Different Heads, Same Fragile Tasks: A Cross-model Retrieval Head Correspondence

Paper

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

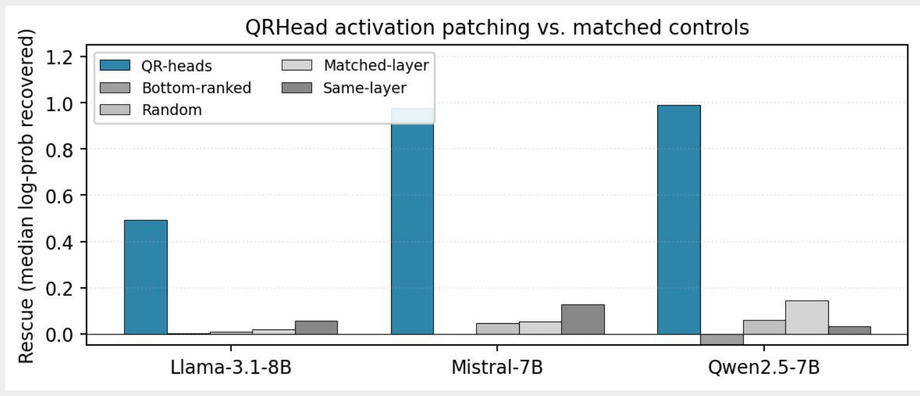
[1] University of California, Los Angeles [2] University of California, Berkeley [3] Georgia Institute Of Technology [4] Stevens Institute Of Technology [5] Brown University

Abstract

Methods for identifying task-associated head groups now exist with high specificity and efficacy, e.g. QRHead.

We characterize how durable task performance is with respect to head group removal, finding certain tasks are fragile across many task-derived QR-head groups.

Are heads and tasks causally linked?



Clean QR-head activations causally recover a substantial-to-near-full fraction of the gold-answer log-probability lost to ablation.

Dataset

We perform recall queries on verified entities from EDGAR-CORPUS [3], a dataset of SEC 10-K filings spanning a broad range of U.S. businesses.

Category	**Tasks (*n* = 192/241)**
Entity/Name	registrant_name, ceo_lastname
Location	headquarters_city, headquarters_state, incorporation_state
Numeric	incorporation_year, employees_count_total, holder_record_amount

Methodology

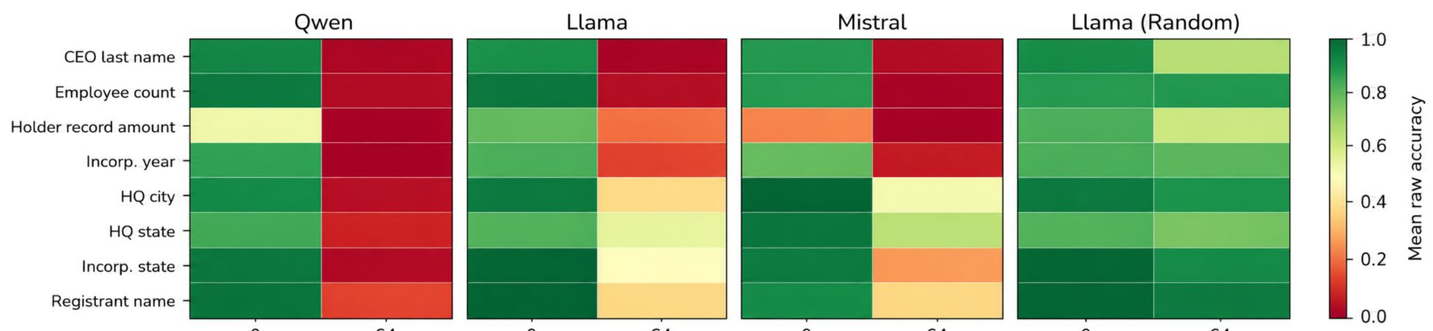
For each source task s, we identify its top-K QR heads, ablate them, and evaluate accuracy on target task t. The resulting accuracy drop is:

$$\Delta_{s \to t}(K) = \mathrm{Acc}_t(0) - \mathrm{Acc}_t(s, K)$$

Target sensitivity, which measures the fragility of task t, averages this drop across all source-task head groups:

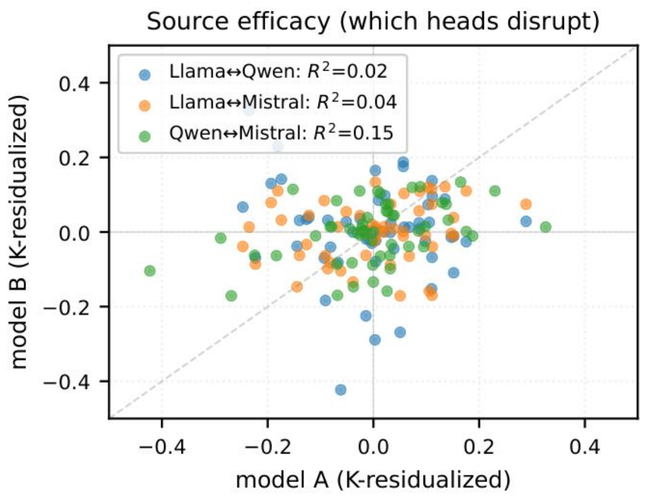
$$\bar{\Delta}_{\cdot \to t}(K) = \frac{1}{|S|} \sum_{s \in S} \Delta_{s \to t}(K)$$

