On Representational and Functional Alignment of Training Data Attribution Methods for Language Models

Anonymous ACL submission

Abstract

Existing approaches to training data attribution leverage different sources of information, including loss-gradients, model-internal states, or surface-level similarity. Current evaluations of alignment between attribution methods are incomplete, often relying on indirect comparisons via downstream task performance. In this work, we propose a framework for directly comparing attribution methods along three dimensions: *functional alignment*, which measures overall correlation of attributions; *representational alignment*, which captures structural similarity of the underlying representations used for attribution; and *cross-method predictability*, which evaluates agreement on local, pairwise ranking decisions. We observe strong representational and moderate functional alignment within paradigms. Cross-paradigm alignment is low and task-dependent, yet stable across models. Our findings suggest that the similar downstream performance across paradigms observed in prior work may be driven by agreement on the relative importance of individual examples, rather than overall agreement on example relevance across the full training data.

1 Introduction

Training data attribution methods assign documents from a language model's training data scores proportional to their contribution to a given generation, to provide insights into model predictions. Three distinct attribution **paradigms** have proven popular, differing in the source of information they base attribution on: some leverage **loss gradients** (e.g., Pruthi et al., 2020; Koh and Liang, 2017), others **model-internal states** (Li et al., 2025), yet others only **lexical overlap** (e.g., Liu et al., 2025), assuming this still provides useful signals for understanding model behavior. For each paradigm, there exist multiple concrete **attribution methods** that instantiate and process this information in different ways. Prior work has largely neglected comparisons across methods and, when conducted, has compared them solely in terms of downstream task performance (e.g., Jiao et al., 2025b) or by simple rank correlation (Pezeshkpour et al., 2021). In this work, we argue that this practice generates an incomplete view of alignment between attribution methods of different paradigms. This is problematic because methods from one paradigm are often used as proxies for others of more expensive paradigms, as heuristics for benchmark construction, or as baselines in evaluations.

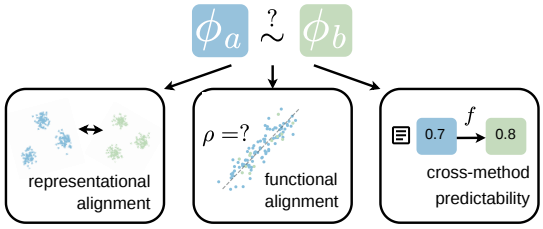

Figure 1: **Analytical Framework.** We directly compare attribution strategies by decomposing alignment into (1) similarity of underlying representations, (2) correlation of attributions, and (3) agreement on pairwise rankings.

We propose a framework for directly comparing the alignment of attribution methods that addresses three main limitations: First, existing comparisons between attribution methods depend heavily on downstream task mechanics, limiting their transferability. For example, methods that work well for question answering may fail to capture aspects relevant to instruction tuning because of differences in learning objectives. Second, current work only considers overall correlation across attributions, which does not adequately reflect functional equivalence in many downstream applications, where agreement on relative relevance judgments is sufficient. Third, existing approaches do not separate representational structure from downstream ranking decisions, making it difficult to determine whether differences between methods arise from how information is represented or from how it is processed.

We accordingly decompose alignment into three dimensions: *functional alignment*, i.e., overall correlation of attributions; *representational alignment*, i.e., structural similarity of the underlying document representations used for attribution; and *cross-method predictability*, i.e., agreement in pairwise relevance decisions, capturing local alignment rather than agreement on example relevance across the full training data. We leverage this framework for a systematic study of 12 attribution methods across four paradigms[1], 5 tasks, and 7 models spanning three model families. Specifically, we provide insights into: (i) the alignment of attribution methods across different language modeling tasks, (ii) the stability of alignment across datasets and base models, (iii) whether methods can be employed as proxies for others, and (iv) factors that impact alignment at the instance level. By providing a framework for directly comparing attribution methods across paradigms, we aim to lay the groundwork for attribution pipelines that combine multiple attribution strategies, whether to make attribution tractable for LLM pre-training corpora or to integrate orthogonal perspectives on data influence.

We find that methods within the same attribution paradigm tend to exhibit high representational but moderate functional alignment. Alignment across paradigms is overall low and highly task-dependent, but consistent across model families and sizes. Our findings suggest that the relatively consistent downstream performance across paradigms in prior work may be driven by agreement on the relative importance of individual examples, rather than overall agreement on example relevance across the full training data. In a qualitative analysis, we find that cross-paradigm alignment appears to be shaped by a mix of prompting artifacts, lexical overlap, and task-dependent structural and behavioral similarities. Overall, our results suggest that methods within a paradigm are partially interchangeable, but that different paradigms capture distinct, only partially aligned notions of data influence, suggesting that only a combination of different sources of information may yield a complete characterization of data influence. Our main contributions are:[2]

(1) A framework for measuring representational and functional alignment.
(2) A test of cross-method predictability that assesses local rather than overall alignment.
(3) A systematic study of alignment across 7 base models, 5 tasks, and 12 attribution methods.
(4) A qualitative analysis of factors of alignment.

[1] gradient-based attribution, model-internal representations, embedding similarity, and lexical overlap

[2] We release our code at anonymous.4open.science/r/alm.

2 Related Work

Indirect evaluation of alignment. Prior work predominantly compares attribution methods based on their benchmark performance on downstream tasks, such as hallucination tracing (Wu et al., 2024), fact tracing (Akyurek et al., 2022), or data cleaning (Pezeshkpour et al., 2021; Jiao et al., 2025a; Yu et al., 2024; Han et al., 2020). Work on example-based explanation compares in terms of application-specific properties such as stability (Gu et al., 2023), faithfulness, or plausibility (Hanawa et al., 2020). Likewise, work on gradient-based attribution compares methods directly with ground-truth measures of influence obtained in training-based experiments (Bae et al., 2022; Grosse et al., 2023; Park et al., 2023; Mlodozeniec et al., 2024; Hong et al., 2025; Ilyas and Engstrom, 2025; Sun et al., 2026). All of these approaches only quantify whether two methods perform similarly on average. They do not consider (1) whether different methods behave similarly on the same instances (which we cover through cross-method predictability and functional alignment), and more generally (2) whether behavior is driven by similar information (which we capture via representational alignment).

Functional alignment across methods. Prior work that *directly* compares attribution methods has characterized alignment only in terms of correlation or set overlap among the top influential training examples, which corresponds to our definition of functional alignment. However, existing comparisons vary substantially in scope: some compare different implementations of the same attribution paradigm (Nguyen et al., 2023; Brophy et al., 2024). Others evaluate the agreement of gradient-based methods with estimates derived from non-traditional attribution strategies, such as in-context probing (Jiao et al., 2025a) and neuron attribution (Yu et al., 2024). Our work is most aligned in scope with Pezeshkpour et al. (2021), who compare embedding similarity with 4 gradient-based attribution methods using BERT models. They find that gradient- and embedding-based methods correlate highly with more expensive Hessian-based methods, and observe only modest correlation between similarity- and gradient-based methods.

We extend this line of work by introducing

Gradient-based representations	
$r^{\mathrm{id}\,x}_{\nabla f}$	Raw model loss gradients
$r^{L2}_{\nabla f}$	L2-normalized loss gradients
$r^{ad}_{\nabla f}$	Adam-adjusted gradients (Xia et al., 2024)
$r_{s_{\text{test}}}$	Hessian-adjusted gradients
Model-internal representations	
r_{repsim}	Final-layer last token hidden state (Li et al., 2025)
r_{mp}	Mean-pooled final-layer hidden states
r_{attn}	Summary of normalized attention maps
Embedding similarity	
qwen	Embeddings of Qwen/Qwen3-Embedding-8B
nemot	Embeddings of nvidia/llama-embed-nemotron-8b
Lexical overlap	
r_{bm25}	Vector of BM25-based relevance scores
r_{jaccard}	Vector of Jaccard similarity scores
r_{tfidf}	Vector of TF-IDF-based similarity scores

Table 1: Summary of attribution methods studied.

two additional perspectives on alignment, for the first time capturing representational alignment and cross-method predictability. We further consider new attribution methods from additional paradigms, new tasks, and more modern model architectures.

3 Methodology and Experimental Setup

We first describe 3 perspectives on alignment and define corresponding metrics. We then integrate different attribution methods under a common evaluation framework, before describing experiments targeting the following main questions: (1) *How aligned are methods of different paradigms?* (2) *Is alignment consistent across models and tasks?* (3) *What drives alignment at the instance level?*

3.1 Measuring Representational Alignment

We aim to capture the similarity between similarity structures induced by different representations used for training data attribution. For this, we adapt methodology from mechanistic interpretability, originally proposed for studying the alignment of model-internal representations. Following Huh et al. (2024), we define representations as functions $r : \mathcal{Z} \to \mathbb{R}^d$ that encode individual training documents as vectors. For illustration, consider two representations r_i and r_j, implemented via two text embedding models (we introduce additional representations in Section 3.4). Given a similarity kernel defined by an inner product over representations, we apply two alignment metrics: Centered Kernel Alignment (Kornblith et al., 2019) yields a measure of *global representational alignment*.

$$\mathrm{CKA}(r_i, r_j) = \frac{\|r_i^\top r_j\|_F^2}{\|r_i^\top r_i\|_F \, \|r_j^\top r_j\|_F}.$$

To assess *local representational alignment*, we calculate Centered Kernel Nearest-Neighbor Alignment (CKNN), which restricts this comparison to the k nearest neighbors (Huh et al., 2024).

3.2 Measuring Functional Alignment

We define functional alignment as the average overall agreement between two attribution methods across test instances. Specifically, given two vectors of pointwise relevance scores over the training data $\phi_{r_i}(z_{\text{test}}), \phi_{r_j}(z_{\text{test}}) \in R^{|\mathcal{D}|}$, we quantify agreement using Spearman's ρ, Kendall's τ, and rank-biased overlap (RBO, prioritizes agreement at the top of ranked lists). All metrics are computed per test instance and then averaged.

3.3 Measuring Cross-Method Predictability

Methods may diverge in their global rankings but still agree on the relative importance of individual examples. We therefore further test whether attributions from one method can be leveraged to emulate those of another method. Specifically, we propose a test based on pairwise comparisons: we evaluate whether the attribution scores from method i can act as features to successfully predict the binary pairwise rankings of method j. Let $s^{(i)}_{z_{test}}(a)$ denote the attribution score by method i for document a given z_{test}. For each z_{test}, we construct a set of 2000 random pairs $(a \neq b)$, subsequently split into train and test sets (80/20). We construct features based on score differences of method i, and labels based on the binary ranking of method j:

$$x_{ab,z_{test}} = s^{(i)}_{z_{test}}(a) - s^{(i)}_{z_{test}}(b)$$
$$y^{(j)}_{ab,z_{test}} = \mathbb{I}\big[s^{(j)}_{z_{test}}(a) > s^{(j)}_{z_{test}}(b)\big].$$

To quantify how well rankings produced by method j can be predicted using features derived from method i (*cross-method predictability*), we report the mean held-out accuracy of instance-specific logistic regression models without intercept $P\Big(y^{(j)}_{ab,z_{test}} = 1\Big) = \frac{1}{1+\exp\big(-w_{z_{test}}\tilde{x}_{ab,z_{test}}\big)}$.

3.4 Representations

We consider attribution methods that span four popular paradigms: *gradient-based attribution*, *model-internal representations*, *embedding similarity*, and *lexical overlap*. See Table 1 for a summary. To enable the study of representational alignment, we require representation functions $r : \mathcal{Z} \to \mathbb{R}^d$, so that documents can be compared via the inner product of their representations (Section 3.1). For the

study of functional alignment, we obtain point-wise scores $\phi(z_{\text{test}}, z)$ by dot product between $r(z_{\text{test}})$ and the representation of the training example $r(z)$ for all methods except $r_{s_{\text{test}}}$ (for which attributions are given by Equation 1). Note that there exist multiple valid definitions of relevance for gradient-based methods: one can retrieve either the *most* or *least influential* samples (i.e., those with the largest or smallest absolute similarity scores) or the *most helpful* or *harmful* samples (those with the most positive or negative scores). We treat each strategy as a distinct attribution method across experiments, but include only the *most helpful* strategy in the plots for simplicity. We select this variant to ease interpretability, given that it is the closest gradient-based analog to the semantic and lexical-similarity-based retrieval logics (examples supporting the test instance that increase loss if removed).

Gradient-based representations. Common to several gradient-based strategies is the definition of influence as the similarity between test-instance gradients and training-instance gradients (Hammoudeh and Lowd, 2024). We define three representations based on formulations in prior work:

$$r^{id_X}_{\nabla f}(z) = \nabla \mathcal{L}(z, \theta)$$

$$r^{L2}_{\nabla f}(z) = \frac{\nabla \mathcal{L}(z, \theta)}{\|\nabla \mathcal{L}(z, \theta)\|_2} \qquad r^{aa}_{\nabla f}(z) = aa(\nabla \mathcal{L}(z, \theta))$$

Representation $r^{id_X}_{\nabla f}$ is based on raw gradients, as used for estimation by Pruthi et al. (2020). $r^{L2}_{\nabla f}$ represents unit-norm gradients, as utilized in Yeh et al. (2022). In $r^{aa}_{\nabla f}$, we use *Adam-adjusted* (normalized) gradients, as in Xia et al. (2024).

