Comparing Perceptual Judgments in Large Multimodal Models and Humans

Billy Dickson[1*†], Sahaj Singh Maini[1†], Robert Nosofsky[2], Zoran Tiganj[1,2*]

[1]Department of Computer Science, Indiana University Bloomington
[2]Department of Psychological and Brain Sciences, Indiana University Bloomington

*Corresponding authors: Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, 700 N Woodlawn Ave, Bloomington, IN 47408, USA.
E-mail: dicksonb@iu.edu, ztiganj@iu.edu.
[†]These authors contributed equally to this work.

Abstract

Cognitive scientists commonly collect participants' judgments regarding perceptual characteristics of stimuli to develop and evaluate memory, learning, and decision-making models. For instance, to model human responses in tasks of category learning and item recognition, researchers often have to collect perceptual judgments of images to embed the images in multidimensional feature spaces. This process is time-consuming and costly. Recent advancements in Large Multimodal Models (LMMs) provide a potential alternative since such models can respond to prompts that include both text and images and could potentially replace human participants. To test whether the available LMMs can indeed be useful for this purpose, we evaluated their judgments on a dataset consisting of rock images that has been widely used by cognitive scientists. The dataset includes human perceptual judgments along ten dimensions considered important for classifying rock images. While the models exhibited a strong positive correlation with human responses, we found that they fell short in replacing an average of a set of judgments from human participants. The models provided correlations with these averaged data that were roughly the same magnitude as observed for individual participants, especially for dimensions that are relatively general (such as lightness and chromaticity) as opposed to domain-specific dimensions (such as pegmatitic structure), where they struggled more. We also found that modifying prompts and providing additional examples of images with corresponding ratings had a positive but relatively modest impact on model performance. Our study provides a benchmark for evaluating future LMMs on human perceptual judgment data.

Introduction

Among the fundamental goals in the computational modeling of cognitive processes such as categorization, old-new recognition, and decision making is to account for human performance at the level of individual items (Love et al., 2004; Nosofsky, 1986, 1991; Nosofsky et al., 2011; Shepard, 1957). With respect to this goal, a major current trend in the psychological and cognitive sciences has involved the scaling up of the models from applications in simple, highly controlled, low-dimensional stimulus domains to complex, real-world high-dimensional ones (Bainbridge, 2019; Battleday et al., 2020; Hebart et al., 2020; Meagher & Nosofsky, 2023; Nosofsky, Sanders, & McDaniel, 2018; Storms et al., 2000). In addition, a current trend is to scale up the application of the models to cases in which performance is modeled for very large numbers of individual items from the domains of interest (Battleday et al., 2020; Hebart et al., 2020; Kramer et al., 2023; Meagher & Nosofsky, 2023).

Applying the models for predicting individual-item performance requires the specification of a psychological "feature space" in which the items are embedded. Historically, one of the major approaches to deriving such an input space for the cognitive models has been to conduct independent tasks involving the collection of varied forms of "similarity-scaling" data (Nosofsky, 1992; Roads & Love, 2024). For example, observers might be required to make direct judgments of the similarity between different pairs of items (Shepard, 1962); to judge whether pairs of presented items are "same" or "different" (Kruskal & Wish, 1978; Rothkopf, 1957); to choose which item among a triad of presented items is the "odd-one-out" (Hebart et al., 2020; Roads & Mozer, 2019); or to form spatial arrangements of entire sets of objects in which the distance between objects is proportional to their judged dissimilarity (Goldstone, 1994; Hout, Goldinger, & Ferguson, 2013). Formal statistical models can then be applied for modeling the obtained similarity-scaling data by embedding the judged items in derived multidimensional feature spaces (Hout, Papesh, & Goldinger, 2012; Lee, 2001; Shepard, 1962, 1980, 1987).

In modern work, however, a limitation of this classic approach is that the techniques eventually become intractable when the number of to-be-scaled items is extremely large and/or when the dimensionality of the feature space is large and complex (but for modern efforts along these lines, see, Hebart et al., 2020; Marjieh et al., 2024; Nosofsky, Sanders, Meagher, et al., 2018. For example, to fill out a similarity-judgment matrix with just a single judgment for all pairs of n items requires something on the order of n^2 similarity judgments. If n=1,000, then something on the order of 1,000,000 judgments are needed.

Therefore, as an alternative to this similarity-scaling approach, a major current trend is for researchers to use a variety of machine-based deep-learning network approaches for deriving the feature space to be used as input to the cognitive-process models (Annis et al., 2021; Battleday et al., 2020; Lake et al., 2015; Peterson et al., 2018; Roads & Love, 2021). The general approach here is to train deep-learning models to learn to classify large sets of items into different categories in a domain of interest. By using appropriate validation and generalization techniques to avoid overfitting the noise in the data, the respective models learn a set of weights that can classify with high accuracy both the trained items and novel items from the relevant domains. Once the training has been completed, the deep-learning model can then be used as a mechanism for embedding the items in a candidate feature space. In particular, the presentation of any given item will give rise to a set of activations of the nodes in the network. Oftentimes, some transformation of the node activations at the penultimate layer of the network, immediately prior to the final classification layer, are chosen as the candidate features. These candidate features are then used as inputs to varieties of cognitive-process models used for predicting human performance in related tasks and domains.

Although researchers have achieved some impressive successes with these deep-learning approaches, these approaches also have potential limitations. One limitation is the lack of assurance that all stimulus features detected by the machine learning models are similarly apprehended by humans. To take a hypothetical example, suppose that a particular machine-

learning model can apprehend wavelengths of light or sound that go beyond human sensitivities and that these “extra-sensory” features turn out to be highly diagnostic for purposes of classification. In that case, the candidate feature space derived from the deep-learning model would include non-human features, so the use of that feature space in combination with cognitive models could lead to misleading conclusions involving the nature of human performance. A second limitation is that there is an enormous set of highly complex nonlinear transformations that take place in translating elementary input features in the networks into the patterns of activations at the penultimate layers. Oftentimes, it is extremely difficult to place a psychological interpretation on the patterns of penultimate-layer activations. Thus, the machine-learning-based feature space that is used as an input to the cognitive models may not correspond to the types of foundational psychological building-block features that are supposed by the cognitive models in the first place.

To address the above-stated limitations, Sanders & Nosofsky (2018, 2020) proposed a hybrid two-step approach (see also Rumelhart & Todd, 1993; Steyvers & Busey, 2000). The approach was intended to combine the strengths of the similarity-scaling and deep-learning approaches to deriving psychological feature spaces for large numbers of real-world high-dimensional objects. In the first step, one uses traditional similarity-scaling methods for deriving a feature-space representation for a representative *subset* of the objects in the domain of interest. In the second step, rather than training a deep-learning network to classify objects in this domain into categories, one instead trains the deep-learning network to reproduce the actual feature-space representation for the subset of objects that was derived from the *human* judgments. Again, by using appropriate validation and generalization techniques to avoid overfitting the training-item representation, the approach could allow for the embedding of an essentially infinite number of objects from the domain of interest in a high-dimensional *psychological* feature space.