R1: Certain tasks are far more fragile

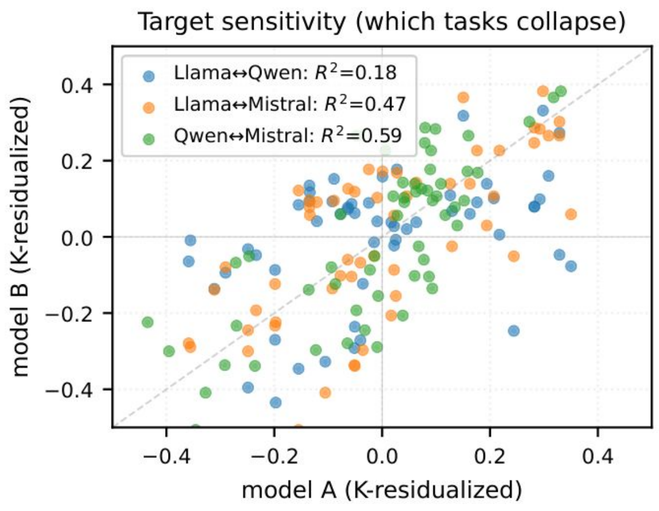


Four recall tasks showed high fragility. Task fragility varies substantially, with several name and numeric tasks collapsing most sharply.

R2: Target fragility is partially conserved across models

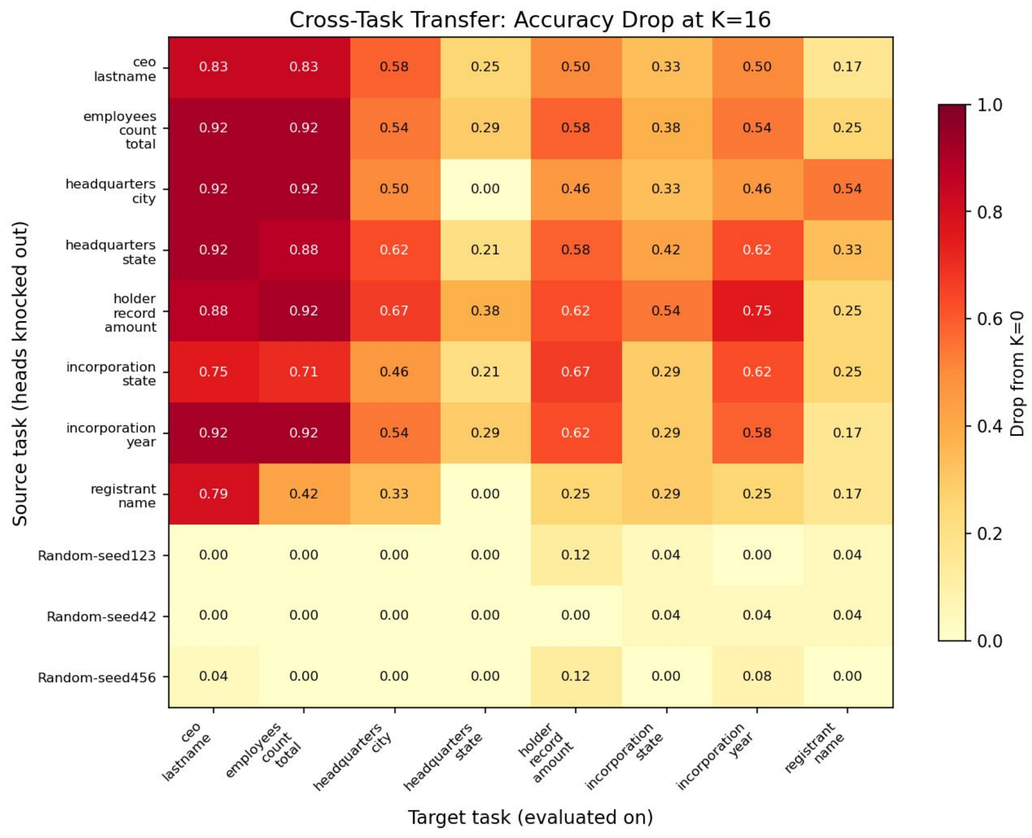


Models do not agree on which source-task head sets are most disruptive.



Models partially agree on which target tasks are fragile.

R3: Head overlap is high across datasets



Future Work

Model Scaling: Test if task fragility and head correspondence hold beyond 7B scale models.

Family Consistency: Track if functional heads emerge at consistent relative depths across a model family's scales.

Task & Domain Diversity: Expand beyond 8 SEC tasks into legal/medical domains to test generalization

References

[1] Zhang, W. et al. Query-Focused Retrieval Heads Improve Long-Context Reasoning and Re-ranking. EMNLP 2025.
[2] Wu, W. et al. Retrieval Head Mechanistically Explains Long-Context Factuality. ICLR 2025.
[3] Loukas, L. et al. EDGAR-CORPUS: Billions of Tokens Make the World Go Round. ECONLP 2021.
[4] Wu, D. et al. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. ICLR 2025.

Acknowledgements

We would like to thank Lambda for providing GPU credits.
We are grateful to Kevin Zhu and the Algoverse AI Research program for providing essential compute resources and logistical support that made this work possible.