NSS

Neutrosophic Sets and Systems, Vol. 100, 2026

University of New Mexico

TINF-Consensus for Multi-Annotator Learning: Separating Neutral Responses from Polarized Disagreement

Mukhtar Ahmad[1,*]

[1]Doctoral Program of Mathematics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jalan Ganesha No. 10, Bandung 40132, Indonesia; Email: mukhkhu0786@gmail.com; 30125701@mahasiswa.itb.ac.id; ORCID: 0009-0006-4597-2278

*Correspondence: mukhkhu0786@gmail.com; Mobile: +62 822-4677-7260

Abstract. Aggregating multiple annotations into one label usually compresses two different phenomena into the same low-confidence outcome: explicit neutrality, where annotators decline to support either class, and polarized disagreement, where decisive annotators support opposite classes. The distinction matters operationally. Neutral items may need clarification or additional evidence, whereas polarized items may require adjudication, subgroup analysis, or retention of plural labels. We introduce TINF-Consensus, a reliability-weighted binary label-fusion framework built from the truth, indeterminacy, neutrality, falsehood (TINF) representation. For each item, weighted positive, neutral, and negative response masses define truth T, neutrality N, and falsehood F. Pure indeterminacy is then $I = 4TF/(T + F)$, which measures polarization only among decisive responses. A derived consensus coordinate $C = (T - F)^2/(T + F)$ yields the exact decomposition $C + I + N = 1$. Thus, all-neutral and evenly split decisive crowds—which have the same signed mean—receive different profiles. Gold questions provide Beta-smoothed annotator orientation and reliability weights. We establish permutation invariance, an adversarial mass bound, consistency of gold-based orientation, and an exponential conditional error bound.

Deterministic semi-synthetic experiments use 21 annotators, 40 repetitions, and four public binary datasets. The contrast $I - N$ distinguishes constructed polarization from explicit neutrality with macro-average AUROC 0.99999, and the rule $I > N$ routes the reason for review with 0.9928 accuracy. With a nominal 40% adversarial setting (eight of 21 annotators), gold-weighted TINF-Consensus retains 0.9942 macro-average accuracy, compared with 0.5930 for majority vote and 0.9949 for a gold-anchored Dawid–Skene model. Importantly, weighted margin deficit, not indeterminacy, is best for predicting aggregation errors (AUROC 0.9552). The results support a two-output design: margin answers *whether the fused label is fragile*, whereas TINF answers *why the annotation process is unresolved*. The study is a controlled methodological validation; real multi-rater and domain-specific evaluation remains necessary.

Keywords: multi-annotator learning; crowdsourcing; neutrosophic logic; label aggregation; annotator disagreement; abstention; human label variation; robust consensus.

1. Introduction

Supervised learning pipelines commonly treat a dataset label as a primitive fact. In many datasets, however, that label is the endpoint of a social and statistical process: several people inspect an item, some select a class, some abstain, and others disagree. Majority vote reduces this process to one bit. Probabilistic aggregation

models improve the estimate of a latent class by learning annotator error rates [4, 12], item difficulty [19], or joint worker–label models [21, 8]. Yet a scalar class posterior can still obscure the mechanism that produced uncertainty.

Consider two panels of equal size. In the first, every panelist chooses an explicit neutral response. In the second, half choose positive and half choose negative. Both panels have signed mean zero; both may produce a 0.5 class probability under a symmetric model. They should not automatically trigger the same intervention. The first may indicate insufficient evidence, an ill-posed question, or an appropriate middle position. The second may indicate competing interpretations, population heterogeneity, ambiguous guidelines, or a genuinely contested construct. Conflating them can waste review budgets and, in subjective tasks, erase minority perspectives [16, 11, 3].

This paper develops a narrow operational specialization of the TINF representation-truth, indeterminacy, neutrality, and falsehood [15]. TINF refines a truth assessment into support for truth, pure indeterminacy, neutrality, and falsehood. Its explicit separation of indeterminacy from neutrality is particularly well matched to multi-annotator data, provided that the mapping from responses to coordinates is defined rather than asserted. Our construction makes this mapping observable: positive, neutral, and negative response masses determine T, N, F, while I measures the balance of the two decisive masses.

The resulting method, TINF-Consensus, is designed around two different outputs:

(i) a reliability-weighted signed margin for label fusion and error screening; and

(ii) a TINF profile for explaining whether unresolved items are dominated by explicit neutrality or decisive conflict.

This separation is deliberate. A disputed item is not necessarily mislabelled, and a confident consensus is not necessarily correct. We therefore resist the tempting claim that one uncertainty score should optimize both error detection and disagreement diagnosis.

Contributions. The paper makes five contributions.

(1) We define a reliability-weighted response profile (T, I, N, F) for binary multi-annotator data in which a neutral annotation is a substantive response and missingness remains separate.

(2) We derive the exact identity $C + I + N = 1$, where C is a consensus coordinate, and prove extreme-case, invariance, and distinguishability properties.

(3) We give a gold-anchored fusion rule that can identify and invert consistently adversarial annotators, together with finite-sample orientation and label-error bounds.

(4) We introduce a two-stage review policy: weighted margin deficit ranks likely label errors, while I versus N routes items to conflict-aware or neutrality-aware review.

(5) We provide a deterministic, executable benchmark with four datasets, five adversarial crowd conditions, 40 repetitions, complete CSV outputs, and a comparison with majority vote and gold-anchored Dawid–Skene.

The experiments are semi-synthetic by design. Ground truth and response mechanisms are known, allowing the diagnostic claims to be tested exactly. They do not replace evaluation on naturally collected annotations.

This boundary is important: the paper establishes a mathematically coherent representation and controlled evidence for its behavior, not universal validity in every annotation domain.

2. Related Work

2.1. *Label aggregation and annotator reliability*

Dawid and Skene's latent-class model estimates annotator-specific confusion matrices by expectation–maximization [4]. Later work jointly learned classifiers and annotator expertise [12], modeled both annotator competence and item difficulty [19], and developed statistically efficient spectral initialization and EM procedures [21]. Bayesian comparisons show that modeling choices and priors matter in realistic annotation settings [8]. Deep models can incorporate annotator-specific noise layers and learn an end-to-end classifier from crowds [13].