Next, we consider point-wise influence via *influence functions* as in Koh and Liang (2017):

$$IF(z, z_{\text{test}}) = -\underbrace{H^{-1}_{\hat{\theta}} \nabla \mathcal{L}(z_{\text{test}}, \hat{\theta})}_{s_{\text{test}}} \cdot \nabla \mathcal{L}(z, \hat{\theta}). \quad (1)$$

Conceptually, s_{test} incorporates second-order information on how changes in parameters affect test loss. To study the effect of this re-weighting, we define the following representation. For implementation details, see Appendix E:

$$r_{s_{\text{test}}}(z) = H^{-1}_{\hat{\theta}} \nabla_\theta L(z, \hat{\theta}).$$

Model-internal representations. We define three representations based on internal model states r_{repsim}, r_{attn}, and r_{mp}. We define r_{repsim} based on the influence estimation strategy in Li et al. (2025):

$$r_{\text{repsim}}(z) = H^{(L)}_{T,:},$$

where $H^{(L)} \in \mathbb{R}^{n \times d}$ denotes the hidden states of the final transformer layer for an input sequence z, and T is the index of the last non-padding token.

We obtain r_{mp} by mean pooling over the non-padding tokens in the final layer. Let $m \in \{0, 1\}^n$ specify the attention mask, then:

$$r_{\text{mp}}(z) = (m^\top H^{(L)})/(\mathbf{1}^\top m).$$

Next, r_{attn} is based on attention patterns. Here, documents are considered similar if they induce similar *distributions* of attention values. Kobayashi et al. (2020) define a token-to-token interaction map $M \in \mathbb{R}^{n \times n}$, $M_{i,j} := \|\sum_h \alpha^h_{i,j} f^h(x_j)\|$, which we adapt to support grouped-query attention (GQA) utilized by Qwen models here (Appendix F). To obtain sequence-length invariant vectors suited for comparison across documents, we pad to the maximum sequence length S, and then compute the singular value decompositions $M_l = U \Sigma V^\top, \Sigma_l = \text{diag}(\sigma^{(l)}_1, \ldots, \sigma^{(l)}_S)$. The attention-based representation is the concatenation of the top $k = 128$ singular values across all layers:

$$\mathbf{r}_{\text{attn}}(z) = \left[\sigma^{(1)}_1, \ldots, \sigma^{(1)}_k, \ldots, \sigma^{(L)}_1, \ldots, \sigma^{(L)}_k\right].$$

Embedding similarity-based representations. We define two representations on embedding similarity using external models, r_{nemot}: (nvidia/llama-embed-nemotron-8b: Babakhin et al., 2025) and r_{qwen}: (Qwen/Qwen3-Embedding-8B: Zhang et al., 2025), selected based on performance on the MMTEB benchmark (Enevoldsen et al., 2024).

Representations based on lexical overlap. Recent work (e.g., Liu et al., 2025) retrieves lexically related training examples to address LLM-scale datasets. We therefore define a set of representations in which two documents will be assigned high similarity scores if they are lexically related to the same documents in the training data. Our approach is inspired by the representation strategy in Lagus et al. (2019), who show that by representing documents as matrices $A \in \mathbb{R}^{n \times d}$, where each row corresponds to a word vector, their similarity can be measured in terms of the squared Frobenius norm of their pairwise word similarity matrices ($\|AB^\top\|^2_F$). This approach compares documents based on the covariance of word vectors. To represent each document in terms of its similarity to other documents in the training data while allowing comparisons via inner-product similarity, we first generate rankings for a given test instance

z_{test} using Jaccard similarity (for lexical overlap), TF-IDF (weighted lexical similarity), and BM25 (probabilistic relevance). Let d_i be the i-th training example, and $\phi(z, z_i)$ denote the relevance of z_i for z. Then r_{bm25}, r_{tfidf} and $r_{jaccard}$ are of the form

$$r_\phi(z) = \big[\phi(z, z_i), \phi(z, z_i), \ldots, \phi(z, z_N)\big].$$

3.5 Modeling Alignment Across Paradigms

We cluster representations using agglomerative hierarchical clustering based on pairwise alignment scores to characterize alignment structure across attribution methods. Let y_{ij} denote the alignment score (e.g., CKA) between representations i and j. We consider a pseudo-dissimilarity measure $d_{ij} = 1 - y_{ij}$ and perform agglomerative hierarchical clustering of representations using average linkage (UPGMA: Sokal and Michener, 1958). This yields one hierarchical clustering per alignment score, which we visualize as dendrograms.

Stability of clusterings. To assess whether our method groupings are stable across different base models and tasks, and not a product of random variation, we compare dendrograms using cophenetic correlation (Sokal and Rohlf, 1962). For a given clustering, let c_{ij} denote the cophenetic distance between representations r_i and r_j, defined as the dissimilarity value at which r_i and r_j are first merged into the same cluster. Given two clusterings over a set of representations, the cophenetic correlation between dendrograms following Sokal and Rohlf (1962) is $\rho_{\text{coph}}(A, B) = \rho\big(\{c^A_{ij}\}_{i<j}, \{c^B_{ij}\}_{i<j}\big)$. We measure the average stability across factors

$$\bar{\rho}_{d,m} = \frac{1}{\binom{|\mathcal{D}|}{2}\binom{|\mathcal{M}|}{2}} \sum_{d_1<d_2} \sum_{m_1<m_2} \rho_{\text{coph}}\big(C^{d_1,m_1}, C^{d_2,m_2}\big),$$

as well as two conditional averages: (i) fixing the dataset, averaging over base models,[3] and (ii) fixing the base model, averaging over datasets. Lastly, we measure the pairwise average cophenetic stability between individual alignment measures.[4]

3.6 Modelling Factors Affecting Alignment

To assess the robustness of our analysis, specifically whether alignment scores are sensitive to the choice of base model and task, we fit linear mixed-effects models and assess contributions of both factors (and their interaction) via likelihood ratio tests over nested specifications. The full model is given by:

$$\mathcal{M}_{\text{full}} : \begin{array}{c} y \sim C(b) * C(d) \equiv \\ y_{ijbd} = \beta_0 + \beta_b + \beta_d + \beta_{bd} + u_{ij} + \epsilon_{ijbd}, \end{array}$$

where y_{ijbd} denotes the alignment score for representation pair ij, given base model b and dataset d. We assume $u_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{pair}})$, $\epsilon_{ijbd} \sim \mathcal{N}(0, \sigma^2_\epsilon)$. We test for interaction effects via a likelihood ratio test between $\mathcal{M}_{\text{full}}$ and $\mathcal{M}_{\text{add}} : y \sim C(b) + C(d)$ corresponding to $H_0 : \beta_{bd} = 0$. We likewise test the effects of base model and dataset via likelihood ratio tests between full and reduced models ($\mathcal{M}_{-B} : y \sim C(d)$, $\mathcal{M}_{-D} : y \sim C(b)$), corresponding to $H_0 : \beta_b = 0$ and $H_0 : \beta_d = 0$. We test whether each fixed effect improves model fit via likelihood ratio tests and quantify effect sizes as the reductions in marginal and conditional R^2.[5]

3.7 Qualitative Comparison of Local Neighborhoods Across Representations

The proposed measures quantify alignment but do not identify sources of agreement or disagreement at the instance level. We therefore complement them with instance-level qualitative analysis to better understand what drives misalignment. Given two representations r_i, r_j and a document z, we retrieve the top-5 most similar examples to z under each representation. For each dataset and representation pair, we compare the resulting local neighborhoods in two settings: (i) when the neighborhoods diverge most, and (ii) when they align best (see Appendix D for details). We restrict this analysis to 4 of the 12 representations, selecting a representative subset using a greedy farthest-first procedure on the CKA dissimilarity matrix. We inspect 60 pairs across the Olmo3 models.

3.8 Models and Data

We fine tune Qwen3-Base (Yang et al., 2025) at sizes 0.6,1.7,4B,8B; Olmo-3-1025-7B (Ettinger et al., 2025); OLMo-2-0425-1B (Walsh et al., 2025); and Llama-3.2-1B (Dubey et al., 2024) for one epoch using LoRA (Hu et al., 2021) on the following datasets (for details see Appendix G):
SFT: the instruction-tuning dataset for OLMo2 based on *Tülu3* (allenai/tulu-3-sft-olmo-2-mixture-0225: Lambert et al., 2025);

[3] $\bar{\rho}_d = \frac{1}{\binom{|\mathcal{M}|}{2}} \sum_{m_1<m_2} \rho_{\text{coph}}\big(C^{d,m_1}, C^{d,m_2}\big).$

[4] $\bar{\rho}_{a_1,a_2} = \frac{1}{\binom{|\mathcal{D}||\mathcal{M}|}{2}} \sum_{d_1,m_1<d_2,m_2} \rho_{\text{coph}}\big(C^{d_1,m_1}_{a_1}, C^{d_2,m_2}_{a_2}\big)$

[5] $R^2_m = \frac{\sigma^2_f}{\sigma^2_f+\sigma^2_{\text{pair}}+\sigma^2_\epsilon}$, $R^2_c = \frac{\sigma^2_f+\sigma^2_{\text{pair}}}{\sigma^2_f+\sigma^2_{\text{pair}}+\sigma^2_\epsilon}$, where σ^2_f and σ^2_{pair} refer to variance explained by fixed- and random-effects (Nakagawa and Schielzeth, 2013).

math(allenai/tulu-3-sft-personas-math), and **code**: subsets of *Tülu3* targeting math and coding problems (allenai/tulu-3-sft-personas-code);
squad_v2: reading comprehension for extractive question answering (Rajpurkar et al., 2018);
ftrace: benchmark dataset that perturbs documents to increase attribution difficulty (Jiao et al., 2025b).

We fine-tune 35 models on the full datasets and measure alignment across 1000 random documents per dataset and model combination.

4 Results

We report pairwise representational alignment (CKA), functional alignment (Spearman's ρ), and cross-method predictability averaged across models and tasks in Figure 3, for Kendall's τ, RBO, per model- and per task results see Appendix C.

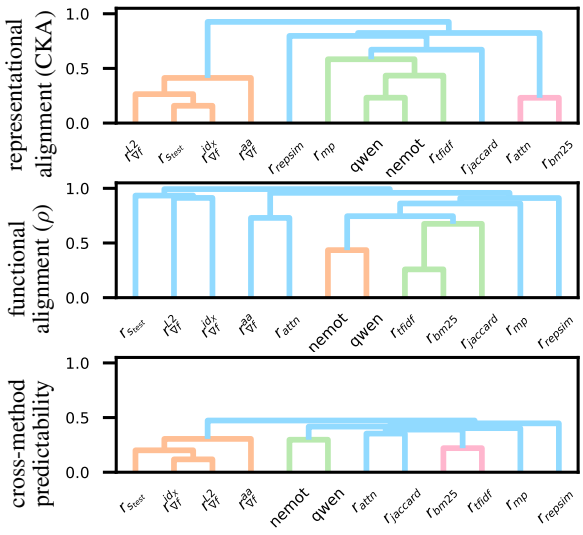

Figure 2: **Hierarchical clustering of attribution methods.** Clusters largely align with attribution paradigms.

4.1 Alignment Across Paradigms

Regarding our first question (*how aligned are methods from different paradigms?*), we find that attribution methods cluster by their underlying paradigms, exhibiting high internal agreement (e.g., $r^{L2}_{\nabla f}$ vs. $r^{idx}_{\nabla f}$) but low cross-paradigm alignment (e.g., $r^{L2}_{\nabla f}$ vs. r_{mp}). This is reflected in the dendrograms in Figure 2, which show clear separation between gradient-based, embedding-based, and lexical similarity methods. Overall, we observe clear within-paradigm alignment, as evidenced by a pronounced block-diagonal structure in the heatmaps of Figure 3 and Appendix C. This is most pronounced for representational alignment (CKA, CKNNA@25) and cross-method predictability within gradient-based ($r_{s_{\text{test}}}$, $r^{aa}_{\nabla f}$, $r^{L2}_{\nabla f}$, $r^{idx}_{\nabla f}$) and embedding-based attribution paradigms (r_{nemot}, r_{qwen}). For functional alignment (ρ, τ, RBO), however, this pattern only holds for embedding-based methods. One notable exception is the pairing of BM25 and the attention-based method, which features relatively high representational alignment (yet still weak functional alignment and low cross-method predictability). Despite their negligible functional alignment, we measure high representational alignment and cross-method predictability between gradient-similarity methods and the Hessian-based $r_{s_{\text{test}}}$. Per-task predictability results (Appendix C) show substantial variation within the same paradigm, for example, alignment between $r^{aa}_{\nabla f}$ and $r_{s_{\text{test}}}$ is about 11-20 pt lower on *SFT* (0.56) than on other tasks. Alignment across paradigms, however, appears largely invariant to task.

