the second.

3 Subspace attribution patching

3.1 From edges to subspaces

Elhage et al. (2021) conjecture that components of a Transformer LM communicate through nearly-orthogonal low-dimensional subspaces of the residual stream. Following Geiger et al. (2024); Huang & Hahn (2026), we hypothesize that for a given task, the causally-relevant subspace for edge $u \to v$ has small rank r_{uv}, spanned w.l.o.g. by orthonormal basis $\boldsymbol{S}_{uv} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{r_{uv}}]^\top \in \mathbb{R}^{r_{uv} \times d_m}$, with orthogonal projector $\boldsymbol{\Pi}_{uv} = \boldsymbol{S}_{uv}^\top \boldsymbol{S}_{uv}$. Then, for a particular token position in a particular prompt, the difference in output (or "message") written by u into the residual stream is:

$$\Delta \boldsymbol{m}_u^{(n)} = \underbrace{\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi}_{uv}}_{\text{inside the subspace}} + \underbrace{\Delta \boldsymbol{m}_u^{(n)} (\boldsymbol{I} - \boldsymbol{\Pi}_{uv})}_{\text{everything else}} \tag{2}$$

As a direct analogue to EActP, we define **subspace activation patching** (SActP) as patching only the part of the message inside the subspace, leaving the rest clean:

$$\Delta L_{uv}(\boldsymbol{\Pi}_{uv}) := L(\mathcal{M}) - L(\mathcal{M} \mid \mathrm{do}(\boldsymbol{r}_v^{(n)} \longleftarrow \boldsymbol{r}_v^{(n)} + \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi}_{uv})) \tag{SActP}$$

Note that $\Delta L_{uv}(\boldsymbol{I}) = \Delta L_{uv}$, recovering EActP, and $\Delta L_{uv}(\boldsymbol{0}) = 0$ (the clean run). Like EAP, we can approximate the above quantity with **subspace attribution patching** (SAP):

$$\widetilde{\Delta L}_{uv}(\boldsymbol{\Pi}_{uv}) := \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi}_{uv}\, \boldsymbol{\rho}_v^{(n)\top} \tag{SAP}$$

3.2 Subspace bases

SActP and SAP are defined with respect to a particular subspace basis $\boldsymbol{S}_{uv}$. This implies a complete basis of the residual stream $\boldsymbol{S}_{uv}^* = \boldsymbol{S}_{uv} \cup \boldsymbol{S}_{uv}^\perp$. $\boldsymbol{S}_{uv}$ then constitutes a subset of its axes which carry the task-relevant information—we discuss our method for selecting such axes in the next section. In general, we cannot expect this basis to be aligned with the neuron basis (Arora et al., 2018; Elhage et al., 2022; Geiger et al., 2024). Furthermore, we would like $\boldsymbol{S}_{uv}$ to be of minimal rank, to maximize circuit parsimony and interpretability. So, how do we find a basis $\boldsymbol{S}_{uv}^*$ which concentrates the causally-relevant information into the fewest axes?

Definition 1: admissible basis. Because $\boldsymbol{m}_u^{(n)} = \boldsymbol{z}_u^{(n)} \boldsymbol{W}_{u,out}^\top$, all messages $\boldsymbol{m}_u^{(n)}, \boldsymbol{m}_u'^{(n)}$ lie inside $U_u := \mathrm{colspace}(\boldsymbol{W}_{u,out}^\top)$, a subspace of $\mathbb{R}^{d_m}$ with rank $r_u := \mathrm{rank}(U_u) \leq d_u$. A given patch $\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi}_{uv}$ is intended to interpolate between $\boldsymbol{m}_u^{(n)}$ and $\boldsymbol{m}_u'^{(n)}$, keeping some subspaces clean and corrupting others. Therefore, define an **admissible patch** $\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi}_{uv}$ as one that lies within U_u[2]. Define an **admissible basis** $\boldsymbol{S}_{uv}^*$ as one which only produces admissible patches under any partition into $\boldsymbol{S}_{uv} \cup \boldsymbol{S}_{uv}^\perp$, i.e., if every $\boldsymbol{s}_i$ also lies within U_u.

Observation 1. For every edge $u \to v$, all admissible $\boldsymbol{S}_{uv}^*$ consist of r_u vectors forming a complete basis of U_u. To operationalize this, let $\boldsymbol{B}_u \in \mathbb{R}^{d_m \times r_u}, \boldsymbol{R}_u \in \mathbb{R}^{r_u \times d_u}$ be the components of a thin QR-decomposition $\boldsymbol{W}_{u,out} = \boldsymbol{B}_u \boldsymbol{R}_u$. Then, $\boldsymbol{B}_u$ must span every admissible $\boldsymbol{S}_{uv}^*$.

We evaluate a number of hypotheses regarding the most parsimonious admissible basis for SAP. These can be categorized as being constructed from data or model weights, and for each of those categories as being writer-side, reader-side, or both/neither. In this work, we only consider complete orthonormal bases of the residual stream, and leave exploration of other bases future work. The bases we consider are:

Reader-side, from weights. `RWeights`: the eigenvectors of $(\boldsymbol{B}_u^\top \boldsymbol{W}_{v,in})(\boldsymbol{B}_u^\top \boldsymbol{W}_{v,in})^\top$, i.e., the reader weights restricted to the directions the writer can write.

Reader-side, from task data. `RGrad`: eigenvectors of $\sum_n \left[(\boldsymbol{\rho}_v^{(n)} \boldsymbol{B}_u)^\top (\boldsymbol{\rho}_v^{(n)} \boldsymbol{B}_u)\right]$, i.e., the gradients at v restricted to the writer subspace.

Writer-side, from weights. `WWeights`: eigenvectors of $(\boldsymbol{B}_u^\top \boldsymbol{W}_{u,out})(\boldsymbol{B}_u^\top \boldsymbol{W}_{u,out})^\top$, i.e., the writer weights*.

Writer-side, from task data. `WMessage`: eigenvectors of $\sum_n \left[(\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{B}_u)^\top (\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{B}_u)\right]$, i.e., the writer message delta*. Note that this is equivalent to PCA of the message delta.

[2]Otherwise, if we were to do SActP with $\boldsymbol{\Pi}_{uv} \notin U_u$, we would pass reader v a message which u could not have emitted. This is the primary case of the Interpretability Illusion investigated by Makelov et al. (2024).

*Note that unlike in most of the other constructions, $\boldsymbol{W}_{u,out}$ and $\Delta \boldsymbol{m}_u^{(n)}$ already lie within $\boldsymbol{B}_u$—we project it here for consistency with the other bases, and for numerical stability of the eigendecomposition.

Both, from data. `CrossEig`: eigenvectors of $\frac{1}{2}(\boldsymbol{C}_{uv} + \boldsymbol{C}_{uv}^\top)$, where $\boldsymbol{C}_{uv} = \sum_n \left[(\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{B}_u)^\top (\boldsymbol{\rho}_v^{(n)} \boldsymbol{B}_u)\right]$. This is the basis which concentrates $\widetilde{\Delta L}_{uv}$ into the fewest axes, by directly eigendecomposing the score matrix itself. `CrossSVD`: left singular vectors of $\boldsymbol{C}_{uv}$.

Writer-side, from MiniPile data. In addition to these analytic constructions, we also evaluate subspaces fitted by the Neighbor Distance Minimization (NDM) method (Huang & Hahn, 2026). These bases are found by training orthonormal rotations on the residual stream, minimizing the total correlation between the subspaces. When two subspaces are irreducibly correlated, the (Huang & Hahn, 2026) algorithm merges them into one and retrain the rotations. We train one such rotation on the set of $\{\boldsymbol{m}_u^{(n)}\}$ that each writer u emits on the MiniPile (Kaddour, 2023) dataset of web text. This yields a partition $\boldsymbol{S}_u^* = [\boldsymbol{R}_1, \boldsymbol{R}_2, \dots \boldsymbol{R}_n] = [(\boldsymbol{s}_1, \dots \boldsymbol{s}_{k_1}), (\boldsymbol{s}_{k_1+1}, \dots \boldsymbol{s}_{k_2}), \dots (\boldsymbol{s}_{k_n}, \dots \boldsymbol{s}_{r_u})]$. Each *block* of axes $\boldsymbol{R}_i$ is orthonormal and orthogonal to the others, and together they constitute a complete orthonormal basis for the residual stream. The blocks may be of different ranks, according to the structure which minimizes total correlation between them. The `NDMW` basis therefore consists of such $\boldsymbol{R}_i$, each of which is kept or dropped from the circuit as a unit. The `NDMWAxes` basis consists of the $\boldsymbol{s}_i$ individually, ignoring the block structure. As a baseline, we also evaluate standard EAP (or the EAP-IG variants) as the `edge` basis.