Reliability weighting is especially relevant when some workers are careless or adversarial. Methods for detecting unreliable and adversarial workers often exploit latent structure across tasks and workers [7]. TINF-Consensus occupies a simpler operating point. It uses a small set of known gold items to estimate the sign and magnitude of each annotator's decisive accuracy, then retains an interpretable item-level decomposition. It is therefore not a replacement for a fully generative crowd model. Rather, it adds a diagnostic representation that standard latent-truth models do not natively expose.

2.2. *Disagreement as signal*

Research on human label variation argues that disagreement is not always annotation noise. Surveys distinguish annotator effects, item ambiguity, task design, and legitimate variation [16]. In natural-language processing, preserving soft labels or annotator distributions can improve learning and analysis [6, 18], perspectivist work warns that majority labels can suppress valid minority judgments [5]. Subjective annotation studies have likewise advocated looking beyond majority vote [3], and Plank [11] frames variation as a property to be modeled rather than merely eliminated. Human-provided class distributions can also improve robustness relative to one-hot labels [10].

Our focus is complementary. We do not infer the sociological cause of a disagreement from votes alone. We preserve a narrower, directly observable distinction between (a) explicit middle or non-committal responses and (b) polarization among decisive responses. Domain interpretation still requires metadata, qualitative investigation, and appropriate governance.

2.3. *From fuzzy and neutrosophic representations to TINF*

Fuzzy sets assign a degree of membership [20]; intuitionistic fuzzy sets add non-membership and hesitation [2]. Neutrosophic representations distinguish truth, indeterminacy, and falsehood and allow their degrees to vary without the normalization imposed by an ordinary probability simplex [14, 17]. The TINF refinement

further separates neutrality from pure indeterminacy [15]. This distinction is the conceptual starting point of our work.

Our contribution is not another free-form semantic score. We define every coordinate from observable response masses, state the assumptions under which those masses are reliability-weighted, and identify the restricted identity satisfied by the resulting operational profile. We found no prior multi-annotator construction with this particular decomposition; the novelty claim is intentionally limited to the construction and analysis given here, not to the broad use of fuzzy or neutrosophic ideas in decision systems.

3. Problem Setting

Let items be indexed by $i \in \{1, \dots, m\}$ and annotators by $j \in \{1, \dots, n\}$. For a binary question, an observed response is

$$R_{ij} \in \{-1, 0, +1\}, \tag{1}$$

where $+1$ supports the candidate-positive label, -1 supports its alternative, and 0 is an *explicit neutral response*. A response that was never collected is denoted by $\perp$ and is excluded from all normalizations. This separation is essential: coding missing values as zero would manufacture neutrality.

Each annotator has a positive weight w_j. If gold questions are available, we estimate these weights as described in Section 5; otherwise equal or externally specified weights can be used. Let

$$\mathcal{A}_i = \{j : R_{ij} \neq \perp\}, \qquad W_i = \sum_{j \in \mathcal{A}_i} w_j.$$

An item with $W_i = 0$ has no annotation evidence and is deferred. This case is different from an item for which all observed responses are explicitly neutral.

We address three research questions:

RQ1:
Can an observable TINF profile distinguish explicit neutrality from decisive polarization when both have zero signed margin?

RQ2:
Can simple gold-based orientation and reliability weighting preserve label accuracy in the presence of consistently adversarial annotators?

RQ3:
Should the same scalar be used both to detect likely label errors and to diagnose the reason an item needs review?

4. A TINF Profile for Crowd Responses

4.1. *Response masses*

Suppose for the moment that responses have already been corrected for an annotator's estimated orientation, giving $X_{ij} \in \{-1, 0, +1\}$. For item i, define

$$T_i = \frac{1}{W_i} \sum_{j \in \mathcal{A}_i} w_j \mathbb{I}\{X_{ij} = +1\}, \tag{2}$$

$$N_i = \frac{1}{W_i} \sum_{j \in \mathcal{A}_i} w_j \mathbb{I}\{X_{ij} = 0\}, \tag{3}$$

$$F_i = \frac{1}{W_i} \sum_{j \in \mathcal{A}_i} w_j \mathbb{I}\{X_{ij} = -1\}. \tag{4}$$

Thus $T_i + N_i + F_i = 1$. The decisive mass and signed margin are

$$D_i = T_i + F_i = 1 - N_i, \qquad M_i = T_i - F_i. \tag{5}$$

Definition 4.1 (Pure decisive indeterminacy)**.** For $D_i > 0$, the item's pure indeterminacy is

$$I_i = \frac{4T_iF_i}{D_i}. \tag{6}$$

If $D_i = 0$, set $I_i = 0$.

The numerator in Equation (6) is large only when both decisive alternatives receive substantial mass. Dividing by D_i keeps I_i on the same item-level scale as the response masses. Unlike entropy over (T_i, N_i, F_i), I_i does not count neutrality itself as polarization.

Definition 4.2 (Consensus coordinate)**.** For $D_i > 0$, define

$$C_i = \frac{M_i^2}{D_i} = \frac{(T_i - F_i)^2}{T_i + F_i}. \tag{7}$$

If $D_i = 0$, set $C_i = 0$.

C_i is not a fifth neutrosophic component. It is a derived certificate of agreement that makes the allocation of one unit of response evidence explicit.

Theorem 4.3 (Exact evidence decomposition)**.** *For every item with at least one observed response,*

$$C_i + I_i + N_i = 1. \tag{8}$$

Moreover, $T_i, I_i, N_i, F_i, C_i \in [0, 1]$ *and* $I_i \le D_i$.

Proof. If $D_i = 0$, then $(T_i, N_i, F_i) = (0, 1, 0)$, so the identity is immediate. If $D_i > 0$, then

$$C_i + I_i = \frac{(T_i - F_i)^2 + 4T_iF_i}{D_i} = \frac{(T_i + F_i)^2}{D_i} = D_i.$$

Adding $N_i = 1 - D_i$ gives Equation (8). The response masses are normalized nonnegative sums. Finally, $4T_iF_i \le (T_i + F_i)^2 = D_i^2$, hence $0 \le I_i \le D_i \le 1$; the same identity establishes the bound for C_i. □

TABLE 1. Canonical equal-weight response configurations. Both the all-neutral and balanced-decisive panels have signed margin $M = 0$, but the TINF profile identifies different mechanisms.

