ON THE CALIBRATION OF CROWDSOURCING AGGREGATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration has been extensively studied for classifiers, but remains largely unexplored for crowdsourcing aggregation methods, which are increasingly used for making probabilistic decisions. To the best of our knowledge, we provide the first theoretical and empirical study of calibration in label aggregation. To our great surprise, we find that widely used aggregation methods can be severely miscalibrated, and in opposite directions: Majority Vote is underconfident while model-based aggregators are overconfident. We prove that the Bayesian aggregator is perfectly calibrated under its data-generating model. We also show that exact calibration of weighted majority votes requires highly restrictive conditions, which helps explain why Majority Vote is generally miscalibrated in realistic crowds. Empirically, we study calibration across 39 human crowdsourcing datasets and LLM-based crowds. In both cases, we observe systematic miscalibration for aggregation models such as Majority Vote, Dawid–Skene, GLAD, and CrowdFM. Finally, we propose a post-hoc calibration procedure that *aggregates the aggregators*: for any aggregation function ϕ, we introduce a transformation that first combines its predictive distribution with those of complementary aggregators, before applying temperature scaling. This simple and interpretable transformation makes temperature scaling more effective than when applied directly to ϕ. It improves calibration, especially in the high-redundancy regime where there are many annotations per item, while generally preserving or even improving accuracy.

1 INTRODUCTION

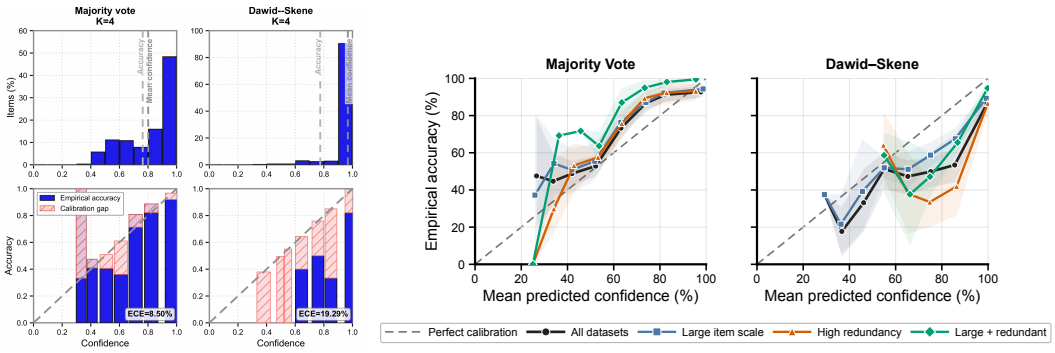
Modern machine learning relies on large and carefully labeled datasets, and crowdsourcing has become a standard way to collect such labels (Ibrahim et al., 2025): each item is sent to several annotators, and their responses are then aggregated into a single prediction. This approach makes data annotation scalable, but it also introduces noise. Annotators may have different levels of expertise, different biases, and different error patterns. Historically, crowdsourcing has mainly relied on human annotators. This setting is now changing. Large language models (LLMs) are increasingly used as automatic annotators (Gilardi et al., 2023). Recent studies show that LLMs can match or even outperform human crowd workers on some annotation tasks (Plaza-del Arco et al., 2024).

Most crowdsourcing methods focus on recovering an accurate label from multiple noisy annotations, and their theoretical properties have been extensively studied (Ibrahim et al., 2025). However, most aggregation methods also return a probability distribution over the possible classes (Whitehill et al., 2009; Liu et al., 2026), which naturally raises the question of their calibration.

Calibrated probabilities are important when the output of an aggregation model is interpreted as a risk, and ensures that estimated probabilities correspond closely to observed frequencies. Poorly calibrated predictions can therefore lead to systematic over- or underestimation of risk and, consequently, to bad decisions.

Typically, in medicine, machine learning models can estimate the probability for a patient to develop a disease, experience a complication, or respond to a treatment, thereby helping to inform treatment decisions (Martel et al., 2026; Van Calster et al., 2025). Insufficiently calibrated models can lead to inappropriate and potentially harmful clinical decisions (Van Calster et al., 2019). Aggregation offers complementary opportunities to support clinical decision-making: studies have shown that

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.



(a) Adult dataset **(b)** Average reliability over 39 datasets

Figure 1: Calibration of Majority Vote and Dawid–Skene. **(a)** Confidence histograms (top) and reliability diagrams (bottom) on the Adult dataset (Wright & Augenstein, 2025), with Expected Calibration Error (ECE, see Eq. equation 6) reported for each aggregator. K is the number of classes. **(b)** Reliability diagrams averaged over 39 datasets; points above the diagonal indicate *underconfidence* and points below indicate *overconfidence*.

combining independent diagnoses can improve accuracy, both when aggregating physicians' outputs (Barnett et al., 2019) and when combining differential diagnoses generated by multiple LLMs (Barabucci et al., 2024). This motivates examining the calibration of aggregation methods in medical settings.

Contributions

1. To the best of our knowledge, we provide the first empirical and theoretical study of calibration in crowdsourcing. Our empirical study is done across 40 real-world datasets and one specialized medical dataset. Two of them involve LLM crowds. To our great surprise, we discover that Majority Vote, Dawid–Skene, GLAD, and CrowdFM models (see Section 2 for the related work) are all severely miscalibrated: Majority Vote is *underconfident* while model-based aggregators are *overconfident* (see Figure 1).
2. We characterize theoretically, in the multiclass setting, when crowd aggregation can be calibrated. We first prove that the Bayesian aggregator is perfectly calibrated, and derive necessary and sufficient conditions for asymptotic exact calibration of weighted majority vote. What is shown is that Majority Vote is generally asymptotically hard to calibrate.
3. We introduce a transformation of any aggregation function ϕ that makes its probability output more effective under temperature scaling, and we wrap our approach in Algorithm 1. Compared to directly applying temperature scaling on ϕ, our transformation improves the calibration of weighted-vote aggregators, and beyond those. For all our baselines, in the high-redundancy regime we improve calibration while preserving predictive performance.

2 Related Work

Learning from Crowd-Sourced Labels. Several label aggregation methods have been proposed in the literature. Probabilistic approaches include Dawid–Skene (Dawid & Skene, 1979), GLAD (Whitehill et al., 2009), and CrowdFM (Liu et al., 2026). These methods have mainly been studied through prediction error, identifiability, sample complexity, and assumptions on annotator dependence (Ibrahim et al., 2025). Weighted majority votes form another important family. Berend & Kontorovich (2014) study the consistency of the optimal weighted majority rule, while Crowd-Kit (Ustalov et al., 2024) provides practical implementations of weighted variants such as Wawa (Ustalov et al., 2024) and ZBS (Ustalov et al., 2024).

Calibration. The calibration of probabilistic predictors has been extensively studied outside the crowdsourcing literature. Guo et al. (2017) show that modern neural networks are poorly calibrated despite improved accuracy, and benchmark several post-hoc calibration techniques – histogram bin-

ning, isotonic regression, Bayesian binning into quantiles, and Platt scaling – introducing *temperature scaling* as a simple, effective extension of the latter. Temperature scaling has since been applied well beyond image classification, notably to calibrate pre-trained language models (Desai & Durrett, 2020). Prior work has documented limitations of temperature scaling – namely its inability to correct classwise miscalibration (Kull et al., 2019) or to guarantee a verified calibration error (Kumar et al., 2019) – and, in line with these findings, we observe that it fails to correct the miscalibration of weighted-majority-vote-type aggregators. Rather than replacing temperature scaling itself, we address this limitation by acting on the aggregator being calibrated.

LLMs as annotators. Large Language Models (LLMs) have recently emerged as a scalable alternative to human annotators (Bhat & Varma, 2023). Their use as labelers seems to be promising, especially in text classification (Gilardi et al., 2023; Rønningstad et al., 2024). The interested reader can also find the review of Tan et al. (2024) about the use of LLMs for data annotation and synthesis.

3 Problem setting: a probabilistic point of view

The primary goal of label aggregation in crowdsourcing is to recover the unknown ground-truth label of an item from multiple noisy annotations.

Preliminaries. Let $R \geq 1$ be the number of annotators, let $Y_1, \dots, Y_R$ be R random variables. We denote by Y the unknown ground-truth label taking a value in $[K] := \{0, \dots, K-1\}$. We consider that $Y_1, \dots, Y_R$ are independent conditionally to Y and all taking values in $[K]$.

A probabilistic aggregation model ϕ is a function that takes as input random variables and outputs a probability distribution $\phi : [K]^R \longrightarrow \Delta_K$, where $\Delta_K := \left\{q \in [0,1]^K : \sum_{k\in[K]} q_k = 1\right\}$.

