Uncovering the Computational Ingredients of Human-Like Representations in Large Language Models

Zach Studdiford[1] **Timothy T. Rogers**[1] **Kushin Mukherjee**[*2] **Siddharth Suresh**[*1]
[1]University of Wisconsin–Madison [2]Stanford University
{zstuddiford, ttrogers, siddharth.suresh}@wisc.edu
kushinm@cs.stanford.edu

Abstract

The human ability to translate diverse perceptual and linguistic inputs into structured behavior has been thought to rest on learning robust representations of concepts. The rapid advancement of transformer-based large language models (LLMs) has surfaced a diversity of computational ingredients relevant for model building — architectures, fine-tuning methods, and training datasets among others — yet it remains unclear which are most crucial for developing human-like conceptual representations. Further, most current benchmarks are ill-suited to measuring *representational alignment*, making LLMs' scores on them unreliable for assessing whether they are progressing as cognitive models. We address these limitations by evaluating over 75 models on a triplet similarity task, a method well established in cognitive science for measuring conceptual representations, using concepts from the THINGS database. We find that instruction fine-tuning and larger attention head dimensionality are among the strongest predictors of human alignment, while activation function choice, multimodal pretraining, and parameter size have limited influence on alignment. Correlations between alignment scores and existing benchmark scores reveal that while some benchmarks (e.g., BigBenchHard) better capture representational alignment than others (e.g., MUSR), none fully accounts for the variance in human-model alignment, demonstrating their insufficiency. Taken together, our findings highlight key computational ingredients for advancing LLMs as models of human conceptual representation and address a key gap in LLM evaluation.

1 Introduction

The success of deep neural network models in domains ranging from perception (Yamins et al., 2014; Yamins & DiCarlo, 2016), to robotic manipulation (Finn et al., 2016), to language understanding (Tuckute et al., 2024) has partly been attributed to their ability to learn powerful *representations* that support translating diverse inputs into appropriate outputs. The continued development of transformer-based large language models (LLMs) and their capabilities on naturalistic human tasks might imply that, through training on sufficiently large corpora and with the correct architectural ingredients, these models come to possess representations largely isomorphic to human mental representations. This implication is critical for cognitive science, which has long sought to characterize human conceptual representations and the operations performed over them that give rise to naturalistic behavior. Indeed, there is a rich tradition of using deep neural networks as *computational cognitive models* — treating their learned representations as proxies for human representations, and mapping their evolution to the development of human conceptual knowledge (Jackson et al., 2022; Warstadt et al., 2025; Rogers & McClelland, 2004; Cichy &

[*]Equal contribution.

Kaiser, 2019; Saxe et al., 2019; Vong et al., 2024). Yet, while evidence across domains has revealed concordances between LLM representations and both human behavior and neural activity, it remains unclear which design decisions relating to building LLMs best predict strong human-model alignment — that is, which *computational ingredients* (architecture, scale, instruction-tuning, activation functions, etc.) give rise to human-aligned conceptual representations across a large and diverse set of concepts.

Identifying these computational ingredients is critical for several reasons. While frontier models have achieved remarkable performance on diverse benchmarks spanning scientific reasoning (Rein et al., 2024; Suzgun et al., 2022), software engineering (Chen et al., 2021; Jimenez et al., 2023), and even graphics and humor understanding (Masry et al., 2022; Methani et al., 2020; Mukherjee et al., 2025; Zhou et al., 2025; Hessel et al., 2022) – domains thought to require 'human-like' thinking – we lack thorough accounts explaining what properties of models account for these successes. Recent work has shown that optimizing for benchmark accuracy alone is insufficient for building models that fail in human-like ways and that are generally aligned with humans (Fel et al., 2022; Ying et al., 2025). We argue that prioritizing representational alignment (Sucholutsky et al., 2023; Sucholutsky & Griffiths, 2023b; Muttenthaler et al., 2024) offers a promising path forward. Measuring the similarity between model and human internal representations can provide a more unified account of model capabilities, guide the development of more robustly aligned systems (Collins et al., 2024; Muttenthaler et al., 2024; Peterson et al., 2018), and accelerate cognitive science by yielding more faithful computational models of the human mind.

2 Related work

Quantifying Human and Model Representations using Large-Scale Concept Datasets Mapping the geometry of human conceptual knowledge has long been a central goal of cognitive science (Rogers & McClelland, 2004; Shepard, 1980; Tversky, 1977; Rumelhart et al., 1986). Early attempts relied on explicit feature ratings and pairwise similarity judgments, which scaled poorly and thus limited their utility for capturing the vast inventory of human concepts (Rosch, 1975; Nosofsky, 1984; Shepard, 1980). Structured probabilistic models, while useful for explaining aspects of concept learning, have similarly been constrained by their reliance on specific relational structures, limiting scope and generalizability (Zhao et al., 2024; Wong et al., 2022; Kemp & Tenenbaum, 2008). Recent years have seen a resurgence of similarity-based approaches, driven by scalable algorithms that convert human judgments into expressive high-dimensional *semantic embeddings* capable of capturing the latent dimensions organizing human semantic knowledge (King et al., 2019; Mukherjee & Rogers, 2025; Hebart et al., 2019; Suresh et al., 2023b; Peterson et al., 2018; Kriegeskorte & Mur, 2012; Cichy et al., 2019; Hebart et al., 2023; Muttenthaler et al., 2022b; Jamieson et al., 2015; Sievert et al., 2023). Among these, triadic similarity judgment tasks — where participants select the odd item from a triplet — are a particularly powerful representational learning paradigm (Tamuz et al., 2011; Wah et al., 2014; Kleindessner & von Luxburg, 2014), with derived embeddings proving reliable predictors of both neural and behavioral patterns across a variety of semantic tasks (Mukherjee & Rogers, 2025; Mukherjee et al., 2026; Hebart et al., 2023; Colon & Rogers, 2023; Suresh et al., 2023b; Marjieh et al., 2022; Mukherjee et al., 2022; Mur et al., 2013; Suresh et al., 2024).

A complementary advance that has recently galvanized research in representational alignment is the development of concept datasets that aim to capture the diversity of object concepts humans reason about (Mehrer et al., 2021; Hebart et al., 2019; Giallanza et al., 2024; Suresh et al., 2025). Datasets like THINGS (Hebart et al., 2019) and ECOSET (Mehrer et al., 2021) consist of concept sets and associated images along with rich psychological and neural metadata (Hebart et al., 2023) that allow for comparisons between human and model representations along multiple levels of analysis — behavioral, neural, and computational.

Measurement of Representational Alignment Several complementary measures for assessing representational alignment have been developed over recent years (Kriegeskorte et al., 2008; Sucholutsky et al., 2023; Barbosa et al., 2025; Oswal et al., 2016). Core measurement techniques include Representational Similarity Analysis (RSA) using correlation between similarity matrices derived from behavioral, neural, or neural network activation data (Kriegeskorte et al., 2008; Nili et al., 2014), Procrustes analysis for finding

optimal linear alignments (Schönemann, 1966), and Centered Kernel Alignment (CKA) for comparing representations invariant to linear transformations (Kornblith et al., 2019; Williams et al., 2021). Recent work has further developed more sensitive metrics including shape-based metrics (Williams et al., 2021), topological measures (Barannikov et al., 2021), information-theoretic approaches (Bansal et al., 2021), and regularized regression-based methods (Oswal et al., 2016) each of which iteratively address issues relating to model regularization, generalization, and error-bounds. Measuring alignment is useful because models with higher human alignment show improved few-shot learning (Sucholutsky & Griffiths, 2023a; Huh et al., 2024), better out-of-distribution generalization (Moschella et al., 2023; Norelli et al., 2023), enhanced adversarial robustness (Engstrom et al., 2019), and more human-like error patterns (Geirhos et al., 2021; Fel et al., 2022). These findings suggest that human-aligned models capture meaningful invariances that support robust intelligence.

Computational Ingredients Driving Alignment Understanding which design choices yield human-aligned representations in foundation models can help uncover general principles undergirding human intelligence. In vision, work has shown that optimizing models for categorization (Yamins et al., 2014) or contrastive objectives (Konkle & Alvarez, 2022; Zhuang et al., 2021) produces representations closely aligned with human visual cortex, offering insight into the objective functions the brain may solve to construct visual representations. Extending this line of inquiry to over 75 models, Muttenthaler et al. (2022a) found that training data and objectives strongly predicted human-model alignment, whereas architecture and scale had minimal impact — a finding that challenges scaling law accounts (Kaplan et al., 2020; Cherti et al., 2023) predicting that larger models should broadly perform better on most tasks. Other key results regarding the importance of different computational ingredients include the finding that self-supervised contrastive methods align more closely with human vision than non-contrastive approaches (Chen et al., 2020; Zbontar et al., 2021; Caron et al., 2020); multimodal image–text contrastive training improves alignment over vision-only training (Radford et al., 2021; Jia et al., 2021; Pham et al., 2022); training on diverse, naturalistic datasets such as `JFT-3B` improves alignment beyond ImageNet (Zhai et al., 2022; Dehghani et al., 2023); enforcing shape bias on models enhances alignment while texture bias impairs it (Geirhos et al., 2019; Hermann et al., 2020); and intermediate layers often show stronger perceptual alignment than final layers (Berardino et al., 2017; Kumar et al., 2022). Similar approaches, dissecting the predictive power of different model building choices, have not been brought to bear in investigations into conceptual structure in LLMs-leaving unclear what the relative importance of different computational ingredients are when building foundation models of human semantic knowledge.

While early studies showed that contextualized embeddings from models like BERT and GPT-2 could predict human similarity judgments (Bommasani et al., 2020; Grand et al., 2022), more recent work demonstrates that LLMs can achieve remarkably high alignment with human conceptual structure (Marjieh et al., 2024a;b; Mukherjee et al., 2024; Suresh et al., 2025; 2023a; Mukherjee et al., 2023b), with instruction-tuned models showing particular advantages (Tong et al., 2024; Gurnee & Tegmark, 2024). Several studies have used triplet tasks specifically to probe LLM representations (Nam et al., 2025; Studdiford et al., 2025; Rathi et al., 2024; Suresh et al., 2023b), finding that larger models and those trained on more diverse data show better alignment. However, systematic analyses disentangling the effects of scale, architecture, training data, and objectives remain limited. Work on multimodal models suggests that vision-language training improves conceptual alignment beyond either modality alone (Yuksekgonul et al., 2023; Thrush et al., 2022; Qin et al., 2025), though the precise mechanisms remain unclear.

3 Methods

Summary of key contributions. We leverage the open-weight model ecosystem to construct a suite of 75+ models spanning diverse computational ingredients. Using concepts and semantic embeddings from THINGS dataset (Hebart et al., 2019) in conjunction with a triadic similarity judgment paradigm (Jamieson et al., 2015; Hebart et al., 2023; Sievert et al., 2023), we obtain both human and model-specific conceptual embeddings using analogous methods to ensure model-fair comparisons (Firestone, 2020). We then compare these embeddings to human data and apply statistical models to identify the ingredients most predictive of

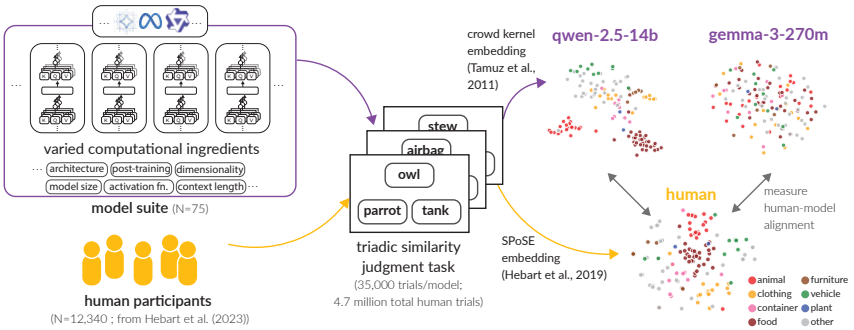


Figure 1: *Method for estimating human-model alignment.* We collected 35k triplet similarity judgments from each of the 77 models in our suite. We computed semantic embeddings based on these judgments and compared the representational geometry of model embeddings to human embeddings derived from Hebart et al. (2023).

human-model alignment. Lastly, we relate alignment to performance on standard LLM benchmarks, highlighting points of divergence between representational alignment and task capabilities.

Concept Dataset To investigate how various LLM characteristics influence human alignment, we carefully curated a set of test concepts that would provide robust and interpretable results. Starting from the full set of 1,854 THINGS object concepts (Hebart et al., 2019; 2023), we subsampled 128 concrete, real-world object concepts (e.g., "lion", "banjo", "car") that spanned four main cognitive axes of variation: familiarity, artificiality, animacy, and size based on prior work (Mukherjee et al., 2023a) [1]. This sampling strategy ensured that we had adequate semantic diversity in our concept dataset (representative of the full THINGS database) while also ensuring that we could generate semantic embeddings for a large suite of models in a tractable manner. See Appendix A.4 for the complete list of concepts with detailed descriptions of concept selection criteria and validation procedures.

Model Suite We evaluate a range of open-source models varying across several key computational ingredients: model architecture, primary activation functions used, pre-training modality, whether models were instruction fine-tuned or not, scale (both in terms of model parameters and training tokens), context length, dimensionality of attention heads/MLPs/residual streams. While in principle it is possible to offer even more granular analyses of computational ingredients, these factors constitute major architectural and training decisions when building foundation models, which we believe provide adequate resolution to investigate the drivers of human-model alignment. We evaluated a set of 77 open-weight transformer models including models from popular families (Llama, Qwen, Gemma, etc.). The full set of models can be seen in Table 4. Notably, we included models that varied in size from 100M parameters to 42B parameters, ranged in the degree of post-training (from base to instruction fine-tuned), and architecture-style (decoder-only, encoder-decoder, and state space models (SSMs) (Gu et al., 2022).

Triadic similarity judgment task The triadic similarity judgment task can be framed in two equivalent ways: (1) **anchored similarity judgments** where participants are presented with a triplet of items (images or words depicting different concepts) and select which of two option items is most similar to an anchored reference item, or **odd-one-out judgments** where they identify which of the three presented item is the odd-one-out. Both framings have been successfully employed to study conceptual representations in humans and models (Hebart et al., 2019; 2023; Mukherjee & Rogers, 2025; Colon & Rogers, 2023; Suresh et al., 2023b; Studdiford et al., 2025; Mukherjee et al., 2026; Muttenthaler et al., 2022b). In this work, we leverage existing human similarity ratings data (Hebart et al., 2023) and corresponding

[1] We empirically validated the consistency of this subsample with the full 1,854 concept set, finding negligible differences in the human-alignment of large and small models between embeddings computed from full and subsamples, see Appendix A.14.

embeddings, which were collected using the odd-one-out paradigm and embedded using the sparse positive similarity embedding (SPoSE) method, respectively. For model evaluations, we adopted the anchored similarity judgment paradigmn for computational efficiency, deriving embeddings using ordinal embedding techniques (Jamieson et al., 2015; Tamuz et al., 2011) following recent work (Suresh et al., 2023b; Mukherjee et al., 2026; Colon & Rogers, 2023). Crucially, the ordinal embedding approach provides theoretical bounds on sample complexity (Jamieson et al., 2015), allowing us to determine the number of triplet comparisons needed for faithful representations a priori. While these approaches differ in framing and embedding computation, they yield comparable representational structures, as we demonstrate empirically in Appendix A.12. This methodological choice allows us to maximize the use of existing high-quality, validated human data while efficiently collecting model responses with known sampling requirements in a scalable manner, ensuring robust human-model comparisons across a large suite of models.

