A SPARSE CIRCUIT FOR CONTEXT-TO-QUERY INFORMATION TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

A number of head families have been identified for in-context learning (ICL), ranging from heads that promote token and semantic-level copying to heads that perform retrieval and task-level functions. While several of these head families *encode* dedicated functions that generalize across ICL tasks, less is known about which heads causally mediate information flow *between the context and the query*. Using a causal intervention that patches *only* the attention connections between query tokens and context tokens, we first identify attention heads that mediate this flow directly. We then identify a shared set of 20 such heads, dubbed ICL20, that supports a diverse range of in-context learning functions. Ablating ICL20 degrades performance across eight diverse evaluation tasks. Moreover, individual ICL20 heads adopt different functional roles by task. Our approach recovers compact, general-purpose circuitry underlying several previously studied ICL functions and reveals two new roles assumed by attention heads during ICL: verbalizing relations between context segments and the query, and using chat-template markers as retrieval anchors.

1 INTRODUCTION

Consider a large language model poised to predict the label of a novel query using a handful of labeled examples specified in the context. To succeed, it must compose computations across multiple circuits: first, it must parse the prompt to identify examples and labels; next, it must infer which features determine an example's label; third it must synthesize this information determine the query's label; finally, it must map that decision to the correct output token in the vocabulary (Wang et al., 2023b; Pan et al., 2023; Hendel et al., 2023; Todd et al., 2024; Yu & Ananiadou, 2024). Broadly, several language-model behaviors can be viewed as compositions of these primitives: low-level circuitry for parsing and copying, combined with higher-level circuitry for transferring and synthesizing contextual information (Olsson et al., 2022; Yin & Steinhardt, 2025; Hendel et al., 2023; Todd et al., 2024).

Prior work characterizes several mechanisms supporting this process. When models continue contextual sequences, (Elhage et al., 2021; Olsson et al., 2022), they rely on specialized token (Elhage et al., 2021; Olsson et al., 2022) and concept *induction* heads (Feucht et al., 2025)[1]. These circuits mediate both verbatim ($[A][B]\ldots[A] \rightarrow [B]$) and concept-based sequence copying ($on..a..whi.m....on..a.. \rightarrow [whi.m]$), and are widely viewed as core mechanisms underlying in-context learning. Related work has also identified heads that retrieve relevant information from context (Wu et al., 2024). What remains less understood is how models synthesize and transfer higher-level information like evidence/labels across examples. For example, in the classification task '`I had a wonderful day` $\mapsto$ `happy`; `He lost his wallet that evening` $\mapsto$ `sad`; `Her dog was diagnosed with cancer` $\mapsto$ `sad`; `They loved ice cream` $\mapsto$ `happy`; `She loved kayaking` $\mapsto$?', it is unclear whether information is transferred primarily through content representations, through token or concept copying (Olsson et al., 2022; Feucht et al., 2025), or through relational representations, such as the similarity between `She loved kayaking` and `They loved ice cream` (Yu & Ananiadou, 2024).

[1]We call these token copiers and concept copiers (Feucht et al., 2025) since induction can have conflicting meanings.

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.

To study this information flow directly, we develop a variant of Desiderata-based Component Masking (Davies et al., 2023b) that patches only attention edges between context and query tokens, allowing us to identify heads that mediate contextual information flow. We find that this flow is language-agnostic, generalizes across tasks, and is carried by a sparse set of specialized heads. In particular, we identify a cross-section of 20 heads (the top 2% of attention heads) selected across the natural-language reasoning tasks, which we call ICL20, that are causal mediators of important ICL functions for a diverse set of tasks like in-context classification, reading comprehension, translation, token copying, relational association, and free-form reasoning. In `Llama-3.1-8B-Instruct` and `Llama-3.2-3B-Instruct`, ICL20 is largely disjoint from known token-copying, concept-copying, and retrieval heads. Ablating ICL20 degrades performance across ICL tasks, including pure sequence copying, even when strong token copier heads remain intact (Yin & Steinhardt, 2025). Beyond copying, ICL20 contains heads that verbalize relational information between contextual segments and the query, as well as heads that attend to chat-template markers at the beginning and end of the prompt when supporting retrieval, suggesting that these are important for in-context learning. Together, these results reveal a compact, general-purpose circuit for contextual information flow across diverse ICL tasks.

Organization We begin by describing our masking and patching approach to identify heads that causally mediate context $\rightarrow$ query information transfer in §2, and experiments that describe the behavior of these heads in §3. Lastly we describe our procedure for finding ICL20, and create a taxonomy of heads contained in it in §4. All the results in the main paper are described for `Llama-3.1-8B-Instruct`. The main results for `Llama-3.2-3B-Instruct` are in the Appendix A.1.

2 Desiderata-Based Masking on Attention Connections

We use Desiderata-based Component Masking to isolate heads that causally mediate contextual information flow between the context and the query. DBM directly optimizes to find the smallest causal circuit that can explain a behavior at any given level of circuit sparsity. We follow the standard recipe used in causal interpretability Geiger et al. (2021); Vig et al. (2020); Meng et al. (2022) to locate these circuits, as described below:

Task setup. An input $x = c \oplus u$ contains a context c followed by a question segment u, which can include a situation. In a mock ROPES example, c is *'For these plants, more water leads to faster growth.'* The question segment is *'Plant A receives more water than plant B. Which plant grows faster?'* The model predicts an answer after the final question token. We restrict and renormalize its answer distribution over a candidate set $\mathcal{Y}$. Here, $\mathcal{Y} = \{\mathrm{A}, \mathrm{B}\}$, with correct answer A.

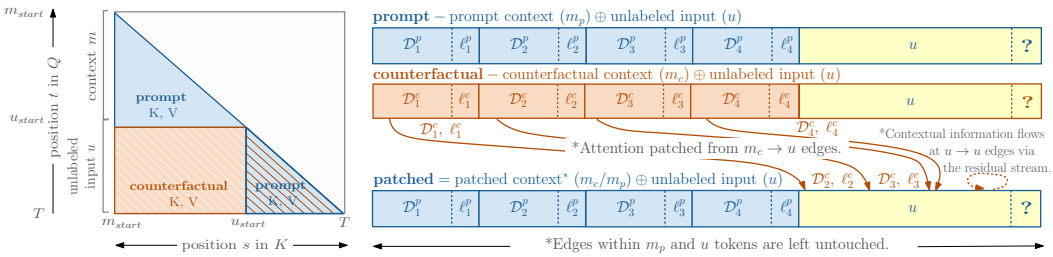
Counterfactual inputs. To identify attention heads that are sensitive to minimal changes in the input, we pair x with its counterfactual, $x^{\mathrm{cf}} = c^{\mathrm{cf}} \oplus u$, keeping the question fixed. x^{cf} replaces *faster* with *slower* and this one word change flips the input's answer to B. We retain only pairs with contrasting correct answers for which the unmodified model predicts both answers correctly. Each pair is matched on the overall length as well as the lengths of the context and question spans $T_{\text{context}}, T_{\text{question}} \subseteq \{1, \ldots, n\}$, where n is the padded sequence length and the spans exclude padding. The token sequences c and c^{cf} have length $|T_{\text{context}}|$; u has length $|T_{\text{question}}|$. Thus x and x^{cf} have length $|T_{\text{context}}| + |T_{\text{question}}|$ before padding. All activation shapes below refer to one padded input.