A further advantage of the proposed two-step approach is that the same technique can be applied for positioning objects along additional dimensions not revealed by the similarity-scaling techniques themselves. For example, Sanders & Nosofsky (2020) tested the proposed approach in a domain involving igneous rock categories as defined in the geologic sciences. Using similarity-scaling techniques, Nosofsky, Sanders, Meagher, et al. (2018) had previously found that an 8-dimensional multidimensional scaling (MDS) representation provided an excellent account of similarity-judgment data obtained for a set of 360 rock images. In addition, the derived MDS dimensions had natural psychological interpretations. These included the extent to which the rocks were: i) light or dark, ii) fine- or coarse-grained, iii) smooth or rough, iv) dull or shiny, v) had disorganized vs. organized textures, vi) were achromatic vs. chromatic, and vii) had a green vs. red hue. (A firm interpretation was not obtained for the eighth dimension, but it appeared to have shape-related components.) Crucially, however, Nosofsky et al. (2020) and Sanders and Nosofsky (2020) discovered that in independently conducted classification-learning experiments, observers made use of additional "supplementary" dimensions not revealed by the initial MDS analysis but that were highly diagnostic for purposes of classifying the rock images into their geologically-defined categories. One example included the extent to which the rocks possessed "porphyritic structure" – the embedding of small-sized fragments within a fine-grained groundmass. (We describe several other examples of these supplementary dimensions below.) In previous research projects, Nosofsky et al. (2020) and Nosofsky, Sanders, Meagher, et al. (2018) obtained direct ratings from human subjects of the positions of all the rock images along each of the MDS-derived and supplementary dimensions. Importantly, the two-step approach proposed by Sanders and Nosofsky (2020) for training deep-learning networks to reproduce MDS-derived representations also showed some preliminary success in reproducing the position of the rocks along these supplementary dimensions as well.

Despite the initial promise of the proposed two-step method, Sanders and Nosofsky (2020) acknowledged that improved machine-learning technologies would likely yield more

effective procedures for implementing it. Indeed, in the short time since the original proposal, an explosion of such improved technologies has emerged.

In recent years multimodal vision and language models demonstrated the ability to produce advanced image understanding. These models combine inputs from both visual and textual sources to understand and generate content that reflects a combined understanding of both modalities. Multimodal vision and language architectures such as CLIP (Radford et al., 18-24 Jul 2021) demonstrate powerful zero-shot learning, surpassing humans in zero-shot, one-shot and two-shot learning on the Oxford IIT Pets image classification dataset. When embedded into larger frameworks such as Open AI's GPT4 (OpenAI et al., 2023), Google's Gemini (Gemini Team et al., 2024), or Anthropic's Claude-3 (Anthropic 2024), which all have multiple training components, including reinforcement learning from human feedback (Glaese et al., 2022), these multimodal models gain additional power and flexibility in terms of interaction with the human users. Careful design of prompts, specifically techniques such as chain-of-thought prompting, can further increase the performance of these models (Wei et al., 2022; Lin 2024). Kouwenhoven et al. (2022) highlight the importance of developing shared vocabularies between humans and machines, proposing that natural evolution in communication could significantly improve AI interactions. Recent research investigated relationships between humans more systematically by quantifying differences across specific perceptual features (Geirhos et al., 2021; Sheybani et al., 2024) and visual illusions (Shahgir et al., 2024), complemented by an overview by de Kleijn (2022) outlining the evolution of AI and neural networks in relation to the human brain.

The central purpose of the present work was to investigate the effectiveness of the current technology using Large Multimodal Models (LMMs) for positioning large sets of complex, real-world objects in high-dimensional feature spaces that align with human judgments. Following Sanders and Nosofsky (2020), our example target domain involves the set of 360 rock images used in their earlier study. This domain is particularly suitable for our current

investigation due to the substantial amount of previously collected psychological-scaling and direct dimension-rating data for these rock images. Consequently, the predictions generated by LMMs can be rigorously compared to the extensive existing sets of human judgments.

In the present work, we conduct an experimental investigation of the performance of LMMs that accept image and text inputs, specifically OpenAI GPT4 and Anthropic Claude-3 model family, for reproducing human dimension ratings within the rock domain[1]. Our investigation is structured around three primary implementation conditions. In each condition, we evaluate the correspondence between the ratings produced by LMMs and human ratings along a set of 7 MDS-derived dimensions and 3 supplementary dimensions. In the first condition, for each to-be-judged image, we provide LMMs with verbal prompts that are identical to the verbal prompts that had been provided to human subjects in the previous dimension-rating studies. This first condition has certain similarities to an important investigation reported recently by Marjieh et al. (2023). Using only verbal prompts, these researchers asked GPT4 and some related machine-learning systems to make judgments of similarity for stimuli in six domains, such as tones varying in pitch or colors varying in wavelength. They then obtained MDS solutions for the GPT4-produced similarity ratings and found that the recovered MDS solutions corresponded well with ones derived in past studies in which humans provided similarity judgments of the actual perceptual stimuli. By contrast, in Condition 1 of the present study, on each trial we present to the machine-learning system an actual rock image to be judged. We then use verbal prompts to ask the machine-learning algorithm to produce its dimension ratings for the image. The system is asked to produce direct ratings along 10 different dimensions that are components of a single set of complex rock-image stimuli.

[1] In independent work conducted in parallel with the present work, Sanders (2024) has been exploring the performance of GPT4 in reproducing human judgments for a subset of 30 rock images from the current rocks-360 set. He reports results similar to those that we report in this article.

In the second condition of our investigation, we supplement the prompts with a set of images that illustrate examples with low, medium, and high values along the rated dimensions. These “anchor images” had in fact been used in the original studies conducted by Nosofsky, Sanders, Meagher, et al. (2018, 2020) in which the human dimension ratings for the rocks had originally been collected. We hypothesized that the correspondence between the multimodal machine-learning ratings and the human ratings would be significantly improved when the system was provided with the example anchor images.

Finally, we followed the investigations in these two main conditions with some further explorations on a subset of the dimensions. Although these additional explorations did not involve comprehensive tests across all the dimensions and machine-learning systems, for simplicity we will refer to them as the third condition. Specifically, we explored whether the models might yield better correspondences with the human ratings on some challenging dimensions if we queried the system with additional anchor images and also provided more detailed prompts that contained additional information about the judged dimensions.

Methods

We used previously collected data from human participants who performed dimension-ratings of 360 rock images across a set of continuous and present-absent dimensions (Meagher & Nosofsky, 2023; Nosofsky, Sanders, Meagher, et al., 2018). The complete dataset including rock images and participant ratings is publicly available through OSF: https://osf.io/cvwu9/.

Stimuli

The stimuli were 360 images of rocks obtained from the web and processed to remove background objects and idiosyncratic markings such as text labels. The 360 images included 120 rocks from each of three main divisions: igneous, metamorphic, and sedimentary. Each main division included 12 rocks from ten major subtype categories (30 subtypes in total), such as granite, marble, sandstone, and so forth.