Human Semantic Embeddings Human semantic embeddings were obtained by Hebart et al. (2023) from the behavioral judgments of $N = 12,340$ online participants in a triplet odd-one-out task. On each trial, participants were shown three items from the THINGS database and were prompted to select the most dissimilar object (*"Which is the odd one out?"*). In total, 4.7 million behavioral triplet judgments were obtained for the full set of 1,854 concepts. To estimate an embedding space from the set of individual triplet judgments, Hebart et al. (2023) used the SPoSE algorithm (Hebart et al., 2023), which learns low-dimensional vectors for each object such that (1) pairs of objects judged as similar are placed closer in the embedding space, and (2) dimensions are constrained to be sparse in order to yield human-interpretable properties. The resulting space captured meaningful representational structure from the individual judgments of participants, revealing core dimensions along which human semantic representations are organized.

Model Semantic Embeddings While much work on representational similarity extracts activation vectors from models and computes their similarity to human behavior-based representations via methods like RSA (Kriegeskorte et al., 2008), we instead estimated model semantic embeddings from similarity judgments using a triadic comparison task akin to those used to estimate human embeddings. We did so for two principled reasons. First, although computations in transformer layers and attention heads have been linked to human language processing (Tuckute et al., 2024), no general framework exists for identifying which parts of a model carry signatures of human semantic knowledge across model families, making fair comparison of activations to human representations difficult. Second, the ability of modern LLMs to perform structured tasks via natural language prompting — combined with the relative simplicity of the triadic judgment task — means that administering the same test to both humans and models constitutes a species-fair (Firestone, 2020) comparison, ensuring that alignment discrepancies reflect genuine representational differences rather than artifacts of differing embedding methods.

For each model in our suite, we collected 35k triadic judgment responses. Using the human semantic embeddings, we estimated that a rank 30 space explains over 95% of the variance in embeddings (see A.8). Based on known bounds (Sievert et al., 2023), we needed ($n \cdot d \cdot log_2(n)$[2]) at least 26,900 judgments on random sets of triplets to estimate a robust embedding. Each triplet trial was randomly sampled from the $\binom{128}{3}$ possible triplet trials. We evaluated each model using its default huggingface sampling settings using a common system-level prompt — `"You are a helpful assistant who gives responses to questions"`. We used the following prompt for each trial for each model — `QUESTION: Which item is most similar to item_x: item_y or item_z? Respond only with the item name.` If we were evaluating a base model (not instruction-tuned), we appended `Answer:` after the initial prompt. Instruction template formats were applied where appropriate and each prompt was evaluated independently.

Using these similarity judgments, we fit an ordinal embedding algorithm (Tamuz et al., 2011; Sievert et al., 2023) that organized the semantic embedding space such that items

[2] n is the number of items and d is the expected dimensionality of the space

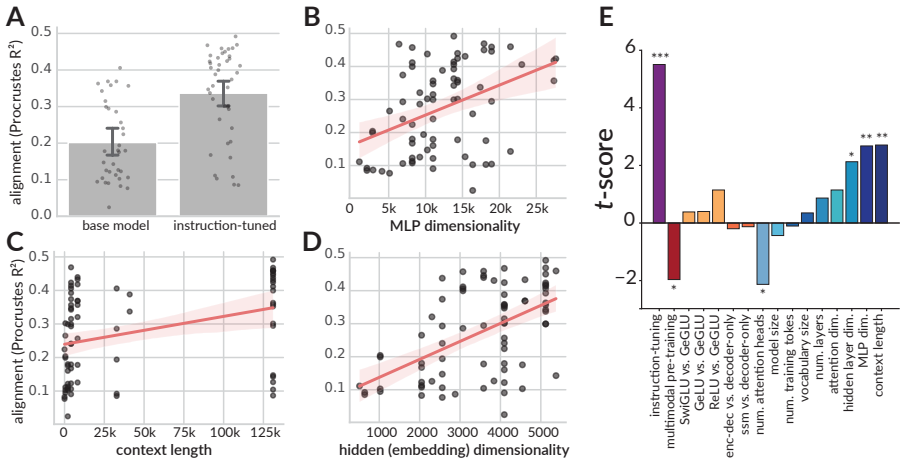


Figure 2: **Computational ingredients predictive of human model alignment**. **A.**, **B.**, **C.**, and **D.** show the relationship between instruction-tuning, MLP dimensionality, context length, and embedding dimensionality on Procrustes R^2. These were the four ingredients most predictive in the mixed-effects regression model. **E.** t-scores for each predictor, which highlight the relative contribution of each ingredient towards alignment when considered in a single statistical model.

that were judged to be similar more often were pulled closer together in this space. We fit 30D embeddings based on the dimensionality of human semantic embeddings. We also held out 20% of the data during the embedding fitting process and evaluated the quality of the embeddings by monitoring the crowd-kernel loss (Tamuz et al., 2011) to ensure the embeddings were of high quality.

4 Results and Discussion

We first report how well different computational ingredients predicted human-model alignment, initially considering each ingredient separately via individual correlations or t-tests (Kim, 2015), then assessing relative contributions of each via mixed-effect regression models (Lindstrom & Bates, 1988). Finally, we evaluate the relationship between human-model alignment and model performance on other key LLM benchmarks (Sprague et al., 2023; Rein et al., 2024; Srivastava et al., 2023; Zhou et al., 2024; Hendrycks et al., 2020). Our primary alignment measure was the variance in human semantic embeddings explained by model embeddings (R^2) after Procrustes alignment (Gower, 1975). Procrustes transforms find the optimal linear transformation (rotation, scaling, translation) between two vector spaces of equal dimensionality to minimize squared distances between corresponding points. Using human embedding variance (human SS) as the baseline, we compute $R^2 = 1 - \frac{\text{residual SSE}}{\text{human SSE}}$. Other alignment measures such as CKA (Kornblith et al., 2019) and RSA(Kriegeskorte et al., 2008) were highly correlated with Procrustes R^2 (Appendix A.13). We also replicated the full suite of analyses reported below on a set of 213 abstract emotion words, details of which can be seen in Appendix Section A.13.

4.1 Contributions of Model Computational Ingredients to Alignment

Instruction tuning leads to greater alignment. Models that were instruction fine-tuned were significantly better aligned than those that were not ($t = 5.11$, $p < 0.001$; see Figure 2A). Qualitatively, instruction fine-tuning clustered representations of semantic categories, such that within-category items were closer in representation space and between-category items were more distal (see Appendix 15). Thus, the same post-training techniques that enable models to be more adept at following prompt instructions also bring their representations closer into alignment with humans.

Model architecture dimensionality corresponds to greater alignment. We found a strong positive relationship between the dimensionality of individual model architecture components — MLPs and attention heads — and the degree of model-human alignment. Specifically, the number of per-layer attention heads ($r = 0.52, p < 0.001$), attention matrix dimensionality ($r = 0.30, p = 0.026$), embeddings dimensionality ($r = 0.54, p < 0.001$), and MLP dimensionality ($r = 0.40, p < 0.001$) each correlated positively with human representational alignment. Figure 2B and D show the individual correlations between MLP and embedding dimensionality and model-human Procrustes R^2 (the attention dimensionality was not significant in multiple regression models reported in subsequent sections). Thus, the greater expressivity granted by larger latent activation spaces and more attention heads predicts greater human alignment.

Multi-modal pretraining has no effect on alignment. While much of human concept learning is inherently multimodal (Rogers & McClelland, 2004; Patterson et al., 2007), there is limited work showing whether vision-language pretraining (the most prominent class of multimodal training) leads to more aligned semantic representations (cf. Qin et al. (2025)). Here, multimodal (image) pre-training had no independent effect on human-model alignment relative to text-only models ($t = 0.34, p > 0.05$; Appendix Figure A.15).

Larger model sizes show greater human alignment. For visual representations, Muttenthaler et al. (2022a) found that larger models need not yield more human-like representations. Neural scaling laws (Kaplan et al., 2020), however, might predict that larger and deeper models should be more performant, and perhaps more aligned. In accordance with this prediction, human-model alignment scaled approximately linearly with both model parameter count and the number of model layers ($r_{param} = 0.45, r_{layers} = 0.41, p < 0.001$; Appendix Figure A.15). Thus, larger scale predicts greater human-model alignment.

Alignment increased with amount of training data. Models exposed to more tokens in pretraining generally exhibited more human-like representations ($r = 0.33$, $p < 0.01$; Appendix Figure A.15). Although model training data also likely varied in quality and composition–thus limiting strong conclusions about which properties matter most–our results indicate that the *amount* of training data can predict alignment.

Alignment scales with context length. The triadic similarity task does not require a long context length, but we nevertheless found that models with a greater maximum context-length also demonstrated greater human alignment ($r = 0.35, p < 0.01$; Figure 2). This finding could be linked to post-training regimes that unlock long-context reasoning, which consequently lead models to perform well on short-context tasks as well.

Choice of activation function does not affect alignment. Activation functions have seen iterative development over the years with recent variants (e.g., SwiGLU (Shazeer, 2020)) being particularly well-suited to managing gradient flows during training and finetuning experiments leading to faster and stabler training regimes. But do choices in activation functions ground out in alignment? We found no advantage for any particular function in increasing human alignment ($F(3, 64) = 1.36, p = .26$; Appendix Figure A.15).

Larger vocabulary size does not increase alignment. One might expect that larger model vocabularies unlock greater expressivity for models, leading them to be more human-aligned. Against this prediction, we found no relationship between model vocabulary size and human-alignment ($r = 0.10$; Appendix Figure A.15).

Which computational ingredients have the greatest relative contribution towards human-model alignment? Thus far, we have investigated how different computational ingredients impact human-model alignment in isolation. In reality, many of these ingredients vary across models at the same time, leaving open the questions of (1) which ingredients are most important *relative to others* and (2) which ingredients account for unique variance after others are considered.

To answer these questions, we fit a mixed-effects multiple regression model (Lindstrom & Bates, 1990) predicting Procrustes R^2 from *all* computational ingredients. Mixed-effects models combine the interpretability of OLS with random effects to capture item-level

variance (Barr et al., 2013)[3]. We included random intercepts for model family (e.g., Llama, Qwen) to capture variance related to these families not grounded out in our set of ingredients, and fixed effects for all ingredients. The fitted model explained substantial variance in alignment ($R^2 = 0.78$; Fig. 2E). Instruction fine-tuning emerged as the strongest predictor ($\beta = 0.13$, $SE = 0.024$, $p < 0.001$). Dimensionality measures — MLP ($\beta = 0.049$, $SE = 0.018$, $p < 0.01$), embedding/hidden layers ($\beta = 0.048$, $SE = 0.022$, $p < 0.05$) and context length ($\beta = 0.042$, $SE = 0.015$, $p < 0.01$) — also accounted for significant unique variance. In contrast to the independent analyses, attention dimensionality was not significant, and more attention heads predicted worse alignment after accounting for other factors ($\beta = -0.045$, $p < 0.05$). Likewise, multimodal pretraining, which showed no independent effect on alignment, predicted reliably lower alignment after accounting for other factors ($\beta = -0.102$,$SE = 0.052$, $p < 0.05$). Thus, current multimodal training regimes may actually limit rather than improve human-model alignment. No other ingredient explained significant unique variance when others were included in the model (see effect sizes in Figure 2E).

Predictors of human alignment are robust across different concept sets. A potential concern is the possibility that our results are sensitive to experimenter decisions regarding the specific concepts tested. We addressed the possibility that these results are specific to the THINGS items evaluated, repeating the analyses above on a set of abstract EMOTIONS concepts (Shaver et al., 1987) for our full model suite. We observe similar findings across both stimulus sets: individual correlations between alignment and MLP dimensionality ($r = 0.42$, $p < 0.001$), context length ($r = 0.31$, $p < 0.01$), and hidden layer dimensionality ($r = 0.56$, $p < 0.001$) are replicated in our EMOTIONS stimuli, while the relative contributions of instruction tuning (objects: $\beta = 0.135$, $t = 5.73$, $p < 0.001$; emotions: $\beta = 0.235$, $t = 5.09$, $p < 0.001$) and number of attention heads (objects: $\beta = -0.048$, $t = -2.27$, $p < 0.05$; emotions: $\beta = -0.178$, $t = -2.76$, $p < 0.01$) are reflected in both of the mixed-effects models. Full results for the EMOTIONS stimuli are included in Appendix A.13. We also find that human-model alignment is consistent when computed over both the subsample and complete set of THINGS concepts (Appendix A.14) and that randomly sampling different sets of LLMs produces results consistent with our main findings (Appendix A.17). In sum, our findings show a fair degree of consistency across different kinds of concepts.

Which concept dimensions drive human alignment? While our analyses capture human-model alignment across the subset of THINGS concepts, it remains unclear whether certain kinds of concepts are more aligned than others. That is, do certain concept attributes predict more or less item-level human alignment? Leveraging attribute metadata included for each concept in THINGS, we first grouped each of the 128 concepts based on superordinate category labels (e.g *furniture, clothing, food*) and computed the Procrustes R^2 within each category group. We found that alignment was significantly higher for inanimate objects (furniture, container) relative to animate objects (animals, food) (Appendix Figure 5), and that these category-level differences were relatively consistent across models (Kendall's $\tau = 0.45$, $ps < .05$). Next, we asked whether certain item-level attributes (e.g familiarity, concreteness) predicted item-level differences in alignment (measured as the cosine distance between the model embedding and the ground-truth human embedding). We fit a mixed-effects regression predicting these item-level divergences from nine different attributes, and found that items judged as more concrete or typical of their category tended to be more aligned, while more polysemous concepts (words with multiple meanings) were less aligned (see Appendix Figure 6).

Effect of different post-training regimens on human alignment. A small number of open-source model families expose checkpoints across various stages of pre-training, allowing for a direct comparison of different post-training methods on human-alignment. Figure 3 shows changes in human alignment when SFT, DPO, and RLVR are applied in OLMo, Instella, and Llama models. While we caveat that the limited number of available checkpoints prevents a more systematic analysis, we qualitatively observe that SFT and DPO post-training stages led to improvements in alignment whereas the effect of RLVR was more muted.

[3]We additionally check the collinearity of model predictors and find low to intermediate levels of collinearity across all predictors, see Appendix A.16

4.2 Relationship Between Alignment and Established Benchmarks

It is possible that representational alignment is already tracked by existing benchmarks. If so, it would provide a lens into what aspects of behavior most closely track representational alignment (e.g., grammaticality judgments). If not, then standard benchmarks overlook representational alignment as an axis of model competence. To understand the relationship between alignment and performance on standard LLM benchmarks we estimated the correlation between alignment, measured using three metrics (Procrustes R^2, CKA, and RSA correlation), and model scores on five key benchmarks – MMLU (Hendrycks et al., 2020), IFEval (Zhou et al., 2023), BigBenchHard (BBH)(Suzgun et al., 2022), GPQA (Rein et al., 2023) and MUSR (Sprague et al., 2024). In addition to the open-source models included in our evaluation, we compute the Procrustes R^2 for GPT-5.4 in order to obtain an approximate empirical upper-bound on human-model alignment in models with high benchmark scores. We also included the Procrustes R^2 obtained from FastText embeddings of concepts to measure how well aligned concepts are between people and modern, static word embedding models. Despite a reported BigBenchHard score of 90.9, we observe that the human-alignment of GPT-5.4 is exceeded by open-source LLMs with significantly lower benchmark scores. Conversely, we observe notably high alignment in FastText embeddings despite this model being comparatively much smaller than any of the LLMs tested. Figure 4A shows the relationship between alignment (Procrustes R^2) and scores on BBH, the benchmark most strongly correlated with alignment. While models that performed well on BBH generally had higher alignment, we found some notable divergences especially for base models that scored lower on alignment than one might expect based on their BBH score. Figure 4B shows the correlations between alignment and all the established benchmarks. While all three alignment metrics were in agreement, we found that the overall correlations varied ($r_{min} = 0.36, p < 0.05; r_{max} = 0.74, p < 0.001$). We focus on Procrustes R^2 moving forward.

