Different Heads, Same Fragile Tasks: Cross-Model Correspondence in Retrieval-Head Effects

Ananya Anand [*1] **Harry Ilanyan** [*2] **Ira Pathak** [3] **Derrick Sual** [4] **Eric Xia** [5]

Abstract

Long-context language models rely on a sparse subset of attention heads for factual retrieval, but the extent to which their causal effects are shared across tasks and model families is not known. We identify query-focused retrieval heads separately for eight SEC 10-K fact-extraction tasks, ablate each source task's top-K heads, and measure the effects on every target task in three open-weight 7–8B instruction-tuned models. Within a model, the selected head sets overlap and their ablation effects transfer across tasks. Across models, source disruptivity has weak task-level correspondence ($R^2 \leq 0.15$ after removing the average effect of ablation size), whereas target sensitivity has partial, model-pair-dependent correspondence (R^2 from 0.20 to 0.60). In four-task activation-patching experiments, QR-head patching has the largest approximate model-level median rescue in all three models (0.50–1.11 nats, versus at most 0.32 nats for control medians), but one Llama task has a large negative rescue. The cross-model result is therefore about task-level effects: target vulnerability corresponds more consistently than the disruptivity of source-task head sets.

1. Introduction

Recent work identifies sparse subsets of attention heads, called *retrieval heads*, whose ablation harms long-context factual retrieval (Wu et al., 2025b; Zhang et al., 2025). Retrieval-head scores are also used in methods for retrieval and KV-cache compression (Fu et al., 2025; Zhang et al., 2025). Whether heads selected on one task remain causally important on another, and whether the same task-level pattern appears across model families, is not known. This is one instance of the broader problem of comparing causal structure across models (Sharkey et al., 2025).

We study this question on a controlled SEC fact-extraction benchmark built from EDGAR-CORPUS (Loukas et al., 2021). Eight long-context extraction tasks (`registrant_name`, `ceo_lastname`, `employees_count_total`, and others) share a document domain and prompt format but differ in answer type and query-to-answer distance. For each (task, model) pair, we identify query-focused retrieval heads with QRScore (Zhang et al., 2025), ablate the top-K heads from one source task, and evaluate every target task. We also test whether restoring clean QR-head activations rescues answer log-probability after ablation. We apply both analyses to Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3.

Two summaries of the cross-task ablation matrix appear throughout. *Target sensitivity* of a task t measures how much its accuracy drops when heads from *other* tasks are ablated; it describes the task's vulnerability. *Source disruptivity* of a task s measures how much ablating s's top-K heads damages *other* tasks. We report three main results:

1. **Within a model, QR-ranked head sets overlap and their ablation effects transfer across tasks.** Pairwise within-model Jaccard between task head sets ranges from 0.14 to 0.88 at K=16 (mean $\approx$ 0.45), well above the expected Jaccard for independent random 16-head subsets (0.008 for the 1024-head models, 0.011 for Qwen's 784). The `ceo_lastname` sets have consistently lower overlap with the other tasks. Ablating heads selected for one task generally harms accuracy on other tasks, consistent with shared retrieval dependencies (Zhang et al., 2025) (Section 4.2).

2. **Target sensitivity and source disruptivity vary by task.** At K=16, per-task target sensitivity on Llama spans 0.20 (`headquarters_state`) to 0.87 (`ceo_lastname`), and source disruptivity also varies

[*]Equal contribution [1]University of California, Los Angeles, Los Angeles, California, USA [2]University of California, Berkeley, Berkeley, California, USA [3]Georgia Institute of Technology, Atlanta, Georgia, USA [4]Stevens Institute of Technology, Hoboken, New Jersey, USA [5]Brown University, Providence, Rhode Island, USA. Correspondence to: Harry Ilanyan <harry_ila@berkeley.edu>, Ananya Anand <ananya04@g.ucla.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

across source tasks. For comparison, pooled random-head ablation at the same K reduces accuracy by 2.4pp.

3. **At the point-estimate level, target sensitivity corresponds more strongly across models than source disruptivity.** After K-residualization, target-sensitivity correspondence is model-pair dependent (Pearson R^2 from 0.20 for Llama–Qwen to 0.60 for Qwen–Mistral), while source disruptivity stays at $R^2 \leq 0.15$ for all three pairs. The bootstrap intervals are wide, so this comparison is descriptive.

Different head-ranking procedures recover overlapping but non-identical sets, so our claims concern only QRScore-selected heads. The same-layer activation-patching control (Section 4.4) tests whether the median rescue is specific to QR-ranked heads rather than generic to the layers in which they occur.

2. Related Work

Retrieval heads. Wu et al. (2025b) characterize retrieval heads as universal, sparse, and intrinsic, identified by copy-paste attention patterns on Needle-in-a-Haystack inputs. Zhang et al. (2025) introduce query-focused retrieval heads (QRHead) and an attention-mass retriever (QRRetriever), validated on BEIR and LongMemEval. Fu et al. (2025) use retrieval-importance scores for KV-cache compression. Lin et al. (2026) show that retrieval-head rankings shift with context; our dataset-aggregated rankings should likewise not be read as fixed model properties. We use the QRScore detector of Zhang et al. (2025) and study the resulting heads under ablation between tasks and models.

Causal head ablation. Attention-head ablation has long been used to identify task-relevant components (Michel et al., 2019; Voita et al., 2019; Olsson et al., 2022; Wang et al., 2023). We use ablation for cross-task transfer and activation patching with same-layer controls to test whether detected heads carry answer information.

