Smarter Sampling for LLM Judges: Reliable Evaluation on a Budget

Anonymous Author(s)
Affiliation
Address
email

Abstract

LLM-as-a-judge is increasingly dominant as a framework for scalable evaluation of artificial intelligence (AI) systems and agents. The technique involves prompting a large language model (LLM) to assess the capabilities of another AI model. Although the system reduces human annotation requirements, the need for human oversight is still required to gauge the performance of the judge LLM. However, human annotations can be expensive to obtain, particularly in domains that require expert annotations, such as clinical text generation. Thus, the problem drives the questions: (1) Can we bound the number of human annotations necessary to gauge the performance of our judge LLM? and (2) Can we curate the subset of data for human annotation in a principled way? In this paper, we answer (1) through a Chernoff bound for intraclass correlation coefficient (ICC), the primary metric for measuring LLM-as-judge performance relative to human labels. To explore (2), we propose 7 sampling methods and demonstrate the utility of these algorithms relative to random sampling in simulated and real-world data. We show tighter bounds for sampling requirements and up to a 41% relative improvement in ICC precision compared to random baselines.

1 Introduction

As large language models (LLMs) increase in prevalence for various tasks, particularly in text generation, the task of scalable evaluation of LLM output increases in importance [Thirunavukarasu et al., 2023, Meyer et al., 2023, Yuan et al., 2021, Celikyilmaz et al., 2021]. The LLM-as-a-judge framework – in which an LLM evaluates another artificial intelligence (AI) agent – is becoming increasingly accepted as an effective and scalable method for evaluation [Gu et al., 2025]. As such, considerable attention has been devoted to the assessment of judge LLMs, including human-labeled benchmarks [Dubois et al., 2024], and reference-free evaluations [Tan et al., 2025].

In the case of text generation by a subject LLM [1], we consider human-annotated scores to be gold-standard. Previous research constructing evaluation frameworks and benchmarks highlight the challenges of costly and slow human evaluation pipelines [Liang et al., 2022, Kiela et al., 2021]. These efforts further motivate the need for scalable alternatives such as LLM-as-a-judge, where annotation impacts the practicality of evaluation at scale. To reduce unnecessary annotation collection and human labor, we consider the following question in our paper: can we derive a minimum number of annotations required such that we are guaranteed with high probability an *accurate* measure of performance of our judge LLM?

Different metrics have been proposed to measure performance of judge LLMs relative to human labels, largely originating from classic statistics literature, such as intra-class correlation coefficient

[1] We refer to the LLM to be evaluated as the subject LLM.

(ICC), Cohen's kappa, Cronbach's alpha, etc. [Shrout and Fleiss, 1979, Cohen, 1960, Cronbach, 1951]. Due to its relation to Pearson's correlation coefficient, and previous work [Salnikov, 2024], we utilize ICC as the primary metric in our theoretical and experimental results.

To address our initial question, we leverage the classic Chernoff bound technique [Chernoff, 1952] on intra-class correlation coefficient between LLM-generated and human annotations. Thus, we provide a simple concentration inequality for intra-class correlation under some limiting assumptions and approximations. With the concentration inequality, we derive an approximate lower bound on the number of annotations necessary to guarantee with high probability that the measured (sample) ICC is ε-close to the population ICC (see Section 2.1).

With a bound on sample size for i.i.d. samples, we extend the question and ask if we can reduce the required number of samples further if we curate a subset of data for human annotation in a principled manner. To interrogate this empirically, we formulate the problem of curating a subset of data for human annotation as an optimization problem. We assume a fully LLM-annotated dataset and assume in-distribution data (out-of-distribution generalization is beyond our scope). With this foundation, we build on previous methods in statistical sampling, clustering, and active learning [Cochran, 1977, Lloyd, 1982, Settles and Craven, 2008], and extend them to the emerging challenge of scalable LLM-as-a-judge evaluation. We propose 7 sampling methods (in addition to random sampling as a baseline) for subset selection, and study how these strategies impact reliability estimates and annotation efficiency.

We evaluate each of our provided algorithms on simulation data, in which we demonstrate the significant utility of 4 sampling methods over random sampling. In real-world text datasets, we find that all proposed sampling methods outperform random selection under extreme annotation constraints, with the best strategy outperforming random selection by a relative improvement of 41%.

