INVESTIGATING THE LINK BETWEEN REPRESENTATIONAL SIMILARITY AND MODEL INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Researchers have shown that neural similarity among humans predicts social closeness and cooperative success, whereas innovation often emerges from interactions among dissimilar individuals. We investigate whether these principles extend to artificial intelligence by examining interactions between large language models (LLMs). In our experiments, 276 model pairs interact across eight collaborative tasks spanning both cooperation and novelty. We find that pairs with more similar representation spaces achieve significantly higher cooperation but exhibit reduced novelty and creativity. These findings suggest that representational similarity can be an important consideration in multi-agent system design.

1 INTRODUCTION

The deployment of multiple LLMs in multi-turn, multi-agent interactions has progressed rapidly from concept to practice, with recent investigations in applications to social simulations (Park et al., 2023; Xie et al., 2024; Zhou et al., 2023), coding (Wu et al., 2024; Ishibashi & Nishimura, 2024), and a range of creative tasks such as brainstorming and scientific idea generation (Fukumura & Ito, 2025; Su et al., 2024). In many collaborative tasks, prior work has found that interaction between multiple agents facilitates stronger performance than single-agent systems (Talebirad & Nadiri, 2023; Zhuge et al., 2023). Beyond treating multi-agent systems as tools, some have even proposed evolving LLMs through multi-agent interaction (Lai et al., 2024; Eisenstein et al., 2025; Wu et al., 2025).

On the other hand, by their very nature, multi-agent systems are more complex than single-agent systems, increasing the potential for unexpected behaviors (Piatti et al., 2024; Hammond et al., 2025; de Witt, 2025). One central concern is whether agents can reliably cooperate with one another, since many multi-agent applications depend on effective collaboration. Being able to understand and predict the dynamics of multi-agent systems is therefore essential. Yet, most efforts to date have focused on single-agent cases, while studies of multi-agent systems have primarily focused on output-level behaviors rather than internal mechanisms.

This work provides an initial exploration of multi-agent interaction through the lens of representational alignment. Specifically, we ask:

What is the relationship between representational similarity and interactive behavior of models?

Evidence from neuroscience and social sciences suggests that similar neural responses among humans are significantly associated with their social closeness and cooperative performance (Parkinson et al., 2018; Thornton & Mitchell, 2017; Shen et al., 2025b; Reinero et al., 2021), while interaction between dissimilar individuals often sparks innovation (Hewlett et al., 2013; Østergaard et al., 2011). Analogously, we hypothesize that *models with higher representational similarity are more likely to cooperate and predict one another, but exhibit reduced collective novelty and creativity.*

To test this, we conduct experiments involving 276 model pairs spanning 23 open-weight LLMs from eight model families. Specifically, we examine cooperation through four games: word guessing, public good, divide-a-dollar, and the Keynesian Beauty Contest (KBC); and assess creativity and novelty through four generative tasks: story writing, fictional biography, haiku composition, and vacation benefit brainstorming.

Our experiments reveal that representational similarity is a strong predictor of interactive outcomes. Figure 1 illustrates how these outcomes vary with increasing internal similarity across scenarios: cooperation performance rises significantly as representational similarity increases. For example, in the word-guessing game

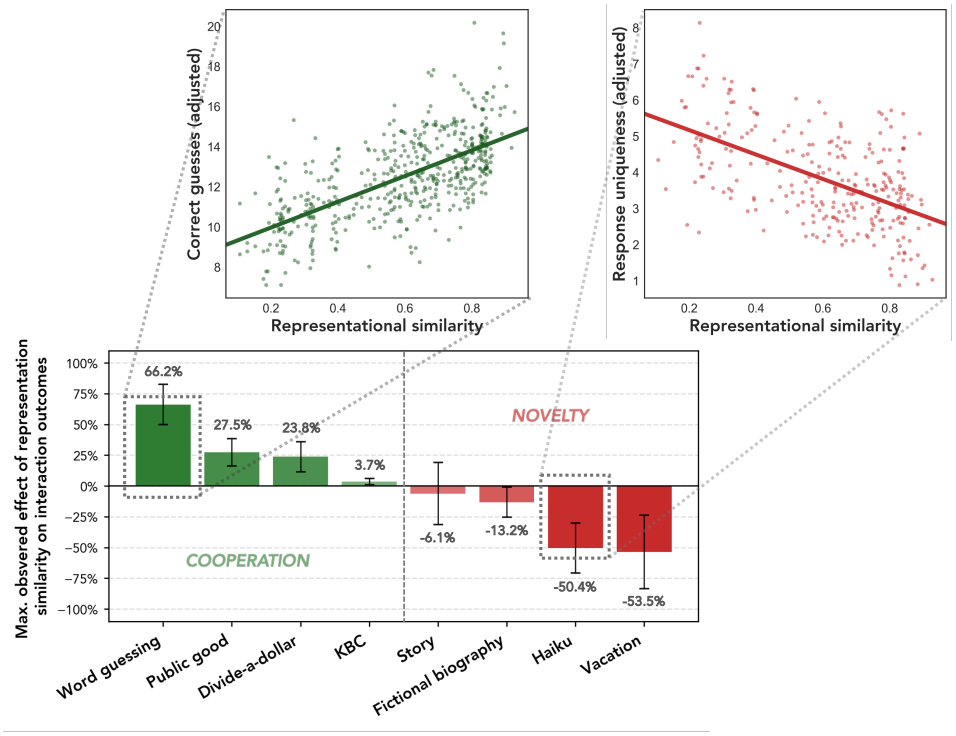

Figure 1: The effect of representational similarity on each game outcome. Representational similarity is quantified using linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019) with WikiText (Merity et al., 2016). In the bar graph, the effect size reflects the relative change (%) in predicted outcomes between the lowest and the highest observed value of representational similarity, with error bars indicating 95% confidence intervals. In the scatter plots, each point represents a model pair, and the y values are adjusted via mixed-effects regression to control for model-specific tendencies, thereby isolating the effect of representational similarity on interaction outcomes. Overall, greater similarity corresponds to higher cooperation but lower novelty.

where one player attempts to identify their partner's secret word, correct guesses increase by roughly 66.2% (relative change) as representational similarity rises from the minimum to the maximum observed values. By contrast, novelty declines consistently across the four creative tasks, though the magnitude and statistical significance vary. These findings suggest a likely tradeoff: model pairs with higher representational similarity tend to cooperate better, but also manifest reduced collective novelty. These results provide new insights into the design of multi-agent systems, where single-model deployment is currently the dominant paradigm.

2 Related Work

Neural Similarity as a Predictor of Interaction in Humans and Models. In neuroscience, similar neural responses between humans significantly predict social dynamics: Parkinson et al. (2018) found that friendship formation is predicted by similarity of neural response patterns to videos, as measured by fMRI. Shen et al. (2025b) extended this analysis to the similarity of neural activations during story-listening. Others have also shown that consistent neural activity patterns appear among personally familiar individuals (Thornton & Mitchell, 2017; Hyon et al., 2020). Such consistent findings suggest that greater neural similarity may facilitate stronger social bonds. Moreover, a related body of research has investigated the relationship between neural similarity and cooperative performance (Cui et al., 2012; Hu et al., 2018; Reinero et al., 2021; Réveillé et al., 2024), where it has been consistently found that higher interbrain synchrony is positively associated with cooperation.

As AI models scale in size and improve in performance, their internal representations increasingly align with human neural activity patterns (Goldstein et al., 2022; Schrimpf et al., 2021; Caucheteux & King, 2022; Shen et al., 2025a; Gurnee et al., 2023). For example, Caucheteux & King (2022) showed

that language algorithms predicting words exhibit representational patterns similar to brain responses to sentences. Shen et al. (2025a) reported a strong correlation between brain-model similarity scores and model performance across both language models and vision models. This growing evidence of brain–model alignment motivates our central hypothesis. It raises the possibility that principles observed in human cognition and social behavior may also extend to advanced artificial systems.

Representational Similarity in Neural Networks. Researchers have long sought to understand the behavior of neural networks by comparing their internal representations. A variety of metrics for representational similarity in artificial neural networks have been proposed (Kornblith et al., 2019; Hotelling, 1992; Morcos et al., 2018; Raghu et al., 2017; Kriegeskorte et al., 2008). One widely used method is Centered Kernel Alignment (CKA; Kornblith et al., 2019), which enables comparison of representations between models regardless of their architecture or layer count. Several studies have investigated the nuances of applying these metrics. Ding et al. (2021) evaluated the sensitivity of similarity measures to changes in model behavior and showed that different metrics exhibit distinct weaknesses. Moschella et al. (2023) demonstrated that representational similarity can serve as a strong predictor of model performance, in tasks such as classification with vision models.

Beyond standard similarity metrics, new approaches have been proposed for comparing representation spaces. For example, *model stitching*—connecting two neural networks—has been argued to capture aspects of representational structure that metrics like CKA cannot (Lenc & Vedaldi, 2015; Bansal et al., 2021). In this view, models with greater similarity are expected to achieve higher stitching success. Hacohen & Weinshall (2020) proposed comparing the similarity of classification predictions in vision models as an alternative perspective on model comparison.

Diversity, Creativity, and Collective Intelligence. Behavioral research on innovation finds that higher diversity within a group of collaborators leads to increased novelty in their creations. For example, Uzzi et al. (2013) analyzed millions of scientific papers and found that the highest-impact science often arises from groups that combined existing research in novel ways. Page (2019) formalizes this and proves that functionally diverse groups outperform homogeneous ones on complex problems, demonstrating superior problem-solving, innovation, and prediction accuracy. Similarly, Paulus (2000) showed that the effectiveness of brainstorming depends on cognitive diversity—that is, differences in how individuals perceive and think. Our results empirically explore this human-inspired principle: Does representational diversity within sets of LLM agents predict greater novelty in multi-agent creative tasks?

3 Representational Similarity of LLMs

To test our hypothesis: whether there is a relationship between representational similarity and interactive behavior of models, we first need a way to measure representational similarity of LLMs. In this section, we describe how we compute this similarity. *It is important to note that CKA computation is conducted independently of model interaction.*

3.1 Similarity metrics

Representational similarity quantifies how similarly two neural models embed the same inputs. Measuring similarity involves two steps: **1)** extracting representational vectors from each model using a probe dataset (i.e., a set of prompts) and **2)** computing a similarity score between the extracted representations using a metric.

Step 1. Extracting representations. The first step can be formalized as follows. The probe dataset $\mathcal{D} \subset \mathcal{X}$ contains m texts $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of all possible texts. Thus, $\mathcal{D} = \{x_i\}_{i=1}^{m}$. For a neural model with parameters θ, we define $f_\theta^k : \mathcal{X} \to \mathbb{R}^n$ as the mapping from a text $x \in \mathcal{X}$ to an n-dimensional activation at the k-th layer, where $1 \leq k \leq l$ and the model has l layers. Stacking the embeddings for all $x \in \mathcal{D}$ yields a matrix $R_\theta^k \in \mathbb{R}^{m \times n}$, with the i-th row equal to $f_\theta^k(x_i)$.

Step 2. Computing similarity. The next step is to compute similarity between the representational spaces $\{R_{\theta_1}^i\}_{1 \leq i \leq l_1}$ and $\{R_{\theta_2}^j\}_{1 \leq j \leq l_2}$, for two models with parameters θ_1, θ_2, depths l_1, l_2, and hidden dimensions n_1, n_2. A variety of similarity metrics (M) have been proposed, including Centered Kernel Alignment (CKA; Kornblith et al., 2019), Canonical Correlation Analysis (CCA; Hotelling, 1992; Morcos et al., 2018), Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017),

and Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008). Each defines a function $M : \mathbb{R}^{m \times n_1} \times \mathbb{R}^{m \times n_2} \to \mathbb{R}$ that takes two matrices (i.e., $R^i_{\theta_1}$ and $R^j_{\theta_2}$) as input.

We use CKA (Kornblith et al., 2019) given its popularity of use in prior work (Ciernik et al., 2024; Shen et al., 2025a; Liu et al., 2025). CKA enables the comparison between two models with different architectures and different numbers of layers. There are four common CKA variants: linear CKA, RBF CKA, unbiased linear CKA, and unbiased RBF CKA. These CKA values range in $[0,1]$, with higher values indicating greater similarity. Following prior work (Liu et al., 2025; Shen et al., 2025a; Zou et al., 2023; Raffel et al., 2020), we obtain $f^i_{\theta_1}(x)$ and $f^j_{\theta_2}(x)$ from the activation of the last token of each input x at their respective layers. CKA scores are then calculated for every layer pair of the two models. That is, $\text{CKA}(R^i_{\theta_1}, R^j_{\theta_2})$ for all $1 \leq i \leq l_1$ and $1 \leq j \leq l_2$, producing an $l_1 \times l_2$ grid of scores.

To summarize similarity with a single score per model pair, there are multiple approaches. The first approach is to average the CKA scores (i.e., global average): $\frac{\sum_{i,j} \text{CKA}(R^i_{\theta_1}, R^j_{\theta_2})}{l_1 \cdot l_2}$. This captures overall similarity between all layers of the two models. Please note that identical model pairs can score below 1, since off-diagonal layer pairs ($i \neq j$) yield values less than 1. An alternative summary measure of CKA is the layer-wise maximum-aligned average, which captures how well each layer aligns with its best-matching layer in the other model. That is,

$$\frac{1}{2} \times \left(\frac{\sum_i \max_j \text{CKA}(R^i_{\theta_1}, R^j_{\theta_2})}{l_1} + \frac{\sum_j \max_i \text{CKA}(R^i_{\theta_1}, R^j_{\theta_2})}{l_2} \right).$$

With this measure, identical model pairs always achieve 1, since each layer's best match is itself and $\text{CKA}(R^i_\theta, R^i_\theta) = 1$.

We observe consistent trends between representational similarity and interactive behavior across both aggregation methods and all four CKA variants. Unless otherwise noted, CKA refers to the global averages of linear CKA. Results for other variants appear in Appendix D.

3.2 Probe dataset

A recent study (Ciernik et al., 2024) shows that representational similarity can depend on the choice of probe dataset. To examine whether the relationship between representational similarity and model interactions depends on probe dataset, we use four probe datasets spanning different domains, from which we compute a CKA score for each: WikiText (Merity et al., 2016) for general language, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) for mathematics, and TruthfulQA (Lin et al., 2021) for truthfulness. From WikiText and MATH, we randomly sample 1000 prompts each. For GSM8K, we use the entire test set (1319 prompts), and for TruthfulQA, we use the full dataset (817 prompts).

3.3 Representational similarity range

We consider 23 open-weight LLMs spanning eight families and sizes ranging from 1B to 72B parameters, yielding 276 model pairs. The full model list is provided in Table 1 in Appendix. These models exhibit a wide range of representational similarity. For example, using the global average score with WikiText as the probe dataset, values range from 0.106 (`gemma-3-4b-it` vs. `gemma-3-12b-it`) to 0.92 (`phi-4` vs. `phi-4`). Using the average of maximum-aligned scores, values range from 0.288 (`gemma-3-4b-it` vs. `gemma-3-12b-it`) to 1 (for all identical model pairs). The complete set of CKA scores for all 276 pairs is shown in Figures 6 and 7. We find that the Gemma family (Team et al., 2025) generally exhibits lower similarity to other models, while pairs within the same family tend to show higher similarity. Figure 8 also reports correlations across different CKA variants, where similarities computed with GSM8K and WikiText display relatively lower agreement.

4 Cooperation increases when similar models meet

Building on evidence that greater interbrain synchrony among humans is strongly linked to increased cooperation, we test the hypothesis that model pairs with higher representational similarity will demonstrate increased cooperative behavior.