Cross-task head reuse. Merullo et al. (2024) find substantial circuit-component reuse between two dissimilar tasks in GPT-2, whereas Chowdhary et al. (2025) report task-dependent "super-heads" with weak overlap across syntactic and numerical task families. We measure all pairwise transfers among eight related retrieval tasks and distinguish source disruptivity from target sensitivity across three model families.

3. Methods

Benchmark. From EDGAR-CORPUS (Loukas et al., 2021) we derive a long-context fact-extraction benchmark with eight tasks: `registrant_name`, `headquarters_city`, `headquarters_state`, `incorporation_state`, `incorporation_year`, `employees_count_total`, `ceo_lastname`, and `holder_record_amount`. Each evaluation instance pairs a query with one gold section and distractor sections from the same filing. We place the gold section at the document midpoint to hold its position constant. Detection and evaluation contain no filings in common. Llama and Mistral are evaluated on n=192 instances, with 24 per task. Qwen is evaluated on n=241 instances, with per-task counts from 24 to 41 (Section A).

Models. We evaluate Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023; Grattafiori et al., 2024; Qwen Team, 2024). For every model, we run head detection, ablation, cross-task transfer, and activation patching. Architectural details and head counts are in Section A.

Detection. For each (task, model) pair, we aggregate the post-softmax attention mass from query tokens to gold-evidence chunks and rank all heads by QRScore (Zhang et al., 2025). The top-K slice defines the task-specific set $\mathcal{H}_K^{(t)}$. We also form a combined SEC ranking by pooling detection examples across the eight tasks. Cross-task transfer uses the task-specific sets; aggregate ablation curves, layer histograms, and activation patching use the combined ranking.

Ablation protocol. Forward pre-hooks on each layer's o_{proj} zero out the slices corresponding to ablated heads. We sweep $K \in \{0, 8, 16, 32, 48, 64, 96, 128\}$. Three independent random ablations (seeds $42, 123, 456$) and the head sets detected on Natural Questions (Kwiatkowski et al., 2019) and LongMemEval (Wu et al., 2025a) serve as baselines.

Cross-task transfer. For source task s and target task t, we ablate $\mathcal{H}_K^{(s)}$ and evaluate on t, recording the drop $\Delta_{s\to t}(K) = \text{Acc}_t(0) - \text{Acc}_t(K)$. The $8 \times 8 \times |K|$ tensor is summarized by per-task target sensitivity $\bar{\Delta}_{\cdot\to t}(K) = \frac{1}{|S|-1}\sum_{s\neq t}\Delta_{s\to t}(K)$ (the off-diagonal column mean) and source disruptivity $\bar{\Delta}_{s\to\cdot}(K) = \frac{1}{|T|-1}\sum_{t\neq s}\Delta_{s\to t}(K)$ (the off-target row mean). Both statistics exclude within-task cells and therefore summarize transfer between distinct tasks. Section B reports a diagonal-inclusive sensitivity variant. The specificity index is the on-target drop minus the off-target mean drop; its intervals use 10,000 parametric-bootstrap draws (Section C).

K-fixed effects (residualization). To remove the average effect of K on accuracy drop, we subtract the within-K task mean from each (task, K) cell and flatten the result

to a 56-element vector (8 tasks $\times$ 7 non-zero K values). Cross-model correspondence is the Pearson R^2 between two models' residualized vectors.

Activation patching. We use the top-16 heads from the combined SEC ranking and cache each head's clean activation immediately before the attention output projection. On the same example, we ablate those heads and patch their clean outputs back at answer-prediction positions. Teacher-forced gold-answer log-probability avoids changes caused by generation branching. The *rescue* is patched-minus-ablated gold-answer log-probability, summed over answer tokens (nats). The experiment covers four tasks: `ceo_lastname`, `employees_count_total`, `headquarters_state`, and `registrant_name`. We first average rescue within task, then use the median of the four task means as a descriptive model-level summary; all task means appear in Section I. QR-head patches are compared with bottom-ranked, random, and same-layer non-QR controls.

4. Results

4.1. QRScore-Selected Heads Are Causally Relevant

The top-16 heads in each model's combined SEC ranking lie mainly in the middle layers, extending later in Qwen and Mistral (Figure 1). At K=16, ablating these sets sharply reduces accuracy and drives some tasks to near zero; random and NQ-selected ablations at the same K are much smaller. QRScore ranks heads by attention mass rather than ablation effect, so this degradation is not guaranteed by the selection rule. We therefore also examine held-out cross-task transfer and activation-patching rescue (Section 4.4). The large ablation effects are concentrated in only 16 of 784–1024 query heads ($\approx$ 1.6–2.0%).

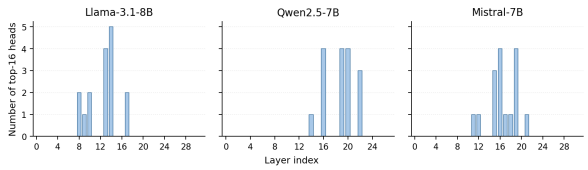

Figure 1. Layer distribution of the top-16 combined SEC QR-heads per model. The heads lie mainly in middle layers, with later-layer mass in Qwen and Mistral.

