

---

# Quantifying Uncertainty in Error Consistency: Towards Reliable Behavioral Comparison of Classifiers

---

Thomas Klein<sup>1,2,3,4\*</sup>

Sascha Meyen<sup>2</sup>

Wieland Brendel<sup>†1,2,3,4</sup>

Felix A. Wichmann<sup>‡2</sup>

Kristof Meding<sup>‡2,5</sup>

## Abstract

Benchmarking models is a key factor for the rapid progress in machine learning (ML) research. Thus, further progress depends on improving benchmarking metrics. A standard metric to measure the behavioral alignment between ML models and human observers is *error consistency* (EC). EC allows for more fine-grained comparisons of behavior than other metrics such as accuracy, and has been used in the influential Brain-Score benchmark to rank different DNNs by their behavioral consistency with humans. Previously, EC values have been reported without confidence intervals. However, empirically measured EC values are typically noisy—thus, without confidence intervals, valid benchmarking conclusions are problematic. Here we improve on standard EC in two ways: First, we show how to obtain confidence intervals for EC using a bootstrapping technique, allowing us to derive significance tests for EC. Second, we propose a new computational model relating the EC between two classifiers to the implicit probability that one of them copies responses from the other. This view of EC allows us to give practical guidance to scientists regarding the number of trials required for sufficiently powerful, conclusive experiments. Finally, we use our methodology to revisit popular NeuroAI-results. We find that while the general trend of behavioral differences between humans and machines holds up to scrutiny, many reported differences between deep vision models are statistically insignificant. Our methodology enables researchers to design adequately powered experiments that can reliably detect behavioral differences between models, providing a foundation for more rigorous benchmarking of behavioral alignment.

## 1 Introduction

Consider the following problem: *Given two classifiers operating on the same domain, how should we quantify their similarity?* This abstract problem has become highly practically relevant in the context of cognitive science and beyond. In cognitive science, deep neural networks (DNNs) have been proposed as models of the human visual system [Doerig et al., 2023, Kriegeskorte, 2015, Cichy and Kaiser, 2019, Kietzmann et al., 2017, Yamins et al., 2014], and thus the question arises how similarly they behave to human observers. In other domains, e.g. law or medicine, it is also crucial to know whether DNNs interpret and judge information similarly to humans. Thus, a metric is needed that can reliably quantify the degree of behavioral similarity between DNNs and human decision makers.

The standard method proposed for this purpose is *error consistency* (EC) [Geirhos et al., 2020], which has seen wide application in the context of human-machine comparisons, e.g. in Geirhos et al. [2021],

\*<sup>1</sup> Max Planck Institute for Intelligent Systems, Tübingen, Germany <sup>2</sup> University of Tübingen, <sup>3</sup> Tübingen AI Center <sup>4</sup> ELLIS Institute Tübingen. <sup>5</sup> CSZ Institute for Artificial Intelligence and Law. <sup>†</sup> Joint supervision. Correspondence to: t.klein@uni-tuebingen.de. Code: [github.com/wichmann-lab/error\\_consistency](https://github.com/wichmann-lab/error_consistency).

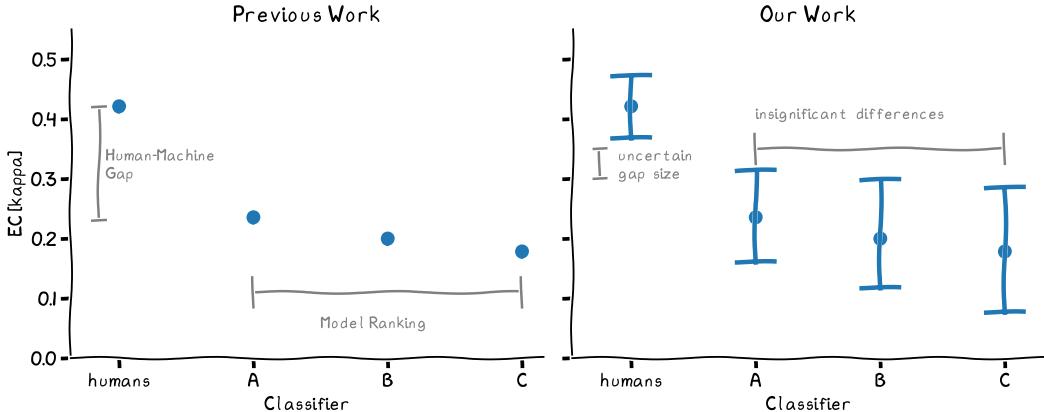


Figure 1: **Schematic: Error Consistency is noisy.** The gist of our findings is that by calculating confidence intervals for empirical EC values, we show that while the human-machine gap is probably real, the measure is not stable enough to resolve differences between models.

Ollikka et al. [2024], Li et al. [2025], Parthasarathy et al. [2023]. Similarity metrics based on EC have also seen application in the behavioral similarity section of the vision-branch of Brain-Score [Schrimpf et al., 2018]. More recently, a variation of EC—sharing many of the relevant properties, so that our considerations apply—has been proposed as a metric to compare Large Language Models (LLMs) [Goel et al., 2025].

EC builds on Cohen’s  $\kappa$  [Cohen, 1960], which considers the trial-level agreement observed for a pair of classifiers to compute a scalar similarity score, which is 1 if the two models are behaviorally indistinguishable and  $-1$  if they are as different as they could possibly be. A value of 0 indicates independence. In practice, EC is reported as a point estimate of a unit-less scalar value, based on classification decisions obtained on a single set of test stimuli, like in Figure 2a. However, since EC relies on trial-by-trial similarity, the measure is inherently noisy, as individual trials can have a large impact on the final score, especially when the total number of trials is small—as is often the case when conducting experiments with humans.

Our contributions are as follows:

- We explain how confidence intervals for empirical EC values can be estimated on any dataset on which EC may be calculated using bootstrapping [Efron and Tibshirani, 1994].
- We propose a novel method of modeling classifier consistency, allowing us to simulate classifiers with a specific consistency and thus to plan sufficiently powerful, conclusive experiments.
- We apply our methods to two influential benchmarks (Model-vs-Human [Geirhos et al., 2021] and Brain-Score [Schrimpf et al., 2018]), showing that the currently available amount of human reference data is not sufficient to draw conclusions about model differences in many cases, as illustrated in Figure 1.
- We provide a python package containing utility functions for obtaining confidence intervals around error consistency values, calculating p-values, planning experiments, and testing results for significance.

We will begin by providing a brief introduction to EC in Section 2, to then explain how confidence intervals can be calculated in Section 3. A detailed Related Work section is provided in Appendix A since we discuss related work on measuring behavioral alignment, Error Consistency, and benchmarking.

## 2 Error Consistency

First, we provide a brief introduction to error consistency and introduce our notation; see Geirhos et al. [2020] for an in-depth description. EC is measured by applying Cohen’s  $\kappa$  to evaluate whether two classifiers’ responses are jointly correct or incorrect—that is, whether they are consistent about when they make errors. The experimental setting in which EC can be calculated is the following: Two classifiers categorize each of  $N$  samples (e.g., natural images) into one of  $K$  classes (e.g., the 1,000 ImageNet classes). Each sample,  $x_i$ , has a ground-truth label  $y_i \in \{1, \dots, K\}$  and receives classifications,  $\hat{y}_i^{(j)} \in \{1, \dots, K\}$ , with  $j \in \{1, 2\}$ . Whether these classifications are correct or not is coded as  $r_i^{(j)} \in \{0, 1\}$  where  $r_i^{(j)} = 1$  if and only if the  $j$ -th classifier responded with the true label,  $\hat{y}_i^{(j)} = y_i$ . This renders EC agnostic to the number of classes of the classification task because it only considers agreement on which samples are jointly classified correctly and incorrectly.

In this setting, EC is measuring the agreement between  $r^{(1)}$  and  $r^{(2)}$  using Cohen’s  $\kappa$  (which is typically applied to  $\hat{y}^{(1)}$  and  $\hat{y}^{(2)}$ ):

$$\kappa = \frac{p_{obs} - p_{exp}}{1 - p_{exp}} \quad (1)$$

where  $p_{obs}$  is the observed agreement, that is, the proportion of trials which are jointly classified correctly or incorrectly. This quantity is normalized by the amount of agreement expected from independent classifiers,  $p_{exp}$ . This normalization accounts for the fact that classifiers with high accuracies are a priori expected to agree more often in their classifications than classifiers with low accuracies.

However, the normalization does not render EC fully orthogonal to the accuracies of the classifiers, because Cohen’s  $\kappa$  depends on the marginal distributions of classifier responses [Falotico and Quatto, 2014, Grant et al., 2016]. This is easy to see in the limiting case where two classifiers have perfect accuracies ( $p_1 = p_2 = 1$ ): Division by zero renders EC undefined because, intuitively, error consistency can not be evaluated in the absence of errors. Moreover, if exactly one classifier has perfect accuracy, EC will be 0 *irrespective of the accuracy of the other classifier*. This phenomenon, which we prove in Appendix C, may be counterintuitive and can lead to estimation instabilities, see below.

EC is bounded by the mismatch of the two classifiers’ accuracies [Geirhos et al., 2020]. We provide an instructive visualization of these asymmetric bounds in Figure 2a. Only when the accuracies of both classifiers are equal can all values in  $[-1, 1]$  be achieved. This implies that EC can never be high between two classifiers if their accuracies are substantially different. In practice, one can therefore not take EC at face value: EC has to be interpreted with the accuracies of the two classifiers in mind, and is, unfortunately, strongly susceptible to both floor- and ceiling-effects. The precondition for a straightforward interpretation of EC is thus a good match of classifiers’ accuracies.

However, previous work has evaluated EC in conditions in which accuracies of the two classifiers were vastly different, see Figure 2b. Even without considering their joint probabilities, EC between classifiers must be low in these cases because their accuracy mismatch imposes an upper bound on the observable EC values.

To formalize this dependence of EC on classifiers’ accuracy match, we present two propositions. First, when the accuracies of the two classifiers are equal, EC can best be understood as the proportion of copied responses: In a model in which the second classifier copies the responses of the first with probability  $p_{copy}$  and responds independently at random otherwise, EC is related to that copy probability. In Proposition 1, we present this result as a general property of Cohen’s  $\kappa$  (for relating  $\hat{y}^{(1)}$  and  $\hat{y}^{(2)}$ ) which immediately translates to the application as EC (relating  $r^{(1)}$  and  $r^{(2)}$ ), where equal marginal distributions correspond to equal accuracies.

