THE FUNCTIONAL ORGANIZATION OF ATTENTION HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

A Transformer spreads its computation over many attention heads, and interpretability work now ties individual heads to specific functions. Induction heads copy repeated patterns, retrieval heads fetch information from long contexts, and truthfulness heads separate true from false answers. These labels are used as if they transferred across models, settled early in training, and could be turned up at will. Each label, though, was found in one model, with one detector and one intervention. This paper tests the assumptions together. We use six head-level signals, five base models, two Pythia pretraining trajectories, and matched-control interventions on Qwen, Llama, and Mistral. One answer recurs: a repeatable signal fixes a head type's coarse structure, not its fine identity. (1) Distribution: local-pattern heads precede induction heads in every model. No complete layer order is shared. (2) Attribution: induction and retrieval detectors pick largely the same heads, 20 and 19 of 30 on Qwen and Llama. The retrieval cost of an "induction" mask falls on those shared heads, not on either label's exclusive heads. (3) Formation: detector strength saturates thousands of steps before top-head membership settles. Fine-tuning changes membership without moving layer distributions. (4) Control: patching the selected retrieval heads restores most of the retrieval loss, and the labels improve KV-cache allocation. Doubling a head's output does not improve its task. A label therefore transfers three things: the relative order of head types, the task contribution of the selected set, and a ranking for memory allocation. It does not transfer absolute depth, an exclusive function, head membership, or gains from amplification. Reuse a label's coarse structure, and re-derive its membership and coefficients for each model.

1 INTRODUCTION

Interpretability research on language models has progressed from probing representations to naming the parts. Individual attention heads in a Transformer (Vaswani et al., 2017) now carry functional labels. Previous-token and duplicate-token heads track local structure (Elhage et al., 2021; Wang et al., 2023). Induction heads copy repeated sequences (Olsson et al., 2022; Chen et al., 2024), and copy-suppression heads damp that copying (McDougall et al., 2024). Retrieval heads fetch information from long contexts (Wu et al., 2025), function-vector heads drive in-context learning (Yin & Steinhardt, 2025), and truthfulness heads separate true from false answers (Li et al., 2023). Heads are convenient units: each is detected by a score, and each can be suppressed, replaced, or scaled without retraining. The labels are already in use. Retrieval heads are described as universal across models (Wu et al., 2025), their scores decide how much long-context memory each head keeps (Fu et al., 2025), and truthfulness heads are steered at inference time (Li et al., 2023).

Each label, however, was established one at a time, in one model, with one detector and one intervention. A catalogue of parts is where a mechanistic account begins, not where it ends. Neuroscience makes the same move after a cell-type atlas. It asks where each type sits, whether two markers label the same cells, when the types develop, and whether driving a type changes behavior, because those answers decide what the atlas can be used for. For attention heads these questions are open. Whether heads of one type sit at the same depth in different models is unknown. So is whether the effect of ablating a labeled head belongs to that label when the same head carries another, whether heads form in the order they depend on one another, and whether strengthening a detected head strengthens its function. Every practical use above assumes an answer.

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

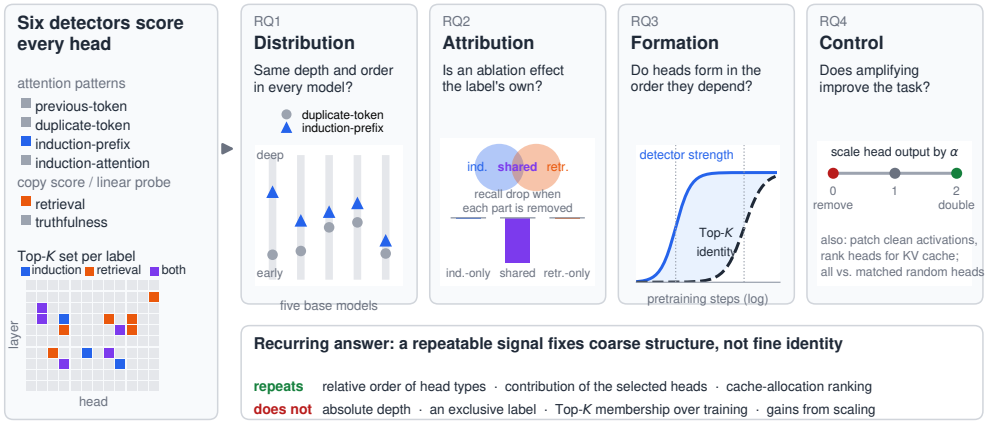

Figure 1: **Overview.** Six detectors score every head and select a Top-K set per label (left). The sets are tested with four questions, each against matched random heads (center), and the same answer recurs (bottom).

We ask the four questions together, with one detector suite and matched random controls:

1. **Distribution.** Early layers are assumed to hold local patterns and later layers the computations that compose them. Do five head types sit in that order, and at the same depth, in every model?
2. **Attribution.** A functional head is validated by ablating it on the matched task. Can the effect be attributed to the label when the same heads also carry another label?
3. **Formation.** If heads depend on one another, the heads that are read should form first. Does formation follow that order, when does head identity settle, and does fine-tuning move it?
4. **Control.** Steering assumes that strengthening a detected head strengthens its function. Does amplification, tested against matched controls, improve the matched task?

Together they form a systematic study of functional heads: six head-level signals on five base models, two Pythia pretraining trajectories, supplied Pythia-6.9B pretraining and supervised fine-tuning (SFT) checkpoints, and interventions on Qwen, Llama, and Mistral. The detectors measure specific patterns or predictive signals rather than assume one exclusive function per head. Induction-prefix matching on repeated sequences is kept separate from an induction-attention pattern on general prompts. RQ1 and RQ3 use the four attention detectors and the truthfulness probe; RQ2 uses induction-prefix and retrieval; the held-out test of RQ4 uses retrieval, induction-attention, and truthfulness. Coverage differs across questions, and a two-model result is never treated as a five-model finding (Table 1).

Our contributions are the answers, one per question. *Distribution* (Section 3.2): detector rankings are highly repeatable within a model, and local-pattern heads precede the induction heads that read them in every model, but no complete layer order is shared. Absolute depth does not transfer. *Attribution* (Section 3.3): induction-prefix and retrieval detectors select largely the same heads, and the retrieval cost of removing either label's heads is carried by the shared heads rather than by the heads exclusive to either label. A label is not an exclusive function. *Formation* (Section 3.4): detector strength saturates thousands of steps before top-head membership settles, and fine-tuning changes membership without moving layer distributions. A head's function is stable long before its identity is. *Control* (Section 3.5): patching the selected retrieval heads recovers most of the retrieval loss, and the labels rank heads usefully for KV-cache allocation, but doubling whole-head outputs does not improve the matched task. A label identifies what to keep rather than what to turn up. The same pattern recurs across the four answers: a repeatable signal fixes coarse structure but not fine identity. The practical rule that follows is to reuse a label's coarse structure and to re-derive its membership and coefficients for each model.

2 What We Measure and How We Test It

We use *functional head* as shorthand for a head that scores highly on a specified detector. This label does not mean that the head has only one function, or that the detected pattern is necessary for a behavior. We therefore measure three things separately: the pattern detected in a head, whether the same heads are selected across settings, and what happens when their outputs are changed. Appendix A defines the head output; Appendix C.1 gives the detailed checks used to limit each conclusion.

2.1 Six head-level signals

Four detectors score local attention patterns. For a prompt $p = (t_0, \ldots, t_{T-1})$, let $A^p_{\ell,h}[i, j]$ be the attention weight from position i to position j. A binary mask P^p_c marks the positions targeted by detector c. The score is the average fraction of attention assigned to these positions:

$$s_c(\ell, h) = \frac{1}{|\mathcal{P}_c|} \sum_p \frac{\sum_{i,j} A^p_{\ell,h}[i, j] P^p_c[i, j]}{\sum_{i,j} A^p_{\ell,h}[i, j]}. \tag{1}$$

Previous-token marks the preceding token; *duplicate-token* marks earlier occurrences of the current token. *Induction-prefix* uses sequences of unique tokens repeated four times and marks the token after the previous occurrence (Olsson et al., 2022; Crosbie & Shutova, 2025). *Induction-attention* instead tests the pattern $t_i = t_{j-1}$ on general prompts. These two induction detectors use different inputs and are reported separately.

The remaining detectors use different signals. *Retrieval* measures copying from a supplied context in grid cells where retrieval succeeds (Wu et al., 2025). *Truthfulness* is a linear probe on final-token head activations on TruthfulQA (Li et al., 2023; Lin et al., 2022), repeated over three split seeds; it measures predictability, not the effect of an intervention. Exact definitions are in Sections A and C.4; raw scores are compared only within a detector.

2.2 Layer position, repeatability, and overlap

For each detector we select the K highest-scoring heads, $\mathcal{H}^K_c$, and record their mean relative layer position

$$\mu_c = \frac{1}{K} \sum_{(\ell,h) \in \mathcal{H}^K_c} \frac{\ell + 0.5}{L}, \tag{2}$$

with L the number of layers. Cross-model comparisons use this normalized depth and the relative order of two types, never raw layer indices. We also compare eight-band depth histograms by cosine similarity, which needs no head-to-head alignment across architectures.

Within a model, Spearman correlation compares whole-head rankings and Jaccard overlap compares selected sets,

$$J_K(a, b) = \frac{|\mathcal{H}^K_a \cap \mathcal{H}^K_b|}{|\mathcal{H}^K_a \cup \mathcal{H}^K_b|}. \tag{3}$$

both across repeated detections and across labels; overlap is checked against random sets that preserve per-layer head counts. The null calibrations were added after the original results and are exploratory (Section D.1).

2.3 Testing behavioral effects

Changing a detector score is not the same as changing behavior, so we measure both. Detector dependence masks a source set, rescores a target detector on the unmasked heads, and subtracts the change under matched random sets:

$$I(a \to b) = \Delta s_b(\mathcal{H}^K_a) - \mathbb{E}_r[\Delta s_b(\mathcal{R}^K_{a,r})]. \tag{4}$$

Behavior is measured on induction, retrieval, or truthfulness tasks under zero ablation, mean ablation on a calibration set, or clean-to-corrupted patching. These operators answer different questions, and their effect sizes are not interchangeable.

Each selected set is compared with five random sets matched on per-layer head counts and kept clear of the detectors' Top-50 union. Intervals that resample only these five controls describe variation over controls, not over examples or head selection; we state which is which with each result.

A separate experiment selects Top-30 heads first and then evaluates them on held-out prompts, retrieval needles, and 128 TruthfulQA questions at $\alpha \in \{0, .5, 1, 1.5, 2\}$, where $\alpha = 0$ removes the selected output. The controls are five layer-matched random sets and five sets matched on activation size before W_O; neither includes the selected heads.

The unchanged model and the two control types give 18 Holm-corrected comparisons across six model–detector runs. Retrieval uses an exact sign test over 12 needles; induction and truthfulness use sign-flip tests with 10,000 draws. Heads, inputs, and interventions were fixed in advance, but the final statistical rule was reconstructed after the outcomes, so these results are exploratory. They cover retrieval, induction-attention, and truthfulness, not induction-prefix.

2.4 When a signal becomes strong and when its heads stabilize

We track detector strength and head membership along pretraining checkpoints, then compare supplied pretraining and SFT endpoints. For checkpoint t, let $m_c(t)$ be the maximum detector score across heads. We record the first sampled checkpoint where it reaches half its final value:

$$t_c^{1/2} = \min\{t : m_c(t) \geq 0.5 m_c(T)\}. \tag{5}$$

This is a strength threshold, not the first moment a capability appears. We do not interpret it when the final signal is zero or when the probe has a high untrained baseline.

Head identity is said to have settled at the first checkpoint after which the ranking's Spearman correlation with the final checkpoint stays at least .8 and Top-30 Jaccard stays at least .5. Ranks are computed over heads active at either checkpoint. Both are threshold summaries of a sampled trajectory, and checkpoint spacing limits how sharply a rise can be dated.