For human experiments, the stimuli were presented on a 23-in. LCD computer screen. The stimuli were displayed on a white background. Each rock picture was approximately 2.1 in. wide and 1.7 in. tall. Subjects sat approximately 20 in. from the computer screen, so each rock picture subtended a visual angle of approximately 6.0° x 4.9°. Images were selected or digitally manipulated to have similar levels of resolution of the salient features that may be used to identify and classify the particular rock types. All of the images were photographed in a field setting and had not been modified in any way other than the removal of other portions of the original image.

Human dimension ratings

After excluding dimensions that did not have complete participant data, we focused our analysis on ten dimensions, seven derived from MDS analysis of similarity judgments (chromaticity, darkness/lightness, disorganized/organized, dull/shiny, fine/coarse grain, red/green, and smooth/rough) and 3 “supplementary” dimensions (conchoidal fracture, pegmatitic structure, and porphyritic texture) that were not revealed by the MDS analysis but that were added due to being highly diagnostic for purposes of classifying the rock images into their geologically-defined categories. Each of these dimensions had a continuous rating scale from 1 to 9.

Multimodal vision and language models

We employed state-of-the-art LMMs, GPT4 and Claude-3 model family. The Claude-3 model family includes the three models - Opus, Sonnet, and Haiku, with Opus stated to be the most intelligent model and Haiku being the least intelligent but the fastest model amongst the three models (Anthropic 2024). We conducted the analysis in three different conditions. All the models used in the experiment were prompted using API calls. In order to minimize randomness in the responses, the temperature parameter was set to 0 for all the models. Prompts to the models were identical to those provided to human subjects aside from minor modifications

instructing the model to return a numerical dimension rating to the hundredths place[2] and changing plural references to describe a single trial, as each API call to the model was independent. Furthermore, in conditions without anchor images (1 and 3a), any text mentioning the presence of example images was omitted from the prompt. RGB images were used to prompt the models in conditions where anchor images were part of the prompt. All prompts are listed as part of the Supplementary Information available at https://osf.io/za847/.

In the first condition, 360 images were provided to each of the four models for each of the ten different dimensions. No anchor images were provided to the model. For GPT4 specifically, the text 'Do not respond with I'm sorry…' was added to the prompt of several features (fine/coarse grain, red/green hue, porphyritic texture, pegmatitic structure, and conchoidal fracture) in order to minimize redundant responses from the model. An example of this condition is displayed in Figure 1 below.

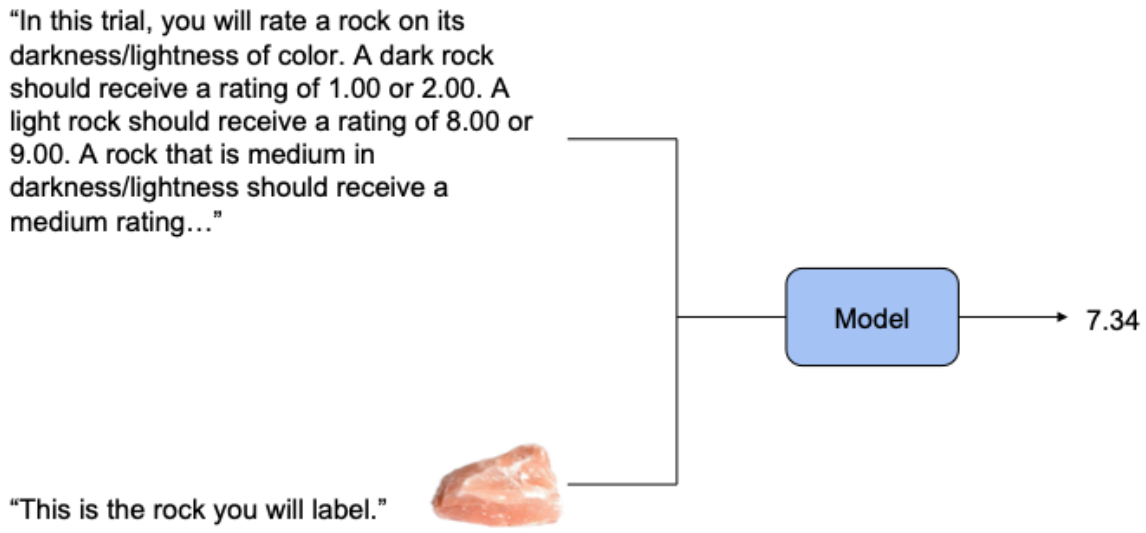

Figure 1. Prompting pipeline for Condition 1.

In this example, the prompt for darkness/lightness is passed to the model along with the rock image to be rated on a scale of 1 to 9. The model produces one numeric value.

[2] Despite the instructions for the model to return numerical dimensions to the hundredths place and to use the full array of available values, the outputs most predominantly appeared in increments of 0.05.

In the second condition, we added three anchor images to each prompt. The anchor images were combined with verbal indicators for the feature being tested (e.g., “This is an example of a dark rock”). Anchor images included examples with low, medium, and high values along the rated dimensions. We used the same anchor images that were presented to the participants in the previous human dimension-rating studies from Nosofsky, Sanders, Meagher, et al. (2018, 2020). An example of this condition with anchor images is provided in Figure 2 below.

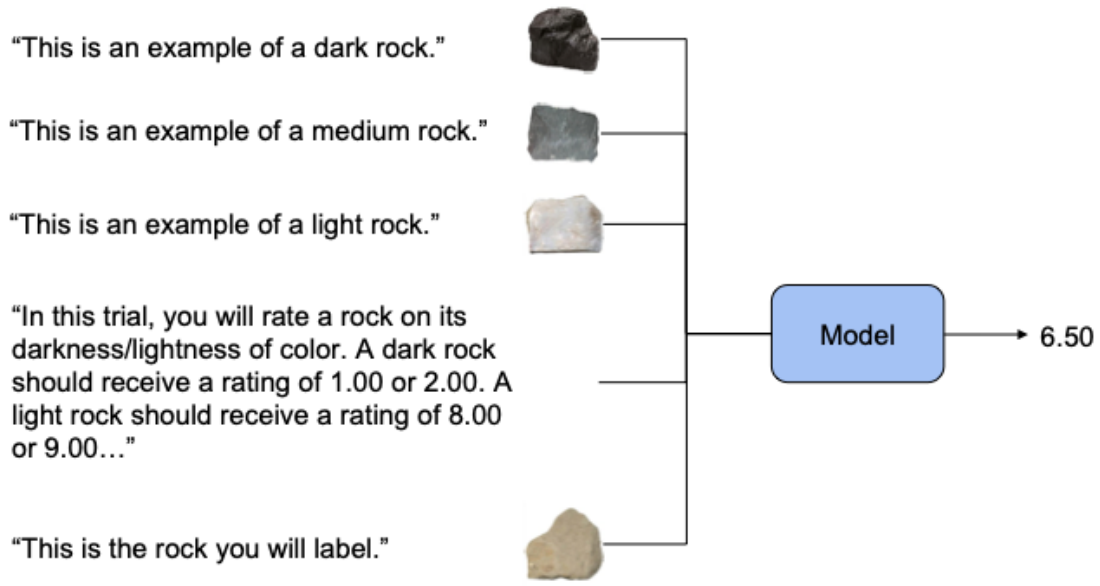

Figure 2. Prompting pipeline for Condition 2.