**Proposition 1.** *Let a classifier give responses according to some marginal distribution  $\hat{y}^{(1)} \sim \tilde{P}^{(1)}$ . If another classifier copies that response  $\hat{y}^{(2)} = \hat{y}^{(1)}$  with probability  $p_{copy}$  and otherwise draws independently from a marginal distribution  $\hat{y}^{(2)} \sim \tilde{P}^{(2)}$  that is the same as the first  $\tilde{P}^{(1)} = \tilde{P}^{(2)}$ , then*

$$\text{Cohen's } \kappa = p_{copy}.$$

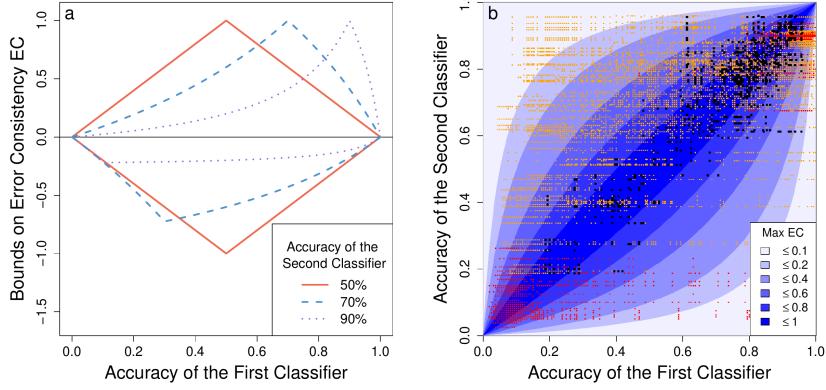


Figure 2: **Theoretical issues with EC.** (a) The bounds on error consistency (EC) depend on the mismatch between the accuracies of the two compared classifiers. (b) Similar to (a) but with accuracies on the x and y axes. Here, each orange dot corresponds to one Model-vs-Human comparison and each black dot to one human-vs-human comparison in the analysis of Geirhos et al. [2021]. Red dots correspond to conditions that were sampled but excluded for the analysis by the original authors. Human-vs-human comparisons have lower accuracy mismatches, so EC values are expected to be higher a priori.

*Proof.* The underlying distributions translate into the observed distributions  $P^{(1)} = \tilde{P}^{(1)}$  and  $P^{(2)} = p_{copy}\tilde{P}^{(1)} + (1 - p_{copy})\tilde{P}^{(2)}$ . With  $\tilde{P}^{(1)} = \tilde{P}^{(2)}$  we have  $P^{(1)} = P^{(2)} = \tilde{P}^{(1)} = \tilde{P}^{(2)}$ .

For the joint distribution, we have  $P^{(1,2)}(y, y) = \tilde{P}^{(1)}(y)(p_{copy} + (1 - p_{copy})\tilde{P}^{(2)}(y))$  because the second classifier makes the same prediction as the first not only in all copy cases ( $p_{copy}$ ), but also in those cases in which it does not copy but coincidentally makes the same prediction  $((1 - p_{copy})P^{(2)}(y))$ . With that,  $\kappa$  simplifies to  $p_{copy}$ .

$$\begin{aligned}
\kappa &= \frac{p_{obs} - p_{exp}}{1 - p_{exp}} \\
&= \frac{\sum_{y=1}^K P^{(1,2)}(y, y) - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= \frac{\sum_{y=1}^K \left( P^{(1)}(y)(p_{copy} + (1 - p_{copy})\tilde{P}^{(2)}(y)) - P^{(1)}(y)P^{(2)}(y) \right)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= \frac{\sum_{y=1}^K \left( P^{(1)}(y)(p_{copy} + (1 - p_{copy})P^{(1)}(y)) - P^{(1)}(y)P^{(2)}(y) \right)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= \frac{\sum_{y=1}^K \left( P^{(1)}(y)p_{copy} + P^{(1)}(y)P^{(1)}(y) - P^{(1)}(y)P^{(1)}(y)p_{copy} - P^{(1)}(y)P^{(1)}(y) \right)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(1)}(y)} \\
&= p_{copy} \frac{\sum_{y=1}^K P^{(1)}(y) - \sum_{y=1}^K P^{(1)}(y)P^{(1)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(1)}(y)} \\
&= p_{copy} \frac{1 - \sum_{y=1}^K P^{(1)}(y)^2}{1 - \sum_{y=1}^K P^{(1)}(y)^2} \\
&= p_{copy}
\end{aligned}$$

□

Second, in the general case when the two classifiers have different accuracies, EC measures two aspects of similarity between the two classifiers:

- (1) the proportion of copied responses ( $p_{copy}$ ) and
- (2) the (mis-)match between their accuracies (factor  $f$  in Proposition 2).

Both are intuitively valid determinants of the similarity between two classifiers. We formalize the interplay between these two aspects in Proposition 2. In contrast, Safak [2020] tackled the same problem but proposed a linear scaling whereas our result reflects a more interpretable multiplicative scaling. Again, this result holds for Cohen’s  $\kappa$  in general and, therefore, also when it is applied to measure EC (for which  $y$  would be replaced by  $r$  and different marginals correspond to different accuracies).

**Proposition 2.** *Let a classifier give correct responses according to some marginal distribution  $\hat{y}^{(1)} \sim \tilde{P}^{(1)}$ . If another classifier copies that response  $\hat{y}^{(2)} = \hat{y}^{(1)}$  with probability  $p_{copy}$  and otherwise draws independently from its (potentially different) marginal distribution  $\hat{y}^{(2)} \sim \tilde{P}^{(2)}$ , then*

$$\text{Cohen's } \kappa = p_{copy} \cdot \underbrace{\frac{1 - \sum_y P^{(1)}(y)^2}{1 - \sum_y P^{(1)}(y)P^{(2)}(y)}}_{\text{factor } f \text{ due to mismatching marginals}}$$

with observed marginal distributions  $P^{(1)} = \tilde{P}^{(1)}$  and  $P^{(2)} = p_{copy}\tilde{P}^{(1)} + (1 - p_{copy})\tilde{P}^{(2)}$ .

See Appendix B for the proof. Crucially, this view of Cohen’s  $\kappa$  holds with generality as long as  $K = 2$ , without limiting assumptions about the marginal distributions. Note that the relation between EC and  $p_{copy}$  is not symmetric: Whether classifier A is considered to copy from classifier B or vice versa results in different mismatch factors  $f$ . But note also that this factor vanishes ( $f = 1$ ) if the reference classifier, from which responses are assumed to be copied, has a uniform marginal distribution.

This inspires a strong recommendation for applications of EC: When comparing multiple DNNs with different accuracies to humans, ensure that the *human responses* have uniform marginal distributions (i.e., have equally many correct as incorrect responses). In these cases, EC becomes interpretable as the proportion of responses that the DNNs copied from the human responses,  $\text{EC} = p_{copy}$ , regardless of the overall accuracy of the different DNNs. Without a uniform marginal distribution, DNNs with a higher-than-human accuracy may be evaluated as less similar to humans because their classification process is more accurate, not because it is functionally different. To be clear, we still believe that it is meaningful to interpret differences in accuracies between classifiers. But, ideally, EC would offer additional, orthogonal information beyond that accuracy mismatch—which is ensured by keeping the marginal distribution of the reference classifier (humans) uniform.

Beyond this contribution to interpreting EC, the generative model in Proposition 2 contributes in another way: It allows simulating data to predict the stability of real data. We will heavily use the latter to give guidance to practitioners planning how many samples they should present to real human participants in their experiments.

### 3 Calculating confidence intervals for EC values

Whenever an aggregate measure of empirical data is reported, it is best practice to also quantify the degree of uncertainty about the final value. Historically, this has not always been done for EC, e.g. none of Geirhos et al. [2018], Ollikka et al. [2024], Li et al. [2025] report confidence intervals. While [Geirhos et al., 2020] provide a basic variant of confidence intervals, these were computed in a suboptimal way: They were based on the expected agreement  $p_{exp}$  rather than the actual accuracies of the two classifiers,  $p_1$  and  $p_2$ . This creates an ambiguity because, for example,  $(p_1 = 0.5, p_2 = 0.5)$  as well as  $(p_2 = 0.5, p_1 = 0.99)$  both produce the same expected agreement of  $p_{exp} = 0.5$  but result in different confidence interval widths. Moreover, these confidence intervals were necessarily centered around 0, because they were constructed only for independent observers. Our approach enables calculating confidence intervals for two dependent observers by making use of the observed accuracies. We begin by outlining how confidence intervals can be obtained for existing empirical data, using a straight-forward bootstrapping approach. After that, we make use of the generative model implied by Proposition 2 to derive confidence intervals ahead of time to plan sufficiently powerful experiments.

### 3.1 Using existing data for bootstrapping

Assuming that we already have access to two sequences of binary responses,  $r^{(1)}$  and  $r^{(2)}$  of length  $N$ , we can obtain a measure of uncertainty simply by *bootstrapping* [Efron and Tibshirani, 1994]. To do so, we sample  $N$  elements with uniform probability and replacement from the sequences,  $((r_i^{(1)}, r_i^{(2)}) \mid i \in \{1, \dots, N\})$ , and re-calculate EC based on the sampled trials. This procedure is established in other fields using consistency measures [McKenzie et al., 1996, Vanbelle and Albert, 2008]. Doing this repeatedly yields an empirical distribution of EC values, for which we can report a 95% confidence interval by taking the central 95% of the posterior mass, i.e. the interval  $[q_{0.025}, q_{0.975}]$ , where  $q_p$  denotes the  $p$ -th quantile of the posterior distribution. Alternatively, one could report the Highest Posterior Density Interval (HDPI), which is the narrowest interval containing a certain amount (e.g. 95%) of the posterior mass. The only parameter of this procedure is the number of bootstraps,  $M$ , which we typically set to  $M = 10,000$ , see Figure 8 for a justification of this choice. For readers unfamiliar with the bootstrap approach, we provide a more detailed description in Appendix F and an explicit algorithm in Algorithm 1.

