EVALUATING STEERING TECHNIQUES USING HUMAN SIMILARITY JUDGMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Current evaluations of Large Language Model (LLM) steering techniques focus on task-specific performance, overlooking how well steered representations align with human cognition. Using a well-established triadic similarity judgment task, we assessed steered LLMs on their ability to flexibly judge similarity between concepts based on size or kind, two central dimensions organizing human mental representations. We found that prompt-based steering methods outperformed other methods both in terms of steering accuracy and model-to-human alignment. We also found LLMs were biased towards 'kind' similarity and struggled with 'size' alignment. This evaluation approach, grounded in human cognition, adds further support to the efficacy of prompt-based steering and reveals privileged representational axes in LLMs prior to steering.

1 INTRODUCTION

Central to the flexibility of human cognition is the ability to marshal both one's knowledge of the world and the current context to guide behavior (Wong et al., 2025). Different aspects of a learned concept can be preferentially activated to execute different tasks. For example, the *size* of an orange might be important when a shopper is deciding how many can fit inside their shopping basket, whereas the *kind* of produce it is (i.e., fruit) might be important when a grocer is arranging their wares on a shelf. Such flexible behavior is facilitated by learning robust representations of concepts, captured by theories and models of semantic knowledge (Rogers & McClelland, 2004; Rogers, 2024; Saxe et al., 2019), and by 'guiding' these representations through context-sensitive mechanisms, captured by accounts of cognitive/semantic control (Cohen et al., 1990; Miller & Cohen, 2001; Ralph et al., 2017; Giallanza et al., 2024).

These aspects of human semantic cognition—representation and control—provide strong analogies to two important axes of evaluation for large language models (LLMs): (1) representation learning through pre-training and (2) intervention-based steering. This isomorphism is important insofar as cognitive scientists have developed methods for characterizing human mental representations under varied contexts for naturalistic tasks that underpin a variety of human behaviors. Here, we focus on evaluating current LLM steering techniques through the lens of cognitive science-inspired methods, which constitute a critical complement to efforts in interpretability research, primarily stemming from the computer sciences. This allows for evaluating steering techniques not just in terms of performance, but in terms of how aligned the resultant model behaviors and representations are to those of humans when presented with qualitatively similar interventions.

In the present work, we used a triadic similarity judgment task (Sievert et al., 2023; Hebart et al., 2020), a technique that has been shown to be effective at characterizing human mental representations, where humans judged the similarity between concepts in terms of either **size** or **kind**. We then tested a suite of LLM intervention steering techniques on `gemma2-27b` and `gemma2-9b` and assessed each method's (1) *accuracy*, measured as the number of 'correct' judgments on the size and kind tasks, and (2) *human alignment*, measured as the Procrustes correlation between embeddings generated from human judgments versus those generated from model judgments. Model accuracy differed dramatically across steering methods, with some approaches yielding human-level performance; but surprisingly, human alignment was poor, especially for size judgments across all steering methods. These results suggest critical differences between semantic control in humans and large language models.

2 RELATED WORK

2.1 TRIADIC JUDGMENT TASKS

Triadic judgment tasks, coupled with advances in embedding algorithms (Jamieson et al., 2015; Sievert et al., 2023; Muttenthaler et al., 2022) have enabled the characterization of the representational geometry underlying human concepts (Jamieson et al., 2015; Sievert et al., 2023; Muttenthaler et al., 2022; Hebart et al., 2020; Muttenthaler et al., 2023b; Suresh et al., 2024; Giallanza et al., 2024). The general approach is to present participants with three concepts and to ask them to indicate either the odd one out or which of two options is most similar to a designated target (Sievert et al., 2023; Muttenthaler et al., 2022). These judgments are used to estimate a low-dimensional embedding where similar concepts are closer together. These techniques have been extended to AI systems, specifically vision models (Muttenthaler et al., 2023a; Mukherjee & Rogers, 2025), to estimate how human-like neural network representations are (see Sucholutsky et al., 2023). Judgments made by LLMs on such tasks can also be used to estimate model embeddings, helping uncover representational structures in otherwise opaque systems (Hebart et al., 2020; Sucholutsky et al., 2023). This approach has revealed fundamental differences: human conceptual structures remain consistent across cultures, while LLMs exhibit task-dependent variation (Suresh et al., 2023). Applications extend to standardized similarity norms (Hout et al., 2022) and semantic organization principles (Mirman et al., 2017).

2.2 MODEL STEERING METHODS

Although language models are already context-sensitive due to their outputs being conditioned on prior text, steering methods further control behavior via internal interventions or structured prompts. Because these methods allow for fine-grained control over LLM behavior at inference time, they offer a potential alternative to more expensive methods–such as supervised fine tuning (SFT)–which directly modify model weights (Wu et al., 2025).

Prompting In many instances, model outputs can be steered directly through prompt instructions provided in natural language (Sahoo et al., 2025). *Zero-shot prompting* involves providing a single task instruction or example to direct LLM outputs, and is capable of shaping responses to support emergent behaviors including multi-step reasoning (Kojima et al., 2023), social simulation (Chuang et al., 2024) and image classification (Abdelhamed et al., 2025). Even without explicit instruction, LLMs can induce general rules and abstractions from minimal examples via *in-context learning*. This is sufficient to steer model behavior for translation tasks, (Brown et al., 2020), logical reasoning (Yin et al., 2024), and task induction (Honovich et al., 2022).

Task vectors Hendel et al. (2023) find that LLMs encode in-context task instructions as a linear direction in intermediate layers. This direction, θ, is a compressed representation of the abstract rule underlying the sequence. When θ is extracted from the forward pass on a rule-based prompt x and patched into the forward pass of an empty prompt x' with no task instructions, the model is able to produce the correct sequence completion for the empty prompt. Initial work by Hendel et al. (2023) demonstrates the efficacy of task vectors for steering model behavior in domains including translation, semantic knowledge (country-capitol mappings) and syntactic rules. Todd et al. (2024) extends this general paradigm to recover finer-grained representations from the activations of individual attention heads. Subsequent work has since applied task vectors to causal interventions in a variety of applications from visual image-masking tasks (Hojel et al., 2024) to more naturalistic domains (Kang et al., 2025), while improvements to the paradigm itself include training mechanism modifications to induce task vectors (Yang et al., 2025) and the superposition of multiple task vectors during in-context learning (Xiong et al., 2024).

Difference-in-means (DiffMean) While task vectors demonstrate, in-practice, that transformer hidden layers encode abstract rules for task representations, these representations are often noisy as a result of extracting the *entire* hidden state at a given token position (Hendel et al., 2023). Marks & Tegmark (2024) introduce a fine-grained method for extracting the steering vector as a linear direction θ in the activation space that represents the difference between the mean hidden states of positive and negative prompt examples. This vector, represented as θ, consistently shifts model

activations along the direction θ, from positive to negative or negative to positive (where *negative* and *positive* are examples along some dimension). Marks & Tegmark (2024) demonstrate that these vectors capture significantly less noise compared to simple linear probes, and generalize to a range of naturalistic domains.

Sparse autoencoders A key difficulty in extracting steering directions from model activations is the dense nature of these activations: a single neuron will activate for many different tasks and concepts as a part of the distributions representing those concepts. *Sparse autoencoders* attempt to solve this issue by learning a compressed set of activations $\mathbf{z}$ corresponding to the activations in a given transformer layer l, with a sparsity penalty applied $\mathcal{L}_{\text{sparse}}$ (Cunningham et al., 2023). This allows for the unsupervised discovery of human-interpretable concepts, represented by the activations of single SAE neurons. Formally, the SAE objective is

$$\mathcal{L} = \|\mathbf{h}^{(l)} - D(E(\mathbf{h}^{(l)}))\|_2^2 + \lambda \mathcal{L}_{\text{sparse}},$$

where $\mathbf{h}^{(l)}$ are the transformer activations, E and D are the encoder and decoder, and λ is the sparsity penalty. The concepts identified by SAEs range from world knowledge and semantic properties to syntactic features and compositional abstractions (Shu et al., 2025). Interestingly, SAE neurons can be used to steer model outputs along dimensions of interest, such as translating LLM outputs to French given the activation of a corresponding latent feature in the sparse autoencoder (He et al., 2025). While the extent to which SAEs provide a practical means of model steering and interpretability is disputed (Wu et al., 2025), subsequent work provides methods for obtaining more interpretable, coherent concepts (Bussmann et al., 2025; Rajamanoharan et al., 2024).