Stability of clusterings. The average cophenetic stability $\bar{\rho}_{d,m}$ is high across metrics (CKA: 0.80, CKNNA@25: 0.75, ρ: 0.71, τ: 0.75, RBO: 0.72, acc_{pred}: 0.79), indicating that dendrogram structure is largely preserved across conditions. Stability across base models $\bar{\rho}_b$ and datasets $\bar{\rho}_d$ is consistently high as well (> 0.7, Appendix B). Regarding cophenetic stability between alignment metrics (CKA, ρ, etc.), $\bar{\rho}_{a_1,a_2}$, we find high agreement between measures targeting the same quantity (representational alignment, functional alignment, or cross-method predictability). Measures targeting representational alignment (CKA, CKNNA@25) show strong consistency ($\bar{\rho}_{a_1,a_2} \approx 0.70$), while functional alignment measures (Kendall's τ, RBO, Spearman's ρ) also exhibit moderate-to-high agreement ($\approx$ 0.57–0.61). In contrast, correlations across metrics targeting different quantities (e.g., CKA vs. ρ) are low ($\approx$ 0.02–0.30). Weak correlation to overall functional alignment confirms that cross-method predictability captures a distinct notion of alignment.

4.2 Factors Affecting Alignment.

To answer our second question (*is alignment consistent across models and tasks?*), we find that alignment patterns (rather than the clusterings discussed earlier) are broadly stable with respect to base model and dataset choice: while the choice of dataset and base model yield statistically significant effects in our likelihood-ratio tests (Appendix A), their practical effect sizes (ΔR^2, Table 2) are minor. This demonstrates that the specific alignment

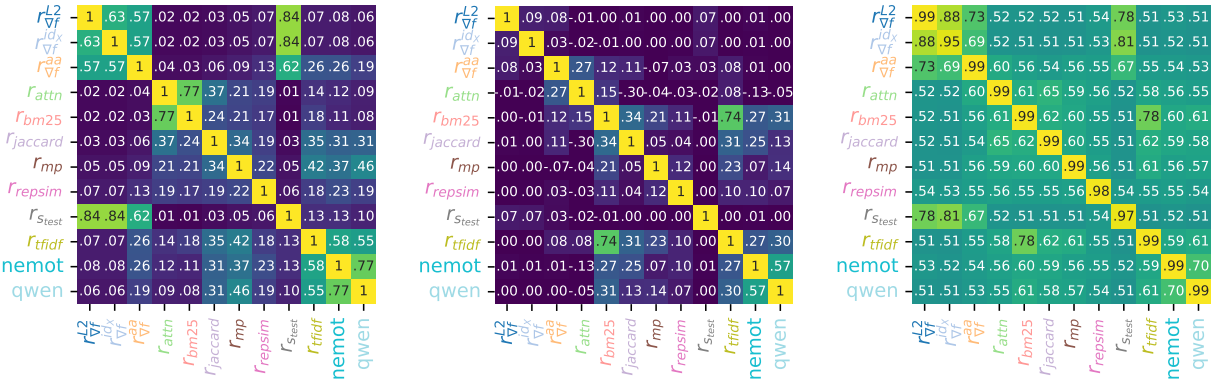

Figure 3: **Pairwise alignment.** Left: representational alignment (CKA $\in [0, 1]$), center: functional alignment (Spearman's $\rho \in [-1, 1]$), and right: cross-method predictability (acc $\in [0, 1]$) averaged across tasks and models. Higher values suggest better alignment. For CKNNA@25, RBO, and Kendall's τ, see Appendix C.

Term →	$\Delta R^2_m(\mathcal{M}_{-B})$	$\Delta R^2_c(\mathcal{M}_{-B})$	$\Delta R^2_m(\mathcal{M}_{-D})$	$\Delta R^2_c(\mathcal{M}_{-D})$	$\Delta R^2_m(\mathcal{M}_{add})$	$\Delta R^2_c(\mathcal{M}_{add})$
CKA	9.88e-03	1.01e-02	6.89e-03	7.07e-03	2.25e-03	2.31e-03
CKNNA@25	4.64e-03	4.77e-03	1.33e-02	1.36e-02	1.59e-03	1.64e-03
τ	5.16e-05	5.30e-05	1.21e-04	1.24e-04	1.29e-05	1.32e-05
acc_{pred}	6.74e-03	6.92e-03	7.67e-03	7.87e-03	3.72e-03	3.82e-03
RBO	1.04e-04	1.05e-04	2.89e-04	2.94e-04	4.95e-05	5.02e-05
ρ	9.74e-05	9.96e-05	2.07e-04	2.12e-04	2.40e-05	2.45e-05

Table 2: **Ablation analysis.** Reductions in marginal and conditional R^2 (ΔR^2_m, ΔR^2_c) when removing base model-, dataset-, or interaction terms from the full mixed-effects model. Although likelihood-ratio tests indicate several significant terms, the effect sizes are small, suggesting that our analysis is robust to changes in the base model and task.

properties (or lack thereof) between paradigms are fundamental, rather than artifacts of a particular experimental setup. Among these secondary factors, removing the dataset factor yields the largest decrease in explained variance across all metrics, except CKA. Conversely, the base model choice shows a stronger effect only on global representational alignment (CKA), but not on its local counterpart (CKNNA@25). Interaction terms between the base model and dataset contribute comparatively little additional explanatory power.

4.3 Qualitative Insights on Alignment

To address our final question (*what drives alignment at the instance level?*), we manually compare local neighborhoods of different representations to understand the qualitative sources of agreement and disagreement. We examine whether local neighborhoods differ across attribution representations in terms of (1) prompt and completion format, such as prompting style or step-by-step thinking, (2) keyword overlap, and (3) shared topics or themes. For each pairing of attribution representations (e.g., $r^{idx}_{\nabla f}$ vs. r_{bm25}), we analyze one test instance with minimal overlap and one test instance with maximal overlap between neighborhoods. Below we summarize our results; full annotations are made available in Appendix D.

Alignment depends on instruction tuning goals. Alignment patterns vary across tasks. For the datasets *math* and *code*, neighborhoods tend to align in terms of example structure, such as shared problem-decomposition patterns or prompting styles. In contrast, neighborhoods for *squad_v2* and *ftrace* frequently align at the level of high-level question type, or semantic category (e.g., incomplete-information handling). When BM25-based neighborhoods lack keyword overlap with the test instance, they also tend to have low overlap with other strategies. In contrast, neighborhoods of gradient- and embedding-based methods often consist of instances with similar structures or completion behaviors, despite limited keyword overlap.

Prompting artifacts drive misalignment. Recurring instruction specifications and persona-prompting templates introduce repeated structure across all datasets, increasing lexical overlap. BM25 appears to be frequently driven by these artifacts. In contrast, representation-based methods appear to exploit more meaningful task-related similarities between instances (from an instruction-tuning perspective). On the other hand, test instances targeting abilities not directly related to

task solving (such as safety refusals, indications of missing information, or reproduction of hard-coded facts such as knowledge-cutoff dates) often appear in the same neighborhoods under the representation-based methods ($r^{id_X}_{\nabla f}$, r_{repsim}, and r_{nemot}). This occurs even when the instances are neither semantically nor topically related: for example, some neighborhoods contain *both* safety-related examples and examples on the cutoff date.

5 Discussion

Task dependency of alignment and limitations of lexical-matching. The observed divergence across representation paradigms appears to stem from differences in what each method encodes as similarity. Lexical methods (e.g., BM25) primarily capture surface form and instruction overlap, and therefore diverge in their attributions from other strategies when structure is shared, but wording differs. In contrast, gradient-based and embedding methods likely better capture structural and behavioral similarity, grouping instances by shared instruction and completion patterns rather than keyword overlap. Given that learning behavior rather than memorization is the central goal of instruction fine-tuning tasks, this suggests that relying on lexical overlap alone may be a weak basis for attribution in instruction fine-tuning.

Measuring functional alignment is insufficient. Our results show that the three proposed evaluation protocols capture distinct and only partially overlapping aspects of alignment. We further show that methods may exhibit low correlation in attributions across the full dataset yet strong pairwise agreement on example relevance, as well as high representational alignment. This may partially explain the apparent mismatch between prior findings, where some report low cross-paradigm correlation, while others independently observe similar downstream-task performance of methods belonging to different paradigms. To assess whether this hypothesis holds, future work should design evaluation protocols spanning downstream tasks that vary in their reliance on local versus global agreement, and test whether cross-method predictability scores predict performance on individual tasks.

Alignment deteriorates cross-paradigm. While most pairings achieve above-chance cross-method predictability, accuracy above 0.75 is consistently reached only for within-paradigm comparisons, whereas the only cross-paradigm pairings scoring above 0.65 are those between lexical- and embedding-based methods. Representational and functional alignment across paradigms is likewise weak. Our findings suggest that practitioners seeking to use one method as a proxy for another (e.g., to pre-filter relevant training examples before applying more computationally expensive but also more accurate methods) are advised to select methods within the same attribution paradigm. At the same time, the low alignment across paradigms suggests that combining multiple attribution strategies could potentially yield more comprehensive characterizations of data influence.

6 Conclusion

We introduce a framework for directly measuring alignment between training data attribution methods, which captures overall correlation of attributions (*functional alignment*), assesses structural similarity of the underlying representation spaces used for attribution (*representational alignment*), and evaluates local agreement on pairwise ranking decisions (*cross-method predictability*). We show that this decomposition is useful for disentangling different notions of alignment that are not captured by existing evaluations based on measures of downstream performance. Comparing 12 attribution methods spanning four paradigms, we find high alignment within individual attribution paradigms (gradient similarity, model-internal representations, embedding similarity, and lexical overlap). In contrast, cross-paradigm alignment is consistently lower and task-dependent. This suggests that methods within a paradigm are partially interchangeable, whereas those across paradigms are not reliable proxies for one another. Qualitatively, we find that this divergence stems from a fundamental difference in how paradigms define similarity, with lexical methods capturing surface-level prompting artifacts, while representation-based methods capture underlying behavioral structure.

Overall, our results show that representational alignment, functional alignment, and cross-method predictability do not necessarily covary, emphasizing the need to distinguish between local and overall agreement in evaluation. To this end, we recommend augmenting future work on attribution through alignment analysis to better contextualize new approaches within known paradigms.

Limitations

Our experiments consider language generation tasks. Alignment between specific paradigms may differ for other task formats, such as simpler classification setups. However, our framework is fully compatible with other task definitions and models outside NLP contexts.

In contrast to prior work, we intentionally forgo evaluating downstream performance in favor of directly testing alignment, which is required to establish a more fundamental understanding. Our focus was on developing complementary perspectives on alignment, rather than studying failure modes specific to individual tasks or replicating prior downstream evaluations.

We restrict our analysis to base models with a maximum of 8B parameters due to the computational costs involved with obtaining and analyzing gradient-based influence estimates. We ensure robustness by validating our findings across seven models spanning parameter sizes from 0.6B to 8B and three model families.

While future work may explore alternative training data attribution paradigms, we have compiled a set of 12 methods across four paradigms, which represent a cross-section of the current literature.

Ethical Considerations

Given that we document patterns of alignment across attribution paradigms, there is a risk that actors acting in bad faith could misuse this information to select an attribution strategy that aligns with their desired claims of data influence, despite prior negative results obtained with other methods. Our work mitigates this risk by demonstrating that single-paradigm attribution may provide an incomplete view of data influence.

References

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards Tracing Knowledge in Language Models Back to the Training Data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-Embed-Nemotron-8B: A Universal Text Embedding Model for Multilingual and Cross-Lingual Tasks. *arXiv preprint*.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. 2022. If Influence Functions are the Answer, Then What is the Question? *Advances in Neural Information Processing Systems*, 35:17953–17967.

Jonathan Brophy, Zayd Hammoudeh, and Daniel Lowd. 2024. Adapting and evaluating influence-estimation methods for gradient-boosted decision trees. *J. Mach. Learn. Res.*, 24(1):154:7325–154:7372.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2024. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, and 48 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying Large Language Model Generalization with Influence Functions. *Preprint*, arXiv:2308.03296.

Peijian Gu, Yaozong Shen, Lijie Wang, Quan Wang, Hua Wu, and Zhendong Mao. 2023. IAEval: A Comprehensive Evaluation of Instance Attribution on Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11966–11977, Singapore. Association for Computational Linguistics.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5553–5563, Online. Association for Computational Linguistics.

Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. 2020. Evaluation of Similarity-based Explanations. In *International Conference on Learning Representations*.

Dat Minh Hong, Bruno Kacper Mlodozeniec, Runa Eschenhagen, and Richard E. Turner. 2025. Better Hessians Matter: Studying the Impact of Curvature Approximations in Influence Functions. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20617–20642. PMLR.

Andrew Ilyas and Logan Engstrom. 2025. MAGIC: Near-Optimal Data Attribution for Deep Learning. *Preprint*, arXiv:2504.16430.

Cathy Jiao, Weizhen Gao, Aditi Raghunathan, and Chenyan Xiong. 2025a. On the Feasibility of In-Context Probing for Data Attribution. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5140–5155, Albuquerque, New Mexico. Association for Computational Linguistics.

Cathy Jiao, Yijun Pan, Emily Xiao, Daisy Sheng, Niket Jain, Hanzhang Zhao, Ishita Dasgupta, Jiaqi W. Ma, and Chenyan Xiong. 2025b. DATE-LM: Benchmarking Data Attribution Evaluation for Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2023. DataInf: Efficiently Estimating Data Influence in LoRA-tuned LLMs and Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Jarkko Lagus, Janne Sinkkonen, and Arto Klami. 2019. Low-Rank Approximations of Second-Order Document Representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 634–644, Hong Kong, China. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Second Conference on Language Modeling*.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. 2025. Do Influence Functions Work on Large Language Models? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14367–14382, Suzhou, China. Association for Computational Linguistics.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, and 11 others. 2025. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188, Vienna, Austria. Association for Computational Linguistics.