4.1 GAME SETTINGS & ANALYSIS

We use four game settings that involve cooperation: word guessing (Gero et al., 2020; Shaikh et al., 2023), public goods (Hauert et al., 2006), divide-a-dollar (Kalai, 1977), and the Keynesian Beauty Contest (KBC) (Duffy & Nagel, 1997). Word games have been used to examine how players infer their partners' mental models. The latter three games have been widely adopted in economics and social science to study cooperative dynamics. The following presents a description of each game rule along with associated outcome metrics, which capture the extent to which the two agents cooperated with one another during a game:

- **Word Guessing:** In the game, one player chooses their own target word that begins with a given letter ("a" to "z") and provides one-word hint to the other player. Here, the player is instructed to make the hint different from the target word. The other player should guess that secret word based on the hint and the given starting letter. Each round is one-shot and independent. We use the number of correct guesses over 26 rounds, one for each letter in the alphabet, as the outcome metric.
- **Public Good:** The game repeats for five rounds and shows an agents ability to reason over individual and group incentives. At the beginning of the game, each agent begins with \$100 of their own money and decides how much to contribute to a public pot every round. After their contribution is collected into the public pot each round, the value increases by 30% and is evenly redistributed back to each agent. We use each agent's total asset value accumulated over five rounds as the outcome metric.
- **Divide a Dollar:** The game repeats for five rounds, and players must make collaborative decisions in order to maximize self-gain. Each round, \$1 is available, and players should demand how much of the \$1 they want. If the total amount requested is not above \$1, players receive the amount they requested. If the total amount requested exceeds \$1, agents don't receive anything. We use each agent's total asset value accumulated over five rounds as the outcome metric.
- **KBC:** The game repeats for five rounds and incentivizes recursively reasoning about the other player's reasoning process and decisions. At the beginning of each round, players choose a number between 0 and 100, guessing the closest number to $2/3$ of the average of the numbers from both agents. The score is based on how close their number is to $2/3$ of the average: $100 - |\text{their number} - 2/3 \times \text{average}|$. We use the total score of each player over the five rounds as the outcome metric.

In all games except the word guessing game where each round is one-shot and independent, players are shown the other's choice and reasoning at the end of each round. A higher game outcome value indicates stronger cooperation in that game. For example, in the word guessing game, performance depends on how accurately each agent guesses the other's secret words—reflecting their ability to interpret their partner and infer unknown information. In the public goods game, achieving high returns requires both cooperation and alignment: purely selfish strategies yield low payoffs, and exploitation due to misunderstanding also reduces outcomes.

We evaluate all 276 possible pairs of the 23 models listed in Table 1. Because the word guessing game is asymmetric, we consider ordered pairs, resulting in 529 model pairings. Each pair interacts across all four games, with temperature set to 0.7 and at least 4 independent samples collected per pair for each game. The average game outcome for each model is presented in Figure 9 in Appendix C.

To analyze the relationship between representational similarity and interaction outcomes, we fit a mixed-effects linear regression model (Bates et al., 2015). In our experimental setup, using a simple linear regression or Pearson correlation would be inappropriate because these tests assume independent data points, whereas our setup produces multiple samples per model pair, and each model appears in multiple pairs. Mixed-effects regression is the standard approach for handling such non-independence (Brown, 2021). In particular, it allows us to account for variance attributable to individual models (e.g., differences in capability) by including model-specific random effects, thereby isolating the effect of representational similarity on interactive outcomes. Specifically, we estimate the following mixed-effects regression:

$$Y_{ij} = \alpha + \beta \cdot \text{CKA}_{ij} + u_i + v_j + \epsilon_{ij},$$

where Y_{ij} is the interactive outcome of interest, and CKA_{ij} is the similarity measure between models i and j. The terms u_i and v_j represent random effects associated with models i and j, respectively, where these terms capture unobserved heterogeneity at the level of model i and model j, respectively. Lastly, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is an error term.

To evaluate whether similarity predicts the interactive outcome, the key quantities are the estimated slope of CKA_{ij} (i.e., β) and its statistical significance. We therefore report β with its p-value throughout the paper.

4.2 Does representational similarity predict cooperation?

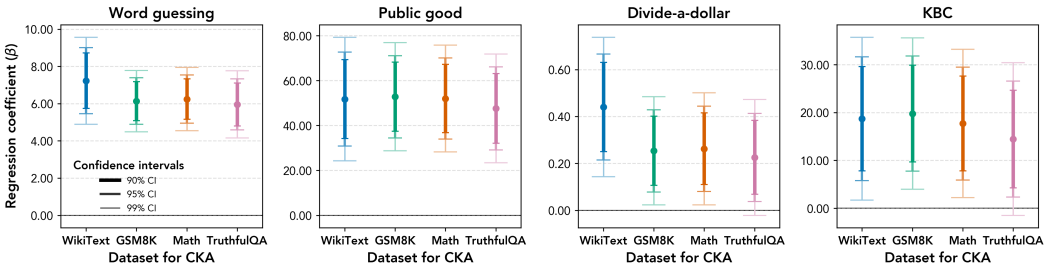

Figure 2: Regression coefficient of representational similarity on game outcomes. Error bars denote 90%, 95%, and 99% confidence intervals. Across games and datasets, the graphs show a positive effect of similarity on outcomes, with WikiText-based similarity exhibiting the strongest effect.

Our results reveal that representational similarity can predict cooperative outcomes. In the word guessing game, correct guesses increase by approximately 88.2, 58.0, 63.3, and 59.4% (relative changes) for each unit increase in representational similarity (i.e., from 0 to 1) measured with WikiText, GSM8K, MATH, and TruthfulQA, respectively. In the public good game, each player's total asset value rises by 34.8, 32.4, 33.0, and 29.8% across the four probe datasets. For divide-a-dollar, asset values increase by 29.9, 15.3, 16.2, and 13.7%. Finally, in KBC, scores increase significantly but modestly—4.5, 4.7, 4.2, and 3.4%. All effects are found to be statistically significant (Figure 2). Among the four games, KBC shows the weakest effect. This is expected: the game has a unique Nash equilibrium in which both players always choose zero, which makes the optimal strategy fixed regardless of representational similarity. For instance, we observe that a certain model such as `GPT-OSS-20B` always chooses 0 regardless of the partner's decision. Nevertheless, even here, we observe a significant upward trend with increasing similarity.

The pattern persists across probe datasets, implying its generalizability. Moreover, we find no difference in effect size across datasets (please refer to Figure 2). This contrasts with a previous finding Ciernik et al. (2024), which showed that the correspondence between representational similarity and task behavior depends on the dataset. The same trend holds across other CKA variants as well (see Appendix D.1).

5 Novelty decreases when similar models meet

Next, we examine whether representational similarity predicts novelty in collaborative generative tasks. For this purpose, we adapt four tasks—story writing, fictional biography, haiku composition, and vacation benefit brainstorming—from NoveltyBench (Zhang et al., 2025), a benchmark originally designed to evaluate an individual model's ability to produce high-quality and original ideas. Because NoveltyBench tasks are defined for single-agent settings, we extend them to the multi-agent case: each of the two models first generates a set of brainstorming ideas, after which each model produces a final output based on the combined brainstorms. The four generative tasks are as follows:

- **Story Writing:** Players brainstorm an outline of a story about a girl and her dog, then individually write a five-sentence story after reviewing the combined brainstorm.
- **Biography Writing:** Players brainstorm an outline for a short biography of a fictional person, then individually write a biography based on the combined brainstorm.
- **Haiku Writing:** Players brainstorm a plot for a haiku about a whale and a walnut tree, then individually compose a haiku after reviewing the combined brainstorm.
- **Vacation Benefit Brainstorming:** Players brainstorm possible benefits of going on vacation, then individually write one best aspect after reviewing the combined brainstorm.

As with the cooperative games, we evaluate all 276 pairs, using a temperature of 0.7. For each pair, we sample 10 generations in accordance with NoveltyBench. We also conduct mixed-effects regression to identify whether representational similarity can predict novelty.

Because novelty encompasses multiple dimensions, we evaluate it using several metrics: the number of distinct responses produced, the quality of those responses, and the extent to which outputs differ from those generated without interaction with another agent. The first two, response uniqueness and quality,

are assessed using NoveltyBench's proposed metrics, while the last is measured as the mutual information between outputs produced through joint brainstorming and those generated without interaction. We describe the evaluation methodologies in more detail below.

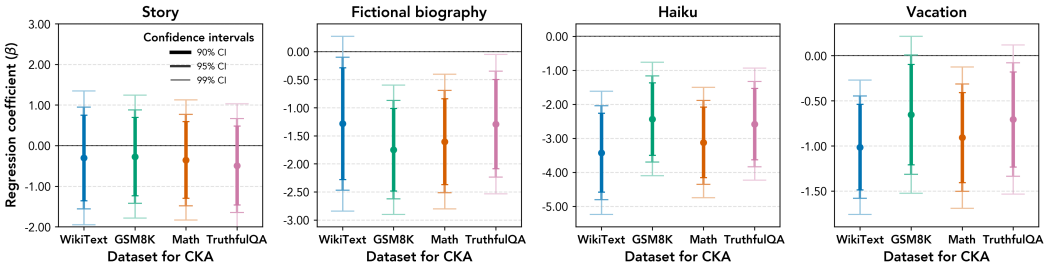

Figure 3: Regression coefficient of representational similarity on response uniqueness. Error bars denote 90%, 95%, and 99% confidence intervals. The graphs reveal a consistent downward trend: as models become more similar, response uniqueness declines. The strongest effect is observed in the haiku task.

5.1 Does representational similarity predict uniqueness and quality?

First, to assess response uniqueness and quality, we use the NoveltyBench evaluation pipeline using autoraters (Zhang et al., 2025). NoveltyBench defines the two measures, uniqueness and quality, over a set of samples. For uniqueness, the benchmark clusters 10 generations using a fine-tuned `deberta-v3-large` model according to content distinctiveness and then counts the number of clusters, which serves as the uniqueness metric. A higher cluster count indicates that models are able to generate more diverse ideas. For response quality, the benchmark relies on `Skywork-Reward-Gemma-2-27B-v0.2` (Liu et al., 2024), with outputs rescaled to a $1-10$ range for a more interpretable score.[1]

As shown in Figure 3, response uniqueness decreases consistently with increasing representational similarity across all tasks and probe datasets. The effect is strongest in haiku composition (coeff $=-3.425$, CI $=[-4.803,-2.047]$, $p<.001$). By contrast, response quality shows no systematic trend with similarity. Fictional biography and haiku tasks exhibit nonsignificant negative slopes of similarity (coeff $=-0.397$, $p=.456$ for biography; coeff $=-0.115$, $p=.724$ for haiku), while story writing and vacation tasks show a nonsignificant positive slope (coeff $=0.279$, $p=.420$ for story; coeff $=0.039$, $p=.901$ for vacation). This implies that interaction with dissimilar models tend to generate more diverse responses, without reducing quality.

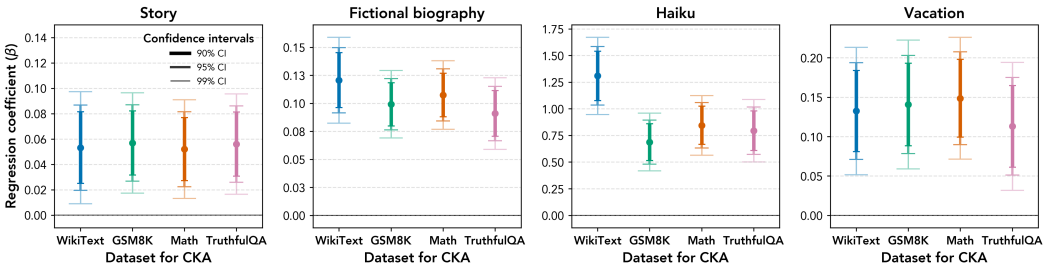

Figure 4: Regression coefficients of representational similarity on mutual information. Error bars denote 90%, 95%, and 99% confidence intervals. The graphs reveal a decreasing trend in novelty with increasing similarity: as models become more similar, the shared information between the amalgam response (i.e., response generated after joint brainstorming) and the individual response (i.e., response generated after solo brainstorming) increases.

5.2 Does representational similarity predict mutual information?

We next examine whether representational similarity has a significant effect on output novelty—specifically, how far a model's responses generated after joint brainstorming ("amalgam response") deviate from the

[1] None of the helper models used in this section are reused as players in the games.

model's outputs conditioned only on its individual brainstorm ("individual response"). To capture this, we compare amalgam and individual responses using mutual information (Kraskov et al., 2004), which quantifies how much information individual responses share with those produced in the joint setting. Such information-theoretic approaches have recently been applied to investigate textual characteristics (e.g., information distribution across paragraphs) (Venkatraman et al., 2023; Clark et al., 2023; Mu et al., 2025).

To calculate mutual information, we follow the method from Mu et al. (2025). Formally, let S_A denote an amalgam response and S_I denote an individual response. We compute the mutual information $I(S_A;S_I)$ as $H_\theta(S_A) - H_\theta(S_A \mid S_I)$. $H_\theta(S_A)$ denotes the total information content of the amalgam response, and $H_\theta(S_A \mid S_I)$ denotes the residual information of the amalgam response given the individual response, both measured under a reference language model with parameters θ. To calculate $H_\theta(S_A)$, we sum the cross-entropy over all tokens in the amalgam response under the model with parameters θ. The cross-entropy of a token quantifies the model's prediction error for that token given its preceding context, thereby reflecting its uncertainty. Similarly, $H_\theta(S_A \mid S_I)$ is computed by summing the cross-entropy of each token in the amalgam response when the individual response is prefixed to the amalgam response. A smaller difference between $H_\theta(S_A)$ and $H_\theta(S_A \mid S_I)$ indicates that the amalgam response deviates more from the individual response, thereby reflecting higher novelty. Following Mu et al. (2025), we use `Llama-3.1-8B-Instruct` as the reference model.[2]

Our analysis shows a significant positive effect of representational similarity on mutual information, which suggests that interactions between more similar models generate less novel outputs with respect to the individual model responses. The trend appears across all tasks and probe datasets (Figure 4). In particular, the haiku task exhibits the strongest effect of representational similarity on mutual information (coeff $= 1.310$, CI $= [1.034, 1.585]$, $p < .001$).

6 Why does the trend appear?

So far, we have identified a strong trend between representational similarity and interactive behaviors of models. This naturally raises the question of why such a trend emerges. In this section, we test several hypotheses regarding what drives the trend.

Confounding Effects of Behavioral Similarity. Models with higher representational similarity may behave more similarly (e.g., bid the same amount in divide-a-dollar), and this behavioral similarity might have led directly to greater measured cooperation. To test this, we conducted a mixed-effects regression controlling for behavioral differences in the public goods, divide-a-dollar, and KBC games. This allows us to isolate the effect of representational similarity from behavioral similarity. Because these games instruct models to make numerical choices, it is straightforward to estimate behavioral difference as the absolute gap between the two models' choices.

Our analysis shows that behavioral difference alone cannot explain the observed trends. In both the public good and divide-a-dollar games, representational similarity remains a significant predictor, while behavioral difference is insignificant (coeff. rep. sim. $= 52.118$, $p < .001$, coeff. beh. diff. $= -0.036$, $p = .086$ for public good; coeff. rep. sim. $= 0.435$, $p < .001$, coeff. beh. diff. $= -0.020$, $p = .281$ for divide-a-dollar). This suggests that behavioral similarity is not what drives the trend. By contrast, in the KBC game, behavioral difference shows a significant effect, while representational similarity does not (coeff. rep. sim. $= 9.024$, $p = .178$, coeff. beh. diff. $= -0.327$, $p < .001$). As discussed in Section 4, KBC has a unique Nash equilibrium in which both players choose 0, which leads to convergence in choices. This structural property of the game likely explains why behavioral difference dominates in this case.

