Measuring all the noises of LLM Evals

[Draft – feedback welcome]
Sida Wang
sida@meta.com

Abstract

As LLM benchmark questions grow more complex, evaluation sample sizes have decreased, heightening the risk of being fooled by randomness. Well-established statistical methods are often misapplied or omitted in LLM research. In this work, we apply the more powerful paired methods to many models and agents evaluated on many benchmarks, making these statistical methods useful to practitioners without much additional effort. Conceptually, we adapt them to LLM evaluations by clearly defining and measuring the prediction noise, the data noise, and their combined total noise. The resulting sample estimators are tested extensively on generate models and are shown to be consistent with the bootstrap and sign test. By analyzing millions of question-level results across hundreds of LLMs and agents on popular benchmarks, we find that each benchmark exhibits a characteristic and highly predictable total noise level on all model pairs. This finding allows practitioners to benefit from rigorous statistical analysis without having to do it themselves and even to infer the likely significance of the results reported by others.

1 Introduction

Inferring the signal from noisy measurements is a key of experimental science. The AI effort is now a leading di AI benchmarks have grown much harder yet are much smaller from 10000 examples on MNIST to 500 on SWEBench-Verified, which heightens the risk of being fooled by randomness. Well-established methods from statistics *can increase the rigor of the conclusions drawn from data* (Benjamini et al., 2021). Even though these methods are found in landmark textbooks and is described clearly for LLM evaluations (Miller, 2024), they are not used by important papers, used incorrectly, or unnecessarily loose (Example 5). While these carefully considered and well-studied methods have long been understood by some experts, they are not understood widely enough to be used correctly and commonly. We apply these well-studied methods to all models and agents evaluated by many benchmarks, and output the resulting noise measurements. Due to the predictability in these measurements, practitioners can use well-tested and accurate-enough noise measurements without running a program or understanding the details of the methods. Besides better usability, this approach also allow us to infer the likely significance of results reported by others.

To do this, we define and measure the *prediction noise* when answering a given question differently in different prediction samples, the *data noise* due to using a finite sample from the population of possible questions, and the *total noise* following the law of total variance. The prediction noise can be observed directly and is significant in LLMs at a higher temperature. The data noise is often ignored in LLM development (Example 1), perhaps because it cannot be measured directly by fixed evaluations. For most cases we argue for the use of total noise as the best standard.

As more resources are devoted to model capability, it becomes more expensive to reject significant results due to a weak analysis. While improving models is hard, the more powerful paired analysis is a known method that should be used. The paired methods considers how model predictions are different from each other. A difficulty is they only apply to pairs of results rather than directly to a table or leaderboard. Our all-pairs-paired method applies the paired method to all pairs of LLMs anyway –

Preprint. Under review.

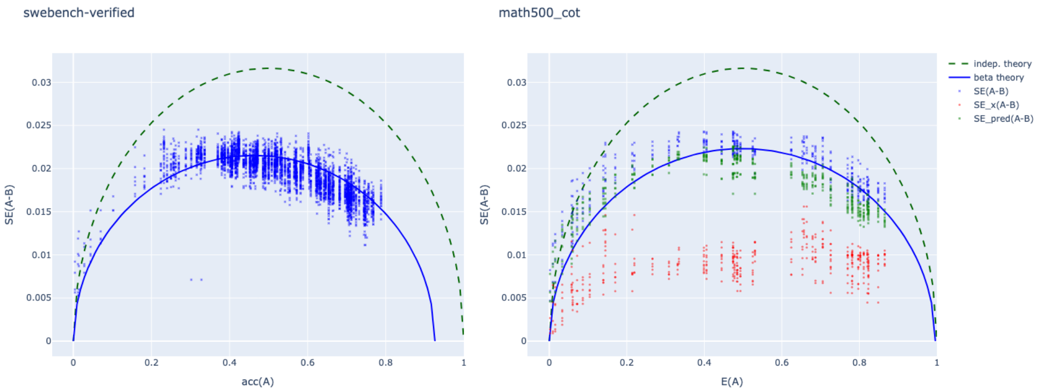

Figure 1: paired total standard errors vs. the accuracy, empirical results agrees well with the Beta theory prediction. Left: on SWEbench-verified 1 prediction per example so only the total noise is estimated, right: MATH500 with 1000 predictions per question and estimated data noise SE_x, the prediction noise SE_pred, and total noise SE. More details in Section 4.

.

we compute the paired variance $\mathrm{Var}[A - B]$ for all model pairs A, B based on the question-level predictions on each benchmark. We find that each benchmark has a characteristic noise level for all pairs of LLM or agents, so without doing separate statistical testing for each new experiment, we can predict the paired variance of the benchmark used. In particular, across benchmarks and LLM/agents, the more powerful paired test usually gives similar total variance as the basic variance predicted by the accuracy p_A of model A, giving this rule of thumb on correctness evaluations

$$\mathrm{Var}[A - B] \approx \mathrm{Var}[A] = p_A(1 - p_A),$$
$$\mathrm{SE}[A - B] \approx \mathrm{SE}[A] = N^{-1/2}\mathrm{Var}[A]^{1/2}.$$

This allows us to meaningfully use a characteristic noise levels per benchmark, thus we have a good idea of the noise levels without running this analysis for each result reported by others or from our own experiments. In addition, Figure 1 shows an example of reference measurements of all noise types that are carefully validated for correctness. With a decoupling of the noise analysis from particular experiments, it is easier to obtain correct variance values and confidence intervals for LLM evaluations, instead of relying on the experimenter to provide their own analysis for each pair of result, which is often done incorrectly in a hurry as after-thoughts to their main results.

The basic analysis method is presented from first principles (Section 3), validated by comparing with well-established bootstrap and sign-test (Section 3.8), the implementation (Section 3.7) is tested against several generative models of the ground truth as well as on many common LLM benchmarks and model combinations, where millions of question level results are aggregated. Section **??** describes data visualization methods that were useful for gaining a qualitative understanding of the model predictions. We show that the noise level is predictable on all the benchmarks, under different temperatures and from millions of question level responses based on public leaderboards as well as on controlled experiments. In discussions (Section 5), we describe a simple generative model that explains the observed data based on the $\mathrm{Beta}(1 - p_A, p_A)$ distribution, why difficult questions cannot yet make up for decreased sample size, how to combine multiple evaluations based on noise, illustrative examples about noise, and some suggestions for current and future benchmarks.