2 Related Work

Large language models (LLMs) are increasingly used as automatic evaluators of other AI systems, offering a scalable alternative to costly human assessment. Early studies benchmarked LLM judgments against human annotations in tasks such as summarization, dialogue, and reasoning [Gilardi et al., 2023, Zheng et al., 2023], while more recent frameworks like AlpacaEval 2.0 and Arena-Hard integrate human and LLM judgments or introduce more challenging comparative tasks [Li et al., 2023, 2024]. These works highlight the importance of measuring judge reliability, often using metrics such as accuracy, agreement, or correlation with human preferences; intraclass correlation coefficient (ICC) has emerged as a standard for evaluating continuous or ordinal judgments [Bedi et al., 2025].

The statistical foundation for ICC estimation and sample size planning is well-established [Fisher, 1925, Bonett, 2002, Zou, 2012], motivating the use of concentration inequalities such as Chernoff bounds to provide formal guarantees on judge reliability. In parallel, reducing reliance on human labels has been extensively studied in active learning and sample-efficient annotation, where the goal is to identify the most informative examples [Settles, 2009, Wei et al., 2022]. Strategies such as uncertainty sampling [Roy and McCallum, 2001], core-set selection [Sener and Savarese, 2018], and diversity-driven sampling [Brinker, 2003] demonstrate that principled subset selection can achieve reliable evaluation with fewer annotations. These directions connect naturally to deeper statistical and machine learning traditions: classical sampling theory provides foundations for stratified and variance-weighted designs [Cochran, 1977, Fedorov, 1972], while clustering [Lloyd, 1982] and density-based approaches [Silverman, 1986] ensure representativeness in diverse datasets. Active learning has likewise explored uncertainty- and density-driven criteria [Nguyen and Smeulders, 2004, Settles and Craven, 2008], which we adapt to LLM-as-a-judge pipelines to design resource-aware evaluation strategies that balance annotation cost with statistical reliability.

2.1 Theoretical Foundations

Here, we derive a simple and loose concentration inequality for the intra-class correlation coefficient (ICC) with some limiting assumptions and approximations. The expression for ICC was originally proposed by Fisher [1925], as an extension of Pearson's correlation coefficient, but since has been revised under the random effects model and several other formulas have been proposed (see Appendix A.1 for definitions, and correct formula).

Given annotation distributions, H and G, for human and LLM-generated, respectively we assume a bivariate normal joint distribution, such that $H_i, G_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ independently and identically distributed (i.i.d.), and $i \in \{1, \ldots, n\}$, where n number of samples. Denote the population ICC between H and G as ρ and the observed (sample) ICC at n samples, $\hat{\rho}_n$. Fisher [1925] showed that under certain assumptions, the distribution of $\hat{\rho}_n$ approaches Gaussian asymptotically (see Appendix A.1.2 for relevant parameters). We therefore obtain the following lemma (see Appendix A.1.2 for proof and necessary assumptions).

Lemma 1 (Chernoff bound for approximate ICC) *Given $\varepsilon, \delta > 0$, under critical assumptions on n and ρ,*

$$\Pr[|\hat{\rho}_n - \rho| \geq \varepsilon] \lesssim 2 \exp\left(- \frac{(n-1)\varepsilon^2}{2(1-\rho^2)^2} \right)$$

Therefore, with probability $1 - \delta$, the sample and population ICC are guaranteed to be ε-close if

$$n \gtrsim 1 + \frac{2(1-\rho^2)^2}{\varepsilon^2} \log\left(\frac{2}{\delta}\right)$$

3 Methods

3.1 Problem Formulation

We study the problem of evaluating LLM judges under limited annotation budgets. Let $\mathcal{N} = \{1, \ldots, n\}$ denote the set of items, with gold labels $H = \{h_i\}$ from humans and inexpensive labels $G = \{g_i\}$ from an LLM judge. Reliability is measured using the Intraclass Correlation Coefficient (ICC(3,k)), which captures absolute agreement.