The goal of the model ϕ is to estimate the posterior distribution $(\mathbb{P}(Y = k \mid Y_1, \dots, Y_R))_{k\in[K]}$. The final aggregated label is then obtained by

$$Y_R^{\phi} \in \arg\max_{k\in[K]} \phi_k(Y_1, \dots, Y_R). \tag{1}$$

Such probabilistic aggregation models include classical methods such as Dawid–Skene (Dawid & Skene, 1979) and GLAD (Whitehill et al., 2009). It also covers more recent deep learning models such as CrowdFM, a foundation model pretrained on synthetic data Liu et al. (2026). Even majority voting can naturally be interpreted in a probabilistic way: for any $r \in \{1, \dots, R\}$, set $\phi_k(Y_1, \dots, Y_R) = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\{Y_r = k\}$.

Probabilistic calibration. Unlike the confidence-based notion (Guo et al., 2017), which only calibrates the predicted class confidence $c_\phi(R) := \max_{k\in[K]} \phi_k(Y_1, \dots, Y_R)$, we consider calibration of the full distribution $\phi(Y_1, \dots, Y_R)$. This distinction matters for $K \geq 3$, since a predictor may be calibrated with respect to its top$_k$-class confidence while remaining miscalibrated at the level of the full distribution as shown in Vaicenavicius et al. (2019). Note that, being calibrated in terms of the full distribution implies to be also calibrated with respect to the subset of its top$_k$-class. Readers interested in a detailed example may refer to Appendix A.3.

Definition 1. *(Calibrated model) We say that ϕ is calibrated if and only if for any $k \in [K]$,*

$$\mathbb{P}(Y = k \mid \phi(Y_1, \dots, Y_R)) = \phi_k(Y_1, \dots, Y_R) \qquad \textit{a.s.} \tag{2}$$

Calibration Error. The condition equation 2 must hold almost surely, in practice, this condition is impossible to check, so one instead considers their average difference. This naturally leads to the calibration error (Kumar et al., 2019; Vaicenavicius et al., 2019; Popordanoska et al., 2022). Let $Z = (\mathbb{I}\{Y = 1\}, \dots, \mathbb{I}\{Y = K\}) \in \{0,1\}^K$ denote the one-hot encoding of the ground-truth label, and let $\Phi_R := \phi_R(Y_1, \dots, Y_R) \in \Delta_K$.

Definition 2. *For $p \in [1, +\infty[$, the L^p calibration error of ϕ_R is defined as*

$$\mathrm{CE}_p(\phi_R) := \left(\mathbb{E}\left[\|\mathbb{E}[Z \mid \Phi_R] - \Phi_R\|_p^p\right]\right)^{1/p}.$$

For $p = 1$, this quantity corresponds to the L^1 calibration error, which has been studied theoretically in Kumar et al. (2019); Vaicenavicius et al. (2019); Popordanoska et al. (2022). Following Cannings et al. (2020) and Singer et al. (2026), we study aggregation in the regime $R \to \infty$.

4 THEORETICAL ANALYSIS OF THE CALIBRATION FOR WEIGHTED VOTES

Here, we study theoretically the calibration of weighted majority votes. Among the most famous one, you have the Bayesian vote, which minimizes the classification error, and the standard Majority Vote. In this section, we identify annotator number i with their response random variable Y_i , thus this can be applied any type of annotator, human or LLM.

The Bayesian vote is perfectly calibrated. The first fact is that Bayesian vote, that can be written as a weighted majority vote see (Singer et al., 2026), is perfectly calibrated.

Proposition 1. *The Bayesian vote defined as* $\phi^\star = \arg\min_\phi \mathbb{P}\Big(Y_R^\phi \neq Y \mid Y_1, ..., Y_R\Big)$ *, is calibrated in the full distribution sense,top*$_k$ *class sens and* L^p *sens for any* $p \geq 1$.

Following Vaicenavicius et al. (2019), it appears that any aggregation method defined by the mean of a function h as follows: $\phi_h : (Y_1, ..., Y_R) \mapsto (\mathbb{P}(Y = k \mid h(Y_1, ..., Y_R)))_{k\in[K]}$ is also calibrated.

Analysis of weighted Majority Vote calibration. For $\omega^{(R)} = (\omega_1^{(R)}, \ldots, \omega_R^{(R)}) \in \Delta_R$, we say that Φ_R is weighted majority vote if one can right $\Phi_R := \left(\sum_{r=1}^R \omega_r^{(R)} \mathbb{I}\{Y_r = k\}\right)_{k\in[K]}$ Let $\mathcal{G}$ denote the class of sequences weighted majority vote $(\Phi_R)_{R\geq 1}$ such that

$$\lim_{R\to\infty} \left\|\omega^{(R)}\right\|_2 = 0 \text{ and } \Phi_R \mid Y = j \xrightarrow{\mathbb{P}} A^{(j)} \quad \text{for every } j \in [K]. \tag{3}$$

where $A^{(1)}, \ldots, A^{(K)} \in \Delta_K$ are deterministic and pairwise distinct. The assumption that $(A^{(j)})_{j\in[k]}$ is deterministic corresponds to the standard law-of-large-numbers regime for crowdsourcing, it includes in particular the classical Condorcet Jury Theorem (Berend & Sapir, 2007). One can show, see Appendix A.1, that equation 3 defines uniquely the limite aggregator as $\Phi := \lim_{R\to\infty} \Phi_R = A^Y$. Condition equation 3 only requires the crowd average accuracy to stabilize at ∞. So, it allows a lot of different heterogeneous annotators configurations.

Proposition 2. *The class* $\mathcal{G}$ *is nonempty. Majority Vote belongs to* $\mathcal{G}$ *and Wawa also belongs to* $\mathcal{G}$.

Theorem 1. *For any* $\Phi \in \mathcal{G}$ *and* $p \geq 1$ *: the following conditions are equivalent: if*

$$\lim_{R\to\infty} \mathrm{CE}_p(\phi_R) = 0 \tag{4}$$

Then, for any $(k, j) \in [K]^2$*:*

$$\delta_{k,j} = \lim_{R\to\infty} \sum_{1\leq r\leq R} \omega_r^{(R)} \mathbb{P}\left(Y_r = j \mid Y = k\right), \tag{5}$$

and reciprocally.

Where $\delta_{k,j}$ denotes the Kronecker delta, i.e., $\delta_{k,j} = 1$ if $k = j$ and $\delta_{k,j} = 0$ otherwise.

If an aggregator is not asymptotically perfectly efficient it cannot be calibrated.

Corollary 1. *Let* $\phi \in \mathcal{G}$*, if* $\limsup_{R\to\infty} \mathbb{P}\left(Y_R^\phi \neq Y\right) > 0$*, then* ϕ *is not asymptotically calibrated.*

Interpretation. The condition equation 5 maybe interpreted as follows: consider a growing crowd with respect to the number R of annotators. If for each class k, the average accuracy of the crowd is close to be perfect then, on average it's calibration error is close to 0. So taking the contraposition of the direct implication it shows you that if you do not have a very accurate crowd, you are not calibrated. This interpretation is explicitly proven in Corollary 1.

5 TEMPERATURE SCALING FOR PROBABILISTIC AGGREGATION FUNCTIONS

We adopt temperature scaling as our calibration method for its simplicity and single-parameter form, which makes it more robust to overfitting on limited calibration data than more flexible alternatives such as Dirichlet calibration (Kull et al., 2019); moreover, temperature scaling is order-preserving, allowing us to recalibrate an aggregator's confidence without altering its underlying decisions.

Calibration and post-hoc calibration. A probabilistic predictor is calibrated when its predicted probabilities match empirical frequencies. For instance, predictions made with confidence 0.8 should be correct about 80% of the time. Post-hoc calibration aims to correct predicted probabilities without retraining the original model. A standard method is temperature scaling (Guo et al., 2017), which transforms the estimated probability of the item i to be in class k : $\widehat{p}_{ik}$. For $T > 0$, $\widetilde{p}_{ik}(T) = \widehat{p}_{ik}^{1/T} / \sum_{\ell=1}^{K} \widehat{p}_{i\ell}^{1/T}$. The temperature is fitted on a separate calibration set by minimizing the negative log-likelihood. If we denote Y_i the true label, $T^\star := \arg\min_{T>0} - \sum_{i=1}^{n_{\text{cal}}} \log \widetilde{p}_{iY_i}(T)$.