Attention connections from contexts to questions. The model has L layers and H query heads per layer, with head dimension d_h and residual dimension d_{model}. For one head, let $Q, K, V \in \mathbb{R}^{n \times d_h}$ denote its query, key, and value projections. Its attention output $A \in \mathbb{R}^{n \times d_h}$ is

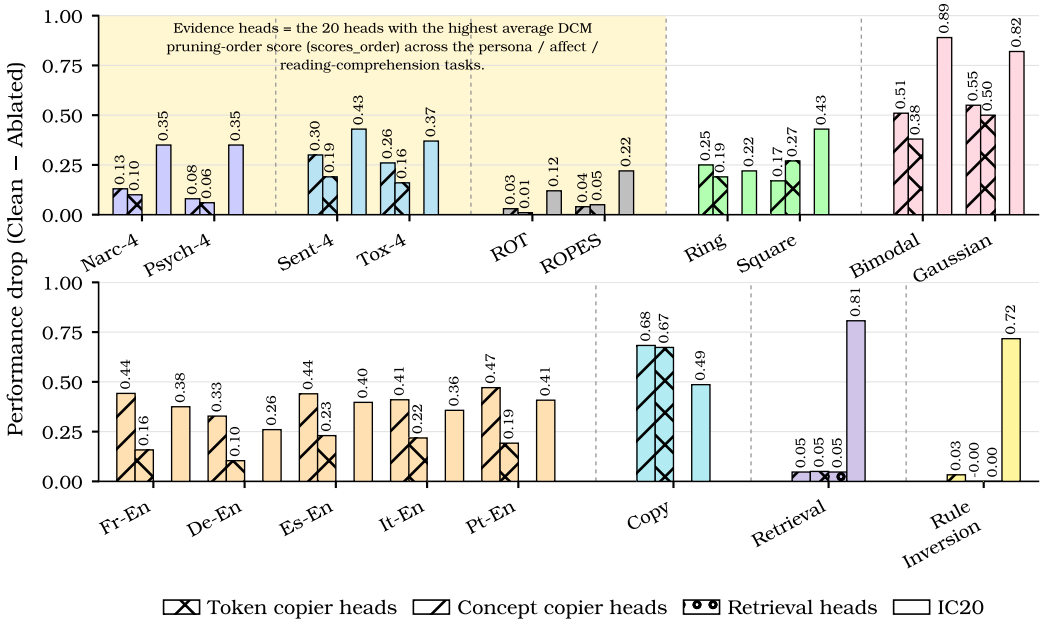
$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + B\right) V, \tag{1}$$

where $B \in \{0, -\infty\}^{n \times n}$ applies the causal and padding masks, and softmax acts across key positions. Rotary position embeddings are applied to Q and K. For grouped-query attention, keys and values are expanded to query heads.

Context tokens attend within the context, while question tokens can attend to all preceding tokens and themselves. A head can carry context information to the final question token or to an earlier



(a) **Identifying attention heads that causally mediate context → query information transfer through Desiderata-Based Masking on Attention Connections** We learn a mask over all attention heads using pairs of inputs that share the same query, u, but contain contrasting contexts, x and x^{cf}. The objective is to identify a sparse set of heads such that patching them during the forward pass on $x \oplus u$ causes the model to produce the output associated with $x^{\text{cf}} \oplus u$. During mask training, for each attention head we recompute the query-region attention using the K/V states from x^{cf} in place of those from x. Attention within the context, as well as attention among tokens in the query region, is left unchanged. A learned scalar mask for each (ℓ, h) then interpolates between the clean and fully patched head outputs across all query positions. Sparsity-regularized optimization identifies the smallest set of heads whose reading of contextual K/V states causally determines the model's prediction.



(b) **A sparse circuit for context → query information transfer** We use the approach in Fig. 1a to train masks on 3 classes of natural language reasoning tasks, and identify a cross section of 20 attention heads in `Meta-Llama-3.1-8B-Instruct` across the tasks, ICL20, which appear to be a general circuit for mediating information transfer from the context to the query across 8 different task families. The ablation of the ICL20 heads causes inconsistent reasoning during the model's generation. ICL20 is disjoint from other important ICL head families like token and concept copier heads (1 shared head), as well as retrieval heads.

Figure 1: Identifying a sparse circuit of attention heads that mediate information between tokens in the context and query.

question token, where later heads can access it. Residual connections carry information across layers at each position.

Patching in counterfactual connections. For a data pair, x and x^{cf}, we first run the model $\mathcal{M}$ on x^{cf} and cache the resulting key and value activations. At layer ℓ and head h, let $K, V \in \mathbb{R}^{T \times d_h}$ denote the key and value matrices computed during the clean run; let $\widetilde{K}, \widetilde{V}$ denote their cached counterparts from the counterfactual run. We rewrite the model's forward pass to perform a surgical interchange

intervention that replaces the keys and values corresponding to the context tokens in the input. The patched output $\widetilde{A} \in \mathbb{R}^{n \times d_h}$ uses these projections at question positions.

$$(\hat{K}_t, \hat{V}_t) = \begin{cases} (\tilde{K}_t, \tilde{V}_t), & t \in T_{\text{context}}, \\ (K_t, V_t), & t \notin T_{\text{context}}, \end{cases} \qquad , \widetilde{A}_{T_{\text{question}}} = \text{softmax}\left(\frac{Q_{T_{\text{question}}} \widetilde{K}^\top}{\sqrt{d_h}} + B_{T_{\text{question}}} \right) \widetilde{V}.$$

Here $Q_{T_{\text{question}}}$ and $\widetilde{A}_{T_{\text{question}}}$ have shape $|T_{\text{question}}| \times d_h$, while $B_{T_{\text{question}},:}$ has shape $|T_{\text{question}}| \times n$. At all other positions, $\widetilde{A}_t = A_t$, preserving context $\rightarrow$ context attention outputs. Recomputing softmax at question positions can also change weights on question tokens. In the example, the question receives context keys and values computed with *slower*. Since later layers use the resulting residual stream, the patch can affect subsequent attention between question tokens.

Desiderata-based masking. We wish to identify causally relevant context-transfer circuits at different levels of attention head sparsity. For each task, we learn one mask $m \in [0,1]^{L \times H}$ across its training pairs, with every gate initialized to 1. At head (ℓ, h), we interpolate the ordinary and patched outputs before applying the output projection:

$$\widehat{A}_{\ell h} = (1 - m_{\ell h}) A_{\ell h} + m_{\ell h} \widetilde{A}_{\ell h}, \qquad O_\ell = \sum_h \widehat{A}_{\ell h} W^O_{\ell h}. \tag{2}$$

Here $A_{\ell h}, \widetilde{A}_{\ell h}, \widehat{A}_{\ell h} \in \mathbb{R}^{n \times d_h}$, $W^O_{\ell h} \in \mathbb{R}^{d_h \times d_{\text{model}}}$, and $O_\ell \in \mathbb{R}^{n \times d_{\text{model}}}$. The scalar gate $m_{\ell h}$ controls the patch: 1 selects the counterfactual attention, and 0 uses ordinary attention on the current residual stream. Heads with gate 0 can propagate effects of earlier interventions.

All other model parameters remain fixed while we optimize the mask. A head masked out at an earlier step may re-enter the circuit later if its interaction with the currently active heads becomes important for minimizing the task loss.

For example i in a batch of N pairs, let $p_m^{(i)}, p_{\text{cf}}^{(i)} \in [0,1]^{|\mathcal{Y}_i|}$ be probability vectors for the intervened and unmodified counterfactual runs, restricted and renormalized over that pair's candidate set $\mathcal{Y}_i$. Define

$$\mathcal{L}_{\text{task}}(m) = \frac{1}{N} \sum_{i=1}^{N} \text{JS}\left(p_m^{(i)}, p_{\text{cf}}^{(i)} \right), \qquad \mathcal{L}_{\text{sparsity}}(m) = \frac{1}{LH} \sum_{\ell,h} m_{\ell h}. \tag{3}$$

Here JS is Jensen–Shannon divergence and both losses are scalars. We minimize $\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{sparsity}}$ with scalar coefficient λ, balancing agreement with the counterfactual distribution against the mean mask value. In the example, the target distribution assigns its highest probability to B.

Controlling the sparsity and intervention effect trade-off. During each optimization run, a proportional–integral (PI) controller adjusts the sparsity penalty, $\lambda \in [10^{-8}, 10^{8}]$, to make the selected head count track a linear decay, which removes the need to tune λ. We save binary masks during training for evaluation across sparsity levels.