Given a budget $z < n$, we seek a subset $S^* \subseteq \mathcal{N}$ of size z such that the agreement computed on (H_S, G_S) closely approximates the agreement on (H, G):

$$S^* = \arg \min_{|S|=z} |I(H_S, G_S) - I(H, G)| \, .$$

3.2 Sampling Methods

We compare eight strategies for selecting S^*:

- **Random:** Uniform baseline.
- **Stratified:** Quantile-based partitioning of G.
- **Disagreement:** Prioritizes items where multiple LLM judges diverge most.
- **Hybrid:** Combines stratified and disagreement sampling.
- **Active:** Greedy selection maximizing coverage and diversity.
- **Cluster-based:** K-means centroids in cheap rating space.
- **Variance-weighted:** Maximizes variance and range coverage to preserve ICC sensitivity.
- **Density-based:** Balances common and rare cases using kernel density estimates.

Formal definitions and derivations of each sampling strategy are provided in Appendix A.2. Additional information for judge model parameters is given in Appendix E.

4 Results

4.1 Simulation Data

We simulate ICC estimation under a 10% budget ($z = 30$ of $N = 300$ items) with a true ICC of 0.71, using Gaussian-distributed synthetic data. Figure 1 shows the average error across eight selection strategies. Several methods outperform random sampling, with **active**, **variance-weighted**, **cluster**, and **stratified** selection yielding the most accurate and stable ICC estimates. Confidence intervals are more narrow for these methods as compared to random selection, which can be found in Appendix C. Results are also robust across variations in dataset size, budget ratio, and true ICC values, as detailed in Appendix D.

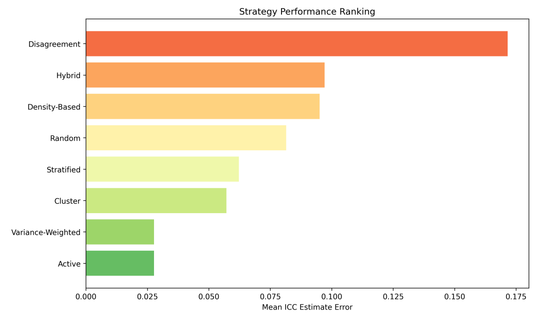

Figure 1: Difference between true ICC and predicted ICC by sampling method for Gaussian-distributed data scores.

4.2 Real-World Data

We next evaluate on three real-world datasets:
MSLR [Wang et al., 2023], HANNAStories [Chhun et al., 2024], and SummEval [Fabbri et al., 2021], each annotated along multiple axes (e.g., faithfulness, accuracy, creativity) with a total of 15 sets of human annotations by which to evaluate algorithmic performance. As shown in Table 1, the **cluster** method achieves the lowest mean error and reduced standard error compared to random, confirming that simulation findings generalize to real data. Across all datasets and tasks, cluster-based selection consistently provides more sample-efficient ICC estimation, with up to 41% relative improvement over random sampling in settings with <5% of annotation budget available.

Interestingly, all methods outperform random sampling with limited budget but different methods plateau at different rates. The clustering approach outperforms all methods (including random) at all evaluated budget ratios, indicating that it is more reliable as a selection proxy.

Budget	Active	Cluster	Density	Disagree	Hybrid	Random	Stratified	Var-Weighted
0.033	.276±.118	**.184±.122**	.286±.395	.240±.142	.226±.196	.311±.449	.211±.190	.265±.125
0.067	.226±.105	**.134±.106**	.194±.188	.187±.095	.178±.149	.188±.170	.154±.148	.229±.095
0.100	.181±.080	**.121±.095**	.161±.135	.145±.082	.150±.109	.131±.131	.144±.145	.184±.071
0.133	.160±.056	**.102±.098**	.157±.158	.134±.113	.143±.101	.108±.095	.111±.099	.158±.064
0.167	.155±.055	**.082±.059**	.130±.155	.141±.148	.115±.074	.095±.089	.101±.092	.153±.060
0.200	.135±.046	**.078±.060**	.114±.120	.118±.104	.114±.076	.085±.066	.088±.076	.132±.051
0.233	.119±.045	**.070±.047**	.093±.087	.103±.109	.103±.088	.080±.064	.080±.080	.118±.043
0.267	.107±.041	**.067±.061**	.070±.067	.096±.089	.098±.076	.069±.053	.086±.078	.107±.041
0.300	.098±.041	**.060±.051**	.076±.067	.086±.080	.087±.063	.064±.053	.081±.076	.098±.040
Avg	.162±.065	**.100±.078**	.142±.152	.139±.107	.135±.103	.126±.130	.117±.109	.161±.066

Table 1: Performance of sampling strategies across annotation budgets, where budget is fraction of total data allocated for human annotation. Bold indicates lowest mean error per row. The final row reports average mean error and standard error across budgets.