2.3 Performance versus alignment

Task performance and representational alignment are distinct axes of evaluation. Linsley et al. (2023) found that higher object recognition accuracy led to poorer alignment with human visual cortex. Liu et al. (2023) reported tradeoffs between alignment and task performance. This suggests that models can match human performance but rely on different internal representations (Piantadosi & Hill, 2021). As De Bruin et al. (2024) argue, evaluations should assess not only what systems do, but how. Thus, we evaluate both competence and alignment.

3 Methods

3.1 Dataset

We use the Round Things Dataset (Giallanza et al., 2024) to evaluate these aspects. This dataset comprises 46 concrete objects: 25 human-made artifacts and 21 natural kinds (fruits and vegetables). All items are roughly spherical in shape, varying along a discrete *kind* dimension (artifacts versus plants) and a continuous *size* dimension.

Dataset design. The Round Things Dataset was specifically designed to investigate how semantic representations can be contextually modulated along orthogonal dimensions. The choice of roughly spherical objects serves multiple methodological purposes: (1) it facilitates size comparisons by minimizing shape-based confounds, (2) it ensures that size judgments are based primarily on scale rather than geometric complexity, and (3) it creates a controlled stimulus set where the two key dimensions of interest—kind and size—are orthogonal rather than confounded.

The dataset was constructed to include commonly known objects to ensure high familiarity across participants. Each item possesses an empirically-derived ground-truth average diameter obtained through systematic internet searches. Specifically, the authors conducted Google searches using the phrase "average diameter of a _____," recording the answer yielded by search engine summaries. When searches returned size ranges, the midpoint was taken as the median diameter. For items that naturally occur in multiple sizes (e.g., pumpkins, yoga balls), searches were refined by adding "medium-size" to obtain more standardized measurements.

Data composition. The 46 items are distributed across two semantic categories: 25 human-made artifacts (such as balls, spherical tools, and round household objects) and 21 fruits and vegetables

(such as apples, oranges, and other roughly spherical produce). Importantly, object sizes are approximately balanced across the two categories, ensuring that neither semantic domain is systematically associated with larger or smaller items. This balanced distribution is crucial for investigating how semantic category membership interacts with size-based similarity judgments without confounding effects.

The dataset thus provides stimuli that differ along two clear but unrelated abstract semantic dimensions: the discrete taxonomic dimension of kind (artifact versus natural) and the continuous physical dimension of size. This orthogonal structure allows for controlled investigation of how task context can selectively emphasize one dimension while de-emphasizing another, and whether such emphasis preserves or eliminates information from the non-target dimension.

Human behavioral data collected using this dataset has been used to derive two-dimensional embeddings via crowd kernel ordinal embedding algorithms, allowing for quantitative analysis of how different task contexts reshape the geometric organization of semantic space while preserving cross-context similarities.

3.2 Triadic similarity judgment task

We employ a triadic judgment task where both humans and LLMs identify which of two items x_1 or x_2 is most similar to a reference item x_{ref} along a specified semantic dimension. Participants are instructed to judge which of the two options is most similar to the target item either "in terms of size" or as "a more similar kind of thing."

This paradigm is motivated by the hypothesis that when people estimate similarity along specific dimensions, they use task instructions to steer their semantic representations toward task-relevant features. We test whether different LLM steering methods can similarly modulate representations to achieve both high accuracy (competence) and human-like similarity patterns (alignment).

The dataset's balanced design—with sizes approximately distributed across both semantic categories—enables investigation of how task instructions influence the relative weighting of size versus kind information in similarity computations. This provides insights into whether context-dependent steering preserves or eliminates information from non-target dimensions, revealing the flexibility and constraints of semantic representations under controlled processing conditions.

3.3 Steering methods

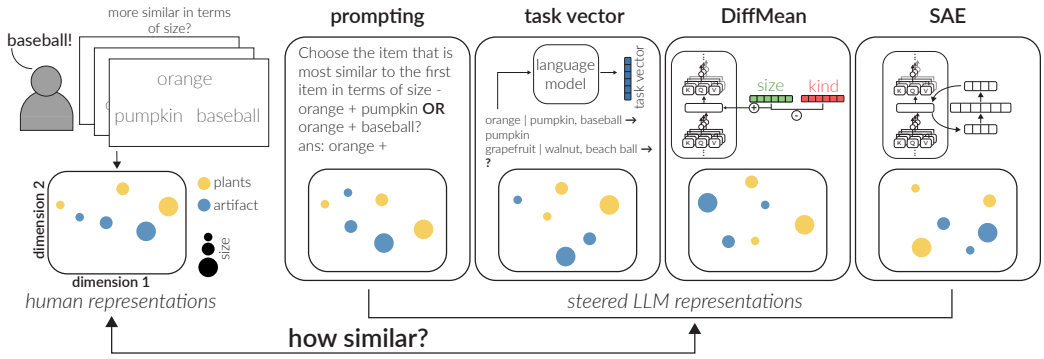


Figure 1: General workflow for deriving human and steered model embeddings.

We evaluated four different steering methods—prompting, task vectors, DiffMean, and SAEs—alongside two prompting baselines (zero-shot and in-context) for comparison. All methods were tested on two open-source LLMs capable of running on a single NVIDIA H100 GPU: `gemma2-27b` and `gemma2-9b`. Figure 4 shows the general workflow for applying these steering methods to the triadic judgment task.

For the prompting-based approaches, different steering conditions required tailored prompts to elicit appropriate similarity judgments along specific semantic dimensions. Each prompt was designed to

direct the model's attention toward either kind-based or size-based similarity, or to allow for general similarity assessment in the neutral condition, as shown in Table 1. All trials received independent prompt evaluation to ensure consistent application of each steering method.

For the activation-based steering methods (task vectors, DiffMean, and SAEs), we applied the corresponding interventions to the model's internal representations while using a consistent neutral prompt format across all trials. This approach allows us to isolate the effects of representational steering from prompt-based steering, providing a comprehensive comparison of different steering mechanisms.

Detailed implementation specifics for each steering method are provided in Appendix A.1.

Table 1: Prompt templates used for different steering conditions. The placeholder {p} is replaced with the triplet comparison question: "Which item is most similar to {item_x}: {item_y} or {item_z}?"

Condition	**Prompt Template**
Kind	`Choose the second item that is most similar to the first item in terms of the KIND of thing it is. Respond only with the name of the item exactly as written. {p}`
Size	`Choose the second item that is most similar to the first item in terms of SIZE. Respond only with the name of the item exactly as written. {p}`
Neutral	`Choose the second item that is most similar to the first item. Respond only with the name of the item exactly as written. {p}`

3.4 Analysis Approach

Embedding algorithm. We applied an ordinal embedding algorithm (Tamuz et al., 2011; Sievert et al., 2023) to similarity judgments from both humans and models, creating semantic spaces where frequently co-judged similar items were positioned closer together. The embedding construction process minimized crowd-kernel triplet loss (Tamuz et al., 2011) by optimizing Euclidean distances between word pairs. To ensure reliable estimates, we reserved 20% of our data for validation purposes and monitored crowd-kernel loss throughout the fitting procedure.

Under this framework, each triplet judgment yields an ordinal constraint of the form "item i is closer to item j than to item k." Let $D^\star$ be the squared Euclidean distance matrix of unknown embedding points $\{x_i^\star\}_{i=1}^n \subset \mathbb{R}^d$. The ordinal embedding model assumes a monotone link function (e.g., logistic):

$$\Pr\left[y_{(i;j,k)} = 1\right] \;=\; f\big(D^\star_{ik} - D^\star_{ij}\big), \qquad f(0) = \tfrac{1}{2}. \tag{1}$$

The ordinal embedding algorithm minimizes an empirical surrogate of the negative log-likelihood using crowd-kernel triplet loss over a low-rank parameterization.