4.2. Within a Model, Ablation Effects Transfer Across Tasks and Detection Benchmarks

Using the combined SEC ranking, Llama accuracy falls by 99/192 instances (51.6pp) at K=16; the rounded endpoint accuracies are 91.1% and 39.6%. Random ablation at the same K averages a 2.4pp drop across three seeds (Figure 2). The corresponding combined-ranking drops are 55.2pp for Qwen2.5-7B-Instruct (n=241) and 37.0pp for Mistral-7B-Instruct-v0.3 (n=192). Thus, combined-SEC head ablation has a large aggregate effect in all three models.

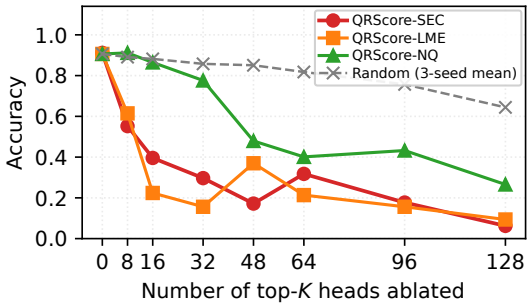

Figure 2. Per-method accuracy curves on the Llama-3.1-8B-Instruct SEC subset (n=192). QRScore-SEC and QRScore-LME ablations cause large drops; QRScore-NQ stays near the random baseline at small K.

The task-specific transfer matrix shows that most source-task head sets affect several targets. Specificity (on-target minus off-target drop) is non-positive for 5 of 8 Llama sources at K=16; for 2 of those, the unadjusted bootstrap 95% CI excludes zero (Section C). The CIs for `ceo_lastname` and `employees_count_total` are the only ones that exclude zero on the positive side. Their on-target drops exceed their off-target means, although both sets still affect other tasks. For most sources, ablating the top-16 heads harms other tasks as much as or more than the source task itself, which is consistent with a shared within-model head pool.

We also ablate heads selected on other retrieval benchmarks. On Llama, LME-selected ablation reduces SEC accuracy by 68.2pp at K=16, compared with 51.6pp for the combined SEC ranking; the NQ-selected drop is 4.2pp, close to the 2.4pp random baseline (Section G). The transfer from LME is not consistent across models, so we do not interpret it as general cross-domain correspondence. A possible explanation for `ceo_lastname`'s distinct profile is that it requires role-to-person resolution in variably phrased narrative text and surname normalization, unlike more templated fields; we did not test this explanation.

4.3. Cross-Model Correspondence Differs Between Sensitivity and Disruptivity

Cross-model correspondence here refers to task-level effects, not to alignment of individual head indices. After K-residualization, target sensitivity has $R^2 = 0.20$ (Llama–Qwen), 0.48 (Llama–Mistral), and 0.60 (Qwen–Mistral). Source disruptivity has $R^2 = 0.02$, 0.04, and 0.15. At the point-estimate level, sensitivity R^2 is higher for each pair (0.20 vs. 0.02, 0.48 vs. 0.04, and 0.60 vs. 0.15).

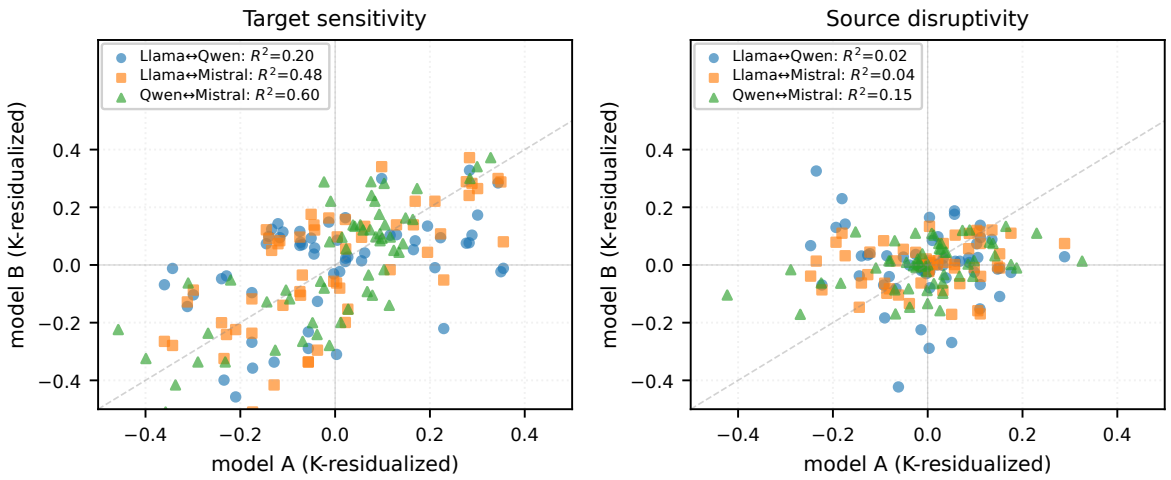

Figure 3. K-residualized cross-model correlations. Left: target sensitivity (per-task off-diagonal column means of the drop matrix). Right: source disruptivity (per-task off-target row means). Each point is one (task, K) cell; marker shapes distinguish model pairs. Target-sensitivity correspondence ranges from R^2=0.20 (Llama–Qwen) to 0.60 (Qwen–Mistral), while source-disruptivity correspondence is at most 0.15. At the point-estimate level, sensitivity R^2 is higher in every pair. Bootstrap intervals are in Section B.