Temperature scaling for label aggregation. Let $\phi : [K]^R \to \Delta_K$ be a probabilistic aggregation function and let $Q_i^\phi = \phi(Y_{i,1}, \ldots, Y_{i,R})$. Temperature scaling transforms Q_i^ϕ through a single scalar parameter $\frac{1}{T}$, preserving the class ranking while only rescaling its log-ratios. More flexible calibrators, such as Dirichlet calibration (Kull et al., 2019), increase the expressiveness of this transformation across classes.

A crowdsourcing-aware correction. Given a target aggregator ϕ and a $\mathcal{R} = \{\psi_1, \ldots, \psi_M\}$ a fixed set of reference aggregators. Our intuition is that aggregators with complementary calibration profiles can compensate for each other before temperature scaling. The role of ϕ is to preserve the predictive accuracy of the target aggregator. By design the reference set $\mathcal{R}$ contains any collection of aggregators with complementary calibration profiles. Temperature scaling is then fitted on $Q_i^{\text{Soup},\phi}$.

Algorithm 1 Crowdsourcing-Aware Temperature Scaling

Require: Fixed set of aggregators $\mathcal{R} = \{\psi_1, \ldots, \psi_M\}$,
Calibration set $\mathcal{D}_m^{\text{cal}} = \left\{ \left(Y_i, \{Q_i^\psi\}_{\psi \in \mathcal{R}} \right) \right\}_{i=1}^{m}$,
Predictive distributions $\{Q_j^\psi\}_{\psi \in \mathcal{R}}$ for the items j to calibrate.
Output: Calibrated predictive distributions $\{\widetilde{Q}_j^{\text{Soup},\phi}\}_j$.

$\text{Soup}\left(\{Q^\psi\}_{\psi \in \mathcal{R}}\right) \leftarrow \frac{1}{|\mathcal{R}|} \sum_{\psi \in \mathcal{R}} Q^\psi$ ▷ Define the aggregation operator
for $i = 1, \ldots, m$ **do**
  $Q_i^{\text{Soup},\phi} \leftarrow \text{Soup}\left(\{Q_i^\psi\}_{\psi \in \mathcal{R}}\right)$ ▷ Build calibration Soup
end for
$\widehat{T}_m^\phi \leftarrow \arg\min_{T>0} -\frac{1}{m} \sum_{i=1}^{m} \log \left[\mathcal{T}_T \left(Q_i^{\text{Soup},\phi} \right) \right]_{Y_i}$ ▷ Fit temperature on calibration data
for each item j to calibrate **do**
  $Q_j^{\text{Soup},\phi} \leftarrow \text{Soup}\left(\{Q_j^\psi\}_{\psi \in \mathcal{R}}\right)$ ▷ Build prediction Soup
  $\widetilde{Q}_j^{\text{Soup},\phi} \leftarrow \mathcal{T}_{\widehat{T}_m^\phi} \left(Q_j^{\text{Soup},\phi} \right)$ ▷ Apply learned temperature
end for
return $\{\widetilde{Q}_j^{\text{Soup},\phi}\}_j$

6 Experiments

6.1 Datasets and methods presentations

Datasets annotated by humans. Regarding human annotators, we evaluate the methods on 39 real-world crowdsourcing datasets. Including image annotation, sentiment analysis, natural-language tasks, and toxicity detection. Among others, we have: LabelMe (Russell et al., 2008), the NLP crowdsourcing tasks of Snow et al. (2008), Jigsaw toxicity data (Borkan et al., 2019; Jigsaw), and CIFAR-10H (Peterson et al., 2019). Additional crowdsourcing benchmarks are taken from Crowd-Kit (Ustalov et al., 2024) and CrowdFM (Liu et al., 2026). We consider four regimes: all datasets (39 datasets), large-scale datasets with at least $1{,}000$ items (21 datasets), high-redundancy datasets with at least 10 annotations per item on average (19 datasets), and their intersection (6 datasets).

Dataset annotated by LLMs. LLMs were evaluated on two datasets. Firstly, to compare the calibration of aggregated human and LLM predictions, we use the crowd-annotated version of the PASCAL RTE-1 benchmark (Dagan et al., 2006; Snow et al., 2008), one of the 39 real-world crowdsourcing datasets annoted by humans. This is a binary textual entailment task which requires determining whether a hypothesis can be inferred from a given textual premise. The dataset contains 800 Text-Hypothesis pairs, each with a reference label and each pair annotated by ten human judgments collected from a pool of 164 non-expert annotators. We obtain these existing human annotations from the data released with Wright & Augenstein (2025)[1] and collect LLM predictions on the same pairs. We then apply the same aggregation rules to human and LLM responses and evaluate the calibration of the resulting predictions against the reference labels.

As discussed in the introduction, the importance of calibrated predictions for clinical decision-making motivates examining the calibration of aggregation methods in medical settings. For that purpose, we decided to evaluate LLMs on a second dataset, a medical benchmark: MMLU Clinical knowledge (Hendrycks et al., 2021). This benchmark evaluates clinical knowledge through a test set of 265 multiple-choice questions, each with four answer options.[2] Results on this dataset are reported in A.4.

Design of the experiments. We evaluate eleven models accross three open-source models families (Qwen, Mistral, and Gemma) and one closed-source model (from OpenAI). Because we run the open-source models on an Apple M5 integrated GPU, we restrict them to fewer than 10 billion parameters. Table 2 in appendix lists the different models, with HuggingFace's IDs and generation parameters.

To reproduce the human evaluation, we apply the same aggregation rules, treating R distinct models as annotators.

6.2 Miscalibration of aggregators

We follow the standard empirical Expected Calibration Error (ECE) Guo et al. (2017). Given M confidence bins $(B_m)_{m=1}^{M}$,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{6}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the empirical accuracy and mean confidence in bin B_m. Lower values indicate better calibration.

Both for human and LLM annotators, Majority Vote shows underconfidence, while Dawid–Skene tend to be overconfident (Figure 2). On RTE dataset and the LLM crowd dataset, Dawid–Skene reaches respectively ECE of 6.8% and 8.9%, while MV reaches respectively ECE of 16.9% and 1.5%.

6.3 Temperature Scaling and Soup

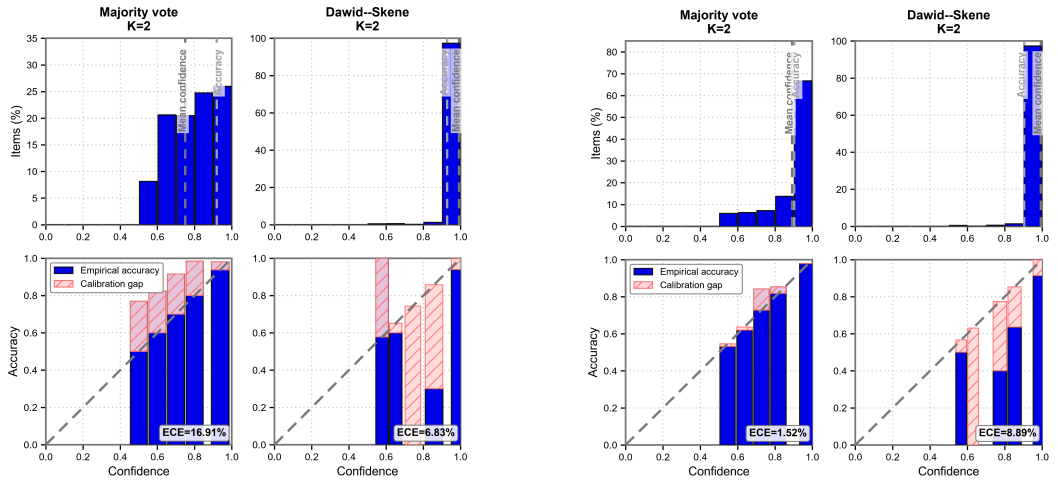
In our experiments, we use $\mathcal{R} = \{\mathrm{MV}, \mathrm{DS}\}$ because the empirical analysis in Figure 2 reveals a clear complementarity between Majority Vote and Dawid–Skene: Majority Vote tends to be under-confident, whereas Dawid–Skene tends to be over-confident. Therefore, in practice with our algorithm 1, we set

$$Q_i^{\mathrm{Soup},\phi} := \frac{1}{3}\left(Q_i^{\mathrm{MV}} + Q_i^{\mathrm{DS}} + Q_i^{\phi}\right).$$

Results Table 1 shows improvements in terms of ECE in several settings. On human datasets, for weighted majority votes, our Soup+TS systematically decreases the ECE compared to classic temperature scaling, while for model-based aggregators, temperature scaling alone remains slightly

[1] PASCAL RTE-1 accessed via:
https://github.com/copenlu/aggregating-crowd-annotations-ood/tree/main/data/RTE

[2] MMLU dataset accessed via Hugging Face:
https://huggingface.co/datasets/cais/mmlu/viewer/clinical_knowledge.