In this example, 3 anchor images depicting dark, medium, and light rocks are passed to the model with the darkness/lightness feature prompt and rock to be rated. The model again produces one numeric value.

In the third condition, we made several exploratory manipulations in an attempt to improve the performance of the models for two dimensions that had a relatively low correlation with human ratings, specifically organized/disorganized and pegmatitic structure. First, we modified each prompt to include what we thought to be a more precise description of each dimension (e.g., providing more information about what it means for a rock to rank high or low

along the respective dimension).[3] We then conducted this revised evaluation without the anchor images (Condition 3a) and with anchor images (Condition 3b). Second, we provided a set of nine additional anchor images for each of the two dimensions. The anchor images spanned from low to high ratings along the two feature dimensions. The rating for each anchor image provided to the model was computed as the average rating obtained from the four authors independently (Condition 3c). An example of Condition 3c is provided in Figure 3 below.

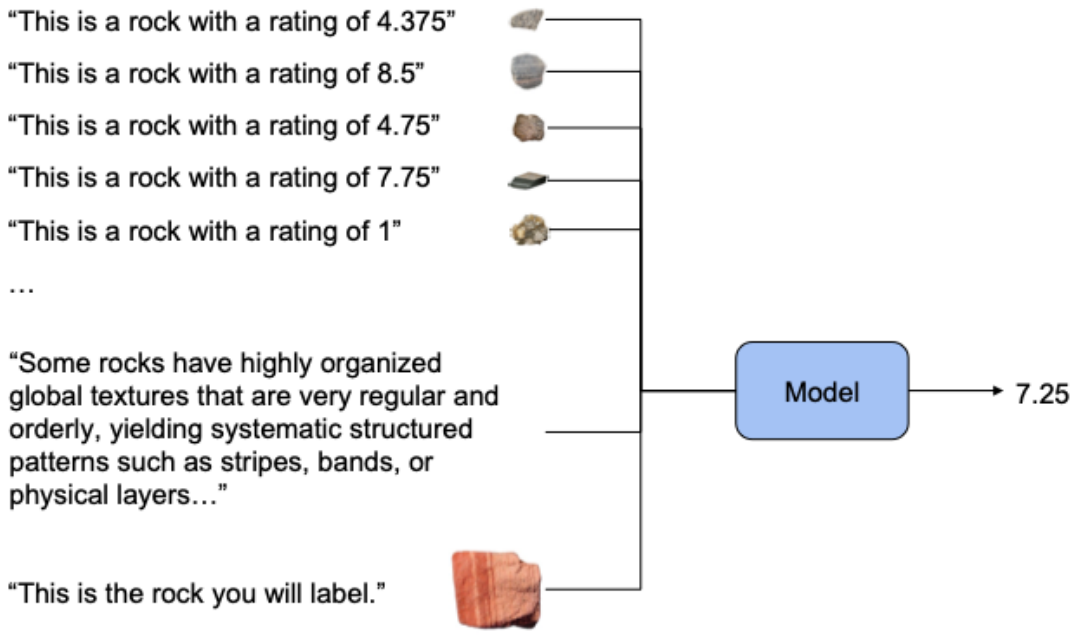

Figure 3. Prompting pipeline for Condition 3c.

In this example, nine anchor images with ratings of average author scores are passed to the model along with a new disorganized/organized feature prompt and the rock to be rated. The model produces one numeric value.

[3] For example, using the original prompt, GPT4 gave high ratings on the organization dimension to rocks with square crystals that were glued together in haphazard fashion, whereas humans had given low ratings to such images. Our modified prompt attempted to emphasize "global organization" rather than whether local components of the rock were well organized. As will be seen, these exploratory modifications had little effect. The wording for the modified prompts is reported in our Supplementary Materials.

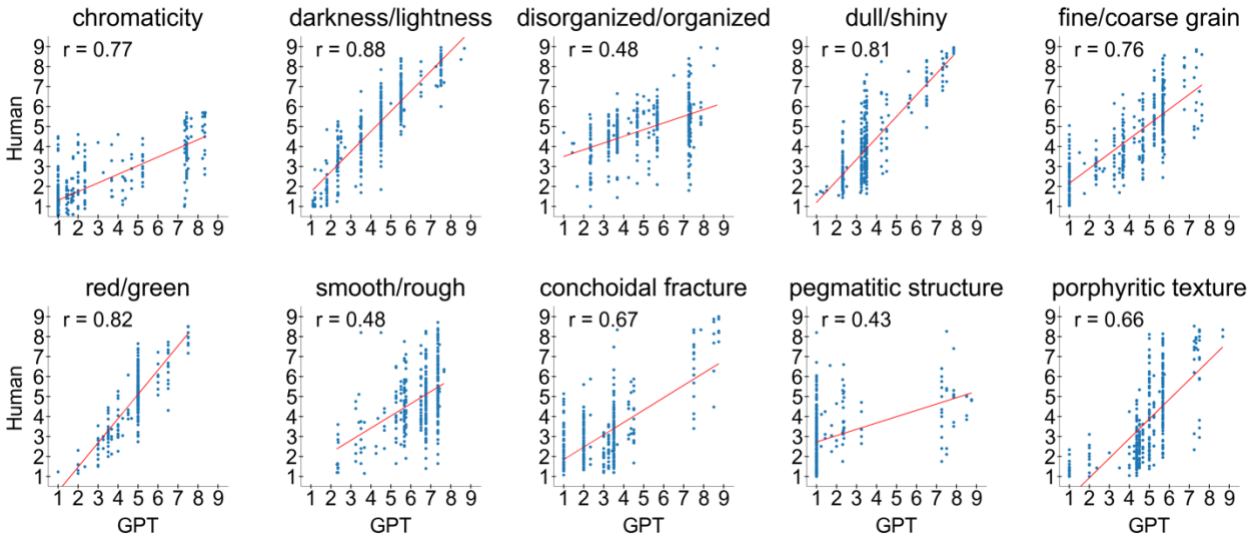

Figure 4. Scatterplots of human and GPT4 ratings in Condition 1 (without anchor images). The Pearson correlation coefficients are shown in the upper left and the regression lines are shown in red.

Results

Preliminary summary of the main pattern of results

Before providing a detailed report of comparisons among models across the conditions, we start by presenting the results from GPT4 in Condition 1. In Condition 1, GPT4 was the best performing model overall. In addition, as will be seen, adding anchor images led to relatively small changes in GPT4's performance relative to the purely verbal prompts used in Condition 1. Claude-3 based models benefited more from adding anchor images, in some cases slightly surpassing the performance of GPT4. The initial report that we provide in this section is intended to give the reader an immediate sense of both the strengths and weaknesses of these multimodal systems in reproducing human judgments along these varied dimensions of the rock images.

We display in Figure 4 scatterplots of GPT4's Condition-1 results for each of the 10 dimensions. Each panel plots the mean observed human ratings for the 360 rock images

against the ratings produced by GPT4. The correlation between the observed and predicted ratings is also reported. As explained later in our article, we have also constructed interactive versions of each scatterplot with more detailed information than shown here. The interactive scatterplots are available at cognlp.com.