We apply this method of estimating CIs via bootstrapping to the data from Geirhos et al. [2021] in Section 4. In their work, several human observers (typically four to six) and 52 models were evaluated on corrupted natural images, which had to be classified into sixteen classes. There are 17 different corruptions, most of which are parameterized by a scalar intensity level, leading to multiple conditions within each corruption. Each condition contains 160 to 640 images. For every model, its EC to human observers is calculated in a hierarchical fashion: First, within each condition, the ECs to all human observers are calculated and then averaged. Then, the conditions of each corruption are averaged again, before calculating the final “human-machine error consistency” as an average over corruptions. To accurately reflect this procedure, we bootstrap through the entire calculation, obtaining a posterior over the final average itself and plot 95% percentile intervals of these averages in Figure 4.

### 3.2 Using Proposition 2 for simulations

But what about the scenario in which no data is available yet, e.g. when planning an experiment? Without data, we first need a stochastic generative model that can produce trial sequences resulting in a target EC, given two classifiers characterized by their marginals, i.e. their accuracies.

Proposition 2 provides such a model in which the underlying marginals and copy probability can be specified to generate data. Note that, depending on the copy probability, the observed marginals will deviate from the underlying marginal: The more responses the second classifier copies from the first, the more its marginal distribution adapts. To simulate data with specific marginal distributions  $P^{(1)}$  and  $P^{(2)}$  as well as EC, one can simulate data with copy probability and underlying marginals

$$p_{copy} = \text{EC} \cdot \left( \frac{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)^2} \right),$$

$$\tilde{P}^{(1)}(y) = P^{(1)}(y), \text{ and}$$

$$\tilde{P}^{(2)}(y) = \frac{P^{(2)}(y) - p_{copy} \cdot P^{(1)}(y)}{1 - p_{copy}}.$$

Equipped with this generative model, we can estimate confidence intervals for a pair of classifiers as follows: We first sample  $N$  trials for the first classifier by drawing from a Bernoulli distribution  $r_i^{(1)} \sim \tilde{P}^{(1)}$ . Next, we copy  $p_{copy} \cdot N$  responses from the first classifier to the second. For the remaining  $(1 - p_{copy}) \cdot N$  responses, we sample independently from a Bernoulli distribution from the underlying marginal  $r_i^{(2)} \sim \tilde{P}^{(2)}$ . This yields one data set based on which we estimate EC. Repeating this process then yields an estimate for the variability of these EC values. We provide an algorithmic description of this approach in Algorithm 2.

We plot the resulting confidence intervals in Figure 3, and include sample sizes used in the literature [Geirhos et al., 2021, Ollikka et al., 2024, Li et al., 2025, Wiles et al., 2024] as reference points. We focus on the most severe case where  $\kappa = 0.5$ , but discuss the effect of  $\kappa$  on CI size in Appendix E. Note that the x-axis is on a logarithmic scale and reaches values that are very difficult to meet in experiments with human observers. Even so, the width of the confidence intervals (especially that

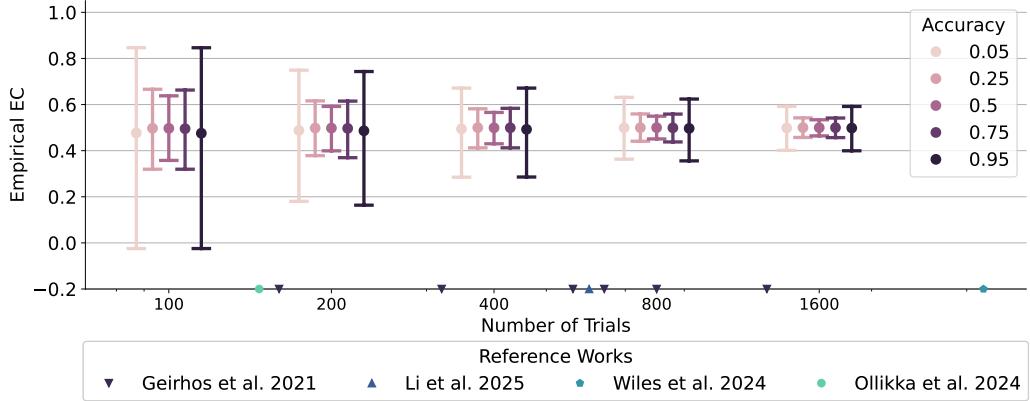


Figure 3: **95% Confidence Interval sizes as a function of trial number.** We simulate data demonstrating how the size of EC confidence intervals depends on both the accuracy of the classifiers and the number of trials. The ground truth EC is 0.5, and we bootstrap 1,000 times for each pair of observers, which are set to the same accuracy for simplicity. We also include the trial numbers employed in the literature for reference (symbols on the x-axis).

of classifiers close to floor / ceiling performance) is very large and shrinks only slowly, covering 0.2 units of  $\kappa$  even at  $N = 400$  trials, which is roughly the size of the difference found between humans and machines by Geirhos et al. [2021]. Especially noteworthy is the fact that at lower  $N$  and performance close to floor / ceiling, the bootstrapped ECs are biased and undershoot the true value. This happens for a simple reason: If one classifier is always correct (or always wrong) the EC will be zero irrespective of the accuracy of the other one. In the critical performance regimes and at a low number of trials, sub-sampling trial sequences for which these edge cases occur becomes likely enough to bias results.

**Limitations.** A core limitation of our work is that we do not present an analytic solution for confidence intervals, but rely on bootstrapping instead. This also limits the numerical accuracy of the significance tests, which cannot resolve p-value differences  $\leq \frac{1}{N}$ . While the analytic solutions (along the lines of Gwet [2016], Fleiss et al. [1969], and Vanbelle et al. [2024]) could improve this slightly, we believe that for the discussed purpose here, our methodology is the pragmatic choice that comes with fewer restrictions (and is certainly much better than not reporting confidence intervals at all).

### 3.3 Significance Tests for Error Consistency

Now that we can quantify the degree of uncertainty, we can also derive significance tests for error consistency. Significance tests are an invaluable tool for empirical scientists because they provide a standardized way of checking if results are statistically reliable. Given two sequences of binary trials of length  $N$  which lead to an EC of  $\kappa$ , we want to calculate a p-value for  $\kappa$ . The p-value is the probability of observing an event at least as extreme as the observed one under the null hypothesis of two independent binomial observers, i.e.  $\kappa = 0$ .

**Basic Significance Test.** To obtain the distribution of EC values obtained under the null hypothesis  $H_0 : \kappa = 0$ , we first need to model the independent binomial observers that constitute the null hypothesis. They are fully characterized by their individual accuracies, of which the empirically observed accuracies provide an estimate. The reliability of this estimate depends on the number of trials. If we assume uniform priors over these accuracies, the posteriors are beta-distributions characterized by the number of (in)correctly solved trials,  $\text{Beta}(N - \sum_i r_i^{(j)}, \sum_i r_i^{(j)})$ .

We can now sample the two individual accuracies from their respective posteriors, sample  $N$  trials from independent binomial distributions and calculate the resulting EC. Repeating this procedure  $M$  times leads to a distribution of EC values expected under the null hypothesis. From this distribution,

we then calculate how many times the absolute value of the simulated EC exceeded the absolute value of the empirically measured EC, for a two-sided significance test. We demonstrate this method by applying it to data from Model-vs-Human in Appendix H to obtain p-values of the error consistency between a human subject and the best-performing model, a CLIP-trained ViT [Radford et al., 2021].

**Advanced Significance Test.** The empirical scenario which we investigate in Section 4 demands a slightly different significance test: Given a fixed reference observer (e.g. a human) and two candidate observers (e.g. two DNNs), is the difference of ECs to the reference observer found between the two candidates statistically significant?

Such a test is also possible. Let  $\kappa_{c_1}, \kappa_{c_2}$  be the error consistencies of the two candidates. We consider the null hypothesis of the candidates having the same error consistency to the reference classifier,  $H_0 : \kappa_{c_1} = \kappa_{c_2}$ . We first obtain a distribution of error consistency differences via bootstrapping as above: First, we randomly sample  $N$  trial indices  $i$  with replacement and uniform probability. We select the corresponding response-correctness-triplets  $(r_i^{(ref)}, r_i^{(c_1)}, r_i^{(c_2)}) \in \{0, 1\}^3$  and re-calculate EC values and their difference. To obtain the distribution of differences one would expect under  $H_0$ , we simply exchange  $r_i^{(c_1)}, r_i^{(c_2)}$  with probability  $p = 0.5$ , as if there had been only one underlying model. We can then perform a t-test between the distributions of differences to compute p-values.

## 4 Revisiting earlier work

### 4.1 Model-vs-Human

Next, we re-analyze well-known literature results by calculating confidence intervals around their empirically measured values. At the focus of our attention is the work by Geirhos et al. [2021], where a cohort of human observers was asked to classify corrupted natural images. The same images were also shown to 52 DNNs, to evaluate their alignment to human behavior. EC values were calculated per condition and averaged as described in Section 3. From our analysis in Figure 3, we already know that some of the error consistencies were calculated in a low-data regime that will result in large confidence intervals around the individual EC values, but we now compute confidence intervals specifically for their empirical data. We do this by bootstrapping through the entire calculation, to arrive at an empirical distribution of final scores, which we plot for humans and every model in Figure 4. The best model is a CLIP-trained ViT[Radford et al., 2021] which performs much better on the distribution shifts like sketches and stylized images, achieving an EC to humans of 0.279. The second-best model, BiTM-ResnetV2-101x1, achieves an EC of 0.251, resulting in a difference of 0.028, which, as a two-sided t-test confirms, is significantly different from 0 ( $p < 1 \times 10^{-10}$ ). But note how model 24 (VGG-19 as per Table 2) has overlapping CIs with all models from 8 to 45, more than 70% of all models. It is conceivable that if these experiments were done again, VGG-19 might jump up to rank 8, or down to rank 45. It just so happens that the very best and very worst models broaden the range we have to plot, thus rendering this effect less obvious than in Figure 1. Testing the differences for all neighboring models for significance, we find no other significant result, even without correcting for multiple comparisons.