2.5 Testing whether stronger head outputs improve the task

We multiply selected head outputs, before W_O, by $\alpha \in \{0, .5, .75, 1, 1.25, 1.5, 2\}$ and test $K \in \{5, 10, 20, 30, 50\}$. Here $\alpha > 1$ strengthens the selected outputs, $\alpha = 1$ leaves them unchanged, and $\alpha < 1$ suppresses them. Improvement is a positive change in the matched task metric against the unchanged model, with matched-head controls as a specificity check. A separate performance screen asks only whether a setting avoids large losses: matched and transfer metrics within 2%, one target metric non-decreasing, NLL within 5%, and control tasks within 5%. We report consecutive passing values around $\alpha = 1$. Passing the screen is not improvement.

A secondary, retrospective test asks whether a coefficient chosen on one model can be reused on another, against fixed coefficients that use no source information (Appendix G.4).

3 Experiments

3.1 Setup and scope

RQ1 compares Qwen2.5-7B, Llama-2-7B, Llama-3.1-8B, Mistral-7B, and Pythia-6.9B. The overlap, patching, and held-out tests of RQ2 and RQ4 use Qwen2.5-7B-Instruct and Llama-2-7B-Chat; the three-way ablation of RQ2 uses Llama-3-8B-Instruct, Mistral-7B-Instruct, and Qwen2.5-7B-Instruct; the KV-cache control uses Llama-3-8B-Instruct. RQ3 follows Pythia-1.4B and Pythia-2.8B checkpoints and compares supplied base and tuned models. No single model enters every question, and no finding is pooled across questions as if one did.

Repeated detection uses different input samples (three disjoint sets of 203 prompts; three blocks of five synthetic sequences for induction-prefix), not independently trained models. Unless stated otherwise we select Top-30 heads and use five layer-matched control sets (seeds 42–46). The held-out experiment fixed selection, inputs, and controls before intervention but reconstructed its statistical rule afterward, so it is exploratory. Data sizes, grids, and hardware are listed in Section C.

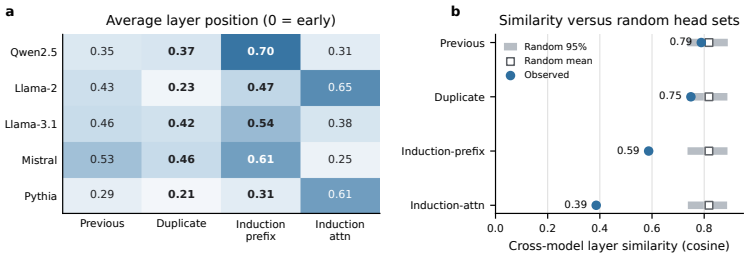

Figure 2: **Some layer relationships repeat, but the full ordering changes.** (a) Average relative layer position of the Top-30 heads, over three detector-input samples. Duplicate-token heads are earlier than induction-prefix heads in every model. (b) Cross-model layer similarity compared with random Top-30 locations that preserve model geometry. Squares and bars show the random mean and 95% range from 100,000 shared-coordinate permutations; no observed similarity exceeds it (post-hoc).

3.2 RQ1: Do layer locations and relative order repeat across models?

A widely held picture of depth in language models is that early layers handle local, surface patterns and later layers handle computations that compose them or read task-level information (Tenney et al., 2019; Jin et al., 2025; Lad et al., 2024). That picture comes from probes on the residual stream; whether it holds at the level of attention heads, and across models, has not been tested. We test it with the positions of five head types in five models. Previous-token and duplicate-token heads read only local patterns and should be earliest. Induction heads read the output of a previous-token head (Elhage et al., 2021) and should follow. Retrieval and truthfulness heads read long-context or task-level information and should be deeper. If the division of labor is structural, the order should hold in every model.

A detector's ranking is reproducible within a model, but its Top-K set need not be. Rankings and Top-30 sets of the four attention-pattern detectors agree closely between input samples on all five models, and no Top-30 depth centroid moves appreciably (Table 9). The truthfulness probe separates the two properties: its rankings agree while its selected sets do not, so a stable score does not fix membership at the cutoff.

Two relative orders survive across all five models. Duplicate-token heads sit earlier on average than induction-prefix heads for every tested $K \in \{10, 20, 30\}$ and detector-input sample, and previous-token heads precede induction-prefix heads at $K = 30$, though not in every cell. Only the second order carries a mechanistic reason, since the composition requires an induction head to sit above the previous-token head it reads (Elhage et al., 2021); the centroid comparison is consistent with it on average rather than establishing it.

Beyond those orders, depth does not transfer. No complete ordering is shared (Figure 2 and Table 6). Induction-attention heads are early in Qwen2.5, Llama-3.1, and Mistral but late in Llama-2 and Pythia, and retrieval and truthfulness heads sit at mid-depth rather than at the end of the network (Table 7). Cross-model cosine similarity of depth histograms is no higher than what random Top-30 locations produce under the same model geometries, and only a label-shuffling test, the second of two post-hoc checks, retains a small signal (Tables 10 and 11). What repeats is thus an order between two head types, not a depth: a label constrains which heads read which, and not where they sit.

3.3 RQ2: Can an ablation effect be attributed to a functional label?

The standard test of a functional head masks the detector's top-k heads, compares with random heads, and reads the drop on the matched task. That shows the heads matter; it does not show they carry the labeled function, because one head can score highly on several detectors. Induction-prefix and retrieval share about two thirds of their Top-30 heads on Qwen2.5 and Llama-2, so a retrieval drop under an "induction" mask may come entirely from the shared heads. We ask whether an ablation effect can be attributed to a label at all, using the archived overlap and interventions, a patching experiment whose rule was fixed in advance, and separate masks for the shared and label-exclusive heads.

The overlap is not a by-product of a shared layer preference. Top-30 Jaccard between the two detectors exceeds layer-matched random overlap by an order of magnitude, and the two rankings also agree within layers (Table 13). Two labels on one head still do not establish two functions: the pair was chosen after inspecting results, and the executed masks of the historical interventions were not saved.

The selected heads are causally involved in both matched tasks. Under zero and mean ablation and under clean patching, all 12 archived induction effects move in the expected direction (Table 14), although their paired outputs are unavailable, which makes them supporting rather than confirmatory. For retrieval, a pre-registered experiment patches the clean activations of the frozen Top-30 retrieval heads into a corrupted run on 36 token-aligned needle pairs per model. The patch recovers most of the clean–corrupt target log-probability gap on both Qwen2.5-Instruct and Llama-2-Chat, ablating the same heads costs more than either family of matched random sets, and all fourteen intervals exclude zero (Section M). The recovery is partial, and the same ablation also lowers induction log-probability and fixed-corpus NLL, so these heads carry a matched-task contribution on both models without carrying it exclusively.

Because the two labels select overlapping sets, the archived masks cannot separate them: masking the induction Top-30 and masking the retrieval Top-30 cost retrieval ROUGE-1 about equally relative to controls (Table 12). We therefore remove the induction-only, shared, and retrieval-only heads separately, with the $\alpha = 0$ operator of RQ4, on Llama-3-8B-Instruct, Mistral-7B-Instruct, and Qwen2.5-7B-Instruct, against five layer-matched random sets per part (Section L).

The effect is confined to the heads both detectors select. Removing either label's exclusive heads leaves needle recall within or above the random controls on every model and grid. Removing the shared heads costs Llama-3 twenty-five points against its controls, and costs more as the context grows. Mistral falls on both grids, with intervals that touch zero. Qwen is insensitive on this needle, although its recall does fall under the same operator on the held-out needles of RQ4. Adding an exclusive set back to the shared set moves recall by a few points in either direction. In the reverse order, the shared set costs as much after an exclusive set is removed as it does alone (Table 29), and the pattern is unchanged at $K = 20$ and $K = 40$ (Table 30). An ablation effect therefore cannot be credited to either label. The conditional effects are not exactly zero, so the exclusive heads may act as partial backups. The clean effect is also limited to one model, and exclusivity is defined at the cutoff of a single detection sample.

3.4 RQ3: Do heads form in the order they depend on each other?

If functional heads depend on one another, as RQ1 and RQ2 suggest, the heads that are read should form before the heads that read them. Formation order is therefore a consistency check on that dependence. Prior work tracked the emergence of induction heads and 1-back heads separately (Olsson et al., 2022; Trott, 2025), but did not track several types together or separate a signal's strength from the identity of its top heads. We track both for five head types along two Pythia pretraining trajectories and ask whether formation follows the dependence order, when head identity settles, and whether fine-tuning changes it.

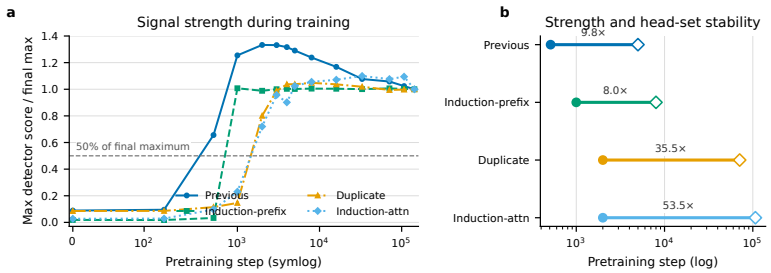

Figure 3: **Signal strength rises before head rankings and membership settle in one Pythia trajectory.** (a) Maximum detector scores at 14 Pythia-1.4B checkpoints, divided by their final values. Values above one are temporary overshoots. (b) Filled circles mark the first half-final crossing; open diamonds mark the first checkpoint after which both ranking and Top-30 agreement meet the stability thresholds at every later sampled checkpoint. Labels give the ratio of the two step indices, not a ratio of learning speeds.

Detector strength and head identity settle at different stages of pretraining. Strength crosses half-final in the order the dependence predicts: previous-token first, then induction-prefix, and duplicate-token, on which induction does not depend, last. Sustained agreement with the final ranking and Top-30 set arrives one to two orders of magnitude later in step index (Figure 3 and Table 16). The separation is robust to the choice of strength and agreement thresholds, although exact times and ties change (Tables 17 and 18). These crossings are reference levels rather than first appearances, and fourteen sampled checkpoints cannot distinguish an abrupt rise from gradual growth.

The observed order is consistent with dependence but does not isolate it. It is also the order of pattern simplicity, and the step at which induction heads emerge is predictable from batch size, context length, and the bigram statistics of the pretraining data (Aoyama et al., 2026). A second trajectory, six checkpoints of Pythia-2.8B scored with the same detectors and thresholds, reproduces the separation between strength and identity. It does not reproduce the order, because previous-token and induction-prefix strength already exceed their final values at the first scored step (Section J). The two runs share data order, batch size, and context length, so the second replicates the measurement rather than the training recipe. Retrieval is excluded from temporal ordering because the successful examples used to score it change by checkpoint, and the truthfulness probe's high untrained baseline is not read as a formed capability.

Fine-tuning moves head membership without moving the layer distribution. In Qwen base-to-instruct and Llama base-to-chat comparisons, the layer distribution is nearly unchanged while the selected heads differ substantially (Table 15). In the supplied Pythia parent and Dolly-SFT pairs, the induction effect at $K = 30$ shrinks under zero and mean ablation but grows under clean patching (Table 19), so the answer depends on the operator. Without a complete training recipe these pairs do not isolate the effect of SFT. A detector therefore reaches its final strength long before it settles on a set of heads, and a later fine-tune can replace that set while the distribution stays put.

3.5 RQ4: Does amplifying functional heads improve their tasks?

Steering methods assume that strengthening a detected head strengthens its function (Li et al., 2023). The usual validation is ablation, which tests necessity; amplification is rarely tested against matched controls. We test two operators: scaling whole selected heads by $\alpha > 1$ against layer-matched and activation-matched random controls, and the directional shift that steering methods apply. In each case we ask whether the matched task improves, and we note where a task is already at ceiling and cannot.

Scaling a whole head's output does not improve its task. At $\alpha = 2$ in the held-out test, one of six comparisons has a positive point estimate with an interval excluding zero; the remaining five are small and negative (Section G.1). The comparisons use different metrics, and the Qwen retrieval task begins at ceiling, leaving no room to improve. This is one local gain at one coefficient on two models. It does not rule out gains at other coefficients or on tasks with headroom, but it does rule out the assumption as a general rule.

