DARWINIAN COMPETITION AMONG TRANSFORMER ATTENTION HEADS: GEOMETRIC FITNESS AND FUNCTIONAL INHERITANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how Transformer components shape representation geometry requires causal analyses beyond changes in task performance. We introduce the Causal Competition Framework (CCF), which quantifies a unit's geometric fitness through intervention-induced changes in representation distributions, measured here by squared Wasserstein distance. We instantiate CCF as Representation-level Darwinian Head Competition (RDHC), coupling this measurement to an explicit population process with fitness-guided replication, parameter inheritance, and mutation. Recorded genealogies show branching reproduction across generations, and offspring exhibit greater functional similarity to their designated parents than to shuffled parents under mutation, supporting functional inheritance within the constructed process. Across text classification benchmarks, attention heads exhibit heterogeneous geometric influence, while progressive removal of high-fitness heads leads to nonlinear degradation in classification performance. To examine geometric compensation, we ablate one head and amplify the output of another, defining geometric rescue as a reduction in distributional distance to the intact model. We observe heterogeneous rescue effects across damaged–rescuer pairs, with some rescuer heads providing positive geometric rescue for multiple damaged heads and some damaged heads exhibiting limited rescue across candidate rescuers. Domain-wise analyses further reveal variation in rescue effects across semantic domains. These improvements concern representation distributions. Together, these results characterize heterogeneous head influence and geometric rescue, alongside evidence of replication and functional inheritance in explicitly constructed head populations.

1 INTRODUCTION

Transformer-based language models underpin a wide range of NLP tasks, from text classification to large-scale generation (Vaswani et al., 2017), with advances in scaling and pretraining improving generalization (Schaeffer et al., 2026; Subramanyam et al., 2026). Understanding how their components causally shape representation geometry remains a central challenge. Mechanistic interpretability seeks to identify functional circuits (Nanda et al., 2023), including attention heads associated with algorithmic or linguistic behaviors (Kawata et al., 2025; Zhong & Andreas, 2024) and sparse circuits supporting model predictions (Wang et al., 2023). Ablations, activation patching, and causal abstraction enable intervention-based analyses of component influence (Meng et al., 2022; Geiger et al., 2021; Conmy et al., 2023), while pruning and sparsification reveal substantial redundancy alongside heterogeneous head importance (Michel et al., 2019; Voita et al., 2019; Frantar & Alistarh, 2023). A remaining problem is to characterize how this influence is distributed across representation geometry and how compensation depends on the damaged–rescuer pair and input domain. An evolutionary account introduces a further question: whether head functions can be transmitted across generations through explicit replication and mutation. Addressing these questions calls for a framework that connects causal geometric measurements with a specified population process and direct tests of functional inheritance.

We introduce the Causal Competition Framework (CCF), which measures a unit's geometric fitness through the squared Wasserstein discrepancy between intact and intervened representation

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.

distributions. By suppressing a head's module-output contribution and recomputing downstream representations with other parameters fixed, CCF quantifies its context-dependent geometric influence. Joint interventions characterize geometric non-additivity, while domain-conditional fitness and entropy-based measures describe specialization across semantic categories. A directed rescue analysis further tests whether amplifying a surviving head reduces the distributional damage caused by another head's removal. The accompanying theory distinguishes general discrepancy properties from bounds requiring additional geometric assumptions, providing a precise basis for interpreting these measurements. We instantiate CCF as Representation-level Darwinian Head Competition (RDHC), an explicit population process over independently stored head programs occupying compatible slots. Geometric fitness guides parent sampling for a fixed number of offspring slots; offspring inherit copied parental parameters followed by mutation and optional learning. Archived identities and parentage track reproduction across generations and enable functional comparisons between parents and offspring in a common reference context. Figure 1 shows branching reproduction and lower offspring functional distances to designated than to shuffled parents under both RDHC and neutral replication, whereas independent initialization shows little separation between the pairing conditions. These observations provide evidence of multiplication and functional heredity within the RDHC population.

Our contributions combine an intervention-defined measure of geometric influence, an explicit population process with evidence of multiplication and functional heredity, and an empirical characterization of attention-head organization. Across DBpedia, AG News, and Yahoo! Answers Topics, geometric fitness reveals uneven head influence and domain-sensitive contributions, while progressive removal of high-fitness heads produces sharply nonlinear accuracy degradation. Rescue interventions reveal a non-uniform directed topology: compensation varies across damaged–rescuer pairs, structured residuals remain after controlling for marginal head effects, and representative pairs exhibit domain-dependent rescue. Together, these results connect causal analysis of representation geometry with the study of functional transmission and targeted geometric compensation in Transformer head populations.

2 Related Work

2.1 Interpretability of Attention in Language Models

Attention interpretability has progressed from visualization toward causal and mechanistic analysis. Studies identify structured token-level information flow and vocabulary-aligned attention projections (Katz & Belinkov, 2023; Zimerman et al., 2025); intervention frameworks emphasize functional over correlational assessment of head importance (Nam et al., 2025; Basile et al., 2025; Achtibat et al., 2024), complemented by principled relevance propagation and controlled mechanistic benchmarks (Mueller et al., 2025; Yin & Steinhardt, 2025). Empirical and theoretical evidence indicates that only some heads implement induction or function-vector transformations in in-context learning and reasoning, while others serve structural or redundant roles (Saponati et al., 2025; Jin et al., 2025; Ortu et al., 2024). Architecture and objectives shape attention structure and interpretability (Zheng et al., 2025; Gopalani et al., 2024). Value concentration, logit competition, and pathway interactions suggest that attention works jointly with MLPs and residual streams, motivating circuit-level rather than head-level analysis (Ahmad et al., 2025; Calderon & Reichart, 2025). LLM explainability taxonomies and open challenges (Sharkey et al., 2025; Lai-Dang et al., 2025), attention variants and expressivity (He et al., 2025; Ustaomeroglu & Qu, 2025; Wen et al., 2025), and automated circuit discovery for scalable mechanistic tracing (Conmy et al., 2023) complement this shift toward causally grounded interpretability.

2.2 Mechanistic Interpretability and Causal Interventions

Mechanistic interpretability (MI) identifies Transformer subcircuits whose activations causally drive behavior. Early work identified induction and copying heads (D'Angelo et al., 2025); subsequent analyses localized indirect object identification and syntactic binding in GPT-style models, validating necessity and sufficiency through ablation and activation patching (Wang et al., 2023). Automated discovery searches for minimal causal subgraphs through counterfactual interventions (Conmy et al., 2023). Related approaches include causal scrubbing to test algorithmic hypotheses with distribution-preserving counterfactuals (Geiger et al., 2025), causal abstraction linking high-level causal models

to neural mechanisms (Geiger et al., 2021), and causal tracing of factual recall through layerwise activation restoration (Meng et al., 2022). ROME and extensions selectively edit factual associations through localized parameter updates, supporting identifiable memory circuits (Fang et al., 2025; Meng et al., 2023). In reasoning and in-context learning, induction heads approximate gradient-like forward-pass updates through sequence matching (D'Angelo et al., 2025; Akyürek et al., 2023); structured ablations and path patching validate layerwise attention and MLP subroutines for multistep arithmetic and composition (Nikankin et al., 2025; Wu et al., 2023). Causal mediation measures attention-mediated information flow in language and reasoning (Nam et al., 2025). Sparse autoencoders reveal approximately monosemantic residual-stream features (Cunningham et al., 2024; Pach et al., 2025), complementing pruning and masking evidence of heavy-tailed head importance across tasks and scales (Michel et al., 2019; Voita et al., 2019). Path attribution and attention flow estimate global influence but require interventions for causal validation (Abnar & Zuidema, 2020). Scaling studies report partial invariance of induction and factual recall circuits and task-dependent redistribution of causal influence (Sun, 2025; Wang et al., 2025b); benchmarks emphasize intervention-based metrics to distinguish causal mechanisms from spurious attribution (Mueller et al., 2025). Despite these advances, systematic intervention-based characterization of the global distribution of causal influence across heads and residual pathways remains underexplored.

2.3 Functional Specialization and Redundancy in Transformers

Transformers combine functional specialization with substantial redundancy. Individual heads and MLP neurons encode interpretable syntactic dependencies, positional templates, and long-range semantic relations, while many components contribute little to task performance (Wang et al., 2023; Kim et al., 2025; Nanda et al., 2023; Meng et al., 2022). Causal tracing and activation patching locate algorithmic circuit motifs in small subnetworks (Wang et al., 2023; Marks et al., 2025; Conmy et al., 2023); probes distinguish heads consistently capturing linguistic regularities across layers and seeds from those with diffuse or redundant representations (Kim et al., 2025; Zhao et al., 2025; Geva et al., 2021). Pruning and sparsification can remove large fractions of heads, MLP blocks, or entire layers with negligible performance loss, revealing overparameterization and heterogeneous contributions (Michel et al., 2019; Behnke & Heafield, 2020; Fan et al., 2020; Frantar & Alistarh, 2023). Redundancy is often layer- or module-specific (Michel et al., 2019; Lagunas et al., 2021), while lottery-ticket analyses identify sparse subnetworks that preserve performance and sometimes improve interpretability (Chen et al., 2020; Paul et al., 2023). Dynamic routing in sparse and mixture-of-experts Transformers induces conditional computation and emergent expert specialization (Fedus et al., 2022; Zhou et al., 2022; Lewis et al., 2021); optimization can break symmetry and spontaneously differentiate the roles of initially identical heads (Conmy et al., 2023; Wang et al., 2025a). Geometrically, redundancy appears as subspace overlap, feature collapse, and low-rank head structure (Nanda et al., 2023; Gurnee et al., 2023). However, existing work largely describes components or pruning robustness without explicitly modeling the causal competition among heads governing specialization's emergence, stability, and persistence. A unified causal account of how optimization allocates functional roles among competing components remains underdeveloped.

3 Methodology

The Causal Competition Framework (CCF) measures functional contributions through intervention-defined changes in representation geometry. We define Representation-level Darwinian Head Competition (RDHC) by coupling this measurement to explicit replication, inheritance, and variation of head programs. Static measurements characterize functional organization; evolutionary claims concern the specified population dynamics.

3.1 Evolutionary Units and Representations

Let $\mathcal{H} = \{1, \ldots, N\}$, $N \geq 2$, index compatible head slots, and let $\mathcal{U}_t = \{u_{h,t} : h \in \mathcal{H}\}$ be their occupants at generation t.

Definition 3.1 (Evolutionary Individuals)**.** A slot is an architectural position; an evolutionary individual is an independently stored head-program instance with a distinct identity, parameter state $\theta_h(t) \in \mathbb{R}^P$, and recorded parentage. The inherited state includes query, key, and value projections, the corresponding output-projection block, and head-specific biases. Its phenotype is its functional

response under a common network context and probe distribution. Compatible slots have matching parameter structures and computational interfaces.

For a concrete transformer instantiation, the population occupies one layer, and all parameters outside these heads remain fixed. At the selected attention-module output,

$$z_t(x) = z_0(x) + \sum_{h=1}^{N} \Delta_h(x; \theta_h(t)), \qquad \Phi_t(x) = \Psi(z_t(x)) \in \mathbb{R}^d, \tag{1}$$

where z_0 contains fixed contributions and Ψ is the fixed downstream map. Suppressing heads in $\mathcal{S} \subseteq \mathcal{H}$ gives $\Phi_t^{(-\mathcal{S})}$ by removing their module-output contributions and recomputing the forward pass. Additivity is assumed only at the module output; Ψ may be nonlinear.

3.2 Causal Representation Fitness

Let $\mathcal{X}_{\text{sel}}$ contain M selection examples. The empirical measures before and after intervention are $\mu_t = \frac{1}{M}\sum_{x\in\mathcal{X}_{\text{sel}}} \delta_{\Phi_t(x)}, \qquad \mu_t^{(-\mathcal{S})} = \frac{1}{M}\sum_{x\in\mathcal{X}_{\text{sel}}} \delta_{\Phi_t^{(-\mathcal{S})}(x)}$. (2) Here δ_z is a Dirac measure. CCF permits a nonnegative discrepancy D between representation measures; we use $D = W_2^2$, with squared Euclidean transport cost. The geometric fitness criterion is

$$F_h(t) = D\Big(\mu_t, \mu_t^{(-h)}\Big), \tag{3}$$

evaluated before reproduction with all remaining parameters fixed. This measures context-dependent representational influence, not an intrinsic task benefit: a large removal effect need not make replication beneficial. Representation coordinates and preprocessing remain fixed across generations; rescaling Φ by c rescales F_h by c^2. The theoretical criterion uses exact W_2^2; approximate implementations additionally specify the OT estimator and numerical parameters.

Joint interventions define $J_{i,j}(t) = D\Big(\mu_t, \mu_t^{(-\{i,j\})}\Big) - F_i(t) - F_j(t)$. (4) This describes geometric non-additivity. Squared costs can produce $J_{i,j} \neq 0$ even for additive head outputs, so this quantity alone does not establish functional cooperation or redundancy.

3.3 Darwinian Head Competition

Each of N offspring independently samples a parent with replacement:

$$q_h(t) = \frac{\exp\{\beta F_h(t)\}}{\sum_{i=1}^{N}\exp\{\beta F_i(t)\}}, \qquad A_t(j) \overset{\text{i.i.d.}}{\sim} \text{Categorical}(\mathbf{q}(t)), \tag{5}$$

where $\beta \geq 0$. The realized offspring count is $K_h(t) = \sum_{j=1}^{N} \mathbf{1}\{A_t(j) = h\}$. Conditional on the pre-reproduction network state and environment, denoted Ξ_t, $\mathbb{E}[K_h(t) \mid \Xi_t] = N q_h(t)$, $\sum_{h=1}^{N} K_h(t) = N$. (6) Thus, competition allocates a fixed number of offspring slots, with larger scores giving higher expected reproductive success when $\beta > 0$.

All offspring copy a frozen parental snapshot into separate storage before replacing the population. Mutation acts on the copied state:

$$\widetilde{\theta}_j(t+1) = \theta_{A_t(j)}(t) + \sigma\epsilon_j(t), \qquad \epsilon_j(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_P), \tag{7}$$

where $\sigma \geq 0$ is the per-coordinate standard deviation; noise affects all copied coordinates without projection. Optional learning updates the offspring population jointly: $\Theta_{t+1} = \mathcal{G}_L(\widetilde{\Theta}_{t+1}), \qquad L \geq 0$, (8) where $\mathcal{G}_L$ performs L gradient steps on training data, updates only offspring parameters, and resets optimizer states at birth; $\mathcal{G}_0$ is the identity. Offspring can reproduce subsequently, and lineage records store $(A_t(j), t) \to (j, t+1)$. Copying provides multiplication and parameter inheritance, while mutation introduces variation. Fixed population size does not prevent lineage expansion.

