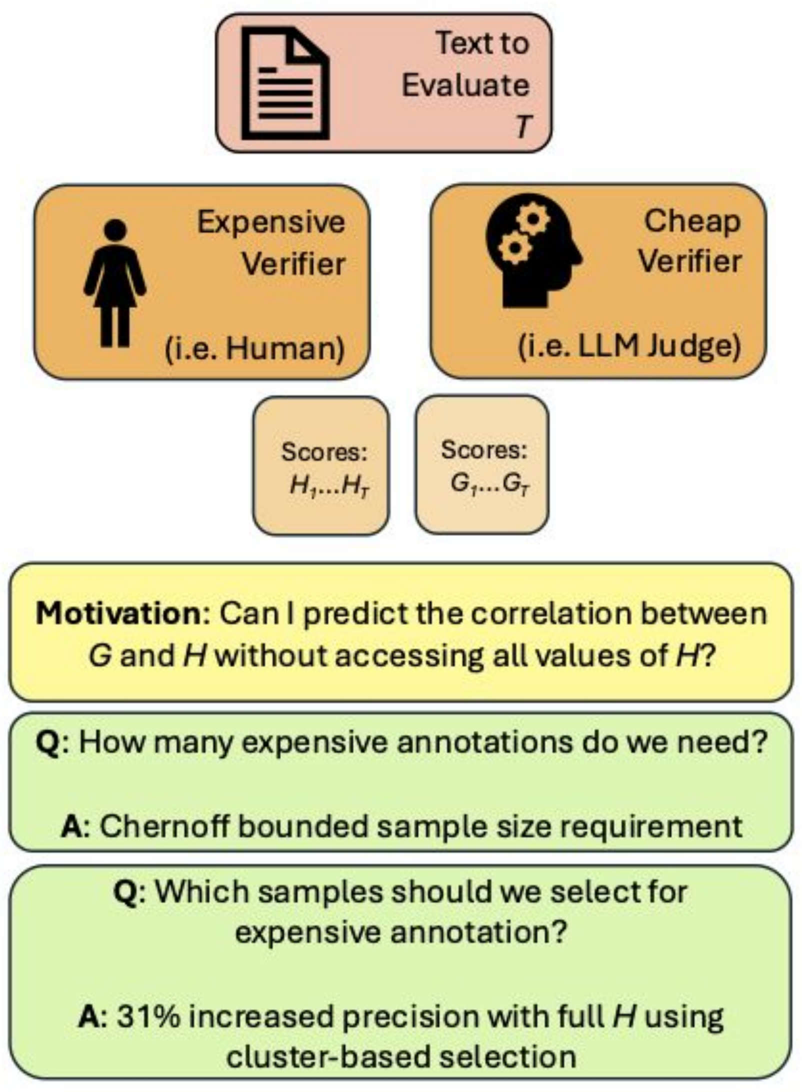
Smarter Sampling for LLM Judges: Reliable Evaluation on a Budget

Alyssa Unell*, Natalie Dullerud*, Nigam Shah, Sanmi Koyejo

Background



- LLM-as-a-judge offers options for scalable AI evaluation but reliability depends on human alignment.
- Expert annotation is costly, especially in specialized domains.
- We establish a lower bound on the number of annotations needed to accurately measure reliability.
- We present an initial panel of annotation sampling methods.

Methods

Measurement of reliability: Intraclass correlation coefficient (ICC)

In order to quantify reliability of LLM scores relative to human annotations, we rely on intraclass correlation coefficient. Under random effects model,

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

X_{ij} : rating *j* (LLM or human) on text *i*
μ : population mean
α_i : random effect on all ratings on text *i*
c_j : random effect on all texts from rater *j*
ε_{ij} : random noise term

All the random effect terms are assumed unobserved. Thus, population ICC defined as:

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

We estimate ICC as:

$$\hat{\rho} = \frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R}}$$

$\mathrm{MS_R}$: mean square error over rows (text)
$\mathrm{MS_E}$: residual mean square error