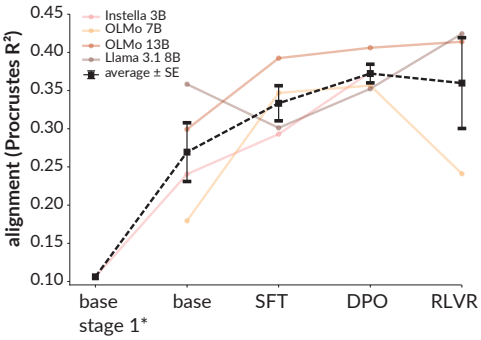


Figure 3: **Effects of post-training on alignment**. The black points show mean alignment (across models) at each post-training stage.*Instella only.

Performance on benchmarks that probe probe domain-general knowledge (BBH; $r = 0.74, p < 0.001$; MMLU; $r = 0.71, p < 0.001$) and instruction-following ability (IFEval; $r = 0.68, p < 0.001$) were more correlated with alignment than benchmarks testing more specific domain knowledge and reasoning ability (Math; $r = 0.61, p < 0.001$, MUSR; $r = 0.36, p < 0.05$). This finding is sensible given that alignment with human semantic judgments requires representing semantic *knowledge* broadly in human-like ways (captured by BBH and MMLU) and *deploying* that knowledge in context-sensitive ways according to instructions (captured by IFEval). We also found that the correlation between BBH and IFEval was significantly weaker than the correlation between these evaluations and our measure of representational alignment($r_{BBH,IFEval} = 0.44, p < 0.01$), suggesting our alignment measure captures separate competencies unique to each benchmark.

5 Conclusion

The rapid advancement of language models in scale and diversity, alongside their varied, often unpredictable, benchmark performance, leaves open a fundamental question: what does it take to build human-like models? A gold standard for "human-like" requires not only human-like task performance but also internal representations and computations that mirror that of humans. Here, we attempt to identify the computational ingredients most predictive of human-model representational alignment, leveraging large-scale human triadic similarity judgments as our method and semantic embeddings of the THINGS concepts as our target.

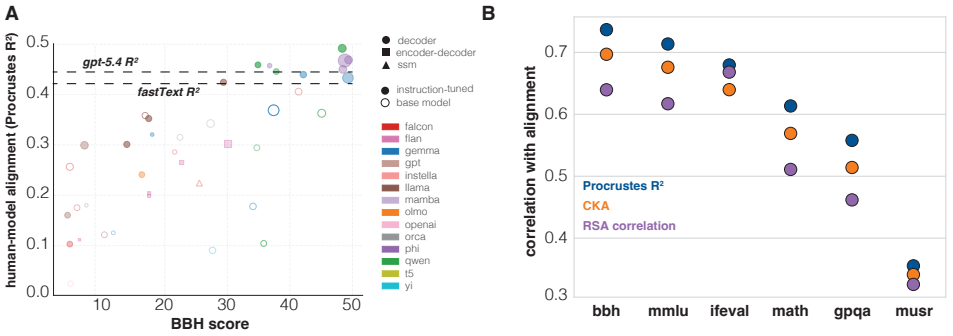


Figure 4: **Relationship between alignment and other LLM benchmarks**. **A.** `BigBenchHard` scores for each model in our suite vs. their Procrustes R^2 value w.r.t. human embeddings. `gpt-5.4` and `fastText` R^2 values indicated by dotted lines. **B.** Correlation between alignment and the six LLM benchmarks evaluated.

We carefully constructed a suite of 77 open-weight language models that varied in several key computational ingredients including model scale, training dataset size, dimensionality of model embeddings, and post-training regimes, among others. By comparing human-model semantic embedding alignment as a function of these model characteristics, we arrived at several insights.

First, we found two convergent pieces of evidence that modern post-training techniques are central to human-model alignment: (1) instruction-finetuning was the strongest predictor of alignment even after controlling for variance explained by other ingredients (Figure 2) and (2) alignment tended to increase across different stages of post-training (Figure 3). The importance of context length in predicting alignment further suggests that the methods that enable long context-reasoning might bring models into closer alignment with humans. Second, we observed that architectural dimensionality (embedding layers, MLP) were relatively more important than overall model size and training corpus size in predicting alignment, especially in the mixed-effects model suggesting that representational capacity is key to developing models that align with human representations. Third, we found that while some existing benchmarks were correlated with our measure of alignment (e.g., *MMLU, BBH*), no single benchmark captured all the variance in representational alignment. Notably, benchmarks emphasizing broad semantic knowledge and its flexible deployment showed the strongest correlations, consistent with theories of how humans represent semantic knowledge (Rogers & McClelland, 2004; Saxe et al., 2019).

Limitations To our knowledge, this is the first attempt to isolate the computational ingredients predictive of human–LLM conceptual alignment, though several limitations remain. First, while we evaluated many core ingredients, training dataset composition (e.g., text–code balance, language coverage) could not be assessed due to limited documentation across models; future work should clarify which dataset properties enable semantic alignment. Second, certain model classes — SSMs, multimodal models, and mixture-of-experts — were sparsely represented in our suite, limiting the reliability of conclusions about those architectures. Third, triadic similarity judgments, while well-validated and scalable, capture only one facet of semantic knowledge; future work should examine how models *deploy* such knowledge in open-ended reasoning tasks (Wong et al., 2025; Mahowald et al., 2024). Fourth, here we only present correlational evidence. To establish true causal validity of our claims, future work should seek to leverage large-scale, open, community-drive model-training efforts (such as Marin) actively train models while varying core computational ingredients.

Acknowledgments. We thank members of the Knowledge and Concepts Lab and Lupyan Lab for helpful feedback. We also thank the Center for High Throughput Computing for providing the compute resources for this project.

Code and materials: https://github.com/Knowledge-and-Concepts-Lab/llm-alignment-benchmark.

References

01.AI Team. Yi: Open foundation models. arXiv:2403.04652, March 2024.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. *arXiv preprint arXiv:2201.00058*, 2021.

Joao Barbosa, Amin Nejatbakhsh, Lyndon Duong, Sarah E Harvey, Scott L Brincat, Markus Siegel, Earl K Miller, and Alex H Williams. Quantifying differences in neural population activity with shape metrics. *bioRxiv*, pp. 2025–01, 2025.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3): 255–278, 2013.

Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

S. Black, S. Biderman, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of Machine Learning and Systems*, April 2022. [Online]. Available: https://aclanthology.org/2022.mlsys-1.72.

Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting pretrained contextualized representations via reductions to static embeddings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4758–4781, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

H. W. Chung, L. Hou, et al. Scaling instruction-finetuned language models. arXiv:2210.11416, December 2022. [Online]. Available: https://arxiv.org/abs/2210.11416.

Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

Radoslaw M Cichy, Nikolaus Kriegeskorte, Kamila M Jozwik, Jasper JF van den Bosch, and Ian Charest. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24, 2019.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature human behaviour*, 8(10):1851–1863, 2024.

Y Ivette Colon and Timothy T Rogers. Yours and ours: Individual and group differences in semantic organization from triplet judgements of faces. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023.

EleutherAI. Gpt-j-6b: 6 billion parameter open source gpt model. EleutherAI Blog, June 2021. [Online]. Available: https://6b.eleuther.ai.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Falcon-LLM Team. The falcon 3 family of open models. Published Dec 2024. [Online]. Available: https://huggingface.co/blog/falcon3, December 2024.

Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35:9432–9446, 2022.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2019.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34: 23885–23899, 2021.

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

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 6(7):975–987, 2022.

D. Groeneveld, I. Beltagy, et al. Olmo: Accelerating the science of language models. arXiv:2402.00838, June 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

S. Gunasekar, Y. Zhang, J. Aneja, et al. Textbooks are all you need. arXiv:2306.11644, October 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2024.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Rebecca L Jackson, Matthew A Lambon Ralph, and Timothy T Rogers. Late maturation of executive control promotes conceptual development. *bioRxiv*, pp. 2022–03, 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 and Robert D. Nowak. Active ranking using pairwise comparisons, 2011. URL https://arxiv.org/abs/1109.3701.

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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.

Marcie L King, Iris IA Groen, Adam Steel, Dwight J Kravitz, and Chris I Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019.

Matthäus Kleindessner and Ulrike von Luxburg. Uniqueness of ordinal embedding. *Conference on Learning Theory*, pp. 40–67, 2014.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

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.

Nikolaus Kriegeskorte and Marieke Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245, 2012.

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

Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin D Cubuk. Do better imagenet classifiers assess perceptual similarity better? *Transactions on Machine Learning Research*, 2022.

Mary J Lindstrom and Douglas M Bates. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pp. 673–687, 1990.

J. Liu, J. Wu, et al. Introducing instella: New state-of-the-art fully open 3b language models. AMD ROCm Blog, March 2025. [Online]. Available: https://rocm.blogs.amd.com/artificial-intelligence/introducing-instella-3B/.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models, 2024. URL https://arxiv.org/abs/2301.06627.

Raja Marjieh, Ilia Sucholutsky, Theodore R Sumers, Nori Jacoby, and Thomas L Griffiths. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*, 2022.

Raja Marjieh, Ilia Sucholutsky, Theodore R Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. Do large language models predict human similarity judgments? *Cognitive Science*, 48(8):e13509, 2024a.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2402.10910*, 2024b.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Meta AI. Llama 3.1 8b instruct (2024 release). Internal technical report, Meta, 2024.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 1527–1536, 2020.

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.

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.

Kushin Mukherjee, Karen Schloss, Laurent Lessard, Michael Gleicher, and Timothy Rogers. Color-concept associations reveal an abstract conceptual space. *Journal of Vision*, 22(14): 4408–4408, 2022.

Kushin Mukherjee, Holly Huey, Xuanchen Lu, Yael Vinker, Rio Aguina-Kang, Ariel Shamir, and Judith Fan. Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, 36:67138–67155, 2023a.

Kushin Mukherjee, Siddharth Suresh, and Timothy T. Rogers. Human-machine cooperation for semantic feature listing, 2023b. URL https://arxiv.org/abs/2304.05012. ICLR 2023 Tiny Papers.

Kushin Mukherjee, Timothy T. Rogers, and Karen B. Schloss. Large language models estimate fine-grained human color-concept associations, 2024. URL https://arxiv.org/abs/2406.17781.

Kushin Mukherjee, Donghao Ren, Dominik Moritz, and Yannick Assogba. Encqa: Benchmarking vision-language models on visual encodings for charts. *arXiv preprint arXiv:2508.04650*, 2025.

Kushin Mukherjee, Holly Huey, Laura M Stoinski, Martin N Hebart, Judith E Fan, and Wilma A Bainbridge. Drawings of things: A large-scale drawing dataset of 1854 object concepts. *Behavior Research Methods*, 58(2):57, 2026.

S. Mukherjee, A. Mitra, G. Jawahar, et al. Orca: Progressive learning from complex explanation traces of gpt-4. arXiv:2306.02707, June 2023c.

Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in psychology*, 4:128, 2013.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*, 2022a.

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, 2022b.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert MÃžller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.

Hyunji Nam, Lucia Langlois, James Malamut, Mei Tan, and Dorottya Demszky. Idealign: Comparing large language models to human experts in open-ended interpretive annotations. *arXiv preprint arXiv:2509.02855*, 2025.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: coupled data turns unimodal models to multimodal without training. *arXiv preprint arXiv:2210.01738*, 2023.

Robert M Nosofsky. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, 10(1):104, 1984.

Urvashi Oswal, Christopher Cox, Matthew Lambon-Ralph, Timothy Rogers, and Robert Nowak. Representational similarity learning with application to brain networks. In *International Conference on Machine Learning*, pp. 1041–1049. PMLR, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2022.

Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it. *arXiv preprint arXiv:2507.13328*, 2025.

Qwen Team. Qwen technical report (tongyi qianwen). Alibaba Cloud, August 2023. [Online]. Available: https://github.com/QwenLM.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International conference on machine learning*, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33499–33530, 2023.

C. Raffel, N. Shazeer, A. Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 2020.

Vedant Rathi, Abhishek Kumar, et al. Can language models learn to skip steps? *arXiv preprint arXiv:2411.01855*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*, 2024.

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

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

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.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52 (6):1061, 1987.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468): 390–398, 1980.

Yann-Yann Shieh and Rachel T Fouladi. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and psychological measurement*, 63(6):951–985, 2003.

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

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14670–14728, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh

Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

Zach Studdiford, Timothy T Rogers, Siddharth Suresh, and Kushin Mukherjee. Evaluating steering techniques using human similarity judgments. *arXiv preprint arXiv:2505.19333*,

2025.

Ilia Sucholutsky and Thomas L. Griffiths. Alignment with human representations supports robust few-shot learning, 2023a. URL https://arxiv.org/abs/2301.11990.

Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36:73464–73479, 2023b.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin J Achterberg, Christiane Fellbaum, Marek Grzes, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Siddharth Suresh, Kushin Mukherjee, and Timothy T. Rogers. Semantic feature verification in FLAN-T5, 2023a. URL https://arxiv.org/abs/2304.05591. ICLR 2023 Tiny Papers.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 722–738, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.47. URL https://aclanthology.org/2023.emnlp-main.47/.

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.

Siddharth Suresh, Kushin Mukherjee, Tyler Giallanza, Xizheng Yu, Mia Patil, Jonathan D. Cohen, and Timothy T. Rogers. Ai-enhanced semantic feature norms for 786 concepts, 2025. URL https://arxiv.org/abs/2505.10718. Also presented at the ICLR 2025 Workshop on Bidirectional Human-AI Alignment.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. In *International Conference on Machine Learning*, pp. 673–680, 2011.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Erik Ferrara, Jie Wu, et al. Mass-producing failures of multimodal systems with language models. *arXiv preprint arXiv:2306.12105*, 2024.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(2024):277–301, 2024.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Similarity comparisons for interactive fine-grained categorization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 859–866, 2014.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, et al. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2504.08165*, 2025.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Ankur Pasupat, Stella Wang, Quoc Le, and Denny Zhou. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott W Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34: 4738–4750, 2021.

Catherine Wong, William P McCarthy, Gabriel Grand, Yoni Friedman, Joshua B Tenenbaum, Jacob Andreas, Robert D Hawkins, and Judith E Fan. Identifying concept libraries from language about object structure. *arXiv preprint arXiv:2205.05666*, 2022.

