How Many Raters Do You Need? Power Analysis for Foundation Models

Christopher M. Homan
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14607
cmh@cs.rit.edu

Shira Wein
Department of Computer Science
Georgetown University
Washington, DC 20057
sw1158@georgetown.edu

Lora M. Aroyo, Chris Welty
Google
New York, NY 10011
loraa|welty@google.com

Abstract

Due to their highly stochastic nature, as well as the complexity of the tasks they can perform, foundation models (large machine learning models) are poorly suited for conventional machine learning evaluation methods. This is because machine learning evaluation methods typically assume behavior to be deterministic and simple enough to be measured against gold standard data with unitary, authoritative, "correct" answers using straightforward metrics such as accuracy, precision, and recall. In this work, we propose an evaluation framework suitable for foundation models, which takes into account variance in the responses of both machine model and human rater. Utilizing recent advances in p-value estimation, we investigate the trade-offs between the number of items in a test set, the number of responses per item, the sampling method, and the metric, when measuring the comparative differences between two hypothetical foundation models at various degrees of similarity. When two models are very far apart in their predictive performance, fewer raters are needed to confidently compare them, as expected. However, as the models draw closer, we find that a larger number of annotators than are currently typical in annotation collection are needed to ensure the power analysis correctly reflects the difference in performance.

1 Introduction

Foundation models are capable of performing tasks that appear to be much more sophisticated than earlier families of machine learning models, and to respond differently each time even when presented with the same prompts (i.e., inputs). Yet these models are capable of failing in surprising ways by providing offensive, inappropriate, or inaccurate content. Moreover, compared to models with simpler outputs, such as classifiers, it is harder for humans to agree on whether a particular generated utterance generated by a foundation model is safe or not.

Conventional machine learning evaluation methods, by contrast, assume that performance can be measured against a gold standard, where each input is associated with a single correct answer. In one-on-one comparisons of such models—needed for leaderboards and scientific papers where such comparisons are performed—it is important for there to be enough inputs in the training data to ensure with some measure of confidence that the best model wins.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

In this work, we propose an evaluation framework suitable for foundation models, in which both model and gold-standard responses are taken to be stochastic, and investigate a crucial yet unexplored question: what is the trade-off between not just the number of items (N), but also *the number of rater responses per item* (K) in a human annotated test-set used for evaluation, under the assumption that annotation costs scale linearly with $N \times K$? Given the associated costs with either collecting judgments for more items or collecting a greater number of responses per item, we assess how many items and raters are needed in order to provide statistical guarantees to model comparisons.

Inspired by recent progress in *p*-value estimation [Wein et al., 2023], we approach this item/response trade-off from the novel perspective of *instrument sensitivity*, which can be defined to be the smallest absolute amount of change that can be detected by a measurement ("instrument"). A major complication of foundation models is that they are typically used stochastically. For example, large language model (LLM) outputs are typically generating by randomly choosing each successive world, according to a distribution based on the LLM and a temperature parameter. We explore how this trade-off affects the power of statistical hypothesis testing for comparing machine learning models where response variance is present. Therefore, we can examine the relationship between number of items/responses under different model conditions (i.e. difference in model performance, metric, and sampling methods) to the measurement's ability to accurately identify whether one model is closer to the human judgments.

To exploit the notion of instrument sensitivity, we follow previous work Wein et al. [2023] and utilize a simulator to produce model predictions for two models, drawn from similar random distributions. We then perturb the second model's simulated predictions by injecting noise, and vary the noise rate for the second model to progressively decrease the similarity between the two models. Finally, we evaluate the progressively dissimilar models and estimate the *p*-value for the differences between their metric scores. We expect that, when the models are more similar, the *p*-value will be higher, but that as we increase $N \times K$, the *p*-value will be lower, because there are more ratings to compare the models to. This yields a quantification of this trade-off which we explore.

We make the following contributions:

- Analysis of the trade-off between the number of *items* (N) and the number of rater *responses* per item (K) in a test set using the *p*-values of model comparisons as a measure of sensitivity;
- Analysis of the sensitivity of several metrics, including mean absolute error (MAE), Earth Mover's Distance (EMD), and Wins (a new metric), as well as the effect of the N/K trade-off on them;
- First results showing the number of total ratings required in a test set to make significant comparisons between two model's metric scores, based on how close they are.