The measurement system is $\mathcal{C}_t = (\mathcal{U}_t, \Phi_t, \mathcal{I}, D)$, where $\mathcal{I}$ specifies interventions. Its Darwinian extension is $\mathcal{C}_D = \big(\{\mathcal{C}_t\}_{t\geq 0}, \mathcal{R}, \mathcal{M}, \mathcal{G}_L, \mathcal{L}\big)$, (9) where $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{L}$ denote replication, mutation, and genealogy, respectively.

Fitness–offspring associations follow by construction. Independent validation compares fitness-guided replication with uniform replication ($\beta = 0$), reinitialized-offspring controls, and learning without replication under matched budgets. Functional inheritance compares parent–offspring resemblance against unrelated controls by replaying archived parameters in the same reference network and slot. Evaluation data are excluded from training, selection, and calibration. Fitness concentration, heritable specialization, and recovery are empirical hypotheses; neither heavy tails nor monotonic degradation under successive ablations follows from these definitions.

3.4 Domain-Conditional Competition

Partition $\mathcal{X}_{\mathrm{sel}}$ into $K \geq 2$ nonempty domains. Let $\mu_{t,k}$ and $\mu_{t,k}^{(-h)}$ be the empirical measures obtained by restricting Equation 2 to domain k. For environment weights $\omega_k \geq 0$, $\sum_k \omega_k = 1$, define $F_{h,k}(t) = D\Big(\mu_{t,k}, \mu_{t,k}^{(-h)}\Big)$, $\quad s_h(t;\omega) = \sum_{k=1}^{K} \omega_k F_{h,k}(t)$. (10) Environment-dependent replication replaces $F_h(t)$ in Equation 5 with $s_h(t;\omega)$, giving $\log \frac{q_h(t;\omega)}{q_i(t;\omega)} = \beta \sum_{k=1}^{K} \omega_k \big(F_{h,k}(t) - F_{i,k}(t)\big)$. (11) The weights need not match domain frequencies, and the weighted conditional score generally differs from OT on the pooled distribution.

For heads with positive total domain fitness, specialization is $\pi_h(k;t) = \frac{F_{h,k}(t)}{\sum_{\ell=1}^{K} F_{h,\ell}(t)}$, $\quad S_h(t) = 1 + \frac{\sum_{k=1}^{K} \pi_h(k;t) \log \pi_h(k;t)}{\log K}$. (12) With $0 \log 0 = 0$, $S_h(t) \in [0,1]$ measures concentration across the chosen domains; it is undefined at zero total fitness. Domain heatmaps describe conditional influence. Darwinian specialization additionally requires lineage-frequency changes and parent–offspring transmission of domain profiles under controlled changes in ω.

3.5 Rescue Topology and Prospective Recovery

Fix a population state and suppress t. For damaged head d and rescuer $r \neq d$, we specify an output-amplification intervention:

$$\begin{aligned} z^{(-d\to r;\lambda)}(x) &= z_0(x) + \sum_{h \neq d} \Delta_h(x) + \lambda \Delta_r(x), \\ \Phi^{(-d\to r;\lambda)}(x) &= \Psi\Big(z^{(-d\to r;\lambda)}(x)\Big). \end{aligned} \tag{13}$$

Thus, the surviving rescuer is scaled by $1 + \lambda$. For each pair, choose λ from a prespecified finite grid $\Lambda \subseteq [0, \infty)$ containing zero to minimize discrepancy from the intact model on calibration data. Freeze this choice before evaluation; all pairs use the same grid and tuning budget.

Below, μ denotes the intact-model measure on held-out data, and intervened measures use the same inputs and the calibrated rescue parameter. The directed rescue effect is

$$R_{d,r} = D(\mu, \mu^{(-d)}) - D(\mu, \mu^{(-d\to r)}). \tag{14}$$

Positive values indicate reduced geometric damage; negative values indicate additional damage. To isolate pair-specific structure, fit an additive model over the observed valid pairs Ω:

$$\begin{aligned} (\hat{c}, \hat{a}, \hat{b}) &\in \arg\min_{c,a,b} \sum_{(d,r)\in\Omega} \left(R_{d,r} - c - a_d - b_r\right)^2, \\ \widetilde{R}_{d,r} &= R_{d,r} - \hat{c} - \hat{a}_d - \hat{b}_r. \end{aligned} \tag{15}$$

Fitted residuals remain well defined when Ω excludes diagonal pairs; ordinary double centering applies only to complete rectangular, equally weighted panels. Domain-specific rescue and its deviation are

$$\begin{aligned} R_{d,r,k} &= D(\mu_k, \mu_k^{(-d)}) - D(\mu_k, \mu_k^{(-d\to r)}), \\ G_{d,r,k} &= R_{d,r,k} - \frac{1}{K} \sum_{\ell=1}^{K} R_{d,r,\ell}. \end{aligned} \tag{16}$$

Here μ_k uses a nonempty held-out sample from domain k. $G_{d,r,k} > 0$ means stronger rescue than the pair's domain average; positive rescue itself requires $R_{d,r,k} > 0$.

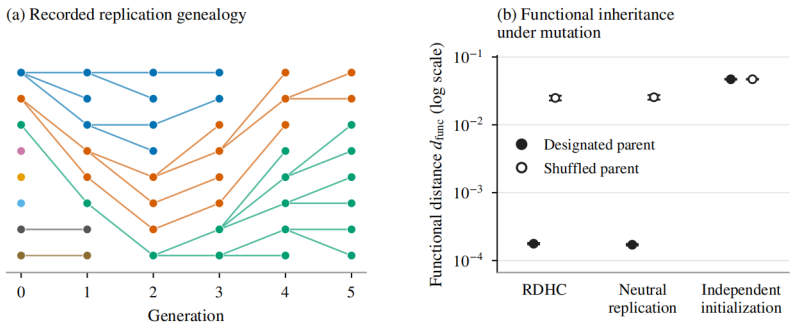

Figure 1: Multiplication and functional heredity in RDHC.

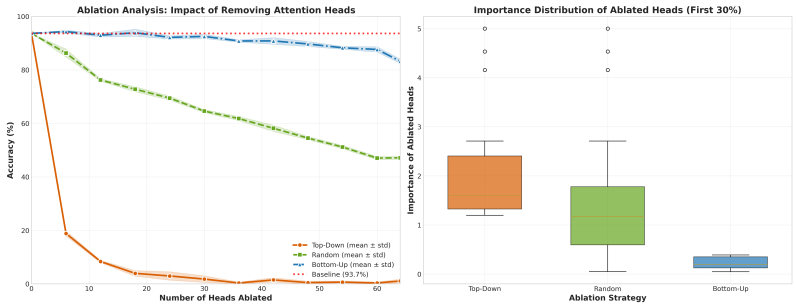

Figure 2: Causal ablation analysis of attention heads under RDHC.

4 Experiments

We empirically evaluate the Causal Competition Framework (CCF) and its Transformer instantiation, Representation-level Darwinian Head Competition (RDHC), on the DBpedia text classification dataset. Additional results on AG_News and Yahoo! Answers Topics are reported in the Appendix.

4.1 Multiplication and Functional Heredity in RDHC

Figure 1 presents the recorded replication genealogy and functional inheritance under mutation. The left panel traces parent–offspring connections from generation zero to generation five. Several parent nodes produce multiple offspring, and their descendants continue to reproduce in subsequent generations. Some lineages terminate, whereas others persist through the final displayed generation, providing a direct record of branching reproduction across generations. The right panel compares offspring functional distances to designated and shuffled parents on a logarithmic scale. Under both RDHC and neutral replication, the filled points lie well below the corresponding open points, indicating that offspring are functionally closer to their designated parents than to shuffled parents. For independent initialization, the two distances are nearly equal, showing little separation between the two pairing conditions. These patterns are consistent with repeated replication and functional heredity in RDHC.

4.2 Experiments on DBpedia

4.2.1 Sparse Dominance and Nonlinear Degradation under Geometric Fitness Ranking

In Figure 2, the baseline model achieves 93.7% accuracy. We perform Top-Down, Random, and Bottom-Up ablations based on geometric fitness ranking to assess the functional contribution of attention heads. Under the Top-Down strategy, performance collapses nonlinearly: removing only six heads reduces accuracy to 19%, and ablating 18–24 heads drives performance to near-zero levels. This abrupt early-phase degradation indicates sparse dominance, where a small subset of high-fitness heads carries a disproportionate share of functional capacity. In contrast, Random removal leads

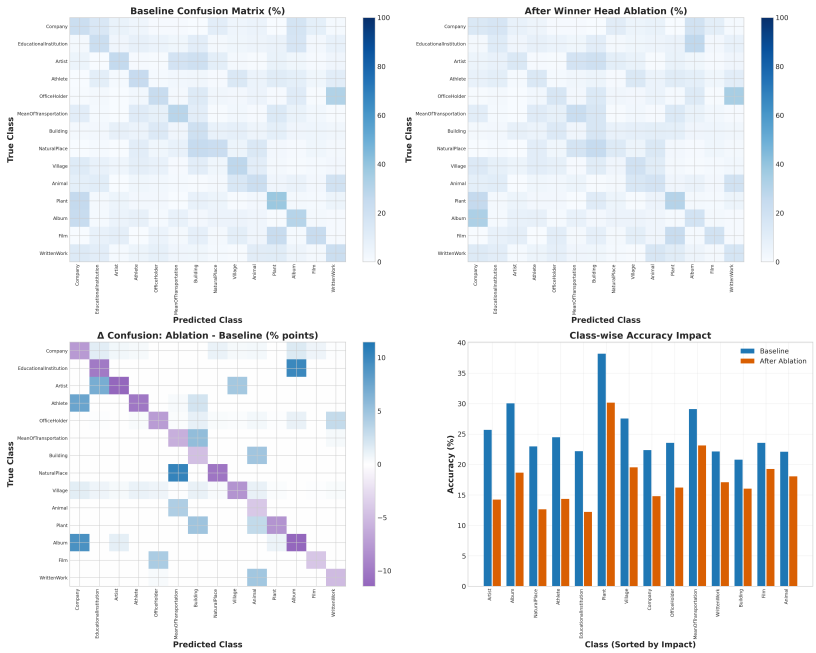

Figure 3: Confusion analysis under winner-head ablation.

to a gradual decline. Even after ablating 60 heads, the model retains 47% accuracy, suggesting substantial redundancy across the attention head population. Bottom-Up removal demonstrates even stronger robustness: eliminating up to 50 low-fitness heads preserves 88% accuracy, and performance still exceeds 83% after ablating 75 heads. Variance across runs remains minimal, confirming the stability of this behavior. Figure 2 (right) further reveals a strongly skewed importance distribution within the first 30% of removed heads. The Top-Down strategy systematically targets heads from a high-magnitude, heavy-tailed regime, whereas Bottom-Up removal primarily eliminates near-zero heads. Together, these observations provide quantitative evidence supporting the heavy-tailed fitness hypothesis underlying geometric ranking.

4.2.2 Domain-Conditional Specialization

In Figure 3, we analyze the post-ablation confusion structure after removing the highest-fitness head. Removing this head leads to heterogeneous performance degradation across classes. Artist, Album, and NaturalPlace show the largest declines, whereas other categories experience more moderate drops. Given that a single head accounts for only a small fraction of the total model parameters, this asymmetric degradation indicates that its functional contribution is not uniformly distributed across domains. Comparing the baseline and post-ablation confusion matrices reveals a global reduction in diagonal mass, accompanied by non-uniform increases in off-diagonal entries. Several semantically related categories exhibit elevated cross-misclassification rates, suggesting that the removed head plays a role in preserving separability among these groups. Overall, these results support the hypothesis that certain attention heads encode domain-sensitive representational features rather than contributing diffusely across all classes.

4.2.3 Attention Head Specialization Across Entity Classes

Figure 4 (top) presents the head–class specialization matrix sorted by global geometric fitness. Only a sparse subset of attention heads exhibits strong class-selective influence, while the majority remain weak and broadly distributed across categories. This pattern suggests that entity-related knowledge is not uniformly encoded but instead concentrated in a limited number of specialized components. Aggregating heads by their dominant semantic category (Figure 4, bottom) further highlights pronounced heterogeneity. Cultural and creative categories—including Artist, Album, and WrittenWork—display relatively high median fitness with low dispersion, indicating consistent

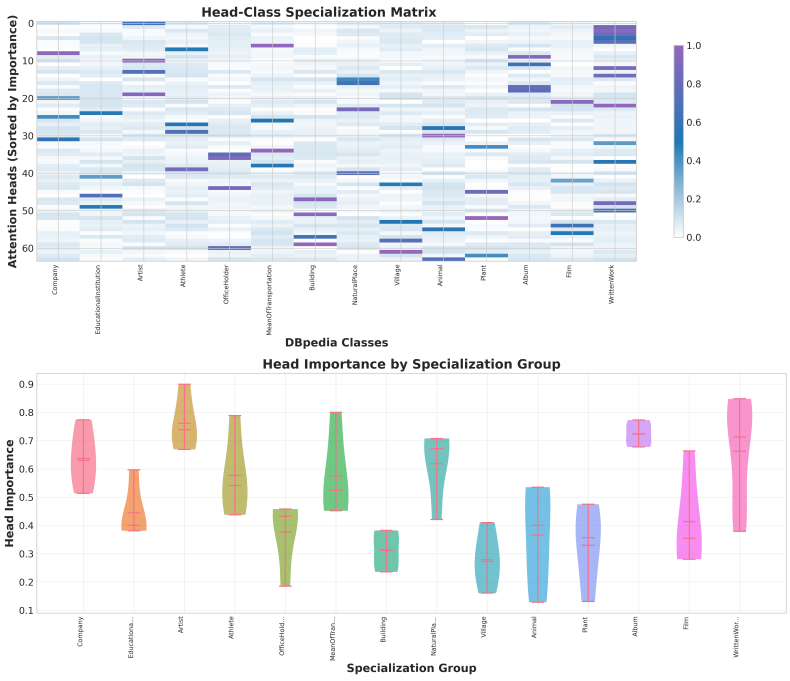

Figure 4: Head-class specialization matrix and importance distributions on DBpedia.

and stable dominance of a small number of heads. In contrast, categories such as Village, Building, and Plant show lower median fitness values and substantially higher variance, suggesting that their representations are more diffuse and distributed across multiple heads. Consistent with this interpretation, selective Top-k head removal produces abrupt nonlinear performance degradation, whereas ablating low-fitness heads results in negligible impact. Beyond aggregate accuracy, removing dominant heads also induces structured shifts in token-level prediction probabilities, particularly within the semantic domains aligned with the heads' specialization. This behavior provides further evidence that certain attention heads function as targeted carriers of entity-class-specific knowledge rather than contributing uniformly to model predictions.