Factors Underlying Representational Similarity. Representational similarity is influenced by several architectural and design-related components, including whether two models are identical, belong to the same model family, share the same tokenizer, or differ in size. Any of these factors could potentially have driven the observed behavioral trends by influencing similarity. To investigate this, we conducted a mixed-effects regression controlling for four key factors: (1) identical model pairing, (2) within-family

[2]They select the model under the requirement that the mutual information values satisfy symmetry and non-negativity. For robustness, we additionally compute mutual information with the base model `Llama-3.1-8B`, identified by Mu et al. (2025) as a strong alternative reference model. Results, shown in Appendix D.4, continue to show a significant association with representational similarity. One might further suspect a same-family bias when the reference model is used to evaluate Llama models. To address this, we analyze model pairs excluding the Llama family and report the results in Appendix D.5. The results still show the significant effect of similarity.

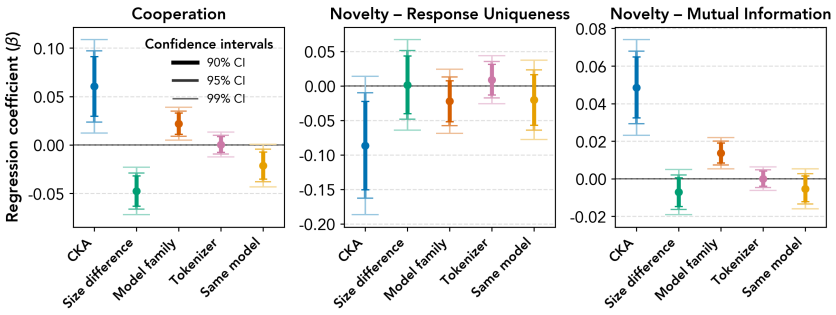

Figure 5: Regression coefficients of representational similarity (CKA) and four other factors across cooperation and novelty games. To enable comparison across predictors, all variables were rescaled to $[0,1]$ in the regression. The graphs show that representational similarity is the strongest predictor.

pairing, (3) shared tokenizer, and (4) model size difference. In this analysis, all predictors and outcome variables were rescaled to $[0,1]$ to allow comparison of effect sizes of predictors. If these factors were mainly responsible for the trend, we would expect representational similarity to lose significance once they were controlled for, while the factors themselves would show significant effects.

Our analysis finds that representational similarity is the strongest predictor on cooperation and novelty, compared to the four factors (Figure 5). In cooperation games, all predictors except tokenizer are significant, with representational similarity showing the largest effect (coeff $= 0.060$, $p = .001$). For response uniqueness, none of the four factors are significant, while representational similarity shows a significant effect (coeff $= -0.087$, $p = .026$). For mutual information, only representational similarity and within-family are significant, with similarity again showing the stronger impact (coeff $= 0.049$, $p < .001$). Taken together, these findings suggest that the four examined factors do not fully explain the trend. Instead, representational similarity itself—likely influenced by deeper, unmeasured aspects of model design and training—remains the primary driver of the observed behavioral patterns. Further exploration of the reason for the trend is left for future work.

7 Future Directions and Open Questions

Existing multi-agent system designs often rely on a single model without exploring the optimal combination of models (Lai et al., 2024; Park et al., 2023; Xie et al., 2024; Zhou et al., 2023; Wu et al., 2024; Ishibashi & Nishimura, 2024). Our findings suggest that which models are combined has a significant effect on their interactions. In neuroscience and social science, researchers have long studied the nature of human social dynamics (Parkinson et al., 2018; Thornton & Mitchell, 2017; Shen et al., 2025b; Reinero et al., 2021; Page, 2019; Paulus, 2000). We argue that such efforts should also be made in the AI community, and our experiments provide an initial step in that direction.

The relationship between representational similarity and model interaction is likely context-dependent. We already observed that the effect size of similarity varies across games. For instance, in KBC, which has a unique Nash equilibrium, the link between similarity and interaction becomes weaker. Other evidence is also found in neuroscience and social science. Some studies show that diversity can foster cooperation (Santos et al., 2008; 2012; Wang et al., 2025), and certain creativity research suggests that greater similarity can yield higher originality (Koo et al., 2024; Bastian et al., 2018; Miura & Hida, 2004). These findings imply that there might be no universal relationship between similarity and interactive dynamics. Understanding when the trend emerges, when it disappears, and when it reverses will require further research. Such insights will be crucial for improving multi-agent system design in diverse application domains.

Another direction is to investigate the mechanisms underlying these trends. In this work, we used CKA as our measure of representational similarity. However, metrics like CKA capture only limited aspects of representational spaces, making it difficult to pinpoint which specific features of representations drive the trends. Future work can examine this at the neuron level—e.g., which neurons are preferentially activated when a model interacts with another model of higher representational similarity. Such analyses could enable us to deliberately steer cooperation or collective novelty through targeted activation steering.

Ethics statement & Reproducibility statement

This paper follows the ICLR Code of Ethics. Our goal is to contribute to the design of multi-agent systems, which we believe can benefit society by enabling more effective and cooperative AI applications. We also used LLMs for typo and grammar checks during manuscript preparation. To support reproducibility, we will publicly release all datasets, code, and evaluation scripts used in this work upon acceptance.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl. Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL https://arxiv.org/abs/2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann. Phi-4 technical report, 2024b. URL https://arxiv.org/abs/2412.08905.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic. The falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Yamini Bansal, Preetum Nakkiran and Boaz Barak. Revisiting Model Stitching to Compare Neural Representations. *Advances in Neural Information Processing Systems*, 34:225–236, 2021.

Brock Bastian, Jolanda Jetten, Hannibal A Thai and Niklas K Steffens. Shared adversity increases team creativity through fostering supportive interaction. *Frontiers in Psychology*, 9:2309, 2018.

Douglas Bates, Martin Mächler, Ben Bolker and Steve Walker. Fitting Linear Mixed-Effects Models using lme4. *Journal of statistical software*, 67:1–48, 2015.

Violet A Brown. An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351, 2021.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Forty-second International Conference on Machine Learning*, 2024.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell and Roger Levy. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Xu Cui, Daniel M Bryant and Allan L Reiss. NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *Neuroimage*, 59(3):2430–2437, 2012.

Christian Schroeder de Witt. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. *arXiv preprint arXiv:2505.02077*, 2025.

Frances Ding, Jean-Stanislas Denain and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

John Duffy and Rosemarie Nagel. On the Robustness of Behaviour in Experimental 'Beauty Contest' Games. *The Economic Journal*, 107(445):1684–1700, 1997.

Jacob Eisenstein, Reza Aghajani, Adam Fisch, Dheeru Dua, Fantine Huot, Mirella Lapata, Vicky Zayats and Jonathan Berant. Don't lie to your friends: Learning what you know from collaborative self-play. *arXiv preprint arXiv:2503.14481*, 2025.

Kazuma Fukumura and Takayuki Ito. Can LLM-Powered Multi-Agent Systems Augment Human Creativity? Evidence from Brainstorming Tasks. In *Proceedings of the ACM Collective Intelligence Conference*, pp. 20–29, 2025.

Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1–12, 2020.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Amir Cohen. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Guy Hacohen and Daphna Weinshall. Let's Agree to Agree: Neural Networks Share Classification Order on Real Datasets. *International Conference on Machine Learning*, pp. 3950–3960, 2020.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak. Multi-Agent Risks from Advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.

Christoph Hauert, Miranda Holmes and Michael Doebeli. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2565–2571, 2006.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Sylvia Ann Hewlett, Melinda Marshall, Laura Sherbin and others. How diversity can drive innovation. *Harvard business review*, 91(12):30–30, 2013.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Yi Hu, Yinying Hu, Xianchun Li, Yafeng Pan and Xiaojun Cheng. Brain-to-brain synchronization across two persons predicts mutual prosociality. *Social Cognitive and Affective Neuroscience*, 13(12): 1225–1233, 2018.

Ryan Hyon, Yoosik Youm, Junsol Kim, Jeanyung Chey, Seyul Kwak and Carolyn Parkinson. Similarity in functional brain connectivity at rest predicts interpersonal closeness in the social network of an entire village. *Proceedings of the National Academy of Sciences*, 117(52):33149–33160, 2020.

Yoichi Ishibashi and Yoshimasa Nishimura. Self-Organized Agents: A LLM Multi-Agent Framework toward Ultra Large-Scale Code Generation and Optimization. *arXiv preprint arXiv:2404.02183*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Ehud Kalai. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society*, pp. 1623–1630, 1977.

Chanhee Koo, Honghong Bai, Aoxin Luo and Stella Christie. How does comparing (dis) similar objects affect young children's creative idea generation? Exploring the role of diversity in facilitating creativity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Alexander Kraskov, Harald Stögbauer and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Nikolaus Kriegeskorte, Marieke Mur and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song and James Evans. Position: Evolving AI collectives enhance human diversity and enable self-regulation. In *Forty-first International Conference on Machine Learning*, 2024.

Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Stephanie Lin, Jacob Hilton and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu and Yahui Zhou. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024.

Xuyuan Liu, Lei Hsiung, Yaoqing Yang and Yujun Yan. Spectral Insights into Data-Oblivious Critical Layers in Large Language Models. *arXiv preprint arXiv:2506.00382*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury and Richard Socher. Pointer Sentinel Mixture Models, 2016.

Asako Miura and Misao Hida. Synergy between diversity and similarity in group-idea generation. *Small Group Research*, 35(5):540–564, 2004.

Ari Morcos, Maithra Raghu and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2023.

J. Mu, A. R. Preston and A. G. Huth. Efficient Uniform Sampling Explains Non-Uniform Memory of Narrative Stories. *bioRxiv*, pp. 2025.07.31.667952, 2025. doi: 10.1101/2025.07.31.667952. URL https://doi.org/10.1101/2025.07.31.667952. preprint.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Christian R Østergaard, Bram Timmermans and Kari Kristinsson. Does a different view create something new? The effect of employee diversity on innovation. *Research policy*, 40(3):500–509, 2011.

Scott E Page. *The diversity bonus: How great teams pay off in the knowledge economy*. Princeton University Press, 2019.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Carolyn Parkinson, Adam M Kleinbaum and Thalia Wheatley. Similar neural responses predict friendship. *Nature communications*, 9(1):332, 2018.

Paul Paulus. Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied psychology*, 49(2):237–262, 2000.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan and Rada Mihalcea. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *Advances in neural information processing systems*, 30, 2017.

Diego A Reinero, Suzanne Dikker and Jay J Van Bavel. Inter-brain synchrony in teams predicts collective performance. *Social cognitive and affective neuroscience*, 16(1-2):43–57, 2021.

Coralie Réveillé, Grégoire Vergotte, Stéphane Perrey and Grégoire Bosselut. Using interbrain synchrony to study teamwork: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 159:105593, 2024.

Francisco C Santos, Marta D Santos and Jorge M Pacheco. Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201):213–216, 2008.

Francisco C Santos, Flavio L Pinheiro, Tom Lenaerts and Jorge M Pacheco. The role of diversity in the evolution of cooperation. *Journal of theoretical biology*, 299:88–96, 2012.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45): e2105646118, 2021.

Omar Shaikh, Caleb Ziems, William Held, Aryan J Pariani, Fred Morstatter and Diyi Yang. Modeling Cross-Cultural Pragmatic Inference with Codenames Duet. *arXiv preprint arXiv:2306.02475*, 2023.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Qian Zhang and Yi Zeng. Alignment between Brains and AI: Evidence for Convergent Evolution across Modalities, Scales and Training Trajectories. *arXiv preprint arXiv:2507.01966*, 2025a.

Yixuan Lisa Shen, Ryan Hyon, Thalia Wheatley, Adam M Kleinbaum, Christopher L Welker and Carolyn Parkinson. Neural similarity predicts whether strangers become friends. *Nature Human Behaviour*, pp. 1–14, 2025b.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *arXiv preprint arXiv:2410.09403*, 2024.

Yashar Talebirad and Amirhossein Nadiri. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *arXiv preprint arXiv:2306.03314*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Mark A Thornton and Jason P Mitchell. Consistent neural activity patterns represent personally familiar people. *Journal of cognitive neuroscience*, 29(9):1583–1594, 2017.

Brian Uzzi, Satyam Mukherjee, Michael Stringer and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Saranya Venkatraman, Adaku Uchendu and Dongwon Lee. GPT-who: An Information Density-based Machine-Generated Text Detector. *arXiv preprint arXiv:2310.06202*, 2023.

Ben Wang, Linjiang Yang, Xinguo He, Yichen Yang and Haochun Yang. Interactive diversity promotes cooperation in multi-games. *The European Physical Journal B*, 98(5):94, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec and Jianfeng Gao. CollabLLM: From Passive Responders to Active Collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens. Can Large Language Model Agents Simulate Human Trust Behavior? *Advances in neural information processing systems*, 37:15674–15729, 2024.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang and Daphne Ippolito. NoveltyBench: Evaluating Language Models for Humanlike Diversity. *arXiv preprint arXiv:2504.05228*, 2025.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A MODEL LIST

Table 1: Full list of models involved in the interaction experiments

Family	Model Name	Checkpoint / Repo	Size	Tokenizer	Reference
Qwen	Qwen2.5-3B-Instruct	Qwen/Qwen2.5-3B-Instruct	3.09B	BPE	Team (2024)
	Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct	7.61B	BPE	Team (2024)
	Qwen2.5-14B-Instruct	Qwen/Qwen2.5-14B-Instruct	14.7B	BPE	Team (2024)
	Qwen2.5-72B-Instruct	Qwen/Qwen2.5-72B-Instruct	72.7B	BPE	Team (2024)
Llama	Llama-3.2-3B-Instruct	meta-llama/Llama-3.2-3B-Instruct	3.21B	tiktoken	Grattafiori et al. (2024)
	Llama-3.2-11B-Vision-Instruct	meta-llama/Llama-3.2-11B-Vision-Instruct	10.6B	tiktoken	Grattafiori et al. (2024)
	Llama-3.3-70B-Instruct	meta-llama/Llama-3.3-70B-Instruct	70.6B	tiktoken	Grattafiori et al. (2024)
Gemma	Gemma-3-1B-IT	google/gemma-3-1b-it	1.0B	SentencePiece	Team et al. (2025)
	Gemma-3-4B-IT	google/gemma-3-4b-it	4.0B	SentencePiece	Team et al. (2025)
	Gemma-3-12B-IT	google/gemma-3-12b-it	12.2B	SentencePiece	Team et al. (2025)
	Gemma-3-27B-IT	google/gemma-3-27b-it	27.0B	SentencePiece	Team et al. (2025)
Falcon	Falcon3-3B-Instruct	tiiuae/Falcon3-3B-Instruct	3.23B	BPE	Almazrouei et al. (2023)
	Falcon3-7B-Instruct	tiiuae/Falcon3-7B-Instruct	7.46B	BPE	Almazrouei et al. (2023)
	Falcon3-10B-Instruct	tiiuae/Falcon3-10B-Instruct	10.3B	BPE	Almazrouei et al. (2023)
Phi	Phi-3.5-mini-instruct	Lexius/Phi-3.5-mini-instruct	3.8B	SentencePiece	Abdin et al. (2024a)
	Phi-3-medium-128k-instruct	microsoft/Phi-3-medium-128k-instruct	14B	SentencePiece	Abdin et al. (2024a)
	Phi-4-mini-instruct	microsoft/Phi-4-mini-instruct	3.8B	tiktoken	Abdin et al. (2024b)
	Phi-4	microsoft/phi-4	14.7B	tiktoken	Abdin et al. (2024b)
Mistral	Mistral-Nemo-Instruct-2407	mistralai/Mistral-Nemo-Instruct-2407	12.2B	tekken	Jiang et al. (2024)
	Ministral-8B-Instruct-2410	mistralai/Ministral-8B-Instruct-2410	8.02B	tekken	Jiang et al. (2024)
OpenAI	GPT-OSS-20B	openai/gpt-oss-20b	21.5B	o200k_harmony	OpenAI et al. (2025)
OLMo	OLMo-2-1B-Instruct	allenai/OLMo-2-0425-1B-Instruct	1.48B	cl100k	OLMo et al. (2025)
	OLMo-2-13B-Instruct	allenai/OLMo-2-1124-13B-Instruct	13.7B	cl100k	OLMo et al. (2025)

