MEASURING ALL THE NOISES OF LLM EVALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Separating signal from noise is central to experiments. Applying well-established statistical methods effectively to LLM evals requires consideration of their correlations and prediction noises. We clearly define and measure three types of noise: prediction noise from generating different answers on a given question, data noise from sampling questions, and their combined total noise following the law of total variance. To emphasize relative comparisons and gain statistical power, we propose the all-pairs paired method, which applies the paired analysis to all pairs of LLMs and measures all the noise components based on millions of question-level predictions across many evals and settings, revealing clear patterns. First, each eval exhibits a characteristic and highly predictable total noise level across all model pairs. Second, paired prediction noise typically exceeds paired data noise, which means reducing prediction noise by averaging can significantly increase statistical power. By measuring all the noises together, we can assess eval results in context, lowering the barrier of using the best analysis to make sound empirical decisions.

1 INTRODUCTION

Identifying signal from noise is important for interpreting any experiment results. Well-established methods from statistics *can increase the rigor of the conclusions drawn from data* (Benjamini et al., 2021). Even though these methods are in classic textbooks and clearly described for LLM evals (Miller, 2024), there are some choices and confusions which lead to misleading measurements and low statistical power (Example 1). In this work, we take advantage of the global identity of the models and LLM eval questions to do all the noise measurements together to increase their technical quality and accessibility.

To do this, we consider the prediction noise of LLMs. Unlike previous probabilistic models that typically have a clear best prediction, LLMs produce persistently diverse predictions. For a model A evaluated on questions x with prediction seed ϵ, we have

$$\mathrm{Var}_{x,\epsilon}[A(x,\epsilon)] = \mathrm{Var}_x[\mathrm{E}_\epsilon[A]] + \mathrm{E}_x[\mathrm{Var}_\epsilon[A]],$$
$$\text{Total variance} = \text{data variance} + \text{prediction variance}.$$

The *prediction variance* $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ is due to the model generating different answers on a fixed question. The *data variance* $\mathrm{Var}_x[\mathrm{E}_\epsilon[A]]$ measures how the questions vary in difficulty level and thus vary in their expected metrics. Prediction variance can be reduced but usually persists in practice. Data variance matters for reliability beyond the specific question set. While prediction variance can be reduced by averaging, the data variance is irreducible and can be large when the questions vary in their difficulties.

While questions range in their difficulty, similar models tend to do similarly on each question, so a greatly reduced noise is possible if and only if *paired comparison* is used. The combination of paired comparison and prediction noise reduction is 3-6 times and 4 times more powerful in Example 1 and Example 2b. A difficulty is that the paired method only applies to pairs of models rather than a list of models in a table or leaderboard, thus not widely adopted. In this work, we propose the all-pairs paired method to measure all the noises, which uses the paired analysis on all pairs of models to measure the data, prediction, and total paired variances $\mathrm{Var}[A-B]$, for all model pairs A, B, and across many evals and settings, revealing clear patterns applicable to a table of results.

We highlight two observations and their applications. First, each eval has a characteristic total noise level that holds across all pairs of LLMs or agents. Specifically, the paired test usually gives a total paired variance similar to the basic variance predicted by the accuracy p_A of model A, resulting in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

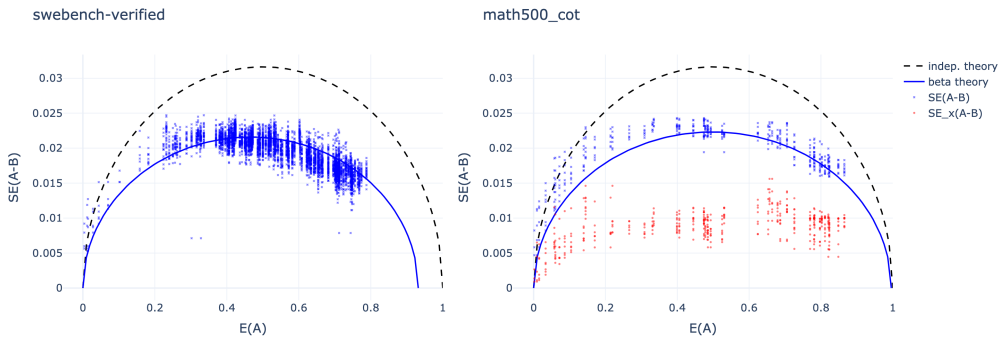

Figure 1: paired total standard errors SE(A-B) vs. the accuracy E(A). Each dot is a paired comparison of model A vs. B, agreeing with the beta theory (trend line). Left: on SWEBench estimated from 1 sample per question, Right: on MATH500 from multiple samples per question and thus also showing the irreducible data noise SE_x(A-B).

this rule of thumb on correctness evals

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

For better accuracy, Figure 1 Left shows the paired total noise on SWEBench-Verified against our beta theory prediction. This allows us to have accurate-enough noise measurements without custom analysis and even to infer the likely significance of results reported by others.

Second, by clearly separating prediction noise from data noise, we find that the paired prediction noise is typically greater than the paired data noise. For example, Figure 1 Right shows that the prediction noise is around two times the data noise on MATH500, which means that reducing the prediction noise can lead to a significant increase in statistical power, or a significant decrease in the minimum detectable effect size (Example 1). Even greater reductions can be expected for controlled experiments between close model pairs. When this condition holds, inferring the noise from the training curve becomes useful (Example 2). The importance of these choices means that we must specify what noise we are measuring: data, prediction, or total noise using paired vs. unpaired method; otherwise we cannot interpret reported noise measurements (Chan et al., 2025; Merrill et al., 2026).

We speculate that the uniformity of these results comes from the relative similarity among LLMs training on similar data using similar methods. Like models, eval questions also have a global identity analogous to the ability to repeat experiments on the same subjects anywhere. By summarizing the noise measurements on many pairs of results, we can put noise measurements in context and infer their likely values, lowering the barrier of the more powerful paired analysis.