Panel	T	I	N	F	M	C
Unanimous positive	1	0	0	0	1	1
Unanimous negative	0	0	0	1	-1	1
All explicitly neutral	0	0	1	0	0	0
Half positive, half negative	0.5	1	0	0.5	0	0
Half positive, half neutral	0.5	0	0.5	0	0.5	0.5

Proposition 4.4 (Extremes and symmetry)**.** *For fixed decisive mass $D_i > 0$, $I_i = 0$ if and only if $T_i F_i = 0$, and $I_i = D_i$ if and only if $T_i = F_i = D_i/2$. Interchanging the two class labels swaps T_i and F_i while leaving I_i, N_i, C_i unchanged.*

Proof. The first statement follows from Equation (6). Equality in $4T_iF_i \le (T_i + F_i)^2$ holds exactly when $T_i = F_i$. The symmetry statement follows because Equation (6) is symmetric in T_i, F_i and Equation (7) squares their difference. □

Proposition 4.5 (Information lost by signed-mean aggregation)**.** *No statistic that is a function only of the weighted signed mean M_i can distinguish an all-neutral panel from an evenly weighted panel split between $+1$ and -1. The profiles in Equations (2) to (4) and (6) distinguish them as $(0, 0, 1, 0)$ and $(1/2, 1, 0, 1/2)$, respectively.*

Proof. Both panels have $M_i = 0$, so every function of M_i takes the same value. Direct substitution gives the two profiles. □

Remark 4.6 (Operational specialization of TINF)**.** The source TINF formalism permits dependent components whose sum need not equal one [15]. Our T, N, F are normalized response masses, while I is derived from T, F; consequently the quadruple is a restricted, data-generating specialization and $T + I + N + F$ can exceed one. Also, binary label reversal maps (T, I, N, F) to (F, I, N, T). This is an operational relabelling map, not the source formalism's logical complement (F, N, I, T); the two should not be conflated.

4.2. *Invariance and corruption stability*

Proposition 4.7 (Invariance)**.** *The profile is invariant to any permutation of annotators and to multiplying all observed weights for an item by the same positive constant.*

Proof. Each coordinate depends only on weighted category sums divided by their common total. Reordering terms or applying common positive scaling leaves the ratios unchanged. □

Theorem 4.8 (Fixed-weight response corruption)**.** *Normalize the observed annotator weights for an item to sum to one. If the observed set is fixed and responses with total normalized weight at most ε are changed arbitrarily among $\{-1, 0, +1\}$ while the weights remain fixed, then*

$$\|(T', N', F') - (T, N, F)\|_1 \le 2\varepsilon, \qquad |M' - M| \le 2\varepsilon. \tag{9}$$

Therefore the fused label $\operatorname{sign}(M)$ *is unchanged whenever* $|M| > 2\varepsilon$.

Proof. Changing one response moves its normalized weight from one of the three categories to another, changing the mass-vector ℓ_1 norm by twice that weight. Summation over changed responses gives the first bound. Since a signed response lies in $[-1, 1]$, changing it alters its contribution to M by at most twice its normalized weight, giving the second bound. A sign can change only if the perturbation reaches or crosses zero. □

The theorem controls the linear masses and decision margin without a distributional assumption. We do not claim a uniform Lipschitz bound for I near $D = 0$; any such bound would require a lower bound on decisive mass. This qualification prevents the nonlinear diagnostic from being mistaken for a generic robust estimator.

5. Gold-Anchored Reliability and Fusion

5.1. *Annotator orientation and weight*

Let $\mathcal{G}$ be a set of gold items with known signed labels $Y_i \in \{-1, +1\}$. For annotator j, count only decisive gold responses:

$$c_j = \sum_{i \in \mathcal{G}} \mathbb{I}\{R_{ij} = Y_i\}, \qquad e_j = \sum_{i \in \mathcal{G}} \mathbb{I}\{R_{ij} = -Y_i\}.$$

Under a Beta$(1, 1)$ prior for decisive accuracy, the posterior mean is

$$\widehat{a}_j = \frac{c_j + 1}{c_j + e_j + 2}. \tag{10}$$

Define orientation and reliability weight by

$$s_j = \begin{cases} +1, & \widehat{a}_j \ge 1/2, \\ -1, & \widehat{a}_j < 1/2, \end{cases} \qquad w_j = \max\{\eta, |2\widehat{a}_j - 1|\}, \tag{11}$$

where $\eta > 0$ is a small floor. Corrected responses are $X_{ij} = s_j R_{ij}$. An annotator who is consistently wrong on objective gold items is therefore inverted rather than merely discarded.

The construction does not punish a legitimate neutral response as an incorrect class. However, a worker with no decisive gold responses has $\widehat{a}_j = 1/2$ and receives only the floor weight. If task designers also need to model propensity to abstain, that behavior should be reported as a separate coverage parameter rather than folded into decisive accuracy.

Theorem 5.1 (Gold-orientation error). *Suppose annotator j gives g_j independent decisive gold responses, each correct with probability $\theta_j \neq 1/2$. Then the probability that Equation (11) selects the orientation opposite to $\operatorname{sign}(\theta_j - 1/2)$ is at most*

$$\exp\left[-2g_j(\theta_j - \tfrac{1}{2})^2\right]. \tag{12}$$

As $g_j \to \infty$, s_j is consistent and $w_j \to \max\{\eta, |2\theta_j - 1|\}$ almost surely.

Proof. The posterior mean is at least $1/2$ exactly when $c_j \geq e_j$. For $\theta_j > 1/2$, a wrong orientation requires the empirical correct fraction to fall below $1/2$; for $\theta_j < 1/2$, it requires the fraction to reach at least $1/2$. Hoeffding's inequality gives Equation (12). The strong law of large numbers yields $c_j/g_j \to \theta_j$, and the remaining limits follow by continuity away from $1/2$. □

5.2. *Prediction and a conditional crowd bound*

TINF-Consensus predicts

$$\widehat{Y}_i = \operatorname{sign}(M_i), \tag{13}$$

with exact ties deferred or resolved using a predeclared class prior. The full (T_i, I_i, N_i, F_i) profile is retained alongside the label.

