Analyzing sensory judgments made by people and machines

Alexander Ku

Department of Psychology

Princeton University

Author Note

Correspondence concerning this article should be addressed to Alexander Ku, Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, NJ 08540. E-mail: alexku@princeton.edu

Abstract

This report reproduces some of the statistical analysis performed by Marjieh et al., 2023, which shows that sensory judgments from a large language model are highly predictive of human sensory judgments across multiple modalities. Furthermore, they also find no stronger correlation for visual judgments over non-visual judgments, despite the language model they use (GPT-4) being trained on both visual and linguistic data. Note that I've slightly modified the analysis as to better fit the content of this class.

Keywords: APA style, statistics, reproduction, similarity, perception

Analyzing sensory judgments made by people and machines

Method

We begin with a concise exposition of the hypothesis and data collection procedure. A thorough methodological description can be found in the original paper. The objective of this section is parsimony: to provide minimally sufficient information to understand the statistical analysis being performed.

Hypotheses

A key question of interest to cognitive scientists and philosophers alike is the extent to which linguistic description can express and convey sensory experience. The authors posit a philosophical thought experiment: can an alien anthropologist (oxymoron?) infer the idiosyncrasies of an alien species' perceptual experience from the language they speak? For example, a species that can see infrared may have words to describe colors that are indistinguishable to people. In the setting we are interested in, the large language model is the anthropologist, and we are the alien species.

Thus, the hypothesis we will be testing is as follows: are large language models predictive of human sensory experience? Concretely, are perceptual similarity judgments from a large language model, solicited via linguistic description, correlated with human similarity judgments, solicited via sensory stimuli.

Marjieh et al., 2023 test this hypothesis for both visual and non-visual stimuli. They find that the similarity judgments from a large language model are significantly and strongly correlated with human similarity judgments. Furthermore, they show that, despite the large language model they use (GPT-4) being trained on both visual an linguistic data, there is no significant difference between the correlation in similarity judgments for visual and non-visual stimuli.

In order to test these hypotheses the authors must have aligned human and model data. That is, human and model similarity judgments for the same stimulus pair. The authors use both preexisting data as well as collect their own using an online crowd-sourcing platform (Amazon Mechanical Turk).

Collecting human data

The authors consider 6 sensory dimension (color, loudness, timbre, taste, pitch, and vocal consonants) across 3 sensory modalities (visual, auditory, gustatory). Human similarity judgments for color, loudness, timbre, taste are pubicly available from previous studies. These data were collected using both online crowd-sourcing platforms as well as in the lab. The authors also collected additional similarity judgments for pitch and vocal consonants via online crowd-sourcing platform. Their detailed methodology can be found in the original manuscript.

These data contain pairwise similarity judgments for n stimuli (per sensory dimension). This yields an $n \times n$ similarity matrix. Because similarity is a subjective measure, each entry in the matrix is averaged over 5 subjects. Entries of this matrix are normalized such that similarity is between 0 (‘completely dissimilar’) and 1 (‘completely similar’). See Figure 1 for a visualization.

Collecting machine data

To collect aligned similarity judgments from a large language model (GPT-4), the authors must translate the sensory stimuli used by the human data into linguistic form. For example, color can be encoded as a hexadecimal code. Similarity judgments are solicited using in-context learning. Here is an example of a prompt the model sees for color:

```
People described pairs of colors using their hex
codes. How similar are the two colors in each
pair on a scale of 0-1 where 0 is completely
dissimilar and 1 is completely similar? Respond
only with the numerical similarity rating.
Color 1: #ff5700 Color 2: #ff9b00 Rating 0.76
Color 1: #b3ff00 Color 2: #00ff61 Rating: 0.45
Color 1: #FF0000 Color 2: #00b2ff Rating: 0.02
Color 1: <hex-code1> Color 2: <hex-code2>
Rating:
```

As with people, model similarity judgments for each stimulus pair are collected multiple times. This is because large language models are stochastic—they decode text using a method called temperature sampling. The temperature indicates how much variability you have in your decoded samples. In theory, you can make decoding deterministic by setting the temperature to 0 or some epsilon value (i.e., greedy decoding), however this often degrades the quality of the samples. Thus, the authors decide to use the default temperature parameter and solicit 10 similarity judgments per stimulus pair, and average them to construct the similarity matrices. See Figure 1 for a visualization.