Contributions. We clearly define and measure prediction noise, data noise, and total noise on many models and evals. Our general findings enable practitioners to assess significance without custom statistical testing, and often allow us to detect much smaller effects.

The data and analysis can be found at `https://all-the-noises-anony.github.io`, which shows the results of this paper in interactive figures across evals, temperatures, based on millions of question-level metrics from public leaderboards and controlled experiments.

2 Related Work

As LLM evals became important itself, there is interest to put error bars on leaderboard and results, Chatbot arena (Chiang et al., 2024) and CRUXEval (Gu et al., 2024) computed confidence intervals using bootstrap with a fixed reference model, which allowed them to obtain per-model confidence intervals but this is impossible in general (Example 4). For example, Chatbot arena aggregates comparisons with all other models to compute their confidence intervals, which necessarily loses any special properties of a particular pair of models. Gu et al. (2024) uses a fixed reference model which may also fail to capture any special pairs. Other important model (Grattafiori et al., 2024)

and benchmarking work report unpaired standard errors which is unnecessarily loose and sometimes just incorrect (Merrill et al., 2026). This work actually compares all pairs of models, and uses the summary statistics of all the comparisons to get meaningful noise levels. More importantly, we show that it is possible to get meaningful graphical summaries and do this in a separable way (§D).

Madaan et al. (2024) measured the noise due to random seeds, which can be seen as a type of prediction noise. Any direct approach to measure the noise necessarily ignores the critical data noise (Example 2). 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 1) and does not separate the data noise and the prediction noise. Hägele et al. (2026) propose a bias-variance decomposition on KL, and studied the ratio of prediction noise to total noise. Recent works (Romanou et al., 2026; Haase et al., 2026; Messing, 2026, among others) proposed more detailed decompositions and study sources of variations such as prompt and judge. Miller (2024) considered a impactful clustered correction. This work attempts to keep the core decomposition method as simple as possible to focus on what can be learned from the empirical all-pairs results and correctness (§A,F).

(Lehmann and Romano, 2005; Dror et al., 2018; Hermann et al., 2024; Card et al., 2020, among others) study noise more generally. We adopt the approach of Miller (2024) to estimate the variance directly. While using samples and paired comparisons is standard statistics, their approach is clarifying. The basic paired method is described by Wasserman (2004) as *paired comparison* in Wald test examples. This method is also consistent with the bootstrap (Efron, 1979) and the sign test (Dixon and Mood, 1946) when applicable (§E). Efron and Tibshirani (1986) clearly described estimating the standard error as the example of not needing bootstrap. The concept of total variance decomposition is present in 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. They also advocated for averaging (resample), but do not provide any data on how much this matters. In this work, we develop and test the methods of estimating noise types from samples and actually measure these variance components on many pairs of models across many evals.

3 Main concepts and examples

3.1 Types of noise

We should clearly distinguish three types of noise: the prediction sampling noise when answering a given question, the data sampling noise due to using a finite sample from the population of possible questions, and their sum total noise.

Prediction noise. The prediction noise comes from generating different answers on a given question with the same model. LLMs have the remarkable ability to sample independent predictions that are especially diverse on reasoning and agentic questions. In the physical world, the prediction noise typically cannot be measured directly — humans cannot completely forget their previous thoughts and medical treatments cannot be undone to get more independent samples. Previous probabilistic ML models such as classifiers can typically make their best prediction directly and do not have interesting prediction noise.

The prediction noise is also persistent even though we can try to reduce it by fixing the random seed, decreasing the temperature, averaging over many samples per question, or applying an aggregation method such as majority voting and verification. Fixing the random seed or lowering the temperature is brittle, which only works on the exact same model. Trying to produce the best answer by aggregation is perhaps the most sound method, though this still leaves some prediction noise due to the diversity of samples. Among these noise reductions, only averaging and fixing the random seed gives the same mean while others change the mean as well. Averaging is the most general and effective, which we will focus on.

Other than sampling, additional prediction noise can come from inference-time choices such as prompt formatting, hyperparameters, and answer extraction; and training-time choices such as the random seeds, hyperparameters, and training steps. All of these can in theory be measured directly by changing the seed (for each question) then drawing a new prediction. While training is not typically considered a noise, it can result in more noise than signal over a small number of steps.

Data noise. The eval data sampling noise comes from sampling a particular set of N questions instead of another equally good N questions from the unobservable population of questions. As long

as the model has different expected accuracy on different questions (i.e. some questions are harder), the mean will vary when a different set of questions is sampled. Unlike the prediction noise, the data noise cannot be directly measured by evaluating on the fixed, given set of questions; and it cannot be reduced. Though the data noise can be inferred using the bootstrap or the variance estimator, it is not observed in experiments such as the training curve (Example 2) and varying the training seeds (Madaan et al., 2024).

To see the presence of data noise, suppose that out of 100 True and False questions, model A is correct on 50 random questions, B is correct on 52 random questions out of the same 100 questions. Then B is not reliably better than A even when each question is answered deterministically with no prediction noise. Suppose we have $p = 0.5$ chance of sampling a question that is answered correctly by A, then the number of correctly answered questions has a standard deviation of $(100p(1-p))^{1/2} = 5$, and thus it is typical to see a difference of more than 2 questions. Statistical methods such as computing the standard error or bootstrap focus on the data noise and will show that A is not reliably better in this case.

If we only care about a specific set of questions, then we can ignore the data noise. There are real life situations when the goal is to memorize a specific list of questions such as a driving knowledge test where the set of questions is published, or to solve specific important problems (AGI/ASI, P vs. NP). On most evals, we care about the underlying ability measured by the eval rather than the specific questions of the eval, so the data noise must not be ignored.

Total noise = prediction + data noise. We can make this relation precise using the law of total variance. The extensive literature about noise focuses on the total noise or the data noise, perhaps because the prediction noise is uninteresting as in classification, or not measurable directly as in most non-digital experiments.