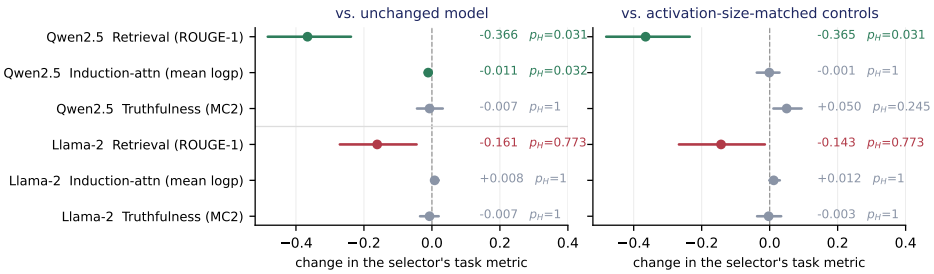

Figure 4: **Only the Qwen retrieval set passes every held-out comparison.** Change in each detector's task metric after removing its pre-selected Top-30 heads, against the unchanged model (left) and activation-size-matched random controls (right); bars are bootstrap 95% intervals, labels give the Holm-corrected sign-test p_H. Green passes ($p_H < .05$); red is the retrieval set that does not; grey detectors are not significant. Negative is worse. Full values in Table 27.

A directional intervention helps on one of the two models. Truthfulness ITI, which shifts activity along a learned direction rather than scaling the whole head, raises both Llama truthfulness scores against random directions at the same setting, while Qwen falls on all three metrics (Figure 7 and Table 22). The coefficient is not norm-normalized and other settings are untested, so the asymmetry between the two models is described rather than explained.

Necessity and benefit come apart on the same head set. Removing the selected retrieval heads lowers needle recall on both models. On Qwen the drop separates from the unchanged model and from both control families; on Llama it does not pass the corrected tests (Figure 4). The pre-registered patching experiment of Section 3.3 finds the same heads sufficient and necessary on Llama under a log-probability measure, so the null is specific to the recall metric and the sign test. Induction-attention is less harmful than layer-matched controls but not clearly different from the activation-size controls, and truthfulness results are inconsistent. This is exploratory evidence about a preselected set, not a test of the difference between models. Its analysis rule was reconstructed after outcomes, and activation-size matching applies before W_O rather than to the residual contribution (Section C.5). Passing a loss-tolerance screen is weaker still: fewer than half of the non-unit conditions qualify on either metric scale, no coefficient passes in all 60 cases, and a coefficient chosen on another model never beats a fixed $\alpha = .75$ (Figure 6 and Tables 23 and 24).

A label remains useful as a ranking where it fails as a control. We feed each head score into the head-level KV-cache allocation of HeadKV (Fu et al., 2025), with its allocation rule fixed and only the importance signal changed, on six LongBench tasks. At the tightest budget our induction-prefix signal outranks HeadKV's own retrieval signal, and the retrieval–induction union of RQ2 falls between that and HeadKV's purpose-built reasoning signal. Previous-token heads score below a random allocation (Figure 8). The ordering narrows as the budget grows, and HeadKV's retrieval signal overtakes induction-prefix. Detection therefore identifies heads worth keeping rather than a dial worth turning.

4 Related Work

Previous work connects attention heads to induction (Olsson et al., 2022), copy suppression (McDougall et al., 2024), long-context retrieval (Wu et al., 2025), and truthfulness (Li et al., 2023). Heads differ in importance under pruning (Voita et al., 2019; Michel et al., 2019), and attention patterns are not causal roles (Jain & Wallace, 2019). Recent studies assign several roles per head and test them by intervention (Ma et al., 2025; Nam et al., 2025; Su et al., 2026), and cross-architecture studies ask whether mechanisms correspond (Wang et al., 2025; Bhardia et al., 2026). We compare several functional categories across models, training stages, and interventions.

Trott (2025) discusses how mechanistic findings generalize, and Bali et al. (2026) studies attention-head stability across refits. We connect four forms of stability that need not agree: repeatable scores, similar layer distributions, stable selected sets, and consistent intervention effects. Sparse autoencoders give a finer view of attention outputs (Kissane et al., 2024; Marks et al., 2025; Thasarathan

et al., 2025) and could separate the features behind a doubly labelled head; we do not evaluate that extension.

5 Discussion, Limitations, and Conclusion

A repeatable label licenses coarse claims and not fine ones. It identifies heads that carry the matched task. Masking them costs the task, patching them back restores most of the loss on both tested models, and ranking heads by the label tells a memory allocator which heads to keep. It fixes the relative order of local-pattern heads and the induction heads that read them. It does not fix where a head type sits in absolute depth or which heads make up the Top-K set as training or fine-tuning proceeds. Nor does it fix which of two overlapping labels a head belongs to, or whether scaling the selected heads improves anything.

A label therefore transfers as a procedure rather than as a set of heads. Running the detector again on the target model costs little and is the only step that recovers the right heads, since membership moves under fine-tuning and across models. Where a second label describes a nearby task, its overlap with the first should be measured before an ablation effect is credited to either, and the shared and exclusive parts reported separately. A ranking transfers even where a set does not, which is what an allocator needs. Coefficients have to be re-derived on the target model, and an amplification has to be checked against matched random heads on a task with room to improve, because necessity under ablation says nothing about benefit under scaling.

Several limitations apply. The detectors measure selected attention patterns or probe predictiveness, not every computation a head performs. Five base models enter the layer comparison, two tuned families the interventions, and two Pythia runs the timing analysis, with the denser run sampled 14 times. The three-way ablation uses one detection sample of three instruct models, with the cutoff varied on one of them; the amplification test uses one coefficient. Supplied base and tuned checkpoints do not isolate SFT. Some archived experiments lack executed head masks or per-example outputs, the held-out study reconstructed its statistical rule after outcomes, and coefficient reuse was assessed retrospectively.

Functional labels for attention heads are reproducible enough to be useful and too loose to be mechanistic identities. Across four assumptions that current practice already relies on, the same boundary appeared: what a detector fixes about a head type survives the move to another model, another checkpoint, or another intervention, while the identity of the heads it selects does not. Reading a label as a procedure to repeat, rather than a set of heads to reuse, keeps the part that transfers and discards the part that does not. Two experiments would sharpen the boundary. Running all four questions on a single model would show whether one head population carries the coarse structure throughout, and sweeping a directional coefficient on a task with headroom would test amplification where a ceiling cannot hide it.

ETHICS STATEMENT

This work uses publicly available model checkpoints and evaluation datasets. It does not involve human subjects or private data. Head-level interventions can help diagnose or edit models, but a functional label does not guarantee a particular behavior or a safe intervention. We report negative results and the limits of our controls. Passing the performance checks in this paper is not evidence that a model is safe to deploy.

REPRODUCIBILITY STATEMENT

Appendix C lists model revisions, data splits, intervention settings, statistical tests, and checks on the saved results. The anonymized supplement contains the available full revisions, individual intervention records, result tables, and code to reproduce the 18 held-out comparisons, the coefficient-reuse analysis, and the five paper figures. Running the models again requires the external weights and datasets listed there.

AI USE STATEMENT

Generative AI tools assisted with conceptual framing, hypothesis refinement, experiment-design feedback, analysis and plotting-code implementation and review, recomputation from author-supplied frozen artifacts, preliminary interpretation, literature search, and manuscript drafting and clarity editing. They did not generate the study's synthetic prompts or evaluation datasets, produce the reported model outputs, construct or prove theoretical results, translate text, clean datasets, or perform qualitative coding. Synthetic prompts were produced by author-specified deterministic generators. AI tools did not replace any recorded resampling or multiple-testing procedure. The authors reran deterministic analyses, checked numerical claims against frozen artifacts within their stated provenance scope, reviewed AI-assisted text and code, and take responsibility for the final content.

REFERENCES

Tatsuya Aoyama, Ethan Gotlieb Wilcox, and Nathan Schneider. Predicting the emergence of induction heads in language model pretraining. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2511.16893.

Karan Bali, Jack Stanley, Praneet Suresh, and Danilo Bzdok. Quantifying LLM attention-head stability: Implications for circuit universality, 2026. URL https://arxiv.org/abs/2602.16740.

Arya Bhardia, Julian Ramirez, Siddhanta Verma, and Karen Mkrtchyan. Mechanistic analysis of universality: Numerical comparison circuits across transformer architectures. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 951–967, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-srw.84. URL https://aclanthology.org/2026.acl-srw.84/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *Advances in Neural Information Processing Systems*, 37:66479–66567, 2024.

Joy Crosbie and Ekaterina Shutova. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5049–5111, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-naacl.283/.

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

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.19258.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, 2025. URL https://arxiv.org/abs/2404.07066.

Shruti Joshi, Aaron Mueller, David Klindt, Wieland Brendel, Patrik Reizinger, and Dhanya Sridhar. Causality is key for interpretability claims to generalise, 2026. URL https://arxiv.org/abs/2602.16698.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024. URL https://arxiv.org/abs/2406.17759.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. The remarkable robustness of LLMs: Stages of inference?, 2024. URL https://arxiv.org/abs/2406.19384.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Zezheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims, 2026. URL https://arxiv.org/abs/2605.08012.

Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Monazam Erfani, Tongliang Liu, and James Bailey. Cognitive mirrors: Exploring the diverse functional roles of attention heads in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/22813e26903a56a0966a3d9c45e139ca-Abstract-Conference.html.

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

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Miami, Florida, US, November 2024. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a7f530e11fa19e9551b7a51dbd0f336f-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Jingtong Su, Julia Kempe, and Karen Ullrich. From concepts to components: Concept-agnostic attention module discovery in transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2b25c39788e5cf11d3541de433ebf4c0-Abstract-Conference.html.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59304–59325. PMLR, 2025. URL https://proceedings.mlr.press/v267/thasarathan25a.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Sean Trott. Toward a theory of generalizability in LLM mechanistic interpretability research, 2025. URL https://arxiv.org/abs/2509.22831.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2J18i8T0oI.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

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.

A Background and Problem Setup

Let $X \in \mathbb{R}^{T \times d}$ be the residual-stream input to an attention layer. The contribution of head h is

$$O^h(X) = \mathrm{softmax}\left(\frac{XW_Q^h(XW_K^h)^\top}{\sqrt{d_h}} + M\right) XW_V^h W_O^h, \tag{6}$$

Multi-head attention adds these contributions. The attention matrix describes *where* a head reads, while $W_V^h W_O^h$ determines *what* it writes to the residual stream. Each O^h can be changed before this sum, letting us suppress, replace, patch, or scale selected query heads while leaving the rest of the model unchanged.

We call a head *functional with respect to detector* c when it has a high, repeatable score $s_c(h)$ on prompts designed to measure computation c. This is a measurement-based label. A high score alone does not show that the head is needed for a behavior. Across the four questions in Section 1, we distinguish three kinds of evidence:

1. **Observations:** repeated detector scores, depth distributions, and agreement between Top-K head sets.
2. **Detector changes after intervention:** whether changing one set of heads changes another measured signal. This can test a directed dependence, but does not establish a complete circuit.
3. **Behavior changes after intervention:** performance changes under ablation, patching, or scaling, compared with stated controls. These test sensitivity to the selected heads and whether a scaling coefficient passes the same performance screen in another setting.

Agreement between these measurements strengthens the case for a head's role. To identify a circuit, we would also need evidence about the paths between components or how they work together. When measurements disagree, the functional label alone cannot resolve that disagreement.

B Summary of findings

C Reproducibility Details

C.1 Evidence checks: what a result can establish

For each result, we record what was measured, what changed, and how the result was analyzed:

$$\mathcal{A} = (c, \mathcal{D}, M, \mathcal{S}, \mathcal{I}, \mathcal{B}, Y; U, \Pi). \tag{7}$$

Here c is the detector and its input realization; $\mathcal{D}$ is the selection/evaluation data and split; M is the resolved model checkpoint; $\mathcal{S}$ is the head set; $\mathcal{I}$ is the intervention; $\mathcal{B}$ is the comparison condition; and Y is the outcome. Together these fields define the quantity being estimated. U is the independent unit, and Π is the analysis rule, including corrections for multiple tests and decision thresholds. Changing one of the first seven fields calls for a new test. Changing U or Π changes what the evidence can support.