As can be seen in Figure 4, there is a good deal of variation in the quality of GPT4's predictions across the different dimensions, with the correlations ranging from r=.43 for pegmatitic structure to r=.88 for darkness/lightness. In general, GPT4 performs best in matching the human ratings for what seem to us to be "elementary" visual dimensions such as darkness/lightness, chromaticity, and red-green hue. The correlations for these elementary dimensions are reasonably high and the scatterplots illustrate that there is appropriate variation across the entire range of ratings. (In other words, the high correlations are not the result of the model predicting only very low ratings for one subset of stimuli and only very high ratings for a second set.) However, GPT4's performance declines for what seem to us to be more abstract and emergent dimensions such as organization and pegmatitic structure. In our Discussion, we consider other possible reasons for the large variation in GPT4's performance across the different dimensions. Interestingly, we will see that humans also show less agreement among themselves in their ratings of the more abstract dimensions compared to the elementary ones.

In testing these multimodal systems, our hope was that their ratings might show sufficient reliability and correspondence with the mean human ratings so as to potentially obviate the need for collecting extensive human ratings in future research. As we will see in our detailed report below, although the systems show promise, this rather ambitious goal was not achieved. Instead, we will see that the best-performing systems provide dimension ratings with reliability that is roughly in the neighborhood of a single subject. Whether or not this degree of reliability is sufficient for one's research goals will likely depend on the specific nature of one's research project.

Detailed evaluation of the LMMs across the three conditions

We first evaluated each of the four models: GPT4 (OpenAI et al., 2023) and Haiku, Sonnet, and Opus (three variants of the Claude-3 model from Anthropic 2024). The results of the evaluation on Condition 1 (which included only verbal prompts) and Condition 2 (which included anchor images) are shown in Table 1.

Table 1. Correlations between multimodal model ratings and averaged human ratings in Conditions 1 and 2 (with and without anchor images). For each dimension, the highest correlation among the four models in both conditions is indicated in boldface font.

dimension/model	GPT4	GPT4 with anchor	Haiku	Haiku with anchor	Sonnet	Sonnet with anchor	Opus	Opus with anchor
chromaticity	0.77	0.80	0.81	**0.82**	0.80	0.77	0.72	0.80
darkness/lightness	0.88	0.85	0.86	0.87	0.84	**0.89**	0.79	0.83
disorganized/organized	0.48	0.42	0.25	0.12	0.45	**0.57**	0.28	0.33
dull/shiny	**0.81**	0.72	0.40	0.48	0.75	0.61	0.33	0.47
fine/coarse grain	0.76	**0.80**	-0.02	0.53	0.61	0.76	0.25	0.46
red/green	0.82	0.78	0.53	0.71	0.67	**0.83**	0.59	0.82
smooth/rough	0.48	0.63	0.04	0.24	0.32	**0.68**	0.19	0.42
conchoidal fracture	0.67	**0.71**	0.30	0.53	0.55	0.70	0.29	0.43
pegmatitic structure	0.43	0.56	0.28	0.32	0.38	**0.66**	0.20	0.57
porphyritic texture	0.66	**0.67**	0.59	0.43	0.59	0.65	0.58	0.55

Overall, Sonnet and GPT4 performed best, with Sonnet scoring highest on five of ten dimensions (darkness/lightness, disorganized/organized, red/green, smooth/rough, and pegmatitic structure), and GPT4 scoring highest on four of ten (dull/shiny, fine/coarse grain, conchoidal fracture, and porphyritic texture). Within the Claude-3 family, the cheapest to use model, Haiku, performed best on chromaticity, and the medium-priced model, Sonnet, outperformed Haiku and Opus on the remaining nine dimensions. Adding anchor images yielded mixed performance improvements across models. For GPT4, performance increased in six of ten features with decreased performance in four. For Haiku, performance increased in eight of ten features with decreases in disorganized/organized and porphyritic texture. For Sonnet, performance improved across all features except dull/shiny. For Opus, performance increased across all features except porphyritic texture. The only top score without using anchor images was GPT4 dull/shiny.

The results of modifying the prompts to include more detailed information and descriptions of dimensions that had a low correlation between human and model data (disorganized/organized and pegmatitic structure) are shown in Tables 2 and 3. Table 2 presents results without using anchor images (Condition 3a). Using the modified prompt without anchor images led to poorer performance across all models.

Table 2. Condition 3a: Modified prompt without anchor images.

dimension/model	GPT4 original prompt	GPT4 longer prompt	Haiku original prompt	Haiku longer prompt	Sonnet original prompt	Sonnet longer prompt	Opus original prompt	Opus longer prompt
disorganized/organized	**0.48**	0.46	**0.25**	0.20	**0.45**	0.44	**0.28**	0.23
pegmatitic structure	**0.43**	0.29	**0.28**	0.25	**0.38**	0.34	**0.20**	0.17

Table 3 presents results using anchor images (Condition 3b). Performance increased for GPT4 in the disorganized/organized dimension but decreased for all other models in all dimensions. In no case was a high correlation achieved.

Table 3. Condition 3b: Longer prompt with anchor images.

dimension/model	GPT4 original prompt	GPT4 longer prompt	Haiku original prompt	Haiku longer prompt	Sonnet original prompt	Sonnet longer prompt	Opus original prompt	Opus longer prompt
disorganized/organized	0.42	**0.47**	**0.12**	0.10	**0.57**	0.55	**0.33**	0.32
pegmatitic structure	**0.56**	0.50	**0.32**	0.25	**0.66**	0.56	**0.57**	0.55

The results of using author-rated anchor images in combination with the longer prompts are shown in Table 4. Including the additional anchor images improved the GPT4 correlation for organized/disorganized from 0.42 to 0.56, but it decreased the correlation for pegmatitic structure from 0.56 to 0.47. Similarly mixed results were obtained for Haiku, and performance decreased in both dimensions for Sonnet. In no case was a very high correlation achieved for either of these complex dimensions. In general, our efforts to improve the performance of the multimodal models on these dimensions by using the modified prompts and more extended set of anchor images were not very successful.

Table 4. Condition 3c: Longer prompt with author-rated images.

dimension/model	GPT4 original prompt	GPT4 longer prompt	Haiku original prompt	Haiku longer prompt	Sonnet original prompt	Sonnet longer prompt	Opus original prompt	Opus longer prompt
disorganized/organized	0.42	**0.56**	**0.12**	0.05	**0.57**	0.55	0.33	**0.47**
pegmatitic structure	**0.56**	0.47	0.32	**0.50**	**0.66**	0.40	**0.57**	0.55

Figure 5 shows for each multimodal model in Conditions 1 and 2 the mean correlation across all the rock dimensions between the ratings of the human subjects and the models. Adding anchor images increases the mean correlation and decreases the standard deviation of the correlations across all models. However, the overall improvement yielded by use of the anchors in the case of GPT4 is rather small.