For each bootstrap, we also obtain a ranking of models. Given that the CIs are fairly wide (for example, no model’s confidence interval is separated from that of its neighbors), we wonder whether different bootstraps imply different model rankings. To check this, we calculate Kendall’s  $\tau$  between the original model ranking and the ranking implied by each of the 10k bootstraps.  $\tau$  is a statistical measure of the rank correlation between two ordered sequences, similar to Spearman’s rank correlation, but with the benefit of an intuitive interpretation:  $\tau$  measures the difference between the probability that a pair of observations is concordant and the probability that it is discordant. The average  $\tau$  is 0.88, which means that two randomly sampled pairs have an 88% chance of being concordant. Evidently, the main conclusion drawn by Geirhos et al. [2021] clearly still holds: Humans are more consistent among themselves than models are to humans. Nevertheless, adding confidence intervals to the reported EC values better quantifies the degree of uncertainty about the measurements for individual models, and the 95% CIs do indeed overlap, meaning that differences between models are not resolved.

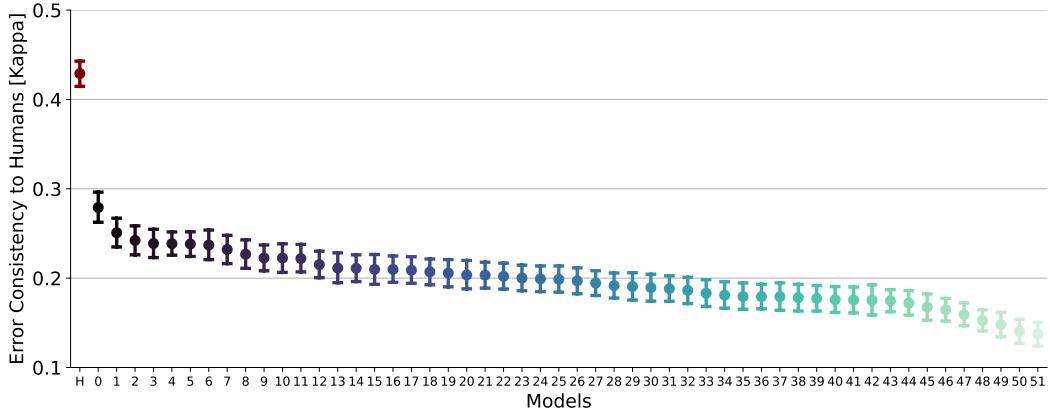


Figure 4: **Models from MvH ranked by their EC to humans.** Inter-human EC is plotted in red. We plot the 95% percentile interval for the mean EC obtained via bootstrapping, using the standard exclusion criteria of the benchmark (orange and black points in Figure 2b). For the mapping of model indices to names, see Table 2. Note that while the gap to humans is large, confidence intervals of many models overlap.

#### 4.2 Brain-Score

One of the most widely-known applications of EC is the influential Brain-Score benchmark<sup>2</sup> [Schrimpf et al., 2018], which evaluates arbitrary DNNs by how much they resemble the brain. One dimension of this evaluation is the behavioral similarity to humans, including error consistency as measured via the data by Geirhos et al. [2021]. Here we assess how reliable the ranking of models implied by their EC to humans really is.

In Brain-Score, the EC for each condition is first divided by the human-to-human EC for this condition, to express values relative to a ceiling. We begin by reverting this operation to obtain raw EC values. To properly bootstrap EC values for these models, we would need access to the complete sequences of trials generated by the models, but this data is not stored anywhere, and re-running the models (not all of which are publicly available anyway) seems excessive. Hence, we instead make mild assumptions to estimate conservative CIs around the empirical values, demonstrating the utility of our copy-model. (See Appendix I for details on this procedure.) We obtain the result depicted in Figure 5 for the top-30 models (by EC to humans). Again, the best model is a CLIP-trained ViT. For the other models, the CIs are so large that any of the top ten models could be the second-best one, meaning that the true rank order between models remains unclear.

We believe that the issue we raise here generalizes beyond just EC and Brain-Score: Any metric that is used to derive rankings of models in a benchmark should come with a method for calculating confidence intervals. Likewise, any benchmark that provides a ranking should quantify the degree of uncertainty surrounding individual values and the stability of the resulting ranking.

### 5 Discussion

In this work, we have built on the error consistency metric proposed by Geirhos et al. [2020], demonstrating how confidence intervals and significance tests for empirical measurements of EC can be calculated using bootstrapping techniques. Additionally, we have presented a new computational model of EC for planning sufficiently powerful, conclusive experiments. We have used our methodology to revisit influential Neuro-AI results Schrimpf et al. [2018], finding that many results are not statistically reliable, because previous studies were underpowered. In the following, we recommend best practices for future benchmarking of behavioral alignment, and facilitate their implementation by making our python package publicly available. A core contribution of our work is an improved understanding of error consistency through a new computational model: EC is broken down into

<sup>2</sup><https://www.brain-score.org>

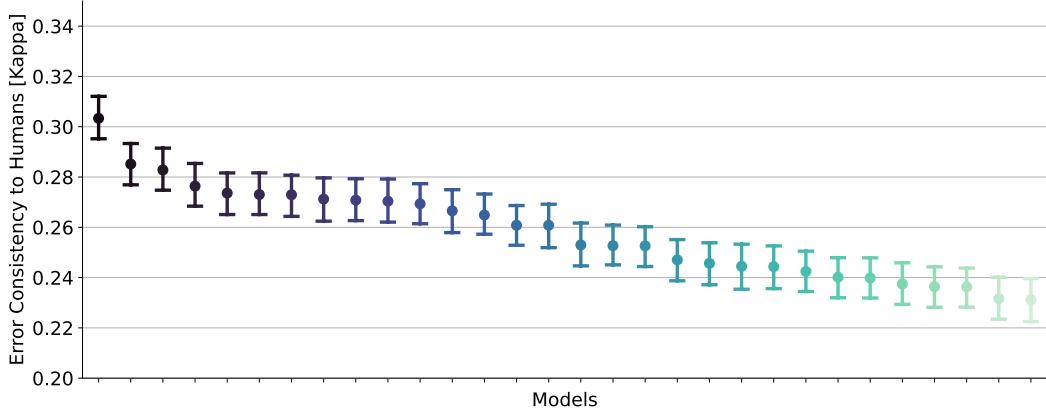


Figure 5: **Confidence Intervals for Brain-Score ECs.** Like in Figure 4, we plot the 95% confidence interval of the EC of the top 30 Brain-Score models. Note that the confidence intervals of many mid-range models overlap.

two interpretable components, copy probability  $p_{copy}$  and marginal mismatch  $f$ . As an analogy, this is similar to breaking down accuracy into true positive and false positive rates—a necessary shift in perspective that will improve interpretations in future studies measuring similarities between classifiers.

### 5.1 Practical Recommendations: How to reliably benchmark behavioral alignment

From our analysis, we can derive concrete recommendations for practitioners. As shown in Figure 3, a sufficiently high number of trials is required to obtain small confidence intervals around EC values. As a rule of thumb, we suggest to collect at least 1000 trials per classifier to balance this requirement against practical constraints. In experiments with human observers, achieving very high resolution of EC differences could become prohibitively expensive, underlining the need for a good selection of stimuli, which probe differences efficiently. Since the size of CIs also depends on the accuracy mismatch between classifiers, we suggest to aim for an accuracy around 75%, but not more than 90% to avoid ceiling-effects, which can be severe. The same holds for floor-effects. This is in agreement with standard best practices from psychophysics, where one aims for a similar accuracy level in human observers to sustain motivation whilst avoiding ceiling effects. Finally, future work should always report confidence intervals and check results for statistical significance, which can easily be done with the python package we provide.

### 5.2 Conclusion

The field of ML research relies heavily on the powerful machinery of benchmarking [Hardt, 2025]. Hence, the correctness and reliability of benchmarks is crucial for progress, since ill-calibrated benchmarks will lead us astray. The field is also notorious for not quantifying uncertainty properly [Pineau et al., 2021, Miller, 2024, Lehmann and Paromau, 2025, Bouthillier et al., 2021], which is especially important in the context of benchmarks because of their large influence on other developments. Our work raises general questions about how benchmarks should aggregate results. For benchmarks that explicitly compute rankings, we give the concrete recommendations of (a) providing confidence intervals around individual estimates and (b) quantifying the stability of the resulting rankings as we did with error consistency.

In this work, we have revisited results by Geirhos et al. [2021] which influence the Brain-Score benchmark Schrimpf et al. [2018], revealing that many differences reported between deep vision models are not statistically reliable. The error consistency metric is particularly susceptible to such issues, because it is inherently noisy, so the requirements for stable measurements are hard to meet in practice. We envision that our work will help the field to overcome these issues, by providing tools to plan sufficiently powerful experiments.

## Acknowledgments and Disclosure of Funding

Funded, in part, by the Collaborative Research Centre (CRC) “Robust Vision – Inference Principles and Neural Mechanisms” of the German Research Foundation (DFG; SFB 1233), project number 276693517. FAW acknowledges funding by the BBVA Foundation Programme Grant “Harnessing Vision Science to Overcome the Critical Limitations of Artificial Neural Networks”. This work was additionally supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1 and via the Open Philanthropy Foundation funded by the Good Ventures Foundation. WB, FAW and KM are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. KM is supported by the Carl Zeiss Foundation through the CZS Institute for Artificial Intelligence and Law. This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Thomas Klein.

We would like to thank Robert Geirhos for helpful discussions, and for providing us with his raw data for Model-vs-Human. We also thank all anonymous reviewers, who provided valuable feedback during the rebuttal period, which improved the final manuscript.

## References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, 2024.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

Jeffrey S. Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfo, John E. Hummel, Rachel F. Heaton, Benjamin D. Evans, Jeffrey Mitchell, and Ryan Blything. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46, 2022. doi: <https://doi.org/10.1017/s0140525x22002813>.

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

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Rosa Falotico and Piero Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470, 2014. doi: 10.1007/s11135-014-0003-1. URL <http://dx.doi.org/10.1007/s11135-014-0003-1>.

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

Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969. doi: 10.1037/h0028106. URL <http://dx.doi.org/10.1037/h0028106>.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

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. In *Advances in Neural Information Processing Systems* 34, 2021.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313*, 2025.

Malcolm J. Grant, Cathryn M. Button, and Brent Snook. An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement*, 41(4):264–276, 2016. doi: <https://doi.org/10.1177/0146621616684584>.

David M Green. Consistency of auditory-detection judgments. *The Journal of the Acoustical Society of America*, 35(5\_Supplement):788–788, 1963.