We write reuse from a source setting to a target setting as $\mathcal{A}_s \to \mathcal{A}_t$. We check it in three steps:

1. **Is the comparison identified?** Check that the model and data records, intervention, and controls establish which contrast was tested.
2. **What is its scope?** Check the independent unit U and whether the analysis rule Π was fixed before target outcomes were available.
3. **Does it pass the stated rule?** Only after the first two checks do we apply Π. A test of the value of source information must include a comparison rule that never uses the source.

The first failed check determines the conclusion. Missing information about the contrast makes it *non-identifying*. An analysis chosen after outcomes were available, or a claim beyond the sampled

Table 1: **The four questions and their main answers.** Each finding is limited to the models, data, and interventions tested for that question.

Question	Main finding	What remains uncertain
RQ1: Do layer locations and relative order repeat?	Normalized layer locations vary across models. Duplicate-token heads precede induction-prefix heads in every cell and previous-token heads do so at $K = 30$; retrieval and truthfulness heads sit at mid-depth rather than deep. No complete early–late ordering is shared.	The previous-token order is required by the induction composition. Retrieval and truthfulness positions come from a single-sample local supplement. Five models do not establish a layer-organization law.
RQ2: Can an ablation effect be attributed to a label?	Induction-prefix and retrieval detectors share 19–22 of their Top-30 heads on five checkpoints. Patching the frozen retrieval heads recovers $.77/.83$ of the retrieval loss on Qwen/Llama. Removing the shared heads is the only intervention that costs needle recall; the exclusive heads cost nothing alone and at most a few points conditionally.	The clean shared-head effect is on Llama-3; Mistral's intervals touch zero and Qwen is insensitive on this needle. Exclusivity is defined on one detection sample; at $K = 20$ and 40 the exclusive heads show conditional effects of up to 9 points. Historical intervention masks are incompletely recorded.
RQ3: Do heads form in the order they depend on each other?	In Pythia-1.4B, previous-token strength crosses half-final before induction-prefix and duplicate-token after it. In both Pythia trajectories, strength saturates thousands of steps before rankings and Top-K membership stabilize. Fine-tuning can preserve layer distributions while changing membership.	The observed order is also the order of pattern simplicity, and the second run shares the training recipe, so dependence is not isolated. Sampled checkpoints cannot pinpoint abrupt versus gradual formation. Supplied tuned checkpoints do not isolate SFT.
RQ4: Does amplifying heads help?	Doubling whole-head outputs gives one local gain on two models. Directional truthfulness ITI helps Llama but not Qwen at the tested setting. The labels rank heads usefully for KV-cache allocation.	One coefficient, two models, and a retrieval task at ceiling; gains elsewhere are not ruled out. Ablation effects and passing a screen do not show that amplification helps. The held-out statistical rules are retrospective.

units, is *descriptive*. Only an identified test with appropriate scope and a rule fixed before target outcomes can be called *supported* or *not supported* under Π. "Not supported" is not a refutation; that would require a suitably powered equivalence test. Testing every member of a finite set supports a conclusion about that set, not a broader population. These are checks on reusing empirical claims, not a formula for causal transport. Table 1 summarizes the four research questions; Table 4 gives the setup and scope of the main comparisons.

C.2 MODELS AND FROZEN ENDPOINTS

Table 2: Evaluation scope. Conclusions are restricted to the models listed for each research question.

Scope	Models and purpose
RQ1	Qwen2.5-7B (Qwen Team, 2024), Llama-2-7B (Touvron et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.1 (Jiang et al., 2023), and Pythia-6.9B (Biderman et al., 2023); five base models spanning four broad lineages.
RQ2	Qwen2.5-7B-Instruct and Llama-2-7B-Chat for detector overlap, archived induction interventions, and retrieval patching; Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.2, and Qwen2.5-7B-Instruct for the three-way ablation.
RQ3	Pythia-1.4B (14 early checkpoints), Pythia-2.8B (6 checkpoints), Pythia-6.9B PT/SFT trajectories, plus Qwen/Llama base-to-tuned endpoints.
RQ4	Qwen2.5/Llama-2 base and tuned endpoints for static scaling; selection-frozen intervention contrasts on the two tuned endpoints only; Llama-3-8B-Instruct for the HeadKV allocation control.

Table 3: Fixed model checkpoints used in the main analyses. The supplement records full 40-character Hub revisions and all 14 Pythia-1.4B commits. The local Pythia-6.9B archive retains its checkpoint label.

Endpoint	Model	Revision (prefix)	Geometry
Qwen base	Qwen2.5-7B	`d1497293`	28×28 query, 4 KV
Qwen tuned	Qwen2.5-7B-Instruct	`a09a3545`	28×28 query, 4 KV
Llama base	Llama-2-7B	`01c7f73d`	32×32
Llama tuned	Llama-2-7B-Chat	`f5db02db`	32×32
Llama-3.1	Meta-Llama-3.1-8B	`48d6d0fc`	32×32 query, 8 KV
Mistral	Mistral-7B-v0.1	`27d67f1b`	32×32 query, 8 KV
Pythia	Pythia-6.9B-deduped	`step143000`	32×32

The Pythia-1.4B study follows 14 checkpoints from step 0 to 143,000, with more closely spaced early checkpoints at 128, 512, 1,000, 2,000, 3,000, 4,000, 5,000, and 8,000. Each checkpoint was resolved to an immutable 40-character commit before the run. All 14 have the same model-configuration hash. The RQ1 archive contains 63,440 per-head scores from five base models and 15 complete model–input records.

Table 4: **What the main comparisons measure.** Each row covers a family of comparisons across models, tasks, or K values, rather than every individual $\mathcal{A}_s \to \mathcal{A}_t$ record. The second column names the main field that changes. The third lists $(c, \mathcal{D}, M, \mathcal{S}, \mathcal{I}, \mathcal{B}, Y)$; the last gives U and Π, as defined in Section C.1. Descriptive results do not establish population-level significance.

Result	Main change	Comparison setup	Analysis and scope
RQ1 repeatability	c realization	Each stochastic detector; fixed prompts; each base checkpoint; all heads or Top-K; no intervention; paired detector seed; Spearman or Jaccard.	U: no population unit; three overlapping seed pairs. Π: mean over pairs, descriptive within a checkpoint.
RQ1 organization	M	Fixed-population attention detector; fixed prompts; five base checkpoints; Top-K; no intervention; paired endpoint; relative-depth centroid/order or eight-band cosine.	U: no model-population unit; ten overlapping pairs from five checkpoints. Π: descriptive mean over all pairs.
RQ2 structural	$c, \mathcal{S}, \mathcal{B}$	Induction-prefix and retrieval detectors; fixed Qwen/Llama inputs; tuned checkpoints; Top-30 sets; no intervention; layer-conditioned permutation null; overlap or score correlation.	U: head assignment within layer. Π: fixed permutation tests reported separately from behavioral evidence.
RQ2 one-way behavior	$\mathcal{D}, Y$	Induction-prefix detector; archived static-grid inputs; tuned endpoints; Top-30; pre-W_O zero scaling; unchanged model and exact-layer low-important controls; retrieval recall.	U: no population unit; five fixed controls are reweighted. Π: 2,000 draws; descriptive, excludes head-selection uncertainty, and cannot separate shared from category-specific heads.
Selection-frozen sensitivity	$c, \mathcal{S}, \mathcal{B}, Y$	Retrieval, induction-attention, or truthfulness detector; disjoint selection and evaluation splits; tuned endpoints; Top-30; pre-W_O zero scaling; unchanged, exact-layer, or pre-W_O RMS reference; matched task metric.	U: needle, induction example, or question. Π: reconstructed 10,000-draw bootstrap/sign procedures and one 18-test Holm family; descriptive because final U, Π were specified after outcomes.
RQ3 timing	M	Each detector; fixed checkpoint prompts; one Pythia-1.4B trajectory; all heads or Top-30; no intervention; final checkpoint; half-final signal and sustained rank/set agreement.	U: checkpoint trajectory. Π: fixed crossing/agreement thresholds; descriptive case study.
RQ4 guardrail-screen reuse	$M, \mathcal{S}, \mathcal{I}$	Each selector; archived static-grid inputs; tuned endpoints; endpoint-specific Top-K sets; source-selected or fixed scalar on a pre-W_O scaling grid; unchanged model comparator; oriented target/null/guardrail metrics.	U: a source–target cell is a reporting unit; cells are correlated. Π: source-only selection, fixed baselines, and a target screen for each metric scale; no population inference.

C.3 Fixed data and statistical controls

Previous-token, duplicate-token, and induction-attention checkpoint evaluations each use 203 prompts, with a separate fixed hash for each set. Induction-prefix uses five independently sampled repeated-token sequences. Retrieval uses the official English needle-in-a-haystack data; needles used in control experiments are held out from head selection. Truthfulness uses the first 790 of 817 TruthfulQA rows because the original manifest set this cap. The other 27 questions were never scored; their exclusion did not depend on outcomes. Splitting by question keeps all true/false answers together: 316 questions are used for training, 79 for validation, and 395 for testing. The selection-frozen evaluation fixes 128 questions from the test split before intervention. The static-control grid uses $K \in \{5, 10, 20, 30, 50\}$ and $\alpha \in \{0, 0.5, 0.75, 1, 1.25, 1.5, 2\}$. Each functional intervention in the original grid has five controls with the same number of heads in each layer. These controls minimize overlap with the union of all selectors' Top-50 heads, subject to that layer constraint. Before producing summary tables, validators check model identity, geometry, input hashes, complete grids, valid head indices, control overlap, and a shared baseline hash.

C.4 Retrieval score implementation

At generation step t, let $j_t^* = \arg\max_j A_{\ell,h}[t, j]$ be the position receiving the most attention. For a successful retrieval cell, N_q is the needle span and L_q its length; y_t and x_j are the generated and context tokens. The recorded implementation scores each head as

$$r_q(\ell, h) = \frac{1}{L_q} \sum_t \mathbb{1}[j_t^* \in N_q \ \wedge \ y_t = x_{j_t^*}]. \tag{8}$$

The final score averages r_q over cells with answer ROUGE-1 recall above 50%. A needle position can be copied at several generation steps, so r_q can exceed one.

In the selection-frozen evaluation, a registry fixes the Top-30 heads using only selection data (internal semantic SHA-256 prefix `eb215e49`; archived file SHA-256 prefix `ba176c23`). Validation and test rankings cannot change these sets. Evaluation uses 12 held-out retrieval needles at three depths (36 cells), 64 induction cases, and 128 fixed TruthfulQA questions. The retrieval needles are held out from head selection, but come from the same fixed source as the RQ4 static grid; they are not a second independent task corpus. Each of the six runs tests five functional strengths, five exact-layer ablations, and five ablations matched by exact layer and approximate pre-W_O activation RMS. All 90 conditions and 990 metric rows pass checks of source records, completeness, and control construction.

C.5 Controls matched by layer and activation size

Let $z_{\ell hptk}$ be coordinate k of head h's output before W_O, at token t of profiling prompt p. Let N_ℓ count all unpadded token positions at layer ℓ. We measure the output's root mean square (RMS):

$$R_{\ell h} = \sqrt{\frac{1}{N_\ell d_h} \sum_{p,t,k} z^2_{\ell hptk}}, \tag{9}$$

We use eight selection-split TruthfulQA questions and eight independently seeded synthetic induction prompts, including all unpadded positions up to 512 tokens. This measures activation size before projection, not the norm of $z_h W_O^h$ written to the residual stream. Control heads exclude the selected functional set and match its head count in every layer. Within each layer, linear-sum assignment minimizes $|\log R_{\ell h} - \log R_{\ell h'}|$ plus seeded jitter bounded by 0.03 times the median base cost. Other known Top-50 functional heads receive a soft penalty equal to 2.0 times that median base cost. No head is repeated within a control set, but sets from different seeds may share heads. Each exact-layer control family has five distinct sets. The RMS families have 5/1/3 distinct sets for Qwen retrieval/induction-attention/truthfulness and 5/5/5 for Llama. The reported control mean includes all five seeded assignments. They are not five independent statistical units: bootstrap and sign-flip tests resample needles, induction examples, or questions while keeping this control mean fixed. Matching has no hard tolerance and can be imperfect. The two largest run-level maximum log distances, $.927$ and 1.140, correspond to RMS ratios of roughly $2.5\times$ and $3.1\times$. In Qwen truthfulness, 5 of 150 matched slots are drawn from the "other functional" pool.