5 Discussion

Given a desired tolerance ε and probabilistic guarantee δ, a practitioner can derive an approximate bound on the minimum number of human annotations necessary to ascertain performance from Lemma 1. Future work can focus on removing the limiting assumptions and tighter bound techniques. We see from Table 1 that cluster-based selection consistently provides the most-sample efficient estimation of true ICC value across a large range of potential "budget" ratios. There is greatest improvememnt in low-budget settings, with relative improvement of 41% compared to random in settings where annotation budget is <5% of total samples. This allows model practitioners to iterate on their LLM judge methodology with higher fidelity without wasting annotation budget. Future work can explore extensions of these algorithms in order to further reduce expensive annotation requirements and exploring associated bounds with specific data selection mechanisms.

References

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. Medhelm: Holistic evaluation of large language models for medical tasks, 2025. URL `https://arxiv.org/abs/2505.23802`.

Douglas G. Bonett. Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21(9):1331–1335, 2002.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *International Conference on Machine Learning (ICML)*, pages 59–66, 2003.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. URL `https://arxiv.org/abs/2006.14799`.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952. doi: 10.1214/aoms/1177729330.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142, 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00689. URL `https://doi.org/10.1162/tacl_a_00689`.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3 edition, 1977.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. doi: 10.1007/BF02310555.

Yann Dubois et al. Alpacaeval 2.0: Automatic evaluation of instruction-following models. *arXiv preprint arXiv:2401.04088*, 2024.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021. URL `https://arxiv.org/abs/2007.12626`.

Valerii V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

Ronald A. Fisher. Statistical methods for research workers. 1925.

Fabrizio Gilardi, Meysam Alizadeh, and Matthias Kubli. Chatgpt outperforms crowdworkers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(23), 2023.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Douwe Kiela, Max Bartolo, et al. Dynabench: Rethinking benchmarking in nlp. In *NAACL*, 2021.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. 2024. URL https://arxiv.org/abs/2406.11939.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Percy Liang, Rishi Bommasani, et al. Holistic evaluation of language models. In *NeurIPS*, 2022.

David Liljequist, Britt Elfving, and Kirsti Skavberg Roaldsen. Intraclass correlation – a discussion and demonstration of basic features. *PLOS ONE*, 14:1–35, 07 2019. doi: 10.1371/journal.pone.0219854. URL https://doi.org/10.1371/journal.pone.0219854.

Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

Jayson G. Meyer, Ryan J. Urbanowicz, Pedro C. N. Martin, et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023. doi: 10.1186/s13040-023-00339-9.

Hieu T. Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In *International Conference on Machine Learning (ICML)*, pages 623–630, 2004.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

Daniel Salnikov. Concentration inequalities for the sample correlation coefficient, 2024. URL https://arxiv.org/abs/2401.12190.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

Burr Settles. *Active learning literature survey*. PhD thesis, University of Wisconsin-Madison, 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges, 2025. URL https://arxiv.org/abs/2410.12784.

Anuraj J. Thirunavukarasu, Daniel S. W. Ting, Kishore Elangovan, et al. Large language models in medicine. *Nature Medicine*, 29(9):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E Kuehl, Erin Bransom, and Byron C Wallace. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Toronto, Canada, 2023. Association for Computational Linguistics.

Alexander Wei et al. Active evaluation: Cost-efficient model selection and evaluation via adaptive sampling. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

Liqun Zheng, Wei-Lin Chiang, Yingfan Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

G.Y. Zou. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in medicine*, 31(29):3972—3981, December 2012. ISSN 0277-6715. doi: 10.1002/sim.5466. URL https://doi.org/10.1002/sim.5466.

A Appendix

A.1 Intra-class Correlation Coefficient Theoretical Analysis

A.1.1 Intra-class Correlation Coefficient Definitions

Intra-class Correlation Coefficient (ICC) was originally proposed by Fisher [1925] as an extension to *interclass* correlation coefficient (Pearson's correlation coefficient (PCC)), and measures the extent to which the total variance in observed data is due to differences between groups, rather than within groups. In this perspective, the ICC is understood within the analysis of variance (ANOVA) framework. As opposed to PCC, the data are pooled in the mean calculation.