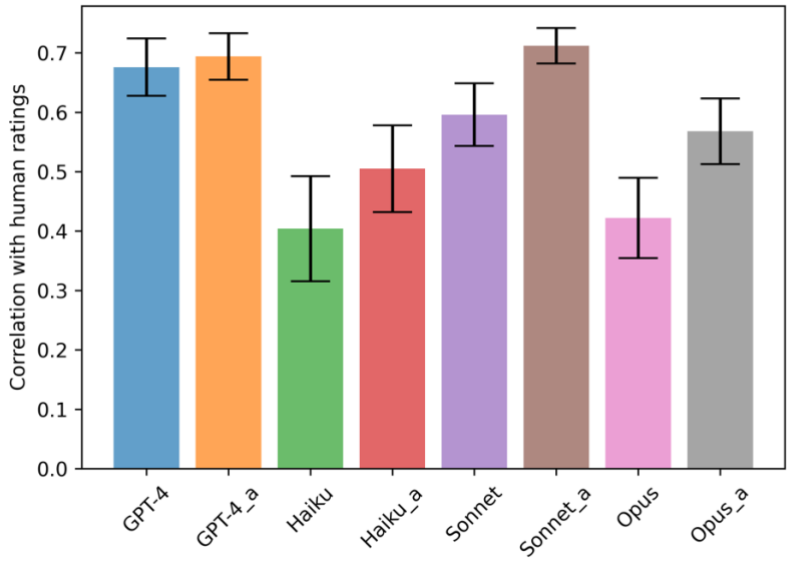

Figure 5. Mean and standard deviation of correlation between the mean rock ratings of human subjects and LMMs in conditions 1 and 2 ('_a' refers to prompts with anchor images).

Variance in human perceptual judgment data

To better understand the performance of the LMMs in relation to the ratings data obtained from human participants, we conducted two analyses to quantify the variance in the human perceptual judgment data. In the first analysis, for each dimension, we computed the correlation coefficient of the 360 ratings between each of the 20 individual subjects and the

mean rating produced by the remaining 19 subjects. For example, we computed the correlation between subject 1's ratings and the mean ratings of subjects 2-20; the correlation between subject 2's ratings and the mean ratings of subjects 1 and 3-20; and so forth. We then computed the mean and standard deviation of each of these 20 individual-subject/mean-rating correlations. The results are given in the first column of Table 5. For ease of comparison, we also re-present the best-performing runs (i.e., with or without use of anchors) for GPT4 and Sonnet. In general, the good-fitting models' correlations are in roughly the same neighborhood as are the correlations of the individual subjects' ratings with the mean ratings.

In a second analysis, we computed split-half correlations for each dimension to quantify reliability at the level of the mean ratings themselves. In this analysis, we conducted 1,000 simulations for each dimension. For each simulation, we divided the 20 subjects who gave ratings into two equally-sized random groups. We then computed the mean rating of each of the 360 rocks on that dimension for each group, and finally computed the correlation between the 360 ratings across the two groups. In Column 2 of Table 5 we report the mean and standard deviation (across the 1000 simulations) of these split-half correlations for each of the dimensions. As can be seen, these split-half correlations are considerably higher than the ones produced by the multimodal models. In sum, the short story is that using the current technology and methods, the multimodal models appear to behave similarly to that of a single participant producing the ratings, but do not provide nearly the same reliability as is obtained when computing the ratings averaged across 20 participants.

Interactive visualization for individual rock scores

For individual rock image results, we have developed an interactive visualization available at cognlp.com where users can view correlation plots for each feature and each condition and hover over each data point to view the specific human and model ratings for that rock. Figure 6 below shows an example of the interactive plots.

Table 5. Correlation in human ratings (first two columns) and correlation between human ratings and best performing LMM runs (columns three to six).

dimension	Mean individual-subject rating correlations +/- SD	Mean split-half group-rating correlations +/- SD	GPT4	Haiku	Sonnet	Opus
chromaticity	0.79 +/- 0.07	0.94 +/- 0.01	0.80	0.82	0.80	0.80
darkness/lightness	0.87 +/- 0.05	0.97 +/- 0.00	0.88	0.87	0.89	0.83
disorganized/organized	0.66 +/- 0.09	0.89 +/- 0.02	0.56	0.25	0.57	0.33
dull/shiny	0.80 +/- 0.06	0.95 +/- 0.01	0.81	0.48	0.75	0.47
fine/coarse grain	0.76 +/- 0.09	0.94 +/- 0.01	0.80	0.53	0.76	0.46
red/green	0.85 +/- 0.03	0.97 +/- 0.00	0.82	0.71	0.83	0.82
smooth/rough	0.74 +/- 0.07	0.93 +/- 0.01	0.63	0.24	0.68	0.42
conchoidal fracture	0.68 +/- 0.10	0.86 +/- 0.02	0.71	0.53	0.70	0.43
pegmatitic structure	0.67 +/- 0.07	0.90 +/- 0.02	0.56	0.32	0.66	0.57
porphyritic texture	0.79 +/- 0.08	0.95 +/- 0.01	0.67	0.59	0.65	0.58

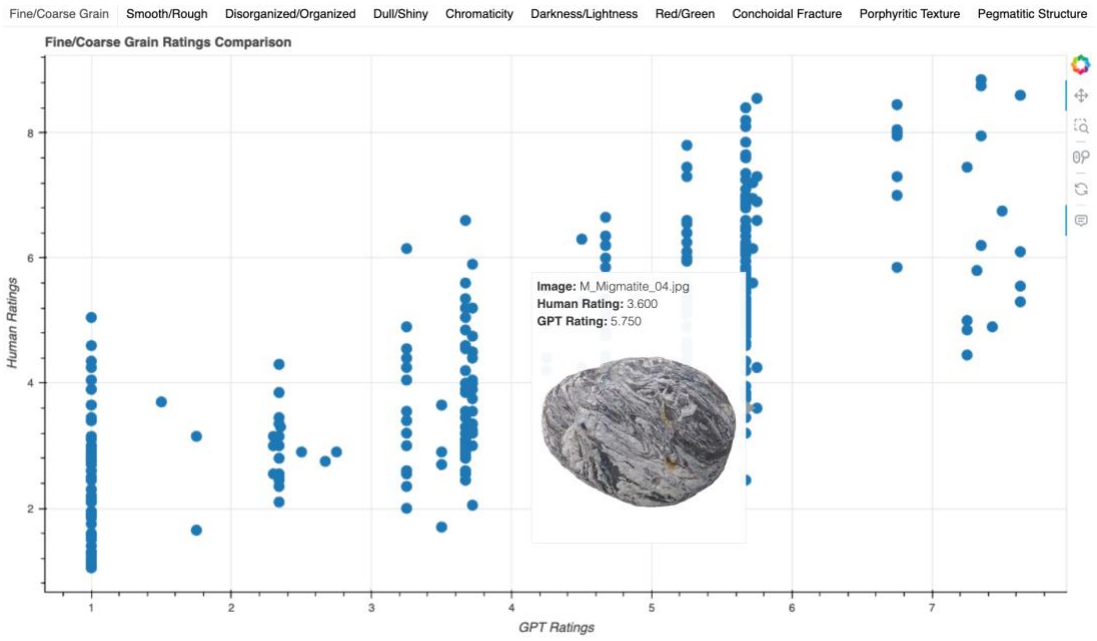

Figure 6. Example of an interactive plot.

