THE SAUSSUREAN REPRESENTATION HYPOTHESIS: A CAUSAL ACCOUNT OF REPRESENTATIONAL STRUCTURES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic Interpretability represents a dominant set of approaches and often seeks to localize model responses to particular input features in specific attention heads, neurons, or circuits, typically validating such localizations by ablating the identified model components and measuring degradation of the targeted capability. This target-centric approach frequently fails to engage with a core tenet of semiotics: that meaning arises from the relations among concepts rather than the concepts themselves. Here we examine collateral damage, investigating how ablating one concept within a model affects performance on other similar concepts. We introduce a non-parametric, information-theoretic method for performing ablations. By estimating attention-head-level mutual information, we reveal which components of a model encode information related to specific features in a data distribution. Examining languages and WordNet-derived knowledge domains, we show that model representations recover independently defined similarity structure in both, placing languages of the same subfamily, and topically related domains, closer together in their representation space. Ablating heads with the highest pointwise mutual information for a given language causes significant damage to the model's performance on that language, and also affects members of the same language subfamily. When we apply this non-parametric method to 51 WordNet domains derived from Wikipedia across Qwen, Llama, OLMo, and Gemma models, we find that collateral damage across concepts consistently correlates with the semantic distance between them. We term this the Saussurean Representation Hypothesis: in LLMs, the functional representation of a language or conceptual domain is constituted in substantial part by its relations to neighboring representations. Consequently, causal localization cannot be evaluated only by what an intervention removes; what it removes in parallel with the target reveals the structure of the representation itself.

1 INTRODUCTION

Large language models (LLMs) acquire representations of a broad range of concepts during training. This has led to a growing literature analyzing how models represent information. Mechanistic interpretability represents a dominant set of approaches, a core aim of which has been to localize representations within specific parts of a model such as individual neurons, attention heads, circuits, or directions in activation space (Olah et al., 2020; Bau et al., 2017; Dai et al., 2022; Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023). Claims are typically validated by intervention: the identified components are ablated or steered, and a claim about the role of that component is supported to the extent that performance on the targeted capability degrades. Implicit in this approach is an assumption about the nature of representation that we refer to as *atomism*: the view that the content of a concept is intrinsic to it and can therefore be identified, localized, and removed independently of other concepts.

This position is most explicit in work on polysemanticity, the observation that individual parameters in a model often respond to several seemingly unrelated properties of the input. Elhage et al. (2022) attribute polysemanticity to superposition, in which a network represents more features than it has dimensions by assigning them non-orthogonal directions. On this account, a trained network ap-

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.

proximates a hypothetical larger network in which each feature is represented without interference from the others. Overlap between features is thus treated as the price paid for a compressed model of high-dimensional data, and the aim of interpretation becomes the recovery of monosemantic features, each corresponding to a single interpretable property. Sparse dictionary learning methods, in particular sparse autoencoders (SAEs), were developed for this purpose - a model that tries to decompose dense representations into a dictionary of monosemantic features (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024).

Subsequent work applying these methods at scale has, however, repeatedly documented relational structure among the recovered features. In SAEs trained on an LLM, cosine similarity between SAE representations for different features approximately tracks their semantic relatedness. Features learned by smaller dictionaries also divide into a number of geometrically proximate and semantically related features in larger ones (Templeton et al., 2024). Leask et al. (2025) describe how features in larger SAEs are frequently decomposable into combinations of features from smaller SAEs; for example, when a larger model represents a dedicated feature for "Einstein," a smaller model represents this as a composition of representations for "scientist," "Germany," and "famous person," each of which itself shares relational structure with still other concepts. Outside of mechanistic approaches, Huh et al. (2024) characterize representations by the similarity structure they induce over inputs, and report that this structure is increasingly aligned across models, architectures, and modalities; they interpret this convergence as evidence that models approximate a statistical model of a shared reality. In each case, the observed similarity structure has been interpreted as a byproduct of compression, an aid to navigating large feature sets, a consequence of dictionary size, or a reflection of the statistical structure of the world. The possibility that it reflects the nature of meaningful representations themselves has received comparatively little attention.

The cognitive sciences have long studied how meaning is represented, largely rejecting atomism in favor of *structuralism*: the view that the content of a representation is determined in substantial part by its relations to other representations. The position most notably originates with de Saussure (1916), for whom the value of a representation derives not from intrinsic content but from how it contrasts with other representations in the same system. The classic illustration is that French *mouton* and English *sheep* only partially overlap in meaning because English also has *mutton* while French does not; each representation depends on its neighbors. Note that Saussure emphasized difference whereas our measure is proximity; but these are complementary, since nearest neighbors are similar but distinguishable. Within the literature on concepts, prototype and exemplar theories in which category membership is graded by similarity (Wittgenstein, 1953; Rosch, 1975; Medin & Schaffer, 1978) supplanted the classical view that had defined concepts by necessary and sufficient conditions. Similarity structure also governs generalization: the probability of generalizing from one stimulus to another decreases systematically with their distance in psychological space (Shepard, 1987; Tenenbaum & Griffiths, 2001).

Of particular relevance to the present work, neuropsychology has inferred the organization of semantic memory from the *pattern* of deficits produced by brain damage, rather than from the loss of individual targets. Semantic dementia erodes distinctions among related concepts before broader category knowledge (Hodges et al., 1992). Lesioned distributed network models reproduce these patterns (Hinton & Shallice, 1991; Farah & McClelland, 1991; Rogers & McClelland, 2004). Under a structuralist account, damage that extends beyond the target of an intervention is informative about how that target is represented.

We formalize this account for LLMs as the **Saussurean Representation Hypothesis (SRH)**: the functional representation of a language or concept is based on its relations to neighboring representations. Not out of convenience, or capacity limitations, but because the relations themselves are what encode meaning. The name is chosen in part in contrast to the Platonic Representation Hypothesis where Huh et al. (2024) note representational similarity structure aligns across models but attribute this to a correspondence with a shared external reality. We argue instead that models converge to particular similarity structures not as convergence to a platonic ideal of a concept, but because the similarity structures themselves are what encodes meaning.

Here we ask whether functional representations in LLMs show a signature of the relational structure Saussure described: if the representations of a full language and an individual concept are constituted by their relations to their neighbors, then removing the components that support the target representation will also degrade its neighbors, in proportion to their proximity. That is, if

the learned relational organization matters to the model's downstream computations, intervening on components informative about one category should produce a structured profile of effects on others. This prediction distinguishes the SRH from three alternatives:

1. *Localization* predicts negligible effects on concepts other than the target.
2. *Generic capacity loss* predicts effects distributed uniformly across concepts.
3. *Interference due to compression* predicts effects organized by co-occurrence rather than similarity. In models of superposition, features that frequently co-occur tend toward orthogonal representations, whereas features that rarely co-occur tend to share dimensions (Elhage et al., 2022). Because distinct languages seldom co-occur within a document, this account does not predict effects graded by linguistic relatedness.

To test the SRH, we introduce a training-free, fast-to-compute information-theoretic ablation method. We use soft density estimates of attention-head activations to compute the pointwise mutual information (PMI) between each head and a property of the input distribution, such as its language or knowledge domain. The method requires no trained probes, sparse dictionaries, or labeled contrast pairs, only samples conditioned on the property of interest. We apply it in two domains with independently established similarity structures: the Indo-European (IE) languages, using the phylogeny tree of Heggarty et al. (2023), and 51 WordNet-derived knowledge domains (Miller, 1995), using a reference we build from Wikipedia's link graph.