Table 5: Quality of pre-W_O activation-RMS matching across five seeded assignments per run. Mean is the average absolute log-RMS distance across sets; Max is the largest matched-pair distance. "Other functional" counts selections from the soft-penalty pool across all assignments. "Unique" counts distinct head sets.

Model	Selector	Mean	Max	Other functional	Unique
Qwen2.5	Retrieval	.121	.927	0	5
	Ind.-attn.	.173	1.140	0	1
	Truth.	.113	.504	5	3
Llama-2	Retrieval	.084	.479	0	5
	Ind.-attn.	.089	.437	0	5
	Truth.	.032	.151	0	5

Giving equal weight to the three distinct Qwen truthfulness RMS sets changes the contrast from .0504 to .0466 and its 18-family Holm value from .245 to .356. It remains non-significant. All Qwen induction-attention RMS assignments are identical, so removing duplicates does not change their control mean.

D ADDITIONAL RQ1 RESULTS

Table 6: Average relative layer position of Top-30 heads (lower is earlier), averaged over three detector-input samples. Duplicate-token heads precede induction-prefix in all five models; other layer relationships vary by model.

Base model	Prev.	Dup.	Ind.-P	Ind.-A
Qwen2.5-7B	.353	.375	.702	.312
Llama-2-7B	.434	**.229**	.472	.650
Llama-3.1-8B	.461	**.416**	.539	.380
Mistral-7B	.526	**.463**	.606	.253
Pythia-6.9B	.288	**.212**	.313	.609

The frozen RQ1 archive scores the four attention detectors and the truthfulness probe on two models. Tables 7 and 8 add a single-sample local supplement that also places retrieval and truthfulness heads; its attention-detector centroids agree with the frozen table to within .02 where the models coincide.

Table 7: **Local supplement: Top-30 relative-depth centroid for all five signals** (front/middle/back share of the Top-30 in parentheses). Single detector seed, local model set; `analysis/rq1_local_centroids.csv`. Llama-2 and Llama-3 centroids for the three attention signals agree with the frozen table above to within .02.

Model	Prev	Dup	Ind-P	Retr	Truth
Llama-2-7B	.439 (40/40/20)	.216 (80/10/10)	.465 (30/53/17)	.481 (27/50/23)	—
Llama-3-8B	.452 (30/50/20)	.426 (53/17/30)	.543 (27/43/30)	.505 (27/53/20)	—
Llama-3-8B-Inst	.451 (30/47/23)	.397 (53/23/23)	.535 (30/37/33)	.552 (20/53/27)	.418 (10/90/0)
Mistral-7B-Inst	.514 (27/37/37)	.468 (53/20/27)	—	—	—
Pythia-6.9B	.350 (63/13/23)	.193 (90/10/0)	.297 (77/17/7)	.351 (60/33/7)	.501 (30/47/23)
Pythia-1.4B	.319 (67/13/20)	.338 (47/47/7)	.456 (27/70/3)	.494 (23/63/13)	.482 (27/57/17)

Table 8: **Pairwise depth order on the local supplement**, counted over $K \in \{10, 20, 30\}$ (cells where centroid$(A) <$ centroid(B)). Only the two orders with a composition rationale hold in every model at $K \geq 20$.

Order	Llama-2	Llama-3	Llama-3-Inst	Mistral-Inst	Pythia-6.9B	Pythia-1.4B
Dup < Ind-P	3/3	2/3	2/3	—	3/3	3/3
Dup < Retr	3/3	3/3	3/3	—	3/3	3/3
Prev < Ind-P	3/3	2/3	2/3	—	0/3	2/3
Dup < Prev	3/3	2/3	3/3	2/3	3/3	2/3
Ind-P < Retr	2/3	1/3	3/3	—	1/3	3/3
Prev < Retr	3/3	3/3	3/3	—	1/3	3/3
Truth < Retr	—	—	3/3	—	0/3	1/3
Ind-P < Truth	—	—	0/3	—	3/3	3/3

Table 9: Mean agreement between detector seeds, averaged over the four stochastic attention detectors. Retrieval is deterministic and is excluded.

Base model	All-head Spearman	Top-30 Jaccard
Qwen2.5-7B	.998	.937
Llama-2-7B	.998	.936
Llama-3.1-8B	.999	.917
Mistral-7B	.999	.909
Pythia-6.9B	.994	.942

Table 10: Cross-model similarity of Top-30 depth distributions for the fixed-population attention detectors. Cosine is computed over eight depth bands for all ten model pairs.

Head type	Mean	Min	Max
Previous-token	.789	.537	.936
Duplicate-token	.749	.587	.909
Induction-prefix	.587	.124	.839
Induction-attention	.386	.074	.858

The same broad pattern appears beyond Top-30. At $K = 10/20/30$, mean cross-model eight-band cosine is $.724/.783/.789$ for previous-token, $.737/.750/.749$ for duplicate-token, $.484/.602/.587$ for induction-prefix, and $.350/.346/.386$ for induction-attention. Across the same grid, the duplicate-token mean depth is earlier than induction-prefix in every model, K, and detector-input setting.

D.1 Exploratory checks against random depth patterns

Table 11 gives two checks designed after inspecting the results. The geometry null randomly permutes head coordinates within each model 100,000 times. Each permutation is shared across detectors, inputs, and K. This preserves model geometry, nested Top-K sets, overlap, and seed dependence, but breaks the link between score and depth. Its plus-one upper-tail p-values are exploratory diagnostics, not a confirmatory test family. The label null shuffles the four detector labels independently within each model, using the same shuffle for all inputs and K. Fixing one common relabeling leaves $24^4 = 331{,}776$ distinct cases, all of which are enumerated. This asks whether matching detector names gives more similar distributions than arbitrary cross-detector matches.

Table 11: **Exploratory RQ1 comparisons with random patterns.** "Geom." randomizes Top-K locations while preserving model geometry. "Label" averages the four detectors under exact label permutations. Intervals are the null 2.5–97.5% quantiles. Upper-tail p-values are not adjusted for multiple tests.

Null	Scope	K	Observed	Null mean	Null 95%	p
Geom.	Previous-token	30	.789	.818	[.735,.891]	.775
Geom.	Duplicate-token	30	.749	.818	[.736,.889]	.953
Geom.	Induction-prefix	30	.587	.818	[.736,.889]	1.000
Geom.	Induction-attention	30	.386	.818	[.737,.888]	1.000
Label	Four-detector mean	10	.574	.508	[.460,.574]	.0247
Label	Four-detector mean	20	.620	.543	[.493,.611]	.0150
Label	Four-detector mean	30	.628	.567	[.517,.636]	.0378
Label	All K	10/20/30	.607	.539	[.493,.603]	.0198

E ADDITIONAL RQ2 RESULTS

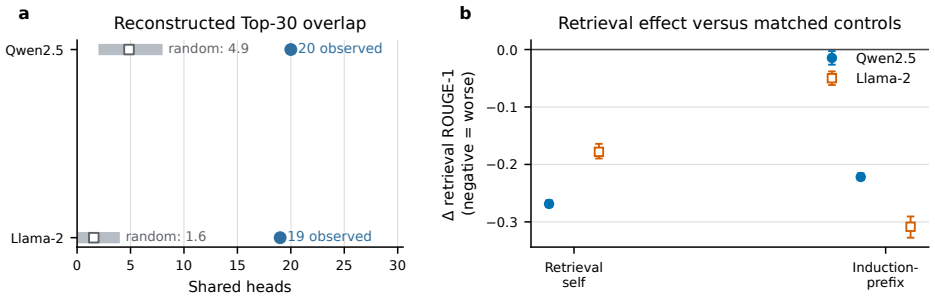

Figure 5: **Induction-prefix and retrieval detectors select many of the same heads.** (a) Reconstructed Top-30 overlap versus a random comparison that preserves layer counts. (b) Archived retrieval effects after ablation, relative to matched controls; negative means a larger performance loss. Bars show quantiles from 2,000 reweightings of five fixed controls, not confidence intervals. Exact historical executed masks cannot be verified, and the comparison does not separate shared heads from category-specific heads.

Table 12: Retrieval ROUGE-1 change in archived $\alpha = 0$ conditions labelled by a Top-30 functional selector, relative to five exact-layer controls. The registry reconstructs expected masks but does not verify historical executed masks. Controls minimize feasible overlap with the union of selectors' Top-50 heads; brackets are empirical 2.5–97.5% quantiles from reweighting only these five fixed controls while holding the functional estimate fixed; they are not confidence intervals, exclude example and head-selection uncertainty, and are not multiplicity corrected. Base rows are descriptive same-sign endpoint contrasts.

Endpoint	Retrieval self	Induction-prefix	Induction-attn
Qwen2.5-Instruct	−.269 [−.274, −.263]	−.222 [−.227, −.215]	−.017 [−.022, −.009]
Llama-2-Chat	−.178 [−.190, −.164]	−.308 [−.327, −.291]	.011 [−.006, .028]
Qwen2.5-Base	−.394 [−.406, −.385]	−.443 [−.457, −.426]	−.136 [−.151, −.124]
Llama-2-Base	−.171 [−.220, −.119]	−.172 [−.226, −.107]	.168 [.114, .230]

Table 13: Induction-prefix–retrieval overlap, examined after choosing this detector pair from the results. The null keeps both Top-30 layer histograms fixed. Intervals are quantiles from one million random draws; p is the exact upper tail of the hypergeometric convolution. ρ_{within} is the pooled within-layer rank correlation. Tests are exploratory and unadjusted, and do not control for retrieval scoring only on successful cells.

Model	Observed J	Null mean	95% null	Exact p	ρ_{within}
Qwen2.5	.500	.089	[.034,.154]	9.6×10^{-15}	.579
Llama-2	.463	.027	[.000,.071]	3.1×10^{-21}	.517

E.1 MATCHED INDUCTION INTERVENTIONS

Table 14: Raw induction-task effects and effects relative to controls. Controls match layer counts and minimize feasible overlap with all selectors' Top-50 heads. Functional intervals are archived example-bootstrap summaries; the paired example vectors needed to reproduce them are unavailable. Fixed-control brackets are 2.5–97.5% quantiles from reweighting five controls while keeping the functional estimate fixed, not confidence intervals. Neither interval includes head-selection uncertainty or jointly models example and control uncertainty. Patching measures a different contrast from ablation.

Model	Modality	K	Raw functional [archived example CI]	Functional−low-important [fixed-control q.]
Qwen2.5	zero	10	1.405 [1.061, 1.776]	1.405 [1.385, 1.425]
	mean	10	.697 [.485, .925]	.709 [.690, .729]
	patch	10	11.692 [10.765, 12.587]	11.367 [11.124, 11.610]
	zero	30	2.894 [2.392, 3.447]	2.858 [2.823, 2.894]
	mean	30	1.592 [1.218, 1.991]	1.578 [1.557, 1.599]
	patch	30	11.974 [11.066, 12.856]	11.295 [10.949, 11.641]
Llama-2	zero	10	.892 [.683, 1.122]	.874 [.820, .932]
	mean	10	.706 [.515, .916]	.719 [.682, .757]
	patch	10	6.106 [5.454, 6.767]	6.045 [5.964, 6.123]
	zero	30	1.769 [1.425, 2.146]	1.716 [1.599, 1.813]
	mean	30	1.413 [1.093, 1.764]	1.404 [1.336, 1.477]
	patch	30	7.681 [6.909, 8.458]	7.491 [7.343, 7.621]

F ADDITIONAL RQ3 RESULTS

F.1 BASE-TO-TUNED ENDPOINT STABILITY

Table 15: Depth distributions change less across tuning than exact Top-30 head membership. These checkpoint comparisons contain detector scores, not changes in task performance.