Sample complexity considerations. If the true embedding has rank d and triplet judgments are sampled approximately uniformly at random, then with high probability the out-of-sample prediction error is

$$\mathcal{E}_{\text{pred}} \;=\; \tilde{O}\left(\sqrt{\frac{d\,n\log n}{|S|}}\right), \qquad \Rightarrow \quad |S| \;=\; \tilde{\Theta}\big(d\,n\log n\big) \tag{2}$$

Thus, $\tilde{\Theta}(nd\log n)$ triplet judgments suffice for accurate prediction of new comparisons, and at least $\Omega(d\,n\log n)$ ordinal comparisons are information-theoretically necessary (Jain et al., 2016). For our implementation with $n = 46$ items and $d = 2$ dimensions, we collected $N_{\text{triplets}} = c \cdot d \cdot n\log n$ triplet judgments per steering method (where $c \approx 5$ based on our 2,500 judgments), balancing statistical efficiency with computational constraints.

Accuracy measurement. We evaluated the competence of each steering method by measuring LLM accuracy on the triadic judgment task compared to ground-truth and human performance. Human accuracy benchmarks were derived from human embeddings by evaluating the same 2,400 triplets used in model assessments: when the Euclidean distance from x_{ref} to x_1 was shorter than from x_{ref} to x_2, we classified x_1 as the correctly identified more similar item. This approach enables direct comparison between human and model performance on identical triplet sets, providing a unified measure of task competence across all steering conditions.

Alignment analysis. To quantify alignment between human and LLM representations, our main measure of human-model alignment was the amount of variance in human semantic embeddings explained by model-based semantic embeddings (R^2) after Procrustes aligning the two spaces (**?**). Procrustes transformation aligns two vector spaces of equal dimensionality by finding the optimal set of linear transformations (rotations, scalings, and translations) to minimize the sum of squared distances between corresponding points in the two spaces (residual SSE). Using the sum of squared distances in the human embeddings (human SSE) as the target, we computed an R^2 metric as:

$$R^2 = 1 - \frac{\text{residual SSE}}{\text{human SSE}}$$

This approach measures how well variations in pairwise distances in LLM-derived representations correspond to those in human representations after allowing for optimal linear transformations in the representational geometry. Since both embedding spaces are of the same dimensionality and permitting these linear transformations makes our alignment estimate maximally generous, this provides a robust upper bound on the correspondence between human and model semantic representations.

4 Results

4.1 Model representations show default kind alignment

Figure 3 shows how successful different steering methods were in guiding LLMs to make accurate kind and size judgments. Neutral prompts that simply asked LLMs to indicate which of the two options was most similar to the target, without additional context, were better aligned to kind judgments than to size judgments. This is evidenced by the high accuracy of neutral prompts for kind judgments relative to their lower accuracy for size judgments ($\beta = 0.187$, $p < 0.001$). A similar pattern arose for human alignment (bottom row of Figure 3): neutral prompt embeddings were moderately aligned with human kind embeddings but weakly aligned with human size embeddings ($R^2_{\text{neutral, kind}} = 0.50$ vs. $R^2_{\text{neutral, size}} = 0.02$, $p < 0.001$).

4.2 Prompting outperforms other steering methods at predicting ground truth

Intervention methods (SAEs, task vectors, and DiffMean) consistently performed worse than prompting ($\beta_{\text{TV}} = -0.29$, $\beta_{\text{DM}} = -0.30$, $\beta_{\text{SAE}} = -0.29$, $p < 0.001$), particularly for kind judgments ($\beta_{\text{kind}} = 0.19$, $p < 0.001$). For size judgments, although prompting methods result in higher accuracies than non-prompting methods, only the zero-shot size prompt produced higher representational alignment than the other methods. Interestingly, we observed higher alignment with human representations in both task vector conditions for the smaller-parameter `gemma2-9b`,

5 Conclusion

We evaluated a set of popular LLM steering techniques using a well-established method in the cognitive sciences—triadic similarity judgments. By using both steered LLM and human judgments as input into embedding algorithms that capture an agent's representational geometry, we applied a commensurate standard to evaluate how human-like different steering methods are. Critically, we looked beyond raw accuracy (competence) and evaluated steered models on how strongly their embeddings aligned with human embeddings. Similar to prior work (Wu et al., 2025), we found that prompting methods tended to outperform other steering techniques in terms of both accuracy and

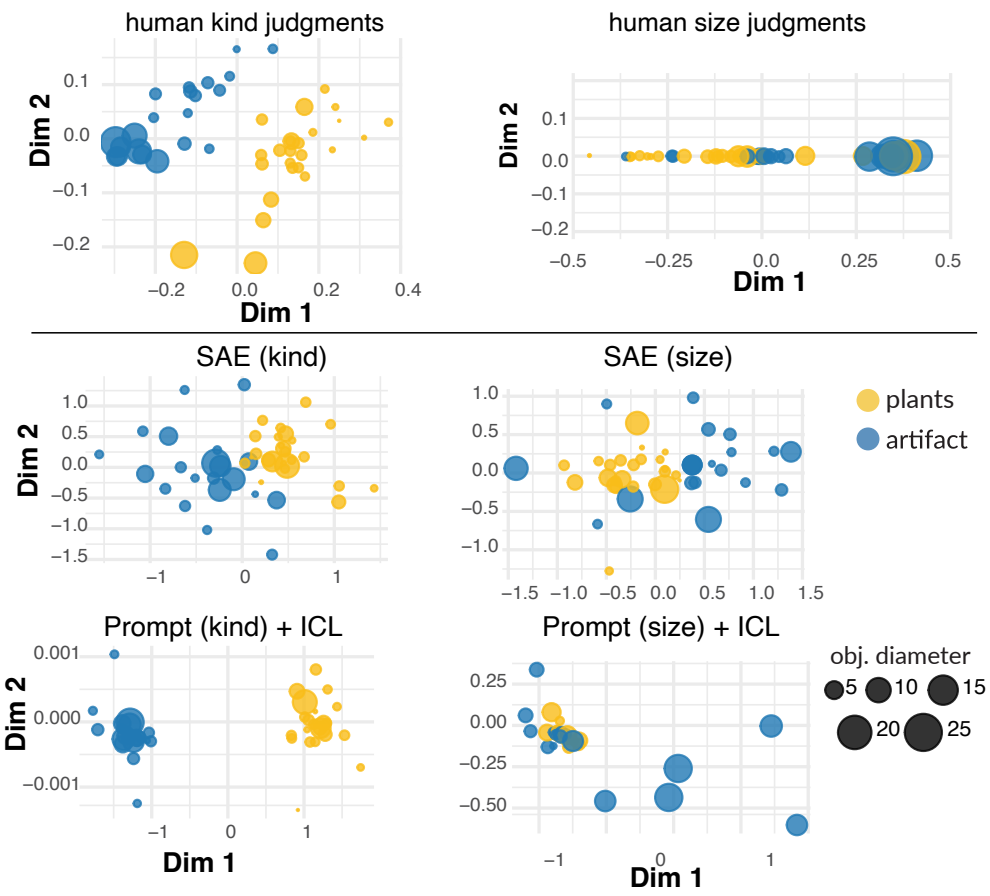


Figure 2: Representational geometry of the concepts in the Round Things Dataset based on embeddings derived from triadic judgments for humans (top row) and `gemma2-27b` (bottom 2 rows) using two steering methods. Plots for all methods can be seen in A.5

alignment. We also found that LLMs, without any steering, tended to be predisposed to privilege the *kind* dimension of similarity over the *size* dimension, indicated both by the stronger alignment of the neutral prompt condition with embeddings from human kind judgments and by the difficulty steered models had aligning with human size representations.

Taken together, these findings provide a novel perspective on how steering methods can and should be evaluated in order to assess how aligned steered model representations are to those of people. We find converging evidence that prompt-based steering is currently the best route for both accurate and aligned steering, and that some axes such as kind are privileged over size. Future work should seek to further integrate insights from controlled semantic cognition (Giallanza et al., 2024) to uncover the basis of prompting's success in guiding LLMs' learned representations in a context-sensitive manner, and to test a wider variety of contexts beyond size and kind judgments.