In both domains, model representations recover the structure of the reference; for languages, this extends prior work (Rakshit et al., 2026), and for WordNet, it is a new result. Ablating the heads with the highest PMI for a given language significantly degrades performance on that language and on other members of its subfamily. For same-script Indo-European language pairs, collateral damage from PMI based ablation decreases with phylogenetic distance from the ablated language (pooled within- ablation arms Spearman $\rho = -0.53$; $p < .001$). Similarly across 51 knowledge domains and four model families (seven checkpoints) collateral damage to a concept is predicted by Wikipedia link-profile distance from the ablated concept (pooled within- ablation arms Spearman $\rho = -0.336$, $p < .001$).

Our contributions are as follows:

1. We state the Saussurean Representation Hypothesis as a testable claim about the causal organization of representations in LLMs, and situate it relative to atomist, superposition-based, and convergence-based accounts.
2. We introduce a non-parametric method, based on head-level mutual information, for identifying and ablating the components that encode a given property of a data distribution.
3. We show that LLM representations recover the topical similarity structure of WordNet-derived knowledge domains, measured independently from Wikipedia's link graph, complementing prior results on language phylogeny.
4. We provide causal evidence, across two domains and four model families, that the collateral damage produced by ablation is graded by similarity to the ablated target.

2 Background and Related Work

Localization and head-level attribution. Localization methods are typically evaluated against a single target; circuit discovery, for example, asks whether a subgraph is faithful, complete, and minimal for one task (Wang et al., 2023; Conmy et al., 2023). Several findings complicate this procedure: layers identified by causal tracing poorly predict where edits to a fact are effective (Hase et al., 2023), and ablating one component can induce compensatory changes in others (McGrath et al., 2023). Attention heads are a natural unit for such analyses, and early ablation studies found that most can be removed with little loss, while a few are critical (Michel et al., 2019; Voita et al., 2019). Causal head gating (CHG; Nam et al., 2025) scales head attribution to large models by learning soft gates over heads under next-token prediction, classifying each head as facilitating, interfering, or irrelevant to performance on a dataset. Our method addresses the same problem but

estimates each head's relevance from the mutual information between its activations and a property of the data, without optimizing any parameters.

Relational structure among learned features. Beyond the findings reviewed in Section 1, SAE features exhibit absorption, in which a general feature fails to activate where a more specific related feature does (Chanin et al., 2024), and some features are irreducibly multidimensional, such as days of the week arranged on a circle (Engels et al., 2025). Attribution-graph analyses group features sharing a computational role into composite nodes (Ameisen et al., 2025). Such structure is treated either as a property of the analysis method or as a departure from an ideal of independent features, not as part of what the features encode.

Representational geometry and convergence. Representational similarity analysis (Kriegeskorte et al., 2008) and kernel alignment (Kornblith et al., 2019) compare representations through the similarity structure they induce over inputs. Huh et al. (2024) use this relational characterization to show that representations converge across models and modalities, and that alignment predicts downstream performance, paralleling the link between similarity structure and cross-lingual transfer in Rakshit et al. (2026). The SRH differs in concerning structure within a single model rather than across models, in resting on intervention rather than correlation, and in treating relational structure as constitutive of content rather than as a reflection of external reality. Closest to our WordNet analysis, Park et al. (2025) show that concepts are represented as simplices and hierarchically related concepts as orthogonal, using over 900 WordNet concepts; we ask whether such geometry determines how interventions on one concept affect others.

Collateral effects have been treated as failure criterion. Where interventions are evaluated beyond their target, effects on related concepts are typically counted as errors. Model-editing benchmarks penalize changes to facts about related subjects (Meng et al., 2022), even as Cohen et al. (2024) argue that edits ought to propagate to logically implied facts. In unlearning, Li et al. (2024) observe the largest losses on the general-knowledge topics closest to the removed knowledge, and treat this as a problem for future methods; concept erasure likewise seeks minimal change outside the target (Ravfogel et al., 2020; Belrose et al., 2023). On the other hand, the SRH implies that the specificity of any such intervention is bounded by the target's relations to other concepts. We argue that some amount of collateral damage is an inevitable effect of distributed relational representations. Such collateral damage in fact reveals meaningful aspects of model representations.

Structuralism and holism. Atomist theories hold that a concept's content is fixed independently of other concepts (Fodor, 1998; Wierzbicka, 1996), whereas conceptual role semantics holds that content is determined by relations among concepts (Block, 1986), a view with a direct lineage in NLP through the distributional hypothesis (Harris, 1954; Firth, 1957). The principal objection to relational theories is that, if content depends on all relations, no two systems can share content (Fodor & Lepore, 1992). The SRH claims graded, local dependence, whose extent our experiments measure, and cross-model convergence (Huh et al., 2024) suggests relationally constituted content is in practice sufficiently shared.

Multilingual representations. Multilingual models interpret, generate, and translate text across a broadening array of languages (Qin et al., 2025; Pires et al., 2019; Conneau et al., 2020), with transfer being stronger between languages sharing vocabulary and typology (Wu & Dredze, 2019), as representations reflect historical (phylogenetic) relations (Rama et al., 2020; Rakshit et al., 2026). Mechanistic studies have aimed to identify neurons that selectively impair only one language when deactivated (Tang et al., 2024; Kojima et al., 2024). Here we ask whether components informative about one language also support other languages in proportion to their phylogenetic proximity.

3 METHOD

We test whether a target's causal dependencies are structured by its relations to other targets. An *arm* a is the language or knowledge domain selected for intervention; E is an evaluation category. We split each category's text into selection and held-out evaluation material. Selection text determines all head scores and substitution means; evaluation text measures loss only. For attention head $c =$

(ℓ, h), let $z_t^{(c)} \in \mathbb{R}^{d_c}$ be its output at token t, immediately before the attention output projection W_O. All model weights remain frozen.

Estimating category information. We subtract a calibration mean from each $z_t^{(c)}$, normalize it to unit length, and assign it softly to a fixed random codebook of M unit vectors $e_m^{(c)}$ (uniformly sampled over the unit hypersphere S^{d_c-1}):

$$q_t(m \mid c) = \mathrm{softmax}_m\left(\frac{\langle \hat{z}_t^{(c)}, e_m^{(c)} \rangle}{\tau}\right), \quad P_L(m \mid c) = \frac{1}{|\mathcal{T}_L|} \sum_{t \in \mathcal{T}_L} q_t(m \mid c). \tag{1}$$

Here $\mathcal{T}_L$ contains category L's selection tokens. Counts are accumulated during forward passes; token activations need not be cached. With N equally weighted categories and $\bar{P}_c = N^{-1} \sum_L P_L(\cdot \mid c)$, define

$$S_{cL} = D_{\mathrm{KL}}\big(P_L(\cdot \mid c) \big\| \bar{P}_c\big), \qquad \frac{1}{N} \sum_L S_{cL} = I(M_c; L). \tag{2}$$

M_c is the sampled codeword of head c under a uniform category prior. The equality follows from the definition of mutual information: S_{cL} apportions the information of head c across categories.

Head ranking and intervention. Raw S_{ca} can be large because head c may distinguish many categories, or because arm a is broadly distinguishable throughout the model. To further pinpoint how much a given head's information is allocated to the arm, let $R_c = \sum_L S_{cL}$, $C_a = \sum_c S_{ca}$, and $Z = \sum_c R_c$. Normalizing the allocation table gives $Q(c, a) = S_{ca}/Z$, with marginals $Q(c) = R_c/Z$ and $Q(a) = C_a/Z$. We score each head–arm pair by the pointwise mutual information under this joint distribution:

$$\Phi_{ca} = \log \frac{Q(c,a)}{Q(c)Q(a)} = \underbrace{\log S_{ca}}_{\text{arm-specific term}} - \underbrace{\log R_c}_{\text{head marginal}} - \underbrace{\log C_a}_{\text{arm marginal}} + \log Z \tag{3}$$

The head marginal R_c discounts heads with large distinctiveness across *all* categories. The arm marginal C_a calibrates the pair's PMI against how much distinctiveness the arm elicits across *all* heads. For a fixed arm, C_a and Z are constant across heads, so only S_{ca}/R_c changes their ranking; the arm-marginal correction affects the PMI value, not which heads are selected. We select the k highest-ranked heads $T_k(a)$ and, at the input to W_O, apply

$$z_t^{(c)} \longmapsto z_t^{(c)} - \mu_a^{(c)} + \bar{\mu}^{(c)} \quad (c \in T_k(a)), \tag{4}$$

where $\mu_a^{(c)}$ is the head's mean output on arm a's selection text and $\bar{\mu}^{(c)} = N^{-1} \sum_L \mu_L^{(c)}$. In effect, we subtract the arm-specific offset $\mu_a^{(c)} - \bar{\mu}^{(c)}$ from every token, including tokens in *other* evaluation categories. This replaces the target arm's mean with the cross-category mean and leaves token-to-token deviations intact.