Paired analysis. Whereas insights, skills, and resources are required to improve LLM ability, the paired methods can increase the signal significantly. Thus they should be recommended, especially given the enormous resources devoted to improving models. The paired method makes use of prediction differences on the same question to gain more information. To see why the paired method is better on real data, Figure 3 shows the heatmap of correctness where models at a similar overall accuracy (close in x-axis) also do similarly on the individual questions. Thus, the paired method has the potential to reduce the data noise significantly. Examples 2b and 1 show the potential increase in power using paired methods, especially when prediction noise > data noise, leading to the ability to detect differences several times smaller. Example 3 shows the basic idea of paired analysis.

Example 1 (HumanEval)**.** HumanEval (Chen et al., 2021) has $N = 164$ questions, if a model has $\bar{A} = 0.5$ accuracy, then the unpaired standard error (SE) is $\sqrt{1/164 \times 0.5(1-0.5)} = 4\%$. Since a factor of $\sqrt{2}$ is needed when comparing pairs of models, we need $4\% \times \sqrt{2} \times 1.96 = 11\%$ to get 0.05 p-value. The table below shows the reductions with paired and averaging, and how the combination can significantly reduce the minimum detectable difference between typical pairs of LLMs.

Test Type	Standard error of $\bar{A} - \bar{B}$	Difference for $p < 0.05$
Unpaired	6%	12%
Unpaired + averaging	4%	8%
Paired	4%	8%
Paired + averaging	1–2%	2-4%

Table 10 of Llama 3 (Grattafiori et al., 2024) reports the unpaired 95% confidence intervals of around 7%, thus many of the comparisons do not clear the $\sqrt{2} \times 7\%$ difference with their method.

Example 2. a) The training curve is invalid in theory. The *training curve* plots the eval result as a function of training steps and graphically shows the variance of eval results as a function of training steps (Figure 2, col 1). While separated training curves are intuitively needed for reliable results, it is not sufficient. The training curve is insufficient because it only shows the prediction noise and completely ignores the data noise as it evaluates on the same questions repeatedly. In theory, the data noise can be much higher than the prediction noise. The prediction noise can be zero with deterministic predictions and convergent training.

b) The training curve is valid when prediction noise is dominant. In practice, if the prediction noise dominates the data noise, then the training curve method is valid provided there is a margin of separation. Figure 2 shows increasingly powerful analysis on the training curve (cite anon.) visualized by the bootstrap.

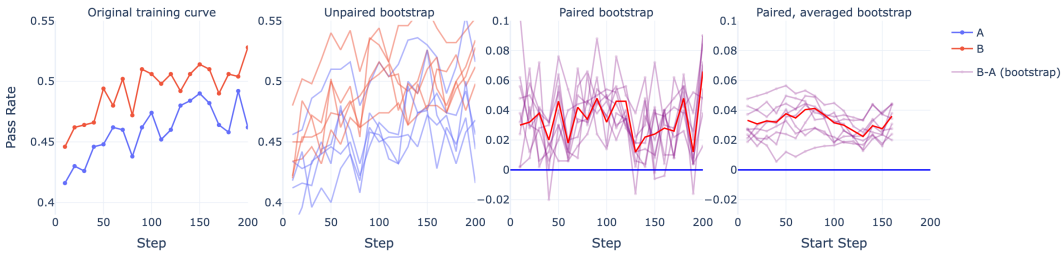

Figure 2: Increasingly sensitive training curves on SWEBench-Verified with $N = 500$, A and B both start from the same checkpoint at step 0. **Column 1) Original training curve.** One prediction accuracy, showing B is better than A over all steps. **2) Unpaired bootstrap.** A is better than B on many resampled set of 500 questions, showing a meaningless difference using this analysis method. **3) Paired bootstrap.** Slightly significant with a z-score of 1.7 at 1 prediction per question. **4) Paired, averaged bootstrap.** The result is very significant with a z-score of 3.5 when the question level predictions are averaged over 5 consecutive training checkpoints to reduce the prediction noise.

Example 3 (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 contained easier questions. In the paired case, suppose B scored 60% on the same exam as A containing exactly the same questions, then the same 10% difference is more significant.

All-pairs paired method. We measure the paired standard error $\mathrm{SE}(A-B)$ of all model pairs A and B, where the total noise is always measured, and the data and prediction noises are measured whenever possible. While this approach can be inconclusive in general (Example 4), our measurements show that on LLM evals this turned out to be quite informative, so we highlight the following two findings.

Finding: predictable total noise. Figure 1 shows the total noise vs. the accuracy of A. We observe that the total noise is quite predictable and agrees well with a theoretical model. This means each eval has a characteristic total noise as a function of the accuracy. This is the right noise number to use for experiments where one prediction is made per question.

Finding: typically prediction noise $>$ data noise. When this holds, we can gain much statistical power by using the expected metric where the prediction noise is averaged away to be comparable to the data noise remaining. Figure 1 Right shows that on MATH500, the data standard error (SE) is typically less than $1/2$ of the prediction SE between different final models. Example 2b and Figure 2 show training curves evaluated on SWEBench-verified starting with the same checkpoint, the data SE is $\sim 1/6$ of the prediction SE. If we average over 36 predictions, then the total SE would be $\sqrt{2}/6 = 0.24$, allowing us to detect effects that are 0.24 the size with the same p-value and power. Example 1 shows that the reduction on HumanEval is smaller but still a factor of 2. For controlled experiments, this allows us to detect differences several times smaller at the same significance level and power. Important caveats are discussed in §C.1.

Example 4 (Error bars for leaderboards)**.** Gu et al. (2024); Chiang et al. (2024) tried to put error bars on a leaderboard. To do this, they use a fixed baseline for paired comparison. However, this is 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, we can summarize the results of all the paired comparisons like in Figure 1. For the total noise of LLM evals, this work shows that there are very few surprises, and thus we can compute meaningful error bars. When there are exceptions, the proposed figure still shows them clearly.