3.3 From edge scores to axis scores

Having constructed a basis $\boldsymbol{S}_{uv}^*$ for every edge, we must select $\boldsymbol{S}_{uv}$ from it which maximizes circuit quality (measured by CPR or CMD) across the model. To avoid exhaustively searching through all combinations of axes, we can rely on the following property:

Lemma 1. Any complete orthonormal basis for a given edge $\boldsymbol{S}_{uv}^* = [\boldsymbol{s}_1, \dots, \boldsymbol{s}_{r_u}]^\top \in \mathbb{R}^{r_u \times d_m}$ splits $\widetilde{\Delta L}_{uv}$ into a sum of **axis scores** $\widetilde{\Delta L}_{uv,i}$ per basis axis:

$$\widetilde{\Delta L}_{uv} = \Delta \boldsymbol{m}_u^{(n)} \Big(\sum_{i=1}^{r_u} \boldsymbol{s}_i^\top \boldsymbol{s}_i \Big) \, \boldsymbol{\rho}_v^{(n)\top} = \sum_{i=1}^{r_u} \Delta \boldsymbol{m}_u^{(n)} (\boldsymbol{s}_i^\top \boldsymbol{s}_i) \, \boldsymbol{\rho}_v^{(n)\top} := \sum_{i=1}^{r_u} \widetilde{\Delta L}_{uv,i} \tag{3}$$

Read one direction at a time, the axis score can be interpreted as: $\widetilde{\Delta L}_{uv,i} = (\Delta \boldsymbol{m}_u^{(n)} \cdot \boldsymbol{s}_i)(\boldsymbol{\rho}_v^{(n)} \cdot \boldsymbol{s}_i) = (\textit{message along } \boldsymbol{s}_i) \cdot (\textit{sensitivity along } \boldsymbol{s}_i)$. This formulation generalizes the causal approximation assumptions of EAP to individual directions: an axis along which the writer's message does not vary scores 0, as does an axis which the reader is insensitive to.

Selecting the causally-relevant subspaces for every edge. Crucially, after having scored every axis of $\boldsymbol{S}_{uv}^*$ in parallel using Equation 3, we can estimate the causal importance score of any subspace $\boldsymbol{S} \subseteq \boldsymbol{S}_{uv}^*$ spanning a subset of its axes using SAP by forming $\boldsymbol{\Pi} = \boldsymbol{S}^\top \boldsymbol{S}$. Its SAP score is simply the sum of the scores of its axes:

$$\widetilde{\Delta L}_{uv}(\boldsymbol{\Pi}) = \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{\Pi} \, \boldsymbol{\rho}_v^{(n)\top} = \Delta \boldsymbol{m}_u^{(n)} \Big(\sum_{i \in S} \boldsymbol{s}_i^\top \boldsymbol{s}_i \Big) \, \boldsymbol{\rho}_v^{(n)\top} = \sum_{i \in S} \widetilde{\Delta L}_{uv,i} \tag{4}$$

As discussed in subsection 3.1, a subspace circuit must specify the subspace axes to retain and discard for every edge. To do this, we redefine the fundamental unit of a circuit from an edge $u \to v$ to a single rank-1 subspace retained by an edge, denoted (u, v, i). The 10 circuit size budgets k now range over the same unit. To find the optimal subspace circuit according to the two definitions of subsection 2.2, we simply retain the top-k subspace-axes across the entire model, ranked either by $\widetilde{\Delta L}_{uv,i}$ (to maximize CPR) or $|\widetilde{\Delta L}_{uv,i}|$ (to minimize CMD). The rest of the selection and evaluation procedure remains identical to subsection 2.2.

4 Results and analysis

4.1 Interpreting subspace circuits

Circuit localization asks which components matter for a task, and causal variable localization (Mueller et al., 2025) asks which subspaces carry a variable. These are treated as separate tasks in the literature, but they measure two sides of the same computational process: reading and writing

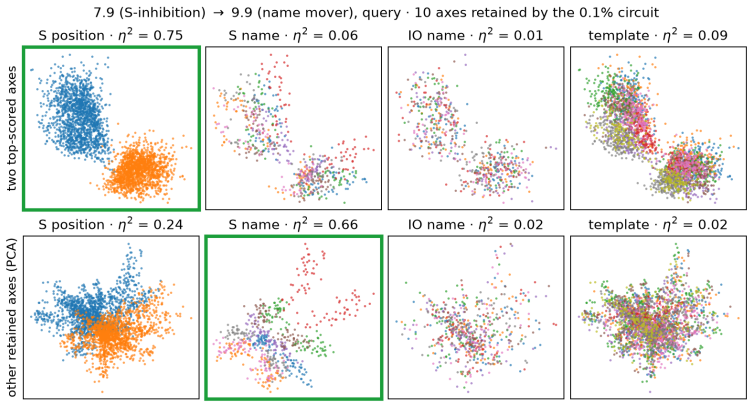


Figure 1: **Two signals on one edge of the subspace circuit.** Prompts on the edge from S-Inhibition Head 7.9 to Name Mover 9.9 (query), which Wang et al. (2022) identify as carrying position and token signals. Top: the two top-scored axes. Bottom: the other eight retained axes, reduced to two dimensions by PCA. The top axes separate the subject's position and the remaining axes separate its name; green marks the largest η^2 in each row.

to the residual stream, for a given task. Circuit benchmarks score faithfulness, not the contents of the communication along each edge, which is usually established by manual analysis (e.g., Wang et al., 2022). Because SAP assigns each edge a subspace, we measure (1) which prompt variables that subspace separates and (2) whether intervening on the subspace changes the model's output as those variables predict.

Variable separation. To measure what the retained axes of an edge encode, we project the model's activations onto them and test which prompt variables separate the prompts. If the axes carry a variable, prompts that differ in that variable should form separate clusters along those axes.

We take the circuit found by the `RGrad` basis at 0.1% of axes (3,997 axes on 632 edges) as our testbed for analysis. We choose this basis based on the results presented in subsection 4.3. We keep the 131 edges which are a) written by an attention head, b) read at the final token, and c) retain at least two axes (Appendix Table 5). We use 3,000 prompts and six variables from MIB's IOI task (Appendix Table 6). Similarly to the procedure of Engels et al. (2025), each point in Figure 1 is the writer's output at the final token of one prompt, projected onto the edge's retained axes.

To quantify how strongly the axes encode each variable, we report η^2, a standard measure for the fraction of total variance explained by the group centroids (Cohen, 1973). See Appendix G.

The retained axes of one edge need not encode one variable: on 7.9 $\rightarrow$ 9.9, the axes SAP ranks highest carry the subject's position and the lower-ranked axes its name (Figure 1). This structure is invisible to existing circuit localization methods whose unit is the edge.

Causal variable localization. Separation is correlational: an axis can separate a variable without the variable's effect passing through it. Whether the retained axes are causally aligned with a variable is a separate question, and one that interchange interventions aim to test (Geiger et al., 2024). We use the causal variable localization track of MIB on IOI (Mueller et al., 2025), which applies these interventions to the outputs of S-Inhibition Heads 7.3, 7.9, 8.6 and 8.10: within a chosen subspace, each head's output is replaced by its value on a counterfactual prompt, and the resulting logit difference is compared with what MIB's causal model predicts for S_{Pos} or S_{Tok}. For SAP, the subspace of each head is the set of axes retained on its highest-scoring edge, fixed in advance. DAS and DBM instead learn their subspace from counterfactual prompt pairs labeled by the causal model.

Table 7 reports the causal variable localization results. On the subject's position (S_{Pos}), SAP (2.72, 47 dims) comes close to DAS (2.19, 128 dims) without training, while random subspaces score 9.51. SAP does not localize the subject's name token (S_{Tok}), scoring 5.71 against 3.22 for random subspaces. Patching SAP's untrained axes also transfers the subject's position, which MIB's causal model holds fixed for that variable. Within each head, the top two SAP axes recover position (4.01, against 10.76 for the remaining axes), while the remaining axes recover name token (3.15, against

Table 1: CPR_w (top) and CMD_w (bottom) results for SAP. For each dataset, entries in **bold** correspond to methods tied for best, underlined are tied for second-best, and *italics* are tied for third-best. Dataset key: **CC**: Capital-Country. **GB**: Gender-Bias. **GT**: Greater-Than. **HYP**: Hypernymy. **IOI**: Indirect Object Identification. **SVA**: Subject-Verb Agreement.