These estimates are descriptive. With only eight tasks, the task-block bootstrap intervals are wide: sensitivity R^2 CIs are 0.01–0.55, 0.14–0.76, and 0.23–0.83 across the three pairs. The sensitivity and disruptivity intervals overlap, so the data do not establish a statistically resolved difference between them. The signed Llama–Qwen disruptivity correlation is negative ($r = -0.13$). Overall, target sensitivity shows partial, model-pair-dependent correspondence, while source-disruptivity correspondence is weak.

An exploratory eight-task analysis in Section D estimates negative correlations between mean cross-model fragility and both query-to-answer distance (r=−0.72) and answer length (r=−0.64); the baseline-accuracy estimate is r=−0.09. The reported p-values are unadjusted for the three comparisons. We also detect no layer-position difference between fragile and robust tasks (Section F).

The qualitative pattern also appears under the two definitions that place little or no weight on the diagonal. `off_target_mean` excludes the on-target cell, whereas `full_row_mean` includes it at $1/8$ weight; both have cross-model $R^2 \leq 0.16$ (Section B). When the on-target drop enters at full weight, the resulting measures also contain target-sensitivity information and their R^2 values rise. We use `off_target_mean` throughout; it reproduces the reported disruptivity values within $|\Delta| \leq 0.0035$.

4.4. QR-Head Patching Produces Larger Median Rescues Than Controls

We test whether restoring clean QR-head activations recovers gold-answer log-probability after ablation. The comparison uses the combined SEC top-16 set and the three control patches defined in Section 3.

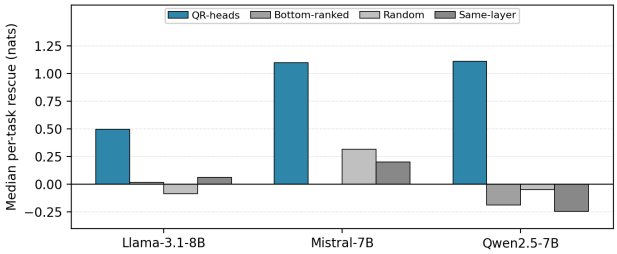

Figure 4. Activation-patching rescue for QR heads and three controls (bottom-ranked, random, and same-layer non-QR heads). Bars are approximate medians of four per-task means: 0.50 nats for Llama, 1.10 for Mistral, and 1.11 for Qwen under QR-head patching. All task-level values appear in Section I.

Under the median-of-four-tasks summary, QR-head patching has the largest rescue in every model (approximately 0.50, 1.10, and 1.11 nats for Llama, Mistral, and Qwen). The largest control median is approximately 0.32 nats, but individual control cells range from −2.26 to +2.16 nats (Section I). At the model-median level, the same-layer control suggests that layer position alone does not explain the QR result. The task-level variation is much larger than the differences between these medians.

This conclusion depends on the aggregation. Three Llama tasks have positive QR rescue (+0.35 to +1.02 nats), but `employees_count_total` has −6.08 nats, making the four-task mean negative even though the median is positive. This negative rescue remains unexplained and prevents a claim that patching reliably recovers Llama performance across tasks.

5. Limitations

Population-level evidence and statistical power. We intervene on attention-head populations and measure changes in output accuracy and gold-answer log-probability. This provides evidence about causal effects at the head-population level, not a feature-level decomposition or an alignment of individual heads across models. The cross-model analysis contains only eight tasks; its bootstrap intervals are wide, and the exploratory task-attribute correlations are not adjusted for multiple comparisons.

Activation patching. Patching covers four of the eight tasks, with $n{=}4$ to $n{=}16$ retained instances per task and model. The four per-task means were retained only to two decimal places, not as raw instance-level log-probabilities, so the reported medians are approximate and no uncertainty interval or normalized recovery can be reconstructed. Rescue is summed over answer tokens and therefore depends on answer length and tokenization; magnitudes should be compared with controls within a model and task, not directly across model families. The unexplained negative Llama result further limits this evidence.

Scope. The experiments cover eight related SEC extraction tasks, three 7–8B instruction-tuned models, one prompt and distractor construction, and an unbalanced Qwen evaluation set. Architecture, pretraining data, optimization, and instruction tuning vary jointly, so cross-model differences cannot be assigned to any one factor. The LongMemEval and Natural Questions results are model-specific baseline experiments, not evidence of cross-domain generalization. We did not vary model scale within a family.

6. Conclusion

Within each model, QR-head sets detected separately for different tasks overlap substantially, and ablating one task's heads often harms other tasks. This supports within-model sharing among QRScore-selected head populations.

Across models, source disruptivity has weak task-level correspondence, whereas target sensitivity has partial, model-pair-dependent correspondence after controlling for K. The point estimates differ in every model pair, but the wide bootstrap intervals do not resolve that contrast statistically. QR-head patching has the largest median rescue under the four-task aggregation, although the negative Llama result and large control cells limit the inference. The supported cross-model result is therefore a difference between two task-level summaries: target vulnerability corresponds more consistently than source-task ablation effects. Individual heads were not aligned across models.

References

Chowdhary, P., Chin, P., and Chakrabarty, D. K-MSHC: Unmasking minimally sufficient head circuits in large language models with experiments on syntactic classification tasks. *arXiv preprint arXiv:2505.12268*, 2025.

Fu, Y., Cai, Z., Asi, A., Xiong, W., Dong, Y., and Xiao, W. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.19258.

Grattafiori, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276.