2 Related Work

Lin et al. [2014] have suggested that response variance is less important than item variance – at least for training data. Specifically, they suggest that collecting more items with a single response is more valuable than collecting multiple responses per item. We set out to examine the effect of number of responses and items on power analyses for model comparisons under various conditions, and to do so, we perform power analyses on large-scale simulated test data. This type of simulated test data production is enabled by recent work developing a simulator to estimate the "true" *p*-value of model comparisons under different metrics and sampling methods [Wein et al., 2023]. This method incorporates rater and item variance into test set items and responses, and then produces a "reference test set" that mixes the responses to produce metric score distributions of (simulated) human raters and two AI systems. Basing our analysis on simulated results allows us to experiment with widely different N/K scenarios that would be impractical if we used actual raters.

Related crowdsourcing trade-offs have examined the balance between cost and quality of annotation collection [Snow et al., 2008], as well as recommendations for which crowdsourcing platforms and protocols to use [Wang et al., 2013]. Chau et al. [2020] explored the use of peer-review and self-review in order to resolve disagreement in annotations, and Hovy et al. [2013] developed an unsupervised model to identify which Mechanical Turk annotators are reliable. Recent assessments of leaderboard practices have also led to models able to indicate which items are most useful to annotate

for evaluation purposes [Rodriguez et al., 2021]. Welinder and Perona [2010] developed a system to select the most useful/informative labels to collect, which can lead to a reduction in annotation cost.

Wein et al. [2023] use null hypothesis significance tests (NHST) on simulated item/response distributions to compare two AI systems and estimate a *p*-value. Statistical testing of system performance is critical to the understanding of state-of-the-art performance on a task or within a domain, in particular due to the flawed nature of benchmarking practices in machine learning evaluation [Ethayarajh and Jurafsky, 2020, Raji et al., 2021, Rodriguez et al., 2021, Hernandez-Orallo, 2020].

Existing metrics such as Student's t-test [Student, 1908] are based on the assumption that the datasets are normally distributed [Søgaard et al., 2014], which is not the case for ML metrics and are therefore not applicable, in particular when testing the system on new datasets [Søgaard, 2013].

Our approach incorporates response variance from human raters [R Artstein, 2008, Plank et al., 2014] and from AI models [Szymański and Gorman, 2020]. Human rater response variance on individual items is most often due to measurable differences in perspective or ambiguity of the item, as opposed to noise. AI models vary in their responses on individual items due to stochastic initial states and gradient descent, as well as changes in training data such as cross-validation.

Dietterich [1998] applied hypothesis testing to machine learning systems and Dror et al. [2020] provide a survey and guide to state-of-the-art techniques for statistical significance testing in AI systems. Deutsch et al. [2021] study permutation and bootstrapping methods for computing significance tests and confidence intervals for text summarization evaluation metrics. In their setting two evaluation metrics are paired and permutation sampling is used to evaluate them over multiple documents and summarization models.

As we do in this work, Søgaard et al. [2014] examine the effect of a number of variables (including variance, effects of sample size, and covariates) on *p*-values.

3 Methodology

In order to investigate the effect of the number of *items* (N) and the number of rater *responses* per item (K) on *p*-value-based model comparisons, we use a simulator provided by Wein et al. [2023] to produce human and model predictions for individual items by modeling the items as random distributions. For each of N items in the test set, the simulator randomly draws a mean and standard deviation $\{(\mu_i, \sigma_i)\ i \in [1, N]\}$, where $\mu_i \sim \mathcal{U}[0, 1]$, $\sigma_i \sim \mathcal{U}[0, .3]$, $\mathcal{U}$ is the uniform distribution and $\sim$ indicates a random draw from a distribution. It then draws from the resulting normal distribution $\mathcal{N}(\mu_i, \sigma_i)$ K times to produce the gold standard set of responses G_i (clipping values outside of $[0, 1]$). This per-item draw of K responses is repeated to produce the machine predictions A_i. This models the idealized situation in which machine system A is a perfect representation of the gold standard, since *it is drawn from the same distribution*.

A third set of item responses for machine system B is then drawn by injecting random noise into the base distribution of each item. For a given perturbation level ϵ, which we choose, a noise parameter is randomly drawn for each item $\epsilon_i \sim \mathcal{U}[-\epsilon, +\epsilon]$, and then K responses for each B_i are drawn from $\mathcal{N}(\mu_i + \epsilon_i, \sigma_i)$.