Theorem 5.2 (Conditional fusion error). *For a fixed item with true signed label Y, suppose corrected annotator responses $X_j \in [-1, 1]$ are conditionally independent and satisfy $\mathbb{E}[YX_j] \geq \gamma_j > 0$. Then*

$$\mathbb{P}\left(\operatorname{sign}\left(\sum_j w_j X_j\right) \neq Y\right) \leq \exp\left[-\frac{\left(\sum_j w_j\gamma_j\right)^2}{2\sum_j w_j^2}\right]. \tag{14}$$

Proof. Let $Z_j = w_jYX_j \in [-w_j, w_j]$ and $\mu = \sum_j \mathbb{E}Z_j \geq \sum_j w_j\gamma_j$. A label error implies $\sum_j Z_j \leq 0$. Hoeffding's inequality gives

$$\mathbb{P}\left(\sum_j (Z_j - \mathbb{E}Z_j) \leq -\mu\right) \leq \exp\left(-\frac{2\mu^2}{\sum_j (2w_j)^2}\right),$$

which yields Equation (14). □

Conditional independence is a transparent analytical assumption, not a claim about all human panels. Shared training, communication, or common cultural perspectives can correlate responses and weaken the bound. The empirical profile remains computable under dependence.

Algorithm 1 (TINF-Consensus with objective gold items). *(1)* ***Input:*** *response matrix R with entries in $\{-1, 0, +1, \perp\}$; gold set $\mathcal{G}$ and signed labels $Y_{\mathcal{G}}$; weight floor η.*

(2) For each annotator j, count decisive gold agreements c_j and errors e_j; set $\widehat{a}_j \leftarrow (c_j + 1)/(c_j + e_j + 2)$, $s_j \leftarrow +1$ if $\widehat{a}_j \geq 1/2$ and -1 otherwise, and $w_j \leftarrow \max\{\eta, |2\widehat{a}_j - 1|\}$.

(3) For each non-gold item i, set $X_{ij} \leftarrow s_jR_{ij}$ for observed responses and compute T_i, N_i, F_i using Equations (2) to (4).

(4) Set $D_i \leftarrow T_i + F_i$ *and* $M_i \leftarrow T_i - F_i$*; compute* I_i *and* C_i *by Equations (6) and (7).*

(5) ***Return:*** $\widehat{Y}_i = \operatorname{sign}(M_i)$ *and* $(T_i, I_i, N_i, F_i, C_i, M_i)$.

6. Two-Stage Review: Whether and Why

Label triage involves at least two decisions. First, which fused labels are most likely to be wrong? Second, what kind of review should those items receive? TINF-Consensus uses separate statistics.

Error screening. The weighted margin deficit

$$E_i = 1 - |M_i| \tag{15}$$

ranks items whose weighted class decision is fragile. A low decisive mass or a close decisive split can both reduce $|M_i|$, which is appropriate when the goal is to predict an error in Equation (13).

Reason routing. Among items selected for process review, use

$$\rho_i = \begin{cases} \text{conflict route}, & I_i > N_i, \\ \text{neutrality route}, & I_i \leq N_i. \end{cases} \tag{16}$$

A conflict route may request independent adjudication, inspect annotator subgroups, or retain a distribution of labels. A neutrality route may request additional evidence, revise instructions, or explicitly keep the neutral class. If a single budget score is required, a declared utility parameter $\beta \in [0, 1]$ gives

$$U_i(\beta) = \beta I_i + (1 - \beta) N_i. \tag{17}$$

The parameter is a policy choice, not an estimated universal constant.

For objective gold tasks, error screening can precede reason routing. For subjective or value-laden tasks, "error" may be the wrong construct: equal/contextual weights and distribution-preserving outputs may be preferable. In particular, inverting an annotator because that person differs from a majority-defined "gold" label would erase perspective rather than correct noise. The gold procedure is therefore restricted to settings with defensible external labels.

7. Experimental Evaluation

7.1. *Goals and datasets*

The experiments test controlled mechanisms, not application-specific performance. Four binary tasks bundled with scikit-learn were selected so the study can run without downloading data [9]. Their diversity in sample size, dimension, and class balance provides several response difficulty landscapes. Dataset features are used only to construct a fixed simulator difficulty; none of the aggregation methods observes them.

7.2. *Semi-synthetic annotator model*

For each task, standardized features are fitted with logistic regression. If p_i is its fitted positive probability, boundary proximity is

$$q_i = 1 - |2p_i - 1| \in [0, 1]. \tag{18}$$

TABLE 2. Binary benchmark tasks. Positive and negative counts refer to the class encoding used in the experiment.

Task	Items	Features	Positive	Negative
Breast cancer	569	30	357	212
Digits 3 versus 8	357	64	174	183
Iris versicolor versus virginica	100	4	50	50
Wine class 0 versus rest	178	13	59	119

This quantity is available only to the response simulator. The crowd has 21 annotators with base decisive accuracies

$$\boldsymbol{\theta} = (0.95 \times 3,\ 0.85 \times 5,\ 0.72 \times 8,\ 0.58 \times 5).$$

On item i, annotator j has effective decisive accuracy

$$\theta_{ij}^{\mathrm{eff}} = \tfrac{1}{2} + (\theta_j - \tfrac{1}{2})(1 - 0.55q_i), \tag{19}$$

so hard items shrink all annotators toward chance.

The neutral-response probability depends on a known experimental regime:

$$\begin{aligned} p_i^{(0)} &= \min\{0.65,\ 0.03 + 0.35q_i\} && \text{(regular)}, \\ p_i^{(N)} &= \min\{0.90,\ 0.65 + 0.20q_i\} && \text{(explicit neutrality)}, \\ p_i^{(I)} &= \min\{0.15,\ 0.02 + 0.08q_i\} && \text{(perspective conflict)}. \end{aligned} \tag{20}$$

For conflict items, ten annotators target the dataset label and eleven target its opposite, after which their individual decisive error is sampled by Equation (19). This generates polarized panels without asserting that one perspective is normatively invalid.

Each repetition uses a stratified 30% gold split and a 70% evaluation split. All reported summaries use 40 deterministic repetitions with predeclared seeds.

7.3. *Experiment A: adversarial robustness*

We use nominal adversarial settings $a \in \{0, 0.1, 0.2, 0.3, 0.4\}$ and replace the nearest whole number of annotators with workers whose decisive accuracy is 0.15. For 21 annotators, the settings correspond to $0, 2, 4, 6$, and 8 workers. The evaluated methods are:

Majority vote::

neutral responses contribute zero; exact ties use the majority class of the gold split.