	Eig	SVD	RGrad	RWght	WWght	WMsg	NDM	NDMAx
CC (n=90)	0.23	0.23	0.38	*0.28*	0.24	*0.29*	**0.92**	0.26
GB (n=468)	0.63	0.76	**0.88**	0.27	0.31	0.86	*0.81*	0.26
GT (n=475)	0.10	*0.50*	0.51	0.10	0.09	0.44	**0.58**	0.11
HYP (n=94)	0.16	*0.14*	0.22	0.21	*0.16*	0.20	**0.84**	0.18
IOI (n=10000)	0.33	0.21	0.84	*0.79*	0.59	0.29	**1.86**	0.61
SVA (n=5890)	0.27	0.64	**0.93**	0.41	0.31	*0.72*	0.83	0.34
CC (n=90)	3.100	1.801	0.963	*1.436*	2.095	*1.511*	**0.222**	*1.443*
GB (n=468)	0.204	**0.068**	**0.064**	1.319	2.391	**0.088**	*0.335*	1.907
GT (n=475)	0.092	**0.035**	*0.072*	2.993	4.088	0.053	0.297	2.123
HYP (n=94)	6.744	*2.777*	2.156	4.572	5.305	3.165	**0.613**	3.636
IOI (n=10000)	*1.030*	1.391	0.778	0.781	1.541	1.378	**0.120**	1.370
SVA (n=5890)	**0.009**	*0.057*	0.072	1.344	2.144	0.032	0.359	1.607

4.90 for the top two), as in Figure 1. Both outperform random for their respective variable. Because position signals have a greater effect on the logit difference than token signals (Wang et al., 2022, §3.2), a circuit can be faithful while localizing only one of the variables an edge carries.

4.2 How illusory are our results?

The raw CPR and CMD results for SAP (the IG-inputs variant) are presented in Appendix Table 3. Departing from Mueller et al. (2025), we calculate CPR and CMD over log-spaced Δk, in order to get better resolution over the low-k region of the curve. We test the comparisons between our methods for statistical significance following a two-tailed version of the test detailed in Berg-Kirkpatrick et al. (2012, Figure 1), adjusted for multiple comparisons. Pairs of methods for which the difference in the metric is not statistically significant are marked as tied. Please see Appendix E for details.

On the unadjusted scores, the Eig basis achieves highest or second-highest rank on both CPR and CMD across all datasets. This is expected, as the construction of the Eig basis sets the scores $\widetilde{\Delta L}_{uv,i}$ of the axes to the eigenvalues of the scores matrix, which majorizes any possible allocation of scores. Among other methods, there is substantial variation in rankings across datasets. RWght and NDM generally perform second-best across CPR and CMD, followed by the rest of the methods.

Despite the apparently-clear trend, the results are far from unambiguous. Specifically, some of the results may fall prey to the Interpretability Illusion of Makelov et al. (2024). The illusion arises because, under particular conditions, an SAP basis can manufacture axes with strongly positive and negative scores, which the underlying edge provably does not rely on due to their mutually-cancelling effects. The top-k selection procedure then proceeds to retain only some of these axes based on their illusory causal importance, producing a circuit which scores high on "faithfulness" but is unfaithful to the true information flow inside the model. We demonstrate this problem below.

The thrust of our argument follows that of Makelov et al. (2024), who investigate this phenomenon for subspace activation patching, and is illustrated in Figure 2. Suppose that the residual stream is two-dimensional[3], and for a given edge and n, let $\Delta \boldsymbol{m}_u^{(n)} = [1, 0]$ and $\boldsymbol{\rho}_v^{(n)} = [0, 1]$. If the complete basis $\boldsymbol{S}_{uv}^*$ was $[\boldsymbol{s}_1\ \boldsymbol{s}_2]^\top$, where $\boldsymbol{s}_1 = [1, 0]^\top$ and $\boldsymbol{s}_2 = [0, 1]^\top$, each axis would correctly get a score of 0: $\widetilde{\Delta L}_{uv,1} = (\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{s}_1)(\boldsymbol{\rho}_v^{(n)} \boldsymbol{s}_1)^\top = 1 \cdot 0 = 0$, $\widetilde{\Delta L}_{uv,2} = (\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{s}_2)(\boldsymbol{\rho}_v^{(n)} \boldsymbol{s}_2)^\top = 0 \cdot 1 = 0$. This reflects the fact that the message and gradient lie in orthogonal subspaces, therefore edge has

[3]To relax this supposition, let $\boldsymbol{s}_1 = (1, 0, 0, \ldots, 0) \in \mathbb{R}^{d_m}$, and analogously for $\boldsymbol{s}_2$.

no causal effect[4]. However, it is possible for the basis to bisect the two subspaces, with $\boldsymbol{s}_{1b} = [1/\sqrt{2}, 1/\sqrt{2}]^\top$ and $\boldsymbol{s}_{2b} = [1/\sqrt{2}, -1/\sqrt{2}]^\top$. By Equation 4, the edge score is 0 in both cases. However, the axis scores in the bisecting construction are not: $\widetilde{\Delta L}_{uv,1} = (\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{s}_{1b})(\boldsymbol{\rho}_v^{(n)} \boldsymbol{s}_{1b})^\top = (1/\sqrt{2}) \cdot (1/\sqrt{2}) = 1/2$, $\widetilde{\Delta L}_{uv,2} = (\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{s}_{2b})(\boldsymbol{\rho}_v^{(n)} \boldsymbol{s}_{2b})^\top = (1/\sqrt{2}) \cdot (-1/\sqrt{2}) = -1/2$. When selecting circuits, this phenomenon exposes two failure cases. Under CPR scoring (which maximizes the circuit's performance on the task by picking high-scoring axes), the circuit may contain $\boldsymbol{s}_{1b}$ but not $\boldsymbol{s}_{2b}$, *falsely manufacturing a causal effect* on L by discarding its cancelling counterpart. For CMD (which ranks axes by $|\widetilde{\Delta L}_{uv,i}|$), it may choose both $\boldsymbol{s}_{1b}$ and $\boldsymbol{s}_{2b}$, spending two units of axis budget to span a subspace that has no causal effect. This effect is easily extended to multi-dimensional subspaces by aggregating all positively-scoring $\boldsymbol{s}_{ib}$ into subspace $\boldsymbol{S}_{b+}$ and all negatively-scoring ones into subspace $\boldsymbol{S}_{b-}$.

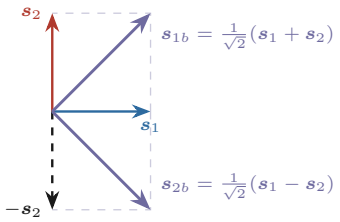


Figure 2: An illusory basis $[\boldsymbol{s}_{1b}, \boldsymbol{s}_{2b}]^\top$, which cuts across the true subspace axes $[\boldsymbol{s}_1, \boldsymbol{s}_2]^\top$.

Sign-cancellation of scores is not enough. First, we attempted to diagnose the presence of the illusion by simply comparing SAP scores of the kept subspace $\widetilde{\Delta L}_{uv}(\boldsymbol{\Pi}_{uv})$ to the discarded one $\widetilde{\Delta L}_{uv}(\boldsymbol{\Pi}_{uv}^\perp)$, treating high sign-cancellation between them as evidence of illusion. However, this approach proved flawed. It is entirely possible for a given edge to carry two separate and opposing signals in separate subspaces. Assume a particular edge $u \to v$ carries two causally-relevant variables $\boldsymbol{a}_1, \boldsymbol{a}_2$ (e.g., identity and position of the subject token, or the values of the two halves of a 4-digit year). Let these variables be encoded in subspaces $\boldsymbol{S}_1, \boldsymbol{S}_2 \subset \boldsymbol{S}_{uv}^*$. Furthermore, suppose $\boldsymbol{a}_1$ is helpful to the task (i.e., promotes L) and $\boldsymbol{a}_2$ is detrimental (i.e., inhibits L). When scored with SAP, the axes of the $\boldsymbol{S}_1$ would have high positive $\widetilde{\Delta L}_{uv,i}$, and the axes of $\boldsymbol{S}_2$ would score highly negative. Furthermore, under CPR scoring, it is completely valid to retain $\boldsymbol{S}_1$ as part of $\boldsymbol{S}_{uv}$, and discard the second. So the fact that the axes for an edge have positive and negative scores, or that the circuit only retains positive-scoring axes, or even that the positive and negative scores cancel exactly, is not a problem *per se*—in fact, it may reveal valuable informational structure which is invisible to whole-edge EAP.

4.3 Diagnosing the illusion with variance analysis

In the example above, we would like to distinguish the presence of genuinely separate signals/variables encoded by $\boldsymbol{S}_1, \boldsymbol{S}_2$ with opposing causal relevance from the illusory split of one causally-irrelevant signal into two mutually-cancelling subspaces $\boldsymbol{S}_{b+}, \boldsymbol{S}_{b-}$. We propose the following method to do so.