2 Related work

While much work deal with noise in science and machine learning (Lehmann and Romano, 2005; Dror et al., 2018; Hermann et al., 2024, among many others), we adopt the approach of Miller (2024) to estimate the variance directly. While using samples to estimate the population is basic statistics, their approach is clarifying, has unique considerations for LLMs, and is directly generalizable all metric functions. Their method is also consistent with the bootstrap (Efron, 1979) and the sign test (Dixon and Mood, 1946) when applicable (Section 3.8). Efron and Tibshirani (1986) clearly

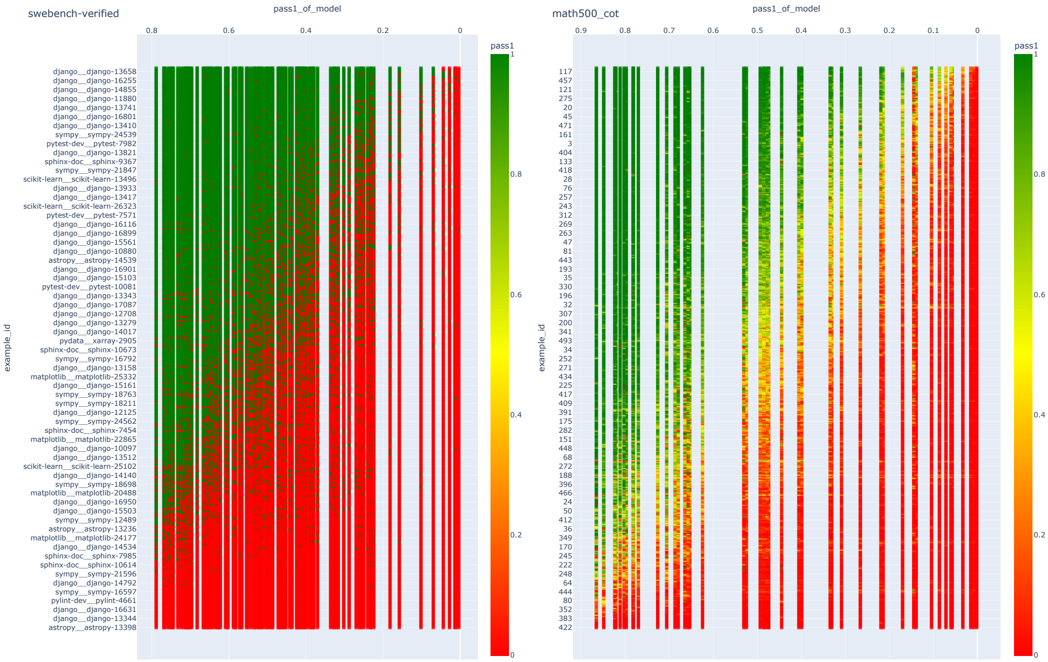

Figure 2: Results heatmap. Each row is a different example, sorted from the easiest to the hardest. Each column is a model whose x-coordinate is the overall accuracy. Left: SWEbench-verified, 1 prediction per question; Right: MATH500, 1000 predictions per question. See Section 4.1 for more details.

described estimating the standard error as the example of not needing bootstrap. The concept of total variance decomposition is present the *analysis of variance*, but Section 3 of Miller (2024) applied this concept to LLMs under *variance of the conditional mean* and *mean conditional variance*, which we call the prediction noise and the data noise. In this work, we develop and test the methods of estimating noise types from samples and then empirically measure these variance components on all pairs of models and draw some general conclusions from the observations.

For leaderboards, Chatbot arena (Chiang et al., 2024) and CRUXEval (Gu et al., 2024) computed confidence intervals using bootstrap with a fixed reference model. Their approach is needed for per-model confidence intervals but is incorrect (Example 7). This work actually compares all close pairs, and summarize all the comparisons to show that the total noise level is predictable on LLM benchmarks.

Madaan et al. (2024) measured the noise due to random seeds. Any direct approach to measure the noise necessarily ignores the data noise (Example 1). For the data noise, they also proposed using unpaired Bernoulli confidence intervals. This method was adopted by Llama 3 (Grattafiori et al., 2024) but is much looser than the paired methods (Example 5) and does not separate the data noise and the prediction noise. Improving model ability is hard and expensive while tightening the analysis is well-established. This work shows it can be done for all results together and decoupled from the main experiments.

[TODO: examples of confusion or wrong, baseline bigram, multitoken prediction]

3 Methods

3.1 Type of noise

In this section, we adopt well-studied statistical methods to estimate variance and standard errors, consistent with Miller (2024), but with emphasis on the ability to draw independent samples and the more powerful paired methods. We should clearly distinguish three types of noise: the prediction noise when answering a given question, the data noise due to using a finite sample from the population of possible questions, and their sum total noise.

Prediction noise. On any given question, LLMs have the remarkable ability of sampling many independent and diverse predictions. On long reasoning and agentic generation tasks, one sample can solve an impressive problem correctly while another sample is completely off-track, with interesting diversity in between. In contrast, humans cannot completely forget their previous thoughts and medical treatments cannot be undone to get more independent samples. While prediction noise is intrinsic to LLMs, they can be reduced by averaging over many samples per question, decreasing the temperature, fixing the random seed, or apply an aggregation method such as majority voting or verification. While averaging can reduce the prediction noise very generally, it also represents a weaker standard and is unlikely to be what you want for LLM evals.

More generally, any inference time inputs to the model, such as the choice of prompts, answer extraction process, scaffolding; and even training time inputs such as the random seed, hyperparameter, or even training steps can be sources of the prediction noise. All of which can theoretically be measured directly by repeating the process of varying the input, train the model, and then draw predictions on a given question.

[TODO: discuss how pred noise got more interesting for LLMs]

Data noise. The data noise is the noise from sampling a particular set of N questions from the population of questions. Unlike the prediction noise, the data noise can only be inferred but cannot be measured directly. Perhaps as a result, the data noise is often ignored in practice when using the training curves as a measure of noise (Example 1) and other repeated experiments on the same evaluation set such as varying the seeds (Madaan et al., 2024).

To see the importance of data noise, suppose that out of 100 True and False questions, model A is correct on 51 question, B is correct on 50 questions. Then intuitively A is not reliably better than B even if there is no prediction noise, especially if the correct answers appears at random. The bootstrap and the super-population viewpoints focus on the data noise and will show that A is not reliably better. Most work on noise before LLMs focus on the total noise or the data noise, perhaps because the prediction noise is uninteresting as in classification, or not measurable like in most non-digital experiments.