4.2.4 Pair-Specific Rescue Topology and Domain-Dependent Compensation

Figure 5 reveals a strongly non-uniform rescue topology. In the tuned rescue matrix, positive repair effects are concentrated in sparse bands and blocks rather than distributed uniformly across all damaged–rescuer combinations. Some rescuer heads compensate for multiple damaged heads, indicating broad positive rescue capacity, whereas other rescuer heads fail to restore the damaged representation. Conversely, some damaged heads remain difficult to compensate for even when paired with high-capacity rescuers. This pattern is inconsistent with a homogeneous redundancy account in which all heads are approximately interchangeable. More importantly, the row–column controlled residual matrix preserves clear structured patterns after removing marginal damaged-head and rescuer-head effects. If rescue were determined only by the average difficulty of the damaged head and the average capacity of the rescuer head, the residual matrix would be approximately diffuse and centered near zero. Instead, coherent positive and negative residual regions remain, indicating that compensation depends on specific head-pair compatibility. RDHC therefore does not merely rank heads by their individual fitness; it also exposes a directed rescue topology over damaged–rescuer pairs. The domain-wise analysis further shows that this pair-specific rescue structure is semantically conditioned. The best global pair achieves strong raw rescue across multiple domains, whereas the vehicle-specific pair exhibits its strongest positive deviation in the vehicle domain after pair mean removal. In contrast, the failed control pair remains near zero across domains. These results demonstrate that functional substitution among attention heads is not only pair-specific but also domain-dependent. Thus, RDHC characterizes both the global rescue capacity of heads and the finer-grained semantic conditions under which one head can functionally compensate for another.

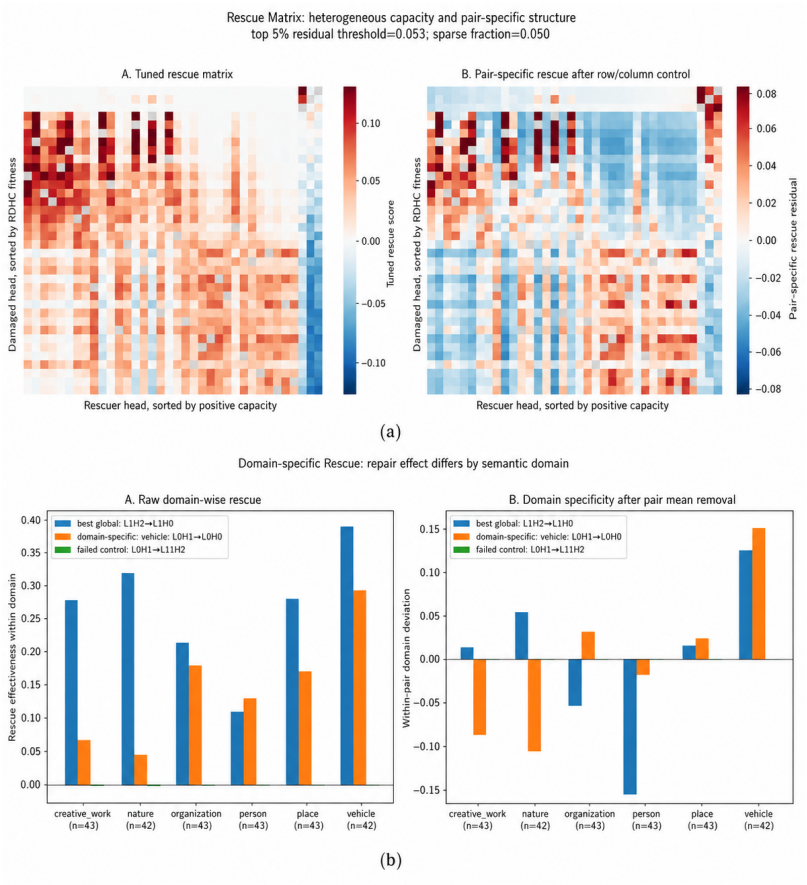

Figure 5: Pair-specific and domain-dependent rescue topology under RDHC.

5 Conclusion

We introduced the Causal Competition Framework (CCF), which characterizes computational units through intervention-induced changes in representation geometry, and instantiated it as Representation-level Darwinian Head Competition (RDHC). RDHC couples geometric fitness to explicit replication, parameter inheritance, and mutation, enabling a direct study of functional transmission across generations. Recorded genealogies reveal branching reproduction and lineage persistence, while offspring are functionally closer to their designated parents than to shuffled parents under mutation. These results establish repeated multiplication and support functional heredity within the RDHC population.

Across DBpedia, AG News, and Yahoo! Answers Topics, our experiments reveal sparse dominance in head influence: removing high-fitness heads causes sharply nonlinear accuracy degradation, whereas removing low-fitness heads preserves much of the model's accuracy. Domain-conditioned analyses further expose differentiated head contributions across semantic categories. Rescue interventions uncover a directed, non-uniform pattern of geometric compensation, with structured residuals after controlling for damaged-head and rescuer-head effects. These patterns show that geometric compensation depends jointly on head pairing and semantic domain.

CCF thus brings geometric influence, domain specialization, and directed rescue into a common intervention-based framework, while RDHC supplies an explicit population process in which multiplication and functional inheritance can be observed. This connection provides a concrete basis for studying how computational functions are distributed across components, transmitted through replication, and compensated through targeted interventions. Future work can extend this framework to generative language models and larger computational circuits, and investigate how functional inheritance shapes specialization and adaptation over longer evolutionary timescales.

AI USE STATEMENT

Disclosure of LLM usage. We used large language models (LLMs) only for polishing the writing; all scientific content is the authors' own.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Areeb Ahmad, Abhinav Joshi, and Ashutosh Modi. Beyond components: Singular vector-based interpretability of transformer circuits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

Maximiliana Behnke and Kenneth Heafield. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2664–2674, 2020.

Nitay Calderon and Roi Reichart. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 656–693, 2025.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33:15834–15846, 2020.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Francesco D'Angelo, Francesco Croce, and Nicolas Flammarion. Selective induction heads: How transformers select causal structures in context. In *The Thirteenth International Conference on Learning Representations*, 2025.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Pulkit Gopalani, Ekdeep S Lubana, and Wei Hu. Abrupt learning in transformers: A case study on matrix completion. *Advances in Neural Information Processing Systems*, 37:55053–55085, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Forty-second International Conference on Machine Learning*, 2025.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Forty-second International Conference on Machine Learning*, 2025.

Shahar Katz and Yonatan Belinkov. Visit: Visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14094–14113, 2023.

Ryotaro Kawata, Yujin Song, Alberto Bietti, Naoki Nishikawa, Taiji Suzuki, Samuel Vaiter, and Denny Wu. From shortcut to induction head: How data diversity shapes algorithm selection in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision–language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10619–10629, 2021.

Quoc-Vinh Lai-Dang, Taemin Kang, and Seungah Son. Adaptive transformer programs: Bridging the gap between performance and interpretability in transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pp. 6265–6274. PMLR, 2021.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*, 2025.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations*, 2025.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8420–8436, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mansheej Paul, Feng Chen, Brett W Larsen, Jonathan Frankle, Surya Ganguli, and Gintare Karolina Dziugaite. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? In *The Eleventh International Conference on Learning Representations*, 2023.

Matteo Saponati, Pascal Sager, Pau Vilimelis Aceituno, Thilo Stadelmann, and Benjamin Grewe. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in transformer training. In *Forty-second International Conference on Machine Learning*, 2025.

Rylan Schaeffer, Noam Levi, Brando Miranda, and Sanmi Koyejo. Pretraining scaling laws for generative evaluations of language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Anirudh Subramanyam, Yuxin Chen, and Robert L Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. In *The Fourteenth International Conference on Learning Representations*, 2026.

Alan Sun. Circuit stability characterizes language model generalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9025–9040, 2025.

Muhammed Ustaomeroglu and Guannan Qu. A theoretical study of (hyper) self-attention through the lens of interactions: Representation, training, generalization. In *Forty-second International Conference on Machine Learning*, 2025.

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, 2019.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Forty-second International Conference on Machine Learning*, 2025b.

Qishuai Wen, Zhiyuan Huang, and Chun-Guang Li. Towards interpretable and efficient attention: Compressing all by contracting a few. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226, 2023.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*, 2025.

Dezhi Zhao, Xin Liu, Xiaocheng Feng, Hui Wang, and Bing Qin. Probing and boosting large language models capabilities via attention heads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28518–28532, 2025.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models. *Patterns*, 6(2), 2025.

Ziqian Zhong and Jacob Andreas. Algorithmic capabilities of random transformers. *Advances in Neural Information Processing Systems*, 37:104357–104382, 2024.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Itamar Zimerman, Ameen Ali, and Lior Wolf. Explaining modern gated-linear RNNs via a unified implicit attention formulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Appendix

A.1 Additional Theoretical Results

This appendix distinguishes general properties of the intervention-defined discrepancy from conditional results that require additional geometric assumptions. Unless stated otherwise, we fix one generation and suppress its time index. All transport results concern exact squared Wasserstein discrepancies with squared Euclidean cost. The dimension d is the dimension of the vector actually supplied to the transport computation, after the specified readout and preprocessing; it need not equal the dimension of an individual attention head.

A.1.1 Optimal Transport Preliminaries

Let

$$\mathcal{P}_2(\mathbb{R}^d) = \left\{ \mu : \mu \text{ is a probability measure on } \mathbb{R}^d,\ \int \|z\|_2^2 \, d\mu(z) < \infty \right\}.$$

All measures below are assumed to belong to this space. Empirical measures with finitely many finite-valued representation vectors satisfy this condition automatically.

Definition A.1 (Squared Wasserstein discrepancy)**.** For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, define

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|z - z'\|_2^2 \, d\gamma(z, z'), \tag{17}$$

where $\Pi(\mu, \nu)$ is the set of probability couplings with marginals μ and ν.

The quantity $W_2 = \sqrt{W_2^2}$ is a metric. We refer to W_2^2 as a *squared distance* or *discrepancy*; it does not in general satisfy the triangle inequality.

Proposition A.2 (Discrete transport and normalization)**.** *Suppose*

$$\mu = \frac{1}{M} \sum_{i=1}^{M} \delta_{z_i}, \qquad \nu = \frac{1}{M} \sum_{j=1}^{M} \delta_{z'_j}, \qquad M \geq 1.$$

Let

$$\mathcal{U}(M) = \{P \in \mathbb{R}_+^{M \times M} : P\mathbf{1} = \mathbf{1},\ P^\top \mathbf{1} = \mathbf{1}\}.$$

Then

$$W_2^2(\mu, \nu) = \frac{1}{M} \min_{P \in \mathcal{U}(M)} \sum_{i,j=1}^{M} P_{ij} \|z_i - z'_j\|_2^2. \tag{18}$$

Equivalently, with

$$\mathcal{C}(M) = \{\Gamma \in \mathbb{R}_+^{M \times M} : \Gamma\mathbf{1} = M^{-1}\mathbf{1},\ \Gamma^\top \mathbf{1} = M^{-1}\mathbf{1}\},$$

we have

$$W_2^2(\mu, \nu) = \min_{\Gamma \in \mathcal{C}(M)} \sum_{i,j=1}^{M} \Gamma_{ij} \|z_i - z'_j\|_2^2. \tag{19}$$

Proof. An indexed empirical coupling assigns mass Γ_{ij} to (z_i, z'_j). Its marginal constraints are precisely those defining $\mathcal{C}(M)$. The substitution $P = M\Gamma$ gives Equation 18. Repeated support points can be treated as separately indexed atoms without changing the optimum. □

Thus, the row- and column-sum-one convention in Equation 18 is valid. Code using a probability coupling Γ must not divide its transport objective by M again.

A.1.2 General Intervention Model

Let $\mathcal{H} = \{1, \dots, N\}$, $N \geq 2$, denote the compatible head slots of Section 3.1. At a fixed generation, the module activation and measured representation are

$$z(x) = z_0(x) + \sum_{h \in \mathcal{H}} \Delta_h(x; \theta_h), \qquad \Phi(x) = \Psi(z(x)) \in \mathbb{R}^d. \tag{20}$$

The module activation belongs to its own activation space; the downstream map Ψ includes all subsequent computation, readout, and fixed preprocessing. It may be nonlinear. Suppressing a set $\mathcal{S} \subseteq \mathcal{H}$ gives

$$\Phi^{(-\mathcal{S})}(x) = \Psi\left(z_0(x) + \sum_{h \in \mathcal{H} \setminus \mathcal{S}} \Delta_h(x; \theta_h)\right). \tag{21}$$

The downstream computation is recomputed after intervention, with all parameters fixed. In general, the displacement in the measured representation space is not the sum of the deleted module contributions.

For an indexed sample $\mathcal{X} = (x_1, \dots, x_M)$, define

$$\mu = \frac{1}{M} \sum_{i=1}^{M} \delta_{\Phi(x_i)}, \qquad \mu^{(-\mathcal{S})} = \frac{1}{M} \sum_{i=1}^{M} \delta_{\Phi^{(-\mathcal{S})}(x_i)}, \qquad F_h = W_2^2(\mu, \mu^{(-h)}). \tag{22}$$

Here $\mu^{(-h)} = \mu^{(-\{h\})}$. For reproductive selection, $\mathcal{X} = \mathcal{X}_{\text{sel}}$, as in Section 3.2. Statements about evaluation data use separately specified empirical measures.

Additive measured-representation special case. Some results below additionally assume that there exist $\phi_0(x) \in \mathbb{R}^d$ and $\Delta_h^{\Phi}(x) \in \mathbb{R}^d$ such that, for the interventions under consideration,

$$\begin{aligned} \Phi(x) &= \phi_0(x) + \sum_{h \in \mathcal{H}} \Delta_h^{\Phi}(x), \\ \Phi^{(-\mathcal{S})}(x) &= \Phi(x) - \Delta_{\mathcal{S}}^{\Phi}(x), \qquad \Delta_{\mathcal{S}}^{\Phi}(x) = \sum_{h \in \mathcal{S}} \Delta_h^{\Phi}(x). \end{aligned} \tag{23}$$

This is an additional assumption in the *measured representation space*. For an affine downstream map $\Psi(z) = Az + b$, it holds with $\Delta_h^{\Phi}(x) = A\Delta_h(x; \theta_h)$. Even in this case, transport costs use the transformed displacements $A\Delta_h$, not the original module displacements.