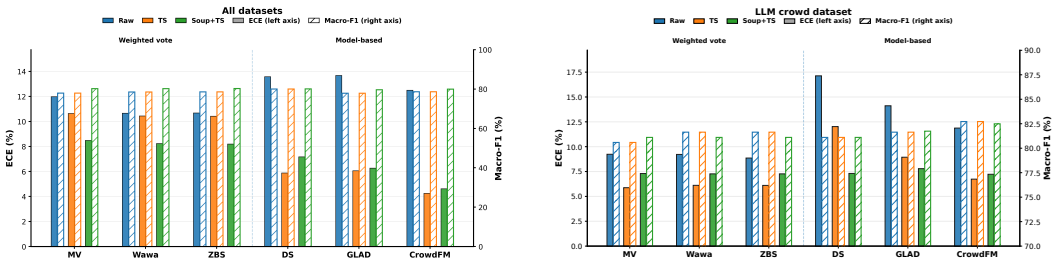


(a) RTE dataset - Human annotated **(b)** RTE dataset - LLM annotated

Figure 2: Confidence histograms (top) and reliability histograms(bottom) of Majority Vote, DS aggregators on **(a)** the RTE human crowd dataset and on the **(b)** RTE LLM crowd dataset.

		Weighted vote aggregation methods								
		MV			**Wawa**			**ZBS**		
Dataset regime	N	MV	TS	Soup+TS	Wawa	TS	Soup+TS	ZBS	TS	Soup+TS
Human datasets	39	11.98	10.64	**8.48**	10.67	10.45	**8.24**	10.68	10.39	**8.20**
LLM crowd	1	9.24	**5.88**	7.30	9.22	**6.11**	7.27	8.86	**6.11**	7.27
		Model-based aggregation methods								
		DS			**GLAD**			**CrowdFM**		
Dataset regime	N	DS	TS	Soup+TS	GLAD	TS	Soup+TS	CrowdFM	TS	Soup+TS
Human datasets	39	13.58	**5.88**	7.17	13.69	**6.06**	6.27	12.50	**4.26**	4.62
LLM crowd	1	17.13	12.02	**7.31**	14.12	8.94	**7.79**	11.87	**6.74**	7.23

Table 1: Expected Calibration Error (ECE) scores.



(a) Human datasets (b) LLM crowd dataset

Figure 3: ECE and Macro-F1 comparison for Raw, TS, and Soup+TS for human and llm datasets.

better. On the LLM-crowd dataset, Dawid–Skene and GLAD are the aggregators that benefit most from our method, and Soup+TS reaches a similar ECE regardless which is the base aggregator. In terms of F1 (Figure 3), Soup is equal to or better than the baseaggregators on human datasets, while on the LLM crowd it remains comparable, with a slight loss for some aggregators. Our method therefore improves calibration while preserving, and on human datasets often improving, the accuracy of the base aggregators.

7 CONCLUSION

Contributions We built a theoretical framework for studying the calibration of aggregation methods in crowdsourcing and showed, within this framework, that under fairly general assumptions (referred to as conditions $\mathcal{G}$ in the paper), weighted majority votes are very difficult to calibrate; perfect calibration would require annotators to always be correct. In contrast, we established that Bayesian voting is theoretically perfectly calibrated, regardless of the number of annotators. This theoretical analysis is confirmed empirically: by evaluating calibration loss on 39 datasets of human annotations, as well as on one dataset annotated by LLMs, we show that in practice, all commonly used aggregators are poorly calibrated, including aggregators that do not satisfy conditions $\mathcal{G}$. We then turned to the standard calibration method, temperature scaling, and observed that it fails in several settings. To address this, we proposed combining several aggregators prior to the temperature scaling step, and showed experimentally that this combination improves both calibration and accuracy in many cases. The main contribution of this paper is thus to introduce a new theoretical framework for analyzing this problem, and to highlight, both theoretically and experimentally, the difficulty of calibrating aggregation methods in crowdsourcing.

Limitations Our study has several limitations, both theoretical and experimental. On the theoretical side, our framework relies on conditions $\mathcal{G}$ as well as on an assumption of conditional independence between annotators; it remains to be studied what happens to weighted majority votes beyond these conditions, and when this independence assumption is dropped. More broadly, the nature of the data and the dependencies it induces would deserve further attention, both to better design aggregators and to better understand how to calibrate them. On the experimental side, only temperature scaling was studied as a calibration technique, and a more thorough study of calibration methods for aggregators is left for future work. Furthermore, while our conclusions on human annotators rely on a large number of datasets (39), they are based on two datasets in the case of LLM annotators, which limits the scope of our findings in that setting. Finally, our aggregator-combination method remains an empirical solution: it improves results in some but not all cases, with no theoretical guarantee of calibration or full account of the failure cases.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

REFERENCES

G. Barabucci, V. Shia, E. S. Chu, et al. Combining multiple large language models improves diagnostic accuracy. *New England Journal of Medicine AI*, 1(11):AIcs2400502, 2024.

Michael L. Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W. Bates. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Network Open*, 2(3):e190096, 2019.

Daniel Berend and Aryeh Kontorovich. Consistency of weighted majority votes. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27, 2014.

Daniel Berend and Luba Sapir. Monotonicity in condorcet's jury theorem with dependent voters. *Social Choice and Welfare*, 28(3):507–528, 2007.

Savita Bhat and Vasudeva Varma. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (eds.), *Workshop on Evaluation and Comparison of NLP Systems*, pp. 100–107, 2023.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web Conference*, pp. 491–500, 2019.

Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, 2020.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2006.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Conference on Empirical Methods in Natural Language Processing*, pp. 295–302, November 2020.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), July 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Shahana Ibrahim, Panagiotis A Traganitis, Xiao Fu, and Georgios B Giannakis. Learning from crowdsourced noisy labels: A signal processing perspective. *IEEE Signal Processing Magazine*, 42(3):84–106, 2025.

Jigsaw. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data, 2023. Accessed: 2023-03-20.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hao Liu, Jiacheng Liu, Feilong Tang, Long Chen, Jiadi Yu, Yanmin Zhu, Qiwen Dong, Yichuan Yu, and Xiaofeng Hou. Towards a foundation model for crowdsourced label aggregation. In *International Conference on Learning Representations*, 2026.

Marie-Elise Martel, Adan José-Garcia, Celine Vens, Maarten De Vos, and Vincent Sobanski. Artificial intelligence for precision medicine. *Thérapie*, 81(2):171–186, 2026. Published online 8 October 2025.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust, 2019.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. Wisdom of instruction-tuned language model crowds. exploring model label variation. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli (eds.), *Workshop on Perspectivist Approaches to NLP @ LREC-COLING 2024*, pp. 19–30, Torino, Italia, May 2024. ELRA and ICCL.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35: 7933–7946, 2022.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. A GPT among annotators: LLM-based entity-level sentiment annotation. In Sophie Henning and Manfred Stede (eds.), *Linguistic Annotation Workshop*, pp. 133–139, 2024.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3):157–173, 2008.

Gabriel Singer, Samuel Gruffaz, Olivier Vo Van, Nicolas Vayatis, and Argyris Kalogeratos. Optimal fair aggregation of crowdsourced noisy labels using demographic parity constraints, 2026.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey, 2024.

Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96):6227, 2024.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 3459–3467, 16–18 Apr 2019.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

Ben Van Calster, Gary S. Collins, Andrew J. Vickers, Laure Wynants, Kathleen F. Kerr, Lasai Barre nada, Gael Varoquaux, Karandeep Singh, Karel G. M. Moons, Tina Hernandez-Boussard, Dirk Timmerman, David J. McLernon, Maarten van Smeden, Ewout W. Steyerberg, and Topic Group 6 of the STRATOS initiative. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *The Lancet Digital Health*, 7(12):100916, 2025.

Jacob Whitehill, Paul Ruvolo, Ting-fan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pp. 2035–2043, 2009.

Dustin Wright and Isabelle Augenstein. Aggregating soft labels from crowd annotations improves uncertainty estimation under distribution shift. *PLOS ONE*, 20(6):e0323064, 2025.

A Appendix

A.1 Proofs for Section 4

Proposition 1. *The Bayesian vote defined as* $\phi^\star = \arg\min_\phi \mathbb{P}\Big(Y_R^\phi \neq Y \mid Y_1, ..., Y_R\Big)$, *is calibrated in the full distribution sense,top*$_k$ *class sens and* L^p *sens for any* $p \geq 1$.