For any given selection of N, K, ϵ, we have a matrix of responses $G^{N,K}$, $A^{N,K}$, and a matrix $B^{N,K,\epsilon}$ for each ϵ. We then seek to compare A and B to each other to determine which is better (the answer should always be A unless $\epsilon = 0$). When evaluating AI systems, the comparison of A and B involves comparing each of their item responses to those of G, using a suitable metric such as error or correlation, which is then aggregated across the items.

The simulator allows us to generate many test sets to extrapolate patterns beyond one domain or system. By holding the item distributions for A, B and G fixed, we can draw from them repeatedly to generate millions of possible test sets, and truly measure the variance of the metric scores, which would be infeasible with actual human ratings. With this variance, we can also construct a null hypothesis set and measure how likely it is that an observed difference between the two metric scores could have occurred by chance. We perform 36 experiments on different simulated datasets for every combination of $N \in \{25, 50, 100, 250, 500, 1000\}$ and $K \in \{1, 5, 10, 25, 50, 100\}$.

Metrics also play a key role in our study; as Wein et al. [2023] exposes, both metrics and test-set sampling methods can affect the power analysis. While the latter did not turn out to be important in

our study, metrics play a key role in our study. They model a metric as a function $\Gamma(M, G)$, where M is a matrix of model predictions (e.g. A or B) and G is the matrix of gold standard responses, which returns a score for M. Clearly each different metric, e.g. mean average error (MAE) or correlation, will produce a different score for the same matrix of responses, so it stands to reason that any comparison $\Gamma(A, G) > \Gamma(B, G)$ will have different *p*-values for different Γ. Wein et al. analyze several metrics, suggesting a few that give the lowest *p*-values. We chose three of the best performing metrics (Γ): mean absolute error (MAE), item-wise wins (Wins), and Mean EMD (MEMD). Further details on these metrics are included in the appendix.

4 Results

		Perturbation rate (ϵ)			
N	K	0.005	0.01	0.02	0.1
100	10	0.4428	0.4414	0.3073	0.0179
1000	1	0.4403	0.4059	0.3274	0.0052
25	100	0.4460	0.4009	0.2514	0.0059
100	25	0.4192	0.3587	0.2481	0.0004
500	5	0.4308	0.3406	0.2009	0.0002
50	100	0.3972	0.3185	0.1611	0.0001
1000	5	0.3797	0.2608	0.1183	0.0000
100	100	0.3463	0.2161	0.0689	0.0000
1000	10	0.3173	0.2030	0.0508	0.0000
250	100	0.2687	0.1030	0.0076	0.0000
1000	25	0.2414	0.0859	0.0051	0.0000
500	100	0.1823	0.0368	0.0002	0.0000
1000	50	0.1748	0.0330	0.0003	0.0000

Table 1: P-values for Γ_{Wins}, in groups with equal $N \times K$, showing lower *p*-value in each group as N increases.

		Perturbation rate (ϵ)			
N	K	0.005	0.01	0.02	0.1
100	10	0.4658	0.4947	0.4233	0.0119
1000	1	0.4806	0.4947	0.4941	0.1193
25	100	0.4587	0.4186	0.2713	0.0000
100	25	0.4669	0.4458	0.3561	0.0000
500	5	0.4864	0.4884	0.4053	0.0009
50	100	0.4541	0.3798	0.2091	0.0000
1000	5	0.5048	0.4671	0.3824	0.0000
100	100	0.4336	0.3216	0.1186	0.0000
1000	10	0.4702	0.4248	0.2897	0.0000
250	100	0.4395	0.2412	0.0270	0.0000
1000	25	0.4304	0.3377	0.1017	0.0000
500	100	0.3862	0.1832	0.0031	0.0000
1000	50	0.4059	0.2385	0.0123	0.0000

Table 2: P-values for Γ_{MAE}, in groups with equal $N \times K$, showing lower *p*-value in each group as K increases.

For each value of Γ and ϵ there are 36 experiments, one for each N, K pair, comparing the two simulated systems and generating a *p*-value. We use three different Γ values and four ϵ, yielding 12 sets of 36 experiments each (full results of each of these experiments included in the appendix).