4 METHOD

4.1 SETUP AND NOTATIONS

To make these concepts precise, suppose there are N eval questions $\{x_1, \ldots, x_N\}$, which are typically text prompts. The model makes prediction $\hat{y}(x)$ and we get $\text{metric}(\hat{y}, x)$, typically 0 for incorrect and 1 for correct but can be a real number or an aggregate metric. Any information needed to compute the metric such as the ground truth answer or agent environments are included implicitly. We use $A(x) := \text{metric}(\hat{y}(x), x)$ to denote the metric function evaluating model A on the question x.

On an eval, we compute the average of all N questions to get the mean metric

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

The estimated variance of A and standard error (SE) of the mean are respectively

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

where the standard error is normalized to the right scale for comparison to the effect size. We just say *noise* when there is no need to distinguish the variance and the standard error. We avoid the convention that corrects the bias and increase the error of $\mathrm{Var}_x[A]$ using $1/(N-1)$.

With models A and B, the score difference $\bar{A} - \bar{B}$ can be compared to $\mathrm{SE}(A - B) = (\mathrm{SE}(A)^2 + \mathrm{SE}(B)^2)^{1/2}$ to determine if the difference is likely due to chance. While this is simple and correct, it is also unnecessarily weak (Example 1).

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.

From the standard errors we can obtain the z-score $z = \frac{\bar{A}-\bar{B}}{\mathrm{SE}[A-B]}$. When $N > 100$ is moderately large, then the Central Limit Theorem applies and the z-scores can be converted to p-values. 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.

4.2 SAMPLING MULTIPLE PREDICTIONS ON EACH QUESTION

To capture this setting, we use ϵ for the random seed that generates the prediction, 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

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

The estimated variance is then

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

While we can use $\bar{A}_K$ for the metric function, we decompose into components that are independent of K but can be estimated using K samples per question,

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

$\mathrm{Var}_{x,\epsilon}[A]$ 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 the variance of the expected score $\mathrm{E}_\epsilon[A]$, which we call the data variance. $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ is the variance due to sampling from the model, which we call the prediction variance. See §A.1 for why $\mathrm{E}_x[\mathrm{Var}_\epsilon[A]]$ is the same as the prediction variance that can be directly measured on a fixed eval.

All the variances can be normalized to be their respective standard errors,

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

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

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

See §A for how to estimate from samples, the importance of small-K correction, and implementation details.

5 Experiments

The experimental results of this paper can be found at `https://all-the-noises-anony.github.io`, where the method is tested on more evals and where the figures are interactive to support more detailed investigations and case studies. The readers are invited to take a sample and check that the main observations hold on most of the evals. See §C.2 for a discussion of some exceptions. The rest of this section discuss a few data visualization methods that are helpful for understanding noise.

5.1 Exploratory analysis and findings

5.1.1 Data heatmap

To get an overview of our data, we use a heatmap shown in Figure 3. Each row is a different question, 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 questions with y-coordinate at their average accuracy instead of rank, but that would depend on which models are included, and less intrinsic to the eval. We immediately observe that models at a similar overall accuracy also do similarly on individual questions. Around 1000 samples are drawn for each question and model for MATH500 in Figure 3. The SWEbench figure is based on the leaderboard data, which had to commit to single answers, but the same overall pattern holds, though with more noise. On MATH500, most of the red regions of bad models are still non-zero.

5.2 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 5.2.1 and the range of the data noise and prediction noises. Only model 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. Figure 4 plots the paired SE vs $E(A - B)$ to show the narrow range where the stat. sig. is questionable.

The noise components can be traded off to some extent so they can depend on the setting and not as predictable as the total. We find at more natural temperatures of $[0.7, 1]$, the prediction noise tends to be higher than the data noise across evals. Figure 5 shows the noise components at two different temperatures on CRUXEval, where the breakdown differs but the total noise remains the same.

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

The $\mathrm{Beta}(p, 1-p)$ distribution, where p is the model's mean accuracy, fits the empirical distribution of question-level expected accuracies well and predicts a paired prediction variance of $p(1-p)$, explaining the observed trend in Figure 1. The details are in §B.

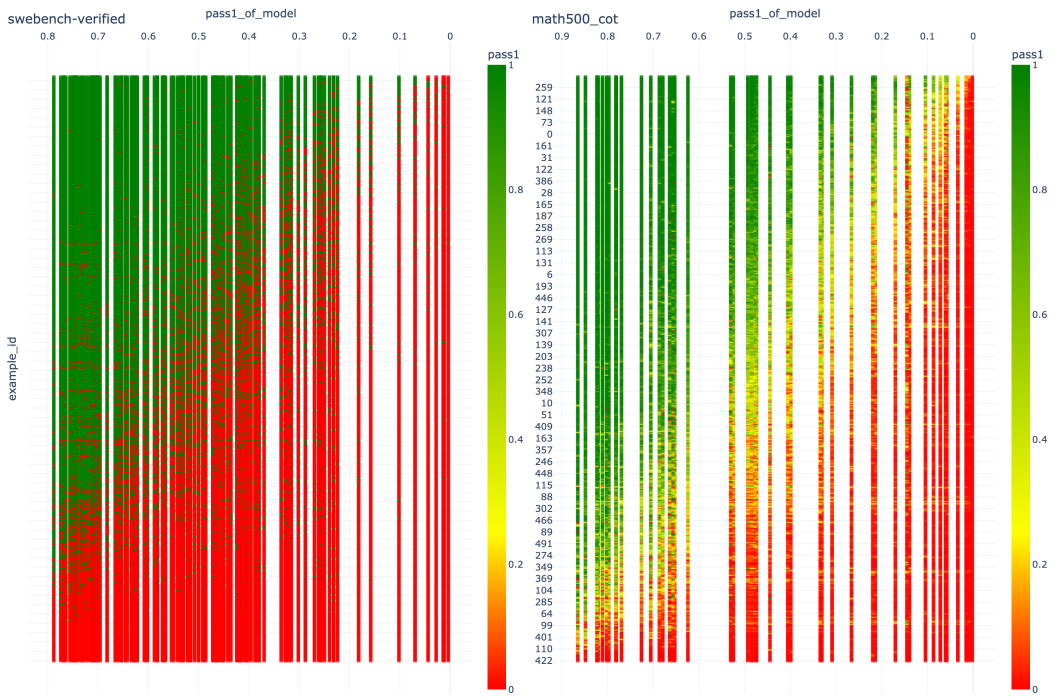

Figure 3: Results heatmap. Each row is a different question, sorted from the easiest to the hardest. Each column is a model whose x-coordinate is the overall accuracy. Left: SWEbench-verified, 1 prediction per question; Right: MATH500, 1000 predictions per question. See §5.1 for more details.