6 Limitations

Model suite. Due to compute constraints we only evaluated models that could be run on consumer-grade GPUs (`gemma2-9b` and `gemma2-27b`). Further, we did not evaluate `gemma2-27b` on the SAE method due to the lack of available trained SAEs for that model. Despite this, we believe the presented results constitute a fair comparison between methods. Future work can scale up this approach to larger models. The efficacy of different steering methods for both accuracy and alignment

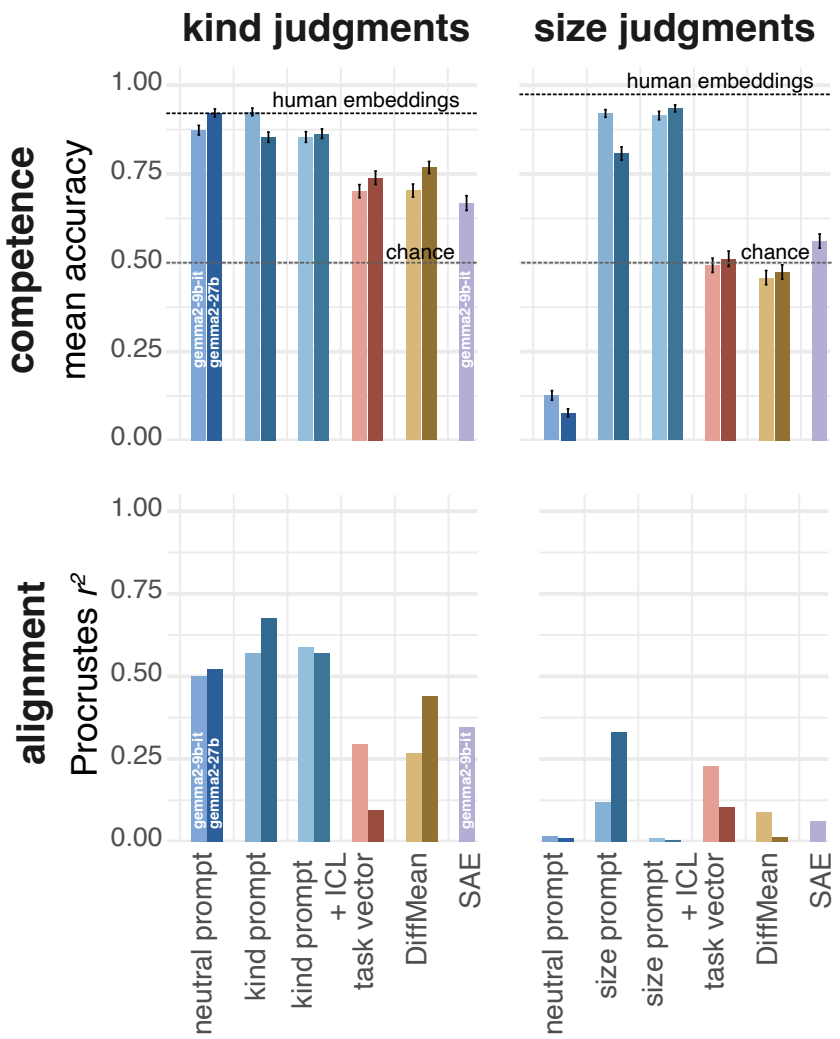


Figure 3: Steering accuracy (top row) and alignment of steered LLM representations to human representations (bottom row) for each steering technique. the dashed line labeled 'human embeddings' corresponds to how accurately human judgments can be predicted from human embeddings. SAE results are reported only for `gemma-9b-it`.

may scale differently with model parameter size. Between the 9B–27B scale we did not observe drastic changes, but this could change at larger scales.

Method suite. While we evaluated a representative set of steering methods for 'online' steering based on prompts or activation changes, we did not exhaustively test alternative 'in-weight' methods of changing model behavior (Anand et al., 2024), including supervised fine-tuning (SFT), low-rank adaptation (LoRA), and others. Since we are modeling human steering, which is thought to invoke online 'control' processes as opposed to in-weight learning (Giallanza et al., 2024), we argue that our chosen set of methods provided fair comparisons to the human data. Nevertheless, future work should compare the relative efficacy of in-weight vs. in-context styles of steering.

Order effects. While we randomized the order in which concepts were presented to the models for the triplet task—a common method in the behavioral sciences to protect against order effects—we note that we only ran the LLM evaluations with a single ordering. Since humans were also only presented with a single overall ordering, and our goal was to make commensurate comparisons across agent types, we chose to trade off robustness against order effects for fairer comparisons.

Limited evaluation methods. We relied on the triadic similarity judgment task since it is a well-established and validated method for evaluating human mental representations. In practice, there are more complex and naturalistic behaviors that humans perform that require 'steering' semantic representations (Giallanza et al., 2024). In the future, it will be crucial to incorporate such tasks when benchmarking different language model steering methods.

REFERENCES

Abdelrahman Abdelhamed, Mahmoud Afifi, and Alec Go. What do you see? enhancing zero-shot image classification with multimodal large language models, 2025. URL https://arxiv.org/abs/2405.15668.

Suraj Anand, Michael A Lepori, Jack Merullo, and Ellie Pavlick. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. *arXiv preprint arXiv:2406.00053*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders, 2025. URL https://arxiv.org/abs/2503.17547.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Beyond demographics: Aligning role-playing llm-based agents using human belief networks, 2024. URL https://arxiv.org/abs/2406.17232.

Jonathan D Cohen, Kevin Dunbar, and James L McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, 97(3):332, 1990.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. doi: 10.48550/arXiv.2309.08600.

Jos De Bruin, Thomas Bourguignon, Mouhamadou Biran, Walid Saoud, Pierre Morizet-Mahoudeaux, Karine Tasso, Nicolas Chanez, Laurent Perrinet, Kathinka Evers, Claire Montfroy, et al. Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1):17428, 2024.

Tyler Giallanza, Declan Campbell, Jonathan D Cohen, and Timothy T Rogers. An integrated model of semantics and control. *Psychological Review*, 2024.

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models, 2025. URL https://arxiv.org/abs/2502.11356.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris Ian Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature Human Behaviour*, 4(11):1173–1185, 2020.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023. URL https://doi.org/10.48550/arXiv.2310.15916. Accepted at Findings of EMNLP 2023.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors, 2024. URL https://arxiv.org/abs/2404.05729.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions, 2022. URL https://arxiv.org/abs/2205.10782.

Michael C Hout, Arryn Robbins, Hayward J Godwin, Gemma Fitzsimmons, and Collin Scarince. Visual and semantic similarity norms for a photographic image stimulus set containing recognizable objects, animals and scenes. *Journal of Open Psychology Data*, 10(1), 2022.

Lalit Jain, Kevin Jamieson, and Robert Nowak. Finite sample prediction and recovery bounds for ordinal embedding, 2016. URL https://arxiv.org/abs/1606.07081.

Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. Next: A system for real-world development, evaluation, and application of active learning. *Advances in neural information processing systems*, 28, 2015.

Joonseong Kang, Soojeong Lee, Subeen Park, Sumin Park, Taero Kim, Jihee Kim, Ryunyi Lee, and Kyungwoo Song. Adaptive task vectors for large language models, 2025. URL https://arxiv.org/abs/2506.03426.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Drew Linsley, Ivan F Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *arXiv preprint arXiv:2306.03779*, 2023.

Wenhao Liu, Xiaohua Wang, Zihan Ye, Jingwei Zhang, Hanchao Tang, Zhi Yang, Chuanyang Wang, Zhicheng Xu, Yiqi Zhou, Xiaocheng Wu, et al. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2024.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Daniel Mirman, Jon-Frederick Landrigan, and Allison E Britt. Taxonomic and thematic semantic systems. *Psychological Bulletin*, 143(5):499–520, 2017.

Kushin Mukherjee and Timothy T Rogers. Using drawings and deep neural networks to characterize the building blocks of human visual similarity. *Memory & Cognition*, 53(1):219–241, 2025.

Lukas Muttenthaler, Charles Y Zheng, Patrick McClure, Robert A Vandermeulen, Martin N Hebart, and Francisco Pereira. Vice: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, 35:33661–33675, 2022.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *International Conference on Learning Representations*, 2023a.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Steven T Piantadosi and Felix Hill. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 118(43):e1905334118, 2021.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. URL https://arxiv.org/abs/2407.14435.

Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1):42–55, 2017.

Timothy T Rogers. Generalization and abstraction: Human memory as a magic library. *The Oxford Handbook of Human Memory, Two Volume Pack: Foundations and Applications*, pp. 172, 2024.

Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL https://arxiv.org/abs/2402.07927.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models, 2025. URL https://arxiv.org/abs/2503.05613.

Scott Sievert, Robert Nowak, and Timothy T Rogers. Efficiently learning relative similarity embeddings with crowdsourcing. *Journal of open source software*, 8(84), 2023.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 722–738, 2023.

Siddharth Suresh, Wei-Chun Huang, Kushin Mukherjee, and Timothy T. Rogers. Categories vs semantic features: What shape the similarities people discern in photographs of objects? In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=iE5aXw3RFd.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2024. URL https://arxiv.org/abs/2310.15213.

Lionel Wong, Katherine M Collins, Lance Ying, Cedegao E Zhang, Adrian Weller, Tobias Gerstenberg, Timothy O'Donnell, Alexander K Lew, Jacob D Andreas, Joshua B Tenenbaum, et al. Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv preprint arXiv:2507.12547*, 2025.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

Zheyang Xiong, Ziyang Cai, John Cooper, Albert Ge, Vasilis Papageorgiou, Zack Sifakis, Angeliki Giannou, Ziqian Lin, Liu Yang, Saurabh Agarwal, Grigorios G Chrysos, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Everything everywhere all at once: Llms can in-context learn multiple tasks in superposition, 2024. URL https://arxiv.org/abs/2410.05603.

Liu Yang, Ziqian Lin, Kangwook Lee, Dimitris Papailiopoulos, and Robert Nowak. Task vectors in in-context learning: Emergence, formation, and benefit, 2025. URL https://arxiv.org/abs/2501.09240.

Qingyu Yin, Xuzheng He, Luoao Deng, Chak Tou Leong, Fan Wang, Yanzhao Yan, Xiaoyu Shen, and Qiang Zhang. Deeper insights without updates: The power of in-context learning over fine-tuning, 2024. URL https://arxiv.org/abs/2410.04691.

A STEERING METHOD AND EMBEDDING METHOD IMPLEMENTATIONS

A.1 DETAILED IMPLEMENTATION OF STEERING METHODS

Let f be a decoder-only language model with L layers and hidden size d. Each triplet comparison is denoted as $t_i = (x_{\text{ref},i}, x_{1,i}, x_{2,i})$ for $i = 1, \ldots, n$. Each prompt consists of a sequence of n triplets $[t_1, \ldots, t_n]$, serialized into a token sequence $x = [x_1, \ldots, x_T]$. The model produces hidden states $h_j^l \in \mathbb{R}^d$ at each token position x_j and layer l.

For all prompting, prompts are formatted as natural language strings of the form:

```
Choose the item that is most similar to the first item
in terms of <d>.  Respond only with the name of the
item exactly as written.
<x_ref> + <x_1> OR <x_ref> + <x_2>?
answer:  <x_ref> +
```

For steering methods, training examples consist of triplets without a natural language instruction like so:

```
<x_ref> + <x_1> OR <x_ref> + <x_2>?
<x_ref> + <x_answer>
```

Where examples are concatenated by commas. Each training example consisting of n triplets has a final incomplete training instance like so:

```
<x_ref> + <x_1> OR <x_ref> + <x_2>?
<x_ref> +
```

This "+" token is used to extract and steer representations for a corresponding "+" token in the zero-shot test example, which is identical to the final training example:

```
<x_ref> + <x_1> OR <x_ref> + <x_2>?
<x_ref> +
```

Fields are interpolated for some $d \in \{\textit{size}, \textit{kind}, \textit{neutral}\}$, triplet $t_n = (x_{\text{ref}}, x_1, x_2)$, and answer t_{answer} given the dimension d. We extract activations, logits, and apply all steering methods at the final input token $x_T =$ + in the last triplet t_n, depending on each method.

ZERO-SHOT PROMPT

In the zero-shot condition, the model is given a single triplet $t_n = (x_{\text{ref}}, x_1, x_2)$ and is asked to make a discrimination along a semantic dimension $d \in \{\textit{size}, \textit{kind}, \textit{neutral}\}$.

PROMPT WITH IN-CONTEXT EXAMPLES

In the in-context condition, the model is given a sequence of $n = 15$ complete triplets $[t_1, \ldots, t_{15}]$ and is asked to make a discrimination for the final triplet $t_{15} = (x_{\text{ref}}, x_1, x_2)$ along semantic dimension $d \in \{\textit{size}, \textit{kind}\}$.

TASK VECTOR

Following Hendel et al. (2023), we extract *task vectors* for the KIND and SIZE conditions by first constructing two prompts organized along each condition: x_{train}, containing 14 complete triplet examples and one final incomplete example (with `"+"`), and x_{test}, containing a single zero-shot incomplete triplet.

For each layer $\ell \in \{0, \ldots, L\}$, we extract the hidden activation in the residual stream at the final token position of x_{train} (i.e., the `"+"`) and patch it into the corresponding position in x_{test}. The language model f then autoregressively generates a sequence $x_0, \ldots, x_k$ until a complete output is produced.

We repeat this procedure over 200 randomly generated $(x_{\text{train}}, x_{\text{test}})$ pairs, selecting the layer ℓ_d^* that yields the highest accuracy. Finally, using this optimal layer ℓ_d^*, we repeat the procedure across 2400 additional prompt pairs to generate task vector embeddings for both the SIZE and KIND conditions.

DIFFMEAN

DIFFMEAN constructs a steering vector by computing the average difference between latent representations of *positive* and *negative* examples along a target output dimension. Specifically, the mean latent representation of the positive examples is subtracted from that of the negative examples, producing a steering vector. This vector is then *added* (as opposed to task vectors, where it is patched) to the latent representation of a held-out prompt x_{test} in order to steer the model's output generation.

For a target steering dimension $d \in \{\textit{size}, \textit{kind}\}$ and its contrast d', we generate 15 triplet examples organized along d, and 15 along d', where the final triplet in each set is incomplete and ends with the `"+"` token.

Then, for a given layer $\ell \in \{0, \ldots, L\}$, we extract the residual stream representation r_T^ℓ at the final token position T from both $x_{\text{train},d}$ and $x_{\text{train},d'}$. The DIFFMEAN steering vector is then computed as the difference:

$$v_{\text{diff}} = r_T^\ell(x_{\text{train},d}) - r_T^\ell(x_{\text{train},d'})$$

This resulting vector v_{diff} defines a direction in the residual stream corresponding to the contrast between the dimensions $d \in \{\textit{size}, \textit{kind}\}$.

Similar to the task vector condition, We repeat this procedure over 200 randomly generated $(x_{\text{train}}, x_{\text{test}})$ pairs, selecting the layer ℓ_d^* that yields the highest accuracy. Finally, using this optimal layer ℓ_d^*, we repeat the procedure across 2400 additional prompt pairs to generate DiffMean embeddings for both the SIZE and KIND conditions.

SAEs

We use sparse autoencoders (SAEs) from GEMMASCOPE (Lieberum et al., 2024) to steer the model along interpretable directions in residual space. An SAE is a linear model that decomposes a residual stream vector $r \in \mathbb{R}^d$ as a sparse linear combination of features, where $W \in \mathbb{R}^{d \times k}$ is a learned feature dictionary and $z \in \mathbb{R}^k$ is a sparse activation vector. Each column of W defines a directional feature in residual space, and only a small number of features are active for any given input.

For each semantic dimension $d \in \{\textit{size}, \textit{kind}\}$, we construct 20 prompts and identify the feature $f \in \mathbb{R}^d$ with the highest average activation at layer $\ell = 20$. We steer the model by injecting $c \cdot f$ (with $c = 50$) into the residual stream at layer 20 (the only available layer for `gemma-2-9b-it` on GEMMASCOPE), and generate 2400 zero-shot completions from held-out prompts x_{test}. These completions are used to construct semantic embeddings for each SAE-steered dimension.