To begin with, our results replicate the previous results of Wein et al. [2023] across metrics, perturbations, and $N \times K$ that Γ_{Wins} has the lowest *p*-values of any metric, when other parameters are equal, and is the only metric to find significant difference with $\epsilon = 0.01$, given at least 50,000 ratings. Γ_{Wins} is not a well-known metric, and appears to have been introduced by Wein et al. [2023].

Across all the experiments, the impact of $N \times K$ is consistent: the more overall ratings, the lower the *p*-value, and thus the more sensitive the evaluation instrument. However, the N vs. K question is not as consistent, we see similar *p*-values in experiments with the same $N \times K$, and the full story is more nuanced.

We begin our N vs. K analysis with Γ_{Wins}, the most sensitive metric configuration we tested, shown in Table 1. Since $N \times K$ is the predominant signal in lowering *p*-values, we group experiments by that value. The *p*-value decreases with increasing N when $N \times K$ is constant.

The value of increasing N vs. K does appear to depend on the metric, however. For Γ_{MAE} and Γ_{MEMD}, there is a consistent decrease in *p*-values across all values of ϵ when K is increased over N (Table 2). Notably, the Γ_{MAE} metric shows generally better *p*-values than Γ_{MEMD}, posting significant a difference between A and B at $\epsilon = 0.02$ when $(N, K) = (250, 100)$, whereas Γ_{MEMD} comes very close with $p = 0.0502$ in the same setting (Table 14).

To understand the impact of the perturbation values (ϵ) on the metric scores, we show in Table 3 the relationship between different perturbation rates and the difference (Δ) in metric scores between A and B, for each metric. Γ_{MEMD} shows the largest absolute increase in B's scores, however it is a non-normalized score. For Γ_{Wins}, the most sensitive metric, we can cross reference with Table 1 to see that with a difference in metric scores of 0.0146 ($\epsilon = 0.005$), we were not able to generate enough ratings to claim significance. At a metric difference of 0.0237 ($\epsilon = 0.01$), we need 50,000 ratings to claim significance ($p < 0.05$), and with a difference of 0.0438 ($\epsilon = 0.02$), 25,000 ratings are required. At the highest perturbation rate we tested, $\epsilon = 0.1$, corresponding to a 0.1631 difference in Γ_{Wins}, all the (N, K) settings in Table 1 are significant. In Table 7, we see that this metric is capable of showing significant difference between A and B with as few as 500 ratings when the score difference is greater than 0.1631.

For the more familiar Γ_{MAE} metric, shown in Table 2 and cross-referencing with Table 3, 25,000 ratings can power a significant difference between two models that differ by 0.0139 ($\epsilon = 0.02$), but only at $(N, K) = (250, 100)$. We tested one other configuration with the same number of ratings, $(N, K) = (500, 50)$, with $p = 0.0624$ (Table 10). We found that we were not able to generate enough ratings for the smaller Γ_{MAE} differences to be significant. For the largest difference we tested, 0.0243 ($\epsilon = 0.1$), all (N, K) combinations are significant except for $(1000, 1)$. In Table 11, we see that with $(N, K) = (25, 25)$ this metric can power a significant measurement at this distance apart.

Γ_{MEMD} shows similar behavior to Γ_{MAE}, with a metric difference of 0.1773 ($\epsilon = 0.02$) requiring 50,000 ratings to be significant (Table 14). At the highest pertubation ($\epsilon = 0.1$), a metric difference of 0.3232 requires 2500 ratings when $(N, K) = (250, 10)$ (Table 15).

		Γ_{MAE}	Γ_{MEMD}	Γ_{Wins}
	A	0.0677	1.2091	0.5073
$\epsilon = \mathbf{0.005}$	B	0.0701	1.2055	0.4927
	Δ	0.0024	0.0036	0.0146
$\epsilon = \mathbf{0.01}$	B	0.0782	1.2657	0.4836
	Δ	0.0106	0.0566	0.0237
$\epsilon = \mathbf{0.02}$	B	0.0816	1.3864	0.4635
	Δ	0.0139	0.1773	0.0438
$\epsilon = \mathbf{0.1}$	B	0.0920	1.5323	0.3442
	Δ	0.0243	0.3232	0.1631

Table 3: Mean metric scores for B shown for each value of ϵ, and the increasing Δ to A.

5 Discussion

Our results indicate that the number of raters and items do have a notable impact on *p*-value estimation, to different degrees depending on the metric, and leave us with two questions:

- Why does the behavior of Γ_{Wins} differ from the other two metrics with respect to N vs K?
- What effect do the metric methods have on our results?