Example 1 (insufficient training curves)**.** [TODO: a figure for this] The *training curve* plots the evaluation result as a function of training steps and graphically shows the variance of evaluation results as a function of training steps. While separated training curves is a necessary condition for models to be statistically different, it is not sufficient. The training curve only shows the prediction noise due to sampling and due to training, it completely ignores the data noise since the exact same evaluation data is used. Suppose the training data of models A and B consists of a random sample of 100 evaluation questions, each included with probability 0.5. Suppose that B is 3% more contaminated than A just due to noise. By design, the procedure leading to B is not better than A, however B may have a consistently higher training curve if both A and B memorize all evaluation questions in their training data and guess randomly on the rest. In this example, the paired testing method will likely tell us that B is not better than A if we also consider the data noise.

3.2 Standards of significance

The following example shows how different predictions corresponds to different standards of significance and noise types. Suppose we have 100 True/False questions, model A has an expected accuracy of 60% and Model B 61% on each question.

Best prediction. Since the probability of correct is greater than half, both models can achieve 100% accuracy if they predict by majority voting or thresholding. **Conclusion**: A is the same as B. This is used by most classification evaluations and corresponds to the *total noise*.

One prediction. One sampled prediction per question is evaluated, so A and B do about 60% with a 7% standard error on the difference. **Conclusion**: A is not significantly different from B. This standard is based on a single answer per question, is used in most of the sciences and tests for humans, and corresponds to the *total noise* as well.

Expected prediction. We draw enough samples per question to be sure that A is 60% and that B is 61%. **Conclusion**: A and B are significantly different. This weaker standard is used by many papers of the area, and corresponds to the *data noise* after averaging away the prediction noise.

The weaker expected prediction standard allows us to more easily get statistically significant results if we fail the two stronger standards. While the underlying distributions are indeed different by design, this difference is not detectable by an user who only gets one prediction per question from the model. This standard is legit when there are a lot of users / predictions, when each prediction corresponds to utility such as revenue or user satisfaction, but it seems too weak for LLM evals. Elaborating on the best prediction, examples 2 and 3 show when the expected accuracy is not the most meaningful metric and the best prediction should be obtained whenever possible. Example 4 show that arbitrarily small but consistent improvements are significant under this standard.

For completeness, if we only care about the specific questions in an evaluation, then we would ignore the data noise altogether and just use the prediction noise. There are real life examples when the goal is to memorize a specific list of questions (driving knowledge test), or to solve specific important problems (AGI/ASI, P vs. NP). For LLMs evals, this standard does not seem to have a legit use.

Example 2 (majority voting)**.** Suppose the expected accuracies on all questions are the same with $\mathrm{E}[A] = 0.6$, $\mathrm{E}[B] = 0.9$ where the expectation averages out the prediction noise. This big improvement in expected accuracy might be trivially achievable by majority voting if we try to put forward the best prediction instead. Whenever the equivalence of sampled predictions can be checked, majority voting can increase the accuracy to 100% if the equivalence class with the correct answer has the highest probability, which is the case when $\mathrm{E}[A] > 0.5$.

Example 3 (reliable verification)**.** Suppose that on half of the questions $\mathrm{E}[A] = 0.1$ and $\mathrm{E}[B] = 1$, but on the other half of questions $\mathrm{E}[A] = 0.1$ and $\mathrm{E}[B] = 0$. While $\mathrm{E}[B] = 0.5 > \mathrm{E}[A] = 0.1$ with B having a much higher average, A is better if the predictions can be verified (search problems or formal proofs) where A is 100% with a verifier but B is only 50%. If a reliable verifier is available, then the *pass-at-large-K* metric is a more meaningful metric than the average.

Example 4 (weak expected prediction standard)**.** Arbitrarily small improvement is significant under this standard. Suppose that on question x_i, $\mathrm{E}[A(x_i)] = p_i$, $\mathrm{E}[B(x_i)] = p_i + e$, then with enough samples we can conclude that A and B are significantly different no matter how small e is. Such difference might be easy to get when the baseline did not try hard to produce its best prediction, then sharpening the distribution, deterministic filter on observed error types, a small amount of in-domain data are all likely to produce very small but consistent improvements.

Paired analysis. Whereas insights, skills and resources are required to improve LLM ability, the paired methods can tighten the analysis and should be recommended as enormous resources are devoted to improving models. The paired method makes use of how different the predictions are to gain some information on model differences. To see why paired method is better on real data, Figure 2 shows the heatmap of correctness. We can observe that models at a similar overall accuracy (close in x-axis) also does similarly on the individual questions and thus the paired method has the potential to reduce the data noise significantly. Example 6 gives the intuition on how paired comparison works.

Example 5 (HumanEval)**.** For the popular HumanEval dataset, $N = 164$, say a model has $\bar{A} = 0.5$ accuracy. The unpaired standard error (SE) is just a function of the accuracy $\sqrt{1/164 \times 0.5(1-0.5)} = 4\%$. Thus we need $4\% \times \sqrt{2} \times 1.96 = 11\%$ to get 0.05 p-value. With a paired test, the total SE is typically still 4%, so a difference of $4\% \times 1.96 = 8\%$ is needed to produce a 0.05 p-value. Both are larger than the gains reported by many papers. For example, Grattafiori et al. (2024) reports the unpaired 95% confidence intervals, but a factor of $\sqrt{2}$ is needed when comparing pairs of models, leading to many insignificant results because of a weak analysis method. The data SE can be as small as 1% between pairs of models in the same series.

Example 6 (Paired vs unpaired)**.** In the unpaired case, suppose A scored 50% on this year's exam, B scored 60% on last year's exam where each exam contains different questions drawn from the same distribution. So B might have done better because last year's exam was easier due to randomness. In the paired case, suppose B scored 60% on the same exact exam as A with exactly the same questions, then the same 10% difference is more indicative of the better performance by B.

All-pairs paired analysis. We measure the paired standard error $\mathrm{SE}(A - B)$ of all model pairs A and B, where the total noise is measured, and the data and prediction noise measured whenever possible. Example 7 show that this method could be uninformative given arbitrary result pairs, but for LLM evals this is useful. Figure 1 shows that these measurements vs the accuracy of A. We observe the range of each type of noise and that the total noise is quite predictable.

Example 7 (Error bars for leaderboards)**.** Gu et al. (2024); Chiang et al. (2024) tried to put error bars on a leaderboard. To do this, one can use the unpaired method, or a fixed baseline for paired comparison. However, this not generally meaningful: let A be the baseline, B and B' are a pair of similar predictions very different from A, so $\mathrm{Var}[A-B] \gg \mathrm{Var}[B-B']$. This example shows it's not in general possible to have one error bar per model, or per accuracy with the paired method. However, at least for LLM evals, this work shows that there is almost no such surprises, thus we can compute meaningful error bars per benchmark, with just dependence on the accuracy.