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.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *International Conference on Learning Representations*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning*, pp. 12310–12320, 2021.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Bonan Zhao, Christopher G Lucas, and Neil R Bramley. A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 8(1):125–136, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. *arXiv preprint arXiv:2502.20356*, 2025.

Yixin Zhou, Le Zhang, Chengyu Liu, Yinsong Yao, Yansong Feng, Yunzhe Wang, Linyi Zheng, Chenghua Zhu, Jiayi Li, Binyuan Li, Libin Wang, Ruyi Chen, Jia Guo, and Lifeng Li. IFEval: A new rigorous evaluation for instruction-following models. *arXiv preprint arXiv:2403.09886*, 2024.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A Appendix

A.1 Odd-One-Out Judgments → SPoSE Estimation (Human Semantic Embeddings)

To establish a ground-truth human semantic space, we employ the SPoSE (Sparse Positive Similarity Embedding) framework applied to odd-one-out behavioral data from the THINGS dataset (Hebart et al., 2023).

SPoSE Algorithm. Let $\mathcal{X} = \{1, \ldots, n\}$ index n items and let $X \in \mathbb{R}^{n \times d}_{\geq 0}$ be a nonnegative, low-dimensional embedding with rows $x_i^\top \in \mathbb{R}^d_{\geq 0}$. We define pairwise similarities using the inner product $s_{ij} = x_i^\top x_j$, yielding similarity matrix $S = XX^\top$. In an odd-one-out trial with unordered triplet $\{i, j, k\}$, participants select the *most similar pair* among $\{(i, j), (i, k), (j, k)\}$. SPoSE models this choice behavior with a multinomial logit over pairwise similarities:

$$\Pr\big[(i,j) \text{ chosen} \mid \{i,j,k\}, X\big] = \frac{\exp(s_{ij})}{\exp(s_{ij}) + \exp(s_{ik}) + \exp(s_{jk})}. \tag{1}$$

Estimation and Implementation. The SPoSE estimate $\hat{X}$ is obtained via a sparsity-promoting MAP program that encourages both nonnegativity and interpretability:

$$\hat{X} \in \arg\max_{X \geq 0} \Big\{ \underbrace{\sum_{t=1}^{T} \log p\big(y_t \mid \{i_t, j_t, k_t\}, X\big)}_{\text{log-likelihood}} - \lambda \|X\|_1 \Big\}, \tag{2}$$

where y_t is the chosen pair on trial t, and $\lambda > 0$ controls sparsity.[4] We use the publicly released SPoSE embeddings as our human reference space.

Dimensionality Selection. To determine the appropriate embedding dimension, we apply PCA to the released SPoSE matrix and retain the smallest d that explains at least 95% of the variance (Appendix A.8). For our corpus of $n = 128$ concepts, this yields $d = \hat{d}$ dimensions (*30*), which we use as the target dimensionality for model embeddings. Additionally, we validate this dimensionality empirically by running models in the range $\hat{d} = 20...30$ and observe relative stability in embeddings loss for all $\hat{d}$) (see Appendix A.11).

A.2 Anchored Similarity Judgments → Ordinal Embedding (Model Semantic Embeddings)

To obtain comparable model-based semantic embeddings, we apply an ordinal embedding algorithm (Tamuz et al., 2011; Sievert et al., 2023) to model similarity judgments, creating semantic spaces where frequently co-judged similar items are positioned closer together.

Embedding Algorithm. Each triplet $\{i, j, k\}$ from model judgments yields an ordinal constraint: "i is closer to j than to k." Let $\{x_i^\star\}_{i=1}^n \subset \mathbb{R}^d$ denote the unknown true point locations we seek to estimate, with corresponding (squared) Euclidean distance matrix $D^\star$. We model the probability of observing each ordinal constraint using a standard noisy triplet model with monotone link function:

$$\Pr\left[y_{(i;j,k)} = 1\right] \;=\; f\big(D^\star_{ik} - D^\star_{ij}\big), \qquad f(0) = \tfrac{1}{2}. \tag{3}$$

Estimation and Implementation. The ordinal embedding algorithm minimizes crowd-kernel triplet loss (Tamuz et al., 2011) by optimizing Euclidean distances between item pairs. We minimize an empirical surrogate of the negative log-likelihood using a low-rank Gram matrix parameterization. To ensure reliable estimates, we reserve 20% of our triplet data for validation purposes and monitor crowd-kernel loss throughout the fitting procedure.

[4] The nonnegativity constraint and elementwise ℓ_1 penalty are equivalent to an exponential prior on factors and empirically yield interpretable semantic dimensions.

Sample Complexity Considerations. When the true embedding has rank d and triplet judgments are sampled approximately uniformly at random, finite-sample theory provides the out-of-sample prediction error bound:

$$\mathcal{E}_{\text{pred}} \;=\; \tilde{O}\left(\sqrt{\frac{d\,n\log n}{|S|}}\right), \qquad \Rightarrow \quad |S| \;=\; \tilde{\Theta}\big(d\,n\log n\big). \tag{4}$$

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). Guided by this theory and setting the model embedding dimension to match the human-derived $\hat{d}$, we collect

$$N_{\text{triplets}} \;=\; c\,d\,n\log n \tag{5}$$

triplet judgments per model, where c is chosen to balance statistical efficiency with computational constraints.

This parallel approach provides a principled framework for human-model semantic comparison. Both methods yield d-dimensional embeddings where geometric relationships reflect semantic similarities, with the SPoSE framework extracting interpretable human semantic structure and ordinal embedding converting model judgments into geometrically comparable representations with theoretically grounded sample complexity $\tilde{\Theta}(d\,n\log n)$.

A.3 Effect of post-training techniques on model alignment

Table 1: Model Alignment Across Training Stages

Model	Training Stage	Alignment (R^2)
Instella 3B	Base Stage 1	0.106
Instella 3B	Base	0.241
Instella 3B	SFT	0.293
Instella 3B	DPO	0.374
Llama 3.1 8B	Base	0.358
Llama 3.1 8B	SFT	0.301
Llama 3.1 8B	DPO	0.352
Llama 3.1 8B	RLVR	0.424
OLMo 13B	Base	0.299
OLMo 13B	SFT	0.392
OLMo 13B	DPO	0.406
OLMo 13B	RLVR	0.414
OLMo 7B	Base	0.179
OLMo 7B	SFT	0.347
OLMo 7B	DPO	0.357
OLMo 7B	RLVR	0.241
Average	Base Stage 1	0.106
Average	Base	0.269
Average	SFT	0.333
Average	DPO	0.372
Average	RLVR	0.360

A.4 Object Concepts

The following table presents the 128 concrete object concepts from the THINGS dataset used in our experiments, along with their semantic categories and definitions.

Table 2: Complete list of object concepts with definitions.

Concept	Category	Definition
badger	animal	sturdy carnivorous burrowing mammal with strong claws; widely distributed in the northern hemisphere
bear	animal	massive plantigrade carnivorous or omnivorous mammals with long shaggy coats and strong claws
butterfly	animal	diurnal insect typically having a slender body with knobbed antennae and broad colorful wings
chipmunk	animal	a burrowing ground squirrel of western America and Asia; has cheek pouches and a light and dark stripe running down the body
cow	animal	female of domestic cattle: "'moo-cow' is a child's term"
crow	animal	black birds having a raucous call
fly	animal	two-winged insects characterized by active flight
giraffe	animal	tallest living quadruped; having a spotted coat and small horns and very long neck and legs; of savannahs of tropical Africa
grasshopper	animal	terrestrial plant-eating insect with hind legs adapted for leaping
hyena	animal	doglike nocturnal mammal of Africa and southern Asia that feeds chiefly on carrion
iguana	animal	large herbivorous tropical American arboreal lizards with a spiny crest along the back; used as human food in Central America and South America
lion	animal	large gregarious predatory feline of Africa and India having a tawny coat with a shaggy mane in the male
mosquito	animal	two-winged insect whose female has a long proboscis to pierce the skin and suck the blood of humans and animals
mouse	animal	any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails
owl	animal	nocturnal bird of prey with hawk-like beak and claws and large head with front-facing eyes
parrot	animal	usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds
puffin	animal	any of two genera of northern seabirds having short necks and brightly colored compressed bills
snail	animal	freshwater or marine or terrestrial gastropod mollusk usually having an external enclosing spiral shell
snake	animal	limbless scaly elongate reptile; some are venomous
tarantula	animal	large hairy tropical spider with fangs that can inflict painful but not highly venomous bites
bathrobe	clothing	a loose-fitting robe of towelling; worn after a bath or swim
beanie	clothing	a small skullcap; formerly worn by schoolboys and college freshmen
bow	clothing	a knot with two loops and loose ends; used to tie shoelaces
bracelet	clothing	jewelry worn around the wrist for decoration
button	clothing	a round fastener sewn to shirts and coats etc to fit through buttonholes
chaps	clothing	(usually in the plural) leather leggings without a seat; joined by a belt; often have flared outer flaps; worn over trousers by cowboys to protect their legs
hat	clothing	headdress that protects the head from bad weather; has shaped crown and usually a brim

continued on next page

Table 2 continued from previous page

Concept	Category	Definition
jeans	clothing	(usually plural) close-fitting trousers of heavy denim for manual work or casual wear
kilt	clothing	a knee-length pleated tartan skirt worn by men as part of the traditional dress in the Highlands of northern Scotland
kimono	clothing	a loose robe; imitated from robes originally worn by Japanese
kneepad	clothing	protective garment consisting of a pad worn by football or baseball or hockey players
tuxedo	clothing	semiformal evening dress for men
uniform	clothing	clothing of distinctive design worn by members of a particular group as a means of identification
veil	clothing	a garment that covers the head and face
visor	clothing	a brim that projects to the front to shade the eyes
bag	container	a flexible container with a single opening
cooker	container	a utensil for cooking
doily	container	a small round piece of linen placed under a dish or bowl
fishbowl	container	a transparent bowl in which small fish are kept
honeypot	container	South African shrub whose flowers when open are cup-shaped resembling artichokes
thermos	container	vacuum flask that preserves temperature of hot or cold drinks
vase	container	an open jar of glass or porcelain used as an ornament or to hold flowers
appetizer	food	food or drink to stimulate the appetite (usually served before a meal or as the first course)
applesauce	food	puree of stewed apples usually sweetened and spiced
baklava	food	rich Middle Eastern cake made of thin layers of flaky pastry filled with nuts and honey
beer	food	a general name for alcoholic beverages made by fermenting a cereal (or mixture of cereals) flavored with hops
bok choy	food	Asiatic plant grown for its cluster of edible white stalks with dark green leaves
bread	food	food made from dough of flour or meal and usually raised with yeast or baking powder and then baked
crepe	food	small very thin pancake
cupcake	food	small cake baked in a muffin tin
dessert	food	a dish served as the last course of a meal
dough	food	a flour mixture stiff enough to knead or roll
enchilada	food	tortilla with meat filling baked in tomato sauce seasoned with chili
grapefruit	food	large yellow fruit with somewhat acid juicy pulp; usual serving consists of a half
gravy	food	a sauce made by adding stock, flour, or other ingredients to the juice and fat that drips from cooking meats
hummus	food	a thick spread made from mashed chickpeas, tahini, lemon juice and garlic; used especially as a dip for pita; originated in the Middle East
leek	food	plant having a large slender white bulb and flat overlapping dark green leaves; used in cooking; believed derived from the wild Allium ampeloprasum
mango	food	large oval tropical fruit having smooth skin, juicy aromatic pulp, and a large hairy seed

continued on next page

Table 2 continued from previous page

Concept	Category	Definition
margarita	food	a cocktail made of tequila and triple sec with lime and lemon juice
mashed potato	food	potato that has been peeled and boiled and then mashed
milkshake	food	frothy drink of milk and flavoring and sometimes fruit or ice cream
oatmeal	food	porridge made of rolled oats
parfait	food	layers of ice cream and syrup and whipped cream
pumpkin	food	usually large pulpy deep-yellow round fruit of the squash family maturing in late summer or early autumn
quesadilla	food	a tortilla that is filled with cheese and heated
quiche	food	a tart filled with rich unsweetened custard; often contains other ingredients (as cheese or ham or seafood or vegetables)
ravioli	food	small circular or square cases of dough with savory fillings
sea urchin	food	shallow-water echinoderms having soft bodies enclosed in thin spiny globular shells
souffle	food	light fluffy dish of egg yolks and stiffly beaten egg whites mixed with e.g. cheese or fish or fruit
stew	food	food prepared by stewing especially meat or fish with vegetables
tortellini	food	small ring-shaped stuffed pasta
waffle	food	pancake batter baked in a waffle iron
wrap	food	a sandwich in which the filling is rolled up in a soft tortilla
bassinet	furniture	a basket (usually hooded) used as a baby's bed
bathtub	furniture	a relatively large open container that you fill with water and use to wash the body
beanbag	furniture	a small cloth bag filled with dried beans; thrown in games
bed	furniture	a piece of furniture that provides a place to sleep
bench	furniture	a long seat for more than one person
coaster	furniture	a covering (plate or mat) that protects the surface of a table (i.e., from the condensation on a cold glass or bottle)
computer screen	furniture	a screen used to display the output of a computer to the user
cot	furniture	a small bed that folds up for storage or transport
couch	furniture	an upholstered seat for more than one person
crib	furniture	baby bed with high sides made of slats
album	other	a book of blank pages with pockets or envelopes; for organizing photographs or stamp collections etc
backscratcher	other	a long-handled scratcher for scratching your back
banjo	other	a stringed instrument of the guitar family that has long neck and circular body
baseball	other	a ball used in playing baseball
baseball glove	other	the handwear used by fielders in playing baseball
bassoon	other	a double-reed instrument; the tenor of the oboe family
beachball	other	large and light ball; for play at the seaside
blower	other	a device that produces a current of air
bongo	other	a small drum; played with the hands
cannonball	other	a solid projectile that in former times was fired from a cannon
canvas	other	an oil painting on canvas fabric
chainsaw	other	portable power saw; teeth linked to form an endless chain

continued on next page

Table 2 continued from previous page

Concept	Category	Definition
cymbal	other	a percussion instrument consisting of a concave brass disk; makes a loud crashing sound when hit with a drumstick or when two are struck together
doorknob	other	a knob used to release the catch when opening a door (often called 'doorhandle' in Great Britain)
doorknocker	other	a device (usually metal and ornamental) attached by a hinge to a door
extinguisher	other	a manually operated device for extinguishing small fires
fire alarm	other	a device that makes a loud sound to warn people when there is a fire
guitar	other	a stringed instrument usually having six strings; played by strumming or plucking
hatchet	other	a small ax with a short handle used with one hand (usually to chop wood)
hula hoop	other	a large hoop spun around the body by gyrating the hips, for play or exercise.
knitting needle	other	needle consisting of a slender rod with pointed ends; usually used in pairs
knob	other	a round handle
lawnmower	other	garden tool for mowing grass on lawns
pinball	other	a game played on a sloping board; the object is to propel marbles against pins or into pockets
pocket watch	other	a watch that is carried in a small watch pocket
shuffleboard	other	a game in which players use long sticks to shove wooden disks onto the scoring area marked on a smooth surface
skeleton	other	the hard structure (bones and cartilages) that provides a frame for the body of an animal
swab	other	implement consisting of a small piece of cotton that is used to apply medication or cleanse a wound or obtain a specimen of a secretion
tennis ball	other	ball about the size of a fist used in playing tennis
toilet paper	other	a soft thin absorbent paper for use in toilets
treadmill	other	an exercise device consisting of an endless belt on which a person can walk or jog without changing place
trowel	other	a small hand tool with a handle and flat metal blade; used for scooping or spreading plaster or similar materials
volleyball	other	an inflated ball used in playing volleyball
flower	plant	a plant cultivated for its blooms or blossoms
gourd	plant	any of numerous inedible fruits with hard rinds
airbag	vehicle	a safety restraint in an automobile; the bag inflates on collision and prevents the driver or passenger from being thrown forward
carriage	vehicle	a vehicle with wheels drawn by one or more horses
exhaust pipe	vehicle	a pipe through which burned gases travel from the exhaust manifold to the muffler
hot-air balloon	vehicle	balloon for travel through the air in a basket suspended below a large bag of heated air
humvee	vehicle	a high mobility, multipurpose, military vehicle with four-wheel drive
missile	vehicle	a rocket carrying a warhead of conventional or nuclear explosives; may be ballistic or directed by remote control
odometer	vehicle	a meter that shows mileage traversed
taillight	vehicle	lamp (usually red) mounted at the rear of a motor vehicle
tank	vehicle	an enclosed armored military vehicle; has a cannon and moves on caterpillar treads

continued on next page

Table 2 continued from previous page

Concept	Category	Definition
taxi	vehicle	a car driven by a person whose job is to take passengers where they want to go in exchange for money