Evaluating masks across sparsity levels. For each task, we evaluate the learned binary mask on different levels of sparsity, on held-out input pairs using interchange intervention accuracy (IIA) (Geiger et al., 2021). IIA measures whether replacing an attention representation for input x with the corresponding contrastive representation from x' produces the contrastive output:

$$\text{IIA} = \mathbb{E}_{x,x'} \left[\mathbf{1}(f_{x \leftarrow x'}(x) = g_{x \leftarrow x'}(x)) \right].$$

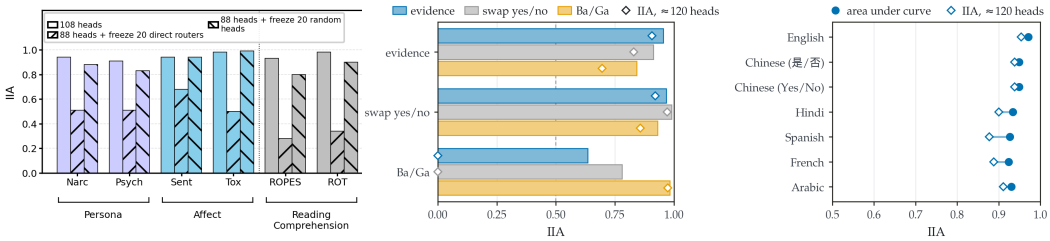
Random masks intervening on the same fraction of heads serve as controls.

Attention Head Ranking. We use the binary masks saved throughout training to construct a head-ranking matrix $R \in \mathbb{R}^{L \times H}$ for each task, where larger entries correspond to heads that remain active (mask $= 1$) for a greater fraction of the optimization process.

3 Attention Heads that Connect the Context to the Query

3.1 Approach

Datasets To investigate circuits that are specialized for transferring information between the query and the context, we train attention masks using the approach from § 2 on a class of natural language



(a) A subset of context → query heads directly write context information to the final token, freezing these heads can diminish the effects of patching the context transfer circuits.

(b) Context and label information are transported through distinct circuits.

(c) Context information circuits are language agnostic.

Figure 2: We study the properties of context → query heads along the above axes.

reasoning tasks that comprise classification and reading comprehension. The classification tasks, Persona (Bigelow et al., 2025) and Affect, use labeled examples to classify a new input for narcissistic or psychopathic traits and sentiment or toxicity, respectively. Reading comprehension includes ROPES (Lin et al., 2019), which asks questions about a situation using a background passage, and ROT (Rao), which asks whether an action is morally acceptable in a given situation. Across both groups, each task provides 4,000 training pairs and 750 test pairs. Within each pair, we use the 'contrastive counterfactuals' design described above, i.e. the question segment stays fixed while the context changes to support a different answer.

Setup In the following experiments, we train the DCM mask in one task setting and evaluate its transfer to a new setting across circuits extracted at different sparsity levels. For each evaluation, we patch the attention edges between the context and the query and measure the IIA(See ¶ 2) as a metric of successful context transfer.

3.2 Results

Semantically similar tasks share their context transfer circuits. We find that the top-10, top-20, and top-65 ranked attention heads for the Affect, Reading Comprehension, and Persona tasks are sufficient to yield an IIA of at least 80% on the corresponding task. Masks from the same task families well onto other tasks from the same families (Also see A.3).

Context information has direct and indirect pathways to the answer token We consider the top 108 attention heads (10% of all heads), ranked by the 4-shot Narcissism mask. We then identify the 20 heads that most directly amplify

$$\Delta_{\text{logit}} = \text{logit}(y^{cf}) - \text{logit}(y),$$

where y and y^{cf} are the clean and counterfactual outputs at the final token. To do so, we apply the logit lens (nostalgebraist, 2020) to each head's o_{proj} input over 200 random samples and rank heads by their Δ_{logit} value. We then patch the o_{proj} inputs of the remaining heads while freezing the outputs of these direct context routers. This recovers a substantial fraction of the original task behavior relative to a random baseline.

Context information and label information are transported through distinct circuits. We ask whether heads that carry contextual information also carry the final answer label. In addition to the original 8-shot Narcissism counterfactual context, x^{cf}, we construct two variants: (1) x^{cf} with swapped labels (`Yes` → `No`, `No` → `Yes`), and (2) x^{cf} with a new label vocabulary (`Yes` → `Ba`, `No` → `Ga`). We learn one DCM mask per setting. The context-derived and swap-derived masks transfer strongly to one another (IIA AUC 0.97 and 0.91), but both fail to transfer to the Ba/Ga setting once the circuit is pruned to ≈ 120 heads, continuing to output their original labels, `Yes` and `No`. On the other hand, the Ba/Ga dataset achieves an IIA AUC of 0.97, suggesting that mapping the answer into this new label space requires additional heads beyond those mediating contextual transfer. On a joint context-and-label intervention, adding the Ba/Ga-specific heads recovers the new

label tokens; however, removing heads shared between the Ba/Ga and the original context-transfer circuit eliminates this effect. The Ba/Ga label-token heads also occur later in the network, primarily in layers 8–31(Also see Appendix B.1.1).

Context information transfer is language-agnostic. We translate every statement and instruction from the 8-shot Narcissism task into Chinese, Hindi, Spanish, French and Arabic and evaluate them on the mask derived from the English evidence. The mask transfers well to all five language settings – the area under the IIA curve stays at 0.92–0.95 in every language, against 0.97 for English. On a smaller circuit in which only the top ≈ 120 heads are patches, IIA stays at 0.88–0.94. The Chinese test set uses the native yes and no characters as labels; a copy of the same pairs with Yes/No labels gives the same curve (0.95), so the label tokens play no part in the transfer. The same heads carry in-context evidence in all six languages.

While the above circuits exhibit interesting properties, they are quite large. On free form tasks, patching these circuits leads to incoherent outputs. In the following section we describe our procedure for locating a highly sparse and general circuit that mediates context $\rightarrow$ query connections, comprising just 20 attention heads in the model, which surgically removes contextual information across 8 diverse task families (See Fig. 1b).

4 ICL20: A Sparse Circuit for Context $\rightarrow$ Query Transfer

We rank heads across masks trained on the natural-language reasoning tasks using their pruning order (see ¶ 2). For each task $j \in \mathcal{T}$, $R^{(j)} \in \mathbb{R}^{L \times H}$ assigns larger values to heads pruned later. We average these rankings,

$$\overline{R} = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} R^{(j)},$$

and retain the 20 highest-scoring heads for subsequent experiments, dubbing them ICL20. The persona and affect tasks use a shared prompt format in the structural tokens that wrap their exemplars and labels, while the two reading comprehension tasks use a different format.

4.1 Comparisons with other ICL Heads

ICL20 constitutes a distinct, general purpose circuit, that is different from head families that perform other functions during in-context learning. Particularly, we compare ICL20 with three head families: token copiers, concept copiers and retrieval heads.

Concept Copier Heads. Concept copier heads (Feucht et al., 2025) attend to the ends of a multi-token semantic unit, and our identified by their ability to predict the 'next-next-token' in a sequence. For e.g., in the sequence `[random`$_1$`] Ed.mont.on.[random`$_1$`]Ed`, a concept copier head amplifies the likelihood of the token `on`. We rank the top 20 concept heads using this score.

Token copier heads. We identify token copier heads over repeating random strings, using the approach documented by Olsson et al.. Signatures are averaged over 16 random sequences per head, and compared against the null rate expected by chance. We select the top 20 token copier heads using this score.

Retrieval heads. We use the approach in Wu et al. (2024) to identify retrieval heads, which copy a segment placed at any depth in a long context into the output. During decoding, each head is scored by whether its top attended position and the generated token falls inside the needle. We average this score across a grid of context lengths and depths, rank all heads, and select the top-20.