First, the Wins metric provides a discrete decision for each *item*, counting those decisions (i.e."wins") across the test set and normalizing by the number of items, making it similar in this respect to accuracy, which classifies each discrete prediction as a true or false positive or negative. Wins is also presented as a meta-metric of sorts, it can use any item-level metric, with absolute error being used here, and requires both model's predictions as input, in order to directly compare their predictions at the item level.

In general, increasing N (number of test set items) increases the statistical power of any measurement by simply providing more scores to base the final metric score on. The more scores there are, the more stable the variance across simulation runs will be, and the lower the *p*-value. All examined metrics respond well to increasing N.

Increasing K (number of responses per item) increases the statistical power of each *item level score*. As K increases, the more stable the variance of an individual item's score will be across simulation runs, thereby lowering the *p*-value. All tested metrics also respond well to increasing K.

The difference between the metrics lies in the way the item-level scores are used. For Wins, which responds better to increasing N, the A's and B's item-level scores are directly compared. In each run, these item-level scores will vary, but in many cases that variance won't change the pairwise comparison. For example, if A_i's metric score is 0.10 and B_i's is 0.12 on the first simulation, a win is recorded for A. In the next simulation, if the scores are 0.11 and 0.13, respectively, this score change does not change the Win, as A_i's score is still lower. This indicates the item-level variance in the discrete win decision is far lower than the score variance - so adding more responses is less likely to further reduce the variance than adding items.

By contrast, for the MAE and MEMD metrics, any changes in item-level metric scores do impact the variance, both at the item and test-set level. Since the item-level scores come from the response distribution, adding more responses stabilizes the simulated distributions under repeated test set generation, reducing the metric variance across simulations and lowering the *p*-value.

The implications of these results are that the item/response trade-off should be handled differently depending on the metric itself, and the demands on number of raters and items are high for all metrics in order to provide statistical guarantees.

6 Conclusion

In this work, we experimented with simulated data in order to examine the trade-off between number of items and number of ratings per item (aka responses) necessary to compare two systems against human judgments with statistical significance ($p < 0.05$). As expected, we see that when two systems are more similar in performance, a greater number of annotations is required to achieve significance on their comparison. Further, the metric itself affects the utility of an increase in either items or responses.

We find that in order to provide statistical guarantees on model comparisons, many more raters and items are needed than are typically used. Specifically, for systems with similar performance, e.g. a difference in MAE scores of 0.0139, 25,000-50,000 total ratings are needed to achieve $p < 0.05$, with *at least* 250 items and 100 responses per item. Previous work has identified a metric that counts item-level wins, which can more reliably provide significant comparisons with a metric difference of 0.0237. However, the community does not have a lot of experience with this metric.

These results suggest that current evaluation practices are not sufficient to confidently assess two systems' performance against gold judgments, as using 25,000-50,000 ratings in a test set is rarely seen. Even when using 1000 items, at least 25 raters are needed for systems with an MAE difference of 0.0139 to achieve significance.

Additionally, we found that the trade-off between number of items and number of responses per item, depended on the metric. For two of our tested metrics, MAE and mean EMD, adding more responses than items is a more optimal division to achieve lower p-values. For the Wins metric, the opposite is true: more items and fewer responses per item lead to lower p-values. Still, in all cases for all metrics, increasing the total number of responses consistently lowers p-values, and thereby increases the sensitivity of the evaluation instrument.

References

Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.law-1.7.

Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146, 2021.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2): 1–116, 2020.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, 2020.

Jose Hernandez-Orallo. Ai evaluation: On broken yardsticks and measurement scales. In *Workshop on Evaluating Evaluation of Ai Systems at AAAI*, 2020.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1132.

Christopher H. Lin, Mausam, and Daniel S. Weld. To re(label), or not to re(label). In *HCOMP 2014*, 2014.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2083. URL https://aclanthology.org/P14-2083.

M Poesio R Artstein. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. 2021.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1027.

Anders Søgaard. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1068.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1601. URL https://aclanthology.org/W14-1601.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Piotr Szymański and Kyle Gorman. Is the best better? Bayesian statistical model comparison for natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2203–2212, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.172. URL https://aclanthology.org/2020.emnlp-main.172.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47:9–31, 2013.