Users select a plot for a specific condition and then select a specific feature dimension to view the correlation plot. When a specific datapoint is hovered over, rocks with that combination of scores appear, displaying the name of the specific rock image and the human and model rating for that rock.

Discussion

Recent rapid advancements in large multimodal neural networks led to the development of models that can respond to verbal prompts about image properties, demonstrating zero-shot learning. These models can also take images as a part of the input prompt, enabling one-shot and few-shot learning. It has been demonstrated that model performance can exceed human performance at many tasks, such as image classification, image captioning and visual question answering (He et al., 2015; Li et al., 23--29 Jul 2023; J. Wang et al., 2022; Weihan Wang et al., 2023; Wenhai Wang et al., 2022; Yang et al., 2022). Here we investigated whether the models are capable of mimicking humans in providing ratings of visual perceptual dimensions composing complex objects.

We compared the performance of multimodal vision and language models to human performance in a perceptual judgment task. We used an existing dataset where human participants were asked to make dimension ratings for a series of rocks along different dimensions that were found either through MDS scaling of human similarity judgment data or were deemed important for classifying the rock images into their geologically-defined categories. This dataset has been immensely valuable in cognitive science for understanding human category learning (Cerone et al., 2022; Lu, Penney, & Kang, 2021; Meagher & Nosofsky, 2023; Nosofsky et al., 2020, 2022; Nosofsky, Sanders, Meagher, et al., 2018). Its application with LMMs presented here constitutes a novel approach to evaluating how close the models are to human performance in terms of judging perceptual features.

Our work provides a starting tool that can be used by the research community to quantify the performance of multimodal vision and language models in communicating with human users about perceptual dimensions. Our finding that GPT4 and Sonnet models overall provided clearly better results than Haiku and Opus provides an example of the utility of this benchmark score. Our results indicate, however, that there is room for improvement and that the benchmark is far from saturation.

A specific purpose of the present investigation was to shed light on the possibility of replacing human participants' perceptual judgment collection with ratings provided by LMMs. This line of work has recently attracted a significant amount of attention (Aher et al., 2022; Argyle et al., 2023; Demszky et al., 2023; Dillion et al., 2023; Marjieh et al., 2023). Our study is novel in that it focuses on perceptual judgments of LMMs of a specific psychological dataset composed of visual stimuli and human ratings along a large set of carefully chosen perceptual dimensions. On the one hand, our results indicated that the correlations between the good-performing models' ratings and the mean human ratings were in roughly the same ballpark as the correlation between individual subjects' ratings and the mean human ratings. For certain research purposes, this level of performance may provide a useful starting point, such as for curating stimulus sets for use in pilot studies. However, using our current methods and technologies, the models did not produce correlations with the human data at the same level as measured in terms of split-half correlations between group means. Thus, it seems unlikely that the dimension ratings yielded by the multimodal models would provide the same predictive power of performance in independent tasks as achieved when using actual human mean ratings (Meagher & Nosofsky, 2023; Sanders & Nosofsky, 2020).

Inevitably, the development of LMMs will advance, and we expect their perceptual judgments to get closer to those of humans. However, our results indicate that general model capability is not a good predictor of the correlation. In particular, for Claude-3 models, previous evaluations found that Opus performs best at various benchmarks (Opus is also the most expensive of the three Claude-3 models), but in our experiments its performance was worse than that of Sonnet. However, the performance of Sonnet was better than the performance of Haiku, which is considered to be the least *intelligent* by Anthropic 2024. Unfortunately, because OpenAI and Anthropic did not release the exact number of parameters in their models, we were unable to perform a quantitative analysis comparing the achieved correlations and model size.

In general, the ratings obtained from multimodal models had the highest correlation with

human ratings for visual dimensions that could be considered elementary, such as darkness/lightness, red/green hue, chromaticity and dull/shiny (all with the correlation above 0.8). The mean correlation of individual subject ratings with mean ratings had similar values for these dimensions (Table 5). However, model correlations significantly dropped for more abstract and emergent dimensions such as organization and pegmatitic structure to around 0.5 and for conchoidal fracture, porphyritic texture and smoothness to around 0.7. For these more abstract and emergent dimensions, correlations of individual subject ratings with mean ratings were also lower than for more elementary dimensions. Whereas elementary dimensions like lightness or color hue might be consistent across various domains, making it easier for both humans and models to learn and transfer knowledge, more abstract dimensions might be specific to certain contexts or domains. For example, the texture of a rock might not have a direct analog in other domains that humans and models have been exposed to, leading to more variable performance in those specific dimensions.

Our results demonstrated only marginal improvement when anchor images were provided as a part of the prompt (Condition 2). For instance, the average correlation of GPT4 increased from 0.68+/-0.16 to only 0.69+/-0.13. In addition, in our explorations with two of the challenging dimensions, providing more details about the dimensions with enhanced prompts (Condition 3b) also did not lead to a notable improvement. When providing additional anchor images to GPT4 (Condition 3c) the average correlation between the two dimensions increased from only 0.49 to 0.52. This shows that improving the correlation performance for the abstract, emergent dimensions is rather challenging and that providing more information about these dimensions through enhanced prompts and additional examples does not strongly impact the results.

As technology continues to advance, and the LMMs approach human performance in perceptual judgments, they might have a transformative impact on cognitive modeling. Rigorous tests of cognitive models of memory, learning, and decision-making are often limited to

simplistic stimuli spaces since modeling realistic visual inputs is rather challenging. Using the LMMs to serve as perceptual modules and extract features that a human would extract given the same visual input could enable the scaling of cognitive models to more realistic settings. Advancements in deep neural networks have already been applied to cognitive science, leading to more general models useful in practical applications, such as decision-making in medicine (Hasan et al., 2022; Holmes et al., 2020; Rahgooy et al., 2022) and economics (Horton, 2023; Korinek, 2023). The present research path of having the multimodal models reproduce human perceptual judgments should enhance the applications of formal cognitive models to these real-world settings even more.

Declarations

Funding: No external funding was used.

Conflicts of interest/Competing: All authors report no conflicts of interest.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Authors' contributions: BD, SSM, RN, and ZT contributed to study design and conceptualization, data analysis and interpretation, and drafting and editing the manuscript. BD and SSM contributed to study implementation.

Open Practices Statement

Availability of data and materials: Supplemental material, including all the prompts, is available at https://osf.io/za847/. The complete dataset, including rock images and participant ratings, is available at https://osf.io/cvwu9/. None of the reported studies were preregistered.

Code availability: The code is available at https://github.com/cogneuroai/multimodal-models-rock

References

Aher, G., Arriaga, R., & Kalai, A. (2022). Using large language models to simulate multiple humans. *ArXiv*, *abs/2208.10264*. https://doi.org/10.48550/arXiv.2208.10264

Annis, J., Gauthier, I., & Palmeri, T. J. (2021). Combining convolutional neural networks and cognitive models to predict novel object recognition in humans. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *47*(5), 785–807.