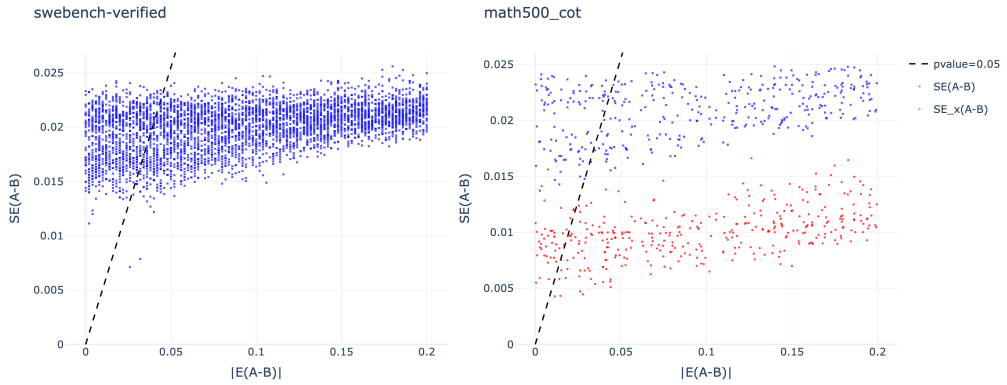

Figure 4: SE(A-B) vs. the difference E(A-B). constant p-values are straight-lines on this figure and every model pair to the right of the line are more significant.

6 Discussions

6.1 Remarks on the main findings

Predictable total noise. When the prediction noise is much greater than the data noise, then total noise $\approx$ prediction noise, and prediction noise is predicted by the Beta theory, so the predictable total noise is a consequence of prediction noise $\gg$ data noise. However, Figure 5 shows non-negligible data noise, yet the total noise still follows the Beta theory prediction even when the data noise exceeds the prediction noise at low temperature, except for chain-of-thought predictions. More data is needed to further test if this continues to hold, and if low temperature can reduce prediction noise as the model is training.

Prediction noise $>$ data noise. It is not surprising that current LLMs are similar to each other on question level metrics, as they are trained on similar data using similar architecture and learning

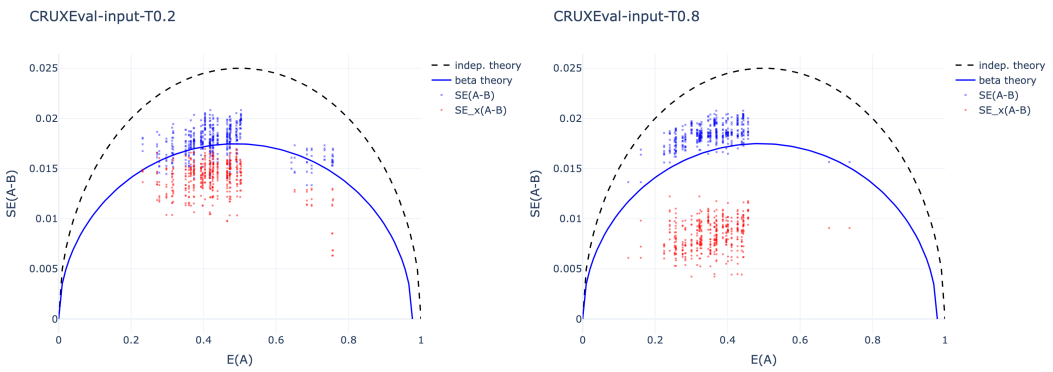

Figure 5: All noise components on CRUXEval at temperature 0.8 (left) and 0.2 (right). The prediction noise dominates at 0.8 temperature, whereas the data noise dominates at 0.2 temperature, while both still yield about the same total noise.

methods. So it is expected that LLMs perform similarly on each question if they have similar overall abilities. In theory, the prediction noise can still be arbitrarily small, but LLMs seem to be both positively correlated and have sufficient prediction noise for this result to hold.

Another implication is that modeling question difficulty using the Item Response Theory (IRT) or filtering questions (Polo et al., 2024) should deal with the prediction noise first. Section 5 of Madaan et al. (2024) and (anonymous) find that IRT does not work well. The data filtering method $\tau-$ of (anonymous) finds questions that are inversely correlated with the overall model performance. When there is high prediction noise, there is no way to get a reliable signal from a small number of questions or to reliably blame the questions for noise.

6.2 Solving hard and special problems

Leading agentic evals often have fewer than 1000 questions, with terminal bench (Merrill et al., 2026) having fewer than 100 questions. In contrast, 100 multiple choice questions clearly have a lot of noise when questions are guessed. Perhaps the reason for believing that very small evals are okay in the agentic setting is that the setting can capture very hard problems that are not guessable. For example, if a model solved an important open problem then we would not object to the sample size of one. Here the prior probability of success is near 0, so while observing a success gives a lot of information, we are expected to fail and get no information. In other words, either we get no information or the problems are not that special. More directly, one can still use the mixed predictor (3) to generate variance from the differences in model predictions, just like guessing. Such evals still need to satisfy the regular statistical standard to be meaningful.

In human evaluations, we seem to get a lot of signal by evaluating on hard problems with a number of steps (e.g. interviews, math). The key is to reflect on the details of how the problem was solved, and if it seems to be memorized. In contrast, LLM evals and RLVR settings only give 1 bit back after doing a lot of work. If we continue in the direction of smaller and harder evals, we probably need to also generate more information per question.