3.3 Setup and notations

To make these concepts precise, there are N evaluations *questions* $\{x_1, \ldots, x_N\}$, which are most commonly just text prompts but may include complex tasks with multi-modal data or environments for the model and agent scaffold. The model makes prediction $\hat{y}(x)$ and we get $\mathrm{metric}(\hat{y}, x)$, typically 0 for incorrect and 1 for correct but can be the mean or a real number. We use $A(x) := \mathrm{metric}(\hat{y}(x), x)$ to denote the metric function evaluating model A on the question x.

On a evaluation, we compute the average of all N questions to get the mean accuracy

$$\bar{A} := \frac{1}{N}\sum_{i=1}^{N} A(x_i) \approx \mathrm{E}_x[A(x)].$$

The estimated variance of A and standard error (SE) of the mean $\bar{A}$ are respectively

$$\mathrm{Var}_x[A(x)] \approx \frac{1}{N}\sum_{i=1}^{N}\left(A(x_i) - \bar{A}\right)^2,$$
$$\mathrm{SE}(A, N) = N^{-1/2} \cdot \mathrm{Var}[A]^{1/2}.$$

SE decreases at a rate of $N^{-1/2}$ and also depends on the population variance $\mathrm{Var}[A]$ where $\mathrm{Var}[A] \leq 1/4$ for binary correctness $A \in \{0, 1\}$ and can usually be estimated accurately. N should be large enough for estimating the variance accurately, so a bias correction on $\mathrm{Var}_x[A]$ using $1/(N-1)$ is not useful.

With models A and B, the score difference $\bar{A} - \bar{B}$ can be compared to $\mathrm{SE}(A-B)^2 = \mathrm{SE}(A)^2 + \mathrm{SE}(B)^2$ ($\approx 2\,\mathrm{SE}[A]^2$ if A, B are close) to determine if the difference is likely due to chance. While this is simple and correct, it is also unnecessarily weak (Example 5).

Paired comparison. Whereas innovation, insight, or compute is required to get real signal, tightening the analysis is easy with the paired methods and should be recommended generally. $A(x)$ and $B(x)$ are correlated when the same question is used to evaluate both $A(x)$ and $B(x)$. To be more powerful for free, we can use the paired variance

$$\mathrm{Var}_x[A(x) - B(x)] = \mathrm{Var}_x[A] + \mathrm{Var}_x[B] - 2\mathrm{Cov}_x[A(x), B(x)].$$

The Cov term may potentially reduce the paired variance to 0 when A and B are perfectly correlated.

Once we have the standard error, we obtain the z-score $z = \frac{\bar{A}-\bar{B}}{\mathrm{SE}\,\mathrm{Var}[A-B]}$ which follows a standard normal distribution. For example, $\Pr[|z| > 1] = 0.32$ is very weak, $\Pr[|z| > 5] < 10^{-6}$ is beyond doubt, and $\Pr[|z| > 1.96] = 0.05$ for a p-value of 0.05 is a reasonable conventional standard.

3.4 Sampling multiple predictions on each question

To capture this setting, we use ϵ for the seed generating the noise, which is independent of the question x, thus allowing the concise notation $\mathrm{Var}_x[\mathrm{E}_\epsilon[A]] = \mathrm{Var}_x[E_{\epsilon|x}[A(x,\epsilon) \mid x]]$. As before, but now also averaging over the K samples for each question x, we consider the average score

$$\bar{A} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K} A(x_i, \epsilon_{ij}) \approx \mathrm{E}_{x,\epsilon}[A(x,\epsilon)],$$

The estimated variance is then

$$\mathrm{Var}_{x,\epsilon}[A(x,\epsilon)] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K}(A(x_i, \epsilon_{ij}) - \bar{A})^2.$$

We have the law of total variance satisfied by the components

$$\mathrm{Var}_{x,\epsilon}[A] = \mathrm{Var}_x[\mathrm{E}_\epsilon[A]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[A]].$$

$\mathrm{Var}_{x,\epsilon}[A]$ is the total variance of first drawing a question x from the pool of questions, and then drawing a sample answer for this question. $\mathrm{Var}_x[\mathrm{E}_\epsilon[A]]$ is variance of the expected score $\mathrm{E}_\epsilon[A]$, which we call the data noise. $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ is the variance due to sampling from the model, which we call the prediction noise. See A.1 for a discussion on why $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ is the correct prediction variance and its other decomposition $\mathrm{Var}_\epsilon[\mathrm{E}_x[A]]$.

Thus all noises can be normalized to be the standard errors,

$$\begin{aligned} \mathrm{SE}(A,N) &= N^{-1/2}(\mathrm{Var}_{x,\epsilon}[A])^{1/2}, \\ \mathrm{SE}_x(A,N) &= N^{-1/2}(\mathrm{Var}_x[\mathrm{E}_\epsilon[A]])^{1/2}, \\ \mathrm{SE}_{\text{pred}}(A,N) &= N^{-1/2}(\mathrm{E}_x[\mathrm{Var}_\epsilon[A]])^{1/2}. \end{aligned}$$

Paired comparison. When we have two models A and B, we can consider the paired standard deviation $\mathrm{Var}_{x,\epsilon}[A(x,\epsilon) - B(x,\epsilon)]$, which also satisfy the law of total variance on $A - B$,

$$\mathrm{Var}_{x,\epsilon}[A - B] = \mathrm{Var}_x[\mathrm{E}_\epsilon[A - B]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[A - B]]. \tag{1}$$

3.5 Standards of significance

[TODO: discuss trade offs of the noise types] 1) We can reduce the prediction noise $\mathrm{E}_x[\mathrm{Var}_\epsilon[A - B]]$ to 0 and keeping the same $\mathrm{E}[A]$ by fixing the random seed $A'(x,\epsilon) = A(x,\epsilon_0)$, which keeps both the total variance (1) and $\mathrm{E}[A]$ the same. 2) We can decrease the temperature, which tend to increase $\mathrm{E}[A]$ as well. 3) We can average over many predictions $A'(x,\epsilon) = \frac{1}{K}\sum_{i=1}^{K} A(x_i,\epsilon_i)$, which reduces the total variance by changing the metric from binary correctness to probability of correct. This can allow a more accurate measurement at a cost of more samples, but focusing on this has drawbacks demonstrated by Examples **??**, 2, 3. Committing to the best single answer does not have these issues.