In the generic version of our use case, ICC is considered a measure that quantifies inter-rater reliability between k raters on n subjects. ICC measures reliability by decomposing the total variance in human evaluations into between-subjects variance and within-subjects error variance. The ICC determines the reliability of ratings by comparing the variability of different ratings of the same individuals to the total variation across all ratings and all individuals. As we only consider two raters, the human and the LLM, we consider the case $k = 2$.

Analogously, modern ICC estimators derive ICC through the random effects model framework. In the random effects model, X_{ij}, rating j on subject i, $i \in [n], j \in [k]$, is modeled as

$$X_{ij} = \mu + \alpha_i + c_j + \varepsilon_{ij}$$

such that μ is an unobserved overall mean, α_i is an unobserved random effect shared by all ratings on subject i, c_j is an unobserved random effect shared by all ratings by subject j and ε_{ij} is an unobserved noise term. Each class of terms is assumed to be respectively identically distributed with expected value 0, and the terms are assumed to be uncorrelated. For certain random effects models, either α_i or c_j is neglected or considered fixed. We refer to Liljequist et al. [2019] for a comprehensive overview of ICC definitions and derivations relating classical estimators to random effects model. See table below for reproduced formulas.

Name	Notation	Rater Model	Use Case	Formula
One-way single	ICC(1,1)	Random	Agreement of 1 random rater	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R} + (k-1)\text{MS}_\text{E}}$
One-way average	ICC(1,k)	Random	Agreement of average random raters	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R}}$
Two-way absolute single	ICC(2,1)	Random	Absolute agreement of 1 random rater	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R} + (k-1)\text{MS}_\text{E} + \frac{k}{n}(\text{MS}_\text{C} - \text{MS}_\text{E})}$
Two-way absolute average	ICC(2,k)	Random	Absolute agreement of average raters	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R} + \frac{1}{n}(\text{MS}_\text{C} - \text{MS}_\text{E})}$
Two-way consistency single	ICC(3,1)	Fixed	Consistency of 1 fixed rater	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R} + (k-1)\text{MS}_\text{E}}$
Two-way consistency average	ICC(3,k)	Fixed	Consistency of average fixed raters	$\frac{\text{MS}_\text{R} - \text{MS}_\text{E}}{\text{MS}_\text{R}}$
Pearson correlation	r	N/A	Correlation only (not agreement)	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

Notation:

- MS_R: Mean square between targets (rows)
- MS_C: Mean square between raters (columns)
- MS_E: Residual mean square (error)
- n: Number of targets
- k: Number of raters

Formulas:

$$\mathrm{MS_R} = \frac{k}{n-1}\sum_{i=1}^{n}(S_i - \overline{X}_{\mathrm{tot}})^2$$

$$\mathrm{MS_C} = \frac{n}{k-1}\sum_{j=1}^{k}(M_j - \overline{X}_{\mathrm{tot}})^2$$

$$\mathrm{MS_E} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - M_j)^2 - k\sum_{i=1}^{n}(S_i - \overline{X}_{\mathrm{tot}})^2}{(n-1)(k-1)}$$

$$S_i = \frac{1}{k}\sum_{j=1}^{k} x_{ij}$$

$$M_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\overline{X}_{\mathrm{tot}} = \frac{1}{k \cdot n}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$$

In our specific use case, we use a two-way consistency average, i.e. $\mathrm{ICC}(3,k)$ this formulation treats *raters* as fixed effects, (i.e. c_j is fixed), meaning the same evaluation panel assesses all LLM outputs, and estimates reliability for the average rating across k evaluators rather than individual rater consistency. The numerator $(MS_R - MS_E)$ captures the true variance between different LLM responses after removing measurement error, while the denominator represents the total variance in averaged ratings, making $\mathrm{ICC}(3,k)$ particularly sensitive to systematic differences in how evaluators rate different model outputs while accounting for random measurement error within the evaluation process. With random effects model for $\mathrm{ICC}(3,k)$, the population ICC

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2/k}$$

We utilize the associated formula as the ICC metric for our experiments due to the appropriateness of the setting and random effects model. In our theoretical analysis, we provide bounds with $\mathrm{ICC}(3,1)$, as the expression resembles Fisher's original proposal for ICC and follows previous theoretical work [Zou, 2012]. Under $\mathrm{ICC}(3,1)$, the associated random effects model dictates that the population ICC

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

This is additionally the more commonly stated population ICC. As previously stated, we consider $k = 2$ only in both our empirical and theoretical results.