A.1.3 General Coupling Upper Bound

Proposition A.3 (General coupling upper bound)**.** *For intact and intervened empirical measures constructed from the same inputs,*

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) \leq \frac{1}{M} \sum_{i=1}^{M} \|\Phi(x_i) - \Phi^{(-\mathcal{S})}(x_i)\|_2^2. \tag{24}$$

This statement holds for the nonlinear downstream map in Section 3. Expressing the right-hand side through summed head displacements requires the additional additive-representation assumption in Equation 23.

Proof. The input-matched coupling

$$\gamma_{\mathcal{S}} = \frac{1}{M} \sum_{i=1}^{M} \delta_{(\Phi(x_i),\, \Phi^{(-\mathcal{S})}(x_i))}$$

has marginals μ and $\mu^{(-\mathcal{S})}$. Its squared transport cost is the right-hand side of Equation equation 24. The infimum over all feasible couplings cannot exceed the cost of this coupling. No linearity or additive decomposition of the final representation is used. □

Under Equation 23, this specializes to

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) \leq \frac{1}{M} \sum_{i=1}^{M} \left\| \sum_{h \in \mathcal{S}} \Delta_h^{\Phi}(x_i) \right\|_2^2. \tag{25}$$

A.1.4 ALIGNED TRANSPORT OPTIMALITY

Proposition A.4 (Optimality under aligned, order-preserving displacements)**.** *Assume Equation 23. Fix a nonempty intervention set $\mathcal{S}$. Suppose there is a fixed nonzero $v \in \mathbb{R}^d$ such that*

$$\Delta_h^\Phi(x_i) = a_h(x_i)v, \qquad a_h(x_i) \geq 0, \qquad h \in \mathcal{S}, \quad i = 1, \ldots, M.$$

Set $e = v/\|v\|_2$, $a_\mathcal{S} = \sum_{h\in\mathcal{S}} a_h$, and

$$t_i = \langle \Phi(x_i), e\rangle, \qquad s_i = \langle \Phi^{(-\mathcal{S})}(x_i), e\rangle = t_i - \|v\|_2 a_\mathcal{S}(x_i).$$

Assume that the projected input pairing for this joint intervention preserves order:

$$(t_i - t_j)(s_i - s_j) \geq 0 \qquad \textit{for all } i, j. \tag{26}$$

Then the input-matched coupling is globally optimal in the full space, and

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) = \frac{1}{M}\sum_{i=1}^M \|\Delta_\mathcal{S}^\Phi(x_i)\|_2^2. \tag{27}$$

Proof. Let $p_e(z) = \langle z, e\rangle$. Every full-space coupling projects to a coupling of $(p_e)_\#\mu$ and $(p_e)_\#\mu^{(-\mathcal{S})}$, and $\|z - z'\|_2^2 \geq |p_e(z) - p_e(z')|^2$. Therefore,

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) \geq W_2^2((p_e)_\#\mu, (p_e)_\#\mu^{(-\mathcal{S})}). \tag{28}$$

For equal-weight empirical measures on the real line, matching sorted coordinates minimizes squared cost. Indeed, replacing a crossed pairing by an ordered pairing cannot increase cost, since for $t_i \leq t_j$ and $s_i \leq s_j$,

$$(t_i - s_j)^2 + (t_j - s_i)^2 - (t_i - s_i)^2 - (t_j - s_j)^2 = 2(t_j - t_i)(s_j - s_i) \geq 0.$$

A minimum over doubly stochastic matrices is attained at a permutation, so uncrossing proves the claim for all empirical couplings. Condition Equation 26 makes the input-matched projected coupling one such sorted matching; ties can be ordered within equal-coordinate blocks. Hence the right-hand side of Equation 28 equals

$$\frac{1}{M}\sum_{i=1}^M (t_i - s_i)^2 = \frac{1}{M}\sum_{i=1}^M \|v\|_2^2 a_\mathcal{S}(x_i)^2.$$

The full-space input-matched coupling leaves every orthogonal component unchanged, so its cost equals this lower bound. Its feasibility also gives the reverse inequality by Proposition A.3, establishing global optimality. □

This proposition concerns an additive displacement model in the measured space. Its assumptions are not asserted for arbitrary nonlinear Transformer representations or for every RDHC population state.

A.1.5 CONDITIONAL MONOTONE INTERVENTION BOUND

Corollary A.5 (Conditional singleton lower bound)**.** *For a nonempty set $\mathcal{S}$ satisfying the assumptions of Proposition A.4,*

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) = \frac{1}{M}\sum_{i=1}^M \|\Delta_\mathcal{S}^\Phi(x_i)\|_2^2 \geq \max_{h\in\mathcal{S}} F_h. \tag{29}$$

Proof. For each $h \in \mathcal{S}$, alignment and nonnegativity imply

$$\|\Delta_\mathcal{S}^\Phi(x_i)\|_2^2 = \|v\|_2^2 \left(\sum_{\ell\in\mathcal{S}} a_\ell(x_i)\right)^2 \geq \|v\|_2^2 a_h(x_i)^2 = \|\Delta_h^\Phi(x_i)\|_2^2.$$

Consequently,

$$W_2^2(\mu, \mu^{(-\mathcal{S})}) \geq \frac{1}{M}\sum_{i=1}^M \|\Delta_h^\Phi(x_i)\|_2^2 \geq F_h,$$

where the final inequality is the singleton coupling upper bound. Taking the maximum proves the result. Order preservation is required for the specified joint intervention; it need not be separately assumed for each singleton. □

Counterexample without same-direction displacements. Consider a single input, a one-dimensional additive representation, and two head displacements $\Delta_1^\Phi = a$, $\Delta_2^\Phi = -a$, where $a \neq 0$. With zero baseline, the intact representation is zero. Thus,

$$F_1 = F_2 = a^2, \qquad W_2^2(\mu, \mu^{(-\{1,2\})}) = 0. \tag{30}$$

Joint deletion cancels the singleton displacements. Therefore, Equation 29 is a conditional geometric statement, not a universal prediction of CCF or RDHC. It does not imply monotonic degradation of classification accuracy.

A.1.6 Geometric Additivity under Factorwise Interventions

Proposition A.6 (Geometric additivity under factorwise interventions)**.** *Suppose the measured space admits an orthogonal decomposition*

$$\mathbb{R}^d = V_i \oplus V_j \oplus V_{\mathrm{rest}}, \qquad \mu = \mu_i \otimes \mu_j \otimes \mu_{\mathrm{rest}}.$$

Let the single-head interventions be measurable maps

$$T_i(z_i, z_j, z_{\mathrm{rest}}) = (t_i(z_i), z_j, z_{\mathrm{rest}}), \tag{31}$$

$$T_j(z_i, z_j, z_{\mathrm{rest}}) = (z_i, t_j(z_j), z_{\mathrm{rest}}), \tag{32}$$

and let the joint intervention be

$$T_{ij} = T_i \circ T_j = T_j \circ T_i.$$

Assume that the intervened measures are the corresponding pushforwards and have finite second moments. With squared Euclidean cost on this orthogonal product space,

$$W_2^2(\mu, \mu^{(-\{i,j\})}) = F_i + F_j. \tag{33}$$

Proof. Writing $\mu_i' = (t_i)_\# \mu_i$ and $\mu_j' = (t_j)_\# \mu_j$, factor-local dependence gives

$$\mu^{(-i)} = \mu_i' \otimes \mu_j \otimes \mu_{\mathrm{rest}},$$
$$\mu^{(-j)} = \mu_i \otimes \mu_j' \otimes \mu_{\mathrm{rest}},$$
$$\mu^{(-\{i,j\})} = \mu_i' \otimes \mu_j' \otimes \mu_{\mathrm{rest}}.$$

For any coupling of two such product measures, its projection onto each factor is a feasible factor coupling. Because the squared cost adds over orthogonal factors, its total cost is at least the sum of the optimal factor costs. Conversely, the product of optimal factor couplings is feasible and attains that sum. The unchanged factor can use the identity coupling at zero cost. Thus,

$$W_2^2(\mu, \mu^{(-\{i,j\})}) = W_2^2(\mu_i, \mu_i') + W_2^2(\mu_j, \mu_j') = F_i + F_j.$$

□

The factorization assumptions apply to the measures being compared. Independence of population coordinates alone does not guarantee exact factorization of a finite empirical sample.

Why changing only one coordinate is insufficient. Let Z_1, Z_2 be independent and uniform on $\{0, 1\}$, and let μ be their joint law. Consider

$$T_i(z_1, z_2) = (z_2, z_2), \qquad T_j(z_1, z_2) = (z_1, z_1).$$

Each map changes only its designated coordinate, but uses the other coordinate to do so. Each single intervention produces $\frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}$, with squared transport cost $1/2$: the original off-diagonal mass is $1/2$, each such point must move by squared distance at least one, and this cost is attainable. If both coordinate updates are evaluated simultaneously on the original input, their joint intervention is

$$T_{ij}(z_1, z_2) = (z_2, z_1).$$

Equivalently, the singleton displacements are $(z_1 - z_2, 0)$ and $(0, z_2 - z_1)$, and joint deletion subtracts both. The joint law is unchanged, giving

$$F_i = F_j = \tfrac{1}{2}, \qquad W_2^2(\mu, (T_{ij})_\# \mu) = 0.$$

This simultaneous joint map is not the composition of factor-local maps required above. The example demonstrates why coordinate support alone cannot establish geometric additivity.

Definition A.7 (Geometric non-additivity)**.** Consistent with Equation (4) in the main text, define

$$J_{i,j} = W_2^2(\mu, \mu^{(-\{i,j\})}) - F_i - F_j. \tag{34}$$

Proposition A.6 gives $J_{i,j} = 0$ under its stated conditions. Positive and negative values describe, respectively, *super-additive geometric deformation* and *sub-additive geometric deformation*. Negative values can arise from cancellation, dependence, or changes in optimal matching; they do not by themselves establish functional redundancy. The symbol $J_{i,j}$ replaces the earlier interaction notation $\Omega_{i,j}$, leaving Ω available for the valid rescue-pair set in Section 3.5.

A.1.7 Conditional Fitness Concentration in a Finite Population

Proposition A.8 (Finite power-law tail bound)**.** *Suppose the N ordered singleton fitness scores satisfy*

$$F_{(k)} = Ck^{-\alpha}, \qquad k = 1, \ldots, N, \qquad C > 0, \quad \alpha > 1,$$

and write $H_{N,\alpha} = \sum_{k=1}^{N} k^{-\alpha}$. For each integer $m \in \{1, \ldots, N\}$,

$$\frac{\sum_{k=m+1}^{N} F_{(k)}}{\sum_{k=1}^{N} F_{(k)}} \leq \frac{m^{1-\alpha} - N^{1-\alpha}}{(\alpha - 1)H_{N,\alpha}}. \tag{35}$$

Proof. Since $x^{-\alpha}$ decreases on $(0, \infty)$,

$$\sum_{k=m+1}^{N} k^{-\alpha} \leq \int_{m}^{N} x^{-\alpha}\, dx = \frac{m^{1-\alpha} - N^{1-\alpha}}{\alpha - 1}.$$

Divide by $H_{N,\alpha}$. The empty tail at $m = N$ is zero. □

For example, if the right-hand side of Equation 35 is at most $1 - \rho$, the first m heads account for at least a fraction ρ of the sum of singleton scores. This is an algebraic property conditional on a given finite score distribution. It neither establishes that RDHC generates a power law nor converts cumulative singleton fitness into joint ablation damage or a decrease in task performance.

A.1.8 Specialization Bounds

Partition the sample into $K \geq 2$ nonempty domains $\mathcal{X}_1, \ldots, \mathcal{X}_K$, with $M_k = |\mathcal{X}_k| > 0$. Define

$$\begin{aligned} \mu_k &= \frac{1}{M_k} \sum_{x \in \mathcal{X}_k} \delta_{\Phi(x)}, \quad \mu_k^{(-h)} = \frac{1}{M_k} \sum_{x \in \mathcal{X}_k} \delta_{\Phi^{(-h)}(x)}, \\ F_{h,k} &= W_2^2(\mu_k, \mu_k^{(-h)}). \end{aligned} \tag{36}$$

For a head with $\sum_{k=1}^{K} F_{h,k} > 0$, set

$$\pi_h(k) = \frac{F_{h,k}}{\sum_{\ell=1}^{K} F_{h,\ell}}, \qquad H(\pi_h) = -\sum_{k=1}^{K} \pi_h(k) \log \pi_h(k), \qquad S_h = 1 - \frac{H(\pi_h)}{\log K}. \tag{37}$$

All logarithms are natural and $0 \log 0 = 0$. Then $S_h \in [0, 1]$. When total domain fitness is zero, π_h and S_h are undefined.

Lemma A.9 (Variance–distance identity)**.** *Let*

$$\bar{F}_h = \frac{1}{K} \sum_{k=1}^{K} F_{h,k} > 0, \qquad \mathrm{Var}_k(F_{h,k}) = \frac{1}{K} \sum_{k=1}^{K} (F_{h,k} - \bar{F}_h)^2, \qquad u(k) = \frac{1}{K}.$$

Then

$$\|\pi_h - u\|_2^2 = \frac{\mathrm{Var}_k(F_{h,k})}{K \bar{F}_h^2}. \tag{38}$$

Proof. Since $\pi_h(k) - K^{-1} = (F_{h,k} - \bar{F}_h)/(K\bar{F}_h)$, squaring and summing over domains gives Equation equation 38. □

Lemma A.10 (A universal Euclidean lower bound on relative entropy)**.** *For every probability vector $p \in \mathbb{R}^K$ and the uniform vector u,*

$$D_{\mathrm{KL}}(p\|u) \geq \frac{1}{2}\|p-u\|_2^2. \tag{39}$$

Proof. First suppose p has strictly positive coordinates. Let $f(p) = \sum_k p_k \log p_k$, $v = p - u$, and $g(s) = f(u + sv)$ for $s \in [0, 1]$. In the interior of the probability simplex,

$$\nabla^2 f(p) = \mathrm{diag}(1/p_1, \ldots, 1/p_K).$$

Every coordinate of $u + sv$ is at most one, so

$$g''(s) = \sum_{k=1}^{K} \frac{v_k^2}{u_k + sv_k} \geq \|v\|_2^2.$$

Also, $g'(0) = (1 + \log(1/K)) \sum_k v_k = 0$. Integrating twice yields

$$D_{\mathrm{KL}}(p\|u) = f(p) - f(u) = \int_0^1 (1-s)g''(s)\, ds \geq \frac{1}{2}\|p-u\|_2^2.$$

For a vector with zero coordinates, apply the result to $(1-\delta)p + \delta u$, $\delta > 0$, and let $\delta \downarrow 0$. Continuity under the convention $0 \log 0 = 0$ gives the boundary case. □

Proposition A.11 (Specialization lower bound)**.** *Under the preceding conditions, if* $\mathrm{Var}_k(F_{h,k}) \geq \epsilon > 0$*, then*

$$S_h = \frac{D_{\mathrm{KL}}(\pi_h\|u)}{\log K} \geq \frac{\mathrm{Var}_k(F_{h,k})}{2K\bar{F}_h^2 \log K} \geq \frac{\epsilon}{2K\bar{F}_h^2 \log K}. \tag{40}$$

Proof. The identity $D_{\mathrm{KL}}(\pi_h\|u) = \log K - H(\pi_h)$, followed by Lemmas A.10 and A.9, gives the result. □

The factor $\bar{F}_h^2$ is necessary for invariance to a common rescaling of all domain fitness scores. A bound obtained by simply dropping K from Equation equation 40 is not valid in general: with $K = 10$ and fitness vector $(1, 0, \ldots, 0)$, $S_h = 1$, whereas that proposed bound would be $9/(2 \log 10) > 1$. The valid result preserves the qualitative implication that nonzero domain-fitness variance implies nonzero specialization. It does not explain how specialization arises or establish its inheritance.