A.2 PROCRUSTES CORRELATIONS FOR ALL PROMPT AND STEERING METHODS

The comprehensive set of procrustes correlations for all steering methods, for `gemma-2-9b-it` and `gemma-2-27b-it`.

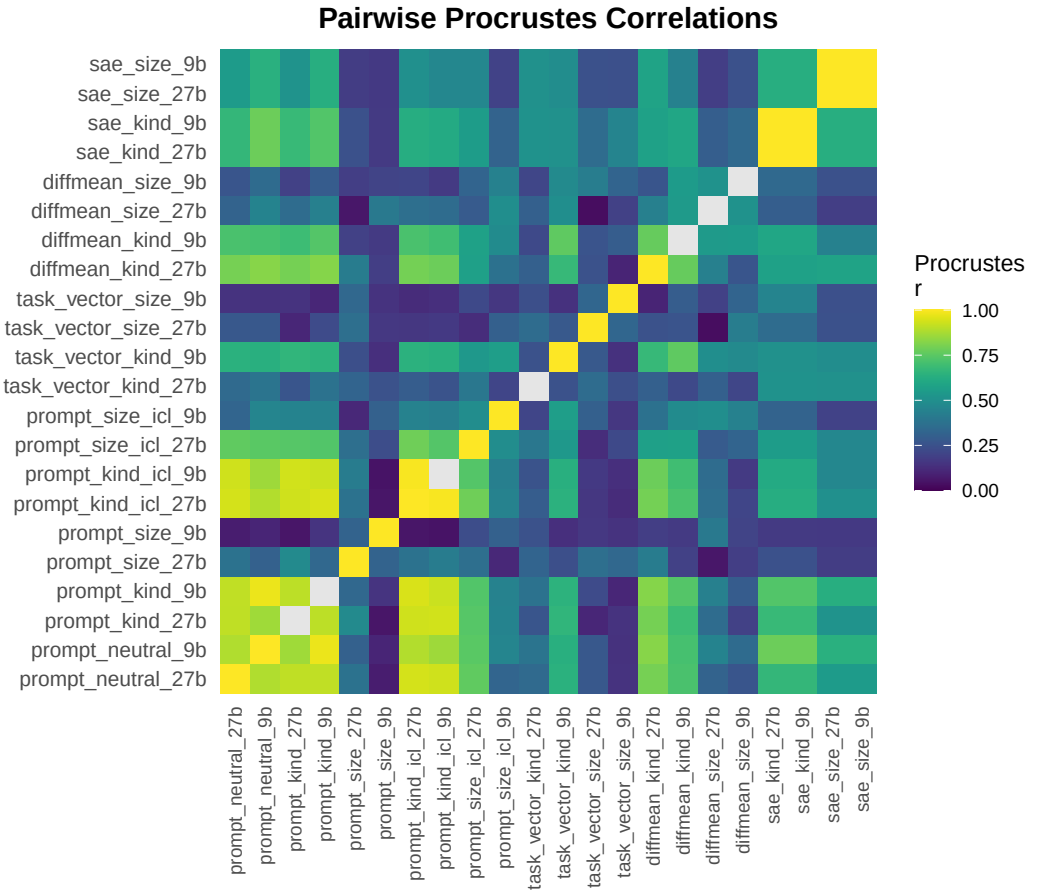
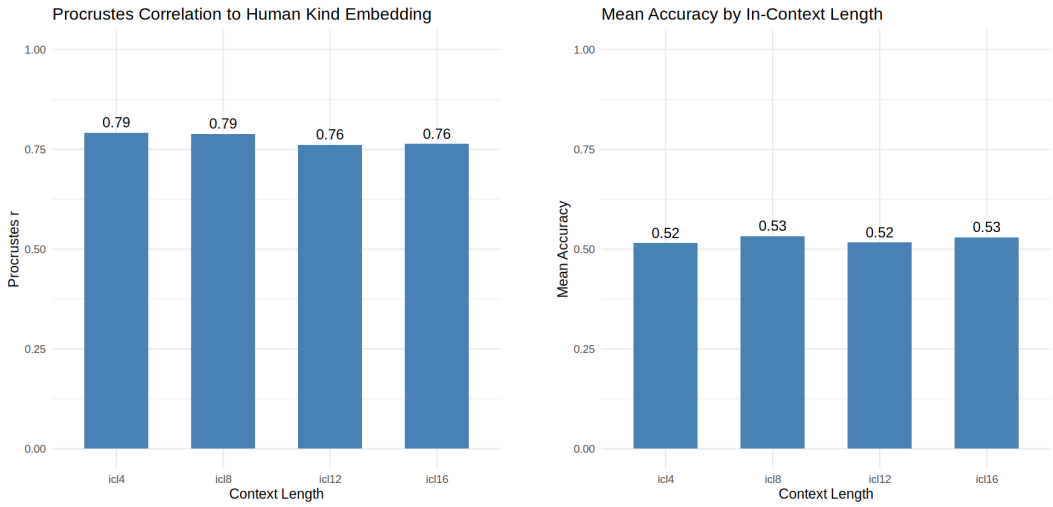


Figure 4: Full procrustes correlations for all methods

A.3 ICL PROMPT ANALYSIS

We systematically vary the number of example triplet pairs included in the KIND condition for `gemma-2-9b-it` to examine how changes to the input prompt affect model accuracy and human alignment. We find that there is no significant impact of the number of ICL examples on accuracy or alignment.



(a) Procrustes correlations with varied ICL examples

(b) Mean accuracy with varied ICL examples

Figure 5: Impact of varying the number of ICL examples on model performance

A.4 EMBEDDINGS DIMENSIONS ANALYSIS

Cumulative variance explained by the first k dimensions of the embeddings for `gemma-2-27b-it`, size condition. The first two dimensions of all embeddings were used for comparisons of representations.

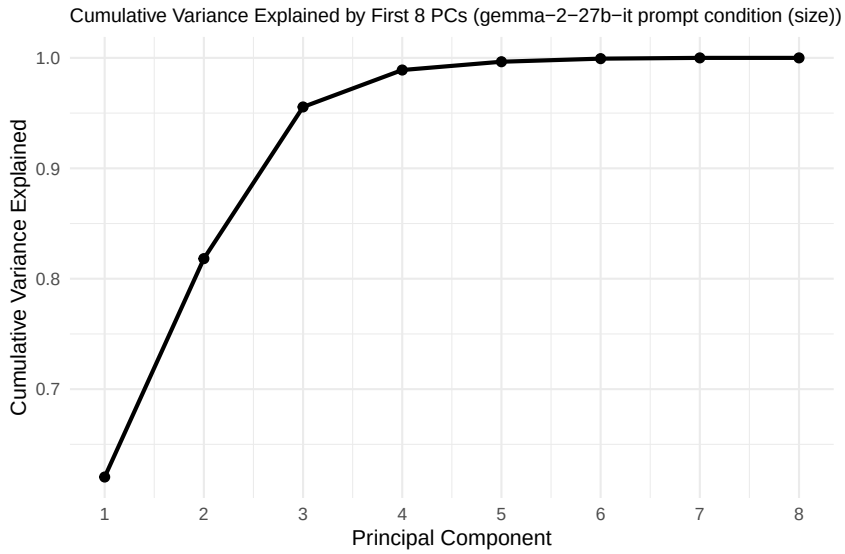
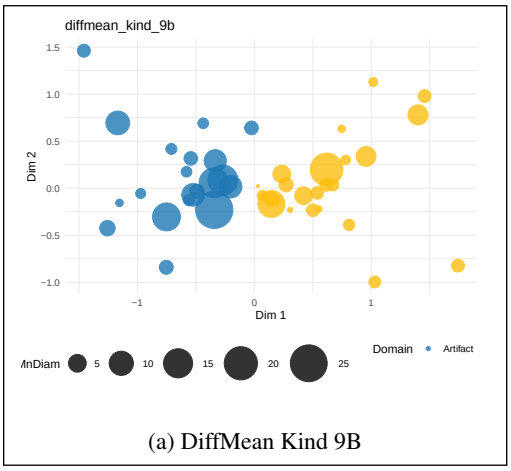
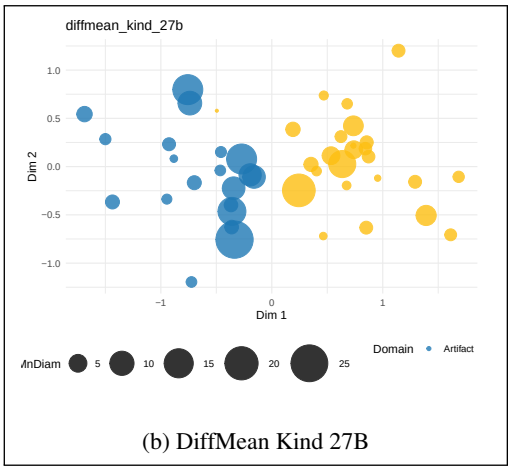