6.3 Applications and Recommendations

Practical recommendations for practitioners and eval builders, including a noise-aware method for meta-analysis across multiple evals, are in §D.

6.4 Caveats of the expected metric

We recommend averaging for most controlled experiments not expected to have a large effect. However, there are caveats when targeting a particular eval metric or when the variable affects the distribution's sharpness. See §C.1 for detailed examples.

6.5 Limitations

While the method is general, all the empirical data are from correctness evals, so whether the empirical findings generalize to other evals remains to be tested. Bowyer et al. (2025) show that CLT does not work well for less than a few hundred questions, though with 100 questions the errors are only significant in the tails. For simplicity, we did not consider the potentially impactful clustered correction of Miller (2024).

Reproducibility statement

This works aims for a high reproducibility by releasing all the code, data, and interactive figures on all of the evals included under various experiment settings. This is describe in the end of the introduction section and the experiment section. This is also updated as more data and settings become available.

AI use statement

In this work, we used generative AI tools for coding, debugging, checking the math, grammar, moving paragraphs, formatting references. It was helpful debugging some tricky issues to get a unbiased estimator to pass the total variance test. An initial draft of of the estimator testing appendix section was generated by AI from the code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. 2025. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *ICML*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *EMNLP*.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2025. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv:2410.07095*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. 2021. Evaluating large language models trained on code. *arXiv:2107.03374*.

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. *arXiv:2403.04132*.

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 *ACL*.

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. *arXiv:2407.21783*.

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:2401.03065*.

Jennifer Haase, Jana Gonnermann-Müller, Paul H. P. Hanel, Nicolas Leins, Thomas Kosch, Jan Mendling, and Sebastian Pokutta. 2026. Within-model vs between-prompt variability in large language models for creative tasks.

Katherine Hermann, Jennifer Hu, and Michael Mozer. 2024. Experimental design and analysis for AI researchers. Tutorial at NeurIPS.

Alexander Hägele, Aryo Pradipta Gema, Henry Sleight, Ethan Perez, and Jascha Sohl-Dickstein. 2026. The hot mess of ai: How does misalignment scale with model intelligence and task complexity?

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: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 *ICLR*.

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

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:2406.10229*.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, et al. 2026. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv:2601.11868*.

Solomon Messing. 2026. Hidden measurement error in llm pipelines distorts annotation, evaluation, and benchmarking.

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

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *ICML*, volume 235.

Angelika Romanou, Mark Ibrahim, Candace Ross, Chantal Shaib, Kerem Oktar, Sam Bell, Elia Ovalle, Jesse Dodge, Antoine Bosselut, Koustuv Sinha, and Adina Williams. 2026. Brittlebench: Quantifying LLM robustness via prompt sensitivity. *arXiv:2603.13285*.

Larry Wasserman. 2004. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, New York, NY.

Chengxing Xie, Bowen Li, Chang Gao, He Du, Wai Lam, Difan Zou, and Kai Chen. 2025. Swe-fixer: Training open-source llms for effective and efficient github issue resolution. *arXiv:2501.05040*.

Dmitri V. Zaykin. 2011. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, 24(8):1836–1841.

A METHOD DETAILS

A.1 ESTIMATING THE PREDICTION NOISE TWO WAYS

For a infinite set of questions, presumably $\mathrm{Var}_\epsilon[\mathrm{E}_x[A]] = 0$, so we consider N specific questions, $\mathrm{E}_x[\mathrm{Var}_\epsilon[A(x,\epsilon)]$ corresponds to the prediction noise that we can measure directly

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

Here $\mathrm{E}_x[A]$ is taken over the uniform distribution of a finite N. While the two sides are equal, estimating the LHS directly with a finite number of random seeds ϵ can have a much higher error. The estimator corresponding to the LHS is running the eval K times and measure the variance of the K final results $\mathrm{var}(\bar{A}_1,\ldots,\bar{A}_K)$ where $\bar{A}_j = \frac{1}{N}\sum_{i=1}^{N} A(x_i,\epsilon_j)$. This fails to use the fact that random samples are draw on each question independently.

The following example shows this concretely, where the variance of the LHS is about 500 times bigger than the RHS.

```
Let X be a 100x5 matrix of Bernoulli(0.5) random variables.
Draw 10000 samples of X to calculate
LHS = var(mean(X, dim=0), dim=1),
RHS = 1/N * mean(var(X, dim=1, ddof=1), dim=0)
Show the mean, variance and the first 5 elements of the two lists
```

A.2 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,

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

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 but it matters the most in the paired case. 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 is 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 6 shows this observation on MATH500, where the small-K correction is needed to reduce the error to an acceptable level even with 2000 bootstrapped 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 due to non-zero prediction noise.

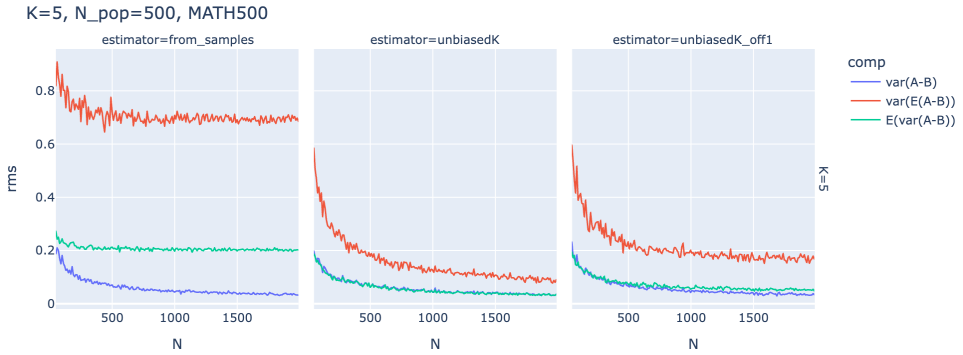

Figure 6: Relative errors of the paired variances. $K = 5$ samples are drawn from the 500 questions from MATH500, where 1000 real predictions are drawn from models A and B on each question, which is treated as the ground truth. The root mean squared relative errors (rms) of variance components are plotted vs. bootstrap sample size N, extending beyond the population size 500. The left figure without the correction on K has an unacceptably high 70% relative error even with 2000 data points. The middle figure is with the correction of §A.2 and the right figure uses $1/K$ instead of $1/(K-1)$.