Bruno Kacper Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard E. Turner. 2024. Influence Functions for Scalable Data Attribution in Diffusion Models. In *The Thirteenth International Conference on Learning Representations*.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. 2023. A Bayesian Approach To Analysing Training Data Attribution In Deep Learning. *Advances in Neural Information Processing Systems*, 36:64155–64180.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27074–27113. PMLR.

Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. 2021. An Empirical Comparison of Instance Attribution Methods for NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 967–975, Online. Association for Computational Linguistics.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

R. Sokal and C. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*.

Robert R. Sokal and F. James Rohlf. 1962. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40.

Weiwei Sun, Haokun Liu, Nikhil Kandpal, Colin Raffel, and Yiming Yang. 2026. Enhancing Training Data Attribution with Representational Optimization. *Advances in Neural Information Processing Systems*, 38:102274–102311.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 Furious (COLM's Version). In *Second Conference on Language Modeling*.

Kangxi Wu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. Enhancing Training Data Attribution for Large Language Models with Fitting Error Consideration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14131–14143, Miami, Florida, USA. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54104–54132. PMLR.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, and Pradeep Ravikumar. 2022. First is Better Than Last for Language Data Influence. *Preprint*, arXiv:2202.11844.

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok, Thailand. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint*.

A Likelihood Ratio Tests

Term → Metric ↴	Base model		Dataset		Interaction	
ρ	13.1		27.9		3.2	
τ	10.1		23.5		2.5	
RBO	49.2	*	136.7	***	23.5	
acc_{pred}	733.9	***	831.1	***	411.0	***
CKA	105.1	***	73.8	***	24.4	
CKNNA@25	50.9	**	142.7	***	17.6	

Table 3: **Likelihood-ratio tests.** Results of likelihood-ratio tests comparing the full mixed-effects model against reduced variants with base model, dataset, or interaction terms removed. See Section 3.6.

B Cophenetic Stability

	Llama3.2-1B	OLMo2-1B	OLMo3-7B	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
CKA	0.80	0.82	0.80	0.84	0.83	0.81	0.79
CKNNA@25	0.86	0.76	0.74	0.76	0.74	0.72	0.75
τ	0.74	0.81	0.76	0.83	0.81	0.79	0.75
acc_{pred}	0.79	0.79	0.80	0.83	0.83	0.80	0.79
RBO	0.74	0.78	0.73	0.76	0.78	0.75	0.72
ρ	0.72	0.78	0.71	0.79	0.76	0.75	0.71

Table 4: Average cophenetic correlation across tasks per base model.

	code	ftrace	math	sft	squad_v2
CKA	0.92	0.95	0.84	0.95	0.97
CKNNA@25	0.87	0.96	0.92	0.95	0.98
τ	0.92	0.90	0.77	0.90	0.96
acc_{pred}	0.99	0.90	0.80	0.93	0.96
RBO	0.91	0.86	0.78	0.90	0.94
ρ	0.89	0.88	0.74	0.89	0.95

Table 5: Average cophenetic correlation across base models per task.

	CKA	CKNNA@25	τ	acc_{pred}	RBO	ρ
CKA	0.73	0.70	0.05	0.03	0.02	0.05
CKNNA@25		0.67	0.09	0.07	0.06	0.09
τ			0.61	0.30	0.57	0.61
acc_{pred}				0.16	0.29	0.30
RBO					0.52	0.57
ρ						0.61

Table 6: Average cophenetic correlation between alignment measures.

C Full Alignment Scores

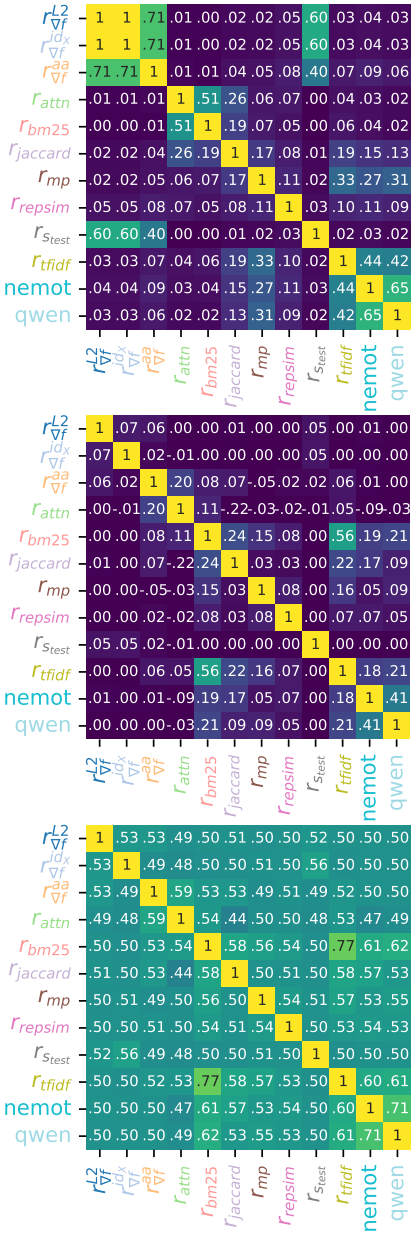

Figure 4: Representational alignment (CKNNA@25: top), and functional alignment (Kendall's τ: center, RBO: bottom), averaged across tasks and models.

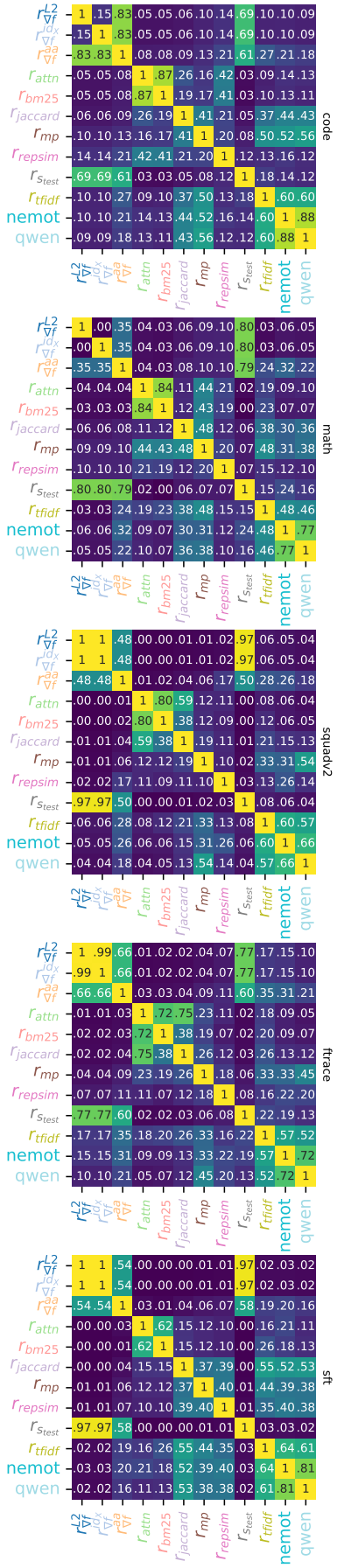

Figure 5: Per-task CKA results.

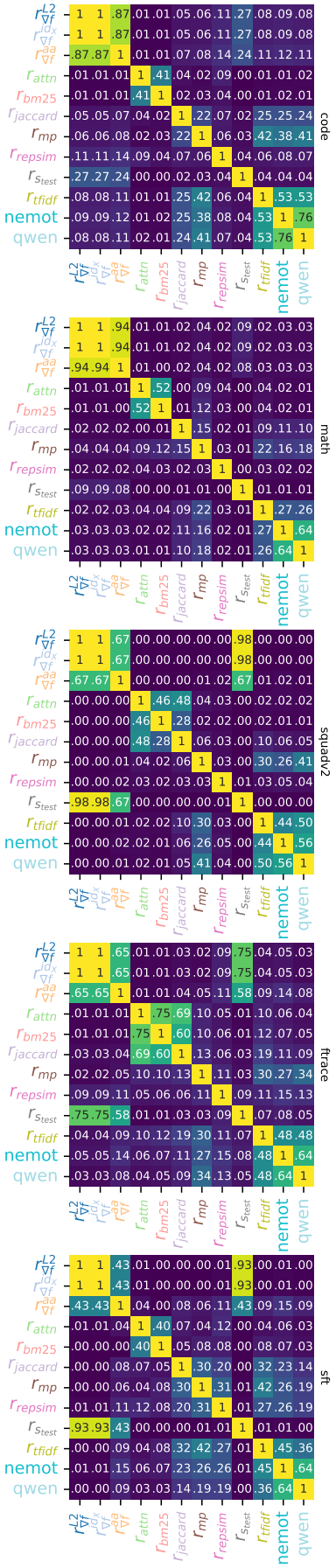

Figure 6: Per-task CKNNA@25 results.

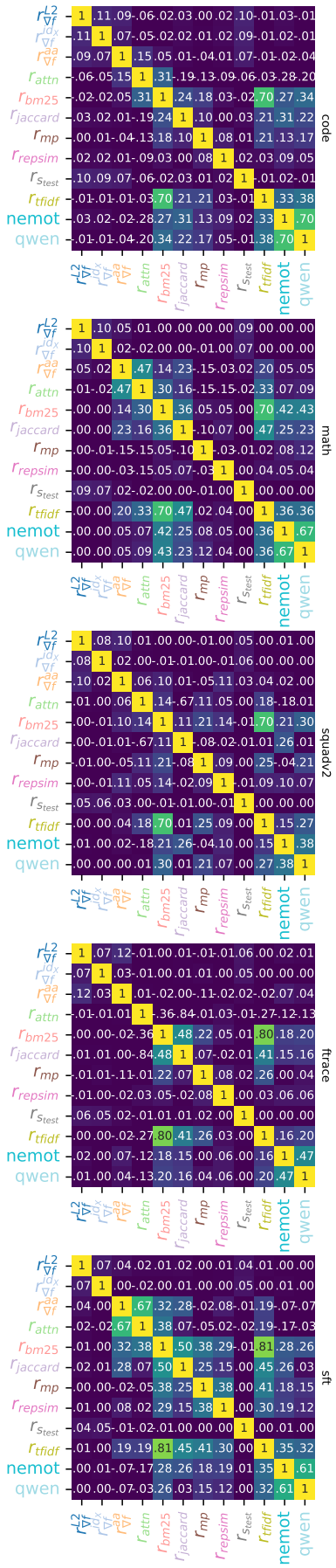

Figure 7: Per-task Spearman's ρ results.

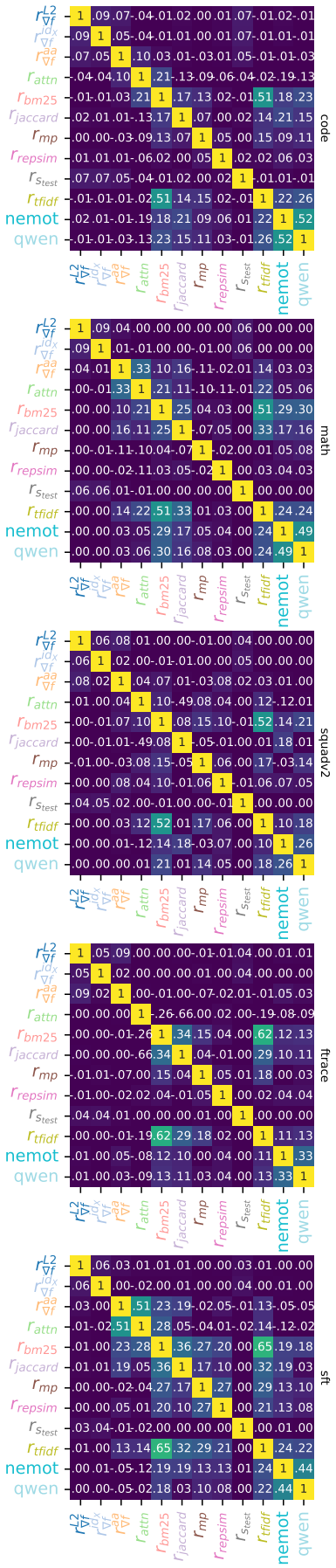

Figure 8: Per-task Kendall's τ results.