Shira Wein, Christopher Homan, Lora Aroyo, and Chris Welty. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3138–3161, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.196. URL https://aclanthology.org/2023.findings-acl.196.

Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32. IEEE, 2010.

7 Appendix

For our experimentation, we chose three of the best performing metrics (Γ) from Wein et al. [2023]:

- *Mean absolute error* (MAE). The absolute value of the difference (error) from the mean gold responses per item to the mean system responses, and then take the mean of that item-wise error:
$$\Gamma_{\mathrm{MAE}}(M, G) = \mu_{i \in N}\left[|\mu(M_i) - \mu(G_i)|\right]$$
- *Item-wise wins* (Wins). The fraction of items in the test set for which the absolute error of A is smaller than B:
$$\Gamma_{\mathrm{Wins}}((A, B), G) = N^{-1} \sum_{i \in N} \mathbf{1}_{<}(|\mu(A_i) - \mu(G_i)|, |\mu(B_i) - \mu(G_i)|)$$
- *Mean EMD* (MEMD). The Earth mover's distance for each item between the system and the gold standard responses, and then take the mean of those item-wise EMDs:
$$\Gamma_{\mathrm{MEMD}}(M, G) = \mu_{i \in N}\left[\mathrm{EMD}(M_i, G_i)\right]$$

	GT P-score					
Responses/Item	1	5	10	25	50	100
Items						
25	0.532728	0.532937	0.504313	0.504714	0.493821	0.445982
50	0.483674	0.506107	0.466705	0.470275	0.443130	0.397206
100	0.500946	0.491946	0.442751	0.419190	0.406609	0.346261
250	0.474844	0.467297	0.420981	0.383181	0.346522	0.268676
500	0.470714	0.430802	0.380449	0.331669	0.262719	0.182288
1000	0.440303	0.379661	0.317261	0.241430	0.174794	0.101379

Table 4: P-value under the null hypothesis for Γ_{Wins}, $\epsilon = \mathbf{0.005}$

	GT P-score					
Responses/Item	1	5	10	25	50	100
Items						
25	0.533728	0.512096	0.526453	0.471878	0.449856	0.400890
50	0.523777	0.489725	0.483119	0.409684	0.370236	0.318485
100	0.496893	0.446792	0.441404	0.358724	0.298661	0.216086
250	0.459792	0.395051	0.368231	0.268284	0.202322	0.102964
500	0.424356	0.340563	0.291115	0.175208	0.096934	0.036776
1000	0.405872	0.260752	0.203025	0.085938	0.033042	0.005474

Table 5: P-value under the null hypothesis for Γ_{Wins}, $\epsilon = \mathbf{0.01}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.530782	0.465481	0.440375	0.404518	0.330468	0.251441
50	0.493435	0.426151	0.385814	0.328747	0.240567	0.161137
100	0.466991	0.359638	0.307280	0.248111	0.141226	0.068867
250	0.420139	0.286696	0.203605	0.102247	0.040013	0.007554
500	0.380634	0.200943	0.125768	0.036828	0.007439	0.000218
1000	0.327381	0.118287	0.050823	0.005102	0.000272	0.000000

Table 6: P-value under the null hypothesis for Γ_{Wins}, $\epsilon = \mathbf{0.02}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.4163	0.2568	0.1674	0.0671	0.0211	0.0059
50	0.3284	0.1409	0.0710	0.0138	0.0020	0.0001
100	0.2421	0.0571	0.0179	0.0004	0.0001	0.0000
250	0.1192	0.0036	0.0003	0.0000	0.0000	0.0000
500	0.0428	0.0002	0.0000	0.0000	0.0000	0.0000
1000	0.0052	0.0000	0.0000	0.0000	0.0000	0.0000

Table 7: P-value under the null hypothesis for Γ_{Wins}, $\epsilon = \mathbf{0.1}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.495551	0.493076	0.477210	0.474897	0.483937	0.458687
50	0.485007	0.501319	0.471535	0.471248	0.466577	0.454079
100	0.491444	0.491754	0.465804	0.466923	0.450338	0.433630
250	0.499223	0.481230	0.486174	0.475731	0.467507	0.439462
500	0.506496	0.486368	0.475118	0.451602	0.437663	0.386180
1000	0.480614	0.504753	0.470196	0.430420	0.405941	0.353477