Do machine judgments predict human judgments?

Now that we understand the type of data we are working with (i.e., pairwise similarity judgments across sensory dimensions), we will shift our focus to the first of the author's questions: do machine judgments predict human judgments?

Before we begin with the analysis, I will describe some preprocessing we do to the data. While similarity need not be commutative (as Tversky famously showed), for the stimuli we use, it is. Thus the resulting similarity matrices are also symmetric. For this reason, we will deduplicate (a.k.a., dedupe) our data. That is, when computing statistical measures, such as correlation, we will only use the upper triangular entries of the similarity matrix. Doing so leaves us with a dataframe containing 650 rows. Each row contains the index of the stimulus pair, the human similarity judgment for that pair, the machine similarity judgment for that pair, and the sensory dimension.

First, let's visualize the data using a scatter plot. As we can see in Figure 2 there seems to be a relationship between human similarity and machine similarity. Running a correlation test tells us that there is a significant and strong positive correlation between the two variables, r = 0.79, 95% CI [0.76, 0.82], t(64) = 33.08, p < .001. That is to say, when human similarity increase so does machine similarity, and vice versa.

Now that we have established that the two variables are correlated, let's fit a linear regression. We will let machine similarity be the independent variable and human similarity be the dependent variable—that is we want to see if machine similarity is predictive of human similarity. The slope of the regression line is 0.759. This means that for every unit increase in machine similarity, there is a 0.759 increase in human similarity. However, because similarity is bounded between 0 and 1, this interpretation does not make sense. It makes more sense to say: for every 0.1 increase in machine similarity, human similarity increases by 0.0759. The intercept of the regression line is 0.0203. This means that when machine similarity is 0, the mean human similarity is 0.0203. The model explains a large and significant portion of the variation in the data (R2 = 0.628, F(1, 64) = 1095., p < .001, adj. R2 = 0.628). F-test tells us that it is significantly better than the empty model. The R^2 value tells us that 62.7% of the variance in human similarity is explained by model similarity. The adjusted R^2 is the same as the regular R^2, since we only have one predictor. Figure 3 plots the regression line with 95% confidence intervals. Figure 4 tells us residuals of this model are normally distributed and homoskedastic.[1]

Does the effect differ between visual and non-visual stimuli?

The large language model the authors use (GPT-4) is trained on both visual an linguistic data, yet they find there is no significant difference in the relationship between human similarity and machine similarity for visual and non-visual stimuli. Let's perform a moderation analysis to determine whether visual vs. non-visual is a moderator. To facilitate this analysis we add a factor to our dataframe indicating whether the similarity is being measured for visual or non-visual stimuli. First, we will plot regression lines to visualize the interaction. As we can see in in Figure 5 there does not seem to be an

[1] The residuals are slightly misbehaved around 0 and 1 (i.e., where similarity saturates). Thus, it may be worth using a transformation (e.g., logit, probit), however, when I tried, the residuals were even more misbehaved. For now, we will proceed assuming the linear model is appropriate.

interaction.

Let’s fit a multiple regression model where the dependent variable is human similarity and the independent variables are mean centered machine similarity, dummy coded visual factor, and an interaction term (i.e., product of machine similarity and visual factor). This results in a model that explains a statistically significant and substantial proportion of variance (R2 = 0.65, F(3, 646) = 408.60, p < .001, adj. R2 = 0.65).