A.1.9 Rescue Bounds and Calibration

Fix a population state and distinct heads $d, r \in \mathcal{H}$, using d here as the damaged-head index. As in Section 3.5, define

$$\begin{aligned} z^{(-d\to r;\lambda)}(x) &= z_0(x) + \sum_{h\neq d} \Delta_h(x;\theta_h) + \lambda\Delta_r(x;\theta_r), \\ \Phi^{(-d\to r;\lambda)}(x) &= \Psi(z^{(-d\to r;\lambda)}(x)). \end{aligned} \tag{41}$$

The surviving rescuer is scaled by $1 + \lambda$. Let D be a finite, nonnegative discrepancy on the representation measures being compared. For a calibrated coefficient $\widehat{\lambda}_{d,r}$, suppress the coefficient in the notation and define, on the same evaluation inputs,

$$R_{d,r} = D(\mu, \mu^{(-d)}) - D(\mu, \mu^{(-d\to r)}). \tag{42}$$

Positive values indicate reduced distributional damage, zero indicates no change in this discrepancy, and negative values indicate additional damage.

Proposition A.12 (Basic rescue bound)**.** *For every valid pair,*

$$R_{d,r} \leq D(\mu, \mu^{(-d)}). \tag{43}$$

If, additionally, $D(\nu, \nu') = 0$ if and only if $\nu = \nu'$, equality holds precisely when $\mu^{(-d\to r)} = \mu$.

Proof. Subtract the nonnegative quantity $D(\mu, \mu^{(-d\to r)})$ from $D(\mu, \mu^{(-d)})$. Equality is equivalent to the subtracted quantity being zero, and the additional separation property gives the claim. □

Exact W_2^2 has this separation property. No triangle inequality is needed. Distributional restoration need not restore individual inputs: permuting intact representation vectors across inputs leaves an empirical measure unchanged while potentially changing their semantic assignments and predictions. Thus, equality in Equation equation 43 certifies restoration only at the representation-distribution level.

Proposition A.13 (Nonnegative calibration rescue)**.** *Let $\Lambda \subseteq [0, \infty)$ be a prespecified finite grid containing zero. On a fixed calibration sample, define the deterministic objective*

$$\mathcal{L}_{d,r}^{\mathrm{cal}}(\lambda) = D(\mu_{\mathrm{cal}}, \mu_{\mathrm{cal}}^{(-d\to r;\lambda)}), \qquad \widehat{\lambda}_{d,r} \in \arg\min_{\lambda\in\Lambda} \mathcal{L}_{d,r}^{\mathrm{cal}}(\lambda).$$

If the damaged-model discrepancy is evaluated with the same objective, then

$$R_{d,r}^{\mathrm{cal}} = \mathcal{L}_{d,r}^{\mathrm{cal}}(0) - \mathcal{L}_{d,r}^{\mathrm{cal}}(\widehat{\lambda}_{d,r}) \geq 0. \tag{44}$$

Proof. At $\lambda = 0$, Equation equation 41 equals the damaged-model intervention. Since zero is a candidate in the same minimization, the selected objective value is no larger than its value at zero. □

This is a calibration identity, not a guarantee on held-out data. After freezing $\widehat{\lambda}_{d,r}$, evaluation rescue can be negative. Such values must be retained; their existence does not contradict Proposition A.13. If the objective is estimated, the calibration statement requires one fixed deterministic evaluation rule, including fixed estimator randomness, rather than independently recomputed noisy objectives.

A.1.10 Row and Column Effects with Missing Pairs

Let $\Omega \subseteq \mathcal{D} \times \mathcal{R}$ be the set of observed valid damaged–rescuer pairs, possibly excluding all self-pairs. The sets $\mathcal{D}$ and $\mathcal{R}$ contain only observed rows and columns. Following Equation (15), define

$$\begin{aligned} (\widehat{c}, \widehat{a}, \widehat{b}) &\in \arg\min_{c,a,b} \sum_{(d,r)\in\Omega} (R_{d,r} - c - a_d - b_r)^2, \\ \widetilde{R}_{d,r} &= R_{d,r} - \widehat{c} - \widehat{a}_d - \widehat{b}_r. \end{aligned} \tag{45}$$

Proposition A.14 (Unique residuals and zero observed marginal sums)**.** *The fitted values and residuals in Equation equation 45 are unique, even if the parameter encoding is not. Moreover,*

$$\sum_{r:(d,r)\in\Omega} \widetilde{R}_{d,r} = 0, \qquad \sum_{d:(d,r)\in\Omega} \widetilde{R}_{d,r} = 0 \tag{46}$$

for every observed row d and column r, respectively.

Proof. Stack the observed scores into $y \in \mathbb{R}^{|\Omega|}$. Let B_Ω have an intercept column and indicator columns for the observed row and column memberships. Least squares fits the unique orthogonal projection $\widehat{y}$ of y onto $\mathrm{col}(B_\Omega)$. Thus both $\widehat{y}$ and $e = y - \widehat{y}$ are unique, including when the design is rank deficient or its observation graph is disconnected. The normal equations $B_\Omega^\top e = 0$ yield the two identities in Equation equation 46. □

Corollary A.15 (Additive observed scores have zero residual)**.** *If $R_{d,r} = c + a_d + b_r$ for every $(d, r) \in \Omega$, then $\widetilde{R}_{d,r} = 0$ on every observed pair.*

Proof. In this case $y \in \mathrm{col}(B_\Omega)$, so its projection is itself. □

Complete equally weighted rectangular panels. Only when $\Omega = \mathcal{D} \times \mathcal{R}$ with equal weights does the preceding definition reduce to ordinary double centering. Define

$$\bar{R}_{d,\cdot} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} R_{d,r}, \qquad \bar{R}_{\cdot,r} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} R_{d,r}, \tag{47}$$

$$\bar{R}_{\cdot,\cdot} = \frac{1}{|\mathcal{D}||\mathcal{R}|} \sum_{d \in \mathcal{D}} \sum_{r \in \mathcal{R}} R_{d,r}. \tag{48}$$

Then

$$\widetilde{R}_{d,r} = R_{d,r} - \bar{R}_{d,\cdot} - \bar{R}_{\cdot,r} + \bar{R}_{\cdot,\cdot}. \tag{49}$$

Indeed, the proposed fitted values are additive and their residuals have zero row and column sums, hence satisfy the least-squares normal equations. Missing self-pairs are not zero-valued observations and must not be filled with zeros to apply this formula.

Weighted fitting. With prespecified positive weights $w_{d,r}$, minimize $\sum_{\Omega} w_{d,r}(R_{d,r} - c - a_d - b_r)^2$. The fitted values are unique under the weighted inner product, and the normal equations give

$$\sum_{r:(d,r)\in\Omega} w_{d,r}\widetilde{R}_{d,r} = 0, \qquad \sum_{d:(d,r)\in\Omega} w_{d,r}\widetilde{R}_{d,r} = 0.$$

The unweighted mean formulas do not apply in general.

A nonzero residual describes a departure from the fitted additive row and column model on the observed panel. Establishing a reliable pair-specific effect requires uncertainty estimates or repeated runs; the projection identities alone do not establish statistical reliability of a finite-sample heatmap.

A.1.11 Domain-Wise Rescue and Mean Deviations

Fix a common prespecified set of valid evaluation domains, relabeled $\{1, \ldots, K\}$, with $K \geq 2$ and nonempty samples $\mathcal{X}_{\mathrm{eval},k}$ of size M_k. For each valid pair, use the same frozen calibration coefficient $\widehat{\lambda}_{d,r}$ in every domain. Define

$$\mu_k = \frac{1}{M_k} \sum_{x \in \mathcal{X}_{\mathrm{eval},k}} \delta_{\Phi(x)}, \qquad \mu_k^{(-d)} = \frac{1}{M_k} \sum_{x \in \mathcal{X}_{\mathrm{eval},k}} \delta_{\Phi^{(-d)}(x)}, \tag{50}$$

$$\mu_k^{(-d\to r)} = \frac{1}{M_k} \sum_{x \in \mathcal{X}_{\mathrm{eval},k}} \delta_{\Phi^{(-d\to r;\widehat{\lambda}_{d,r})}(x)}. \tag{51}$$

Consistent with Equation (16), set

$$\begin{aligned} R_{d,r,k} &= D(\mu_k, \mu_k^{(-d)}) - D(\mu_k, \mu_k^{(-d\to r)}), \\ \bar{R}_{d,r}^{\mathrm{dom}} &= \frac{1}{K} \sum_{\ell=1}^{K} R_{d,r,\ell}, \qquad G_{d,r,k} = R_{d,r,k} - \bar{R}_{d,r}^{\mathrm{dom}}. \end{aligned} \tag{52}$$

Proposition A.16 (Algebraic properties of domain deviations)**.** *For each valid pair,*

$$\sum_{k=1}^{K} G_{d,r,k} = 0. \tag{53}$$

If $R_{d,r,k} = c_{d,r}$ for all valid domains, then $G_{d,r,k} = 0$ for every k. Conversely, zero deviations in all domains imply equal observed rescue scores across those domains.

Proof. Summing Equation equation 52 gives

$$\sum_{k=1}^{K} G_{d,r,k} = \sum_{k=1}^{K} R_{d,r,k} - K\left(\frac{1}{K}\sum_{\ell=1}^{K} R_{d,r,\ell}\right) = 0.$$

If all raw scores equal $c_{d,r}$, their mean is $c_{d,r}$, so each deviation is zero. Conversely, $G_{d,r,k} = 0$ implies $R_{d,r,k} = \bar{R}_{d,r}^{\mathrm{dom}}$ for every k. □

The signs of raw rescue and centered deviation answer different questions. A positive $R_{d,r,k}$ indicates reduced geometric damage in domain k; a positive $G_{d,r,k}$ indicates a score above the pair's domain average, even if its raw rescue remains negative. The zero-sum identity holds by construction and is not evidence of semantic organization. Likewise, nonzero sample deviations do not alone establish population-level domain differences.

The mean in Equation equation 52 assigns equal weight to each valid domain, including when the M_k differ, because the estimand compares domains rather than a frequency-weighted mixture of examples. It is not silently replaced by a sample-size-weighted mean. Empty or unavailable domains must be handled through an explicitly reported valid-domain set, not by inserting zero rescue scores.

A.2 Replication and Functional Inheritance Protocol

This appendix makes the population process in Section 3.3 explicit. The algorithms specify a protocol; the sampling identities below are implementation checks, not additional empirical findings.

A.2.1 Individuals, Data Roles, and Archived States

An individual consists of a unique identity, a slot assignment, its parameter vector $\theta \in \mathbb{R}^P$, and a parent record. As in Section 3.1, inherited coordinates include the query, key, and value projections, the corresponding output-projection block, and any head-specific biases. Compatible slots have the same parameter structure and interfaces. All other network parameters remain fixed. Independent storage is required at initialization and after every birth; two individuals must not share mutable parameter buffers. The initialization rule is recorded explicitly. Randomly initialized individuals use independent draws; checkpoint-based initialization records the source head for each separately stored instance.

Use four disjoint data splits, fixed before the population runs:

Split	Role in the protocol
$\mathcal{X}_{\text{train}}$	Optional gradient updates to offspring parameters only.
$\mathcal{X}_{\text{sel}}$	Pre-reproduction singleton fitness and, when used, domain-conditional selection scores.
$\mathcal{X}_{\text{cal}}$	Choice of rescue coefficients on the prespecified grid.
$\mathcal{X}_{\text{eval}}$	Frozen-protocol evaluation of rescue, task performance, and functional inheritance in a common reference context.

Evaluation data do not select parents, tune coefficients, update parameters, or choose the reported generation or other hyperparameters. Generation schedules, controls, and evaluation statistics are specified in advance.

A.2.2 Population Update

Algorithm 1 follows Equations (1)–(8). It is written for the exact criterion $D = W_2^2$. If an implementation uses a numerical estimator $\widehat{D}$, its scores are recorded as $\widehat{F}_h$; they are not identified with exact F_h without qualification. The operator $\mathsf{Measure}(\Phi; \mathcal{X})$ computes the empirical measure of a fresh forward pass on $\mathcal{X}$. Every singleton ablation starts from the same intact population, so interventions do not accumulate during fitness computation. For domain-weighted selection, the prescribed environment weights satisfy $\omega_k(t) \geq 0$ and $\sum_k \omega_k(t) = 1$.

Mutation uses per-coordinate standard deviation σ, with no projection, clipping, or normalization of the Gaussian perturbation in the specified process. All offspring are copied from the same frozen parental snapshot before population replacement. This prevents within-generation overwriting from changing later offspring's parents. The learning operator satisfies $G_0 = \mathrm{Id}$; optimizer moments are not inherited. A stable implementation computes the softmax after subtracting $\max_h \beta s_h(t)$, which leaves $q(t)$ unchanged.