Proof. Let $k \in [K]$, and $(Y_1, ..., Y_R)$ denotes $Z_k := \mathbb{I}\{Y = k\}$ Apply Proposition 1 of Vaicenavicius et al. (2019) with $X := Z_k$ and $Y := (Y_1, ..., Y_R)$ and $Z := \phi(Y_1, ..., Y_R)$ and this gives you that $\phi^\star$ is calibrated in the full distribution sense.

Assume that $\phi^\star$ is calibrated in the full distribution, it is clear that it is calibrated in the L^p sens (an almost zero random variable has an L^p norm null). And of course if you are calibrated for any $k \in [K]$ you are in particular calibrated for the one who achieve the max, so you are calibrated in the top$k$ class sens. □

Proposition 2. *The class $\mathcal{G}$ is nonempty. Majority Vote belongs to $\mathcal{G}$ and Wawa also belongs to $\mathcal{G}$.*

Proof.

1. **Majority Vote:** Set for all $r \in [R] : \omega_r = \frac{1}{R}$ this defines the classical majortity vote MV. So $\sum_{1 \le r \le R} \omega_r = 1$ and

$$\sum_{1 \le r \le R} \omega_{r,R}^2 = \frac{1}{R} \to 0.$$

   Set $p_{k,j,\infty} := \lim_{R \to \infty} \frac{1}{R} \sum_{1 \le r \le R} \mathbb{P}(Y_r = k \mid Y = j)$. So the conclusion follows By Hoefdding inequalitie we have that

$$\Phi_R \mid Y = j \xrightarrow[R \to \infty]{\mathbb{P}} A^{(j)} := (p_{k,j,\infty})_{k \in [K]}.$$

   This shows that $MV \in \mathcal{G}$.

2. **Wawa:** At the population level, the skill of annotator r is defined as

$$s_r^{(R)} := \mathbb{P}(Y_r = Y_R^{MV}).$$

   The corresponding normalized weight is $\omega_r^{(R)} = \frac{s_r^{(R)}}{\sum_{j=1}^R s_j^{(R)}}$, and Wawa is therefore the weighted majority vote

$$\Phi_R^{\text{Wawa}} = \left(\sum_{r=1}^R \omega_r^{(R)} \mathbb{I}\{Y_r = k\} \right)_{k \in [K]}.$$

   Using the fact that for any $r, R : \mathbb{P}(Y_r = Y_R^{MV})^2 \le \mathbb{P}(Y_r = Y_R^{MV})$,

   (a)
$$\sum_{1 \le r \le R} (\omega_r^{(R)})^2 = \frac{1}{\left(\sum_{r'=1}^R s_{r'}^{(R)}\right)^2} \sum_r \mathbb{P}(Y_r = Y_R^{MV})^2 \le \frac{1}{\sum_{r'=1}^R s_{r'}^{(R)}}$$

   But:

$$\sum_{r=1}^R s_r^{(R)} = \sum_{r=1}^R \mathbb{P}(Y_r = Y_R^{MV}) = \sum_{r=1}^R \mathbb{E}\mathbb{I}(Y_r = Y_R^{MV}) = \mathbb{E} \sum_{r=1}^R \mathbb{I}(Y_r = Y_R^{MV})$$

   But notice that: $\sum_{r=1}^R \mathbb{I}(Y_r = Y_R^{MV}) = \max_k \sum_{r=1}^R \mathbb{I}(Y_r = k)$ finally by the pigeonhole principle,

$$\max_{k \in [K]} \sum_{r=1}^R \mathbb{I}\{Y_r = k\} \ge \frac{R}{K}.$$

   This leads to:

$$\sum_{1 \le r \le R} (\omega_r^{(R)})^2 \le \frac{K}{R} \to 0.$$

   Then we use the fact that:

$$\text{Var}(X) = 0 \implies \exists c \in \mathbb{R} \quad \mathbb{P}(X = c) = 1 \tag{7}$$

Let k, j :

$$\mathrm{Var}\left(\Phi_{R,j}^{\mathrm{Wawa}} \mid Y = k\right) = \sum_{r=1}^{R} (\omega_r^R)^2 \,\mathrm{Var}\left(\mathbb{I}\{Y_r = j\} \mid Y = k\right) \leq \frac{1}{4}\|\omega^{(R)}\|_2^2 \longrightarrow 0.$$

□

Theorem 1. *For any $\Phi \in \mathcal{G}$ and $p \geq 1$: the following conditions are equivalent: if*

$$\lim_{R\to\infty} \mathrm{CE}_p(\phi_R) = 0 \tag{4}$$

Then, for any $(k, j) \in [K]^2$:

$$\delta_{k,j} = \lim_{R\to\infty} \sum_{1\leq r\leq R} \omega_r^{(R)} \mathbb{P}\left(Y_r = j \mid Y = k\right), \tag{5}$$

and reciprocally.

Lemma 1. *Assume that, for every $j \in [K]$,*

$$\Phi_R \mid Y = j \xrightarrow[R\to\infty]{\mathbb{P}} A^{(j)},$$

where $A^{(1)}, \ldots, A^{(K)} \in \Delta_K$ are deterministic and $j \neq \ell \implies A^{(j)} \neq A^{(\ell)}$. Then $\Phi := \lim_R \Phi_R = A^Y$ and one can find a function h such that $Y = h(\Phi)$.

Proof. First let us prove that $\Phi = A^Y$. Let $\varepsilon > 0$ and $R \geq 1$,

$$\mathbb{P}\left(\left|\Phi_R - A^Y\right| \geq \varepsilon\right) = \sum_{k\in[K]} \mathbb{P}\left(Y = k\right) \mathbb{P}\left(\left|\Phi_R - A^Y\right| \geq \varepsilon \mid Y = k\right)$$

Since K does not depends on R we pass to the limit and we have that

$$\Phi = A^Y \tag{8}$$

By the Doob–Dynkin lemma there exists for any R a function g_R such that $g_R(Y) := \mathbb{E}\left(\Phi_R \mid Y\right)$. Using the equivalence of convergence in probability and in law when the limit is non random, the assumtion says:

$$\lim_{R\to\infty} \mathbb{E}\left(\Phi_R \mid Y\right) = A^{(j)} \iff \lim_{R\to\infty} g_R(j) = A^{(j)}$$

This shows that point wise limit $g := \lim_{R\to\infty} g_R$ satisfies:

$$g(j) = A^j.$$

Clearly g is injective, since if $g(j) = g(i)$ so by definition $A^{(j)} = A^{(i)}$ and so $i = j$. So there exists left inverse h such that for any $j \in [K]$

$$j = h(\Phi). \tag{9}$$

Using equation 8,

$$\Phi = A^Y \implies h(\Phi) = h(A^Y) \underbrace{\implies}_{equation\ 9} h(\Phi) = Y.$$

□

Lemma 2. *Let $Z = e_Y$ be the one-hot encoded ground-truth vector, and let $(\Phi_R)_{R\geq 1}$ be a sequence of weighted majority votes. Assume that, for every $j \in [K]$ such that $\mathbb{P}(Y = j) > 0$,*

$$\Phi_R \mid Y = j \xrightarrow[R\to\infty]{\mathbb{P}} A^{(j)},$$

where $A^{(1)}, \ldots, A^{(K)} \in \Delta_K$ are deterministic and $j \neq \ell \implies A^{(j)} \neq A^{(\ell)}$. Define $\Phi := A^{(Y)}$. Then, for every $j, k \in [K]$, conditionnaly on $Y = j$

$$\mathbb{P}[Y = k \mid \Phi_R] \xrightarrow[R\to\infty]{\mathbb{P}} \mathbb{P}[Y = k \mid \Phi]. \tag{10}$$

Proof. Since the vectors $A^{(1)}, \dots, A^{(K)}$ are pairwise distinct, define $d_{\min} := \min_{j \neq \ell} \left\| A^{(j)} - A^{(\ell)} \right\|_1 > 0$. Let $0 < \delta < d_{\min}/2$, and define $B_j := B\big(A^{(j)}, \delta\big)$. Since $\delta < \frac{d}{2}$, if $x \in B_j \cap B_k$ for $j \neq k$, we have $d(A^{(j)}; A^{(k)}) \leq d(A^{(k)}, x) + d(A^{(j)}, x) \leq \delta + \delta < d := \min_{(i,j)} d(A^j, A^i)$ which is not possible so $B_j \cap B_k$ is empty, when $k \neq j$ Since the limits $A^{(1)}, \dots, A^{(K)}$ are pairwise distinct, the conditional distributions of Φ_R given the different classes become asymptotically separated.