The model’s intercept is 0.40 (95% CI [0.39, 0.41], t(646) = 65.34, p < .001) and indicates the predicted human similarity when machine similarity is at its mean and the stimuli are non-visual. The effect of machine similarity is statistically significant and positive (beta = 0.73, 95% CI [0.69, 0.78], t(646) = 29.98, p < .001; Std. beta = 0.77, 95% CI [0.72, 0.81]); every 0.1 increase in machine similarity results in a 0.073 increase in human similarity when the stimuli are non-visual. The effect of the visual factor is statistically significant and negative (beta = -0.11, 95% CI [-0.15, -0.08], t(646) = -6.75, p < .001; Std. beta = -0.16, 95% CI [-0.21, -0.12]); the intercept is 0.11 lower (i.e., the predicted change in human similarity) for visual stimuli when machine similarity is at its mean—this offset can be seen in Figure 5. Finally, the interaction effect is not statistically significant whatsoever (beta = 0.02, 95% CI [-0.10, 0.14], t(646) = 0.34, p = 0.733; Std. beta = 7.45e-03, 95% CI [-0.04, 0.05]); consistent with what the authors report. Finally, we’ll calculate power. The smallest effect size we can observe with 99% power is 0.02—which means we can observe small effects given the sample size (n = 650).

Discussion

To summarize, our analysis is consistent with claims of the authors: (1) machine similarity judgments are highly predictive of human similarity judgments; and (2) whether the similarity is measured for visual or non-visual stimuli does not change this interaction, despite the model being trained on visual data.

Finally, I would like to end with a discussion of future directions this line research may pursue; as it is related to my own interests.

As cognitive psychologists, we infer mental processes and representations from behavior, often assuming that the behavior people exhibit is the (resource-)rational solution to a computational task. If the behavioral pattern of people and machines are aligned, and consistent with a mathematical theory (e.g., Bayesianism, connectionism), it is evidence that both systems have learned a similar solution and or internal representation.

Rather than inferring representations from behavior, cognitive neuroscientists look inside the system itself. While brain imaging techniques (e.g., fMRI, EEG) have spatial or temporal resolution constraints, observing the numerical processing that happens inside of a large language model (a deep neural network) does not.

Many theories in cognitive psychology posit that stimuli are represented as points in a psychological metric space (e.g., Euclidean, hyperbolic). By studying the internal activations of a large language model, we can explicitly assess the embedding geometry of each layer of the network. And, like cognitive neuroscientists, we can develop mechanistic theories of perceptual semanticity that are consistent with behavioral ones, without the resolutional limitations of brain imaging.

References

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*.

