THREE IS NOT A CROWD: A DECISION THEORY FOR LABEL VALIDITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Annotators disagree, and in many annotation tasks more than one label for the same object may be genuinely valid. Label aggregation must therefore address two questions simultaneously: which labels are sufficiently supported to be treated as valid, and how reliable is the resulting set of selections? Standard approaches do not generally provide both. Majority vote and classical latent-label models such as Dawid–Skene typically return a single aggregate label, often without an explicit mechanism for withholding a decision when evidence is weak. Distribution-preserving approaches take a different view, representing disagreement through soft labels rather than determining which individual labels have sufficient evidence to be selected as valid. More broadly, existing methods typically quantify uncertainty through global performance measures or model-dependent quantities, rather than providing an explicit error-rate guarantee over the labels ultimately selected. We recast label recovery as a multiple-testing selection problem. Each object–label pair defines a statistical hypothesis, and labels are selected only when the annotation evidence is sufficiently strong; otherwise, they remain unresolved. Using false discovery rate (FDR) control, the procedure bounds the expected proportion of invalid labels among the selected object–label pairs at a user-specified level. The resulting guarantee is therefore not merely a post-hoc evaluation metric, but an error-control criterion built directly into the recovery procedure. Our framework requires neither a predefined number of valid labels per object nor an explicit model of annotator reliability, and avoids combinatorial search over candidate label subsets. This reframing turns label aggregation from forced estimation into statistically controlled selection under disagreement.

1 INTRODUCTION

Crowd labels often disagree, and that disagreement is not always noise. Ambiguity in the object, differences in social position or expertise, and underspecified instructions can make multiple labels defensible for the same item Plank (2022). A comment may be both threatening and identity-directed, and an image may contain several species. In such settings, collapsing all reports to a single label throws away information where the target is plural. Perspectivist approaches therefore preserve individual annotations and treat their variation as potentially meaningful signal rather than mere noise (Aroyo & Welty, 2015).

Aggregation calls one label wrong; perspectivism calls neither wrong; both answer a question: *how many labels of each object are actually valid?* Classical aggregation methods—majority vote, Dawid–Skene (DS), and their many extensions (Dawid & Skene, 1979)—sidestep this question by assuming a single gold label per object. Accommodating multiple valid labels, however, is far from straightforward: when the valid annotation of an object is an unknown subset of K candidate labels rather than a single class, a fully joint latent-variable model must in principle distinguish among 2^K possible label sets. Existing multi-label and multi-annotator methods tame this combinatorial explosion through factorization, structural restrictions, or an assumed upper bound on the number of valid labels per object. These assumptions can become restrictive as the label space grows, especially in modern crowdsourcing and LLM-assisted annotation pipelines.

More consequential than the question of how many labels are valid is the question of *how they should be recovered* — and on *what grounds the accuracy of that recovery can be guaranteed*. Prevailing methods answer both questions with a single device: a probabilistic latent-variable model that encodes annotator reliability and label noise, fitted to the observed votes and then decoded — typically by maximum a posteriori (MAP) inference — into the most probable label set. Within this paradigm,

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.

the accuracy of recovery is only as good as the model's fit: nothing in the procedure asks whether the evidence for a particular label is actually sufficient, and with as few as three annotators per object the fitted model will still return a definite answer. Even the simplest of these models—a two-component mixture—rests on demanding fitting assumptions: annotators conditionally independent given latent true label, and confusion matrices informative enough to guarantee identifiability (Zhang et al., 2016). In practice, annotator behavior is heterogeneous, context-dependent, and correlated, so fitted reliability parameters absorb model misspecification rather than reflect true annotator quality (Balasubramanian et al., 2026).

Finally, multi-label annotation is sparse in two senses: labeling is expensive, so objects receive votes from only a few of the available annotators; and among the many candidate object–label pairs, only a small fraction is valid. Recovering valid labels is therefore an *imbalanced discovery problem*, so Hamming loss can be misleading, because a model that predicts no labels incurs little error by agreeing with the overwhelming number of negative pairs, despite recovering none of the valid ones. Moreover, it weights false inclusions and missed valid labels equally, even though false inclusions can direct costly downstream review or validation toward spurious labels. FDR instead targets the reliability of the reported discoveries: it controls the expected proportion of invalid labels among those declared valid. Thus, FDR is a natural error criterion for sparse label recovery, while recall or power should be reported separately to quantify missed valid labels.

1.1 Our Contributions

We therefore argue for a paradigm shift: away from point estimation and toward a decision-theoretic view of annotation, in which each object–label pair is treated as a statistical hypothesis and only pairs supported by sufficient evidence are released. In this approach, Recovery does not rely on the usual generative assumptions: a single true class, conditionally independent workers, or an identifiable confusion matrix for each annotator. It also does not require knowing how many labels are valid, enumerating 2^K label subsets, or filling a complete worker-object table. The procedure uses only aggregate support counts, needs no gold negatives or annotator identities, and may abstain when evidence is weak rather than forcing a MAP include decision.

Because invalid pairs are unobserved, we learn what null support looks like from the aggregation table. A conservative lower-tail screen retains null-like pairs; we prove it is pure when valid labels sit above that tail. We Poisson-thin training counts into independent screening and fitting streams, fit a right-truncated Poisson null to the latter, and cross-fit at the object level to obtain held-out p-values. Under explicit screen-purity, identifiability, and regularity conditions, we establish conditional consistency of the screened estimator and asymptotic validity of the resulting p-values; then a standard multiple-testing rule (Benjamini–Hochberg (BH) (Benjamini & Hochberg, 1995), or a weighted variant) chooses which claims to release. Building this null evidence is the contribution

On simulations, known valid-label ground truth is used only to evaluate false discoveries: BH and weighted BH stay under the target FDR, whereas majority and MAP rules do not. Diagnostics probe the estimated null separately. On six multi-valid crowd tables, the selectors give descriptive FDP–power points, not a certificate of prospective control.

1.2 Related Work

Crowd aggregation and multiple valid labels. Classical aggregation methods infer a single latent class under worker-specific noise (Dawid & Skene, 1979; Zhang et al., 2016). Later work considers sparse assignments and correlated annotators, including dependent LLM judges (Karger et al., 2011; Traganitis & Giannakis, 2022; Balasubramanian et al., 2026). Multi-label extensions instead estimate label sets using joint latent models, structural reductions, or annotator transition matrices (Duan et al., 2014; Zhang & Wu, 2018; Yin et al., 2021; Xia et al., 2024; Wang et al., 2025). These approaches target consensus, prediction, or annotation efficiency, but do not control the false fraction among the object–label pairs ultimately selected.

Disagreement as signal. Perspectivist research interprets disagreement as possible ambiguity or population variation rather than only annotation error (Aroyo & Welty, 2015; Plank, 2022; Mostafazadeh Davani et al., 2022). Recent methods separate genuine variation from noise using additional validity judgments (VariErr) or annotator-background models (NUTMEG) (Weber-Genzel et al., 2024; Ivey et al., 2025). These methods characterize or denoise disagreement; our goal is complementary: selecting a variable-size set of supported labels with an explicit error guarantee.

Error-controlled selection. FDR has been applied to detecting biased annotators (Xu et al., 2016), while conformal and learn-then-test methods control predictive errors using labeled calibration or reference data (Angelopoulos et al., 2025; Marandon, 2024). Recent work similarly certifies model-generated labels or selects AI labels under statistical guarantees (Candès et al., 2025). In our setting, however, every candidate's validity is latent and no trusted null examples are available. The main contribution is therefore constructing valid null evidence from a sparse, contaminated crowd table and using it to control FDR over the released object–label pairs.

2 Problem Formulation

There are N objects and a vocabulary of K labels; write $[K] := \{1, \dots, K\}$. Object i is offered its own candidate set $\mathcal{K}_i$, and M_i annotators each mark some of those candidates. The support S_{ik} is how many of them mark label k, and y_{ik} is that count as observed. Objects and labels also have feature vectors x_i and c_k. The unknown target for object i is its valid set $\mathbf{Z}_i^* \subseteq \mathcal{K}_i$: the labels that are genuinely associated with the object. The indicator Z_{ik}^* equals one when label k belongs to that set. Inside $\mathcal{K}_i$, write $\mathcal{D}_i$ for the valid candidates and $\mathcal{N}_i$ for the invalid ones. Every candidate pair (i, k) is a hypothesis. These pairs form the family $\mathcal{H} := \{(i, k) : k \in \mathcal{K}_i\}$, of size $H := |\mathcal{H}|$. We test the pairs one at a time, $H_{ik}^0 : Z_{ik}^* = 0$ against $H_{ik}^1 : Z_{ik}^* = 1$. The null says label k is invalid for object i; rejecting it declares the label valid. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be the truly invalid and truly valid pairs. A procedure returns the rejection set $\widehat{\mathcal{R}} \subseteq \mathcal{H}$, the labels it is willing to release. We measure false selections by FDR and recovery by power:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\widehat{\mathcal{R}} \cap \mathcal{H}_0|}{|\widehat{\mathcal{R}}| \vee 1}\right], \qquad \mathrm{Power} = \mathbb{E}\left[\frac{|\widehat{\mathcal{R}} \cap \mathcal{H}_1|}{|\mathcal{H}_1| \vee 1}\right].$$

FDR is the expected fraction of released labels that are invalid; power is expected recall among the valid pairs. The goal is to control global FDR while keeping power high. The predefined FDR level q is part of the rule that builds $\widehat{\mathcal{R}}$.

Validity in our approach differs from that of majority-vote and Dawid–Skene-style models. We suppose there exists a set $\mathbf{Z}_i^* \subseteq \mathcal{K}_i$ containing every label that is defensible under the perspectives in the reference population. However, we do not attempt to recover all of these labels: based on the annotations, every label is presumed invalid until the evidence against that presumption is strong enough to reject it. Our goal is to select exactly those labels whose evidence clears this bar, while controlling the rate of false rejections.

2.1 Why Sparse Multi-Valid Annotation Calls for Selective Recov7ry

In § 1 we noted that human annotations are sparse in two ways. A review of the most widely used annotated dataset (Fig. 1, and 6) shows that many objects are annotated by only a few votes. We therefore wish to emphasize the side effects of sparse annotation — in particular, what happens to the valid-set policies of previous works when they are forced to select valid labels even though the underlying model is far too unreliable. To this end, we derive information-theoretic bounds for a very simple annotation model, characterizing the minimum number of annotations required to recover the true label.

Proposition 1 (**Sparse-annotation lower bound**). *Fix $\eta \in (0, \frac{1}{2})$ and $\eta \le p_0 < p_1 \le 1-\eta$, and let $\Delta := p_1 - p_0$. Let $C \subseteq [K]$ be a consensus set, $B = [K] \setminus C$, and $m := |B| \ge 2$. For an instance $A \subseteq B$ with $|A| = a \in 1, \dots, m-1$, the valid set is $Z = C \cup A$. Suppose A is endorsed only within a stratum of mass $w \in (0, 1)$, while C is unanimously endorsed. The M annotators select labels independently: labels in C are selected with probability $p_1 - w\Delta$, labels in A with probability $p_0 + w\Delta$, and all other labels with probability p_0. Thus, all selection probabilities lie in $[\eta, 1-\eta]$. Let $\widehat{Z}$ be any estimator of Z, even one that knows C, a, w, p_0, p_1, and define $L := \max a, m - a + 1$. Then achieving worst-case exact recovery with error at most $\varepsilon \in (0, \frac{1}{2})$ requires*

$$M \ge \frac{2(1-2\varepsilon)^2 \eta(1-\eta)}{w^2\Delta^2}, \qquad M \ge \frac{\eta(1-\eta)}{w^2\Delta^2}\big[(1-\varepsilon)\log L - \log 2\big]_+. \tag{1}$$

Prop. 1 shows that exact recovery demands annotation depth of order $(w\Delta)^{-2}$, with an additional logarithmic cost for searching the residual label space. A small minority mass therefore inflates the needed depth by w^{-2} relative to a homogeneous gap Δ. Without further pooling assumptions,

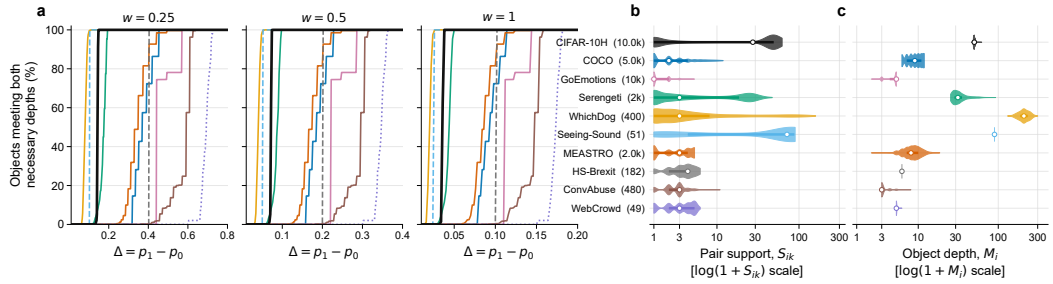

Figure 1: **Are real annotation tables deep enough to identify valid labels? a**, Fraction of objects whose depth meets both necessary conditions of Prop. 1 ($\varepsilon = 0.10$, $\eta = 0.05$), plotted against the endorsement gap $\Delta = p_1 - p_0$ at minority masses $w \in \{0.25, 0.5, 1\}$. Each object uses its candidate count K and gold valid-set size z with $C = \varnothing$. Because the bound depends on $(w\Delta)^{-2}$, the $w = 0.25$ and $w = 0.5$ panels extend Δ to 0.80 and 0.40 so they cover the same effective-gap window as $w = 1$. **b–c**, Nonzero pair support S_{ik} and object depth M_i on $\log(1 + x)$ scales; segments and points mark interquartile ranges and medians. The figure is a sensitivity analysis, not an estimate of (w, Δ), and meeting a necessary condition is not sufficient for recovery.

adding more shallowly annotated objects cannot resolve uncertainty for a particular object. We next examine whether common annotation datasets meet these necessary depth requirements.

Implications for real multi-valid data. Fig. 1 shows that depths meeting the necessary conditions under homogeneous support may fall short for minority-label recovery. At $\Delta = 0.05$, all four deeper datasets satisfy both conditions when $w = 1$, but only WhichDog does so at $w = 0.5$, and none do at $w = 0.25$. None of the six sparser datasets satisfy both conditions at this gap for any displayed minority mass. Thus, even comparatively well-annotated datasets can lack the depth required to recover the full valid-label set. These findings concern the limits of exact recovery, not the calibration of the fitted null model.