4.2 Results

Unlike token and concept copier heads as well as retrieval heads, ICL20 heads occur in the early to middle layers of the model. Ablating them degrades performance by gating downstream information that is read by token/concept copier heads as well as output token heads. They have weak inductive but strong previous-token head and duplicate-token head like signatures(See Appendix C). They're also disjoint from the top 20 token copier/retrieval heads and share only one head with the top 20 concept copier heads. As shown in Fig. 1b, ablating them causes a higher drop in performance on

average ($p < 0.001$, $n = 18$), as compared to token copier, concept copier, as well as retrieval heads. Surprisingly, although ICL20 is identified from multi-shot classification tasks, ablating these heads also degrades performance on differently structured ICL tasks such as translation and copying, as well as on distinct settings involving free-form reasoning and retrieval. This breadth suggests that ICL20 supports a general computation shared across many forms of language processing.

5 The Effects of Lesioning ICL20

5.1 Approach

We evaluate the effects of ablating the ICL20 heads on the following classes of data in addition to the natural language reasoning tasks they were trained on. To ablate an ICL20 head, we overwrite the head's slice of the o_{proj} input with zero, i.e. we remove that head from the computational flow.

Synthetic tasks. In addition to the natural language reasoning tasks that were used to identify ICL20, we evaluate them on two synthetic tasks: (1) Distribution tasks, which require predictions based on integers sampled from Gaussian or bimodal distributions (Liu et al., 2025). (2) Grid tasks, in which the context contains tokens generated by a walk on a square grid or ring (Park et al., 2025). Grid tasks contain strong n-gram motifs, since random graph walks contain repeating nodes and path segments, resembling n-grams (n=128).

Translation and Sequence Copying Tasks Following the task design of Feucht et al. (2025), we use word-pair data from Conneau et al. (2017) to construct translation and copying tasks. Each task presents $n = 10$ words in one of five languages or in English, followed by a parallel list containing either their English translations or exact copies, and asks the model to complete the final translation or copy corresponding to the last word in the first list ($n = 1000$).

Free-form generation tasks To evaluate effects on free-form generation, we construct two synthetic datasets ($n = 300$), which measure contextual retrieval and contextual rule inversion. In the former, the model has to answer a retrieval query, like *Tobias resigned in March and Ines took over in April. In which month did Ines take over the role?* and in the latter, it has to answer a query based on a contextual rule where the semantic meaning is inverted, for eg. saying that a component is tagged *'grounded'* if it passed its inspection and *'airworthy'* if it failed it, and asking the model to tag the component based on the context.

5.2 Results

Ablating ICL20 heads suppresses capabilities across several tasks Zero ablations of the ICL20 heads cause the model to output the contrastive response (i.e. treat the prompt as if it were the counterfactual) on the natural language reasoning tasks. On the free-form generation tasks, the model exhibits an internal inconsistency in its own reasoning; on several inputs it is able to situate itself in the context, even state the query being asked, and yet unable to answer the question. Though the ICL20 heads have been extracted from masks trained on the persona, affect, and reading comprehension tasks alone, their behavior transfers to completely out of distribution in-context learning and free-form generation tasks.

Qualitative results Ablating the top-20 ICL20 heads obscures contextual information from the model and produces a spectrum of failures. At one extreme, the model can still recover and situate itself in the relevant context but reasons inconsistently during generation; at the other, the relevant context is effectively hidden, leading the model to explicitly report that it lacks the information needed to answer the query.

Question: Copy this sentence: "Science is really awesome"

Clean continuation: "Science is really awesome."

Ablated continuation: I cannot copy the sentence 'Science is really awesome'.

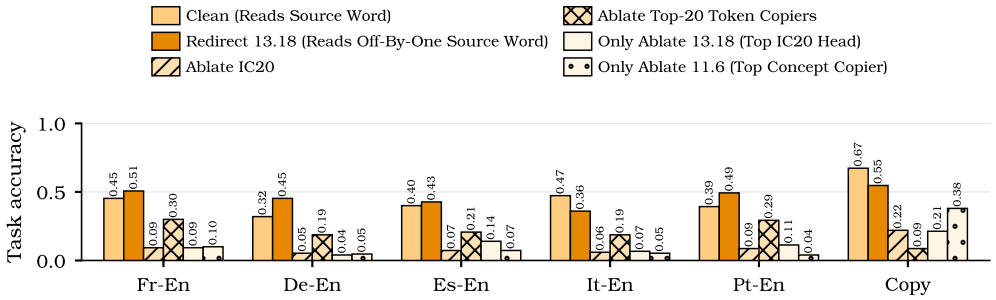


Figure 3: Ablating one ICL20 head can degrade translation abilities more strongly than ablatng the top-20 token copier heads.

Question: Parking costs two dollars per hour. The car stayed for five hours. What is the hourly parking rate?

Clean continuation: The statement "Parking costs two dollars per hour" already gives us the hourly rate. It's $2 per hour. The car staying for 5 hours is additional information, but it doesn't change the hourly rate. The hourly rate remains $2 per hour.

Ablated continuation: The information provided does not give enough details to determine the hourly parking rate. However, it does mention that the car stayed for five hours.

5.2.1 A TAXONOMY OF ICL20 HEADS

ICL20 heads contain a concept copier head for translation tasks. We first replicate results from (Feucht et al., 2025) on the translation tasks (Conneau et al., 2017) to show that ablating the top-20 concept copier heads is destructive on translation tasks. However we find that ICL20 heads are more influential than classic token copier heads. In fact, `13.18`, a top-20 ICL20 head with a strong duplicate-token function gates information flowing to downstream induction head. On the forward pass of the clean model, `13.18` shows a large attention weight on the last sub-token of the source word for the translation, behaving in this task as the previous token head for the downstream induction head (Olsson et al., 2022; Feucht et al., 2025). Ablating `13.18` alone has the effect of diminishing accuracy by more than 75% across all translation tasks, and is more destructive than ablating the top-20 token copier heads. Qualitatively, ablating `13.18`'s attention redirects the model to translate a different word without impacting fluency or relevance, suggesting that this head is an information reference for a downstream head. In fact, our investigations suggest that `13.18` seems to be a source head for several downstream heads on this task.

Decodings of ICL20 head representations contain interpretable tokens for predicting the output label We find 6 heads (`11.22, 11.27, 10.28, 13.20, 11.13, 10.24`) among the top-20 ICL20 heads that degrade performance on the persona and affect tasks (narcissism, psychopathy, sentiment, toxicity) when their representations are zero-ablated or negated, causing drops in the range of 47 to 80% across the 4 tasks. Uniquely, one head in that set, 13.20, has a decoding that is verbalized on forward passes for the classification-style ICL tasks (4-shot narcissism, sentiment-chat, and psychopathy). At intermediate points in the processing, when its O-V representation is projected onto the unembedding, it increases the probability of human-interpretable labels like 'opposite', 'unlikely' etc. to track whether the most recent labeled example agrees or disagrees with its prediction for the output, essentially creating a shortcut to the prediction (Gurnee et al., 2026). A search through the full set of attention heads finds similar heads (head 21.21 is one, which express the same behavior on the toxicity-4 task).

Models rely on contextual anchors; ablating heads that attend to these anchors can degrade performance. We select 9 ICL20 heads (`14.24, 13.18, 12.15, 11.27, 13.20, 13.21, 4.16, 10.26, 8.11`) whose attention on chat-formatted prompts falls mostly on template markers and BOS, not on the item's words. Six attend to the tokens that close the user turn (`14.24, 13.18, 12.15, 11.27, 13.20 and 13.21`) and open the reply

(<|eot_id|>, the assistant header, the final newline), and to the last token in the input, a ?; the other three (4.16, 10.26 and 8.11) attend to BOS, system-header tokens (i.e tokens that open the user turn), and the newline after earlier assistant headers. On two synthetically generated QA sets (Retrieval and Rule Inversion, 300 items each), zero-ablating all 9 drops accuracy from 0.89 to 0.37 and from 0.85 to 0.34 (vs. 0.08 and 0.13 for all 20 ICL20 heads). The damage comes almost entirely from the six question-side heads. While the absence of these heads degrades performance, they cannot be patched to steer the output. These heads suggest that the model uses contextual anchors while processing the input; removing these anchors can degrade performance on generations.