Gold-anchored Dawid–Skene::

a binary EM model with unit pseudocounts; neutral responses are treated as missing and gold labels are held fixed.

Gold-TINF::

Algorithm 1 with $\eta = 0.02$.

Oracle-TINF::

the same profile using the simulator's true θ_j for orientation and weights; this is an unattainable reference.

We report accuracy and balanced accuracy on non-gold items. Gold-TINF and Dawid–Skene receive the same gold subset. The unweighted majority receives only its class prior from that split, while Oracle-TINF receives the simulator's true annotator parameters and is not a deployable competitor.

7.4. *Experiment B: mechanism diagnosis and selective review*

In each evaluation split, the hardest 25% of items are assigned the neutral regime, a random 25% of the remaining items are assigned perspective conflict, and the rest are regular. Gold items remain regular and objective. We evaluate conflict-versus-neutral routing among the two mechanism groups using AUROC and the reason-rule accuracy in Equation (16).

For error screening, AUROC measures discrimination of incorrectly fused labels. Area under the risk–coverage curve (AURC; lower is better) evaluates selective retention when high-score items are deferred. We compare raw and weighted margin deficit, three-way entropy over (T, N, F), pure indeterminacy I, and $U_i(0.75)$.

Finally, a transparent decision-analysis stress test assigns potential review benefit 1 to a conflict item, 0.25 to a neutral item, and 0 to a regular item. For a top-20% review budget, selected benefit is divided by the maximum available benefit. This *normalized review utility* is not a measured causal effect or a universal cost model; it only tests ranking behavior under declared priorities.

7.5. *Implementation and reproducibility*

The reference implementation is a single Python program. It uses Python 3.12.13, NumPy 2.3.5, and scikit-learn 1.8.0 in the reported run. The simulator, aggregation methods, metrics, seeds, and CSV writers are all included. No network access, unpublished data, or manual result editing is required.

8. **Results**

8.1. *The profile separates neutrality from conflict*

Table 3 reports macro-averages of the dataset-level repetition means. Neutral items allocate 0.7002 of evidence to N, while conflict items allocate 0.8130 to I. Both have low consensus, yet for different reasons. Regular items also show nonzero indeterminacy because the simulated crowd is heterogeneous and hard items induce errors; the diagnostic claim is comparative, not that ordinary crowds have $I = 0$.

The TINF contrast $I - N$ achieves macro-average conflict AUROC 0.99999 among neutral and conflict items. The simple rule $I > N$ routes the mechanism with 0.9928 accuracy (Table 4). In contrast, absolute signed mean and three-way entropy do not identify which mechanism generated uncertainty. Their AUROCs

TABLE 3. Macro-average reliability-weighted profiles over four datasets and 40 repetitions. The identity error for $C + I + N = 1$ was below 4×10^{-16} in every dataset summary.

Regime	T	I	N	F	C
Regular	0.4619	0.5728	0.0363	0.5018	0.3910
Explicit neutrality	0.1543	0.1711	0.7002	0.1454	0.1287
Perspective conflict	0.4834	0.8130	0.0223	0.4942	0.1646

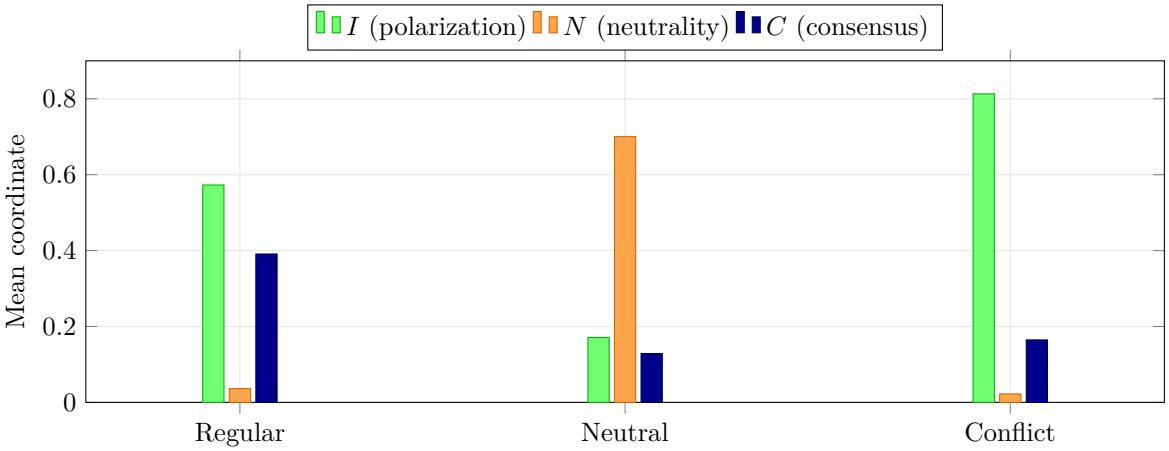

FIGURE 1. The decomposition assigns uncertainty to different mechanisms. Explicit-neutral items are dominated by N; conflict items are dominated by I.

TABLE 4. Mechanism diagnosis by dataset (mean over 40 repetitions). AUROC compares conflict with explicit-neutral items; accuracy evaluates Equation (16).

Dataset	$-\lvert M_{\text{raw}}\rvert$	Entropy	$I - N$	Rule accuracy
Breast cancer	0.4223	0.3917	1.0000	0.9929
Digits 3 versus 8	0.4613	0.3207	1.0000	0.9915
Iris versicolor versus virginica	0.3468	0.5688	1.0000	0.9965
Wine class 0 versus rest	0.4232	0.3763	1.0000	0.9903
Macro-average	0.4134	0.4143	1.0000	0.9928

below 0.5 are not typographical errors: neutral items often have an even smaller signed margin than polarized items, so an undirected more uncertain ranking reverses the desired conflict ordering.

8.2. *Gold weighting neutralizes consistent adversaries*

Table 5 and Figure 2 show macro-average accuracy. Majority vote deteriorates from 0.9864 with no injected adversaries to 0.5930 at the nominal 40% setting (eight of 21). Gold-TINF remains above 0.991 throughout

TABLE 5. Macro-average evaluation accuracy across four datasets. Each cell averages dataset-level means from 40 repetitions. Adversary percentages are nominal and correspond to $0, 2, 4, 6$, and 8 of 21 workers.