Since Prop. 1 and our dataset analysis show that Hamming loss conceals the substantial error of forced label-set commitment, we reframe label recovery from point estimation to selection.

3 Calibrated Evidence and FDR Control

To select the best candidates for true labels, we calibrate the evidence against the annotators' noisy behavior. The obstacle is that invalid object–label pairs are unknown, so we cannot directly calibrate against the support an invalid label typically receives. The pipeline therefore proceeds in three steps: *Screen.* Identify low-support pairs that are likely to be invalid. *Debias.* Restricting to low-support pairs biases their count distribution; correct this by fitting a right-truncated likelihood. *Test and select.* Fit the null model and test object–label claims on held-out objects via cross-fitting, yielding valid p-values; BH then selects the supported labels while controlling FDR. The method is thus a sequence of calibration steps — not BH applied directly to raw vote counts. In this work, we avoid conformal p-values, which presuppose exchangeable null scores, an assumption that oversimplifies worker behavior.

3.1 Null Calibration and p-Values

We begin by modeling the support count of an invalid label, giving a reference for judging whether the observed support for label k on object i is unusually high. Under H^0_{ik}, S_{ik} is a sum of rare false-positive selections, hence admits a Poisson approximation under null independence Barbour et al. (1992). This obviates per-annotator false-positive rates, which are unreliable under sparse or highly uneven annotation budgets and undefinable when annotator identities are unrecorded; we instead model the aggregate null support directly:

$$S_{ik} \mid H^0_{ik} \sim \text{Poisson}(\lambda_{ik}), \qquad \lambda_{ik} := \mathbb{E}[S_{ik} \mid H^0_{ik}],$$

To estimate λ_{ik}, we need a parametric model family that is simple yet rich. We use polynomial regression, which includes $s_{ik} := x_i^\top c_k$, a measure of object–label similarity, and M_i, the annotation depth, as features. We obtain x_i and c_k from a fixed multimodal embedding layer that maps text, images, and audio into a shared space; after ℓ_2 normalization, s_{ik} is cosine similarity. The embedding layer is not fitted using crowd annotations or gold labels. Thus,

$$\log \lambda_{ik} = \mu + \beta_k + \sum_{p=1}^{P} \gamma_p s_{ik}^p + \delta \log M_i + \delta_2 (\log M_i)^2 + \rho\, s_{ik} \log M_i. \tag{2}$$

The intensity remains object-dependent through $s_{ik} = x_i^\top c_k$ and M_i, while β_k captures label-specific background support. The remaining terms capture nonlinear effects of similarity, annotation depth, and their interaction. This model includes classical log-exposure models (Cameron & Trivedi, 2013) and simpler label-only models as special cases. We select the active terms in Equation 2 separately for each dataset using nested QAIC comparisons, a standard approach to model selection in statistics. On five of the six real tables the fitted null is conservative or near-uniform in the BH left tail; ConvAbuse is anti-conservative (Fig. 2). Fitting this model requires a set of candidate null pairs, which we construct using the screening procedure described next.

Screening candidate nulls. To fit this model, we require a pure set of null labels. We construct a null-enriched training set by selecting the candidate labels with the lowest support counts. To do that, we need the following assumption on our dataset.

Assumption (**Lower-tail separation**). For each object i:

I. **Null independence.** The null counts $\{S_{ik} : k \in \mathcal{N}_i\}$ are mutually independent.

II. **Enough invalid candidates.** At least $q_i := \lceil \alpha |\mathcal{K}_i| \rceil$ candidate labels are invalid: $|\mathcal{N}_i| \geq q_i$.

III. **Lower-tail separation.** Let $\theta_{i,\alpha} := Q_\alpha(\{\mathbb{E}[S_{i\ell}] : \ell \in \mathcal{K}_i\})$, where Q_α is the q_ith order statistic. For some $\Delta_i > 0$,

$$\mathbb{E}[S_{ik}] \geq \theta_{i,\alpha} + \Delta_i \qquad \text{for every } k \in \mathcal{D}_i.$$

Unlike standard latent crowd models, this assumption does not model dependence among valid labels. Part (III) ensures the lower tail is null-enriched without requiring every valid label to dominate every invalid label, paralleling purity conditions in large-scale testing (Efron, 2008). We next construct a finite-sample pure null screening set.

Within each training fold, we independently thin every count as $S_{ik}^{\mathrm{scr}} \mid S_{ik} \sim \mathrm{Bin}(S_{ik}, 1/2)$ and $S_{ik}^{\mathrm{fit}} = S_{ik} - S_{ik}^{\mathrm{scr}}$. Marginally, the two pieces are independent $\mathrm{Pois}(\lambda_{ik}/2)$ variables. For the likelihood argument, we additionally require each null fitting count to be independent of its object's screening vector; this holds for independent Poisson pair counts or genuinely independent annotation splits. The screening stream determines L_i, while only the fitting stream enters the screen and likelihood; App. B.1 gives the derivation.

Theorem 2 (**Purity of Quantile-LCB Screening**). *Under Assumption 3.1, suppose the null counts are mutually independent and $S_{ik} \sim \mathrm{Pois}(\lambda_{ik})$ for $k \in \mathcal{N}_i$. For $\delta \in (0,1)$, define the union-bound error for valid labels entering the population lower tail by $\varepsilon_{\mathrm{sep}} := \sum_{(i,k)\in\mathcal{H}_1} \Pr(S_{ik}^{\mathrm{fit}} \leq \theta_{i,\alpha}/2)$. Order the screening counts as $S_{i[1]}^{\mathrm{scr}} \leq \cdots \leq S_{i[|\mathcal{K}_i|]}^{\mathrm{scr}}$ and define*

$$A_i := \sum_{r=1}^{q_i} S_{i[r]}^{\mathrm{scr}}, \qquad L_i := \frac{1}{q_i} \begin{cases} 0, & A_i = 0, \\ \frac{1}{2}\chi^2_{2A_i,\delta/N}, & A_i \geq 1, \end{cases} \qquad \widehat{\mathcal{N}}_\tau := \{(i,k) \in \mathcal{H} : S_{ik}^{\mathrm{fit}} \leq L_i\},$$

where $\chi^2_{\nu,p}$ is the lower p-quantile of χ^2_ν. Then all thresholds are simultaneously conservative and the screened set contains only null pairs, with

$$\Pr\Big(L_i \leq \theta_{i,\alpha}/2 \;\forall i, \quad \widehat{\mathcal{N}}_\tau \subseteq \mathcal{H}_0\Big) \geq 1 - \delta - \varepsilon_{\mathrm{sep}}.$$

For an asymptotic sequence, $\delta = \delta_n \to 0$ and $\varepsilon_{\mathrm{sep}} = \varepsilon_{\mathrm{sep},n} \to 0$ make the purity event required by Theorem 5 hold with probability tending to one.

Large-scale multiple-testing methods commonly assume the existence of a null-dominated region containing little alternative mass (Storey, 2002). Our lower-support screen plays the same role, which motivates treating $\varepsilon_{\mathrm{sep}}$ as small when valid labels are well separated from annotation noise. Some weakly supported valid labels may nevertheless enter the screened region. Such contamination will typically increase the fitted null intensity and reduce power rather than create false discoveries.

Fitting the null model on screened pairs. Screening changes the distribution of the retained counts. In particular, because a pair is retained only when $S_{ik}^{\mathrm{fit}} \leq L_i$, fitting an ordinary Poisson

likelihood to the screened pairs would tend to underestimate the null intensity and produce anti-conservative upper-tail p-values. Following the general selection-adjustment principle of conditioning on the event that made an observation available (Lee et al., 2016), we condition the fitting count on $S_{ik}^{\text{fit}} \le \lfloor L_i \rfloor$. Under the stated cross-stream independence condition, this is the exact conditional likelihood rather than a fixed-threshold approximation. Let $F_\lambda(t) := \Pr\{\text{Pois}(\lambda) \le t\}$, and let $\boldsymbol{\theta}$ collect the active parameters in Equation equation 2. Subject to $\sum_k \beta_k = 0$, we estimate $\boldsymbol{\theta}$ by

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{(i,k)\in\widehat{\mathcal{N}}_\tau} \left[S_{ik}^{\text{fit}} \log(\lambda_{ik}/2) - \lambda_{ik}/2 - \log F_{\lambda_{ik}/2}(\lfloor L_i \rfloor)\right] - \rho\|\boldsymbol{\theta}_{-\mu}\|_2^2, \tag{3}$$

where terms constant in $\boldsymbol{\theta}$ are omitted and $\boldsymbol{\theta}_{-\mu}$ excludes the intercept. The ridge penalty $\rho \ge 0$ stabilizes estimation under sparse or unbalanced designs. Here λ_{ik} remains the mean of the unthinned count used for held-out testing. Theorem 5 in App. B.1 gives conditions under which the fitted parameters converge to their true values as the sample size grows. Independent thinning yields a centered conditional score; the theorem states the remaining rank, concentration, curvature, and scaling conditions.

Cross-fitting. Double dipping—using the same objects to screen, fit the null, and test—can make p-values anti-conservative. We therefore split by object: each fold screens and estimates $\widehat{\boldsymbol{\theta}}$ on the training objects, then scores held-out counts with the fitted intensities. Given the training fold those intensities are fixed, so the split blocks double dipping but does not itself calibrate. The p-values are conservative when $\widehat{\lambda}_{ik} \ge \lambda_{ik}$ and asymptotically valid under uniform consistency. Because fits from different folds reuse one another's test objects, we allocate the FDR budget across folds rather than assuming pooled cross-fold independence.

Constructing p-values. For each held-out pair, we compare its observed support y_{ik} with the support expected under the fitted null model. Let $\widehat{F}_{ik}$ and $\widehat{f}_{ik}$ be the CDF and PMF of $\text{Pois}(\widehat{\lambda}_{ik})$. For an independent $U_{ik} \sim \text{Unif}(0,1)$, define

$$\hat{p}_{ik} = 1 - \left[\widehat{F}_{ik}(y_{ik}-1) + U_{ik}\widehat{f}_{ik}(y_{ik})\right].$$

If the true null intensity λ_{ik} were known, replacing $\widehat{F}_{ik}$ and $\widehat{f}_{ik}$ by the true Poisson CDF and PMF would give an exactly uniform randomized p-value. With the cross-fitted estimate $\widehat{\lambda}_{ik}$, randomization removes discreteness; the fitted p-values are asymptotically calibrated under consistent null estimation (Prop. 7). The nonrandomized survival probability $1 - \widehat{F}_{ik}(y_{ik}-1)$ is a conservative alternative. Conditional on a training split, independent held-out null counts yield independent null p-values (Prop. 9).

3.2 Covariate-Adaptive FDR Control

To prioritize promising pairs and boost power, we apply weighted Benjamini-Hochberg (WBH) (Genovese et al., 2006). We map covariates $\xi_h = (x_i, c_k, \log M_i, \log\widehat{\lambda}_{ik})$ (item/label features, depth, and cross-fitted null intensity) to priority scores a_h. For held-out fold r with family $\mathcal{H}_r$ and size H_r, these scores are normalized within the fold:

$$w_h = \frac{H_r \exp(\kappa a_h)}{\sum_{h'\in\mathcal{H}_r}\exp(\kappa a_{h'})}, \qquad \sum_{h\in\mathcal{H}_r} w_h = H_r, \quad h \in \mathcal{H}_r. \tag{4}$$

Setting $\kappa = 0$ recovers ordinary BH. We apply WBH within fold r at a prespecified level q_r, using $\tilde{p}_h = \hat{p}_h / w_h$, and release the union of the foldwise rejection sets. Any allocation with $\sum_r q_r \le q$ is valid; we use $q_r = qH_r/H$. Prop. 10 proves global FDR control without cross-fold independence, provided each fold satisfies Prop. 3.

Proposition 3 (Conditional weighted-BH control for held-out annotations). *For fold r, let $\mathcal{A}$ collect the design covariates, training-fold screen, fitted null model, and resulting weights, all fixed before evaluating the held-out counts. If, conditional on $\mathcal{A}$, the null p-values are super-uniform and the foldwise family is mutually independent or PRDS on the nulls, then weighted BH at level q satisfies*

$$\text{FDR}(\widehat{\mathcal{R}} \mid \mathcal{A}) \le q, \qquad \text{FDR}(\widehat{\mathcal{R}}) \le q.$$

4 Empirical Evaluation

We organize the numerical evaluation around three questions. **(Q1)** Are Poisson regression models rich enough to capture the nuances of noisy worker behavior, particularly when the Poisson p-values are super-uniform as required by BH? **(Q2)** On synthetic tables with known ground truth,

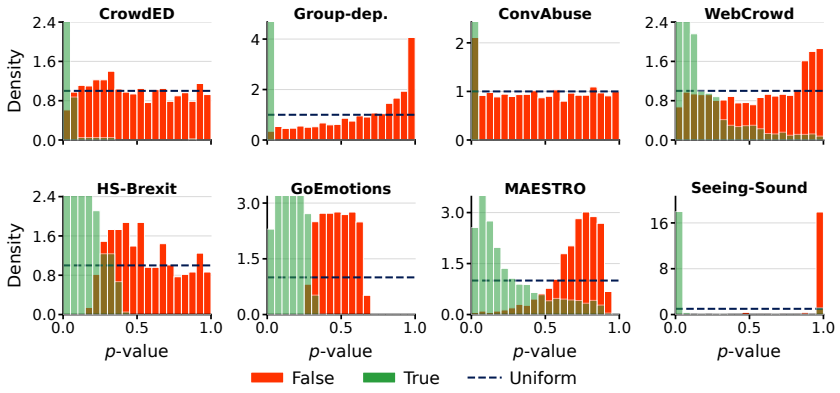

Figure 2: **BH needs super-uniform null p-values.** Red: density of false (null) pairs; green: true pairs; dashed: Unif$(0, 1)$. Super-uniformity is the CDF bound $\Pr(\hat{p} \leq t) \leq t$, so BH cares about the red *left* tail: excess red below the dashed line at small p is anti-conservative, whereas under super-uniformity the red mass piles toward the *right* of the histogram. The randomized Poisson PIT is Uniform at the correct null mean, conservative if that mean is too large, and anti-conservative if it is too small. Most panels pass the left-tail check, indicating that the model captures the nuances of the data well enough to keep null p-values super-uniform; ConvAbuse does not.

can BH and weighted BH recover multi-valid labels at level q (rather than via consensus or crowd models), and is that advantage largest under grouped dependence with sparse votes? **(Q3)** On real crowd tables, do the selectors keep gold-relative FDP near the target across different data sectors and perform competitively with other baselines?