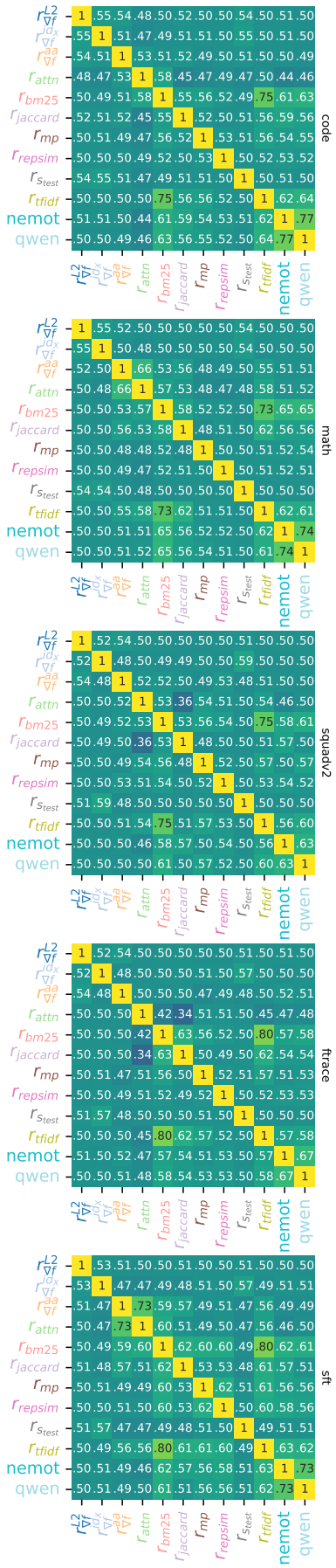

Figure 9: Per-task RBO results.

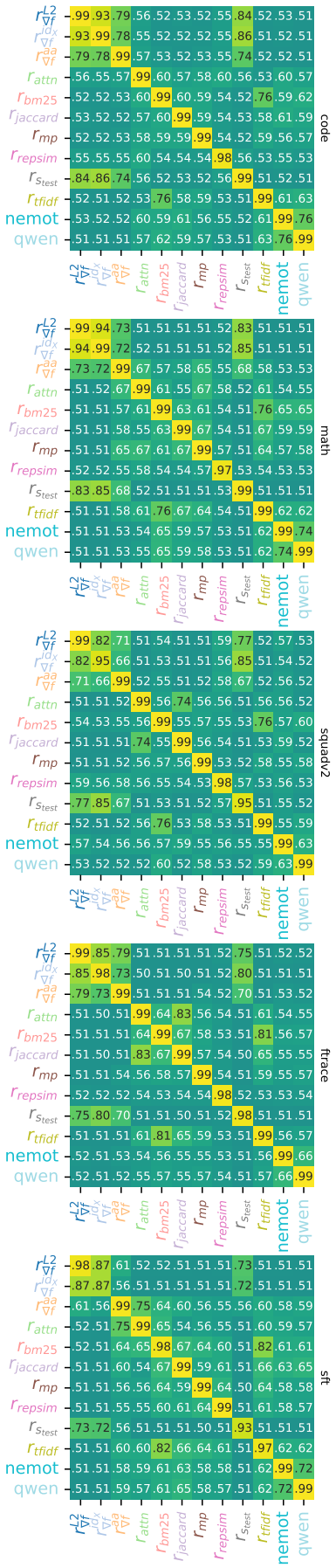

Figure 10: Per-task cross-method predictability results.

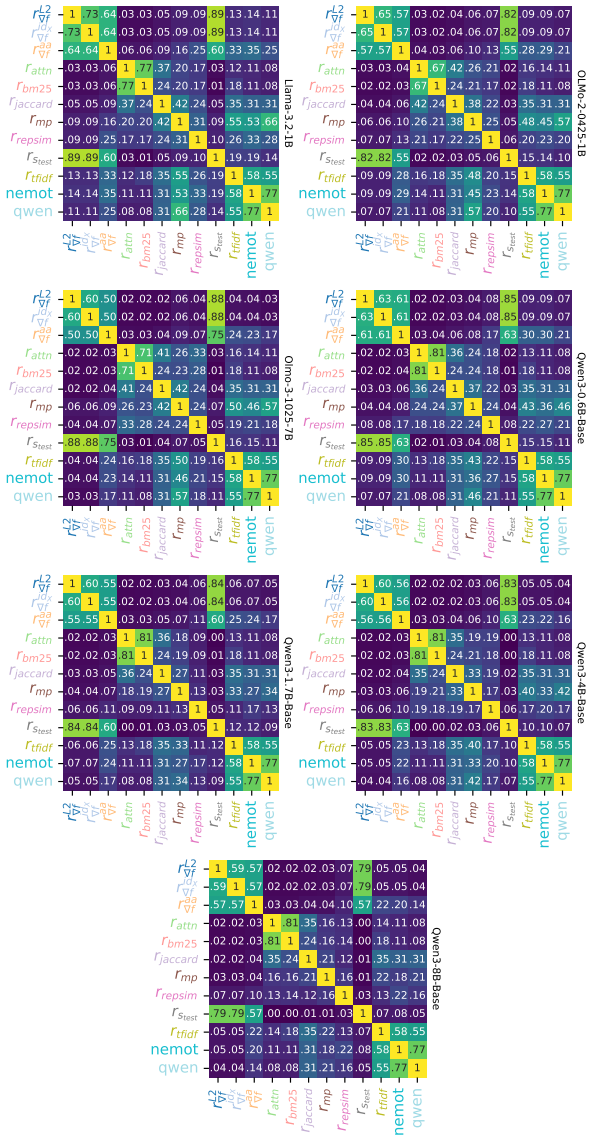

Figure 11: Per-task CKA results.

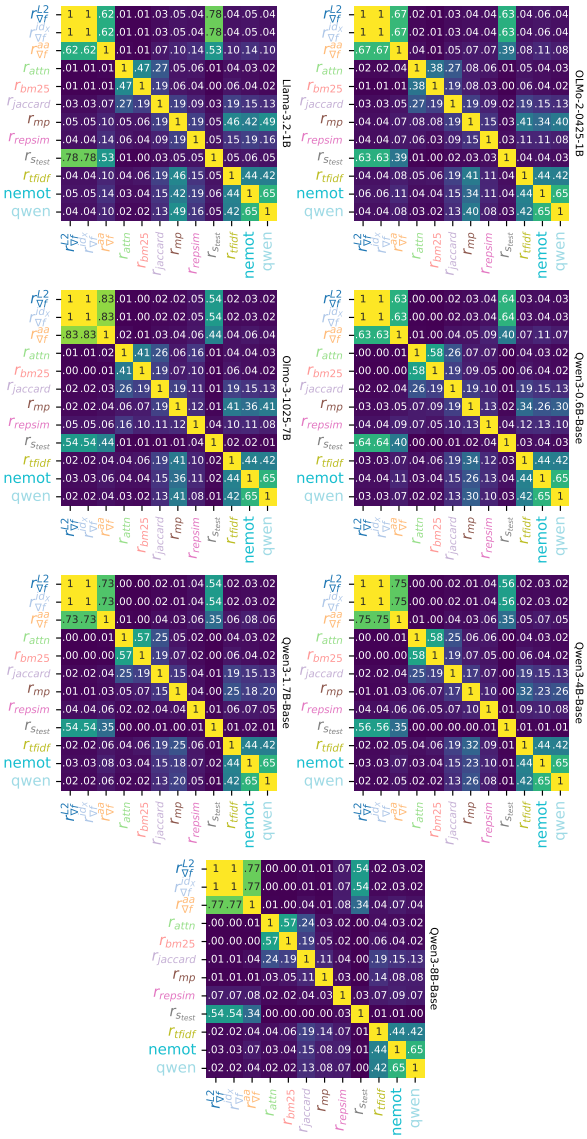

Figure 12: Per-base-model CKNNA@25 results.

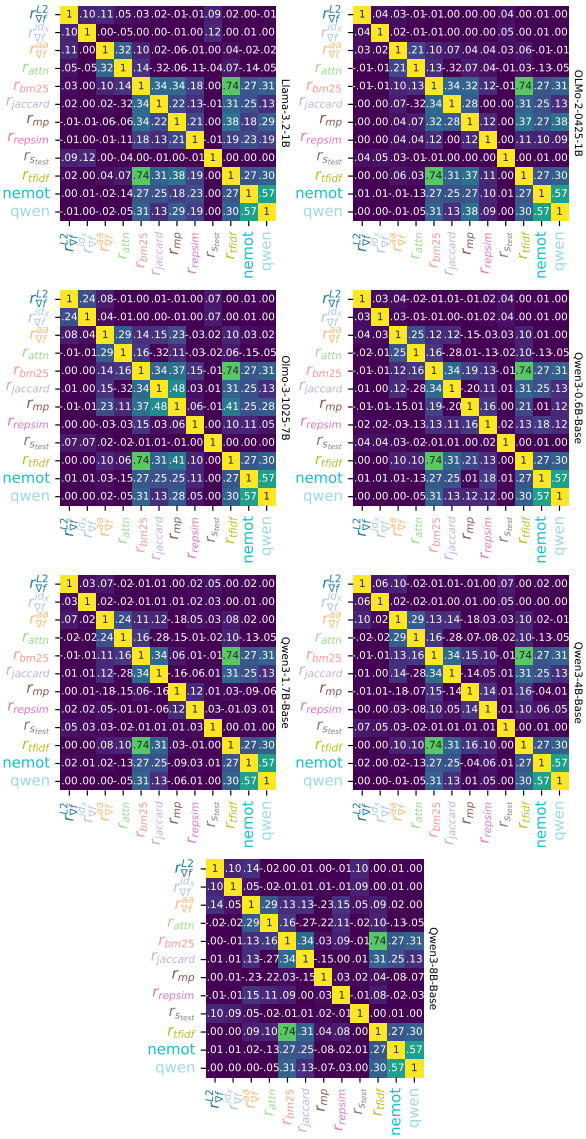

Figure 13: Per-task Spearman's ρ results.

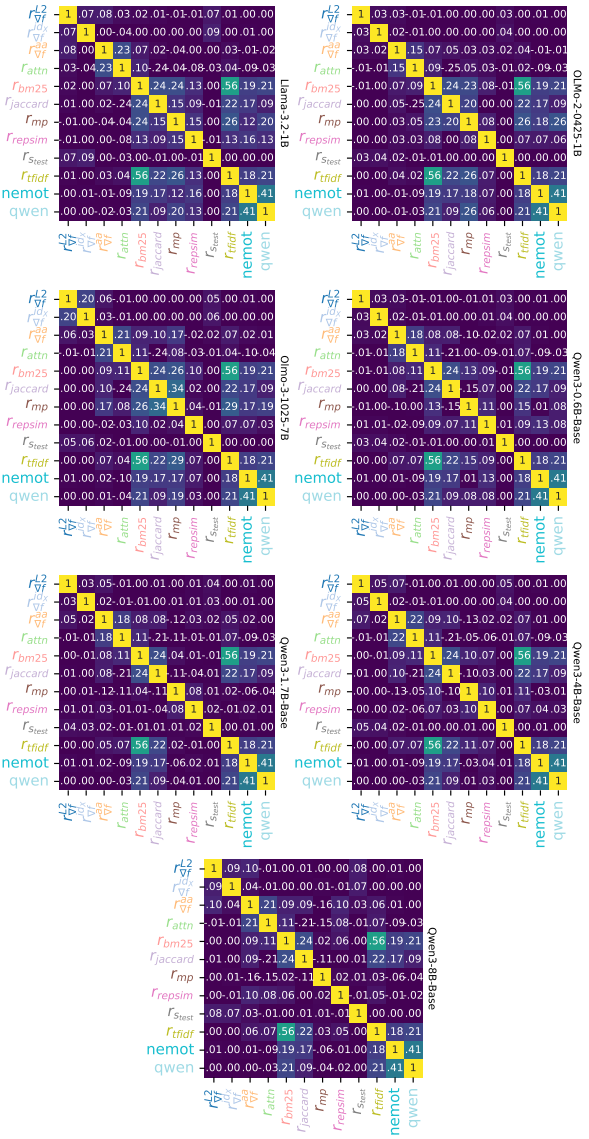

Figure 14: Per-base-model Kendall's τ results.

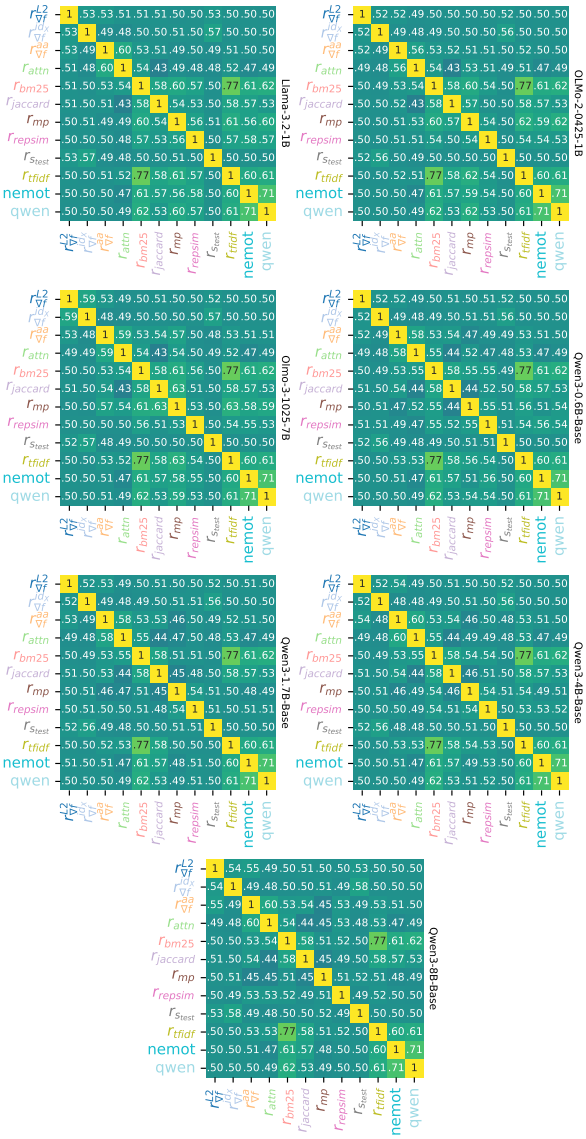

Figure 15: Per-task RBO results.