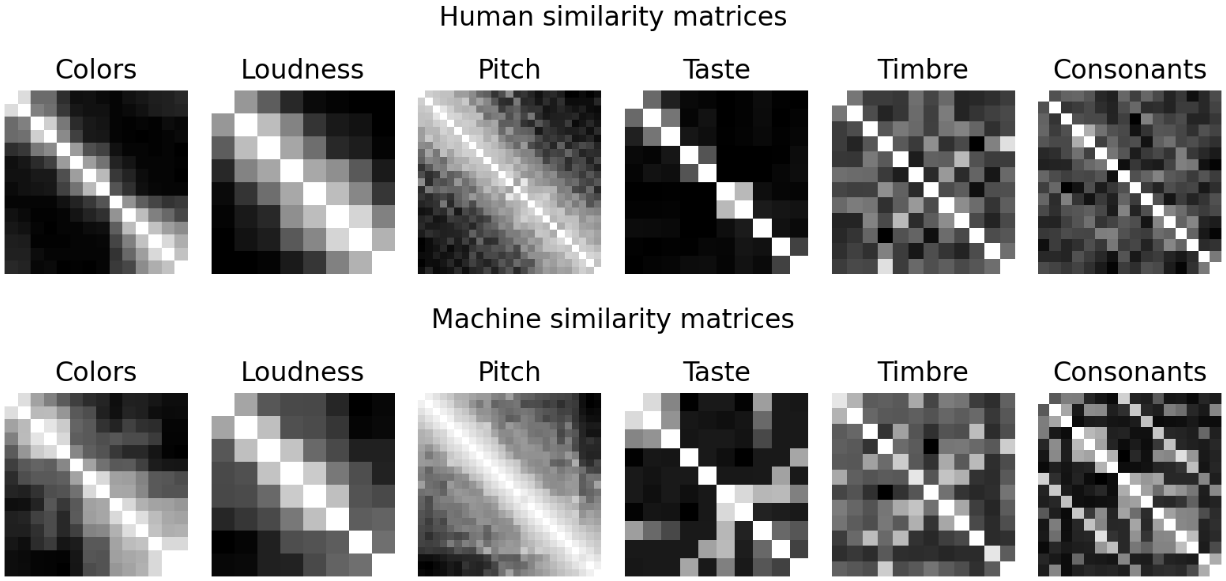

Figure 1

Human and machine similarity matrices for each sensory dimension. Light pixels indicate high similarity, dark pixels indicate low similarity.