Proposition A.17 (Conditional offspring-count distribution)**.** *Let Ξ_t include the pre-reproduction population, environment, selection data, and any realized score-estimation randomness, so that $q(t)$*

Algorithm 1 RDHC population update with archived inheritance

Require: $\mathcal{H}, T, \beta, \sigma, L$, fixed network, initialization rule, and random seeds
Require: $\mathcal{X}_{\mathrm{sel}}, \mathcal{X}_{\mathrm{train}}$, $D = W_2^2$, and optional prescribed $\omega(t)$
Initialize Θ_0 in independent storage; assign unique $\mathrm{id}_{h,0}$ and null parents.
Archive initial states, identities, slots, and configuration.
for $t = 0, \dots, T-1$ **do**
  Freeze the population during selection-score computation.
  $\mu_t \leftarrow \mathsf{Measure}(\Phi_t; \mathcal{X}_{\mathrm{sel}})$.
  for $h \in \mathcal{H}$ **do**
    $\mu_t^{(-h)} \leftarrow \mathsf{Measure}(\Phi_t^{(-h)}; \mathcal{X}_{\mathrm{sel}})$.
    $F_h(t) \leftarrow D(\mu_t, \mu_t^{(-h)})$.
  end for
  $s_h(t) \leftarrow F_h(t)$ for all h.
  For domain-weighted selection, measure $F_{h,k}(t)$ on $\mathcal{X}_{\mathrm{sel},k}$ and set $s_h(t) \leftarrow \sum_k \omega_k(t) F_{h,k}(t)$.
  $q_h(t) \leftarrow \exp(\beta s_h(t)) / \sum_i \exp(\beta s_i(t))$.
  $\Theta_t^{\mathrm{snap}} \leftarrow \mathsf{DeepCopy}(\Theta_t)$; freeze this snapshot.
  Archive the snapshot, parent identities, scores, weights, and $q(t)$.
  for $j = 1, \dots, N$ **do**
    Draw $A_t(j) \sim \mathrm{Categorical}(q(t))$ independently **with replacement**.
    $\theta_j^{\mathrm{copy}} \leftarrow \mathsf{DeepCopy}(\theta_{A_t(j)}^{\mathrm{snap}}(t))$ in a new buffer.
    Draw $\epsilon_j(t) \sim \mathcal{N}(0, I_P)$ independently.
    $\widetilde{\theta}_j(t+1) \leftarrow \theta_j^{\mathrm{copy}} + \sigma \epsilon_j(t)$ in all P coordinates.
    Assign a fresh $\mathrm{id}_{j,t+1}$ and record $\mathrm{id}_{A_t(j),t} \to \mathrm{id}_{j,t+1}$.
  end for
  Replace the population simultaneously by $\widetilde{\Theta}_{t+1}$.
  Reset offspring optimizer states; archive pre-learning parameters.
  $\Theta_{t+1} \leftarrow G_L(\widetilde{\Theta}_{t+1}; \mathcal{X}_{\mathrm{train}})$ (offspring parameters only).
  Archive post-learning states, genealogy, mutation seeds, and training settings.
end for
return archived populations, scores, and genealogy.

is fixed conditional on Ξ_t. Define

$$K_h(t) = \sum_{j=1}^{N} \mathbf{1}\{A_t(j) = h\}, \qquad K(t) = (K_1(t), \dots, K_N(t)).$$

Then

$$K(t) \mid \Xi_t \sim \mathrm{Multinomial}(N, q(t)), \qquad \mathbb{E}[K_h(t) \mid \Xi_t] = N q_h(t), \qquad \sum_{h=1}^{N} K_h(t) = N. \tag{54}$$

Proof. The N offspring make conditionally independent categorical draws with the same probability vector. Their category counts therefore have the stated multinomial law. Summing the indicators gives the expectation and the fixed total. □

Under neutral replication, $\beta = 0$ and $q_h = 1/N$. Nevertheless,

$$\mathrm{Var}(K_h \mid \Xi_t) = 1 - \frac{1}{N} > 0, \qquad \Pr(K_h = 0 \mid \Xi_t) = \left(1 - \frac{1}{N}\right)^N > 0.$$

Neutral replication therefore still permits stochastic drift, lineage expansion, and lineage loss. A branching genealogy or an association between prescribed fitness and offspring counts is not, by itself, evidence of an additional emergent mechanism.

A.2.3 Calibration and Evaluation

Common reference context. Let $f_{\mathrm{ref}}(x; \theta)$ denote the functional response obtained by inserting θ into the fixed reference slot while holding every other reference-network parameter fixed. Comparisons use the same input order, readout, preprocessing, and evaluation mode. The distance on

Algorithm 2 Frozen calibration and common-context functional evaluation

Require: Archived states, prespecified generations and valid pairs Ω, calibration data $\mathcal{X}_{\mathrm{cal}}$, evaluation data $\mathcal{X}_{\mathrm{eval}}$, and a finite grid $\Lambda \subseteq [0, \infty)$ containing zero
Require: A fixed reference network, one compatible reference slot, a functional response map, and a prespecified functional distance
for each prespecified population state **do**
  for $(d, r) \in \Omega$ **do**
    On $\mathcal{X}_{\mathrm{cal}}$, evaluate intact, damaged, and rescued measures for every $\lambda \in \Lambda$.
    Select a grid minimizer $\widehat{\lambda}_{d,r}$, using the prespecified tie rule; freeze this coefficient.
    On $\mathcal{X}_{\mathrm{eval}}$, compute $R_{d,r}$ and all valid $R_{d,r,k}$ using that same coefficient.
  end for
  Fit the additive row–column model on Ω and compute $\widetilde{R}_{d,r}$ and $G_{d,r,k}$.
  Retain and report positive, zero, and negative evaluation scores.
end for
for each prespecified parent–offspring comparison **do**
  Replay the archived parent and offspring separately in the same reference network and reference slot.
  Evaluate their responses on the same $\mathcal{X}_{\mathrm{eval}}$ inputs in deterministic evaluation mode.
  Compute the prespecified functional distance to the designated parent and to prespecified shuffled-parent controls.
  Record generation, identities, comparison stage, and repeat seed.
end for
Aggregate over prespecified independent runs and report uncertainty.

these response arrays must be specified before evaluation. For example, an input-matched squared response difference is

$$d_{\mathrm{func}}(\theta, \theta') = \frac{1}{|\mathcal{X}_{\mathrm{eval}}|} \sum_{x \in \mathcal{X}_{\mathrm{eval}}} \| f_{\mathrm{ref}}(x; \theta) - f_{\mathrm{ref}}(x; \theta') \|_2^2. \tag{55}$$

The exact response map, distance, normalization, and reference checkpoint used for the reported experiment must be recorded. Replaying both individuals in a common context avoids conflating parameter inheritance with a change in the surrounding population.

Comparison stages and controls. The parent reference is its archived pre-reproduction snapshot. Offspring are evaluated separately before optional learning ($\widetilde{\theta}_j(t+1)$) and, when applicable, after learning ($\theta_j(t+1)$); these stages must not be pooled without disclosure. Shuffled-parent controls use a prespecified randomization within the same generation and compatible parent set, independent of measured functional similarity. The shuffle rule, repetitions, and treatment of accidental identity matches are reported. A fixed-point-free permutation of parent identities is one valid rule.

Fitness-guided replication is compared with uniform replication ($\beta = 0$), reinitialized offspring, and learning without replication under matched population sizes, data access, evaluation schedules, and training budgets. Reinitialized offspring use the declared initialization rule rather than copying their assigned parents. Learning without replication retains the population and applies the matched training budget. Repeated-run uncertainty must respect dependence among offspring from the same population run.

Parameter copying establishes inheritance of the stored state. Functional inheritance is an empirical outcome under specified mutation, learning, and reference-context choices. It does not follow unconditionally for arbitrary mutation scales, and lineage diagrams alone do not establish functional preservation.

A.3 More Experiments

A.3.1 Experiment Setup

The experiments were conducted on an institutional high-performance computing cluster designed to support high-performance and high-throughput computational workloads. The facility provides large-scale CPU and GPU-accelerated computing resources, including multi-core CPU nodes, high-memory compute nodes, and GPU-enabled nodes. It also provides shared parallel storage and large-capacity

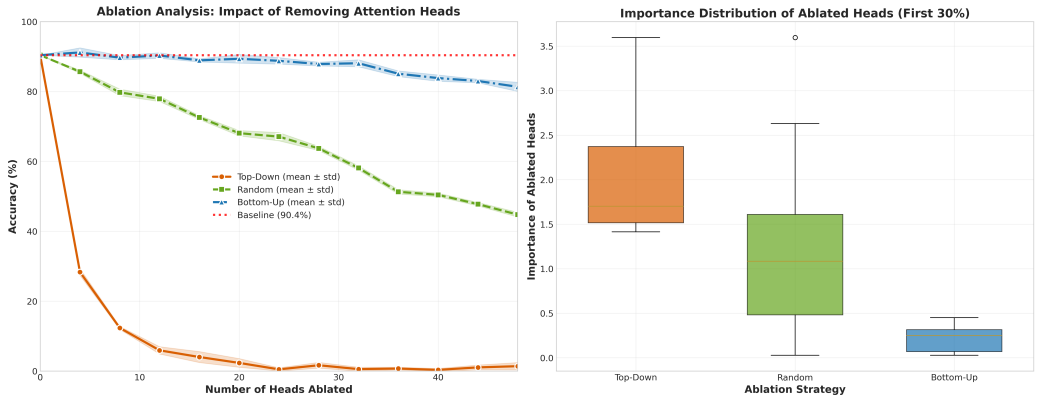

Figure 6: Causal ablation analysis of attention heads under RDHC, illustrating sparse dominance and nonlinear degradation under geometric fitness ranking. Downstream accuracy is shown as a function of the number of ablated heads for three removal strategies: RDHC-guided Top-Down, Random, and Bottom-Up. Shaded regions denote standard deviation across runs. The right figure shows the importance distribution of the first 30% of ablated heads under each strategy.

storage for data-intensive workloads. The cluster runs a Linux-based operating system, and job submission and scheduling are managed using the SLURM Workload Manager.

A.3.2 Experiments on AG_News

Sparse Dominance and Nonlinear Degradation under Geometric Fitness Ranking We begin with the AG_News dataset. The unablated model achieves 90.4% accuracy. We compare three ablation strategies: Top-Down, Random, and Bottom-Up. Figure 6 (left) plots accuracy as a function of the number of removed heads. The degradation regimes differ qualitatively across strategies. Under Top-Down removal, performance collapses sharply and nonlinearly: removing only 4 heads reduces accuracy from 90.4% to approximately 28%, ablating 8 heads pushes performance below 15%, and after 16–20 heads, accuracy approaches near-zero levels. This rapid early-phase collapse is consistent with the sparse dominance emergence proposition, suggesting that a small subset of high-fitness heads governs most of the representation geometry. Under Random removal, accuracy declines gradually and approximately linearly, reaching roughly 45% after ablating 48 heads. This regime reflects distributed but non-uniform contribution: most heads are not individually decisive, yet collectively influence performance. In contrast, Bottom-Up removal leaves the model highly robust when low-fitness heads are removed first. Accuracy remains near baseline for the first 20 removals and stays above approximately 80% even after ablating 48 heads. This behavior indicates substantial redundancy among low-fitness heads and supports the claim that Wasserstein fitness meaningfully stratifies geometric contribution. Figure 6 (right) shows the distribution of fitness values among the first 30% of ablated heads. The Top-Down group occupies a concentrated high-magnitude regime, Bottom-Up heads cluster near zero, and Random spans a broad intermediate range. Together, these results provide empirical evidence for sparse dominance and nonlinear degradation under geometric fitness ranking in representation space.

Domain-Conditional Specialization To analyze geometric sensitivity at the behavioral level, we ablate RDHC-identified high-fitness (”winner”) heads and examine confusion matrices before and after intervention in Figure 7. Winner-head removal induces coherent and directional confusion shifts rather than isotropic collapse. These transitions reflect structured redistribution of representation geometry under intervention. Class-wise accuracy declines heterogeneously: Sci/Tech and World degrade most sharply, while Sports and Business show moderate drops. This asymmetry mirrors conditional fitness heterogeneity and supports sparse dominance at the domain level. In contrast, ablating low-fitness heads produces negligible confusion reorganization. Overall, winner-head removal confirms that a small elite subset governs inter-class representational routing.

Attention Head Specialization Across Entity Classes We next evaluate class-conditional competition among attention heads. Figure 8 (top) presents the normalized class-conditional fitness across

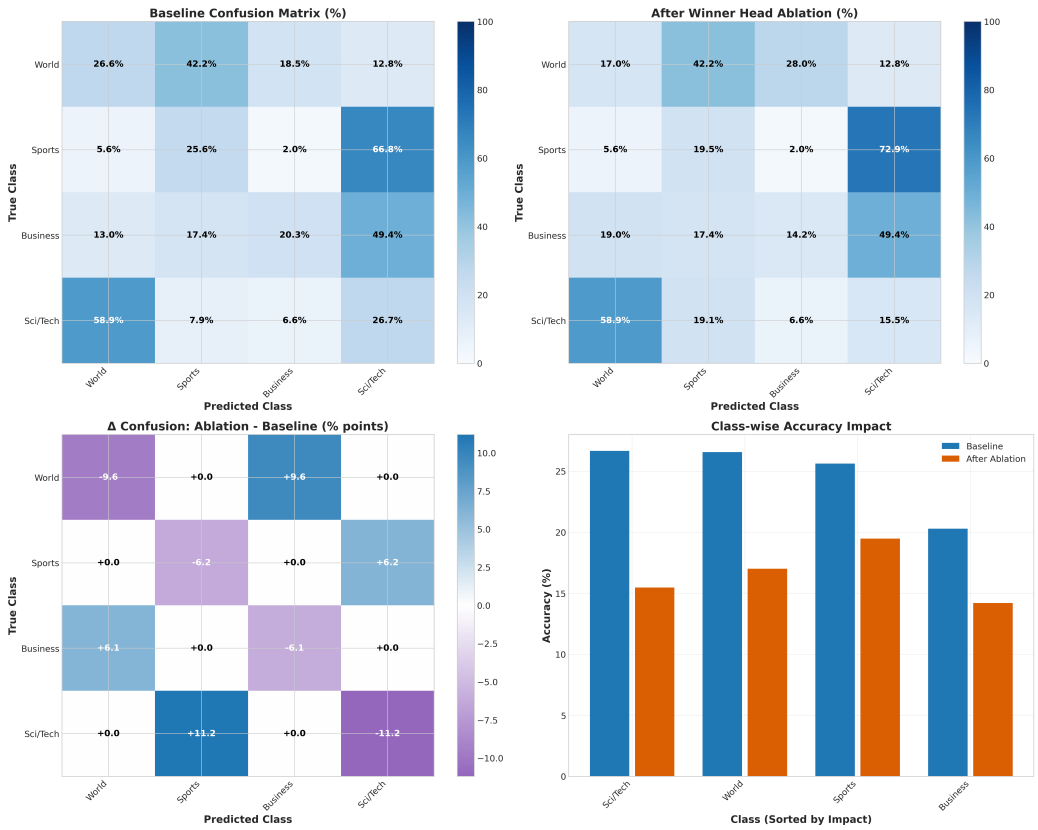

Figure 7: Causal impact of RDHC-identified winner-head ablation on class-level behavior. Top-left: baseline confusion matrix (%). Top-right: confusion matrix after ablating high-fitness attention heads selected by RDHC. Bottom-left: differential confusion matrix, highlighting structured redistribution of prediction mass across semantically related categories. Bottom-right: class-wise accuracy comparison between baseline and ablated models, sorted by impact magnitude.

attention heads. The distribution is highly heterogeneous: only a small subset of heads exerts strong influence for specific entity classes, while the majority contribute weakly and relatively uniformly. Clear vertical banding patterns emerge for certain classes, suggesting a structured allocation of representational influence rather than diffuse sharing. This indicates that class-relevant control is concentrated in a limited number of competitive heads. Figure 8 (bottom) aggregates heads according to their dominant entity class and summarizes their empirical fitness distributions. Heads associated with Sci/Tech display the highest median fitness along with a pronounced upper tail, consistent with the presence of a small number of highly dominant specialists. Business-oriented heads exhibit moderate concentration and intermediate dispersion. In contrast, World and Sports classes show lower central tendency and broader spread, reflecting weaker yet still observable specialization. Overall, these results demonstrate that entity-class specialization emerges at the population level through uneven competitive dynamics. The observed variance in head influence naturally gives rise to differentiated roles, supporting the view that specialization is an emergent consequence of geometric competition in representation space rather than an explicit architectural constraint.