3.6 Estimating from samples

While (1) allows us to estimate the variance components directly from K paired samples $A(x,\epsilon_j) - B(x,\epsilon_j)$ for $j = 1,\ldots,K$, this is inaccurate since the seeds or predictions are interchangeable. We can use this to derive formulas for estimating from samples accurately, as if using all $K \times K'$ pairs of collected predictions $A(x,\epsilon_j) - B(x,\epsilon'_k)$ for $j = 1,\ldots,K$ and $k \in 1,\ldots,K'$. First, the prediction variance $\mathrm{E}_x[\mathrm{Var}_\epsilon[A - B]]$ decomposes by independence on each question x,

$$\begin{aligned} \mathrm{Var}_\epsilon[A(x,\epsilon) - B(x,\epsilon')] &= \mathrm{Var}_\epsilon[A(x,\epsilon)] + \mathrm{Var}_\epsilon[B(x,\epsilon)], \\ \mathrm{E}_x[\mathrm{Var}_\epsilon[A(x,\epsilon) - B(x,\epsilon')]] &= \mathrm{E}_x[\mathrm{Var}_\epsilon[A(x,\epsilon)]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[B(x,\epsilon)]]. \end{aligned}$$

Next, the data variance $\mathrm{Var}_x[\mathrm{E}_\epsilon[A - B]] = \mathrm{Var}_x[\mathrm{E}_\epsilon[A] - \mathrm{E}_\epsilon[B]]$ decomposes by linearity of expectation. Finally, the total variance is slightly tricky, where

$$\begin{aligned} \mathrm{Var}_{x,\epsilon}[A - B] &= \mathrm{Var}_{x,\epsilon}[A] + \mathrm{Var}_{x,\epsilon}[B] - 2\mathrm{Cov}_{x,\epsilon}[A,B] \\ &= \mathrm{Var}_{x,\epsilon}[A] + \mathrm{Var}_{x,\epsilon}[B] - 2\mathrm{Cov}_x[\mathrm{E}_\epsilon[A], \mathrm{E}_\epsilon[B]]. \end{aligned} \tag{2}$$

The equality is by the law of total covariance, $\mathrm{Cov}_{x,\epsilon}[A,B] = \mathrm{Cov}_x[\mathrm{E}_\epsilon[A], \mathrm{E}_\epsilon[B]] + \mathrm{E}_x[\mathrm{Cov}_\epsilon[A,B]]$ where the second term is 0 since different random predictions on a given x are independent.

Small K correction. To estimate $\mathrm{Var}_x[\mathrm{E}_\epsilon[A]]$ from K sampled predictions per question, a correction from the direct estimator can significantly increase the accuracy. We describe the concept for the unpaired case for simplicity. By the law of total variance on $\bar{A}_K(x,\epsilon) = \frac{1}{K}\sum_{j=1}^{K} A(x,\epsilon_j)$, we have

$$\begin{aligned} \mathrm{Var}_{x,\epsilon}[\bar{A}_K] &= \mathrm{Var}_x[\mathrm{E}_\epsilon[\bar{A}_K] + \mathrm{E}_x[\mathrm{Var}_\epsilon[\bar{A}_K]] \\ &= \mathrm{Var}_x[\mathrm{E}_\epsilon[A]] + \frac{1}{K}\mathrm{E}_x[\mathrm{Var}_\epsilon[A]] \\ \implies \mathrm{Var}_x[\mathrm{E}_\epsilon[A]] &= \mathrm{Var}_x[\bar{A}_K] - \frac{1}{K}\mathrm{E}_x[\mathrm{Var}_\epsilon[A]] \\ &\approx \hat{\mathrm{Var}}_x[\bar{A}_K] - \frac{1}{K-1}\hat{\mathrm{E}}_x\hat{\mathrm{Var}}_\epsilon[\bar{A}_K] \end{aligned}$$

where the hats mean sample estimate. This correction for small K can be important because we will average over N questions. If the estimate on each question is biased by $1/(K-1)$, the average error will still be off by the bias no matter how big N is. Whereas if the estimate is unbiased on each question, then the average error can decrease as N increases. Figure 3 shows this observation on MATH500, where the small-K correction is needed to reduce the error to an acceptable level even with $N = 500$ questions. The high relative error is because the paired data noise is much smaller than the prediction noise. An instructive example is to consider when $\mathrm{E}_\epsilon[A - B] = 0$ but an estimate using K samples will not be 0 with non-zero prediction noise.

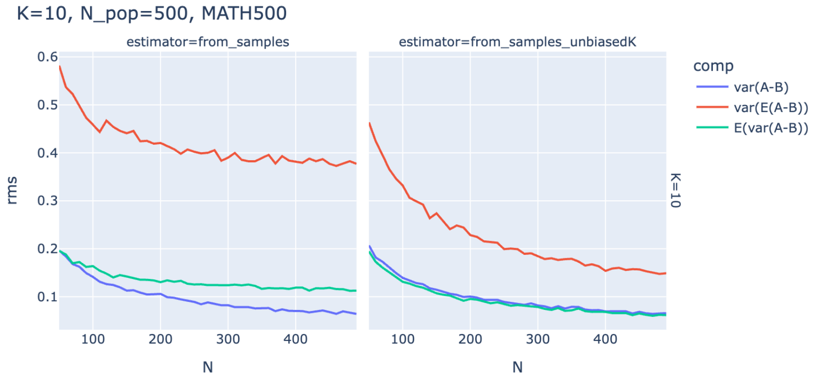

Figure 3: Relative errors of the paired noises. The left figure without the correction on K has an unacceptably high 40% relative error with 500 data points. 1000 predictions are drawn from model A and B on each of 500 questions of MATH500, which is treated as the ground truth. $K = 10$ samples are drawn from the 1000 samples and the root mean squared relative errors of variance components are plotted vs. N.

3.7 Implementation in array notation

In an experiment, we typically evaluate the model on all N questions, and may draw K answers for each question. In this section, we overload our notation with $A, B \in \mathbb{R}^{N\times K}$ and present the formulas in Section 3.6 in numpy-style code. For simplicity, we use the same number of predictions $K_i = K$ on all questions x_i. For the unpaired case, we have Table 1. For paired, we have Table 3.

Name	Formula	Code	Bernoulli
total variance	$\mathrm{Var}_{x,\epsilon}[A]$	`var(A)`	$\bar{p}(1-\bar{p})$
data variance	$\mathrm{Var}_x[\mathrm{E}_\epsilon[A]]$	`var(mean(A, axis=1))-b`	$\frac{1}{N}\sum_i (p_i - \bar{p})^2$
prediction variance	$\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$	`mean(var(A, axis=1))+b`	$\frac{1}{N}\sum_i p_i(1-p_i)$

Table 1: Formulas where $A \in \mathbb{R}^{N\times K}$ with K samples for each of N questions. `b=1/(K-1)*mean(var(A, axis=1))`

Formula	Code
$\mathrm{Var}_{x,\epsilon}[A-B]$	`var(A)+var(B)-2*cov(mean(A,axis=1),mean(B,axis=1))`
$\mathrm{Var}_x[\mathrm{E}_\epsilon[A-B]]$	`var(mean(A,axis=1)-mean(B,axis=1))-b`
$\mathrm{E}_x[\mathrm{Var}_\epsilon[A-B]]$	`mean(var(A,axis=1)+var(B,axis=1)) +b`

Table 2: code for paired estimators
`b=1/(kA-1)*mean(var(A, axis=1))+1/(kB-1)*mean(var(B, axis=1))`

3.8 Equivalence to the bootstrap and the sign test

Two other very well-established methods are the bootstrap (Efron and Tibshirani, 1986) and paired difference tests such as the sign test (Dixon and Mood, 1946). These methods ask if the observed results are likely when the examples are random for bootstrap, and when the comparison outcomes are random for the sign test. In this section we show they all give the same answer.