Suite of sampling Methods derived from active testing & learning

Given dataset $\mathcal{X}$ of size *n*, (observed) cheap labels *G* and (unobserved) human labels *H*, and a budget *b* < *n*, we seek a subset S^*, $|S^*| = b$, such that the ICC estimand on S^*, $\hat{\rho}_b$, closely approximates the ICC estimand on $\mathcal{X}$, $\hat{\rho}_n (\approx \rho)$

$$S^* = \arg\min_{S \subseteq \mathcal{X}, |S|=b} |\hat{\rho}_b(H_S, G_S) - \hat{\rho}_n(H, G)|$$

Methods (Cont.)

The following sampling strategies are investigated:

- Random
  $$S_{\text{rand}} = \text{UniformSample}(\mathcal{N}, k)$$
- Stratified
  $$S_{\text{strat}} = \bigcup_{j=1}^{k} \text{UniformSample}(\text{Stratum}_j, 1)$$
- QBC
  $$S_{\text{QBC}} = \arg\max_{|S|=k} \sum_{i \in S} |g_i^{(1)} - g_i^{(2)}|$$
- Stratified QBC
  $$S_{\text{sQBC}} = S_{\text{strat}}^{(k/2)} \cup S_{\text{QBC}}^{(k/2)}$$
- Cluster
  $$S_{\text{clust}} = \{\arg\min_{i \in C_j} |g_i - c_j|\}_{j=1}^{k}$$
- Maximum-variation
  $$S_{t+1} = S_t \cup \{\arg\max_{i \notin S_t} \dots\}$$
- Density-based
  $$S_{\text{dens}} = S_{\text{high}} \cup S_{\text{low}}$$

Theoretical Results

Chernoff Bound for ICC

Given *H*, *G*, random variables s.t. $H_i, G_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ i.i.d. Let ρ denote the population ICC, $\hat{\rho}_n$ denote the estimated ICC on sample size *n*. Given ε > 0, *n* sufficiently large for CLT and $|\rho|$ not close to 1,

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

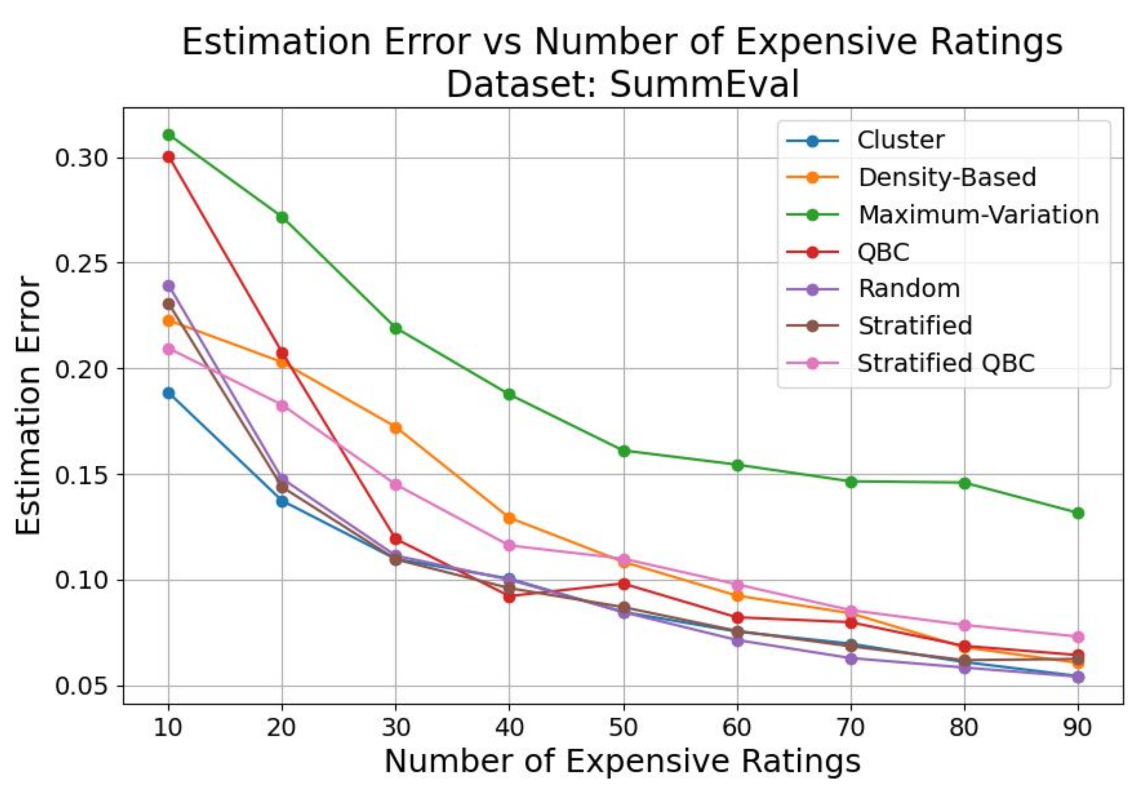
Therefore, given δ > 0, with probability 1-δ, the population and estimated ICC are guaranteed to be ε - close if