Baselines. We compare BH and weighted BH with vote rules and latent crowd models. Multi-label majority keeps every label with support strictly above one half; single-label majority keeps only the top-supported label on each object. The crowd models are objectwise Dawid–Skene (Obj-DS) (Dawid & Skene, 1979), a GLAD-style fit (Obj-GLAD) (Whitehill et al., 2009), spectral DS (Zhang et al., 2016), and a MACE-style competence/spam model (Obj-MACE) (Hovy et al., 2013). Baselines and full protocols are given in App. D. For synthetic data, the annotation process and ground truth structure follow the CrowdED setting (Zaveri et al., 2024), described below.

Protocol. All methods are evaluated at $q \in \{0.10, 0.20\}$ on synthetic and real data. Tables for every method under the grouped power-law and CrowdED settings use $B = 1000$ replicates. On real data, gold-relative FDP and power are averaged over B=1000 perturbations retaining a random 95% of the annotated pairs. These perturbations gauge subsampling stability rather than frequentist FDR. Power is the fraction of gold-valid pairs selected, and FDP on an empty selection is reported as `n/a`; consequently, the synthetic bars labeled FDR average FDP over nonempty selections. In the synthetic experiments, where larger annotation sets allow more reliable comparison, we use $q = 0.1$. For the real-world sparse datasets, we use $q = 0.2$ to permit more exploratory discoveries.

Q1: Super-uniform p-values (Fig. 2). BH needs false-pair p-values that are super-uniform, $\Pr(\hat{p} \leq t) \leq t$. Fig. 2 plots the randomized Poisson p-value of every object–label pair, with the red density estimated from false pairs only. True pairs (green) should pile on the left. The BH diagnostic is the red left tail: it should not sit above the dashed Uniform line at small p. That check holds on both synthetic tables and on five of the six real tables; ConvAbuse is anti-conservative there. The other real tables that depart from Uniform are conservative (too few small p-values).

Q2: Selection with known synthetic truth (Figure 3). The generalized CrowdED experiment (Zaveri et al., 2024) assigns one or two uniformly drawn gold labels to each task. As the hard-task share changes from 80% to 50% to zero, BH and weighted BH have displayed FDR 0.09/0.10/0.07 and power 0.52/0.55/0.62. Per-label and spectral DS are similarly conservative; majority and most multi-valid/objectwise models gain power but exceed q. Thus, CrowdED shows that the selection method maintains error control while achieving good power (Tab. 2).

The grouped power-law stress test assigns $z_i \geq 3$ true labels across latent perspective groups and varies whether the highest-support label is false. At N=250, R=400, and Y=100, BH and weighted

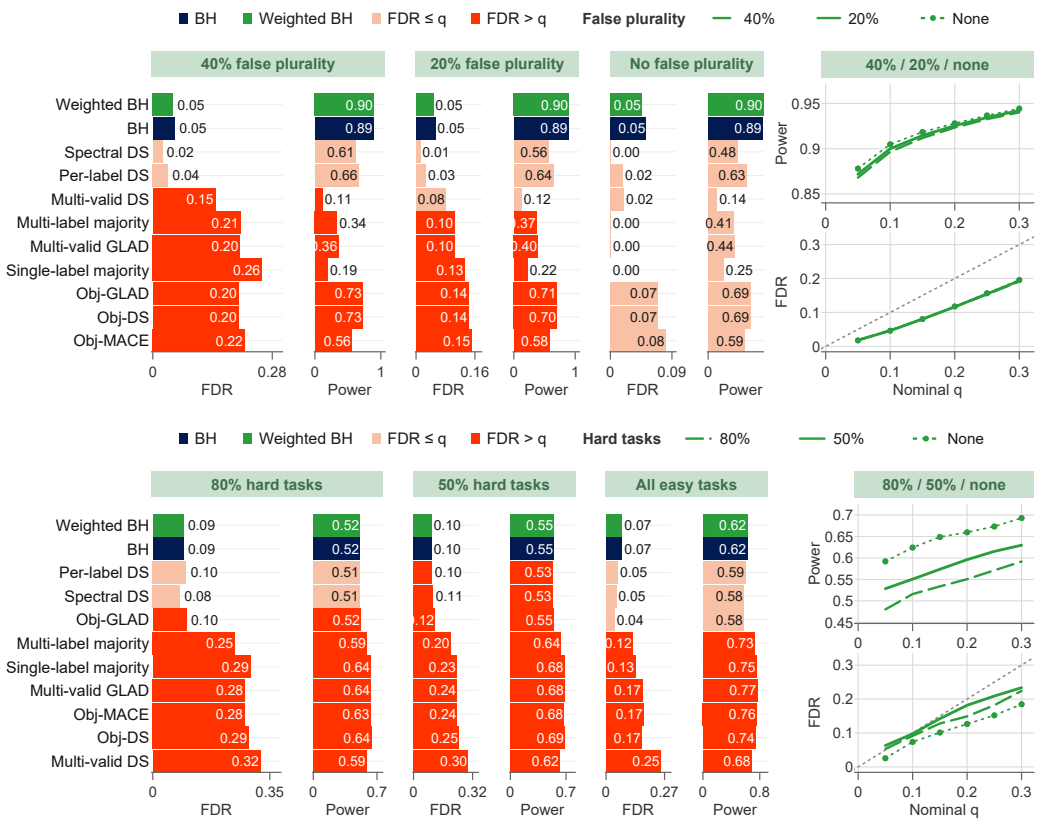

Figure 3: **Selection with known synthetic truth at q=0.10.** Each regime has an FDR and power panel; printed values are means. Navy and green identify BH and weighted BH. For other methods, pale bars meet the target and red bars exceed it. Right: weighted-BH power and FDR as the nominal q varies; the diagonal is $\mathrm{FDR} = q$. **Top (grouped power law):** 40%, 20%, or no false-plurality objects (an invalid label has the largest support). we report results averaged over B=1000 replicates for every method. **Bottom (CrowdED Zaveri et al. (2024)):** 80%, 50%, or no hard tasks, with one or two gold labels per task and B=1000 tables for every method. Hard/easy task accuracies are drawn from $U(0, 0.5)$/$U(0.5, 1)$; long-dashed, solid, and dotted curves denote the three regimes. Among methods meeting the FDR target, weighted BH has the highest power.

BH attain FDR 0.05 and power 0.89–0.90 in all three regimes using randomized Poisson p-values at the known Zipf null mean. When false plurality rises from zero to 40%, high-power Obj-DS/GLAD retain power near 0.7, but their FDR rises from 0.07 to 0.20; per-label and spectral DS remain conservative at lower power. The grouped power-law captures the strong dependencies between group votes, exposing the weakness of other methods that fail to model them, while BH still controls the error. Simulator details and q=0.20 companions are in App. D.7.

Q3: Human-only selection on real tables (Fig. 4). We use six gold-labeled crowd tables that challenge multi-valid recovery and admit fair baselines: stable worker IDs, an external (or release) gold catalogue, and multi-label truth ($\bar{Y}$>1.5 on every table). The cohort covers sound events (MAESTRO, Seeing-Sound), emotions (GoEmotions), abuse (ConvAbuse, HS-Brexit), and document relevance (WebCrowd) (Martín-Morató et al., 2021; Cartwright et al., 2017; Demszky et al., 2020; Cercas Curry et al., 2021; Akhtar et al., 2020; Goyal et al., 2018). In the Fig. 4 BH and weighted BH stay at or below q on every table. Power is high except on HS-Brexit and WebCrowd, where weighted BH has a small edge. Crowd models often match or beat that power but then exceed q, except on Seeing-Sound. Majority stays under q where gold tracks the vote and is weak on MAESTRO. Weighted BH is the most reliable method in this comparison. Because it reweights only through design covariates—depth and null intensity—its guarantees depend neither on a well-specified worker model nor on the gold label coinciding with the majority, and its realized FDR stays below the budget across the $\bar{Y}$ range. Majority and latent-crowd methods can win on tables that match their assumptions, but their FDR does not remain close to the nominal level q over the

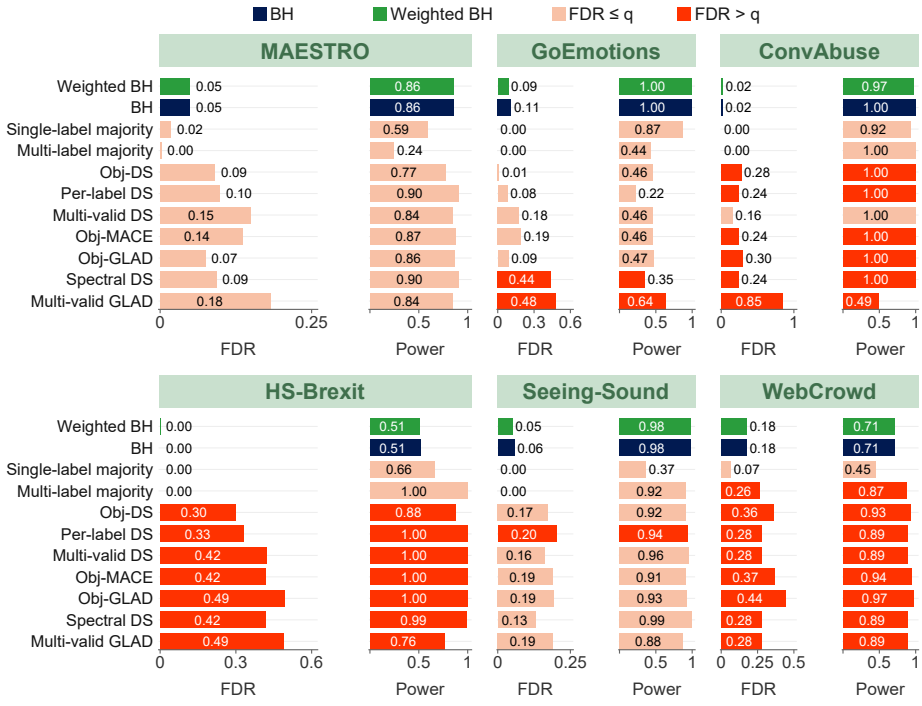

Figure 4: **Human-only selection on six crowd datasets at q=0.20.** Navy and green denote BH and weighted BH; pale and red denote mean FDP $\leq q$ and $> q$. Bars: mean gold-relative FDP and power over B=1000 overlapping 95% subsamples (perturbation stability, not frequentist FDR), using randomized Poisson p-values at the cross-fitted truncated-Poisson means $\widehat{\lambda}_{ik}$ for BH and weighted BH. Axis limits differ; compare printed values. Weighted BH holds the steadiest FDP–power balance; some baselines gain power where assumptions fit (majority rules: majority-derived ConvAbuse gold and strong crowd–gold agreement on HS-Brexit; worker models: repeated, structured sound-event judgments). These gains do not imply uniform FDP control.

full range. The same cohort at q=0.10 appears in Fig. 9. We further evaluate our method on Many-Names Silberer et al. (2020), which lacks worker identifiers, and Snapshot Serengeti Swanson et al. (2015) , whose objects carry on average only 1.06 true labels on average. As these datasets differ from the main analysis cohort, we report their results in Fig. 10, and Tab. 8.

5 Discussion

Unlike latent point-estimation methods such as Dawid–Skene and GLAD, our selection framework accommodates sparse and irregular annotation datasets, including those with aggregate counts or no stable worker identities. It operates directly on label-support counts, allows the number of valid labels to vary across objects, and abstains when evidence is insufficient. Theoretical and empirical results demonstrate its advantages under sparse annotation and its ability to maintain low false-discovery proportions across diverse datasets.

Despite relying on weaker assumptions than existing worker-based aggregation methods, our framework has important limitations. Its validity depends on accurate null calibration, a sufficiently pure screening set, and an adequate model for crowd-support counts. More fundamentally, most annotation datasets are not designed to capture the full range of perspectives that complex or subjective objects may elicit. Because annotation is costly, each object is typically evaluated by only a few workers drawn from a limited population. Valid but underrepresented perspectives may therefore receive little or no support and remain unrecoverable by any aggregation method. A central direction for future work is to improve perspective coverage through targeted human annotation or complementary sources of evidence, including carefully calibrated LLM annotations. Such augmentation must preserve the distinction between generated evidence and genuine human perspectives rather than treating LLM outputs as substitutes for missing populations.

REPRODUCIBILITY STATEMENT

To support reproduction of the empirical results and independent checking of the theory, we provide two supplements as attachments on OpenReview: the project source code, and a machine-checked Lean 4 formalization of the proofs. The code covers data preparation, null calibration, the Monte Carlo studies, and the baselines, and it reproduces the figures and tables in Section 4 and Appendix D. The experimental protocol, hyperparameters, and random seeds are stated in Appendix D; the same seeds and per-run configurations are stored with the saved artifacts. The Lean development (Lean 4.34.0) formalizes the arguments for the screening, consistency, calibration, and weighted Benjamini–Hochberg guarantees used in the paper.

ETHICS STATEMENT

Our method controls the expected fraction of invalid labels among those it selects. It does not establish fairness, cultural validity, or equal recovery across groups. Low support can reflect noise or a genuine minority view. When the two are not separated, the lower-tail screen can fold rare valid labels into the fitted null, and a global FDR guarantee can still hide poor coverage for a particular language, culture, or viewpoint. False or missed labels can suppress speech, invite unwarranted psychological inference, or change what information is shown, so the method should not be used as an unattended release rule in high-stakes settings. Real-data operating points in this paper are descriptive: they do not certify prospective control on new tasks or populations.

AI USE STATEMENT

Generative-AI tools assisted with language editing, proof checking, simulation implementation, and reviewer-style critique. The authors verified all arguments, code, outputs, and citations and take responsibility

REFERENCES

Sohail Akhtar, Valerio Basile, and Viviana Patti. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154, 2020.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Lora Aroyo and Chris Welty. Truth is a lie: CrowdTruth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015. doi: 10.1609/aimag.v36i1.2564.