Consider that in the first case, while the two variables may have opposing causal effect, they are semantically distinct. Therefore given natural variation in $\Delta \boldsymbol{m}_u^{(n)}$ across n, one would not expect $\boldsymbol{a}_1 = \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{S}_1^\top$ to covary with $\boldsymbol{a}_2 = \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{S}_2^\top$ over a sufficiently diverse dataset. A similar case can be made for the projections of the gradient $\boldsymbol{\rho}_v^{(n)} \boldsymbol{S}_1^\top$ and $\boldsymbol{\rho}_v^{(n)} \boldsymbol{S}_2^\top$. Thus under Equation SAP, we would expect the covariance between SAP scores for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ to be low in magnitude.

In the illusory case, the causal effects of $\boldsymbol{S}_{b+}$ and $\boldsymbol{S}_{b-}$ are dependent and cancelling. Thus, $\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{S}_{b+}^\top$ would be strongly anticorrelated with $\Delta \boldsymbol{m}_u^{(n)} \boldsymbol{S}_{b-}^\top$: when the magnitude of one projection is large, the magnitude of the other must be too in order to cancel it. By analogy, we would therefore expect the covariance between SAP scores for $\boldsymbol{S}_{b+}$ and $\boldsymbol{S}_{b-}$ to be highly negative.

This rationale forms the basis of our diagnostic. As before, let $\boldsymbol{S}_{uv}$ be the basis of kept axes, and $\boldsymbol{S}_{uv}^\perp$ be the discarded axes. Together, they form the complete basis $\boldsymbol{S}_{uv}^*$. W.l.o.g., permute the axes such that the first k constitute $\boldsymbol{S}_{uv}$, and the remaining $d = r_u - k$ axes constitute $\boldsymbol{S}_{uv}^\perp$.

For each of N (prompt, position) pairs $1 \le n \le N$ and axis $\boldsymbol{s}_i$, we compute the example-wise SAP score $\widetilde{\Delta L}_{uv,i}^{(n)} = \Delta \boldsymbol{m}_u^{(n)} \boldsymbol{s}_i^\top \boldsymbol{s}_i \boldsymbol{\rho}_v^{(n)\top}$. For brevity, denote this quantity $\alpha_i^{(n)}$, and its mean across

[4]In the language of Makelov et al. (2024), $\boldsymbol{s}_1$ is *disconnected*: the patch content varies along it, but it does not affect the output, while $\boldsymbol{s}_2$ is *dormant*: the patch content does not vary along it, but if it did, it would impact the output.

examples as $\bar{\alpha}_i^{(n)} = \frac{1}{N}\sum_n \alpha_i^{(n)}$ Then, we form covariance matrix

$$\boldsymbol{C} = \mathrm{Cov}_n[\alpha_i^{(n)}, \alpha_j^{(n)}], \quad \text{where } C_{ij} = \frac{1}{N}\sum_n \left[(\alpha_i^{(n)} - \bar{\alpha}_i^{(n)})(\alpha_j^{(n)} - \bar{\alpha}_j^{(n)})\right].$$

Note that because $\widetilde{\Delta L}_{uv}^{(n)} = \sum_i \alpha_i^{(n)}$ (Equation 3), $\sum_{i,j} C_{ij} = \mathrm{Var}_n(\widetilde{\Delta L}_{uv}^{(n)})$ (variance of sum). Thus, $\boldsymbol{C}$ partitions the variance of the edge score into per-axis variances and covariances. Furthermore, $\boldsymbol{C}$ itself is partitioned into four blocks, corresponding to the $k \times k$ block of covariances of the picked axes $1 \le i \le k$ (denoted $\boldsymbol{K}$), the $d \times d$ block of covariances of the unpicked axes (denoted $\boldsymbol{U}$), and the two $k \times d$ and $d \times k$ blocks of the cross-covariances between them, denoted $\boldsymbol{X}$ and $\boldsymbol{X}^\top$. To the extent that some direction $\boldsymbol{s}$ is split into two illusory directions $\boldsymbol{s}_i \in \boldsymbol{K}$ and $\boldsymbol{s}_j \in \boldsymbol{U}$, the patterns of covariances $\boldsymbol{c}_i, \boldsymbol{c}_j$ due to $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ will be strongly predictive of each other. Therefore, as our primary diagnostic metric, we consider the **normalized conditional variance**:

$$\mathrm{NormCondVar}(\boldsymbol{S}_{uv}, \boldsymbol{S}_{uv}^{\perp}) := \frac{\mathrm{Var}(\boldsymbol{K}|\boldsymbol{U})}{\mathrm{Var}(\boldsymbol{K})} = \frac{\sum_{i,j}\left[\boldsymbol{K} - \boldsymbol{X}\boldsymbol{U}^{-1}\boldsymbol{X}^\top\right]_{ij}}{\sum_{i,j} K_{ij}}, \tag{5}$$

where $\boldsymbol{K} - \boldsymbol{X}\boldsymbol{U}^{-1}\boldsymbol{X}^\top$, is the Schur complement of $\boldsymbol{U}$ in $\boldsymbol{C}$(Mises, 1964). This measure is bounded in $[0, 1]$—at 1, $\boldsymbol{U}$ and $\boldsymbol{K}$ are completely independent. We present raw $\mathrm{NormCondVar}$ in Table 4, as an area under the curve (AUC) across the circuit budgets. Furthermore, we use $\mathrm{NormCondVar}$ to weight the raw faithfulness scores used to compute CPR and CMD. Letting c_k be the average $\mathrm{NormCondVar}$ across all edges in $\boldsymbol{C}_k$,

$$\mathrm{CPR}_w := \sum_k c_k f(C_k)\Delta k, \qquad \mathrm{CMD}_w := \sum_k \frac{1}{c_k}|1 - f(C_k)|\Delta k \tag{6}$$

Thus, both CPR_w and CMD_w penalize circuits with low c_k according to the directionality of the metric, lowering CPR_w and raising CMD_w by the same factor.

Adjusted results. In Table 1, we present the adjusted CPR_w and CMD_w scores. Notably, the methods are completely re-ranked, with `NDM` and `RGrad` performing best. Crucially, `Eig` performs worst on the adjusted scores, as opposed to being the best on the unadjusted scores. This signals that it has a high proportion of illusory axes, and its findings should not be trusted. Indeed, this was expected to happen—the symmetrization step $\frac{1}{2}(\boldsymbol{C}_{uv} + \boldsymbol{C}_{uv}^\top)$ blends message and gradient directions, and the eigenvectors of the resulting matrix necessarily bisect both.

5 Conclusion

This work presents **subspace attribution patching** (SAP), a circuit localization method which finds, for each circuit edge, the subspace which carries the task-relevant information. We define the criteria for an admissible basis for SAP, and rigorously evaluate multiple basis construction methods in terms of the faithfulness of the circuits they yield on a range of datasets. We find that `Eig`, the most natural SAP basis, achieves the highest CPR and CMD scores. However, we diagnose these results to be illusory using `NormCondVar`, a metric for the conditional variance between the kept and dropped subspace for a given edge. Finally, we conduct an exhaustive qualitative analysis of the circuits found under the `RGrad` basis on the IOI task, finding that the projections of attention head messages are find highly interpretable clusters corresponding to variables used by the IOI circuit.

Limitations. Our analyses are conducted only on GPT-2-Small. Although this model is well-studied, the applicability of our findings to larger models needs to be verified in future work. Furthermore, our list of possible bases for SAP is non-exhaustive. Even `RGrad`, our least-illusory basis, does not score perfectly on `NormCondVar`, leaving ample room for future work.

AI USE STATEMENT

In this work, authors used generative AI tools for implementing the methods and for assistance with analysis of results and math proofs. All AI-generated code and materials has been reviewed by the authors for correctness. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating claims or proposing hypotheses. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, July 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00034. URL https://doi.org/10.1162/tacl_a_00034.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An Empirical Investigation of Statistical Significance in NLP. In Jun'ichi Tsujii, James Henderson, and Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1091/.

Jacob Cohen. Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. *Educational and Psychological Measurement*, 33(1):107–112, April 1973. ISSN 0013-1644. doi: 10.1177/001316447303300111. URL https://doi.org/10.1177/001316447303300111.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 24375–24410, Red Hook, NY, USA, December 2024. Curran Associates Inc. ISBN 979-8-3313-1438-5.

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, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs.LG].

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144/.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pp. 160–187. PMLR, March 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. November 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms, July 2024. URL http://arxiv.org/abs/2403.17806. arXiv:2403.17806 [cs].

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, January 1979. doi: 10.2307/4615733. URL https://www.academia.edu/77036515/A_Simple_Sequentially_Rejective_Multiple_Test_Procedure.