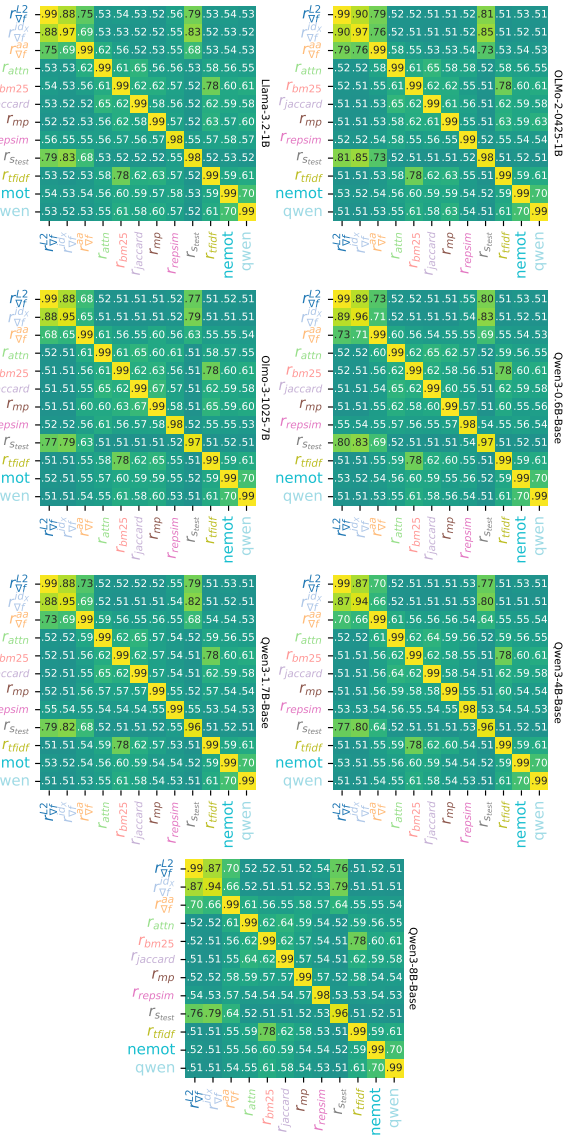

Figure 16: Per-base model cross-method predictability.

D Qualitative Study of Local Neighborhoods

We provide observations for all 60 studied examples in Table 7. Figures 17 to 21 depict a subset of examples selected as follows: For a given representation r_i, let $\mathcal{N}_5^{(r_i)}(z)$ denote the set of top-5 training examples ranked by similarity to document z. For two representations (r_i, r_j), we quantify overlap between their local neighborhoods around a common z: $\frac{|\mathcal{N}_5^{(r_i)}(z) \cap \mathcal{N}_5^{(r_j)}(z)|}{5}$.

For each dataset, we retrieve and study neighborhoods of two documents z_{max} and z_{min}: one with maximal overlap among the top-5 examples (cf Figure 17: top), and one with minimal overlap (bottom). We break ties by random selection. Prior to computing inner-product similarity, we mean-center representations to remove global offsets in the embedding space, ensuring that neighborhoods reflect relative structure rather than absolute activation magnitudes.

Choice of representations. We restrict this analysis to 4 of the 12 representations. Given a CKA dissimilarity matrix over all candidates, we apply a greedy farthest-first selection procedure: starting from the most central representation, we iteratively add the most dissimilar remaining representation. This results in

- r_{nemot} (embedding model),
- $r_{\nabla f}^{id_X}$ (raw loss gradients),
- r_{bm25}, and
- r_{repsim} (model-internal embeddings).

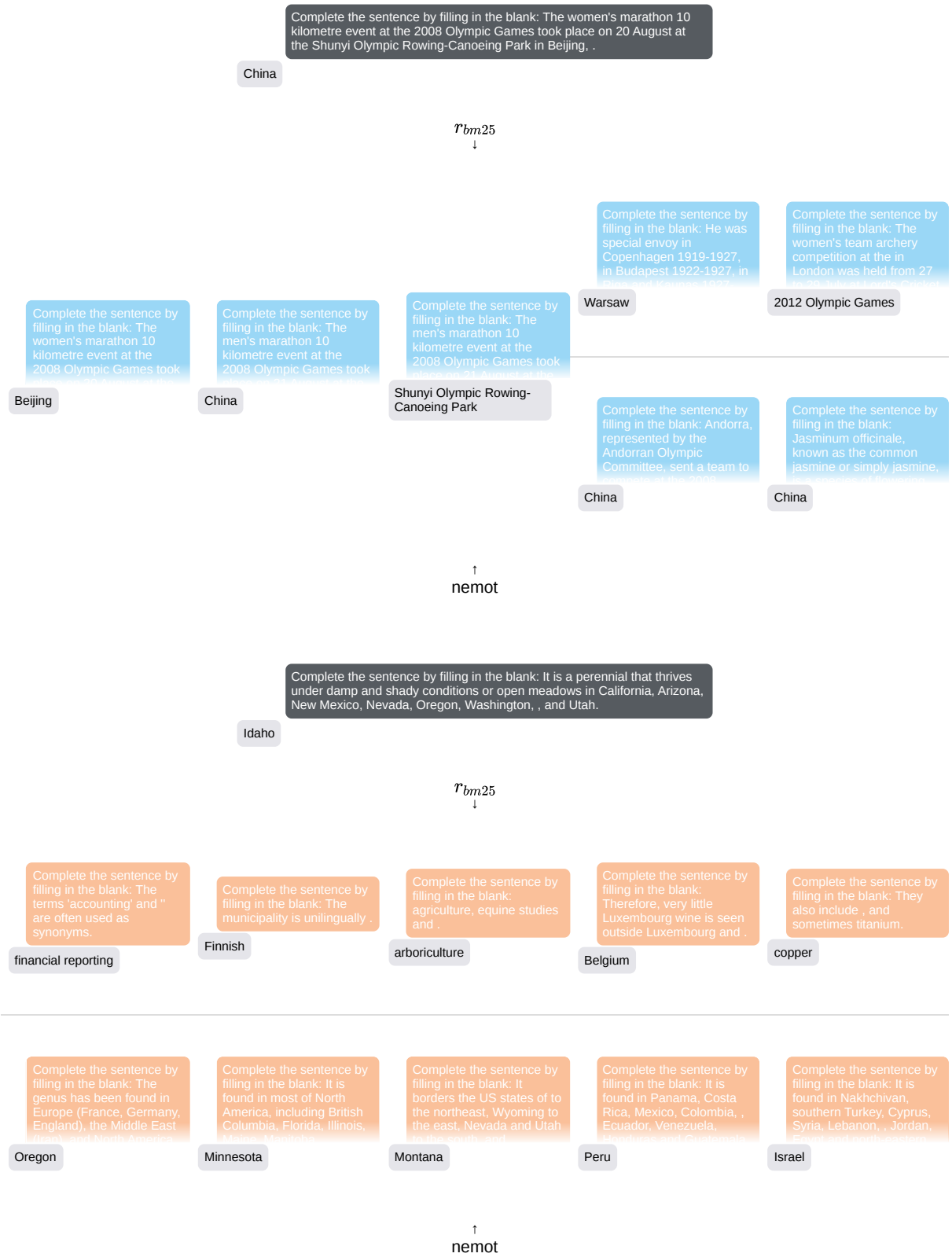

Figure 17: Instances with highest (top) and lowest (bottom) 5-nearest neighbor overlap across representations r_{bm25} and nemot for `Olmo-3-7B_ftrace_date-lm_conversational`.

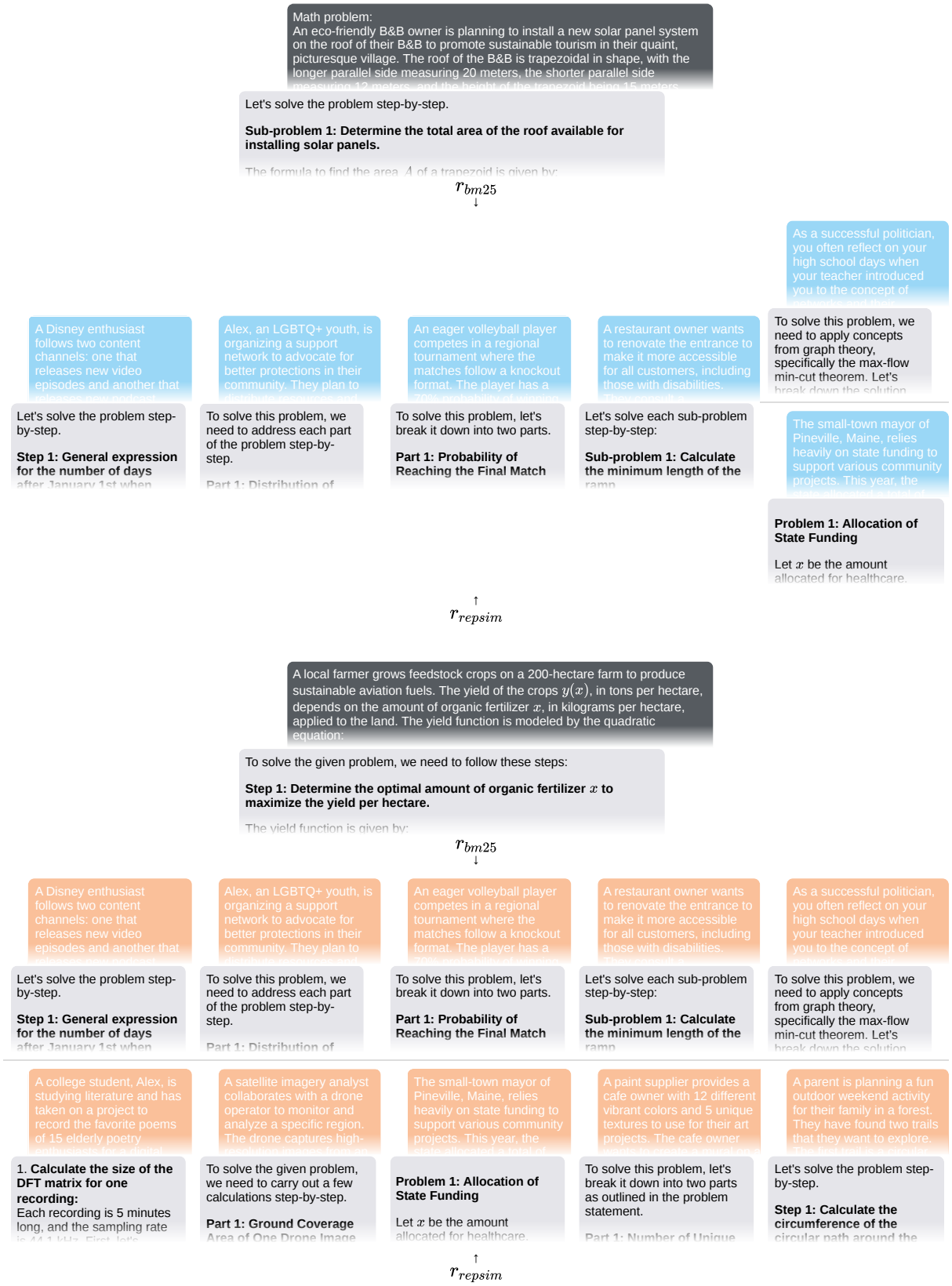

Figure 18: Instances with highest (top) and lowest (bottom) 5-nearest neighbor overlap across representations r_{bm25} and r_{repsim} for `Olmo-3-7B_tulu-3-sft-personas-math`.