6 Related Work

Attention mechanisms for in-context learning. Studies of transformer circuits identify attention heads that copy continuations from earlier token or concept occurrences and retrieve information from long contexts (Elhage et al., 2021; Olsson et al., 2022; Feucht et al., 2025; Wu et al., 2024). Function and task vectors encode information inferred from demonstrations that can be transferred to new inputs, with ablation studies testing the contributions of function-vector and induction heads to task performance (Todd et al., 2024; Hendel et al., 2023; Yin & Steinhardt, 2025).

Binding and retrieval in context. Causal studies identify representations that bind entities to attributes and circuits that retrieve those bindings from context (Feng & Steinhardt, 2024; Prakash et al., 2024; Dai et al., 2024; Gur-Arieh et al., 2025). Work on entity and belief tracking shows how information about position directs attention to earlier context when the model retrieves an answer (Prakash et al., 2024; Gur-Arieh et al., 2025; Prakash et al., 2025).

Causal analysis of internal computations. Causal mediation analysis measures how input changes affect model outputs through internal components (Imai et al., 2010; Hicks & Tingley, 2011; Vig et al., 2020; Finlayson et al., 2021; Stolfo et al., 2023; Mueller et al., 2026). To assess whether a model implements an interpretable computation, causal abstraction compares interventions on neural representations with interventions on an interpretable causal model (Geiger et al., 2021; 2025b;a). Activation and path patching locate representations and information routes by replacing activations and controlling how their effects reach the output (Meng et al., 2022; Wang et al., 2023a; Goldowsky-Dill et al., 2023; Zhang & Nanda, 2024).

Learning sparse interventions. Desiderata-based masking uses behavioral objectives to learn masks over model parameters or activations (De Cao et al., 2020; Csordás et al., 2021; Cao et al., 2021; Davies et al., 2023a; Lepori et al., 2023). Distributed alignment search learns subspaces through counterfactual objectives, and circuit discovery methods search for subgraphs that preserve model behavior under interventions (Geiger et al., 2024; Wu et al., 2023; Conmy et al., 2023; Mueller et al., 2025). Here, we learn head masks for patching context keys and values at question positions, with a controller that varies the sparsity penalty during training.

6.1 Discussion and Conclusion

We identify a sparse circuit of 20 attention heads, ICL20, which causally mediates the transfer of information between context and query segments, generalizing to eight diverse evaluation tasks. In order to find these heads, we surgically patch only the attention edges that connect context tokens to query tokens (See Fig. 1). Prior work that identifies similar families of heads does so by scoring the heads on a specialized metric, on tasks involving repeating random strings (Olsson et al., 2022). Our results suggest that the ICL20 heads assume attention functions and signatures that are task-dependent. For example, head 13.18 shows signatures resembling a duplicate-token head when attending to tokens in repeating random strings; however, in every task containing prompts with repeating structure in the form of format markers etc., it attends to later tokens in the query segment, often attending to the last token; on a subset of examples from the ROPES task, it behaves like an attention sink and is largely inert. This suggests that in-context learning functions may be task specific and such attention heads may hold function and task specific subspaces (Todd et al., 2024; Geiger et al., 2024) that compose to determine the behavior of the head. ICL20 may have implications for safety and monitoring, and could be a useful test-bed to discover attention subspaces.

AI USE STATEMENT

In this work, we used generative AI tools for synthetic data generation as well as for code assistance to implement the PI Controller idea on the DCM patching pipeline. An agent also helped build an interface for interacting with the Llama models. We also used coding assistance for help with editing this draft, and identifying related work, though nearly all sections were first drafted by the authors and citations and references were manually copied over. Large parts of the appendix content have been generated with the help of an agent. A generative agent also collaborated through the process of identifying the roles of the different heads in the ICL20 circuit, and helped in creating visualizations. An AI agent also helped brainstorm ideas for testing the functions of the ICL20 heads. Remaining disclosure tasks are not applicable to this work. We have not used generative AI tools for autonomous writing/code generation for the main pipeline. We have reviewed all AI-assisted work, by running sanity tests as well as test-benches on toy examples. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates the role of attention heads in mediating information transfer between the context and query segments of an input. The results and approaches in this paper can be used for malicious purposes in the hands of the wrong actors. However, the results also suggest new avenues for monitoring and understanding models. We will share our methodology, and code to support reproducibility. Ultimately, our goal is to improve understanding of how language models operate and how they can be reliably controlled for benign purposes.

REPRODUCIBILITY STATEMENT

We ran all experiments on two shared clusters with 7 80GB NVIDIA A100 GPUs as well as a mix of 6 H200, H100 and A100s, using the HuggingFace Transformers Library (Wolf et al., 2019) and PyTorch (Paszke et al., 2019). We used NNsight (Fiotto-Kaufman et al., 2024) for our patching experiments. We will release our code if the paper is accepted.

REFERENCES

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

Steven Cao, Victor Sanh, and Alexander Rush. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 960–966. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.74. URL https://aclanthology.org/2021.naacl-main.74/.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=7uVcpu-gMD.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata. *CoRR*, abs/2307.03637, 2023a. doi: 10.48550/ARXIV.2307.03637. URL https://doi.org/10.48550/arXiv.2307.03637.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*, 2023b.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3243–3255. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.262. URL https://aclanthology.org/2020.emnlp-main.262.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. The dual-route model of induction. *arXiv preprint arXiv:2504.03022*, 2025.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144.

Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, et al. Nnsight and ndif: Democratizing access to open-weight foundation model internals. *arXiv preprint arXiv:2407.14561*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the 3rd Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Jacqueline Harding, and Thomas Icard. How causal abstraction underpins computational explanation, 2025a. URL https://arxiv.org/abs/2508.11214.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025b. URL http://jmlr.org/papers/v26/23-0058.html.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. URL https://arxiv.org/abs/2304.05969.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context, 2025. URL https://arxiv.org/abs/2510.06182.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Raymond Hicks and Dustin Tingley. Causal mediation analysis. *The Stata Journal*, 11(4):605–619, 2011. doi: 10.1177/1536867X1201100407. URL https://doi.org/10.1177/1536867X1201100407.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010. doi: 10.1037/a0020761. URL https://doi.org/10.1037/a0020761.

Michael A. Lepori, Ellie Pavlick, and Thomas Serre. Neurosurgeon: A toolkit for subnetwork analysis. *CoRR*, abs/2309.00244, 2023. doi: 10.48550/ARXIV.2309.00244. URL https://doi.org/10.48550/arXiv.2309.00244.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. Reasoning over paragraph effects in situations. In *MRQA@EMNLP*, 2019.

Toni Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher Earls. Density estimation with llms: a geometric investigation of in-context learning trajectories. In *International Conference on Learning Representations*, volume 2025, pp. 22163–22197, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

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*, pp. 1–48, 2026. doi: 10.1162/COLI.a.572. URL https://doi.org/10.1162/COLI.a.572.

nostalgebraist. Interpreting GPT: The logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527/.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In *International Conference on Learning Representations*, volume 2025, pp. 53258–53284, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs, 2025. URL https://arxiv.org/abs/2505.14685.

Kavel Rao. d-rules-of-thumb. Hugging Face Datasets. URL https://huggingface.co/datasets/kavelrao/d-Rules-of-Thumb. Accessed: 2026-09-22.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

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_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

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*, 2023a. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.609. URL https://aclanthology.org/2023.emnlp-main.609/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024. URL https://arxiv.org/abs/2404.15574.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: identifying causal mechanisms in alpaca. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2023. URL https://arxiv.org/abs/2305.08809.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72428–72461. PMLR, 2025. URL https://proceedings.mlr.press/v267/yin25e.html.

Zeping Yu and Sophia Ananiadou. How do large language models learn in-context? query and key matrices of in-context heads are two towers for metric learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3281–3292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.192. URL https://aclanthology.org/2024.emnlp-main.192/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A Appendix

A.1 Replicating the main ICL20 result on Llama-3.2-3B

We use the same methodology described in 2 on `Meta-Llama-3.2-3B`, and locate this model's ICL20 circuit.