Lin, Y., Li, Z., Xing, Y., He, P., Cui, Y., Li, Y., Ding, B., Zhou, J., and Tang, J. Retrieval heads are dynamic. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15710–15729, San Diego, California, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.715.

Loukas, L., Fergadiotis, M., Androutsopoulos, I., and Malakasiotis, P. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pp. 13–18, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.econlp-1.2.

Merullo, J., Eickhoff, C., and Pavlick, E. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=fpoAYV6Wsk.

Michel, P., Levy, O., and Neubig, G. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 14014–14024, 2019.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Saunders, W., Michaud, E. J., Casper, S., Tegmark, M., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. arXiv:2501.16496.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. 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 (ACL)*, pp. 5797–5808, Florence, Italy, 2019. doi: 10.18653/v1/P19-1580.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.10813.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025b. arXiv:2404.15574.

Zhang, W., Yin, F., Yen, H., Chen, D., and Ye, X. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 23791–23805, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1214. arXiv:2506.09944.

A. Models and Benchmark Details

Table 1 reports architectural statistics for the three models studied.

Table 1. Model architectural statistics. Query heads are the units of ablation.

Model	Layers	Q heads/layer	KV heads/layer	Total Q heads
Llama-3.1-8B-Instruct	32	32	8	1024
Mistral-7B-Instruct-v0.3	32	32	8	1024
Qwen2.5-7B-Instruct	28	28	4	784

Per-model evaluation set sizes. Llama and Mistral each use n=192 (24/task). Qwen uses n=241: `registrant_name` 41, `headquarters_city` 28, `headquarters_state` 27, `incorporation_state` 32, `incorporation_year` 30, `employees_count_total` 24, `ceo_lastname` 35, and `holder_record_amount` 24. The comparison in Figure 2 restricts LME and NQ accuracy to the same 192 Llama SEC instances.

B. K-Fixed-Effects Estimates Under Alternative Definitions

Table 2 reports cross-model Pearson R^2 after K-residualization for the off-diagonal sensitivity definition, its diagonal-inclusive variant, and four disruptivity definitions. The two sensitivity variants differ by at most 0.014 in R^2. Only `off_target_mean` reproduces the reported disruptivity values within $|\Delta| \leq 0.0035$.

Table 2. K-residualized cross-model R^2. The off-diagonal column mean is the sensitivity definition used in the body; the diagonal-inclusive variant gives similar values. `off_target_mean` excludes the diagonal, while `full_row_mean` includes the on-target cell at $1/8$ weight; both remain ≤ 0.16. Measures in which $\Delta_{s\to s}$ enters at full weight also contain target-sensitivity information. Parenthetical r is the signed correlation. The 56 flattened points are repeated measures of eight tasks; bracketed intervals are task-block bootstrap 95% CIs (20,000 resamples).

Measure	Definition	Llama–Qwen	Llama–Mistral	Qwen–Mistral
Sensitivity	off-diagonal column mean (canonical)	0.197 ($r = +0.44$) [0.01, 0.55]	0.483 ($r = +0.69$) [0.14, 0.76]	0.596 ($r = +0.77$) [0.23, 0.83]
Sensitivity	column mean (incl. diagonal variant)	0.184	0.472	0.593
Disruptivity	off-target row mean	0.018 ($r = -0.13$) [0.00, 0.25]	0.039 ($r = +0.20$) [0.00, 0.26]	0.153 ($r = +0.39$) [0.01, 0.32]
Disruptivity	full row mean (incl. on-target)	0.054	0.002	0.096
Disruptivity	on-target drop only	0.031	0.332	0.302
Disruptivity	specificity index (−off-target)	0.133	0.466	0.434

C. Per-Task Specificity With Bootstrap CIs (Llama)

Table 3 reports per-task specificity at K=16 on Llama. The 95% CIs use 10,000 parametric-bootstrap draws per source–target cell, treating baseline and ablated correct counts as independent binomials. This independence approximation is conservative because the evaluations use the same examples. 5 of 8 tasks have non-positive specificity, and 2 of those have CIs that exclude zero. The intervals are unadjusted for the eight comparisons.

Table 3. Specificity index (on-target drop minus off-target mean drop) at K=16 on Llama-3.1-8B-Instruct. Negative entries indicate that off-target damage exceeds on-target damage. Bold entries have unadjusted bootstrap 95% CIs that exclude zero.

Source task	On-target drop	Off-target mean	Specificity	95% CI
`employees_count_total`	0.92	0.50	**+0.42**	[+0.27, +0.54]
`ceo_lastname`	0.83	0.45	**+0.38**	[+0.20, +0.54]
`holder_record_amount`	0.63	0.63	0.00	[−0.24, +0.22]
`incorporation_year`	0.58	0.54	+0.05	[−0.20, +0.27]
`headquarters_city`	0.50	0.52	−0.02	[−0.24, +0.20]
`incorporation_state`	0.29	0.52	**−0.23**	[−0.42, −0.03]
`headquarters_state`	0.21	0.63	**−0.42**	[−0.67, −0.17]
`registrant_name`	0.17	0.33	−0.17	[−0.32, +0.01]

D. Task-Level Fragility Correlates

Figure 5 compares cross-model fragility with baseline accuracy, query-to-answer word distance, and answer length. This is an exploratory analysis of eight tasks, not a held-out prediction experiment.