Pair-Specific Rescue Topology and Domain-Dependent Compensation In Figure 9, we further investigate whether RDHC identifies rescue as a global redundancy effect or as a structured interaction between specific damaged and rescuer attention heads. Using the AG_News dataset, we construct a damaged–rescuer rescue matrix in which rows correspond to damaged heads sorted by RDHC fitness and columns correspond to rescuer heads sorted by positive rescue capacity. The resulting matrix shows a highly heterogeneous rescue topology: positive compensation is concentrated in a sparse subset of damaged–rescuer head pairs rather than being uniformly distributed across heads. Several rescuer heads exhibit consistently positive repair effects across multiple damaged heads,

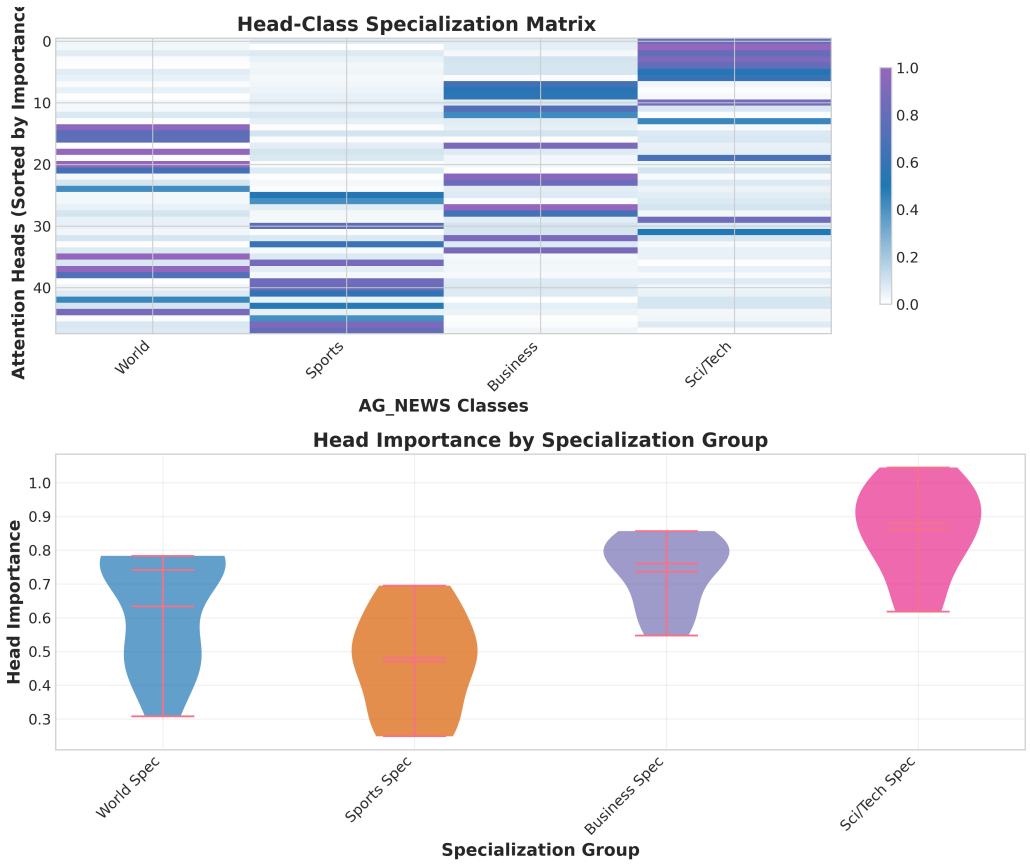

Figure 8: Causal head specialization on AG_NEWS under RDHC. Top: Head–Class Specialization Matrix showing normalized causal fitness scores for all attention heads across the four AG_NEWS classes. Heads are sorted by overall importance. Bottom: Violin plots of head importance grouped by dominant specialization. Distributions highlight hierarchical dominance patterns.

while some damaged heads remain difficult to compensate for, indicating that rescue capacity is unevenly distributed across the attention-head space.

To separate pair-specific structure from marginal head effects, we further remove row and column effects from the rescue matrix. The residual matrix still exhibits clear structured patterns, showing that rescue cannot be fully explained by the average vulnerability of a damaged head or the average capacity of a rescuer head. Instead, successful compensation depends on the compatibility between particular damaged and rescuer head pairs. We also conduct a domain-wise analysis over the four AG_News semantic categories. Representative head pairs display markedly different repair effects across domains, and after pair-mean removal, some pairs retain positive deviations only in specific domains. These results suggest that RDHC not only captures the overall rescue capacity of individual heads, but also reveals pair-specific and domain-dependent functional substitution patterns among attention heads.

A.3.3 More Experimental Results on Yahoo! Answers Topics

Sparse Dominance and Nonlinear Degradation under Geometric Fitness Ranking To investigate how attention head importance affects model robustness, we perform a controlled ablation study under the proposed Geometric Fitness Ranking. Specifically, attention heads are ranked by their geometric fitness scores and progressively removed according to three strategies: Top-Down, Bottom-Up, and Random. We measure the resulting classification accuracy on the evaluation set as the number of ablated heads increases from 0 to 64.

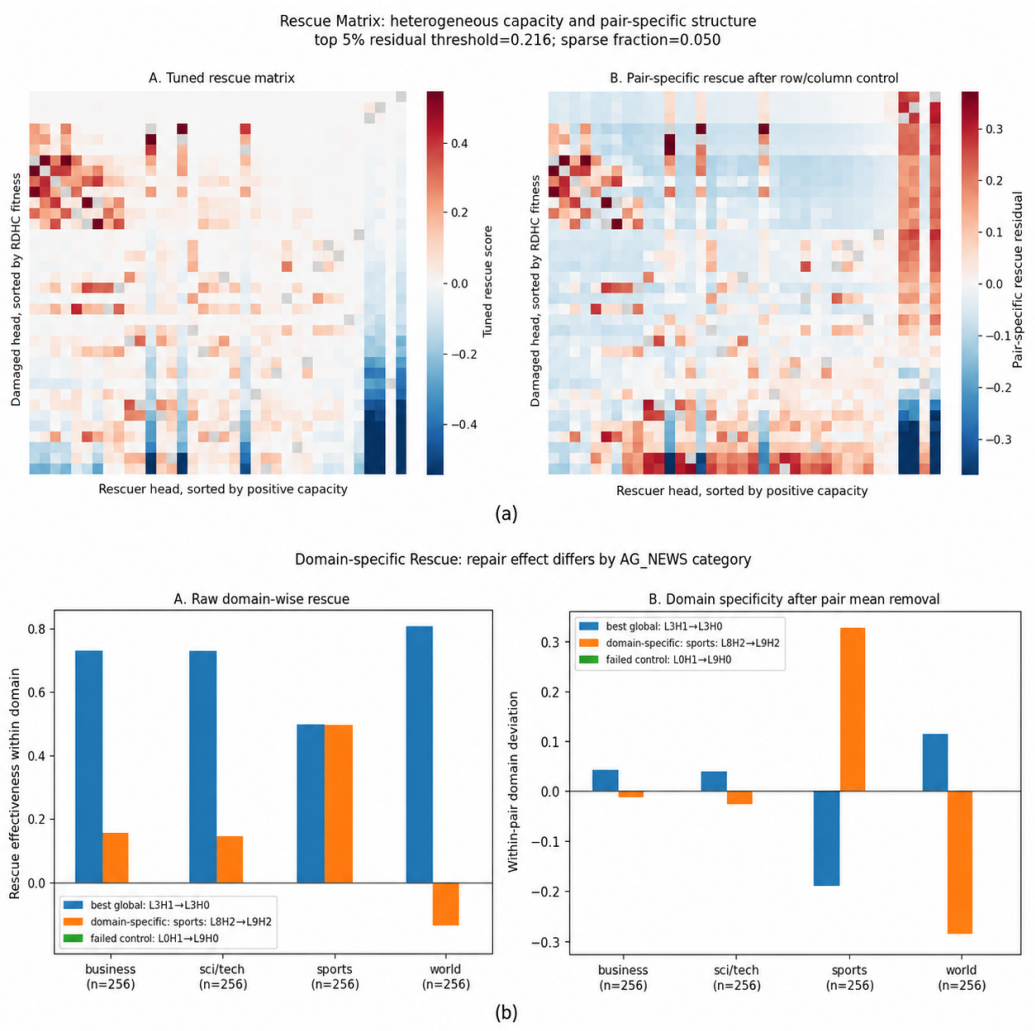

Figure 9: Pair-Specific Rescue Topology and Domain-Dependent Compensation on AG_News. (a) Rescue matrix analysis across damaged and rescuer attention heads. The left heatmap shows the tuned rescue matrix, where rows are damaged heads sorted by RDHC fitness and columns are rescuer heads sorted by positive rescue capacity. Red indicates positive rescue effects and blue indicates negative rescue effects. Rescue is sparse and heterogeneous, with strong compensation concentrated in a limited set of damaged–rescuer head pairs. The right heatmap shows the pair-specific rescue residual after controlling for damaged-head and rescuer-head marginal effects. The remaining structured residual patterns indicate that rescue is not merely determined by global head capacity, but depends on specific head-pair interactions. (b) Domain-wise rescue analysis on AG_News categories. The left panel reports raw within-domain rescue effectiveness for representative head pairs across business, sci/tech, sports, and world domains. The right panel shows domain specificity after removing each pair's mean effect. Certain head pairs exhibit positive deviations only in specific semantic domains, demonstrating that RDHC can reveal domain-dependent compensation in addition to global rescue capability.

Figure 10 (left) shows that removing high-ranked heads leads to an immediate and catastrophic performance drop. Ablating only 6 top-ranked heads reduces accuracy from the baseline 68.1% to approximately 22%, and performance falls below 5% after removing 18 heads, eventually approaching near-zero predictions. In contrast, the Bottom-Up strategy exhibits remarkable robustness: removing up to 50 low-ranked heads causes only a minor degradation, with accuracy remaining above 63%. Even after ablating nearly the entire set of heads, performance only decreases moderately to around 60%. The Random strategy lies between these two extremes, producing a roughly linear degradation

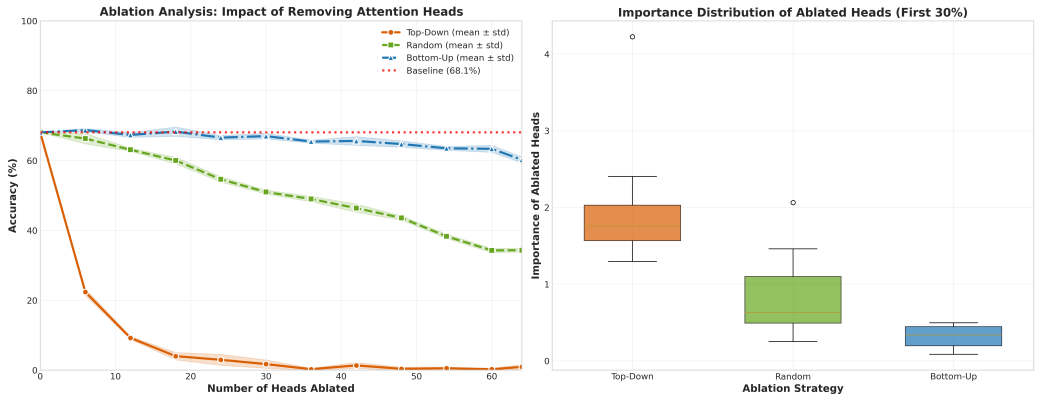

Figure 10: Causal ablation analysis of attention heads under RDHC, illustrating sparse dominance and nonlinear degradation under geometric fitness ranking. Downstream accuracy is shown as a function of the number of ablated heads for three removal strategies: RDHC-guided Top-Down, Random, and Bottom-Up. Shaded regions denote standard deviation across runs. The right figure shows the importance distribution of the first 30% of ablated heads under each strategy.

curve from 68% to approximately 34% as more heads are removed. This asymmetric behavior reveals a sparse dominance phenomenon: a small subset of geometrically high-fitness heads carries a disproportionate amount of functional importance. The performance collapse under Top-Down removal further demonstrates a strongly nonlinear degradation pattern, indicating that model capability depends on a concentrated set of critical heads rather than being evenly distributed.

We further analyze the importance distribution of the first 30% ablated heads (Figure 10, right). The Top-Down strategy consistently removes heads with significantly higher fitness values, while Random sampling exhibits moderate variance and Bottom-Up selects heads with near-zero importance. The clear separation between these distributions provides additional evidence that the proposed geometric ranking successfully identifies structurally critical attention heads. Overall, these results demonstrate that attention importance in language models is highly sparse and structured. The proposed Geometric Fitness Ranking reliably isolates the dominant heads whose removal causes rapid performance collapse, validating its effectiveness as a principled criterion for interpreting and analyzing transformer attention mechanisms.