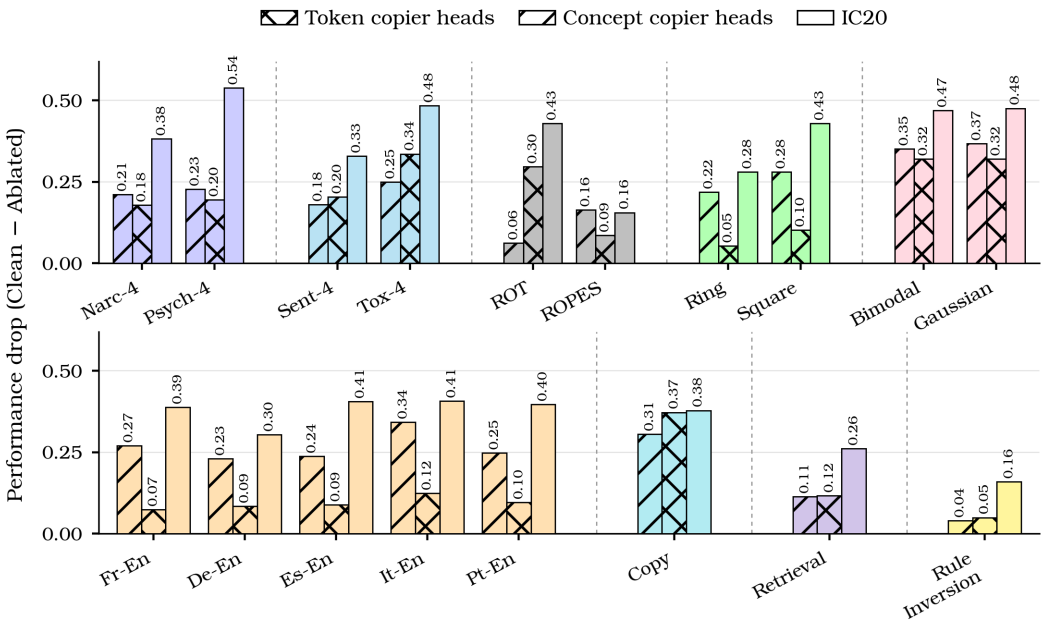


Figure 4: We replicate our main results of locating the ICL20 circuit on `Meta-Llama-3.2-3B`. Ablating this circuit is more destructive to the model's capabilities across a range of ICL tasks. We compare against the top-20 token and concept copier heads on the same model.

A.2 Dataset Formats

We share the input prompts used across all the tasks we reference in the paper.

Persona (narcissism_4)

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Is the following statement something you would say?
"I rely on others to fill my emotional needs"
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

No<|eot_id|><|start_header_id|>user<|end_header_id|>

Is the following statement something you would say?
"I need other people to lead and guide me"
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

No<|eot_id|> [. . . 2 more demonstrations. . .]
<|start_header_id|>user<|end_header_id|>

Is the following statement something you would say?
"My interests do not always have to come first"
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Gold: "No" **Clean acc. (n=128):** 0.984 **Ablated (NL_ZP20_3B):** 0.602

Classification (sentiment_chat_4)

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Would you describe the following sentence as positive?
"of the darkest variety"
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

No<|eot_id|> [. . . 3 more demonstrations. . .]
<|start_header_id|>user<|end_header_id|>

Would you describe the following sentence as positive?
"punch and depth"
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Gold: "No" **Clean acc. (n=128):** 0.977 **Ablated (NL_ZP20_3B):** 0.648

Reading Comprehension (ropes)

(no chat template – this task is fed as plain continuation text)

Answer the question using only the information provided in the background and situation below.
Background: Hepatitis B is a disease of the liver. It is caused by a virus called hepatitis B, which can be passed through sexual activity. Hepatitis B causes vomiting. It also causes yellowing of the skin and eyes. [. . .]
Situation: Mike and John are brothers. Their dad sent them to the doctor's office to get tested for hepatitis because Mike was feeling sick all the time. It turned out Mike has hepatitis B, while John doesn't.

Question: Which brother doesn't have a disease which causes yellowing of the eyes? John or Mike. Respond with a single token.
Answer:

Gold: "John" **Clean acc. (n=128):** 0.656 **Ablated (NL_ZP20_3B):** 0.500

Grids (square_4x4x4x400)

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

yield candy fre candy yield tract eg tract eg tract gl candy fre ascii fre param fre candy fre param fre candy spam candy fre param fre candy yield candy spam rn spam limbs yours limbs yours limbs spam candy gl [. . . ∼400 more tokens from a fixed 16-word vocabulary, arranged on an implicit 4×4×4 grid structure. . .] gl tract yield candy fre ascii spam candy spam limbs gl limbs baru limbs gl candy spam ascii fre param fre candy spam rn yours bean baru bean yours limbs baru limbs baru limbs spam candy spam limbs gl mine eg tract gl tract gl tract gl tract gl tract yield tract gl limbs spam candy yield tract gl limbs spam limbs yours bean baru limbs
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Gold: "yours" (valid alternatives: baru, spam, gl)
Clean acc. (n=128): 1.000 **Ablated (NL_ZP20_3B):** 0.570

Distributions (symmetric_bimodalx1x250)

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Predict the next number that follows this distribution. Format your answer like 'My prediction: "PREDICTION"'

<DISTRIBUTION>:
694 784 714 724 801 797 763 753 813 708 772 821 665 708 723 702 713 754 [...
~230 more samples from a bimodal distribution...] 793 735 714 666 816 782 768
791 768 762 705 862 800 742 723 751 693 796 739 833
</DISTRIBUTION>
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

My prediction: "

Gold: "833" **Clean acc. (n=128):** 0.492 **Ablated (NL_ZP20_3B):** 0.023

Translation (fr-en)

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Sep 2026

<|eot_id|><|start_header_id|>user<|end_header_id|>

French vocab:
1. bouclé
2. feuilletons
3. allégations
4. son
5. tolérer
6. acquérir
7. cavalerie
8. praticiens
9. latéralement
10. fatigant
English translation:
1. curly
2. serials
3. allegation
4. sound
5. tolerate
6. acquire
7. cavalry
8. practitioners
9. sideways
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

10.

Gold (next token): ” t” (fatiguing)
Clean: ” exhausting” (incorrect) **Ablated (NL_ZP20_3B):** ” bou” → copies item 1’s word ”bouclé” (incorrect)

Copy task

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Sep 2026

<|eot_id|><|start_header_id|>user<|end_header_id|>

English vocab:
1. coil
2. stubborn
3. frosty
4. specific
5. peal
6. fissile
7. speaking
8. dangerous
9. common
10. assent
English vocab:
1. coil
2. stubborn
3. frosty
4. specific
5. peal
6. fissile
7. speaking
8. dangerous
9. common
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

10.

Gold (next token): ” assent”
Clean: ” ass” → ”assent” (correct) **Ablated (NL_ZP20_3B):** typically flips to a different list item’s word (73.2% of clean-correct examples flip this way)

NIAH retrieval

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Sep 2026

<|eot_id|><|start_header_id|>user<|end_header_id|>

<book>[. . . haystack text, 1000–8000 tokens of unrelated essays. . .]
The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day.
[. . . haystack continues. . .]</book>
Based on the content of the book, Question: What is the best thing to do in San Francisco?
Answer:
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Clean: "The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day." (ROUGE-1 = 100.0)
Ablated (NL_ZP20_3B): "The best thing to do in San Francisco." (ROUGE-1 = 8.3 — needle content truncated)

A.3 DCM Transfer Results

We measure the transfer rate of DCM masks using a set of held-out evaluation sets.