B REPRESENTATIONAL SIMILARITY OF MODEL PAIRS

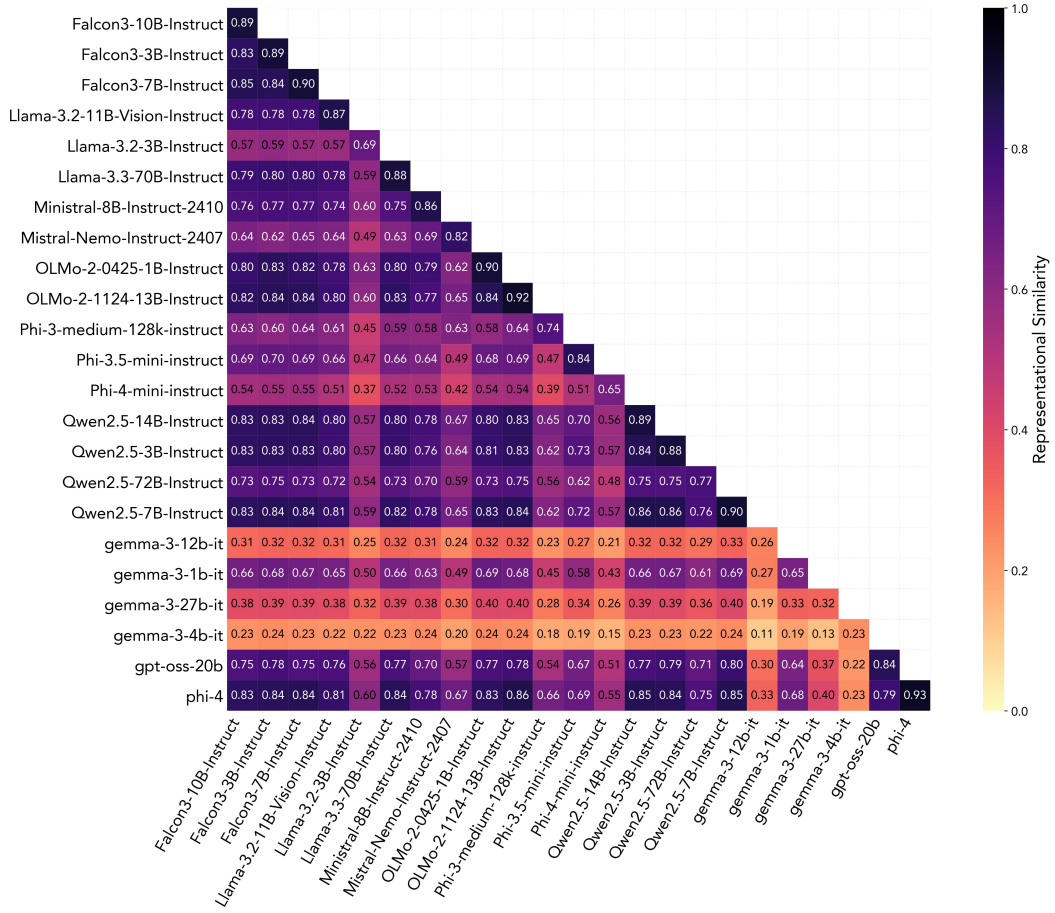

Figure 6: Representational similarity of model pairs using WikiText, where CKA scores across layers are aggregated into a global average.

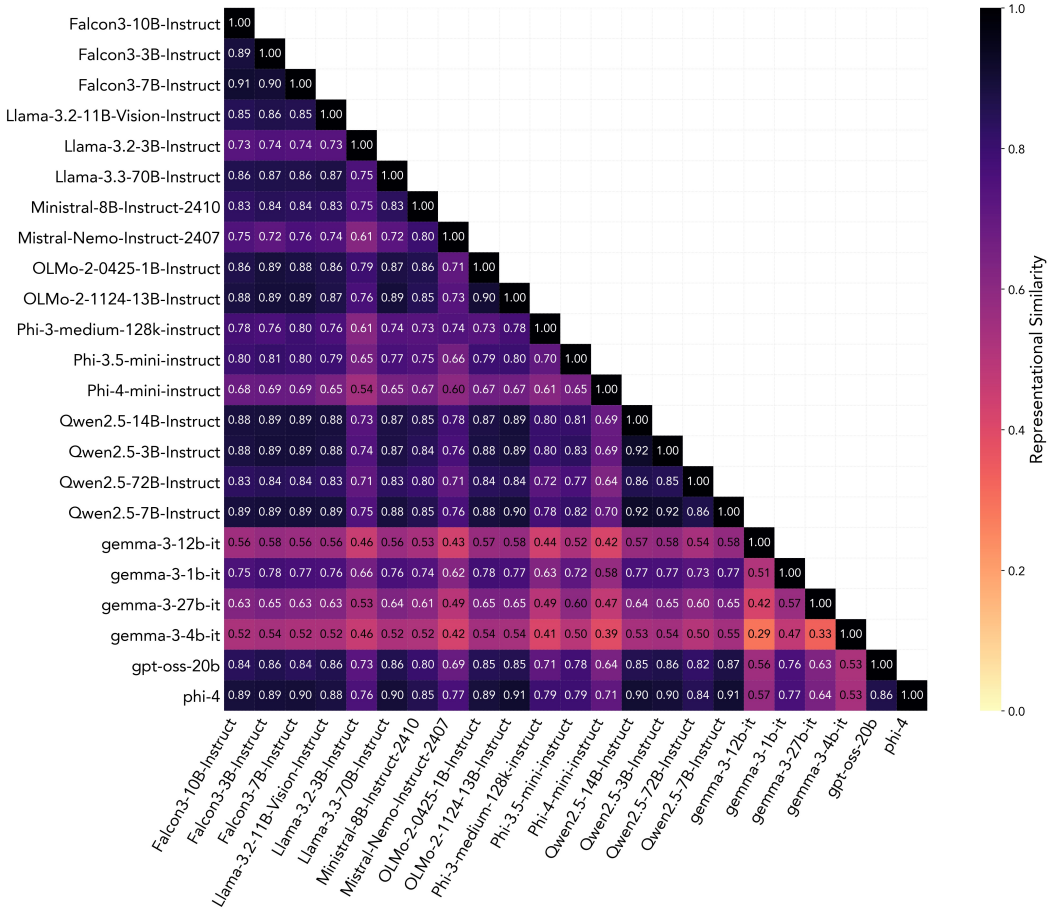

Figure 7: Representational similarity of model pairs using WikiText, where CKA scores across layers are aggregated into a maximum-aligned average.

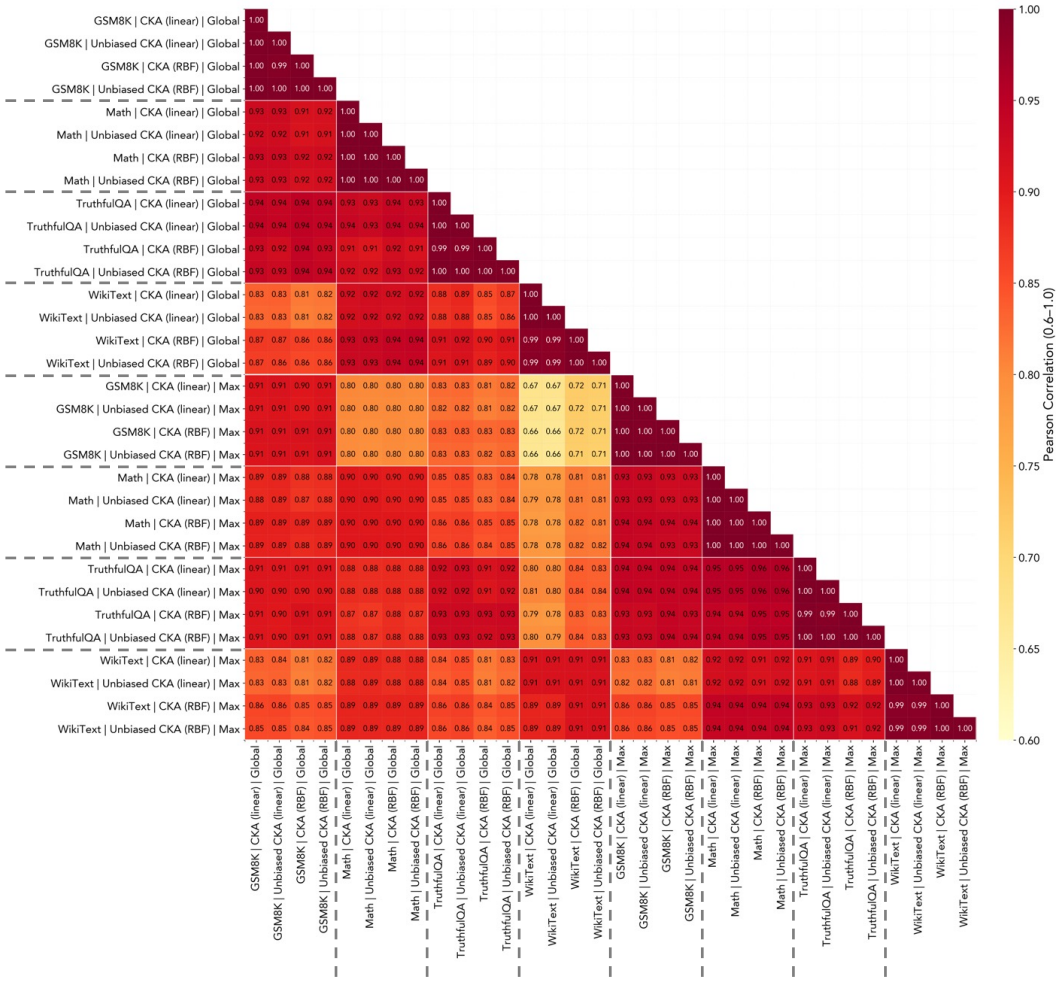

Figure 8: Correlation between different CKA scores

C MODEL PERFORMANCE DURING GAMES

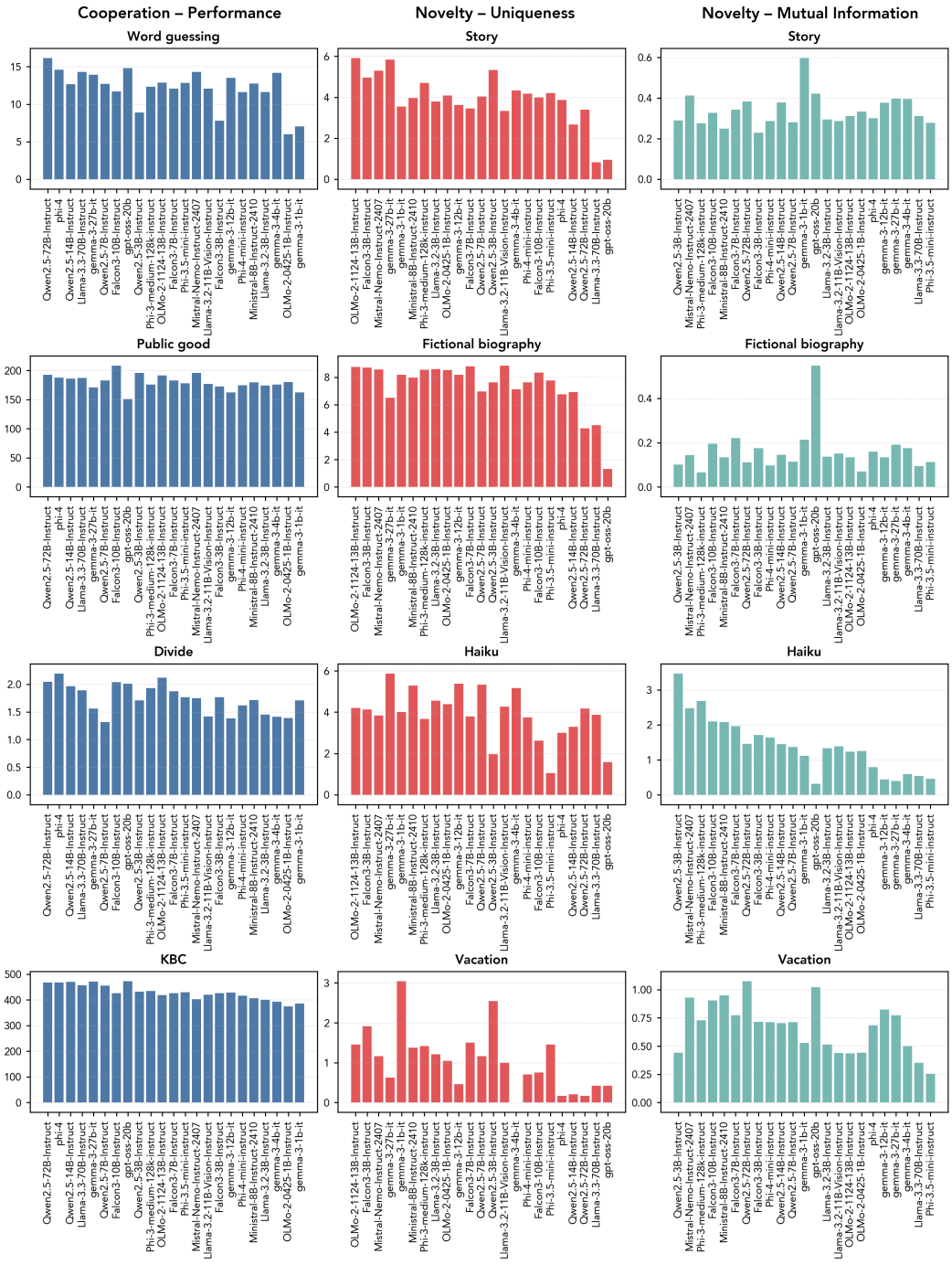

Figure 9: Average game outcomes for each model

D Detailed Results

D.1 Mixed-effects regression results for cooperation

Tables 2∼9 show strong positive trends across all datasets, CKA variants, and games.