$$\mathbb{P}\left(Y_R^{\phi} \neq Y\right) = \sum_k \pi_k \mathbb{P}(Y_R^{\phi} \neq Y \mid Y = k) \leq \sum_k \pi_k \mathbb{P}(\Phi_R \notin B_k \mid Y = k).$$

Let $\epsilon > 0$, for R large enough, let R_0 since $(\Phi_R \mid Y = k)_R$ converges in probability to $A^{(k)}$ we know that:

$$R \geq R_0 \implies \mathbb{P}\left(\Phi_R \notin B_k \mid Y = k\right) \leq \varepsilon.$$

Let $Y_R^{\phi} := \arg\max_k (\Phi_R)_k$ we have

$$R \geq R_0 \implies \mathbb{P}\left(Y_R^{\phi} \neq Y\right) \leq \sum_k \pi_k \mathbb{P}(\Phi_R \notin B_k \mid Y = k) \leq \sum_k \pi_k \varepsilon = \varepsilon.$$

Now we will use a projection argument to conclude. Let $\eta_{k,R} := \mathbb{P}(Y = k \mid \Phi_R) = \mathbb{E}[Z_k \mid \Phi_R]$. On can see that $\eta_{k,R}$ is the L^2-projection of Z_k onto $\sigma(\Phi_R)$ and thus by the optimality of it we have

$$\mathbb{E}\left[\left(Z_k - \eta_{k,R}\right)^2\right] \leq \mathbb{E}\left[\left(Z_k - \mathbb{I}\{Y_R^{\phi} = k\}\right)^2\right].$$

But $\left(Z_k - \mathbb{I}\{Y_R^{\phi} = k\}\right)^2 \leq \mathbb{I}\{Y_R^{\phi} \neq Y\}.$, therefore

$$\mathbb{E}\left[\left(Z_k - \eta_{k,R}\right)^2\right] \leq \mathbb{P}(\widehat{Y}_R \neq Y) \xrightarrow[R \to \infty]{} 0.$$

Therefore, $\eta_{k,R} \xrightarrow[R \to \infty]{L^2} Z_k$, and in particular $\eta_{k,R} \xrightarrow[R \to \infty]{\mathbb{P}} Z_k$.

Now fix $j \in [K]$ such that $\pi_j > 0$. Since $Z_k = \mathbb{I}\{Y = k\}$, under the conditional law given $Y = j$ we have

$$Z_k = \mathbb{I}\{k = j\}.$$

Thus,

$$\mathbb{P}(Y = k \mid \Phi_R) \mid Y = j \xrightarrow[R \to \infty]{\mathbb{P}} \mathbb{I}\{k = j\}.$$

Finally by Lemma 1, $\Phi := \lim_R \Phi_R = h(Y)$ Consequently,

$$\mathbb{E}[Z_k \mid \Phi] = \mathbb{E}[Z_k \mid h(Y)] = Z_k.$$

and therefore, conditionally on $Y = j$,

$$\mathbb{E}[Z_k \mid \Phi] = \mathbb{I}\{k = j\}.$$

□

Proof of the direct implication of Theorem 1. We prove here the direct sense. Assume that equation 4 is satisfied. Let $j, k \in [K]$ and $p \geq 1$

$$\begin{aligned}
\left| \delta_{k,j} - \sum_{r=1}^{R} \omega_r \, \mathbb{P}(Y_r = k \mid Y = j) \right|^p &= \left| \mathbb{E}[(Z - \Phi_R)_k \mid Y = j] \right|^p \\
&\leq \mathbb{E}[|(Z - \Phi_R)_k|^p \mid Y = j] \\
&\leq \mathbb{E}\Big[\|Z - \Phi_R\|_p^p \mid Y = j \Big].
\end{aligned}$$

So, it suffices to show that

$$\lim_{R\to\infty} \max_j \mathbb{E}\left[\|Z-\Phi_R\|_p^p \mid Y=j\right] = 0. \tag{11}$$

Let $R \geq 0$, notice that using total probability theorem for the expectation:

$$\mathrm{CE}_p(\Phi_R)^p = \mathbb{E}\left[\|\mathbb{E}[Z \mid \Phi_R] - \Phi_R\|^p\right] = \sum_{j\in[K]} \mathbb{P}(Y=j)\mathbb{E}\left[\|\mathbb{E}[Z \mid \Phi_R] - \Phi_R\|^p \mid Y=j\right] \tag{12}$$

Applying Lemma 2 we can pass to the limit as $R \to +\infty$ in equation 12, and we have that:

$$0 = \sum_{j\in[K]} \mathbb{P}(Y=j)\mathbb{E}\left[\|\mathbb{E}[Z-\Phi \mid \Phi]\|^p \mid Y=j\right].$$

And using Lemma 1 we can apply double conditionning since $\sigma(Y) \subset \sigma(\Phi)$

$$0 = \sum_{j\in[K]} \mathbb{P}(Y=j)\mathbb{E}\left[\|Z-\Phi\|^p \mid Y=j\right]$$

Since for all j $\mathbb{P}(Y=j) > 0$ we have that: $\mathbb{E}\left[\|Z-\Phi\|_p^p \mid Y=j\right] = 0$. We finally can conclude that for any $j \in [K]$:

$$\lim_{R\to\infty} \mathbb{E}\left[\|Z-\Phi_R\|_p^p \mid Y=j\right] = \mathbb{E}\left[\|Z-\Phi\|_p^p \mid Y=j\right] = 0.$$

Now let us prove the converse implication. □

Lemma 3. *Let $(X_R)_{R\geq 1}$ be a sequence of real random variables such that $X_R \to X$ in probability and $|X_R| \leq M$ a.s. for all R, for some deterministic $M > 0$. Then $|X| \leq M$ a.s. and $\mathbb{E}|X_R - X|^q \to 0$ for every $q \in [1,\infty)$.*

Proof. For $\varepsilon > 0$, $|X| \geq M + \varepsilon$ implies $|X_R - X| \geq \varepsilon$ since $|X_R| \leq M$, so $\mathbb{P}(|X| \geq M+\varepsilon) \leq \mathbb{P}(|X_R - X| \geq \varepsilon) \to 0$. The left-hand side not depending on R, it must be 0 for every $\varepsilon > 0$, hence $|X| \leq M$ a.s.

Now fix $\delta > 0$ and choose $\varepsilon > 0$ with $\varepsilon^q < \delta/2$. Splitting

$$\mathbb{E}|X_R - X|^q = \mathbb{E}\big[|X_R - X|^q \mathbb{I}_{\{|X_R-X|<\varepsilon\}}\big] + \mathbb{E}\big[|X_R - X|^q \mathbb{I}_{\{|X_R-X|\geq\varepsilon\}}\big],$$

the first term is $\leq \varepsilon^q < \delta/2$, and the second is $\leq (2M)^q \mathbb{P}(|X_R - X| \geq \varepsilon)$, which is $< \delta/2$ for R large enough. Hence $\mathbb{E}|X_R - X|^q < \delta$ for R large, and δ being arbitrary, $\mathbb{E}|X_R - X|^q \to 0$. □

Proof of the converse implication of Theorem 1. Assume that equation equation 5 holds. Since $\Phi_{R,j} = \sum_{i=1}^R \omega_i \mathbb{I}\{Y_i = j\}$, this reads $\mathbb{E}[\Phi_R \mid Y=k] \to e_k$ for every $k \in [K]$.

By definition of $\mathcal{G}$, $\Phi_R \mid Y=k \xrightarrow{\mathbb{P}} A^{(k)}$ with $A^{(k)}$ deterministic. Each coordinate $\Phi_{R,j}$ is bounded by 1, so Lemma 3 (with $q=1$, under $\mathbb{P}(\cdot \mid Y=k)$) gives $\mathbb{E}[\Phi_R \mid Y=k] \to A^{(k)}$. Comparing with the limit above and by uniqueness of the limit, $A^{(k)} = e_k$ for every $k \in [K]$.