Same model correction. 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 $A = A(x,\epsilon_1)$, $B = A(x,\epsilon_2)$.

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

Equivalently, we can still use (2) and Table 2 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 $\texttt{mean}_{i,j\neq i} A_i A_j = \frac{1}{K^2-K}\sum_{i,j\neq i} A_i A_j = \frac{1}{K^2-K}[(\sum_i A_i)^2 - \sum_i A_i^2]$.

A.3 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 actual samples $A, B \in \mathbb{R}^{N\times K}$ and present the formulas in §A.2 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 2.

B The Beta theory predicts the noise and fits the distribution

Figure 7 shows the comparison between a $\mathrm{Beta}(a,b)$ distribution and the empirical cumulative distribution (CDF) of the metrics, where the models are binned by range of accuracy. Then $\hat{p}_1 = c_1/K, \ldots, \hat{p}_N = c_N/K$ is computed for each question i across the K combined predictions among the binned models, c_i is the total number of predictions that evaluated as correct on question i. 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 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 has a single mode at $1/C$.

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`	$\frac{1}{N}\sum_i p_i(1-p_i)$

where `b=1/(K-1)*mean(var(A, axis=1))`

Table 1: Unpaired estimators where $A \in \mathbb{R}^{N\times K}$ with K samples for each of N questions.

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`

where `b=1/(kA-1)*mean(var(A, axis=1))+1/(kB-1)*mean(var(B, axis=1))`

Table 2: Paired estimators where $A \in \mathbb{R}^{N\times K_A}$, $B \in \mathbb{R}^{N\times K_B}$.

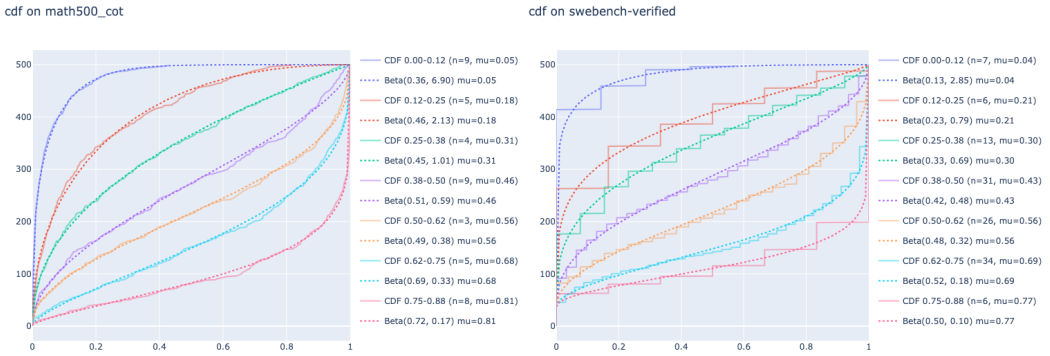

Figure 7: Empirical cumulative distribution curves (CDF, solid) and their respective Beta models (dots). Each color represents a different bin by model accuracy. The average accuracy of each question in each bin of models is computed and sorted to produce the CDF curves. The fit is good.

The Beta theoretical model says that the expected accuracy of each question 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 questions actually have $u_i = 0$ instead of $\mathrm{Beta}(p, 1-p)$.

C Caveats and Exceptions

C.1 Caveats of the Expected Metric

With the impressive increase in power, we recommend averaging for most controlled experiments on model architecture, optimization, loss function, etc. that are not expected to have a large effect on the eval and do not target the eval specifically. However, there are some caveats if our goal is to improve a particular eval metric using all the best available methods, and when the independent experiment variable is expected to affect the "sharpness" of the distribution. Example 5 shows how improvement in expected accuracy leads to no increase after aggregation, and how sharpening the distribution on good-enough model can lead to much better expected accuracy. Example 6 shows how increased expected accuracy leads to a decrease the final performance when the trade-off is less diversity, which can happen in post-training. Example 7 compares the expected metric vs. alternatives and shows that

arbitrarily small but consistent changes is significant using the expected metric, clarifying that an independent standard of effect size is also needed.

Example 5 (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 6 (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 7 (Expected metric and alternatives)**.** Suppose we have 100 True/False questions, model A has an expected accuracy of 60% and Model B 61% on each question.

Expected metric. 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 metric is used by many papers of the area, and corresponds to the *data noise* after averaging away some prediction noise.

Best prediction metric. 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 evals and should be used when pushing the eval.

One prediction metric. 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 metric 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*.

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. Still, pure noise without the bias is not expected to produce a significant result.

C.2 Exceptions to main findings

We note a few exceptions to the general findings. On guessable evals such as MMLU, when the performance is near random chance, the observed noise is closer to the independent prediction rather than the Beta prediction. In Figure 1, swebench-verified, we see 2 outlier results near $x = 0.3$ that come from SWE-Fixer (Xie et al., 2025), due to using a deterministic filter instead of drawing another prediction from the model. The predictions of Llama on vLLM is more deterministic than expected and thus has higher data noise than others, likely due to faulty settings. Distilled models tend to have less sharp distribution at the same performance level.

D Applications and Recommendations

In addition to the recommendation of Miller (2024), we advocate for more global noise measurements to decouple noise analysis from most experiments, so practitioners can focus on their experiments instead of the noise analysis. For some motivating examples, in the Llama 3 report many tables show the unpaired standard errors, which is too loose and gives no additional information since they are just a simple function of the accuracy. Merrill et al. (2026) measured the prediction noise (incorrect) and Chan et al. (2025) measured the total noise with different number of seeds (correct but under-specified). We recommend that eval and leaderboard builders run all-pairs-paired method to measure all the noise types on their eval, thus providing well-specified noise measurements and taking responsibility for the statistical reliability of their evals. This can even be done by a third-party as in this work, as long as leaderboard maintainers release their question level results like Jimenez et al. (2024); Jain et al. (2024), ideally with multiple predictions per question so the noise components can be estimated. This might be a unique possibility in the LLM paradigm, where both models and evals have global identities to enable this approach.