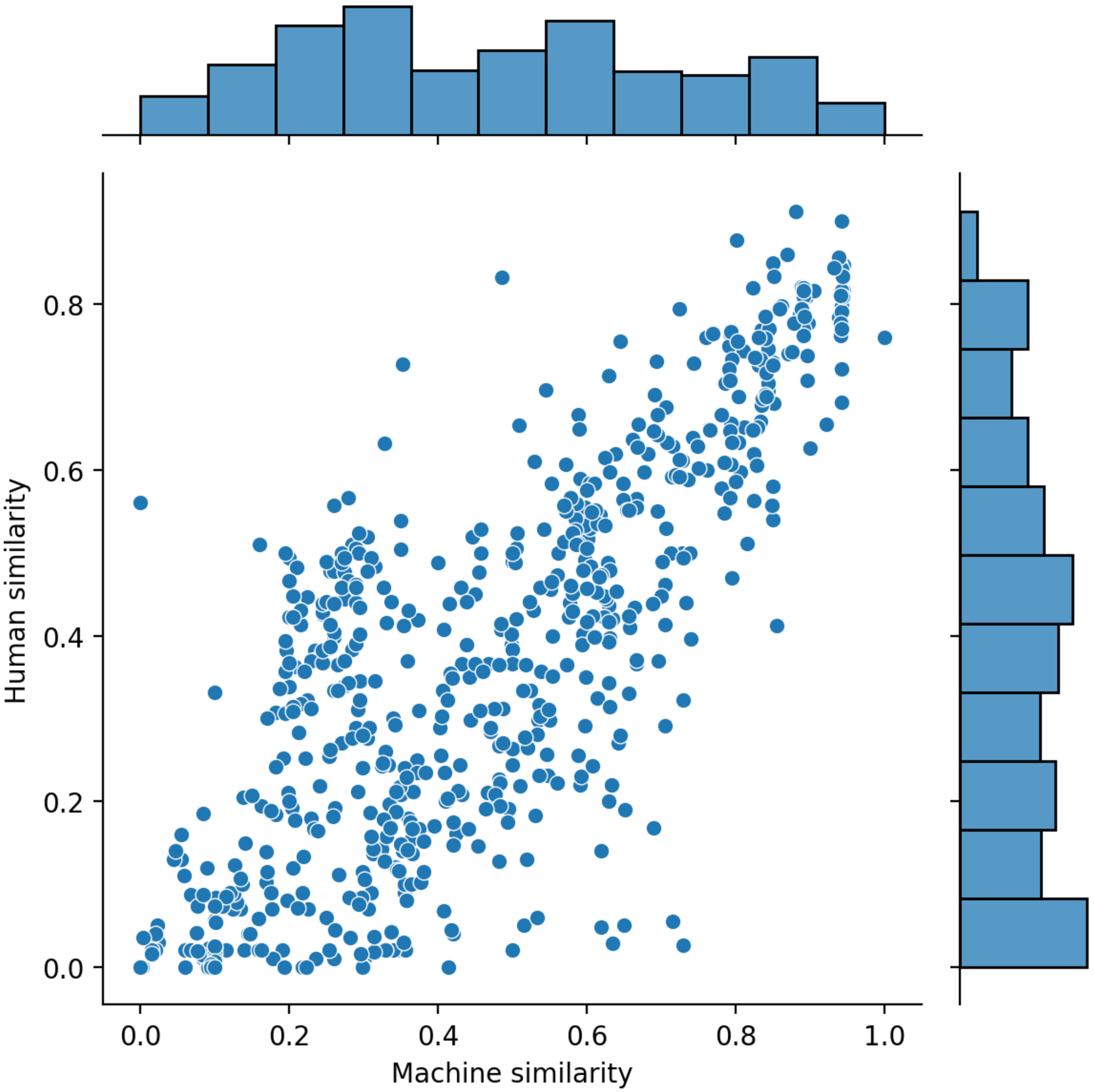

Figure 2

Scatter plot of human similarity judgments against machine similarity judgments. Marginal densities are visualized on the top and right as histograms.

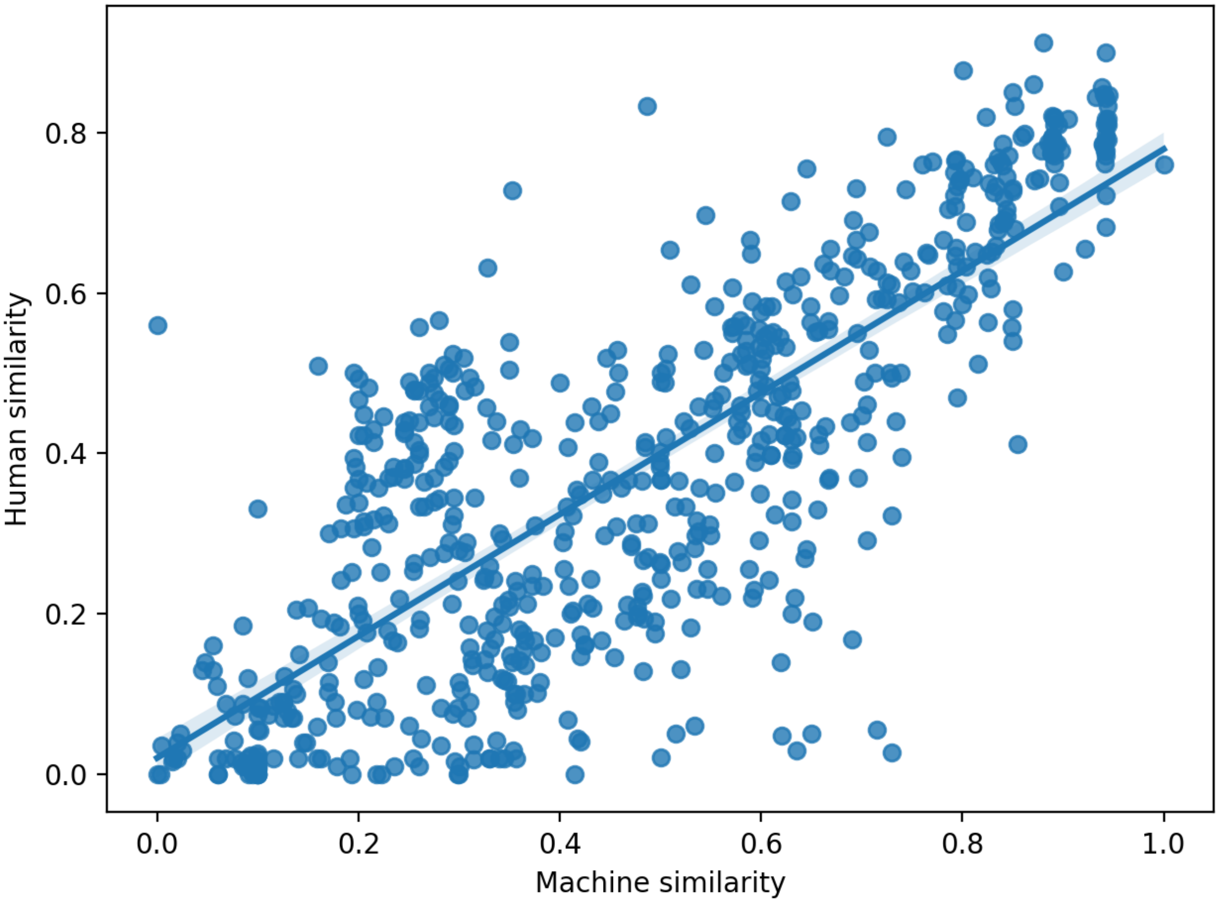

Figure 3

Linear regression line with 95% confidence intervals.