A.1.2 Chernoff Bound on Intra-class Correlation Coefficient

Given the population ICC stated in the previous section, we denote the sample ICC of n samples as $\hat{\rho}_n$ and calculate as the formula listed in our table for $\mathrm{ICC}(3,1)$. In Fisher [1925], Fisher demonstrates that with the assumption of sufficiently large number of samples, and given that $\hat{\rho}_n$ is not close to -1 nor 1, the distribution of $\hat{\rho}_n$ on bivariate Gaussian random variables asymptotically approaches Gaussian with parameters $\mathbb{E}[\hat{\rho}_n] = \rho$ and $\mathrm{Var}(\hat{\rho}_n) = \frac{(1-\rho^2)^2}{n-1}$. As earlier work on sample bounds for ICC leverage this approximation and associated assumptions [Zou, 2012], we consider these assumptions and approximations reasonable. As stated in Section 2.1, we assume a bivariate normal distribution for LLM-annotated and human annotated samples.

Thus, we only require a few additional (already extant) propositions to derive an approximate Chernoff bound.

Proposition 2 *[Chernoff, 1952] For any random variable X, the Chernoff bound dictates that*

$$\Pr(X \geq \varepsilon) \leq \inf_{\lambda \geq 0} \varphi_X(\lambda) e^{-\lambda\varepsilon}$$

where $\varphi_X(\lambda)$ is the moment generating function for X.

For a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, the moment generating function $\varphi_X(\lambda) = \exp(\mu\lambda + \sigma^2\lambda^2/2)$. Due to linearity of the Gaussian distribution, $X - \mu \sim \mathcal{N}(0, \sigma^2)$, and the moment generating function is $\varphi_X(\lambda) = \exp(\sigma^2\lambda^2/2)$. Thus, the Chernoff bound for a Gaussian random variable is

$$\Pr(X - \mu \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

Analogously,

$$\Pr(\mu - X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

Therefore, as the above events are mutually exclusive,

$$\Pr(|X - \mu| \geq \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

Combining this with the fact that $\hat{\rho_n}$ approaches Gaussian asymptotically with variance $\frac{(1-\rho^2)^2}{n-1}$, we obtain our desired approximate bound

$$\Pr[|\hat{\rho}_n - \rho| \geq \varepsilon] \lesssim 2\exp\left(-\frac{(n-1)\varepsilon^2}{2(1-\rho^2)^2}\right)$$

As standard with concentration inequalities, we can derive the necessary n such that $|\hat{\rho}_n - \rho| \geq \varepsilon$ with at most probability δ by setting δ equal to our bound and solving for n.

$$\begin{aligned}
\delta &= 2\exp\left(-\frac{(n-1)\varepsilon^2}{2(1-\rho^2)^2}\right) \\
\log\left(\frac{2}{\delta}\right) &= \frac{(n-1)\varepsilon^2}{2(1-\rho^2)^2} \\
\frac{2(1-\rho^2)^2}{\varepsilon^2}\log\left(\frac{2}{\delta}\right) &= (n-1) \\
1 + \frac{2(1-\rho^2)^2}{\varepsilon^2}\log\left(\frac{2}{\delta}\right) &= n
\end{aligned}$$

A.2 Selection Methods

A.2.1 Random Selection

The baseline random selection strategy serves as our control method:

$$S_{\text{random}} = \text{UniformSample}(\mathcal{N}, k)$$

where items are selected uniformly at random from the full set $\mathcal{N}$ without replacement.

A.2.2 Stratified Selection

Stratified selection partitions the cheap ratings into k quantile-based strata and selects one representative from each stratum:

$$Q_j = \text{Quantile}(G, \frac{j}{k}) \quad \text{for } j = 0, 1, \ldots, k$$

$$\text{Stratum}_j = \{i \in \mathcal{N} : Q_{j-1} \leq g_i \leq Q_j\}$$

$$S_{\text{stratified}} = \bigcup_{j=1}^{k} \text{UniformSample}(\text{Stratum}_j, 1)$$

A.2.3 Disagreement Selection

This strategy prioritizes items where multiple cheap raters exhibit maximum disagreement, under the hypothesis that such items are most informative:

$$d_i = |g_i^{(1)} - g_i^{(2)}| \quad \text{for } i \in \mathcal{N}$$

$$S_{\text{disagreement}} = \arg\max_{|S|=k} \sum_{i \in S} d_i$$

where $g_i^{(1)}$ and $g_i^{(2)}$ represent ratings from two different cheap judges.