Domain-Conditional Specialization To examine whether individual heads develop domain-specific expertise, we conduct a winner-head ablation analysis and measure the resulting changes in class-wise confusion patterns in Figure 11. For each input, we first identify the head with the highest routing the winner head and remove its contribution at inference time. Figure 11 compares the baseline confusion matrix with the matrix obtained after ablation, together with the difference map and the class-wise accuracy changes. The results show a systematic degradation concentrated along the diagonal entries of the confusion matrix, indicating that the removed head provides class-specific discriminative signals. Several domains exhibit substantial drops in accuracy. In particular, Science and Mathematics decreases from approximately 55% to 35%, Education and Reference from 41% to 27%, and Family and Relationships from 39% to 29%. Similar reductions are observed for Sports and Business and Finance. The difference matrix further reveals that these losses are accompanied by increased confusion with semantically related domains. For example, Science and Mathematics becomes more frequently confused with Health and Education and Reference, while Business and Finance shows increased confusion with Politics and Government. Importantly, the degradation is not uniform across classes but concentrated in a subset of domains. These observations suggest that certain heads encode domain-specialized decision boundaries rather than generic features. Overall, the results support our hypothesis that the model implicitly learns domain-conditional specialization, where specific heads become responsible for resolving fine-grained distinctions within particular semantic regions of the label space.

Attention Head Specialization Across Entity Classes To investigate whether attention heads exhibit functional specialization across semantic domains, we analyze head-level importance with

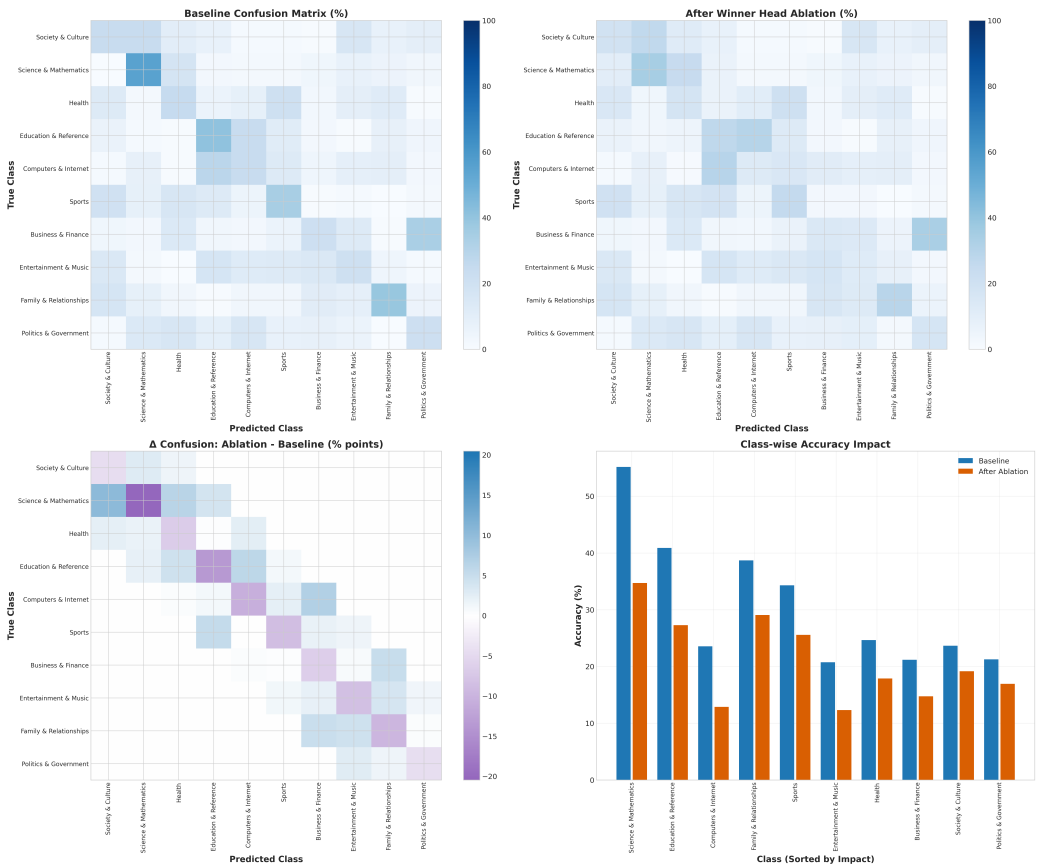

Figure 11: Causal impact of RDHC-identified winner-head ablation on class-level behavior. Top-left: baseline confusion matrix (%). Top-right: confusion matrix after ablating high-fitness attention heads selected by RDHC. Bottom-left: differential confusion matrix, highlighting structured redistribution of prediction mass across semantically related categories. Bottom-right: class-wise accuracy comparison between baseline and ablated models, sorted by impact magnitude.

respect to the ten category labels in the Yahoo! Answers dataset. Figure 12 visualizes a head–class specialization matrix, where each row corresponds to an attention head and each column represents a semantic class. The heatmap reveals a sparse but structured activation pattern: while many heads distribute their attention broadly, a subset of heads demonstrates strong class-specific responses, suggesting emergent specialization. For example, several high-importance heads show concentrated activation for knowledge-intensive categories such as Science and Mathematics, Education and Reference, and Business and Finance. In contrast, socially oriented categories such as Family and Relationships and Entertainment and Music display more distributed attention across heads, indicating that these topics rely on more shared linguistic patterns rather than tightly specialized representations.

To further quantify these effects, we group heads by their dominant specialization and examine their importance distributions (Figure 12, bottom). The violin plots show clear variability in head importance across categories. Knowledge-centric classes—including Science and Mathematics, Education and Reference, and Politics and Government—tend to be associated with higher median head importance and tighter distributions, indicating the presence of a small number of highly influential specialized heads. In contrast, categories such as Computers and Internet and Entertainment and Music exhibit lower median importance and broader distributions, suggesting a more diffuse allocation of attention across multiple heads. Overall, these findings support the hypothesis that transformer attention heads develop partial semantic specialization aligned with topical structure. Rather than contributing uniformly across all classes, certain heads behave as domain-sensitive detectors that capture features particularly relevant to specific knowledge areas. This emergent

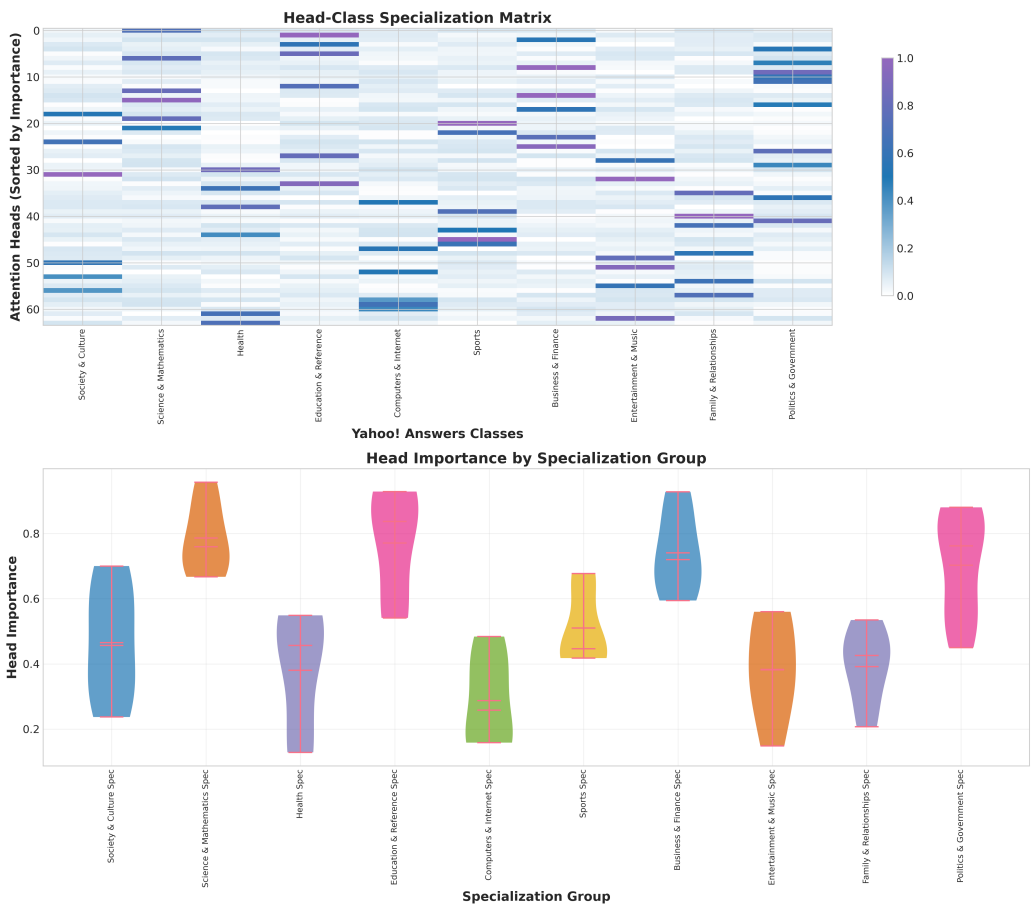

Figure 12: Causal head specialization on Yahoo! Answers Topics under RDHC. Top: Head–Class Specialization Matrix showing normalized causal fitness scores for all attention heads across the ten Yahoo! Answers Topics classes. Bottom: Violin plots of head importance grouped by dominant specialization. Distributions highlight hierarchical dominance patterns.

specialization provides further evidence that attention mechanisms encode structured semantic roles even without explicit supervision at the head level.

Pair-Specific Rescue Topology and Domain-Dependent Compensation We further evaluate whether RDHC reveals structured rescue relationships among attention heads on the Yahoo! Answers Topics dataset. Figure 13 shows that the rescue behavior is highly non-uniform across damaged–rescuer head pairs. In the tuned rescue matrix, positive repair effects are concentrated in a sparse subset of pairs rather than being uniformly distributed, indicating that compensation is not simply a consequence of global redundancy among attention heads. Some rescuer heads consistently improve multiple damaged heads, while some damaged heads remain difficult to compensate for across most rescuers. More importantly, after controlling for the marginal effects of damaged heads and rescuer heads through row/column mean removal, the residual rescue matrix still exhibits clear structured patterns. This suggests that the repair effect cannot be fully explained by the average rescue capacity of an individual head; instead, it depends on the compatibility of specific damaged–rescuer head pairs.

We also conduct a domain-wise analysis over Yahoo! Answers topics to examine whether rescue effects are semantically conditional. Representative head pairs show substantially different repair strengths across domains. After removing each pair's mean rescue effect, several pairs still exhibit positive deviations only on specific domains, suggesting that functional compensation is not only pair-specific but also domain-dependent. These results indicate that RDHC can characterize both the global

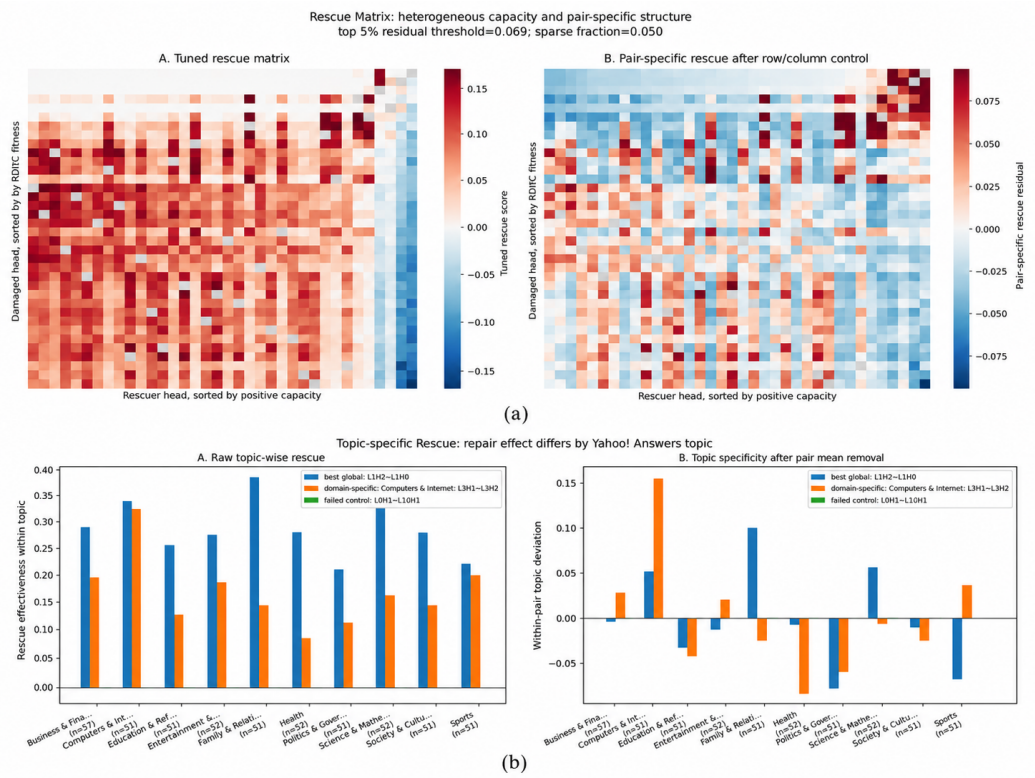

Figure 13: Pair-Specific Rescue Topology and Domain-Dependent Compensation on Yahoo! Answers Topics. (a) The tuned rescue matrix reveals heterogeneous compensation patterns among attention heads. Rows correspond to damaged heads sorted by RDHC fitness, and columns correspond to rescuer heads sorted by positive rescue capacity. Positive rescue effects are concentrated in a sparse set of damaged–rescuer pairs, showing that compensation is not uniformly distributed across heads. After controlling for damaged-head and rescuer-head marginal effects, the pair-specific rescue residual still preserves visible structure, indicating that repair cannot be explained solely by the average capacity of individual heads. (b) Topic-wise rescue analysis for representative head pairs on Yahoo! Answers Topics. Raw topic-wise repair effects vary substantially across semantic domains. After pair mean removal, several head pairs retain positive deviations only for particular topics, demonstrating that rescue behavior is domain-dependent. Together, these results show that RDHC uncovers a non-uniform rescue topology in which attention-head compensation depends on both pair-specific compatibility and semantic domain.

rescue capacity of attention heads and the finer-grained topology of pair-specific, domain-sensitive functional substitution.