Bootstrap is a general method where resampling the given examples uniformly with replacement is shown to give the right answer for many estimation problems, including estimating the standard error. That is, resampling the given samples uniformly with replacement gives the same answer as getting more real samples from the population. Applied to the problem of noise, the natural question is how likely are we to get the opposite result $\mathrm{E}[A] < \mathrm{E}[B]$ as opposed to the observed result $\mathrm{E}[A] > \mathrm{E}[B]$. If the opposite outcome also has significant probability, then the conclusion is not statistically significant.

The sign test predates bootstrap and instead of sampling the questions, it makes the comparison outcome random. It supposes that any difference actually happens with a random probability $\Pr[A(x) \neq B(x)] = 0.5$ (null hypothesis) and asks how likely we are to get a more extreme outcome than what we actually observe. If it is quite likely to observe a more extreme outcome under the null hypothesis, then the conclusion is not statistically significant.

The key step in both methods are estimating the variance of their respective random procedure. If the variances are the same, then the probability to observe the opposite result from an observed mean (bootstrap) is also the same as the probability of observing a more extreme result if the real mean is 0 (sign test). The details are deferred to B.

4 Experiments

The experimental results of this paper can be found at `https://all-the-noises.github.io`, where the method is tested on more benchmarks and where the figures are interactive. We can use this to check that the main conclusion indeed holds on all the evaluations.

4.1 Exploratory analysis and findings

4.1.1 Data heatmap

To get an overview of our data, we use a heatmap shown in Figure 2. Each row is a different example, sorted from the easiest to the hardest. Each column is a model whose x-coordinate is the overall accuracy $\mathrm{E}[A]$ of this model. Like models, we might consider showing examples with y-coordinate at their average accuracy instead of rank, but that would depend on which models are included, and less intrinsic to the benchmark. We immediately observe that models at a similar overall accuracy also does similarly on individual questions. Around 1000 samples are drawn for each question and model for MATH500 in Figure 2. The SWEbench figure is based on the leaderboard, which had to commit to single answers so it is binary, but the same overall pattern holds, though with more noise. On MATH500, most of the red regions of bad models are not 0, but are rather a few percent accuracy.

4.1.2 All-pairs paired noises and predictable total noise

Figure 1 shows that the total SE agrees well with the $\mathrm{Beta}(p, 1-p)$ theory predictions described in 4.1.3 and the range of the data noise and prediction noises. Only models pairs that are close in performance are plotted $E[A] - E[B] < 5\,\mathrm{SE}(A - B)$ to avoid pointless comparisons between models whose difference is not in doubt.

The noise components can be traded off so they can depend on the setting and not as predictable. We find at more natural temperatures of 0.6-1, the prediction noise tend to be higher than the data noise. Figure 4 shows that the prediction noise dominates at 0.8 temperature, whereas the data noise dominates at 0.2 temperature, and yields about the same total noise.

4.1.3 The Beta theory makes good noise predictions and fits the data

The Beta theoretical model says that the expected accuracy each question of a model follows the $\mathrm{Beta}(p, 1-p)$ distribution where p is the mean of the model. More precisely, the expected accuracy of question x_i is $u_i \sim \mathrm{Beta}(p, 1-p)$ so that $A(x_i), B(x_i) \sim \mathrm{Bernoulli}(u_i)$ and $\mathrm{E}[A(x_i)] = u_i$. Too see that this leads to the observed SE, the models draw independent predictions on each x_i so $\mathrm{Cov}_\epsilon[A, B \mid x_i] = 0$. Thus we have $\mathrm{E}_x[\mathrm{Var}_\epsilon[A - B]] = \mathrm{E}_x[2\mathrm{Var}_\epsilon[A]]$, which integrates to

$$\int_0^1 2u(1-u)\frac{u^{p-1}(1-u)^{(1-p)-1}}{B(p, 1-p)}du = \frac{B(p+1, (1-p)+1)}{B(p, (1-p))} = p(1-p),$$

where $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}du$ is the normalization constant for the $\mathrm{Beta}(a, b)$ distribution. One consideration to fit the data better is that all models fail on some questions, so we suppose those examples actually have $u_i = 0$ instead of $\mathrm{Beta}(p, 1-p)$. Figure 5 shows the quality of fit using $\mathrm{Beta}(a, b)$. Smaller $a + b$ means more bimodal where $a + b$ tend to decrease when the model accuracy is higher, showing that better models become more bimodal whereas very bad models is

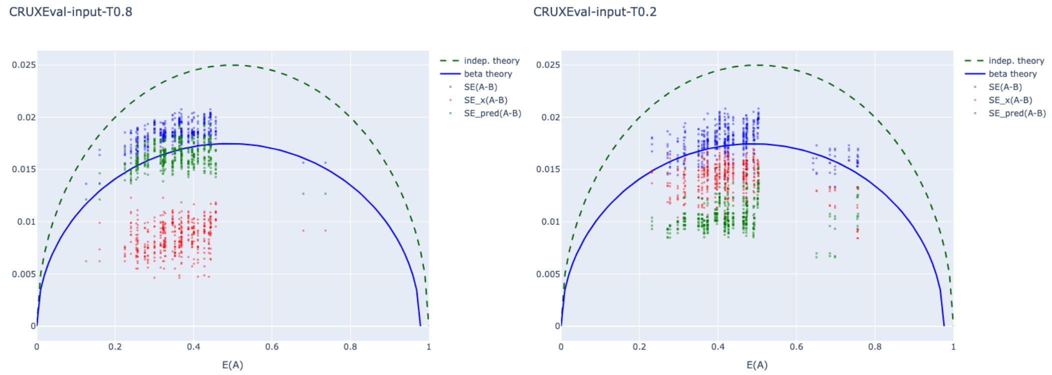

Figure 4: All noise components on CRUXEval at temperature 0.8 (left) and 0.2 (right). The prediction tend to be higher at natural temperatures 0.7-1.