D.1.1 When using the global average

Table 2: Word guessing game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	8.19	7.23	[5.45, 9.01]	1.5×10^{-15}
	unbiased CKA (linear)	8.26	7.16	[5.40, 8.92]	1.6×10^{-15}
	CKA (RBF)	8.26	7.33	[5.60, 9.06]	9.0×10^{-17}
	unbiased CKA (RBF)	7.76	7.98	[6.02, 9.94]	1.5×10^{-15}
GSM8K	CKA (linear)	10.58	6.13	[4.88, 7.38]	6.9×10^{-22}
	unbiased CKA (linear)	10.67	6.03	[4.80, 7.26]	9.3×10^{-22}
	CKA (RBF)	10.63	5.90	[4.68, 7.12]	3.4×10^{-21}
	unbiased CKA (RBF)	10.55	6.27	[4.97, 7.56]	2.4×10^{-21}
MATH	CKA (linear)	9.85	6.24	[4.94, 7.53]	3.2×10^{-21}
	unbiased CKA (linear)	9.95	6.13	[4.86, 7.41]	4.3×10^{-21}
	CKA (RBF)	9.71	6.25	[4.92, 7.57]	2.5×10^{-20}
	unbiased CKA (RBF)	9.75	6.51	[5.16, 7.87]	4.7×10^{-21}
TruthfulQA	CKA (linear)	10.03	5.96	[4.58, 7.33]	1.8×10^{-17}
	unbiased CKA (linear)	10.14	5.90	[4.54, 7.25]	1.3×10^{-17}
	CKA (RBF)	10.11	5.75	[4.39, 7.12]	1.6×10^{-16}
	unbiased CKA (RBF)	9.93	6.21	[4.76, 7.66]	4.8×10^{-17}

Table 3: Public good game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	148.95	51.77	[30.82, 72.73]	1.3×10^{-6}
	unbiased CKA (linear)	149.6	51.03	[30.44, 71.62]	1.2×10^{-6}
	CKA (RBF)	148.3	54.36	[32.95, 75.77]	6.5×10^{-7}
	unbiased CKA (RBF)	149.4	51.31	[29.66, 72.96]	3.4×10^{-6}
GSM8K	CKA (linear)	162.99	52.81	[34.48, 71.14]	1.6×10^{-8}
	unbiased CKA (linear)	163.8	52.03	[33.84, 70.21]	2.1×10^{-8}
	CKA (RBF)	162.9	52.44	[34.45, 70.44]	1.1×10^{-8}
	unbiased CKA (RBF)	163.1	52.82	[34.14, 71.51]	3.0×10^{-8}
MATH	CKA (linear)	157.53	52.01	[33.91, 70.10]	1.8×10^{-8}
	unbiased CKA (linear)	158.3	51.31	[33.42, 69.21]	1.9×10^{-8}
	CKA (RBF)	156.0	52.85	[34.48, 71.23]	1.7×10^{-8}
	unbiased CKA (RBF)	157.6	52.25	[33.65, 70.84]	3.6×10^{-8}
TruthfulQA	CKA (linear)	159.93	47.62	[29.16, 66.08]	4.3×10^{-7}
	unbiased CKA (linear)	160.8	47.06	[28.78, 65.34]	4.5×10^{-7}
	CKA (RBF)	160.7	45.83	[27.64, 64.02]	7.9×10^{-7}
	unbiased CKA (RBF)	160.1	47.55	[28.61, 66.49]	8.7×10^{-7}

Table 4: Divide-a-dollar game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	1.47	0.44	[0.21, 0.67]	0.00014
	unbiased CKA (linear)	1.47	0.44	[0.22, 0.67]	0.00011
	CKA (RBF)	1.48	0.45	[0.23, 0.67]	7.7×10^{-5}
	unbiased CKA (RBF)	1.46	0.46	[0.22, 0.70]	0.00017
GSM8K	CKA (linear)	1.66	0.25	[0.08, 0.43]	0.0046
	unbiased CKA (linear)	1.66	0.25	[0.08, 0.43]	0.0040
	CKA (RBF)	1.65	0.26	[0.09, 0.44]	0.0027
	unbiased CKA (RBF)	1.66	0.25	[0.07, 0.43]	0.0074
MATH	CKA (linear)	1.63	0.26	[0.08, 0.44]	0.0046
	unbiased CKA (linear)	1.63	0.27	[0.09, 0.45]	0.0036
	CKA (RBF)	1.61	0.28	[0.10, 0.47]	0.0027
	unbiased CKA (RBF)	1.63	0.26	[0.07, 0.45]	0.0066
TruthfulQA	CKA (linear)	1.64	0.23	[0.04, 0.41]	0.0185
	unbiased CKA (linear)	1.64	0.23	[0.05, 0.42]	0.0140
	CKA (RBF)	1.63	0.25	[0.06, 0.43]	0.0090
	unbiased CKA (RBF)	1.65	0.21	[0.01, 0.41]	0.0352

Table 5: KBC game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	418.64	18.72	[5.75, 31.68]	0.0047
	unbiased CKA (linear)	418.8	18.53	[5.70, 31.36]	0.0046
	CKA (RBF)	419.4	17.94	[4.77, 31.10]	0.0076
	unbiased CKA (RBF)	419.3	17.77	[4.47, 31.07]	0.0088
GSM8K	CKA (linear)	423.49	19.78	[7.73, 31.84]	0.0013
	unbiased CKA (linear)	423.7	19.82	[7.89, 31.76]	0.0011
	CKA (RBF)	423.8	18.60	[6.76, 30.44]	0.0021
	unbiased CKA (RBF)	423.7	19.30	[6.99, 31.60]	0.0021
MATH	CKA (linear)	422.22	17.71	[5.90, 29.52]	0.0033
	unbiased CKA (linear)	422.4	17.68	[5.84, 29.51]	0.0034
	CKA (RBF)	422.0	17.31	[5.34, 29.27]	0.0046
	unbiased CKA (RBF)	422.5	17.15	[5.06, 29.24]	0.0054
TruthfulQA	CKA (linear)	423.78	14.46	[2.33, 26.60]	0.0195
	unbiased CKA (linear)	423.9	14.69	[2.69, 26.69]	0.0164
	CKA (RBF)	424.1	13.68	[1.75, 25.61]	0.0246
	unbiased CKA (RBF)	424.3	13.35	[0.90, 25.80]	0.0356

D.1.2 When using the average of maximum-aligned scores

Table 6: Word guessing game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	8.37	5.58	[4.14, 7.02]	3.5×10^{-14}
	unbiased CKA (linear)	5.78	9.10	[7.15, 11.05]	6.3×10^{-20}
	CKA (RBF)	8.31	5.76	[4.36, 7.16]	6.4×10^{-16}
	unbiased CKA (RBF)	8.44	5.64	[4.27, 7.01]	7.1×10^{-16}
GSM8K	CKA (linear)	10.64	3.79	[3.00, 4.58]	3.6×10^{-21}
	unbiased CKA (linear)	10.30	4.49	[3.62, 5.37]	1.3×10^{-23}
	CKA (RBF)	10.50	3.92	[3.11, 4.73]	2.5×10^{-21}
	unbiased CKA (RBF)	10.61	3.80	[3.02, 4.59]	3.2×10^{-21}
MATH	CKA (linear)	9.96	4.32	[3.36, 5.27]	8.1×10^{-19}
	unbiased CKA (linear)	9.27	5.51	[4.40, 6.62]	1.6×10^{-22}
	CKA (RBF)	9.68	4.56	[3.53, 5.59]	4.2×10^{-18}
	unbiased CKA (RBF)	9.79	4.45	[3.45, 5.46]	5.1×10^{-18}
TruthfulQA	CKA (linear)	10.24	3.84	[2.92, 4.76]	3.9×10^{-16}
	unbiased CKA (linear)	9.53	5.05	[3.95, 6.16]	3.2×10^{-19}
	CKA (RBF)	10.12	3.93	[2.97, 4.89]	1.0×10^{-15}
	unbiased CKA (RBF)	10.28	3.79	[2.87, 4.71]	7.8×10^{-16}

Table 7: Public good game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	150.11	39.99	[20.48, 59.50]	5.9×10^{-5}
	unbiased CKA (linear)	139.0	55.08	[30.95, 79.20]	7.7×10^{-6}
	CKA (RBF)	149.9	40.93	[21.96, 59.90]	2.4×10^{-5}
	unbiased CKA (RBF)	150.9	39.97	[21.39, 58.54]	2.5×10^{-5}
GSM8K	CKA (linear)	164.33	30.77	[19.18, 42.37]	2.0×10^{-7}
	unbiased CKA (linear)	162.1	35.51	[22.64, 48.38]	6.4×10^{-8}
	CKA (RBF)	162.8	32.58	[20.68, 44.49]	8.2×10^{-8}
	unbiased CKA (RBF)	163.9	31.47	[19.86, 43.08]	1.1×10^{-7}
MATH	CKA (linear)	159.77	33.66	[19.91, 47.41]	1.6×10^{-6}
	unbiased CKA (linear)	155.9	40.41	[24.75, 56.07]	4.2×10^{-7}
	CKA (RBF)	156.7	37.00	[22.26, 51.74]	8.7×10^{-7}
	unbiased CKA (RBF)	157.6	36.08	[21.68, 50.48]	9.1×10^{-7}
TruthfulQA	CKA (linear)	161.06	31.43	[18.21, 44.65]	3.2×10^{-6}
	unbiased CKA (linear)	156.8	38.85	[23.57, 54.13]	6.2×10^{-7}
	CKA (RBF)	159.8	32.59	[19.07, 46.12]	2.3×10^{-6}
	unbiased CKA (RBF)	161.5	30.82	[17.83, 43.81]	3.3×10^{-6}

Table 8: Divide-a-dollar game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	1.52	0.29	[0.11, 0.48]	0.00170
	unbiased CKA (linear)	1.46	0.38	[0.13, 0.62]	0.0025
	CKA (RBF)	1.53	0.29	[0.11, 0.47]	0.0015
	unbiased CKA (RBF)	1.53	0.28	[0.11, 0.46]	0.0015
GSM8K	CKA (linear)	1.67	0.14	[0.03, 0.24]	0.0089
	unbiased CKA (linear)	1.67	0.14	[0.03, 0.26]	0.0161
	CKA (RBF)	1.66	0.15	[0.04, 0.25]	0.0079
	unbiased CKA (RBF)	1.67	0.14	[0.04, 0.25]	0.0078
MATH	CKA (linear)	1.64	0.16	[0.03, 0.29]	0.0132
	unbiased CKA (linear)	1.63	0.17	[0.03, 0.32]	0.0204
	CKA (RBF)	1.63	0.18	[0.04, 0.31]	0.0107
	unbiased CKA (RBF)	1.63	0.17	[0.04, 0.30]	0.0105
TruthfulQA	CKA (linear)	1.64	0.16	[0.04, 0.29]	0.0082
	unbiased CKA (linear)	1.63	0.18	[0.03, 0.32]	0.0187
	CKA (RBF)	1.63	0.18	[0.06, 0.31]	0.0045
	unbiased CKA (RBF)	1.64	0.17	[0.05, 0.29]	0.0050

Table 9: KBC game. We fit a mixed-effects regression between game outcome and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	413.25	22.15	[9.64, 34.65]	0.00052
	unbiased CKA (linear)	411.2	25.12	[9.80, 40.44]	0.00131
	CKA (RBF)	414.1	21.43	[9.13, 33.72]	0.00064
	unbiased CKA (RBF)	414.5	21.01	[8.96, 33.07]	0.00063
GSM8K	CKA (linear)	422.86	13.71	[6.19, 21.24]	0.00036
	unbiased CKA (linear)	422.4	14.82	[6.46, 23.18]	0.00051
	CKA (RBF)	422.8	13.44	[5.70, 21.17]	0.00066
	unbiased CKA (RBF)	423.0	13.35	[5.81, 20.89]	0.00052
MATH	CKA (linear)	420.55	15.44	[6.50, 24.39]	0.00072
	unbiased CKA (linear)	419.6	17.18	[6.99, 27.37]	0.00095
	CKA (RBF)	419.8	15.88	[6.29, 25.47]	0.00117
	unbiased CKA (RBF)	420.1	15.75	[6.38, 25.12]	0.00098
TruthfulQA	CKA (linear)	421.48	13.88	[5.30, 22.46]	0.00152
	unbiased CKA (linear)	420.9	14.97	[5.02, 24.91]	0.00318
	CKA (RBF)	421.4	13.67	[4.88, 22.46]	0.00230
	unbiased CKA (RBF)	421.8	13.49	[5.05, 21.92]	0.00172

D.2 MIXED-EFFECTS REGRESSION RESULTS FOR UNIQUENESS

Tables 10∼17 consistently show negative trends across all datasets, CKA variants, and games.

D.2.1 WHEN USING THE GLOBAL AVERAGE

Table 10: Story writing task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	4.12	-0.31	[-1.56, 0.95]	0.633
	unbiased CKA (linear)	4.12	-0.30	[-1.55, 0.94]	0.630
	CKA (RBF)	4.14	-0.34	[-1.63, 0.96]	0.610
	unbiased CKA (RBF)	4.14	-0.33	[-1.60, 0.93]	0.608
GSM8K	CKA (linear)	4.03	-0.27	[-1.42, 0.88]	0.643
	unbiased CKA (linear)	4.02	-0.26	[-1.40, 0.88]	0.659
	CKA (RBF)	4.04	-0.29	[-1.44, 0.86]	0.626
	unbiased CKA (RBF)	4.03	-0.26	[-1.39, 0.87]	0.655
MATH	CKA (linear)	4.10	-0.36	[-1.48, 0.77]	0.536
	unbiased CKA (linear)	4.09	-0.34	[-1.45, 0.77]	0.548
	CKA (RBF)	4.12	-0.37	[-1.54, 0.79]	0.532
	unbiased CKA (RBF)	4.10	-0.35	[-1.49, 0.80]	0.553
TruthfulQA	CKA (linear)	4.15	-0.49	[-1.65, 0.66]	0.403
	unbiased CKA (linear)	4.13	-0.46	[-1.61, 0.68]	0.428
	CKA (RBF)	4.14	-0.45	[-1.61, 0.72]	0.453
	unbiased CKA (RBF)	4.11	-0.40	[-1.54, 0.74]	0.490

Table 11: Fictional biography generation task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	8.14	-1.28	[-2.47, -0.10]	0.034
	unbiased CKA (linear)	8.13	-1.28	[-2.45, -0.11]	0.0327
	CKA (RBF)	8.21	-1.42	[-2.59, -0.24]	0.0181
	unbiased CKA (RBF)	8.19	-1.40	[-2.55, -0.25]	0.0169
GSM8K	CKA (linear)	7.94	-1.75	[-2.63, -0.87]	9.3×10^{-5}
	unbiased CKA (linear)	7.92	-1.76	[-2.63, -0.90]	6.7×10^{-5}
	CKA (RBF)	7.95	-1.67	[-2.55, -0.79]	1.89×10^{-4}
	unbiased CKA (RBF)	7.92	-1.69	[-2.55, -0.83]	1.16×10^{-4}
MATH	CKA (linear)	8.07	-1.60	[-2.52, -0.69]	5.7×10^{-4}
	unbiased CKA (linear)	8.05	-1.61	[-2.51, -0.71]	4.60×10^{-4}
	CKA (RBF)	8.12	-1.61	[-2.56, -0.65]	9.47×10^{-4}
	unbiased CKA (RBF)	8.10	-1.61	[-2.54, -0.68]	7.17×10^{-4}
TruthfulQA	CKA (linear)	7.92	-1.29	[-2.24, -0.35]	0.0073
	unbiased CKA (linear)	7.91	-1.33	[-2.27, -0.40]	0.00506
	CKA (RBF)	7.91	-1.21	[-2.17, -0.25]	0.0134
	unbiased CKA (RBF)	7.90	-1.26	[-2.19, -0.32]	0.00824

Table 12: Haiku composition task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	5.95	-3.42	[-4.80, -2.05]	1.1×10^{-6}
	unbiased CKA (linear)	5.93	-3.42	[-4.77, -2.07]	6.71×10^{-7}
	CKA (RBF)	5.74	-3.13	[-4.56, -1.70]	1.71×10^{-5}
	unbiased CKA (RBF)	5.70	-3.10	[-4.49, -1.70]	1.39×10^{-5}
GSM8K	CKA (linear)	4.67	-2.43	[-3.70, -1.16]	1.7×10^{-4}
	unbiased CKA (linear)	4.64	-2.42	[-3.68, -1.17]	0.000158
	CKA (RBF)	4.65	-2.22	[-3.49, -0.96]	0.000589
	unbiased CKA (RBF)	4.61	-2.22	[-3.46, -0.97]	0.000477
MATH	CKA (linear)	5.24	-3.12	[-4.36, -1.89]	7.2×10^{-7}
	unbiased CKA (linear)	5.22	-3.13	[-4.35, -1.92]	4.48×10^{-7}
	CKA (RBF)	5.39	-3.19	[-4.45, -1.93]	6.88×10^{-7}
	unbiased CKA (RBF)	5.33	-3.17	[-4.40, -1.94]	4.60×10^{-7}
TruthfulQA	CKA (linear)	4.98	-2.58	[-3.84, -1.33]	5.2×10^{-5}
	unbiased CKA (linear)	4.94	-2.58	[-3.83, -1.34]	4.75×10^{-5}
	CKA (RBF)	4.84	-2.12	[-3.43, -0.81]	0.00149
	unbiased CKA (RBF)	4.95	-2.19	[-3.83, -0.91]	0.000794