Kilem L. Gwet. Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement*, 76(4):609–637, 2016. doi: 10.1177/0013164415596420. URL <http://dx.doi.org/10.1177/0013164415596420>.

Moritz Hardt. The emerging science of machine learning benchmarks. Online at <https://mlbenchmarks.org>, 2025. Manuscript.

Tim C Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep neural networks in computational neuroscience. *BioRxiv*, page 133504, 2017.

Bernard J Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. *arXiv preprint arXiv:2404.06647*, 2024.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

Christoph Lehmann and Yahor Paromau. Quantifying uncertainty and variability in machine learning: Confidence intervals for quantiles in performance metric distributions. *arXiv preprint arXiv:2501.16931*, 2025.

Fanfei Li, Thomas Klein, Wieland Brendel, Robert Geirhos, and Roland Zimmermann. Laion-c: An out-of-distribution benchmark for web-scale vision models. In *ICLR 2025 Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Ben Lonnqvist, Alban Bornet, Adrien Doerig, and Michael H Herzog. A comparative biology approach to dnn modeling of vision: A focus on differences, not similarities. *Journal of vision*, 21(10):17–17, 2021.

Tim Maniquet, Hans Op de Beeck, and Andrea Ivan Costantino. Recurrent issues with deep neural network models of visual recognition. *bioRxiv*, pages 2024–04, 2024.

A. Martín Andrés and M. Álvarez Hernández. Estimators of various kappa coefficients based on the unbiased estimator of the expected index of agreements. *Advances in Data Analysis and Classification*, 19(1):177–207, 2024. doi: 10.1007/s11634-024-00581-x. URL <http://dx.doi.org/10.1007/s11634-024-00581-x>.

Dean P. McKenzie, Andrew J. Mackinnon, Normand Péladeau, Patrick Ongena, Peter C. Bruce, David M. Clarke, Susan Harrigan, and Patrick D. McGorry. Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *Journal of Psychiatric Research*, 30(6):483–492, 1996. doi: 10.1016/s0022-3956(96)00033-7. URL [http://dx.doi.org/10.1016/S0022-3956\(96\)00033-7](http://dx.doi.org/10.1016/S0022-3956(96)00033-7).

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

Netta Ollikka, Amro Abbas, Andrea Perin, Markku Kilpeläinen, and Stéphane Deny. Humans beat deep networks at recognizing objects in unusual poses, given enough time. *arXiv preprint arXiv:2402.03973*, 2024.

Will Orr and Edward B Kang. Ai as a sport: On the competitive epistemologies of benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1875–1884, 2024.

Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765, 2023.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

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. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Veli Safak. Min-mid-max scaling, limits of agreement, and agreement score, 2020. URL <https://arxiv.org/abs/2006.12904>.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner’s curse? on pace, progress, and empirical rigor. 2018.

Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, 5: 399–426, 2019.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D’Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scientific machine learning benchmarks. *Nature Reviews Physics*, 4(6):413–420, 2022.

Rachel Thomas and David Uminsky. The problem with metrics is a fundamental problem for ai. *arXiv preprint arXiv:2002.08512*, 2020.

S. Vanbelle and A. Albert. A bootstrap method for comparing correlated kappa coefficients. *Journal of Statistical Computation and Simulation*, 78(11):1009–1015, 2008. doi: 10.1080/00949650701410249. URL <http://dx.doi.org/10.1080/00949650701410249>.

Sophie Vanbelle, Christina Hernandez Engelhart, and Ellen Blix. A comprehensive guide to study the agreement and reliability of multi-observer ordinal data. *BMC Medical Research Methodology*, 24(1), 2024. doi: 10.1186/s12874-024-02431-y. URL <http://dx.doi.org/10.1186/s12874-024-02431-y>.

Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9, 2023.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Chris Knutsen, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024.

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.

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 Related Work

**DNNs as vision models.** Within the last decade, Deep Neural Networks (DNNs) have been proposed as models of the human visual system [Doerig et al., 2023, Kriegeskorte, 2015, Cichy and Kaiser, 2019, Kietzmann et al., 2017], not only because they form the only class of image-computable models that achieves human-level performance on benchmark tasks like ImageNet [Russakovsky et al., 2015], but also because performance-optimized DNNs are the best predictors of biological neural activations [Yamins et al., 2014, Kubilius et al., 2019, Zhuang et al., 2021]. However, the suitability of DNNs as models of vision is debated [Bowers et al., 2022, Wichmann and Geirhos, 2023, Maniquet et al., 2024, Serre, 2019], as is the question of how comparisons should best be conducted [Firestone, 2020, Lonnqvist et al., 2021, Wichmann and Geirhos, 2023].

**Error Consistency.** The underlying idea motivating error consistency is that of “molecular psychophysics” [Green, 1963], which is to go beyond aggregate measures like accuracy, and instead compare behavior on a trial-by-trial basis. EC promises to achieve this by building on Cohen’s  $\kappa$  [Cohen, 1960, Martín Andrés and Álvarez Hernández, 2024], which considers the trial-level agreement observed for a pair of classifiers. Geirhos et al. [2020] proposed EC in the context of human-machine comparison, but in principle, the method applies to the general setting of comparing arbitrary classifiers. Martín Andrés and Álvarez Hernández [2024] observe that  $\kappa$  is a biased estimator and should be corrected, which we support in our implementation. Safak [2020] also observe the problem of  $\kappa$ ’s dependency on the marginals, which we outline in Section 2. They propose to scale  $\kappa$  between its theoretical limits in an attempt to obtain a measure of error consistency that is orthogonal to accuracy. Goel et al. [2025] propose to apply Cohen’s  $\kappa$  directly to the responses given by both classifiers  $\hat{y}_i$  rather than the correctness values  $r_i$ , and adapt the marginal probabilities of independent classifiers to reflect their accuracy. We still focus on the standard error consistency proposed by Geirhos et al. [2020] because it has seen the widest application. In principle, our bootstrapping methodology also works for these other metrics.

**Uncertainties in Benchmarking.** Benchmarking has a long history and is central to modern machine learning [Koch and Peterson, 2024, Singh et al., 2025, Hardt, 2025, Orr and Kang, 2024]. For example, the popular Papers with Code database<sup>3</sup> lists approximately 14,000 different benchmarks. With the emergence of modern benchmarks and leaderboards, issues such as overfitting to benchmarks [Singh et al., 2025], the limitations of metrics [Thomas and Umansky, 2020], and various sources of uncertainty have been identified and challenged [Lehmann and Paromau, 2025, Bouthillier et al., 2021]. In general, it has been criticized that benchmarks should not be the primary goal of scientific machine learning research [Alzahrani et al., 2024]. At the same time, it has been pointed out that benchmarking itself requires a scientific approach and some kind of standard methodology [Thiyagalingam et al., 2022, Sculley et al., 2018, Hardt, 2025]. Similar to our work, Nado et al. [2021] argued that uncertainty is an important aspect to consider for benchmarks. In contrast to previous work, we focus on uncertainty in the behavioral comparison of classifiers using error consistency.

## B Proof of Proposition 2

*Proof.* The underlying distributions translate into the observed distributions  $P_1 = \tilde{P}_1$  and  $P^{(2)} = p_{copy}\tilde{P}^{(1)} + (1 - p_{copy})\tilde{P}^{(2)}$ . Inverting the marginal  $P^{(2)}$ , we get  $\tilde{P}^{(2)}(y) = \frac{P^{(2)}(y) - p_{copy} \cdot P^{(1)}(y)}{1 - p_{copy}}$ .

For the joint distribution, we have  $P(y, y) = \tilde{P}^{(1)}(y)(p_{copy} + (1 - p_{copy})\tilde{P}_2(y))$  because the second classifier makes the same prediction as the first in all copy cases ( $p_{copy}$ ) and in those cases, in which it does not copy but coincidentally makes the same prediction ( $(1 - p_{copy})P^{(2)}(y)$ ).

<sup>3</sup><https://paperswithcode.com/sota>

$$\begin{aligned}
\kappa &= \frac{p_{obs} - p_{exp}}{1 - p_{exp}} \\
&= \frac{\sum_{y=1}^K P^{(1,2)}(y, y) - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= \frac{\sum_{y=1}^K \left( P^{(1)}(y)(p_{copy} + (1 - p_{copy})\tilde{P}_2(y)) \right) - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= \frac{\sum_{y=1}^K \left( P^{(1)}(y)(p_{copy} + P^{(2)}(y) - p_{copy} \cdot P^{(1)}(y)) \right) - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= p_{copy} \frac{\sum_{y=1}^K \left( P^{(1)}(y) - P^{(1)}(y) \cdot P^{(1)}(y) \right)}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \\
&= p_{copy} \cdot \left( \frac{1 - \sum_{y=1}^K P^{(1)}(y)^2}{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)} \right)
\end{aligned}$$

□

## C Proof of Ceiling Effects

In Section 2, we claim that as soon as one classifier achieves perfect accuracy on the investigated sequence of trials, the EC of the pair will be 0 irrespective of the accuracy of the other one. Here, we proof this claim, which is easiest to derive from a confusion matrix as shown in Table 1. The observed agreement is found on the diagonal,  $p_{obs} = a + d$ , while the expected agreement is what would be on the diagonal if observers were independent ( $p_{exp} = (a + b)(a + c) + (c + d)(b + d)$ ). For the proof, consider without loss of generality that classifier 2 was perfect, i.e.  $a + c = 1$ .

$r^{(2)} = 1$	$r^{(2)} = 0$	
$r^{(1)} = 1$	$a$	
$r^{(1)} = 0$	$d$	
$a + c$	$b + d$	1.0

Table 1: Error consistency matrix.

*Proof.* Let  $a + c = 1$ , then  $b + d = 0$ , so  $p_{exp} = (a + b)(a + c) + (c + d)(b + d) = (a + b)$ . Since  $b \geq 0$  and  $d \geq 0$ ,  $b = d = 0$ .

$$\begin{aligned}
\kappa &= \frac{p_{obs} - p_{exp}}{1 - p_{exp}} \\
&= \frac{(a + d) - (a + b)}{1 - (a + b)} \\
&= \frac{d - b}{1 - (a + b)} \\
&= 0
\end{aligned}$$

□

By symmetry, the same argument holds for the case where one classifier gives incorrect responses on all trials.