Matched controls and effect size. For each arm, we draw 30 random k-head sets with *that arm's own* number of selected heads per layer, excluding $T_k(a)$; each receives the same arm-specific substitution rule in Eq. (4). Let n_{aE}^* be the evaluation negative log-likelihood (NLL) per character under $T_k(a)$, $n_{aE}^{(j)}$ the NLL under random set j, and n_E^0 the unmodified base model's NLL. Writing $\bar{n}_{aE}$ and s_{aE} for the sample mean and standard deviation of the 30 controls, we report

$$z_{aE} = \frac{n_{aE}^* - \bar{n}_{aE}}{s_{aE}}, \qquad \mathrm{exc}_{aE} = \frac{n_{aE}^* - \bar{n}_{aE}}{n_E^0} \tag{5}$$

z measures departure from this empirical null; where excess gives the additional damage in percent of the evaluation category's baseline loss. The main heatmaps display $\log_2(\mathrm{exc}_{aE} / f_a)$, where f_a is the median excess of non-target categories outside the arm's reference subfamily or domain family. This ensures that the structure of the collateral damages post-ablation becomes more salient in the visualizations.

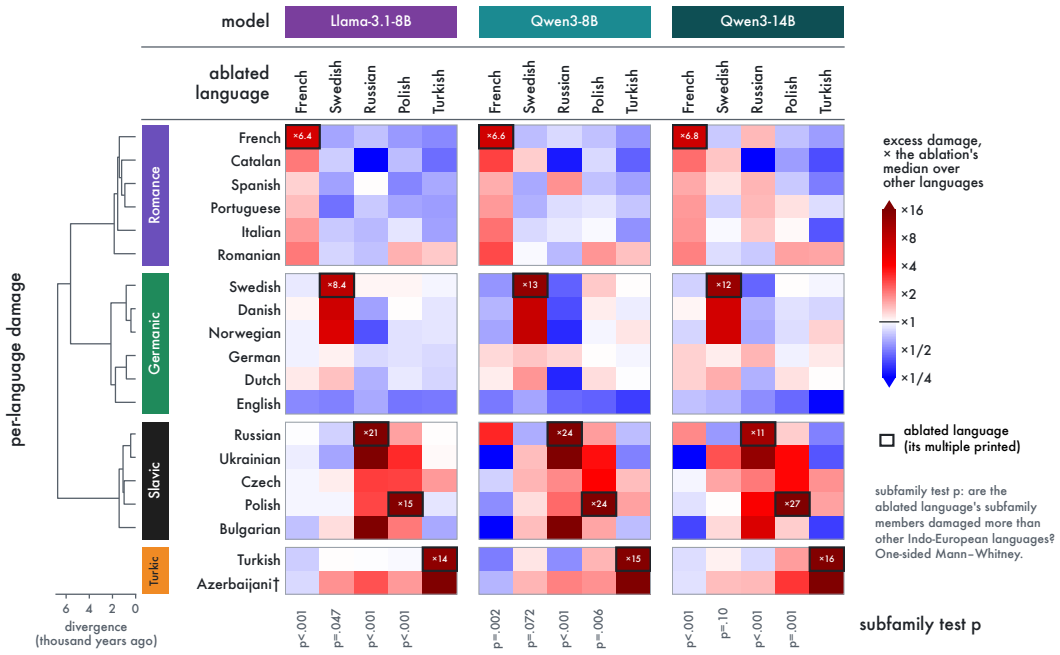

Figure 1: **Ablating one language's heads damages its relatives most.** For each model we remove the k=256 attention heads most specific to one language, chosen without training by PMI (columns: French, Swedish, Russian, Polish, Turkish), and measure the loss on every evaluated language (rows). Colour shows each cell's damage as a multiple of that ablation's median over the other 27 languages of the 28-language common set (log scale; red: damaged more than the median, blue: less). Outlined cells are the ablated language, with its multiple printed. Rows are grouped by family and ordered by the Heggarty et al. (2023) tree (left; UPGMA on patristic distances, axis in thousands of years). In all three models, damage concentrates in the ablated language's own subfamily for French (Romance) and for Russian and Polish (Slavic). For all 28 languages, see Figure 8

3.1 Experimental setup

Languages: We evaluate Qwen3-8B-Base, Qwen3-14B-Base, Apertus-8B-2509, and Llama-3.1-8B. The estimator uses 51 categories: 49 languages from mC4 (Xue et al., 2021) plus Serbian written natively in Cyrillic and in Latin (FineWeb-2; Penedo et al., 2025). Azerbaijani (FineWeb-2) is used for evaluation only. The eight intervention arms are French, Spanish, German, Swedish, Russian, Polish, Hindi, and Turkish. We compare held-out collateral effects with subfamily membership and patristic distance on the independent Indo-European phylogeny of (Heggarty et al., 2023) (Figure 6).

Knowledge domains: The identical PMI pipeline runs on 51 WordNet-derived domains (Miller, 1995) comprising 5,769 cleaned English Wikipedia articles, split by article into selection and evaluation halves. We test eleven arms (aviation, maritime, zoology, biology, textiles, sports, music, law, astronomy, philosophy, crime) in Qwen3-8B-Base, Qwen3-14B-Base, Llama-3.1-8B, OLMo-2-7B, and Gemma-2-9B. For display, we arranged the domains into four Dewey Decimal groups (Dewey, 2011) (life sciences, technology, social sciences, arts). To evaluate the model-derived conceptual structure, we construct an independent, model-agnostic reference tree using pairwise semantic proximities from Wikipedia link profiles, serving as the topical analogue to the Heggarty language phylogeny (see Appendix A.2 for tree derivation and topology).

For languages we sweep $k \in \{32, 64, 128, 256, 576\}$ (512 instead of 576 for Apertus-8B) and report $k = 256$. For domains we set k to 22% of a model's heads (and 11% as a check): $k = 256$ in Qwen3-8B, 228 in Llama-3.1-8B and OLMo-2-7B, and 356 in Qwen3-14B. Heads are ranked by Eq. (3).

4 RESULTS

4.1 REPRESENTATIONAL GEOMETRY OF MODELS BEFORE INTERVENTION

Before any intervention, the models' representations already separate and organize categories within their representation space in ways that track independently motivated similarity. The Jensen–Shannon divergence between languages' distributions of hidden-state codewords, averaged across layers, correlates with Heggarty et al.'s Indo-European patristic distance over 37 languages (Mantel $\rho = 0.64$ in Qwen3-8B, 0.60 in Llama-3.1-8B, 0.65 in OLMo-2-7B; each permutation $p < .001$). For the 51 WordNet knowledge-domains, the corresponding correlations with Wikipedia link-profile distance are 0.69, 0.72 and 0.72 (each $p < .001$). Tree-level agreement is also high: cophenetic correlation with the reference tree is 0.69, 0.61 and 0.70 for languages and 0.35, 0.43 and 0.40 for domains (relabel-null $z = 11$–17). The model-derived dendrograms, for both languages and knowledge-domains, can be found at Appendix A.3.