Table 8: P-value under the null hypothesis for Γ_{MAE}, $\epsilon = \mathbf{0.005}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.496683	0.484496	0.485500	0.491932	0.460635	0.418577
50	0.496567	0.487428	0.497806	0.463185	0.425495	0.379793
100	0.489275	0.492621	0.494663	0.445780	0.401342	0.321604
250	0.483415	0.508406	0.479275	0.422221	0.360559	0.241187
500	0.492526	0.488373	0.454868	0.391084	0.307984	0.183233
1000	0.494677	0.467117	0.424842	0.337664	0.238503	0.102904

Table 9: P-value under the null hypothesis for Γ_{MAE}, $\epsilon = \mathbf{0.01}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.501454	0.468317	0.467570	0.431753	0.353261	0.271256
50	0.502779	0.473373	0.449794	0.396709	0.315623	0.209136
100	0.502380	0.449883	0.423328	0.356146	0.248200	0.118621
250	0.473889	0.444882	0.390885	0.265469	0.137652	0.026972
500	0.491067	0.405322	0.345264	0.192024	0.062419	0.003117
1000	0.494113	0.382415	0.289692	0.101698	0.012317	0.000126

Table 10: P-value under the null hypothesis for Γ_{MAE}, $\epsilon = \mathbf{0.02}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.4283	0.2470	0.1247	0.0188	0.0009	0.0000
50	0.3913	0.1641	0.0527	0.0027	0.0000	0.0000
100	0.3492	0.0845	0.0119	0.0000	0.0000	0.0000
250	0.2776	0.0106	0.0001	0.0000	0.0000	0.0000
500	0.2039	0.0009	0.0000	0.0000	0.0000	0.0000
1000	0.1193	0.0000	0.0000	0.0000	0.0000	0.0000

Table 11: P-value under the null hypothesis for Γ_{MAE}, $\epsilon = \mathbf{0.1}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.539872	0.504037	0.496508	0.487312	0.501847	0.482049
50	0.517989	0.503496	0.508247	0.496522	0.495213	0.467237
100	0.501885	0.498564	0.509060	0.489665	0.487306	0.464313
250	0.497131	0.498128	0.489341	0.480132	0.469489	0.441288
500	0.496065	0.495014	0.476651	0.456579	0.436668	0.381117
1000	0.494769	0.492469	0.480331	0.430268	0.402919	0.338026

Table 12: P-value under the null hypothesis for Γ_{MEMD}, $\epsilon = \mathbf{0.005}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.572866	0.503486	0.494903	0.470426	0.438353	0.411324
50	0.540708	0.491227	0.499775	0.469750	0.436827	0.399017
100	0.526373	0.503497	0.489841	0.444308	0.405353	0.351879
250	0.501497	0.485039	0.480643	0.415959	0.353553	0.272446
500	0.501425	0.463417	0.478026	0.386038	0.305701	0.198424
1000	0.500922	0.463818	0.458767	0.352803	0.253529	0.112516

Table 13: P-value under the null hypothesis for Γ_{MEMD}, $\epsilon = \mathbf{0.01}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.538096	0.490751	0.482491	0.443054	0.377346	0.309274
50	0.514710	0.478230	0.459458	0.410965	0.342481	0.245174
100	0.507034	0.468373	0.453770	0.372558	0.280805	0.154227
250	0.504493	0.451870	0.404244	0.283895	0.158546	0.050219
500	0.503959	0.440916	0.383145	0.227624	0.086505	0.012174
1000	0.502619	0.421632	0.334548	0.144549	0.024115	0.001115

Table 14: P-value under the null hypothesis for Γ_{MEMD}, $\epsilon = \mathbf{0.02}$

Responses/Item Items	GT P-score 1	5	10	25	50	100
25	0.5671	0.3862	0.3000	0.1442	0.0527	0.0072
50	0.5482	0.3368	0.2352	0.0576	0.0094	0.0000
100	0.4659	0.2728	0.1348	0.0111	0.0001	0.0000
250	0.4336	0.1779	0.0417	0.0002	0.0000	0.0000
500	0.3998	0.0897	0.0060	0.0000	0.0000	0.0000
1000	0.3488	0.0297	0.0003	0.0000	0.0000	0.0000

Table 15: P-value under the null hypothesis for Γ_{MEMD}, $\epsilon = \mathbf{0.1}$