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

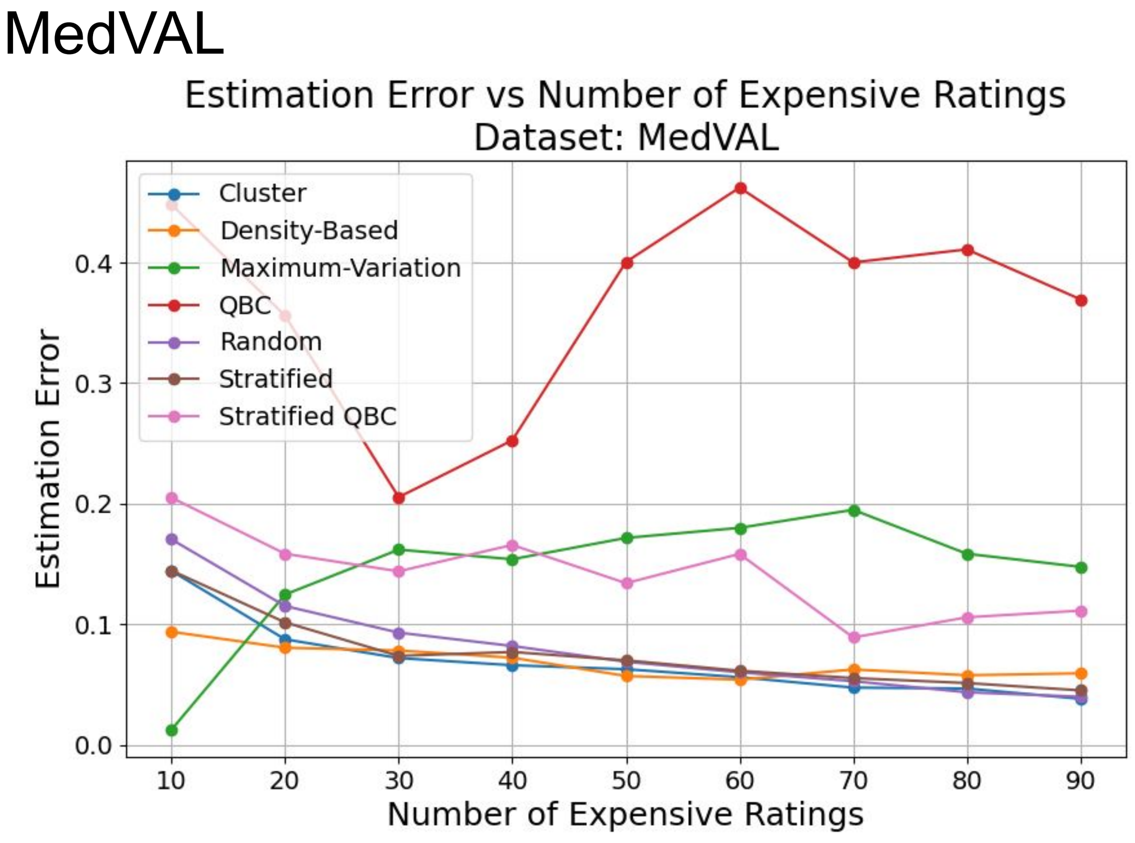
Empirical Results

Panel of Sampling Methods: Results

SummEval

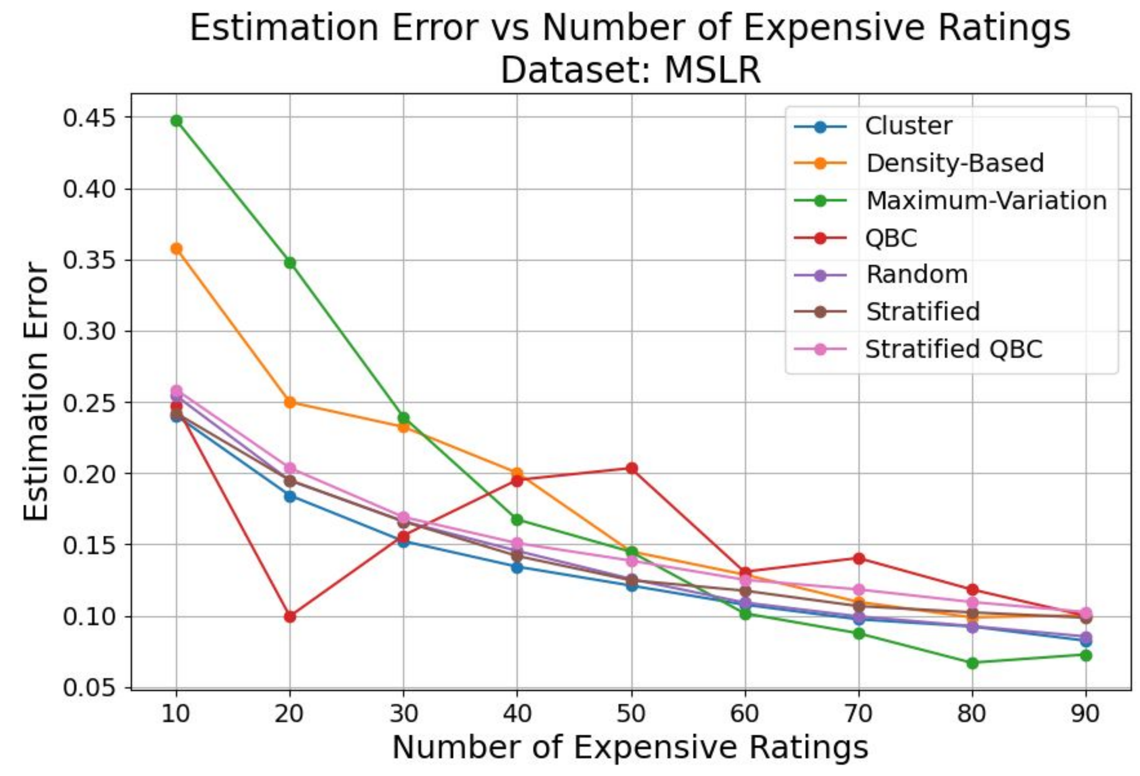


MedVAL

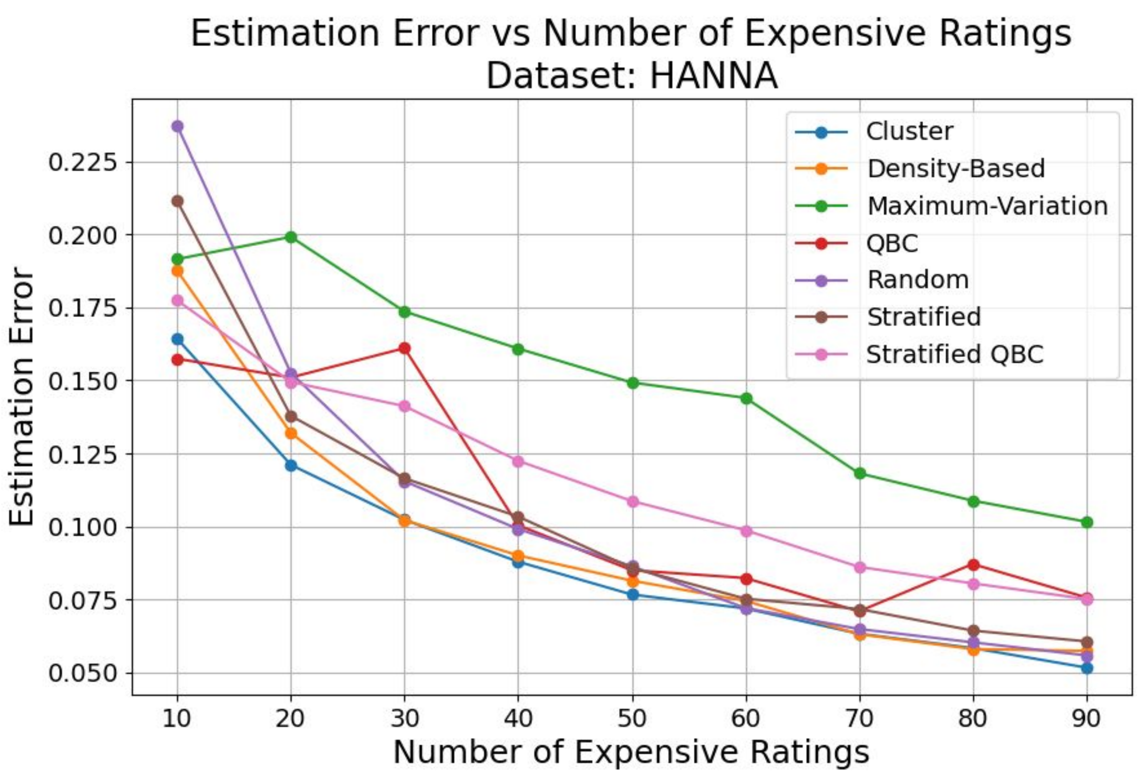


Empirical Results

MSLR



HANNA



Best Sampling Method: Cluster

$n_{\text{expensive}}$	Mean ICC Improvement over Random (%)			
	HANNA	MedVAL	MSLR	SummEval
10	31.0%	15.0%	5.5%	21.0%
20	21.0%	24.0%	5.4%	7.1%
30	11.0%	23.0%	8.4%	1.4%
40	11.0%	20.0%	7.6%	0.0%