A.5 Model Suite

Table 3: List of all the models used

Model	Attention Type	Training Tokens	#Heads	Hidden Dim
Falcon3-3B-Base Falcon-LLM Team (2024)	GQA	100B	12 (4 KV)	3072
tiiuae/falcon-7b Falcon-LLM Team (2024)	MQA	1.5T	71	4544
tiiuae/falcon-7b-instruct Falcon-LLM Team (2024)	MQA		71	4544
Falcon3-10B-Base Falcon-LLM Team (2024)	GQA	2T	12 (4 KV)	3072
tiiuae/falcon-11b Falcon-LLM Team (2024)	GQA	5T	32 (8 KV)	4096
flan-t5-small Chung et al. (2022)	Multi-Head		8	512
flan-t5-large Chung et al. (2022)	Multi-Head		16	1024
flan-t5-xl Chung et al. (2022)	Multi-Head		32	1024
flan-t5-xxl Chung et al. (2022)	Multi-Head		128	1024
gemma-3-270m Team et al. (2024)	GQA		8 (shared KV)	2048
gemma-3-270m-it Team et al. (2024)	GQA		8 (shared KV)	2048
gemma-2-2b Team et al. (2024)	MQA		8 (shared KV)	2048
gemma-2-2b-it Team et al. (2024)	MQA		8 (shared KV)	2048
gemma-2-9b Team et al. (2024)	MHA		16	4096
gemma-2-9b-it Team et al. (2024)	MHA		16	4096
gemma-2-27b Team et al. (2024)	MHA		16	6144
gemma-2-27b-it Team et al. (2024)	MHA		16	6144
gemma-3-4b-it Team et al. (2024)	GQA		12 (4 KV)	3072
gemma-3-4b-pt Team et al. (2024)	GQA		12 (4 KV)	3072
gemma-3-12b-it Team et al. (2024)	GQA		16 (4 KV)	4096
gemma-3-12b-pt Team et al. (2024)	GQA		16 (4 KV)	4096
gemma-3-27b Team et al. (2024)	GQA		16 (4 KV)	5120
gemma-3-27b-it Team et al. (2024)	GQA		16 (4 KV)	5120
GPT-OSS-20B Black et al. (2022)	Multi-Head	∼300B	64	6144
amd/Instella-3B Liu et al. (2025)	Multi-Head	4.15T	32	2560
amd/Instella-3B-Instruct Liu et al. (2025)	Multi-Head		32	2560

Continued on next page

Table 3 continued from previous page

Model	Attention Type	Training Tokens	#Heads	Hidden Dim
Llama-3.1-8B-Instruct Meta AI (2024)	MHA		32	4096
state-spaces/mamba-1.4b-hf Gu & Dao (2023)	(SSM)			
allenai/OLMo-2-1124-7B Groeneveld et al. (2024)	MHA		32	4096
allenai/OLMo-2-1124-7B-Instruct Groeneveld et al. (2024)	MHA		32	4096
allenai/OLMo-2-1124-13B-Instruct Groeneveld et al. (2024)	MHA		40	5120
OLMo-2-1124-7B-DPO Groeneveld et al. (2024)	MHA		32	4096
OLMo-2-1124-7B-SFT Groeneveld et al. (2024)	MHA		32	4096
AMD-OLMo-1B Groeneveld et al. (2024)	MHA		16	2048
AMD-OLMo-1B-SFT-DPO Groeneveld et al. (2024)	MHA		16	2048
GPT-J-6B EleutherAI (2021)	Multi-Head	∼300B	16	4096
orca-2-7b Mukherjee et al. (2023c)	Multi-Head		32	4096
orca-2-13b Mukherjee et al. (2023c)	Multi-Head		40	5120
phi-3.5-mini-instruct Gunasekar et al. (2023)	Multi-Head		32	2048
phi-3-mini-128k-instruct Gunasekar et al. (2023)	Multi-Head		32	2048
phi-3-medium-128k-instruct Gunasekar et al. (2023)	Multi-Head		40	5120
Phi-3-medium-4k-instruct Gunasekar et al. (2023)	Multi-Head		40	5120
phi-1_5 Gunasekar et al. (2023)	Multi-Head	7B	32	2048
qwen-14b Qwen Team (2023)	GQA	3T	40 (10 KV)	5120
qwen-14b-chat Qwen Team (2023)	GQA		40 (10 KV)	5120
qwen-2-7b Qwen Team (2023)	GQA		32 (8 KV)	4096
qwen-2-7b-instruct Qwen Team (2023)	GQA		32 (8 KV)	4096
qwen-2.5-7b Qwen Team (2023)	GQA		32 (8 KV)	4096
qwen-2.5-7b-instruct Qwen Team (2023)	GQA		32 (8 KV)	4096
qwen-2.5-14b Qwen Team (2023)	GQA		40 (10 KV)	5120
qwen-2.5-14b-instruct Qwen Team (2023)	GQA		40 (10 KV)	5120
t5-3b Raffel et al. (2020)	Multi-Head	∼1T	32	1024
t5-11b Raffel et al. (2020)	Multi-Head	∼1T	128	1024
yi-9B 01.AI Team (2024)	Multi-Head	3.1T	32	6144

Continued on next page

Table 3 continued from previous page

Model	Attention Type	Training Tokens	#Heads	Hidden Dim
yi-9B (coder) 01.AI Team (2024)	Multi-Head	0.8T	32	6144

A.6 Details for SPoSE (Odd-One-Out) Estimation

Given $S = XX^\top$ with $X \geq 0$, the triplet likelihood is the three-way softmax in equation 1. The MAP program equation 2 is optimized by minibatch stochastic gradients; λ is tuned by held-out triplets. The released THINGS-data SPoSE embeddings are derived from large-scale odd-one-out judgments and approach the behavioral noise ceiling in predicting left-out triplets, despite far fewer than the $O(n^3)$ triplets required to fully determine a dense similarity matrix (Hebart et al., 2023).

A.7 Ordinal Embedding: Risk Bounds and Recovery

Let $G^\star$ be the centered Gram matrix of $\{x_i^\star\} \subset \mathbb{R}^d$ and let $\mathcal{L}(G)$ denote the expected logistic triplet loss induced by equation 3. If $G^\star$ has rank d and $|S|$ triplets are drawn uniformly at random, then with probability $1 - \delta$,

$$\mathcal{L}(\hat{G}) - \mathcal{L}(G^\star) \;\leq\; C_1 \sqrt{\frac{d\, n \log n}{|S|}} + C_2 \sqrt{\frac{\log(1/\delta)}{|S|}}, \tag{6}$$

for universal constants C_1, C_2 depending on the loss Lipschitz constant (Jain et al., 2016). Moreover, writing $C^\star$ for the component of the EDM orthogonal to the linear operator's kernel, one can recover the distance structure with

$$\frac{1}{n^2} \|\hat{C} - C^\star\|_F^2 \;=\; \tilde{O}\left(\frac{d\, n \log n}{|S|}\right), \tag{7}$$

which implies equation 7 samples suffice up to logarithmic factors; $\Omega(d\, n \log n)$ lower bounds are known (Jamieson & Nowak, 2011).

A.8 Dimensionality Selection from Human SPoSE

We compute PCA on the released SPoSE embedding matrix (items $\times$ dimensions), retain the top d components explaining at least 95% variance, and set $d = \hat{d}$ for model-side ordinal embeddings. In our corpus ($n = 128$), this yields $\hat{d} = 29$, and therefore $N_{\text{triplets}} = c\, \hat{d}\, n \log n = 31{,}288$ triplets per model where $c = 4$. We increase N_{triplets} to **35,000** to provide a margin of error.

A.9 Category and item-level alignment of THINGS concepts

Superordinate category-level alignment

We first compute mean human-model Procrustes R^2 grouped by superordinate category label (*furniture, container, vehicle, clothing, food, animal, other*) and model size (binned by parameter size $0B$...$1B$, $1B$...$3B$, $3B$...$10B$, and $108B+$). Results are shown in the figure below:

Concept-level alignment We also test whether different concept attributes predict embeddings (cosine) distance between human and model concept representations. We use the following THINGS attribute dimensions as predictors of human-model alignment:

- Typicality
- Familiarity
- Memorability

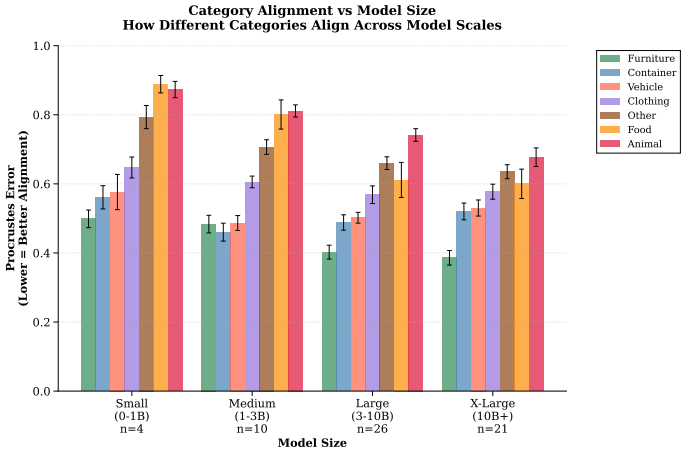


Figure 5: Alignment by superordinate category labels, grouped by model parameter size. Y-axis values indicate Procrustes error.

- Concreteness
- Frequency rank
- Frequency (COCA)
- Frequency (BNC)
- Dispersion
- Polysemy

We fit a mixed-effects regression with model as a random effect and each attribute dimension given as a fixed effect, defined as:

$$d_{\text{cosine}_{ij}} = \beta_0 + \sum_{k=1}^{9} \beta_k X_{ki} + u_j + \epsilon_{ij} \tag{8}$$

where $d_{\text{cosine}_{ij}}$ is the cosine distance between human and model embeddings for item i and model j, β_0 is the intercept, X_{ki} is the value of attribute k for item i, u_j is the random intercept for model j, and ϵ_{ij} is the residual error.

The t-values for all fixed-effects in this regression are included below.

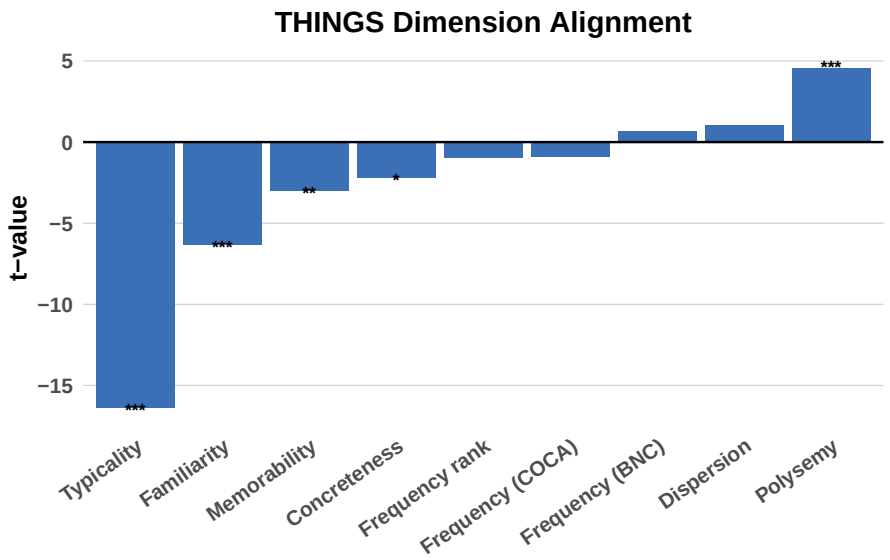


Figure 6: t-values for all THINGS attributes predicting cosine distance between human and model embeddings for each item.

A.10 A case study: the effects of post-training paradigms on alignment

Modern LLM systems undergo various stages of post-training including *supervised fine-tuning* (SFT) (Ouyang et al., 2022), *direct preference optimization* DPO (Rafailov et al., 2023), *reinforcement learning from human feedback* (RLHF) (Ouyang et al., 2022), and *instruction tuning* (IT) (Wei et al., 2021). It remains unclear what impact these post-training steps have on human-model alignment. Progress on this front has been limited due to a lack of open-weight models that have been checkpointed at the various possible stages. `OLMo` (Groeneveld et al., 2024) is a notable exception since it is available at four stages of post-training, including the base model (no post-training), SFT, DPO, and IT and is available at two model sizes (7B and 13B). `Instella` (Liu et al., 2025) and `Llama-3.1` (Meta AI, 2024) also have similar checkpointed model variants available with Instella having an additional stage of pretraining checkpointed.

A.11 Test loss and accuracy comparisons for different embeddings dimensionality

We re-compute embeddings for small, medium and large number parameter models and compare downstream performance of embeddings fit from those models with dimensions 20...30. We find that embeddings are largely similar in test loss and accuracy across different numbers of dimensions.

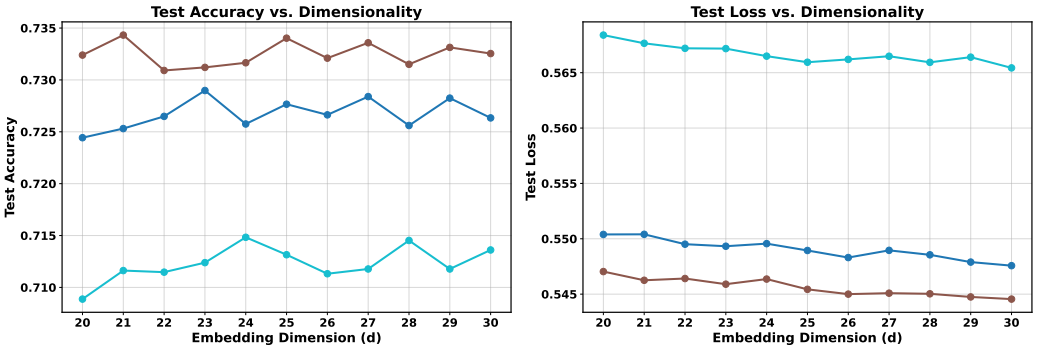


Figure 7: **Embeddings dimensionality comparisons**. Brown, blue, and aqua lines indicate test accuracy/loss for `gemma-3-27b-it`, `gemma-3-12b-it`, and `gemma-3-4b-it` respectively.