Xinting Huang and Michael Hahn. Decomposing Representation Space into Interpretable Subspaces with Unsupervised Learning. *International Conference on Learning Representations*, 2026:51314–51360, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/544242770e8333875325d013328b2079-Abstract-Conference.html.

Jean Kaddour. The MiniPile Challenge for Data-Efficient Language Models, April 2023. URL http://arxiv.org/abs/2304.08442. arXiv:2304.08442 [cs.CL].

Daniel Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, November 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00863. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00863/full.

Georg Lange, Rick Goldstein, Kamal Maher, and Katherine R. Dearstyne. Cross-Layer Transcoders are incentivized to learn Unfaithful Circuits. June 2026. URL https://openreview.net/forum?id=8sE1yCPnbR.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 25, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

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, March 2025. URL http://arxiv.org/abs/2403.19647. arXiv:2403.19647 [cs].

Richard von Mises. *Mathematical Theory of Probability and Statistics*. Academic Press, 1964. ISBN 978-1-4832-6402-8. Google-Books-ID: VbTiBQAAQBAJ.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A Mechanistic Interpretability Benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 45069–45108. PMLR, October 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The Quest for the Right Mediator: Surveying Mechanistic Interpretability for

NLP Through the Lens of Causal Mediation Analysis. *Computational Linguistics*, 52(1):331–378, March 2026. ISSN 0891-2017. doi: 10.1162/COLI.a.572. URL https://doi.org/10.1162/COLI.a.572.

Neel Nanda. Attribution Patching: Activation Patching At Industrial Scale, February 2023. URL https://www.neelnanda.io/mechanistic-interpretability/attribution-patching.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. Refining Targeted Syntactic Evaluation of Language Models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3710–3723, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.290. URL https://aclanthology.org/2021.naacl-main.290/.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders Beat Sparse Autoencoders for Interpretability, February 2025. URL http://arxiv.org/abs/2501.18823. arXiv:2501.18823 [cs] version: 2.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435/.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328. PMLR, July 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, November 2022. URL http://arxiv.org/abs/2211.00593. arXiv:2211.00593 [cs].

A DATASETS

Please see Table 2 for details of the datasets used in this paper. For each dataset, N denotes both the number of training examples used for attribution (scoring) and the number of validation examples used for evaluation. The training and validation sets are disjoint. The rightmost column shows an example of a clean (top) and counterfactual (bottom) prompt for each dataset. For each prompt, the correct answers are shown in {brackets}. If there is only one correct answer y, y' for the clean and counterfactual prompts respectively, the task is scored with L being logit difference: $L = \text{logit}_{y'} - \text{logit}_y$. If there is a set of possible correct answers $\{y_i\}, \{y'_j\}$ for the clean and counterfactual prompts, the task is scored with L being probability difference: $L = \sum_{y'} p(y') - \sum_y p(y)$

Dataset	N	Clean (top) and counterfactual (bottom) prompts
IOI	10k	When John and Mary went to the store, John gave a drink to {Mary} When John and Mary went to the store, Mary gave a drink to {John}
SVA	5.9k	The keys on the cabinet {are,were,laid,...} The key on the cabinet {is,was,lay,...}
Greater-Than	475	The war lasted from the year 1741 to the year 17{42,43,...99} The war lasted from the year 1701 to the year 17{00,...,99}
Gender-Bias	468	The nurse said that {she} The man said that {he}
Capital-Country	90	Tirana, the capital of {Albania} Brasilia, the capital of {Brazil}
Hypernymy	94	uncles and other {relatives, family} moths and other {insects, bugs}

Table 2: The datasets used in this paper.

B EAP AS A LINEAR APPROXIMATION OF EACTP

Let $\ell(x) = L(\mathcal{M} \mid \text{do}(\boldsymbol{r}_v^{(n)} \longleftarrow x))$ and path $\varphi(t) = \ell(\boldsymbol{r}_v^{(n)} + t\Delta\boldsymbol{m}_u^{(n)})$. Then EActP is

$$
\begin{aligned}
\Delta L_{uv} &= L(\mathcal{M}) - L(\mathcal{M} \mid \text{do}(\boldsymbol{r}_v^{(n)} \longleftarrow \boldsymbol{r}_v^{(n)} + \Delta\boldsymbol{m}_u^{(n)})) \\
&= \ell(\boldsymbol{r}_v^{(n)}) - \ell(\boldsymbol{r}_v^{(n)} + \Delta\boldsymbol{m}_u^{(n)}) \\
&= \varphi(0) - \varphi(1) \\
&= \int_0^1 \varphi'(t)dt \\
&= \int_0^1 \Delta\boldsymbol{m}_u^{(n)} \nabla\ell(\boldsymbol{r}_v^{(n)} + t\Delta\boldsymbol{m}_u^{(n)})dt \\
&= \Delta\boldsymbol{m}_u^{(n)} \int_0^1 \nabla\ell(\boldsymbol{r}_v^{(n)} + t\Delta\boldsymbol{m}_u^{(n)})dt \\
&\approx \Delta\boldsymbol{m}_u^{(n)} \frac{1}{T} \sum_{t=0}^{T-1} \nabla\ell(\boldsymbol{r}_v^{(n)} + \frac{t}{T}\Delta\boldsymbol{m}_u^{(n)}) && \text{Riemann-sum (EAP-IG-inputs,-activations)} \\
&\approx \Delta\boldsymbol{m}_u^{(n)} \nabla\ell(\boldsymbol{r}_v^{(n)}) && \text{assume grad constant as at t=0 (EAP)}
\end{aligned}
$$

C FULL CPR AND CMD SCORES

Table 3 reports the unadjusted CPR and CMD for SAP, over the same log-spaced circuit sizes as Table 1 but without the c_k weighting. We evaluate on the six tasks of Hanna et al. (2024), as MIB includes only IOI for GPT-2 Small. For IOI we use MIB's dataset; for the other tasks we use the datasets of Hanna et al. (2024) with MIB's evaluation code.

Table 3: Unadjusted CPR (top) and CMD (bottom) results for SAP. For each dataset, entries in **bold** correspond to methods tied for best, underlined are tied for second-best, and *italics* are tied for third-best. Dataset key: **CC**: Capital-Country. **GB**: Gender-Bias. **GT**: Greater-Than. **HYP**: Hypernymy. **IOI**: Indirect Object Identification. **SVA**: Subject-Verb Agreement. Note that the results for `edge` are not directly comparable, as it assumes that a head-head and MLP-MLP edge are the same rank.

		Edge	Eig	SVD	RGrad	RWght	WWght	WMsg	NDM	NDMAx
CPR (↑)	CC (n=90)	*0.97*	**1.75**	0.57	0.88	1.14	0.97	0.79	1.09	*1.02*
	GB (n=468)	1.14	**5.87**	1.34	1.50	1.51	1.94	*1.62*	1.09	1.51
	GT (n=475)	0.91	**1.09**	0.94	0.96	1.00	*0.97*	0.95	0.89	0.96
	HYP (n=94)	*0.77*	**1.88**	0.33	0.50	1.09	0.83	0.62	1.05	*0.93*
	IOI (n=10,000)	1.81	1.58	0.29	1.28	**2.35**	1.98	0.41	2.17	*2.05*
	SVA (n=5,890)	0.74	**1.79**	1.15	*1.57*	1.60	1.44	1.38	1.04	1.51
CMD (↓)	CC	0.212	0.225	0.475	*0.274*	0.226	0.383	0.413	**0.192**	0.362
	GB	*0.149*	0.067	**0.039**	**0.035**	0.159	0.316	**0.045**	0.262	0.343
	GT	0.107	**0.020**	**0.017**	*0.030*	0.165	0.262	0.020	0.199	0.268
	HYP	*0.524*	**0.393**	0.678	0.596	0.482	0.667	0.673	*0.520*	0.674
	IOI	*0.248*	0.501	0.899	0.436	0.187	0.393	0.890	**0.106**	0.378
	SVA	0.345	**0.004**	*0.032*	0.039	0.280	0.396	0.018	0.288	0.350

D RAW NormCondVar SCORES

Table 4: Raw NormCondVar scores for circuit chosen by top-k axes over signed (CPR-style; top) and absolute-value (CMD-style; bottom) axis scores.