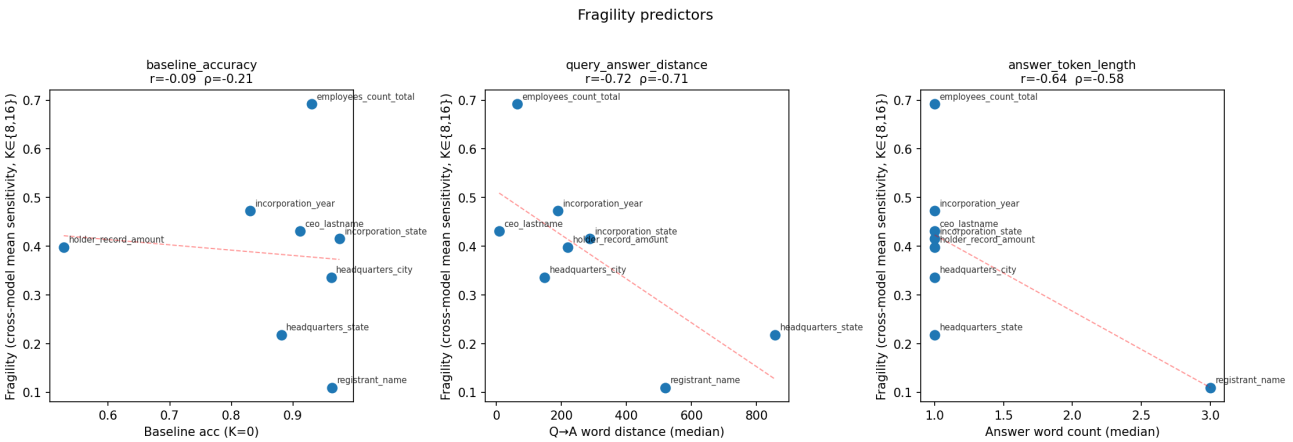

Figure 5. Cross-model fragility (mean target sensitivity across the three models, $K \in \{8, 16\}$) against three task attributes. The estimated Pearson correlations are $r{=}{-}0.09$ for baseline accuracy, $r{=}{-}0.72$ for within-document query-to-answer word distance, and $r{=}{-}0.64$ for answer length. The corresponding unadjusted two-sided p-values are 0.83, 0.04, and 0.09 ($n{=}8$). Panel annotations also report Spearman ρ.

E. Cross-Task Transfer Heatmap and Per-Task Curves

Figure 6 shows the source $\times$ target drop matrix on Llama at $K{=}16$. Figure 7 resolves the same data per target task across the full K sweep.

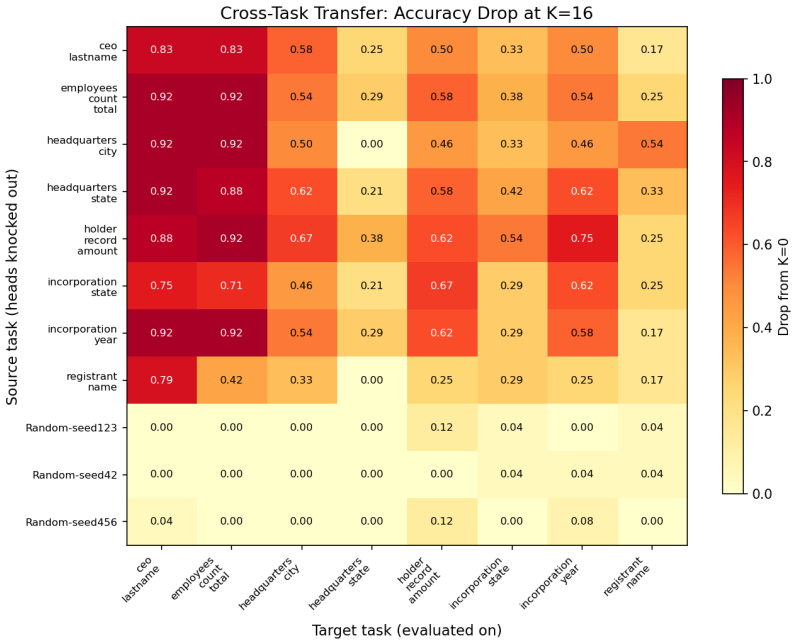

Figure 6. Cross-task transfer drop matrix at $K{=}16$ on Llama. Rows are source tasks (heads ablated); columns are target tasks (evaluation). Random-seed rows at the bottom are the random-ablation controls. Many off-diagonal drops are comparable to their diagonal counterparts.