A.2.4 Hybrid Selection

The hybrid approach combines stratified and disagreement-based selection:

$$S_{\text{hybrid}} = S_{\text{strat}}^{(k/2)} \cup S_{\text{disagree}}^{(k/2)}$$

where $S_{\text{strat}}^{(k/2)}$ contains $k/2$ items selected via stratification and $S_{\text{disagree}}^{(k/2)}$ contains the remaining items selected by disagreement, excluding those already chosen.

A.2.5 Active Selection

Active selection employs a greedy algorithm that iteratively selects items to maximize range coverage and rating diversity:

$$\text{Score}(S, i) = \frac{\max(G_{S \cup \{i\}}) - \min(G_{S \cup \{i\}})}{\max(G) - \min(G)} + 0.3 \cdot \text{Std}(G_{S \cup \{i\}})$$

$$S_{\text{active}} = \text{GreedyMax}(\text{Score}, k)$$

where $G_S = \{g_i : i \in S\}$ denotes the subset of cheap ratings corresponding to selection S.

A.2.6 Cluster-Based Selection

This method applies K-means clustering to identify k clusters in the cheap rating space and selects the item closest to each cluster centroid:

$$\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_z\} = \text{KMeans}(G, z)$$

$$S_{\text{cluster}} = \left\{ \arg\min_{i \in C_j} |g_i - c_j| : j = 1, 2, \ldots, k \right\}$$

where C_j represents the set of items assigned to cluster j and c_j is the corresponding cluster center.

A.2.7 Variance-Weighted Selection

Variance-weighted selection aims to preserve the between-item variance crucial for ICC computation by iteratively selecting items that maximize subset variance:

$$S_0 = \left\{ \arg\min_{i \in \mathcal{N}} |g_i - \text{Median}(G)| \right\}$$

$$S_{t+1} = S_t \cup \left\{ \arg\max_{i \in \mathcal{N} \setminus S_t} \left[\text{Var}(G_{S_t \cup \{i\}}) + 0.1 \cdot \text{Coverage}(S_t \cup \{i\}) \right] \right\}$$

where $\text{Coverage}(S) = \frac{\max(G_S) - \min(G_S)}{\max(G) - \min(G)}$ measures the range coverage of the selected subset.

A.2.8 Density-Based Selection

Density-based selection balances representation between high-density regions (typical cases) and low-density regions (outliers) using kernel density estimation:

$$\rho_i = \text{KDE}(g_i|G) \quad \text{for } i \in \mathcal{N}$$

$$S_{\text{high}} = \text{Sample}\left(\underset{|T|=k}{\arg\max} \sum_{i \in T} \rho_i, k/2\right)$$

$$S_{\text{low}} = \text{Sample}\left(\underset{|T|=k, T \cap S_{\text{high}}=\emptyset}{\arg\min} \sum_{i \in T} \rho_i, k/2\right)$$

$$S_{\text{density}} = S_{\text{high}} \cup S_{\text{low}}$$

where $\text{KDE}(g_i|G)$ represents the kernel density estimate of rating g_i given the distribution of all cheap ratings G.

B Simulation Robustness Analyses

In the main text (Section 4), we reported results using a fixed configuration ($N = 300$, $z = 30$, true ICC = 0.71). Here we provide additional robustness checks. We varied:

- **Subject set size** $N \in \{100, 200, 300, 400\}$,
- **Budget size** $z \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$,
- **True ICC values** across low, medium, and high agreement regimes.

Across all conditions, random sampling was consistently outperformed by cluster, activate, variance-weighted and density-based methods. These methods remained the most sample-efficient in both low- and high-variance regimes.