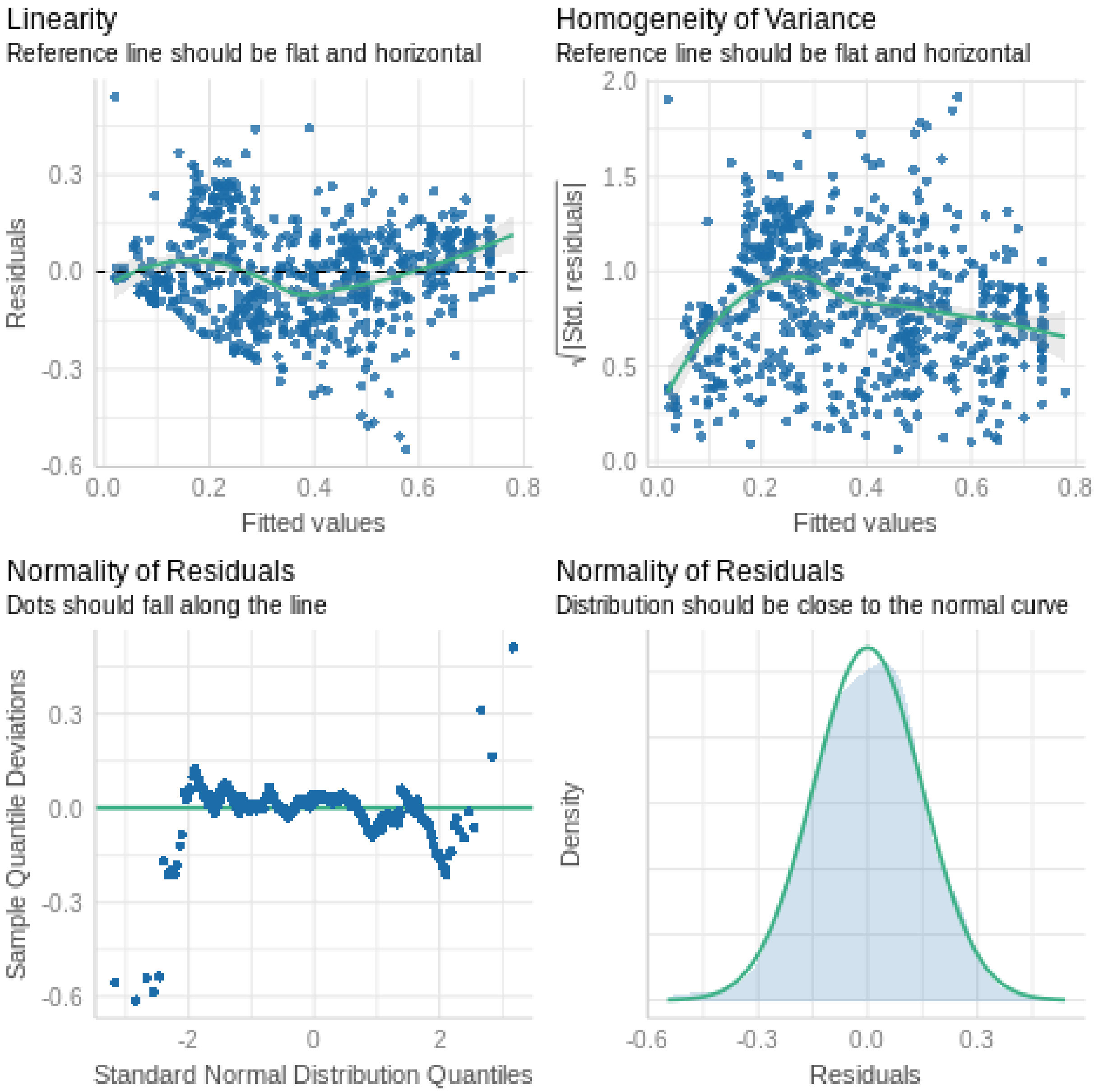

Figure 4

Residual checks for linear regression.