K. Balasubramanian, A. Podkopaev, and S. P. Kasiviswanathan. Dependence-aware label aggregation for llm-as-a-judge via ising models. *arXiv preprint arXiv:2601.22336*, 2026.

A. D. Barbour, Lars Holst, and Svante Janson. *Poisson Approximation*. Oxford University Press, 1992.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 2 edition, 2013. doi: 10.1017/CBO9781139013567.

Emmanuel J. Candès, Andrew Ilyas, and Tijana Zrnic. Probably approximately correct labels, 2025. URL https://arxiv.org/abs/2506.10908.

Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–26, 2017. doi: 10.1145/3134664.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7388–7403. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.587.

A. 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. doi: 10.2307/2346806.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372.

Lei Duan, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara. Separate or joint? estimation of multiple labels from crowdsourced annotations. *Expert Systems with Applications*, 41(13): 5723–5732, 2014. doi: 10.1016/j.eswa.2014.03.048.

Bradley Efron. Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1): 1–22, 2008. doi: 10.1214/07-STS236.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, 2023. doi: 10.18653/v1/2023.emnlp-main.415.

Christopher Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006. doi: 10.1093/biomet/93.3.509.

Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. Your behavior signals your reliability: Modeling crowd behavioral traces to ensure quality relevance annotations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pp. 41–49, 2018.

D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130, 2013.

Jonathan Ivey, Susan Gauch, and David Jurgens. NUTMEG: Separating signal from noise in annotator disagreement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2874–2887, 2025. doi: 10.18653/v1/2025.emnlp-main.144. URL https://aclanthology.org/2025.emnlp-main.144/.

David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems 24*, pp. 1953–1961, 2011.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371.

Ariane Marandon. Conformal link prediction for false discovery rate control. *TEST*, 33:1062–1083, 2024. doi: 10.1007/s11749-024-00934-w.

Irene Martín-Morató, Manu Harju, and Annamaria Mesaros. Crowdsourcing strong labels for sound event detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 246–250, 2021. doi: 10.1109/WASPAA52581.2021.9632761.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Peterson_Human_Uncertainty_Makes_Classification_More_Robust_ICCV_2019_paper.html.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731/.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. Object naming in language and vision: A survey and a new dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5792–5801, 2020.

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002. doi: 10.1111/1467-9868.00346.

Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015. doi: 10.1038/sdata.2015.26.

Panagiotis A Traganitis and Georgios B Giannakis. Identifying dependent annotators in crowdsourcing. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 1276–1280. IEEE, 2022.

Ziquan Wang, Mingxuan Xia, Xiangyu Ren, Jiaqing Zhou, Gengyu Lyu, Tianlei Hu, and Haobo Wang. Multi-instance multi-label classification from crowdsourced labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21438–21446, 2025. doi: 10.1609/aaai.v39i20.35445.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2256–2269, 2024. doi: 10.18653/v1/2024.acl-long.123. URL https://aclanthology.org/2024.acl-long.123/.

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

Mingxuan Xia, Zenan Huang, Runze Wu, Gengyu Lyu, Junbo Zhao, Gang Chen, and Haobo Wang. Unbiased multi-label learning from crowdsourced annotations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54064–54081. PMLR, 2024.

Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False discovery rate control and statistical quality assessment of annotators in crowdsourced ranking. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1282–1291. PMLR, 2016.

Junming Yin, Jerry Luo, and Susan A. Brown. Learning from crowdsourced multi-labeling: A variational bayesian approach. *Information Systems Research*, 32(3):752–773, 2021. doi: 10.1287/isre.2021.1000.

Amrapali Zaveri, Pedro V Hernandez Serrano, and Michel Dumontier. Crowded 1.0: Statistical guideline for designing optimal crowdsourcing experiments. *Human Computation*, 11(1):29–46, 2024.

Jing Zhang and Xindong Wu. Multi-label inference for crowdsourcing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2738–2747, 2018. doi: 10.1145/3219819.3219958.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1): 3537–3580, 2016.

A NOTATION SUMMARY

Throughout the paper, we use uppercase letters for random variables and lowercase letters for their realized values; for example, Y denotes a random variable and y denotes an observed instance. Bold symbols denote vectors or sets, such as $\mathbf{Y}$ for a collection of random variables and $\mathbf{y}$ for its realization. We abbreviate index sets by $[n] := \{1, \ldots, n\}$.

Table 1: Notation for the perspective-aware multi-label annotation problem.

Notation	Description
$[N]$	Index set of objects (or instances) to be annotated.
$[M]$	Index set of annotators.
$[K]$	Index set of candidate labels in the annotation vocabulary.
x_i	Context or covariate vector associated with object $i \in [N]$.
c_k	Context, feature vector, or side information associated with label $k \in [K]$.
$\mathcal{I}_i \subseteq [M]$	Set of annotators who label object i; $M_i = \lvert\mathcal{I}_i\rvert$ is its size.
$M_i = \lvert\mathcal{I}_i\rvert$	Number of annotators who labeled object i.
$\mathcal{K}_i \subseteq [K]$	Candidate label set considered for object i.
$\mathcal{Y}_{ij}^* \subseteq [K]$	Individual perspective set of valid labels for object i under annotator j's perspective.
$y_{ik}^j \in \{0, 1\}$	Observed annotation; $y_{ik}^j = 1$ if annotator j assigns label k to object i, 0 otherwise.
$y_{ik} = \sum_{j \in \mathcal{I}_i} y_{ik}^j$	Total number of annotators in $\mathcal{I}_i$ who assign label k to object i.
$\mathbf{Y}_i = (y_{ik}^j)_{k,j}$	Complete annotation matrix (or tensor) for object i.
$\mathbf{Z}_i^*$	Global truth set for object i, obtained as the union of all individual perspective truth sets.
$Z_{ik}^* = \mathbb{1}\{k \in \mathbf{Z}_i^*\}$	Latent indicator: 1 if label k is truly associated with object i, 0 otherwise.
$H_{ik}^0 : Z_{ik}^* = 0$	Null hypothesis for object–label pair (i, k); label k not truly associated with i.
$H_{ik}^1 : Z_{ik}^* = 1$	Alternative hypothesis; label k is truly associated with i.
$\boldsymbol{\theta}$	Parameter vector for the aggregate null-intensity model.
μ	Global intercept parameter in the model.
β_k	Label-specific effect (fixed or random effect) for label k.
γ_p	Coefficient of s_{ik}^p in the similarity polynomial ($p = 1, \ldots, P$).
δ, δ_2	Coefficients of $\log M_i$ and $(\log M_i)^2$.
$\gamma_\times$	Coefficient of the interaction $s_{ik} \log M_i$.
$W \in \mathbb{R}^{n \times p}$	Design matrix for the linearized model in the consistency theorem.
p	Number of active parameters selected by QAIC (dataset-dependent).
Θ	Parameter space for $\boldsymbol{\theta}$.
$\ell_{\rho_n}(\boldsymbol{\theta})$	Penalized log-likelihood for screened estimator (see equation 3).
ρ_n	Ridge-penalty parameter for regularization.
$\widehat{\boldsymbol{\theta}}$	Estimated parameter vector (maximizing penalized likelihood).
$\widehat{\mathcal{N}}_\tau$	Screened null set (subset of object–label pairs considered null under screening rule).
$n = \lvert\widehat{\mathcal{N}}_\tau\rvert$	Number of screened object–label pairs used in estimation.
L_i	Data-dependent lower confidence threshold used to screen object i.
S_{ik}	Aggregate support count for object–label pair (i, k).
$\mathbb{1}\{\cdot\}$	Indicator function, 1 if condition is true, else 0.
$O_p(\cdot)$	Order in probability notation (big-O in probability).
$\mathbb{E}[\cdot]$	Expectation operator.
$\mathrm{Var}[\cdot]$	Variance operator.
$\preceq$	Loewner ordering for symmetric matrices (for curvature inequalities).

B ADDITIONAL BACKGROUND AND THEORETICAL RESULTS

The complete pipeline of our approach is illustrated in Figure 5.

B.1 INDEPENDENT THINNING AND CONDITIONAL CONSISTENCY

Independent thinning and the selection likelihood. For every pair in a training fold, draw

$$S_{ik}^{\mathrm{scr}} \mid S_{ik} \sim \mathrm{Bin}(S_{ik}, 1/2), \qquad S_{ik}^{\mathrm{fit}} = S_{ik} - S_{ik}^{\mathrm{scr}},$$

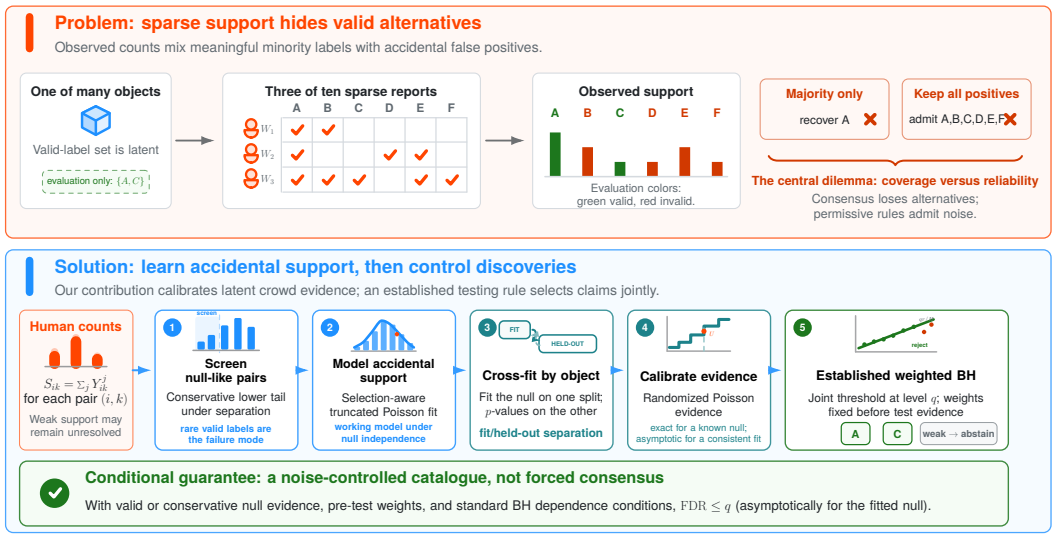

Figure 5: **From Crowd Support to FDR-Controlled Label Selection. Top:** Majority vote can discard valid minority-supported labels, whereas retaining every reported label as a soft annotation can preserve noise. **Bottom:** Our pipeline first estimates the distribution of support under the null and then performs joint label selection. A lower-tail screening step identifies null-like object–label pairs, whose counts are modeled using a truncated Poisson distribution. Object-level cross-fitting constructs held-out p-values, and foldwise weighted Benjamini–Hochberg uses allocated levels whose sum is at most q. The union controls global FDR while abstaining from labels with insufficient evidence.

independently across pairs. The Poisson splitting identity implies that, under the null model,

$$S_{ik}^{\mathrm{scr}} \perp S_{ik}^{\mathrm{fit}}, \qquad S_{ik}^{\mathrm{scr}}, S_{ik}^{\mathrm{fit}} \sim \mathrm{Pois}(\lambda_{ik}/2).$$

Let $\mathcal{G}_{i,\mathrm{scr}}$ contain object i's complete screening-count vector and its covariates. We impose the cross-stream condition

$$S_{ik}^{\mathrm{fit}} \perp \mathcal{G}_{i,\mathrm{scr}} \qquad \text{for every } k \in \mathcal{N}_i.$$

Marginal Poisson thinning guarantees independence from the pair's own screening count, but the displayed condition is additionally needed because L_i uses the other labels. It holds if the pair counts are jointly independent Poisson variables before thinning, or if the two streams come from independent annotation samples. No dependence model among the valid fitting counts is required.

Lemma 4 (Centered score after independent screening)**.** *For a retained null pair, write $m_i = \lfloor L_i \rfloor$. Conditional on $\mathcal{G}_{i,\mathrm{scr}}$ and $S_{ik}^{\mathrm{fit}} \le m_i$,*

$$\Pr_{\boldsymbol{\theta}}\big(S_{ik}^{\mathrm{fit}} = s \mid \mathcal{G}_{i,\mathrm{scr}}, S_{ik}^{\mathrm{fit}} \le m_i\big) = \frac{e^{-\lambda_{ik}/2}(\lambda_{ik}/2)^s/s!}{F_{\lambda_{ik}/2}(m_i)}, \qquad 0 \le s \le m_i.$$

Consequently, the unpenalized log-likelihood contribution in equation 3 satisfies

$$\mathbb{E}_{\boldsymbol{\theta}^*}\left[\nabla_{\boldsymbol{\theta}} \ell_{ik}(\boldsymbol{\theta}^*) \mid \mathcal{G}_{i,\mathrm{scr}}, S_{ik}^{\mathrm{fit}} \le m_i\right] = 0.$$

Proof. The binomial-thinning probability-generating function factorizes into those of two independent $\mathrm{Pois}(\lambda_{ik}/2)$ variables. Because m_i is fixed conditional on $\mathcal{G}_{i,\mathrm{scr}}$, conditioning the fitting count on its retention event gives the displayed truncated mass function. Differentiating its finite, normalized sum with respect to $\boldsymbol{\theta}$ gives the zero conditional expectation of its score. □

Thus score centering is a consequence of the construction, not an assumption about a threshold computed from the same count. The consistency result below remains conditional on screen purity and standard rank, concentration, curvature, and scaling conditions. The main-text ridge level ρ is written ρ_n when indexed by sample size.