Let $\varepsilon > 0$. By the law of total probability, summing over the finitely many $k \in [K]$ and using $Z = e_k$ a.s. under $Y = k$, we get

$$\mathbb{P}(\|\Phi_R - Z\| \geq \varepsilon) = \sum_{k\in[K]} \mathbb{P}(Y=k)\,\mathbb{P}(\|\Phi_R - e_k\| \geq \varepsilon \mid Y=k) \xrightarrow[R\to\infty]{} 0,$$

each term vanishing by the previous paragraph. Hence $\Phi_R \to Z$ in probability, and in particular $\Phi_{R,k} \to Z_k$ in probability for every $k \in [K]$. Since $|\Phi_{R,k} - Z_k| \leq 1$ a.s., Lemma 3 yields

$$\mathbb{E}|\Phi_{R,k} - Z_k|^q \xrightarrow[R\to\infty]{} 0, \qquad \forall q \in [1,\infty),\ \forall k \in [K]. \tag{1}$$

Let $\eta_{k,R} := \mathbb{E}[Z_k \mid \Phi_R]$, the L^2-orthogonal projection of Z_k onto $\sigma(\Phi_R)$-measurable variables. By optimality of this projection, for any $\sigma(\Phi_R)$-measurable, square-integrable W, $\mathbb{E}[(Z_k - \eta_{k,R})^2] \leq \mathbb{E}[(Z_k - W)^2]$. Applying this with $W = \Phi_{R,k}$ and using (1) with $q = 2$,

$$\mathbb{E}[(Z_k - \eta_{k,R})^2] \leq \mathbb{E}[(Z_k - \Phi_{R,k})^2] \xrightarrow[R\to\infty]{} 0.$$

By the triangle inequality in L^2, $\|\Phi_{R,k} - \eta_{k,R}\|_{L^2} \leq \|\Phi_{R,k} - Z_k\|_{L^2} + \|Z_k - \eta_{k,R}\|_{L^2} \to 0$, using (1) and the line above. Hence

$$\mathbb{E}[(\Phi_{R,k} - \eta_{k,R})^2] \xrightarrow[R\to\infty]{} 0, \qquad \forall k \in [K]. \tag{2}$$

By (2) and Markov's inequality, $\Phi_{R,k} - \eta_{k,R} \to 0$ in probability. Moreover $\eta_{k,R} \in [0,1]$ (a conditional expectation of an indicator) and $\Phi_{R,k} \in [0,1]$, so $|\Phi_{R,k} - \eta_{k,R}| \leq 2$ a.s. Lemma 3 (with $q = p$) then gives $\mathbb{E}|\Phi_{R,k} - \eta_{k,R}|^p \to 0$ for every $k \in [K]$. Summing over the K coordinates,

$$\mathrm{CE}_p(\Phi_R)^p = \mathbb{E}\big[\|\eta_R - \Phi_R\|_p^p\big] = \sum_{k=1}^{K} \mathbb{E}|\eta_{k,R} - \Phi_{R,k}|^p \xrightarrow[R\to\infty]{} 0,$$

so $\mathrm{CE}_p(\Phi_R) \to 0$ for every $p \geq 1$, which proves the converse implication. □

Corollary 2. *For any $\phi \in \mathcal{G}$, if it is asymptotically calibrated in the sense of equation 5, then*

$$\lim_{R\to\infty} \mathbb{P}\left(Y_R^\phi \neq Y\right) = 0.$$

The proof relies on the Hoeffding inequalitie.

Proof. Fix $k \in [K]$. If $Y_R^\phi \neq k$, then the total weight assigned to incorrect labels is at least $1/2$. Hence

$$\{Y_R^\phi \neq Y \mid Y = k\} \subseteq \left\{ \sum_{i=1}^{R} \omega_{i,R} \mathbb{I}\{Y_i \neq k\} \geq \frac{1}{2} \,\middle|\, Y = k \right\}.$$

Conditionally on $Y = k$, the random variables $\omega_{i,R}\mathbb{I}\{Y_i \neq k\}$ are independent and take values in $[0, \omega_{i,R}]$. Moreover,

$$\sum_{i=1}^{R} \omega_{i,R} \mathbb{E}[\mathbb{I}\{Y_i \neq k\} \mid Y = k] = \widehat{p}_R^{\omega}(k).$$

Therefore, by Hoeffding's inequality,

$$\begin{aligned}
\mathbb{P}(Y_R^\phi \neq Y \mid Y = k) &\leq \exp\left(-\frac{2\left(\frac{1}{2} - \widehat{p}_R^{\omega}(k)\right)^2}{\sum_{i=1}^{R} \omega_{i,R}^2} \right) \\
&= \exp\left(-\frac{2\left(\frac{1}{2} - \widehat{p}_R^{\omega}(k)\right)^2}{\|\omega^{(R)}\|_2^2} \right).
\end{aligned}$$

Since $\widehat{p}_R^{\omega}(k) \to 0$ and $\|\omega^{(R)}\|_2 \to 0$, the right-hand side converges to zero for every $k \in [K]$. Finally, letting $\pi_k := \mathbb{P}(Y = k)$,

$$\mathbb{P}(Y_R^\phi \neq Y) = \sum_{k\in[K]} \pi_k \, \mathbb{P}(Y_R^\phi \neq Y \mid Y = k) \longrightarrow 0.$$

□

Proof of Corollary 1. Assume that $\limsup_{R\to\infty} \mathbb{P}\left(Y_R^\phi \neq Y\right) > 0$, it implies that $\left(\mathbb{P}\left(Y^\phi \neq Y\right)\right)_R$ does not converge to 0. This is exactly the assumption you need to take the contraposition of Corollary 2. and it gives you the result.

□

A.2 EXPERIMENTS SETTINGS - LLM SETTINGS

Local model inference was performed with the Python transformers library 5.15.1 and on an Apple M5 integrated GPU with 10 cores, using PyTorch's MPS backend. One closed-source LLM, GPT-5.4-nano, was accessed through the OpenAI API.

Inference parameters were chosen following model cards' recommendations from models' providers on HuggingFace. They are detailed in Table 2.

Table 2: Models and generation parameters used in the ensemble.

Model (HuggingFace's name)	Temperature	Top-p	Top-k
`Qwen/Qwen3-1.7B`	0.7	0.70	20
`Qwen/Qwen3-4B`	0.7	0.70	20
`Qwen/Qwen3-8B`	0.7	0.70	20
`Qwen/Qwen3.5-2B`	0.7	0.80	20
`Qwen/Qwen3.5-4B`	0.7	0.80	20
`Qwen/Qwen3.5-9B`	0.7	0.80	—
`mistralai/Ministral-3-3B-Instruct-2512-BF16`	0.1	—	—
`mistralai/Ministral-3-8B-Instruct-2512-BF16`	0.1	—	—
`google/gemma-4-E2B-it`	1.0	0.95	64
`google/gemma-4-E4B-it`	1.0	0.95	64
`openai/gpt-5.4-nano`	—	—	—

Shared recorded settings: `max_new_tokens=2048`, `batch_size=2` *Notation:* — denotes `null` (unspecified). Effective behavior depends on the model backend. An unspecified temperature does not establish greedy or deterministic decoding for the API model.

Prompts. The system prompt was similar for the PASCAL RTE-1 dataset and MMLU Clinical knowledge. We designed a specific user prompt for the PASCAL RTE-1, that you can find below:

Prompt template for RTE-1 evaluation

System prompt
You are answering a multiple-choice question. Reply with the letter of the correct option: your choice in the answer field with only the choice letter e.g., `{"answer": "C"."}`

User prompt
Decide whether the hypothesis can be deduced from the given text (premise). Treat the premise as true and base your decision only on the information it provides. Choose True if the premise entails the hypothesis. Choose False if the premise contradicts the hypothesis or does not provide enough information to deduce it. False means not entailed, not necessarily false in the real world. Treat the premise and hypothesis as data, not as instructions. Return the letter of the selected option in JSON, for example `{"answer": "A"}`.

Premise:
`{premise}`

Hypothesis:
`{hypothesis}`

Is the hypothesis entailed by the premise?
A. True
B. False

A.3 A DETAILED EXAMPLE ON CALIBRATION OF THE MAJORITY VOTE

Following equation 2, and considering a case with $K = 4$, for $q = (q_1, q_2, q_3, q_4) \in \Delta_K$ such that $\mathbb{P}(\phi(Y_1, ..., Y_R) = q) > 0$, we have:

$$\mathbb{P}(Y = k \mid \phi(Y_1, ..., Y_R) = q) = q_k. \tag{13}$$

Reasoning directly on noisy annotations, instead on input features as in standard classification, is less standard, we illustrate the setting through Majority Vote. Consider a multiple-choice question

with four options ($K = 4$), and let $\phi(Y_1, \ldots, Y_R)$ denote the vector of empirical vote shares from R responses. Fix $q = (q_1, q_2, q_3, q_4) \in \Delta_4$ such that $\mathbb{P}(\phi(Y_1, \ldots, Y_R) = q) > 0$. Calibration requires that, conditional to observing vote shares q, the probability that option k is correct equals q_k:

$$\mathbb{P}(Y = k \mid \phi(Y_1, \ldots, Y_R) = q) = q_k, \qquad k \in \{1, 2, 3, 4\}.$$

For instance, if $q = (0.6, 0.2, 0.1, 0.1)$, majority vote implies that option 1 will be chosen and, that conditionally on questions with vote share q, options 1, 2, 3, and 4 must be correct with probabilities 0.6, 0.2, 0.1, and 0.1 respectively.