Anthropic, A. I. (2024). The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *31*(3), 337–351.

Bainbridge, W. A. (2019). Chapter One - Memorability: How what we see influences what we remember. In K. D. Federmeier & D. M. Beck (Eds.), *Psychology of Learning and Motivation* (Vol. 70, pp. 1–27). Academic Press.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 1–14.

Cerone, A., Autili, M., Bucaioni, A., Gomes, C., Graziani, P., Palmieri, M., Temperini, M., & Venture, G. (2022). *Software Engineering and Formal Methods. SEFM 2021 Collocated Workshops: CIFMA, CoSim-CPS, OpenCERT, ASYDE, Virtual Event, December 6–10, 2021, Revised Selected Papers*. Springer Nature.

de Kleijn, R. (2022). Artificial Intelligence Versus Biological Intelligence: A Historical Overview. In *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice* (pp. 29-41). The Hague: TMC Asser Press.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L.,

JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *abs/2106.07411*. https://proceedings.neurips.cc/paper/2021/hash/c8877cff22082a16395a57e97232bb6f-Abstract.html

Gemini Team, Reid, M., Savinov, N., Teplyashin, D., Dmitry, Lepikhin, Lillicrap, T., Alayrac, J.-B., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., … Vinyals, O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2403.05530

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., … Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2209.14375

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology. General*, *123*(2), 178–200.

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C., & Trueblood, J. S. (2022). Improving medical image decision-making by leveraging metacognitive processes and representational similarity. *Topics in Cognitive Science*, *14*(2), 400–413.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-

level performance on ImageNet classification. *IEEE International Conference on Computer Vision*, 1026–1034.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Holmes, W. R., O'Daniels, P., & Trueblood, J. S. (2020). A Joint Deep Neural Network and Evidence Accumulation Modeling Approach to Human Decision-Making with Naturalistic Images. *Computational Brain & Behavior*, *3*(1), 1–12.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? In *https://www.nber.org › papershttps://www.nber.org › papers* (No. 31122). National Bureau of Economic Research. https://doi.org/10.3386/w31122

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256–281. https://doi.org/10.1037/a0028860

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2012). Multidimensional scaling: Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93–103. https://doi.org/10.1002/wcs.1203

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, *61*(4), 1281–1317.

Kouwenhoven T, Verhoef T, de Kleijn R, Raaijmakers S. Emerging Grounded Shared Vocabularies Between Human and Machine, Inspired by Human Language Evolution. Front Artif Intell. 2022 Apr 26;5:886349. doi: 10.3389/frai.2022.886349. PMID: 35558168; PMCID: PMC9087278.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lee, M. D. (2001). Determining the Dimensionality of Multidimensional Scaling Representations for Cognitive Modeling. *Journal of Mathematical Psychology*, *45*(1), 149–166.

Li, J., Li, D., Savarese, S., & Hoi, S. (23--29 Jul 2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 19730–19742). PMLR.

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour*, *8*(4), 611-615.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309–332.

Lu, X., Penney, T. B., & Kang, S. H. (2021). Category similarity affects study choices in self regulated learning. *Memory & Cognition*, *49*, 67-82.

Marjieh, R., Jacoby, N., Peterson, J. C., & Griffiths, T. L. (2024). The universal law of generalization holds for naturalistic stimuli. *Journal of Experimental Psychology. General*, *153*(3), 573–589.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). Large language models predict human sensory judgments across six modalities. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2302.01308

Meagher, B. J., & Nosofsky, R. M. (2023). Testing formal cognitive models of classification and old-new recognition in a real-world high-dimensional category domain. *Cognitive Psychology*, *145*, 101596.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology. General*, *115*(1), 39–57.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and

recognition memory. *Journal of Experimental Psychology. Human Perception and Performance*, *17*(1), 3–27.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315.

Nosofsky, R. M., Meagher, B. J., & Kumar, P. (2022). Contrasting exemplar and prototype models in a natural-science category domain. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *48*(12), 1970–1994.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science*, *27*(2), 129–135.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*(2), 530–556.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2020). Search for the Missing Dimensions: Building a Feature-Space Representation for a Natural-Science Category Domain. *Computational Brain & Behavior*, *3*(1), 13–33.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT4 Technical Report. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2303.08774

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A.,

Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (18-24 Jul 2021). Learning Transferable Visual Models From Natural Language Supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.

Rahgooy, T., Venable, K. B., & Trueblood, J. S. (2022). Integrating Machine Learning and Cognitive Modeling of Decision Making. *Computational Theory of Mind for Human-Machine Teams*, 173–193.

Roads, B. D., & Love, B. C. (2021, June). Enriching ImageNet with human similarity judgments and psychological embeddings. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. https://doi.org/10.1109/cvpr46437.2021.00355

Roads, B. D., & Love, B. C. (2024). Modeling Similarity and Psychological Space. *Annual Review of Psychology*, *75*, 215–240.

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*, *51*(5), 2180–2193.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94–101.

Rumelhart, D., & Todd, P. (1993). Learning and Connectionist Representations. In *Attention and Performance XIV*. The MIT Press.

Sanders, C. A. (2024). Personal communication, May 29, 2024

Sanders, C. A., & Nosofsky, R. M. (2018). Using deep learning representations of complex natural stimuli as input to psychological models of classification. *Proceedings of the 2018 Conference of the Cognitive Science Society.*

Sanders, C. A., & Nosofsky, R. M. (2020). Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain. *Computational Brain & Behavior*,

3(3), 229–251.

Shahgir, H. S., Sayeed, K. S., Bhattacharjee, A., Ahmad, W. U., Dong, Y., & Shahriyar, R. (2024). IllusionVQA: A Challenging Optical Illusion Dataset for Vision Language Models. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2403.15952

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1980). Multidimensional Scaling, Tree-Fitting, and Clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Sheybani, S., Smith, L. B., Tiganj, Z., Maini, S. S., & Dendukuri, A. (2024). ModelVsBaby: A developmentally motivated benchmark of out-of-distribution object recognition. In *PsyArXiv*. https://doi.org/10.31234/osf.io/83gae

Steyvers, M., & Busey, T. (2000). Predicting Similarity Ratings to faces using physical descriptions. *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, 115–146.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. Journal of Memory and Language, 42(1), 51–73. https://doi.org/10.1006/jmla.1999.2669

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., & Wang, L. (2022). GIT: A Generative Image-to-text Transformer for Vision and Language. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2205.14100

Wang, Weihan, Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., & Tang, J. (2023). CogVLM: Visual Expert for

Pretrained Language Models. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2311.03079

Wang, Wenhai, Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X.-H., Lu, T., Lu, L., Li, H., Wang, X., & Qiao, Y. (2022). InternImage: Exploring large-scale vision foundation models with deformable convolutions. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 14408–14419.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Xia, F., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *abs/2201.11903*. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Yang, J., Li, C., & Gao, J. (2022). Focal Modulation Networks. *Advances in Neural Information Processing Systems*, *abs/2203.11926*. https://doi.org/10.48550/arXiv.2203.11926