4.2 TARGETED HEAD ABLATION HAS STRUCTURED COLLATERAL EFFECTS

The pre-intervention dendrograms in Figure 7 establish clear representational structure among languages and concept domains, respectively, but do not reveal whether the structures have functional consequences. They are constructed from hidden-state distributions, whereas the PMI-based intervention ranks and mean-ablates attention-head outputs. We therefore ask whether excess damage on held-out categories follows the model-independent reference relationships used to evaluate the dendrograms. For each target, this yields a directed profile of collateral damage beyond 30 random head sets matched to that target's selected heads per layer. The empirical test is whether the *pattern* of collateral damage across other categories is predicted by the model-agnostic reference structures (Figure 6).

To determine the targets to be ablated, we rank attention heads by the arm-specific information allocation Φ and replace the selected heads' arm-conditioned mean output with the cross-category mean, using (3) and (4). Excess damage and damage from floor are as defined in Section 3; both separate selective collateral effects from loss caused by removing the same number of heads at the same depths. In the language heatmaps (Figure 1), ablating the top-k heads selected for French also damages Spanish and Catalan, removing heads selected for Russian damages other Slavic languages, and so on. A one-sided Mann–Whitney test comparing the arm's own subfamily with the other Indo-European languages is significant for French, Russian and Polish in all three models ($p \leq .006$) and for Swedish only in Llama-3.1-8B (with $p = .047$; while Qwen3-8B .072, Qwen3-14B .10).

The evaluation of the digraphic Serbian text disambiguates the role of script here, since collateral damage on Serbian is enhanced in the Cyrillic script when ablating Russian heads, for instance. The only non-Indo-European arm, Turkish, damages its close relative Azerbaijani 19–61$\times$ its median, while no other Latin-script language exceeds 3.6$\times$. See Appendix A.1.2 for details on how shared ancestry and script modulate collateral damage.

4.3 KNOWLEDGE-DOMAIN ABLATION

We next apply the same pipeline to the 51 knowledge domains, arranged into four Dewey Decimal groups (life sciences, technology, social sciences, arts) of 6–10 domains per group; e.g., life sciences comprises ecology, agriculture, zoology, botany, biology and medicine. We expand the models tested to include OLMo-2-7B, and apply the same ablation of heads based on PMI. Knowledge-domain ablation shows a corresponding pattern of collateral damage on related concepts. Across all four models, damage spreads to related neighbors: biology to zoology and medicine, maritime to aviation, transportation and engineering, law to economics, finance and crime, and music to theater and dance. Figure 2 shows results from four models (OLMo-2-7B, Llama-3.1-8B, Qwen3-8B, Qwen3-14B).

4.4 COLLATERAL DAMAGE IS GRADED BY INDEPENDENT SIMILARITY-DISTANCE METRICS

For same-script Indo-European targets, excess damage falls with patristic distance: pooled within-arm Spearman $\rho = -0.53$ over 19 model–arm profiles in four models (Indo-European languages

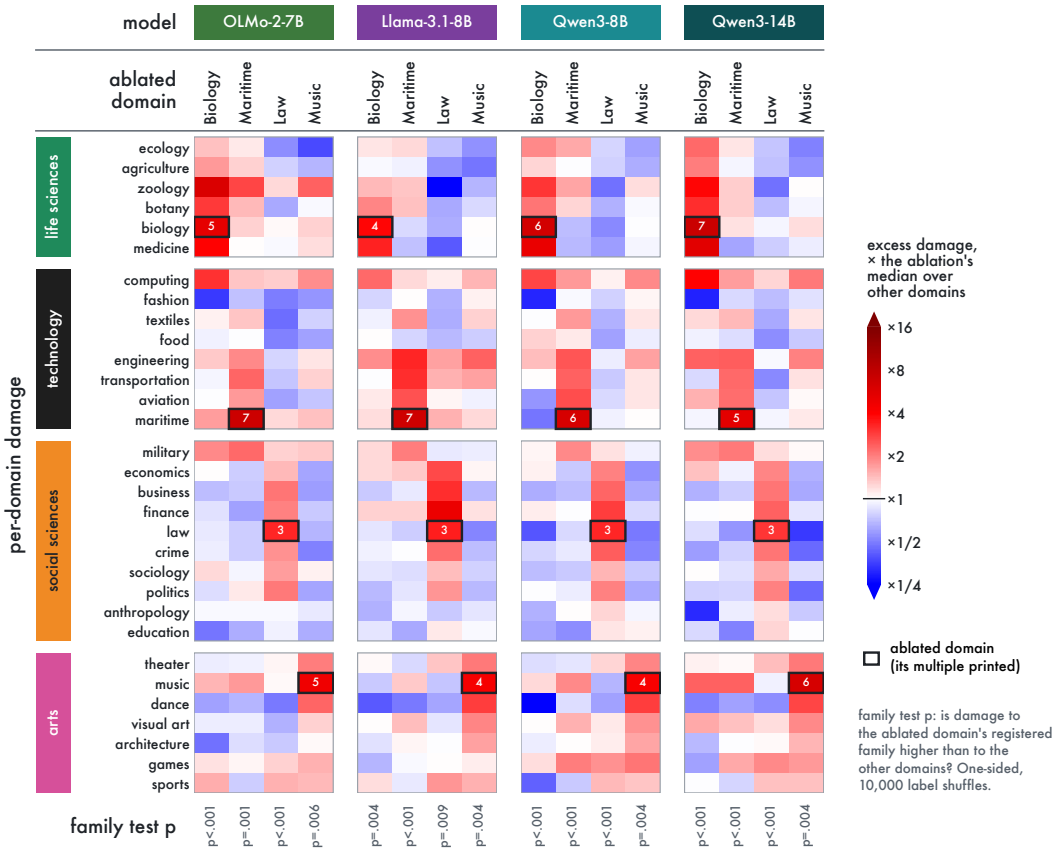

Figure 2: **Ablating one domain's heads damages related domains most.** For each model we remove the k attention heads most specific to one knowledge domain, chosen without training by PMI (k=228 for OLMo-2-7B and Llama-3.1-8B, 256 for Qwen3-8B, 356 for Qwen3-14B; shown domains: biology, maritime, law, music), and measure the loss on every evaluated domain (rows). Colour shows each cell's excess damage as a multiple of that ablation's median over the other 30 domains (rows: 31 of the 51 domains shown for clarity; log scale). Outlined cells are the ablated domain, with its multiple printed. Rows are grouped by Dewey Decimal class and ordered within each class by our Wikipedia link similarity (See Appendix A.2). **Bottom**: family test, i.e. the probability, under 10,000 random relabellings, that the ablated domain's neighbouring members (Section 3.1) are damaged more than the other domains.

of the 28-language set, 13 same-script targets per arm; within-arm permutation $p < .001$). For knowledge domains, excess damage falls with Wikipedia link-profile distance: pooled within-arm $\rho = -0.336$ over 77 model–arm profiles and 3,752 non-self cells in seven models from four families ($p < .001$).