Figure 6: Cumulative variance explained by embedding dimensions

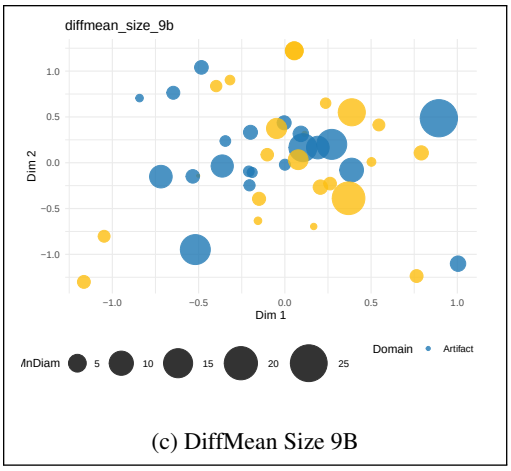
A.5 FULL EMBEDDING PLOTS



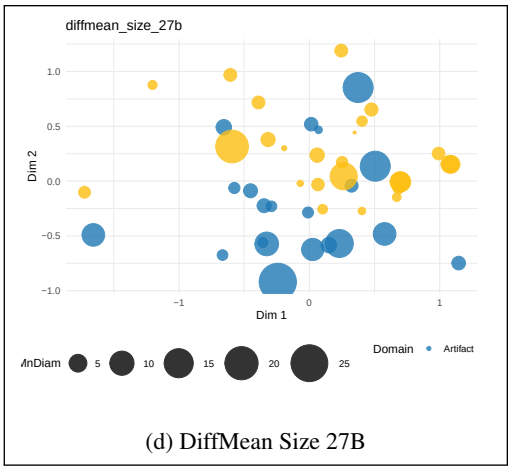
(a) DiffMean Kind 9B



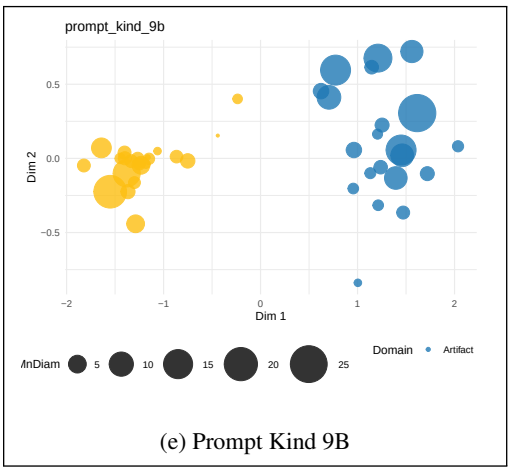
(b) DiffMean Kind 27B



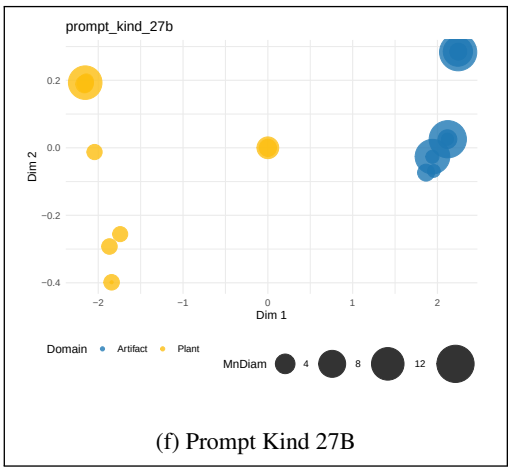
(c) DiffMean Size 9B



(d) DiffMean Size 27B

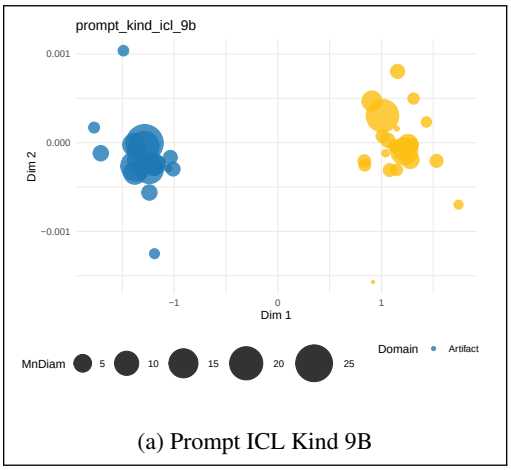


(e) Prompt Kind 9B

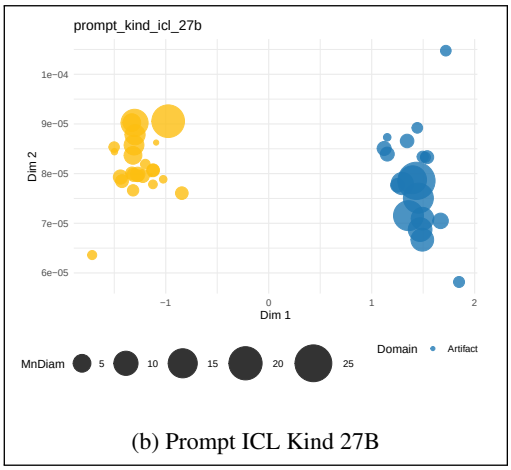


(f) Prompt Kind 27B

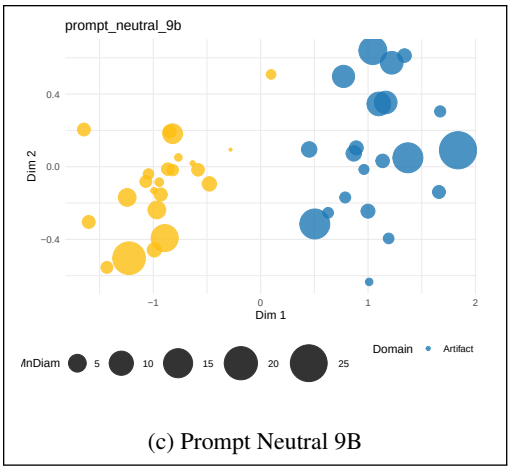
Figure 7: Embedding Plots: DiffMean and Prompt Methods