A.12 Differences in SPoSE and crowd-kernel computed representations

As human embeddings were computed using SPoSE and model embeddings were computed using the crowd-kernel algorithm, it is necessary to ensure that our human-model alignment results are not confounded by differences in implementation between SPoSE and crowd-kernel methods. To this end, we re-compute embeddings for 77 concepts where we have anchored human similarity judgments using both SPoSE and crowd-kernel algorithms, and then compute the Procrustes R^2 between the resulting embedding spaces. We observe extremely high alignment between embeddings computed from SPoSE and crowd-kernel (R^2=0.99) suggesting that differences due to embeddings algorithm are negligible. Comparisons of RSMs for embeddings computed with SPoSE and crowd-kernel are given below.

A.13 Human-Model alignment for abstract concepts

In order to test whether our analyses generalized to non-concrete concepts, we replicated the exact same procedure described in the main text and obtained human and model semantic embeddings for a set of 213 emotion words derived from Shaver et al. (1987). These words included terms like dread, love, alienation, rage, among others. In general, as can be seen in Figure 8, we found thar larger context lengths, MLP and embedding dimensionality were strong predictors of alignment for these concepts as well. We also replicated the core finding that instruction-tuned models were also more aligned. The full set of t-scores for all predictors can be seen in Figure 8E.

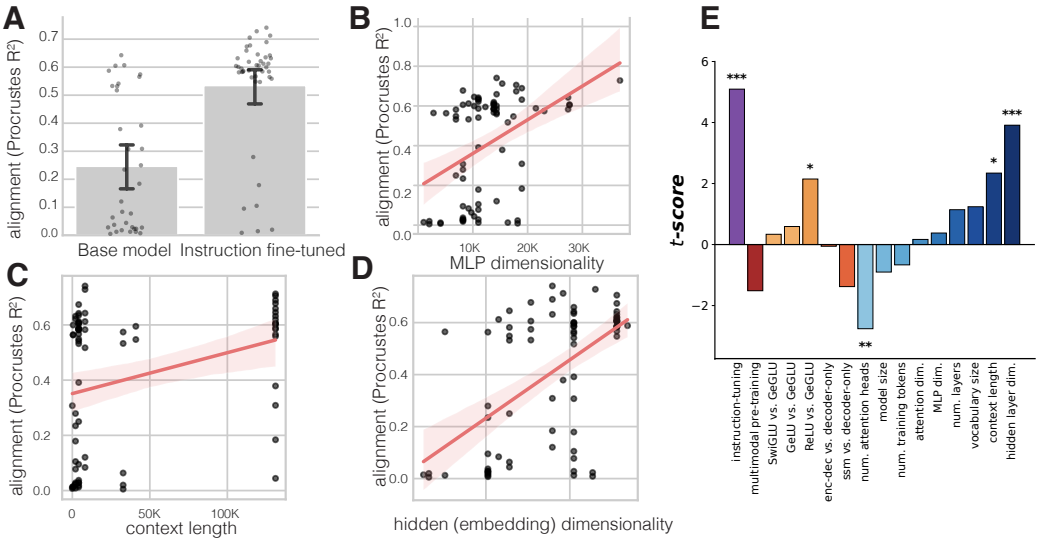


Figure 8: Factors contributing to human-model alignment for abstract emotions concepts.

A.14 Measure correlations

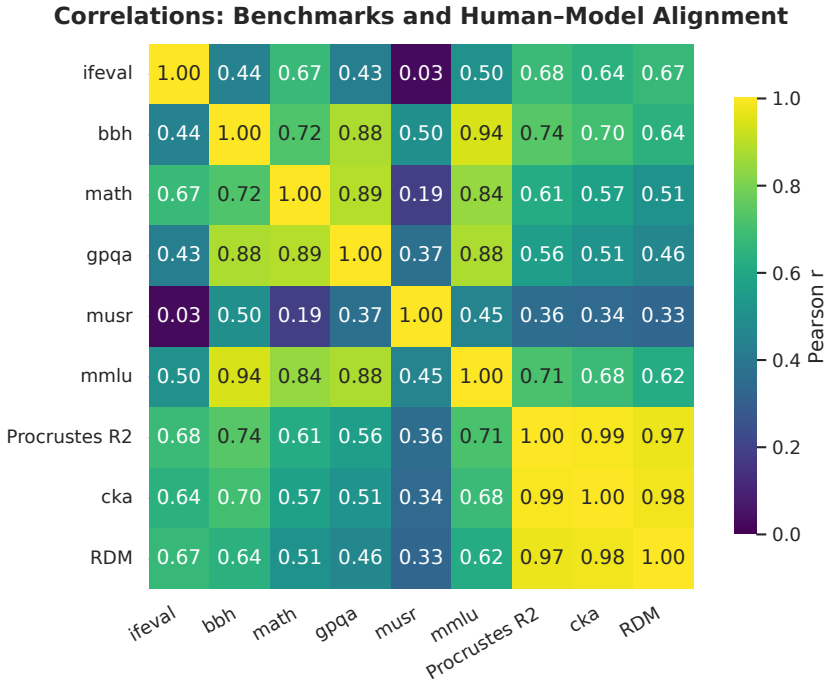


Figure 9: Correlations between measures of human-model alignment and model benchmarks, for models where benchmark scores are publicly are available. (Procrustes R^2, CKA, RSM)

A.15 Consistency in human-alignment for full and subsampled concept embeddings

While our procedure for subsampling embeddings (using n=128 concepts stratified across key object dimensions) is consistent with that of Mukherjee et al. (2023a), it is an empirical question as to whether we would observe the same results if our embeddings were instead sampled using judgments from the full $n = 1,824$ item space. To this end, we obtained 1.2 million triplets on all 1,854 items for the most aligned model (`qwen-2.5-14b`) and one of the least aligned models (`gemma-3-250m-it`) and tested for meaningful differences in the human alignment of the subsampled item embeddings when these embeddings were computed only in the 128 item subsample space, or taken from the full 1824 item space. We computed bootstrapped 95% confidence intervals to test whether the human predictiveness of the subsample was meaningfully different from the full sample, for both `qwen-2.5-14b` and `gemma-3-250m-it`. Figure 10 shows results for t-tests on differences of Procrustes alignment for `qwen-2.5-14b` and `gemma-3-250m-it`. We observe no meaningfull differences in alignment for the larger model `qwen-2.5-14b` ($p > 0.05$), and small differences in alignment for the smallest model in our evaluation `gemma-3-250m-it` ($p < 0.05$).

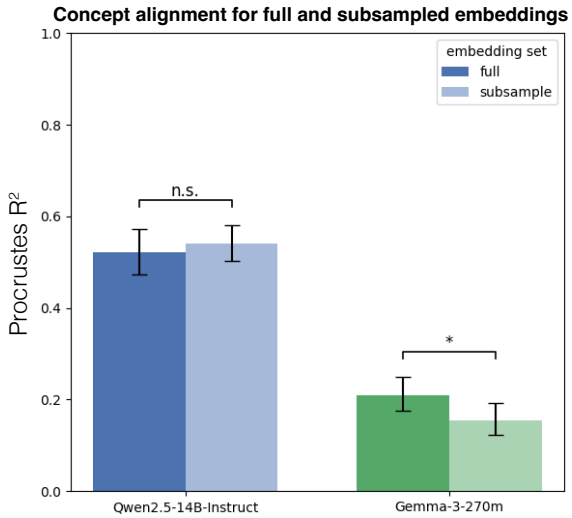


Figure 10: 95% bootstrap CIs for human-alignment of the 128 subsampled concepts, when computed from the subsampled or full embedding space.

A.16 Relationships between computational ingredient properties and human-model alignment

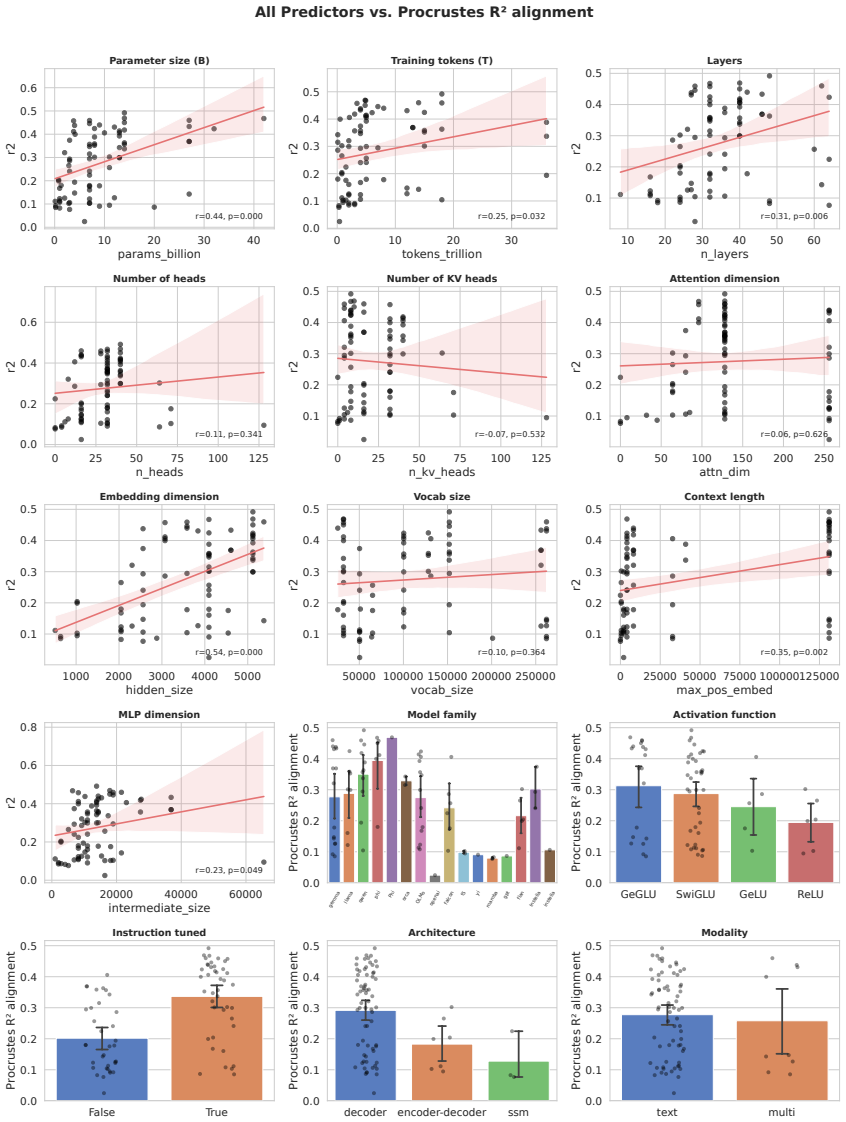


Figure 11: Categorical and continuous between all computational ingredients and human-model alignment. All significant effects in the full mixed linear-model are reported in the main text, Figure 2.

A.17 Testing for multicollinearity in predictors of human-model alignment

Mixed-effects linear models produce unbiased parameter estimates even in cases where level-1 predictors –in our case, computational ingredients– are collinear (Shieh & Fouladi, 2003). However, collinearity among these computational ingredients could present difficulties in future analyses if different statistical models were to be used. We measure the GVIF (generalized variance inflation factor, see Fox & Monette (1992)) values of our fixed effects in both the THINGS and EMOTIONS regressions, and find that none of the significant predictors we report on exceed even an intermediate threshold of multicollinearity. Figure 12 shows GVIF values of parameter estimates for both EMOTIONS and THINGS mixed-effects models. In both cases, we find that GVIF values do not exceed thresholds conventionally agreed upon as indicating high covariance. Additionally, variance explained (R^2) between individual predictors is relatively low for most predictor-predictor relationships, as shown qualitatively by Figure 13.

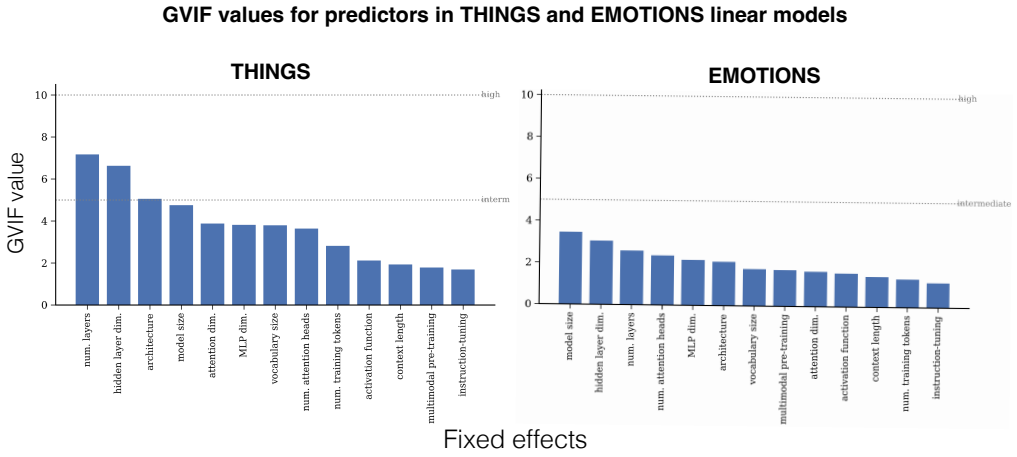


Figure 12: GVIF values for mixed-effects linear models fit from THINGS (top) and EMOTIONS concepts (bottom). Dashed lines indicate GVIF values commonly used to indicate *intermediate* and *high* levels of multicollinearity among predictors.

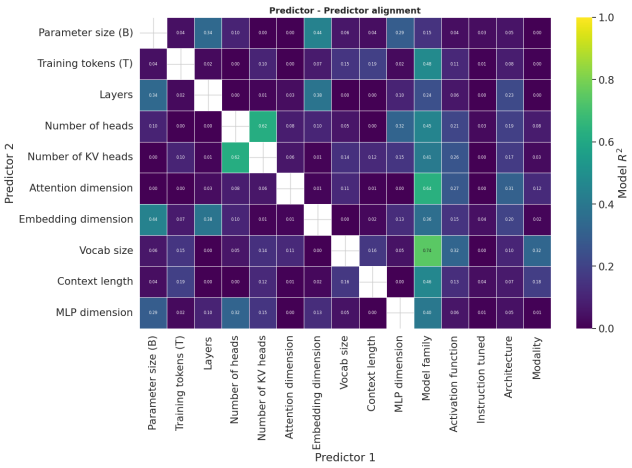


Figure 13: R^2 between individual predictors used in our mixed-effects linear models. For each unique predictor-predictor pair, a linear regression was fit predicting predictor 2 from the values of predictor 1.

A.18 Stability of predictors for different subsamples of models

Would we have observed different computational ingredients predictive of human-model alignment if we had tested a different sample of open source models? We conduct a permutation test to determine whether our observed effects were sensitive to the given models measured. We randomly selected subsamples of $n = 50$ models across $k = 10,000$ iterations, and then used the distributions of observed effects to obtain 95% CIs for each predictor. We find that–even on smaller random subsamples of our models–the effects are commensurate with those we report in the original mixed effects models for THINGS and EMOTIONS. This indicates that significantly smaller subsamples of various open source models would converge to similar findings. Figure 14 shows t-values estimated from these permutation tests for THINGS and EMOTIONS data.