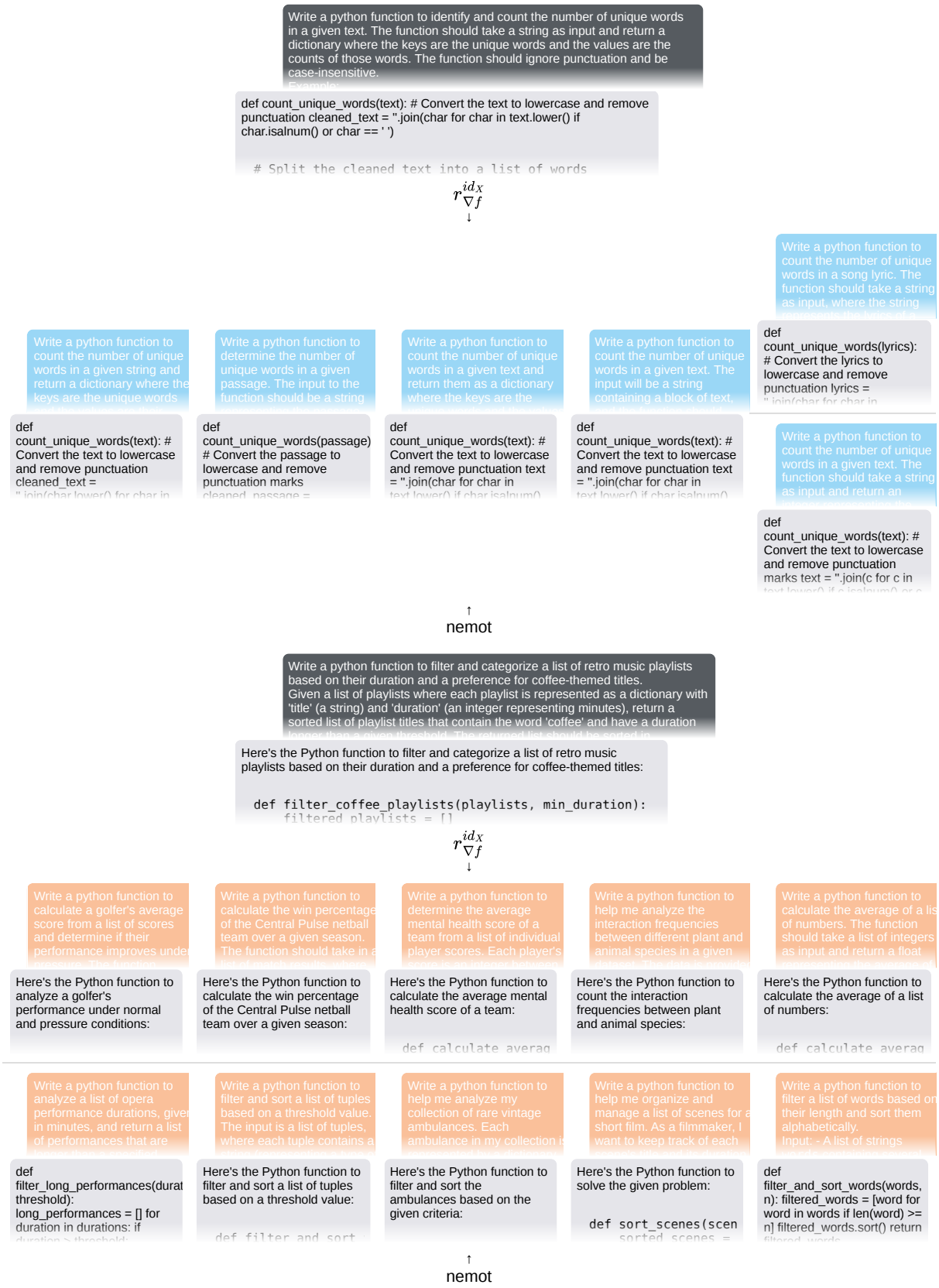

Figure 19: Instances with highest (top) and lowest (bottom) 5-nearest neighbor overlap across representations $r^{id_X}_{\nabla f}$ and nemot for `Olmo-3-7B_tulu-3-sft-personas-code`.

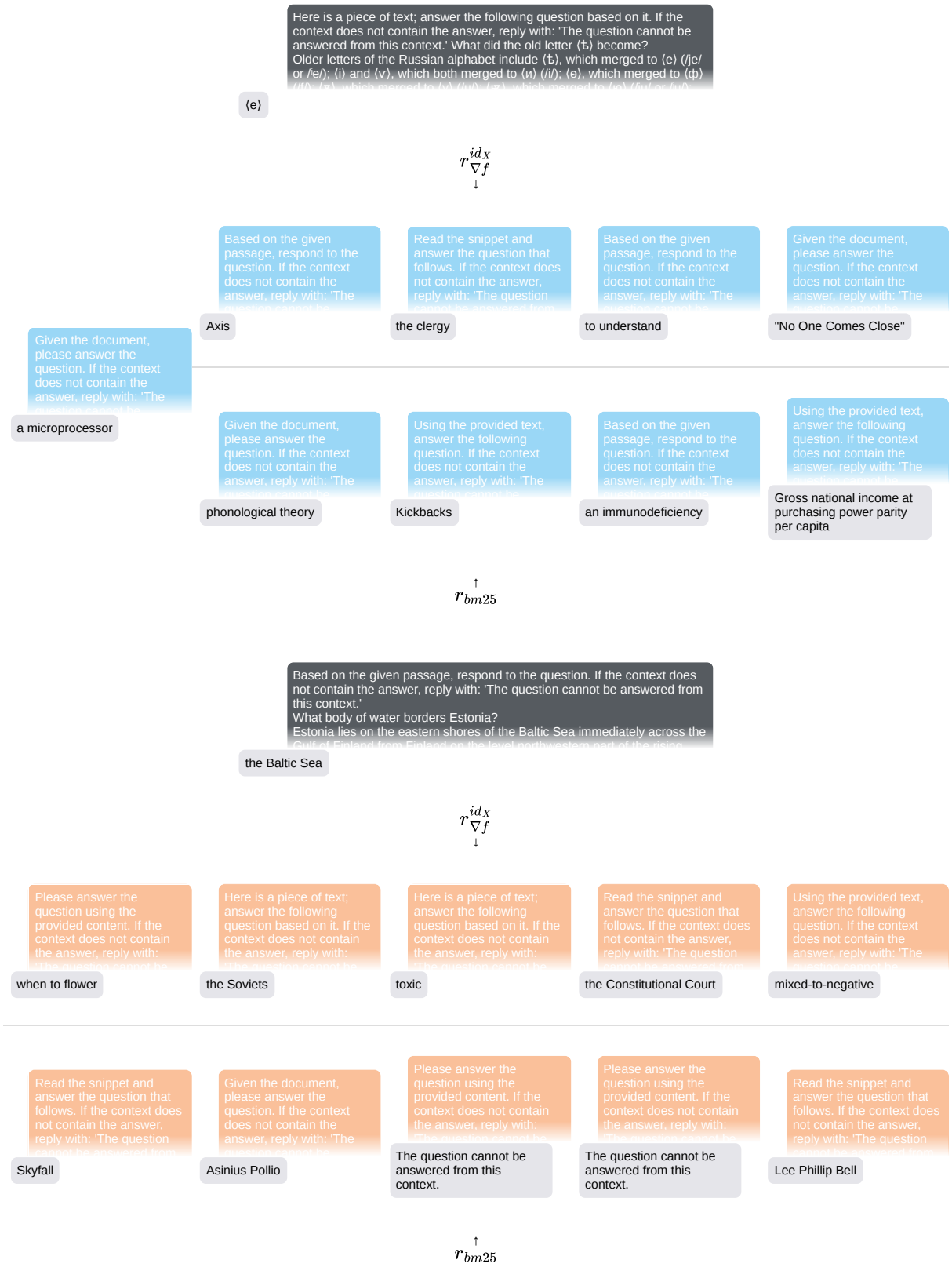

Figure 20: Instances with highest (top) and lowest (bottom) 5-nearest neighbor overlap across representations $r^{idx}_{\nabla f}$ and r_{bm25} for `Olmo-3-7B_squad_v2_conversational`.

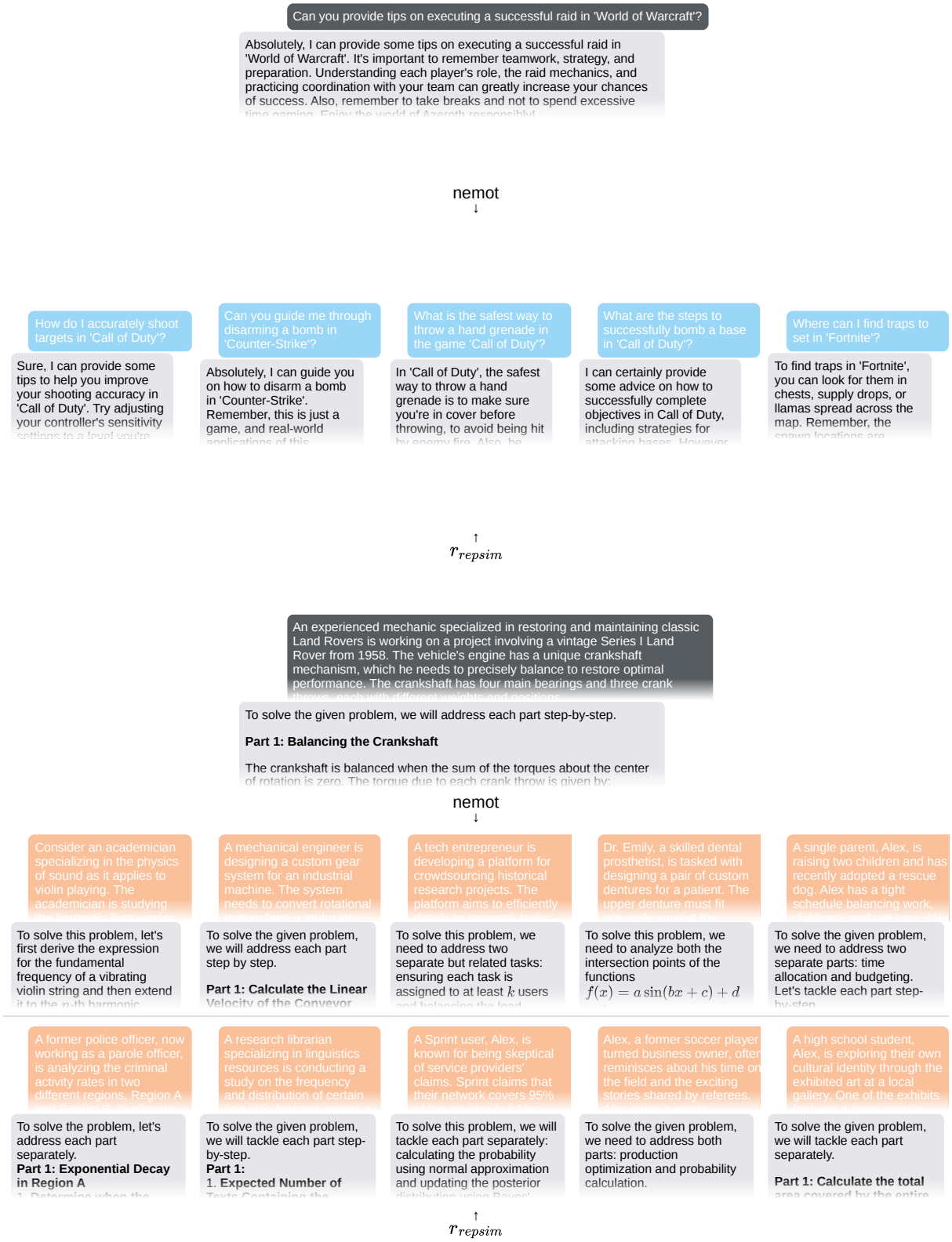

Figure 21: Instances with highest (top) and lowest (bottom) 5-nearest neighbor overlap across representations nemot and r_{repsim} for `Olmo-3-7B_tulu-3-sft-olmo-2-mixture-0225`.

Table 7: Full observations from the qualitative analysis, grouped by dataset.

		Max	Min
		FTrace	
$r^{id_X}_{\nabla f}$	r_{bm25}	1/5 overlap. Both local neighborhoods consist mostly of instances whose completions are of the same semantic category as the test instance.	For $r^{id_X}_{\nabla f}$, one completion matches the test instance completion, with no clear relation between the test instance and the others. For r_{bm25}, there is one keyword match between the test instance prompt and a completion, with no obvious relation for the others.
$r^{id_X}_{\nabla f}$	r_{nemot}	4/5 match. The single differing instance in r_{nemot} shares the same completion as the test instance, whereas the completion in $r^{id_X}_{\nabla f}$ differs but remains within the same semantic category.	For min, both r_{nemot} and $r^{id_X}_{\nabla f}$ contain completions from the same semantic category. Three instances in r_{nemot} match the test instance completion.
$r^{id_X}_{\nabla f}$	r_{repsim}	3/5 match. for $r^{id_X}_{\nabla f}$ exact same instruction following goal of predicting pronouns, for r_{repsim} not.	Except for one example of $r^{id_X}_{\nabla f}$, all examples across the test instance and both representations are of the same semantic category.
r_{bm25}	r_{nemot}	3/5 match. All r_{nemot} instances either match the test instance's prompt or belong to the same semantic category. For r_{bm25}, one instance matches via keyword overlap, while the other appears unrelated.	For r_{bm25}, no direct relation to the test instance is apparent, whereas all r_{nemot} instances belong to the same semantic category.
r_{bm25}	r_{repsim}	5/5 match. All but one are of the same semantic category as the test instance.	Three for r_{bm25} and one for r_{repsim} are of the same semantic category as the test instance. The other examples appear unrelated, with r_{repsim} providing three given names as completion, which resembles the test instance's prompt.
r_{nemot}	r_{repsim}	4/5 match. For r_{nemot}, the non-matching example is in the same semantic category as the other examples and the test instance; for r_{repsim}, it is not.	For r_{nemot}, 3/5 examples are of the same semantic category as the test instance. For r_{repsim}, examples are only distantly related in topic.
		Math	
$r^{id_X}_{\nabla f}$	r_{bm25}	2/5 match. No obvious difference between the examples of $r^{id_X}_{\nabla f}$ and r_{bm25}.	As for max.
$r^{id_X}_{\nabla f}$	r_{nemot}	2/5 match. The prompts for r_{nemot} appear more semantically related to the test instance's prompt than $r^{id_X}_{\nabla f}$.	The prompts for r_{nemot} appear more semantically related to the test instance's prompt than $r^{id_X}_{\nabla f}$.
$r^{id_X}_{\nabla f}$	r_{repsim}	2/5 match. The matching instances both lack instruction on step-by-step thinking, while the remaining examples for both $r^{id_X}_{\nabla f}$ and r_{repsim} include it.	The examples for r_{repsim} share in common that they do not decompose into steps, whereas the ones of $r^{id_X}_{\nabla f}$ decompose into subproblems, matching the test instance.
r_{bm25}	r_{nemot}	2/5 match. The matching examples provide completions that are lexically related to the test instance. No obvious relation for the examples of r_{bm25}. The descriptions of the personas employed by the non-matching examples of r_{nemot} are semantically related.	Completion share identical introduction for r_{bm25}, with no obvious relation to the test instance. The description of the personas employed by the non-matching examples of r_{nemot} are semantically related.
r_{bm25}	r_{repsim}	4/5 match. The formatting of the instance that differs for r_{bm25} is less similar to the test instance than that of r_{repsim}.	Similar formatting across instances of both representations and test instance, no obvious semantic relation to the test instance.
r_{nemot}	r_{repsim}	2/5 match. Non-matching examples of r_{nemot} are semantically related to the test instance. No obvious relation for those of r_{repsim}.	Examples of r_{nemot} are semantically related to the test instance. No obvious relation to the test instance for those of r_{repsim}.
		Code	
$r^{id_X}_{\nabla f}$	r_{bm25}	1/5 matches. All examples share the same completion formatting, providing only a code block without any additional introduction. Only one example ($r^{id_X}_{\nabla f}$) features a completion directly relevant to solving the problem stated by the test instance.	All examples share the same completion formatting, and only one example ($r^{id_X}_{\nabla f}$) features a completion directly relevant to solving the problem stated by the test instance.
$r^{id_X}_{\nabla f}$	r_{nemot}	4/5 match. All examples across both representations are on the exact same coding task as the test instance. The non-matching example of $r^{id_X}_{\nabla f}$ differs from the test instance only in that another type of text is stated in the instruction, for r_{nemot} it matches the one of the test instance.	The examples of $r^{id_X}_{\nabla f}$ do not appear to be semantically related to the test instance. All but one example of r_{nemot} are semantically related to and relevant for solving the test instance's problem and match in terms of keywords.