Pair	Signal	Spearman	J_{30}	Layer cosine
Qwen	Previous	.998	.935	1.000
base→tuned	Duplicate	.991	.978	1.000
	Induction-prefix	.858	.785	.996
	Induction-attn	.993	.819	1.000
Llama	Previous	.978	.748	.999
base→chat	Duplicate	.966	.915	.997
	Induction-prefix	.952	.876	.996
	Induction-attn	.977	.514	.999

Across 13 Pythia-6.9B pretraining/SFT checkpoints and three truthfulness split seeds, mean all-head Spearman is .952 (range .923–.966). Mean Top-10 and Top-30 Jaccard are .460 and .526, while mean layer-distribution cosine is .99994. Thus, overall depth patterns are stable even when the small set of top heads depends on the data split.

F.2 THRESHOLD SENSITIVITY FOR THE PYTHIA-1.4B CASE STUDY

Table 16: When detector strength and head membership meet their thresholds on 14 Pythia-1.4B checkpoints. Strength reaches half its final maximum before both ranking and Top-30 agreement remain above the stability thresholds. Retrieval is excluded because its successful examples change across checkpoints.

Signal	Strength threshold	Stable ranking/set
Previous-token	512	5,000
Induction-prefix	1,000	8,000
Duplicate-token	2,000	71,000
Induction-attention	2,000	107,000

Table 17: First checkpoint at which each detector reaches 25%, 50%, or 75% of its final maximum score. The early and late groups remain separate, but ties and ordering within each group depend on the threshold.

Signal	**25%**	**50%**	**75%**
Previous-token	512	512	1,000
Induction-prefix	1,000	1,000	1,000
Duplicate-token	2,000	2,000	2,000
Induction-attention	2,000	2,000	3,000

Table 18: First checkpoint after which both ranking and Top-30 agreement with the final checkpoint stay above the stated Spearman/Jaccard thresholds. The main setting is $(.8, .5)$. Other thresholds change the exact timing, but stable head identity still follows the rise in signal strength.

Signal	$(.7, .4)$	$(.8, .5)$	$(.9, .6)$
Previous-token	3,000	5,000	33,000
Induction-prefix	3,000	8,000	71,000
Duplicate-token	16,000	71,000	107,000
Induction-attention	71,000	107,000	107,000

We also compare four supplied Pythia parent checkpoints with their Dolly-derived checkpoints, each labeled SFT step 236. The archive does not identify the Dolly release, optimizer settings, training seed, or whether `236` has the same meaning across training systems. These comparisons describe the supplied pairs. They do not estimate variation across SFT runs or isolate the effect of instruction tuning.

Table 19: Change in the conditional induction functional-minus-random effect at $K = 30$, from each supplied parent checkpoint to its Dolly-labeled SFT236 checkpoint (SFT minus parent). Each row uses five paired random-selector seeds. Zero ablation, mean ablation, and clean patching measure different contrasts; their magnitudes should not be compared directly.

Parent PT step	**Zero**	**Mean**	**Clean patch**
36,000	$-.529$	$-.316$	$+2.309$
71,000	$-.812$	$-.588$	$+2.299$
107,000	$-.635$	$-.856$	$+2.640$
143,000	$-.905$	$-.911$	$+1.692$

G ADDITIONAL RQ4 RESULTS

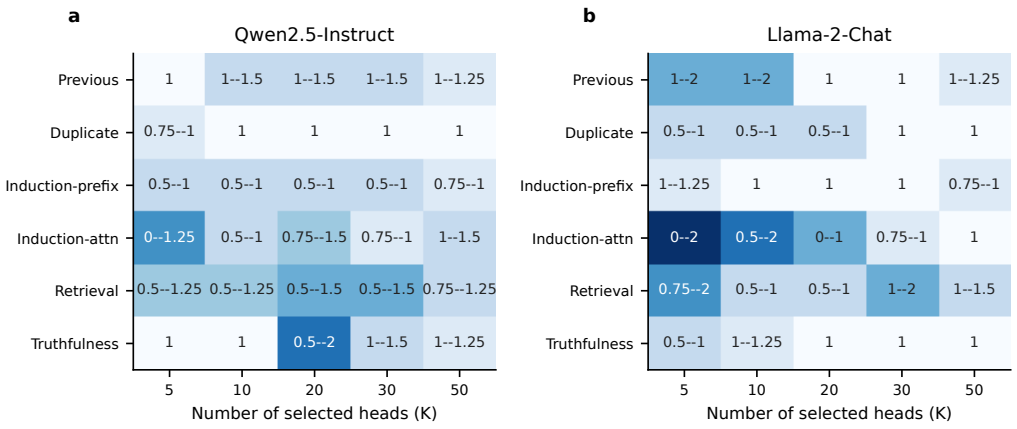

Figure 6: **The range of acceptable scaling coefficients differs between the two tested models.** Each cell shows the consecutive tested α values around $\alpha = 1$ that pass the target and guardrail screen, for one selector and Top-K set. Results use the archived normalized-log scale. Text gives the sampled span; darker cells show wider spans. These are descriptive comparisons.

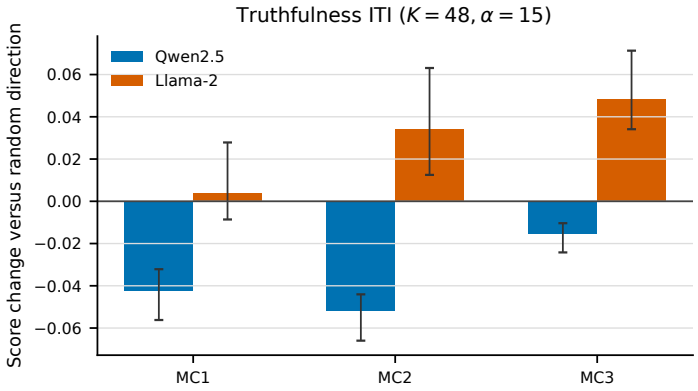

Figure 7: **Truthfulness ITI compared with layer-matched random directions.** Point estimates are positive on Llama-2 MC2/MC3 and negative on all Qwen2.5 metrics. The coefficient α is not normalized by direction norm. Error bars resample three detector splits and are not corrected for multiple tests.

G.1 RANKING AND INTERVENTION RESULTS WITH FIXED HEAD SELECTION

Table 20: Mean ranking agreement across the two models and six signals. The averages hide differences between detectors. One Qwen retrieval test needle gives an all-zero score matrix, so agreement cannot be estimated; it is not zero agreement. Llama retrieval selection-to-test agreement is $\rho = .502$ and $J_{30} = .622$.

Frozen split vs.	**Spearman**	J_{30}	**Min.** J_{30}	**Non-est.**
Validation	.946	.782	.429	0
Test	.938	.818	.395	1
Previous consensus	.959	.876	.539	0

Table 27 reports target effects and Holm-adjusted tests from the reconstructed 18-test family. The original summary treated all 36 retrieval depth cells as independent. The current analysis instead groups the three depths of each needle, giving 12 independent units. The pre-run manifest did not fix this unit, the final test family, or the rule requiring all three comparisons to pass before target outcomes were available. We therefore report these contrasts as descriptive, even though head selection and evaluation inputs were fixed.

The condition records allow the statistical analysis to be reconstructed, but do not retain target-token IDs. A separate tokenizer-only check reconstructs both 64-example seeds. In every example, the suffix found by the legacy longest-common-prefix (LCP) procedure equals the exact target suffix for the pinned Qwen tokenizer and a pinned public Llama-2 mirror. The mirror check is a diagnostic; it does not cryptographically verify the gated endpoint's tokenizer. At $\alpha = 2$, the changes from baseline are -0.003 (Qwen retrieval), 0.014 (Qwen induction-attention), -0.016 (Qwen truthfulness), -0.003 (Llama retrieval), -0.005 (Llama induction-attention), and -0.003 (Llama truthfulness). Only Qwen induction has a confidence interval excluding zero. The six response curves do not establish a common rule that amplification improves performance.

Table 21: The six exact-layer-random comparisons not shown in Table 27. Together with that table's 12 cells, they form the reported 18-comparison Holm family. Each cell gives the oriented effect, pointwise percentile-bootstrap 95% CI, and sign-based Holm-adjusted p-value. A CI excluding zero does not imply Holm-adjusted significance.

Model	Selector	vs. exact-layer random
Qwen2.5	Retr. (ROUGE-1)	$-.362$ [$-.479$, $-.235$], $p_H = .031$
	Ind.-attn. (mean logp)	.041 [.016, .064], $p_H = .010$
	Truth. (MC2)	.029 [$-.007$, .068], $p_H = 1.000$
Llama-2	Retr. (ROUGE-1)	$-.134$ [$-.252$, $-.014$], $p_H = 1.000$
	Ind.-attn. (mean logp)	.052 [.037, .068], $p_H = .0018$
	Truth. (MC2)	.009 [$-.022$, .040], $p_H = 1.000$

Full ablation worsens the mean-NLL guardrail in all six runs: oriented effects range from -0.012 to -0.332. The selection-frozen benign-QA evaluation has only 16 examples, so its changes are diagnostic, not confirmatory. Exact tokenwise $\mathrm{KL}(p_{\text{base}} \| p_{\text{intervened}})$ is zero at $\alpha = 1$ by construction and rises most sharply under full ablation.

G.2 Truthfulness ITI

We also test directional inference-time intervention (ITI) on the Top-48 truthfulness heads at $\alpha = 15$. Each learned probe direction is compared with five layer-matched random directions. ITI changes activity along a direction within each head; static pre-W_O scaling multiplies the head's whole output. The ITI coefficient α is not normalized by direction norm.

Table 22: Directional ITI with $K = 48, \alpha = 15$, averaged over three detector splits and compared with five layer-matched random directions.

Model	Metric	Base	ITI	ITI−random	95% CI
Llama-2	MC1	.322	.321	.004	[$-.009$, .028]
	MC2	.496	.527	.034	[.012, .063]
	MC3	.232	.277	.048	[.034, .071]
Qwen2.5	MC1	.418	.368	$-.042$	[$-.056$, $-.032$]
	MC2	.622	.555	$-.052$	[$-.066$, $-.044$]
	MC3	.304	.284	$-.016$	[$-.024$, $-.010$]

G.3 Static-control coverage and the performance screen

The tuned-model grid contains 2,520 conditions and 14,280 metric rows. The base-model grid contains 2,100 conditions and 10,080 metric rows. For metrics valid in both models, excluding

unchanged $\alpha = 1$ conditions, effect signs agree in 456/870 tuned-model comparisons and 360/600 base-model comparisons. The benign-QA guardrail adds 216 conditions and 144 dose-response rows. Its refusal-rate results describe only 32 benign questions; they do not test safety on harmful requests.

For static-grid effects, we orient each metric so that higher is better. At each α, we subtract the mean change in exact-layer low-important controls from the change in the selected functional heads. The 95% intervals resample only the five controls, using 2,000 draws and keeping the functional estimate fixed. They are pointwise intervals conditional on these controls, without correction for multiple tests.

Table 23: How performance-screen counts change with the thresholds, using the archived normalized-log scale. Non-unit conditions have $\alpha \neq 1$; there are $360 = 2$ models $\times 6$ selectors $\times 5$ values of K $\times 6$ non-unit α values. A cell is counted when the consecutive qualifying values around $\alpha = 1$ include at least one tested non-unit value.

Rule	τ_T	τ_G	Qualifying conditions	Qwen/Llama cells
Strict	1%	2.5%	80/360	21/30 / 15/30
Nominal	2%	5%	105/360	23/30 / 19/30
Lenient	5%	10%	128/360	25/30 / 21/30

The old induction next-token accuracy used the first token of the continuation tokenized in isolation. This can differ from its first token when prompt and continuation are tokenized together. We therefore exclude this metric from the main screen for both models and retain mean target log-probability. This post-hoc correction addresses the token definition, not the direction of the effect. Retaining the old Qwen accuracy as a diagnostic gives 82/108/134 qualifying conditions and 21/24/26 versus 15/19/21 Qwen/Llama cells under the three rules. This diagnostic also finds no non-unit coefficient that passes for every cell.