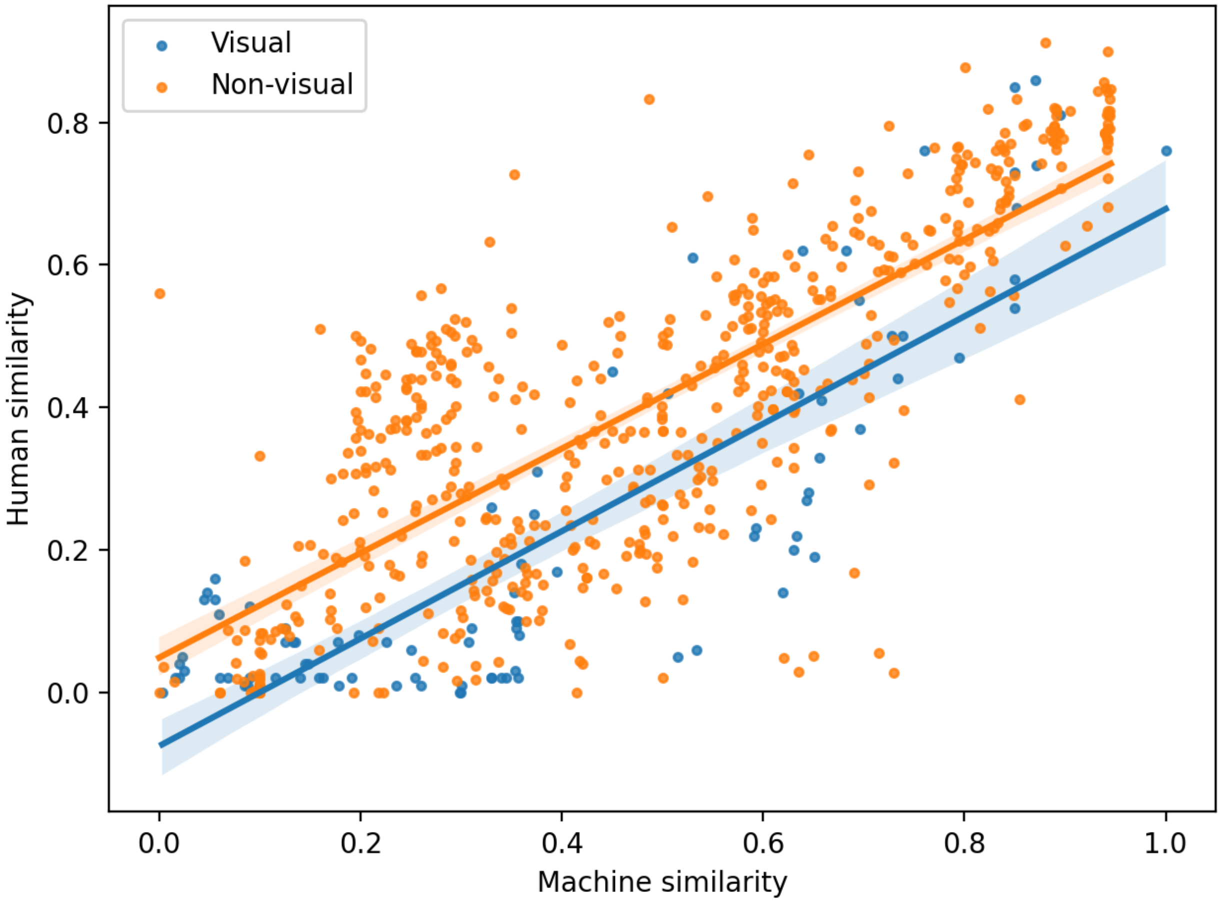

Figure 5

Linear regression lines for similarities of visual and non-visual stimuli.