## D Lower bound on EC

Figure 2b shows how the mismatch in accuracies between two classifiers imposes upper bounds on EC. In extension, Figure 6 shows that there are also non-trivial lower bounds (subplot b). These lower bounds are not a simple function of the upper bounds. Instead, they are inversely related such that when accuracies are inverted (i.e.,  $p_1 = 1 - p_2$  along the off-diagonal) lower values can be attained. But perhaps counterintuitively, even with inverted accuracies,  $EC = -1$  can not always be reached. This is in contrast to the upper bound, where aligned accuracies (i.e.,  $p_1 = p_2$ ) always allows  $EC = +1$ .

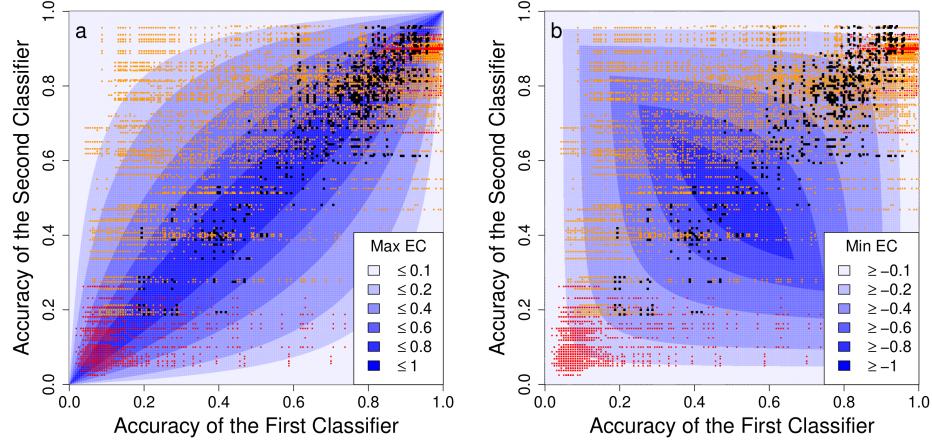


Figure 6: **Lower bound of EC in comparison to upper bounds.** The left subplot (a) is the same as 2b, showing regions in which EC is upper bounded by the mismatch in accuracies of two classifiers (x and y axes). Additionally, the right subplot (b) shows the lower bounds. Again, each point represents a Model-vs-Human comparison in the analysis of Geirhos et al. [2021].

## E Effect of EC on CI size

For Figure 3, we have chosen a ground-truth EC of 0.5 to demonstrate how the EC between two classifiers depends on their accuracy and the number of trials on which EC is measured. However, the choice of the ground-truth EC influences the CI size as well, which we demonstrate in Figure 7. We again generate data using our copy-model, and plot the ground-truth EC (input of the model) against the distribution of empirical ECs (calculated on model outputs). To better show the changes in CI size, we plot the delta between empirical and ground-truth EC rather than the absolute EC. Note how the CI bounds are not symmetric and largest at a ground-truth EC of 0.5.

## F Bootstrapping

For readers unfamiliar with bootstrapping, we provide a basic introduction here, but refer to Efron and Tibshirani [1994] for details. The idea is the following: We want to estimate the uncertainty in a summary statistic (e.g., the mean), which is computed for a dataset  $X$  of size  $N$ . If we had access to the underlying data-generating process, we could simply generate  $M$  more datasets (each of size  $N$ ) and compute the summary statistic for each. This would give us a distribution of values of the statistic, from which we could then compute a measure of uncertainty (e.g. a percentile interval of the distribution of means). Without access to the true data-generating process, we build the best possible parameter-free model: To generate a data point, randomly select one of the data points  $x_i$  with uniform probability. To generate an entire dataset, draw  $N$  data points with uniform probability, with replacement. In the context of error consistency between two observers, one data point corresponds to one trial, i.e. a pair of responses  $r_i = (r_i^{(1)}, r_i^{(2)}) \in \{0, 1\}^2$ . The full algorithm to obtain a confidence interval for two sequences of responses  $r^{(1)}$  and  $r^{(2)}$  can thus be summarized in Algorithm 1.

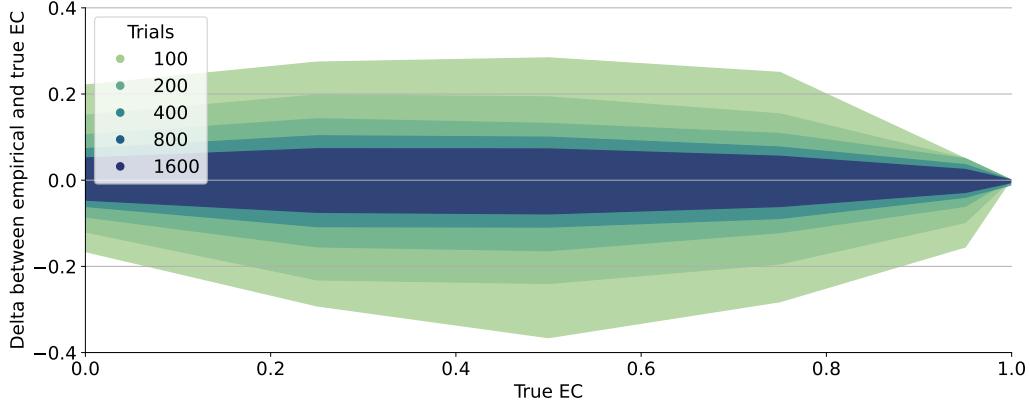


Figure 7: **Ground-Truth EC against CI sizes.** We simulate data with different ground-truth EC values and plot the size of 95% confidence intervals. Evidently, one obtains the largest CIs at ground-truth EC of 0.5.

---

**Algorithm 1** Bootstrapping EC confidence intervals

---

**Require:** Arrays  $r^{(1)}, r^{(2)} \in \{0, 1\}^N$ , number of bootstrap iterations  $M$ , sample size  $N$   
**Ensure:** 95% confidence interval for error consistency

```

1:  $\mathcal{L} \leftarrow \emptyset$                                      ▷ Initialize list of EC values
2: for  $i = 1$  to  $M$  do
3:    $\mathcal{I} \leftarrow$  sample  $N$  indices from  $\{1, \dots, N\}$  uniformly with replacement
4:    $ec \leftarrow \text{EC}(r^{(1)}[\mathcal{I}], r^{(2)}[\mathcal{I}])$            ▷ Compute error consistency
5:    $\mathcal{L} \leftarrow \mathcal{L} \cup \{ec\}$                          ▷ Add to list
6: end for
7:  $[\text{lower}, \text{upper}] \leftarrow \text{Percentile}_{95\%}(\mathcal{L})$            ▷ Compute 95% interval
8: return  $[\text{lower}, \text{upper}]$ 

```

---

Similarly, the algorithm for estimating confidence intervals without access to real trials is described in Algorithm 2.

## G Validating the copy model

As proven by Proposition 2, our copy model is correct in the limit of trials. Additionally, we empirically validate the uncertainty estimates we derive from this model, by first generating some ground-truth data with a fixed error consistency at a fixed accuracy of both classifiers. We then bootstrap this data  $M = 1,000$  times, as we would to obtain percentile intervals. Next, we instead draw  $M$  samples from our copy model, to compare the CIs implied by our copy model to those implied by the bootstrap, which we consider the gold standard. We plot this comparison in Figure 9. Evidently, the distributions we obtain from our copy model match the bootstrapped results quite well, as long as the number of trials is sufficient or the accuracy avoids floor- and ceiling-effects.

## H Details on Model-vs-Human results

An idiosyncrasy of the Model-vs-Human benchmark that we have not expanded on so far is the fact that while the final scores (which are averages across the different corruptions) are stable, the uncertainty within each corruption can be quite large. In Figure 11, we plot the CIs surrounding the EC values obtained for the phase-scrambling corruption as an example. In Figure 12 we plot the pairwise rank correlations between all 17 corruptions, revealing that some, but not all corruptions are highly correlated, i.e. imply similar model rankings. In Figure 10, we plot the confidence intervals one would obtain under the alternative aggregation strategy of simply averaging across conditions,

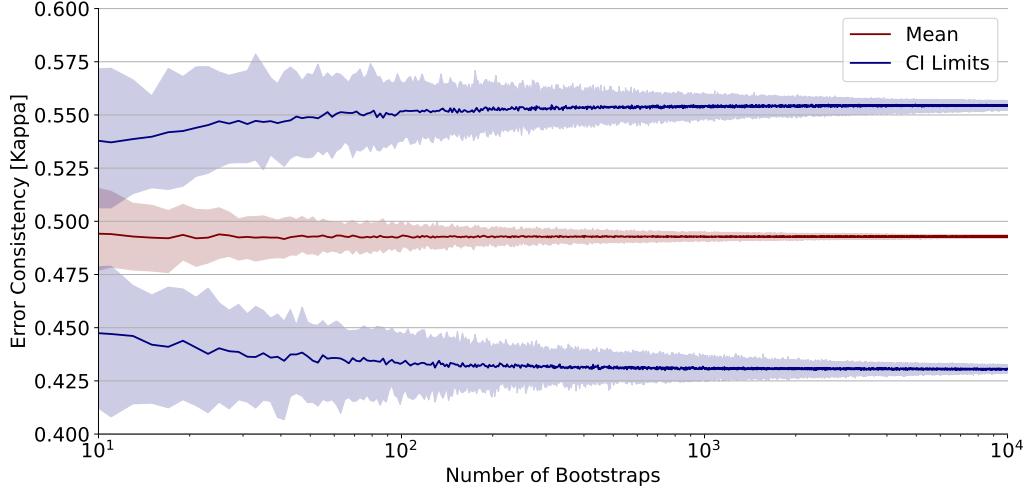


Figure 8: **Convergence of EC estimates as a function of number of bootstraps.** Starting from a single sequence of 1,000 trials generated by our copy model (arbitrary target EC of 0.5, classifier accuracies 0.7 and 0.8 respectively) we chose a number of bootstrap runs and bootstrapped 100 times for each, to obtain estimates of uncertainty. Shaded regions correspond to 95% percentile intervals. Evidently, at the  $M = 10,000$  bootstrap runs we use, the estimates have stabilized, thus justifying our choice of  $M$ .