50	11.0%	9.0%	3.7%	0.0%
60	0.0%	6.8%	1.3%	−5.6%
70	2.5%	10.0%	2.0%	−11.0%
80	3.2%	−7.1%	0.0%	−4.4%
90	7.5%	4.2%	3.4%	0.0%

$n_{\text{expensive}}$	CI Width Improvement of Cluster over Random (%)			
	HANNA	MedVAL	MSLR	SummEval
10	6.4%	32.4%	−0.9%	24.1%
20	7.8%	18.3%	0.1%	21.9%
30	7.1%	13.9%	−2.4%	18.4%
40	2.6%	9.7%	−1.0%	14.6%
50	4.9%	9.4%	−0.7%	14.0%
60	2.9%	2.1%	0.0%	11.3%
70	3.5%	8.0%	−0.4%	9.1%
80	1.5%	5.9%	0.0%	7.8%
90	1.8%	4.5%	−1.3%	6.2%

Table 1. Cluster-based sampling can decrease estimation error and improve confidence intervals in low data settings.

Discussion

- Our **Chernoff bound for ICC** relies on assumption of *normality* of samples, and *sufficient samples* for CLT s.t. distribution of ICC approaches normality. We provide tighter bounds than previous work in most parameter setting – future work could remove assumptions.
- Our **cluster-based sampling approach** can improve ICC estimation at low budget settings by up to 31%. Future work could explore further algorithm adjustments to improve generalizability and gain.