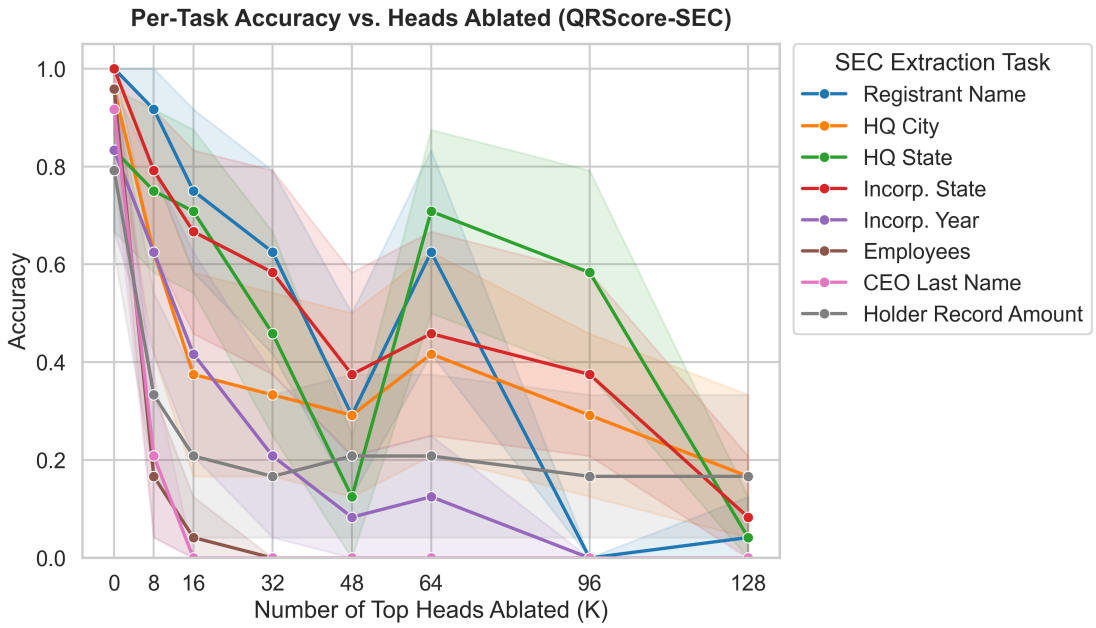

Figure 7. Per-task accuracy curves for combined-SEC QRScore ablation on Llama. Shading denotes the reported bootstrap 95% intervals. `employees_count_total` and `ceo_lastname` decline rapidly, whereas `headquarters_state` and `registrant_name` decline more gradually.

F. Layer-Position Analysis: No Detected Difference Between Task Groups

Table 4 reports per-model permutation and Mann–Whitney tests on the layer-entropy difference between fragile and robust tasks. Neither test rejects equality at conventional thresholds in any model. Figure 8 shows the underlying distributions.

Table 4. Layer-entropy differences between fragile (top-4 by cross-model fragility) and robust (bottom-4) tasks. A permutation test (1000 random group assignments) and a Mann–Whitney U test compare the four-vs-four per-task entropies. Neither test rejects equality at the 0.05 level.

Model	Mean fragile entropy	Mean robust entropy	Permutation p	Mann-Whitney U p
Llama-3.1-8B-Instruct	1.561	1.842	0.293	0.343
Mistral-7B-Instruct-v0.3	1.900	1.882	0.840	0.885
Qwen2.5-7B-Instruct	1.738	1.718	0.942	0.686

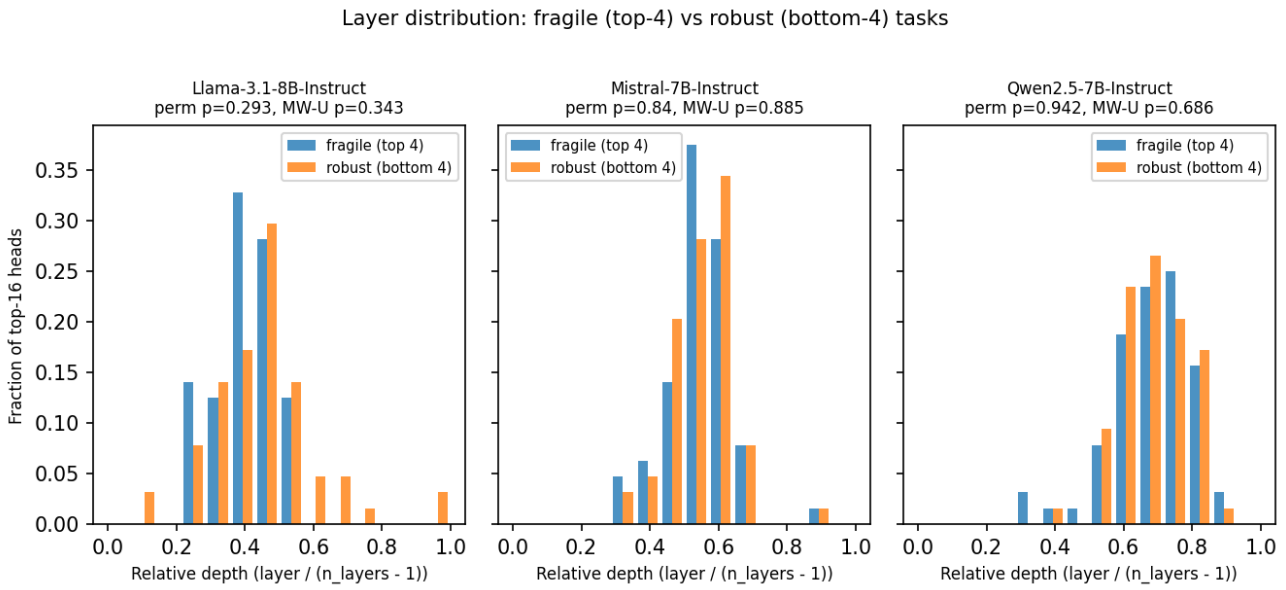

Figure 8. Fragile (blue) and robust (orange) per-task layer distributions on a relative-depth axis. We detect no group difference with the tests in Table 4.

G. LME- and NQ-Selected Head Transfer Is Model-Dependent

Table 5 contrasts the accuracy change at $K{=}16$ on each model's SEC evaluation set for combined SEC, LME-selected, NQ-selected, and random head sets.