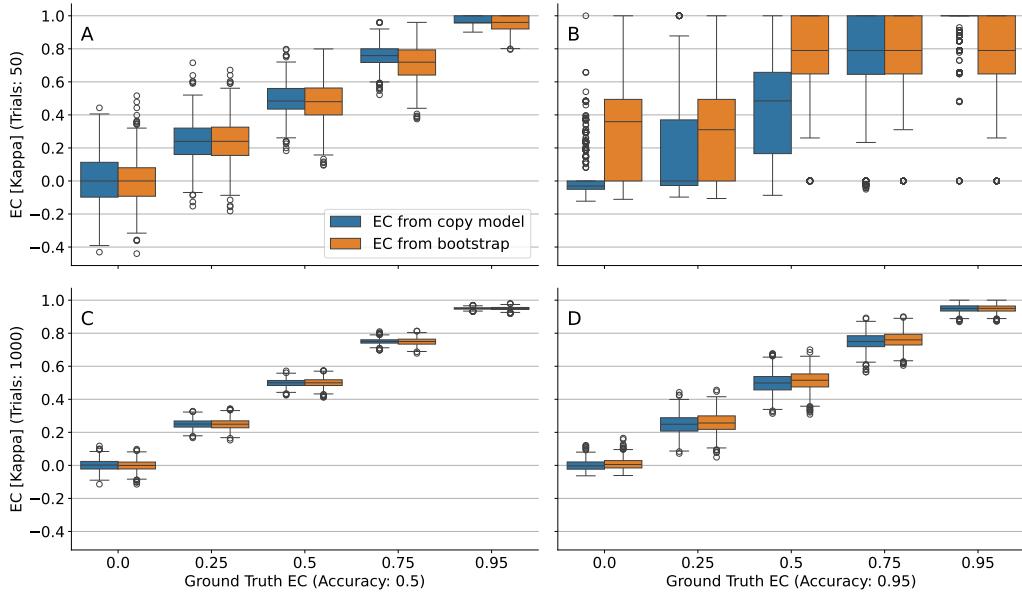


Figure 9: **Validating our copy model.** We plot the distributions of EC estimates from the copy model to bootstrapped (gold standard) estimates of uncertainty. **Top:** 50 trials. **Bottom:** 1000 trials. **Left:** Accuracy 50%. **Right:** Accuracy 95%. Evidently, if both the number of trials is low and accuracy approaches the ceiling of 100%, the model diverges from bootstrapped results, but error consistency is not well-behaved in this case anyway. Ideally, experiments should avoid this regime, which can be achieved by increasing the number of trials and calibrating their difficulty.

---

**Algorithm 2** Bootstrapping EC confidence intervals without real trials

---

**Require:** Target marginals  $P^{(1)}, P^{(2)}$  over  $K$  classes, target error consistency EC, number of trials  $N$ , number of simulations  $M$

**Ensure:** 95% confidence interval for error consistency

- 1: **Compute underlying parameters:**
- 2:  $p_{\text{copy}} \leftarrow \text{EC} \cdot \left( \frac{1 - \sum_{y=1}^K P^{(1)}(y)P^{(2)}(y)}{1 - \sum_{y=1}^K P^{(1)}(y)^2} \right)$
- 3:  $\tilde{P}^{(1)}(y) \leftarrow P^{(1)}(y)$  for all  $y \in \{1, \dots, K\}$
- 4:  $\tilde{P}^{(2)}(y) \leftarrow \frac{P^{(2)}(y) - p_{\text{copy}} \cdot P^{(1)}(y)}{1 - p_{\text{copy}}}$  for all  $y \in \{1, \dots, K\}$
- 5:
- 6:  $\mathcal{L} \leftarrow \emptyset$  ▷ Initialize list of EC values
- 7: **for**  $i = 1$  to  $M$  **do**
- 8:   **Generate responses for first classifier:**
- 9:   **for**  $j = 1$  to  $N$  **do**
- 10:     $r_j^{(1)} \sim \text{Categorical}(\tilde{P}^{(1)})$  ▷ Sample from marginal
- 11:   **end for**
- 12:
- 13:   **Generate responses for second classifier:**
- 14:     $N_{\text{copy}} \leftarrow \lfloor p_{\text{copy}} \cdot N \rfloor$  ▷ Number of responses to copy
- 15:     $\mathcal{I}_{\text{copy}} \leftarrow \text{sample } N_{\text{copy}} \text{ indices from } \{1, \dots, N\} \text{ without replacement}$
- 16:   **for**  $j = 1$  to  $N$  **do**
- 17:     **if**  $j \in \mathcal{I}_{\text{copy}}$  **then**
- 18:        $r_j^{(2)} \leftarrow r_j^{(1)}$  ▷ Copy response
- 19:     **else**
- 20:        $r_j^{(2)} \sim \text{Categorical}(\tilde{P}^{(2)})$  ▷ Sample independently
- 21:     **end if**
- 22:   **end for**
- 23:
- 24:     $ec \leftarrow \text{EC}(r^{(1)}, r^{(2)})$  ▷ Compute error consistency
- 25:     $\mathcal{L} \leftarrow \mathcal{L} \cup \{ec\}$
- 26: **end for**
- 27:  $[\text{lower}, \text{upper}] \leftarrow \text{Percentile}_{95\%}(\mathcal{L})$  ▷ Compute 95% interval
- 28: **return**  $[\text{lower}, \text{upper}]$

---

without grouping by corruption first. Evidently, this would drastically affect the confidence intervals. To be clear, this is a deviation from the MvH protocol and thus “wrong”, but the data itself is the same, thus illustrating the point that there is large variance in the data, which is obfuscated by the hierarchical aggregation.

In Figure 2b, we explain how in many of the conditions within Model-vs-Human, the accuracy mismatch between models and human participants is so large that  $k_{\max}$  takes on very small values. One might wonder how Model-vs-Human scores would change if one excluded conditions not as prescribed by the authors, but based on this mismatch. However, this would render the selection of conditions dependent on the model. Hence, the benchmark would no longer present a fair comparison of models, and final scores would no longer be comparable.

## I Generating Synthetic Brain-Score Data

For the analysis conducted in Section 4.2, we estimate the size of confidence intervals surrounding models on the public Brain-Score benchmark. To estimate a model’s EC to humans within one condition  $c$ , we would technically have to create one set of binary trials that has the property of having the desired average EC to the  $n$  human observers of  $\kappa_{\text{real},c}$ . To simplify, we drop this dependence and generate  $n$  independent sequences of trials, one for each human who saw images in this condition. Generating only one sequence of trials would drastically overestimate the variance, which is reduced by taking the mean. For each of these sequences, we need three inputs for the copy-model: The

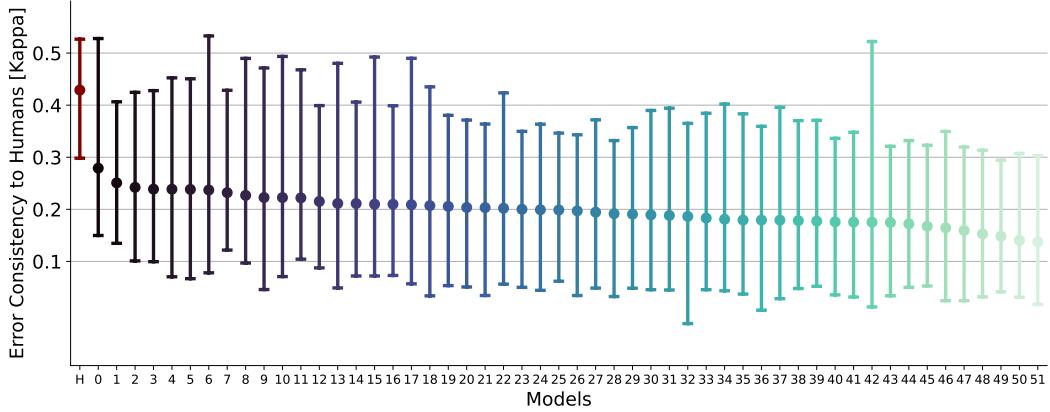


Figure 10: **EC to humans without hierarchical aggregation.** If one were to simply define a model’s EC to humans as its average EC over all conditions (without first grouping by corruption), one would obtain the 95% confidence intervals depicted here, which clearly overlap for all models.

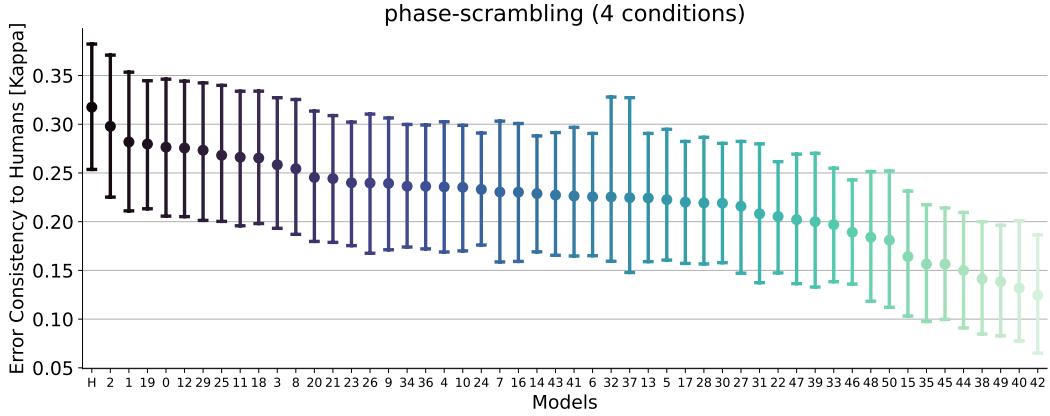


Figure 11: **ECs on phase-scrambled images.** We plot CIs for the EC to humans achieved by all models in the phase-scrambling corruption as an example, analogous to Figure 4

desired kappa (which is given by the Brain-Score data after removing the ceiling correction), the accuracy of the humans (which is given by the Model-vs-Human data) and the accuracy of the model in this condition. The latter is unknown, but bounded mathematically. We obtain the CI sizes in Figure 5 by selecting the accuracy in the middle of the bounds, which seems reasonable. Selecting the lower bounds of the accuracy leads to the most optimistic estimate of CI size, which we plot in Figure 14.