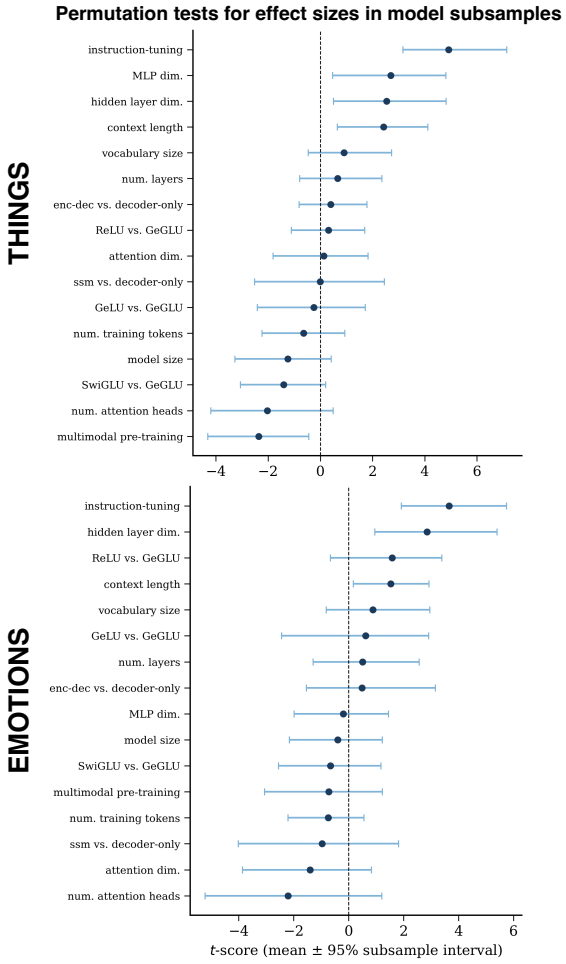


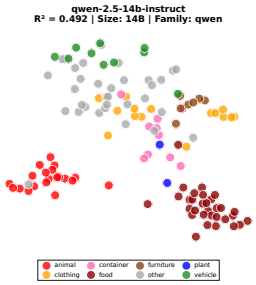
Figure 14: t-values estimated from permutation tests for THINGS and EMOTIONS data, with 95% CIs.

A.19 *t*-SNE Visualizations

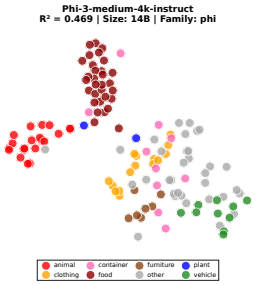
Table 4: Complete ranking of models by R^2 score

Rank	Model Name	R^2 Score	Rank	Model Name	R^2 Score
1	qwen25-14b-instruct	0.492	39	qwen2-7b	0.294
2	Phi-3-medium-4k-instruct	0.469	40	Instella-3B-SFT	0.293
3	Phi-3.5-MoE-instruct	0.468	41	Falcon3-3B-Base	0.286
4	gemma-3-27b-it	0.460	42	flan-t5-xl	0.265
5	qwen25-7b-instruct	0.459	43	falcon-11b	0.257
6	Phi-3.5-mini-instruct	0.457	44	OLMo-2-1124-7B-Instruct	0.241
7	phi-3-medium-128k-instruct	0.450	45	Instella-3B	0.241
8	qwen2-7b-instruct	0.446	46	Falcon3-Mamba-7B-Base	0.224
9	gemma-2-9b-it	0.440	47	flan-t5-large	0.199
10	gemma-3-4b-it	0.438	48	Qwen3-4B-Base	0.194
11	gemma-2-27b-it	0.433	49	phi-1_5	0.180
12	gemma-3-12b-it	0.431	50	OLMo-2-1124-7B	0.179
13	Llama-3.1-8B-Instruct	0.424	51	gemma-2-9b	0.178
14	OLMo-2-0325-32B-Instruct	0.423	52	falcon-7b	0.175
15	qwen-14b-chat	0.418	53	OLMo-2-0425-1B-SFT	0.168
16	OLMo-2-1124-13B-Instruct	0.414	54	llama2-7b-chat	0.160
17	phi-3-mini-128k-instruct	0.412	55	gemma-3-4b-pt	0.147
18	OLMo-2-1124-13B-DPO	0.406	56	gemma-3-27b-pt	0.143
19	Falcon3-10B-Base	0.406	57	gemma-3-12b-pt	0.127
20	Phi-3.5-vision-instruct	0.400	58	gemma-2-2b	0.126
21	OLMo-2-1124-13B-SFT	0.392	59	OLMo-2-0425-1B	0.123
22	qwen3-8b	0.388	60	llama2-7b	0.122
23	Instella-3B-Instruct	0.374	61	AMD-OLMo-1B	0.113
24	gemma-2-27b	0.369	62	flan-t5-small	0.111
25	qwen25-14b	0.363	63	AMD-OLMo-1B-SFT-DPO	0.108
26	Llama-3.1-Tulu-3-8B	0.358	64	Instella-3B-Stage1	0.106
27	qwen-14b	0.357	65	qwen25-7b	0.104
28	OLMo-2-1124-7B-DPO	0.357	66	falcon-7b-instruct	0.103
29	Llama-3.1-Tulu-3-8B-DPO	0.352	67	t5-3b	0.103
30	OLMo-2-1124-7B-SFT	0.347	68	t5-11b	0.095
31	orca-2-13b	0.343	69	gemma-3-270m	0.092
32	qwen3-14b	0.337	70	Yi-9B	0.091
33	gemma-2-2b-it	0.321	71	gpt-oss-20b	0.087
34	orca-2-7b	0.315	72	gemma-3-270m-it	0.085
35	flan-t5-xxl	0.302	73	mamba-1.4b-hf	0.083
36	Llama-3.1-Tulu-3-8B-SFT	0.301	74	mamba-2.8b-hf	0.077
37	llama2-13b-chat	0.299	75	gpt-j-6b	0.025
38	OLMo-2-1124-13B	0.299			

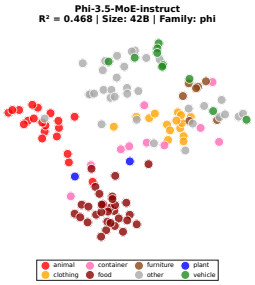
Figure 15: t-SNE visualizations of model embeddings (Part 1). R^2 scores are in parentheses.



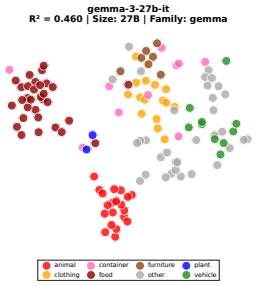
(1) Qwen2.5-14B-Inst (0.492)



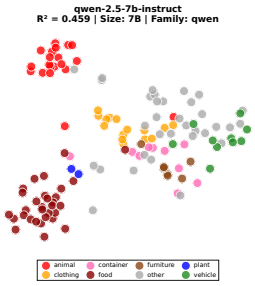
(2) Phi-3-Med-4k (0.469)



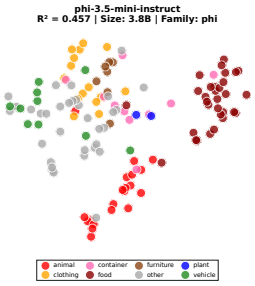
(3) Phi-3.5-MoE (0.468)



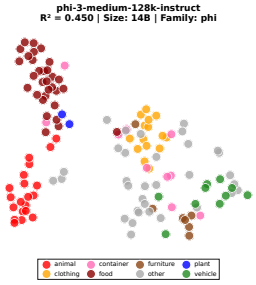
(4) Gemma-3-27B-IT (0.460)



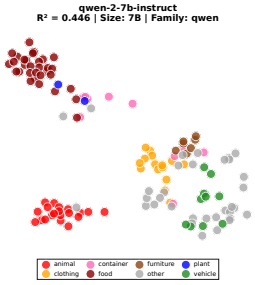
(5) Qwen2.5-7B-Inst (0.459)



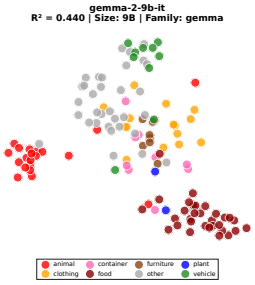
(6) Phi-3.5-Mini (0.457)



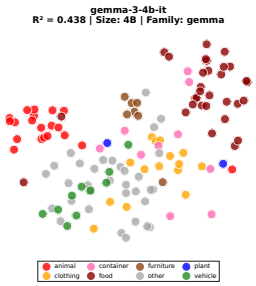
(7) Phi-3-Med-128k (0.450)



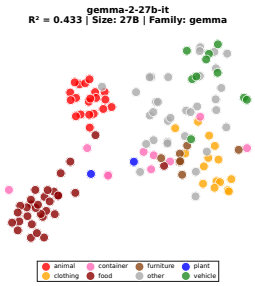
(8) Qwen2-7B-Inst (0.446)



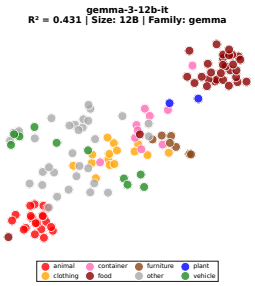
(9) Gemma-2-9B-IT (0.440)



(10) Gemma-3-4B-IT (0.438)

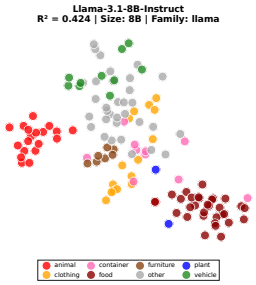


(11) Gemma-2-27B-IT (0.433)

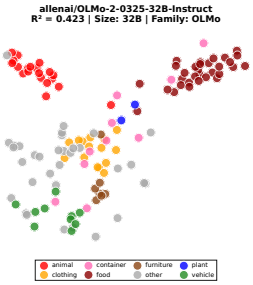


(12) Gemma-3-12B-IT (0.431)

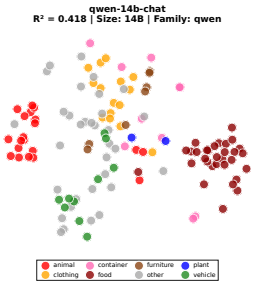
Figure 16: t-SNE visualizations of model embeddings (Part 2). R^2 scores are in parentheses.



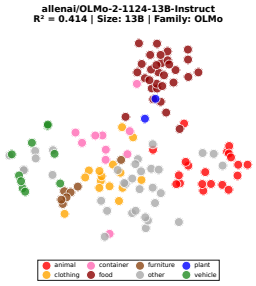
(13) Llama-3.1-8B (0.424)



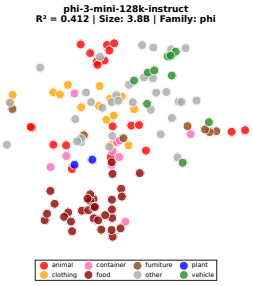
(14) OLMo-2-32B (0.423)



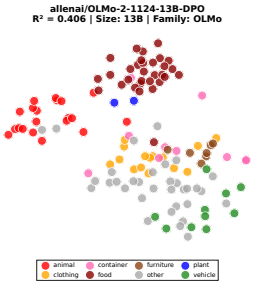
(15) Qwen-14B-Chat (0.418)



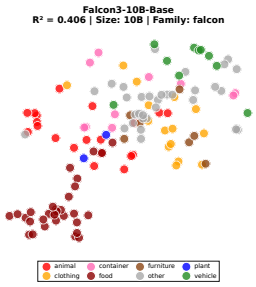
(16) OLMo-2-13B-Inst (0.414)



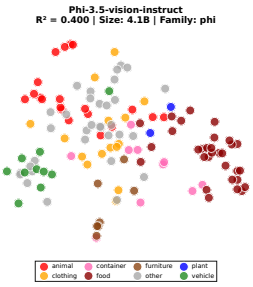
(17) Phi-3-Mini-128k (0.412)



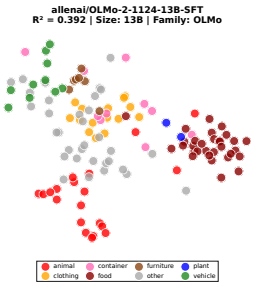
(18) OLMo-2-13B-DPO (0.406)



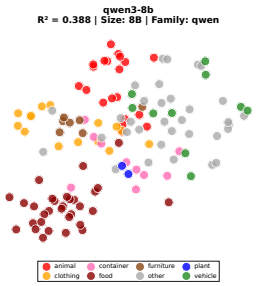
(19) Falcon3-10B (0.406)



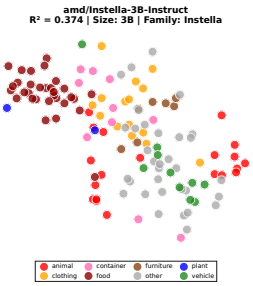
(20) Phi-3.5-Vision (0.400)



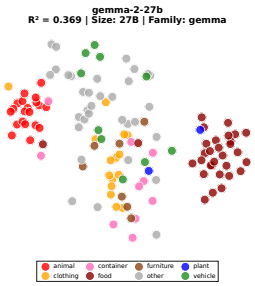
(21) OLMo-2-13B-SFT (0.392)



(22) Qwen3-8B (0.388)

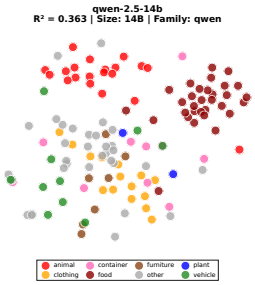


(23) Instella-3B-Inst (0.374)

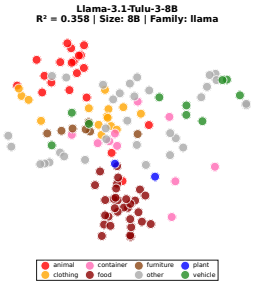


(24) Gemma-2-27B (0.369)

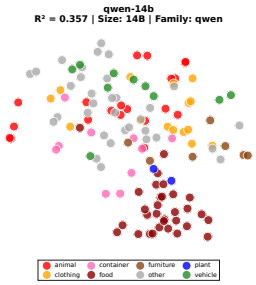
Figure 17: t-SNE visualizations of model embeddings (Part 3). R^2 scores are in parentheses.