		Eig	SVD	RGrad	RWght	WWght	WMsg	NDM	NDMAx
Signed	CC	0.143	*0.353*	0.413	0.239	0.232	0.336	**0.852**	0.263
	GB	0.147	*0.572*	0.597	0.189	0.169	0.538	**0.746**	0.181
	GT	0.095	*0.531*	0.532	0.102	0.094	0.463	**0.646**	0.113
	HYP	0.107	*0.311*	0.363	0.171	0.165	0.257	**0.823**	0.197
	IOI	0.219	*0.650*	0.641	0.344	0.299	0.651	**0.862**	0.302
Absolute	CC	0.319	*0.379*	0.441	0.284	0.267	0.366	**0.876**	0.308
	GB	0.550	0.644	*0.615*	0.278	0.248	0.582	**0.815**	0.272
	GT	0.486	0.605	*0.565*	0.198	0.173	0.521	**0.732**	0.204
	HYP	0.283	*0.355*	0.403	0.223	0.202	0.301	**0.853**	0.247
	IOI	0.606	0.675	0.645	0.402	0.365	*0.673*	**0.897**	0.376
	SVA	0.550	*0.616*	0.622	0.340	0.285	0.582	**0.838**	0.306

E STATISTICAL SIGNIFICANCE TESTING

We test the comparisons between our methods for statistical significance following a two-tailed version of the test detailed in Berg-Kirkpatrick et al. (2012, Figure 1). The procedure is the same for testing CPR, CMD and performance; for expositionary purposes, assume we are testing CPR.

For each pair of methods (i, j) (e.g., `Eig` vs `SVD`), let the observed difference between their CPR scores on the full validation set be $\delta_{i,j}$. We would like to know whether this difference is statistically significant, i.e., the probability of the effect appearing by chance in a sample from a system where no real effect exists. First, we bootstrap $N = 100,000$ draws over the validation set of each dataset, and compute each method's CPR over the examples in the k-th drawn set, yielding CPR_i^k, $k, 1 \leq k \leq 100,000$. Then, for each pair of methods (i, j), we compute $\delta_{i,j}^k = \mathrm{CPR}_i^k - \mathrm{CPR}_j^k$, i.e., the difference between the CPR scores for i and j on each bootstrapped set. The raw p-value of the difference in CPR scores between methods i and j is given by

$$\frac{1 + \mathrm{count}(|\delta_{i,j}^k - \delta_{i,j}| \geq |\delta_{i,j}|)}{1 + k}$$

(i.e., the number of times each bootstrapped difference in CPR matches or exceeds the observed difference). We opt for the two-tailed version of the test because unlike the setting of Berg-Kirkpatrick et al. (2012), we do not have hypotheses about the directionality of the comparisons. To account for multiple comparisons, we adjust the raw p-values for all pairwise comparisons between methods within each dataset using the Bonferroni-Holm method (Holm, 1979).

F ALTERNATIVE MEASURE FOR DIAGNOSING INTERPRETABILITY ILLUSION

We also considered an alternative measure of the Interpretability Illusion. To the extent that some direction $\boldsymbol{s}$ is split into two illusory directions $\boldsymbol{s}_i \in \boldsymbol{K}$ and $\boldsymbol{s}_j \in \boldsymbol{U}$, the corresponding entry X_{ij} will be strongly negative. Therefore, we define **the illusory opposition score** $\mathrm{Opp}(\boldsymbol{S}_{uv}, \boldsymbol{S}_{uv}^{\perp})$ as the mean negative entry of X_{ij}, normalized by the total variance of the scores:

$$\mathrm{Opp}(\boldsymbol{S}_{uv}, \boldsymbol{S}_{uv}^{\perp}) := \frac{\mathbb{1}[X_{ij} < 0] X_{ij}}{\sum_{i,j} \mathbb{1}[X_{ij} < 0] \ \mathrm{Var}_n\big(\widetilde{\Delta L}_{uv}^{(n)}\big)}, \tag{7}$$

where $\boldsymbol{X} = \mathrm{Cov}_n\big[\alpha_i^{(n)}, \alpha_j^{(n)}; \boldsymbol{s}_i \in \boldsymbol{S}_{uv}, \boldsymbol{s}_j \in \boldsymbol{S}_{uv}^{\perp}\big]$. To get the opposition score for an entire circuit at a particular budget k, we can simply average the Opp score for every edge $u \to v$. Finally, to get a measure of Opp across all circuit budgets, we can take the area under the curve (AUC) of the Opp scores.

G DETAILS OF VARIABLE SEPARATION

With x_i the coordinates of prompt i, $\bar{x}$ their mean over all N prompts, and $\bar{x}_g$ the mean over the n_g prompts in group g, Cohen (1973)'s η^2 is given as:

$$\eta^2 = \frac{\sum_g n_g \, \|\bar{x}_g - \bar{x}\|^2}{\sum_{i=1}^{N} \|x_i - \bar{x}\|^2}. \tag{8}$$

η^2 is 1 when every prompt lies at its group's mean and 0 when all groups share one mean. η^2 is equivalent to is equivalent to R^2 (Lakens, 2013), the quantity reported as explanation via regression (EVR) by Engels et al. (2025).

Here we provide additional details on the variable separation analysis of **??**. Table 5 lists how the 131 edges analyzed are selected from the subspace circuit, and Table 6 lists the six prompt variables. Figure 3 reports η^2 for each variable on all selected edges.

Table 5: From the 0.1% subspace circuit to the edges analyzed in **??**.

selection step	count
axes in the model	∼4.1M
circuit: top 0.1% of axes by attribution score	4,120
after removing edges with no path to the logits	3,997
edges with at least one retained axis	632
written by an attention head	436
with at least two retained axes	267
read at the final token (query input or logits)	131
with at least four retained axes (remainder view)	93

Table 6: The six variables, with their values for the example prompt of Wang et al. (2022), "When Mary and John went to the store, John gave a drink to".

variable	source	example	values
S position	MIB causal variable S_{Pos}	S second	2
S name	MIB causal variable S_{Tok}	John	44
IO name	`metadata.indirect_object`	Mary	44
template	`template`	When {A} and {B} went to the {place}…	9
object	`metadata.object`	drink	87
place	`metadata.place`	store	53

H Causal Variable Localization

Table 7: Causal variable localization on the IOI task in GPT-2, following MIB: variables for the position S_{Pos} and token S_{Tok} of the subject, intervening on heads 7.3, 7.9, 8.6 and 8.10. We use MIB's evaluation code and metric: mean-squared error against its causal model's predicted logit difference, averaged over the three counterfactual datasets; lower is better. Full Vector patches the best subset of whole heads for each variable. Random is the mean over ten random subspaces with SAP's dimensions. DAS, DBM+PCA and Full Vector are run with MIB's default hyperparameters. DAS and DBM+PCA are averaged over five training runs.

Method	Dims	Trained	S_{Pos}	S_{Tok}
DAS	128	yes	*2.19*	*2.25*
DBM+PCA	–	yes	2.27	2.40
Full Vector	–	–	2.44	2.83
No intervention	0	–	16.19	5.39
SAP	47	no	2.72	5.71
Random	47	no	9.51	3.22
SAP, top two axes	8	no	4.01	4.90
Random	8	no	14.41	4.78
SAP, remaining axes	39	no	10.76	3.15
Random	39	no	11.63	3.52

I Prompt variables separated per edge

This section illustrates the separation of distinct variables by edge. Against the 131 selected edges (Table 5), the object variable is only separated on four, and the place variable on zero. This is expected, as these variables are not causally-relevant to the IOI task, and would not be expected to be retained in a circuit with a small axis budget (here, $k = 0.001$). We interpret the sparseness as indicating this metric is a strong signal for the variable carried by each edge.

The heatmap is interpretable: it shows that most edges are concentrated, encoding only one or two variables, with strong signal.