Table 13: Vacation brainstorming task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	1.66	-1.01	[-1.58, -0.45]	0.00045
	unbiased CKA (linear)	1.65	-1.00	[-1.56, -0.44]	0.000480
	CKA (RBF)	1.61	-0.94	[-1.55, -0.32]	0.00287
	unbiased CKA (RBF)	1.59	-0.91	[-1.51, -0.31]	0.00308
GSM8K	CKA (linear)	1.26	-0.65	[-1.32, 0.01]	0.053
	unbiased CKA (linear)	1.25	-0.64	[-1.29, 0.02]	0.0584
	CKA (RBF)	1.26	-0.60	[-1.26, 0.05]	0.0721
	unbiased CKA (RBF)	1.24	-0.57	[-1.22, 0.07]	0.0824
MATH	CKA (linear)	1.45	-0.91	[-1.50, -0.31]	0.0028
	unbiased CKA (linear)	1.43	-0.89	[-1.48, -0.30]	0.00298
	CKA (RBF)	1.45	-0.87	[-1.47, -0.27]	0.00468
	unbiased CKA (RBF)	1.45	-0.87	[-1.47, -0.27]	0.00468
TruthfulQA	CKA (linear)	1.35	-0.71	[-1.34, -0.08]	0.0273
	unbiased CKA (linear)	1.34	-0.69	[-1.31, -0.07]	0.0304
	CKA (RBF)	1.34	-0.65	[-1.27, -0.02]	0.0438
	unbiased CKA (RBF)	1.32	-0.62	[-1.24, -0.01]	0.0480

D.2.2 When using the average of maximum-aligned scores

Table 14: Story writing task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	4.31	-0.49	[-1.69, 0.71]	0.422
	unbiased CKA (linear)	4.31	-0.50	[-1.68, 0.69]	0.412
	CKA (RBF)	4.31	-0.51	[-1.66, 0.65]	0.391
	unbiased CKA (RBF)	4.31	-0.50	[-1.68, 0.67]	0.402
GSM8K	CKA (linear)	4.02	-0.16	[-0.88, 0.56]	0.656
	unbiased CKA (linear)	4.02	-0.16	[-0.87, 0.55]	0.662
	CKA (RBF)	4.03	-0.17	[-0.89, 0.55]	0.641
	unbiased CKA (RBF)	4.04	-0.18	[-0.92, 0.56]	0.632
MATH	CKA (linear)	4.11	-0.29	[-1.14, 0.57]	0.512
	unbiased CKA (linear)	4.11	-0.28	[-1.13, 0.56]	0.514
	CKA (RBF)	4.13	-0.31	[-1.20, 0.59]	0.502
	unbiased CKA (RBF)	4.14	-0.32	[-1.23, 0.60]	0.500
TruthfulQA	CKA (linear)	4.06	-0.19	[-1.01, 0.63]	0.644
	unbiased CKA (linear)	4.05	-0.19	[-0.99, 0.61]	0.641
	CKA (RBF)	4.05	-0.19	[-1.00, 0.62]	0.646
	unbiased CKA (RBF)	4.06	-0.19	[-1.03, 0.65]	0.654

Table 15: Fictional biography generation task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	8.53	-1.55	[-2.47, -0.63]	0.00094
	unbiased CKA (linear)	8.52	-1.54	[-2.45, -0.63]	0.001
	CKA (RBF)	8.50	-1.55	[-2.43, -0.68]	0.001
	unbiased CKA (RBF)	8.53	-1.58	[-2.47, -0.68]	0.001
GSM8K	CKA (linear)	7.96	-1.15	[-1.67, -0.62]	1.8×10^{-5}
	unbiased CKA (linear)	7.95	-1.14	[-1.66, -0.63]	1.0×10^{-5}
	CKA (RBF)	7.96	-1.15	[-1.67, -0.62]	2.0×10^{-5}
	unbiased CKA (RBF)	7.98	-1.16	[-1.70, -0.62]	3.0×10^{-5}
MATH	CKA (linear)	8.17	-1.31	[-1.94, -0.69]	4.3×10^{-5}
	unbiased CKA (linear)	8.15	-1.31	[-1.93, -0.69]	4.0×10^{-5}
	CKA (RBF)	8.23	-1.37	[-2.03, -0.71]	5.0×10^{-5}
	unbiased CKA (RBF)	8.25	-1.39	[-2.07, -0.71]	6.0×10^{-5}
TruthfulQA	CKA (linear)	8.08	-1.18	[-1.79, -0.56]	0.00016
	unbiased CKA (linear)	8.06	-1.16	[-1.76, -0.56]	1.0×10^{-4}
	CKA (RBF)	8.04	-1.12	[-1.72, -0.51]	3.0×10^{-4}
	unbiased CKA (RBF)	8.08	-1.14	[-1.78, -0.51]	4.0×10^{-4}

Table 16: Haiku composition task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	5.86	-2.64	[-3.95, -1.32]	8.9×10^{-5}
	unbiased CKA (linear)	5.84	-2.61	[-3.92, -1.31]	8.75×10^{-5}
	CKA (RBF)	5.62	-2.37	[-3.63, -1.11]	2.33×10^{-4}
	unbiased CKA (RBF)	5.65	-2.39	[-3.68, -1.11]	2.67×10^{-4}
GSM8K	CKA (linear)	4.62	-1.44	[-2.22, -0.66]	0.00028
	unbiased CKA (linear)	4.60	-1.43	[-2.19, -0.66]	2.57×10^{-4}
	CKA (RBF)	4.59	-1.38	[-2.16, -0.60]	5.19×10^{-4}
	unbiased CKA (RBF)	4.62	-1.40	[-2.20, -0.60]	6.01×10^{-4}
MATH	CKA (linear)	5.04	-1.92	[-2.84, -1.00]	4.9×10^{-5}
	unbiased CKA (linear)	5.03	-1.91	[-2.83, -1.00]	4.09×10^{-5}
	CKA (RBF)	5.11	-1.97	[-2.94, -1.00]	6.83×10^{-5}
	unbiased CKA (RBF)	5.14	-1.98	[-2.98, -0.99]	9.19×10^{-5}
TruthfulQA	CKA (linear)	4.88	-1.65	[-2.54, -0.76]	0.00030
	unbiased CKA (linear)	4.86	-1.64	[-2.51, -0.76]	2.43×10^{-4}
	CKA (RBF)	4.78	-1.49	[-2.37, -0.61]	8.84×10^{-4}
	unbiased CKA (RBF)	4.82	-1.51	[-2.42, -0.59]	0.0013

Table 17: Vacation brainstorming task. We fit a mixed-effects regression between response uniqueness and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	1.79	-0.97	[-1.65, -0.30]	0.0046
	unbiased CKA (linear)	1.78	-0.97	[-1.63, -0.30]	0.005
	CKA (RBF)	1.69	-0.87	[-1.52, -0.21]	0.009
	unbiased CKA (RBF)	1.70	-0.88	[-1.55, -0.22]	0.009
GSM8K	CKA (linear)	1.25	-0.38	[-0.80, 0.05]	0.081
	unbiased CKA (linear)	1.24	-0.37	[-0.80, 0.05]	0.083
	CKA (RBF)	1.25	-0.37	[-0.80, 0.05]	0.083
	unbiased CKA (RBF)	1.26	-0.39	[-0.82, 0.05]	0.080
MATH	CKA (linear)	1.44	-0.64	[-1.14, -0.14]	0.012
	unbiased CKA (linear)	1.44	-0.64	[-1.13, -0.14]	0.011
	CKA (RBF)	1.48	-0.68	[-1.20, -0.16]	0.011
	unbiased CKA (RBF)	1.49	-0.69	[-1.22, -0.16]	0.011
TruthfulQA	CKA (linear)	1.36	-0.50	[-0.98, -0.03]	0.039
	unbiased CKA (linear)	1.35	-0.50	[-0.96, -0.03]	0.036
	CKA (RBF)	1.36	-0.51	[-0.97, -0.04]	0.032
	unbiased CKA (RBF)	1.37	-0.52	[-1.00, -0.03]	0.036

D.3 MIXED-EFFECTS REGRESSION
RESULTS FOR MUTUAL INFORMATION CALCULATED WITH `LLAMA-3.1-8B-INSTRUCT`

Tables 18∼25 show significant positive trends across all datasets, CKA variants, and games.

D.3.1 WHEN USING THE GLOBAL AVERAGE

Table 18: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.303	0.053	[0.020, 0.087]	0.00194
	unbiased CKA (linear)	0.304	0.053	[0.019, 0.086]	0.00193
	CKA (RBF)	0.306	0.050	[0.016, 0.084]	0.00379
	unbiased CKA (RBF)	0.307	0.049	[0.016, 0.082]	0.00372
GSM8K	CKA (linear)	0.317	0.057	[0.027, 0.087]	0.00021
	unbiased CKA (linear)	0.318	0.057	[0.027, 0.086]	0.00020
	CKA (RBF)	0.316	0.055	[0.025, 0.085]	0.00035
	unbiased CKA (RBF)	0.318	0.054	[0.025, 0.084]	0.00032
MATH	CKA (linear)	0.313	0.052	[0.022, 0.082]	0.00058
	unbiased CKA (linear)	0.313	0.052	[0.022, 0.081]	0.00055
	CKA (RBF)	0.310	0.054	[0.024, 0.085]	0.00051
	unbiased CKA (RBF)	0.311	0.053	[0.023, 0.083]	0.00047
TruthfulQA	CKA (linear)	0.312	0.056	[0.026, 0.086]	0.00026
	unbiased CKA (linear)	0.312	0.057	[0.027, 0.086]	0.00019
	CKA (RBF)	0.312	0.053	[0.023, 0.083]	0.00062
	unbiased CKA (RBF)	0.313	0.053	[0.024, 0.083]	0.00040

Table 19: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.083	0.121	[0.091, 0.150]	5.07×10^{-16}
	unbiased CKA (linear)	0.084	0.120	[0.091, 0.149]	4.71×10^{-16}
	CKA (RBF)	0.084	0.119	[0.090, 0.149]	2.31×10^{-15}
	unbiased CKA (RBF)	0.087	0.117	[0.088, 0.146]	1.96×10^{-15}
GSM8K	CKA (linear)	0.123	0.099	[0.076, 0.122]	2.17×10^{-17}
	unbiased CKA (linear)	0.124	0.098	[0.075, 0.121]	2.08×10^{-17}
	CKA (RBF)	0.121	0.098	[0.075, 0.121]	6.39×10^{-17}
	unbiased CKA (RBF)	0.124	0.096	[0.074, 0.119]	5.27×10^{-17}
MATH	CKA (linear)	0.109	0.107	[0.084, 0.131]	1.39×10^{-19}
	unbiased CKA (linear)	0.110	0.106	[0.083, 0.129]	1.13×10^{-19}
	CKA (RBF)	0.103	0.111	[0.086, 0.135]	5.17×10^{-19}
	unbiased CKA (RBF)	0.106	0.108	[0.085, 0.132]	3.86×10^{-19}
TruthfulQA	CKA (linear)	0.117	0.091	[0.067, 0.115]	2.38×10^{-13}
	unbiased CKA (linear)	0.118	0.091	[0.067, 0.115]	1.03×10^{-13}
	CKA (RBF)	0.115	0.088	[0.064, 0.113]	2.44×10^{-12}
	unbiased CKA (RBF)	0.118	0.088	[0.064, 0.112]	6.81×10^{-13}

Table 20: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.594	1.310	[1.03, 1.59]	1.26×10^{-20}
	unbiased CKA (linear)	0.603	1.302	[1.029, 1.575]	9.20×10^{-21}
	CKA (RBF)	0.672	1.194	[0.919, 1.470]	2.03×10^{-17}
	unbiased CKA (RBF)	0.693	1.177	[0.907, 1.447]	1.28×10^{-17}
GSM8K	CKA (linear)	1.161	0.688	[0.48, 0.89]	6.18×10^{-11}
	unbiased CKA (linear)	1.171	0.679	[0.475, 0.882]	6.20×10^{-11}
	CKA (RBF)	1.152	0.670	[0.463, 0.877]	2.15×10^{-10}
	unbiased CKA (RBF)	1.168	0.656	[0.454, 0.858]	1.88×10^{-10}
MATH	CKA (linear)	1.017	0.845	[0.63, 1.06]	8.56×10^{-15}
	unbiased CKA (linear)	1.025	0.842	[0.632, 1.053]	4.48×10^{-15}
	CKA (RBF)	0.970	0.878	[0.654, 1.101]	1.33×10^{-14}
	unbiased CKA (RBF)	0.986	0.871	[0.653, 1.089]	4.73×10^{-15}
TruthfulQA	CKA (linear)	1.046	0.794	[0.57, 1.02]	2.62×10^{-12}
	unbiased CKA (linear)	1.054	0.804	[0.585, 1.024]	6.80×10^{-13}
	CKA (RBF)	1.039	0.766	[0.539, 0.992]	3.51×10^{-11}
	unbiased CKA (RBF)	1.053	0.777	[0.557, 0.997]	4.51×10^{-12}

Table 21: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.587	0.132	[0.071, 0.194]	2.40×10^{-5}
	unbiased CKA (linear)	0.589	0.131	[0.070, 0.192]	2.43×10^{-5}
	CKA (RBF)	0.592	0.127	[0.061, 0.192]	0.00015
	unbiased CKA (RBF)	0.594	0.125	[0.060, 0.189]	0.00014
GSM8K	CKA (linear)	0.621	0.141	[0.078, 0.203]	9.48×10^{-6}
	unbiased CKA (linear)	0.623	0.140	[0.079, 0.202]	8.27×10^{-6}
	CKA (RBF)	0.619	0.140	[0.077, 0.202]	1.09×10^{-5}
	unbiased CKA (RBF)	0.621	0.139	[0.078, 0.200]	8.45×10^{-6}
MATH	CKA (linear)	0.602	0.149	[0.090, 0.208]	7.62×10^{-7}
	unbiased CKA (linear)	0.604	0.148	[0.090, 0.206]	6.08×10^{-7}
	CKA (RBF)	0.594	0.154	[0.093, 0.215]	7.74×10^{-7}
	unbiased CKA (RBF)	0.597	0.153	[0.093, 0.212]	5.49×10^{-7}
TruthfulQA	CKA (linear)	0.619	0.113	[0.051, 0.175]	0.00034
	unbiased CKA (linear)	0.621	0.112	[0.051, 0.174]	0.00032
	CKA (RBF)	0.618	0.108	[0.046, 0.170]	0.00068
	unbiased CKA (RBF)	0.622	0.107	[0.046, 0.167]	0.00060

D.3.2 When using the average of maximum-aligned scores

Table 22: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.272	0.085	[0.053, 0.117]	2.64×10^{-7}
	unbiased CKA (linear)	0.272	0.084	[0.052, 0.116]	2.62×10^{-7}
	CKA (RBF)	0.276	0.081	[0.049, 0.112]	5.54×10^{-7}
	unbiased CKA (RBF)	0.277	0.079	[0.048, 0.110]	5.23×10^{-7}
GSM8K	CKA (linear)	0.312	0.045	[0.026, 0.065]	4.39×10^{-6}
	unbiased CKA (linear)	0.313	0.045	[0.026, 0.064]	4.35×10^{-6}
	CKA (RBF)	0.311	0.046	[0.026, 0.066]	5.92×10^{-6}
	unbiased CKA (RBF)	0.312	0.045	[0.025, 0.064]	5.76×10^{-6}
MATH	CKA (linear)	0.301	0.057	[0.034, 0.080]	9.67×10^{-7}
	unbiased CKA (linear)	0.302	0.056	[0.034, 0.079]	9.72×10^{-7}
	CKA (RBF)	0.297	0.061	[0.036, 0.085]	1.09×10^{-6}
	unbiased CKA (RBF)	0.298	0.059	[0.036, 0.083]	1.09×10^{-6}
TruthfulQA	CKA (linear)	0.302	0.056	[0.034, 0.078]	5.79×10^{-7}
	unbiased CKA (linear)	0.303	0.055	[0.033, 0.076]	4.89×10^{-7}
	CKA (RBF)	0.301	0.055	[0.033, 0.078]	1.35×10^{-6}
	unbiased CKA (RBF)	0.303	0.054	[0.032, 0.075]	9.85×10^{-7}