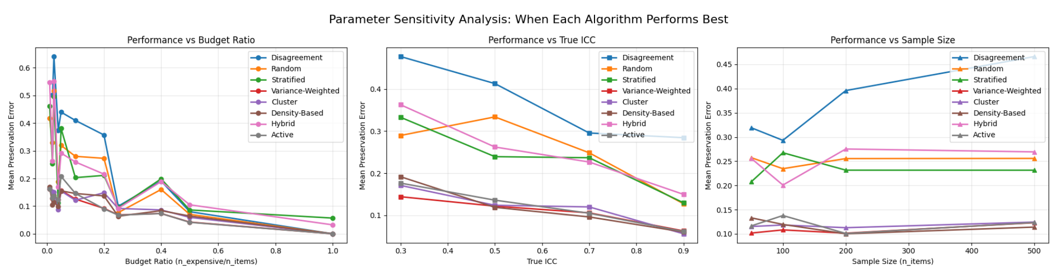

Figure 2: Simulation robustness: performance of sampling strategies across different dataset sizes, budgets, and true ICC values (rollouts = 5).

C Simulation Confidence Intervals

We see that clustering has a more narrow confidence interval as compared to random selection in simulation, supporting the claim that this method allows users to increase confidence in their ICC estimation from selected points. We can also compare other methods against random, identifying that random has the largest confidence interval in simulation outside of the "hybrid" approach. This supports the need for more systematic selection mechanisms to improve confidence of reported ICC score.

D Real Data Visualization

We plot mean preservation error by budget aggregated across all datasets as a corollary to Table 1 such that we can visualize ICC improvement. We see from the below visualization that cluster-based selection for these 15 tasks remains on the pareto-frontier of performance, and additionally that all methods outperform random at low human annotation budget (number of expensive ratings), but cluster-based selection continues to outperform random as budget increases.

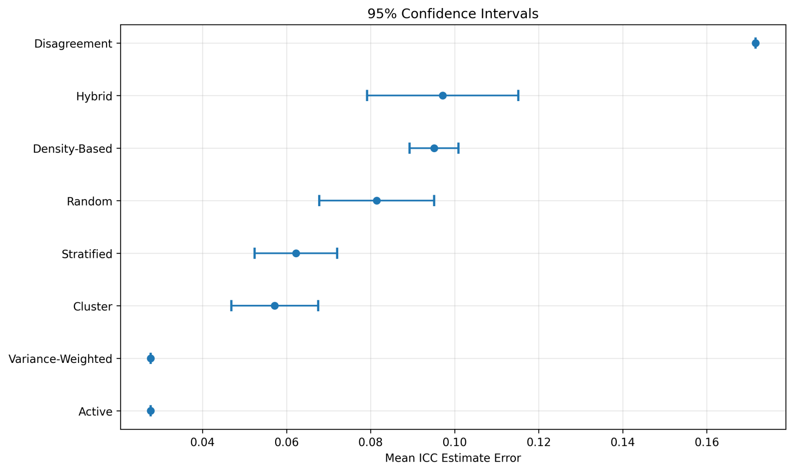

Figure 3: Simulation robustness: performance of sampling strategies across different dataset sizes, budgets, and true ICC values (rollouts = 5).

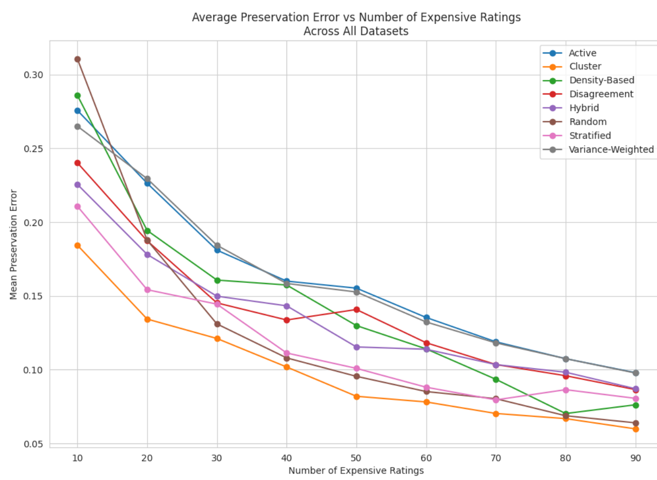

Figure 4: Simulation robustness: performance of sampling strategies across different dataset sizes, budgets, and true ICC values (rollouts = 5).

E Judge Model Information

We use GPT-4o-mini as our model judge for this setting due to the balance of accuracy and cost. We use temperature 0.7, and sample twice from the model when needed for disagreement-based selection methods. Future work should involve exploring generalizability of claims across different judge model architectures and families.