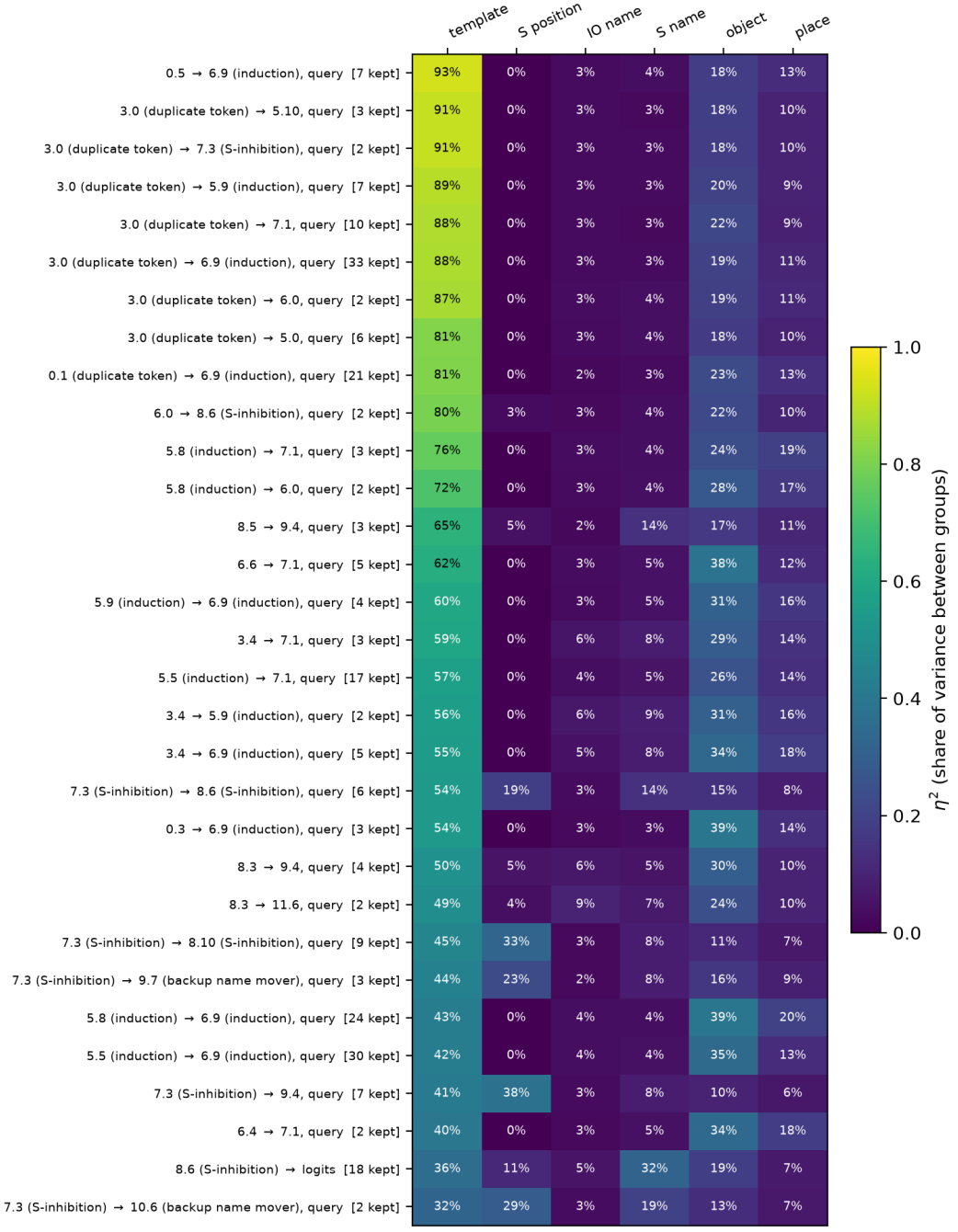


Figure 3: η^2 **on every edge analyzed.** All 131 edges of Table 5, grouped by the variable with the largest η^2 and sorted by it within each group; place has the largest η^2 on no edge. This page: template.

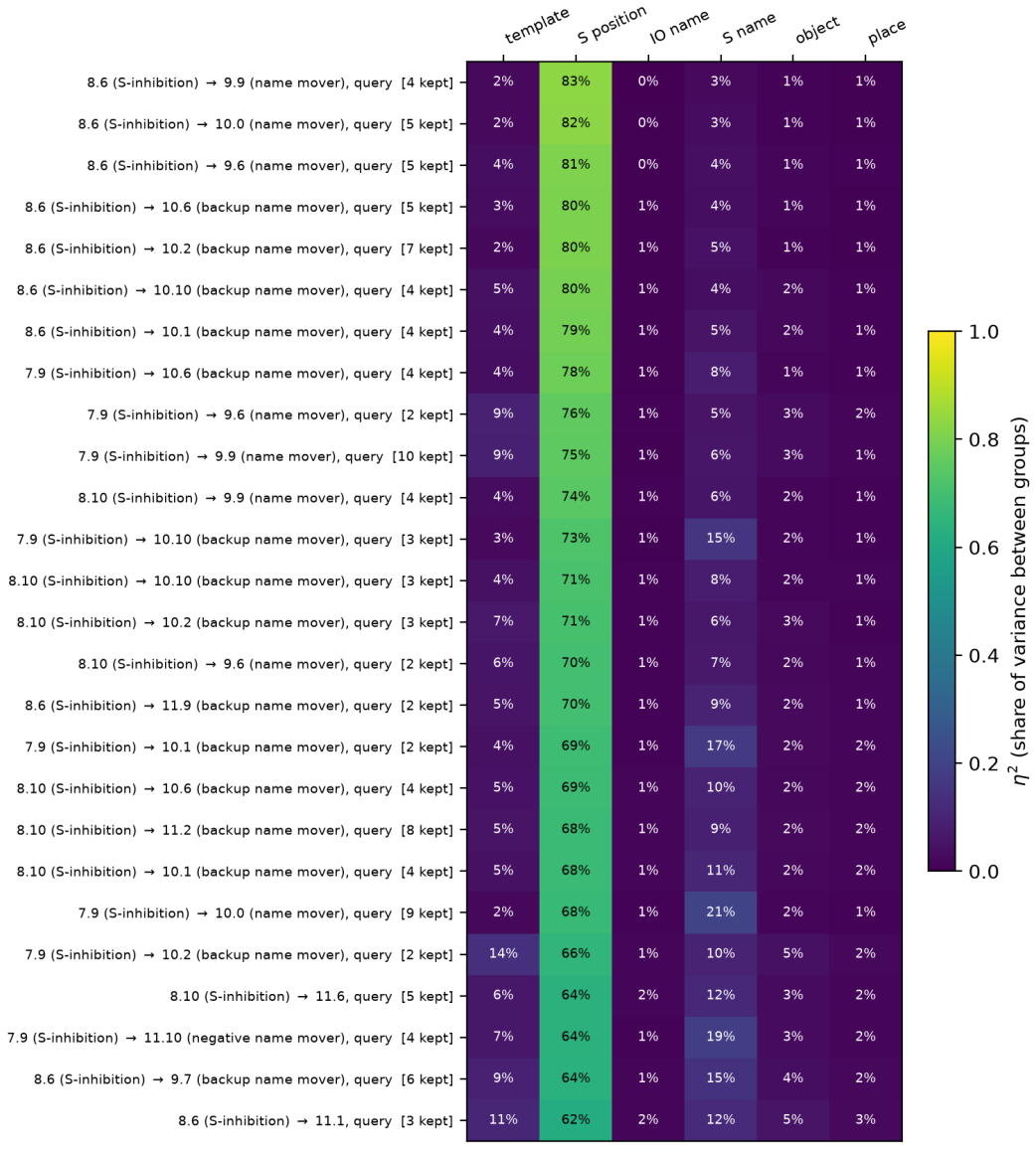


Figure 3: (continued) S position, part one.