5 Discussion

We introduced the Saussurean Representation Hypothesis, so named because of Saussure's central thesis: Representations within a system causally rely on their relationships to other representations within a network of representations (de Saussure, 1916). We have demonstrated that this holds true of modern LLMs in two quite different and broad domains, Languages and Knowledge-domains. In particular, similar, related languages are represented similarly in a model's representation space. And similar topics are likewise represented similarly. The learned Representations are distributed and overlapping, so that damaging any one representation impacts its neighbors. Targeted ablation in both testbeds consistently produced collateral damage beyond the target. Ablating the heads most informative about French damaged other Romance languages more than Germanic or Slavic ones;

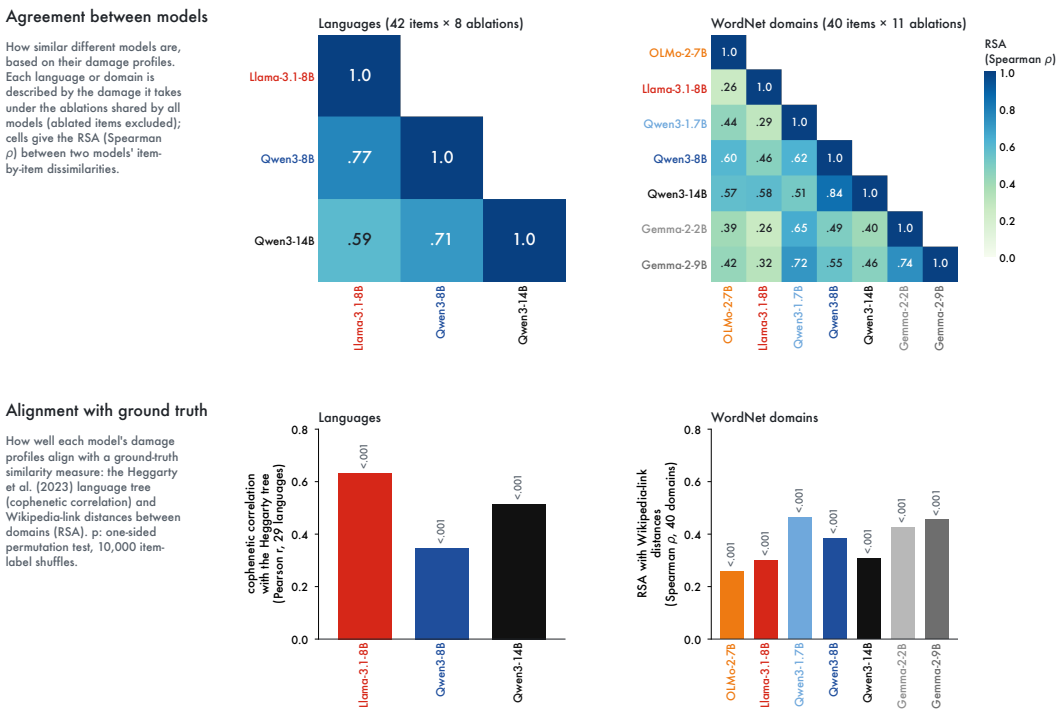

Figure 3: **Damage profiles agree across models and align with linguistic and semantic structure.** Top: RSA between models' damage-profile dissimilarities over languages and WordNet domains. Bottom: alignment with the Heggarty et al. (2023) tree (cophenetic correlation) and Wikipedia-link distances (RSA); all $p < .001$.

ablating biology damaged other life-science domains (e.g., botany, medicine) more than less related ones (e.g., economics, crime, politics).

The heads to be ablated for any given target were identified and ranked by a cheap, training-free approach. We merely calculated the point-wise mutual information (PMI) of all heads to identify those that were most informative for the target to be ablated, followed by a fixed-mean substitution at those heads.

The SRH allows us to recast existing approaches to interpretability and the study of representations in contemporary models. We show that even when considering very broad categories, the core structuralist assertion – that meaning is fundamentally relational – holds. Similar concepts end up with similar representations, not due to an artifact of training, or compression, or because of a shared objective reality, but because relations offer a constitutive account of meaning itself.

AI Use Statement

We used generative AI tools to assist with conceptual and methodological feedback, coding tasks, and editing parts of the manuscript. We did not use generative AI to generate data sets or to formulate the central hypothesis. The authors reviewed the AI-assisted material and take responsibility for the final methods, analyses, citations, quantitative claims, and conclusions.

Ethics Statement

The interventions studied here can degrade performance beyond their intended target, including on related languages and knowledge domains. This matters if similar techniques are used for model editing or unlearning. We report collateral effects against matched controls to make those limitations visible; our experiments do not establish that selective removal of a language or domain is safe or

achievable. The reported loss changes on evaluation corpora should not be equated with measured effects on deployed systems or their users.

Reproducibility Statement

Section 3 specifies the estimator, head ranking, intervention, matched controls and evaluation measures; Section 3.1 lists the checkpoints, corpora, categories and arms; the appendix details the regression, the Serbian and Azerbaijani evaluations, and the Wikipedia-link reference. The supplementary material contains the code for every stage, the derived data behind each reported number, and a script (`reproduce_numbers.py`) that recomputes the main results from those data.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ned Block. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10: 615–678, 1986.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8440–8451, 2020.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.08600.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Paris, 1916.

Melvil Dewey. *Dewey Decimal Classification and Relative Index*. OCLC Online Computer Library Center, Dublin, Ohio, 23rd edition, 2011.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025.

Martha J. Farah and James L. McClelland. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339–357, 1991.

John R. Firth. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, 1957.

Jerry Fodor and Ernest Lepore. *Holism: A Shopper's Guide*. Blackwell, 1992.

Jerry A. Fodor. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press, 1998.

Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 17643–17668, 2023.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, et al. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656):eabg0818, 2023.

Geoffrey E. Hinton and Tim Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

John R. Hodges, Karalyn Patterson, Susan Oxbury, and Elaine Funnell. Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain*, 115(6):1783–1806, 1992.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20739–20764. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of NAACL*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations (ICLR)*, 2025.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Andrew Joohun Nam, Henry Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, et al. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.13737.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Guilherme Penedo, Hynek Kydlíček, et al. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. *arXiv preprint*, 2025.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4996–5001, 2019.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. A survey of multilingual large language models. *Patterns*, 6(1), 2025.

Supantho Rakshit, Adele Goldberg, and Henry Conklin. Similarity all the way up: Multilingual generalization in LLMs relies on language-level similarity structures. *arXiv preprint arXiv:2607.22699*, 2026.

Taraka Rama, Lisa Beinborn, and Steffen Eger. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 1214–1228, 2020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Timothy T. Rogers and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press, 2004.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975.

Roger N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Joshua B. Tenenbaum and Thomas L. Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, 2001.

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 (ACL)*, 2019.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Anna Wierzbicka. *Semantics: Primes and universals*. Oxford University Press, UK, 1996.

Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pp. 833–844, 2019.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4836–4846, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.383. URL https://aclanthology.org.

A Appendix

A.1 Descent and script jointly modulate collateral damage under a given ablation

Contribution of shared descent and script Across five Indo-European arms common to all four models (French, German, Swedish, Russian, Polish) and the 28 evaluation languages, we classify the 135 non-self pairs per model by shared subfamily and shared script (Figure 4). Same-subfamily, same-script targets incur 3.62 times the damage from the floor of targets sharing neither in Qwen3-8B, 2.98 in Qwen3-14B, and 2.97 in Llama-3.1-8B. Pooled over models, the same-subfamily same-script contrast is $2.65\times$ ($p < .001$). Sharing descent without script also raises damage in the first three models, so neither factor is a necessary condition; the result is that joint similarity amplifies damage.