For model developers, we offer in increasing accuracy and overhead:

- use the heuristics $\mathrm{SE}[A-B] \approx \mathrm{SE}[A]$ and read these SEs from our tables
- read our paired total noise curves (Figure 1) to obtain the ranges of each type of noise. Total noise levels can be used when evaluating using one prediction per question and data noise can be used to determine how effective averaging is.
- use method and estimators to measure the noise components on any data, which enables the detection of much smaller signal in controlled experiments.

If it is verified that prediction noise is dominant, then averaging can be used to greatly increase the power, and the usual method of inferring noise from the training curve can also be justified.

Noise-aware Meta-analysis of multiple evals. We have an increasing number of evals of varying quality and size, so we want to get more signal from more evals. Directly averaging MMLU where $N \sim 10k$ and HumanEval where $N \sim 100$ likely leads to worse results than using MMLU alone, so some weighting is necessary. We advocate using the z-score because it is signed and well-behaved. Recall the z-score $z = \frac{\bar{A}-\bar{B}}{\mathrm{SE}[A-B]}$. This work measured the paired and unpaired versions of prediction, data, and total SEs and is suitable here. Inspecting this list of z-scores, one per eval, and checking their consistency is a good start. To get a meta-z-score given the z-score of eval i, an intuitive method is $\bar{z} = \frac{\sum_i w_i z_i}{(\sum_i w_i^2)^{1/2}}$, where some prior beliefs can be incorporated into w_i. For example, $w_i = 1$ means that we believe each eval has equal weight, whereas $w_i = N_i^{1/2}$ means each question has equal weight. While incorporating our prior belief of the importance of each eval into the weight is best suited in LLM evals, $w_i = 1$ is recommended as a baseline. See Zaykin (2011) for more related work and analysis, on which we have two remarks specific to LLMs: 1) the effect size of increasing the model size (scaling law) can be a reference on the expected effect size on an eval if desired; 2) giving each question from different evals the same weight is less valid for evals since different evals may capture different abilities, even if this has the most statistical power for a pool of iid questions.

E 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 questions are random for bootstrap, and when the comparison outcomes are random for the sign test. We show they all give the same answer.

Bootstrap is a general method where resampling the given questions uniformly with replacement is shown to give the right answer for many estimation problems, including estimating the standard error. That is, *resampling the given questions uniformly with replacement gives the same answer as getting more real questions from the population*. Applied to the problem of noise, the natural question is how likely we are 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 randomly resampling the questions, it randomizes the comparison outcomes. 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. The sign test suggests a mix predictor that mixes the predictions of A and B with probability 0.5, which generates variance from different predictions. Let $\hat{y}$ be the prediction to be evaluated then,

$$\hat{y}_{AB}(x) = \begin{cases} \hat{y}_A(x) & \text{with prob } 1/2 \\ \hat{y}_B(x) & \text{with prob } 1/2. \end{cases} \tag{3}$$

If it is quite likely to observe a more extreme outcome under the null hypothesis or equivalently with the mixed predictor, then the conclusion is not statistically significant.

The key step in both methods is estimating the variance of their respective random procedure. Under the same variance, 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 follow.

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 we have the same conclusion $W_A > W_B$ 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 probabilities 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 gives 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 question. 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. Table 3 shows our variance components in terms of the win rates q_A, q_B used in the sign test.

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 $\Pr[A(x_i) > B(x_i)] = q_{A,i}$, $\Pr[A(x_i) < B(x_i)] = q_{B,i}$.

F ESTIMATOR TESTING

We validate our variance estimators by sampling from generative models where the ground truth is known. Our primary focus is measuring estimation accuracy and how it scales with sample size, though we also check the bias and test some properties as well.

For each test scenario, we generate data $A, B \in \mathbb{R}^{N \times K}$ from models with known variance components, apply our estimators, and compare against ground truth over many independent attempts. We use two primary metrics: (1) root mean square (rms) relative error to measure accuracy, and (2) t-scores to detect systematic bias.

We test under three generative processes. The *Bernoulli model* samples N problems with replacement from a population of size N^* parameterized by $p_A \in \mathbb{R}^{N^*}$, then sample K Bernoulli trials for each of the N problems. The *stratified Bernoulli model* evaluates when every question is sampled exactly

once, simulating the case of drawing samples from real eval, breaking independence and introduces tricky effects. The *bootstrap model* resamples from empirical metric matrices, useful for comparisons to bootstrap and testing based on real data.

Accuracy vs. sample size. Our main concern is practical accuracy. Tests establish that estimators achieve rms error below 0.25 for $N = 100$ sample sizes and below 0.13 for larger samples ($N = 400$). We verify this across diverse probability distributions including uniform, Beta, specific distributions and real data. For paired comparisons, tests confirm we get reasonable accuracy (relative rms < 0.1) on the total variance and prediction variance. The data variance is the hardest to estimate since it tend to be the smallest where correcting for the bias in K is critical (Figure 6).

Bias Testing. Deriving and testing unbiased estimators proved useful for understanding estimator for each type of sampling procedure. The process of eliminating bias forced us to carefully account for all sources of estimation error—finite sample corrections in both N and K, paired sampling effects, and stratified sampling adjustments. Moderately large N is necessary for accurate estimation anyway. We test corrections for N and found they always led to larger errors and thus not recommended. On the other hand, small K is common in practice, so we focus on K, and found it to be more than just a bit of bias correction but critical in the case of large N small K.

This testing builds confidence that our estimators behave as predicted and that we understand how to get unbiased estimators in each situation, even as we only recommend two of the simplest estimators.