Continued on next page

		Max	Min
$r^{id_X}_{\nabla f}$	r_{repsim}	3/5 match. The examples of r_{repsim} appear more closely related to the test instance's objective than the ones of $r^{id_X}_{\nabla f}$.	All examples of $r^{id_X}_{\nabla f}$ share a common completion formatting that does not match the test instance nor r_{repsim}. No direct relation to test instance for $r^{id_X}_{\nabla f}$, examples of r_{repsim} are loosely of the same task category as the test instance.
r_{bm25}	r_{nemot}	2/5 match. For r_{nemot}, examples are semantically related to the test instance and match via keywords. Examples for r_{bm25} feature unrelated goals but match in terms of keywords.	Examples for r_{bm25} match via keywords, those of r_{nemot} are more semantically related to the test instance's task goal.
r_{bm25}	r_{repsim}	2/5 match. Examples of r_{bm25} match via keywords, no obvious relation of the examples of r_{repsim} to the test instance beyond similar completion formatting.	All examples of r_{bm25} share a common completion formatting that does not match the test instance nor r_{repsim}. Examples of r_{bm25} match in terms of keywords with the test instance, no obvious relation of the examples of r_{repsim} to the test instance.
r_{nemot}	r_{repsim}	3/5 match. All examples but one for r_{repsim} across both representations are on the exact same coding task as the test instance.	The description of the personas employed by r_{nemot} are semantically related to those in the test instance. For r_{repsim}, all examples share the same completion formatting with the test instance but appear not semantically related to it.
		squad_v2	
$r^{id_X}_{\nabla f}$	r_{bm25}	1/5 matches. No apparent relation of either set of examples to the test instance.	As for max.
$r^{id_X}_{\nabla f}$	r_{nemot}	2/5 match. Examples of $r^{id_X}_{\nabla f}$ appear unrelated to the test instance, only one example for r_{nemot} is semantically related.	All completions for $r^{id_X}_{\nabla f}$ and all but two for r_{nemot} are examples for indicating incomplete information. The example that provides a response for r_{nemot} is semantically related to the test instance.
$r^{id_X}_{\nabla f}$	r_{repsim}	2/5 match. For $r^{id_X}_{\nabla f}$ all completions consist of numbers or indications of years, months or days, for r_{repsim} one example is a geographical location.	All examples are examples for indicating incomplete information, as is the test instance.
r_{bm25}	r_{nemot}	3/5 match. All examples of r_{nemot} are semantically related to the test instance, those of r_{bm25} appear unrelated.	Examples of r_{bm25} appear unrelated, those of r_{nemot} are semantically related to the test instance.
r_{bm25}	r_{repsim}	4/5 match. For both representations, examples either indicate incomplete information like the test instance, or belong to the same semantic task category.	For r_{bm25}, one example features matching keywords with the test instance, two indicate incomplete information (unlike the test instance). The examples in r_{repsim}, do not appear to be semantically related to the test instance, but two completions provide dates like the test instance.
r_{nemot}	r_{repsim}	2/5 match. All but one example (r_{repsim}) belong to the same semantic category.	Two completions for r_{nemot} and two for r_{repsim} belong to the same semantic category. For both representations, all prompts have keyword overlap or semantic similarity.
		SFT	
$r^{id_X}_{\nabla f}$	r_{bm25}	1/5 match. None of the examples from either representation appears relevant or in the format of the test instance. Examples of r_{bm25} are non-english. Two examples of $r^{id_X}_{\nabla f}$ are examples for answer refusal, one is a labeling task for harmful data.	For $r^{id_X}_{\nabla f}$, one is a labeling task for harmful data, two are safety-related, both in line with the instruction tuning goal of the test instance. For r_{bm25}, examples are non-English and do not appear related to the test instance.
$r^{id_X}_{\nabla f}$	r_{nemot}	3/5 match. All examples are of the same semantic category and share the refusal phrasing of the test instance almost verbatim.	For $r^{id_X}_{\nabla f}$, two examples contain information likely intended for the model to memorize verbatim, which appears to align with the instruction-tuning goal of the test instance. The remaining three are safety-related. For r_{nemot}, two examples also contain information likely intended for memorization in line with the test instance. Two examples are non-English.
$r^{id_X}_{\nabla f}$	r_{repsim}	4/5 match. All examples are of the same semantic category and share the refusal phrasing of the test instance almost verbatim.	For $r^{id_X}_{\nabla f}$, three examples are safety-related, and two appear unrelated to the test instance. For r_{repsim}, all examples share the semantic task category and completion format of the test instance.

Continued on next page

		Max	Min
r_{bm25}	r_{nemot}	4/5 match. Two of the matching examples share the same task category with the test instance, an additional one for r_{nemot} does so as well, despite being the only English-language example.	All examples except for one (r_{bm25}) share the same task category with the test instance for both representations.
r_{bm25}	r_{repsim}	4/5 match. All examples are non-English, none appear to be relevant to the test instance.	For r_{bm25}, only one example matches the task category of the test instance, for r_{repsim}, all do.
r_{nemot}	r_{repsim}	5/5 match. All examples match and are in line with the intended safety behavior demonstrated in the test instance.	All examples share the task category of the test instance and follow the same completion format. The examples of r_{nemot} seem to be more semantically related to the test instance than the ones of r_{repsim}.

E Details on $r_{s_{\text{test}}}$

Kwon et al. (2023) note that for negative log-likelihood loss functions, the inverse Hessian part in s_{test} can be approximated with a closed-form expression. Following their approach, we compute:

$$r_{s_{\text{test}}} \approx \frac{1}{n\lambda} \sum_{i=1}^{n} \left(\nabla_\theta L(z, \hat{\theta}) - \frac{\nabla_\theta L(z, \hat{\theta})^\top \nabla_\theta L(z_i, \hat{\theta})}{\lambda + \left\| \nabla_\theta L(z_i, \hat{\theta}) \right\|^2} \nabla_\theta L(z_i, \hat{\theta}) \right)$$

where n is the number of training examples and λ is a damping constant.

F Details on r_{attn}

Let $z = (x_1, \ldots, x_S)$ be a document of length S. Following Kobayashi et al. (2020) we consider a standard multi-head attention layer with output

$$\boldsymbol{y}_i^{\text{integrated}} = \sum_{j=1}^{n} \sum_{h} \alpha_{i,j}^{h} f^h(\boldsymbol{x}_j)$$

where $f^h(\boldsymbol{x}) := \left(\boldsymbol{x}\boldsymbol{W}^{V,h} + \boldsymbol{b}^{V,h}\right) \boldsymbol{W}^{O,h}$. Kobayashi et al. define a token-to-token interaction map $M \in \mathbb{R}^{n \times n}$, $M_{i,j} := || \sum_h \alpha_{i,j}^h f^h(x_j) ||$. To adapt this analysis to modern MHA implementations (e.g., HuggingFace), which use a single fused output projection (o_{proj}) after concatenating all heads, we note that the per-head output projections $W^{O,h}$ can be viewed as a block decomposition of W^O, such that

$$\sum_h (v^h W^{O,h}) \equiv \text{Concat}(v^1, \ldots, v^H) W^O.$$

We utilize the following norm-based measure to capture token-to-token attention at layer l:

$$M_{l,i,j} = \left\| \text{Concat}_h \left(\alpha_{i,j}^h f^{\tilde{h}}(\boldsymbol{x}_j) \right) W^O \right\|, M \in \mathbb{R}^{L \times (N \times N)}$$

where $f^{\tilde{h}}(\cdot)$ denotes the value transformation associated with head h. In standard multi-head attention, $f^{\tilde{h}} = f^h$, while in grouped-query attention (GQA), $f^{\tilde{h}} = f^{g(h)}$ with $g(h)$ mapping query heads to shared value groups.

G Fine-tuning

We fine-tune on NVIDIA H100 and H200 GPUs for a total of 199.22 hours (456.53 GPU hours), adjusting gradient accumulation to ensure consistent effective batch sizes across runs. All models and samples are made available under CC-BY 4.0 in line with their original licenses.

Precision	bfloat16	**LoRA rank (r)**	8	**Optimizer**	AdamW
Learning rate	1×10^{-4}	**LoRA alpha**	32	**LR scheduler**	Linear
Weight decay	0.0	**LoRA dropout**	0.1	**Effective batch size**	64
Max grad norm	1.0	**LoRA bias**	none	**Training epochs**	1
Trainable params	LoRA only	**Max sequence length**	4096	**Gradient checkpointing**	True
Target modules	`q_proj, k_proj, v_proj, o_proj`	**Seed**	42		

Table 8: LoRA fine-tuning hyperparameters

Task Suite Base Model	squad_v2_conv
Llama-3.2-1B	74.72
OLMo-2-0425-1B	77.81
Olmo-3-1025-7B	50.08
Qwen3-0.6B-Base	80.33
Qwen3-1.7B-Base	84.20
Qwen3-4B-Base	87.74
Qwen3-8B-Base	89.26

Table 9: Performance (F1) on held-out data of models fine-tuned on squad_v2.

Task Suite Base Model	ftrace_conv
Llama-3.2-1B	39.10
OLMo-2-0425-1B	36.96
Olmo-3-1025-7B	9.93
Qwen3-0.6B-Base	37.92
Qwen3-1.7B-Base	48.18
Qwen3-4B-Base	59.14
Qwen3-8B-Base	61.80

Table 10: Performance (F1) on held-out data of models fine-tuned on ftrace.

Task Suite Base Model	code_bpb
Llama-3.2-1B	0.42
OLMo-2-0425-1B	0.52
Olmo-3-1025-7B	0.32
Qwen3-0.6B-Base	0.35
Qwen3-1.7B-Base	0.31
Qwen3-4B-Base	0.29
Qwen3-8B-Base	0.28

Table 11: Performance on OLMES task suite `olmo3:base_easy:code_bpb`.

Task Suite Base Model	math_bpb
Llama-3.2-1B	0.62
OLMo-2-0425-1B	0.53
Olmo-3-1025-7B	0.37
Qwen3-0.6B-Base	0.43
Qwen3-1.7B-Base	0.38
Qwen3-4B-Base	0.34
Qwen3-8B-Base	0.33

Table 12: Performance on OLMES task suite`olmo3:base_easy:math_bpb`.

Task Suite Base Model	MMLU	olmo_2_generative	olmo_2_heldout
Llama-3.2-1B	0.29	0.46	0.25
OLMo-2-0425-1B	0.43	0.54	0.26
Olmo-3-1025-7B	0.67	0.71	0.54
Qwen3-0.6B-Base	0.54	0.47	0.37
Qwen3-1.7B-Base	0.65	0.57	0.47
Qwen3-4B-Base	0.75	0.65	0.62
Qwen3-8B-Base	0.79	0.69	0.67

Table 13: Macro-average score across OLMES task suites: `mmlu:mc::olmes`, `olmo_2_generative::olmes`, and `olmo_2_heldout::olmes` for models fine-tuned on `allenai/tulu-3-sft-olmo-2-mixture-0225`.