Theorem 5 (Conditional identifiability and consistency). *Let $\widehat{\mathcal{N}}_\tau$ be the screened null set, let $n = |\widehat{\mathcal{N}}_\tau|$, and let $W \in \mathbb{R}^{n\times p}$ be the design matrix for the constrained parameter*

$$\boldsymbol{\theta} = (\mu, \{\beta_k\}, \{\gamma_p\}_{p=1}^{P}, \delta, \delta_2, \gamma_\times), \qquad \sum_k \beta_k = 0,$$

where p is the number of active parameters selected by QAIC. Let

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_{\rho_n}(\boldsymbol{\theta}),$$

where ℓ_{ρ_n} is the penalized screened log-likelihood in equation 3 with ridge penalty level ρ_n. Assume that, with probability tending to one:

1. ***Screening, independence, and rank.*** *The screened set contains only null pairs, each L_i is $\mathcal{G}_{i,\mathrm{scr}}$-measurable, $S_{ik}^{\mathrm{fit}} \perp \mathcal{G}_{i,\mathrm{scr}}$ for every screened pair, and $\mathrm{rank}(W) = p$.*
2. ***Local curvature.*** *The true parameter $\boldsymbol{\theta}^*$ lies in the interior of Θ, and for some neighborhood Θ_0 of $\boldsymbol{\theta}^*$ and constants $0 < \underline{c} \le \overline{c} < \infty$,*

$$\underline{c} I_p \preceq -\frac{1}{n}\nabla^2 \ell_{\rho_n}(\boldsymbol{\theta}) \preceq \overline{c} I_p, \qquad \forall \boldsymbol{\theta} \in \Theta_0.$$

3. ***Score concentration and parameter size.***

$$\left\| \frac{1}{n}\nabla \ell_{\rho_n}(\boldsymbol{\theta}^*) \right\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right), \qquad \|\boldsymbol{\theta}^*\|_2 = O(\sqrt{p}).$$

4. ***Asymptotic scaling.***

$$\frac{p}{n} \to 0, \qquad \sqrt{p}\,\rho_n \to 0.$$

Then, with probability tending to one, $\widehat{\boldsymbol{\theta}}$ is locally unique and

$$\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 = O_p\left(\sqrt{\frac{p}{n}} + \sqrt{p}\,\rho_n\right) = o_p(1).$$

The rate has the usual high-dimensional form: a variance term of order $\sqrt{p/n}$ and a penalty-bias term of order $\sqrt{p}\,\rho_n$. The result assumes the score and curvature conditions; severe imbalance can violate them or make their constants poor. It is therefore a modular M-estimation result, not an unconditional theorem for the full screening pipeline.

B.2 Poisson p-values

The main text uses the randomized exact Poisson upper tail. Let F_μ and f_μ be the $\mathrm{Pois}(\mu)$ CDF and PMF. For a working mean $\widehat{\lambda}_{ik}$ computed independently of the held-out pair (S_{ik}, U_{ik}),

$$\hat{p}_{ik} = 1 - \Big[F_{\widehat{\lambda}_{ik}}(S_{ik} - 1) + U_{ik}\, f_{\widehat{\lambda}_{ik}}(S_{ik})\Big],$$

with $U_{ik} \sim \mathrm{Unif}(0,1)$ independent of S_{ik}. Lemma 6 and Prop. 7 give exact uniformity at a known mean and asymptotic super-uniformity under uniformly consistent means. Prop. 9 records the within-fold independence used by cross-fitting, and Prop. 10 converts those foldwise guarantees into global FDR control without cross-fold independence. Prop. 8 quantifies the FDR slack when super-uniformity holds only approximately.

Lemma 6 (**Randomized and conservative discrete upper tails**). *Let Y be integer-valued with CDF F and PMF f, let Y' be an independent copy, and let $U \sim \mathrm{Unif}(0,1)$ be independent of Y. Define*

$$p_U(Y) := 1 - \big[F(Y-1) + U f(Y)\big] = \Pr(Y' \ge Y) - U \Pr(Y' = Y), \qquad p_+(Y) := \Pr(Y' \ge Y).$$

Then $p_U(Y) \sim \mathrm{Unif}(0,1)$, while $p_+(Y)$ is super-uniform.

Proposition 7 (Asymptotic validity under estimated parameters). *Suppose the null count model is correctly specified, so that under H_{ik}^0, $S_{ik} \sim \mathrm{Pois}(\lambda_{ik})$, and suppose the fitted means are computed from data independent of each pair's held-out (S_{ik}, U_{ik}) and satisfy*

$$\sup_{(i,k)\in\mathcal{H}_0} |\widehat{\lambda}_{ik} - \lambda_{ik}| \to_p 0.$$

For independent $U_{ik} \sim \mathrm{Unif}(0,1)$, define

$$\hat{p}_{ik} := 1 - \Big[F_{\widehat{\lambda}_{ik}}(S_{ik}-1) + U_{ik}\, f_{\widehat{\lambda}_{ik}}(S_{ik})\Big].$$

Then they are uniformly asymptotically super-uniform over the true nulls:

$$\sup_{(i,k)\in\mathcal{H}_0} \sup_{t\in[0,1]} [\mathbb{P}(\hat{p}_{ik} \le t) - t]_+ \to 0.$$

If $\widehat{\lambda}_{ik} \ge \lambda_{ik}$ almost surely, then $\hat{p}_{ik}$ is finite-sample super-uniform. These p-values are valid inputs to FDR-controlling procedures under their required dependence conditions.

Proposition 8 (Aggregate effect of approximate calibration). *Suppose the conditional weight and independence assumptions of Prop. 3 hold, but null p-values satisfy the approximate bound*

$$\Pr(\hat{p}_h \le t \mid \mathcal{A}) \le t + \varepsilon_H \quad \textit{for all } h \in \mathcal{H}_0,\ t \in [0,1].$$

Then weighted BH obeys $\mathrm{FDR} \le q + H_0\varepsilon_H$*, where $H_0 = |\mathcal{H}_0|$. Hence uniform convergence alone gives asymptotic FDR control for a growing family only when $H_0\varepsilon_H \to 0$ (or when a sharper procedure-specific bound is available).*

Proposition 9 (Cross-fitted null p-values). *For each fold r, estimate the Poisson null model on $\mathcal{D}^{(r)}_{\mathrm{train}}$ to obtain $\widehat{\lambda}^{(r)}_{ik}$, and compute $S^{(r)}_{ik}$ and $\hat{p}^{(r)}_{ik}$ only on $\mathcal{D}^{(r)}_{\mathrm{test}}$, using independent randomizers. Conditional on $\mathcal{D}^{(r)}_{\mathrm{train}}$, if Assumption 3.1 holds and $\{S^{(r)}_{ik} : H^0_{ik}\}$ are independent across object–label pairs, then $\{\hat{p}^{(r)}_{ik} : H^0_{ik}\}$ are conditionally independent. This statement is within fold r and does not establish joint independence across folds, because one fold's test data can enter another fold's fit. The foldwise budget construction below does not require such joint independence.*

Proposition 10 (Global FDR control from foldwise budgets). *Let $\mathcal{H} = \bigsqcup_{r=1}^J \mathcal{H}_r$ be the disjoint held-out families. In fold r, normalize the nonnegative weights to sum to H_r and apply weighted BH at a prespecified level q_r. If the conditional assumptions of Proposition 3 hold within each fold and $\sum_{r=1}^J q_r \le q$, then the union $\widehat{\mathcal{R}} = \bigcup_{r=1}^J \widehat{\mathcal{R}}_r$ satisfies $\mathrm{FDR}(\widehat{\mathcal{R}}) \le q$, with no assumption on dependence across folds.*

Proof. Write $R_r = |\widehat{\mathcal{R}}_r|$, let V_r be its false rejections, and set $R = \sum_r R_r$, $V = \sum_r V_r$. Pointwise,

$$\frac{V}{R \vee 1} = \sum_{r=1}^J \frac{V_r}{R \vee 1} \le \sum_{r=1}^J \frac{V_r}{R_r \vee 1}.$$

Proposition 3, applied conditionally and then averaged within each fold, gives $\mathbb{E}[V_r/(R_r \vee 1)] \le q_r$. Taking expectations and summing yields $\mathrm{FDR}(\widehat{\mathcal{R}}) \le \sum_r q_r \le q$. □

C Proofs of Theoretical Results

Proof of Proposition 1. Logs are natural; $\mathrm{TV}(P,Q) := \sup_E |P(E) - Q(E)|$, so Pinsker's inequality reads $\mathrm{TV}(P,Q) \le \sqrt{\mathrm{KL}(P\|Q)/2}$. A lower bound that holds for estimators that know C remains valid without that side information, so it suffices to recover $A \subseteq B := [K] \setminus C$. By the proposition's coordinatewise-independence assumption, conditional on (C, A) the label indicators within an annotator are independent and the annotator vectors are i.i.d.

Step 0 (observation law; curvature bound). Under (C, A), one annotator is a vector of independent Bernoulli variables with means $(\tilde{p}_k)$, and annotators are i.i.d., so for any two instances

$$\mathrm{KL}(P_{C,A}\|P_{C,A'}) = M \textstyle\sum_k \mathrm{kl}(\tilde{p}_k\|\tilde{p}'_k), \quad \mathrm{kl}(r\|q) := r\log\tfrac{r}{q} + (1-r)\log\tfrac{1-r}{1-q}. \tag{5}$$

With $g(s) := s\log s + (1-s)\log(1-s)$, one has $g''(s) = 1/[s(1-s)]$, and Taylor's formula with integral remainder gives

$$\eta \le q < r \le 1-\eta \Rightarrow \mathrm{kl}(r\|q) = \int_q^r (r-t)\tfrac{dt}{t(1-t)} \le \tfrac{(r-q)^2}{2\eta(1-\eta)}, \tag{6}$$

using $t(1-t) \ge \eta(1-\eta)$ on $[\eta, 1-\eta]$. The same integral with reversed limits yields $\mathrm{kl}(q\|r) \le (r-q)^2/[2\eta(1-\eta)]$. Note $\eta \le p_0 < p_A < p_1 \le 1-\eta$ and $p_A - p_0 = w\Delta$.

Step 1 (two-point term). Fix $A_0 \subseteq B$, $|A_0| = a$, $u \in A_0$, $v \in B \setminus A_0$, and $A_1 := (A_0 \setminus \{u\}) \cup \{v\}$. The instances differ in exactly two coordinates, each with mean gap $w\Delta$; by equation 5–equation 6, $\mathrm{KL}(P_{A_0}\|P_{A_1}) \le Mw^2\Delta^2/[\eta(1-\eta)]$, so Pinsker gives $\mathrm{TV} \le w\Delta\sqrt{M/[2\eta(1-\eta)]}$. Any $\widehat{Z}$ induces a test between these instances with average error $\ge (1-\mathrm{TV})/2$: the first term of equation.

Step 2 (Fano term). Star packing of size L: if $a \le m-a$ take $\{A_0\} \cup \{(A_0 \setminus \{u\}) \cup v : v \in B \setminus A_0\}$ ($1+m-a = L$ members); otherwise $\{A_0\} \cup \{(A_0 \setminus u) \cup v_0 : u \in A_0\}$ for fixed v_0 ($1+a = L$). All members are a-subsets at pairwise symmetric distance 2, so every pair has $\mathrm{KL} \le Mw^2\Delta^2/[\eta(1-\eta)]$. With a uniform prior, Fano with $I \le \max_{A\ne A'} \mathrm{KL}(P_A\|P_{A'})$ (convexity of KL in its second argument) yields the second term; taking positive parts is free.

Step 3 (inversion). If $M < 2(1-2\varepsilon)^2\eta(1-\eta)/(w^2\Delta^2)$ then $w\Delta\sqrt{M/[2\eta(1-\eta)]} < 1-2\varepsilon$, so the first term exceeds ε. If $(1-\varepsilon)\log L > \log 2$ and $M < \eta(1-\eta)[(1-\varepsilon)\log L - \log 2]/(w^2\Delta^2)$, the second term exceeds ε; otherwise that requirement is vacuous, matching $[\cdot]_+$. □

Proof of Theorem 2. Set $\eta = \delta/N$ and write $\ell_\eta(0) = 0$ and $\ell_\eta(s) = \frac{1}{2}\chi^2_{2s,\eta}$ for $s \ge 1$, so that $L_i = \ell_\eta(A_i)/q_i$. Let $T_i \subseteq \mathcal{N}_i$ index the q_i smallest null means and define

$$X_i := \sum_{k\in T_i} S_{ik}^{\mathrm{scr}}, \qquad \Lambda_i := \frac{1}{2}\sum_{k\in T_i}\lambda_{ik} = \frac{1}{2}\sum_{r=1}^{q_i}\lambda^0_{i,(r)} \le \frac{q_i\theta_{i,\alpha}}{2}.$$

Null independence gives $X_i \sim \mathrm{Pois}(\Lambda_i)$. Moreover, $A_i \le X_i$ almost surely, because A_i sums the q_i smallest *screening* counts, and ℓ_η is nondecreasing. The Garwood bound $\Pr\{\ell_\eta(X) > \mu\} \le \eta$ for $X \sim \mathrm{Pois}(\mu)$ therefore yields

$$\Pr(L_i > \theta_{i,\alpha}/2) \le \Pr\{\ell_\eta(X_i) > q_i\theta_{i,\alpha}/2\} \le \Pr\{\ell_\eta(X_i) > \Lambda_i\} \le \frac{\delta}{N}.$$

Hence, by a union bound,

$$\Pr\{L_i \le \theta_{i,\alpha}/2 \ \forall i\} \ge 1-\delta. \tag{7}$$

Similarly, the definition of $\varepsilon_{\mathrm{sep}}$ and a union bound over the alternatives give

$$\Pr\{S_{ik}^{\mathrm{fit}} > \theta_{i,\alpha}/2 \ \forall (i,k) \in \mathcal{H}_1\} \ge 1-\varepsilon_{\mathrm{sep}}. \tag{8}$$

On the intersection of the events in equation 7–equation 8, any $(i,k) \in \widehat{\mathcal{N}}_\tau$ satisfies $S_{ik}^{\mathrm{fit}} \le L_i \le \theta_{i,\alpha}/2$, whereas every alternative satisfies $S_{ik}^{\mathrm{fit}} > \theta_{i,\alpha}/2$. Thus $\widehat{\mathcal{N}}_\tau \subseteq \mathcal{H}_0$; a final union bound gives the claimed probability $1-\delta-\varepsilon_{\mathrm{sep}}$. □

Proof of Theorem 5. All derivatives are restricted to the constrained parameter space $\Theta_{\mathrm{c}} = \{\theta : \sum_k \beta_k = 0\}$, whose dimension is p. Define

$$Q_n(\theta) := \frac{1}{n}\ell_{\rho_n}(\theta), \qquad a_n := \sqrt{\frac{p}{n}} + \frac{\sqrt{p}\,\rho_n}{n} = o(1).$$

For a retained observation with linear predictor ζ and truncation point m, its conditional log-likelihood, up to an additive constant, is

$$s\zeta - \log\left(\sum_{j=0}^{m}\frac{e^{j\zeta}}{j!}\right).$$

The log-sum-exp term is convex in ζ, so the unpenalized criterion is concave in θ. The negative quadratic ridge penalty preserves concavity; hence Q_n is concave on the convex parameter space Θ.