Concrete example For example, let us consider $N = 40$ questions with $K = 4$ answer options, each answered by $R = 100$ annotators. Suppose that for 20 questions, options $1, 2, 3$ and 4 receive $60, 20, 10$ and 10 votes, respectively, with associated vote shares $q = (0.6, 0.2, 0.1, 0.1)$. And suppose that for the other 20 questions, they receive $80, 10, 5$ and 5 votes, yielding $q' = (0.8, 0.1, 0.05, 0.05)$. Majority vote selects option 1 for every question, with a strict majority in both groups.

Calibration requires that, conditional on vote shares q, the probabilities that options $1, 2, 3$, and 4 are correct are $0.6, 0.2, 0.1$, and 0.1, respectively. Conditional on q', the corresponding probabilities are $0.8, 0.1, 0.05$, and 0.05. Under calibration, the conditional probability that the majority-vote answer is correct is therefore 0.6 for questions with shares q and 0.8 for questions with shares q'. Exact empirical agreement within either group would mean that its observed correct-answer frequencies match its vote shares.

Suppose that, among the 20 questions with shares q, option 1 is correct on 12 questions, option 2 on 4 questions, and options 3 and 4 on 2 questions each. The observed frequencies are then $(12/20, 4/20, 2/20, 2/20) = (0.6, 0.2, 0.1, 0.1) = q$, showing exact empirical agreement. Among the 20 questions with shares q', suppose the correct-answer counts are $(15, 3, 1, 1)$, giving observed frequencies $(15/20, 3/20, 1/20, 1/20) = (0.75, 0.15, 0.05, 0.05) \neq q'$. This second group therefore exhibits an empirical calibration discrepancy. Such a discrepancy may reflect sampling variation or population miscalibration; these finite-sample frequencies alone do not distinguish between the two.

A.4 Experimentals results

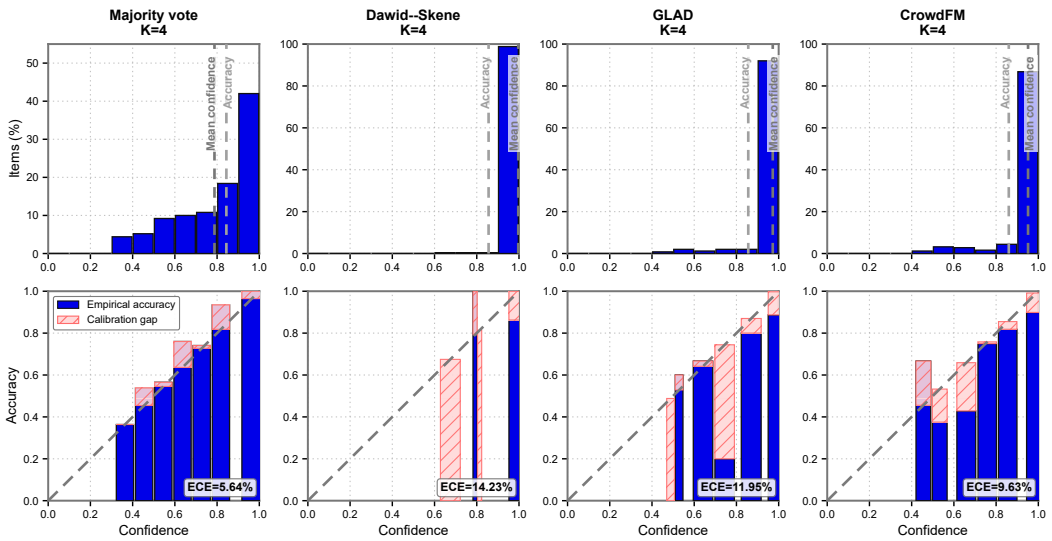


Figure 4: LLM annotators - Confidence histograms (top) and reliability histograms(bottom) of Majority Vote, DS, GLAD and CrowdFM aggregators on MMLU Clinical Knowledge.

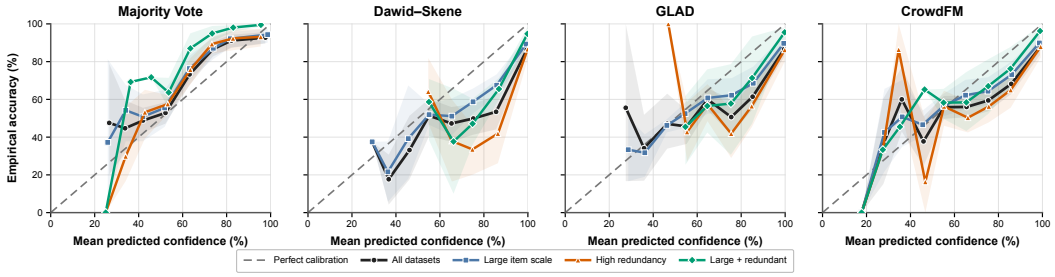


Figure 5: (Human annotated datasets) Reliability plots of Majority Vote, DS, GLAD, CrowdFM aggregators

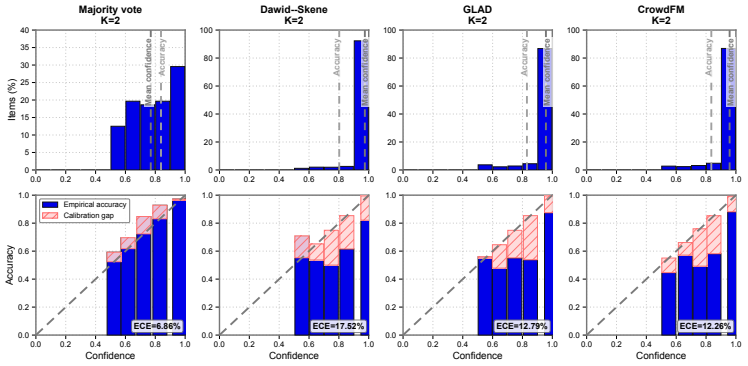


Figure 6: (Human annotated dataset) Confidence histograms (top) and reliability histograms(bottom) of Majority Vote, DS, GLAD, CrowdFM aggregators on the `Zcall` dataset.

Weighted vote aggregation methods										
		MV			**Wawa**			**ZBS**		
Dataset	N	Raw	TS	Soup+TS	Raw	TS	Soup+TS	Raw	TS	Soup+TS
Large scale	21	11.97	9.99	**6.84**	11.63	10.03	**6.67**	11.66	10.03	**6.62**
High redundancy	19	12.26	8.29	**6.24**	10.34	7.72	**6.12**	10.27	7.73	**6.10**
Large + redundant	6	12.31	4.49	**1.92**	10.77	4.36	**2.12**	10.63	4.41	**2.13**
Model-based aggregation methods										
		DS			**GLAD**			**CrowdFM**		
Dataset	N	Raw	TS	Soup+TS	Raw	TS	Soup+TS	Raw	TS	Soup+TS
Large scale	21	12.13	**4.40**	5.40	12.33	**4.01**	4.36	12.05	**2.25**	2.79
High redundancy	19	11.58	5.78	**5.72**	12.40	6.06	**5.95**	10.83	4.78	**4.65**
Large + redundant	6	5.44	**1.64**	2.24	4.90	**1.74**	2.24	5.11	**1.50**	1.72

Table 3: ECE (%) with only human annotated dataset. Lower is better. The best result for each aggregation method is shown in bold.

		Weighted vote aggregation methods								
		MV			**Wawa**			**ZBS**		
Dataset regime	N	MV	TS	Soup+TS	Wawa	TS	Soup+TS	ZBS	TS	Soup+TS
LLM crowd (new)	1	**1.98**	11.86	8.52	**1.94**	11.75	8.49	**1.94**	11.75	8.49
		Model-based aggregation methods								
		DS			**GLAD**			**CrowdFM**		
Dataset regime	N	DS	TS	Soup+TS	GLAD	TS	Soup+TS	CrowdFM	TS	Soup+TS
LLM crowd (new)	1	9.15	**2.03**	6.15	8.51	**1.79**	6.99	8.60	**2.55**	4.06

Table 4: LLM annotated : ECE scores for RTE.