The archived normalization divides $\Delta \log p$ by the absolute baseline mean log probability. This is a normalized log change, not a probability change. Replacing only this scale with $\exp(\Delta \log p) - 1$ gives 125/153/176 qualifying non-unit conditions; 35/42/49 of these have no positive target estimate. The Qwen/Llama cell counts become 25/24, 25/26, and 25/26. The number that pass therefore depends on the metric scale. Neither scale gives a tested non-unit coefficient that passes every cell.

The performance screen (called "guardrail-qualified" in the archive) uses matched and transfer metrics as targets. A condition passes if:

- every target's oriented relative change is at least $-\tau_T$, and at least one is non-negative;
- every NLL/perplexity guardrail's oriented relative change is at least $-\tau_G$; and
- every null task's absolute relative change is at most τ_G.

The archived screen admits 105/360 non-unit conditions, or 153/360 under the probability-ratio scale. Counts differ across the three threshold rules, but no tested non-unit α passes all 60 model–selector–K cells. The screen uses point estimates only at $\alpha \in \{0, .5, .75, 1, 1.25, 1.5, 2\}$. It does not interpolate between values, account for example or head-selection uncertainty, or adjust for multiple tests.

G.4 REUSING A COEFFICIENT ACROSS MODELS

Table 24: **Choosing a coefficient from another model versus fixing it at** .75**.** Each scale contains 60 directed model–detector–K cases. Neither source rule passes more cases overall than fixed .75. W/L counts cases where only the source rule/only the fixed value passes, among cases where the source rule makes a choice. The comparison is retrospective; the appendix gives the full fixed-value sweep and subset counts.

Metric scale	Source rule	Can choose	Pass: source / fixed .75	Paired W/L
Normalized log	Nearest to 1	42/60	23/60 / **28/60** (-5)	4/2
	Maximin	42/60	19/60 / **28/60** (-9)	3/5
Probability ratio	Nearest to 1	51/60	38/60 / **41/60** (-3)	5/6
	Maximin	51/60	32/60 / **41/60** (-9)	3/10

We reuse only the coefficient α. Each target model uses its own Top-K heads for the corresponding detector. For all 30 selector–K settings, we test both Qwen→Llama and Llama→Qwen. The nearest-to-baseline rule chooses the qualifying coefficient closest to one. Ties favor the larger minimum target change on the source, then the lower coefficient. The source-maximin rule first maximizes that minimum source target change, then prefers the coefficient closest to one. Both rules abstain when the source's consecutive qualifying region has no non-unit coefficient. Neither uses target-model results to select the coefficient.

Under the archived normalized-log scale, 42/60 cells allow a source choice and 18 require abstention. In 28 of those 42, at least one non-unit coefficient passes on the target (the target-only oracle). The nearest-to-baseline rule passes the target screen in 23/60 cells (23/42 eligible; 10/23 Qwen→Llama and 13/19 Llama→Qwen), covering 23/28 oracle-eligible cases. Source-maximin passes in 19/60 (19/42; 10/23 and 9/19), or 19/28 oracle-eligible cases. Under the probability-ratio sensitivity, 51/60 cells are eligible and nine require abstention; 46 have a target oracle. The two rules pass in 38/60 (38/51; 17/25 and 21/26; 38/46 oracle-eligible) and 32/60 (32/51; 15/25 and 17/26; 32/46 oracle-eligible).

For a retrospective baseline on the same fixed grid, we also test coefficients chosen without source information. The fixed-coefficient sweep at $\alpha \in \{0, .5, .75, 1.25, 1.5, 2\}$ qualifies $3/20/28/25/16/7$ of 60 cells under the normalized-log scale and $6/30/41/34/23/11$ under the probability-ratio scale. Fixed .75 is therefore the best adjacent setting in this archive, not a preregistered policy. It passes more of the 60 cells than either source rule. On the 42 archived-scale eligible cells, nearest/source-maximin qualify 23/19, compared with 21 for fixed .75; their paired source-only/fixed-only counts are 4/2 and 3/5. On the 51 probability-ratio source-eligible cells, the corresponding counts are 38/32 versus 39, with paired counts 5/6 and 3/10. The nearest rule's two-cell advantage on one scale reverses on the other, while maximin is lower on both. Fixed 1.25 qualifies 17/42 and 31/51 on the same eligible subsets; complete paired contingency tables are supplied as frozen CSVs.

Among source choices that pass on the target, we compare the minimum target change with the best target-only choice. The mean shortfall is .0125/.0105 on the normalized-log scale and .0026/.0028 on the probability-ratio scale for nearest/source-maximin; all four medians are zero. Failed choices receive no shortfall value, and the values are not compared across scales. The target-only choice is a hindsight diagnostic and never enters source selection. The screen compares with the unchanged model. Subtracting exact-layer controls is part of the separate specificity analysis above. The 60 cells share models, selectors, and examples, so they are correlated. We therefore report no binomial interval or population-level efficacy claim in Table 24.

The archived static-grid configuration labels retrieval selectors as matched, induction-prefix and induction-attention selectors as transfer, and truthfulness selectors as null for the retrieval task. Table 12 reports the Top-30, $\alpha = 0$ subset of that grid; it was not assigned a separate confirmatory multiplicity family.

H Scope Comparison with Closest Work

Table 25: **Scope of representative related work.** Entries summarize each paper's main empirical questions. Our study brings together head distributions, overlap, training-time changes, and interventions, while keeping the evidence for each question separate.

Work	Primary scope	Unit and coverage	Primary validation and boundary
Olsson et al. (2022)	Induction formation and in-context learning	Attention heads in decoder-only Transformers from small attention-only models to 13B	Training co-emergence, architecture perturbations, ablation, and reverse engineering; causal evidence is strongest in small models.
Wu et al. (2025)	Long-context retrieval	Retrieval heads across model families, scales, and base/derived variants	Copy-event detection, base–variant similarity, and Top-K versus random masking; broad model coverage centered on retrieval.
Trott (2025)	Axes of mechanistic generalization	1-back heads across pretraining and random refits of four Pythia sizes	Separates functional, developmental, positional, relational, and configurational correspondence; a conceptual and developmental framework.
Bali et al. (2026)	Attention-head stability across refits	Representation-level correspondences across independently trained instances	Quantifies head stability relevant to circuit universality; broader model-refit variation than our within-endpoint detector-input repeats.
Joshi et al. (2026)	Causal calibration of interpretability claims	General observational, interventional, and counterfactual claims	Aligns claims with estimands, causal evidence levels, and assumptions; a diagnostic framework rather than a functional-head benchmark.
Lin & Liu (2026)	Identification-assumption disclosure	Literature audit across mechanistic-interpretability strands	Separates validation from identification and proposes a disclosure norm; our study applies these distinctions to specific head comparisons.
Wang et al. (2025)	Cross-architecture mechanistic similarity	Matched components and computations across language-model architectures	Tests whether mechanisms correspond across architectures, with a focus on explicitly matching mechanisms.
Bhardia et al. (2026)	Universality of numerical-comparison circuits	Numerical-comparison behavior and circuits across Transformer families	Shows that similar behavior need not imply the same mechanism; our work examines related questions across detector signals, models, and interventions.
This work	Systematic comparison of six measured head signals	Five base models, one Pythia-1.4B trajectory, and Qwen/Llama causal endpoints	Studies distributions, overlap, training-time change, and interventions while distinguishing repeatability from exact head identity and coefficient reuse; intervention coverage remains limited.

I Source Records and Limits of the Evidence

The August 12 snapshot contains 14 formal packages. Validators check model revisions, hashes, counts, head indices, source records, controls, and complete grids. Historical runs that fail these checks are excluded. The plotting script rebuilds figures from fixed CSVs and checks row counts and uniqueness. A deterministic script reconstructs every RQ4 source–target cell, abstention, oracle, no-source baseline, paired comparison, and summary.

For the older static RQ2 result, a separate registry gives all four ordered Top-30 layer–head lists, set/order hashes, consensus-detail and consumed-score hashes, and a four-row link to the aggregate behavioral results. It reproduces the 20/19 structural intersections and verifies that each consumed score file has the same Top-30 order as its consensus detail. The static package manifest, not the aggregate tables themselves, records the common RQ1 job-manifest hash. The aggregate archive does not retain the selected-head list for each historical condition. The registry therefore reconstructs the expected sets; it cannot show which mask each condition executed. The later selection-frozen registry is a separate artifact, with 22/20 shared heads, and is checked against all six packaged run records.

The original matched-induction RQ2 snapshot retains 12 aggregate rows but not the paired per-example values. Exactly reproducing its example-bootstrap intervals would require a new forward run. This does not affect the 18 selection-frozen contrasts, which can be reconstructed from packaged condition records. For those induction runs, the archived YAML retains template seed 42, but the execution wrapper used seed 20260811 for the 64-example task. The supplement records this correction and the reconstructed input hashes. Its tokenizer-only check records prompt, full-text, and exact-target token-ID hashes for both seeds. The pinned Qwen tokenizer and a pinned public Llama-2 mirror pass the exact-prefix check. The gated Llama endpoint's tokenizer remains unverified.

Finally, the unvalidated zero induction-to-retrieval *detector rescore* shows neither dependence nor independence. It is separate from the behavioral comparison of overlapping head sets in Table 12. Mixed effects are not exact null effects, and stable detector scores do not show that behavior is preserved. We apply these limits throughout the paper.

J PYTHIA-2.8B SIX-CHECKPOINT REPLICATION

Six immutable checkpoints of `pythia-2.8b-deduped` (steps 1,000, 14,000, 36,000, 71,000, 107,000, 143,000) were scored with the detectors, prompts, seeds, and thresholds of the Pythia-1.4B study. Sustained agreement uses active-union Spearman $\geq .8$ and Top-K Jaccard $\geq .5$ at the checkpoint and every later sampled checkpoint. Intervals are 2,000 paired bootstraps over prompts (induction-prefix over its five seeds; retrieval over common successful cells).

Table 26: **First sustained-agreement checkpoint on Pythia-2.8B** (point estimate; bootstrap median and 95% interval in brackets). Values are the earliest sampled step and are interval-censored by the six checkpoints.

Signal	$K = 10$	$K = 20$	$K = 50$
Previous-token	36k [36k, 36k]	14k [14k, 36k]	14k [14k, 14k]
Duplicate-token	36k [36k, 36k]	36k [36k, 36k]	36k [36k, 36k]
Induction-prefix	14k [14k, 36k]	14k [14k, 36k]	14k [14k, 36k]
Induction-attention	107k [71k, 143k]	71k [36k, 71k]	71k [71k, 107k]
Retrieval	71k [71k, 71k]	71k [71k, 71k]	71k [71k, 71k]

Final-checkpoint maxima are .850 (previous), .547 (duplicate), .957 (induction-prefix), 1.000 (induction-attention), and 1.088 (retrieval); at step 1,000 the corresponding values are .897, .068, .980, .919, and .263.