Table 5. Accuracy change $\mathrm{Acc}(K) - \mathrm{Acc}(0)$ in percentage points at $K{=}16$ ($n{=}192$ for Llama and Mistral; $n{=}241$ for Qwen). Negative values denote degradation. NQ-selected effects are close to the random baseline, while LME-selected effects vary by model family.

Model	QRScore-SEC	QRScore-LME	QRScore-NQ	Random (3-seed)
Llama-3.1-8B-Instruct	-51.6	-68.2	-4.2	-2.4
Qwen2.5-7B-Instruct	-55.2	-2.5	-0.8	-0.4
Mistral-7B-Instruct-v0.3	-37.0	-14.6	$+1.6$	-0.4

H. Geographic Head Subcluster

Figure 9 shows per-task head-set Jaccard at top-K on Qwen2.5-7B-Instruct. The `headquarters_city` / `headquarters_state` / `registrant_name` subcluster (top-16 Jaccard ≥ 0.45) is visible at small K and recurs in Llama (top-16 HQ-city $\leftrightarrow$ HQ-state $J{=}0.78$) and Mistral. These overlaps provide secondary evidence of within-model head sharing.

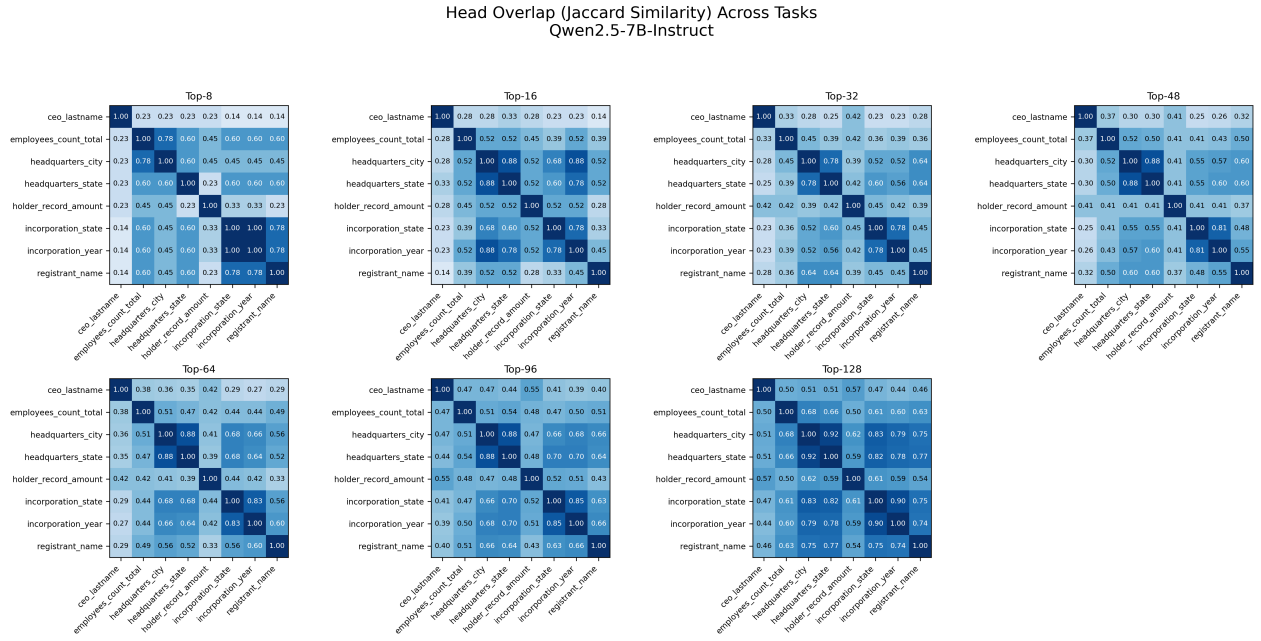

Figure 9. Per-task Jaccard at top-K on Qwen2.5-7B-Instruct.

I. Activation Patching: Per-Task and Per-Model Rescue Values

Table 6 reports the retained per-task mean log-probability rescues for QR-head patching and the three controls. Each cell is patched-minus-ablated teacher-forced gold-answer log-probability, summed over answer tokens and averaged within task. The model summaries in Section 4.4 are medians of the four task means. Positive values indicate rescue.

Table 6. Per-task mean activation-patching rescue in nats. Bold marks the negative Llama `employees_count_total` result discussed in Section 4.4. Values were retained to two decimal places.

Model	Task	patch_qr	patch_bottom	patch_random	patch_same_layer
Llama	ceo_lastname	+0.64	+0.00	−0.19	+0.04
Llama	employees_count_total	**−6.08**	+0.02	+2.16	−0.35
Llama	headquarters_state	+0.35	+0.01	−0.07	+0.08
Llama	registrant_name	+1.02	+0.02	−0.10	+0.13
Qwen	ceo_lastname	+0.74	−0.10	+0.27	+0.19
Qwen	employees_count_total	+1.30	−0.27	−0.05	+0.47
Qwen	headquarters_state	+0.92	−0.20	−0.23	−0.68
Qwen	registrant_name	+2.69	−0.18	−0.05	−2.26
Mistral	ceo_lastname	+1.22	+0.00	+0.37	+0.01
Mistral	employees_count_total	+0.98	+0.00	+0.29	+0.08
Mistral	headquarters_state	+0.95	−0.00	+0.34	+0.32
Mistral	registrant_name	+2.54	−0.36	−1.07	+1.60