Let P denote the coordinate projection that removes the unpenalized intercept. Since $\ell_{\rho_n}(\theta) = \ell_0(\theta) - \rho_n\|P\theta\|_2^2$, Assumption 3 and $\|P\theta^*\|_2 \le \|\theta^*\|_2 = O(\sqrt{p})$ give

$$\|\nabla Q_n(\theta^*)\|_2 \le \left\|\frac{1}{n}\nabla\ell_0(\theta^*)\right\|_2 + \frac{2\rho_n}{n}\|P\theta^*\|_2 = O_p(a_n).$$

Work on the event, of probability tending to one, on which the theorem's rank and curvature conditions hold. Then Q_n is $\underline{c}$-strongly concave on a neighborhood U of θ^*. Consequently, for every h with $\theta^* + h \in U$, Taylor's theorem gives

$$Q_n(\theta^* + h) - Q_n(\theta^*) \le \langle \nabla Q_n(\theta^*), h\rangle - \frac{\underline{c}}{2}\|h\|_2^2. \tag{9}$$

Because $\|\nabla Q_n(\theta^*)\|_2 = O_p(a_n)$ and $a_n \to 0$, one can choose a deterministic sequence $M_n \to \infty$ slowly enough that $M_n a_n \to 0$ and $\Pr\{\|\nabla Q_n(\theta^*)\|_2 \le \underline{c} M_n a_n/4\} \to 1$. Set $r_n = M_n a_n$. Then $B(\theta^*, r_n) \subset U$ eventually and, with probability tending to one, every h with $\|h\|_2 = r_n$ satisfies equation 9 and

$$Q_n(\theta^* + h) - Q_n(\theta^*) \le \|\nabla Q_n(\theta^*)\|_2 r_n - \frac{\underline{c}}{2}r_n^2 < 0.$$

Hence the maximizer of Q_n over the compact ball lies in its interior and is therefore a local stationary point. Because Q_n is concave on Θ, this local maximizer is global. Strong concavity in U and global concavity also rule out a second global maximizer: otherwise the segment joining the two maximizers would consist of maximizers and would contain two distinct points in U. Thus this point is the unique member of $\arg\max_{\theta\in\Theta}\ell_{\rho_n}(\theta)$ and equals the estimator $\widehat{\theta}_n$ defined in the theorem.

Finally, applying the strong-concavity gradient inequality to θ^* and $\widehat{\theta}_n$, and using $\nabla Q_n(\widehat{\theta}_n) = 0$, gives

$$\underline{c}\|\widehat{\theta}_n - \theta^*\|_2^2 \le \langle \nabla Q_n(\theta^*), \widehat{\theta}_n - \theta^*\rangle \le \|\nabla Q_n(\theta^*)\|_2\|\widehat{\theta}_n - \theta^*\|_2.$$

Thus

$$\|\widehat{\theta}_n - \theta^*\|_2 \le \underline{c}^{-1}\|\nabla Q_n(\theta^*)\|_2 = O_p(a_n) = O_p\left(\sqrt{\frac{p}{n}} + \frac{\sqrt{p}\,\rho_n}{n}\right) = o_p(1),$$

proving existence, uniqueness, and consistency. □

Proof of Lemma 6. Conditional on $Y = y$, $p_U(Y)$ is uniform on the interval $[1 - F(y), 1 - F(y-1)]$. These intervals partition $[0,1]$ up to endpoints, and their lengths equal $\Pr(Y = y)$, so the unconditional mixture is uniform. For p_+, fix $t \in [0,1]$. The case $t = 1$ is immediate. For $t < 1$, the function $p_+(y) = 1 - F(y-1)$ is nonincreasing and tends to 1 as $y \to -\infty$. If $\{y : p_+(y) \le t\}$ is empty, the claim is immediate. Otherwise, let c be its smallest element, which exists because this subset of $\mathbb{Z}$ is bounded below. Then

$$\Pr\{p_+(Y) \le t\} = \Pr(Y \ge c) = p_+(c) \le t.$$

□

Proof of Proposition 7. For a working mean $\mu \ge 0$, define

$$p_\mu(S,U) := 1 - \big[F_\mu(S-1) + U\, f_\mu(S)\big].$$

By Lemma 6, $p_\mu(S,U)$ is uniform when $S \sim \mathrm{Pois}(\mu)$. If the true law is $P_\lambda = \mathrm{Pois}(\lambda)$, then, for every $t \in [0,1]$, the total-variation bound (including the common independent randomizer) gives

$$\Pr_\lambda\{p_\mu(S,U) \le t\} - t \le d_{\mathrm{TV}}(\mathrm{Pois}(\lambda), \mathrm{Pois}(\mu)).$$

A coupling with a shared $\mathrm{Pois}(\min\{\lambda,\mu\})$ component and an independent $\mathrm{Pois}(|\lambda - \mu|)$ remainder gives

$$d_{\mathrm{TV}}(\mathrm{Pois}(\lambda), \mathrm{Pois}(\mu)) \le 1 - e^{-|\lambda-\mu|} \le \min\{1, |\lambda - \mu|\}.$$

Fix a null pair $h = (i,k)$ and condition only on its fitted mean $\widehat{\lambda}_h$. By the pairwise sample-splitting assumption, (S_h, U_h) retains law $\mathrm{Pois}(\lambda_h) \otimes \mathrm{Unif}(0,1)$ under this conditioning. Hence, for every $t \in [0,1]$,

$$\Big[\Pr(\hat{p}_h \le t \mid \widehat{\lambda}_h) - t\Big]_+ \le \min\{1, |\widehat{\lambda}_h - \lambda_h|\}.$$

This coordinatewise conditioning is important: the assumptions do not require S_h to remain Poisson after conditioning on the entire vector of fitted means. Define

$$D_n := \sup_{h \in \mathcal{H}_0} |\widehat{\lambda}_h - \lambda_h|.$$

Using $[\mathbb{E}X]_+ \le \mathbb{E}[X_+]$ and then taking the supremum gives

$$\sup_{h \in \mathcal{H}_0} \sup_{t \in [0,1]} [\Pr(\hat{p}_h \le t) - t]_+ \le \mathbb{E}[\min\{1, D_n\}].$$

Since $D_n \to_p 0$, the bounded random variable $\min\{1, D_n\}$ converges to zero in expectation, proving the asserted uniform asymptotic super-uniformity.

Finally, suppose $\widehat{\lambda}_h \ge \lambda_h$ almost surely. Conditional on $\widehat{\lambda}_h$, couple $S_h \sim \mathrm{Pois}(\lambda_h)$ and $S_h^+ \sim \mathrm{Pois}(\widehat{\lambda}_h)$ so that $S_h \le S_h^+$. Because $p_\mu(s,u)$ is nonincreasing in s, $p_{\widehat{\lambda}_h}(S_h, U_h) \ge p_{\widehat{\lambda}_h}(S_h^+, U_h)$, and the latter is exactly uniform. Thus the fitted p-value is conditionally, and therefore unconditionally, super-uniform. □

Proof of Proposition 8. Condition on $\mathcal{A}$. The standard weighted-BH leave-one-out argument used in the proof of Proposition 3 partitions on the number r of step-up rejections after setting p_h to zero. Independence of p_h from that leave-one-out event and approximate super-uniformity give, for each null h,

$$\mathbb{E}\left[\frac{\mathbb{1}\{\hat{p}_h \le w_h q R/H\}}{R \vee 1}\, middle| \, \mathcal{A}\right] \le \frac{q w_h}{H} + \varepsilon_H.$$

Summing over the H_0 nulls yields $\mathrm{FDR}(\widehat{\mathcal{R}} \mid \mathcal{A}) \le q \sum_{h \in \mathcal{H}_0} w_h/H + H_0 \varepsilon_H \le q + H_0 \varepsilon_H$. Averaging over $\mathcal{A}$ proves the claim. □

Proof of Proposition 9. Condition on the training split $\mathcal{D}^{(r)}_{\mathrm{train}}$. Then all fitted means $\widehat{\lambda}^{(r)}_{ik}$ are fixed, and each null p-value is a measurable function of $(S^{(r)}_{ik}, U^{(r)}_{ik})$. The held-out count–randomizer pairs are mutually independent under this conditional law by assumption. Applying separate measurable functions to the coordinates of a conditionally independent family preserves conditional independence, proving the within-fold claim. No assertion is made after jointly conditioning on the training data from all folds. □

Remark 1. If the null p-values are exactly or asymptotically super-uniform and satisfy independence or positive regression dependence, Benjamini–Hochberg-type procedures control the global FDR. Under arbitrary dependence, the Benjamini–Yekutieli correction can be used.

Proof of Proposition 3. Condition on $\mathcal{A}$, so the weights are fixed. Write $R = |\widehat{\mathcal{R}}|$. For a null h with $w_h > 0$, let $R^{(-h)}$ be the number of weighted-BH rejections after replacing $\hat{p}_h$ by zero. On the event that h is rejected, $R = R^{(-h)}$. Therefore, conditional super-uniformity and independence of $\hat{p}_h$ from all remaining p-values give

$$\mathbb{E}\left[\frac{\mathbb{1}\{h \in \widehat{\mathcal{R}}\}}{R \vee 1} \,\middle|\, \mathcal{A}\right] = \sum_{r=1}^{H} \frac{1}{r} \Pr\Big(\hat{p}_h \le \frac{q w_h r}{H},\ R^{(-h)} = r \,\Big|\, \mathcal{A}\Big) \le \frac{q w_h}{H}.$$

If $w_h = 0$, the corresponding contribution is zero. Summing over the true nulls yields

$$\mathrm{FDR}(\widehat{\mathcal{R}} \mid \mathcal{A}) \le \frac{q}{H} \sum_{h \in \mathcal{H}_0} w_h \le q.$$

Under conditional PRDS, the same inequality follows from the standard weighted step-up PRDS argument (Benjamini & Yekutieli, 2001; Genovese et al., 2006). Finally, averaging over $\mathcal{A}$ gives $\mathrm{FDR}(\widehat{\mathcal{R}}) \le q$. □

D EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

D.1 PROTOCOL

Section 4 evaluates object–label hypotheses in three settings.

1. **Sparse-annotation depth diagnostic.** Figure 1 reports the share of objects whose observed depth meets both necessary conditions in Proposition 1 at $\varepsilon = 0.10$ and $\eta = 0.05$, as a function of the endorsement gap Δ and minority mass $w \in \{0.25, 0.5, 1\}$. The Fano term uses each object's candidate count K and gold valid-set size z with $C = \varnothing$, so $a = z$ and $m = K$. The $w = 1$ panel uses $\Delta \in [0.01, 0.20]$; the $w = 0.25$ and $w = 0.5$ panels extend Δ to 0.80 and 0.40, matching the same effective-gap window $w\Delta \leq 0.20$. The diagnostic uses one recorded exposure per object: the dataset-specific depth M_i in the processed pair tables and, for CIFAR-10H, the sum of mutually exclusive class-vote frequencies. It does not estimate p_1, p_0, w, or Δ, and it does not classify an object as recoverable. The violin panels plot $\log(1 + S_{ik})$ and $\log(1 + M_i)$ with interquartile segments and medians.
2. **Synthetic calibration and power.** Figure 3 reports B=1000 CrowdED tables at q=0.10 for hard-task fractions 0.80, 0.50, and 0. BH and weighted BH use randomized Poisson p-values at the CrowdED guess-rate mean. Figure 3 is the grouped power-law stress study in App. D.7: BH and weighted BH use B=1000 tables and the remaining baselines use B=1000, with Poisson p-values at the known Zipf null mean.
3. **Real-data operating points.** Figure 4 evaluates one cleaned table for each of GoEmotions, HS-Brexit, MAESTRO, Seeing-Sound, ConvAbuse, and WebCrowd at q=0.20, ordered by increasing $\bar{Y}$. BH and weighted BH use randomized Poisson p-values at the cross-fitted truncated-Poisson intensities $\widehat{\lambda}_{ik}$. Gold-relative FDP and power are averaged over B=1000 draws that each keep a random 95% of annotated pairs from the same table. The draws are not independent and overlap heavily (any two share about 90% of the pairs). These results are descriptive: the available gold catalogue defines the evaluation truth. Weighted BH uses the shared priority $a_h = \log M_i$ on every table (Section D.5).

On synthetic data, power is averaged over all independent tables, whereas the plotted FDR is the mean replicate FDP among nonempty selections; an empty selection is `n/a`. Whiskers are ± 2 Monte Carlo standard errors. On real data, whiskers are the 2.5th–97.5th percentiles of those overlapping, dependent draws. They measure stability of one table, not uncertainty over new datasets, and the FDP is not an estimate of repeated-study FDR.

D.2 DATASETS

CrowdED tables. The CrowdED Monte Carlo uses `run_crowded_one_or_two.py`. Each replicate cycles through the CrowdED Table 3 worker pools and caps a source dataset at 400 tasks. A task draws one or two gold labels uniformly (two only when $K \geq 3$); each worker returns one label, choosing among the gold labels when correct and among the remaining labels otherwise. Hard/easy task accuracies draw from $U(0, 0.5)$ and $U(0.5, 1)$. BH and weighted BH use randomized Poisson p-values at the known CrowdED guess-rate mean.

Grouped power-law tables. Grouped power-law uses `run_multi_valid_crowded.py`. Each table cycles a CrowdED Table 3 worker pool with $K \geq 6$, draws a valid set of size $2 \leq |Z^\star| < K/2$, and assigns m workers per task. Invalid labels are independent Bernoulli$(1/K)$; valid labels use CrowdED's Beta–Bayes response. Hard and easy tasks draw p_{task} from $U(0, 0.5)$ and $U(0.5, 1)$. The Poisson working mean for BH and weighted BH is the known CrowdED guess rate.