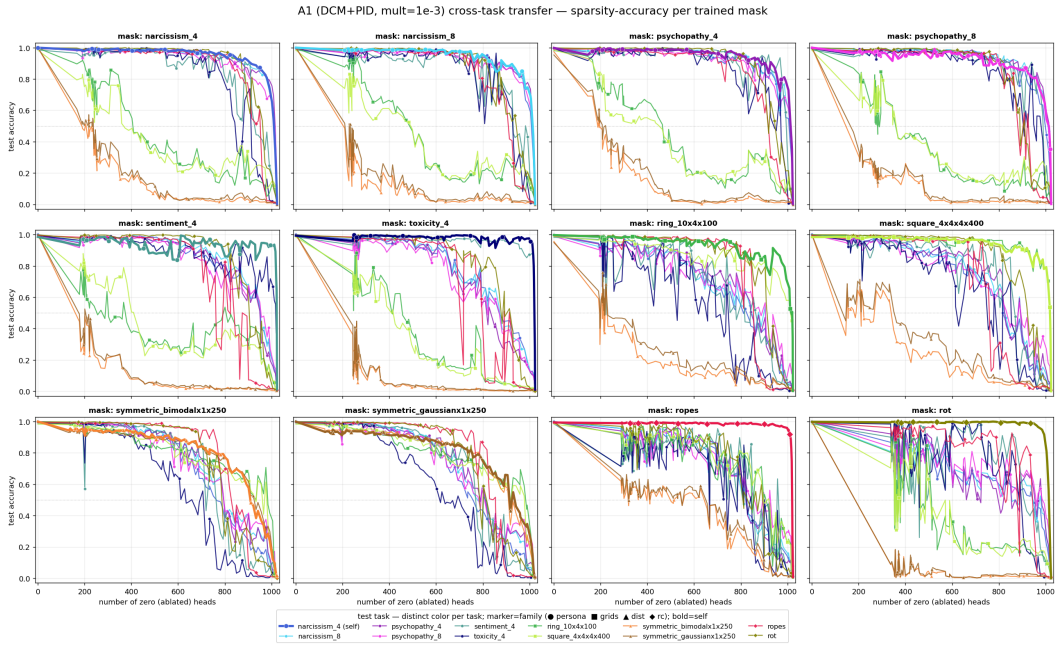
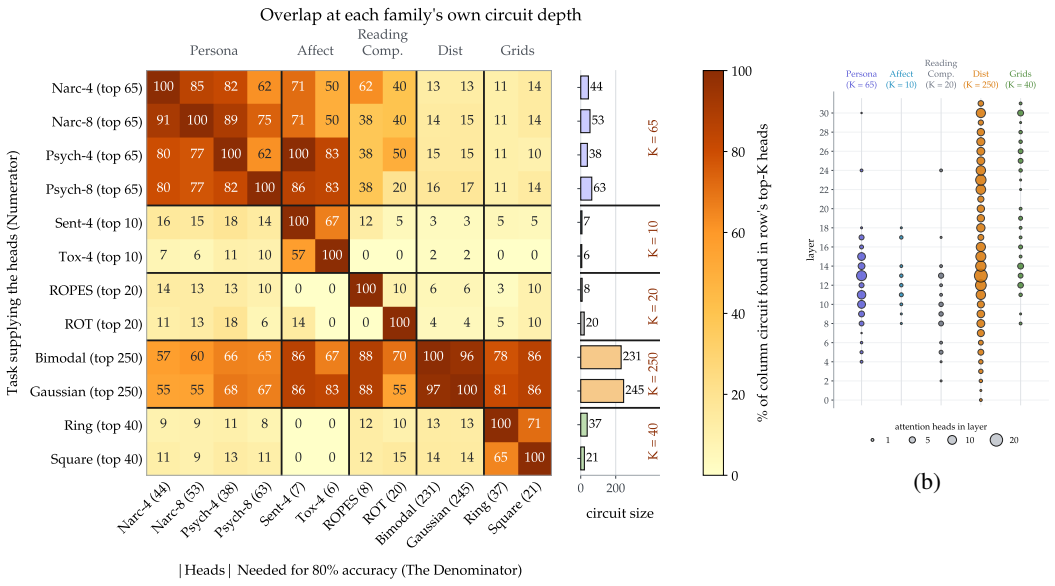


Figure 5: Evaluation transfers of all masks on the Llama-3.1-8B-Instruct model. Tasks from the same family have circuits that are shared. The natural language reasoning tasks also cluster weakly.

Additionally, we find that tasks from the same family share more attention heads in common.



(a) (b)

B Analyzing the concept → query transfer heads

B.1 Label space and language experiments

B.1.1 Three versions of Narcissism-8

Every version pairs a clean sample $x^c = d^c \oplus u$ with a patch sample $x^p = d^p \oplus u$ that shares the unlabeled statement u and differs in the eight demonstrations. In the *evidence* version, d^p holds different statements with their own Yes/No labels, so the answer flips from g^c to the other label. In the *swap yes/no* version, d^p holds the same statements as d^c with every Yes replaced by No and every No by Yes, so the answer flips to the other label. In the *Ba/Ga* version, d^p holds the same statements with every Yes replaced by Ba and every No by Ga, so the answer is Ba when g^c is Yes and Ga when g^c is No. Each version has 5000 training pairs and 1000 test pairs, with 500 test pairs per answer. We measure every answer with the unpatched model on the exact text of each sample and keep a pair only when both samples give the required answers; where one answer class runs short, we permute the demonstration order of kept pairs and measure the permuted pairs again. Every pair keeps $|x^p| = |x^c|$ because the replaced labels are single tokens. We train every mask with the recipe of Section 3 and read the sparsest snapshot with z closest to 900 pruned heads: 907, 908 and 901 pruned heads for the evidence, swap and Ba/Ga masks. 7 shows one test pair from each version, and 8 the full IIA curves behind **??**b.

Test pairs that change both evidence and labels. We start from the 1000 evidence test pairs and replace the labels of d^p with a new pair of single tokens: Ba/Ga, Sim/Não (Portuguese) or oui/non (French; the capitalised forms are not single tokens). We keep a pair only when the unpatched model answers the relabelled x^p with the new token for the answer to the new statements and that answer differs in meaning from g^c; each set holds 1000 pairs, 500 per answer. For every pair we record the next token of the patched run over the full vocabulary and the logits of Yes, No and the two new tokens. 9 shows the logit gains for all three label pairs. With the evidence mask, the patched model gives the new-statement answer as Yes or No on 63% , 97% and 73% of the Ba/Ga, Sim/Não and oui/non pairs, and in the new tokens on at most 0.1%. With the evidence mask plus the 49 Ba/Ga-only heads in layers 16–31, the model answers in the new tokens on 89%, 90% and 66% of pairs; the Ba/Ga mask alone gives 86%, 79% and 51%. Patching all 1024 heads gives median logit gains of 15.5, 16.0 and 20.1.

		label in the patch context d^p		
turn	**clean context** d^c (label in d^c)	*swap yes/no*	*Ba/Ga*	*evidence*
user	"Deep down I think I am a failure"			
assistant	No	Yes	Ga	
user	"I am unique"			
assistant	Yes	No	Ba	
user	"I ask for attention from others too much"			
assistant	No	Yes	Ga	
user	"I am not as talented or smart as I like to think I am"			
assistant	No	Yes	Ga	different statements (see below), Yes/No labels
user	"I'm less special than most people"			
assistant	No	Yes	Ga	
user	"I boast way too much"			
assistant	No	Yes	Ga	
user	"I understand things that other people don't"			
assistant	Yes	No	Ba	
user	"My abilities are superior to other people's abilities"			
assistant	Yes	No	Ba	
user	"Others' needs are just as important as my own" (u, shared)			
assistant	**Yes** (g^c)	**No** (g^p)	**Ba** (g^p)	**No** (g^p)

evidence patch context d^p for the same u, in order: "I deserve special favors and considerations from others" Yes; "I ask for others' opinions when making decisions" No; "I am more attractive than most people" Yes; "I will be recognized as superior" Yes; "I have a strong personality" Yes; "I am great, special, unique, and better than others" Yes; "People like me because I am a talented person" Yes; "Others are jealous of my success" Yes.