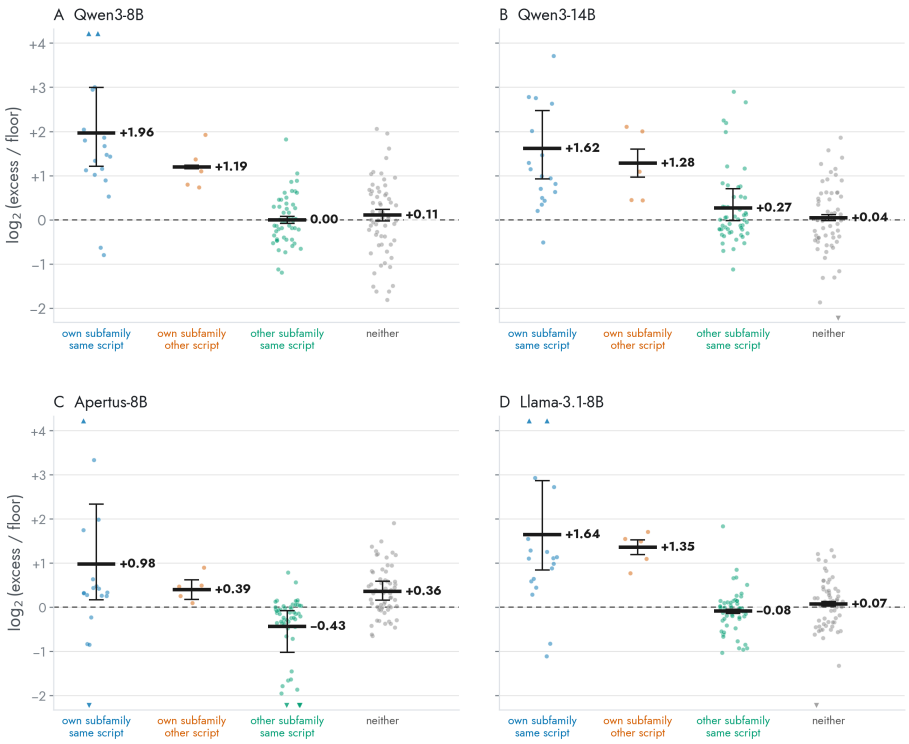

Figure 4: **Shared subfamily and script.** Five ablation arms common to all models (French, German, Swedish, Russian, Polish) and 28 evaluation languages, with the self cells removed, give the same 135 pairs per model: 18 in the own subfamily with the same script, 5 in the own subfamily with a different script, 52 in another subfamily with the same script, and 60 sharing neither. Each dot is one pair's damage from floor, $\log_2(\text{exc}/f_a)$, where f_a is the arm's median excess over languages outside its subfamily (clipped to $[-2, 4]$; triangles mark clipped pairs). Bars are the mean over arms of the arm-level class means, and error bars are 95% bootstrap intervals over arms.

A.1.1 Regression analysis

To test whether writing system changes the relation between ancestry and collateral damage, we regress damage on patristic distance, shared script, and their interaction. Random effects account for repeated ablation arms and evaluation languages. We fit the models separately to PMI excess damage. We exclude the target's own language and pairs outside the Indo-European reference. We fit the same crossed-intercept mixed model on the arm–language cells:

$$y_{aE} = \beta_0 + \beta_1 s_{aE} + \beta_2 d_{aE} + \beta_3 s_{aE} d_{aE} + u_a + v_E + \epsilon_{aE}, \tag{6}$$

where y_{aE} is the excess damage to evaluation language E under ablation arm a (%of E's baseline loss); $s_{aE} = 1$ if E is written in the same script as a and 0 otherwise; d_{aE} is the patristic distance between a and E on the Heggarty tree (Heggarty et al., 2023), in thousands of years; $u_a \sim \mathcal{N}(0, \sigma_a^2)$ absorbs differences in overall damage between arms, $v_E \sim \mathcal{N}(0, \sigma_E^2)$ absorbs differences in overall fragility between evaluation languages, and ϵ_{aE} is residual error. Both predictors are centred, so β_1 is the same-script effect at the mean distance, β_2 the distance slope at the mean script share, and β_3 the difference in distance slope between same-script and different-script targets. The descent $\times$ script conjunction predicts $\beta_3 < 0$: damage should fall with ancestral distance chiefly among languages that also share a script.

The descent-script interaction is negative in every fit (Table 1). Decomposing the PMI fits, excess damage to same-script targets falls by 0.67–1.88 percentage points per thousand years of divergence, whereas for different-script targets the slope is near zero (-0.16 to $+0.52$). The same-script advantage is correspondingly largest for close relatives (+7.8 to +46.5 points at 1,000 years) and shrinks

Table 1: Crossed-intercept mixed model (Eq. 6), y = excess damage (% of baseline loss), REML. Patristic distance in thousands of years; predictors centred. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Model	Script β_1 (SE)	Patristic distance β_2 (SE)	Script × patristic distance β_3 (SE)
Qwen3-8B	$+3.13$ (1.34)*	-0.33 (0.12)**	-1.60 (0.23)***
Qwen3-14B	$+1.51$ (0.84)	-0.33 (0.08)***	-0.59 (0.16)***
Apertus-8B	$+21.07$ (5.39)***	-0.66 (0.30)*	-2.40 (0.68)***
Llama-3.1-8B	$+15.50$ (2.47)***	-0.78 (0.17)***	-1.43 (0.33)***

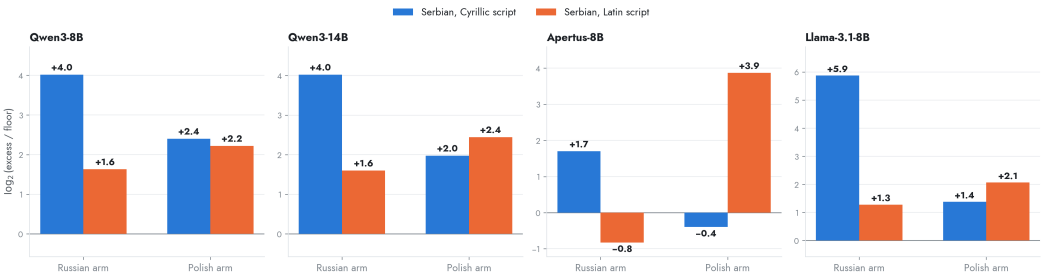

Figure 5: **Serbian digraphia.** Serbian text in Cyrillic and in Latin script, evaluated under the training-free ablation of the Russian (Cyrillic) and Polish (Latin) arms, $k = 256$. The Russian ablation damages Cyrillic Serbian more than Latin Serbian in all four models; the Polish ablation damages Latin Serbian more in three of four (Qwen3-8B is the exception, +2.2 vs. +2.4). Every Serbian cell remains above the random-head null.

with distance (+1.4 to +20.0 points at 12,000 years). Shared script therefore does not act alone: descent predicts damage within a script, and script predicts damage mainly among relatives. In Apertus-8B and Llama-3.1-8B a smaller script effect persists at large distances.

A.1.2 Script disambiguation by testing Serbian and Azerbaijani

These two evaluation-only cases further probe how the language damage profiles depend on both relatedness and writing system.

Serbian in two scripts. We evaluate Serbian text written in Cyrillic and in Latin script (both from FineWeb-2) under the Russian- and Polish-arm ablations based on PMI. Serbian is equally distant from Russian and Polish in the Indo-European reference tree (Heggarty et al., 2023), while Russian is written in Cyrillic and Polish in Latin script. The Russian ablation damages Cyrillic Serbian more than Latin Serbian in all four models; the Polish ablation damages Latin Serbian more in three of four (Qwen3-8B is the exception). Every Serbian cell remains above the random-head null. The result, therefore, shows that surface form does *modulate* damage to a related language, not that different scripts can prevent damage spillover from ablation of a related language, see Figure 5.

Azerbaijani under Turkish ablation. Turkish and Azerbaijani are closely related Oghuz Turkic languages, written in Latin script. The collateral degradation induced by Turkish on its one relative is 9–25× larger than a typical arm's collateral damage on its own subfamily, and well above even the strongest, Russian (+2.08 to +2.99). All the other ablation arms leave Azerbaijani near floor (mean $\log_2 -0.54$ to $+0.83$). Latin-script languages do not show this effect merely by sharing Turkish's script.

Together, these two cases support graded relatedness with an effect of written form, without treating shared script or shared descent as a strict necessary condition.