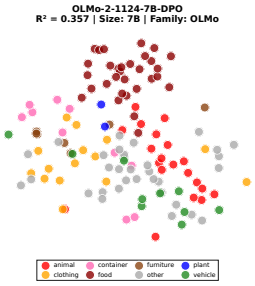
(25) Qwen2.5-14B (0.363)



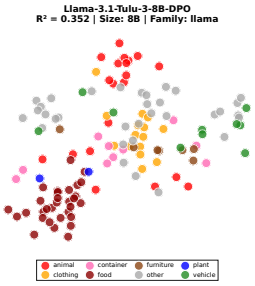
(26) Llama-3.1-Tulu (0.358)



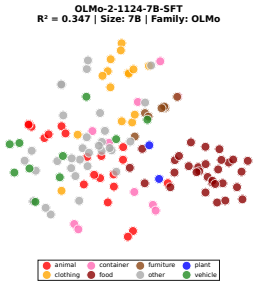
(27) Qwen-14B (0.357)



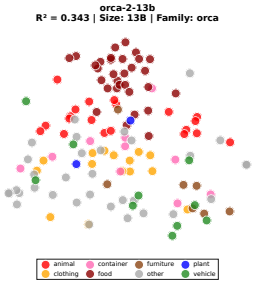
(28) OLMo-2-7B-DPO (0.357)



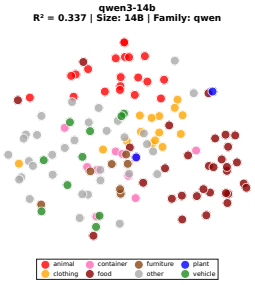
(29) Llama-3.1-Tulu-DPO (0.352)



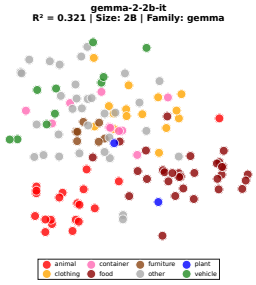
(30) OLMo-2-7B-SFT (0.347)



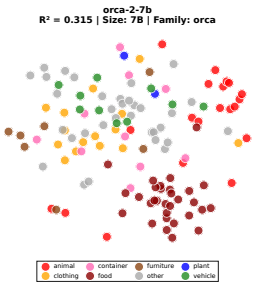
(31) Orca-2-13B (0.343)



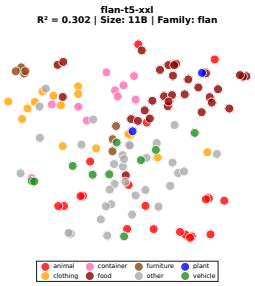
(32) Qwen3-14B (0.337)



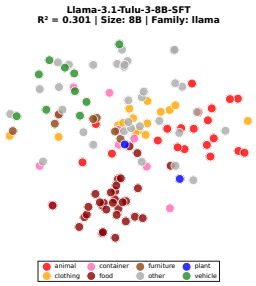
(33) Gemma-2-2B-IT (0.321)



(34) Orca-2-7B (0.315)

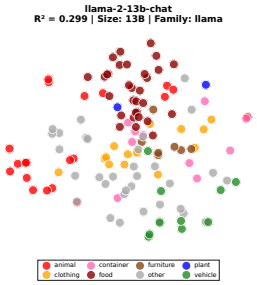


(35) FLAN-T5-XXL (0.302)

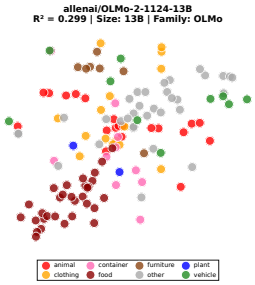


(36) Llama-3.1-Tulu-SFT (0.301)

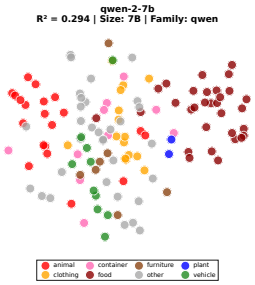
Figure 18: t-SNE visualizations of model embeddings (Part 4). R^2 scores are in parentheses.



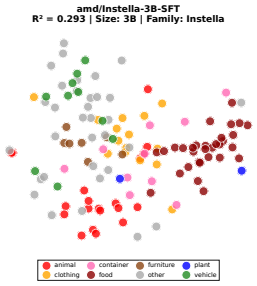
(37) Llama2-13B-Chat (0.299)



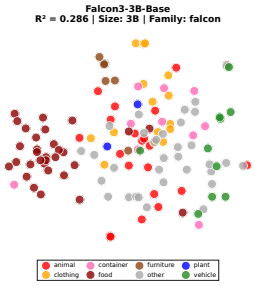
(38) OLMo-2-13B (0.299)



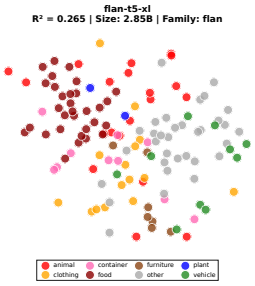
(39) Qwen2-7B (0.294)



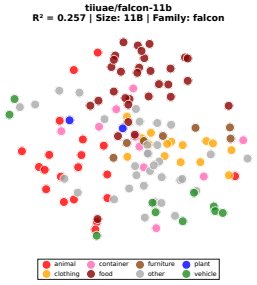
(40) Instella-3B-SFT (0.293)



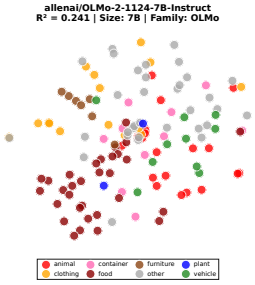
(41) Falcon3-3B (0.286)



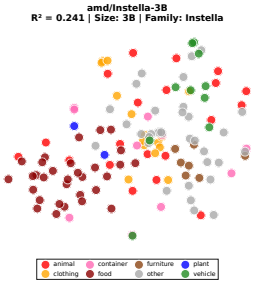
(42) FLAN-T5-XL (0.265)



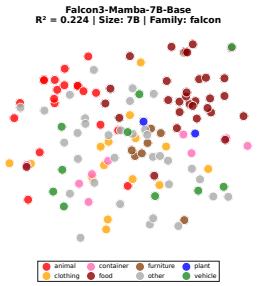
(43) Falcon-11B (0.257)



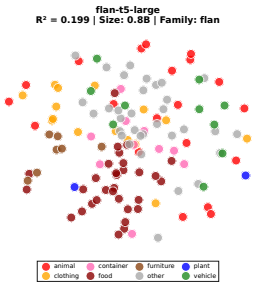
(44) OLMo-2-7B-Inst (0.241)



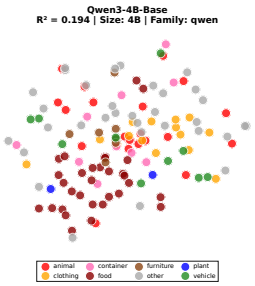
(45) Instella-3B (0.241)



(46) Falcon3-Mamba-7B (0.224)

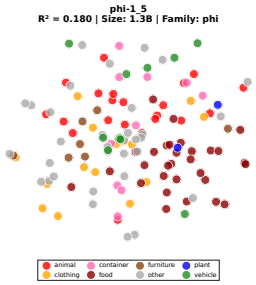


(47) FLAN-T5-Large (0.199)

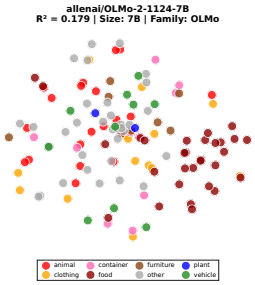


(48) Qwen3-4B (0.194)

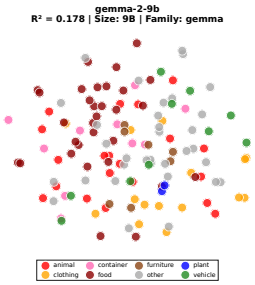
Figure 19: t-SNE visualizations of model embeddings (Part 5). R^2 scores are in parentheses.



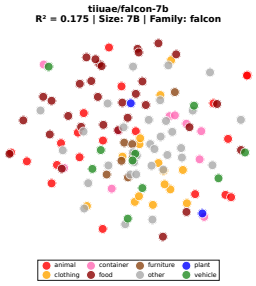
(49) Phi-1.5 (0.180)



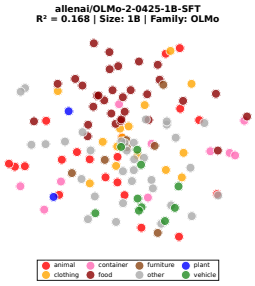
(50) OLMo-2-7B (0.179)



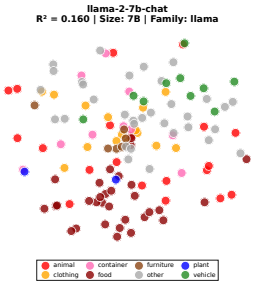
(51) Gemma-2-9B (0.178)



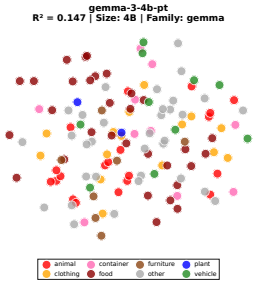
(52) Falcon-7B (0.175)



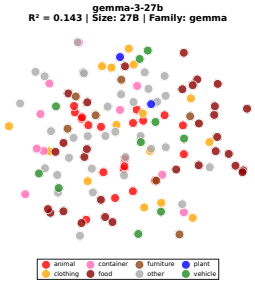
(53) OLMo-2-1B-SFT (0.168)



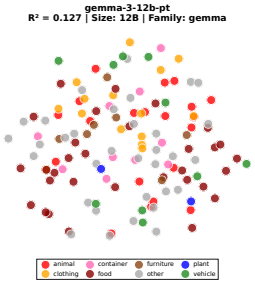
(54) Llama2-7B-Chat (0.160)



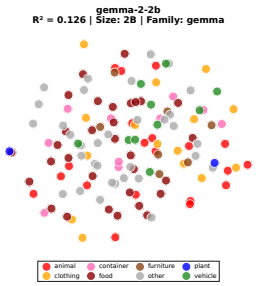
(55) Gemma-3-4B-PT (0.147)



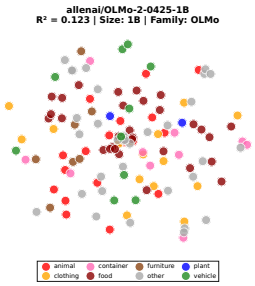
(56) Gemma-3-27B-PT (0.143)



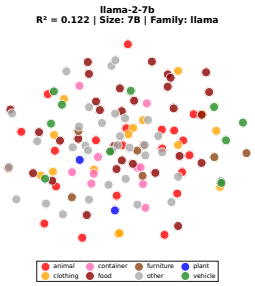
(57) Gemma-3-12B-PT (0.127)



(58) Gemma-2-2B (0.126)

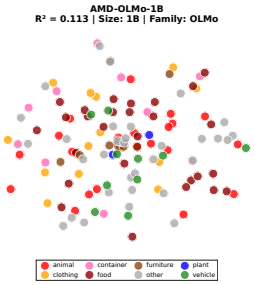


(59) OLMo-2-1B (0.123)

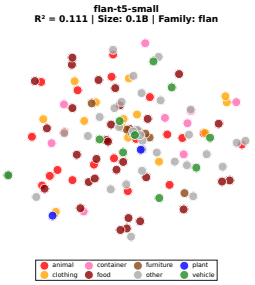


(60) Llama2-7B (0.122)

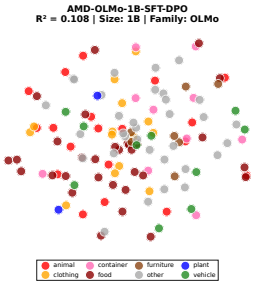
Figure 20: t-SNE visualizations of model embeddings (Part 6). R^2 scores are in parentheses.



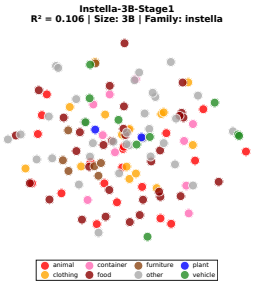
(61) AMD-OLMo-1B (0.113)



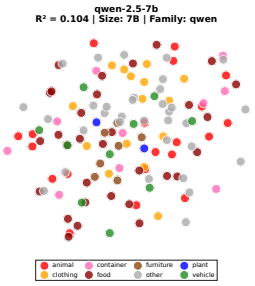
(62) FLAN-T5-Small (0.111)



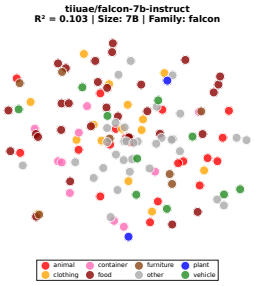
(63) AMD-OLMo-1B-DPO (0.108)



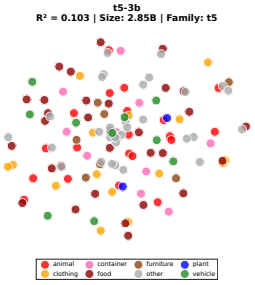
(64) Instella-3B-Stage1 (0.106)



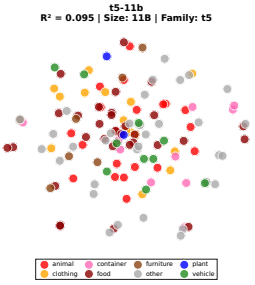
(65) Qwen2.5-7B (0.104)



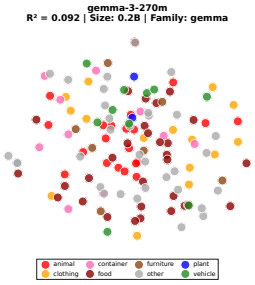
(66) Falcon-7B-Inst (0.103)



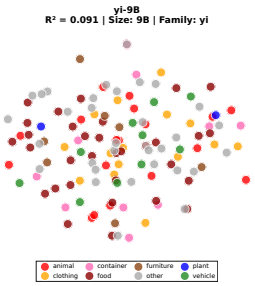
(67) T5-3B (0.103)



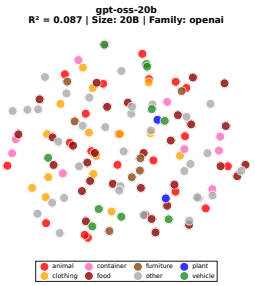
(68) T5-11B (0.095)



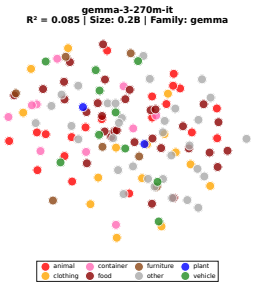
(69) Gemma-3-270M (0.092)



(70) Yi-9B (0.091)

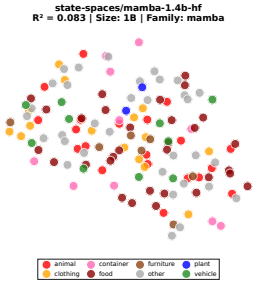


(71) GPT-OSS-20B (0.087)

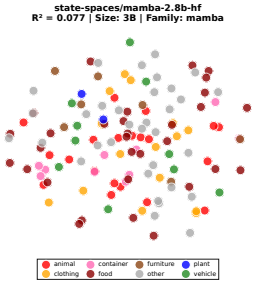


(72) Gemma-3-270M-IT (0.085)

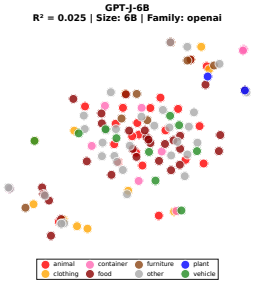
Figure 21: t-SNE visualizations of model embeddings (Part 7). R^2 scores are in parentheses.



(73) Mamba-1.4B (0.083)



(74) Mamba-2.8B (0.077)



(75) GPT-J-6B (0.025)