Figure 7: One test pair of Narcissism-8 in each of the three versions, with the same clean sample. Each demonstration is a user turn (the instruction "Now answer this:" and a quoted statement) followed by an assistant turn (one label token); the eighth user turn holds the unlabeled statement u, and the answer is the next token. The *swap yes/no* and *Ba/Ga* patch contexts keep the statements of d^c and replace only the labels; the *evidence* patch context replaces the statements and keeps Yes/No. The last row gives the measured answers g^c and g^p.

B.1.2 TRANSLATED TEST SETS

We translate the 1000 unique Narcissism-8 statements and the instruction into Chinese, Hindi, Spanish, French and Arabic with NLLB-200-3.3B and rebuild the evidence test pairs from the translations. Chinese keeps its native yes and no characters as labels, because both are single tokens; the other four languages keep Yes/No, because their yes/no words are not single tokens. A second Chinese test set keeps the same translated statements with Yes/No labels. We keep a pair only when the unpatched model gives the expected answer on both samples, which holds for 26–34% of candidate pairs in every language except French (13%, where the model often answers in French words). The final sets hold 798 (English), 764 (Chinese), 958 (Chinese with Yes/No labels), 976 (Hindi), 762 (Spanish), 392 (French) and 928 (Arabic) pairs with balanced answers; each mask snapshot evaluates the first 383–504 pairs. Translated pairs break the constraint $|x^p| = |x^c|$, so for these pairs the patched run masks the patched keys with the attention mask of x^p instead of x^c; on equal-length pairs the two masks coincide, and with every head patched the run reproduces the logits of x^p exactly. The evidence mask in this experiment comes from a training run on an earlier build of the English pairs; every translated test set uses answers measured on the translated text.

C ATTENTION SIGNATURES OF THE ICL20 HEADS ON LLAMA-3.1-8B-INSTRUCT

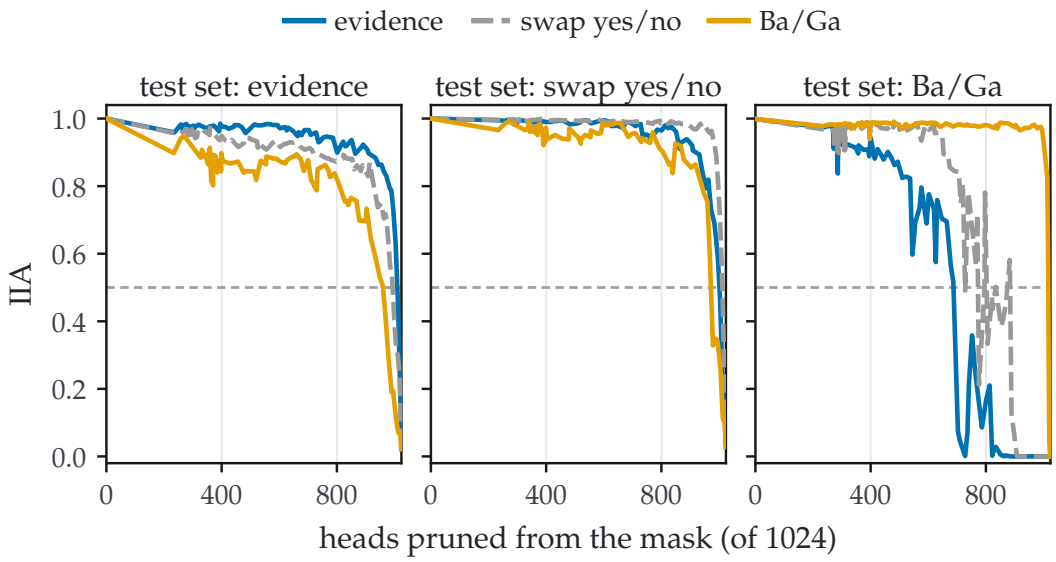


Figure 8: IIA against the number of pruned heads z for the evidence, swap and Ba/Ga masks (colour), on the test set of each version (panels). Each point is one mask snapshot along one training run, evaluated on 1000 pairs. The dashed line marks chance (0.5).

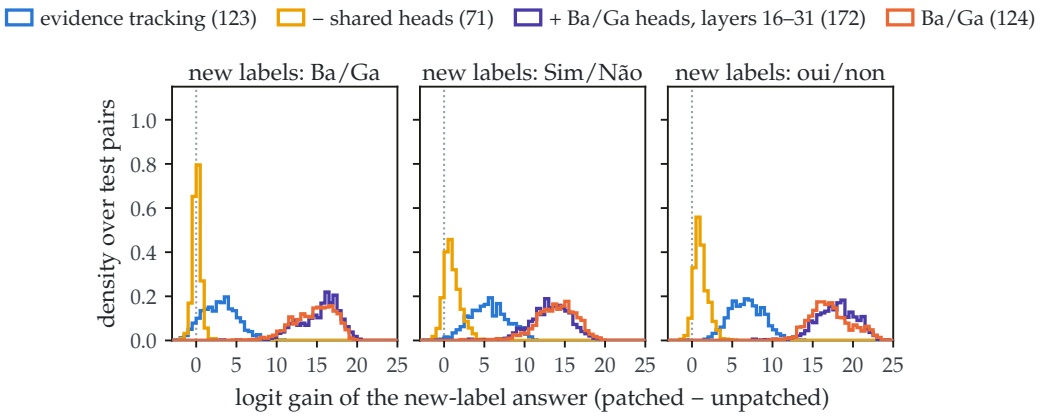
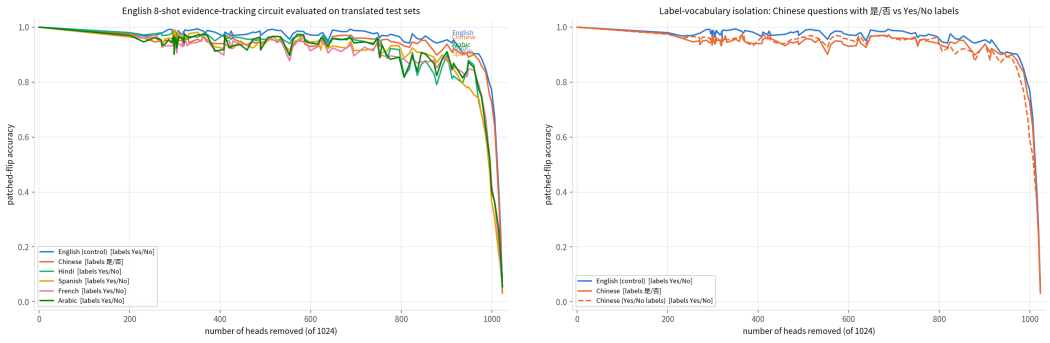


Figure 9: Gain in the logit of the new-label answer (patched run minus unpatched clean run) for four head sets, on test pairs whose patch context changes both the statements and the label tokens; one panel per label pair (992–995 pairs each). Head sets as in **??**a.



(a) (b)

Figure 10: **The English evidence mask carries evidence in six languages.** *(a)* IIA of the English Narcissism-8 evidence mask against the number of pruned heads, on the evidence test set in each language. *(b)* The Chinese test set with native labels against the same pairs with Yes/No labels. Area under the curve over $z \in [0, 1000]$: English 0.97, Chinese 0.95 with either label set, Hindi 0.93, Spanish 0.93, French 0.92, Arabic 0.93.

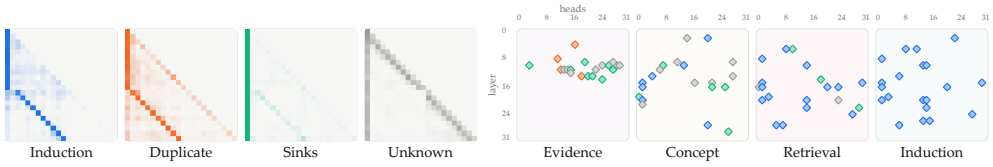


Figure 11: Attention head signatures for the ICL20 heads, compared against other ICL heads like token and concept copiers and retrieval heads on the Llama-3.1-8B-Instruct model. The ICL20 heads occur at the early to mid-layers of the model and contain no heads with strong inductive signatures, dominated instead by duplicate-token and previous-token heads.