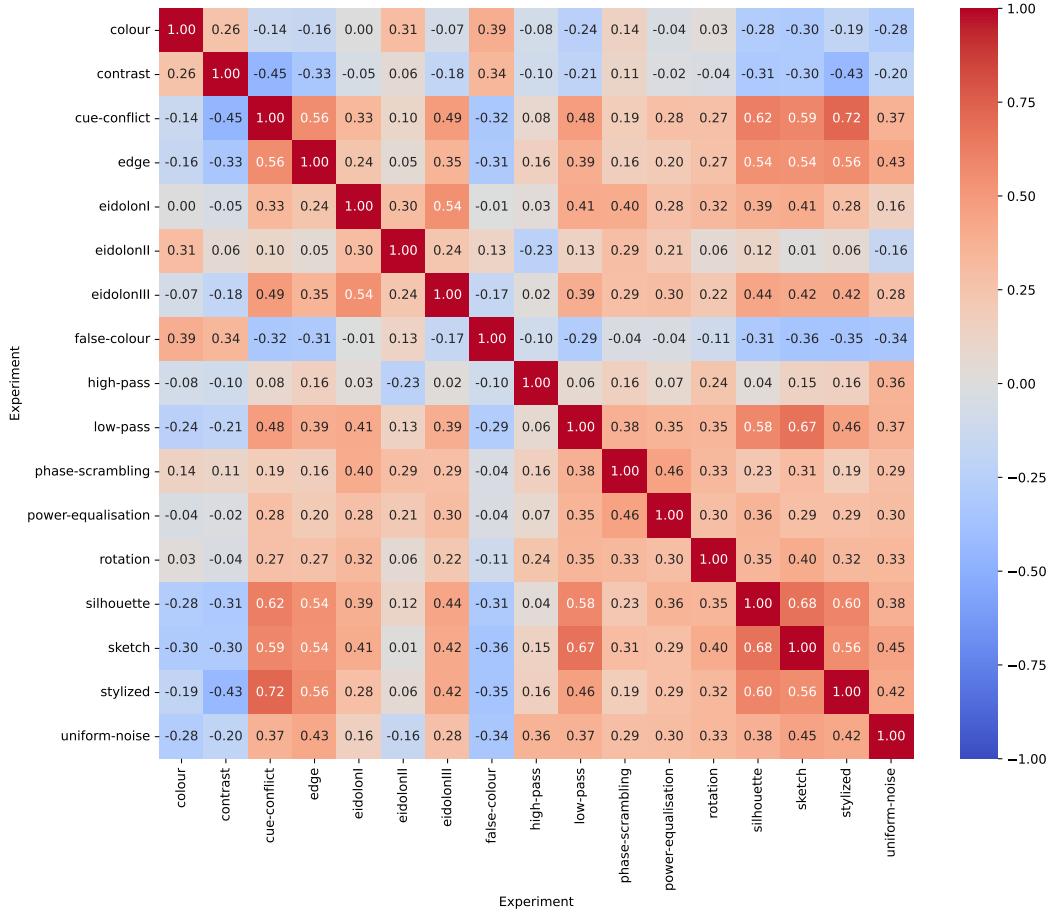


Figure 12: **Spearman Correlation between Experiments.** For all possible pairs formed by the 17 corruptions from Model-vs-Human, we compute the Spearman rank correlation between the rankings of models they imply. All rank correlations are statistically significant at  $\alpha = 0.05$  (without correcting for multiple comparisons).

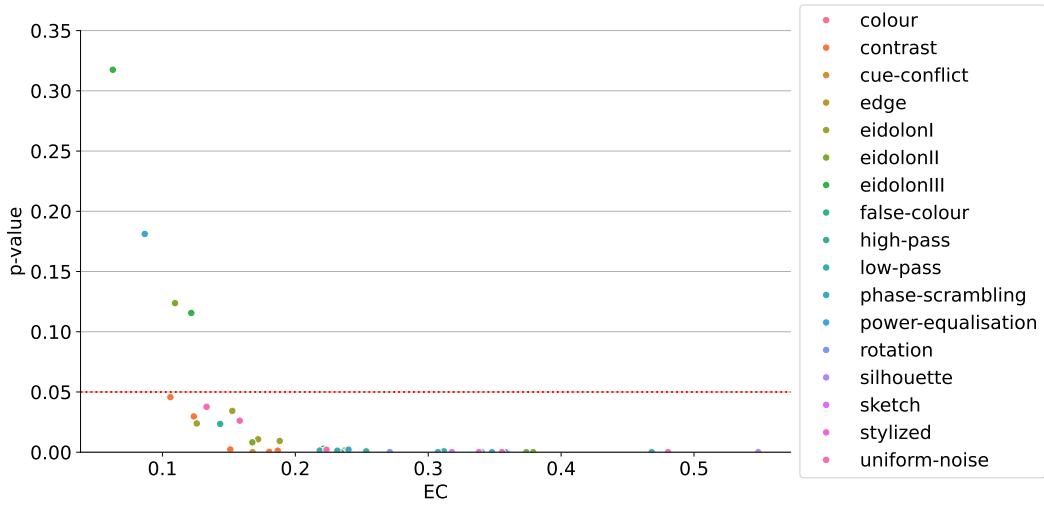


Figure 13: **p-values for the EC between CLIP and human participants.** For all experimental conditions (corruptions at a certain strength) we scatter the EC against the p-value, finding that the majority of conditions is significant, while some values close to 0 are not, which is reasonable.

Index	Model Name	Index	Model Name
0	clip	26	vgg16_bn
1	BiTM_resnetv2_101x1	27	mobilenet_v2
2	BiTM_resnetv2_152x2	28	vgg11_bn
3	BiTM_resnetv2_50x1	29	resnet101
4	resnet50_l2_eps5	30	resnet152
5	resnet50_l2_eps3	31	mnasnet1_0
6	ResNeXt101_32x16d_swsl	32	wide_resnet101_2
7	BiTM_resnetv2_152x4	33	resnext50_32x4d
8	BiTM_resnetv2_50x3	34	resnext101_32x8d
9	resnet50_l2_eps1	35	simclr_resnet50x2
10	vit_large_patch16_224	36	vgg13_bn
11	vit_base_patch16_224	37	simclr_resnet50x4
12	vit_small_patch16_224	38	simclr_resnet50x1
13	transformer_L16_IN21K	39	resnet50_l2_eps0
14	densenet201	40	MoCoV2
15	resnet50_swsl	41	wide_resnet50_2
16	inception_v3	42	efficientnet_l2_noisy_student_475
17	transformer_B16_IN21K	43	squeezezenet1_1
18	resnet50	44	mnasnet0_5
19	densenet169	45	InfoMin
20	BiTM_resnetv2_101x3	46	alexnet
21	resnet34	47	shufflenet_v2_x0_5
22	resnet50_l2_eps0_5	48	squeezezenet1_0
23	resnet18	49	MoCo
24	vgg19_bn	50	PIRL
25	densenet121	51	InsDis

Table 2: **Model Indices for MvH.** We map model indices in Figure 4 to their full names.

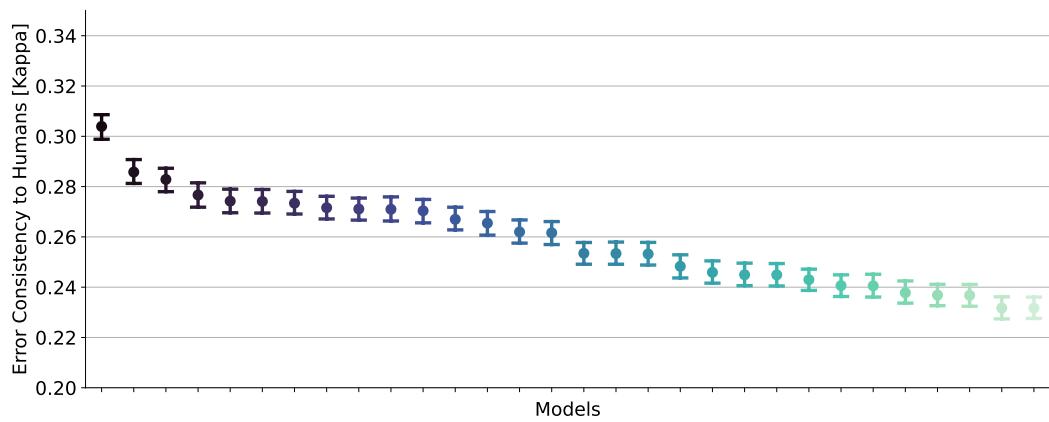


Figure 14: **Most optimistic Confidence Intervals for Brain-Score ECs.** Like in Figure 5, we plot the 95% confidence interval of the EC of the top 30 Brain-Score models, but assume the lowest mathematically possible model accuracy. The true CI size for these models probably looks more like Figure 5, but this is a lower bound on their size.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: In abstract and introduction, we claim to calculate confidence intervals for error consistency and use this to revisit earlier work, which we then do.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations of our work, in particular the fact that we are limited to bootstrapping instead of relying on particular analytical solutions. The computational costs of our algorithms are negligible. There are no privacy and fairness concerns.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: **[Yes]**

Justification: The new theoretical contribution of our paper is the copy model in Section 3. For didactic reasons, we first proof the correctness of the model for a simpler case of equal accuracies, Proposition 1, before moving on to the harder case of unequal accuracies in Proposition 2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: **[Yes]**

Justification: We provide code to reproduce all experiments in the supplemental material. The data sources for empirical experiments (Brain-Score<sup>4</sup> and Model-vs-Human<sup>5</sup>) are publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

---

<sup>4</sup><https://www.brain-score.org>

<sup>5</sup><https://github.com/bethgelab/model-vs-human>

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: See previous question. The submitted code is extensively annotated with instructions and contains an environment specification via requirements.txt.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[NA]**

Justification: We do not train DNNs in the context of this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[Yes]**

Justification: The reporting of error bars is the topic of our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: All experiments (bootstrapping) done for this work are at tiny scale and run on consumer laptops in seconds or a few minutes.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research is not in conflict with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We feel this question is not applicable to our work. We do not release models that could be misused, or directly impact society.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the works we investigate and whose data we use, such as Model-vs-Human.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets. For the code we will release upon acceptance, we will follow standard procedures by releasing via Github etc.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not conduct experiments with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not conduct experiments with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not make use of LLMs beyond standard programming aid for editing purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.