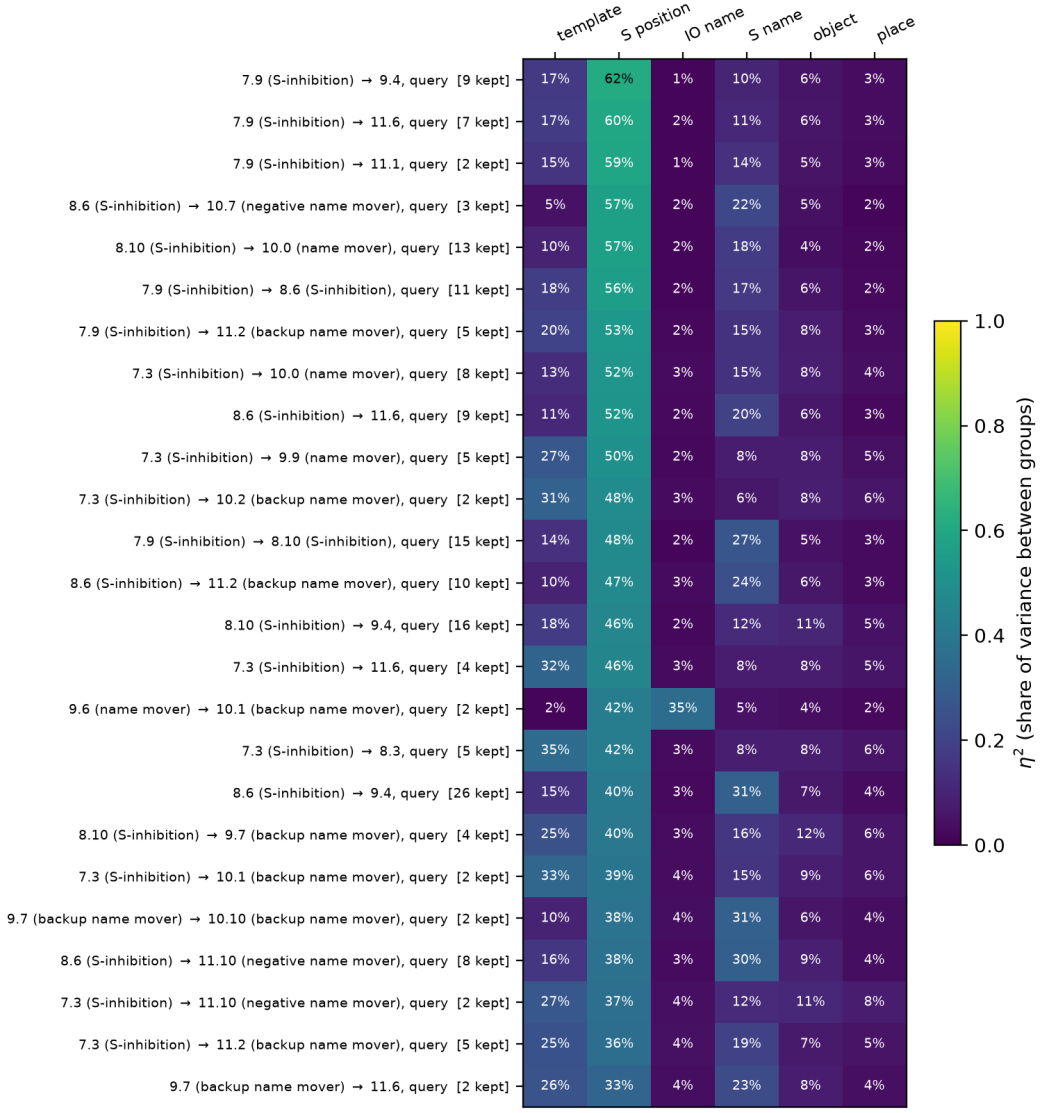


Figure 3: (continued) S position, part two.

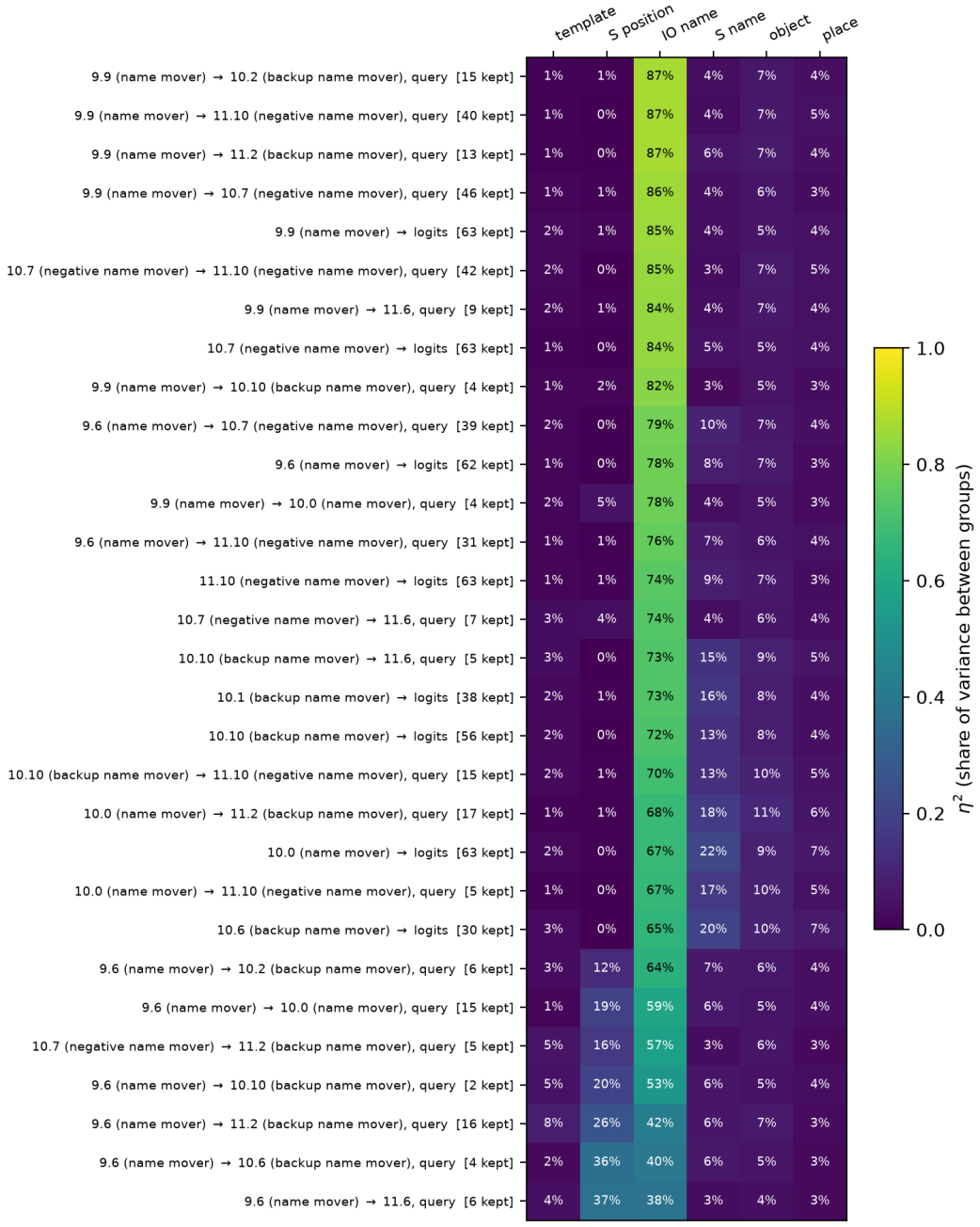


Figure 3: (continued) IO name.

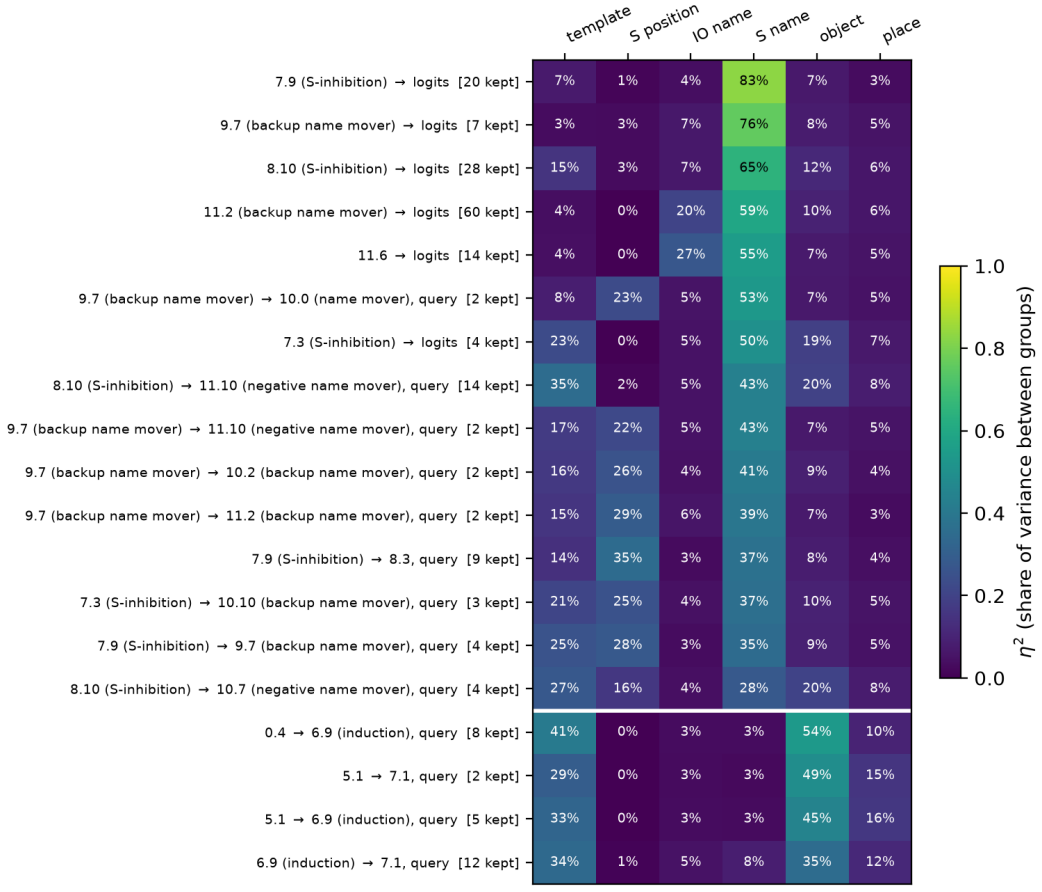


Figure 3: (continued) S name and object.