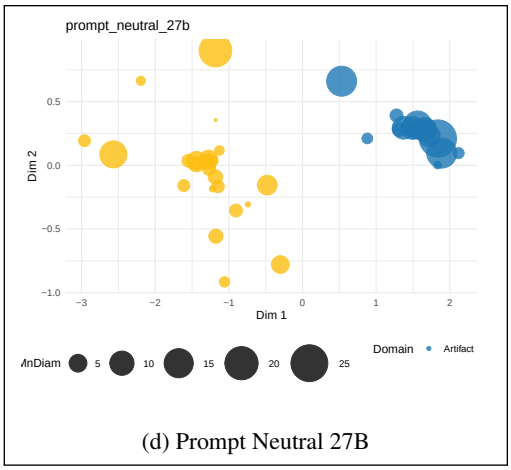
(a) Prompt ICL Kind 9B



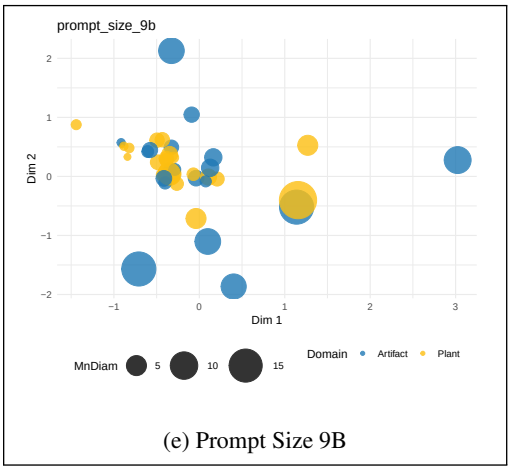
(b) Prompt ICL Kind 27B



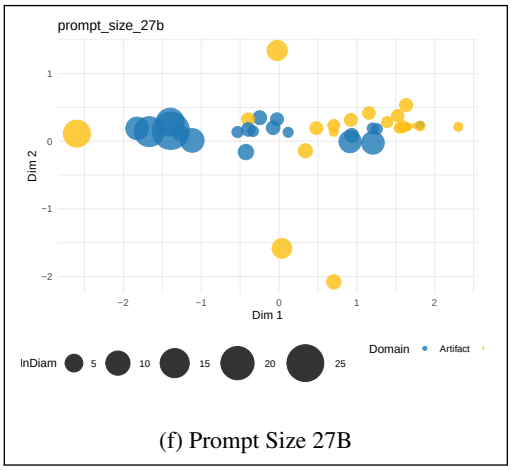
(c) Prompt Neutral 9B



(d) Prompt Neutral 27B

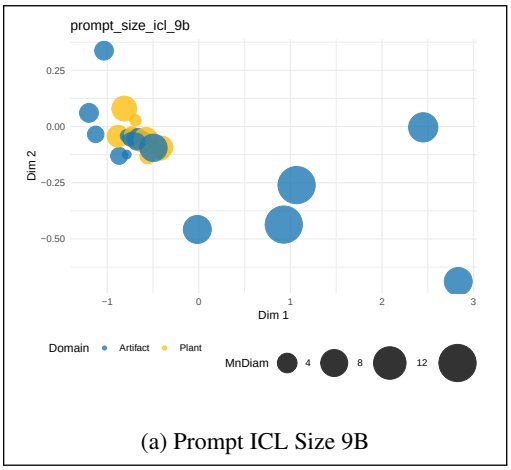


(e) Prompt Size 9B

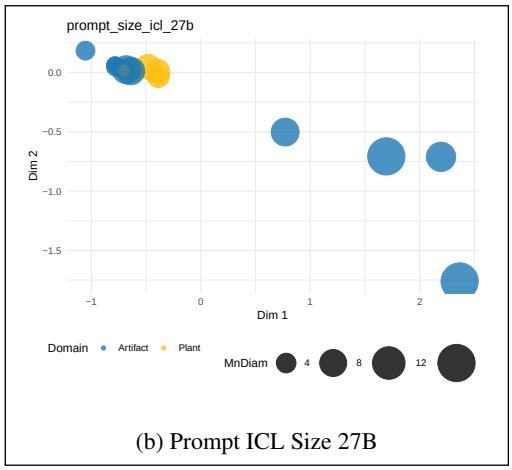


(f) Prompt Size 27B

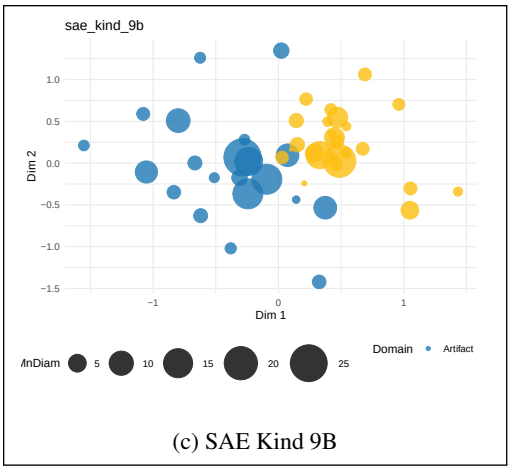
Figure 8: Embedding Plots: Prompt Method Variations



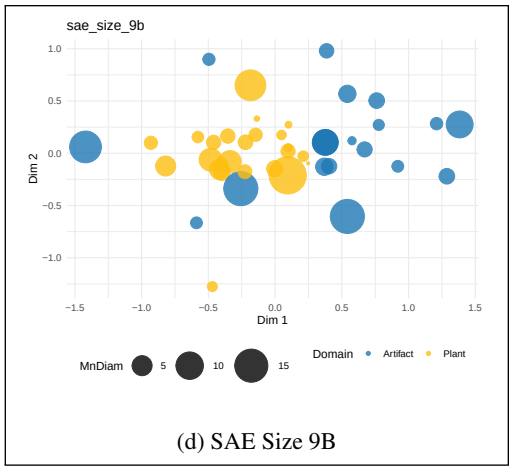
(a) Prompt ICL Size 9B



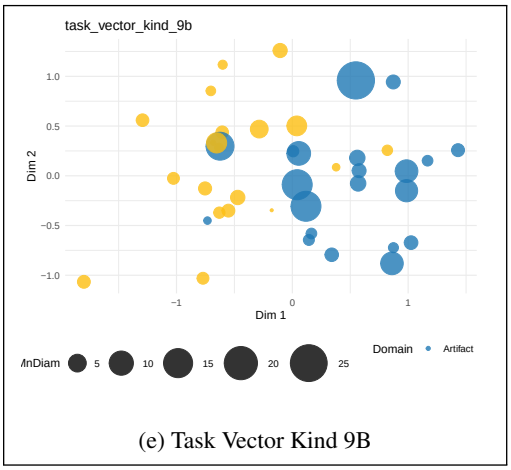
(b) Prompt ICL Size 27B



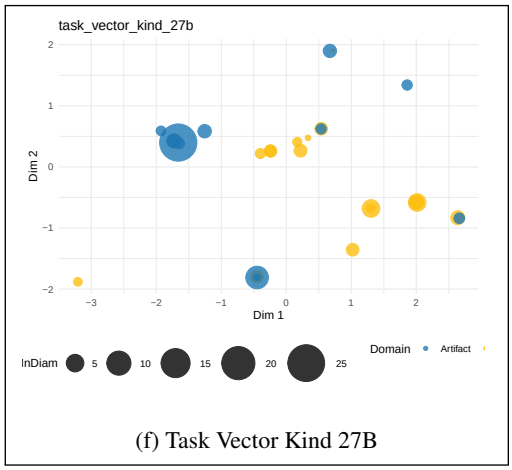
(c) SAE Kind 9B



(d) SAE Size 9B

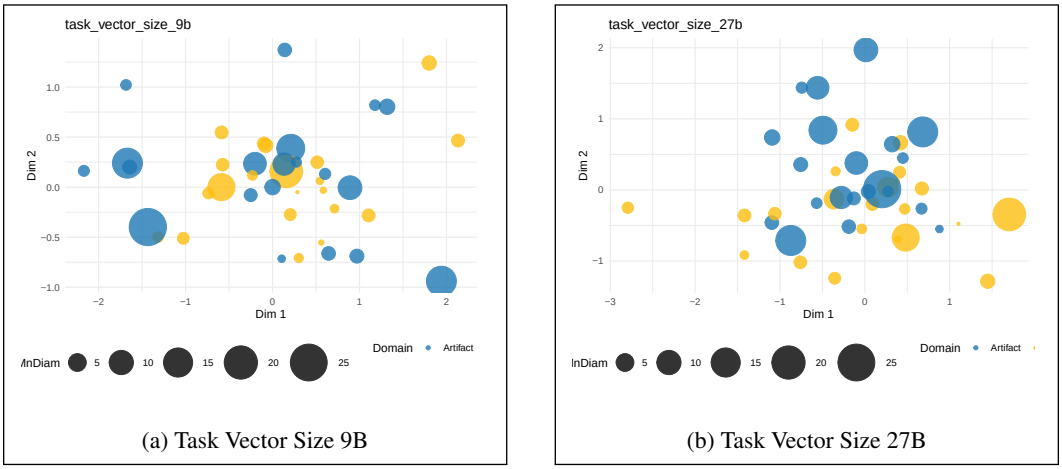


(e) Task Vector Kind 9B



(f) Task Vector Kind 27B

Figure 9: Embedding Plots: Prompt ICL, SAE, and Task Vector Methods



(a) Task Vector Size 9B

(b) Task Vector Size 27B

Figure 10: Embedding Plots: Task Vector Size Condition