Table 23: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.089	0.088	[0.064, 0.112]	6.16×10^{-13}
	unbiased CKA (linear)	0.090	0.087	[0.063, 0.111]	7.49×10^{-13}
	CKA (RBF)	0.091	0.088	[0.064, 0.111]	1.86×10^{-13}
	unbiased CKA (RBF)	0.093	0.086	[0.063, 0.109]	2.40×10^{-13}
GSM8K	CKA (linear)	0.124	0.061	[0.047, 0.075]	2.48×10^{-18}
	unbiased CKA (linear)	0.125	0.060	[0.047, 0.074]	3.41×10^{-18}
	CKA (RBF)	0.121	0.064	[0.050, 0.078]	1.05×10^{-18}
	unbiased CKA (RBF)	0.123	0.062	[0.048, 0.075]	1.70×10^{-18}
MATH	CKA (linear)	0.112	0.071	[0.055, 0.088]	2.68×10^{-17}
	unbiased CKA (linear)	0.113	0.070	[0.054, 0.086]	4.29×10^{-17}
	CKA (RBF)	0.107	0.077	[0.059, 0.094]	2.40×10^{-17}
	unbiased CKA (RBF)	0.109	0.074	[0.057, 0.091]	5.19×10^{-17}
TruthfulQA	CKA (linear)	0.116	0.065	[0.049, 0.081]	2.18×10^{-15}
	unbiased CKA (linear)	0.118	0.063	[0.047, 0.078]	2.94×10^{-15}
	CKA (RBF)	0.114	0.065	[0.049, 0.082]	6.08×10^{-15}
	unbiased CKA (RBF)	0.118	0.063	[0.047, 0.078]	7.99×10^{-15}

Table 24: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.652	0.973	[0.760, 1.187]	4.43×10^{-19}
	unbiased CKA (linear)	0.659	0.966	[0.754, 1.178]	3.84×10^{-19}
	CKA (RBF)	0.714	0.907	[0.699, 1.115]	1.29×10^{-17}
	unbiased CKA (RBF)	0.729	0.893	[0.689, 1.097]	9.53×10^{-18}
GSM8K	CKA (linear)	1.123	0.507	[0.385, 0.629]	3.05×10^{-16}
	unbiased CKA (linear)	1.131	0.501	[0.381, 0.620]	2.77×10^{-16}
	CKA (RBF)	1.111	0.513	[0.387, 0.638]	1.05×10^{-15}
	unbiased CKA (RBF)	1.125	0.500	[0.378, 0.622]	9.96×10^{-16}
MATH	CKA (linear)	0.999	0.635	[0.489, 0.781]	1.83×10^{-17}
	unbiased CKA (linear)	1.006	0.629	[0.485, 0.774]	1.35×10^{-17}
	CKA (RBF)	0.956	0.674	[0.517, 0.832]	4.69×10^{-17}
	unbiased CKA (RBF)	0.970	0.662	[0.508, 0.816]	3.13×10^{-17}
TruthfulQA	CKA (linear)	1.012	0.612	[0.470, 0.754]	2.98×10^{-17}
	unbiased CKA (linear)	1.023	0.604	[0.466, 0.743]	1.34×10^{-17}
	CKA (RBF)	1.005	0.606	[0.460, 0.753]	4.83×10^{-16}
	unbiased CKA (RBF)	1.026	0.592	[0.452, 0.732]	1.37×10^{-16}

Table 25: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.575	0.122	[0.056, 0.188]	0.00031
	unbiased CKA (linear)	0.577	0.121	[0.055, 0.186]	0.00031
	CKA (RBF)	0.586	0.110	[0.045, 0.175]	0.00090
	unbiased CKA (RBF)	0.588	0.108	[0.044, 0.171]	0.00097
GSM8K	CKA (linear)	0.626	0.079	[0.039, 0.119]	9.26×10^{-5}
	unbiased CKA (linear)	0.628	0.077	[0.038, 0.116]	0.00011
	CKA (RBF)	0.623	0.082	[0.042, 0.123]	7.27×10^{-5}
	unbiased CKA (RBF)	0.626	0.079	[0.040, 0.119]	8.69×10^{-5}
MATH	CKA (linear)	0.604	0.105	[0.057, 0.152]	1.38×10^{-5}
	unbiased CKA (linear)	0.605	0.103	[0.057, 0.150]	1.37×10^{-5}
	CKA (RBF)	0.597	0.111	[0.061, 0.161]	1.60×10^{-5}
	unbiased CKA (RBF)	0.599	0.109	[0.060, 0.158]	1.45×10^{-5}
TruthfulQA	CKA (linear)	0.613	0.089	[0.043, 0.134]	0.00012
	unbiased CKA (linear)	0.616	0.086	[0.042, 0.130]	0.00014
	CKA (RBF)	0.612	0.089	[0.042, 0.135]	0.00017
	unbiased CKA (RBF)	0.616	0.085	[0.040, 0.129]	0.00018

D.4 MIXED-EFFECTS REGRESSION RESULTS FOR MUTUAL INFORMATION CALCULATED WITH LLAMA-3.1-8B

Tables 26∼33 show significant positive trends across all datasets, CKA variants, and games.

D.4.1 WHEN USING THE GLOBAL AVERAGE

Table 26: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.363	0.088	[0.037, 0.139]	0.00075
	unbiased CKA (linear)	0.363	0.087	[0.037, 0.138]	0.00074
	CKA (RBF)	0.372	0.072	[0.026, 0.119]	0.00238
	unbiased CKA (RBF)	0.374	0.071	[0.026, 0.117]	0.00227
GSM8K	CKA (linear)	0.389	0.081	[0.045, 0.118]	1.39×10^{-5}
	unbiased CKA (linear)	0.390	0.081	[0.045, 0.117]	1.20×10^{-5}
	CKA (RBF)	0.389	0.077	[0.040, 0.113]	3.60×10^{-5}
	unbiased CKA (RBF)	0.390	0.077	[0.041, 0.113]	2.76×10^{-5}
MATH	CKA (linear)	0.381	0.078	[0.041, 0.116]	4.25×10^{-5}
	unbiased CKA (linear)	0.382	0.078	[0.041, 0.115]	3.90×10^{-5}
	CKA (RBF)	0.378	0.079	[0.040, 0.118]	6.24×10^{-5}
	unbiased CKA (RBF)	0.380	0.078	[0.040, 0.116]	5.45×10^{-5}
TruthfulQA	CKA (linear)	0.382	0.079	[0.040, 0.118]	6.21×10^{-5}
	unbiased CKA (linear)	0.383	0.080	[0.042, 0.119]	4.33×10^{-5}
	CKA (RBF)	0.383	0.071	[0.033, 0.110]	0.00030
	unbiased CKA (RBF)	0.385	0.072	[0.035, 0.110]	0.00018

Table 27: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.135	0.130	[0.098, 0.162]	1.88×10^{-15}
	unbiased CKA (linear)	0.136	0.129	[0.097, 0.161]	1.70×10^{-15}
	CKA (RBF)	0.136	0.130	[0.0981,0.163]	2.96×10^{-15}
	unbiased CKA (RBF)	0.138	0.128	[0.097, 0.160]	2.12×10^{-15}
GSM8K	CKA (linear)	0.175	0.119	[0.094, 0.144]	5.79×10^{-21}
	unbiased CKA (linear)	0.177	0.118	[0.093, 0.142]	5.51×10^{-21}
	CKA (RBF)	0.173	0.118	[0.093, 0.143]	1.74×10^{-20}
	unbiased CKA (RBF)	0.176	0.116	[0.091, 0.140]	1.37×10^{-20}
MATH	CKA (linear)	0.162	0.118	[0.093, 0.144]	7.24×10^{-20}
	unbiased CKA (linear)	0.164	0.117	[0.092, 0.142]	6.08×10^{-20}
	CKA (RBF)	0.155	0.123	[0.097, 0.150]	9.91×10^{-20}
	unbiased CKA (RBF)	0.158	0.121	[0.095, 0.147]	7.60×10^{-20}
TruthfulQA	CKA (linear)	0.169	0.106	[0.079, 0.132]	5.53×10^{-15}
	unbiased CKA (linear)	0.170	0.106	[0.080, 0.132]	2.24×10^{-15}
	CKA (RBF)	0.166	0.105	[0.078, 0.132]	3.05×10^{-14}
	unbiased CKA (RBF)	0.169	0.104	[0.078, 0.130]	7.51×10^{-15}

Table 28: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.628	1.335	[1.033, 1.638]	5.17×10^{-18}
	unbiased CKA (linear)	0.637	1.329	[1.029, 1.629]	3.54×10^{-18}
	CKA (RBF)	0.700	1.230	[0.928, 1.533]	1.71×10^{-15}
	unbiased CKA (RBF)	0.720	1.215	[0.918, 1.511]	1.02×10^{-15}
GSM8K	CKA (linear)	1.209	0.693	[0.466, 0.920]	2.27×10^{-9}
	unbiased CKA (linear)	1.218	0.684	[0.460, 0.909]	2.23×10^{-9}
	CKA (RBF)	1.200	0.675	[0.447, 0.903]	6.47×10^{-9}
	unbiased CKA (RBF)	1.216	0.662	[0.439, 0.884]	5.65×10^{-9}
MATH	CKA (linear)	1.058	0.863	[0.628, 1.098]	6.30×10^{-13}
	unbiased CKA (linear)	1.066	0.861	[0.629, 1.093]	3.39×10^{-13}
	CKA (RBF)	1.011	0.897	[0.651, 1.143]	8.97×10^{-13}
	unbiased CKA (RBF)	1.026	0.891	[0.651, 1.131]	3.44×10^{-13}
TruthfulQA	CKA (linear)	1.098	0.791	[0.545, 1.036]	2.65×10^{-10}
	unbiased CKA (linear)	1.103	0.804	[0.562, 1.046]	7.32×10^{-11}
	CKA (RBF)	1.090	0.762	[0.512, 1.012]	2.22×10^{-9}
	unbiased CKA (RBF)	1.103	0.777	[0.535, 1.020]	3.34×10^{-10}

Table 29: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.696	0.158	[0.094, 0.222]	1.40×10^{-6}
	unbiased CKA (linear)	0.698	0.155	[0.094, 0.215]	6.16×10^{-7}
	CKA (RBF)	0.698	0.156	[0.087, 0.224]	8.20×10^{-6}
	unbiased CKA (RBF)	0.701	0.153	[0.086, 0.220]	7.34×10^{-6}
GSM8K	CKA (linear)	0.733	0.177	[0.112, 0.243]	1.20×10^{-7}
	unbiased CKA (linear)	0.735	0.177	[0.112, 0.242]	9.84×10^{-8}
	CKA (RBF)	0.730	0.175	[0.110, 0.241]	1.83×10^{-7}
	unbiased CKA (RBF)	0.734	0.175	[0.110, 0.239]	1.25×10^{-7}
MATH	CKA (linear)	0.713	0.179	[0.117, 0.241]	1.39×10^{-8}
	unbiased CKA (linear)	0.715	0.179	[0.118, 0.240]	1.01×10^{-8}
	CKA (RBF)	0.703	0.186	[0.121, 0.251]	1.84×10^{-8}
	unbiased CKA (RBF)	0.707	0.184	[0.121, 0.247]	1.13×10^{-8}
TruthfulQA	CKA (linear)	0.731	0.142	[0.078, 0.206]	1.33×10^{-5}
	unbiased CKA (linear)	0.733	0.142	[0.079, 0.206]	1.08×10^{-5}
	CKA (RBF)	0.729	0.138	[0.073, 0.203]	3.04×10^{-5}
	unbiased CKA (RBF)	0.733	0.137	[0.074, 0.200]	2.09×10^{-5}

D.4.2 When using the average of maximum-aligned scores

Table 30: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.317	0.131	[0.090, 0.171]	2.92×10^{-10}
	unbiased CKA (linear)	0.318	0.129	[0.089, 0.169]	3.11×10^{-10}
	CKA (RBF)	0.327	0.120	[0.081, 0.159]	2.00×10^{-9}
	unbiased CKA (RBF)	0.329	0.118	[0.079, 0.156]	1.96×10^{-9}
GSM8K	CKA (linear)	0.380	0.069	[0.046, 0.092]	2.98×10^{-9}
	unbiased CKA (linear)	0.381	0.068	[0.046, 0.091]	2.99×10^{-9}
	CKA (RBF)	0.378	0.070	[0.046, 0.093]	6.20×10^{-9}
	unbiased CKA (RBF)	0.380	0.068	[0.045, 0.091]	5.88×10^{-9}
MATH	CKA (linear)	0.362	0.088	[0.060, 0.115]	4.06×10^{-10}
	unbiased CKA (linear)	0.364	0.086	[0.059, 0.113]	4.40×10^{-10}
	CKA (RBF)	0.357	0.092	[0.063, 0.121]	8.61×10^{-10}
	unbiased CKA (RBF)	0.359	0.090	[0.061, 0.118]	9.81×10^{-10}
TruthfulQA	CKA (linear)	0.363	0.086	[0.059, 0.113]	2.37×10^{-10}
	unbiased CKA (linear)	0.365	0.084	[0.058, 0.110]	2.25×10^{-10}
	CKA (RBF)	0.363	0.085	[0.057, 0.112]	1.46×10^{-9}
	unbiased CKA (RBF)	0.366	0.082	[0.055, 0.108]	1.16×10^{-9}

Table 31: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.143	0.094	[0.068, 0.120]	1.60×10^{-12}
	unbiased CKA (linear)	0.144	0.093	[0.067, 0.119]	1.83×10^{-12}
	CKA (RBF)	0.144	0.095	[0.069, 0.120]	2.72×10^{-13}
	unbiased CKA (RBF)	0.146	0.092	[0.068, 0.117]	3.19×10^{-13}
GSM8K	CKA (linear)	0.178	0.069	[0.054, 0.084]	5.27×10^{-20}
	unbiased CKA (linear)	0.179	0.068	[0.054, 0.083]	6.93×10^{-20}
	CKA (RBF)	0.175	0.072	[0.057, 0.088]	1.65×10^{-20}
	unbiased CKA (RBF)	0.177	0.070	[0.055, 0.085]	2.42×10^{-20}
MATH	CKA (linear)	0.167	0.077	[0.059, 0.095]	3.49×10^{-17}
	unbiased CKA (linear)	0.168	0.075	[0.058, 0.093]	5.06×10^{-17}
	CKA (RBF)	0.161	0.083	[0.064, 0.103]	1.85×10^{-17}
	unbiased CKA (RBF)	0.163	0.081	[0.062, 0.100]	3.36×10^{-17}
TruthfulQA	CKA (linear)	0.170	0.071	[0.054, 0.089]	6.60×10^{-16}
	unbiased CKA (linear)	0.172	0.070	[0.053, 0.086]	7.44×10^{-16}
	CKA (RBF)	0.168	0.072	[0.054, 0.090]	1.93×10^{-15}
	unbiased CKA (RBF)	0.172	0.069	[0.052, 0.086]	1.85×10^{-15}

Table 32: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.664	1.023	[0.788, 1.259]	1.73×10^{-17}
	unbiased CKA (linear)	0.672	1.015	[0.782, 1.249]	1.55×10^{-17}
	CKA (RBF)	0.726	0.957	[0.728, 1.187]	2.83×10^{-16}
	unbiased CKA (RBF)	0.743	0.942	[0.717, 1.167]	2.20×10^{-16}
GSM8K	CKA (linear)	1.161	0.530	[0.395, 0.664]	9.96×10^{-15}
	unbiased CKA (linear)	1.169	0.523	[0.391, 0.655]	8.96×10^{-15}
	CKA (RBF)	1.149	0.535	[0.397, 0.673]	3.24×10^{-14}
	unbiased CKA (RBF)	1.164	0.522	[0.387, 0.656]	3.02×10^{-14}
MATH	CKA (linear)	1.031	0.664	[0.503, 0.826]	7.19×10^{-16}
	unbiased CKA (linear)	1.039	0.658	[0.499, 0.817]	5.55×10^{-16}
	CKA (RBF)	0.985	0.706	[0.532, 0.879]	1.57×10^{-15}
	unbiased CKA (RBF)	1.000	0.693	[0.523, 0.862]	1.14×10^{-15}
TruthfulQA	CKA (linear)	1.049	0.632	[0.475, 0.788]	2.49×10^{-15}
	unbiased CKA (linear)	1.061	0.624	[0.471, 0.777]	1.22×10^{-15}
	CKA (RBF)	1.042	0.628	[0.466, 0.789]	2.55×10^{-14}
	unbiased CKA (RBF)	1.063	0.613	[0.458, 0.768]	8.33×10^{-15}