Adversaries	Majority	Dawid–Skene	Gold-TINF	Oracle-TINF
0%	0.9864	0.9929	0.9919	0.9929
10%	0.9675	0.9945	0.9947	0.9956
20%	0.9163	0.9943	0.9941	0.9946
30%	0.7977	0.9951	0.9958	0.9959
40%	0.5930	0.9949	0.9942	0.9947

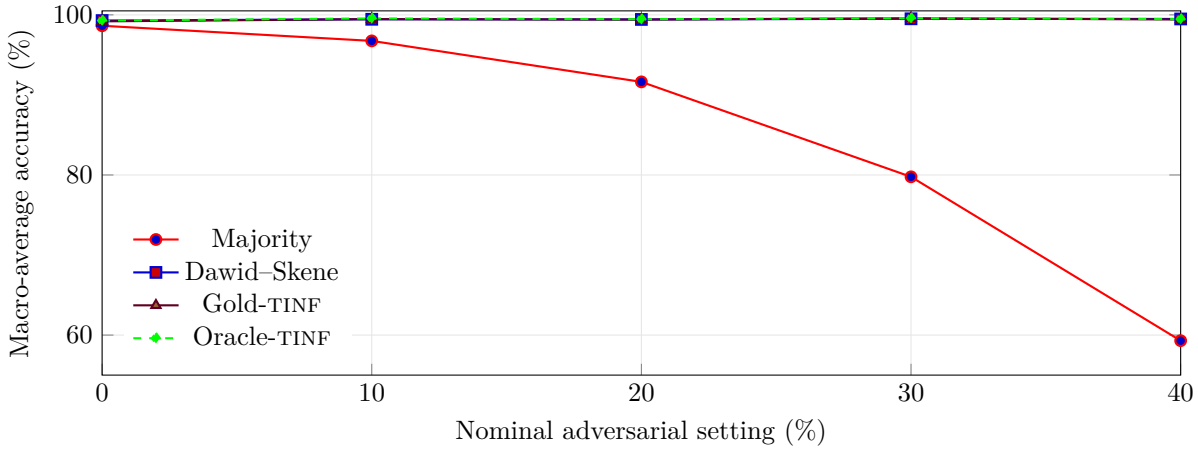

FIGURE 2. Gold information identifies consistently inverted annotators. Gold-TINF and gold-anchored Dawid–Skene remain stable as the unweighted majority collapses.

and reaches 0.9942 in that setting, a difference of 40.1 percentage points from majority vote. Gold-anchored Dawid–Skene performs similarly (0.9949 at the nominal 40% setting), and the oracle profile reaches 0.9947. The correct conclusion is parity with a strong gold-aware latent-class baseline in this simulator, not superiority over Dawid–Skene.

8.3. *Error detection and mechanism routing require different scores*

Table 6 is the key negative as well as positive result. Weighted margin deficit is the best error detector, with macro-average AUROC 0.9552 and AURC 0.0028. Pure indeterminacy and the mixed TINF review score perform poorly for error AUROC because high process disagreement is not equivalent to an incorrect fused label. Conversely, pure I gives the highest normalized mechanism-weighted review utility (0.6996), closely followed by $U(0.75)$ (0.6959). Weighted margin is less aligned with that declared conflict-prioritizing utility (0.3961).

TABLE 6. Macro-average selective-review results. Error AUROC and review utility are higher-is-better; AURC is lower-is-better. Utility uses the declared synthetic benefits in Section 7.

Ranking score	Error AUROC	AURC	Review utility
Raw margin deficit	0.8299	0.0077	0.5419
Weighted margin deficit	**0.9552**	**0.0028**	0.3961
Three-way entropy	0.5883	0.0331	0.3563
Pure indeterminacy I	0.3688	0.0539	**0.6996**
$U(0.75) = 0.75I + 0.25N$	0.4701	0.0361	0.6959

The evidence answers RQ3 in the negative: forcing one scalar to represent both prediction risk and the reason for disagreement loses useful structure. The recommended workflow is to use $1 - |M|$ to allocate a generic error-review budget and then use the pair (I, N) to route or describe the selected cases. If a governance process values contested items independently of label error, I can define a separate review queue.

9. Discussion

9.1. *What the method adds*

TINF-CONSENSUS adds an explanatory layer to weighted voting. The predicted label is intentionally simple: it is the sign of a reliability-weighted response sum. The novelty lies in retaining the response geometry after fusion. Theorem 4.3 says that one unit of item-level evidence is allocated among consensus, decisive polarization, and explicit neutrality. This makes three operational states visible:

(1) high C: decisive support is concentrated on one class;
(2) high I: decisive support is split across classes; and
(3) high N: annotators explicitly occupy or select the neutral option.

The coordinates do not themselves explain *why* a person was neutral or why two people disagreed. They identify an observable pattern that can guide the next inquiry. For example, a conflict route may stratify profiles by annotator role or demographic group, while a neutrality route may inspect item evidence and the wording of the response options.

9.2. *Why entropy and confidence are insufficient*

Entropy over three response categories is useful as a measure of dispersion, but it is symmetric with respect to the names of those categories. A panel distributed between positive and negative and a panel distributed between a decisive class and neutrality can have similar entropy even though they imply different review actions. A scalar posterior confidence has the same general limitation: it indicates that the final decision

is uncertain, not the mechanism of uncertainty. The explicit pair (I, N) restores that mechanism for the particular response protocol considered here.

At the same time, Table 6 cautions against replacing confidence with indeterminacy. A highly polarized crowd can still have a correct fused label, especially after reliable annotators are weighted. This is why margin deficit is empirically better for error prediction. The method's most useful output is therefore not a new universal confidence number but a small structured record:

$$(\widehat{Y}_i,\ M_i,\ T_i,\ I_i,\ N_i,\ F_i).$$

9.3. *Objective and subjective annotation*

The gold-anchored results apply most naturally to tasks with externally defensible answers: image quality checks, diagnostic test interpretation under a fixed standard, transcription, or factual verification. Even there, gold sets should cover the relevant difficulty range and be audited for leakage or systematic bias.