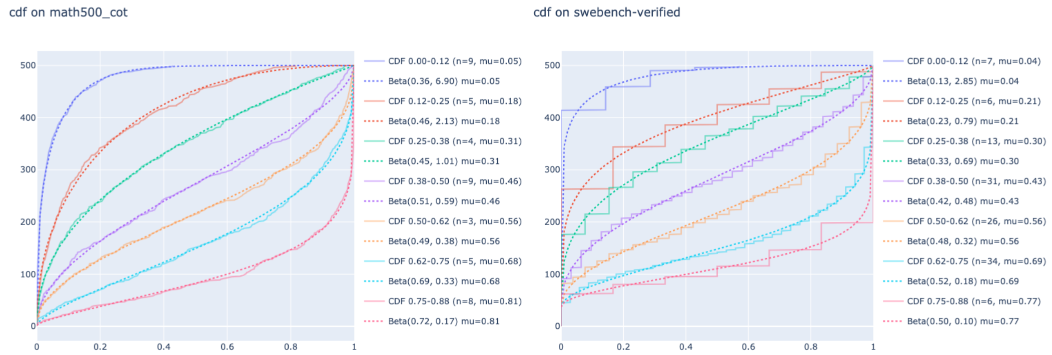

Figure 5: Empirical cummulative distribution curves (CDF, solid) and their respective Beta models (dots). Each color represents a different bin by model accuracy. The average accuracy of each question in each bin of models is computed and sorted to produce the CDF curves. The fit is good.

usually unimodal. For an extreme example, guessing uniformly on multiple choice questions with C choices means the expected accuracy $u_i = 1/C$ for all questions, which is has a single mode at $1/C$.

5 Discussions

5.1 Why is noise predictable

The range of correctness evaluations $A(x) \in [0, 1]$ immediately give an meaningful upperbound on the noise $\mathrm{SE}[A] \leq (\frac{1}{N} 1/4)^{1/2}$ and $\mathrm{SE}[A - B] \leq (\frac{2}{N} 1/4)^{1/2}$. The dependence on the mean accuracy p_A is $\mathrm{SE}[A - B] \leq (\frac{2}{N} p_A(1 - p_A))^{1/2}$, which is shown by the independent theory curve of Figure 1. We should expect LLM predictions to be positively correlated because questions have different intrinsic difficulties and LLMs are also trained on similar data, thus we expect $\mathrm{SE}[A - B]$ to be smaller than the independent curve.

The lowerbound is $\mathrm{SE}[A - B] = 0$ if $A(x) = B(x)$ on all x. For a strict net improvement on k questions bounded by a constant multiple of the standard error, $\mathrm{SE}[A - B] = O(N^{-3/4})$, still much smaller than the actual trend of $\mathrm{SE}[A - B] = O(N^{-1/2})$. In Figure 1, swebench-verified, we have 2 results near $x = 0.3$ that comes from SWE-Fixer (Xie, 2025), due to using a deterministic filter. Another explanation is that the prediction noise alone is also $O(N^{-1/2})$. The Beta model provides a partial explanation too, where all the between close pairs are hypothesized to come from prediction. Still, on Figure 2, it is clear there is some data noise even between close pairs. On leaderboard predictions such as SWEBench (Jimenez et al., 2024), or LiveCodeBench (Jain et al., 2024), there usually is only $K = 1$.

Test-time scaling If the correctness drawn from a distribution $p_i \sim \mathcal{P}$, then the expected correctness is $\mathrm{E}[p]$, whereas the expected majority vote is $> E[p > 1/2]$. For the beta distribution, an improvement is expected for high performing models.

5.2 Solving hard and special problems

For a problem requiring a long answer where guessing correctly is unlikely, answering even 1 problem might be significant and interesting. For example, the problem can ask for the proof of an important open problem and the test checks the proof. If a model (or someone) solves such a hard problem then we should not object to the sample size of 1. We probably don't need to consider this yet. In the empirical data, all pairs of models have enough noisy inconsistencies where model A may beat B on 2 hard examples, but then B beats A on 2 easy examples. If A is so good that it didn't make any mistakes on the long complex answer required to solve the hard problems, why did it fail on some easy ones? Second, problems solvable by mediocore models or where reference solutions can be found on the internet (i.e. training data) is unlikely to deserve special deference, which still applies to most evaluation sets.

In our human experience, we intuitively get a lot of signals by evaluating on hard problems, for example in interviews. The key is perhaps gaining much more information than just one bit of binary correctness judgement from the details of how the problem was solved. In contrast, LLM evaluations and learning settings only give back 1 bit after a lot of work. If we move in the direction of smaller and harder evals, we probably need to also generate more information per question.

[TODO: applications, mentions of appendix]

References

Yoav Benjamini, Richard D De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry I Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M Stigler, Stephen B Vardeman, Christopher K Wikle, Tommy Wright, Linda J Young, and Karen Kafadar. 2021. The asa president's task force statement on statistical significance and replicability. *The Annals of Applied Statistics*, 15(3):1084–1085.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Wilfrid J. Dixon and Alexander M. Mood. 1946. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bradley Efron. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Bradley Efron and Robert Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. 2024. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*.

Katherine Hermann, Jennifer Hu, and Michael Mozer. 2024. Experimental design and analysis for AI researchers. Tutorial at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

E. L. Lehmann and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*, third edition. Springer Texts in Statistics. Springer, New York.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*.

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

Chengxing Xie. 2025. Swe-fixer: Training open-source llms for effective and efficient github issue resolution. *arXiv preprint arXiv:2501.05040*.

A Analysis details

A.1 Alternative decomposition

Taking expectation over either the noise or the question has the same total variance and thus the sums are equal

$$\begin{aligned}\mathrm{Var}_{x,\epsilon}[A] &= \mathrm{Var}_x[\mathrm{E}_\epsilon[A]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[A]] \\ &= \mathrm{Var}_\epsilon[\mathrm{E}_x[A]] + \mathrm{E}_\epsilon[\mathrm{Var}_x[A]]\end{aligned}$$

To see that it indeed corresponds to what we want to measure,

$$\mathrm{Var}_\epsilon\left[\frac{1}{N}\sum_{i=1}^{N} A(x_i, \epsilon_i)\right] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}_\epsilon[A(x_i, \epsilon)] \to \frac{1}{N}\mathrm{E}_x[\mathrm{Var}_\epsilon[A(x, \epsilon)]]. \tag{3}$$

The left hand side is the prediction noise on N particular data points, which approach the mean if they are iid. One can directly evaluate the prediction noise on a particular dataset by running the eval K times and measure the variance of the K results. That is also correct in expectation but could be much more noisy since (3) includes equally valid samples not drawn by any particular evaluation run.