Table 2: Turkish damage to Azerbaijani compared to the mean own-subfamily effect for PMI based top-k head ablation with $k = 256$, reported in $\log_2(\text{excess}/\text{floor})$.

Model	Turkish → Azerbaijani	Mean Own-Subfamily Effect
Qwen3-8B	$+4.85$	$+1.63$
Qwen3-14B	$+5.92$	$+1.30$
Apertus-8B	$+4.49$	$+0.67$
Llama-3.1-8B	$+4.29$	$+1.12$

A.2 Derivation of the knowledge-domain reference tree

The 51 knowledge-domain labels come from WordNet categories and their descendants (Miller, 1995). WordNet's hierarchy classifies concepts by ontological type; it is not a measure of whether domains such as aviation and maritime draw on related bodies of knowledge. We therefore construct a model-independent reference for *topical proximity* from the outgoing links of the 5,769 Wikipedia articles in our domain corpus. These are the articles used to supply selection and evaluation text, so the reference is independent of the tested model, but not of the article collection.

For each source article we collect all outgoing main-namespace article links through the MediaWiki API, resolve redirects, and follow pagination to completion. We keep targets outside the 5,769-article corpus: only about 8% of the 1,909,483 collected links lead back into it. For each source article i, we divide its link counts by its total number of links. A domain profile p_A averages these normalized article profiles, giving every article in domain A equal weight. We then down-weight commonly linked destinations:

$$p_A(t) = \frac{1}{|A|} \sum_{i \in A} \frac{\ell_i(t)}{\sum_u \ell_i(u)}, \qquad w_A(t) = p_A(t) \log \frac{5769}{\mathrm{df}(t)},$$
$$d_{\text{link}}(A, B) = 1 - \frac{\sum_t w_A(t) w_B(t)}{\|w_A\|_2 \|w_B\|_2}. \tag{7}$$

Here $\ell_i(t)$ counts article i's links to destination t, and $\mathrm{df}(t)$ counts corpus articles linking to t. Two domains are close when their articles link to many of the same destinations; direct links between the two domains are unnecessary. The inverse-document-frequency factor prevents ubiquitous destinations from dominating this comparison.

We use average-linkage clustering (UPGMA) on d_{link} to draw the reference tree in Figure 6. Damage-gradient tests use the *distance matrix itself*. We did consider WordNet ancestor distance, links restricted to the corpus, two-hop reachability, and unweighted link-profile divergence. They respectively measure ontological rather than topical relatedness, discard most links, saturate, or overemphasize common destinations. The chosen reference is reliable (mean split-half Spearman 0.79 across random halves of each domain's articles; for 75% of domains the nearest neighbor lies in the same thematic group) and is independent of model activations. It remains a proxy for topical relatedness, not a uniquely correct semantic tree.

A.3 Additional figures

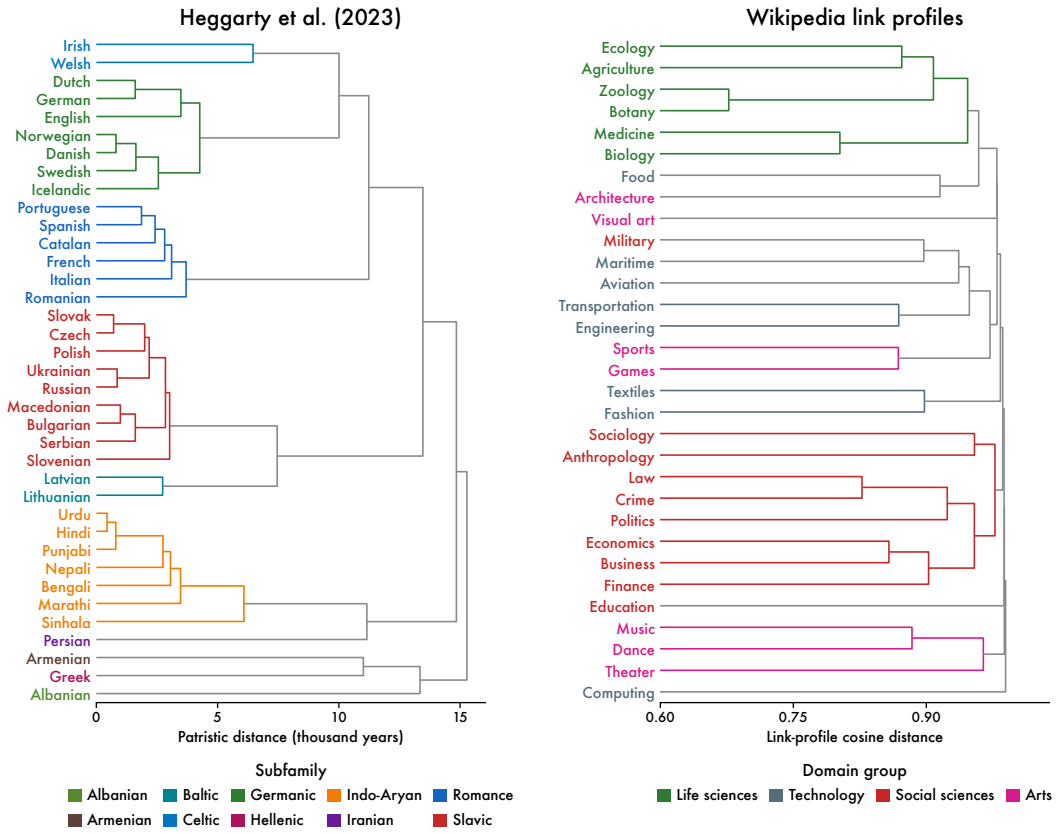

Figure 6: **Independent reference structures.** Left: the (Heggarty et al., 2023) Indo-European reference tree, with branch distances shown in thousands of years. Right: an average-linkage tree of the displayed knowledge domains, constructed from inverse-document-frequency-weighted Wikipedia outgoing-link profiles. The domain tree visualizes topical proximity, not WordNet ancestry. Quantitative comparisons use the underlying pairwise reference distances.