Table 33: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.654	0.182	[0.114, 0.249]	1.36×10^{-7}
	unbiased CKA (linear)	0.656	0.180	[0.113, 0.247]	1.40×10^{-7}
	CKA (RBF)	0.666	0.169	[0.102, 0.236]	7.80×10^{-7}
	unbiased CKA (RBF)	0.670	0.165	[0.099, 0.231]	8.44×10^{-7}
GSM8K	CKA (linear)	0.731	0.117	[0.076, 0.159]	3.98×10^{-8}
	unbiased CKA (linear)	0.733	0.115	[0.074, 0.157]	4.77×10^{-8}
	CKA (RBF)	0.726	0.121	[0.078, 0.165]	3.29×10^{-8}
	unbiased CKA (RBF)	0.730	0.118	[0.075, 0.160]	4.21×10^{-8}
MATH	CKA (linear)	0.701	0.148	[0.098, 0.197]	4.32×10^{-9}
	unbiased CKA (linear)	0.703	0.146	[0.097, 0.195]	4.16×10^{-9}
	CKA (RBF)	0.691	0.158	[0.105, 0.211]	4.75×10^{-9}
	unbiased CKA (RBF)	0.694	0.155	[0.103, 0.207]	4.03×10^{-9}
TruthfulQA	CKA (linear)	0.712	0.131	[0.083, 0.178]	5.71×10^{-8}
	unbiased CKA (linear)	0.715	0.127	[0.081, 0.173]	6.40×10^{-8}
	CKA (RBF)	0.709	0.132	[0.083, 0.180]	8.78×10^{-8}
	unbiased CKA (RBF)	0.715	0.126	[0.080, 0.173]	8.95×10^{-8}

D.5 MIXED-EFFECTS REGRESSION RESULTS FOR MUTUAL INFORMATION, EXCLUDING THE LLAMA FAMILY

Tables 34∼41 show significant positive trends across all datasets, CKA variants, and games.

D.5.1 WHEN USING THE GLOBAL AVERAGE

Table 34: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.307	0.058	[0.020, 0.095]	0.00246
	unbiased CKA (linear)	0.308	0.058	[0.021, 0.095]	0.00232
	CKA (RBF)	0.311	0.052	[0.015, 0.090]	0.00633
	unbiased CKA (RBF)	0.312	0.052	[0.015, 0.089]	0.00584
GSM8K	CKA (linear)	0.323	0.058	[0.024, 0.092]	0.00083
	unbiased CKA (linear)	0.324	0.058	[0.024, 0.092]	0.00076
	CKA (RBF)	0.323	0.056	[0.022, 0.090]	0.00130
	unbiased CKA (RBF)	0.324	0.056	[0.022, 0.089]	0.00113
MATH	CKA (linear)	0.320	0.051	[0.018, 0.083]	0.00212
	unbiased CKA (linear)	0.320	0.051	[0.019, 0.083]	0.00192
	CKA (RBF)	0.317	0.053	[0.019, 0.086]	0.00193
	unbiased CKA (RBF)	0.318	0.053	[0.020, 0.085]	0.00165
TruthfulQA	CKA (linear)	0.318	0.056	[0.023, 0.090]	0.00097
	unbiased CKA (linear)	0.319	0.057	[0.024, 0.090]	0.00067
	CKA (RBF)	0.319	0.053	[0.020, 0.087]	0.00186
	unbiased CKA (RBF)	0.319	0.054	[0.022, 0.087]	0.00116

Table 35: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.078	0.138	[0.105, 0.171]	5.0×10^{-16}
	unbiased CKA (linear)	0.080	0.136	[0.103, 0.169]	4.8×10^{-16}
	CKA (RBF)	0.081	0.135	[0.102, 0.169]	2.2×10^{-15}
	unbiased CKA (RBF)	0.084	0.133	[0.100, 0.166]	1.9×10^{-15}
GSM8K	CKA (linear)	0.121	0.121	[0.095, 0.148]	6.0×10^{-19}
	unbiased CKA (linear)	0.123	0.120	[0.094, 0.147]	5.0×10^{-19}
	CKA (RBF)	0.119	0.120	[0.093, 0.147]	1.7×10^{-18}
	unbiased CKA (RBF)	0.122	0.118	[0.092, 0.145]	1.1×10^{-18}
MATH	CKA (linear)	0.108	0.120	[0.094, 0.146]	3.2×10^{-19}
	unbiased CKA (linear)	0.109	0.119	[0.093, 0.145]	2.5×10^{-19}
	CKA (RBF)	0.102	0.124	[0.097, 0.152]	1.0×10^{-18}
	unbiased CKA (RBF)	0.104	0.122	[0.095, 0.149]	6.8×10^{-19}
TruthfulQA	CKA (linear)	0.116	0.107	[0.079, 0.135]	4.3×10^{-14}
	unbiased CKA (linear)	0.117	0.107	[0.079, 0.134]	2.3×10^{-14}
	CKA (RBF)	0.115	0.103	[0.075, 0.132]	6.4×10^{-13}
	unbiased CKA (RBF)	0.118	0.103	[0.075, 0.130]	2.2×10^{-13}

Table 36: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.497	1.536	[1.242, 1.831]	1.7×10^{-24}
	unbiased CKA (linear)	0.510	1.522	[1.231, 1.814]	1.4×10^{-24}
	CKA (RBF)	0.590	1.404	[1.108, 1.700]	1.5×10^{-20}
	unbiased CKA (RBF)	0.615	1.381	[1.092, 1.670]	7.1×10^{-21}
GSM8K	CKA (linear)	1.106	0.958	[0.729, 1.188]	2.6×10^{-16}
	unbiased CKA (linear)	1.117	0.951	[0.724, 1.178]	2.2×10^{-16}
	CKA (RBF)	1.092	0.941	[0.712, 1.171]	9.8×10^{-16}
	unbiased CKA (RBF)	1.111	0.929	[0.704, 1.154]	6.4×10^{-16}
MATH	CKA (linear)	0.978	1.001	[0.772, 1.230]	1.0×10^{-17}
	unbiased CKA (linear)	0.987	0.996	[0.770, 1.222]	5.8×10^{-18}
	CKA (RBF)	0.913	1.061	[0.820, 1.302]	5.5×10^{-18}
	unbiased CKA (RBF)	0.932	1.048	[0.813, 1.283]	2.2×10^{-18}
TruthfulQA	CKA (linear)	0.991	1.014	[0.772, 1.256]	2.1×10^{-16}
	unbiased CKA (linear)	1.003	1.018	[0.780, 1.257]	5.8×10^{-17}
	CKA (RBF)	0.987	0.972	[0.727, 1.217]	7.6×10^{-15}
	unbiased CKA (RBF)	1.008	0.975	[0.737, 1.213]	9.7×10^{-16}

Table 37: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using a global average. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.592	0.185	[0.119, 0.250]	3.2×10^{-8}
	unbiased CKA (linear)	0.593	0.183	[0.118, 0.248]	3.2×10^{-8}
	CKA (RBF)	0.597	0.179	[0.107, 0.251]	1.1×10^{-6}
	unbiased CKA (RBF)	0.600	0.176	[0.105, 0.246]	1.0×10^{-6}
GSM8K	CKA (linear)	0.644	0.180	[0.108, 0.251]	7.8×10^{-7}
	unbiased CKA (linear)	0.646	0.179	[0.108, 0.250]	7.3×10^{-7}
	CKA (RBF)	0.642	0.176	[0.105, 0.247]	1.2×10^{-6}
	unbiased CKA (RBF)	0.645	0.175	[0.105, 0.246]	9.7×10^{-7}
MATH	CKA (linear)	0.623	0.182	[0.117, 0.247]	4.3×10^{-8}
	unbiased CKA (linear)	0.624	0.181	[0.117, 0.245]	3.3×10^{-8}
	CKA (RBF)	0.613	0.188	[0.120, 0.255]	4.5×10^{-8}
	unbiased CKA (RBF)	0.616	0.186	[0.120, 0.252]	2.8×10^{-8}
TruthfulQA	CKA (linear)	0.638	0.154	[0.083, 0.224]	1.9×10^{-5}
	unbiased CKA (linear)	0.640	0.153	[0.084, 0.223]	1.7×10^{-5}
	CKA (RBF)	0.638	0.146	[0.075, 0.216]	5.6×10^{-5}
	unbiased CKA (RBF)	0.642	0.145	[0.076, 0.214]	4.1×10^{-5}

D.5.2 When using the average of maximum-aligned scores

Table 38: Story writing task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.275	0.090	[0.054, 0.125]	5.87×10^{-7}
	unbiased CKA (linear)	0.275	0.089	[0.054, 0.124]	5.71×10^{-7}
	CKA (RBF)	0.281	0.083	[0.049, 0.118]	1.94×10^{-6}
	unbiased CKA (RBF)	0.282	0.082	[0.048, 0.116]	1.81×10^{-6}
GSM8K	CKA (linear)	0.318	0.047	[0.026, 0.068]	1.40×10^{-5}
	unbiased CKA (linear)	0.318	0.047	[0.026, 0.068]	1.36×10^{-5}
	CKA (RBF)	0.317	0.048	[0.026, 0.069]	1.99×10^{-5}
	unbiased CKA (RBF)	0.318	0.047	[0.025, 0.068]	1.89×10^{-5}
MATH	CKA (linear)	0.306	0.059	[0.034, 0.084]	3.48×10^{-6}
	unbiased CKA (linear)	0.307	0.059	[0.034, 0.083]	3.46×10^{-6}
	CKA (RBF)	0.302	0.063	[0.036, 0.090]	4.00×10^{-6}
	unbiased CKA (RBF)	0.303	0.062	[0.035, 0.088]	3.92×10^{-6}
TruthfulQA	CKA (linear)	0.307	0.058	[0.034, 0.082]	1.75×10^{-6}
	unbiased CKA (linear)	0.308	0.058	[0.034, 0.081]	1.43×10^{-6}
	CKA (RBF)	0.306	0.058	[0.033, 0.082]	3.66×10^{-6}
	unbiased CKA (RBF)	0.308	0.056	[0.033, 0.080]	2.63×10^{-6}

Table 39: Fictional biography generation task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.086	0.099	[0.072, 0.126]	3.39×10^{-13}
	unbiased CKA (linear)	0.087	0.098	[0.072, 0.124]	3.67×10^{-13}
	CKA (RBF)	0.089	0.097	[0.071, 0.123]	1.93×10^{-13}
	unbiased CKA (RBF)	0.091	0.095	[0.070, 0.121]	2.12×10^{-13}
GSM8K	CKA (linear)	0.124	0.070	[0.055, 0.086]	4.29×10^{-19}
	unbiased CKA (linear)	0.125	0.069	[0.054, 0.085]	5.20×10^{-19}
	CKA (RBF)	0.121	0.073	[0.057, 0.089]	1.94×10^{-19}
	unbiased CKA (RBF)	0.123	0.071	[0.056, 0.087]	2.50×10^{-19}
MATH	CKA (linear)	0.112	0.080	[0.062, 0.098]	1.85×10^{-17}
	unbiased CKA (linear)	0.113	0.079	[0.061, 0.097]	2.34×10^{-17}
	CKA (RBF)	0.106	0.086	[0.066, 0.106]	2.03×10^{-17}
	unbiased CKA (RBF)	0.108	0.084	[0.064, 0.103]	3.01×10^{-17}
TruthfulQA	CKA (linear)	0.116	0.073	[0.055, 0.091]	9.57×10^{-16}
	unbiased CKA (linear)	0.118	0.071	[0.054, 0.089]	9.84×10^{-16}
	CKA (RBF)	0.115	0.073	[0.055, 0.091]	4.75×10^{-15}
	unbiased CKA (RBF)	0.118	0.070	[0.053, 0.088]	4.28×10^{-15}

Table 40: Haiku composition task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.628	1.048	[0.823, 1.274]	8.91×10^{-20}
	unbiased CKA (linear)	0.637	1.040	[0.816, 1.264]	8.54×10^{-20}
	CKA (RBF)	0.702	0.970	[0.750, 1.189]	4.29×10^{-18}
	unbiased CKA (RBF)	0.719	0.953	[0.738, 1.168]	3.67×10^{-18}
GSM8K	CKA (linear)	1.104	0.602	[0.472, 0.733]	1.22×10^{-19}
	unbiased CKA (linear)	1.112	0.595	[0.466, 0.723]	1.24×10^{-19}
	CKA (RBF)	1.088	0.613	[0.479, 0.747]	3.39×10^{-19}
	unbiased CKA (RBF)	1.103	0.599	[0.468, 0.730]	3.27×10^{-19}
MATH	CKA (linear)	0.984	0.710	[0.554, 0.865]	3.73×10^{-19}
	unbiased CKA (linear)	0.992	0.701	[0.547, 0.855]	3.93×10^{-19}
	CKA (RBF)	0.934	0.756	[0.589, 0.923]	8.06×10^{-19}
	unbiased CKA (RBF)	0.951	0.739	[0.575, 0.903]	9.36×10^{-19}
TruthfulQA	CKA (linear)	0.990	0.701	[0.551, 0.852]	6.29×10^{-20}
	unbiased CKA (linear)	1.006	0.686	[0.539, 0.833]	6.31×10^{-20}
	CKA (RBF)	0.986	0.693	[0.538, 0.847]	1.48×10^{-18}
	unbiased CKA (RBF)	1.013	0.669	[0.521, 0.817]	9.74×10^{-19}

Table 41: Vacation brainstorming task. We fit a mixed-effects regression between mutual information and representational similarity. Here, the summary similarity score is calculated using the average of maximum-aligned scores. β indicates the coefficient of similarity, and the 95% CI represents the confidence interval of β.

Dataset	Metric	Intercept	β	95% CI	p
WikiText	CKA (linear)	0.582	0.161	[0.086, 0.235]	2.21×10^{-5}
	unbiased CKA (linear)	0.583	0.159	[0.086, 0.233]	2.22×10^{-5}
	CKA (RBF)	0.596	0.144	[0.072, 0.216]	9.66×10^{-5}
	unbiased CKA (RBF)	0.599	0.141	[0.070, 0.212]	9.68×10^{-5}
GSM8K	CKA (linear)	0.651	0.099	[0.055, 0.142]	9.04×10^{-6}
	unbiased CKA (linear)	0.653	0.097	[0.054, 0.140]	1.01×10^{-5}
	CKA (RBF)	0.648	0.102	[0.057, 0.147]	8.23×10^{-6}
	unbiased CKA (RBF)	0.650	0.099	[0.055, 0.143]	9.33×10^{-6}
MATH	CKA (linear)	0.624	0.129	[0.077, 0.181]	1.13×10^{-6}
	unbiased CKA (linear)	0.625	0.128	[0.077, 0.180]	9.97×10^{-7}
	CKA (RBF)	0.615	0.137	[0.081, 0.193]	1.34×10^{-6}
	unbiased CKA (RBF)	0.617	0.136	[0.081, 0.190]	9.86×10^{-7}
TruthfulQA	CKA (linear)	0.633	0.114	[0.064, 0.164]	7.39×10^{-6}
	unbiased CKA (linear)	0.635	0.112	[0.063, 0.161]	7.23×10^{-6}
	CKA (RBF)	0.632	0.113	[0.062, 0.163]	1.43×10^{-5}
	unbiased CKA (RBF)	0.637	0.109	[0.060, 0.158]	1.25×10^{-5}