Real tables. Cleaned tables under `data/clean/` cover sound-event tags (MAESTRO (Martín-Morató et al., 2021) and Seeing-Sound (Cartwright et al., 2017)), Reddit emotions (GoEmotions (Demszky et al., 2020)), conversational abuse (ConvAbuse (Cercas Curry et al., 2021)), hate-speech judgments (HS-Brexit (Akhtar et al., 2020)), and WebCrowd documents (Goyal et al., 2018). Gold labels define $\mathcal{H}_1$; remaining tested candidates are treated as nulls for descriptive FDP and power. Gold is independent of crowd majority on five of the six tables in Figure 4 (Tab. 3): official NLP emotion IDs (GoEmotions), HS-Brexit release labels, Scaper event windows (MAESTRO /

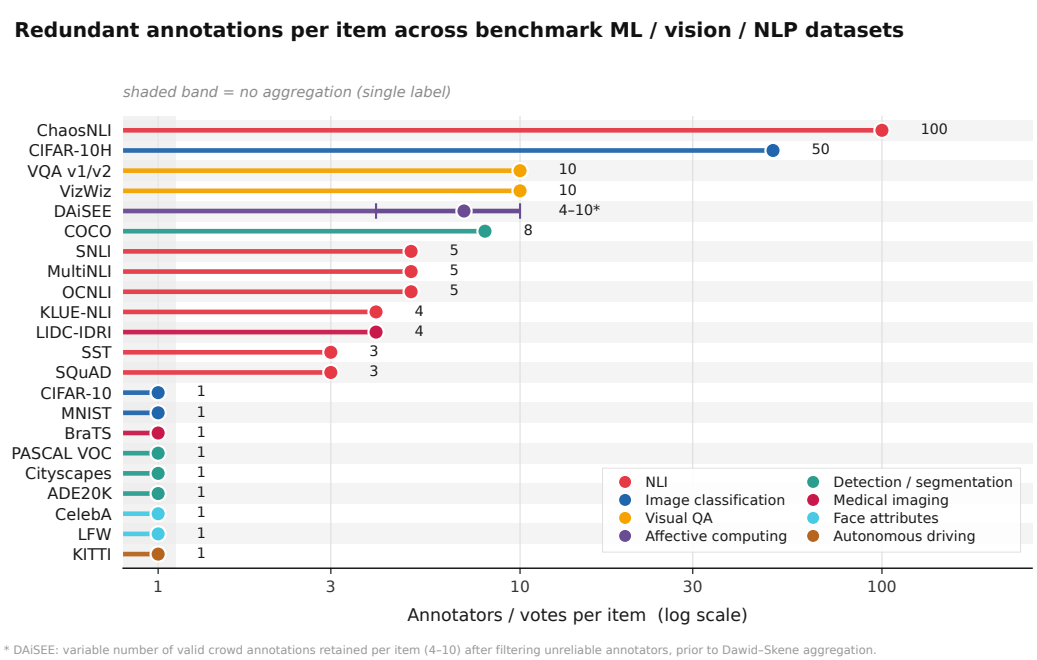

Figure 6: **Annotation redundancy across 22 benchmark datasets.** Each point is the number of independent human annotations per item, sorted and colored by domain. The x-axis is log-scaled. The shaded band at $x = 1$ marks single-annotator datasets. DAiSEE is shown as a range (4–10). Values are taken from the original dataset publications.

Seeing-Sound), and NIST relevance grades (WebCrowd). ConvAbuse uses majority consensus as a gold surrogate because the release provides no separate expert catalogue. Mean gold multiplicity is $\bar{Y} > 1.5$ on all six plotted tables. GoEmotions uses the fixed $10{,}000$-object paper cohort. WebCrowd treats a topic as the object and its independently judged relevant documents as valid labels.

All six tables retain genuine stable worker identifiers from the raw releases: `annotator_id` (MAESTRO, ConvAbuse, HS-Brexit), `rater_id` (GoEmotions), `participant_id` (Seeing-Sound), and `wid` (WebCrowd). Worker-aware methods use distinct-rater depths reconstructed from the raw reports. Fabricated row identifiers are never treated as workers.

Real-data Poisson operating mean. BH and weighted BH on the plotted real tables use randomized Poisson p-values at the cross-fitted truncated-Poisson intensities $\widehat{\lambda}_{ik}$ from equation 2–equation 3: each fold screens and fits on training objects, then scores held-out pairs with the fitted mean (Section 3.1). This is the same null used for screening, QAIC term selection, and the histograms in Figure 2, so the reported selectors exercise the calibrated pipeline rather than a plug-in rate of the form $\lambda_{ik} = \pi_0 M_i$. As a simple conservative baseline one may still take $\pi_0 \approx 1$ (or estimate an accidental rate by other means); we do not use that baseline for Figure 4. Weighted BH uses the prespecified depth score $a_h = \log M_i$ on all tables (Section D.5).

D.3 Annotation redundancy in standard benchmarks

Figure 6 records independent annotations per item for 22 widely used benchmarks. Most crowd-annotated datasets collect 3–10 votes per item. A substantial fraction of heavily cited datasets use a single annotator per item. Dawid–Skene-style methods therefore apply only where redundant labels were collected, and even then typical redundancy is far below disagreement-focused datasets such as ChaosNLI (100 votes) or CIFAR-10H (50 votes; Peterson et al., 2019).

D.4 BASELINES

Every method returns a subset of $\mathcal{H}$. Crowd models produce a posterior $\pi_{ik} = \widehat{\Pr}(Z^*_{ik} = 1 \mid \text{data})$ and are converted by MAP or Bayesian FDR. Figure 4 reports the Bayesian FDR conversion; MAP is compared in Tab. 7. Write $\ell_{ik} := 1 - \pi_{ik}$ for the model local FDR. The conversions are

$$\widehat{\mathcal{R}}^{\mathrm{MAP}} = \{(i,k) \in \mathcal{H} : \pi_{ik} \geq 1/2\}, \tag{10}$$

$$\widehat{\mathcal{R}}^{\mathrm{BFDR}} = \big\{(i,k)_{(t)} : t \leq \widehat{r}\big\}, \quad \widehat{r} = \max\Big\{r : \tfrac{1}{r}\sum_{t=1}^{r} \ell_{(t)} \leq q\Big\}, \tag{11}$$

where $\ell_{(1)} \leq \cdots \leq \ell_{(H)}$ are the ordered local FDRs. MAP is a classification rule. Bayesian FDR controls posterior expected FDP when π_{ik} is well specified (Efron, 2008); it need not control frequentist FDR under misspecification.

BH (Benjamini & Hochberg, 1995). With ordered p-values $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(H)}$,

$$\widehat{r}_{\mathrm{BH}} = \max\big\{r \in [H] : \hat{p}_{(r)} \leq qr/H\big\}, \qquad \widehat{\mathcal{R}}^{\mathrm{BH}} = \big\{h : \hat{p}_h \leq q\,\widehat{r}_{\mathrm{BH}}/H\big\},$$

or empty if no such r exists. Weighted BH uses the main-text weights equation 4 with the shared priority $a_h = \log M_i$ (annotation depth only). The same score, temperature, and cap are used on every real table; no per-dataset retuning and no gold labels enter the weight. For cross-fitted estimated nulls, this rule is applied within fold r after replacing (H, q) by (H_r, q_r), and the foldwise rejection sets are united; known-null experiments may use the single global family directly. This is not the Bioconductor IHW algorithm.

Majority (Fleisig et al., 2023). Multi-label majority keeps every label with $S_{ik}/M_i > 1/2$. Single-label majority keeps the $\arg\max$ support set on each object. Neither rule has an FDR target.

Objectwise crowd models. Obj-DS (Dawid & Skene, 1979) and Obj-GLAD (Whitehill et al., 2009) fit two-coin or difficulty-weighted binomial likelihoods separately within an object from the counts S_{ik}. They do not estimate a persistent worker confusion matrix across the corpus. Obj-MACE (Hovy et al., 2013) is a per-object binary competence/spam adaptation. Spectral DS (Zhang et al., 2016) is Opt-D&S on the worker–item matrix, treating each object–label pair as a binary task; it requires persistent annotator IDs, which all six selected cohorts provide. On synthetic tables we also report per-label DS and powerset multi-valid DS/GLAD when 2^{K_i} is enumerable; otherwise those columns are omitted.

D.5 HYPERPARAMETERS

On all six real tables, weighted BH uses the same priority $a_h = \log M_i$ in equation 4, with temperature κ=0.5 and weight cap 10. Depth M_i is design-fixed, so the weights do not use gold labels or held-out counts. BH and weighted BH use the cross-fitted truncated-Poisson means $\widehat{\lambda}_{ik}$ described above. The principal panel does not retune q. QAIC-selected Poisson GLM terms for the six tables are $\mu + \beta_k + s + s^2 + \log M + (\log M)^2$ (MAESTRO, GoEmotions, ConvAbuse), $\mu + \beta_k + s + s^2 + \log M + s \log M$ (HS-Brexit), $\mu + \beta_k + s + s^2 + \log M$ (Seeing-Sound), and $\mu + s + s^2 + \log M + s \log M$ (WebCrowd). Gold-null PIT histograms for these fitted means appear in Figure 2 and Tab. 4.

D.6 IMPLEMENTATION

Data preparation, calibration, Monte Carlo experiments, and baselines are implemented in Python 3.12 (NumPy, SciPy, pandas, scikit-learn). The B=1000 studies parallelize over replicates on CPU. Seeds and per-run configurations accompany the saved artifacts under `outputs/`; figures are stored under `figs/`.

D.7 GROUPED-PERSPECTIVE POWER-LAW DESIGN

This generator (`synthetic_fdr.beta_binomial`, `Python/run_trap_hardness.py`) is the grouped-dependence setting in Q2. It keeps CrowdED's Beta–Binomial idea (Zaveri et al., 2024)

but recasts it for multi-valid object–label hypotheses with an ICC mixture on true labels. Unless noted, the design uses $N{=}R{=}1000$, $Y{=}1000$, $\alpha{=}2$, and $\mathrm{ICC}{=}0.15$. False-label responses are conditionally independent Bernoulli. True-label responses are exchangeable within latent perspective groups. Marginal alternative means are held at calibrated targets, so dependence and signal strength are decoupled.

Discrete truncated power law. For integers on $\{L, \dots, H\}$,

$$P(X = k) = \frac{k^{-\alpha}}{\sum_{s=L}^{H} s^{-\alpha}}, \qquad k \in \{L, \dots, H\}. \tag{12}$$

Object-level geometry. For each object i,

$$K_i \sim \mathrm{PowerLaw}\big(\lfloor Y/100 \rfloor,\, Y;\, \alpha\big), \tag{13}$$

$$z_i \sim \mathrm{PowerLaw}\big(3,\, \lfloor K_i/3 \rfloor;\, \alpha\big) \quad \text{(resample } K_i \text{ if } \lfloor K_i/3 \rfloor < 3\text{)}, \tag{14}$$

$$G_i \sim \mathrm{PowerLaw}(3,\, 10;\, \alpha), \tag{15}$$

$$M_i \sim \mathrm{PowerLaw}\big(\lfloor R/10 \rfloor,\, R;\, \alpha\big). \tag{16}$$

A candidate set $\mathcal{K}_i \subset [Y]$ of size K_i is drawn without replacement. Workers are partitioned into groups with sizes $(m_{i1}, \dots, m_{iG_i}) \sim \mathrm{Multinomial}(M_i, \pi_i)$ for $\pi_i \sim \mathrm{Dirichlet}(4 \cdot \mathbf{1}_{G_i})$. The truth set $\mathbf{Z}_i^*$ of size z_i is drawn from $\mathcal{K}_i$. Two constraints are then enforced by permuting ranks only: the bottom frequency quartile contains only null labels, and a prescribed fraction of objects (0, 20%, or 40% in Figure 3) have a false most-frequent label.

Null intensity and worker responses. Null means follow a floored Zipf law on popularity rank,

$$\lambda_{ik}^0 = \min\big(b_i + A_i\, r_{ik}^{-\alpha},\; 0.70\, M_i\big), \tag{17}$$

with $b_i \sim \mathrm{Unif}(0.09, 0.15)\, M_i$ and $A_i \sim \mathrm{Unif}(0.35, 0.65)\, M_i$. Under H_{ik}^0,

$$Y_{ij}^k \mid H_{ik}^0 \overset{\text{ind}}{\sim} \mathrm{Bernoulli}(\lambda_{ik}^0/M_i), \qquad S_{ik} \mid H_{ik}^0 \sim \mathrm{Binomial}(M_i, \lambda_{ik}^0/M_i). \tag{18}$$

BH and weighted BH use randomized Poisson p-values at this Zipf mean, which isolates FDR comparisons from null-mean estimation error.

Calibrated alternatives and within-group dependence. Each group g is assigned a recognized subset $\mathcal{P}_{ig} \subseteq \mathbf{Z}_i^*$. Each true label is given to one uniformly chosen group and, independently with probability 0.30, to a second group; empty groups are filled so that $\bigcup_g \mathcal{P}_{ig} = \mathbf{Z}_i^*$. A multiplicative boost $\delta_{ik} \sim \mathrm{Unif}(0.08, 0.32)$ in the design, or $\delta \in [1.2, 2.4]$ in the high-δ Monte Carlo of Figure 3, defines a target alternative mean μ_{ik}^1. An intercept is solved so that size-weighted group rates match this target exactly. Within a group and true label, counts follow a mean–concentration Beta–Binomial mixture at $\mathrm{ICC}_{ig} = 0.15$:

$$\theta_{igk} \sim \mathrm{Beta}\big(q_{igk}\kappa_{ig},\, (1 - q_{igk})\kappa_{ig}\big), \qquad S_{igk} \mid \theta_{igk} \sim \mathrm{Binomial}(m_{ig}, \theta_{igk}), \tag{19}$$

with $\kappa_{ig} = 1/\mathrm{ICC}_{ig} - 1$. The mixture preserves the calibrated mean and inflates the variance by the usual Beta–Binomial factor. $\mathrm{ICC} = 0$ recovers independent Bernoulli alternatives.

E Additional results supporting the main experiments

E.1 CrowdED Monte Carlo at $q = 0.10$

Tab. 2 records the plotted means for the bottom panel of Figure 3. BH and weighted BH use randomized Poisson p-values; all methods use the same $B{=}1000$ one-or-two-truth tables.

E.2 Companion synthetic analysis at $q = 0.20$

Figure 7 repeats the CrowdED comparison at $q{=}0.20$ on $B{=}1000$ one-or-two-truth tables. Across the 80%, 50%, and all-easy regimes, BH and weighted BH have displayed FDR 0.15, 0.18, and 0.13 and power 0.55, 0.60, and 0.66.

Figure 8 is the grouped power-law companion at $q{=}0.20$. BH and weighted BH and the other uses $B{=}1000$ tables.