For subjective constructs—toxicity, political framing, emotion, fairness, or acceptability—a single gold orientation may not exist. In those settings TINF-Consensus can still compute equal-weight or context-specific profiles, but orientation inversion should be disabled. High I may be the desired scientific result, not a defect. Downstream models can consume soft response distributions, annotator-conditioned targets, or multiple perspective labels rather than a forced consensus [3, 18].

9.4. *Practical deployment*

A deployment protocol should declare the neutral option in advance and tell annotators what it means. "Neither class," "insufficient evidence," "both," and "prefer not to answer" are not interchangeable. If these meanings matter, they should be separate response codes and may require more than one neutrality coordinate.

The following sequence is a reasonable objective-task workflow:

(1) collect representative gold items and ordinary items with missingness recorded separately from explicit neutrality;
(2) estimate orientation and reliability with Equations (10) and (11);
(3) fuse labels by weighted margin and retain the complete item profile;
(4) allocate error review with $1 - |M|$;
(5) route reviewed items using I versus N; and
(6) monitor profiles over annotator groups and time rather than auditing aggregate accuracy alone.

The final monitoring step can reveal guideline drift. A rise in N suggests that more annotators are declining to decide; a rise in I suggests that decisive interpretations are diverging. These are diagnostic hypotheses, to be confirmed by qualitative and domain-specific evidence.

9.5. *Ethical and governance considerations*

Annotator weights encode institutional judgments about whose answers count. Even objective gold questions can advantage workers familiar with a dominant terminology or practice. Reliability estimates should therefore be used for quality control with notice, appeal, and subgroup audit, not as opaque worker rankings. Low-weight annotations should remain available for analysis when privacy and consent permit.

Conflict routing can also expose sensitive group differences. Reporting should use aggregation thresholds, access controls, and data-minimization principles. A high-I item should not be described as evidence that a particular group is irrational or malicious. It is evidence only that the weighted decisive responses, under the chosen coding and weights, are split.

10. **Limitations and Future Work**

This study has seven main limitations.

(1) **Semi-synthetic evidence.** The response mechanisms are known by construction, enabling exact diagnosis but limiting external validity. Evaluation on naturally collected multi-rater datasets with an explicit neutral option is the next essential step.

(2) **Binary labels.** One-versus-rest profiles are possible, but they do not by themselves capture conflict among several non-reference classes. A principled multiclass generalization should preserve class geometry and the neutrality distinction.

(3) **Gold dependence.** Orientation and reliability require representative, defensible gold items. Sparse or biased gold sets can misorient annotators. Hierarchical shrinkage and uncertainty intervals for w_j are promising extensions.

(4) **Conditional independence.** The exponential error bound assumes conditionally independent corrected responses. Correlated annotator communities require cluster-aware concentration bounds or explicit dependence models.

(5) **One operational mapping.** The formula $I = 4TF/(T + F)$ is motivated by symmetry, boundedness, and exact decomposition, but it is not the only possible indeterminacy functional. Axiomatic uniqueness results and sensitivity comparisons are open.

(6) **No downstream learner study.** We evaluate aggregation and routing, not whether feeding the profiles to a classifier improves generalization or calibration.

(7) **Synthetic review utility.** The utility experiment verifies a ranking under declared benefits; it does not estimate real review cost or causal benefit. Deployment needs domain-specific elicitation and a prospective study.

Future work should collect explicit-neutral datasets, evaluate annotator-group profiles without exposing identities, and compare the method with modern distribution-preserving and annotator-conditioned neural models. Other directions include time-varying weights, Bayesian posterior profiles, active selection of gold questions, and multiclass conflict graphs that identify which alternatives divide the crowd.

11. Conclusion

Multi-annotator uncertainty is not one phenomenon. An all-neutral panel and a polarized decisive panel can have the same signed mean while demanding different responses. TINF-Consensus preserves this distinction through an observable TINF profile and the exact decomposition $C + I + N = 1$. Gold-based orientation and reliability weighting give robust label fusion in controlled adversarial experiments, with performance comparable to a gold-anchored Dawid–Skene model. The strongest practical result is a division of labor: weighted margin is effective for predicting fusion errors, whereas I and N identify the mechanism of unresolved annotation. This two-output view avoids treating legitimate disagreement as mere noise and turns consensus aggregation into an auditable process.

Declarations

Funding. NA.

Competing interests. NA.

Ethics statement. The experiments use datasets distributed with scikit-learn and simulated annotator responses; no new human participants or identifiable annotator data were collected.

Data and code availability. The accompanying reproducibility package contains the complete Python implementation and all reported CSV summaries. The public datasets are loaded through scikit-learn without network access.

Appendix A. Additional Proof Details

A.1. *Alternative form of the decomposition*

For $D_i > 0$, write decisive balance $b_i = (T_i - F_i)/(T_i + F_i) \in [-1, 1]$. Then

$$C_i = D_i b_i^2, \qquad I_i = D_i(1 - b_i^2), \qquad N_i = 1 - D_i. \tag{21}$$

This form shows that the construction has two stages. First, total evidence is split between neutrality N_i and decisiveness D_i. Second, decisive evidence is split between squared balance b_i^2 (consensus) and its complement $1 - b_i^2$ (polarization). It also makes Equation (8) immediate.

A.2. *Why the factor four is necessary*

Consider the symmetric family $I_i^{(k)} = kT_iF_i/D_i$ for $D_i > 0$. Requiring a perfectly balanced decisive panel, $T_i = F_i = D_i/2$, to have maximal indeterminacy $I_i = D_i$ gives

$$k\frac{(D_i/2)^2}{D_i} = D_i,$$

so $k = 4$. Thus the factor is fixed within this bilinear symmetric family by the scale convention used in Theorem 4.3.

A.3. *Tie handling*

When $M_i = 0$, the profile deliberately declares no directional consensus. The reference experiment resolves a prediction tie using the majority class in the gold subset solely so that accuracy is defined for every item. A production system should instead consider deferral, randomization with a declared prior, or preservation of both labels. Tie handling affects the point prediction but not (T, I, N, F, C).

Appendix B. Additional Experimental Results

TABLE 7. Accuracy at the nominal 40% setting (eight adversarial annotators out of 21) by dataset, averaged over 40 repetitions.