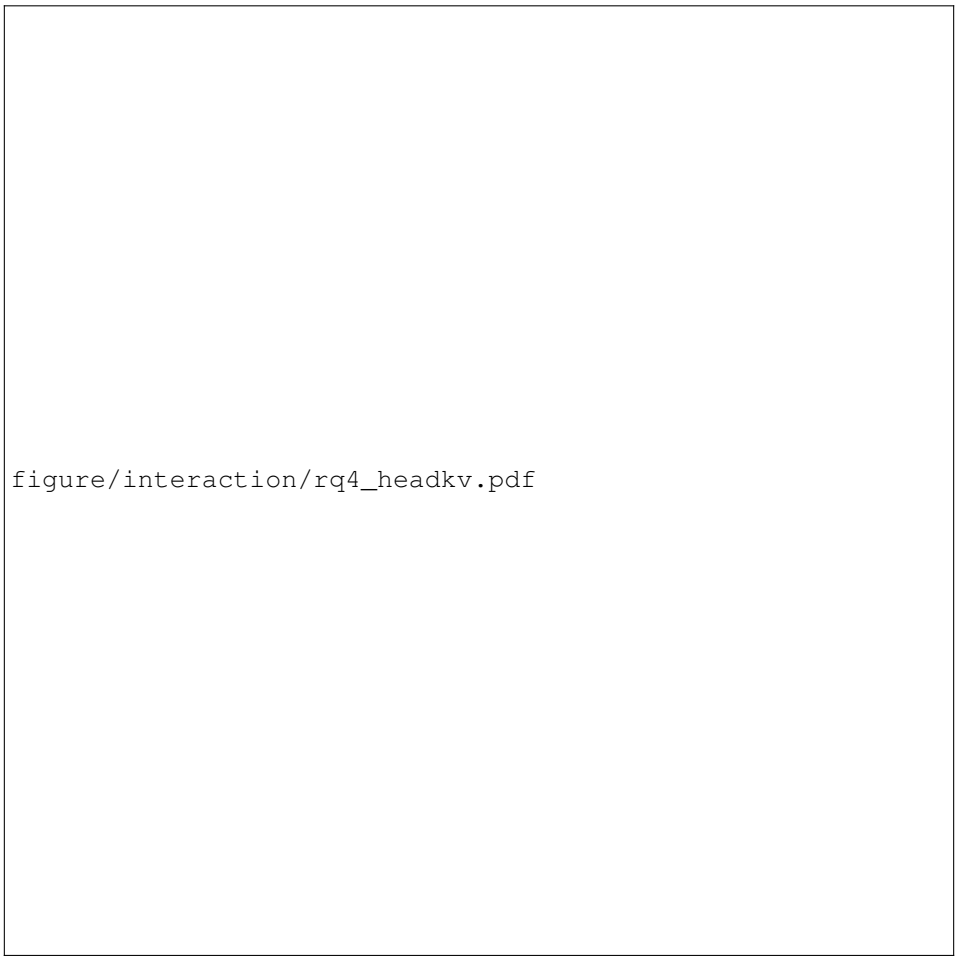

Figure 8: **Copy-circuit labels beat HeadKV's own retrieval signal; previous-token heads score below random.** (a) Six-task LongBench mean versus per-head KV budget for eight importance signals. (b) Ranking at budget 64.

K HELD-OUT ABLATION: FULL VALUES

Table 27: **Removing heads chosen before the held-out test.** Cells show mean performance changes and bootstrap 95% intervals; negative is worse. The separate p_H columns use sign-based tests with Holm correction. The final test rules were reconstructed after outcomes, so the comparisons are exploratory. Activation-size controls match approximate RMS before W_O, not the final residual contribution. Layer-matched results appear in the appendix.

Model	**Selector**	**vs. unchanged**		**vs. activation-size controls**	
		Mean Δ [95% CI]	p_H	**Mean Δ [95% CI]**	p_H
Qwen2.5	Retr. (ROUGE-1)	$-.366$ [$-.483$, $-.238$]	.031	$-.365$ [$-.481$, $-.235$]	.031
	Ind.-attn. (mean logp)	$-.011$ [$-.021$, $-.003$]	.032	$-.001$ [$-.038$, .029]	1.000
	Truth. (MC2)	$-.007$ [$-.044$, .032]	1.000	.050 [.011,.094]	.245
Llama-2	Retr. (ROUGE-1)	$-.161$ [$-.271$, $-.045$]	.773	$-.143$ [$-.267$, $-.014$]	.773
	Ind.-attn. (mean logp)	.008 [$-.001$, .020]	1.000	.012 [$-.001$, .030]	1.000
	Truth. (MC2)	$-.007$ [$-.035$, .020]	1.000	$-.003$ [$-.037$, .034]	1.000

L DECISIVE ABLATION: SHARED VERSUS LABEL-EXCLUSIVE HEADS

For each model the induction-prefix and retrieval Top-30 sets (local detection, single seed) are split into induction-only, shared, and retrieval-only heads. We remove each part, the two full Top-30 sets, their union, and five layer-matched random sets per part (drawn outside the union, seeds 42–46). Removal uses the $\alpha = 0$ operator, which zeroes the selected head outputs before W_O, in the needle tester of Wu et al. (2025). The 1k–8k grid has five context lengths and five depths (25 cells); the long grids have three lengths and three depths (9 cells). Intervals are 10,000 paired bootstraps over cells. None of the three models produces the needle phrase when the needle is absent from the context, so recall at ceiling is not memorization.

Table 28: **Needle recall (ROUGE-1, %) after removing each set with $\alpha = 0$.** The last column is the shared-set recall minus the mean of its five random controls, with a paired cell bootstrap interval.

Model	Grid	None	Ind.-only	Retr.-only	Shared	Ind. Top-30	Retr. Top-30	Shared − random
Llama-3-8B-Instruct	1k–8k	98.7	98.7	97.3	69.7	71.7	66.3	-25.0 [$-31.3, -19.0$]
Mistral-7B-Instruct	1k–8k	100.0	100.0	100.0	92.7	96.7	87.7	-7.3 [$-18.3, 0.0$]
Mistral-7B-Instruct	12k–30k	100.0	100.0	100.0	72.2	86.1	68.5	-25.6 [$-51.1, 0.0$]
Qwen2.5-7B-Instruct	1k–8k	100.0	100.0	100.0	99.7	100.0	95.7	-0.3 [$-1.0, 0.0$]
Qwen2.5-7B-Instruct	10k–16k	100.0	100.0	100.0	100.0	100.0	93.5	0.0 [0.0, 0.0]

Random controls for the exclusive sets are 88.0–98.7 on Llama-3 and 100.0 on Mistral and Qwen. The exclusive sets never fall below them: induction-only $+4.8$ [$1.2, 9.6$] and retrieval-only $+2.1$ [$0.5, 4.3$] against their controls on Llama-3, and 0.0 elsewhere. Shared-set recall by context length on Llama-3 is 86.7, 86.7, 71.7, 58.3, and 45.0 at 1,000, 2,750, 4,500, 6,250, and 8,000 tokens. On Mistral it is 100.0, 100.0, 100.0, 81.7, and 81.7 on the same grid, and 77.8, 69.4, and 69.4 at 12,000, 21,000, and 30,000 tokens. Qwen's 152k-token vocabulary limits the tester to 16,000 tokens on a 46 GB GPU.

Table 29 reports the effect of removing one part after another part is already removed. Adding an exclusive set to the shared set changes recall by at most 5.0 points on every grid except the Mistral long grid, where the induction-only heads add $+13.9$ [$-5.6, 38.0$] on nine cells. The only interval excluding zero is Qwen's retrieval-only set (-4.0 [$-8.0, -0.7$]), a cost the set does not show on its own. In the reverse order the shared set's cost after an exclusive set is removed equals its cost alone on Llama-3 (-27.0 and -31.0 versus -29.0). At $K = 30$ the exclusive sets are therefore not necessary on their own and at most weak backups for the shared set. The local single-sample retrieval Top-30 of Qwen2.5-7B-Instruct shares 24 of 30 heads with the frozen three-sample set of Section M; the other two models have no frozen registry.

Table 29: **Conditional effects of removing one part given another (recall points, paired cell bootstrap 95% intervals).** "$X \mid Y$" is recall after removing $X \cup Y$ minus recall after removing Y; "Shared alone" is against the unchanged model.

Model	Grid	Ind.-only | shared	Retr.-only | shared	Shared | ind.-only	Shared | retr.-only	Shared alone
Llama-3-8B-Instruct	1k–8k	$+2.0$ [$-7.3, 12.0$]	-3.3 [$-9.3, 2.7$]	-27.0 [$-33.7, -20.3$]	-31.0 [$-38.3, -23.7$]	-29.0 [$-37.0, -21.3$]
Mistral-7B-Instruct	1k–8k	$+4.0$ [$0.0, 11.7$]	-5.0 [$-13.3, 0.0$]	-3.3 [$-10.0, 0.0$]	-12.3 [$-25.3, -0.7$]	-7.3 [$-18.3, 0.0$]
Mistral-7B-Instruct	12k–30k	$+13.9$ [$-5.6, 38.0$]	-3.7 [$-11.1, 0.0$]	-13.9 [$-34.3, 0.0$]	-31.5 [$-63.0, 0.0$]	-27.8 [$-55.6, -7.4$]
Qwen2.5-7B-Instruct	1k–8k	$+0.3$ [$0.0, 1.0$]	-4.0 [$-8.0, -0.7$]	0.0 [0.0, 0.0]	-4.3 [$-8.7, -0.7$]	-0.3 [$-1.0, 0.0$]
Qwen2.5-7B-Instruct	10k–16k	0.0 [0.0, 0.0]	-6.5 [$-15.7, 0.0$]	0.0 [0.0, 0.0]	-6.5 [$-15.7, -0.9$]	0.0 [0.0, 0.0]

Removing the induction-only or retrieval-only heads leaves recall at 97.3–100.0 on every model and grid. Shared-set recall is 69.7 on Llama-3, 92.7 and 72.2 on Mistral's two grids, and 99.7–100.0 on Qwen; the induction and retrieval Top-30 sets themselves reach 71.7/66.3, 96.7/87.7, 86.1/68.5, 100.0/95.7, and 100.0/93.5 on the same rows of Table 28. Repeating the split at $K = 20$ and $K = 40$ on Llama-3-8B-Instruct (Table 30) gives the same pattern. The exclusive sets cost at most 1.3 points alone and never fall below their layer-matched controls; the shared set costs 14.1 and 27.9 points against its controls. The conditional effect of an exclusive set once the shared set is removed ranges from -9.0 to $+6.7$ across the three cutoffs, with no consistent sign. Two of the six intervals exclude zero, one in each direction. Exclusive heads are therefore not necessary for retrieval at any tested

cutoff, but their conditional effects are not exactly zero, so they may act as partial backups for the shared set.

Table 30: **The three-way split at three cutoffs on Llama-3-8B-Instruct, 1k–8k grid (recall, %).** Sizes are induction-only/shared/retrieval-only. Intervals are paired cell bootstraps (25 cells).

K	Sizes	None	Ind.-only	Retr.-only	Shared	Ind. Top-K	Retr. Top-K	Shared − random	Ind.-only | shared	Retr.-only | shared
20	7/13/7	98.7	98.7	97.3	80.0	86.7	71.0	$-14.1\,[-23.5, -5.9]$	$+6.7\,[3.0, 10.7]$	$-9.0\,[-17.0, -1.7]$
30	11/19/11	98.7	98.7	97.3	69.7	71.7	66.3	$-25.0\,[-31.3, -19.0]$	$+2.0\,[-7.3, 12.0]$	$-3.3\,[-9.3, 2.7]$
40	15/25/15	98.7	97.3	98.7	63.3	55.0	65.3	$-27.9\,[-36.0, -19.2]$	$-8.3\,[-11.7, -5.3]$	$+2.0\,[-1.0, 5.3]$

The choice of operator matters. The needle tester's own mask zeroes the selected heads' queries, which makes their attention uniform rather than removing their output. Under that operator the same sets behave differently: retrieval-only 100.0, shared 53.3, and induction-only 24.3. Layer-matched random sets, however, range from 6.3 to 100.0, and a single-head sweep shows that two layer-2 heads (L2H25, L2H27) collapse recall to 26.0 and 0.0 on their own. Uniform attention in an early-layer head disrupts the whole forward pass; removing the head's output does not. We report the $\alpha = 0$ operator, which matches RQ4, in the main text.

M Retrieval activation patching

Both models use 12 held-out needles at three insertion depths in 2,048-token contexts, giving 36 token-aligned clean/corrupt pairs that differ only at the answer span. The intervention targets the frozen Top-30 retrieval heads, with five exact-layer random sets and five activation-RMS-matched random sets (seeds 42–46) as controls. We compute exact full-vocabulary KL at every answer-token position and 10,000 paired bootstraps. The patch exceeds the layer-matched random patches by 24.7/17.0 and the activation-size-matched random patches by 21.9/16.2 log-probability units on Qwen/Llama. Ablating the selected heads costs 5.56/1.27 units against the clean run and stays below both control families; exact output KL after patching is 5.2/1.9. The recovery definition, control construction, and guardrails were declared in the run manifest before the formal jobs. Ablation guardrails: induction target log-probability $-2.58\,[-3.12, -2.09]$ / $-1.80\,[-2.14, -1.49]$, benign-QA accuracy 0, fixed-corpus NLL $+.036$ / $+.012$ (Qwen / Llama).