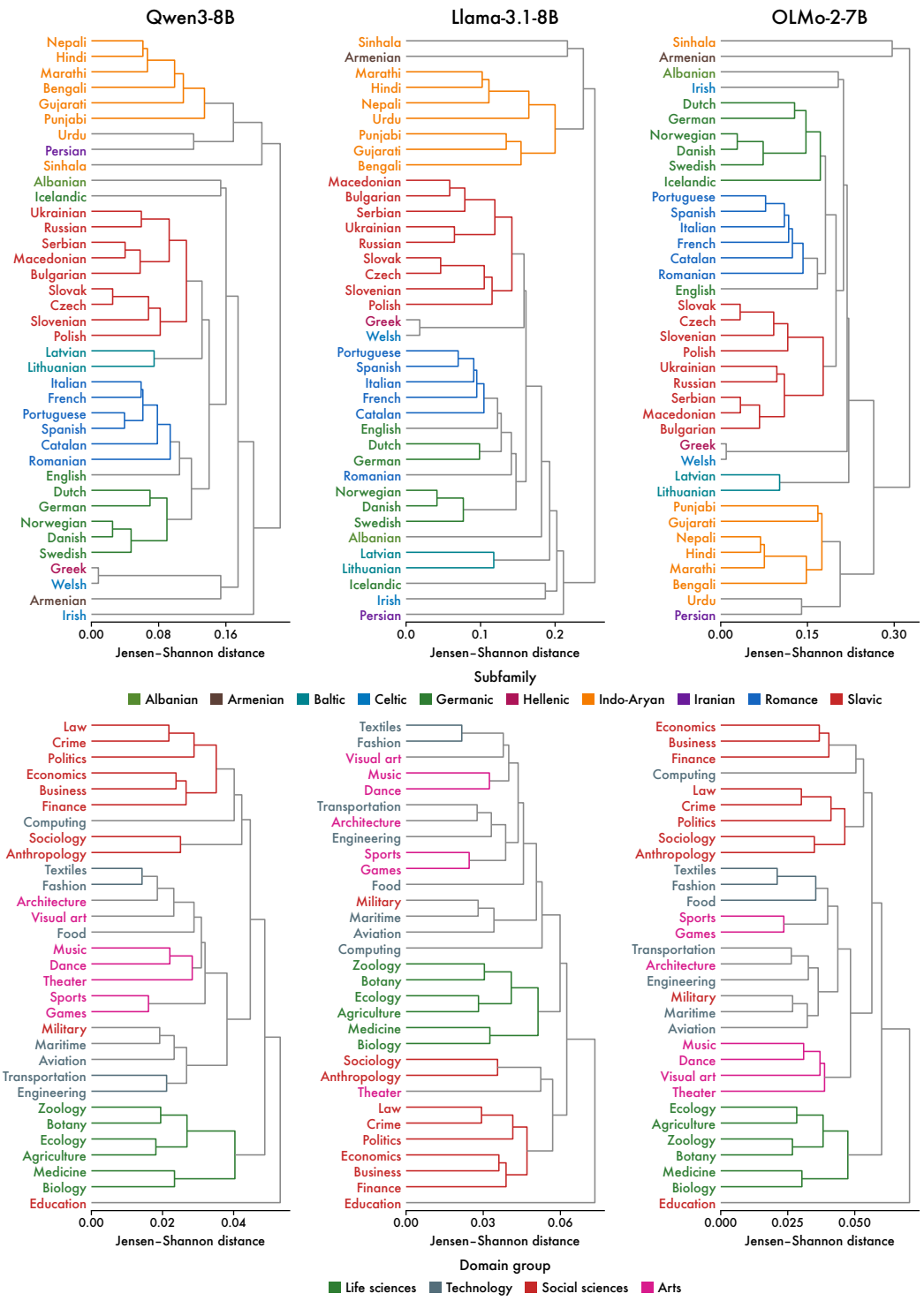

Figure 7: **Model-derived trees before intervention.** Per-layer residual-stream hidden states are softly assigned to $K = 3d$ random unit codewords; the Jensen–Shannon divergence between categories is averaged over layers and clustered with UPGMA. Top: 37 languages (Mantel test; $\rho = 0.64, 0.60, 0.65$). Bottom: 51 domains ($\rho = 0.69, 0.72, 0.72$ against Wikipedia link distance).

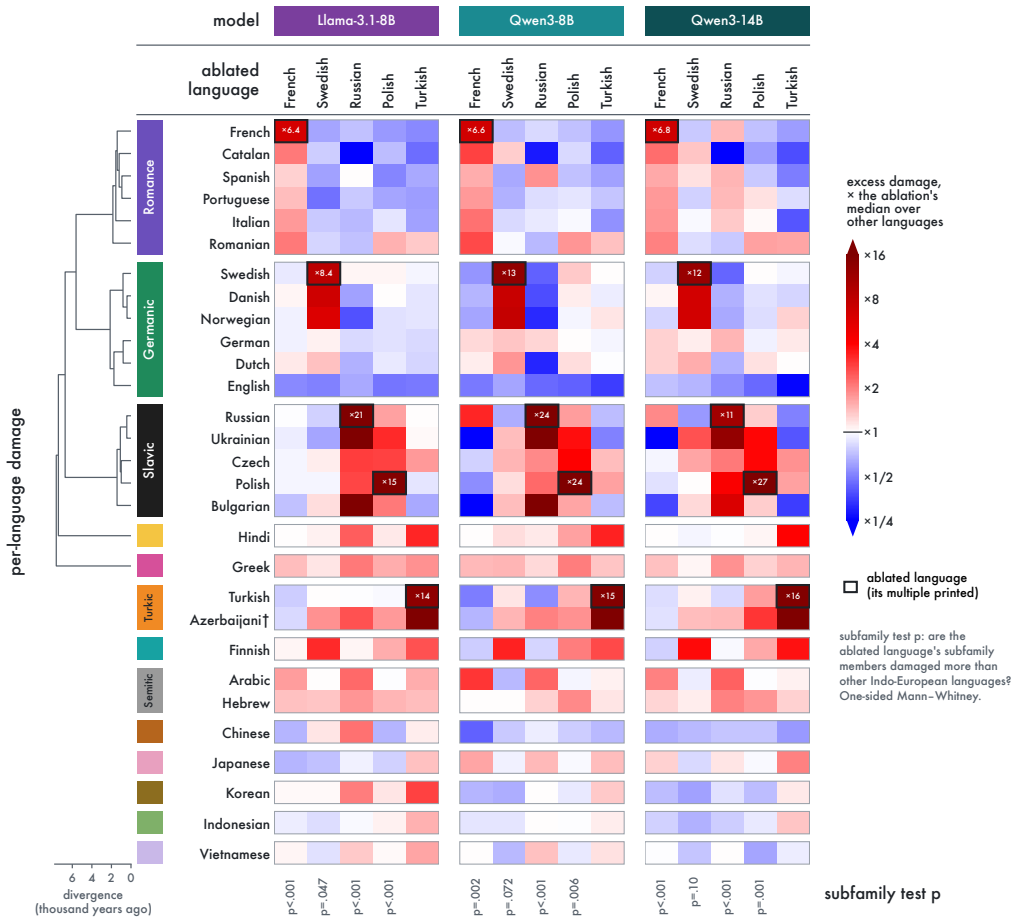

Figure 8: **Language ablation beyond the compact main-text display.** PMI-selected mean substitution for French, Swedish, Russian, Polish, and Turkish in three models, evaluated on the 28-language common set and evaluation-only Azerbaijani. Rows are arranged by linguistic family; outlined cells mark an arm's own language. The colours and within-arm subfamily tests follow Figure 1.

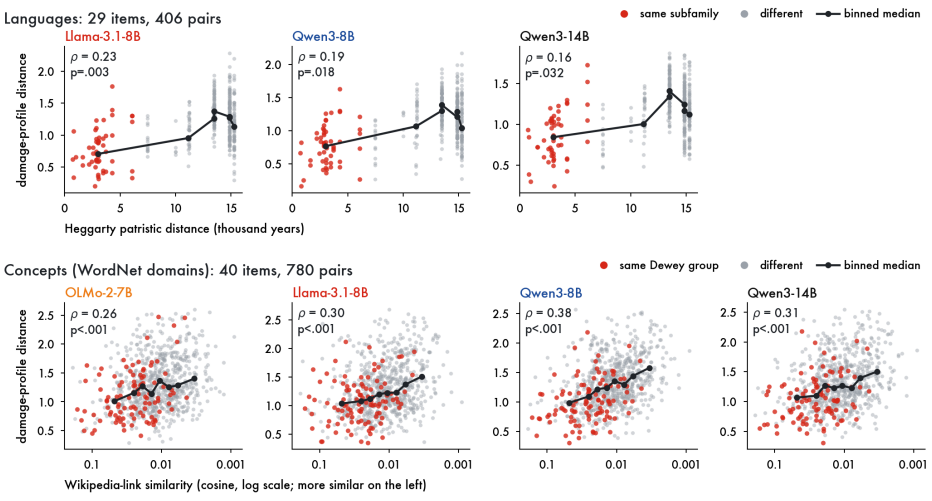

Figure 9: **Concept and Language Similarity Tracks Collateral Damage.** Items that are closer in ground truth dendrograms are damaged more. Each dot is a pair of languages (top; Heggarty et al. (2023) patristic distance) or WordNet domains (bottom; Wikipedia-link similarity, log scale), plotted against their PMI damage-profile distance. Red: same subfamily or Dewey group; line: binned median. Spearman ρ with one-sided permutation p.