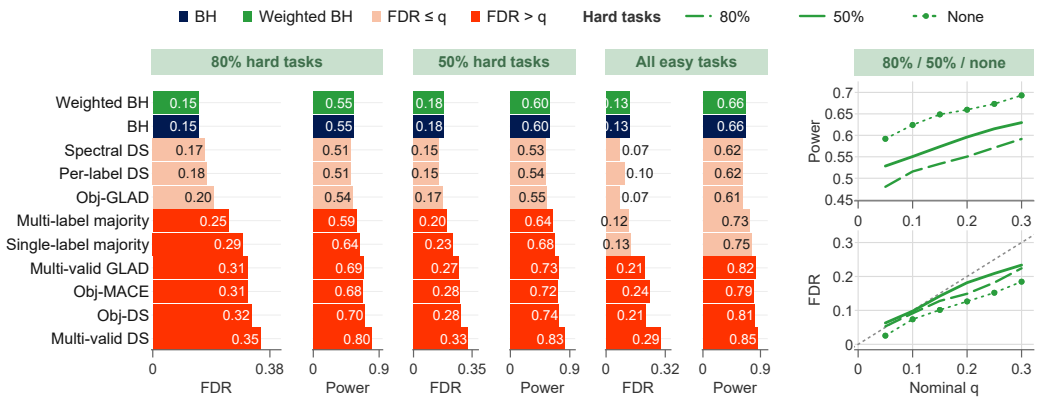

Figure 7: **CrowdED one-or-two-truth selection at** q=0.20**.** Difficulty regimes and methods match Figure 3; every bar averages the same B=1000 tables. BH and weighted BH use randomized Poisson p-values.

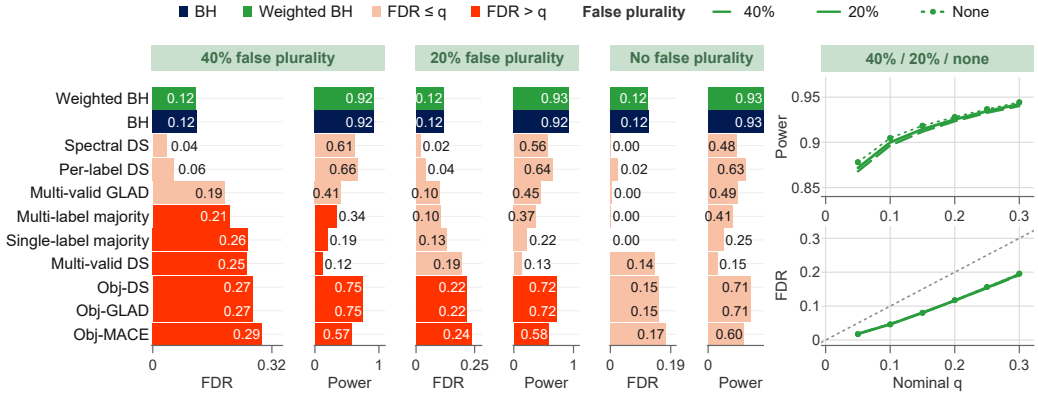

Figure 8: **Grouped power-law selection at** q=0.20**.** False-plurality regimes match Figure 3. BH and weighted BH (navy/green; B=1000) have FDR 0.12 and power 0.92–0.93; other bars use B=1000. High-power objectwise models exceed q under 20% or 40% false plurality, whereas conservative baselines trade power for lower FDR.

Table 2: CrowdED one-or-two-truth Monte Carlo at q=0.10 (B=1000). Entries reproduce the figure's means: power averages all tables; FDR averages replicate FDP over nonempty selections (empty is `n/a`).

	80% hard		50% hard		all easy	
Method	FDR	Power	FDR	Power	FDR	Power
BH	0.09	0.52	0.10	0.55	0.07	0.62
Weighted BH	0.09	0.52	0.10	0.55	0.07	0.62
Per-label DS	0.10	0.51	0.10	0.53	0.06	0.59
Spectral DS	0.08	0.51	0.11	0.53	0.05	0.58
Obj-GLAD	0.10	0.52	0.12	0.55	0.04	0.58
Multi-label majority	0.25	0.59	0.20	0.64	0.12	0.73
Single-label majority	0.29	0.64	0.23	0.68	0.13	0.75
Multi-valid GLAD	0.28	0.64	0.24	0.69	0.17	0.77
Obj-MACE	0.28	0.63	0.24	0.68	0.17	0.76
Obj-DS	0.29	0.64	0.25	0.69	0.17	0.74
Multi-valid DS	0.32	0.59	0.30	0.62	0.25	0.69

E.3 Real-data census, calibration, and $q = 0.20$ numbers

Table 3: **Analysis tables used in Figure 4.** N: objects; K: candidate labels; pairs: object–label hypotheses; true / null: gold membership; $\bar{Y}$: mean gold-true labels per object; $\bar{M}$: mean annotators per object. BH and weighted BH use cross-fitted truncated-Poisson means $\widehat{\lambda}_{ik}$ from equation 2. Weighted BH uses the shared priority $a_h = \log M_i$ (κ=0.5, cap 10) on every table.

Dataset	N	K	Pairs	True	Null	$\bar{Y}$	$\bar{M}$
GoEmotions	10000	28	46532	15193	31339	1.52	4.5
HS-Brexit	201	4	804	388	416	1.93	6.0
MAESTRO	3420	6	9903	6953	2950	2.03	8.3
Seeing-Sound	60	9	294	189	105	3.15	90.0
ConvAbuse	480	13	6240	1663	4577	3.46	3.3
WebCrowd	50	4996	4996	2267	2729	45.34	5.0

E.4 Companion real-data analysis at $q = 0.10$

Figure 9 repeats Figure 4 at q=0.10 on the same six tables. BH and weighted BH recompute randomized Poisson p-values at the cross-fitted truncated-Poisson means $\widehat{\lambda}_{ik}$ on each of the B=1000 overlapping 95% subsamples. Crowd-model posteriors are fitted on the full table and thresholded at q=0.10. BH is empty on HS-Brexit and WebCrowd; weighted BH is empty on WebCrowd. Those empty selections are plotted as `n/a`.

At N=5000, the screen is 0.99997 pure, the truncated fit has mean $\widehat{\lambda}/\lambda = 1.0023$ and relative MAE 0.0436, and every optimizer converges. At N=1000, relative MAE is 0.0872 and FDR is modestly above 0.10. The experiment supports the correctly specified asymptotic mechanism. It

Table 4: **Null calibration on the six plotted tables.** n_0: gold-null pairs. Mean: QAIC GLM family used in Figure 2. BH FDP is the Monte Carlo mean from Tab. 5. Histogram calibration is in Figure 2.

Dataset	n_0	Mean	BH FDP
GoEmotions	31339	$\mu + \beta_k + s + s^2 + \log M + (\log M)^2$	0.195
HS-Brexit	416	$\mu + \beta_k + s + s^2 + \log M + s \log M$	0.000
MAESTRO	2950	$\mu + \beta_k + s + s^2 + \log M + (\log M)^2$	0.046
Seeing-Sound	105	$\mu + \beta_k + s + s^2 + \log M$	0.081
ConvAbuse	4577	$\mu + \beta_k + s + s^2 + \log M + (\log M)^2$	0.012
WebCrowd	2729	$\mu + s + s^2 + \log M + s \log M$	0.122

Table 5: **Real-data FDP / power at** q=0.20 (Monte Carlo means from Figure 4 / `outputs/real_fdr_02.csv`, B=1000 overlapping 95% subsamples, cross-fitted truncated-Poisson p-values for BH and WBH). Bold FDP satisfies FDP $\leq q$. Maj. = multi-label majority; S-maj. = single-label majority; Obj-DS / Obj-GLAD / Obj-MACE = per-object adaptations; Spec. = Spectral DS.

Dataset	BH		WBH		Maj.		S-maj.		Obj-DS		Obj-GLAD		Obj-MACE		Spec.	
	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.
GoEmotions	**0.19**	1.00	**0.16**	1.00	**0.00**	0.44	**0.00**	0.87	**0.01**	0.46	**0.09**	0.47	**0.19**	0.46	0.44	0.35
HS-Brexit	**0.00**	0.59	**0.00**	0.67	**0.00**	1.00	**0.00**	0.66	0.30	0.88	0.49	1.00	0.42	1.00	0.42	0.99
MAESTRO	**0.05**	0.85	**0.05**	0.85	**0.00**	0.24	**0.02**	0.59	**0.09**	0.77	**0.07**	0.86	**0.14**	0.87	**0.09**	0.90
Seeing-Sound	**0.08**	0.99	**0.06**	0.99	**0.00**	0.92	**0.00**	0.37	**0.17**	0.92	**0.19**	0.93	**0.19**	0.91	**0.13**	0.99
ConvAbuse	**0.01**	1.00	**0.01**	1.00	**0.00**	1.00	**0.00**	0.92	0.28	1.00	0.30	1.00	0.24	1.00	0.24	1.00
WebCrowd	**0.12**	0.57	**0.13**	0.58	0.26	0.87	**0.07**	0.45	0.36	0.93	0.44	0.97	0.37	0.94	0.28	0.89

Table 6: **Tables with** FDP $\leq q$ **at** q=0.20**.** Descriptive counts on the six selected tables, not repeated-sampling control guarantees.

Method	Datasets scored	Meets q	Rate
BH	6	6/6	1.00
WBH	6	6/6	1.00
S-maj.	6	6/6	1.00
Maj.	6	5/6	0.83
Obj-DS	6	3/6	0.50
Obj-GLAD	6	3/6	0.50
Obj-MACE	6	3/6	0.50
Spec.	6	2/6	0.33

Table 7: **Bayesian-FDR vs MAP** for worker aggregators at q=0.20 (Monte Carlo means from `outputs/real_fdr_02.csv`). MAP is omitted from Figure 4; neither conversion carries a frequentist FDR guarantee under misspecification. Bold FDP satisfies FDP $\leq q$.

Method	Rule	MAESTRO		GoEmotions		ConvAbuse		HS-Brexit		Seeing-Sound		WebCrowd	
		FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.
Obj-DS	FDR	**0.09**	0.77	**0.01**	0.46	0.28	1.00	0.30	0.88	**0.17**	0.92	0.36	0.93
	MAP	**0.05**	0.69	**0.02**	0.71	**0.08**	1.00	0.25	0.82	**0.00**	0.89	0.27	0.87
Obj-GLAD	FDR	**0.07**	0.86	**0.09**	0.47	0.30	1.00	0.49	1.00	**0.19**	0.93	0.44	0.97
	MAP	**0.03**	0.74	**0.08**	0.63	**0.13**	1.00	0.36	1.00	**0.00**	0.92	0.35	0.92
Obj-MACE	FDR	**0.14**	0.87	**0.19**	0.46	0.24	1.00	0.42	1.00	**0.19**	0.91	0.37	0.94
	MAP	**0.02**	0.79	**0.01**	0.44	**0.06**	1.00	0.27	1.00	**0.00**	0.90	0.28	0.88
Spec.	FDR	**0.09**	0.90	0.44	0.35	0.24	1.00	0.42	0.99	**0.13**	0.99	0.28	0.89
	MAP	**0.01**	0.79	0.57	0.64	**0.06**	1.00	0.33	0.95	**0.00**	0.92	0.21	0.79

Table 8: **ManyNames and Serengeti FDP / power at** q=0.20 (Monte Carlo means from Figure 10, B=1000 overlapping 95% subsamples). Bold FDP satisfies FDP $\leq q$. Maj. = multi-label majority; S-maj. = single-label majority.

Dataset	BH		WBH		Maj.		S-maj.	
	FDP	Pow.	FDP	Pow.	FDP	Pow.	FDP	Pow.
ManyNames	**0.06**	0.47	**0.07**	0.47	**0.00**	0.22	**0.01**	0.30
Serengeti	**0.20**	0.99	**0.19**	0.99	**0.01**	0.78	**0.03**	0.88



Figure 9: **Human-only selection on six crowd tables at q=0.10.** Each panel is one observed table. FDP is gold-relative, averaged over B=1000 dependent 95% pair-retention draws from the same table (any two share about 90% of the pairs), not frequentist FDR. Navy and green denote BH and weighted BH; pale and red denote mean FDP $\leq q$ and $> q$. BH and weighted BH use cross-fitted truncated-Poisson p-values. Layout and methods match Figure 4.

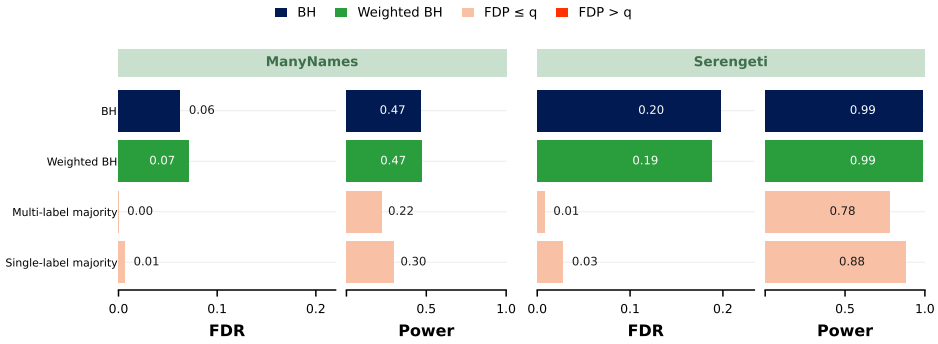

Figure 10: **ManyNames and Serengeti at q=0.20, aggregate methods only.** Navy and green denote BH and weighted BH; pale and red denote mean FDP $\leq q$ and $> q$. Bars: mean gold-relative FDP and power over B=1000 overlapping 95% subsamples (perturbation stability, not frequentist FDR). Worker models are omitted because these tables are not worker-resolved in the analysis files. Layout matches Figure 4.

Table 9: **End-to-end Poisson simulation.** Monte Carlo means over B=1000 independent tables (seed 20260912). "Truncated" is the proposed cross-fitted working likelihood; "naive" ignores selection truncation.

			$q = 0.10$		$q = 0.20$	
N	H	Method	FDR	Power	FDR	Power
1000	20000	Oracle	0.088	0.924	0.176	0.952
		Truncated	0.107	0.917	0.200	0.947
		Naive	0.852	1.000	0.865	1.000
5000	100000	Oracle	0.088	0.924	0.176	0.952
		Truncated	0.093	0.923	0.183	0.951
		Naive	0.856	1.000	0.867	1.000

does not supply a finite-sample or misspecification guarantee, and it is not a substitute for the real-data operating points in Figure 4.