Paired prediction noise from the same model Using (2) to compute $\mathrm{Var}_{x,\epsilon}[A(x, \epsilon_1) - A(x, \epsilon_2)]$ is too small when using the same set of samples for A. Instead, we can use the following equality and an unbiased estimator for $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ with $A = A(x, \epsilon_1)$, $B = A(x, \epsilon_2)$.

$$\begin{aligned}\mathrm{Var}_{x,\epsilon}[A - B] &= \mathrm{Var}_x[\mathrm{E}_\epsilon[A - B]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[A - B]] \\ &= 0 + \mathrm{E}_x[\mathrm{Var}_\epsilon[A] + B]] \\ &= 2\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]\end{aligned}$$

Equivalently, we can still use (2) and Table 3 if we sample without replacement when estimating $\mathrm{E}[A(x, \epsilon_1)A(x, \epsilon_2)]$ to avoid the bias over-estimating the covariance using the same set of samples. Suppose we have K samples for a question x with scores $A_1, \ldots, A_k$. To estimate $\mathrm{E}[A(x, \epsilon_1)A(x, \epsilon_2)]$ we must use $\texttt{mean}_{i,j\neq i} A_i A_j = \frac{1}{K^2-K}\sum_{i,j\neq i} A_i A_j = \frac{1}{K^2-K}[(\sum_i A_i)^2 - \sum_i A_i^2]$.

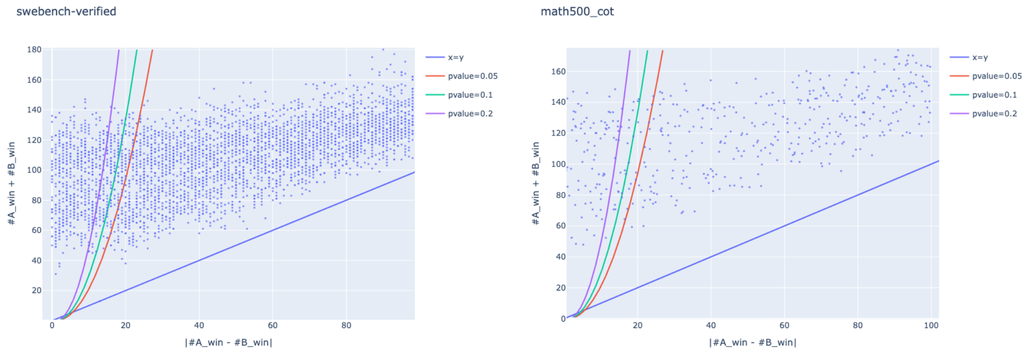

Figure 6: The amount of inconsistency.

Difference in predictions is high. [TODO: signal noise]

B Equivalence to bootstrap and sign test

To setup both methods, let $W_A := \sum_{i=1}^{N} \mathbb{I}[A(x_i) > B(x_i)]$ be the number of times model A wins against model B and vice versa for W_B, and let's assume $W_A > W_B$ for convenience. The question is how likely are we to still observe $W_A > W_B$ under the randomness prescribed by bootstrap or the sign-test.

The bootstrap. We draw another N samples with replacement from the existing samples $x_1, \ldots, x_N$ to obtain $X_A, X_B, X_0 \sim \text{multinomial}(N, q_A, q_B, q_0)$ with $X_A + X_B + X_0 = N$, where the outcome probabilites are $q_A = W_A/N$ for A win, $q_B = W_B/N$ for B win, and $q_0 = 1 - q_A - q_B$ for tie. The mean and variance of the resamples are respectively $E[X_A - X_B] = W_A - W_B > 0$,

$$\begin{aligned}\text{Var}[X_A - X_B] &= \text{Var}[X_A] + \text{Var}[X_B] - 2\text{Cov}[X_A, X_B] \\ &= N\left(q_A(1-q_A) + q_B(1-q_B) + 2q_Aq_B\right) \\ &= N\left(q_A + q_B - (q_A - q_B)^2\right) \\ &\approx N(q_A + q_B) = W_A + W_B.\end{aligned} \tag{4}$$

Under bootstrap, a natural question is how likely we would see the opposite result $\Pr[X_A < X_B]$. With an approximation that $(q_A - q_B)^2 \ll q_A + q_B$ (else the result is probably significant already), the bootstrap asks if $\Pr[z \geq \frac{W_A - W_B}{\sqrt{W_A + W_B}}]$ for standard normal z.

Directly estimating the variance from the samples also give the same answer as (4). While this is a consequence of bootstrap variance equal to the sample variance, a direct calculation is this

$$\begin{aligned}\text{Var}_x[A(x) - B(x)] &= E[(A-B)^2] - E[A-B]^2 \\ &= q_A + q_B - (q_A - q_B)^2.\end{aligned}$$

The sign test. When comparing model A vs. model B, the null-hypothesis is that model A and B each has a 1/2 chance of being better (win) on each example. A wins if A gets a question correct but B does not, tie if A and B are both correct or incorrect. The question is if the observed results are likely to happen under this null-hypothesis.

The p-values is then $\Pr[X \geq W_A]$ for $X \sim \text{binomial}(W_A + W_B, 0.5)$ with $E[X] = \frac{1}{2}(W_A + W_B)$ and $\text{Var}[X] = \frac{1}{4}(W_A + W_B)$. The question is how likely is X to be as extreme as W_A as observed. When $W_A + W_B$ is large, the normal approximation is accurate and reduces to $\Pr[z > \frac{W_A - W_B}{\sqrt{W_A + W_B}}]$ for standard normal z (i.e. one sided). If we also consider how A might be worse, then we should consider the two-sided question.

So the main empirical variance method, the bootstrap, and the sign test give the same answer. A slight approximation was used on the sign test. Without the approximation, $\text{Var}[\text{sign test}] \geq \text{Var}[\text{bootstrap}]$ due to using 0.5 for the null hypothesis instead of the empirical win rate.

Formula	**Win rates**
$\text{Var}_{x,\epsilon}[A - B]$	$q_A + q_B - (q_A - q_B)^2$
$\text{Var}_x[\text{E}_\epsilon[A - B]]$	$\frac{1}{N}\sum_i q_{A,i} + q_{B,i} - (q_A - q_B)^2$
$\text{E}_x[\text{Var}_\epsilon[A - B]]$	$\frac{1}{N}\sum_i q_{A,i} + q_{B,i} - (q_{A,i} - q_{B,i})^2$

Table 3: For where $A, B \in \{0, 1\}$ with $\Pr[A(x_i) > B(x_i)] = q_{A,i}$.