Dataset	Majority	Dawid–Skene	Gold-TINF	Oracle-TINF
Breast cancer	0.6072	0.9964	0.9962	0.9964
Digits 3 versus 8	0.5863	0.9986	0.9987	0.9987
Iris versicolor versus virginica	0.5707	0.9871	0.9839	0.9861
Wine class 0 versus rest	0.6076	0.9976	0.9978	0.9976

TABLE 8. Profile means by dataset and regime. Values are ordered (I, N, C) and averaged over 40 repetitions.

Dataset	Regular	Explicit neutrality	Conflict
Breast cancer	(.567, .031, .402)	(.170, .694, .136)	(.812, .020, .168)
Digits 3 vs. 8	(.567, .030, .403)	(.186, .664, .150)	(.814, .022, .164)
Iris Vc. vs. Vg.	(.587, .050, .363)	(.146, .756, .098)	(.812, .025, .163)
Wine 0 vs. rest	(.571, .034, .395)	(.183, .687, .130)	(.814, .023, .163)

Appendix C. Reproducibility Specification

The program writes four CSV summaries:

`robustness_summary.csv`	adversarial accuracy and balanced accuracy,
`routing_summary.csv`	mechanism AUROC and routing accuracy,
`profile_summary.csv`	regime-specific TINF coordinates,
`selective_summary.csv`	error AUROC, AURC, and review utility.

Standard deviations in those files are sample standard deviations across the 40 repetitions. Macro-averages in the main text are unweighted means of the four dataset-level means, preventing the largest dataset from dominating the summary.

TABLE 9. Parameters fixed before the reported run.

Parameter	Value
Annotators	21
Base decisive accuracies	3×0.95, 5×0.85, 8×0.72, 5×0.58
Gold fraction	0.30, stratified by class
Repetitions	40 per dataset and condition
Adversary grid	Nominal $0, 0.1, 0.2, 0.3, 0.4$, corresponding to $0, 2, 4, 6, 8$ of 21 workers; adversary accuracy 0.15
Difficulty shrinkage	0.55 in Equation (19)
Weight prior and floor	Beta$(1, 1)$; $\eta = 0.02$
Dawid–Skene	Unit cell pseudocount; at most 80 EM iterations; tolerance 10^{-8}; neutral treated as missing
Mechanism allocation	25% explicit neutral, 25% perspective conflict, 50% regular on the evaluation split
Review budget and utility	Top 20%; benefits $(0, 0.25, 1)$ for (regular, neutral, conflict)
Software	Python 3.12.13, NumPy 2.3.5, scikit-learn 1.8.0

Appendix D. Reporting Checklist for a Real-World Extension

Before applying the method to naturally collected annotations, a study should report at least:

(1) the exact wording and semantics of every neutral or abstention option;
(2) how missing responses differ technically from explicit neutrality;
(3) recruitment, training, compensation, and consent procedures;
(4) the construction, representativeness, and audit of gold items;
(5) uncertainty intervals for annotator reliability and subgroup-level profile summaries where ethically permissible;
(6) the tie, deferral, adjudication, and appeal policies;
(7) the review utility or cost model and its stakeholders;
(8) sensitivity to weight floors, gold-set size, and alternative indeterminacy functions; and
(9) whether disagreements are treated as errors, perspectives, or both, with domain justification.

References

[1] Artstein, R.; Poesio, M. Inter-coder agreement for computational linguistics. Computational Linguistics 2008, 34(4), 555–596.
[2] Atanassov, K.T. Intuitionistic fuzzy sets. Fuzzy Sets and Systems 1986, 20(1), 87–96.

[3] Davani, A.M.; Díaz, M.; Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics 2022, 10, 92–110.

[4] Dawid, A.P.; Skene, A.M. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) 1979, 28(1), 20–28.

[5] Fleisig, E.; Abebe, R.; Klein, D. When the majority is wrong: Modeling annotator disagreement for subjective tasks. Proceedings of EMNLP 2023, 6715–6726.

[6] Fornaciari, T.; Uma, A.N.; Paun, S.; Plank, B.; Hovy, D.; Poesio, M. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. Proceedings of NAACL-HLT 2021, 2591–2597.

[7] Jagabathula, S.; Subramanian, L.; Venkataraman, A. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. Journal of Machine Learning Research 2017, 18(93), 1–67.

[8] Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; Poesio, M. Comparing Bayesian models of annotation. Transactions of the Association for Computational Linguistics 2018, 6, 571–585.

[9] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.

[10] Peterson, J.C.; Battleday, R.M.; Griffiths, T.L.; Russakovsky, O. Human uncertainty makes classification more robust. Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, 9617–9626.

[11] Plank, B. The "problem" of human label variation: On ground truth in data, modeling and evaluation. Proceedings of EMNLP 2022, 10671–10682.

[12] Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. Learning from crowds. Journal of Machine Learning Research 2010, 11, 1297–1322.

[13] Rodrigues, F.; Pereira, F. Deep learning from crowds. Proceedings of the AAAI Conference on Artificial Intelligence 2018, 32.

[14] Smarandache, F. Neutrosophy: Neutrosophic Probability, Set, and Logic. American Research Press: Rehoboth, 1998.

[15] Smarandache, F. $(\mathsf{T}, \mathsf{I}, \mathsf{N}, \mathsf{F})$ Neutrosophic Set and Logic. Critical Review 2017, 14, 20–28.

[16] Uma, A.N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M. Learning from disagreement: A survey. Journal of Artificial Intelligence Research 2021, 72, 1385–1470.

[17] Wang, H.; Smarandache, F.; Zhang, Y.-Q.; Sunderraman, R. Single valued neutrosophic sets. Multispace and Multistructure 2010, 4, 410–413.

[18] Weerasooriya, T.; Liu, A.; Homan, C. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with discourse-level annotation. Findings of the Association for Computational Linguistics: ACL 2023 2023, 4679–4695.

[19] Whitehill, J.; Wu, T.-F.; Bergsma, J.; Movellan, J.R.; Ruvolo, P.L. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in Neural Information Processing Systems 2009, 22.

[20] Zadeh, L.A. Fuzzy sets. Information and Control 1965, 8(3), 338–353.

[21] Zhang, Y.; Chen, X.; Zhou, D.; Jordan, M.I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. Advances in Neural Information Processing Systems 2014, 27.

Received: Jan 7, 2026. Accepted: Aug 22, 2026